University of St Andrews | FOUNDED 1413

Understanding sensing from a more formal perspective

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

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

Introduction

- Sensor data is booming
  - Increasing volumes
  - Increasing reliance on what we learn from it
- ...so it's unfortunate that we don't really understand their engineering or interpretation
  - Placement, errors, long-term degradation, …
- This talk
  - Try to clarify my thoughts on how we might improve things with some experiments

Acknowledgements

- Our sponsors:
  - UK EPSRC under grant number EP/N007565/1, "Science of sensor systems software (S4)"

EPSRC
Engineering and Physical Sciences Research Council

University of Glasgow

Imperial College London

University of St Andrews

UNIVERSITY OF LIVERPOOL

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

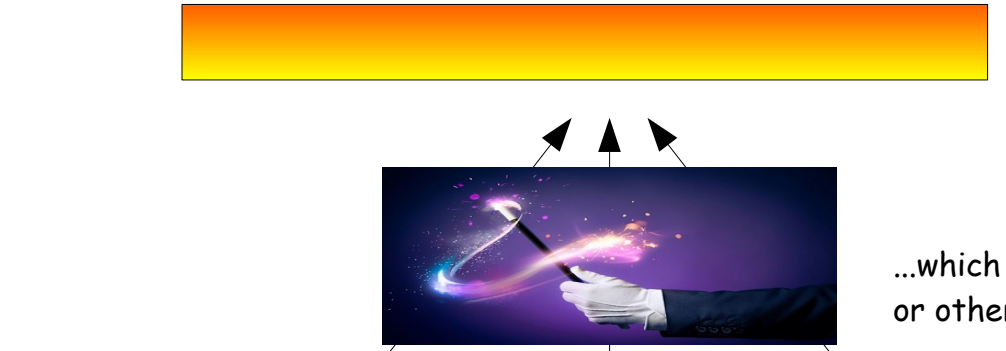
Aggregate programming

- Often a strong relationship between function and physical space
  - For example, crowd steering

Viroli *et alia*. Engineering resilient collective adaptive systems by self-stabilisation. *ACM Trans, Mod. CS.* **28**(2) 2018.

  - Given meaning by a spacetime structure that relates space to observation
  - Want to relate the structure of space and the observations of it directly to control

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

Make observations of some kind, with errors and omissions

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

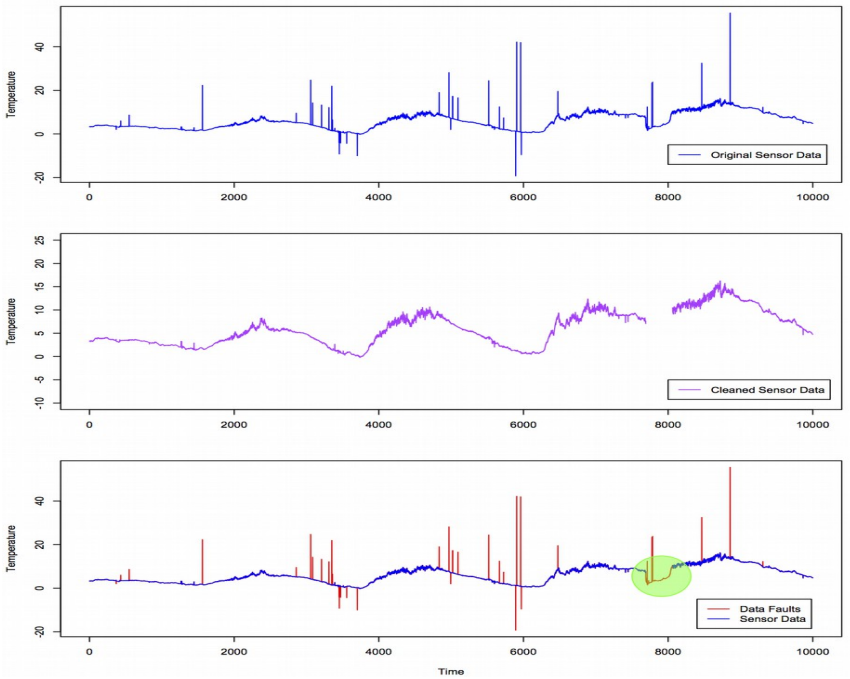
How well does the approximation match reality?

