University of St Andrews | FOUNDED 1413

How good is my dataset?

Simon Dobson
School of Computer Science, University of St Andrews UK

simon.dobson@st-andrews.ac.uk
https://www.simondobson.org @simoninireland

Overview

- We're interested in sensors and their datasets
  - How can sensor networks be engineered?
- Effects of different choices on datasets
  - Sensor placement
  - Error characteristics
  - Analytic approaches

- My aim
  - Introduce this work, and what we hope it will mean for data-driven systems in general

Acknowledgements

- Partially supported by the UK EPSRC under grant number EP/N007565/1, "Science of sensor systems software (S4)"

- The team who really do the work:
  - Lei Fang, Yasmeen Rafiq, Sven Linker, Blair Archibald, Michele Sevegnani, Mike Breza

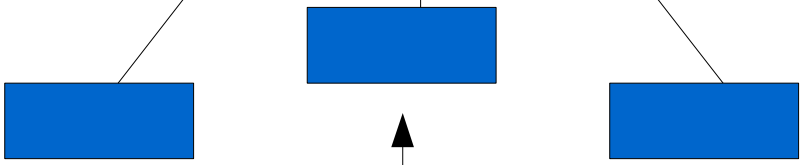
What makes sensing different

- Observing, and often then responding to, physical-world changes
  - Wireless sensor networks: temperature, pressure, humidity, proximity, target-counting, …
  - Internet of Things: the same, with phones!! :-)
- Often building open systems
  - No traditional closed-loop control
  - Mission ~~creep~~ sprint
- Hard to program, in all sorts of ways
  - Errors, decisions, responses, ...

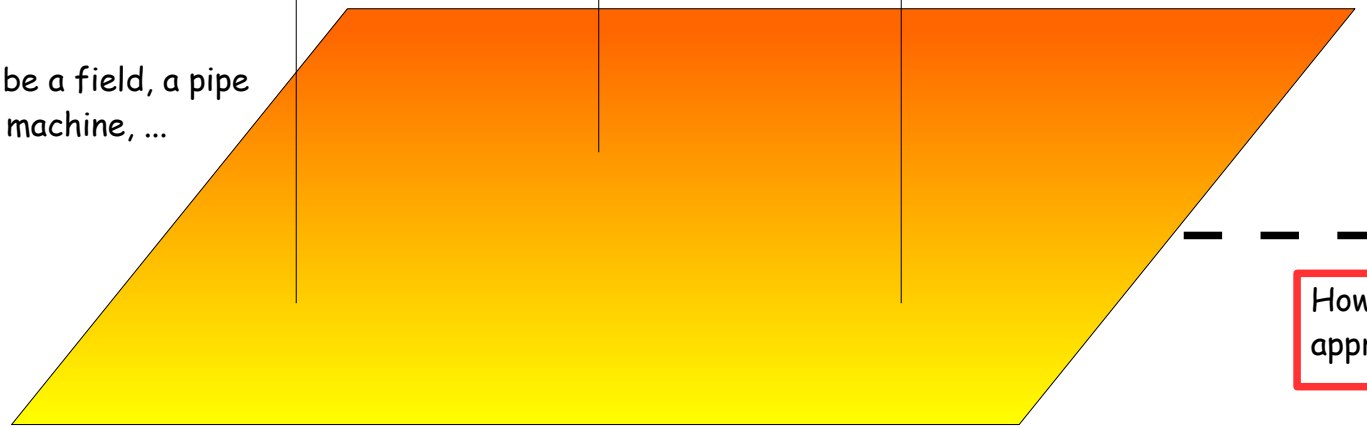
Sensor systems design on one slide

...to form an approximation of the state of the real world being observed

...which are then averaged or otherwise aggregated



Making observations at discrete points



This could be a field, a pipe network, a machine, ...

