The Cosm Sensor Data Set

Data Integration of a Sensor Commons

Martin Sebastian Dittus

Centre for Advanced Spatial Analysis, UCL

MRes Advanced Spatial Analysis & Visualisation

September 9, 2012

I, Martin Dittus confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This dissertation presents the first assessment of a large volume of volunteer sensor data published on Cosm, an online platform for sensor data enthusiasts. This is highly heterogenous, geo-referenced, and tagged time series data, encompassing a large variety of observed phenomena, and its existence reveals a new potential to capture many of the earths's urban and environmental systems at unprecedented levels of detail.

Since this is such a novel data set the assessment is approached in a broad manner. A thorough literature review establishes the general nature of such informal data-gathering activities, including a discussion of the aims and motivations of participants. It also presents some historic context, and argues for the potentially significant role of such activities in society.

A second quantitative component evaluates the quality of the volunteer sensor data with a focus on data heterogeneity. Based on a number of descriptive statistics and data visualisations it is determined to which degree Cosm community sensor data can already be used to build integrated spatiotemporal models of environmental and other phenomena, and methods are developed that allow the identification of distinct groups of related sensors despite the presence of very heterogenous metadata.

A number of key sensor groups identified with these methods are then assessed in some detail. A case study comparing Cosm temperature data with an equivalent high-quality ground truth data set reveals that the volunteer annotations are not sufficiently detailed to support the selection of larger numbers of outdoor temperature sensors in the region of interest.

However it is shown that there is a marked increase in informal sensor data-gathering activities, particularly in the context of contemporary themes such as energy usage, environmental concerns, and climate change. The dissertation concludes with recommendations for better volunteer sensing tools and practices, and suggestions for future research.

Contents

Chapter 1

Introduction

In recent years an increasing number of electronic sensors has begun capturing many of the planet's urban and environmental systems, often in realtime, and with the aim of making this data public. With keen foresight Gross (1999) vividly described some of the many forms this may take:

> *In the next century, planet earth will don an electronic skin. It will use the Internet as a scaffold to support and transmit its sensations. This skin is already being stitched together. It consists of millions of embedded electronic measuring devices: thermostats, pressure gauges, pollution detectors, cameras, microphones, glucose sensors, EKGs, electroencephalographs. These will probe and monitor cities and endangered species, the atmosphere, our ships, highways and fleets of trucks, our conversations, our bodies–even our dreams.*

As this dissertation demonstrates, and as Gross already suggested in 1999, such an extensive monitoring of the planet and its subsystems may not only be performed by well-funded governmental or commercial efforts, but may also be the result of informal activities by private citizens. A driving force behind such informal efforts are newly affordable technologies of information-gathering and communication, including Geographic Information Systems (GIS) that allow to assess and present such data.

This dissertation presents a discussion of the nature and value of informal sensor data-gathering activities, and an evaluation of some of the early outcomes. We describe and analyse a large body of volunteer sensor data published on the community platform Cosm. This is a complex multi-dimensional time series data set that is geo-referenced and tagged by its contributors.

To our knowledge this is the first time this data has been subject to scientific inquiry, and consequently we are presented with a wide range of fundamental questions. What activities take place? How rigorous are the data-gathering practices of predominantly untrained participants? Can the outcomes already be used to build integrated spatiotemporal models of large-scale environmental systems, particularly since the data is predominantly gathered for personal use?

1.1 The Emergence of a Sensor Commons?

Kera and Graham (2010) first introduced the term “Collective Sensor Networks” (CSN) to describe sensing activities involving heterogenous groups of participants that have some agency:

> *These kinds of projects have involved low cost monitoring and sharing of various data to create new types of communities of enthusiasts and citizens that self-organize around various biological and sensor data. …The goal is to create new types of local and global awareness and support various identities and communities in the contexts of our neighborhoods, cities, countries or even the whole planet and the biological habitat.*

However they also point out that this introduces new challenges, among them the problem of integrating sensor data from a large number of heterogenous sources. Fisher (2011) describes a similar vision of ubiquitous sensing infrastructure which he calls “Sensor Commons”:

> *For me the Sensor Commons is a future state whereby we have data available to us, in real time, from a multitude of sensors that are relatively similar in design and method of data acquisition and that data is freely available whether as a data set or by API to use in whatever fashion they like.*

These are two well-articulated visions of the nature of communal sensor infrastructure, and they illustrate some of its potential, but neither offer an analysis of the expected efficacy of such shared monitoring systems. If there is to be such public sensing infrastructure, and such public access to many different sources of sensor data, how can one integrate these disparate sources to produce a coherent signal?

Compared to the popular map-making efforts of Wikimapia or Google Earth there is a higher barrier of entry for the contribution to public sensor data catalogues. Sensor technology is not widespread and often expensive, environmental data is hard to capture and subject to many complex considerations, and sensors need to be monitored and maintained on an ongoing basis. As a result there currently are only few community efforts, and data volumes are comparably low.

However as this dissertation will demonstrate there is increased public interest, particularly in the context of contemporary themes such as energy usage, environmental concerns, and climate change, and there is an increased ability to produce or acquire affordable do-it-yourself (DIY) sensor hardware. As a result there is a marked increase in communal sensor data-gathering activities.

Such efforts often cannot be direct substitutes for the products of existing institutions and their established formats, and cannot be compared with them on the same terms;

there is a stark difference in approach and ability. Instead personal and communal sensing projects can introduce a new potential for understanding and describing the world. A key attribute of such efforts often is that they can draw from an unprecedented number of contributors, and that they integrate the multitude of perspectives that comes with this, supported by modern telecommunications technology which introduces new capabilities for large-scale collaboration. And in fact many of such efforts are only possible and sustainable with the participation of a large number of volunteers.

1.2 The Cosm Sensor Data Platform

The London-based company Cosm, formerly Pachube, was founded in 2008 to provide a public platform and meeting ground for DIY sensor data enthusiasts. One of its key products is a general-purpose sensor data store that aggregates and visualises the data published by Cosm community participants (Cosm, 2012a).

The company provides an Application Programming Interface (API) that allows participants to record and publish sensor data from any Internet-connected device. Once published, sensor data can be browsed and visualised in realtime on the Cosm website, shared, and exported for further analysis. In late 2011 the service was made free to use (Reidy, 2011).

A characteristic property of the service is that it does not propose a singular aim behind sensing efforts, and instead provides general infrastructure and services. The purpose of the platform is determined by its participants, and the history of posts on the company blog reveals a wide range of sensing practices by individuals, institutions and corporations. Activities include the monitoring of temperature, humidity, home energy usage, radiation levels, indoor and outdoor air quality levels, and many other phenomena. (Cosm, 2012c)

It is not the expressed company aim to specifically support the purposeful aggregation of highly consistent sensor data in order to capture large environmental phenomena. However as will be shown in Section 3.1 Cosm as a data hub has a number of properties that may make it suitable for such large-scale collaboration, it already stores many millions of sensor measurements from sensors placed across the globe, and there are indicators that coordinated activities are already taking place.

In addition it is feasible that its large volume of sensor data may already be sufficiently homogeneous to support large-scale environmental modelling efforts even in the absence of explicit collaboration. For this reason the public Cosm sensor data archive can serve as a testing ground to address some of the aforementioned questions: is Cosm an early approximation of a Sensor Commons?

1.3 Research Aims

Despite a marked increase in DIY sensing activity there has been very little formal evaluation of the nature and quality of data produced under the informal circumstances alluded to by Kera et al. Generally the focus of scientific research in this domain is on sensor infrastructure that was produced for very particular aims.

To our knowledge there is no research on the nature of public sensor data repositories. Such data stores are interesting study objects: they are the public record for a large spectrum of activities, and within some limitations their existence allow us to assess the nature and sophistication of certain types of DIY sensing activities.

This work presents such a study based on a subset of the Cosm sensor data archive. At this early stage it is hard to separate mere short-term experiments from practices that will demonstrate long-term merit, but using the published data it is possible to highlight a selection of current sensing practices and their outcomes.

A core aspect of this research is an evaluation of the resulting sensor data. What statements can be made about the quality of sensor data found on Cosm? How good is Cosm's collective spatial coverage for particular types of data? How heterogeneous is the data in its measurements and annotations?

Most Cosm community participants likely do not expect that their data may be useful to others beyond its initial, personal purpose, and as a result annotations are likely sparse and inconsistent. Is it still possible to identify and integrate related sensor feeds in order to capture larger environmental or social phenomena? For example can sensor data published on Cosm be used to enhance or replace other existing data sets, such as official weather sensor data?

To establish this we will assess the nature of the sensor data in a number of analytical and qualitative ways, including some summary statistics to establish volume and breadth of activities, an assessment of the tag distribution and the variance of sensor data measurements to assess data heterogeneity, and an assessment of the spatial structure of the data. We will then identify a number of popular sensing activities, and evaluate their outcomes.

1.4 Research Questions

1. What types of Cosm community activity are there?

We will identify some of the key community activities on Cosm. This entails the acquisition and analysis of a large volume of published sensor data, the development of means of identifying community subgroups based on their activities, and the development of descriptive statistics and visualisations that illustrate the nature of such activities.

We expect to find a broad range of activities with a focus on energy monitoring and environmental sensing, and a broad geographic distribution, with clear centres of activity in economically developed and urban areas.

2. How heterogenous is Cosm sensor data?

In order to be able to integrate and aggregate large amounts of sensors it is necessary that participants follow a number of shared practices. At a basic level this involves the use of shared scales and units of measurement, and of shared annotation schemes. The degree to which this takes place can be taken as an indicator for the level of coordination among participants.

We expect to find a number of popular annotation schemes and measurement scales in use, which would demonstrate a potential for shared practices and more explicit collaboration; but also to find a lack of strong evidence for such explicit collaboration.

3. Can Cosm sensor data be integrated to build larger spatiotemporal models?

How easy is it to identify sensors publishing comparable types of sensor data? What indicators can be used to identify and group data streams of shared activities?

We expect to encounter data heterogeneity problems that affect the ability to integrate many individual sensor data streams, including inconsistent uses of metadata annotation schemes, inconsistent units of measurement, and metadata gaps. We will offer a number of methodical approaches to address this.

4. Case Study: How does Cosm community data compare with data gathered by trained specialists?

To assess Cosm data quality we will contrast a subset of its data with an equivalent high-quality control data set to determine whether sensor data published on Cosm can be used to enhance or replace existing sensor data sets. The selected control is a temperature data set produced by the Met Office, the national weather service for the United Kingdom. The data is used for climate modelling and weather forecasts, and its provenance and quality are well-documented.

We expect that the data quality of Cosm community efforts does not match that of the control data, caused by a lack of strong coordination and a lack of metadata that allows to integrate the individual efforts. However we also expect that the spatial density of Cosm data may already match or exceed that of other data sets.

1.5 Limitations

This study cannot describe the Cosm community in all its aspects, instead it merely identifies a small number of practices that are prominently reflected in the data. It is evident that this is an ever-evolving meeting ground for practitioners, curious newcomers, and one-time visitors, and that the nature of the most popular activities can change within the period of just a few months.

An important limitation of such research is that the aims and means of Cosm community participants are often not made explicit, and the provenance of the resulting data (the devices, installation procedures, and calibration practices) frequently cannot be identified from publicly available information. Many of the activities can be described in terms of their data traces, but not in terms of the human motivations that triggered them, or the technologies used to observe particular phenomena.

This presents a major challenge for the evaluation of the outcomes of such activities. With thousands of participants and very little formal structure around most sensing activities it becomes hard to establish sufficient context for all community activities. Attempts were made to identify documentation that can describe these aspects, but in most cases this was not possible.

Instead the literature review will provide a general overview of the motivations and structures behind such activities based on research of comparable data-gathering communities.

Chapter 2

Literature Review

To establish context for a quantitative analysis of Cosm community data in the following chapters we will present an overview of existing research on comparable data-gathering activities. To the best of our knowledge no scientific research on the activities of the Cosm community has been published, there is little quantitative research on the outcomes of similar DIY environmental data-gathering practices, and many key communities have not been studied.

This chapter is organised in three sections. First some general context is provided on the role of community-based data gathering practices in society, then a number of data-gathering practices and project groups are presented, including an evaluation of their outcomes where this exists. The chapter is concluded with a discussion of the potential aims, motivations, and means of communal data-gathering practice.

2.1 DIY Culture and Communal GIS

Recent years have seen an increase in activities around the informal production of electronic hardware, including DIY sensing devices. What are some of the implications, particularly in the context of DIY sensor data gathering? As the literature shows there are clear social effects when large numbers of citizens gain access to new means of information gathering that previously were only in the hands of specialists.

Towards Community-Driven GIS

According to Goodchild (2007) for centuries key contributions to map-making were provided by untrained private citizens, not specialists, and he suggests that recent years have seen a return to this model of communal gathering of geographic information. Contemporary volunteer mapping efforts include the documentation of our planet and its environments on OpenStreetMap, Google Earth, Weather Underground, and in many other places. In contrast to this increased level of public interest and ability Goodchild identifies a reduced ability to support state-funded mapping and data gathering efforts. In the words of Fisher (2011),

> *...as a population we are deciding that governments and civic planners no longer have the ability to provide meaningful information at a local level.*

Because of this confluence of new volunteer abilities and decreased public funding Goodchild et al. (2012) predict a greater need of integrating a more diverse range of data sources and data collection efforts, a necessary expansion of the means through which geo-referenced data is produced and aggregated.

An early and highly successful example of such relatively new efforts is Weather Underground (2012), a long-running Internet-based weather service with a clear focus and the support of an expert team of meteorologists. It already aggregates data from an international network of more than 22,000 personal weather stations, and produces weather maps and forecasts. The Citizen Weather Observer Program (CWOP) is another large-scale effort to aggregate weather data provided by a large interested public, aggregating sensor data from around 6,000 stations which is then incorporated into weather observation systems operated by the National Oceanic and Atmospheric Administration (NOAA) in the United States (CWOP, 2012).

Emergence of a New DIY Culture?

In September 2011 ABC News reports on the New York Maker Faire, a community festival for DIY enthusiasts and "makers" interested in a wide range of activities from textile production to 3D printers and other low-cost computer-controlled manufacturing equipment (Monroy, 2011). It pronounces a "*New Generation of Innovators*":

> *If these pieces of the puzzle sound unfamiliar, it's because they are very recent innovations. This may account for the pervasive feeling among makers of all ages that they are on the cusp of a revolution.*

Kuznetsov (2010) present a review of the activities of a number of such amateur communities of "*builders, crafters and makers*" in a survey of 2,600 participants of the DIY communities Instructables, Dorkbot, Craftster, Ravelry, Etsy, and Adafruit. DIY is defined "*as any creation, modification or repair of objects without the aid of paid professionals*".

The long-term commercial implications of such activities are yet to be seen, but there are some early indicators that there has been noteworthy impact on existing markets. The large U.S. software vendor Autodesk, producer of a range of professional computer-aided design (CAD) software packages, acquired the DIY community website Instructables in 2001, stating that:

> *Instructables will introduce Autodesk customers to a thriving community of like-minded, smart individuals, with whom they can learn and share their personal inspiration or hobbies. (Wauters, 2011)*

> *Autodesk is teaming up with two companies, Ponoko and Techshop to help everyday 'makers' produce products. Ponoko offers a service where people send their designs to the company, and the company will make the parts and send them back to the consumer for assembly. (Gustin, 2011)*

Also in 2011 Radio Shack took steps to cater to DIY electronics communities, a market which they had abandoned years earlier. They sought customer feedback to improve their DIY offerings, and found among the most requested items the low-cost microchip Arduino, kits for self-assembly, and various basic electronic components that had been taken out of stock (The Shack Blog, 2011).

Many of the sensor data enthusiasts found on Cosm are members of such maker communities, and the Cosm website prominently advertises both "Consumer Products" and "Development / Hackable Platforms", two distinct types of sensor hardware that can integrate with the sensor data hub (Cosm, 2012b).

DIY sensor data communities combine the technical interest in informal means of production and electronic sensor devices with a second second strong interest in the gathering of geo-referenced data, be it for environmental monitoring or self-monitoring, and this particular combination presents a shift in means of knowledge production and in public discourse that may have significant long-term implications.

Social and Political Contexts for GIS Use

In 2004 the French sociologist Bruno Latour embraced emergent communal forms of information discourse as an opportunity to re-evaluate "*one of the most tragic intellectual failures of our age*": to separate science and technology from "*the search for values, meaning and ultimate goals.*" He describes how these "collective experiments" may differ from classical scientific practice: they are not limited to controlled laboratories, not limited to expert participation only, no-one is in control of the full experimental setup, and these activities happen at massive scale: they are often global, distributed, highly participative, more driven by personal curiosity than formal structures of scientific inquiry (Latour, 2004).

Such tensions of methodology are already hinted at in an earlier discussion by Pickles (1995) who contrasts contemporary GIS practices with the perspectives of proponents of critical geography who ascribe GIS a technology-centric view on solving social and political problems, and as a means of a minority to wield power over others. He presents commentary by Taylor (1990) who regards GIS as re-emergence of a positivist approach to cartography, a switch from knowledge to information, "*a most naive empiricism*" (Pickles, 1995, p. 12), but also a counter-argument by Openshaw (1991) who sees in GIS a newly holistic view of geography, a new ability to apply the same methods to many different problems.

More recent commentary has moved on from this old conflict. Warren (2011) discusses GIS as a tool to negotiate social and political concerns, she agrees with Latour in that GIS,

> *…once scorned by critical geographers as irretrievably positivist, is now being reconceptualized as a potentially liberating discourse that situates technology in social context …as one of several communication technologies that have evolved throughout the twentieth century with potentially participatory capabilities.*

In this she references GIS criticism as described in Schuurman (2000) who sees inherent problems in the nature of GIS technology, for example a reliance on established structures of power to the detriment of other perspectives. However she also points out that according to Craig et al. (2002) new participative data gathering activities involving GIS achieve an increase in community participation, particularly in the context of social and environmental issues. Warren introduces two key examples of such contemporary GIS use to negotiate power relationships between unequal opponents: the negotiation of Israel's borders with Palestine, and the negotiation of grasslands use in Western China between local residents and government agencies.

Armstrong et al. (2011) describe recent changes in computer systems and information infrastructure and the ways these affect the collection, preparation, and publication of geospatial data. They highlight that such technical change also brings about a shift in perspective: "*people are able to conceptualize problems and interact with others in ways that they were unable to only a few years ago.*" They refer to a new variety of mobile and stationary computing devices, an expansion of telecommunications networks, and to new online services that provide interfaces and channels of collaboration for communal activities.

This communal access to new means of knowledge production can have significant effects. Malone and Klein (2007) argue that computer technology introduces general new abilities to address complex problems such as climate change using software-based means of communal discourse and large-scale simulations. Longhorn (2011) suggests that there is socio-economic and cultural value of publicly funded GI infrastructure that goes beyond their immediate purpose, provided the resulting data is made available to a wider public.

In this sense collective sensing activities can reflect a desire by communities to be more deeply involved in the assessment of environmental problems that affect them, and indeed there are indications that decision-making by privileged minorities alone may not always be the most efficient method of managing shared resources or addressing complex problems. Ostrom and Nagendra (2006) describe a long-term study of forest harvesting systems, and argue to directly involve communities when addressing environmental concerns that affect them, as opposed to acting on their behalf:

> *Evidence from all three research methods challenges the presumption that a single governance arrangement will control overharvesting in all settings. When users are genuinely engaged in decisions regarding rules affecting their use, the likelihood of them following the rules and monitoring others is much greater than when an authority simply imposes rules.*

A further motivation to encourage participation in communal data gathering projects is their potential to aid in behavioural change. Ehrhardt-Martinez and Laitner (2010) provide an extensive report on the effects of "*people-centred*" approaches to energy saving which rely not on economic incentives, but on didactic methods and forms of more direct involvement:

> *Unfortunately, many analysts continue to suggest that while behavior-oriented programs provide a useful way to help deploy smart technologies, they are best thought of as boutique or niche strategies that can only round out a technology-based deployment of more energy-productive investments. We suggest to the contrary; and in this paper we argue that the social or human dimension may have a surprising scale which rivals a pure technology-based perspective in terms of expected long-term, cost-effective energy savings. (Ehrhardt-Martinez and Laitner, 2010, p. 28)*

Mankoff et al. (2007), Malone and Klein (2007), and others offer studies of the use of social networking systems as a driver of such behavioural change.

2.2 Data Gathering Communities

Since the body of research literature on communal sensor data activities is comparatively small, a number of comparable activities and structures are presented where existing research can aid to understand the breadth of activities. Additionally some research is presented on the outcomes of a number of communal data gathering practices that helps establish what expectations can be made regarding the potential and limitations of such efforts.

Volunteered Geographic Information

A number of terms have been introduced to describe the confluence of new technologies and new abilities for mass-participation in data-gathering efforts, including the outcomes this yields. Some of the key terms are introduced here, further terminology is presented throughout this chapter.

Goodchild (2007) provides a number of key examples of community-based GIS applications that have come about in the last decade, the outcomes of which he terms Volunteered Geographic Information (VGI.) He describes participants of such web-based applications gather, aggregate, and publish geo-referenced data voluntarily and at increasing volumes, particularly using websites and applications such as Wikimapia, OpenStreetMap, and Google Earth, but also publishing geo-annotated photos on Flickr, among other activities.

Wang (2010) introduces the term "CyberGIS" for the infrastructure and methods underlying such activities, a synthesis of Cyberinfrastructure that allows to coordinate distributed information-gathering efforts, GIS, and spatial analysis.

Boulos (2005) first coined the phrase "*Wikification of GIS by the masses*": a process of change where GIS activities become increasingly collaborative, and where contributions by and arguments between peers replace top-down information flows controlled by a privileged minority. Kamel Boulos et al. (2011) provide a review of some the progress made since then:

> *Google Earth is now a full-fledged, crowdsourced 'Wikipedia of the Earth' par excellence, with millions of users contributing their own content to it, attaching photos, videos, notes and even 3-D (three dimensional) models to locations.*

Around the same timeframe that Google Earth evolved to this state, the quality of several purely volunteer-driven efforts has reached that of commercial works, leading up to a present where the popular location-based service foursquare is among the first to replace their use of Google Maps with the volunteer maps produced by OpenStreetMap (Foursquare Blog, 2012).

Communal Mapping

Communal mapping activities are particularly well documented, partially because this is one of the earliest domains of Internet-driven VGI, but also because it is an area where great progress has been made in a short amount of time. The focus is often on evaluating the relative quality of the data produced by volunteer mapping groups.

Neis et al. (2011) present a number of comparative studies that evaluate a subset of OpenStreetMap street network data against that of a commercial provider and find only a 9% difference, and additionally found that the OpenStreetMap street network data exceeds that of the commercial equivalent by 27%.

Välimäki (2011) compares the positional accuracy and annotations of OpenStreetMap data with official data by the National Land Survey of Finland and finds the quality to be "*good*", with a main limitation being a greater degree of uncertainty. It is neither documented how complete OpenStreetMap data is, nor whether all the data is acquired with the same degree of rigour in all locations.

Haklay et al. (2010) find that the number of participating volunteers that produce mapping data for a particular region has a noteworthy impact on the quality of the positional accuracy of the resulting mapping data.

Zielstra and Zipf (2010) find that the combined road length of OpenStreetMap volunteer maps was smaller than the map data of an equivalent TeleAtlas MultiNet dataset, but that this difference was reduced from 29% to 7% in a period of only 8 months.

These are just a few examples of the many evaluations of communal mapping activities, and they clearly demonstrate that not only there is growing public interest in the gathering of geo-referenced data, but that the quality of VGI is steadily increasing. Goodchild et al. (2012) offer a recent survey of further research activities.

Geo-referenced Social Media

Popular social media platforms such as Flickr and Facebook often allow the annotation of contributions with location information, and they are quickly turning into producers of large corpuses of geo-referenced data. Early research suggests that such data may have uses beyond its initial personal purpose.

Kennedy et al. (2007) attempt to algorithmically derive location descriptors from a corpus of annotated photos published on the photo sharing community site Flickr and find that such annotations can be used to navigate large archives of photos. Jaffe et al. (2006) develop comparable algorithmic means of summarising and presenting large collections of geo-referenced photos.

Rattenbury et al. (2007) derive event and place semantics from Flickr tags, and while they identify problems attributable to sparse data in many locations they find that the San Francisco Bay is particularly well-represented, to a point where some areas are well-documented at sub-city and sub-region scales.

A growing body of research evaluates the ability of social media and mobile phone applications to act as monitoring systems and early-warning systems, frequently under the term "citizens as sensors". Sakaki (2009) propose an algorithm for location-specific event detection based on Tweets, and manage to detect 96% of all earthquakes that have been reported by the Japan Meteorological Agency at a seismic intensity of 3 or higher. Demirbas et al. (2010) employ Twitter as a collaborative weather radar and noise mapping application. Sheth (2009) describes a range of Twitter-based event detection applications and some of the technologies and methods used to construct event ontologies from free-form text. More examples are provided by Heinzelman and Waters (2010), and others.

Communal Sensor Data

The preceding sections presented research on a range of activities, however not all areas of communal data gathering activity are equally well understood. Most of the existing research of communal sensor data activities falls into two broad categories: project presentations by project practitioners, and commentary that attempts to describe general trends.

Kamel Boulos et al. (2011) discuss the notion of a "Sensor Web" and citizen sensing in the context of environmental and public health monitoring and disaster response, and introduce some of the key technologies that help integrate increasingly heterogenous citizen sensor networks. The range of projects reviewed in their paper include "citizens as sensors" approaches and mobile sensor applications such as Usahidi as well as fully automated collective sensor networks.

Campbell et al. (2008) introduce MetroSense, an application that combines "citizens as sensors" with more automated mobile sensing activities. Instead of deploying dedicated sensing infrastructure they build on the wide distribution of mobile phones. Their application aims to operate in the background without requiring human input, an approach they term “opportunistic sensing.” Such opportunistic sensing projects have the potential of involving very large numbers of volunteers.

Gouveia and Fonseca (2008) present the Senses@Watch project and show that the sensors suitable for participative projects lack the accuracy and precision of more expensive sensor devices. Consequently they recommend the careful design of data quality evaluation mechanisms.

Cuff and Hansen (2008) review a number of “urban sensing” projects and posit that “*urban sensing shifts focus and control away from the scientist at the center*”, but also highlight two key challenges of participative sensor network models: maintaining data quality, and the observer effect where the act of monitoring a system changes the behaviour of participants.

Research on distributed sensor infrastructure, along with some reflection on the impact of large numbers of participants, is provided by Masser (2011), Davis et al. (2011), and others.

2.3 Community Structures and Motivations

The motivations, processes, and structures that guide contributors all have an important impact on the outcomes of communal data-gathering projects. They can affect the potential geographic distribution of volunteers and sensed data, the technical barriers to participation and need for training of participants, the timeliness of results, and many other aspects. Do the motivations to participate align with the needs for rigorous data collection? Under which conditions can the results be trusted?

Motivations to Contribute

In their survey of contributors to a number of online DIY communities Kuznetsov (2010) assess the structure of stated motivations to participate in communal activities and identify among them an desire to find “*inspiration and new ideas for future projects*”, to “*learn new concepts*”, and to a much lesser extent a utilitarian interest in “*finding employment*” or “*improving online reputation*”. In this they find similarities with prior studies of the motivations of open source software contributors, and contributors to communal online projects such as Wikipedia and SETI@Home.

In a more general context Coleman and Sabone (2010) summarise findings from a range of empirical studies of online volunteer activities ranging from open source software development and Wikipedia contributions to participation in consumer review sites, and others. They produce a consolidated list of potential motivations: professional or personal interest, intellectual stimulation, protection or enhancement of a personal investment, social reward, enhanced personal reputation, having an outlet for creative and independent self-expression, pride of place.

Citizen science research yields further insights into the nature of motivations to specifically join volunteer data-gathering efforts, albeit with a particular focus:

> *A citizen scientist is a volunteer who collects and/or processes data as part of a scientific enquiry. (Silvertown, 2009)*

Wiggins (2010) suggests that

> *...in the right circumstances, citizen science can work on a massive scale and is capable of producing high quality data as well as unexpected insights and innovations.*

In an empirical study of ecological citizen science projects Rotman et al. (2012) find that motivations to contribute change over the duration of a person's participation. In particular, initial motivations are driven by self-interested reasons such as familiarity with or curiosity about the topic, or project engagement as a potential career building block. Other motivations only appear later and drive long-term engagement, including community involvement, the opportunity to learn, and the initiation to scientific practices and modes of inquiry.

Cohn (2008) identifies some of the appealing features of public participation from the perspective of scientists and research institutions: volunteers cost less, and they allow researchers to cover larger geographic scales and longer time periods; but citizen science projects also encourage volunteers to participate in the scientific process, and to interact with their surroundings.

Further research on motivations is frequently based on existing theories in social sciences, such as Klandermans (1987) on participation in social movements and Maslow (1943) on human motivation.

Community Structures

Elwood (2011) introduces different forms of "participative GIS": practices of collective geo-data gathering and usage that invite participation by a public. She distinguishes between Participative GIS (PGIS) projects that are initiated by volunteers, and Public Participation GIS (PPGIS) projects that are initiated by government and other large institutions. Jankowski (2011) suggests that PPGIS projects are well-planned systems, not emergent bottom-up systems.

Elwood also introduces the notion of "grassroots GIS", GIS use by social movement groups, community organisations, and activists, often highly localised and personal, and with a focus on self-representation. The problem statements, tools, and protocols are often developed by participants themselves. Elwood points out that descriptive models of such systems are drawn from political economic and social theory, and that projects are often studied 'in situ' rather than through for example quantitative evaluations of project outcomes.

Malone et al. (2009) develop a framework to help understand Internet-based systems of "collective intelligence" based on a study of more than 300 community-driven projects, including Google, Wikipedia, the t-shirt design community Threadless, and many others. They refer to a number of terms that have been used to describe these systems: "*radical decentralization, crowd-sourcing, wisdom of crowds, peer production, and wikinomics*". They offer categories of inquiry to assess project structures, including questions about authority structures, incentives, the kind of participation from the evaluation of incoming data to the creation of new artefacts and knowledge, and the means of production, which may include self-interested curation or contests, means of collaboration and peer production, or group decision-making using voting mechanisms, consensus, averaging, prediction markets, and others.

There is a growing amount of new terminology to describe structures and processes of communal data gathering, and in many cases the boundaries between the terms are not clear, as for example demonstrated by Tulloch (2008). The structures of a particular project are rarely easily placed within a single category, and frequently change over time.

Additionally many of the presented categories are not concerned with ad-hoc collaboration between DIY practitioners as evident on Cosm, but more frequently focus on formal structures where there is a clear distinction between project initiators and participants.

There is further applicable writing on communal decision-making, the nature of cooperative efforts, communal resource management and more, including writing on organisational structure (Mintzberg, 1978), collective action (Olson, 1965), the management of shared resources (Ostrom, 1990), small group research (Poole et al., 2004), and others.

Credibility and Trust

How reliable is the data produced by volunteer-driven data gathering projects? What measures can be taken to assess and improve data quality? Are the results trusted? The majority of quantitative evaluations and comparative studies of VGI exists either in the domain of community mapping projects or in citizen science research, and a few of the key studies have been introduced in earlier sections.

Wiggins and Newman (2011) identify data quality as a primary concern for citizen science practitioners, and present a survey of 280 projects about their approach to data quality and validation which culminates in a framework of 18 commonly employed mechanisms. They find that 75% of all projects use multiple data validation methods, with an average of 2.5 validation methods per project. Among the most frequent choices are expert reviews of contributions, photo submissions for species identification, submission on paper along with each online entry, replication (making use of redundant labour) or rating of contributions, and training programs for volunteers. They also present criteria to evaluate the applicability and expected efficacy of data quality and validation mechanisms for particular contexts.