...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

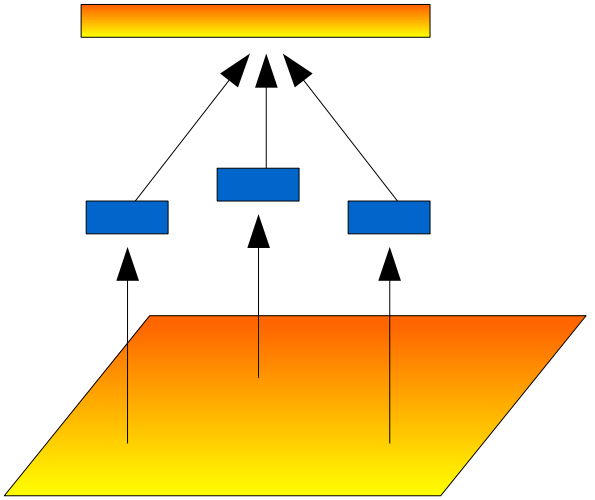
The authors of one famous early experiment (Great Duck Island, 2002) deemed 30—60% of their sensor data to be junk

Image from lighthousefriends.com



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

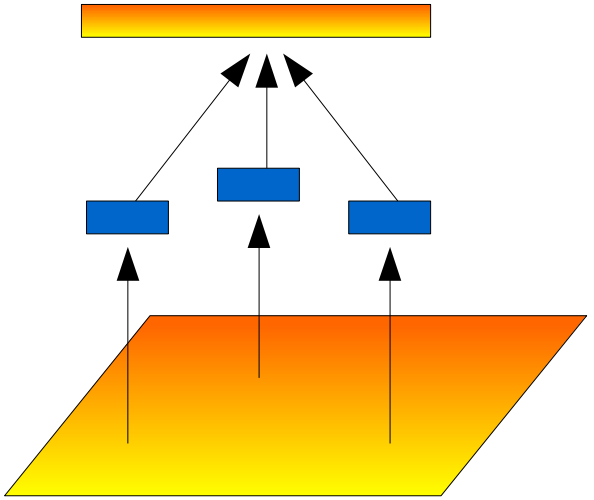
Basic science questions – 1

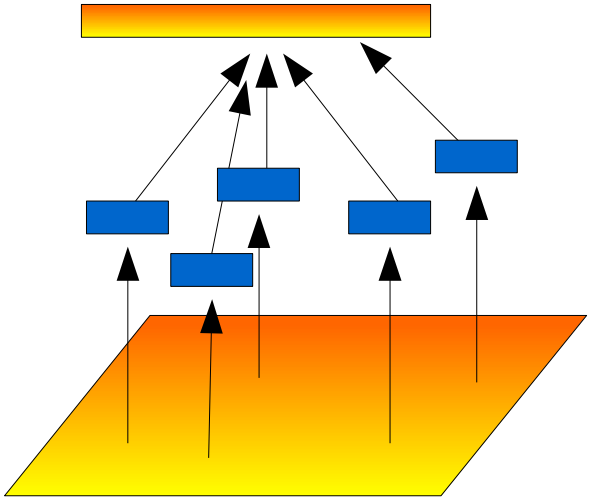


- Given a science (or business) question, what is the right sensor suite to answer it?
  - Choice of sensors
  - Locations of sensors
  - Mapping the data collected to the answer

