ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

@RSS_DSS, #DSSLaunch

The Industrialisation and Professionalisation of Data Science

RSS Data Science Section

Launch Event (Recorded)

19th June 2017

Agenda

- An introduction to the DSS – Andy Garrett
- 12 questions around the industrialisation and the professionalization of Data Science – Harry Powell & Fran Bennett
- An illustration as to how the DSS might address a data science topic – Martin Goodson
- Three short invited responses – Ioanna Manolopoulou
    - Tom Smith (MD, ONS Data Science Campus)
    - Juan Manuel Hernández (Global Analytical Innovations Director, Kantar Millward Brown)
    - Sofia Olhede (Scientific Director, UCL Big Data Institute)
- RSS president, Sir David Spiegelhalter, will respond on behalf of the Society
- Open Q&A session – Ioanna Manolopoulou
- Close and next steps – Andy Garrett

The event will be followed by a Networking Reception sponsored by

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

EvolutionAI

Data Science Section Remit

To be a professional body that represents data scientists in the UK. The section will organise meetings for a broad range of attendees and generate outputs that are aimed at:

- Promoting good practice by addressing what good Data Science looks like (with exemplars) and what it does not look like.
- Promoting the statistical aspects of Data Science / re-enforcing the statistical framework
- Being a trusted voice on Data Science for employers, including inputting to consultation exercises
- Supporting the Data Science community throughout the UK
- Supporting the pipeline and career development of data scientists and statisticians by elevating skill sets to work in the modern world
- Supporting important emerging topics such as ethics, privacy, algorithmic responsibility and personalization - lifting the quality of the conversation
- Fostering multi-disciplinary connections and the exchanging of ideas

DSS Committee Members

Fran Bennett
Simon Briscoe (Council representative)
David van Dyk
Andrew Garrett (Chair)
Martin Goodson
Mark Girolami
Ioanna Manolopoulou
Giles Pavey
Harry Powell
Richard Pugh (Meetings Secretary)
Matthew Upson (Secretary)
Leone Wardman
James Weatherall (Vice Chair)

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

@RSS_DSS, #DSSLaunch

The Industrialisation and Professionalisation of Data Science

RSS Data Science Section

Launch Event (Recorded)

19th June 2017

Industrialisation and Professionalisation

- Data Science is a cottage industry, driven on personal experience, expertise and exploration.
- Data Science transforms analytics from being a peripheral reporting activity and a cost centre to a core revenue-generating part of the business.
- Organisations that succeed with this transformation will need
  - Clarity: Data Science as an institutional activity with a coherent structure and a common understanding of itself.
  - Strategy: Data Science able to credibly show both value and delivery.
  - Implementation: The practical challenges of introducing Data Science to a business.

What is Data Science and how do you do it?

- What does great data science look like?
- What does a good data science workflow look like?
- What kind of problems can be addressed by data science?
- What is a data scientist's responsibility to wider society?

Who are Data Scientists and how do you become one?

- What are the characteristics of the ideal data scientist?
- What career paths are available to a data scientist?
- How can data scientists measure the value they create?

How should Businesses initiate Data Science activity?

- How should an organisation start a data science function?
- How should data science fit into the structure of an organisation?

How should businesses organise Data Science activity?

- How should business practices change to make a success of data science?
- What do executives and managers need to learn about data science?
- How can an organisation build a coherent data science capability from a collection of data science projects?

The industrialisation of the RSS DSS

In proposing these 12 Questions RSS DSS is providing a forum to explore them and around which the industrialisation of Data Science can coalesce to make UK world leader in this growing field.

Get involved on the Slack channel:

rssdatascience.slack.com

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

EvolutionAI

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Preventing simple errors in data science pipelines

Martin Goodson
Chief Scientist, Evolution AI
@martingoodson

EvolutionAI

What is this talk?