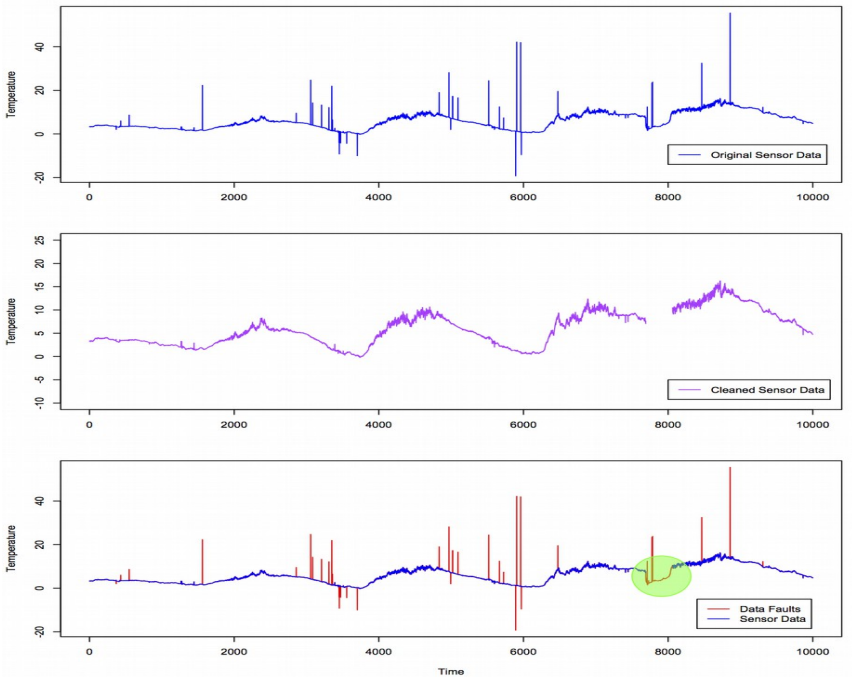
How well does the approximation match reality?

...and we often don't know

- A lot of data collected by sensors is junk
  - An unusual set of failure modes
  - Don't respond to the junk!
- No ground truth
  - Can't compare the *in situ* behaviour
  - Inherent noise

The authors of one famous early experiment (Great Duck Island, 2002) deemed 30—60% of their sensor data to be junk

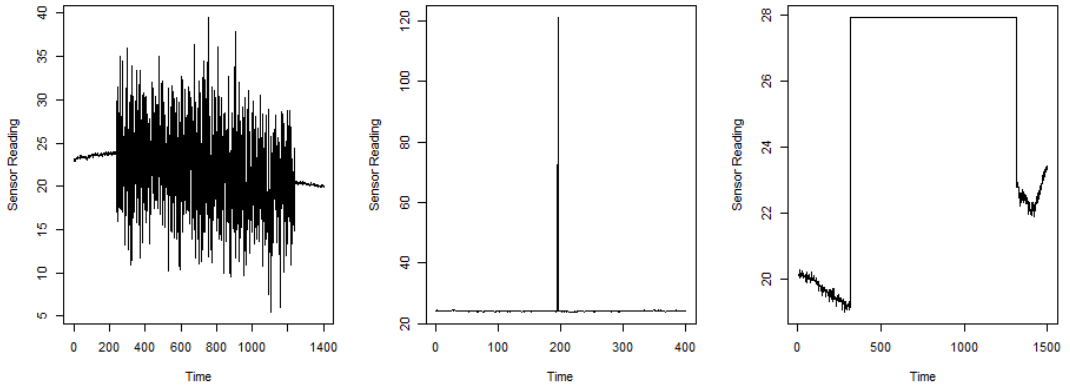
Image from lighthousefriends.com



Data problems

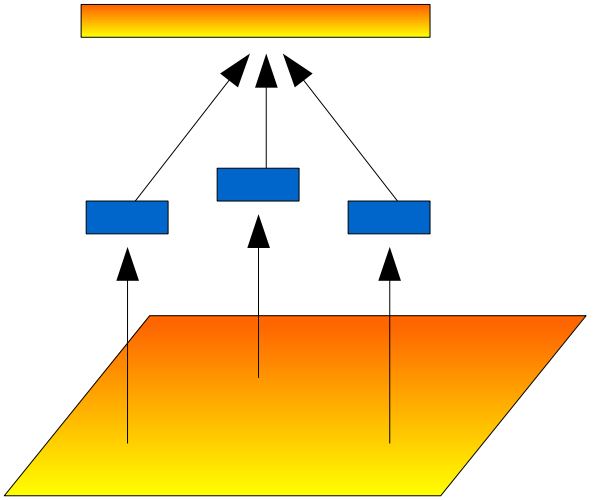
- Errors of different kinds
  - Need to be identified in the data stream