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

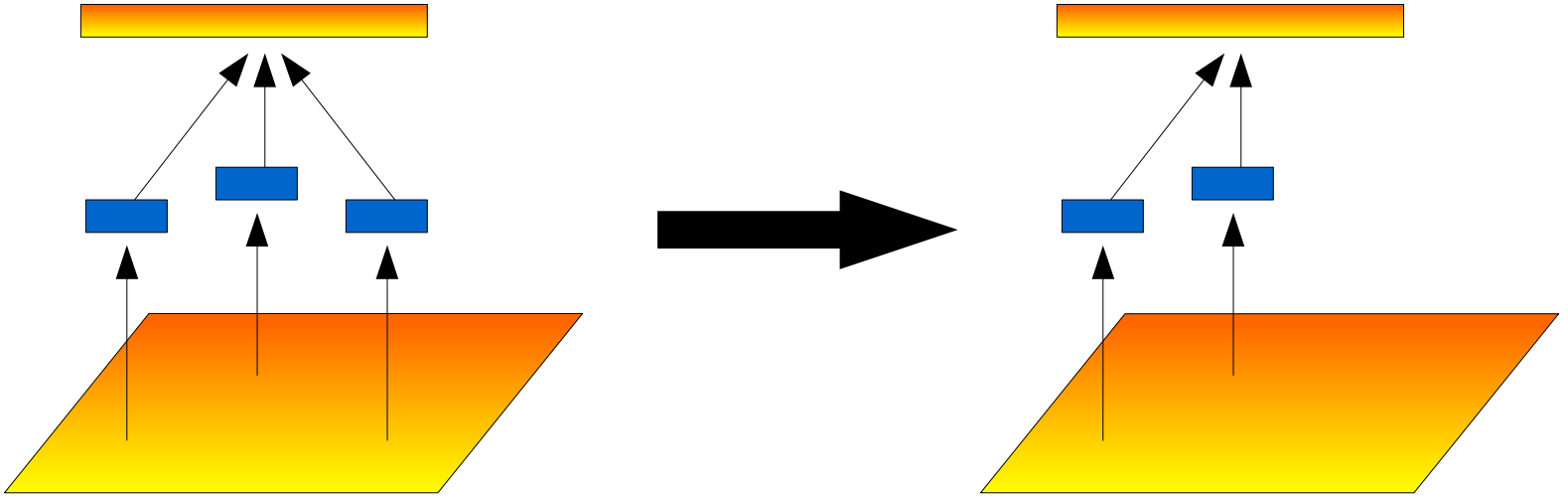
Basic science questions – 2





- Given two sensor suites, which will allow more accurate conclusions?
  - You often don't get to decide anyway
  - Noise and overlap make this hard to answer: more is not always better

Basic science questions – 3

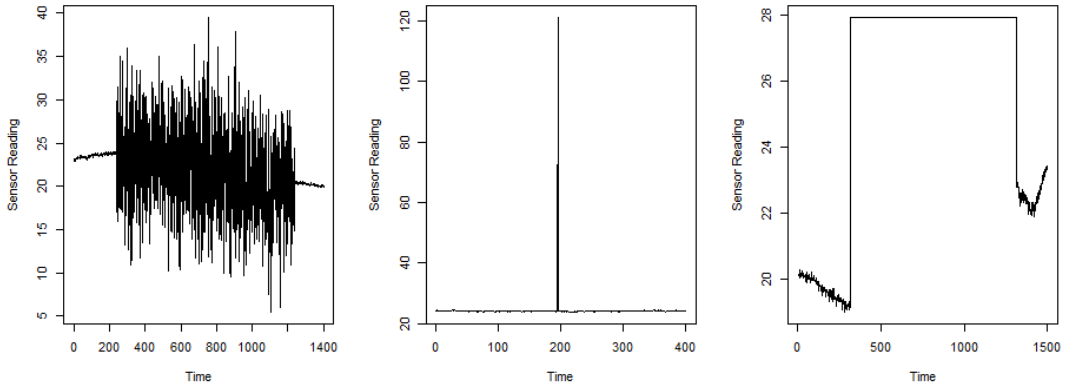


- What happens as the suite degrades?
  - Long lifetimes, partial failure
  - How should confidence change?
  - How do the detectable features change?

The model has no intrinsic value: its value comes from how we can use it, so we need to understand the effects that changes have on interpretation

Data problems

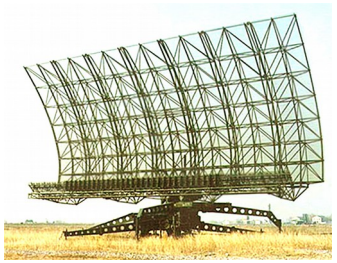
- Errors of different kinds
  - Need to be identified in the data stream



- Physical degradation
  - Decalibration, full and partial failure

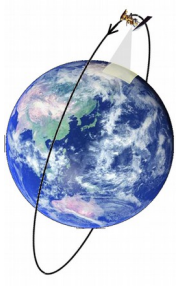
An ontology of sensors – 1

- 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.