There are considerations around the perceived trustworthiness of community-gathered data that go beyond measurable data quality. As Goodchild (2007) observes:

> *VGI is sometimes termed asserted geographic information, in that its content is asserted by its creator without citation, reference, or other authority.*

Bishr (2007) discuss the propensity of publicly accessible online collaborations to attract vandalism or invite the contribution of inaccurate information, and see a need for mechanisms to express and manage varying levels of trust. They suggest that data quality is subjective, and describe quantitative means of assessing data quality as a function of trust between participants.

Flanagin and Metzger (2008) discuss the relationship between VGI quality and its credibility. They posit that traditional gatekeeper models become untenable in this context since any rigorous vetting of large volumes of VGI data would come at a significant cost. As a result they propose that, depending on the context of use of a collectively acquired data set, its credibility can either be a function of its accuracy, or merely of its perceived sufficiency for a particular task. This is particularly true for data that is produced by participants who may have low technical expertise in data-gathering, but a high degree of domain knowledge. The authors suggest:

> *…while credibility-as-accuracy is an appropriate concept for those who have a 'factual' relationship with geospatial information (as do most scientists), credibility-as-perception is more useful for those who use VGI for social, communal, or political purposes.*

Chapter 3

Methodology

This study aims to develop data integration techniques for volunteer sensor data published on Cosm in order to produce large-scale spatiotemporal models of the measured phenomena, and then to test these techniques by identifying a number of key sensor community activities in the data. This is highly heterogenous, geo-referenced, and tagged time series data, complex in structure, and encompassing a large variety of observed phenomena.

In order to accomplish this a large volume of public Cosm sensor data was acquired, covering a period of four months in 2011 and 2012. This chapter describes the general nature of the data, and then introduces a number of methods and visualisations that are used to assess many aspects of this previously unexplored data set, and to identify subgroups of activity based on sensor annotations. This is followed by a summary of the results, and a discussion of the findings.

3.1 The Cosm Data Set

The Cosm service allows its participants to capture and archive realtime data streams from arbitrary sensors. This data can then be displayed on Cosm dashboard pages (see Figure A.1 on page 63), searched, and exported via the Cosm API. Participants can additionally grant data access to applications that read or write sensor data.

In order to be able to assess many Cosm community activities over longer periods, several months worth of historic data were acquired for all public sensor feeds active at that time. An index of all environments and datastreams was retrieved on 27 May 2012, this index contains around 70,000 datastreams for 20,000 environments. In total the acquired data set encompasses around 80 million sensor readings including timestamps, covering August and September 2011, and March and April 2012.

The following sections describe some of the elements of Cosm sensor data in more detail, with a focus on aspects relevant to this study. A more detailed discussion of the data acquisition process is provided in Appendix B on page 65.

The Data Format

Cosm sensor data is geo-referenced and tagged time-series data, with facilities to provide a number of further annotations. It is published using the data format of the Extended Environments Markup Language (EEML), a structured document format that can contain both raw sensor data and metadata with further contextual information (Haque, 2008a). EEML has a number of properties that make it an interesting data interchange format for collaborative sensor projects:

- Its use of tags allows the creation of folksonomies (Wal, 2007).
- Location information can be provided in a detailed and structured manner.
- Units of measurement can be described in a number of standard notations, or as free-form text.
- Sensor data values are not restricted by type: sensors can capture arbitrary character strings as well as numeric data.
- Valid EEML documents can be produced with a minimum of effort since most of the metadata annotations are optional.

The data format attempts to strike a balance between well-structured metadata that uses a controlled vocabulary, and free-form textual metadata that allows data producers to describe new practices. While it offers the ability to describe sensor data with a rich set of annotations, in practice not all published data is equally well described.

Sensor data is organised in "datastreams" which contain the time series data of sensor readings for a single sensor, and "environments" which are collections of one or more datastream, and which can provide further contextual data.

Environments can be annotated with an arbitrary number of tags, and with a detailed description of a location including the location name, geo coordinates, elevation, and its "exposure" (outdoor or indoor placement), "disposition" (fixed or mobile), and "domain" (physical or virtual). Each environment on Cosm has a globally unique identifier in the form of a numeric ID. Environments are sometimes also called "feeds" (Cosm API, 2012d).

Datastreams can be annotated with an arbitrary number of tags and with a description of the sensor's unit of measurement. Every datastream has a user-defined textual identifier in lieu of a title, but no numeric ID. Datastreams are also called streams, data streams, or sensor data streams (Cosm API, 2012c).

Units of measurement for a datastream can be specified as a unit's "name", "type", and "symbol", all of which are optional. Despite the presence of such a thorough classification scheme participants are given little guidance on Cosm when specifying units of measurement. The unit "type" attribute is the only element with a restricted vocabulary, adopted from SI units as described in ISO/DIS 16739 (2012), however it cannot be specified when using the Cosm web interface to create new datastreams. (See Figure A.2 on page 62.)

A datastream stores historic sensor data as "datapoints", a list of values and timestamps. Sensor values can be published by sensor devices to the Cosm API in realtime, in which case the Cosm service archives them along with a server-provided timestamp. Alternatively devices can submit batches of historic values including timestamps.

Measurement values are recorded as character strings, which means there is great freedom in the choice of sensor data formats: numbers, string values, structured data, and others. This makes Cosm effectively a dynamically typed system, and its API a generic web-based data store. However it has the side-effect that it is impossible to describe general rules for validating sensor data. Data capturing errors are not caught by the API, instead preserved as part of the datastream, be they caused by software bugs, sensor problems of embedded devices, or other problems, and it is in the responsibility of the data producer to monitor their sensor measurements for errors. Cosm dashboards only draw charts for sensor data that can be interpreted as numbers (Cosm API, 2012b).

Numeric	Non-numeric
`-0.5`	`17,5`
`15`	`10.30 19.60`
␣`15`	`0.0 cm`
`10,051.9`	`2012-03-09 14:14:00`
`5.0E-5`	`d=0000000000`
`-4.076958E-5`	`** NE **`
`0.0`	`<font color='#00FF00'>220v Nominal</font>`
	`[actual_wind0_speed_kmh]`
	`{"10":7, "11":8, "12":10, "13":11}`
	`-2.-71`
	`22:48:32`
	`00..`
	`Edge Lexington`

Table 3.1: Examples of sensor measurement values encountered in Cosm datastreams. The ␣ symbol indicates the presence of a whitespace character.

A General Overview

After the data was acquired a number of general assessments were made in preparation for the development of data integration techniques. These served to establish the general character of the data set, including its volume and breadth, general spatial distribution, and longevity of activities, but they also established some early insight into the degree of heterogeneity of metadata annotations and sensor measurements. This latter aspect will be discussed in more detail in Section 3.2.

Account Activity

A first general point of interest is the growth rate of participation, and the degree to which activities are sustained over longer periods. Participation requires substantial effort and long-term commitment: it demands the acquisition of sensor hardware, some technical expertise in the setup, and ongoing maintenance to ensure sensors remain active.

As a result we expect there to be low adoption rates, and a very high rate of turnover.

Data derived from the Cosm sensor feed index (Cosm API, 2012f) reveals that the rate at which new environments and datastreams are created has significantly increased over the past three years. Particularly in 2011 and early 2012 there appears to have been great significant in adoption. (Figure 3.1)

The same data shows that among the datastreams that publish any measurements, around 75% are still active after 24 hours, and around half are still active 90 days after their creation. In combination those metrics indicate that there is clear potential for such sensor data communities to sustain long-term monitoring activities. (Figure 3.2)

However as the number of participants increased, the longevity of activities was somewhat reduced. It is hard to draw conclusions from this without further review. A potential cause may be that more recent generations of adopters have different motivations than earlier adopters. For example the presence of existing activity may now attract new contributors who are less committed, but curious enough to join the community.

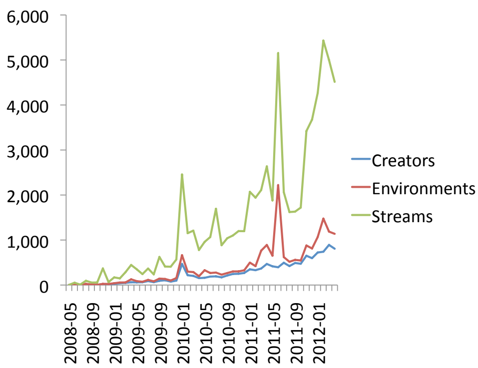


Figure 3.1: The monthly volume of newly created environments and datastreams, and the number of participants who created these new data sets.

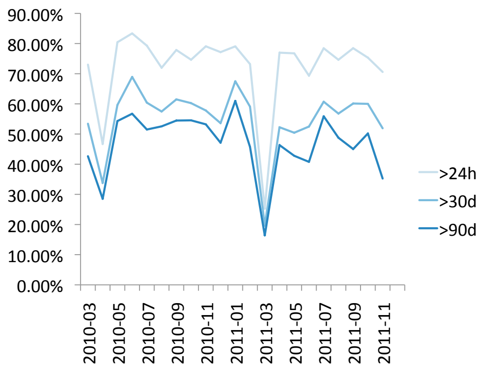


Figure 3.2: Longevity of sensor activities, shown as the percentage of datastreams that are still active 24 hours, 30 days, and 90 days after their creation.

Geographic Coverage

Of all environments in the public Cosm feed index 60% have provided geo coordinates, and 55% have provided a location name. This is a remarkable volume of geo-referenced data, and likely a consequence of Cosm's interface design: during the creation of a new sensor environment, coordinates can easily be specified with a zoomable and clickable map interface as shown in Figure A.2 on page 64. It is unclear how accurate these annotations are.

The key centres of Cosm community activity are in developed countries across the world (see Figure 3.3 on page 27). Many areas of significant community activity are found in Europe, with several strong hotspots of activity in the UK, Holland, Belgium, and Switzerland, and significant levels of activity in most other Western European countries

(Figure 3.4). In Japan activity is clustered around the commercial centres Tokyo and Kyoto/Osaka, but also the coastal region towards Fukushima (Figure 3.5). Similar high activity levels can be found in the United States, where the key hotspots of activity are in the San Francisco, the northern regions of the East Coast around New Jersey, New York, and Boston, but also in Washington State around Seattle (Figure 3.6). In Australia the centres are in Adelaide, Melbourne, Sydney, and Brisbane, but also a number of other larger coastal cities (Figure 3.7).

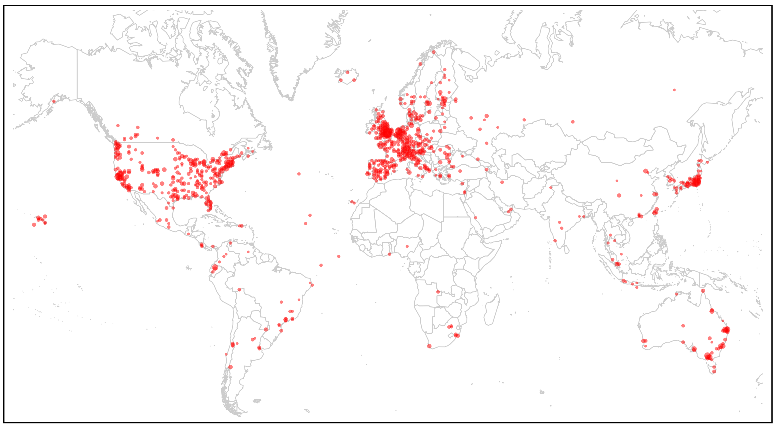


Figure 3.3: Global locations of Cosm sensing activity, as determined by datastream activity for geo-referenced feeds in April 2012.

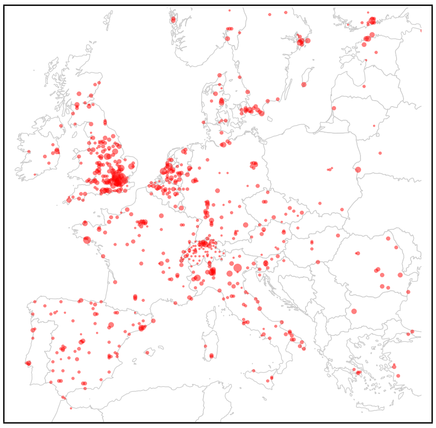


Figure 3.4: Cosm sensing activity in Europe in March 2012.

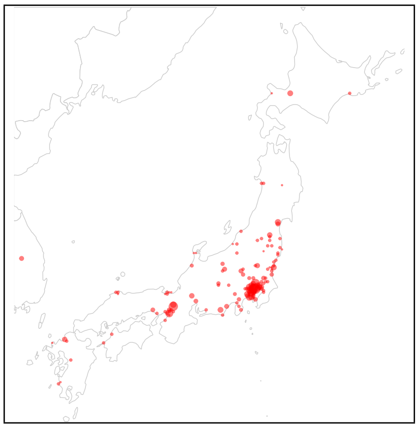


Figure 3.5: Cosm sensing activity in Japan in March 2012.

A breakdown of popular location names in Table 3.2 shows that this attribute may contain personal references as well as city names in a variety of spellings. For the remainder of this study we exclusively rely on geo coordinate annotations, not location names, when geo-referencing data sets since the manual effort required to clean and geo-reference these location names did not yet weigh up the potential benefit. Only 8% of all listed environments have associated location names but no geo coordinates.

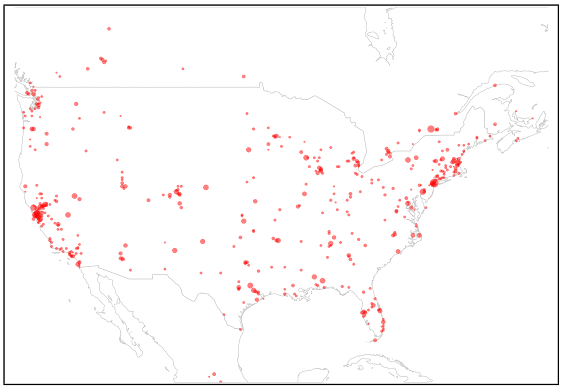


Figure 3.6: Cosm sensing activity in the United States in March 2012.

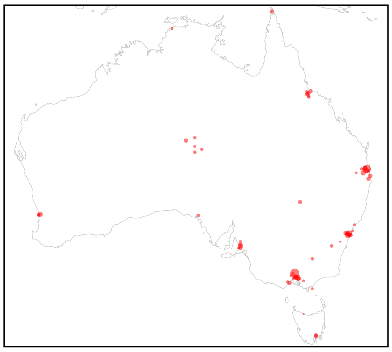


Figure 3.7: Cosm sensing activity in Australia in March 2012.

Home	London	Japan
home	office	Barcelona
UK	Amsterdam	Madrid
Singapore	Paris	UNSW, Sydney, NSW, Australia
japan	Sydney	Tokyo
Toronto	London, UK	Beijing
london	New York	Brisbane
My Room	Brooklyn, NY	Montreal
Eindhoven	Brooklyn	New York, NY

Table 3.2: Some of the most frequently used location names in the Cosm feed index.

Popular Tags

Cosm makes use of a folksonomy, its tag-based annotation system, as sole method of grouping related sensing activities by individuals into a shared corpus. This is a flexible system which allows data producers to create specialised annotations that are specific to their individual activities.

- In total there are around 17,000 distinct tags, 16,000 ignoring capitalisation. These have been applied around 140,000 times.
- Around 24% of tags have been applied to two or more datastreams, 14% to three or more, and 8% to 5 datastreams or more.
- 60% of all datastreams have at least one tag, 20% two or more, and 8% have 3 or more tags.
- 45% of all environments have at least one tag, 30% two or more, and 18% have 3 or more tags.

These numbers suggest the presence of a power-law distribution, this will be discussed in the context of data heterogeneity and term distributions on page 31. Table C.1 and Table C.2 in the appendix on page 69 present visualisations of the number of participants who are contributing sensor data to a selection of tags over the observed period.

temperature	watts	longitude
latitude	altitude	alt
elevation	lat	lon
power	Temperature	electricity
degrees	celsius	coffee
temp	Humidity	humidity
contentment	psychogeography	mood
Lat	Lon	Temp
light	test	LDR
light sensor	Light	analogRead0

Table 3.3: Some of the most frequently used datastream tags.

twitter	twitter_stats	temperature
arduino	dashboard	gsiot
test	Temperature	Arduino
humidity	nSv/h	radiation
Mark2	cpm	light
Geiger	electricity	temp
energy	power	モニタリングポスト
Humidity	sensor:type=radiation	線
weather	Comma-separated descrip …	sensor
ガイガーカウンタ	Test	air quality

Table 3.4: Some of the most frequently used environment tags.

Units of Measurement

Like tags and location names, units of measurement can be provided as free-form text, and the use of this annotation is consequently just as diverse and inconsistent, even when accounting for differences in language and notation. Table C.3 on page 71 presents the popularity over time for a number of units, measured by number of participants.

- In total there are around 2,800 distinct units of measurement, 2,300 ignoring capitalisation.
- Around 30% of units have been applied to two or more datastreams, 23% to three or more, and 19% to 5 datastreams or more.
- Around 50% of all datastreams have been annotated with a unit of measurement.

Celsius	celsius	watts	Watts
followers	tweets	friends	favourites
lists	followers_to_fr…	C	%
°C	Cups	Volt	Fahrenheit
Volts	Celcius	W	F
cpm	V	degrees	kilograms
nSv/h	Degrees	1	Watt
kWh	Percent	Percentage	volts
Relative Humidity	hPa	percent	watt
ºC	c	microsieverts p…	celcius
counts per minute	mm	lux	Lux
RH	Degrees Celsius	m/s	%RH
kW	A	volt	Amps
degree	m	counts/minute	microsieverts/h…
km/h	ppm	cm	Deg C

Table 3.5: Some of the most frequently used units of measurement.

Sensor Values

Datastreams predominantly capture numeric data: of all sensor measurements in March 2012, under 2% could not be interpreted as a number. Manual inspection showed that many of these were the result of data capturing errors or software bugs, as shown in Table C.4 on page 72. Since the focus of this study is on the integration of sensor feeds and not primarily the exhaustive exploration of the data no effort was made to interpret textual sensor information, and instead such values were discarded.

3.2 Assessing Data Heterogeneity

Before attempting to aggregate or integrate the data captured by large numbers of sensors it needs to be ensured that their data is suitably homogeneous. To assess this it may be necessary to identify the specific nature of the measured phenomenon, and the physical properties that are observed by a sensor. The measurement of "air quality" can mean a number of things: it could be described in terms of ozone levels, particulates of varying sizes, nitrogen oxide levels, nitrogen dioxide levels, sulphur dioxide levels, and in other ways.

The scale and units of measurement are a further consideration. Radiation levels can be quantified in "counts per second", "microsievert", "millirem", and any number of additional ways. And all of these entail an additional consideration of notation, which includes the use of multiple languages, of abbreviations, but also spelling errors. The same Celsius temperature scale can be referenced as "degree celsius", "ºC", "Grad Celsius", ambiguously as "º" or "degree", or even as "Celsuis".

Along with this come considerations of precision and accuracy: has the sensor been calibrated? Does it produce repeatable results? Has it been set up in a suitable manner? Have all sensors of the same kind been set up in a consistent manner, at the same elevation above ground, either indoors or outdoors, mobile or in a fixed location, and with consideration for local air flow, temperature, and humidity, all of which may affect measurement results? Have their datastreams been annotated accordingly to reflect all these choices?

In the context of DIY sensing any number of these conventions can be expected to be ignored or violated, and the initial assessment of Cosm data has already yielded indications that this is in fact taking place to a significant degree. However sensor data that has at least a small number of annotations may allow us to draw conclusions about some missing aspects, and the large and growing volume of the Cosm data corpus may allow us to apply statistical methods of inference when attempting to establish the nature of a published sensor feed.

Such an expected variability in data quality poses the main challenge in integrating the outcomes of DIY sensor activities. We will present a number of techniques to assess these aspects of Cosm sensor data where it is possible, and determine some of the implications for the ability to integrate such data.

Tag Term Distribution

In our early assessment of the tagging vocabulary we already have seen indicators of the presence of power-law distributions in the data. Term frequency charts for tags in Figure 3.8 and Figure 3.9 indicate visually that this may indeed be the case, as do the following numbers:

- 8,300 participants have tagged 44,700 datastreams with 13,800 distinct tags.
- About 25% of these tags have been used more than once.
- About 15% of these tags have been used by more than one person.
- 4,600 participants have tagged 9,200 environments with 5,000 distinct tags.
- About 40% of these tags have been used more than once.
- About 20% of these tags have been used by more than one person.

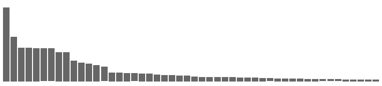


Figure 3.8: Datastream tag term frequency. The top tags show a clear peak in use, and there is a sharp drop around rank 15. (Both charts are limited to the top 50 ranks only.)

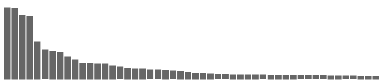


Figure 3.9: Environment tag term frequency. The top 4 tags see a similar level of use. The top 2 tags refer to a Twitter monitoring application, see Table 3.4 on page 29.

Using the maximum-likelihood method of power-law parameter estimation introduced by Clauset et al. (2007) we determined the exponent, or scaling parameter, of this tag frequency distribution and evaluated the result with a Kolmogorov-Smirnov goodness-of-fit metric, and we can demonstrate that there is reasonable fit with a power-law function.

We assume that the choice of units of measurement, location name, and other annotations follow similar power-law distributions.

This term distribution shows that particular tags alone cannot suffice to meaningfully group datastreams by shared activity, as this would not allow use of the majority of the data. At minimum it becomes necessary to attempt to identify synonymous tags, variant articulations of semantically equivalent concepts, in order to aggregate larger numbers of sensor data feeds.

Dataset	**n**	α	**D**
Datastream tags	100	1.750000	0.064893
Datastream tags	1,000	2.030000	0.019593
Environment tags	100	1.970000	0.028414
Environment tags	1,000	2.000000	0.017065

Table 3.6: A number of iterations of power-law coefficient estimates (α) and corresponding Kolmogorov-Smirnov goodness-of-fit metrics (D) on a sampled subset (n) of the tag frequency tables. Both distributions show a good fit for power-law curves.

Data Variance Plots

Statistical variance as measure of heterogeneity of sensor measurements can help establish whether captured sensor values within a particular group of feeds are generally in a comparable range. It is noteworthy when measurements differ by multiple orders of magnitude, or when there is no correlation in relative change over time between any sensors of a nominally comparable type.

Variance plots, as presented in Table C.5 on page 73, allow to visually compare ranges in values over time for groups of sensors. A single plot presents the time series data of multiple feeds (all datastreams of a certain tag) along with an overlaid trend curve of the median sensor value across feeds. In combination those two plots give a simple visual indication of measurement heterogeneity over time.

In order to accommodate the fact that different feeds update their measurements at different times of day the measurements for each datastream are aggregated to one value per day, computed as the median of a all measurements that day. This removes diurnal cycles, but it makes it possible to compare long-term trends.

In addition the coefficient of variation is computed for every day across the values of all sensors in the group. The point in time of the minimum and maximum coefficient of variation are shown along the horizontal axis with a blue and red dot, respectively. This is a simple visual aid to help identify the points of smallest and largest variance. The coefficient of variation is a useful descriptor of variance in this context because it is a relative measure, and comparable across plots even when two data sets have different scales of measurement.[1]

[1] The coefficient of variation does however have a number of caveats. It is not suitable for measurements of fractional values: as measurements approach zero the coefficient of variation approaches infinity. Additionally the coefficient of variation may not yield a meaningful result when values are on an interval scale, not a ratio scale, and consequently can have negative values. This can be seen in the plot for

On the right-hand side of each plot three labels indicate, from top to bottom, the group size (the number of sensors shown), the average value across all measurements, and the average coefficient of variation across all days of a plot. The latter is a basic indicator of the overall measurement variance of the group, with a larger coefficient of variation indicating greater variance.

The data variance plots of a number of selected datastream tags in Table C.5 strongly suggest that while there are some types of activity which have a more homogeneous distribution of measurements, overall there is a great diversity in sensor value distributions and trends. This is attributable to systemic differences between the observed properties, to differences in measurement scales, but also to differences in measurement setup, such as the choice of device and calibration, of whether a sensor is mounted indoors or outdoors, and other factors.

3.3 Identifying Groups of Activity

The power-law distribution of metadata terms makes the integration of large numbers of sensor data feeds challenging, particularly since Cosm provides little guidance to its participants about how to effectively apply metadata. As a result the vast corpus of Cosm sensor metadata is likely to contain inconsistencies in the use of metadata annotations by different individuals, as well as multiple redundant ways of describing the same concepts.

The task of grouping related activities then necessarily entails the act of building term lists of related annotation schemes, and such term lists then are starting points for describing particular activities. For example the selection of all activities to measure outdoor temperature in a particular region may entail the combination of the following constraints:

- Datastream tags include any of `temperature`, `température`, or `temperatura`.
- Location disposition is `outdoor`.
- Unit of measurement is any of `c`, `C`, `°C`, or `Celsius`.
- Geo coordinates are provided, and within the area of interest.

Finding groups of shared activity among all sensor data streams requires the identification of semantic concepts shared by multiple creators, and some care in the selection process, including consideration for different modes of annotation. The following discussion introduces such a selection process for two types of annotation: tags, and units of measurement.

the "longitude" tag in Table C.5 which for March 2012 incorrectly indicates a negative coefficient of variation. For these reasons each plot indicates both the variation and the mean, and the standard deviation of measurements can then be derived as the product of coefficient of variation and mean value. For a negative coefficient of variation it can be assumed that the mean is negative and the standard deviation positive, since the latter is never negative.

Interpreting Metadata Uses

The semantic interpretation of Cosm tags is subject to two main considerations. First it is necessary to identify the tagging domain: has a tag been applied to an environment or to a datastream? One may expect that environment tags are more general, in that they may describe the location and general purpose of a sensing activity, while datastream tags are more specific, in that they describe a particular sensor and the nature of its observed phenomenon. In practice however Cosm provides no official recommendation on this matter. A ranking of popular environment and datastream tags reveals both differences and similarities in the vocabulary used for both domains.

Environment tags	**Datastream tags**
twitter_stats	watts
twitter	temperature
temperature	Temperature
nSv/h	Humidity
Mark2	power
cpm	electricity
Temperature	1 時間移動平均
arduino	humidity
radiation	longitude
humidity	latitude
Geiger	altitude
Arduino	lon
モニタリングポスト	lat
線	alt
electricity	elevation

Table 3.7: The most frequently used environment and datastream tags for sensor data in March 2012 as ranked by the number of tagging participants.

Making use of environment tags as a factor in the identification of comparable activities risks that the term vocabulary for such a grouping gets diluted with unrelated terms. We cannot make the assumption that all participants are disciplined in semantically grouping datastreams, instead it is more likely that many participants may group unrelated streams in the same environment as some initial data exploration suggested. Since the volume of environment tags is comparably low we will not rely on environment tags in this initial study of the data set, and instead focus on the use of datastream tags and units of measurement.

A second consideration in the interpretation of Cosm tags is the choice of tagging vocabulary. Since the vocabulary is not restricted there are many possible ways of describing the same concepts with different words. Interpreting tag vocabulary for the purpose of grouping them by activity entails the detection of tag synonyms, which may include variances in spelling, variances in capitalisation, the use of symbols, abbreviations, or full sentences, spelling errors, and the use of different languages. A ranking of popular datastream tags of two different geographic regions may illustrate some of these variations.

Japan	**United Kingdom**
1 時間移動平均	watts
10 分間移動平均	power
cpm	electricity
1 時間移動平均	temperature
sensor:type=radiation	Temperature
Temperature	Humidity
10 分移動平均	Pressure
humidity	humidity
CPM	light
radiation	Power
Humidity	Outside Temp
sensor:model=lnd-712	latitude
temperature	longitude
sensor:model=SBM-20	Wind Speed
µSv/h	elevation

Table 3.8: The most frequently used tags for datastreams active in March 2012 within the geographic regions of the United Kingdom and Japan.

Similar concerns arise for metadata specifying the units of measurement, as has been discussed in the context of data heterogeneity on page 30. Although Cosm in principle supports the use of a controlled vocabulary to describe such units (refer to the discussion of the EEML data format on page 24) in practice there are no restrictions on the form of unit names, which means they become subject to similar spelling variations as those found in tags. Additionally there are often multiple systems of measurement for the same environmental phenomenon, such as the use of Celsius, Fahrenheit, or Kelvin to describe temperature measurements. This can be illustrated with a ranking of frequently used units for datastreams from different regions or tagged with different languages.

temperature (US)	temperature (UK)	temperatura
Celsius	Celsius	Celsius
Fahrenheit	C	ºC
Degrees	Degrees Celsius	C°
degrees C	degrees C	celsius
Degrees F	DEGC	ªC
degrees Fahrenheit	Degrees C	Centigrados
Degrees Fahrenheit	degrees Celsius	degree Celsius
F	degrees centigrade	Grados
°F	T	°C
Farenheight	Celcius	GRADOS CENTIGRADOS

Table 3.9: Units of measurement for datastreams active in March 2012 for three different groups: sensors tagged with `temperature` in the US and the UK, and sensors tagged with `temperatura`.

Synonym Detection

Clements et al. (2008) observe that tag synonyms are often the result of distinct groups of participants employing different language for the same concepts, such as the different spellings of "humour" and "humor" among English speakers. In such a case these two terms will have a high item correlation (many items are tagged with both terms, few just with one spelling variant) but a negative user correlation (very few participants make use of both spellings.) They present an approach to synonym detection using Pearson correlations of user and tag similarity measures.

We have implemented Clements' synonym detection and present some of our results here. In order to drastically reduce the high computational cost of the method we repeatedly sampled a percentage of all tags, computed tag pair correlation coefficients for each subset, and then combined the results of these subsets for all further processing, which allows to process relatively large data sets. When selecting data for processing we introduced a number of additional thresholds, for example to exclude tags that were used on less than n streams.

Initially the method did not produce good results. The Clements method was designed for tagging systems of shared catalogues, where linguistic sub-communities all annotate the same global items or concepts. In our case however only the creator can tag a datastream, and datastreams are never shared between participants.

What is shared among participants however is metadata vocabulary to annotate a datastream, including tags, location names, and units of measurement. And the existence of these globally shared terms do allow us to establish stronger indirect connections between datastreams: if two datastreams employ the same unit of measurement there is some likelihood that their respective tags may be complementary as well. This probability can be quantified as described by Clements, by determining the same correlation coefficients but substituting units for datastreams. In that sense the correlation coefficient does not describe how participants tag streams, but instead how they tag units of measurement. This approach worked very well.

t_q	t_s	$S_U(t_q, t_s)$	$S_I(t_q, t_s)$
`temperature`	`Temperature`	-0.0006410	0.6913
`Humidity`	`humidity`	-0.0001824	0.6486
`Temperature`	`temperature`	-0.0006252	0.6943
`Temperature`	`temp`	-0.0002286	0.6260
`Temperature`	`Temp`	-0.0002505	0.6637
`temperature`	`temp`	-0.0002734	0.7393
`temperature`	`degrees`	-0.005491	0.6354
1 時間移動平均	`Counts Per Minute`	-0.005332	0.6235

Table 3.10: The Pearson correlation coefficients for the user similarity S_U and item similarity S_I of some query tags t_q and the potential synonyms t_s. This table shows some synonymous tag pairs matching the correlation thresholds suggested by Clements: $S_U < 0$ and $S_I > 0.5$. Note that false positives are still present, but also that several variant spellings and capitalizations are captured. Also note that "1 時間移動平均" labels a one-minute moving average, and not a count per minute. Both are used in the measurement of radiation.

t_q	t_s	$S_U(t_q, t_s)$	$S_I(t_q, t_s)$
temperature	three	-0.01187	-0.0009531
temperature	sensor:model=SBM-20	-0.01187	-0.0009531
temperature	coffee	-0.04724	-0.0006516
temperature	celsius	0.03535	0.6179
temperature	CO2	-0.02378	-0.0008746
temperature	pressure	0.03959	-0.001430
temperature	psychogeography	-0.01455	-0.0009042
three	Master Bedroom	-0.0007147	-0.0007148

Table 3.11: Some tag pairs that did not meet the Clements thresholds for synonymity. There still are some false negatives in this data set.