Not always Gaussian, not always stationary



- Physical degradation
  - Decalibration, full and partial failure

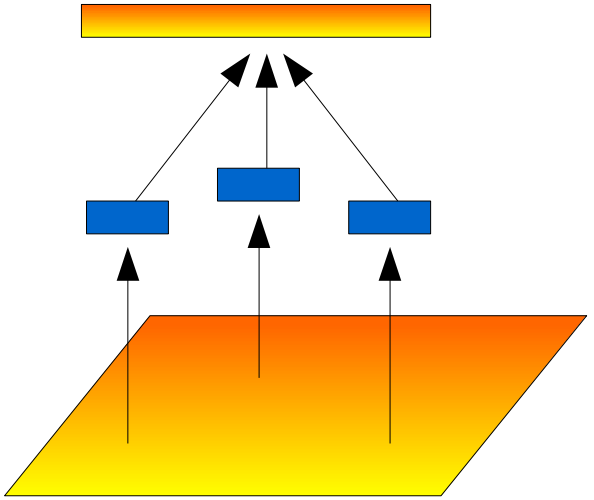
Basic questions – 1

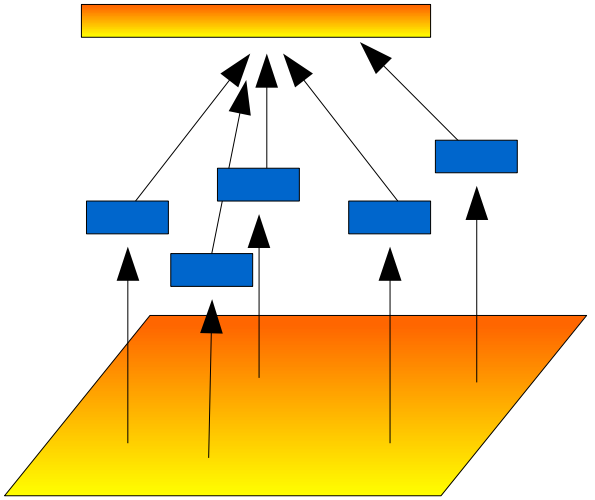


- Given a science or business question, what is the right sensor suite to answer it?
  - Choice of sensors
  - Locations of sensors
  - Mapping the data collected to the answer

Dearle and Dobson. Mission-oriented middleware for sensor-driven scientific systems. J. Int.Serv.App. **3**(1). 2012