Macro properties

- Correlation
  - Observations that are close in space or time will tend to be correlated
  - This introduces notions of *adjacency* into the dataset
- Hysteresis
  - A lot of phenomena change quite slowly relative to the speed of observation

Confusing yourself

- Any kind of dependency introduces the possibility for confusion
  - Two or more plausible, but different, interpretations
- Particularly a problem for sensor fusion
  - Area sensors and/or multiple datasets
  - The correlations can either compensate for errors and omissions, or reduce certainty

The topology of data

Topology! The stratosphere of human thought! In the 24th century it might possibly be of use to someone...

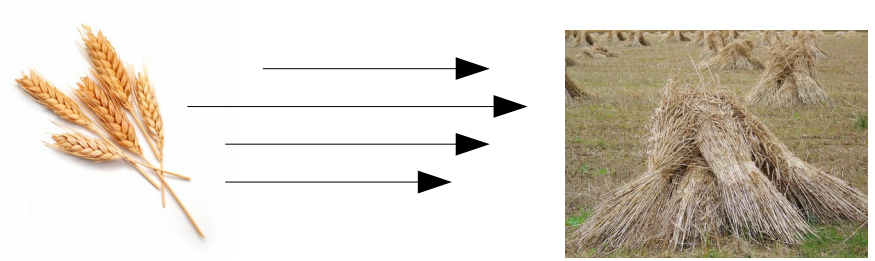
– Alexander Solzhenitsyn, *The First Circle*

- (Of the data, *not* the network)
- A description of the relationships between observations
  - Limits on variability between "adjacent" points
  - Topological data analytics
- A set of tools for analysis
  - Specifically, we've become interested in *sheaf theory*

Gunnar Carlsson. Topology and data. Bull. Amer. Math. Soc. **46**(2). 2009.

You already know sheaf theory – 1

- A sheaf of a bundle of stalks



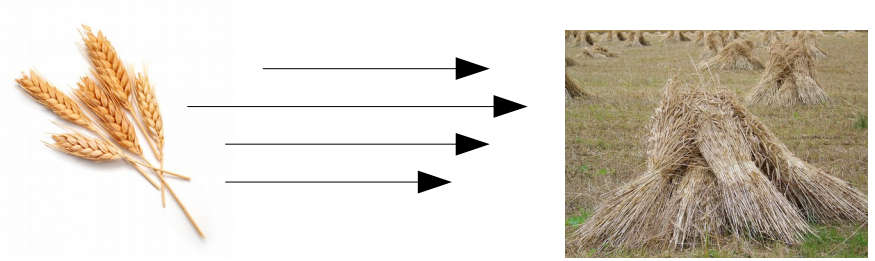
You already know sheaf theory – 2

- A sheaf of a bundle of stalks

The stalks lie in relationship to each other: they don't head off in random directions, they vary smoothly

You already know sheaf theory – 3

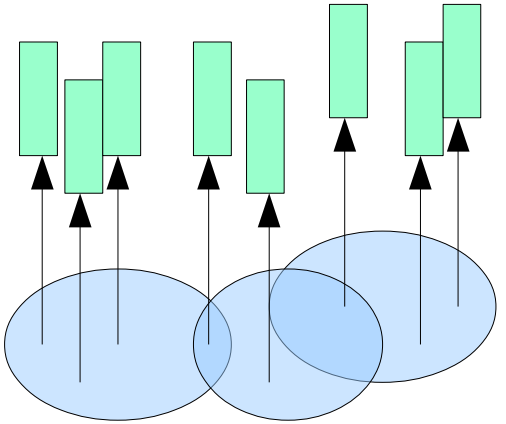
- A sheaf of a bundle of stalks



The stalks lie in relationship to each other: they don't head off in random directions, they vary smoothly

- A sheaf of X over Y
  - To each point in Y, continuously attach some object X
  - Take sections through these objects