Term pairs that were wrongly identified as synonyms are mostly at least topically related, and their relative volume can be controlled with a number of thresholds for tag inclusion. It must however be pointed out that this approach, while it works well, heavily reduces the tag vocabulary from a few thousand down to low hundreds of tags since most tags have only been used once, as demonstrated in Section 3.2, and thus cannot form strong connections with nearby terms. If there is not enough social overlap in the use of rare mis-spellings then we are unlikely to be able to identify them.

The Clements method of identifying tag synonyms offers a trade-off between vocabulary size, computing cost, and relative quality of the result. To assess whether this distance measure can be used to identify groups of synonymous tags three variants were inspected in more detail, and their tag distance graph was visualised in Gephi. These visualisations can be seen in Appendix D.1 on page 75.

Community Detection

In order to develop a good approach for clustering or community detection among the tag pairs of potential synonyms a number of data properties have to be considered:

- The optimal number and size of term clusters is not known and will differ between subsets of the data.
- The network of term relationships has a clear structure, as indicated by the term distance graph visualisations in Appendix D.1. Connections between topically related terms form cliques.
- There are many isolated tags that cannot be meaningfully linked to a group since not every tag has enough contextual information associated with it.

There are many approaches to data clustering and tag partitioning, and a number of these specialise on the inclusion of network topology as a cluster criterion. A fairly recent family of methods aims to identify term communities in such a graph, and Blondel and Guillaume (2008) describe a heuristic that is particularly suitable for large data sets.

The Blondel community detection method is implemented in Gephi as a graph statistic called "Modularity", this was applied it with default parameters to variant B of the tag synonym graphs and the unit synonym graph, the key clusters were extracted, and some badly-fitting terms were manually pruned. Visualisations of these are presented in Appendix D.2 on page 79. These term groups are used to highlight a number of key activities of the Cosm community in the following chapter.

`1 hour average`	10 分移動平均	10 分間移動平均
1 時間移動平均	`Counts Per Minute`	`CPM`
`cpm`	`microsieverts`	`microsieverts/hour`
`Radiation`	`radiation`	`radiation sensor`
`SBM-20`	`sensor`	`sensor:model=lnd-712`
`sensor:model=SBM-20`	`sensor:type=radiation`	`test`
`μSv/h`	1 時間移動平均	

Table 3.12: For radiation-related tags the Blondel community metric worked very well as a topical grouping, though not necessarily as a grouping of perfect synonyms.

3.4 Summary of Results

To conclude this chapter we evaluate the sensor data of a number of exemplary sensing activities as identified by our term clustering approach. Term synonym lists were generated, attributed to particular groups of sensor activity, and then used as selection criteria to extract sensor data. One of the identified groups, sensors of temperature measurements in degree celsius, was then compared with an equivalent high-quality data set. The aim was to identify groups of sensors that may be suitable for building large-scale spatiotemporal models of the observed phenomena.

Data Selection and Evaluation

The methods presented in previous sections yielded synonym lists of datastream tags and units, these were used to identify groups of sensors for particular sensing activities, and to extract the measurements produced by these sensors.

For example the tags `Air Temp`, `sensor:type=temperature`, `température`, and others were identified as synonymous, and the units `°C`, `Celsius Degree`, `Grados C`, and more. In cases where several term lists were generated for apparently equivalent activities, these were merged into a single term list. The final term lists are presented in Appendix E.

In cases where both tag and unit term lists were found for particular activities, such as the two temperature examples just shown, both were used as filter criteria to select a more specific set of datastreams. However where such a selection resulted in a comparatively small number of sensors, only tag lists or only unit lists where used as selection criterion, whichever yielded the larger number of sensors.

The sensor data produced by such activity groups was then assessed in a number of qualitative and quantitative ways in order to establish whether Cosm sensor data identified in such manner is suitable for building integrated spatiotemporal models. The focus was on three aspects: data volume over time, data homogeneity, and geographic distribution.

In this chapter we present a brief summary for each key activity group. Appendix F provides descriptive statistics and visualisations which support the presented assessment.

Energy Usage and Power

Energy monitoring is among the most popular activities on Cosm, and the data shows a steady growth in activity: the number of energy usage sensors approximately doubles from 417 sensors in August 2011 to 725 sensors in April 2012. Many of them are clearly labelled as energy consumption sensor, but most of the datastreams are missing any additional tags that describe the particular sensing context. Almost none of the datastreams in this group have any environment tags specified.

Among the few contextual tags that provide background information are some that refer to Current Cost monitoring stations. These are low-cost home energy use monitoring devices that with Internet connectivity, and some of these devices can publish their data to Cosm (Current Cost, 2012). In principle such popular devices could provide large volumes of consistently annotated monitoring data, but in practice annotations change between device generations and are often not complete, and consequently it is not easily possible to specifically select data produced by these devices.

The most significant share of sensors reports energy usage in watt, often annotated both as a unit of measurement and in the form of a `watts` datastream tag. There is a great range of measurement values across sensors, with the strongest band of measurements in the low thousands of watts, but a number of high outliers. On a number of occurrences at least one sensor reports negative energy usage.

More descriptive statistics and visualisations are provided in Appendix F.1 on page 84.

Radiation

The data shows clear growth of radiation data over time, but also indications that sensors are frequently turned on and off in large groups. This potentially indicates concerted efforts, or at least individuals operating large numbers of sensors.

The radiation activity group identified here is noteworthy for a comparably rigorous use of annotations, including the prominent use of machine tags. The `webscrape` tag in August and September appears to indicate that some data is not from a primary source, but republished from elsewhere. A number of streams have been tagged with device identifiers such as `sensor:model=lnd-712` or `sensor:model=sbm-20`.

A number of different units of measurement feature prominently, and the most frequently used units change between months. In later months `cpm` and `microsieverts/hour` are most popular.

In early 2011, shortly after the Fukushima reactor disaster in Japan, Cosm became a popular meeting ground for concerned citizens gathering radiation data and publishing it online. A number of project groups emerged that attempted to coordinate volunteers and provide guidance in the acquisition of this data. Data gathering practices ranged from the republishing of governmental data to the coordination of large numbers of volunteers by highly skilled specialists, and a range of practices to monitor and increase data quality. Other project groups gathered this data and produced visualisations. The Cosm company blog provides references to many of these projects at Haque (2011), and speaks of “hundreds of radiation-related feeds from Japan”.

As the data shows in late 2011 many of the initial sensor feeds were no longer active, and many of the project groups ceased to operate. But there are clear bursts of new activity, particularly in Japan, but also in Spain (Haque, 2010).

More descriptive statistics and visualisations are provided in Appendix F.2 on page 88.

Humidity

Comparably few humidity sensors could be identified in the observed period. As seen for the other groups, data volume is steadily increasing. About half the identified sensors are placed outdoors. More than 50% of humidity sensors do not have any environment tags (110 of them in April.)

With a few exceptions the sensors generally measure relative humidity, but as many as 20% of sensors (29 in August) do not report a unit. Most sensor values are below 100, as is expected for relative humidity measurements. The coefficient of variation is low, particularly in April, which indicates lower variance in measurement values than those of several other groups.

More descriptive statistics and visualisations are provided in Appendix F.3 on page 92.

Pressure

The selection criteria for this group encompass a number of different measurement units such as `Hectopascal`, `millibar`, and `PSI` (refer to page 82 for the full list), and the selected data could not be used to build integrated models without first converting it to a common reference model.

This group was included regardless in order to determine the purposes to which air pressure is being measured, and to what extent they are made clear in sensor annotations. Unfortunately the metadata gives little context beyond the fact that most sensors appear to measure atmospheric pressure.

There is a consistent almost even split between measurements in pascal and in bar. Because of the use of a number of different measurement scales, variance plots show several bands of activity. They also indicate that these pressure sensors tend to be stable in their reported measurements, there are not many large fluctuations in values.

Many sensors have no environment tags, and those that do frequently have generic tags, or even tags that are not directly related to air pressure, such as `temperature`. Datastream tags employ a number of synonymous terms for `atmospheric pressure`, but may also include device names such as `bmp085`, which is a barometric pressure sensor popular among DIY practitioners.

More descriptive statistics and visualisations are provided in Appendix F.4 on page 96.

Temperature

The temperature group has been selected for a case study to determine how Cosm data compares with high-quality data from other sources. The data is very suitable to build integrated spatial models: in several respects temperature sensing is the most popular activity, so that even fairly constrained selection criteria still yield a comparably large group of sensors. Additionally temperature is an observable phenomenon that changes comparatively little over large spatial distances, which means only a low number of sensors is needed to produce meaningful spatial models.

To prepare the production of such a model the group selection criteria that were generated from the data have additionally been refined manually: several temperature groups had been identified, these were manually merged into one. Additionally all references to measurement scales other than Celsius have been removed in both tag and unit term lists. For the final list of terms refer to Appendix E on page 82.

The daily number of sensors increases by 150% over the observed period. In April, 75% of streams have been tagged with the same `temperature` tag. Most sensors do not provide any environment tags.

Data variance plots clearly indicates that there is a wide range of purposes for temperature measurements: a small number of sensors report temperatures of hundreds of degrees, although most stay at temperatures below 100. The coefficient of variation is fairly high, and the mean value reaches around 45 degrees in August and more than 8,000 in September. As the number of sensors increases the mean temperature value falls to around 26 degrees in April, which is also the month with the lowest variance.

This indicates that in order to make this data comparable with official weather data it needs to be filtered even further.

More descriptive statistics and visualisations are provided in Appendix F.5 on page 100.

Case Study: Comparing Sensor Data

The Cosm temperature data set was considered to be of sufficiently high volume to subject it to more detailed study. It was compared with high-quality data published by the Met Office, the national weather service of the United Kingdom. The aim was to assess to what extent community data published on Cosm is already suitable for building integrated large-scale models of environmental phenomena.

Modelling Environmental Phenomena

In order to model temperature data based on a limited number of sensor points it is necessary to understand the spatial variability of temperature measurements. Sluiter (2009) mentions a number of factors affecting meteorological models such as "*land-sea gradients, altitude, rain radar and yearly trends of environmental factors like circulation patterns and land-use*". Most of these aspects are not captured by Cosm metadata.

Tveito et al. (2008) provide basic guidance on the spatial distribution of some environmental and meteorological phenomena. They observe that temperature is spatially homogeneous and subject to regular seasonal and diurnal cycles, which makes it one of the most easily modelled phenomena. (Tveito et al., 2008, p. 72)

Because of this spatial homogeneity, spatial interpolation models for temperature can be based on a relatively small amount of measurements. Lennon and Turner (1995) evaluate several types of interpolation models to predict the spatial distribution of temperature on the British mainland and conclude that "*a minimum of just over 30 temperature recording stations would generate a satisfactory surface, provided the stations were well spaced.*"

Spatial interpolation can provide quantitative estimates for the distribution of spatial phenomena based on a limited number of distributed measurements, and such predictive methods can be used to build environmental models from any spatially autocorrelated data set where, according to Tobler's first law of geography, "*near things are more related than distant things*" (Tobler, 1970). Atkinson and Lloyd (2009) and others provide a general introduction to the modelling and analysis of spatial phenomena using such statistical approaches.

Met Office Temperature Data

Met temperature data is collected by sensor stations in the UK and comprises hourly temperature measurements, along with a number of other weather phenomena such as humidity, wind direction and strength, and others (UK Meteorological Office, 2012). According to the data, the Met Office station network encompasses around 30,000 stations worldwide (BADC, 2012a). Station locations were derived from BADC (2012c).

The data is highly structured, and annotated with quality control information. Additionally all data is tested for correctness and for consistency with surrounding data points, both manually and by automated processes, and corrections are made where needed. All these steps are recorded in some detail and published along with the resulting temperature data (BADC, 2012b).

Around 300 Met weather stations provided the subset of UK data that was chosen for comparison.

Cosm Temperature Data

The Cosm temperature data set is derived from the temperature sensor group presented above. It does not necessarily contain all suitable sensors in the Cosm archive, but merely those which could be identified from metadata annotations using the automated process presented in Chapter 3.

The preceding section has revealed that this sensor group warrants even stricter selection criteria: many sensors appear to measure hot systems of hundreds of degrees, not necessarily air temperature. As a result there are large fluctuations in values, and a large coefficient of variation. In order to address this before comparing against temperature data provided by the UK Met Office the data has been further processed to only include sensors from “outdoor” environments.

After these additional constraints were introduced only 9 Cosm temperature sensors remained that fit all selection criteria: they had a datastream tag that signified temperature measurements, their unit of measurement was stated in degrees celsius, their location was within the borders of the United Kingdom, and they were situated outdoors.

This low number of sensors makes it doubtful that any comparison between the two data sets would produce a meaningful result. However the preceding section has shown that while less strict selection criteria would have resulted in a somewhat larger data volume, their measurements would have covered a large range of very different phenomena, including hot systems of hundreds of degree celsius.

Comparison

A two-hour window of time on 30 April 2012, from 12:00 noon to 14:00, was selected for both data sets, sensor values were averaged per station, and ArcMap was used to build spatial interpolation models for each data set using ordinary kriging. The semivariograms and cross-validation results for each data set are provided in Appendix G on page 104.

As was expected from the low volume of Cosm data the comparison was not successful: the Cosm data was not only too sparse to build a large-scale model, it was negatively spatially autocorrelated. As a result it was not possible to build a temperature model using kriging, and any further comparison was aborted.

Chapter 4

Discussion

4.1 Findings

Range of Activities

The prominent sensor activity groups identified in this study illustrate that many of the Cosm community activities relate to either personal interests and personal contexts, or larger contemporary issues of shared concern. There is a strong community interest in environmental monitoring and home monitoring, and certain activities produce a large volume of sensor data by hundreds of sensors. This includes measurements of energy usage, temperature, humidity and air pressure levels, rainfall, wind direction and strength, sunlight, and many others.

Radiation measurements have seen a marked increase in activity in 2011, particularly in Japan, but also in large parts of Europe and in other regions. They can be considered a first strong indicator that there is interest in explicit collaboration around the capturing of environmental data, particularly when such activities address shared concerns.

In the context of home monitoring there is a particular focus on the measurement of home energy usage, often using devices with Cosm integration such as CurrentCost energy monitors. The use of such popular monitoring systems introduces an opportunity for consistent metadata annotations, although in the case of CurrentCost it appears annotations practices are not consistent between different versions of their devices.

Additionally Cosm is used for many forms of systems monitoring: voltages, CPU usage, solar panel yield, gas levels, light monitors, and others, and there are a number of official demo applications such as the Cosm mood map and twitter stats that do not require hardware sensors to participate.

The geographic distribution of activities was discussed in Section 3.1, there are particular hotspots of activity in many areas of Europe, but also in Japan, the USA, Australia, and elsewhere.

Heterogeneity

In Cosm sensor metadata there is particularly consistent metadata terminology for basic phenomena such as temperature, humidity, energy use, and radiation levels, and as a result it is possible to establish clear topical relationships between many individual sensors. In the case of radiation data such annotation consistency is frequently a result of explicit collaboration, and this may be true for other types of sensor data. However most of the remaining annotations of Cosm sensor data are very heterogeneous, and often vague or incomplete.

There are many possible ways of describing the same concepts with different words, and many causes for the existence of term synonyms, including the use of different languages, the use of multiple standard notations or abbreviations, but also spelling errors, and others. As much as 75% of all unique tag terms have only been used once, and the term frequency of datastream and environment tags can be shown to follow a power-law distribution.

Furthermore many details of the sensing context are generally not known, such as the type and specific placement of a sensor. This means that very few general statements can be made about the relative quality of sensor measurements.

We have demonstrated that for some basic activity groups data variances can be described and compared using a coefficient of variation, but so far we have not offered a more thorough and exhaustive assessment. For example it was found that measurements in “Dosage Rate”and “Humidity”sensor groups are fairly stable across a large number of sensors, whereas “air quality index”and “ozone”sensor groups show great variance both for individual sensors, and across sensors (Table C.5).

Data Integration

The encountered heterogeneity in sensor annotations poses a key challenge when attempting to integrate the outcomes of these DIY sensor activities. The power-law distribution of tag terms implies that particular tags alone cannot suffice to meaningfully group all datastreams by shared activity, as this would not capture the majority of the data. At minimum it becomes necessary to attempt to identify synonymous tags, variant articulations of semantically equivalent concepts, in order to aggregate larger numbers of sensor data feeds.

It is also necessary to identify the specific nature of the measured phenomenon, and the physical properties that are observed by a sensor. The measurement of “air quality” can mean a number of things: it could be described in terms of ozone levels, particulates of varying sizes, nitrogen oxide levels, nitrogen dioxide levels, sulphur dioxide levels, and in other ways. Many Cosm feeds have not been annotated with a sufficient degree of detail to make such distinctions. Even the scale and units of measurement can differ within groups of sensors that observe the same physical property.

Despite all these challenges we have demonstrated that the identification of groups of activity is possible. We present a heuristic to identify tag and unit synonyms based on term use, and a method of identifying groups of activity from the graph of all term synonym pairs. Many of the annotations are currently too sparse to establish clear links between all key activities, but as the Cosm community grows the automated identification of many more distinct groups of sensor data and sensing activities will become possible.

There are a number of ways in which our current results can be improved. A number of suggestions are provided in Section 4.3.

Case Study

The overall size of the Cosm sensor catalogue, at around 70,000 nodes, already competes with the size of the Met Office station network which aggregates data from 30,000 stations worldwide. In addition the Cosm community is experiencing a period of significant growth, and it does not seem infeasible that within a few years it may reach a multiple of its current size.

However our first evaluation of a subset of Cosm data against an equivalent high-quality data set revealed the big impact of inhomogeneous annotation practices. After applying enough constraints to make the Cosm data sufficiently specific the remaining data volume was very low. It became clear that the general quality of annotations is not good enough to support such detailed constraints. Additionally the data showed negative spatial autocorrelation, indicating that either sensors were not calibrated correctly, or that they were measuring different phenomena.

This result does not put in question that there may be a potential ability of such community sensor activities to yield data suitable for building large-scale spatiotemporal models, but it clearly indicates that such data aggregation is only possible when the metadata supports it, and currently Cosm metadata is too heterogeneous. We provide a number of further recommendations in Section 4.2.

4.2 Recommendations

User Interfaces and Guidance

During our review of Cosm metadata it emerged that a number of essential metadata attributes are provided as free-form text fields. For a number of attributes this is expected and provides great freedom of annotation. However it also limits the degree to which such data can later be aggregated into coherent groups.

Were Cosm to provide means of selecting a particular unit of measurement during sensor setup from a set of standard notations, as considered in the EEML standard, it would greatly increase consistency of annotations. This would not impede the general ability of additionally offering free-form text entry.

Similar techniques could be employed in the use of tags. While there is a wide diversity of sensing practices, a number of these are popular enough that a limited set of standard terminology could describe them in terms that are meaningful and useful to participants, as well as establish clear links between related sensors. Such standard terminology could be determined from observing community vocabulary or through other means.

In these aspects Cosm is presented with an opportunity to develop similarly "opportunistic" models of large-scale environmental monitoring as discussed by Campbell et al. (2008). If user interface changes are made that encourage more homogeneous data annotations it will be possible to integrate the public data of more volunteers without having to request their explicit participation, and without demanding further work on their side.

In order to clarify the semantic meaning of certain annotations there needs to be clearer guidance in the distinction between environment and stream tags. In principle environment tags serve the purpose to describe the broader sensing context, and stream tags describe the observed phenomenon and the sensor technology. In practice these distinctions are not made by practitioners, and frequently environment tags contain metadata that are only true for some streams within the environment. As a result environment tags cannot be used as simple selection criterion.

If communal efforts to gather large amounts of sensor data are to yield meaningful quantitative results it also becomes necessary that practitioners gain a deeper understanding of the many pitfalls of environmental sensing. This could happen in the form of online discussion fora and community meet-ups where practices are discussed and refined, in the development of DIY-friendly sensing practices and sensor devices by professionals and scientists, as a collection of manuals that help newcomers set up their first sensors while educating them about some important details of the systems and observed phenomena, and in other ways.

Channels for Collaboration

In order to address particular problems in a collaborative manner there need to be clear incentives to pool efforts, and these need to be articulated well. In some cases they may be inherent in the problem. Individual radiation measurements cannot be used to establish radiation levels over large areas, and consequently a number of radiation sensing practitioners started to pool their efforts.

Such collaborations often work best when there are clear channels to guide the work of participants. Section 3.4 highlighted a number of examples of the forms this can take. Such channels were frequently set up by individuals or organisations who were already sufficiently familiar with the observed phenomenon and its characteristics so they could provide guidance and establish organisational structures that made it easier for newcomers to participate.

In other cases incentives are created by the systems that are used to capture information, particularly where this can be done with little effort. This is evident from the large number of Cosm environments that have been annotated with geo coordinates as was shown in Section 3.1. Since such annotations have larger communal effects, for example they allow to group the many distributed sensor efforts, conscious annotation practice can be regarded as a less explicit form of collaboration.

It is however also evident that currently not many structures exist that encourage Cosm community participants to work together more explicitly. Practitioners may be interested in shared mapping and data gathering projects, or in shared dashboards that present data gathered by related sensors. These may also serve as a basis for the communal assessment of sensor output, especially when combined with an ability to suggest changes in sensor metadata annotation. Such public aggregation dashboards could provide the means to curate selections of sensors and annotate them further.

With a few exceptions most Cosm data gathering activities are not currently sufficiently developed to have strong means of assessing and asserting data quality. It may however be possible to develop general community mechanisms for such purposes. A variety of general quality control mechanisms have been introduced in the literature review, and surveys as provided by Wiggins and Newman (2011) and others can serve as a source of recommendations.

4.3 Future Research

A large amount of time was spent on the acquisition and preparation of the data which left comparatively little time for data analysis. We will outline a number of additional analyses that could yield particularly interesting results for this data set. First we will outline a number of potential improvements to the data integration techniques that were presented.

Basic Improvements

The ability to determine synonymous expressions in annotation vocabulary is an important prerequisite for identifying shared activities amongst a highly heterogeneous data set such as this, and consequently much time was spent on establishing some initial techniques, as outlined in Section 3.3. More time can be spent on tailoring these to particular characteristics of the data, for example to establish clearer thresholds for data inclusion when building particular synonym lists. Sensible thresholds for a minimum number of tags per item, user per tag, and so on vary by context and by data set.

Much may also be gained by identifying faster means of synonym detection that can operate on larger data sets without a need for sampling techniques. Some of the research on tag similarity measures offers further insight on comparable tag distance metrics. For example Cattuto and Benz (2008) find that the co-sine tag similarity measure is best at identifying synonyms, and Begelman (2006) and others offer further research into tag similarity measures and clustering approaches.

Other clustering approaches could yield further improvements. There is a rich body of research on clustering and graph community detection, graph partitioning, spectral clustering, cut-based graph clustering and other techniques, each addressing different aspects of the general problem of group detection. An example is provided by Mishra and Schreiber (2007), but there are many others.

A key limitation of the current approach to term grouping is that it cannot differentiate between regional terminology and regional practices, for example in the use of distinct units of measurement where Fahrenheit and Celsius units are classified as synonyms. Spatial segmentation of the data may help uncover such regional differences. Language-specific dictionaries and term databases may also assist in this, provided they can cover the highly domain-specific language encountered here. Alternatively such measurements could be converted to a shared reference model.

It may prove fruitful to attempt similar comparative studies as presented with temperature data in Section 3.4, with the caveat that many activities reflected in this data set may not produce sufficient volumes of data to make comparisons possible. Depending on the observed phenomenon it may be possible to address this by aggregating data at different temporal resolutions, for example over a period of days or weeks.

Other Grouping Criteria

Other grouping schemes are possible. There may be further indirect connections between sensors such as shared temporal patterns. However it becomes evident that the aggregation of “equivalent” data sets in the presence of such a rich and inconsistent participant-provided annotation scheme is a major challenge. In addition it is likely impossible to evaluate the veracity of most annotations.

Multi-dimensional grouping approaches could be applied that construct term lists not only for terms in isolation, but combinations of annotation attributes that may establish more specific context for rarely used tags.

A further topic for exploration is the automatic detection of concept hierarchies to establish nested relationships between related terms such as “weather”, “temperature”, and “Celsius”. Heymann and Garcia-molina (2006) offer a simple yet effective method for building a navigable hierarchical taxonomy of participant-provided tags based on tag similarity measures, and the iterative construction of a similarity tree.

Once a group of similar sensors have been identified their metadata may assist in finding additional sensors of the same kind. This may particularly be useful to uncover groups of sensors for activities that have no well-established vocabulary for annotations.

Assessing Selection Quality

Evaluations of candidate activity groups may include the review of basic summary statistics such as data volume, data variance, metadata vocabulary, and spatial distribution for each variant, but also a quantitative comparison against a ground truth data set, a high-quality data set produced under known circumstances.

If the aim of grouping sensor is to identify a coherent and consistent data set of a particular phenomenon then spatial autocorrelation measures can serve as test criterion.

For each particular group selection the degree of spatial autocorrelation describes the level of data heterogeneity within that group, under consideration of its spatial structure. This can also contain a temporal component: do nearby feeds follow similar trends over time?

Such means of evaluation can be used to iteratively test refinements in order to support gradual optimisation of selection thresholds and other grouping parameters.

Qualitative Studies

More time can be spent assessing the aims, motivations, collaboration structures, and practices of participants. These may provide insight into the particular nature of the captured data, and may help understand its shortcomings and potential in more detail. At the moment it is rarely possible to assess suitability of practices: does the particular sensor setup effectively address the participant's aims? How much consideration do DIY practitioners put into the selection of their devices, and their calibration and setup?

Correlated Activities

There may be activities which are highly correlated, such as humidity sensor data which is frequently collected along with temperature data. The identification of these couplings, for example based on tag pair frequencies or spatial correlation, may reveal more information about particular sensing contexts or means of sensor data production.

Data Volume and Data Quality

As presented in our literature review on page 17, Haklay et al. (2010) identified a clear relationship between the number of OpenStreetMap participants in particular regions and the quality of the data their efforts yield. It is feasible that similar relationships exist here, particularly in the context of efforts to gather large amounts of environmental data.

Additionally there is an opportunity to increase data quality by encouraging several participants to monitor the same systems. The data produced by individual participants is shaped by their particular sensing context, and their technologies and procedures, much of which are not publicly documented. Can techniques be found that integrate the data of larger numbers of sensors capturing similar locations, particularly to address potential and unknown shortcomings in individual measurements? Are there possible compensations for the variability of sensors, for example by also considering other nearby sensor data that may be available?

Composite Data Sets

The data and techniques presented here could serve to enhance the spatial resolution of other data sets. For example energy providers frequently publish energy usage data only at national or regional scale, and higher spatial resolutions are not freely available. Volunteer home energy usage data, once identified, could be used to build an integrated spatiotemporal model of the system that provides data aggregated at postcode scale, or higher.

Chapter 5

Conclusion

According to the data presented in this study Cosm is predominantly a meeting ground for DIY sensor data enthusiasts and a general-purpose data store, but not a place of strong collaboration. Its public data store and its structured data formats offer a strong basis for distributed sensing activities, and the significant increase in activities and data volume over the period of study illustrates that the platform addresses the needs of many DIY practitioners.

Beyond that, a few rare examples also illustrate its potential as a collaboration platform, be it the explicit and international collaboration around capturing radiation measurements, or the implicit collaboration around home energy use data, where a growing number of distributed participants captures a particular socio-political data set under fairly homogeneous conditions. There currently are few similarly explicit efforts to set up large-scale environmental monitoring systems with large communities of participants.

In Cosm sensor metadata there is particularly consistent metadata terminology for basic phenomena like temperature, humidity, energy use, and radiation levels, and as a result it is possible to establish clear topical relationships between many individual sensors. In the case of radiation data such annotation consistency is frequently a result of explicit collaboration, and this may be true for other types of sensor data. However most of the annotations of Cosm sensor data are very heterogeneous, and often vague or incomplete.

Furthermore many details of the sensing context are generally not known, such as the type and specific placement of a sensor. This means that very few general statements can be made about the relative quality of sensor measurements.

The encountered variability in sensor data annotations poses a key challenge when attempting to integrate the outcomes of these DIY sensor activities. Power-law distributions of tag terms indicate that particular tags alone cannot suffice to meaningfully group all datastreams by shared activity, as this would not capture the majority of the data. At minimum it becomes necessary to attempt to identify synonymous tags, variant articulations of semantically equivalent concepts, in order to aggregate larger numbers of sensor data feeds.

Many of the annotations are currently too sparse to establish clear links between all key activities, but we demonstrate that the identification of groups of activity is possible. This study presents a number of techniques that assist in identifying larger groups of shared sensor activity. Sensor data that has a minimum degree of annotations can allow us to draw conclusions about some missing aspects, and the large and growing volume of public DIY sensor data allows us to apply statistical methods of inference when attempting to establish the nature of a published data set.

We present an approach to synonym detection for tags and units that is based on a method first presented by Clements et al. (2008) who observed that tag synonyms are often the result of distinct groups of participants employing different language for the same concepts, and who present a statistical measure based on Pearson correlations of tag similarity and user similarity. We have modified their method to adopt it to this particular data set, and implemented a sampling-based approach since the computation of term synonymity for a data set of this large size comes at a considerate computational cost.

Our identification of groups of shared activity makes use of a community detection method by Blondel and Guillaume (2008). This method identifies distinct groups of nodes in a large network by their relationship to the topology of the remaining network: there are strong connections between the nodes of a community, and weaker connections between node communities. The basis for our computation are the term graphs of synonymous tags or units of measurement, although other metadata attributes can be used as well.

However our evaluation of a subset of Cosm data against an equivalent high-quality data set revealed the big impact of inhomogeneous annotation practices. After applying enough constraints to make the Cosm data sufficiently specific the remaining data volume was very low. It became clear that the general quality of annotations is not good enough to support such detailed constraints. Additionally the data showed negative spatial autocorrelation, indicating that either sensors were not calibrated correctly, or that they were measuring different phenomena.

Regardless of this finding, within some limitations Cosm could be considered an early articulation of a Sensor Commons. The wide spectrum of sensing activities and the large data volume produced by thousands of participants form a significant public archive of sensor measurements, much of it updated over longer periods and available in realtime. The technical facilities provided by Cosm provide a flexible foundation for collaborative sensor data-gathering efforts.

It was demonstrated that there is increased public interest in these informal sensor data-gathering practices, particularly in the context of contemporary themes such as energy usage, environmental concerns, and climate change. Some recommendations for the provision of better collaboration tools were made, as well as suggestions for improvements to existing data-gathering practices.

It is important to identify the context under which such DIY sensing data is produced in order to determine what can be expected of it, as for example suggested discussed by

Flanagin and Metzger (2008): is the data used for science, or social or political purposes? Much of the data discussed in this context is primarily acquired for personal use, and data quality expectations by producers of sensor data may vary widely.