An example of a blog post that the DSS could publish

Aimed at junior data scientists

This is a draft (and just an illustration)

With input from

Detlef Nauck *Chief Researcher for Data Science, BT*

Piers Stobbs *Chief Data Officer, MoneySuperMarket.com*

Partha Bose *Data Science Controller, Argos*

Caroline Hargrove *Technical Director, McLaren Applied Technologies*

Which of the 12 questions?

'What does a good data science workflow look like?'

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Structure of MsbA from *E. coli*: A Homolog of the Multidrug Resistance ATP Binding Cassette (ABC) Transporters

Geoffrey Chang* and Christopher B. Roth

Multidrug resistance (MDR) is a serious medical problem and presents a major challenge to the treatment of disease and the development of novel therapeutics. ABC transporters that are associated with multidrug resistance (MDR-ABC transporters) translocate hydrophobic drugs and lipids from the inner to the outer leaflet of the cell membrane. To better elucidate the structural basis for the "flip-flop" mechanism of substrate movement across the lipid bilayer, we have determined the structure of the lipid flippase MsbA from *Escherichia coli* by x-ray crystallography to a resolution of 4.5 angstroms. MsbA is organized as a homodimer with each subunit containing six transmembrane α-helices and a nucleotide-binding domain. The asymmetric distribution of charged residues lining a central chamber suggests a general mechanism for the translocation of substrate by MsbA and other MDR-ABC transporters. The structure of MsbA can serve as a model for the MDR-ABC transporters that confer multidrug resistance to cancer cells and infectious microorganisms.

coproteins, which have
into a single polypepti
codes a half transporte
membrane spanning re
MsbA is assembled as a
molecular mass of 129
ysis indicates six mem
with the NBD located
of the cell membrane (
the transmembrane dor
transport substrates acr
ABC, which is the hall
transporter family and
couples the energy of
strate translocation. Al
tures of the histidine
maltose transporter (M
enzyme (Rad50), and
no acid transporter fro
aschii (MJ1267) have
structural basis for
through the cell membr

The structure of
general architecture
transporter family, a
derstanding of the
mechanism that mo

RETRACTION

Post date 22 December 2006

We wish to retract our Research Article "Structure of MsbA from *E. coli*: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters" and both of our Reports "Structure of the ABC transporter MsbA in complex with ADP•vanadate and lipopolysaccharide" and "X-ray structure of the EmrE multidrug transporter in complex with a substrate" (*1*–*3*).

The recently reported structure of Sav1866 (*4*) indicated that our MsbA structures (*1*, *2*, *5*) were incorrect in both the hand of the structure and the topology. Thus, our biological interpretations based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I–) to (F– and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (*1*–*3*, *5*, *6*) had the wrong hand.