A stalk sits over each point in the base space



We need some additional machinery to make everything glue together consistently

Example

- The rainfall over a landscape
  - A sheaf of subsets of $\mathcal{R}$ over $\mathcal{R}^2$
  - The base space $\mathcal{R}^2$ forms a grid of points
  - Each stalk subset of $\mathcal{R}$ is the possible rainfall in mm
  - A section through the sheaf selects a value from each stalk

Discussion

- This model is clearly un-implementable
  - We don't have measurements at every point
  - We have a sparse dataset of values at specific points within the base space $\mathcal{R}^2$
- Continuous systems are a pain for computers
  - Arbitrary equations glued together
  - Often difficult to deal with effectively – or check

It's still useful to think about this sort of model, though, as the *continuum limit* of the sorts of model we'll discuss

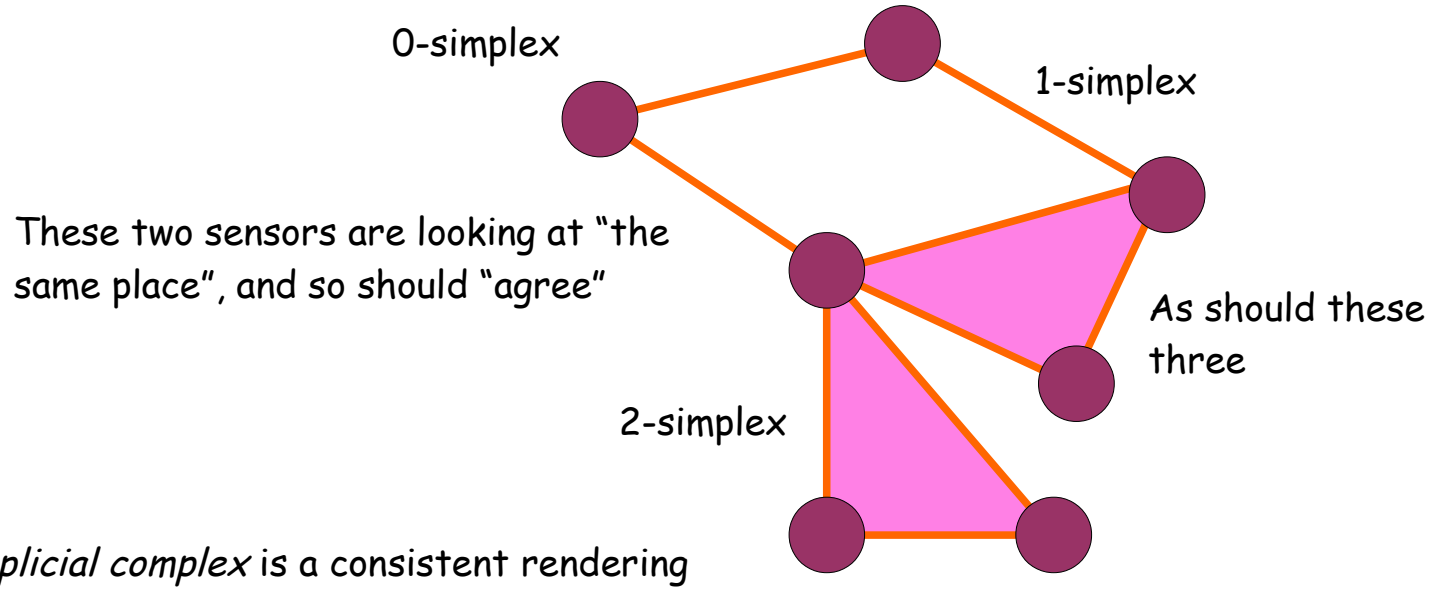
Continuous *vs* discrete

- Fortunately you can discretise continuous sheaf theory very effectively
  - Go from surfaces to *simplices*, a generalisation of graphs that allow higher-dimensional structures
  - A simplex captures the constraints and consistency conditions between observations
  - ...in ways we don't really understand yet, but form the basis for some experiments

How good an approximation a discrete sheaf is to a continuous one, for example, isn't well-studied

Representing dependencies

- The topology of data
  - Connect observations that are "close" – correlated or otherwise constrained
  - Physically close, connected by pipes, separated by mountain ranges, ...