If more explicit collaboration projects are to emerge from this initial set of personal activities it is clear that they need a well-defined purpose that aligns with the aims and motivations of their participants, and the means at their disposal. This also entails consideration of the degree of knowledge and amount of time participants can contribute to acquire, prepare and monitor their sensors.

Such collaborative efforts may not yield a general replacement for more expensive existing infrastructure, but they may serve as a replacement for particular purposes, just as Wikipedia does not provide an encyclopaedia of reliably vetted information, but instead a catalogue of knowledge that is continually updated and refined. The use of similarly discursive models of information acquisition can be particularly beneficial in negotiating the interests and abilities of a great number of willing participants.

As just one final example of the potential differences in perspective regarding the value of volunteer data we provide a quote by Cosm employee Ed Borden from a blog post on the relative merits and shortcomings of DIY radiation sensor data:

> *Is it more useful to know if the value at a particular location is exactly '.075 microsieverts' or if it has been steadily rising over the past 3 days? (Borden, 2011)*

Bibliography

Armstrong, M. P., Nyerges, T. L., Wang, S., and Wright, D. (2011). Connecting Geospatial Information to Society Through Cyberinfrastructure. *The SAGE Handbook of GIS and Society*, pages 109–122.

Atkinson, P. M. and Lloyd, C. D. (2009). Geostatistics and Spatial Interpolation. In Fotheringham, A. S. and Rogerson, P. A., editors, *The SAGE Handbook of Spatial Analysis*, pages 159–181. SAGE.

BADC (2012a). Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853-current). `http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_ukmo-midas`. [Online; accessed 2 September 2012].

BADC (2012b). Met Office Surface Data Users Guide. `http://badc.nerc.ac.uk/data/ukmo-midas/ukmo_guide.html`. [Online; accessed 2 September 2012].

BADC (2012c). Met Office Surface stations on Google Earth. `http://badc.nerc.ac.uk/search/midas_stations/google_earth.html`. [Online; accessed 2 September 2012].

Begelman, G. (2006). Automated tag clustering: Improving search and exploration in the tag space. *Proceedings of the Collaborative Web Tagging Workshop at WWW'06*.

Bishr, M. (2007). Geospatial Information Bottom-Up: A Matter of Trust and Semantics. *The European information society*.

Blondel, V. and Guillaume, J. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, pages 1–12.

Borden, E. (2011). Crowdsourcing data accuracy. `http://blog.pachube.com/2011/06/crowdsourcing-data-accuracy.html`. [Online; accessed 2 September 2012].

Boulos, M. N. K. (2005). Web GIS in practice III: creating a simple interactive map of England's Strategic Health Authorities using Google Maps API, Google Earth KML, and MSN Virtual Earth Map Control. *International journal of health geographics*, 4:22.

Campbell, A. T., Eisenman, S. B., Lane, N. D., Miluzzo, E., Peterson, R. a., Lu, H., Zheng, X., Musolesi, M., Fodor, K., and Ahn, G.-S. (2008). The Rise of People-Centric Sensing. *IEEE Internet Computing*, 12(4):12–21.

Cattuto, C. and Benz, D. (2008). Semantic analysis of tag similarity measures in collaborative tagging systems. *Arxiv preprint arXiv:0805.2045*, pages 3–8.

Clauset, A., Shalizi, C., and Newman, M. (2007). Power-law distributions in empirical data. (1):1–26.

Clements, M., de Vries, A. P., and Reinders, M. J. (2008). Detecting synonyms in social tagging systems to improve content retrieval. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 739.

Cohn, J. P. (2008). Citizen Science : Can Volunteers Do Real Research ? *BioScience*, 58(3):192–197.

Coleman, D. and Sabone, B. (2010). Volunteering Geographic Information to Authoritative Databases: Linking Contributor Motivations to Program Characteristics. *Geomatica*, 64(1):383–396.

Cosm (2012a). Cosm - About. `https://cosm.com/about_us`. [Online; accessed 2 September 2012].

Cosm (2012b). Cosm - Hardware. `https://cosm.com/support/hardware`. [Online; accessed 2 September 2012].

Cosm (2012c). Cosm Blog. `http://blog.cosm.com/`. [Online; accessed 2 September 2012].

Cosm API (2012a). Cosm API Documentation. `http://cosm.com/docs/v2/`. [Online; accessed 2 September 2012].

Cosm API (2012b). Datapoints - Cosm API Documentation. `http://cosm.com/docs/v2/datapoint/`. [Online; accessed 2 September 2012].

Cosm API (2012c). Datastreams - Cosm API Documentation. `http://cosm.com/docs/v2/datastream/`. [Online; accessed 2 September 2012].

Cosm API (2012d). Feeds - Cosm API Documentation. `http://cosm.com/docs/v2/feed/`. [Online; accessed 2 September 2012].

Cosm API (2012e). Historical Queries - Cosm API Documentation. `http://cosm.com/docs/v2/history.html`. [Online; accessed 2 September 2012].

Cosm API (2012f). Listing all feeds - Cosm API Documentation. `http://cosm.com/docs/v2/feed/list.html`. [Online; accessed 2 September 2012].

Craig, W., Harris, T., and Weiner, D. (2002). Community Participation and Geographical Information Systems.

Cuff, B. D. and Hansen, M. (2008). Urban Sensing: Out of the Woods. *Communications of the ACM*, 51(3).

Current Cost (2012). Reducing your energy bills so you can live a greener life! http://www.currentcost.com/product-netsmart.html. [Online; accessed 2 September 2012].

CWOP (2012). Citizen Weather Observer Program. http://www.wxqa.com/. [Online; accessed 2 September 2012].

Davis, C., Fonseca, F., and Camara, G. (2011). Environmental Sustainability: The Role of Geographic Information Science and Spatial Data Infrastructures in the Integration of People and Nature. *The SAGE Handbook of GIS and Society*, pages 123–137.

Demirbas, M., Bayir, M., and Akcora, C. (2010). Crowd-sourced Sensing and Collaboration Using Twitter. *World of Wireless*.

Ehrhardt-Martinez, K. and Laitner, J. A. (2010). People Centered Initiatives for Increasing Energy Savings. Technical report, American Council for an Energy-Efficient Economy.

Elwood, S. (2011). Participatory Approaches in GIS and Society Research: Foundations, Practices, and Future Directions. *The SAGE Handbook of GIS and Society*, pages 381–399.

Fisher, A. (2011). Towards a Sensor Commons. http://ajfisher.me/2011/12/20/towards-a-sensor-commons/. [Online; accessed 2 September 2012].

Flanagin, A. J. and Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72(3-4):137–148.

Foursquare Blog (2012). foursquare is joining the OpenStreetMap movement! http://blog.foursquare.com/2012/02/29/foursquare-is-joining-the-openstreetmap-movement-say-hi-to-pretty-new-maps/. [Online; accessed 2 September 2012].

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, pages 1–15.

Goodchild, M. F., Guo, H., Annoni, A., Bian, L., de Bie, K., Campbell, F., Craglia, M., Ehlers, M., van Genderen, J., Jackson, D., Lewis, A. J., Pesaresi, M., Remetey-Fülöpp, G., Simpson, R., Skidmore, A., Wang, C., and Woodgate, P. (2012). Next-generation Digital Earth. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1–7.

Gouveia, C. and Fonseca, A. (2008). New approaches to environmental monitoring: the use of ICT to explore volunteered geographic information. *GeoJournal*, 72(3-4):185–197.

Gross, N. (1999). The Earth Will Don an Electronic Skin. http://www.businessweek.com/1999/99_35/b3644024.htm. [Online; accessed 2 September 2012].

Gustin, S. (2011). Democratizing Design: Autodesk's CEO Carl Bass Announces 123D. `http://www.wired.com/business/2011/05/democratizing-design/`. [Online; accessed 2 September 2012].

Haklay, M. M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *Cartographic Journal, The*, 47(4):315–322.

Haque, U. (2008a). Extended Environments Markup Language: EEML. `http://www.eeml.org/`. [Online; accessed 2 September 2012].

Haque, U. (2008b). The EEML XML Schema 0.5.1. `http://www.eeml.org/xsd/0.5.1/0.5.1.xsd`. [Online; accessed 2 September 2012].

Haque, U. (2010). Environment tags, machine tags and device classifications. `http://community.cosm.com/taxonomy/term/199`. [Online; accessed 2 September 2012].

Haque, U. (2011). Crowd-sourced realtime radiation monitoring in Japan. `http://community.cosm.com/node/611`. [Online; accessed 2 September 2012].

Heinzelman, J. and Waters, C. (2010). Crowdsourcing crisis information in disaster-affected Haiti. *Uited States Institute of Peace Special Report*.

Heymann, P. and Garcia-molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems.

ISO/DIS 16739 (2012). *Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries.* ISO, Geneva, Switzerland.

Jaffe, A., Naaman, M., Tassa, T., and Davis, M. (2006). Generating summaries and visualization for large collections of geo-referenced photographs. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval - MIR '06*, page 89.

Jankowski, P. (2011). Designing Public Participation Geographic Information Systems. *The SAGE Handbook of GIS and Society*, pages 347–360.

Kamel Boulos, M. N., Resch, B., Crowley, D. N., Breslin, J. G., Sohn, G., Burtner, R., Pike, W. a., Jezierski, E., and Chuang, K.-Y. S. (2011). Crowdsourcing, citizen sensing and Sensor Web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *International journal of health geographics*, 10(1):67.

Kennedy, L., Naaman, M., Ahern, S., Nair, R., and Rattenbury, T. (2007). How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections Categories and Subject Descriptors. In *ACM Multimedia 2007*.

Kera, D. and Graham, C. (2010). Collective sensor networks and future communities: designing interaction across multiple scales. *Proceedings of the 22nd Conference of the*

Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction, pages 396–399.

Klandermans, B. (1987). Potentials, Networks, Motivations, and Barriers: Steps Towards Participation in Social Movements. *American sociological review*, 52(4):519–531.

Kuznetsov, S. (2010). Rise of the expert amateur: DIY projects, communities, and cultures. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pages 295–304.

Latour, B. (2004). Which protocol for the new collective experiments? *Philosophy*, (March 2001).

Lennon, J. and Turner, J. (1995). Predicting the spatial distribution of climate: temperature in Great Britain. *Journal of Animal Ecology*, 64(3):370–392.

Longhorn, R. (2011). Geographic Information Value Assessment. *The SAGE Handbook of GIS and Society*, pages 441–457.

Malone, T. W. and Klein, M. (2007). Harnessing collective intelligence to address global climate change. *Innovations: Technology, Governance, Globalization*, pages 15–26.

Malone, T. W., Laubacher, R., and Dellarocas, C. (2009). Harnessing Crowds : Mapping the Genome of Collective Intelligence. *Elements*, pages 1–20.

Mankoff, J., Matthews, D., Fussell, S., and Johnson, M. (2007). Leveraging Social Networks To Motivate Individuals to Reduce their Ecological Footprints. *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 87–87.

Maslow, A. H. (1943). A Theory of Human Motivation. *Psychological Review*, pages 370–396.

Masser, I. (2011). Emerging Frameworks in the Information Age: The Spatial Data Infrastructure. *The SAGE Handbook of GIS and Society*, pages 271–286.

Mintzberg, H. (1978). *The Structuring of Organizations: A Synthesis of Research*. The theory of management policy series.

Mishra, N. and Schreiber, R. (2007). Clustering social networks. *Proceedings of the 5th international conference on Algorithms and models for the web-graph*, pages 1–12.

Monroy, M. (2011). New York Maker Faire Showcases a New Generation of Innovators. `http://abcnews.go.com/blogs/technology/2011/09/makerfaire/`. [Online; accessed 2 September 2012].

Neis, P., Zielstra, D., and Zipf, A. (2011). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007-2011. *Future Internet*, 4(1):1–21.

Olson, M. (1965). *The Logic of Collective Action: Public Groups and the Theory of Groups*. Harvard University Press.

Openshaw, S. (1991). Commentary: A view on the GIS crisis in geography, or, using GIS to put Humpty-Dumpty back together again. *Environment and Planning*, 23:621–628.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Ostrom, E. and Nagendra, H. (2006). Insights on linking forests, trees, and people from the air, on the ground, and in the laboratory. *Proceedings of the National Academy of Sciences of the United States of America*, 103(51):19224–31.

Pickles, J. (1995). Representations in an Electronic Age: Geography, GIS, and Democracy. In Pickles, J., editor, *Ground Truth - The Social Implications of Geographic Information Systems*, chapter 1, pages 1–30. The Guilford Press.

Poole, M. S., Hollingshead, A. B., Mcgrath, J. E., Moreland, R. L., and Rohrbaugh, J. (2004). Small Group Research: Interdisciplinary Perspectives on Small Groups. *Small Group Research*, 35.

Rattenbury, T., Good, N., and Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 103.

Reidy, C. (2011). Bringing down the barriers: Pachube service goes free! `http://blog.cosm.com/2011/01/bringing-down-barriers-pachube-service.html`. [Online; accessed 2 September 2012].

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., and Jacobs, D. (2012). Dynamic Changes in Motivation in Collaborative Citizen-Science Projects. In *CSCW 2012*.

Sakaki, T. (2009). Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors. *Earthquake*.

Schuurman, N. (2000). Trouble in the Heartland: GIS and its Critics in the 1990s. pages 569–590.

Sheth, A. (2009). Citizen Sensing, Social Signals, and Enriching Human Experience. *IEEE Internet Computing*, 13(4):87–92.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, 24(9):467–71.

Sluiter, R. (2009). Interpolation methods for climate data literature review. Technical report, KNMI, Information and Observation Technology.

Taylor, P. (1990). Editorial comment: GKS. *Political Geography Quarterly*, 9:211–212.

The Shack Blog (2011). Top Ten DIY Suggestions From YOU. http://blog.radioshack.com/2011/07/21/top-ten-diy-suggestions-from-you/. [Online; accessed 2 September 2012].

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 2(46):234–240.

Tulloch, D. L. (2008). Is VGI participation? From vernal pools to video games. *GeoJournal*, 72(3-4):161–171.

Tveito, O. E., Wegehenkel, M., van der Wel, F., and Dobesch, H. (2008). The Use of Geographic Information Systems in Climatology and Meteorology. Technical Report November, Earth System Science and Environmental Management.

UK Meteorological Office (2012). Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853-current). http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_ukmo-midas. [Online; accessed 2 September 2012].

Välimäki, S. (2011). A Case Study of Openstreetmap. 64.

Wal, T. V. (2007). Folksonomy. http://vanderwal.net/folksonomy.html. [Online; accessed 2 September 2012].

Wang, S. (2010). A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis. *Annals of the Association of American Geographers*, 100(3):535–557.

Warren, S. (2011). The Social Potential of GIS. *The SAGE Handbook of GIS and Society*, pages 67–86.

Wauters, R. (2011). Autodesk Acquires DIY Community Instructables. http://techcrunch.com/2011/08/01/autodesk-acquires-diy-community-instructables/. [Online; accessed 2 September 2012].

Weather Underground (2012). Weather Forecast and Reports - Long Range and Local. http://www.wunderground.com/. [Online; accessed 2 September 2012].

Wiggins, A. (2010). Developing a Conceptual Model of Virtual Organizations for Citizen Science. *International Journal of Organisational*.

Wiggins, A. and Newman, G. (2011). Mechanisms for Data Quality and Validation in Citizen Science. *e-Science Workshops*.

Zielstra, D. and Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. *Methods*, 1:1–15.

Appendix A

The Cosm User Interface

A.1 Dashboard with Charts

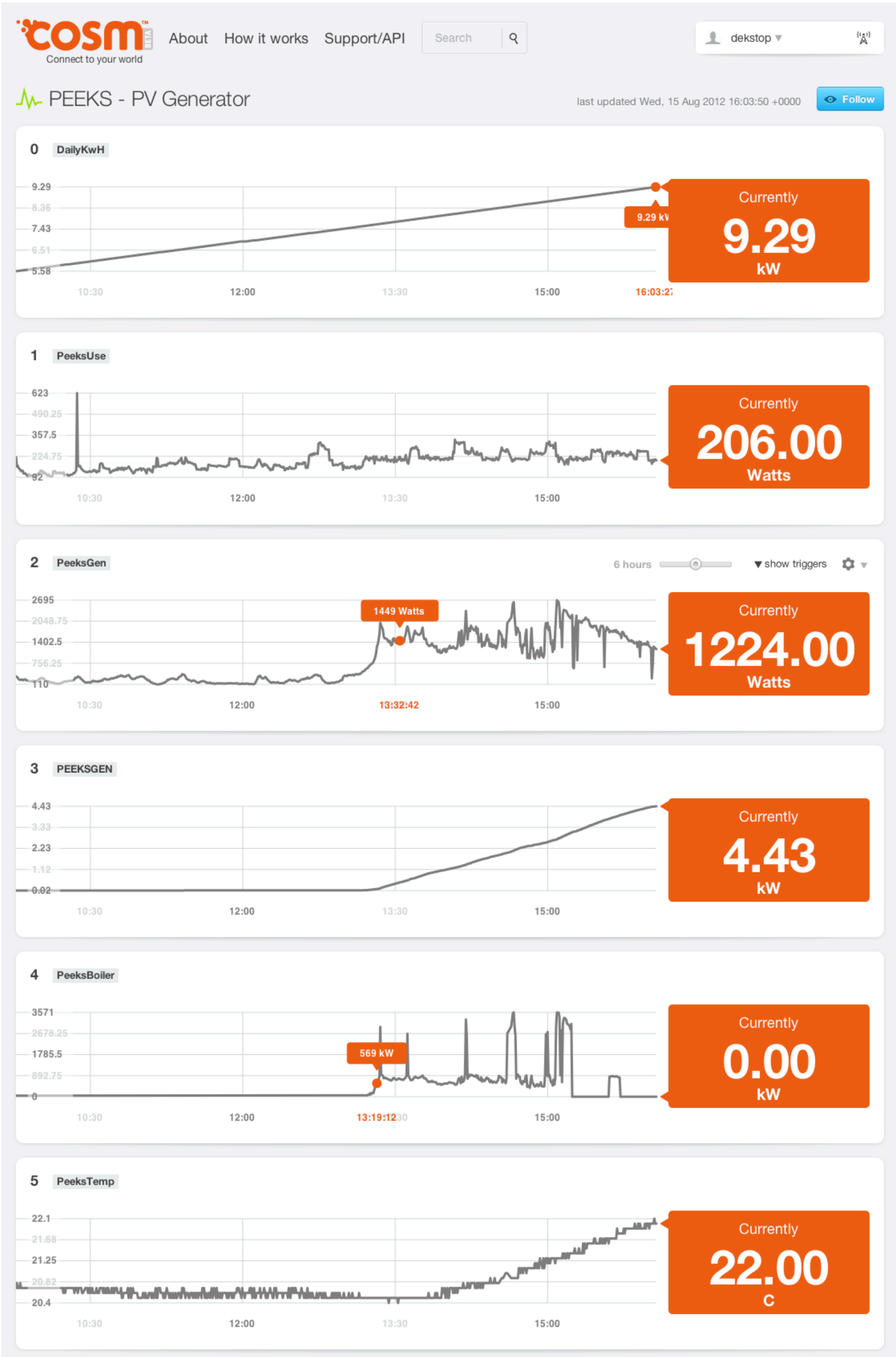


Figure A.1: The Cosm dashboard for an environment, showing charts for all its datastreams.

A.2 Adding New Sensors

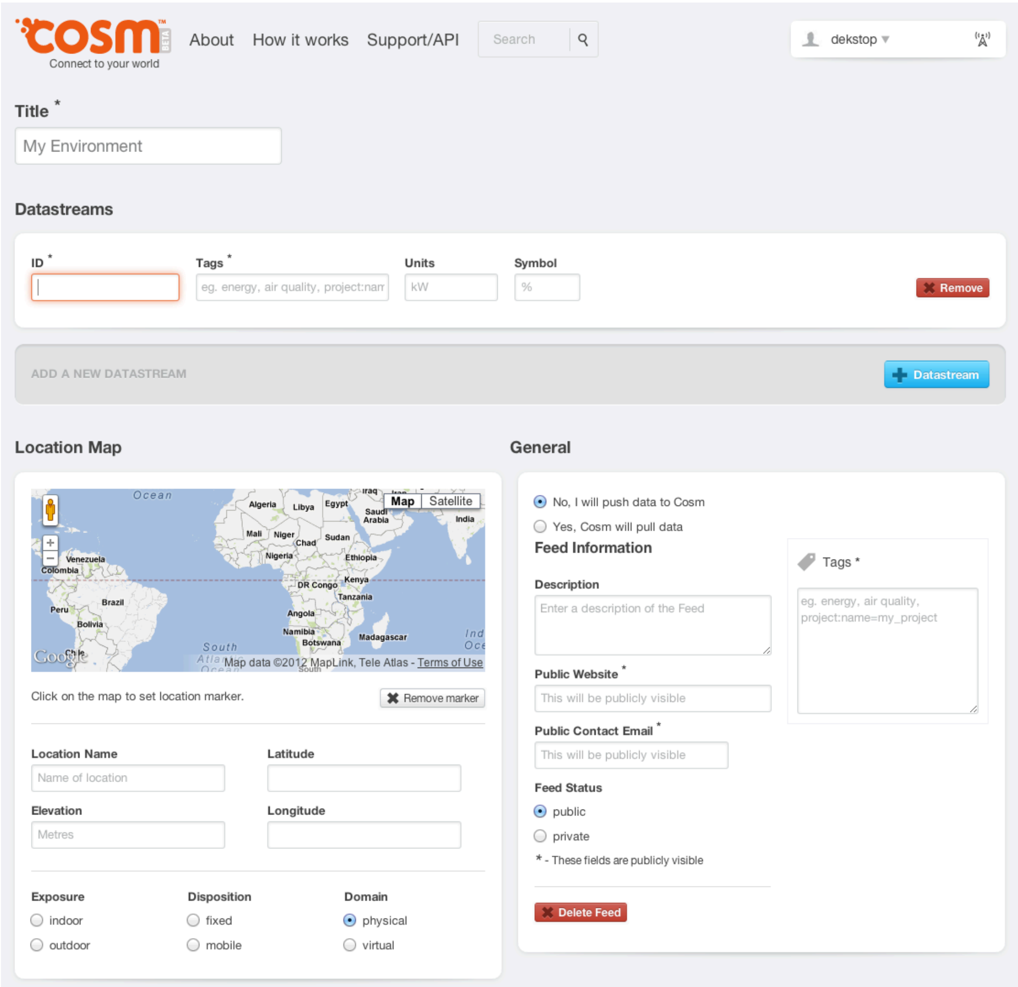


Figure A.2: The Cosm user interface to create a new environment, edit its metadata, and manage its datastreams.

Appendix B

Data Acquisition

The following summarises the software and procedures that were introduced in order to acquire large volumes of Cosm sensor data using the public Cosm API. This includes processes of data acquisition, data cleaning and extraction, and preparation of the data for analyses.

B.1 Cosm API Access

Unless made private by the owner all Cosm sensor data is made public via an Application Programming Interface (API) with facilities to write, update, and read sensor data and metadata in a JSON, XML, or CSV format. (Cosm API, 2012a) Data can be requested for entire environments or for individual datastreams, and API queries can request access to particular historic periods. (Cosm API, 2012e) The Cosm API makes use of the EEML data format in version 0.5.1. (Haque, 2008b)

There are a number of service limitations, among them:

- Per default only 100 data points are returned per request, which can be increased to up to 1,000.
- Data is available in varying temporal resolutions: as the original sequence of sensor measurements and timestamps, or aggregated for times rounded up to the nearest 30 seconds, minute, 5 minutes, and so on.
- These choices of temporal resolution can be requested at different maximum interval sizes, for example sensor data at the greatest temporal resolution can at most be requested for a maximum historic interval of 6 hours at arbitrary points in time. Very active environments can record a few hundred data points in such a period. As a result requesting historic data for a particular datastream necessarily entails breaking up the requested period into batches, and making multiple consecutive API requests.
- Historic API requests are limited to a maximum of one year in the past.
- There is a general limit of 100 API requests made per minute.

B.2 Software Development

As a result of these limiting factors it required a number of attempts to determine an appropriate strategy for acquiring large amounts of data. Considerable time was spent on the software that requests, parses, and cleans this data, then loads it into a relational database.

Over a period of months multiple software components were evolved from a very basic starting point, including:

- Simple shell scripts to retrieve an index of environments and datastreams. This is a fairly basic task that required little control logic.
- An Ruby application to request historic sensor data in an XML format and store them on disk. This needs to make millions of API requests without violating request limits, cope with intermittent networking and API problems, and manage a queue of requests which is scheduled using a relational database.
- Scripts to parse the XML and produce TSV files for all sensor data and metadata.
- A Python application for data loading and database access which reads all TSV data into a relational DB. It needs to process millions of files and gigabytes of data quickly. It also includes a database access layer that allows structured access to the data, which is then used for large-scale analytics queries and data extraction.
- A number of ad hoc shell scripts were used for automation.

The software written for this dissertation is made available on GitHub:

`https://github.com/dekstop/cosm`

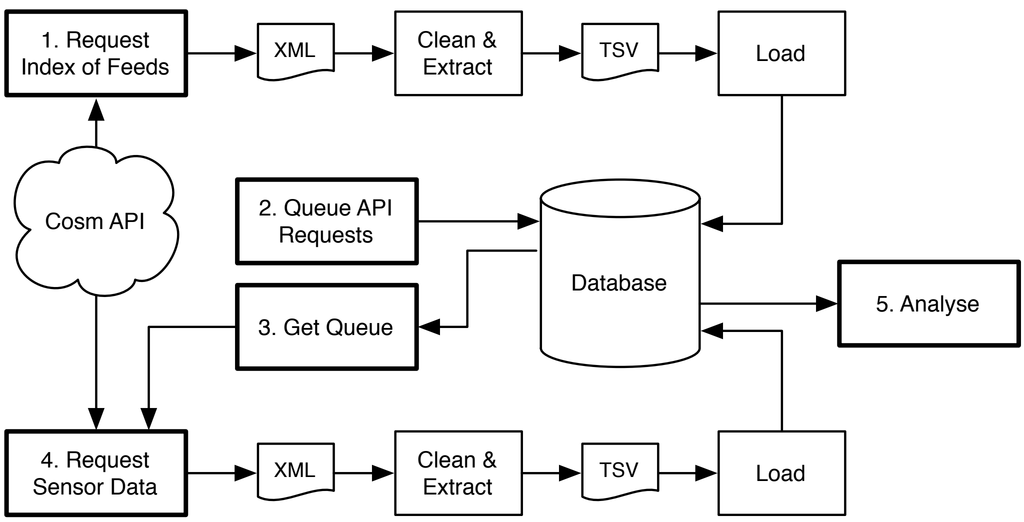


Figure B.1: The data flow chart for the entire process of data acquisition, data cleaning, and data loading. This involves the acquisition of an index of data streams, and the scheduling of API requests for historic data of these streams.

B.3 Outcomes

A number of problems occurred while processing this large volume of data. It soon became clear that the XML documents returned by the API are not always valid, which is a problem frequently encountered with APIs publishing user-provided data. Almost all such problems were in the `<value>…</value>` tags for sensor data – that is, in data produced by devices, not typed by users when setting up their account. Depending on the datastream this field may include control characters which are ASCI characters in the code point range 0 to 32, or undeclared XML entities such as "°" for the degree sign. Such invalid sensor values were all inspected manually and either edited, or in most cases deleted.

At a late stage it also became evident that datastream IDs can change, and that it is virtually impossible to detect the action of renaming a stream. Instead it appears as if an existing stream has ceased to exist and a new stream appeared. The data returned by the Cosm API captures neither the point in time of such a renaming, nor a history of the changed values of a stream identifier, and as a result it becomes impossible to reliably match up the old and new name of a datastream. Even if all remaining metadata of the two versions of such a renamed stream matches it cannot be assumed that they semantically represent the "same" stream. This differs from environments which have both a numeric ID, which is assigned and unchangeable, and a user-provided title.

This caused problems for historic requests. Since all metadata was only extracted from the initial index which represented the state of all environments on 27 May 2012, any earlier versions of metadata states were not captured. This means that the unit of measurement and tags associated with any datastream renamed within the assessed period from August 2011 to May 2012 were lost. This affected about 5% of all datastreams.

A few key data points regarding the data acquisition and extraction:

- An index of all environments and datastreams was retrieved on 27 May 2012.
- This index contains around 70,000 datastreams for 20,000 environments.
- The complete sensor data archive was acquired for August and September 2011, and March and April 2012.
- Data was acquired in 6-hour windows, at maximum time resolution, throttled to the allowed request limit.
- This required millions of API requests which were made over a period of several months.
- The API responses resulted in 16GB of XML files.
- In total the data encompasses around 80 million sensor readings and timestamps.

Appendix C

Data Summaries

C.1 Tag and Unit Popularity

	August 2011	September 2011	March 2012	April 2012
watts	327	353	575	602
temperature	95	90	180	189
humidity	89	85	143	148
power	48	47	73	76
1 時間移動平均	2	22	69	72
radiation	42	42	48	50
wind speed	26	26	38	40
brightness	5	5	9	12
gas	6	6	10	10
failures	3	4	9	9
noise	9	7	4	3

Table C.1: Number of participants contributing sensor data to a datastream tag over time. Labels indicate the peak number of participants in a given month.

	August 2011	**September 2011**	**March 2012**	**April 2012**
`twitter_stats`			268	407
`sensor:type=radiation`	187	193	154	195
`mark2`	5	45	111	114
`cpm`	7	42	104	109
`arduino`	50	50	100	105
線	9	24	53	54
`weather`	18	19	33	35
環境放射線計測	12	19	27	26
`solar`	9	13	25	24
`xbee`	17	15	24	24
`dashboard`	4	4	6	6

Table C.2: Number of participants contributing sensor data to an environment tag over time. Labels indicate the peak number of participants in a given month.

	August 2011	September 2011	March 2012	April 2012
watts	55	408	614	653
followers			268	406
celsius	62	58	127	134
celcius	26	30	47	49
knots	1	30	35	36
percent	16	21	32	33
rh	11	10	14	15
microsieverts per hour	8	11	11	13
lux	5	6	12	12
counts per minute	5	7	7	10
kilograms	105	106	1	1

Table C.3: Number of participants capturing sensor data with a particular unit of measurement. Labels indicate the peak number of participants in a given month.

C.2 Numeric & Non-Numeric Data

	August 2011	September 2011	March 2012	April 2012
Dosage Rate	32 3	34 4	226 0	193 0
temperature	60 2	60 0	179 1	177 5
PM2.5	139 0	142 0	142 0	143 0
1 時間移動平均	2 0	26 0	85 0	90 0
radiation	41 0	40 0	49 0	55 0
power	23 0	22 0	37 2	37 0
Energy	29 0	36 0	33 0	26 0
watts	16 0	17 0	20 0	21 1
Wind Speed	5 0	4 0	8 0	7 0
noise	9 0	8 0	3 0	3 0
last port	0 29	0 29	0 30	0 30

Table C.4: Type of sensor data captured, by number of datastreams. The plots show the volume of both numeric sensor data (above) and sensor data which is not in numeric format (below.) The presence of the latter could indicate data capturing errors, or datastreams capturing textual records.