RETRACTION

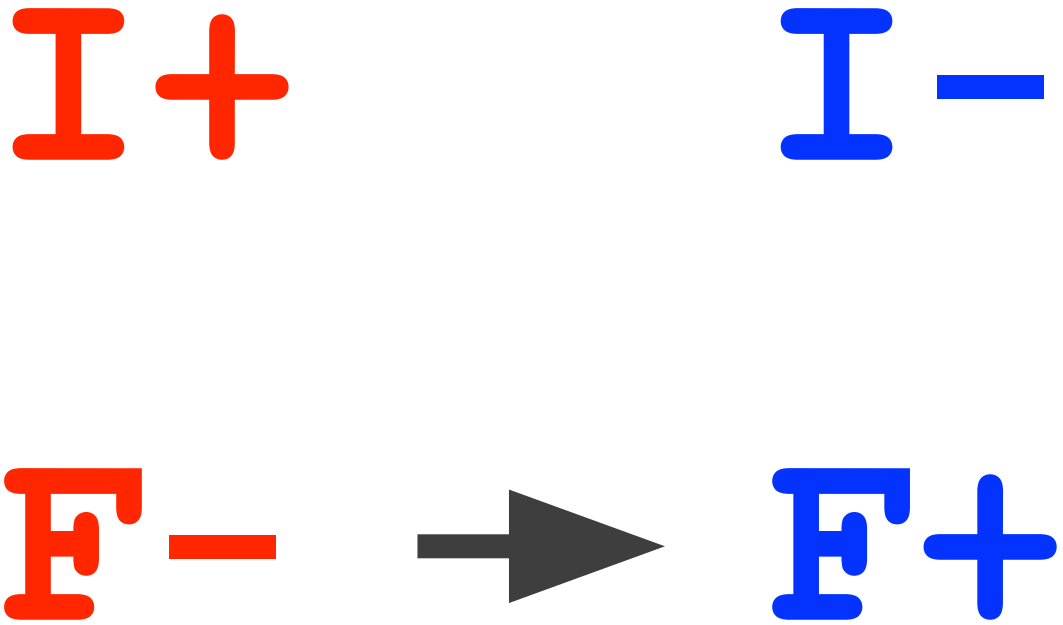
Post date 22 December 2006

‘An in-house data reduction program introduced a change in sign for anomalous differences. This program […] converted the anomalous pairs (I+ and I–) to (F– and F+)…’

tions based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I–) to (F– and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (*1*–*3*, *5*, *6*) had the wrong hand.

What went wrong?



What went wrong?

I+ I−

F+ <=> F−

Paper cited over **750 TIMES**
before its retraction
5 YEARS after initial publication

Baggerly and Coombes

Retraction of **four papers** by Anil Potti, a researcher at Duke University.

Three clinical trials were shut down

Potti was forced to resign.

What went wrong?

'*The list of genes was shifted with respect to the expression data, so that the one did not correspond with the other.*'

'Most mixups involve simple switches or offsets'

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

A real data science project

Aim: discover individuals at high risk of a certain event

The team used a sophisticated algorithm

Very high predictive accuracy

The model was scheduled to go into production

Six months later

They were predicting the wrong outcome variable.

The correct outcome variable only had 35 events

Failed project

What went wrong?

‘It's machine learning - you don’t need to interpret the model’

Didn’t sanity check with with domain expert

Should have kept it simple

Avoiding Failure: Best Practice

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Example of ‘simple errors’

Column switched

Column mislabelled (or misunderstood)