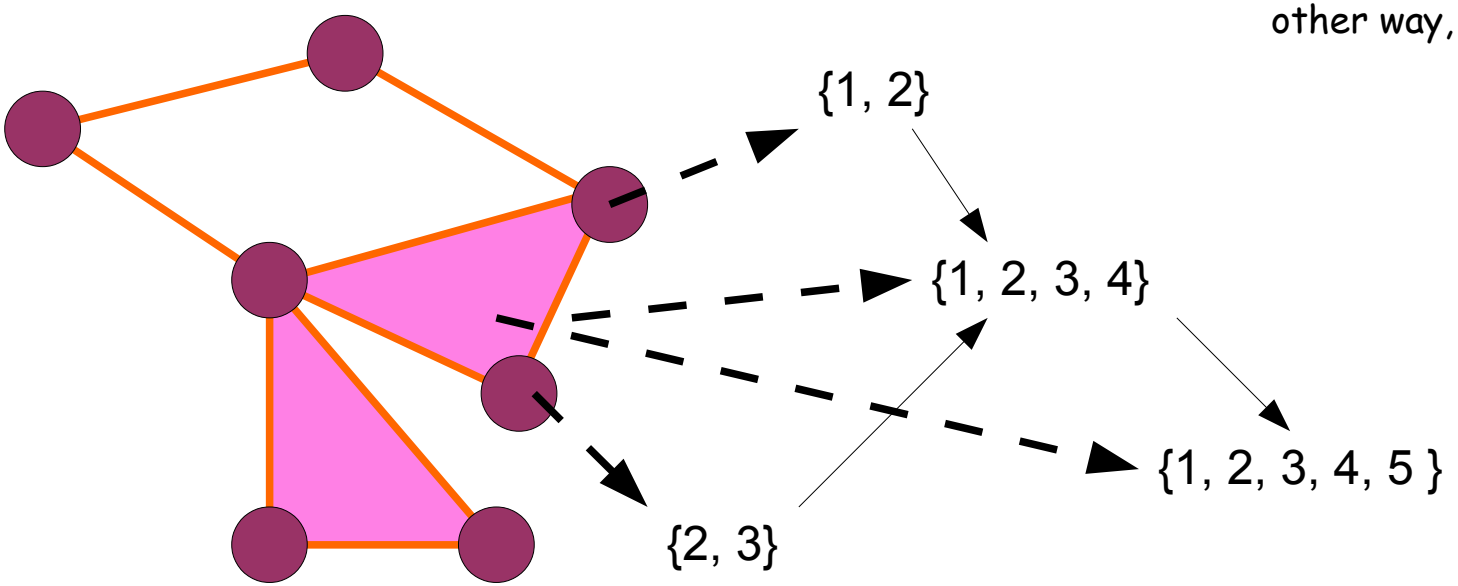


A simplicial complex is a consistent rendering of the sensor observations as simplices

Sheafification

- Associate a stalk to each simplex, together with *attachment maps* from a stalk to the stalks on higher simplices of which it is a face

If the mappings go the other way, we get a *cosheaf*

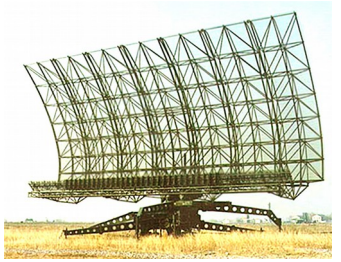


The inability to create a section indicates data inconsistency – which can be fixed by relaxing the interpretation

- A section is a selection of values at 0-simplices such that the diagram commutes under the attachments

A previous experiment – 1

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



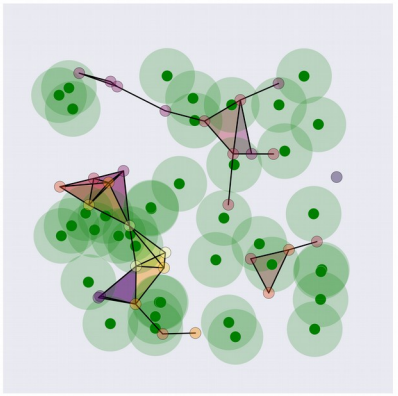
  - The area sensors overlap: how many targets are there?
- Simplicial model
  - Form a simplicial complex from the overlaps in observations
  - "Integrate" the counts to build an estimate