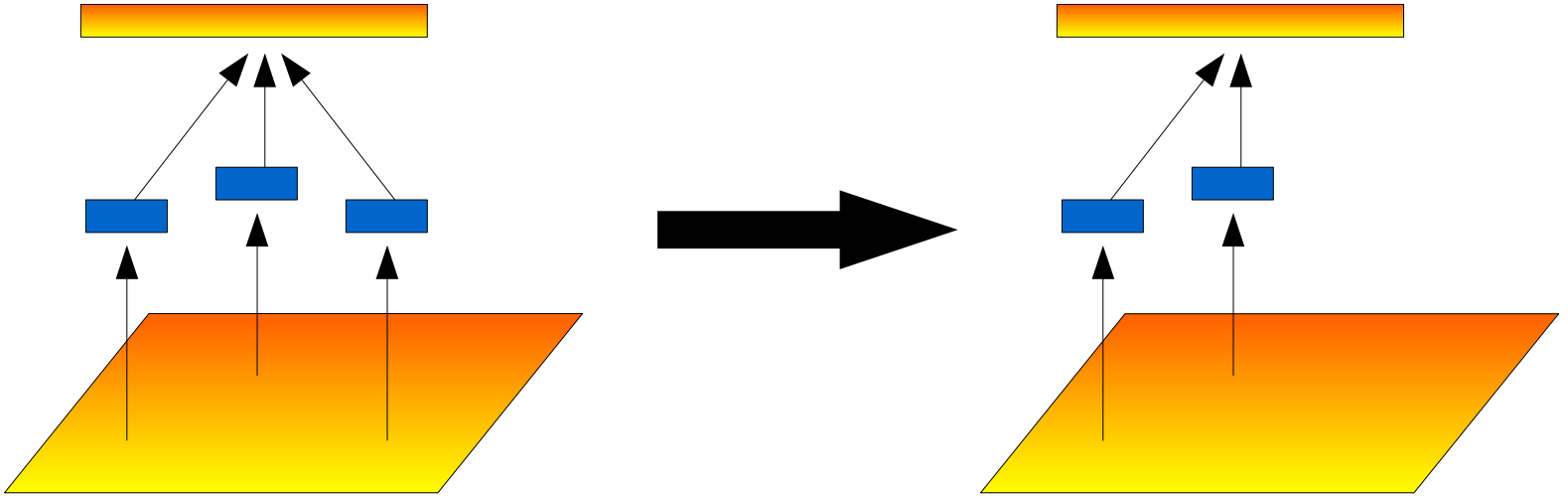
Basic questions – 2





- Given two sensor suites, which will allow more accurate conclusions?
  - Noise and overlap make this hard to answer: more is not always better

Basic questions – 3



- What happens as the suite degrades?
  - Long lifetimes, partial failure
  - How should confidence change?
  - How do the detectable features change?

What we (don't) care about

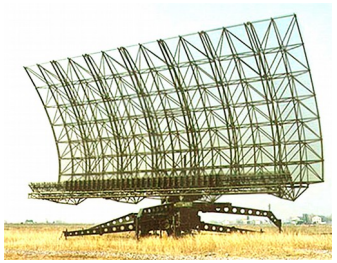
- There is huge prior art in specific domains
  - For example, meteorology
  - Specialised analytics regimes, correction factors derived from long-term observation, …
- We're interested in the general case
  - Inexpert users (in sensing, not in the science)
  - Where we don't have huge surrounding knowledge
  - With things within the designers' control
  - With the data we can observe

So our target user base is the science and industrial communities that don't have access to the scale of funding or sensing of State agencies

science (n): the nagging feeling you get when you realise that that thing you're struggling to understand isn't actually understood by anyone.

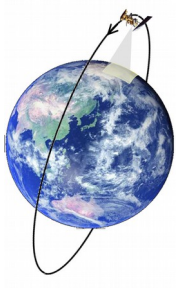
An ontology of sensors – 1

- Point
  - A single value at exactly one point, at an instant
- Pixel
  - A single value for a small uniquely-observed area
- Area
  - A single value for a small area, which might overlap with other observations



An ontology of sensors – 2

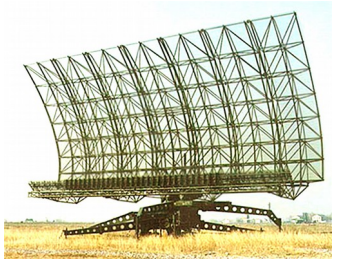
- Temporal behaviour
  - Fixed stream, on request, events, …
- Spatial behaviour
  - Fixed location
  - Trajectory
  - Steerable
  - Attached to something else



Zhang *et alia*. Hardware design experience in ZebraNet. Sensys '04.

A previous experiment – 1

- Target counting
  - A set of sensors that can count (but not identify) "targets" in a space



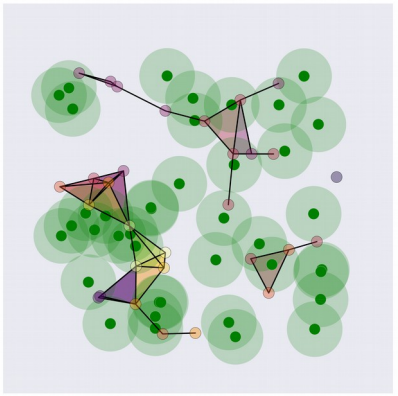
  - The area sensors overlap: how many targets are there?
- Appeal to notions from algebraic topology
  - "Closeness" and "correlation" of observations
  - Form a topological structure that represents these dependencies in the dataset

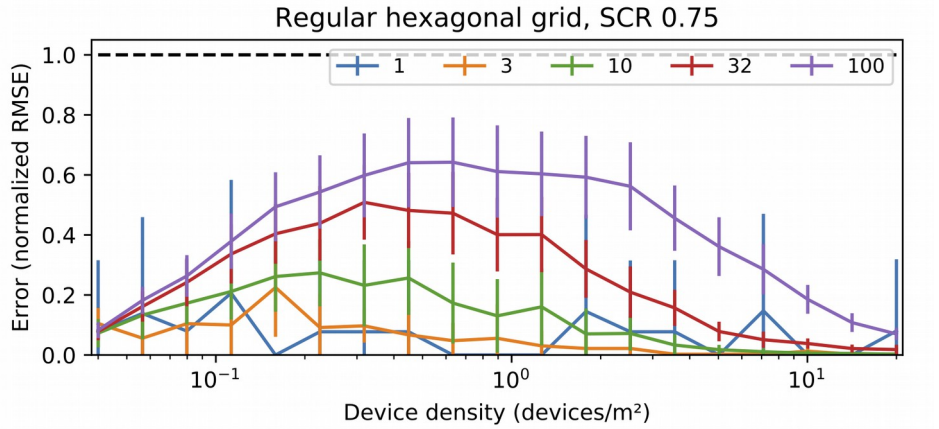
These may or may not correspond to the geometry in which the data was collected

Baryshnikov and Ghrist. Target enumeration *via* Euler characteristic integrals. SIAM J. Appl. Math. **70**(3). 2009.

A previous experiment – 2

- It turns out to be more complicated than this...





- More sensors can *decrease* the accuracy of the count – something the theory didn't predict
  - More data is more confusing – at least until there's "enough" of it

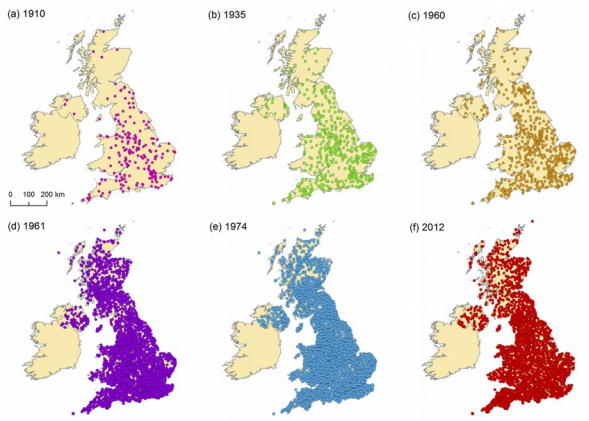
Pianini, Dobson, and Viroli. Self-stabilising target counting in wireless sensor networks using Euler integration. SASO'17.

Limitations

- This work was all done in simulation
  - Hard and expensive to do a real-world experiment at the scale we'd need
- Dataset desiderata
  - Large enough that we can so “knockout” analysis: remove some observations but have enough left
  - Use a similar approach to study different placement strategies
  - Complicated enough to be illustrative
  - A decent understanding of the phenomena underlying the observations

A new dataset: UK rainfall

- An application in environmental sensing
  - Point observations from rain gauges
  - Actual rain gauges have changed over the lifetime of the dataset (1860—present)
  - Not placed for scientific convenience

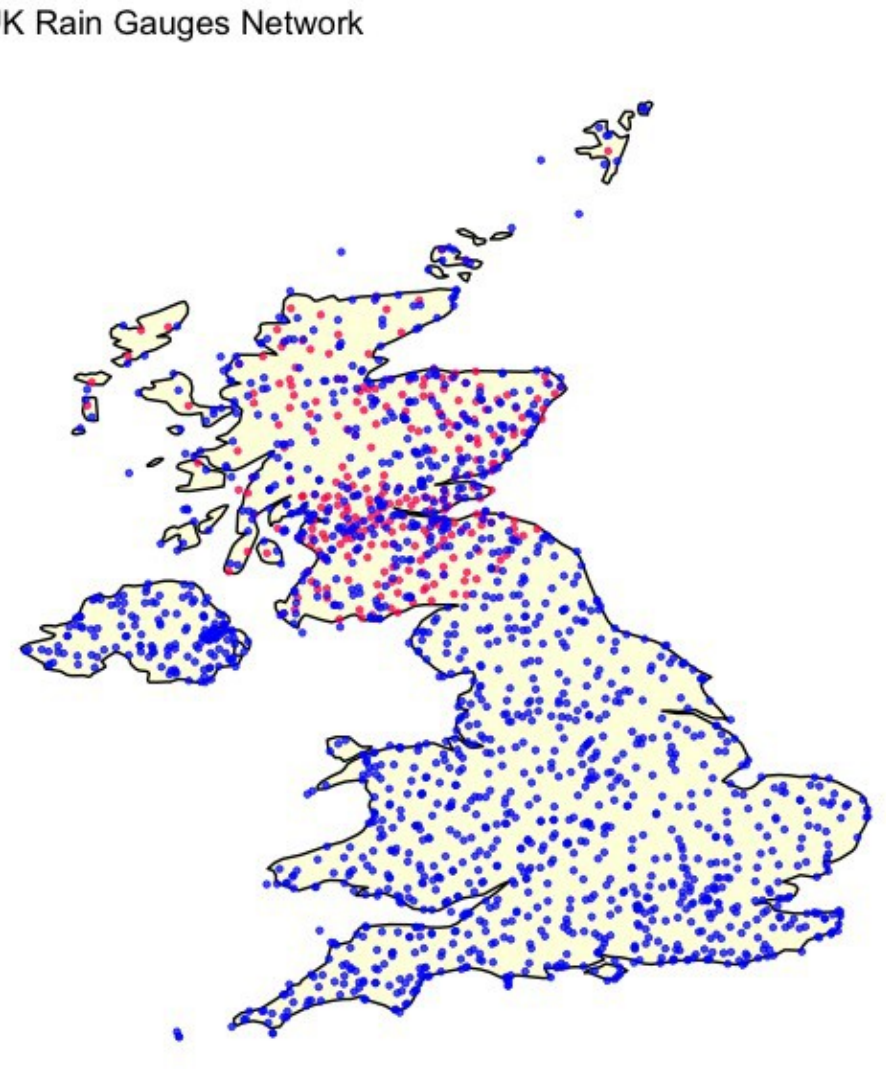


- Want an estimate of rainfall across the UK
  - Interpolation between data points

Keller *et alia*. CEH-GEAR: 1km resolution daily and monthly areal rainfall estimates for the UK for hydrological and other applications. Earth Systems Science Data **7**. 2015.

Richness

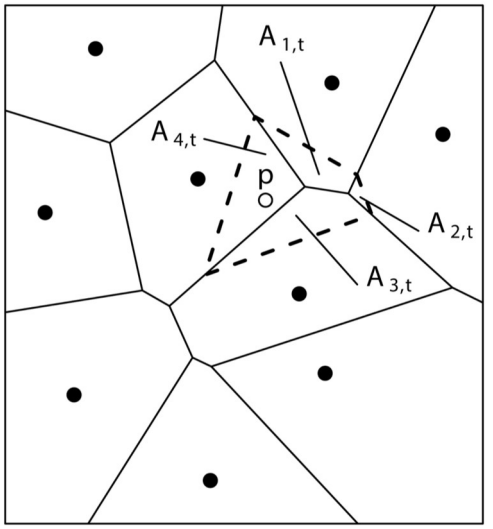
- Scotland has its own, independent network entwined with the larger UK one
- Lots of sample points
- Varying density



The data

- Hourly measurements from several thousand rain gauges
  - Several Gb/day
  - Also some monthly values as checks
- Interpolated between measurements at 1km^2 resolution
  - Weighted average of coverage of Voronoi cells

Often called *tip bucket* gauges

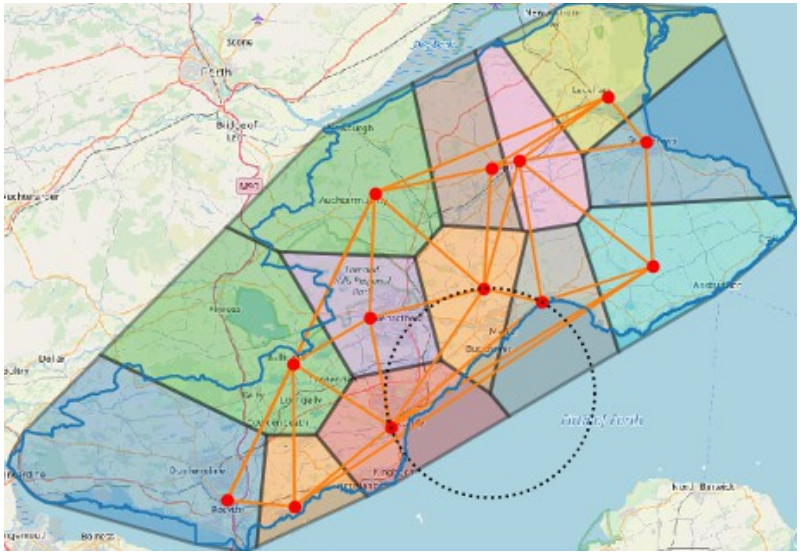


$$w_{i,t}(p) = \frac{\text{area}(T_{i,t} \cap \hat{T}_{p,t})}{\text{area}(\hat{T}_{p,t})}.$$