C.3 Data Variance Plots

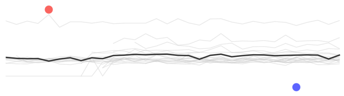
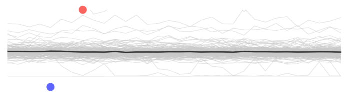
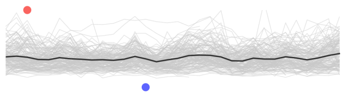
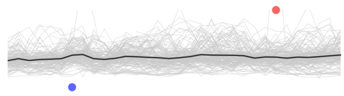
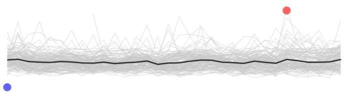
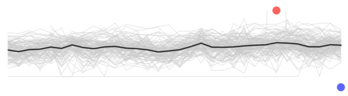
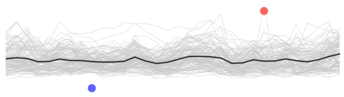
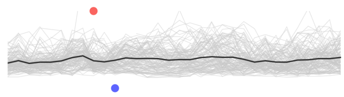
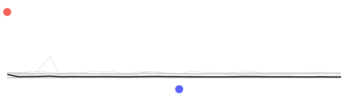
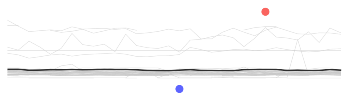
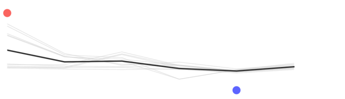
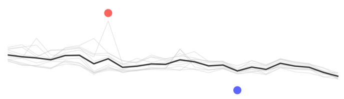
	August 2011		March 2012	
`Dosage Rate`		32 0.109 0.556		226 0.110 0.292
`air quality index`		174 34.67 0.472		175 30.66 0.431
`ozone`		151 27.48 0.417		144 25.85 0.290
`PM2.5`		139 33.94 0.508		142 27.42 0.575
`Humidity`		52 163.8 1.250		67 87.97 1.222
`latitude`		8 48.33 0.067		52 19.21 -0.598
`Price`		11 111.6 0.145		11 98.82 0.144

Table C.5: Variance plots for the sensor values of datastreams grouped with a common tag, for the months of August 2011 and March 2012. Refer to page 32 for a detailed description of this visualisation type. The measures shown to the right of each plot are, from top to bottom: the number of datastreams in this group, the mean of all values in this period, and the mean coefficient of variation over the period. These plots reveal that sensor data sets of different types can have a very different distribution of measurements. Measurements in the `Dosage Rate` and `Humidity` groups are fairly stable across a large number of sensors. The `air quality index` and `ozone` sensor groups on the other hand show great variance.

Appendix D

Term Synonym Detection

D.1 Term Distance Graphs

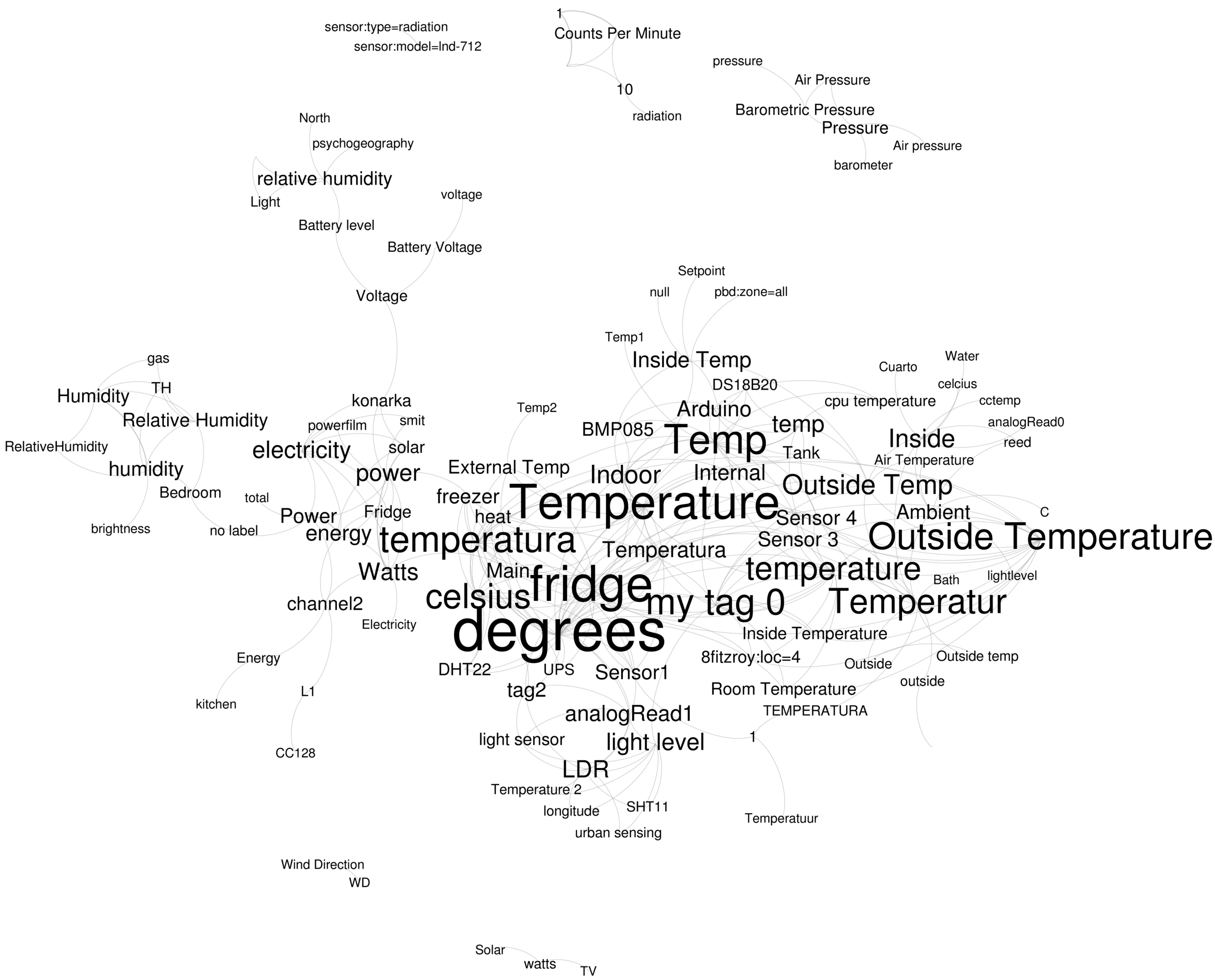


Figure D.1: **Tag Synonym Variant A.** This is a small data set that can be computed very quickly. The computation only includes units of measurement shared between at least 3 users, and tags that have been used on at least 2 datastreams. Based on ten iterations of a 10% sample of all tags. Note that `relative humidity` (top left) and `Relative Humidity` (left) are completely disconnected, both with around the same degree of connections to other tags, and that a `battery/voltage` cluster (top left) is connected to `relative humidity` and other semantically disparate concepts.

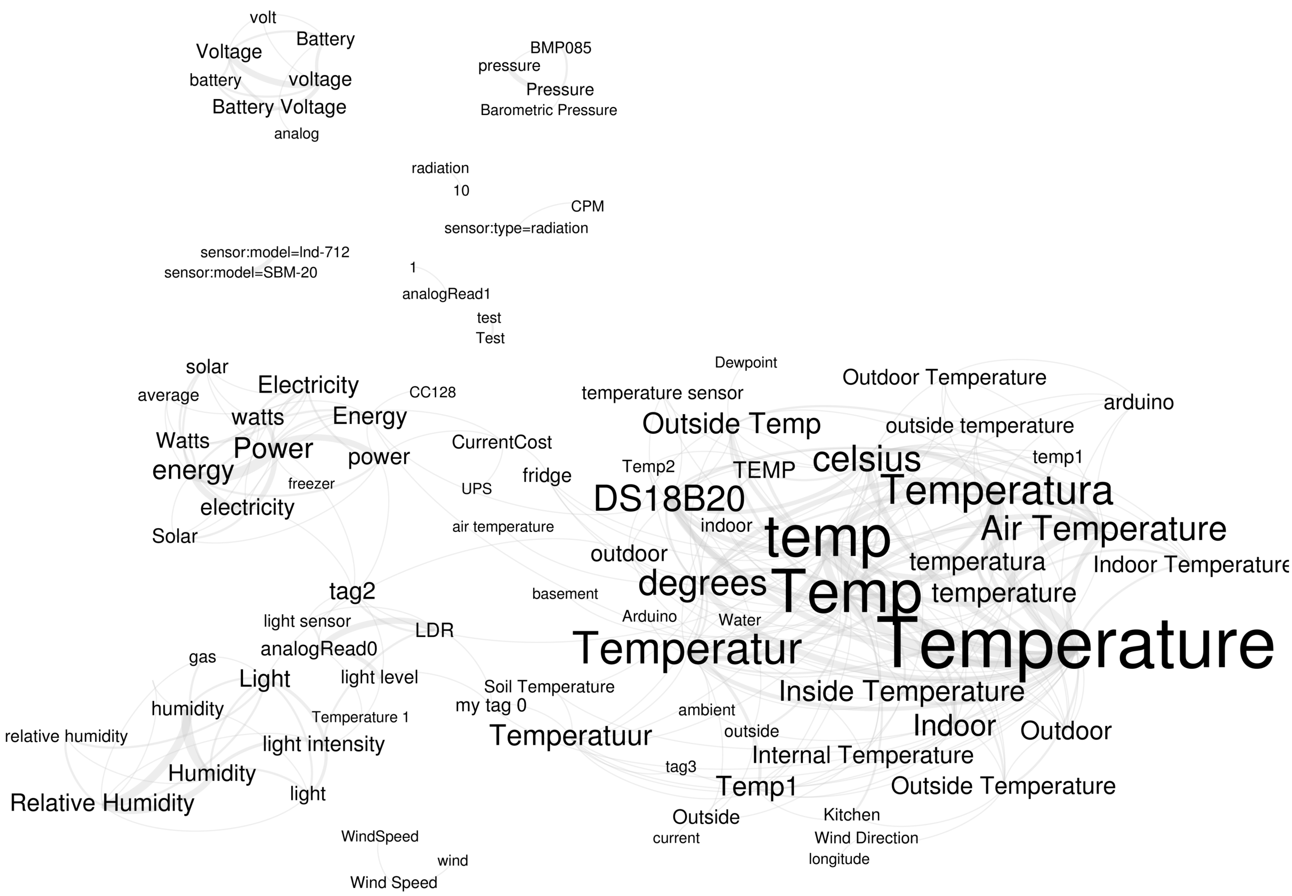


Figure D.2: **Tag Synonym Variant B.** With a larger vocabulary that can still be computed fairly quickly, however it discards a large number of potentially good matches. It includes units used by at least 5 users, tags that have been used on at least 4 datastreams, and is based on 10 iterations of a 20% sample. In this graph `relative humidity`" and `Relative Humidity` are now direct neighbours, and the latter is now much more well-connected. There is also a clearly separate `battery/voltage` cluster (top left) and a new distinct `energy/power` cluster (left).

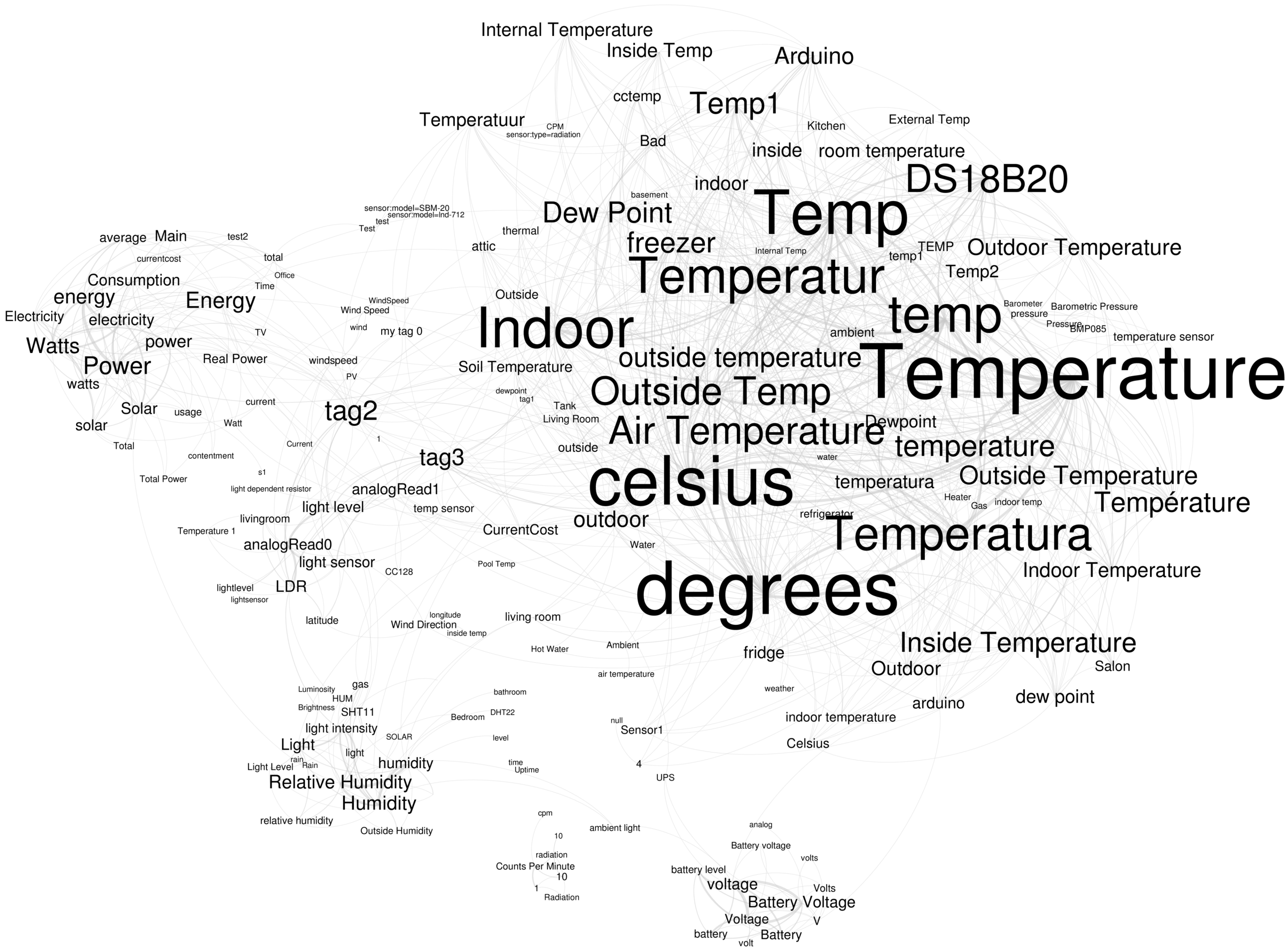


Figure D.3: **Tag Synonym Variant C.** The largest data set version can still be computed fairly quickly, and the resulting term list is the longest without suffering from too many false positives. Data was filtered to include units that have been used by at least 5 users, and tags that have been used by at least two users and applied to at least 3 units, and is based on 10 iterations of a 20% sample. The vocabulary is much richer than before without losing strong existing associations. The `energy/power` cluster (left) gained nuances: `Consumption`, `currentcost`, and other terms.

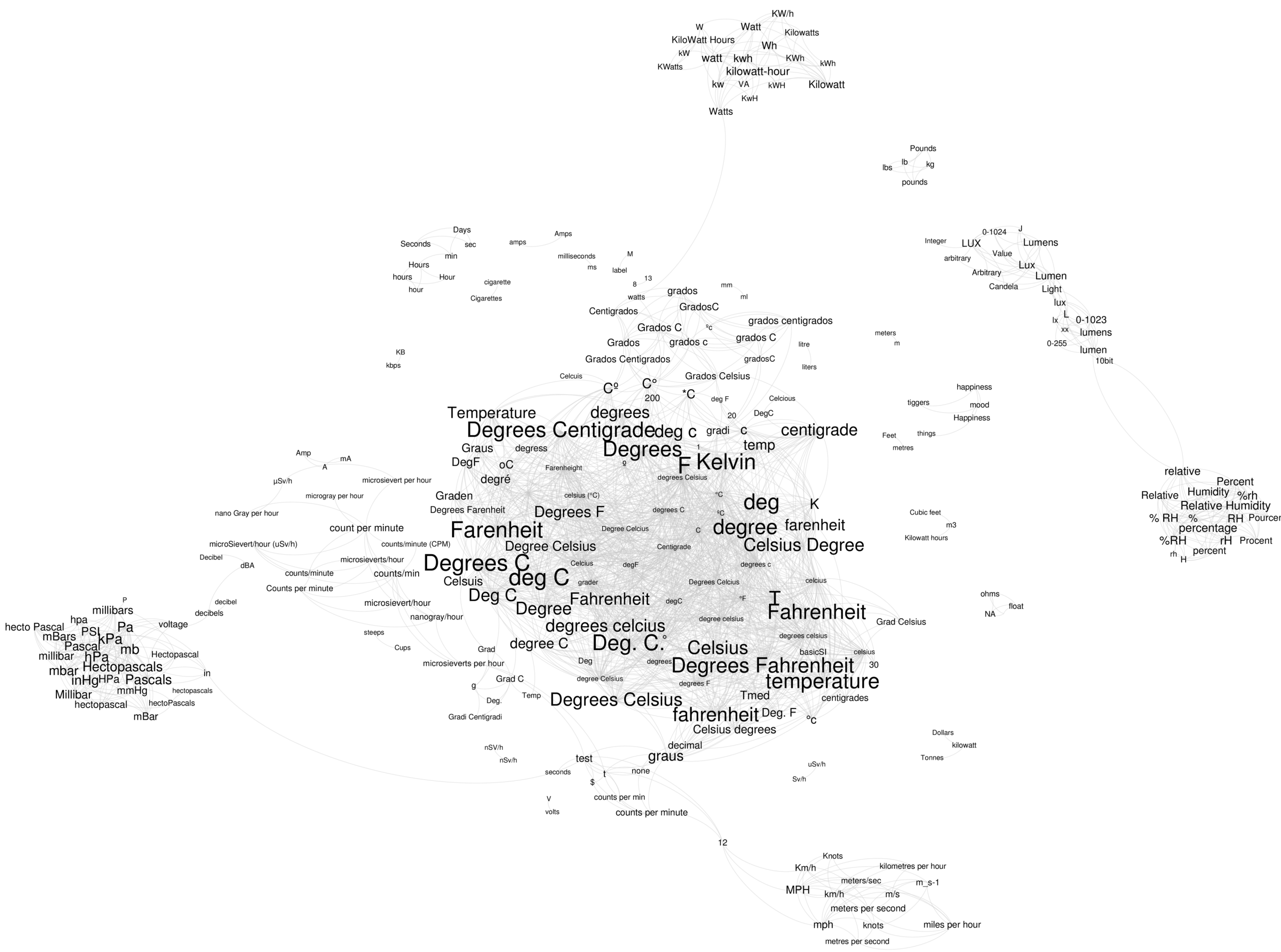


Figure D.4: **Synonyms of Units of Measurement.** This data does not require the use of sampling since the vocabulary for units is smaller than that for tags. Two observations are noteworthy: a) Celsius and Fahrenheit are grouped together since they fit exactly the Clements definition of a synonym, since their use is highly correlated with language use: linguistically separate groups apply these different labels to describe the same concept of a temperature scale. This was an unexpected outcome of an otherwise successful method. b) The topics covered are less diverse compared to the tag synonym sets, most terms relate to temperature scales. This may indicate that there is little general agreement in notation for units other than temperature scales.

D.2 Term Clusters

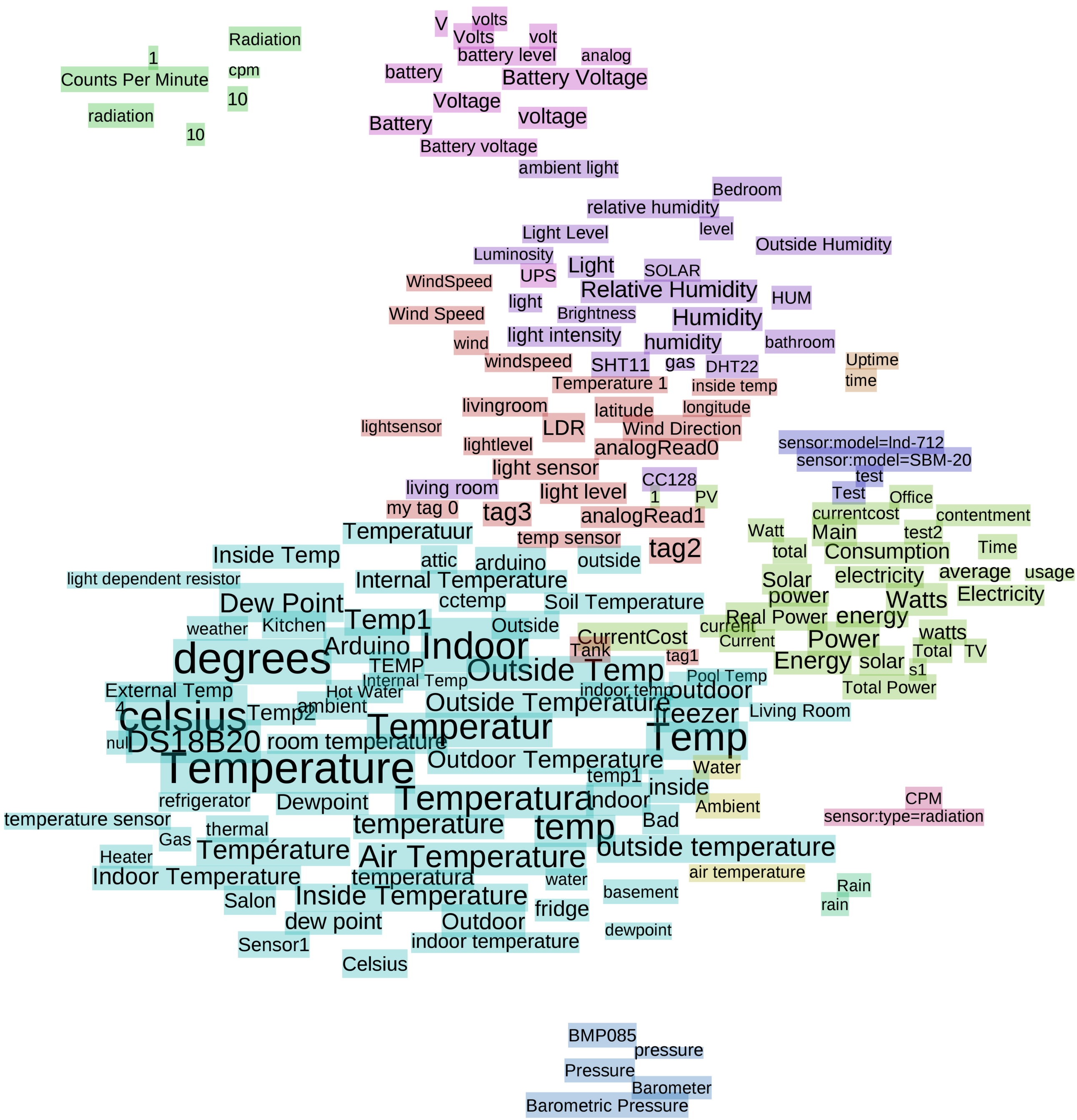


Figure D.5: The term communities of the tag synonym graph variant A.

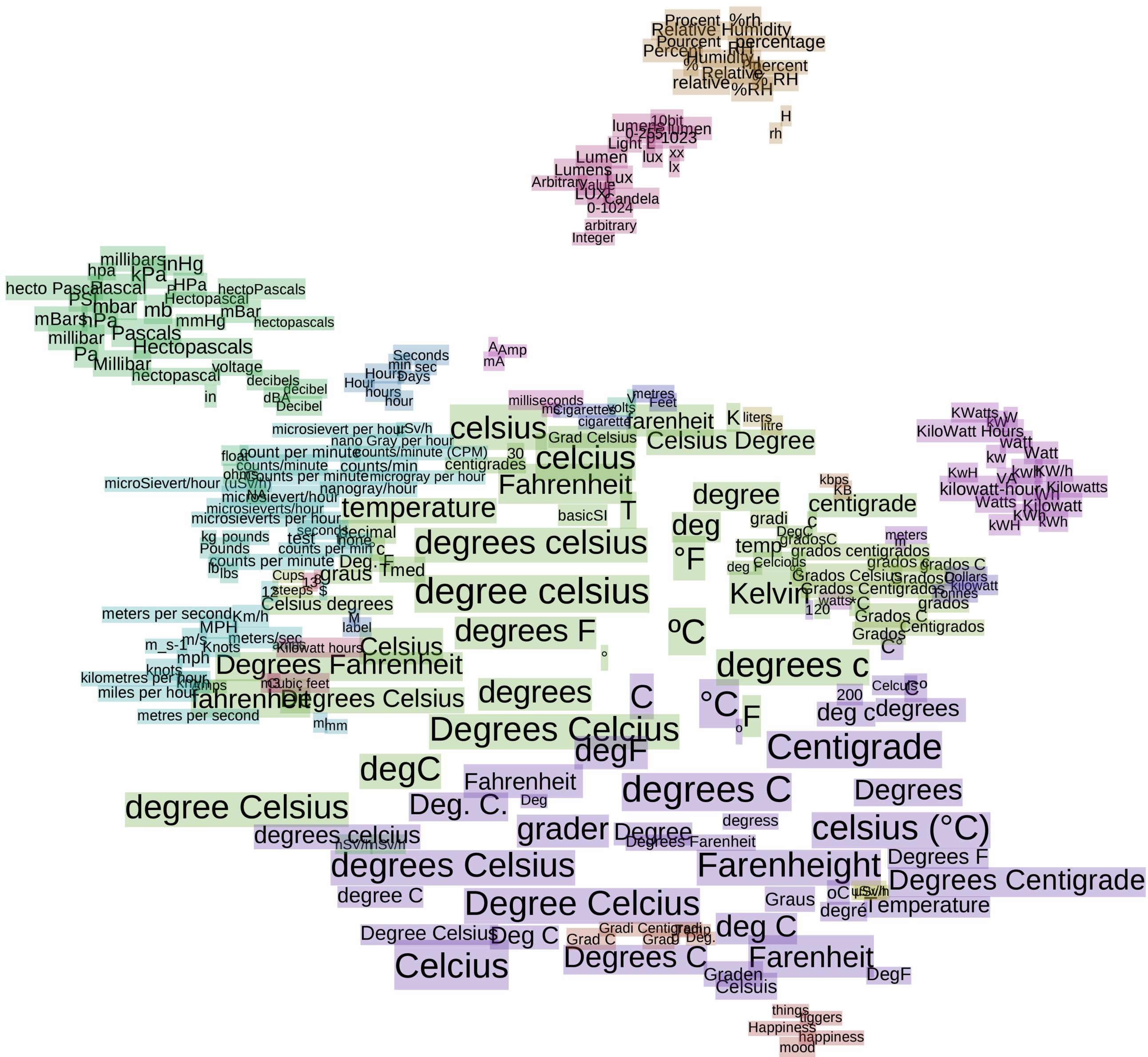


Figure D.6: The term communities of the unit synonym graph. It is striking that temperature-related units were grouped in two communities, with their respective vocabulary showing no clear difference in language or topical focus.

Appendix E

Term Lists

The following terms lists have been generated by the term clustering approach presented in Chapter 3. They were used to select the subsets of Cosm sensor data presented in Section 3.4.

E.1 Energy Usage and Power

This activity group is the set of datastreams that have been annotated with at least one of each of the following:

- Datastream tags: `AC`, `Apparent Power`, `apparent power`, `apparentPower`, `CC128`, `CurrentCost`, `currentcost`, `Din rail 2-Power`, `Din rail 3-Power`, `Electric`, `Electricity`, `electricity`, `Electricity Usage`, `Energy`, `energy`, `generation`, `Instantaneous Power`, `kWh`, `kwh`, `Power`, `power`, `Power Consumption`, `power consumption`, `Power Factor`, `Power Usage`, `Power Use`, `puissance`, `Real Power`, `RealPower`, `realPower`, `W`, `Watt`, `watt`, `Watts`, `watts`
- Units of measurement: `Kilowatt`, `KiloWatt Hours`, `kilowatt-hour`, `Kilowatts`, `kW`, `kw`, `KW/h`, `KWatts`, `KWh`, `KwH`, `kWH`, `kWh`, `kwh`, `VA`, `W`, `Watt`, `watt`, `Watts`, `watts`, `Wh`

E.2 Radiation

This activity group is the set of datastreams that have been annotated with at least one of the following:

- Datastream tags: `1 hour average`, 10 分移動平均, 10 分間移動平均, 1 時間移動平均, `Counts Per Minute`, `CPM`, `cpm`, `microsieverts`, `microsieverts/hour`, `Radiation`, `radiation`, `radiation sensor`, `SBM-20`, `sensor`, `sensor:model=lnd-712`, `sensor:model=SBM-20`, `sensor:type=radiation`, `μSv/h`, 1 時間移動平均

E.3 Humidity

This activity group is the set of datastreams that have been annotated with at least one of the following:

- Datastream tags: `Air Humidity`, `HUM`, `hum`, `HUMIDITY`, `Humidity`, `humidity`, `Humidity Out`, `Humidity Outside`, `Indoor Humidity`, `Inside Humidity`, `Luftfeuchtigkeit`, `Moisture`, `moisture`, `Outside Humidity`, `Relative Humidity`, `Relative humidity`, `relative humidity`, `RelativeHumidity`, `RH`, `Room Humidity`, `SHT11`, `Soil Moisture`, 湿度

E.4 Pressure

This activity group is the set of datastreams that have been annotated with at least one of the following:

- Units of measurement: `hecto Pascal`, `Hectopascal`, `hectopascal`, `Hectopascals`, `hectoPascals`, `hectopascals`, `HPa`, `hPa`, `hpa`, `inHg`, `kPa`, `mb`, `mBar`, `mbar`, `mBars`, `Millibar`, `millibar`, `millibars`, `mmHg`, `P`, `Pa`, `Pascal`, `Pascals`, `PSI`

E.5 Temperature

This activity group is the set of datastreams that have been annotated with at least one of each:

- Datastream tags: `Air Temp`, `Air Temperature`, `air temperature`, `Außentemperatur`, `Average Temperature`, `DS1820`, `DS18B20`, `ds18b20`, `External Temp`, `External Temp.`, `External Temperature`, `LM35`, `lm35`, `Outdoor temp`, `outdoor temp`, `Outdoor Temperature`, `outdoor temperature`, `Outside Temp`, `outside temp`, `Outside Temperature`, `Outside temperature`, `outside temperature`, `OutsideTemp`, `outsideTemp`, `sensor:type=temperature`, `TEMP`, `Temp`, `temp`, `Temp Out`, `Temp Outside`, `temp outside`, `temp sensor`, `Temp1`, `temp1`, `Temp2`, `temp2`, `temp3`, `Temparature`, `TempC`, `Temperatur`, `temperatur`, `Temperatura`, `temperatura`, `Temperatura esterna`, `Temperatura Exterior`, `Temperature`, `temperature`, `Temperature Office`, `Temperature Out`, `temperature sensor`, `Temperature1`, `temperature1`, `Temperature2`, `Temperatuur`, `Temprature`, `tempreture`, `Tempsensor`, `temp_ext`, `Température`, `température`, `teplota`, `thermal`, `Thermistor`, `thermistor`, `Thermometer`, `thermometer`, `TMP36`
- Units of measurement: `*C`, `C`, `c`, `Celcius`, `celcius`, `Celcuis`, `Celsius`, `celsius`, `celsius (°C)`, `Celsius Degree`, `Celsius degrees`, `Celsuis`, `Centigrade`, `centigrade`, `centigrades`, `Centigrados`, `C°`, `Cº`, `Deg C`, `deg C`, `deg c`, `Deg.`, `Deg. C.`, `degC`, `degree C`, `Degree Celcius`, `Degree Celsius`, `degree Celsius`, `degree celsius`, `Degrees C`, `degrees C`, `degrees c`, `Degrees Celcius`, `degrees celcius`, `Degrees Celsius`, `degrees Celsius`, `degrees celsius`, `Degrees Centigrade`, `degré`, `Grad`, `Grad C`, `Grad Celsius`, `Graden`, `grader`, `gradi`, `Gradi Centigradi`, `Grados C`, `grados C`, `grados c`, `Grados Celsius`, `Grados Centigrados`, `grados centigrados`, `GradosC`, `gradosC`, `oC`, `℃`, `°C`, `°c`, `º`, `ºC`, `ºc`

Appendix F

Cosm Community Activities

This appendix presents the data traces of a number of exemplary sensing activities as identified by our term clustering approach, the data was selected using the term lists presented in Appendix E. A summary for each group is provided in Section 3.4.