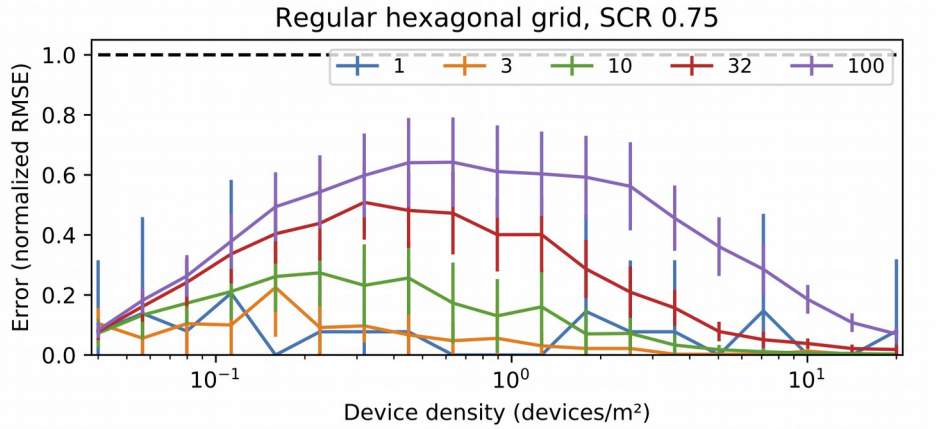
Integration in continuous domain
= summation over simplex values in a discrete domain

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...



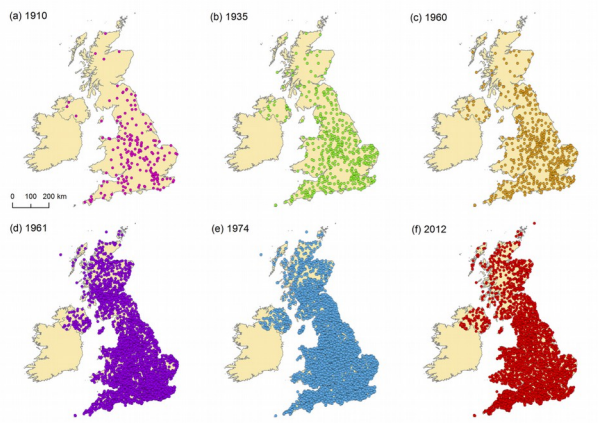


- More sensors can *decrease* the accuracy of the count – something the theory didn't predict
  - More data is more confusing – at least until there's "enough" of it

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

A new experiment

- 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



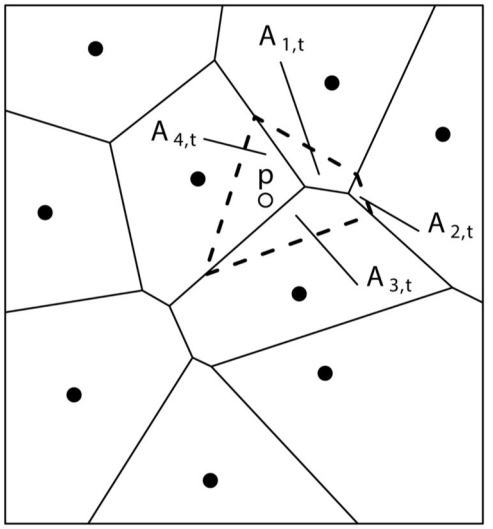
- Want an estimate of rainfall across the UK
  - Interpolation between data points

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.

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



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

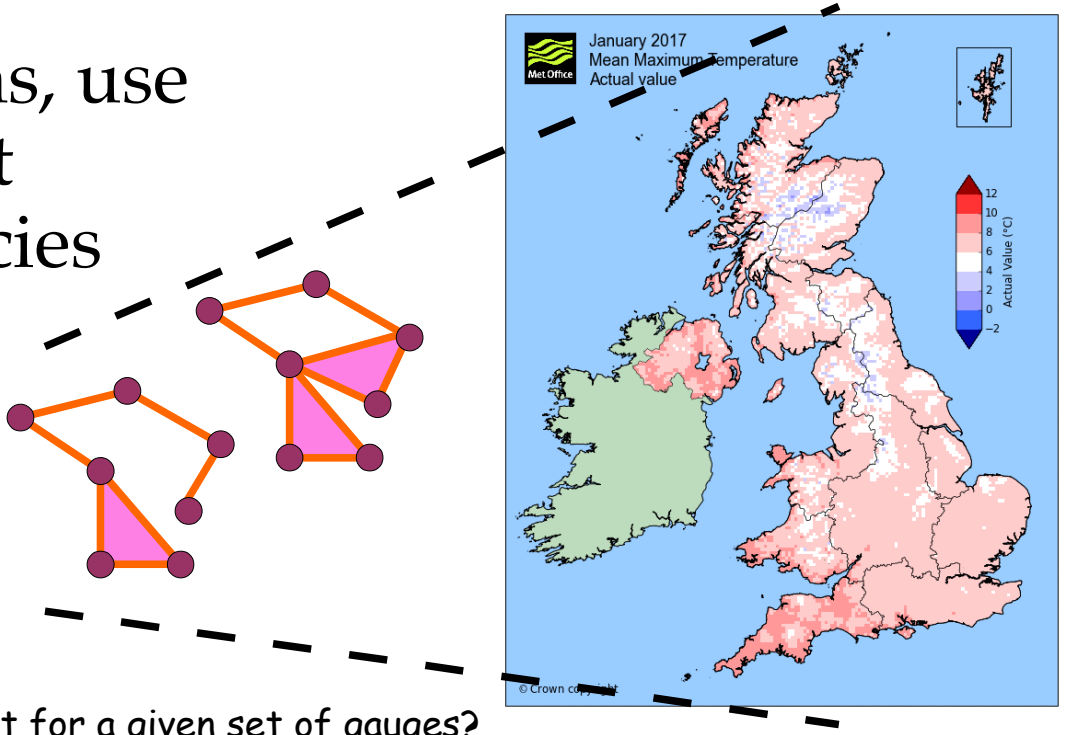
Called a *Thiessen polygon* in atmospheric science

Experimental question: robustness

- What happens as one removes gauges? Or introduces error?
  - These are "point" sensors in our ontology
  - The interpolation process is designed to be smooth
- Reduce sampling from some of the gauges
  - Would expect low impact in areas with dense coverage, larger in sparser areas
  - What is the effect?

Experimental question: interpolate

- The interpolation process is entirely divorced from the underlying landscape
- Can we use dependencies to improve the result?
  - Look for correlations, use form a complex that captures dependencies
  - Use the richer structure to compensate for reduced data



What is the best placement for a given set of gauges?

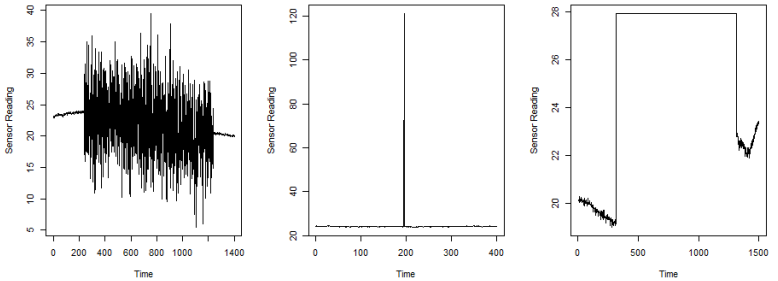
Foundational question: learning

- One can 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.

Foundational question: errors

- Can we build sheaves that let us compensate for errors in data?
  - Noise, transients, swings
  - Adding more structure to the stalks



- We need to look at this because of the non-Gaussian nature of the errors
  - They don't drop out through averaging

Future directions

- Theoretical/experimental
  - Do these techniques work? Do they give us anything (apart from sore heads)?
- Programmatic
  - Can we use this approach to enrich aggregate programming?
  - More structure into the space and its observation might lead to more predictable aggregate programs

Audrito *et alia*. A higher-order calculus of computational fields. ACM Trans. Comp. Logic **20**(1). 2019.