Called a *Thiessen polygon* in atmospheric science

Reconstruction

- We've reconstructed the data analytics infrastructure so we can experiment with it
  - Construct a topology of adjacency (Delaunay triangulation)
  - What other adjacency topologies are there?

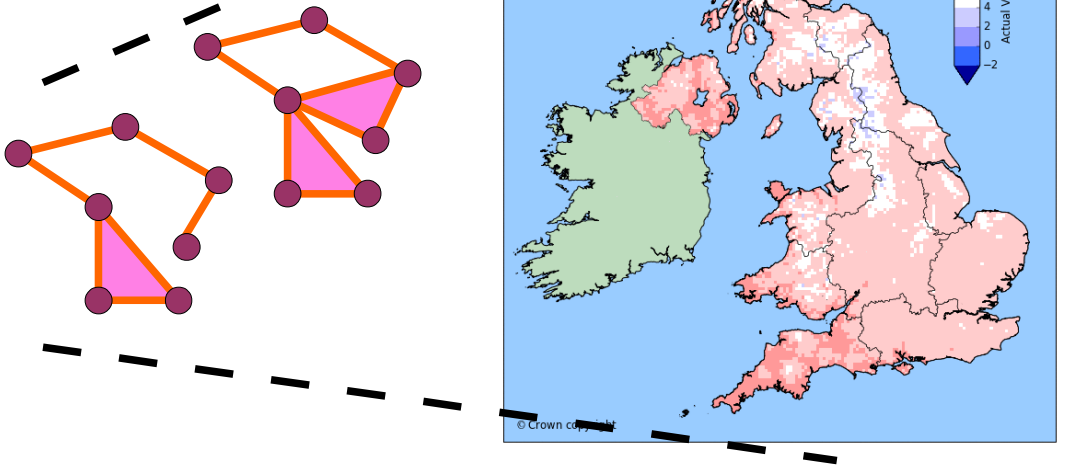


Experimental question: robustness

- What happens when we introduces errors?
  - Amplified or damped?
  - Do different topological calculations matter?
- The different error modes
  - Not Gaussian, don't drop out from averaging
  - Non-stationary, properties change with time
  - Not independent, may depend on topography which isn't captured in the dataset

Experimental question: interpolate

- Reduce sampling from some of the gauges
  - Would expect low impact in areas with dense coverage, larger in sparser areas – wouldn't we?...
- How does placement affect interpolation?
  - The interpolation process is designed to be smooth



The current state of the art is probably that derived from Krause, Singh, and Guestrin. Near-optimal sensor placement in Gaussian processes: theory, efficient algorithms and empirical studies. J. ML. Res **9**. 2008.

Foundational question: learning

- One can also take a Bayesian perspective to this problem
  - Each observation is a sample of the distribution of rainfall
  - Guide sampling to the "most informative" points
  - Use the complex to help decide which points can be sampled (and which inferred)
- An opportunity to explore machine learning in a more structured context, outwith (just) the data

Chandra *et alia*. Bayesland: A Bayesian inference approach for parameter uncertainty quantification in Badlands. Computers and Geosciences **131**. 2019.

Potential impact

- Understand the behaviour of sensor systems more generally
  - Where we don't have the knowledge (or the investment) that we have in meteorology
- If you have a science question:
  - What sensors do you need to answer it?
  - Where do you place them?
  - What analysis techniques can you use to get the most from the data, given its issues?

Current state and future work

- Our analytics infrastructure is just about built
  - Doesn't quite scale to the full dataset at present
  - There are some obvious optimisations
- Knockout and error analysis
  - Failure, placement, decalibration, aggregation, …
- Then look at applying techniques from topological data analysis and machine learning