F.1 Energy Usage and Power

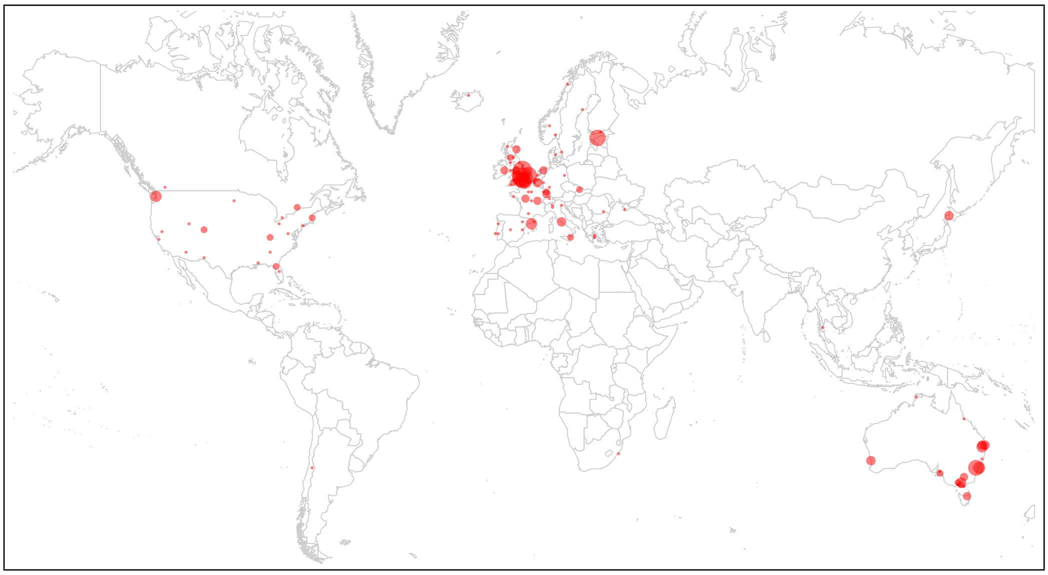


Figure F.1: Global distribution of energy usage sensors in April 2012.

Data Volume

August 2011	September 2011	March 2012	April 2012
417	433	699	725

Table F.1: Number of active sensors in this group. The charts are labelled with the peak number of daily sensors. There is significant and constant growth in activity over the observed period, from 417 sensors in August to 725 sensors in April.

August 2011	September 2011	March 2012	April 2012
417 1	433 1	697 3	723 2

Table F.2: Number of sensors in this group reporting numeric and non-numeric values, respectively. The latter may indicate data capturing errors, or the capturing of textual data or structured records. The charts are labelled with the peak number of daily sensors.

Data Variance

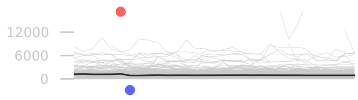
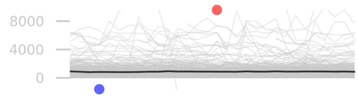
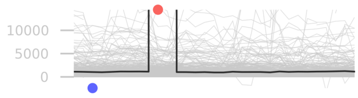
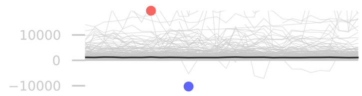
August 2011	September 2011
 523 974.0 2.033	 519 877.9 1.786
March 2012	**April 2012**
 868 165.8k 3.525	 923 1,156 2.417

Table F.3: Data variance plots for data streams, by month. There is a great diversity of values across datastreams, with the strongest band of measurements in the low thousands of watts, but a number of high outliers. On a number of occurrences at least one sensor reports negative energy usage.

Units of Measurement

August 2011	September 2011	March 2012	April 2012
 343	 363	 644	 702
watts	watts	watts	watts
watts	watts	watts	watts
watt	kwh	watt	watt
kwh	watt	kwh	kwh
watt	kwh	kwh	kwh
kwh	kw	kw	kw
kwh	watt	watt	kwh
	kwh	kwh	watt
		kwh	kilowatt-hour
		kw	

Table F.4: Most frequently used units of measurement, by month. The histograms show term frequencies for the entire unit vocabulary used in the respective month. The most significant share of sensors reports energy usage in watt.

Environment Tags

August 2011	September 2011	March 2012	April 2012
 414	 426	 734	 783
(not provided)	(not provided)	(not provided)	(not provided)
energy	energy	energy	electricity
current cost	electricity	electricity	energy
currentcost	power	temperature	temperature
electricity	temperature	power	power
power	small power	currentcost	currentcost
temperature	currentcost	small power	small power
cc128	comma-separated descrip …	solar	solar
comma-separated descrip …	solar	cc128	cc128
solar	current sensor:ctl-6-s3 …	pv	router

Table F.5: Most frequently used environment tags for sensor data, by month. Almost none of the datastreams in this group have any environment tags specified. Among those that do, Current Cost features prominently.

Stream Tags

August 2011	September 2011	March 2012	April 2012
 458	 436	 760	 813
watts	watts	watts	watts
power	power	power	power
electricity	electricity	electricity	electricity
cc128	energy	energy	energy
energy	cc128	cc128	cc128
watt	watt	watt	watt
average	small power	small power	small power
channel1	corecard	average	total
total	power consumption	total	channel1
sum	channel1	channel1	average

Table F.6: Most frequently used environment tags for sensor data, by month. Most of the datastreams have been tagged with their unit of measurement: `watts`. There is no significant other annotation that allows to deduce the sensing context.

Geographic Hotspots

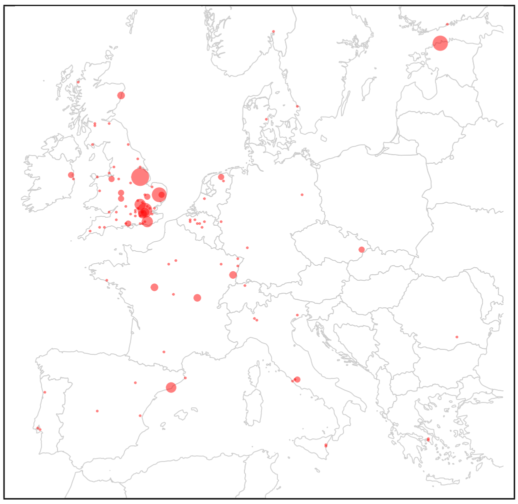


Figure F.2: Geographic distribution of sensors in Europe.

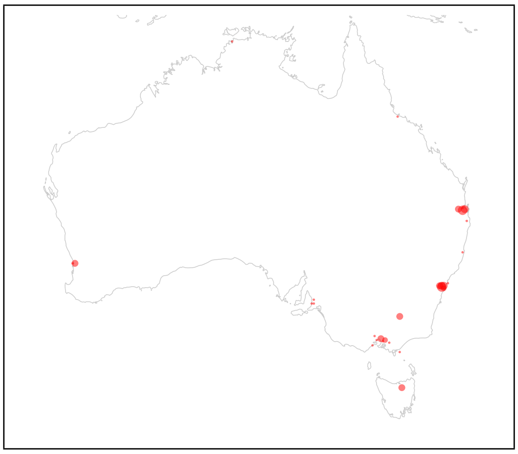


Figure F.3: Geographic distribution of sensors in Australia.

Location Names

August 2011	**September 2011**	**March 2012**	**April 2012**
335	351	613	647
(not provided)	(not provided)	(not provided)	(not provided)
paris	lincoln	lincoln	lincoln
lincoln	eye	tallinn	tallinn
tallinn	8 fitzroy st	eye	eye
eye	eau rouge	8 fitzroy st	8 fitzroy st
home	home	home	dunstable
penthouse	himalayan institute	dunstable	home
roznov pod radhostem	japan	11 camillo st pendle hi …	barcelona
dunstable	dunstable	somewhere and nowhere	seattle
himalayan institute	penthouse	hokkaido, japan	11 camillo st pendle hi …

Table F.7: Most frequently used location names for sensor data, by month. Most of the datastreams have no location name provided (335 in August.)

F.2 Radiation

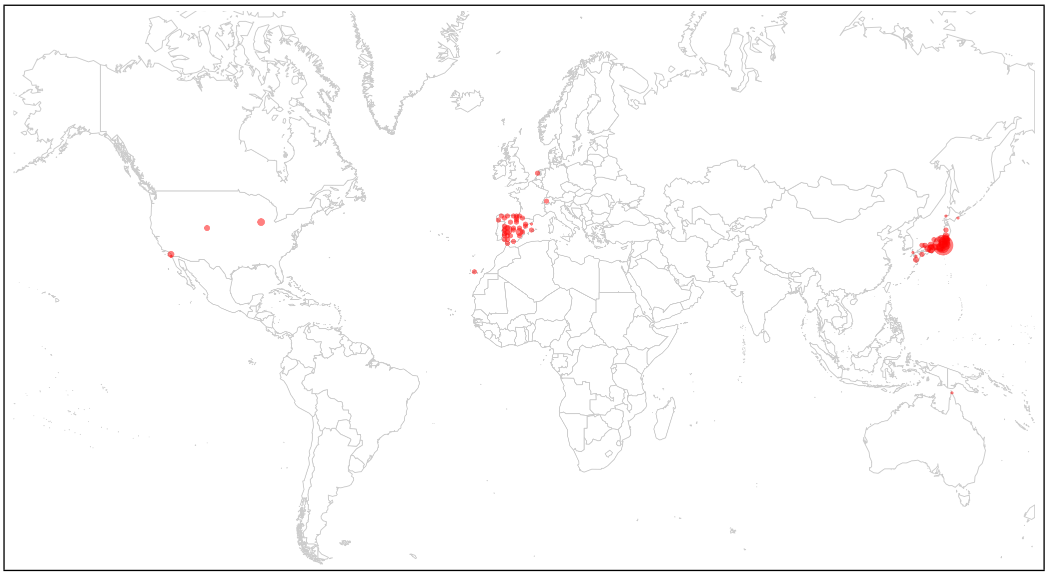


Figure F.4: Global distribution of radiation sensors in April 2012.

Data Volume

August 2011	September 2011	March 2012	April 2012
70	92	243	287

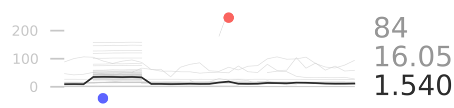
Table F.8: Data volume of radiation sensor activities, in number of sensors. There is clear growth over time, but also clear indications that sensors are frequently turned on and off in large groups. This potentially indicates concerted efforts, or at least individuals operating large numbers of sensors.

August 2011	September 2011	March 2012	April 2012
70 0	92 0	243 0	287 0

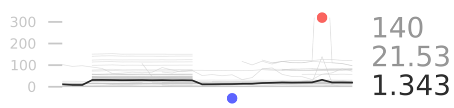
Table F.9: Number of sensors in this group reporting numeric and non-numeric values, respectively. All reported data is in numeric form.

Data Variance

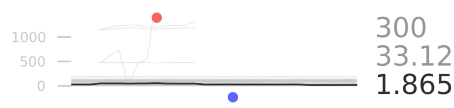
August 2011



September 2011



March 2012



April 2012

Table F.10: Data variance plots for radiation data streams, by month. These plots clearly show that the sensor groups that temporarily operated in August and September reported values in `nano gray per hour`, during their period of activity there are prominent bands in value ranges between 20 and 150.

Units of Measurement

August 2011	September 2011	March 2012	April 2012
 36	 36	 95	 100
`nano gray per hour`	`nano gray per hour`	`cpm`	`cpm`
(not provided)	`cpm`	`microsieverts/hour`	`microsieverts/hour`
`microsieverts/hour`	(not provided)	`nano gray per hour`	`nano gray per hour`
`microsieverts per hour`	`microsieverts/hour`	`nsv/h`	`nsv/h`
`cpm`	`microsieverts per hour`	(not provided)	`microsieverts per hour`
`micro-sieverts per hour`	`nanosieverts per hour`	`nanosieverts per hour`	(not provided)
`usv/h`	`counter per minute`	`micro-sieverts per hour`	`counts per minute`
`counts/minute (cpm)`	`count per minute`	`µsv/h`	`µsv/h`
`counts per minute`	`µsv/h`	`count per minute`	`nanosieverts per hour`
`microsieverts`	`counts/minute (cpm)`	`counts per minute`	`cpm`

Table F.11: Most frequently used units of measurement for radiation data, by month. A number of different units of measurement feature prominently, and the most frequently used units change between months. In later months `cpm` and `microsieverts/hour` are most popular.

Environment Tags

August 2011	September 2011	March 2012	April 2012
 58	 57	 190	 204
sensor:type=radiation	sensor:type=radiation	radiation	radiation
webscrape	mark2	sensor:type=radiation	sensor:type=radiation
radiation	radiation	mark2	mark2
geiger	nsv/h	nsv/h	cpm
(not provided)	cpm	cpm	nsv/h
geiger counter	geiger	geiger	geiger
sensor:model=lnd-712	webscrape	nuclear	nuclear
環境放射線計測	japan	radioactividad	country:spain
japan	線	country:spain	radioactividad
線	wakwak_koba	radioactivity	radioactivity

Table F.12: Most frequently used environment tags for radiation data, by month. Environments of radiation sensors are frequently tagged with machine tags. The `webscrape` tag in August and September appears to indicate that some data is not from a primary source, but republished from elsewhere.

Stream Tags

August 2011	September 2011	March 2012	April 2012
 45	 44	 87	 90
radiation	radiation	radiation sensor	1 時間移動平均
sensor:type=radiation	sensor:type=radiation	1 時間移動平均	radiation sensor
sensor:model=lnd-712	1 時間移動平均	air daily average gamma …	air daily average gamma …
cpm	cpm	radiación gamma diaria …	radiación gamma diaria …
unverified	sensor:model=lnd-712	radiation	radiation
sensor:model=sbm-20	sensor:model=sbm-20	sensor:type=radiation	sensor:type=radiation
microsieverts	1 時間移動平均	10 分間移動平均	10 分間移動平均
counts/minute (cpm)	sensor	cpm	cpm
1 時間移動平均	microsieverts	sensor:model=lnd-712	sensor:model=lnd-712
radiation sensor	10 分移動平均	sensor:model=sbm-20	sensor:model=sbm-20

Table F.13: Most frequently used environment tags for radiation data, by month. A number of streams have been tagged with device identifiers such as `sensor:model=lnd-712` or `sensor:model=sbm-20`.

Geographic Hotspots

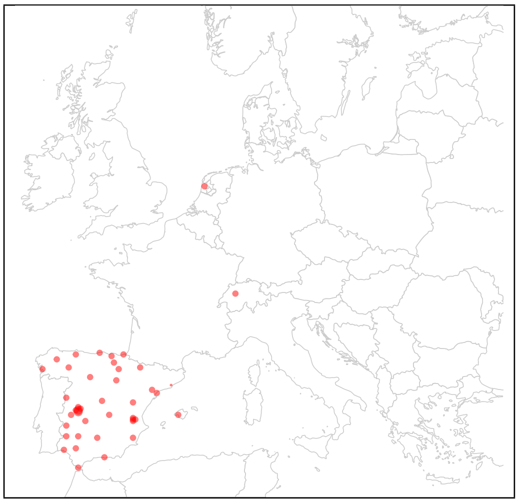


Figure F.5: Geographic distribution of sensors in Europe.

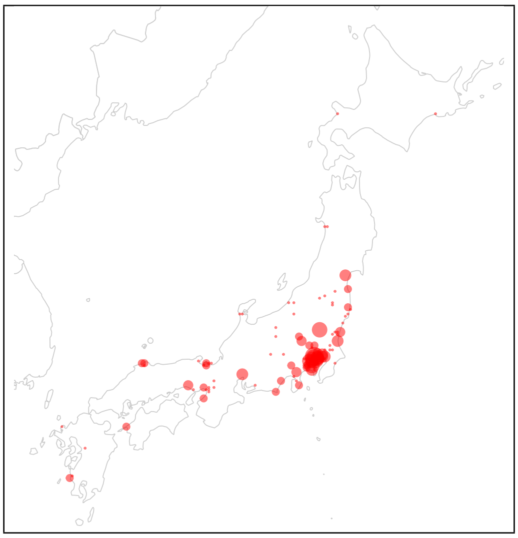


Figure F.6: Geographic distribution of sensors in Japan.

Location Names

August 2011	September 2011	March 2012	April 2012
 9	 9	 29	 32
(not provided)	(not provided)	japan	japan
archetype	japan	(not provided)	(not provided)
longmont	名古屋から概ね 50 …	japan	japan
akadamachigata, niigata	tokyo	japan	japan
araji, ibaraki	software park	tokyo	tokyo
asao kawasaki, japan	himeji	itabashi-ku tokyo japan	demonstration irad geig …
berne	hayamacho,kanagawa-ken, …	名古屋から概ね 50 …	名古屋から概ね 50 …
boulder, co	zushi,kanagawa-ken,japan	machida-city, tokyo, ja …	itabashi-ku tokyo japan
chidori, kanagawa	longmont	himeji	utsunomiya
fukusima,kawamatamachi	shinjyuku east, tokyo, …	chiba, japan	kashiwa

Table F.14: Most frequently used location names for radiation data, by month.

F.3 Humidity

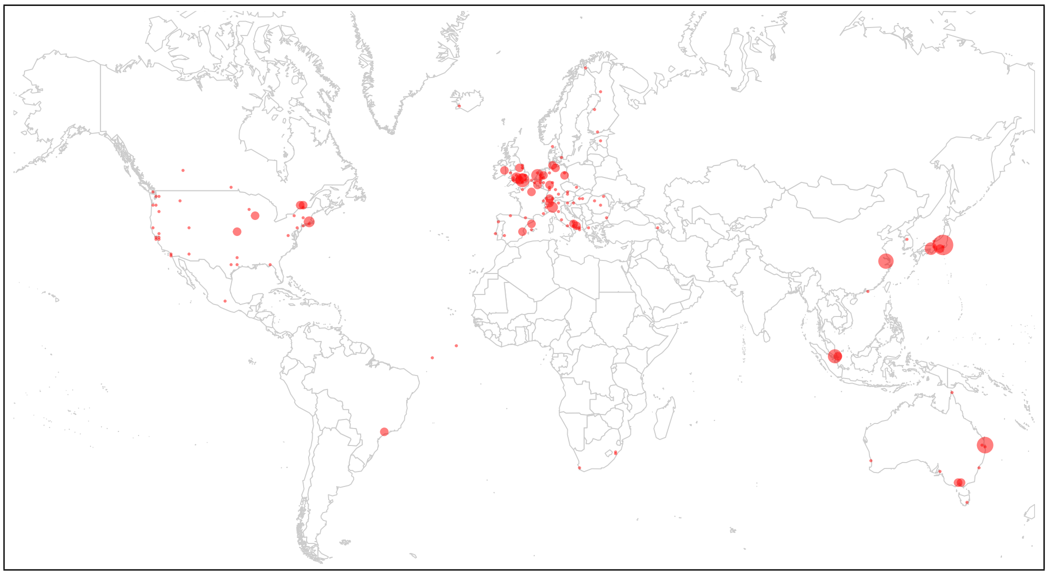


Figure F.7: Global distribution of humidity sensors in April 2012.

Data Volume

August 2011	September 2011	March 2012	April 2012
112	112	187	194

Table F.15: Number of active sensors in this group. As seen for the other groups, data volume is steadily increasing.

August 2011	September 2011	March 2012	April 2012
110 3	110 2	185 3	193 3

Table F.16: Number of sensors in this group reporting numeric and non-numeric values, respectively.

Data Variance

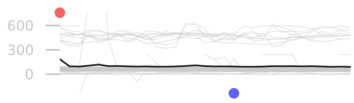
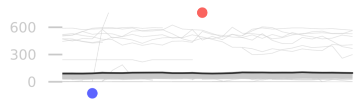
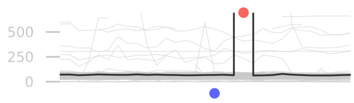
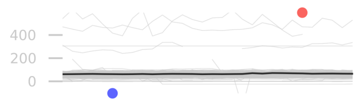
August 2011	September 2011
 136 100.3 1.380	 134 97.57 1.340
March 2012	**April 2012**
 246 14.94k 1.443	 252 64.62 0.959

Table F.17: Data variance plots for data streams, by month. Most sensor values are below 100, as is expected for relative humidity measurements. The coefficient of variation is low, particularly in April, which indicates lower variance in measurement values than those of several other groups.

Units of Measurement

August 2011	September 2011	March 2012	April 2012
			
(not provided)	(not provided)	%	%
%	%	(not provided)	(not provided)
percentage	percent	percent	percent
percent	percent	%rh	%rh
percent	percentage	percent	percent
rh	humidity	percentage	percentage
humidity	relative humidity	rh	rh
relative val	rh	percentage	humidity
relative humidity	relative val	humidity	percentage
(%rh)	(%rh)	(%rh)	(%rh)

Table F.18: Most frequently used units of measurement, by month. Sensors predominantly report relative humidity values, but as many as 20% of sensors (29 in August) do not report a unit.

Environment Tags

August 2011	**September 2011**	**March 2012**	**April 2012**
 75	 71	 102	 110
(not provided)	(not provided)	(not provided)	(not provided)
`temperature`	`temperature`	`temperature`	`temperature`
`humidity`	`humidity`	`humidity`	`humidity`
`arduino`	`arduino`	`light`	`arduino`
`light`	`light`	`arduino`	`light`
`carbon dioxide`	`weather`	`weather`	`weather`
`carbon monoxide`	`carbon monoxide`	`pressure`	`pressure`
`noise`	`monitoring`	`relative humidity`	湿度
`monitoring`	`carbon dioxide`	湿度	温度
`weather tunnel`	`architecture`	温度	`wind`

Table F.19: Most frequently used environment tags for sensor data, by month. More than 50% of humidity sensors do not have any environment tags (110 of them in April.)

Stream Tags

August 2011	**September 2011**	**March 2012**	**April 2012**
 111	 105	 201	 205
`humidity`	`humidity`	`humidity`	`humidity`
`relative humidity`	`relative humidity`	`relative humidity`	`relative humidity`
`hum`	`hih4030`	`hum`	湿度
`indoor`	`hum`	湿度	`hum`
`moisture`	`moisture`	`outdoor`	`outdoor`
`outdoor`	`outdoor`	`relative`	`dht22`
`percents`	`percents`	`bedroom`	`relative`
`realative`	`realative`	`dht22`	`moisture`
`relative`	`relative`	`weather`	`bedroom`
`sensorc`	`sensorc`	`dht11`	室外

Table F.20: Most frequently used environment tags for sensor data, by month. The `humidity` datastream tag is one of the selection criteria, and it has been applied to most of the sensors in this group.

Geographic Hotspots

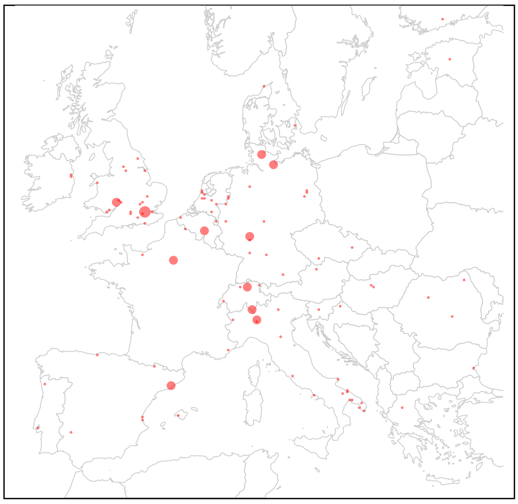


Figure F.8: Geographic distribution of sensors in Europe.

Figure F.9: Geographic distribution of sensors in Japan.

Location Names

August 2011	**September 2011**	**March 2012**	**April 2012**
 23	 19	 36	 37
(not provided)	(not provided)	(not provided)	(not provided)
japan saitama koshigaya	japan saitama koshigaya	singapore	suzhou, jiangsu, china, …
brooklyn ny	brooklyn ny	suzhou, jiangsu, china, …	singapore
osaka, japan	bromsash	japan saitama koshigaya	japan saitama koshigaya
bromsash	hirakawa-shi, aomori, j …	complex urban systems	complex urban systems
nanorite office 202h	zzzinc	london, uk	japan
apartment	osaka, japan	newbury uk	brooklyn ny
zzzinc	apartment	brooklyn ny	london, uk
heemskerk	berlin - weather tunnel	japan	institute of telematics …
bitritto	bitritto	barcelona	osaka, japan

Table F.21: Most frequently used location names for sensor data, by month.

F.4 Pressure

Figure F.10: Global distribution of pressure sensors in April 2012.

Data Volume

August 2011	September 2011	March 2012	April 2012
47	46	86	84

Table F.22: Number of active sensors in this group.

August 2011	September 2011	March 2012	April 2012
47 2	46 2	86 3	84 2

Table F.23: Number of sensors in this group reporting numeric and non-numeric values, respectively.

Data Variance

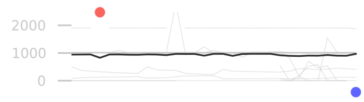
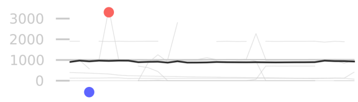
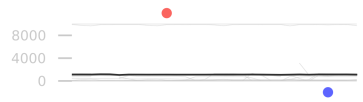
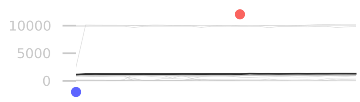
August 2011	September 2011
 53 934.8 0.345	 52 914.2 0.379
March 2012	**April 2012**
 100 1,123 1.301	 97 1,254 1.395

Table F.24: Data variance plots for data streams, by month. Because of the use of a number of different measurement scales these charts show several bands of activity. They also indicate that these pressure sensors tend to be stable in their reported measurements, there are not many large fluctuations.

Units of Measurement

August 2011	September 2011	March 2012	April 2012
 22	 22	 41	 40
hpa	hpa	hpa	hpa
millibar	millibar	mbar	millibar
mbar	mbar	millibar	mbar
hectopascals	mb	mb	mb
mb	hpa	mb	mb
hpa	hectopascal	hpa	hpa
hectopascals	millibar	hectopascal	hectopascal
hectopascal	hectopascals	hectopascals	hectopascals
millibars	millibars	hpa	hpa
hectopascals	hectopascals	mbar	mbar

Table F.25: Most frequently used units of measurement, by month. There is a consistent almost even split between measurements in pascal and in bar, with most measurements in pascal.

Environment Tags

August 2011	**September 2011**	**March 2012**	**April 2012**
 25	 25	 36	 34
(not provided)	(not provided)	(not provided)	(not provided)
`temperature`	`temperature`	`temperature`	`temperature`
`pressure`	`weather`	`humidity`	`humidity`
`weather`	`pressure`	`pressure`	`pressure`
`humidity`	`humidity`	`dilution`	`weather`
`arduino`	`millikelvin`	`helium-3`	`arduino`
`cryoconcept`	`cryoconcept`	`cryogenics`	`air pressure`
`cryomagnetics`	`cryogenics`	`cryomagnetics`	`wind`
`helium-3`	`cryomagnetics`	`cryoconcept`	`millikelvin`
`dilution`	`helium-3`	`weather`	`helium-3`

Table F.26: Most frequently used environment tags for sensor data, by month. Many sensors have no environment tags, and those that do frequently have generic tags, or even tags that are not directly related to air pressure, such as `temperature`.

Stream Tags

August 2011	**September 2011**	**March 2012**	**April 2012**
 11	 11	 22	 24
`atmospheric pressure`	`atmospheric pressure`	`pressure`	`pressure`
`barometric pressure`	`pressure`	`barometric pressure`	`barometric pressure`
`pressure`	`barometric pressure`	`atmospheric pressure`	`atmospheric pressure`
`barometer`	`bmp085`	(not provided)	(not provided)
(not provided)	(not provided)	`barometer`	`barometer`
`bmp085`	`barometer`	`air pressure`	`air pressure`
`pinject t`	`air pressure`	`pvc t`	`bmp085`
`pkeg t`	`press`	`bmp085`	気圧
`presión atmosférica`	`pression`	`pstill t`	`luftdruck`
`press`	`pression relative`	`pinject t`	`external pressure`

Table F.27: Most frequently used datastream tags for sensor data, by month. These employ a number of synonymous terms for `atmospheric pressure`. The `bmp085` is a barometric pressure sensor, an electronics component popular among DIY practitioners.

Geographic Hotspots

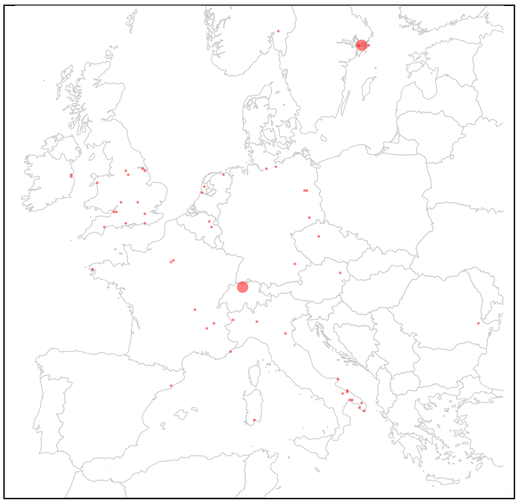


Figure F.11: Geographic distribution of sensors in Europe.

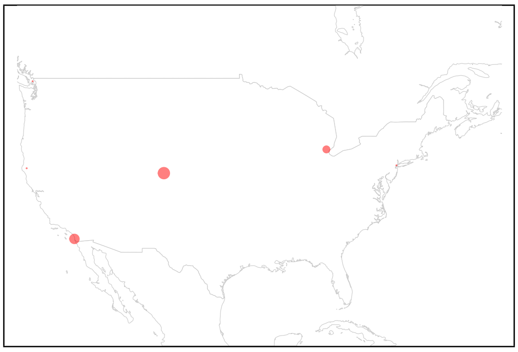


Figure F.12: Geographic distribution of sensors in the United States.

Location Names

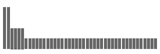
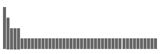
August 2011	**September 2011**	**March 2012**	**April 2012**
 4	 4	 10	 11
boulder, co	boulder, co	(not provided)	(not provided)
(not provided)	(not provided)	325 broadway, boulder, …	boulder, co
brisbane	gap qub	boulder, co	la jolla, ca
france	brisbane	la jolla, ca	brisbane
gap qub	lincoln, ma	lincoln, ma	klara östra kyrkogata …
lincoln, ma	beauregard-baret	dublin	lincoln, ma
berlin - wedding	berlin - wedding	brisbane	plymouth, michigan (cl …
bitritto	bitritto	klara östra kyrkogata …	dublin
asaktoppen leirsund	braila , romania	aarau, switzerland	aarau, switzerland
braila , romania	busware.de offices	plymouth, michigan (cl …	berlin - wedding

Table F.28: Most frequently used location names for sensor data, by month.

F.5 Temperature

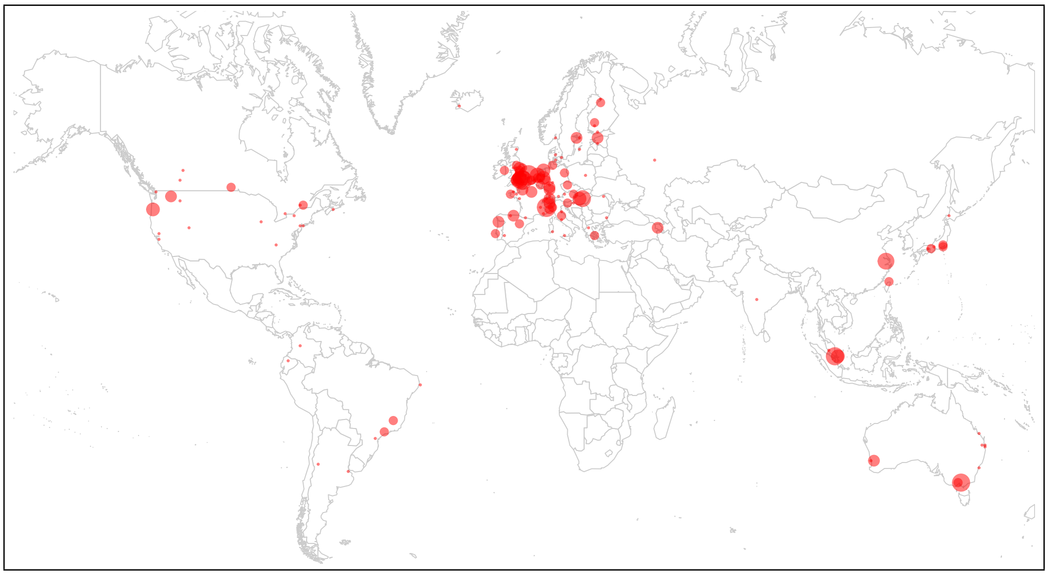


Figure F.13: Global distribution of temperature sensors in April 2012.

Data Volume

August 2011	September 2011	March 2012	April 2012
206	200	330	348

Table F.29: Number of active sensors in this group. The daily number of sensors increases by 150% over the observed period.

August 2011	September 2011	March 2012	April 2012
205 3	200 1	329 2	346 10

Table F.30: Number of sensors in this group reporting numeric and non-numeric values, respectively.

Data Variance

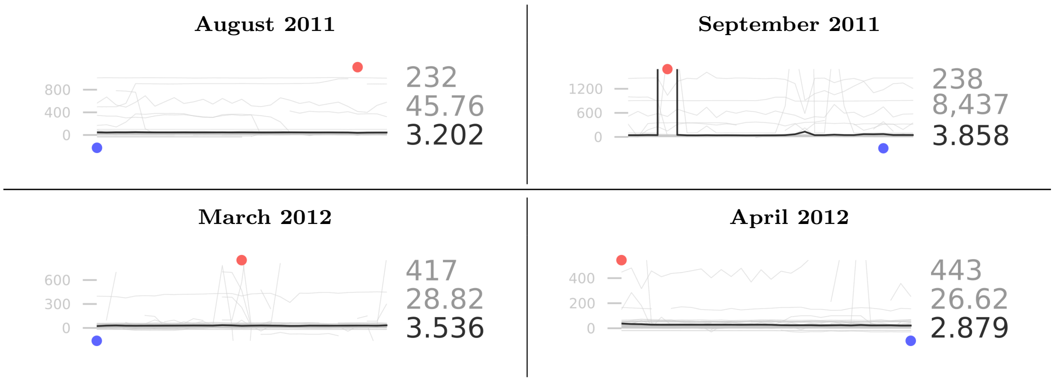


Table F.31: Data variance plots for data streams, by month. This clearly indicates that there is a wide range of purposes for temperature measurements: a small number of sensors report temperatures of hundreds of degrees, although most stay at temperatures below 100. The coefficient of variation is fairly high, and the mean value reaches around 45 degrees in August and 8,000 in September. As the number of sensors increases the mean temperature value falls to around 26 degrees in April, which is also the month with the lowest variance.

Units of Measurement

August 2011	**September 2011**	**March 2012**	**April 2012**
143	147	198	198
celsius	celsius	celsius	celsius
celcius	celcius	c	c
°c	c	°c	°c
c	°c	deg c	celcius
degrees celsius	celsius	celcius	celsius
celsius	degrees celsius	°c	deg c
degrees celsius	°c	celsius	°c
deg c	degrees celsius	degree celsius	degrees c
°c	centigrade	degrees celsius	degrees celsius
degree celsius	degree celsius	degree celsius	celcius

Table F.32: Most frequently used units of measurement, by month.

Environment Tags

August 2011	September 2011	March 2012	April 2012
 174	 164	 211	 245
(not provided)	(not provided)	(not provided)	(not provided)
`temperature`	`temperature`	`temperature`	`temperature`
`humidity`	`humidity`	`humidity`	`humidity`
`currentcost`	`arduino`	`light`	`arduino`
`power`	`bmp085`	`arduino`	`ds18b20`
`weather`	`power`	`pressure`	`weather`
`arduino`	`weather`	`weather`	`light`
`electricity`	`tmp102`	`ds18b20`	`nanode`
`light`	`temt6000`	`smart building`	`pressure`
`energy`	`hih4030`	`ldr`	`smart building`

Table F.33: Most frequently used environment tags for sensor data, by month. Most sensors do not provide any environment tags.

Stream Tags

August 2011	September 2011	March 2012	April 2012
 197	 203	 321	 335
`temperature`	`temperature`	`temperature`	`temperature`
`pbd:type=tix`	`pbd:type=tix`	`pbd:type=tix`	`pbd:type=tix`
`pbd:floor=5`	`pbd:floor=5`	`temp`	`temp`
`pbd:floor=2`	`pbd:floor=4`	`temperatura`	`pbd:floor=5`
`pbd:floor=3`	`pbd:floor=2`	`pbd:floor=5`	`pbd:floor=3`
`pbd:floor=4`	`pbd:floor=3`	`pbd:floor=4`	`temperatura`
`zigbee`	`temp`	`pbd:floor=3`	`pbd:floor=2`
`1 wire`	`arduino`	`pbd:floor=2`	`pbd:floor=4`
`xbee`	`xbee`	`outside temp`	`outside temp`
`arduino`	`zigbee`	`zigbee`	`zigbee`

Table F.34: Most frequently used environment tags for sensor data, by month. In April, 75% of streams have been tagged with the same `temperature` tag.

Geographic Hotspots

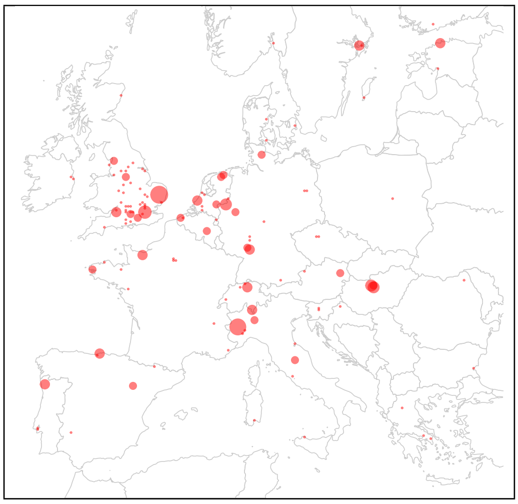


Figure F.14: Geographic distribution of sensors in Europe.

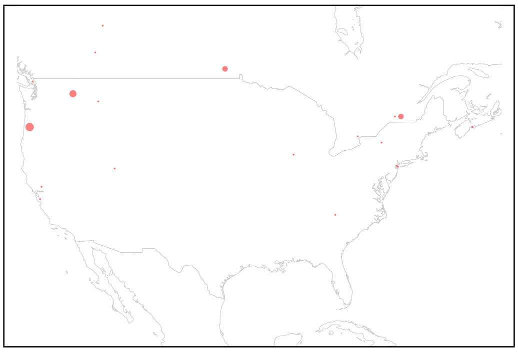


Figure F.15: Geographic distribution of sensors in the United States.

Location Names

August 2011	**September 2011**	**March 2012**	**April 2012**
 98	 98	 98	 98
london	london	london	london
(not provided)	eye	(not provided)	(not provided)
eye	scotland	singapore	singapore
scotland	eau rouge	eye	eye
home	home	turin	turin
waverley	(not provided)	mannheim, germany	suzhou, jiangsu, china, …
caen	waverley	scotland	scotland
tallinn	caen	suzhou, jiangsu, china, …	home
home	yerevan, armenia	nyíregyháza, hungary	london, uk
yerevan, armenia	home	home	waverley

Table F.35: Most frequently used location names for sensor data, by month.

Appendix G

Spatial Temperature Models

G.1 Met Office Temperature Data

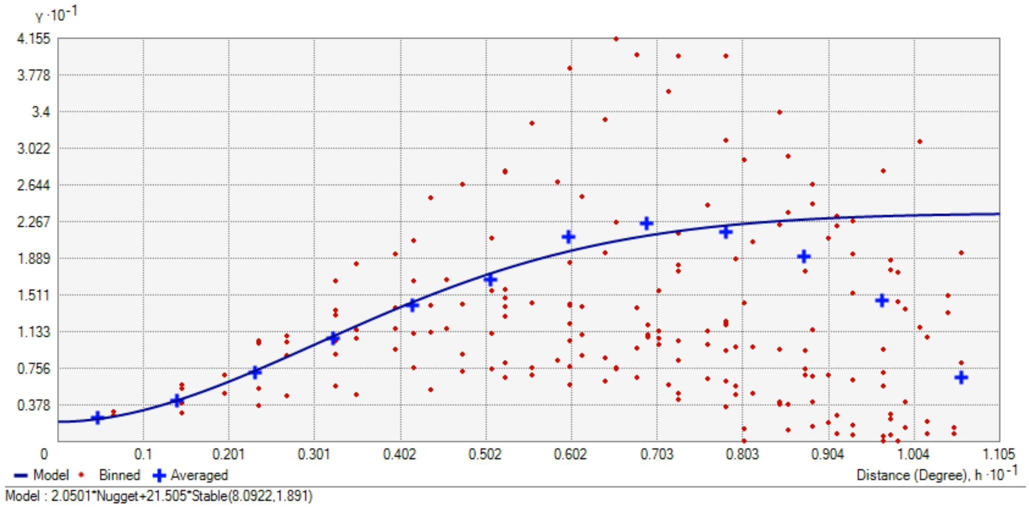


Figure G.1: The semivariogram of the Met temperature data overlaid with a statistical model of its spatial distribution.

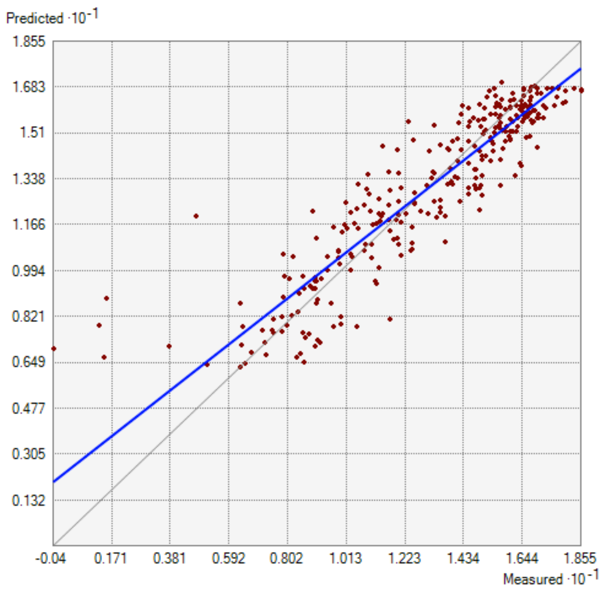


Figure G.2: The Kriging cross-validation of the Met data shows good predictive properties for the generated surface model.

G.2 Cosm Temperature Data

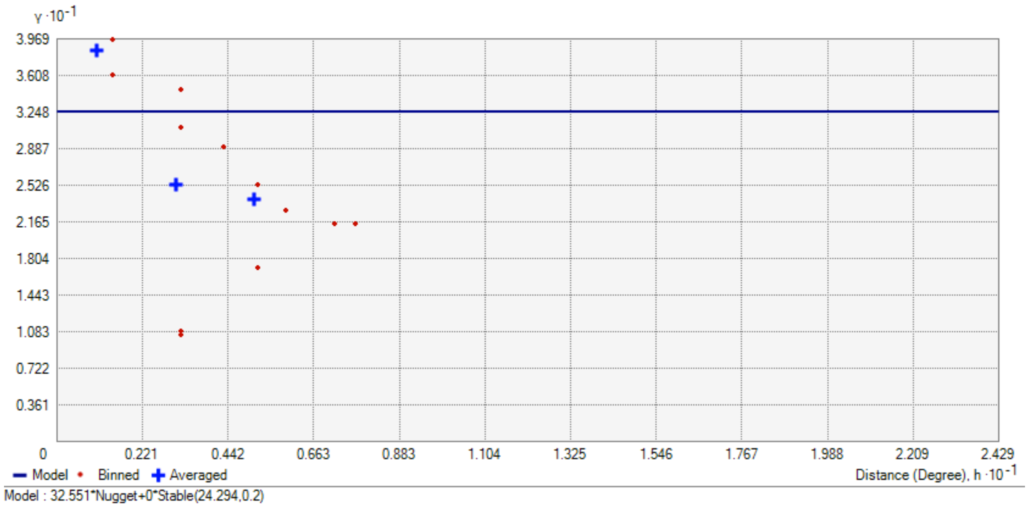


Figure G.3: The semivariogram of the Cosm temperature data. The data set is very small, and the semivariogram indicates negative spatial autocorrelation: it shows that nearby sensors have greater differences than sensors that are far away.

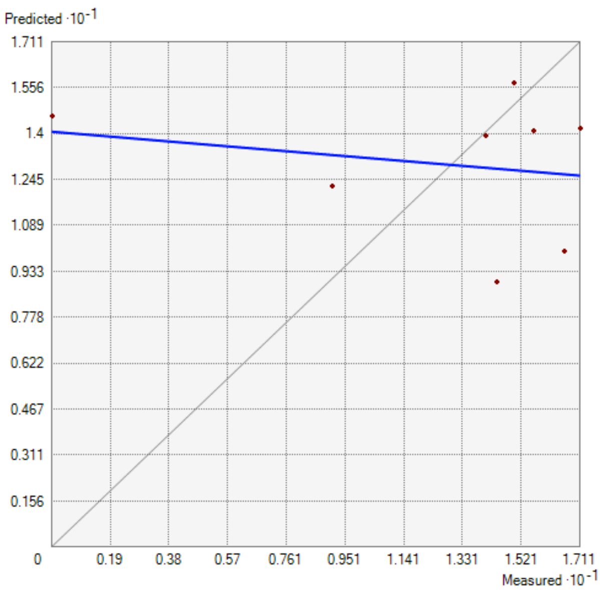


Figure G.4: The Kriging cross-validation confirms the early indicators. The resulting surface model is not a good predictor of the spatial distribution of this Cosm sensor data set. In other words, the data was unsuitable to build a large-scale spatial temperature model.

Appendix H

Data Sources

Basemaps

The World Borders Dataset is provided by Bjorn Sandvik under a Creative Commons Attribution-Share Alike License.
The data is available at `http://thematicmapping.org/downloads/world_borders.php`

It is derived from free mapping data provided by the Mapping Hacks website at `http://www.mappinghacks.com/data/`

Met Office Temperature Data

Temperature data and station location data provided by the UK Meteorological Office.

"Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853-current)"
Available from `http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_ukmo-midas`
Published by the NCAS British Atmospheric Data Centre, 2012.

Met Office weather station locations were derived from:
"Met Office Surface stations on Google Earth"
Available from `http://badc.nerc.ac.uk/search/midas_stations/google_earth.html`

Cosm Sensor Data

The Cosm sensor data analysed in this study is owned and published by the respective creators. Their usernames, in alphabetical order:

007, 0ste00, 100ideas, 1timetraveller, 2010061, 2bitpunk, 2mo_mp, 2v6, 3326192unsw, 3330673ld, 3335373, 3335386, 59kpd, 5ko, 65tux, 6db, _7l4ruc, 7n2atd, 87corey, 8rueducerf, 9600, 99dbspl, a10i, a28s5162, a744517, aaaahanda, aacgood, aadjan, aalborgzoo, aalfaro, aarathy, aardvarklove, aaronds, aaronlewtas, aaronmase, aarphacker, aarrieta, aathanor, abdulrahim, abe00makoto, abebe, abeler, abezencon, ablais, abluesheep, abslikke, abu_uca, accrete, acebillyfr, acefaser, achgarim, aciberlinpid, acidave, acky, acobo, acucos, adamrgreene, adanpinero, addidis, ade, adelaide101, adent, adolfo, adrian, adriantomic, adrionics, aetolos, aewp2, af6bi, afeltham, aferrandosixtac, afigar, agallia, agirod, agrigate, aguillet, aguw, ah01, ahannula, aideen, aigle, air_now, air_variable, aitsmartlab, aivo, aivoa, aivoh, aizu, ajcary, ajfisher, ajpowell, aka, akagi, akartem, akiccyo, akileos, akino_yo, akitakuma, akrauss, aksonlyaks, al1fch, alanarthur, alanmh, alansalter, alaszlo, albert, albertcordero, albertmaeda, albertoib, albertonaranjo, albertopanu, alcor_fr, aleardu, alegomesbr, aleksey, aleksi, aleredondi, alexanderg, alexandre, alexbc, alexlai, alexo, alexp, alfi, alfopine, alfredgunsch, algirdas, alicef_, alienstone, alinapier, alisonw, allanayr, almenny, alpchris, alphamunki, alphapapa, alramedicion, altieri, alvaro, alvydas, alwin, amagro, amalipawater, amcewen, amindlin, amir, amitchell60, amontep, amudle, amvv, ana3mic, anchan, anchunath, andersborg, andersd, anderso, andje50p, andopp, andras, andre, andreas6, andrebea, andrebstv, andreperazzi, andres, andres90125, andrescarceller, andreva, andrew1, andrew_debbie, andrewmovic, andrewn, andreword, andrey, andygodber, andysavery, an_ext, angelnu, angkringan, aniimsaj, aniketos, anko76bg, anmwhite, anniegoh, anno, anonimity80, anon_weather, antelec3, anthonyb, anthonyleung2003, antoinek, antoniokz, antpeng, antroyal, antw001, antw002, antw003, antw004, apduino, apineda, aplusautodoor, appi, apromix, aquamammal, ar, aranondo, arantec, aravintht, archcompsamuel, archimetrics, arcticspike, ard10n, ardentfan, arducrop, arduhome, arduino_mfc, arduinomstr, arduinopraxis, arduku, arfrabarb, ari, armored, arms22, arnaud, aromaoftacoma, arrch, arsalabs, arsenalwei, arthg, arthurmani, artistide, artoor, aruba_first, aruethe2, arutorin, arzhur, ashleyelsdon, ashos, asper, asto, asuka_yao, asve99, asw24b, atennant, atomicdave, atommann, atsushi, attom, atycocene, aubrey, audiobuzz, augie, aung02, aungmyat, austec, austingriffith, authoredsensors, autodomain, aveclaudenum, avneetkalsi, avonelectric, avrnoob, avthart, awaismaqsood, awall, awam, awheeley, awmiller1115, awoogah, awowk, awtracy, axello, axhieb, azabujyuban, azadef, azertyx, azhatoth, azimuth, azrul2506, aztechtest1, azwar10402, azzurra, badbob, badcopnodonut, baddawg, bagustrihatmaja, bakera01, bakerc01, balerion, ballen450, ballpremierlab, balr0g, banzy, barageiger, barbieri, barbosap, barbouri, bardsley, baritb, baronvonsid, barryfarrimond, barryingham, bashford, basten, batkinson, batou, bazee, bbalint01, bbharris, bbm3, bbonfanti, bcb289, bchhotel, bchudkim, bcolcord, bco_sandbox, bdowns, beanpolew, beatkrik, beber7310, bec_3333620, bedri, beduino, beejaye22, beep, beer, bekem, bela54321, ben, benbastian, bencrox, benddennis, benheather, benjamin44, benoitcarl, benoitdr, benpirt, benscroggs, bensnoorz, benthomas, beprojectkjsce, berkinet, bernie, bernix, berry_hvac, bertb, bertbret, bertouttier, besko2504, bettinac, bez002, bgr121, bgunnarb, bhedrick, bid, bigmoe88, biizujc, billinghursts, billyboy, billyd, bionicowl, bionueces, bioregional, bioz, birchthompson, birkjones, bit, bitflops, bitsflew, bizbiz, bjepson, bjmorel, bjoernhoefer, bjpirt, bl00027, black, blackcat22, blankdots, blarran, bld, bligh, blueagle, bluebird, bluecloudpowerco, bluejay117, bluescreen, bluespike, bmjenergy, bmtjpark, bobmclarty, bobz, bodgeit, boerm1, boffman, bogaziciuniversity, bogdanc, boilermaker, bongobbongo, bookie988, bookswapsteve, booyaa, bop, borisftz, boss1968_1, bostoen, bostonenginerd, boubou2005, bovine, boxfullofyertoys, boxingorange, boxysean, boz, bpijls, bradchapin, bradlannon, bradltaylor, bramper, brass, brenndorfler, bretforeman, brian, brian9, brianhouse, brian_huebner, bricogeek, brightstar, britiger, britt, broker, broo2, bruce69abc, bruh5200, brunokruse, brushedmoss, bryanbr, b_saravanan130832, btf83, btorrente, buckidge, budip, buildingbanter, building_banter, bul, bulkfoods, bundara, burkhardwave, burnig, bwstein, bynums, byrsa, c0ba1t, caddtw, cajomferro, cakester, calumscott, calvinbradshaw, canel_x, canfire, caniculari, canssens, caog, capablazab, capaianca, cardini, carefactorzer0, carlin550, carlism, carlitos, carlojav, carlosch, carlosvigo, carrickal, carrieandadrian, casatechnogaian, casestudies, caseyanderson, casperk, cathedral01, cauchyh, caveman, cazeault, cbandarrinha, cbastos, cberg, cbolanos, cburman, cbworks, ccailleux, cclpachube, ccspk, ccurren, cdebusschere, cdswift, ceaton, ced1100, cedriclocqueneux, cedric_willems, ceitmark, cejones2, celia, celioric, cesar, ceskar, cfgproject, cgt_maas, chadoneill, chads, chaidrinker, chalky99, chalyons3, chandra50, chandrase, changhao, charithperera, charlesweir, charlie36, charly_at, chayasorn, cheelo007, cheetoresearch, chellaufs, chemicus, chemoul, chenghsienyu, cheshir, chibiegg, chickabaumbaum, chido, chinesecell, chiro, chitosegeiger, chlost, chocoboy, chongado, chook, chreetz, chris_ebsworth, chrisjx, chrisk2011, chrismalyon, chrismayu, christianfenzl, christianhau, christiank21, christophemathieu, christosk, chriswmoixa, chriszero, chrlilje, chroma, chronosweb, chuckol, chumpster, chunky, chutan, chydenius, cipesokram, ciseco, citizensensor, civicexa, civix, cjaccino, clabarga, claire_w, clarencewollman, clarionwave, class1controls, claudemit, claws, cleeve, clete2, cliensol, c_lobo, clocky, cmalmqui, cmkpl, cml42, cmorton, cnaqphealth, cnetandreaspersson, cocolectric, coconox, codeartisan, cody, coeco, colarsson, coleoptero, colibri, colinmil, congojoe, conrad, consumobosco, contractorwolf, convolution, cooxee, corndogjoe, cornelius, cornholio, cosm, costantino, cotixtof, cottage4woods, cottoa00, couchounou, cougar, cpayne4471, cphmars, cphoenix, cptunderpants, cpvfd, craig, craigburton, craigsibthorpe, craigspry, crazymonster999, creon, cridasensors, crilyx, crispone, crocky, crocodileinside, crouchingbadger, crsherman, crtest, crturner10, cryptoman, crysmoore, csakovacs, csdatenlogger, csi, csorda, csrakowski, cstutz, ctitley, ct_moi, cu, cuborama, cubulet, cuno, cupolo, curi, curlgurlkatharine, curly_cat, cutehong, cverdel, cwharto3, cyberandy, cyberelectronics, cyberiantyger, cybersloth, cybraphon, cyclic, cynthiaquang, cyrl, d33g, d3estudio, daekwonpark, daemach, daibe123, daisen, dajuis, dakeqi, dalelane, damorian, dan, dan3335222, danbuzzo, dancelover, daneelariantho, danejohns, danewatkins, daniel2, danielvg, danilodal1976, danjnes, danlockton, dannyhaak, dannyunsw, dannyx, danward79, dapnet, darinrmcclure, darkmatter2uk, dark_tmkd, darkvenger, darkwolf, darladirladada, darrenjrobinson, datap, datataker, dave905947, davebe, daved, davedee, daveharte, davekelly, davekgen, daverg, davestoweuk, davetimm, davey, david1341, daviddiz, davide___z, davidfrith, david_g, davidillsley, davidm, davidncrowley, davidrei, davis441, dazjwil7, dbenjamin, dberenguer, dbnarduino, dbrgn, dbzoo, dchayes, dcrasch, ddef, ddejean, ddev, dealbu, debian123, debolman, deborahdora, decatel, dedja, deepaktlives, deepc0re, deerhntr, deez2009, deico, dekist, delicioustakoyaki, demetersrest, denkrecht, dennisg, deppie, derekfoster, derekw, derkammi, derraab, dettech, devenj, dewing, dexterindustries, dezellis, dfleck, dgbdgb, dgomes, dgraff, dhandzic, dhunter, diablo666, diasgaspar, diblau, dickinsonlabs, didess, didier12, didierrdv12, digilearner, digitalkatie, dimi86g, dimtemp, dincay, dinothepheno, dinvin, direith, dirk, dirtydeef, djanowski, djdave, djdecaprio, djdunc, djg, djinnantonnix, djrm, djsplice, dkhyde, dkruyt, dky, dl2sba, dlbird, dm3281, dminear, dmos, dna100, dochayashi, docmaly, doctorwho8, doghead, dolinar, doma, dominiccheung, domos, domotiga, domovince, domus, donaldsclark, donblake, dongambas, donger88, donmoylan, donnorbeck, donny, doom, dotmancando, doublespeed, douglascoupar, dovadi, doyoungju, dperfetto, dpeterse1, dpjrodrigues, dpm_ec, dptoinf, dqmarkham, draythomp, drdiode, drecuk, d_renata, drone2, drtobbe, drutherf, dscoular, dsetin19, dskouras, dss49, dsteele, duanelitchfield, duanxin1973, duarteaf, dudeeemc2, duncan_ellison, duppy, durhamtempstudy, dushantsi_reservoir, dustymonkey, dvrepairs, dwelltech, dw_soothill, dwwilde, dylanswift, dynablow, ead3103, eagle_temp, earthwatch, easterro, ebstsk, ebuchap, eburkon, echiara, echnaton, ecodrin, ecolo, edalessio, edbilodeau, edemoit, eden, edfhes, edmilner, edsonpavoni, eduard69, eduardosleal, edufl, edurao, edwar376, edwardrowlands, edy555, eeit, ees_bbonfanti, eettu, efimarket, efrecon, egadenne, egasp, eggplant, ego, ejeklint, ejj, ejsharp, ejvind, ejy, ekenny, ekmel, elbarto71, elchubi, eld_log, electra, electrosoft, elephantman, eli_thomas, elizarov, elliotb, elminister, elog, elringer, elsatch, elsmorian, elson, elwell, elwood, elykahn, elyobelyob, eman5oh, emante, embtrans, emielk, emmcee, empower, emsyslab, emu, emul, en6geh, endernet, energodata, energydesign, energy_elements, energymadeclean, enerwise, engeeknyer, engineering, enigma808, enjoe, ennouna, enri220, entasan, entraya, enviromapdevice1, envyen, eparizher, eraser928, erictj, ericwjanssen, erikendress, erikthompson, ermt, erniberni, erosales, errordeveloper, ersadmin, erspearson, erwinraysparks, esbsensor, esceptico, eseufert, esilvester, esmiddleton, esorense, estimmel, eszespeter, et08499, etemcan, eternity, etmmvdp, etunko, etx, euring, eversion, eviljess, evolvingtech, exentronom, experikus, exs10mh, exsertus, exta77, f0rdprefect, f3rr3t, fa2r, fab29plou, fabeuss67, fabianmartinez, fablabbcn, fadzlitohid, fahadalaieri, failsafedestruct, fali, falkr, fanatic, farmers_in_love, fbpinheiro, fcaeiro, fcarraro, fcint_2011, fclu21, fco, fcontino, feabhas, feddc, fede_lay, feibit, fekalen, felipecaioba, felipermrz, feng4ever, ferdep, fergalom, ferrifabio, feurig, ffaerber, fhug, fibreman, fictoris, fiddian, fikander, filartrix, filipecalegario, filipegandine, fiona_chan, firefoxx74, fishandchipsing, fishandlily, fishbonenetworks, fishy6969, fivefeet, fj604, fklug, fla, flagadajones, flannelators, flapdoodle, flavernhe, flavia, flecro, fleetfootmike, flg8r96, flintcave, fllfreak, florisa, flotess, flotorious, flow, fmepachube, fmeserver, fm_hobby, foliot, fomination, fontofdata, foobar, foolbar, foowho, foxnolds, fparaggio, fpsycho, fracsun_alpha, frafel, fraguada, francescov, francisperea, franckpvp, franklin97355, franzz, fred325i, fredd500, freddotnet, frederih56, fredw, freepower, freimann, frenkystromar, frey110, fribbe, fringui, frivisen, frogdog, frosty, frtklau, frugalnerd, fruita_man, fryo2me, fserbu, fsve, ftopalovski, ftranda, fuat_beaucaire, fukin, fukushimaobserv1, futboltactico, futurefabric, futureshape, futuretab, futurething, futureworx, fvanzati, fyp_ckt2012, g7pkf, gabrielneumueller, gadget3996, gadomer, galexa, galvanist, gambler2073, ganesh, ganis, garak, garlinto, garrettowne, garrydwilms, garyb130, garybailey, gashmore, gasmas, gaspode44, gaston, gavin, gavo2o, gayte, gbme, gbroiles, gbroos, gciuser1, gcm70, gctp, gdandrea, gdawgskillet, gdworks, geckopelt, geekcowboy, gegeoutest, geidy167, geiger_data, geigermonjp, geigertj, geluid, gemeline, genestone, genetikayos, genky, genniferem, geobruce, geodenx, geoffjones, geoffrey_aerts, george_batalinski, georgehorn, georgeofearth, georgepachube, geoweb, geraldkrug, geraldo, german_a, geros, gerymer, getusama, gfuggiano, ggkinuthia, ggrisandi, gham, ghianduinoxx, ghisi, ghubsch, giacomospano, gian2477, gianluca, giantbike, giantmolecules, gibsonmb, gies73, gigamegawatts, gigawatts, gilesdring, ginco70, ginidanhome, ginkgo, gino2206, gipelo, giulianob, gjbroom, gjchester, gjscott75, gknauff, glacier, glaw54, glegge, glenhendry, glenk, glennblackburn, gmartin, gmaxsonic, gmccarthy, gmtest, gnaddl, gnagnu, goatchurch, gockenbr, god, godblezzus, godden, godzero, gogasmart, gogojohn, gogol123, gohealthdemo, goilio, gojabako, goldeye, gonalo, gonzalo, goranmaz, gorath, gorillabuilderz, gorious, gotstu, graemel, granddaddyd, grantwood406, gratzel, gravitatoria, gravy82, greg81uk, gregeigsti, greghall, gregoryfenton, gremlin, greyvador, greywolf, grideyeq, griffithba, griffog, grivvr, groner, groot, grotecam, grumpeh, grunto, gsagmail009, gsgeschke, gsmgprstest, guanguan07, guanilon, gubgift, guidewire, guigui79, guitarmasaki, guitsou, guo, guolivar, guvenaasr, guy, guylevans, guzzydev, gvdw, gvool, gw0udm, gwithers105, gwooden, gzub, h0dges, h2kxyz, h34dup, h3q, ha247, hackerone, hackerspace_ffm1, hadara, haggan, hal0bender, hamargeiger, hamstereater, hanaana, hanscr, happyantix, haqthat, harder, harisdmac, harold65, harper, harry_c, haslar, hdr, he219, headphones, healthy, heater, heatherr, hectorcastrob, hectorguinness, hedi, heflopod, heisterberg, hema_mundkur, hendrik, henk_kuipers, henrywong, hepworthjd, herrprofessor, hervelu, hese, hetdorp, hexagon, hexxanlabs, hfiennes, hidebozu, higo, hillwalker, hilser, himanshu_ag, hinouchi, hiram, hirapon, hisaq, hiyo, hjalmarg, hmano, hnybear, hoangnhat, hodder_fisher, hoegertn, holbage, holcon, holgerc, hollie, home2net, homesys, homunkoloss, honjyogeiger, horward, hosokawa, hoting, houseofvinyl, houshold, howard, howardseymour, howiemnet, hpastorino, hpfuente, hsenot, hubert, hubmartin, hueittest, huffduino, hugozaggo, hum, hummitech, hurgh, hvac_proto, hvymtlihck, hyper2000, i4xcc, iago, iaincaldwell, iamcorey, iam_lab, iamlsh, ianeyon, iangb, ian_harker, ianmbrick, ibehoey, ibot, ibtek, ichilton, icubed, idahobucharest, idurz, iee, iepwong, ifs, igaguri, igfarm, igordutra, igort, igreenmonitor, igy, ijct, ikbenechtben, ikbenra, ikohler, iktp_cobra, ikutest, illrequesens, imanaka, imker, immerito, impbarn, imwinkelried_nacho, ina_geiger, incase, indeed, indiana1977, inejko, infernix, inflame, infracontrol, inigo, innovita, inouk, intrecon, iogate, iot11, iot2020, ipadfanboysf, ipr2603, ipswitch, iradtest, irekh, isaacmalca, ishankarve, ismb_pert, itaico, italianlassie, itinerancia, itsjennkaye, itstpower, ivan, ivanpike, ivorivetta, iw1qif, iwageiger, ixav, iyerx, j02rr, j0nf00, j335, ja2okx, jaap, jaboersma, jackaltx, jack_kelly, jackwu751, jacqueline22, jacques, jaelcio, jagabandhu, jager, jailbreak, jaimedelosrios, jakeloggins, jakevv, jakobmn, jam, jama, jamar, jamesbond007, jamesjl, jameslitton, james_milner, jamil, janadels, janemba, janfer, janisalnis, jansipke, japeq, jarjekk, jarpin, jarrydb, jarsconce, jartweb, jas_, jasonm, jasonpt, jasperp, jasperpants, javalin, jaws75, jayalchemapl, jayblue, jaygill314, jayor, jayyoontokyo, jbranlund, jbridson, jburdy, jcc120102geiger, jchidley, jchristensen, jclarkmm2, jcm, jcmadill1, jcolearyiii, jcorrigan, jcsaez, jcvi19, jcw, jddchatham, jdeprez, jdintenso, jdtoy, jdybowski, je1ldv, jedmonds, jefbear, jefe317, jeffersonjardem, jeffreysun, jelmoni, jengel, jennyb, jensah, jeremychansen, jeremy_jackson, jeremy_lay, jeremyroebuck, jermberg, jeroenjanssen, jerryd, jesmt10, jessie, jesussiero, jf7elg, jfelipe, jflagarde, jflynn129, jfo, jfsolnet, jg1, jgeisler0303, jgorman, jh1mqj, jh2903, jh2pvs, jhardesty, jhbowling, jheerink, jheidtke, jhoekstra, jhojlund, jhorck, jiijiihenkka, jimbobba2773, jimjamjee, jimmycat, jimmyhalstead, jimsr, jim_thorpe, jinosv, jiribarren, jirkas, jjamison52, j_jwalrus, jk7dul, jkaart, jkario, jkassot, jkelner, jl1lcw, jlauber, jlmartinmas, jlv1963, jm1wbbb, jm7ock, jmaxado, jmccrohan, jmcdonald124, jmilkov, jmlambert78, jmourich, jmundt, jn1wyf, jo2vtm, joachim, joakim, joakimekblad, joanlofe, joaquin, jobytaffey, joca133, joe95443, joeandsheila, joebritt, joehays, joeleno, joelfischer, joemoniz, joergpube, john942s, johncfeed, johndaskovsky, johndeere8650, johnmarcel, johnnycirl, johnnyp, johnnywalker18, johnpower, johnprosek, johnr, jojoflyhigh, jollyburnz, jollyjohn, jonandel, jonas4, jonathanfurness, jonathangrinham, jonathanmcgregor, jonbartlett, jonbcmarshall, jondata06, jongeluijs, jongsunlee, jonnyhottub, jonosere, jonsnow, jorge5a, jorgearrua, joris0708, jorngeorg, josefkucera, joseg, joseluisb, josepjc, josetinoco, joshsimerman, journey6061, jp1awp, jpa, jpalombo, jpb123, jpharvey, jpinovations, jposthuma, jq1bwt, jq2gyu, jqnrubio, jr1vmx, jr3qfb, jrdhome, jrfreemantle, jrheling, jrm, jrsikken, jrtappers, jrubio, js1234, js2002, jsaeki, jsc928, jskinit, jsl123, jsokolowski, jswikert13, jt28, jtallak, jth, jtl, jtonline, jtreg, jtuhtan, juanmacuevas, juantomas, juha, juhah, juill, jukka, juliendebeauregard, julietchen, juliogarcia, juliolanas, julpyssel, junhong, juno, justinh24, just_jk, justy, jvanaert, jvangorp, jwalden7, jwaltert, jwgoertz, jwhite, jwpwm28s, k3larra, k4dsp, k4web, k6o, k8, k99, ka00039, ka_32, kaikiwa, kali3146, kalle1303, kalupa, kamageiger, kamakazi34, kamegeiger, kamikazzzi, kaneohe1450, kaoru622, kapros, karaokebats, karelbern, karelv, karlandersson, karli, karlos40, karlp, karpour, karugeiger, kas, kasperkamperman, katherine_sullivan, kattivik76, kawamura, kawatake, kayakpete, kayno, kazekun, kazfmi, kazsan, kazuk777, kbas111, kchassapis, kcr2, kcwerle, kd7eir, kdriver, kdwalton, keaonerock, keith2ei, kenchandler, kencollins, kenji12345, kenmunkholm, kenned, kensama_char, kentropb, kentsigurdsen, kerisaar, kes, kevin, kevinbardos, kevinmoons, keysafe, kfcito, kfury, khai, khaindar, khchiang8, k_housebot, khtong, kiba_geiger, kibu, kidpixo, kiilo, kimjonghun, kimoknowles, kimtec, kimwo, kinetix, kingcutgroover1, kinjinson, kiser89, kitashinjyukump, kitefly82, kitwallace, kiwi64ajs, kjhughes, kjyoo1989, kke_sensor_demo, kkmyktk, kkonis, klausbueker, klbegeberg, kliger, kliux, klouche, klystron, kmcital91, knightmg2012, knobowler, koa16v, koda2, kodamapixel, koitaroh, kolja, koma5, komageiger, korayal, koshigaya, kostasthegreat, kotaro, kouiti, kousukekikuchi, kovalick, kozmix, kpudner, krad, kraushaus, kreggz, krekr, krio, kris0013, kris969, krisirk, kristfin, kristophernguyen, krmped, krogoth, krtrivedi, kryan09, ksr, kstoechkel, ktaylor89, kuato, kuba_feed, kuchida9, kudos1uk, kuju, kuma1950, kupsztal, kuro_pachube, kus, kutterkoeter, kyelsmith, kylegordon, kyomusashi, labjacksupport, lacosaradiactiva, laesq, lahune, laird, lalindra, lam32767, lambda108, lampshade, lapique, larcher, larizgoitia, larmoise, larryfurlong, larrylegend, lars, larsjohan85, lattech, laurensvr, laurentpo, laurlaurland, laust, lawinslow, lawson, lbeck37, lberardinelli, lbergeret, lbergeretjp, lbnl_vandehey, lcflyr, lcltech, lcoudert, lebreeze, leeseunghyun, legalet, legrosteve, leif_karlsson, le_miroir, lemonkey, lenyss, leoboulton, leocolomb, leoferraz, leogaggl, leohernan, leon05, leonardpitzu, leonrische, lermont, leroipele, lesink, lexar, lfleck, lfs, lguezennec, liamfox, lightbulb, lilike31, linusdillon, linuxpunk, liphy, lisax, lissom, lisw2012mashupecologico, lithoijen, littoclime, liveapp5, liveboxhah, lizgreen333, llister61, lmnv, localbay, log320datebs, logicaprogrammabile, logicatronics, loic, lokari, lokomoto, lolg, longarc, lontek, looping, loran133, lord_alan, lorenzobianconi, lostininaka, lovander, lowsky86, lowway, loxone_uk, lp0, lpantunes, lpercifield, lpxav, lsensor, lsimpson, lstg, ltickett, lucaberta, lucavax, luccie_007, lucckini, luigi, luiscano, luispark, luispena15, luisuson, luiz_rodrigues, luk, lukesmith, luminis, lumpmonkey, lunauta, luy_cipher, lvalenty, lvirgos, lykos1986, lytke, m000c400, m1elr, m33s, ma002, ma003, maartenr, maatale, mabikoke, mabru, macascos, macca, macca45, mac_coombe, macegr, macfred, mack, mackeynl, maddox, madelon, madjalapeno, madleech, madshobye, magarcan, magdl, magicmike, magicmonkey, magnoliamanor, magnush, mahieu, maileys, maitake, majland, mak69, malcsmith, mametaro999, mandarinlabs, mandymiur, maneco1227, maniacbug, mani_art, mannel, manpie, mantaraffu, mantra, manu, manugj91, manusvs650, manypoints, manzamb, maptnh, marc0nicole, marcams, marcel, marcelbassman, marcelus, marcioasf, marco64, marcob, marcodomenichetti, marcosfrmn, marcotasto, marcozennaro, margaretgoldsack, mariame, marian, marianoviola, marijosita29, mariorenato, marioum27, markab, markallanson, mark_anderson_us, markbuckley27, markdib, markkropf, mark_mcconnell, marknic, markp, markussurek, markw, markwakeling, markwilson, marnold, marquistaylor, marsflyer, marshman, martgadget, marthaproject, martindamp, martini, martinknutson, martinvilu, martymartin12, martynw, marunga, marupin9, marvings3, marvinlemos, marwoodmetal, masato, masayoshi, mass, masterplanex, mastria, matelodave, mateobernal, materia, matevz, matgyver, matherton, mathew, mathiasw, matjack, matrixtransform, matsekberg, matstherkelsen, matt276, mattbrailsford, mattcallow, matteo, matterfactory, mattganis, matthewp, mattjhonson, mattomatto, mattvenn, mattyw, mauhen, maurocorinaldesi, mausami, mauzer, maverick_56, max256, maxh1981, maximo101, maxsal, mbarker, mbarwell, mbatchelor, mbatnak, mbattig, mbcerro888, mbedmk2r3, mboyd, mbuchmann, mbutz, mcchots, mce, mchavez, mcspidey, mctubster, mcuseeme, mdavid, mdavidlow, mdekruijf, mdknight, mdt005363, mduerig, mechanical_prints2, medicopress, mekatie, melior, melort, mem, memtv, meneer, meneldur, mercedeslanda, mercinat, merlin13, merlin66, merlin_r68, merton, mespet, metawatch, meteo90, meteringmyhouse, mfo, mganucheau, mg_arca, mgenti, mgf909, mgifford, mhanuel, mharizanov, mi4, michael, michaeldu, michaelp, michael_raymond_jones, michaelwilsonnz, michel33, mickess, micky1962, micra, microdev, microera, midcoastmac, midniteandy, migratus, migray, miguelitic, mihajloz, miho, mihovolk, mike2307, mike390, mike57, mikeblanche, mikecando, mikedc, mikeduino, mikefez, mikeinoz, mikejck, mikenl, mikepotts64, mikeshouse, mikethebee, mikhail, mildlymad, miles, milfab2011, milkgirl, millstonebarn, miloslick, min0n, minibart, minorgal, minoru, miquel, mirakonta, mir_guillermo, miska, mistasteve, misuser, mitch, mitreklov, miyasita, mizzen, mjdempseybc, mjh, mjkoster, mjostock, mjpramos, mjrtoo, mjtiern, mkang, mkeim, mkosun, mkrasz, mktacoma, mkufner, ml9aw, mladenjovanovic, mlekwok, mlng, mm989, mmassie, mmosegui, mmr_hrs, mnarita_6, mnicolato, moallen, mobcom, mobilear, moderndevice, mohammedm, mohansen, mohclips, mohsinraza, moisesrc, moixa, moixa_building_banter, moixaoffice, mojganik, mojotexas, moko, molmasse, molnar, monaka, monaka333, monchopena, monitoringdude, monnu, monroyhd, monte1980, monterolabs, montini, monty, moonlighter, moo_wdty, mora, morecat_lab, mores, moritz, mori_uma, morris, mortenbjohansen, mortesq833, mothra78, moumoutea, mpalade, mpaltsacity, mplabs, mp_nara, mra, mranostay, mr_blank, mrevgen, mrfatfingers, mrfmanagement, mrhappy, mrkhnstn, mrod, mrossi, mrparp, mrperez, mrussell42, msantos, msbwilson, mschreuder, mschwanzer, msheridan, mshook, msplan, mspohr, mstaeche, mszabo, mterebessy, mthk1248, mthome316, mths_ie, mtripolt, mudchet, mugi, mungell, muntypro, mutleybones, mvhaperen, mvial, mvtonder, mwaldegg, mwheeler6249, mxjmpbean, myashubh, myougeiger, myowndata, mytenshijp2011, mzt_gig, mzwaterski, _n_, n001, n4spd, n9wxu, nabito, nafis, nagoya, naguara, nakamurabashi, nakazato4gaeiger, nalra, nanana, nanometer, nao_yokohama, narong, narwhal, natamyk, nateful, natewin, nathanknz, nathanvi, natn, naturalfuse, naturalsystems, navarromiguel, navied, nbyoshi, ndygen, neal, necaxa_carlos, nedex, negativeninth, neha, neill, neilnz, nelius, nelly111s, nelsonjt, nemus, neon1, neon_24, neo_tama, nerginer, neroli_w, nestalk, net0040, netcrusher88, netdrag, netduino, neufuture, neuralblade, newmagnetico, newmedia, newquayweather, nezumi, nfcf, nferreyra, ngregory, nhdb, nherriot, ni, nick1020, nickgrant, nickmaynard, nickn5, nickneutrino, nicknick, nickp, nickwallen, nickyte5, nicmarks, nicolaivr, nicolas, nicolepachube, nicomarchais, niconsanchez, niels, nieto0823, nightflash, nightlife31, nikchien, niknkids, niko_sk87, nikospachube, nilivedata, niloxy, nils_westerlund, nim, nimbusgb, ninebladed, niomix, nipunreddevil, nivaone, nj6, n_k, nkildal, nkt002, nkt003, nneil, noahcrowley, noc, nolab, nolonger, nona, nopolabs, norgrenlichfield, norikatsui, normanr, northfield_house, northq, notsosimple, nottingham, novazembla, nrwtaylor, ntaka206, ntshane1, nuclearenergych_berne, nudnik, nuku74, nunonunes, nv0n, nvandegiesen, nwabuking125, nzrob, oas1s, object, oceanwanderlust, odelma, odisej, oe1gca, oeboema, okdkisk, okekubo, okini3939, oktawian, okuyamamasao, olab, olavis69, olbert, oldbooy, olegzee, olicorn, olilam, oliverkellow, olivlogol, olliepalmer, omnomnomsies, omronit, ondo_bot, onefix, ongec1967, onodera_, onuma, opabinia, opprud, optolab, orangemikey, orenopu, oroulet, orsonfrut, osamuh, oseiler, ospdk, osray, osvaldo_hinojosa, otakubicek, otherguy, otomasiku, otter, ottokeyo7, ottoplanet, oum, outdooractivist, ovia, ovow, ovuncozturk, owaisnajam, owi_data, oztds, ozzieg, pa0zhb, pa30ro23, paancrv, pablitogordito, pachkuni, pachube, pachube231, pachube232, pachube323, pachube324, pachube327, pachube329, pachube63, pachubedotcom, pachubehb, pachube_rmd, pachubesfd, pachubesy, pachubez1, pachutest, pagter, pak, paksoft, paladinfello, paleksi, panaut, panobird, pantonvich, paolo_m, pap2k, papaschlumpf, paporas, paracha3, parappa, parkers, parrid, parsonscollab, pascal75, pasrich, passivehousenz, patchman68, patlbs, pat_m_212, patrickc88750, patrickdijusto, paul, paulabellamy, paulb, pauldb38, pauldeng, paulefranz, paulfinch, paulhome, paulisimo, paulmans01, paulnewton, paulnugent, paulstaugaitis, paul_tanner, pavelstorek, pawk, pbcharlie, pbtestvito, pchbmarcaccnt, pculurc, pdarche, pdelprat, pdesrochers01, pdgelling, pe1nsu, peedjy, peerke1969, peggus, peihsienhsu, pema, penguin, pentala, pepahr, pepeceballos, percih70, pereskjo, persidion, pescadito, petecorreia, peteh59, peteratsal, petercrozier, peterhd, peterleong, peter_longley, peterpixel, peterpwc, petertettelaar, peteryuezheng, petescarth, pgeleyn, pgluciano, phelanm, philc2013, philh, philip16888, phillips321, philmonty, philwc, phineasjw, photovor, phubbard, picatchoume, pierreca, pieterjm, pietervanharen, pietroblu, piffle, pikpok, piljapan, pilotos, pim, pineridge, pines, pinfold, pinorabi, pinting, pinyopanpa, pipoa, pippo927, piratstian, pisuke, pitp2, pixldance, pjacob, pjay, pjgonc, pjohns, pk, plan, plcasale, pleech, plego, plenet, pletskud, pllgarcia, plugnfly, plumer, pluvalis, pm2587, pnadal, pochi, pocketscience, poerepoere, poi83, polley, porkoprince, porschefanroel, potents, powerbook, powerman, powermeister, power_meter, poxika, prada, prairiescientist, prengerchristiaan, prensel, preservenergy, prestona, privategeigercountersouthtokyo, priyodas, proq, proteus, prov_000306f16885, provo, prowler, prowntree, psc, psgeiger, pshen3, psi6030, psl, psok, psuen, pu2pe, puccimarco76, pueblo, pulsecloud, punkisgonnagetya, punkracy, pvet, pvkonovalov, pwakeford, pw_ktp, pwsensors, pylonz, pyramidalclark, pyranja, qiktrade, qrper723, quarbe, qubgap, queenvictoria, quflow, quim123, qwbot, r055454ru5, rabeliakoff, rabz, radgoes, radhakanth, radlab, radpa, radsan, raecaldwell, raf, rafaeldenis, rafaeljegundo, rafaellopezgtz, rafalt87, rafta7, raghav, rakesh_roshan, ralight, ralph123, ralphwella, ran123, random, randomrichard, randybal, rangerbob, ranulfo, raputam, rasmusromm, rastef, raster, ratranch, raul7, raulesteves, ravagliano, rbender, rbirkby, rdlabs, rdtn, reciclem, red010b37, redaktore, red_cat, reddata, redjacek, redleg, reggie, remco011, renftel, renlou, resington, restlesslegs0, retofehr, reubenb87, rexkeen, reylan, rfidsolutioncenter, rfmoore, rftechnologies, rga218, rgratwick, rgtenergy, rhihar, rhjacobsen, rhoath, rhughdavies, rhughes103, rhysa, ricardinho26, ricc1170, riccios, richie_sails, richms, richroed, ricklon, ricky, rico, rico83, rido, riduino, rigou, rikp, riotgibbon, riotnrrrd, ris8_allo_zen0, rjdtaylor, r_kamiuma, rkeiner, rlindenberg, rm, rmansfield, rmbranco, rmhoover0, rmitenergymeter, rmotilla, rmp_kamiina01, rmullins, rnakajim, rnitzel, roberthlin, roberto_ifsc, robertson, robhell, robhorton, robinson69, roblewis, robmcgregor, robncd, robo, robokaren, robomotic, roborealm, robot101, robowaz, robreilly, robthompson, robward, rocgiov, rodentnoir, rodomo, rodri16, rodrigo, roeland54, rogerrally, roguelantern, rokokko, rokos, rollos, rolly, romantena, romilbar, romilly, romka0174, ronaldcortez, ronstuff, roowilliams, roryaj, roscoe, rose90m, rosemariestill, rossdargan, rossoreed, rosssharman, rowellw, rowsto, roymurphy, rpkessler, rpm, rportugal, rprinz08, rprolu, rriveros, rsoutlook, rsstaffelstein, rtare, r_test, ruben, rubiojr, rudiratlos, ruk, rulen, rupertb, rupertnoton, rupertprescot, ruralenergy, russinnes, rutgervd, ruy_silva, ruzeki, rw3861, rw950431, rwbshop, rwcherry, ryder, rygh, ryofuzuki, ryuchi, rza_arduino, s2o, s373, sabrinarv, sadesign_energy, sadminko, sai_yaji_post, sakura_gamma, sala_ambiente, salcimmino, sale, salem, salireza, sallyinnorfolk, sam, samih, sampo70, samutley, sandeen, sander, sandoz, san_lic, santi_ago, sanwa, sapeach, sardron, sarosun, saturno, saulcozens, sauliusl, saulwright, saymen, sayrus777, sc84647, scaredycat, scattea, schenck, schneseb, schroederx, scoace, scooter99, scott216, scottadams, scottmc, scrapit85, seacritter, sean, seanmp, seb50, seba73, sebozorde, securitrax, seedpod, seejayess, seeksensor, seewatt, seigel, seiji, sembari, sembee, sendaigeiger, sennet, sensorhome, sensors_wcu_ecet, seppinger, serendip, sergiu, serro, setagayamatsubara, sf324929, sfaweiss, sferber, sferigan, sfewings, sfolea, sg2012, sgrimee, shabubu, shade9, shadow6363, shadowsgo1st, shaf, shaitan, shakespeare, shanemdowd, sheepdog, shigemori, shinbashigeiger, shingeiger, shinjyuku_east_mp, shinkichi, shiogeiger, shore2ship, shortplay, shukong, siberx, siftersam, silasc, silsha, silvanogai, simbarr1, simonbosman, simondaniels, simonem, simonf_, simonf_testing, simongalgut, simonh2o, simonsonjh, simtschitsch, sina, sindono, singman, sintan555, sipem, sirchas, sirius0504, sirkro, sirvieira, sisawat, sistemasorp, sitest69, sitziaimpianti, sjarvis, sjhp, sjhp01, sjmsteyn, sjoerdhuininga, skawakura, skcalanderson, skilr18, skinnyman, sklarm, skpang, skrutor, skunkworksnu, skyer, skyggebjerg, skyley, skysh, slateblue, slavek, sldpachube, sleander, sleepy, sleung, slick, sligotom, slolly, slovon, smart2_ecoeye, smart_ecoeye, smartermeter, smarthomelife, smartimmo_xp, smartlubeck, smartmonspain, smartprojects, smatsung, smazero, smcharg, smithandbyford, smmankad, smokey, smpclough, smuseby, sndrv, snedecor, snim2, snootlab, snow27, snsmoore, sobrighton, soccer, soccou, socratez, soend, softhook, soilboy, sokrcoch, solab, sole69, solspotsystems, somhi, soneca, sonia, soorin, sorenp, sorrow53, southampton_uni, southesk, sovesky, soward, spampaloni, spanos, sparkyichi, sparq, spetz, spir, splunkdev, sprague, spresley, sprokkie, spunjah49, spurgear, spyware, sreckov, sreeramraavi, srg, srnattrass, srogerio, srrobertson, srstiers, ssapporo, ssekim, sstirley, stalvia, staropram, staschler, staticdet5, stationsoftware, statto52, staviloglu, ste, ste1, ste2, steamshift, stefacc, stefano, stefanw, stefbou, ste_furutani, steiner, steliosm, stelna971, stepanb, stephanko, stephenbrown, stephencalcutt, stepir, steve_exton, stevemarple, steven4096, stevenpetros, stevepyle, steve_rhoads, stevieby, stewietee, sthiroux, stijn, stik79, stofakiller, stonehouse, strammelse, stridervc, stroff1, strykejern, strykeroz, stschazs, stuart, stuswen, sugix, sumannoh, sunflower, sunish, sunny_boy, sunohara, superbrudd, supersector, supersidor, susan, sutar, suteakanico, svensimilia, svgribachev, svindler, swalker, swaller, swapper, sweemeng, swemace, swests, swharper, swperry60, syano, sydc, sydkahn, syl954, synekvl, szac, szelma, t2248cf, t4sm, tab00, tacticalspace, tactiledata, tag, tajimi_gifu_mp, taka_otk, takhisos, takissd, tamberg, tamori, tampapoet, tan, tanabe_irb, tandan, taniguchisuguru, tano, tapcity, tarantion, taranulvesel, tarokamo, tasha19903, tasman, tatsu, tautic, tax, tayfly, tbfsolar, tcivs, tcrobotics, tcs, tdrew_data, teamkleijn, techcom_ya, techie2010, technoboy01, technoweenie, techris, techteam, tedc47, teddytel, teemhome, tefono, tegerdine, teggen, tek1229, telejapan, tenas318i, teonebello, teppei, tequilover, teranag, terrarino, terrymic, tervis, test_cosm, testfuji3400, texasdan, tf3hr, tgmaxx, thallaran, tharg0r, thatitguy, thbodart, theador, thebest28, thebigg, thebull, theddle, thegeek, thekidd, thelocks, themayet, theodor_grousopoulos, theorbtwo, theskiff, things44, thirano, thomas, thomaschiroux, thor, thordur, thsha1, tht, thumschi, thwolf, ticiotix_1, tiefpunkt, tigergt, tigoe, timballinger, tim_battersby, timbeach, timdye, timesup, timjanes, timoverwoest, timr788, tingenek, tinodb, tipoca99, tisniwoar, tkampp, tkamsker, tleegaard, tleegwater, tlevine, tlmiii01, tmenting, tmh92, tmiura, tmsmartins, tnp, tobiasd, tobiasmedia, tobor, tobyasz, todbot, todrobbins, tofrisistobee, tokyo_0, tokyohackerspace, tollkuhnator, tomasolsson, tomdudley, tomfanning, tomg, tomm74, tommygun, tomppa, tomrush, tomski777, tomtek, tomvdb, tonimm, tonkos, tonosamajapan, tonyo610, toomey8, toon_c, topcat1001, topsykretz, torakichi, torarne, tornado229, torukitaoka, torvald, toshio_kato_1, tostmann, toto, toveybooks, toxi, toyosugeiger, tp, trachsler, tranceash, trapt, trazzaque19, trecedejunio, treffhetes17, trembo, trentino123, trevorjs, trifouillax, trilix, trinco, trjw, tronoso3, troyb, trystanlea, tshek, tsubasabbs, ttam123, tthath, tulipan81, tumnao, turbotas, turner228, tushara, tuttut, tuuma, tuxtof, tvcable_watching, tweeb, twinkel, tyamageiger, tylerselby, tynetlogging, ubersteve, ucdevel, uchi, uclabpstt, uffy, uh, uh_99, ujjalpathak, ulltin, ultimac, umbrellamap, unforgiven512, unicum, unmaster, uol, upheave, uraiseme, urakarin, urax, urbandrum, urmlife, uscweetie, usuyu, va009039, vahan, valerio, valjean, van, vanepon, vanous, vantoegevoegdewaarde, varmarum, vasqo, vaugrat, vbkolos, vbonogomis, vc0001, vc0002, vddrift, ve3james, ve3suy, vel205, venpixel, vepakari, verodificil, version6net, vicatcu, vicente, victor, victordiazcaballero, videocent, viemedia, vikinge, vikingfr, vikkey, villiers91700, vilts, vincec, vindolin, vintage1951, viosmwsn, visveridis, vjeumont, vk999, vlawson, vmrh, vodka, voodoocr, vores8, vrshow, vtfeecsolarhouse, vvasil01, vvsmartcities, vwtimah, w13700508977, wafflezombie, wakwak_koba, walkerandrew, walterleonardo, walterw, wanderson, wanton, warp, warren333, waruma, waspfeeds6, watchingec, waterforpeople11, watteco, wattson, watzlaf0815, wcshaffer, weatherornot, weathertunnel, webby1234, webdude, webguitou, wecja, weeleebin, weiqi39, wellsuser7, werezak, werner, wes, westfalia1648, wfas303, wgx731, wh00per, whereswally, w hiteblack, whitelabel, whitto, whoun, whrw, wiaker, wificash, wilberforce, wildjoe, wild_luxury, wilfredo, wili98, willbar, willennys, willesks, will_fish, williamchan1110, williamsa3, windfall, windymiller, winitalia, winky, winnaing, winpum, winston, winten, wittgenfrog, wittmer, wjmicrodyne, wjousts, wkoperwhats, wkorfhage, wkshoptest, woberts, wolfbierman, woodenball, woodyhut, workshop2012, wotid, wouter1994, wouterastrid, wovyn, woxunmi, wpiman, wpmax, wrapware, wredgum, wreynoldson, wsmckenz, wsoranno, wtoysaito, wuerml, wujason0416, wvu_cafee, wwogrs, wwolkers, xaky, xarrigo, xaviergiralt, xbillon, xemuel, xianzhu, xiewenqing, xiyentan, xmelekax, xoph, xseadog, xylobot, y2kgray, y2kgrays, yachttech, yaffstone, yamafu, yamaneko, yamashitayuhji, yamusketeer, yan78, yasu, yavi2k, ydirgan, yeison, yeisoneng, yenzamsp, yiannissiakavelis, yk2000, yliu, ym79280, yoda, yokogeiger, yokohamarad2011, yomikdodo, yopero, yoshiakiban, yotommy, youdian001, youdian002, youdian003, youngplanneur, yt_86, ythebaud, ytouchette, yumegahirogaringu, yumoto, yusuf, yvesmonta, z3292328_digitalcomp, z3334844, zachtatum, zagato, zaleo, zalesski, zangai, zanko, zaphod, zaqifathis, zeborgesbh, zecno, zerokol, zerotilt, zeusdog, zhouming, zhulin, zilver, zirogirls, zmanfry, zmatzkin, zoopigzoo, zootman, zurajapara, zuran, zuuted, zws, zzkt, zzyzxx.