Column not aligned properly

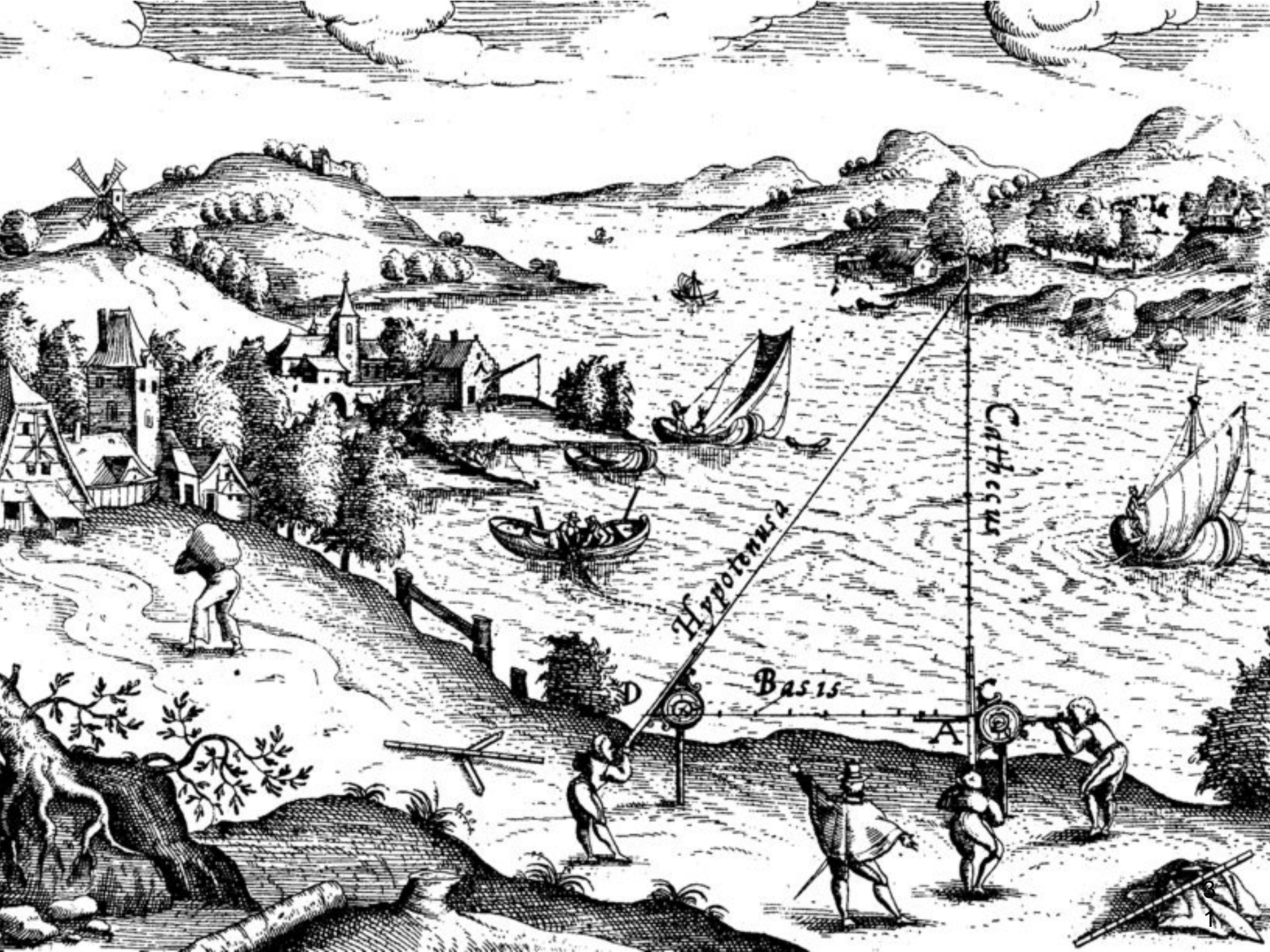
Be paranoid!

Triangulation

Confirm on multiple distinct data sets

Use multiple algorithms

Same algorithm in multiple programming languages


Cathecus
Hypotenusa
Basis
D
C
A
B

Coding standards

Unit tests & code reviews

Test asserted properties of data

Re-run test code every time the data is touched

Data standards

Self-describing data: log your parameters and inputs

After all joins or filters, count and log.

Human-readable data:

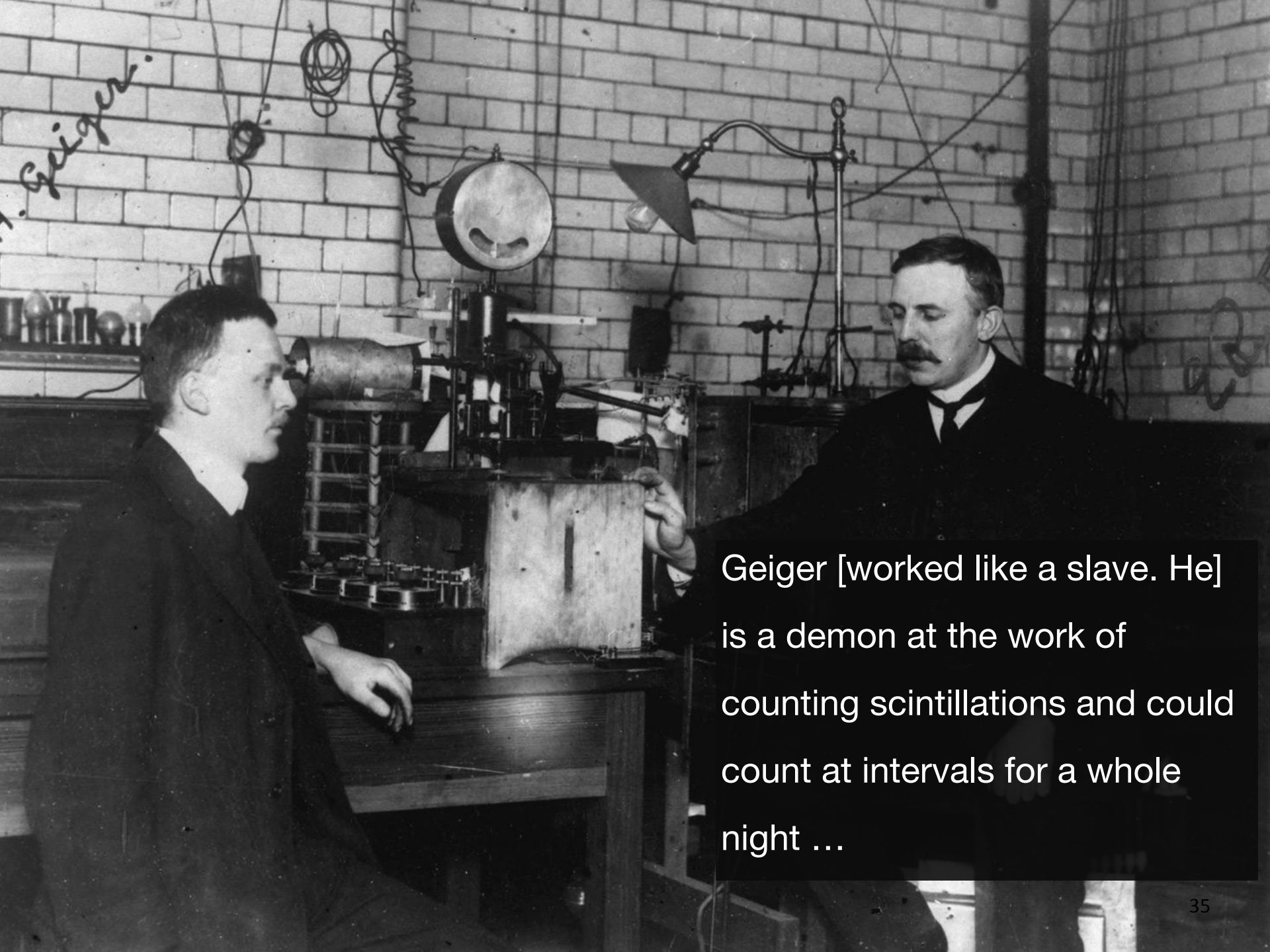
e.g. ‘sensitive’ or ‘resistant’ instead of ‘1’ and ‘0’

Cultural standards

Formal lab meetings - include domain experts

Hand over data sets in a formal handover session

Encourage a culture without blame



Geiger [worked like a slave. He] is a demon at the work of counting scintillations and could count at intervals for a whole night …

That's what **real** scientific work looks like.

Not like….

References

Deriving chemosensitivity from cell lines. Keith A. Baggerly, Kevin R. Coombes (2010)

An array of errors, The Economist (2011)

en.wikipedia.org/wiki/Geoffrey_Chang

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

@RSS_DSS, #DSSLaunch

The Industrialisation and Professionalisation of Data Science

RSS Data Science Section

Launch Event (Recorded)

19th June 2017

Closing remarks and next steps

Thank you for coming to our launch event and to all of those that made it happen

Our aim was to get feedback from the DS community

Our aim is involve the DS community

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

How you can get involved

In the first instance, please comment on our documents over the next 2 weeks (i.e. 3rd July):

- github.com/rssdatascience

We have been taking notes for this evening too and the recording will go on YouTube (now available here)

Based on the feedback, we will select the priority topics

Determine the best means to engage (work groups etc.)

@RSS_DSS

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Networking reception

Please carry on the discussion this evening too

Thank you to our sponsor of the evening event

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS