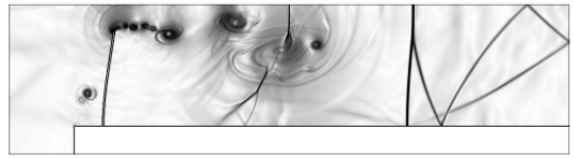
Introductory Lightning Talks

- Two slides
  1. Who are you? What do you do? Your perspectives
  2. An idea at the intersection of RSE and DS

- 3 minutes each
- Timed!
- Rob has the clicker
- Alphabetical order (by last name)

- Questions at the end – if there is time

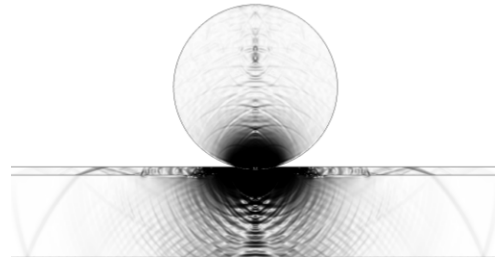
Matthew Archer



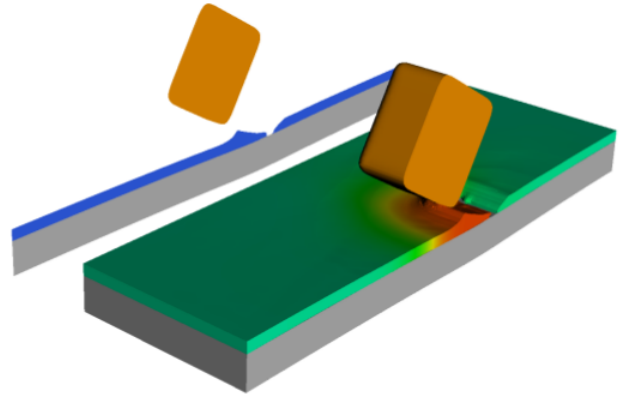
Background

- MEng Chemical Engineering.
- PhD Computational Physics: Solid Mechanics.

Current (Dec. 2017)



- Research Software Engineer specialising
  - Big-Data and Machine Learning
  - Continuum mechanics.
- Help with user support on **CSD3**.



Optimising performance of deep learning frameworks.

Make it easier for users to make the most out of hardware.

- Which **technology** for the job?
- How is **data** stored?
- Does **hardware** work as vendors say?

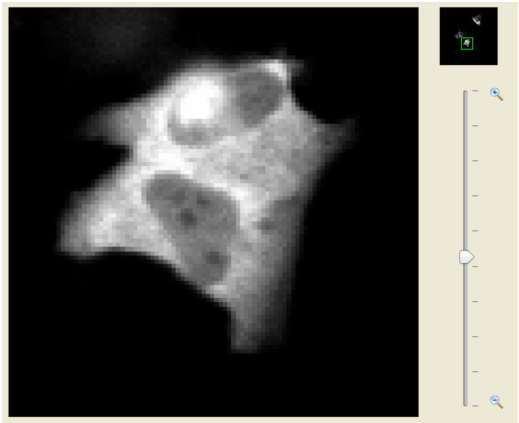
sanaz jabbari bayandor

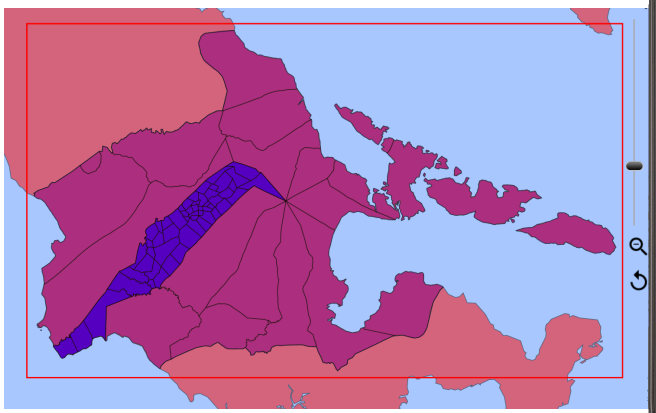
- University yrs: Phd in Computer Science, University Of Sheffield, NLP Group. Worked as an RA.
- Start-up yrs: Head of R&D team at a tech start-up, Fizzback Ltd, Text classification and Sentiment Analysis across industries
- Corporate yrs: Applied computer scientist at Microsoft Ltd, Worked for Bing/Outlook/Query Formulation Team in London
- Back to Uni: Data Science Lead at UCL's Research IT Center
  - Add Data Science consultancy & services to existing software engineering services for UCL researchers and its partners

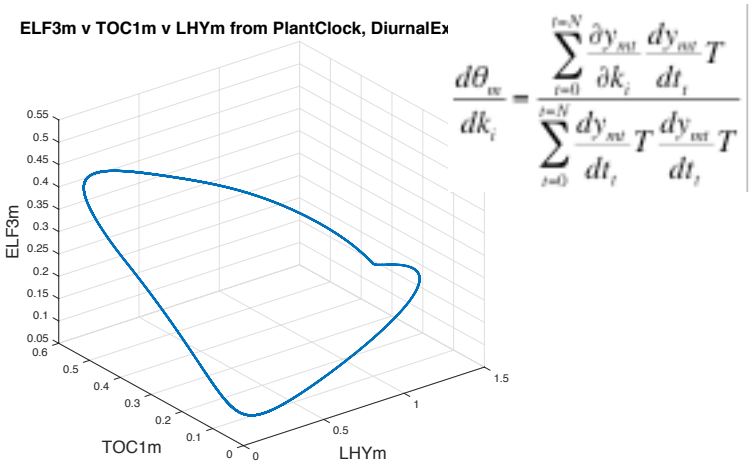
- Portrait of a data scientist as a Research Software Developer
    - Which parts of Research Software Development are uniquely a DS role?
        - Coding for Data Analysis, Computational statistics, ML
    - What are the best practices for Data scientists in a Research software team?
    - Is domain expertise necessary for Data scientists for carrying out data analysis and modeling?
        - How to best collaborate with researchers across different fields

Paul Brown, University of Warwick

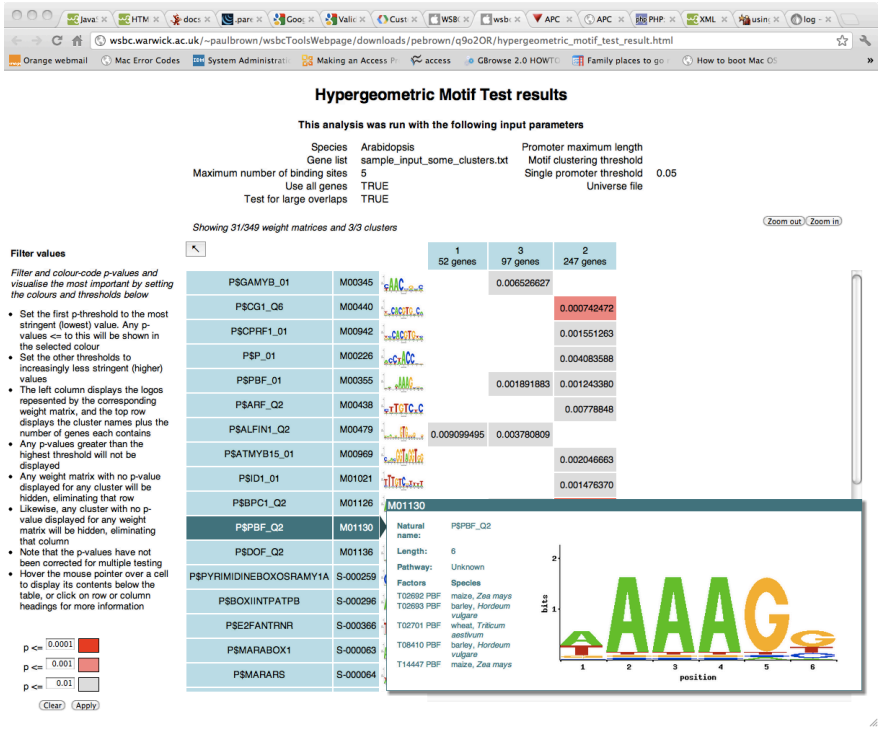
- Background in experimental biology
- ODE models of gene networks
- Image analysis
- Bioinformatics tools
- Infectious disease modeling







$$\frac{d\theta_{wt}}{dk_i} = \frac{\sum_{t=1}^{t=N} \frac{\partial y_{wt}}{\partial k_i} \frac{dy_{wt}}{dt_t} T}{\sum_{t=1}^{t=N} \frac{dy_{wt}}{dt_t} T \frac{dy_{wt}}{dt_t} T}$$



- Multi dimensional models
- Multiple levels of uncertainty
- Database of outputs
- New set of inputs -> interpolate outputs
- Model too complex to be done analytically
- Can we train a neural network to determine relationship between inputs and outputs?

Stephen Dowsland

Newcastle University &
The National Innovation Centre for Data

You can find me at @ste1
or email stephen.dowsland

Newcastle University

Speaking the right language(s) for Data Science

What languages are currently used?

What languages are up and coming?

Who is driving this change?

How should we assess/appraise language choice?

- Conversion of existing code vs. training in new language
- Example Matlab to Python - Licensing problem vs. new code generation

How can this form part of best practice?

Rosa Filgueira

Background:

- **PhD Computer Science** - University Carlos III – **HPC Research**
  - Dynamic optimization techniques for MPI-based applications

- 5 years as a **Postdoc** – University of Edinburgh – **Data Intensive Research**
  - Research activities in Scientific Workflows & Gateways, and Data-Intensive methods. Work in several UK & EU project funded
    - e.g. VERCE and REAR (Amy!)

- 2 years as a **Senior Data Scientist** – BGS – **Geoscience Domains**
  - Data gathering, cleaning, filtering, analysis
  - Parallelisation/optimization of applications
  - Promoting scientific workflows, data-frameworks, containers and reproducibility tools, etc.

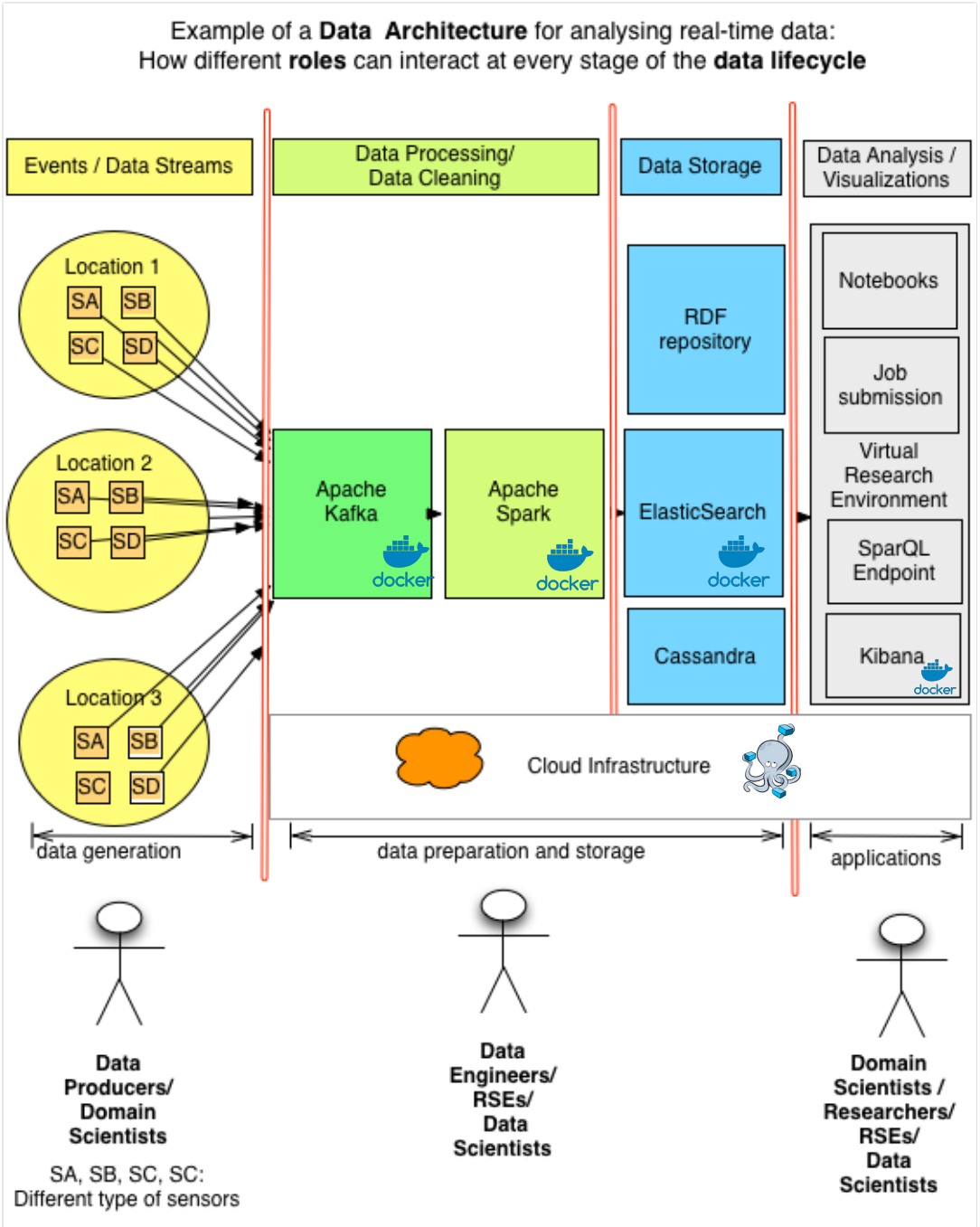
Currently (started two weeks ago!):

- **Data Architect** – EPCC, University of Edinburgh – **Multiple Domains**
  - Focused in data-intensive architectural challenges

epcc

THE UNIVERSITY OF EDINBURGH

https://github.com/rosafilgueira/EMB_datastreaming



Environmental monitoring example
- Datasets:
  - **Water quality**
  - **Seismicity**
  - **Atmospheric composition**

-Multiple locations & data types in real-time

- Roles with **different** set of **skills/expertise**

- Roles **interact** with datasets in different ways**: different needs / interests / frameworks/ computing languages & resources**

- Communications/Collaborations between those roles **are essential**

epcc

THE UNIVERSITY OF EDINBURGH

Stuart Geiger: ethnographer/postdoc
UC-Berkeley Institute for Data Science

Background: History & Philosophy of Science, Organizational Sociology, Information Studies

RQ: How is the way we produce knowledge changing?

Themes: Openness, sustainability, institutions, infrastructure, invisible work, socio-technical systems

WIKIPEDIA
The Free Encyclopedia

Data Science → The Data Sciences

Multiple portraits of what RSEs & data scientists do, versus a one-size-fits-all job description

Possible dimensions:

- Work on: analytics ↔ infrastructure
- Work in: a lab ↔ a cross-campus group
- Work is: project-focused ↔ general purpose
- Work with: large teams ↔ independent
- Output is: publications ↔ software
- Teaching: informal ↔ formal

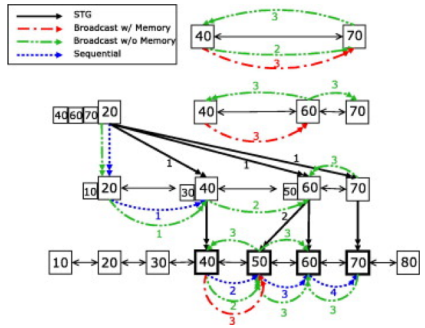
Alejandra Gonzalez-Beltran – Research Lecturer (Research Software Engineer & Data Scientist)

Strong Background in Maths & Computer Science

E-Government apps
Financial web apps
Code Slicing Techniques

Efficient Access to Distributed Information Using Structured Peer-to-Peer Systems

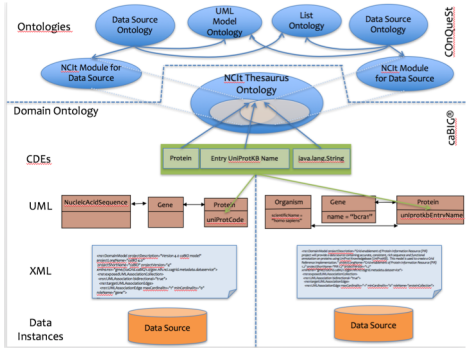
Skip Tree Graph
Probabilistic Data Structure



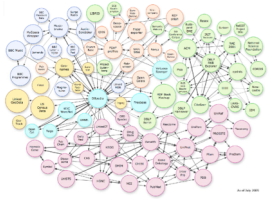
Software Engineer

MSc & PhD Computer Science

conquest
(cancer **on**tology **que**rying **s**ystem)



Data integration &
linked data
for cancer databases



isatools
http://isa-tools.org

FAIRsharing.org
standards, databases, policies

isacreator
OntoMaton
linkedisa
Risa
isaAPI

stato
statistics ontology

DATS
DatA Tag Suite

dataMED
BETA version

Research Software Development
Data Standards; Data Description; Data Provenance
Semantic Web; Linked Data
Ontology Development
Digital Research Objects Dissemination
Reproducibility
Software Engineering and Data Science Teaching
Diversity

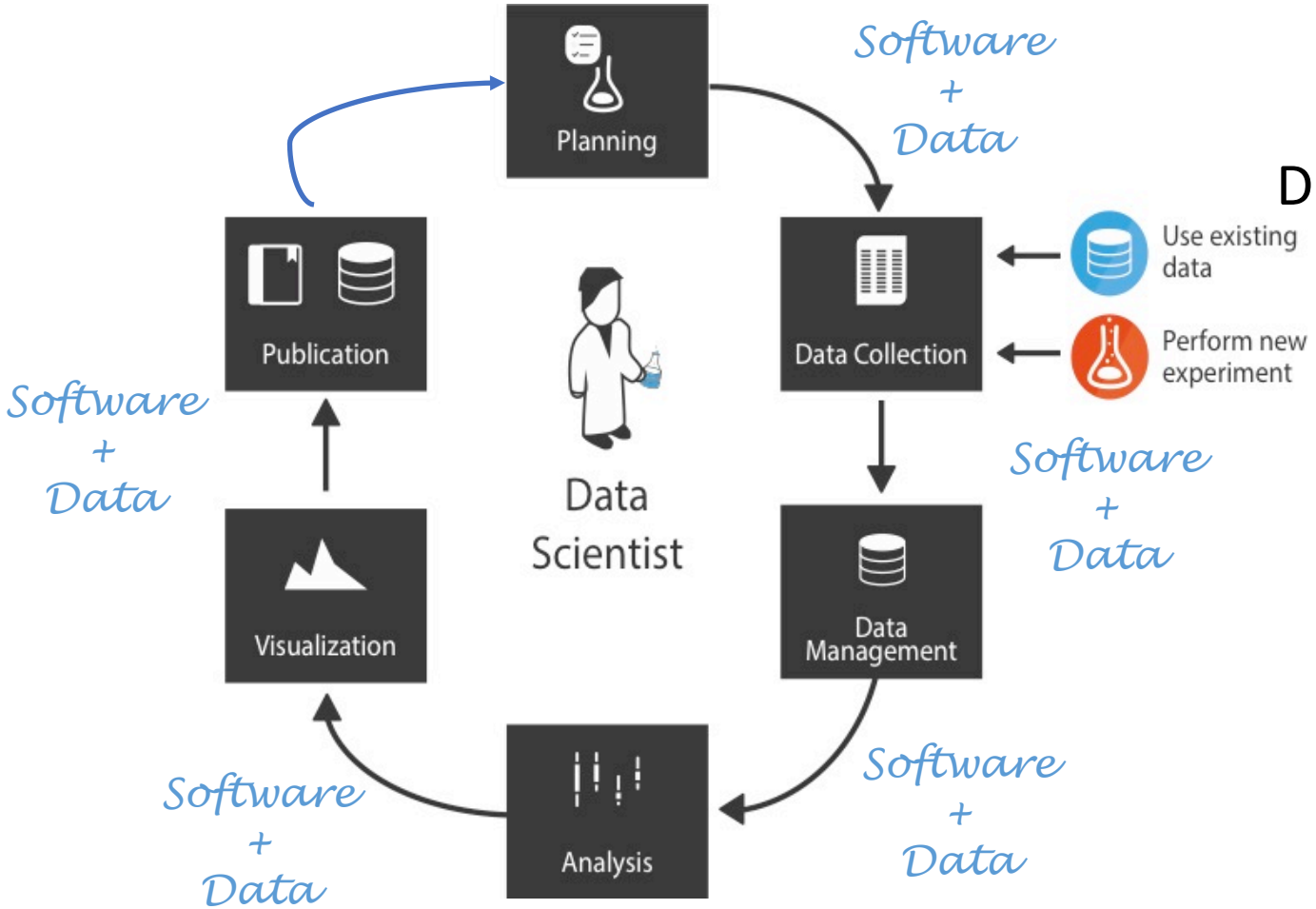
impa
MR
Lifia
ATX
Universidad Nacional de Rosario
Queen's University Belfast
UCL
University of Oxford
Data Carpentry
software carpentry
Software Sustainability Institute
W3C

Pre-Registration

Metadata / Publication / Discovery / Citation

Scholarly Article
Viz

Metadata
Publication
Discovery
Citation



Data Management Plans
Data Collection
Data Curation

Metadata
Publication
Discovery
Citation

Data + Software as part of Reproducible Analysis

Metadata / Publication / Discovery / Citation

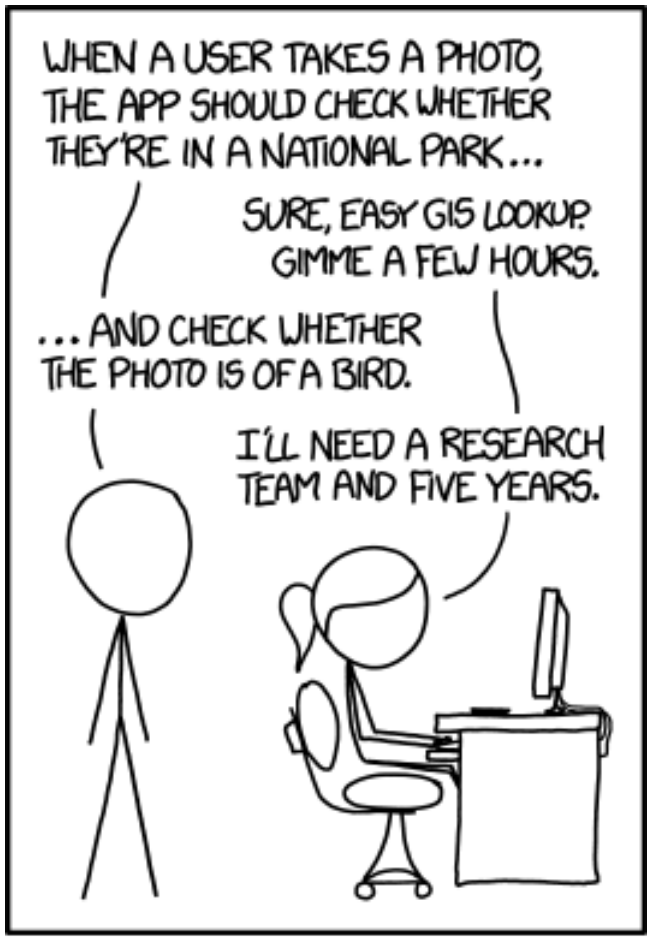
Infrastructure, Decentralization, Education/Skills, Credit, Sustainability

Robert Haines

Head of Research Software Engineering
University of Manchester, UKRSE, SSI
robert.haines@manchester.ac.uk

MANCHESTER 1824
The University of Manchester

- From Research Associate
  - Writing code
  - Writing (some) papers
- To Software Engineer
  - Designing and building systems
  - Writing (fewer) papers
- To Research Software Engineer
  - Collaborating with researchers on projects
  - Writing papers
- To Head of Research Software Engineering
  - Managing expectations



https://xkcd.com/1425/ (CC-BY-NC 2.5)

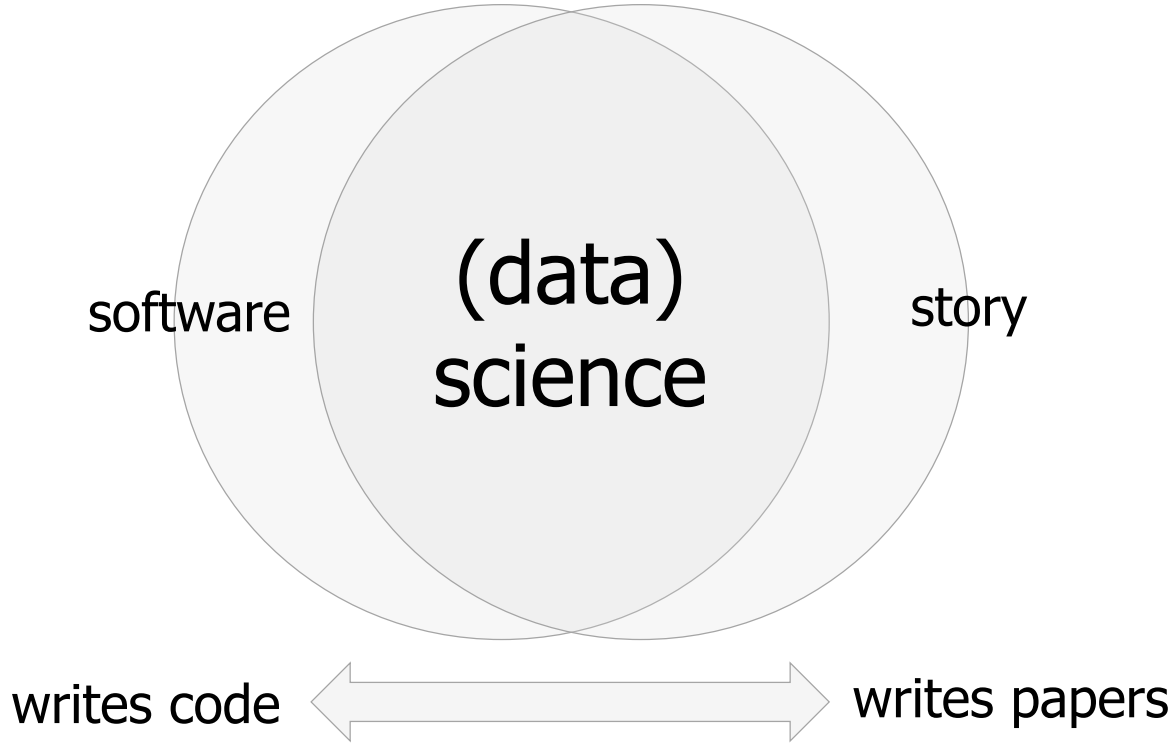
Robert Haines

Head of Research Software Engineering
University of Manchester, UKRSE, SSI
robert.haines@manchester.ac.uk

MANCHESTER 1824
The University of Manchester

*"While some of us work on the software necessary to produce results, and others work on the surrounding story necessary to explain them, in team-based research we are all **scientists**"*

Jay, Haines, Vigo, Matentzoglu, Stevens
10.3897/rio.3.e13236



Can we answer these questions with a combination of RSE and DS:

- How do we translate between scientific theory and code?
- How do we properly test code derived from theory?

CHRIS HOLDGRAF

@choldgraf | choldgraf@berkeley.edu
Berkeley Institute for Data Science
Project Jupyter

"open source neuroscience electrophysiology analysis"

Visualization
Feature extraction
Statistical analysis
Machine learning

"easily connect users to an environment in the cloud"

Online collaboration
Teaching and education
Accessibility to data analytics

"create interactive, reproducible, sharable computational environments"

One-click interactivity
Reproducible publications
Communication and sharing

MY INTERESTS

project-specific

using jupyterhub/binder to make teaching and education more effective

using jupyterhub/binder to facilitate the use of complex hardware/data

connecting with other web/cloud-savvy or interactive open source projects

anything with brains

general and meta

valuing social solutions as much as we value technical solutions

making open, reproducible analysis more accessible and easy

the difference between “technically” open and “practically” open

sustaining open-source culture without sacrificing its core values

Amy Krause

- Research data engineer at EPCC, University of Edinburgh
- Working with industry and research in the area of distributed data integration and analysis
- Over 15 years of experience in software engineering, integration of distributed data, databases, real-time data analysis, large-scale data analytics frameworks, cloud (and what used to be called Grid services!)
- Recent projects:
  - DARE, VERCE, ADMIRE, OGSA-DAI: European Projects; support scientists with distributed data infrastructure and tools
  - REAR: Emergency aftershock response using mobile phone data
  - ThinkAnalytics: Query infrastructure to analyse statistics produced by a TV recommendation engine

epcc | THE UNIVERSITY OF EDINBURGH

Ideas

- I want to support data scientists and help them re-implement their research questions at scale (”big data”)
- Be able to ask bigger questions and study larger datasets
- Pet projects:
  - I’m interested in using the data produced by mobile phones to improve road travel for cyclists and pedestrians
  - An open source framework to collect and analyse step data/movements from horses, to create tools for training, monitoring health (lameness)

David Mawdsley

RSE at **University of Manchester** since 2016.

The route to the north…

- Physics with computational physics degree (C, Fortran)
- PhD in a physics department (C, Perl)
- Data analyst at HEFCE (SAS, R)
- Back to academia via an MSc in medical statistics (Stata, R)
- Postdoc at Bristol (R, WinBUGS, Stan)

- Manchester

Spreading good practice

Data science projects span multiple skills and specialities (no unicorns):

Software Engineering <————————> *Statistics*

What

Sharing good software development practice with statisticians.

Why

Allows easy collaboration and working across the data science spectrum.

How

Carpentry style courses…Suggestions?

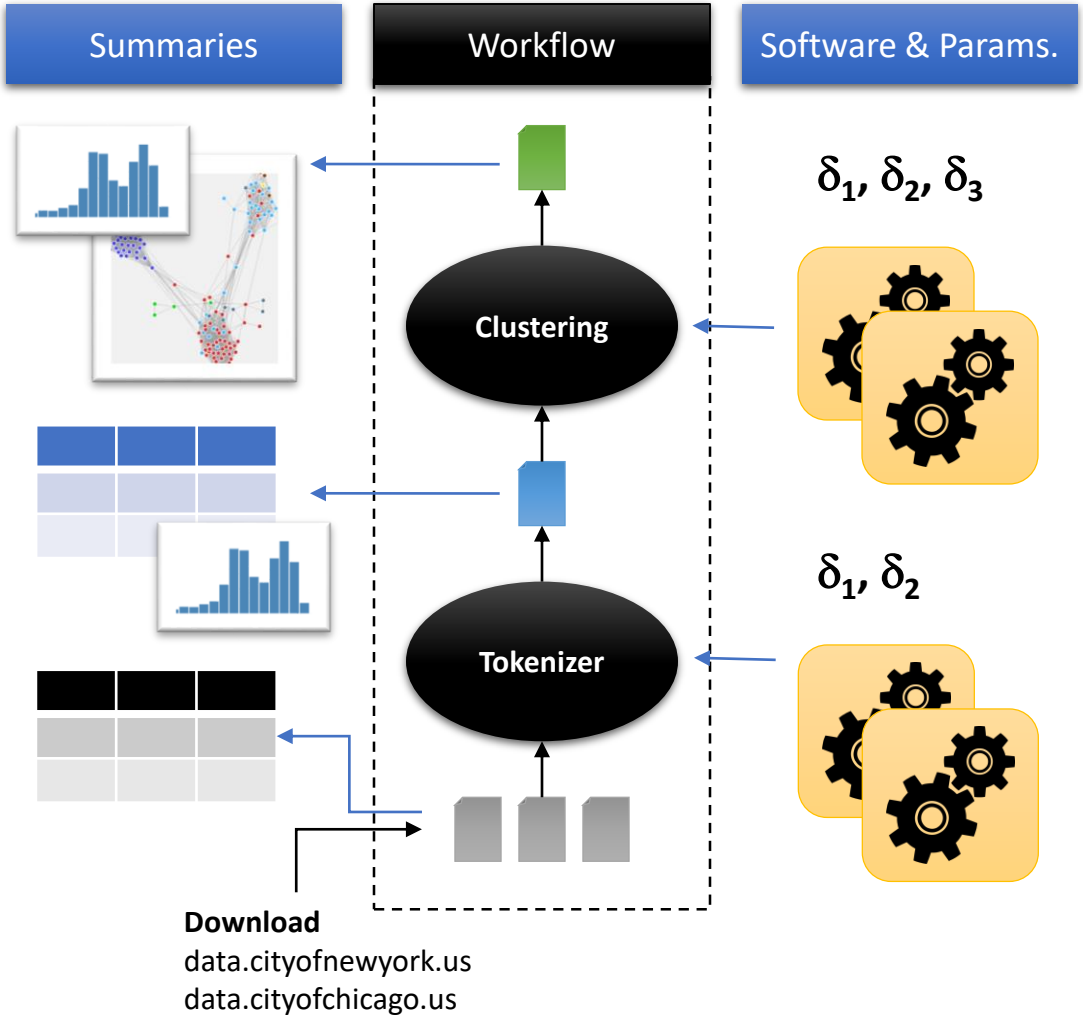
Heiko Mueller

Research Engineer

New York University

Vision for Research Project Management Tool

How to avoid the *WCPM.K3.SIM0.8.APRO.COS.CSW.SIM0.5.ALL.txt* problem



Experiments keep track of their environment (variables)

Data
Workflow & Software
Parameters

'Clone environment' to (re)-run experiment with different values

Keep track of updated files automatically

Allow user to interact with data via Web Interface

The Alan Turing Institute

Martin O'Reilly

Principal Research Software Engineer

Industry: Running a "shadow IT" team of analysts and developers

MSc: Playing with robots, machine learning and understanding the brain by faking it

PhD: Helping computers find neurons in high-resolution microscope images

Industry: Consultant developer and technical project manager

Turing: Lead the RSE team within the Research Engineering Group

Interests: Good working practices for research software development.

Reproducible research. Open science.

Understanding the brain.

UNIVERSITY OF CAMBRIDGE · THE UNIVERSITY OF EDINBURGH · UCL · WARWICK THE UNIVERSITY OF WARWICK · UNIVERSITY OF OXFORD · EPSRC Pioneering research and skills

Ideas

What do good working practices for data science look like?

- What can we use from software engineering practices?
- Are these just good reproducible research practices?

How to make data sources better?

- Data navigable from within my code editor
- Can easily slice the data how I want
- Can easily combine data from different sources
- Data is versioned and supports differential download
- Smooth access to data requiring an access agreement / authentication
- Documentation accessible in context of data presentation

TOMAS PETRICEK

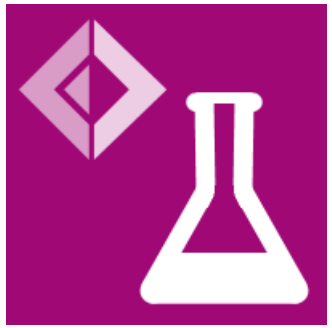
The Alan Turing Institute & University of Kent
@tomaspetricek | tomas@tomasp.net

Programming languages

Theory of context-aware programming languages

Extensions for concurrent & reactive programming

F# and open source



F# tooling including editors & documentation tools

F# libraries for data access and data visualization

Data science & journalism

Tools for transparent data-driven storytelling

Making notebooks smart, reproducible & polyglot

PROGRAMMING RESEARCH meets DATA SCIENCE

Can research engineering become a regular part of data science research?

Interactive programming for data science

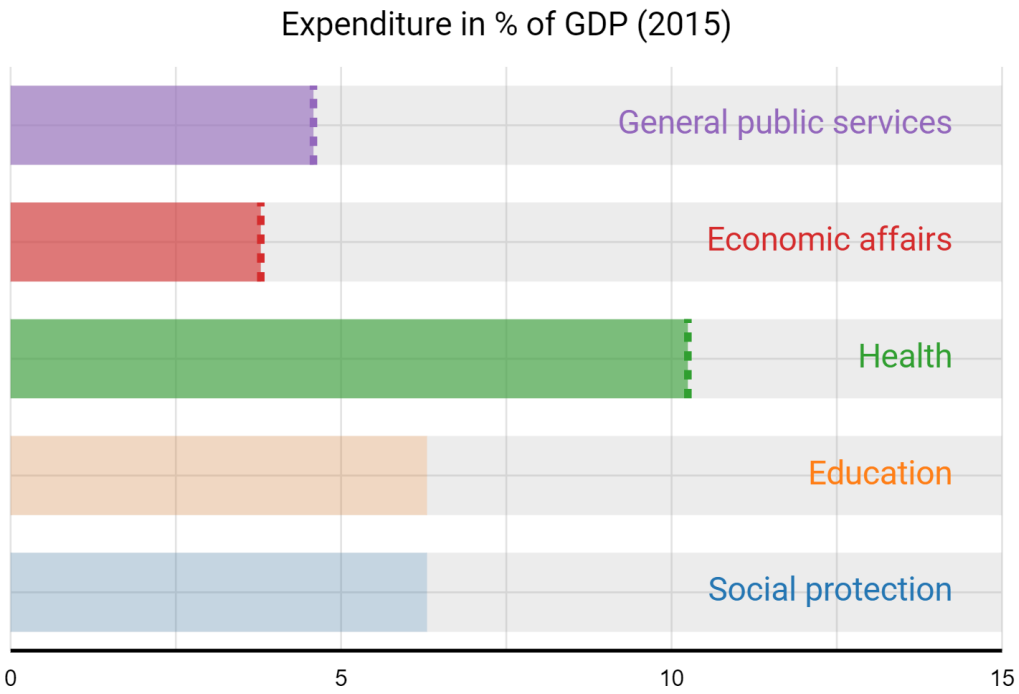
```
let data =
  olympics
    .'filter data'.'Games is'.'Rio (2016)'.then
    .'group data'.'by Athlete'.'sum Gold'.then
    .'sort data'.'by Gold descending'.then
```

olympics | filter by ✕ | group by ✕ | sort by ✕ +

by Gold descending ✕ +

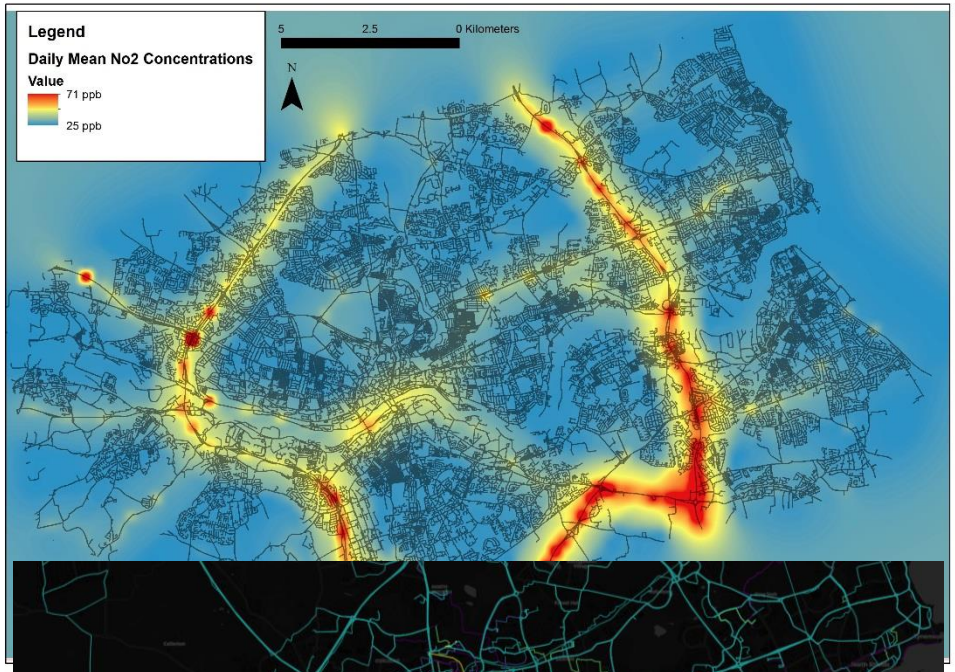
Athlete	Gold
Michael Phelps	5
Katie Ledecky	4
Simone Biles	4
Katinka Hosszu	3
Usain Bolt	3
Jason Kenny	3

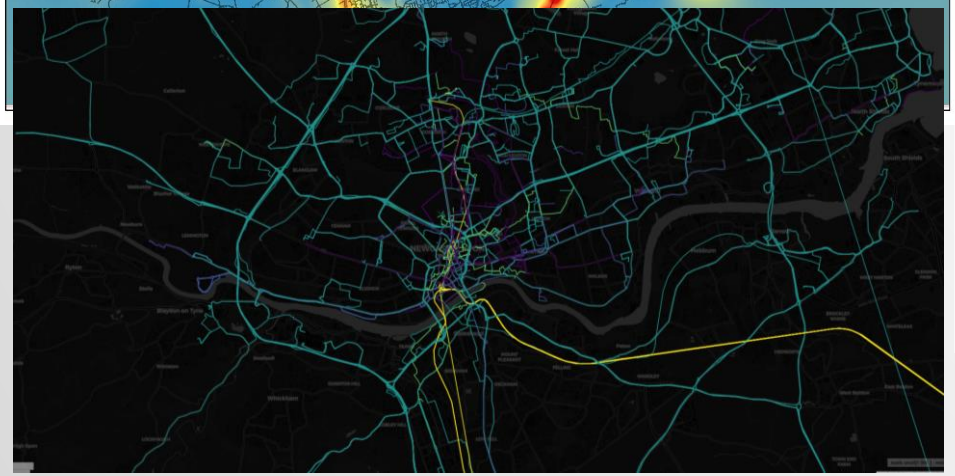
Data visualizations for critical thinking

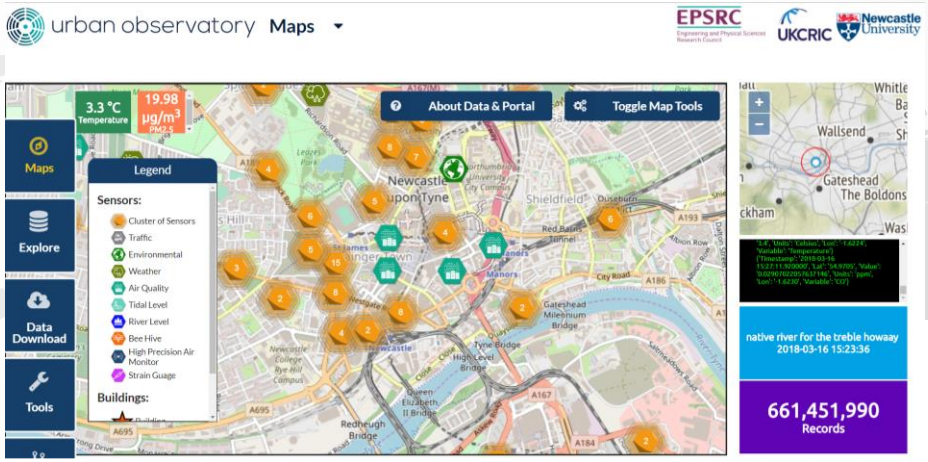


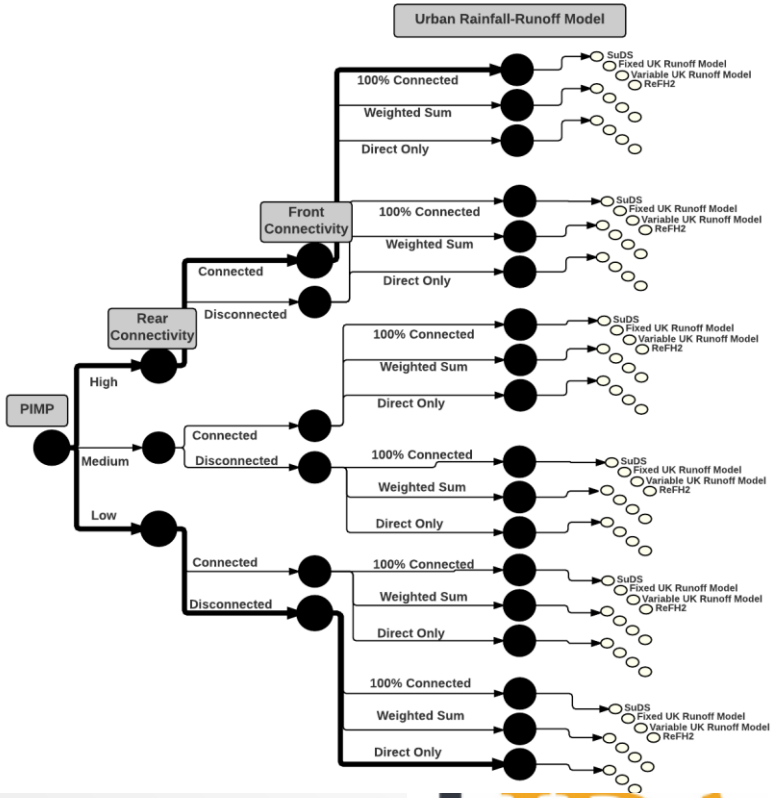
TOM REDFERN

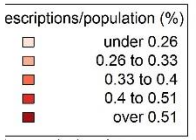
DATA SCIENTIST
LEEDS INSTITUTE FOR DATA ANALYTICS
UNIVERSITY OF LEEDS











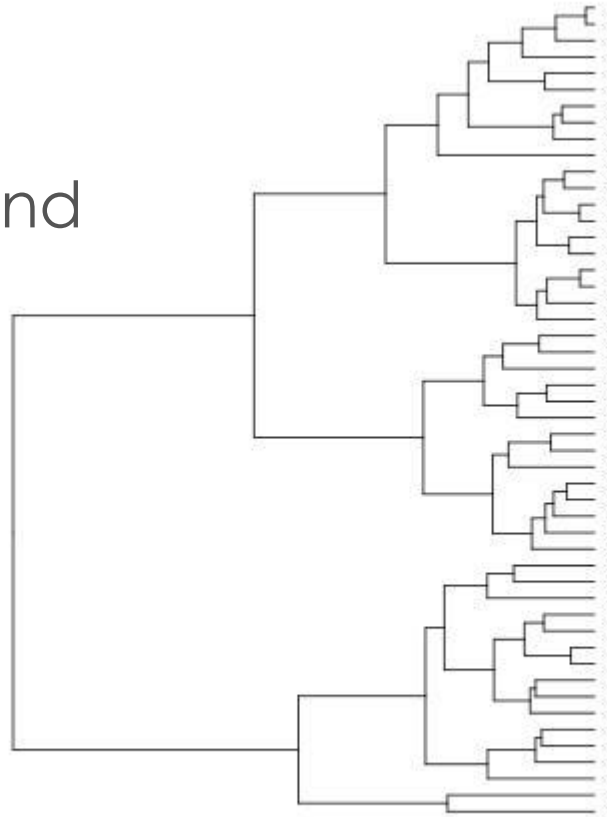
LIDA

UNIVERSITY OF LEEDS

TOM REDFERN

DATA SCIENTIST
LEEDS INSTITUTE FOR DATA ANALYTICS
UNIVERSITY OF LEEDS

"How to identify and support RSEs?



jobs.ac.uk

LIDA

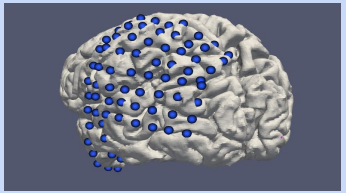
UNIVERSITY OF LEEDS

Valentina Staneva - Data Scientist, eScience Institute, University of Washington

PhD - Object Segmentation and Tracking

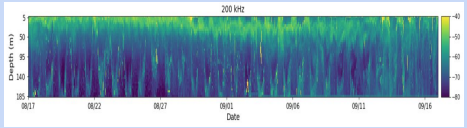
Pipelines for Automatic Cornea Image Segmentation

Spatiotemporal Data
Optimization
Dimensionality
Reduction

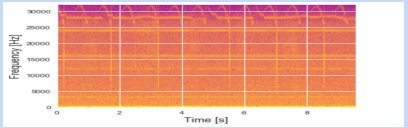


Data Science for Social Good: transportation projects

Multimodal Neural Encoding: ecog + video + audio



Ocean Acoustic Data Mining: echosounder + hydrophone



Software Simplicity vs Scalability

Need to handle large datasets: images, streaming time series, videos, networks, ...

Need to have libraries:

- Easy to use
- Easy to develop for
- Work the same on small and large datasets
- Work the same on laptop, university cluster, cloud, gpu

Some examples: dask, keras, jupyterhub, kubernetes, ...

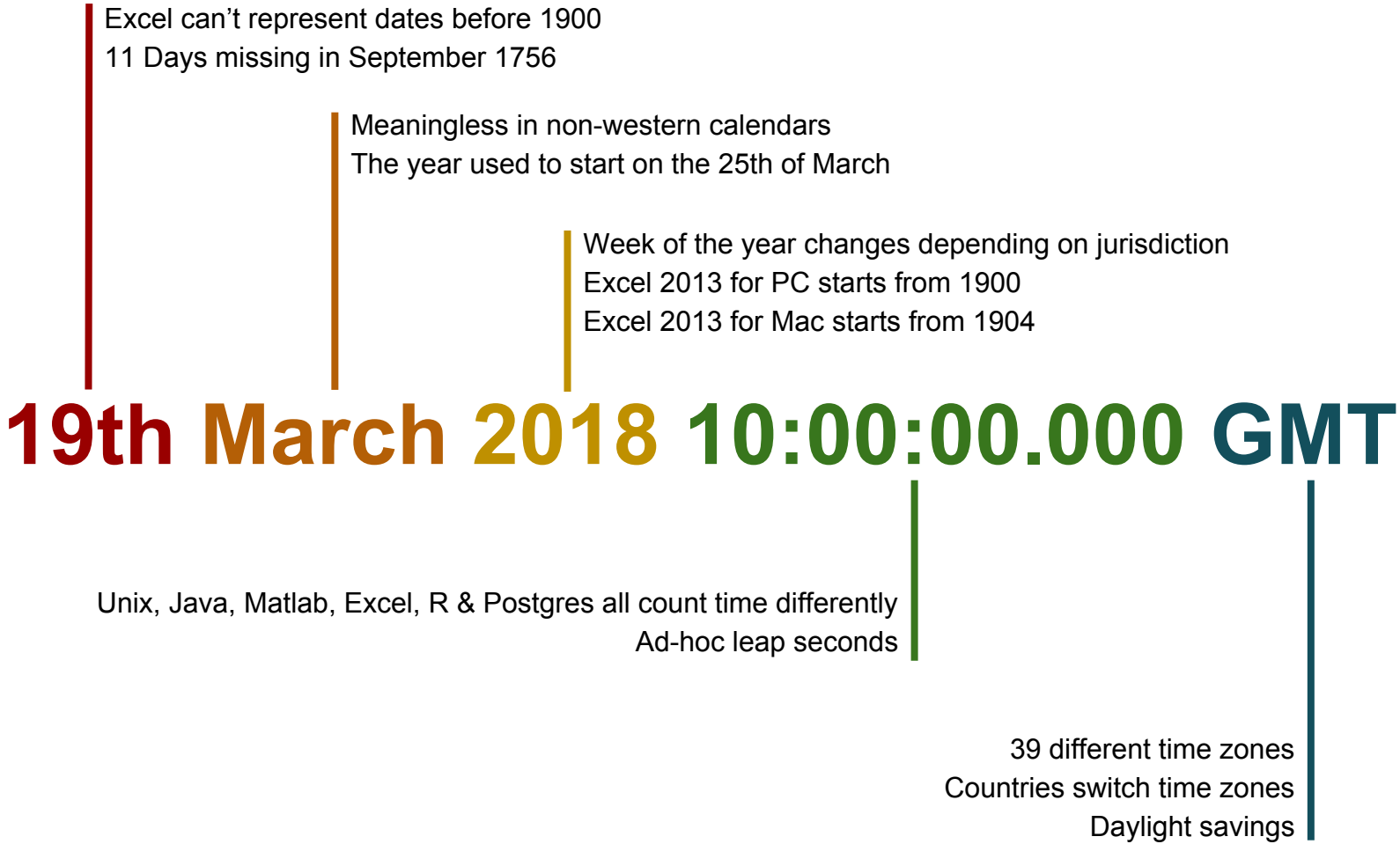
Should we rewrite the libraries, should we change the deployment systems?

Mark Turner

Newcastle University &
The National Innovation Centre for Data

You can find me at @markdturner
or email mark.turner@ncl.ac.uk

Newcastle University


Excel can’t represent dates before 1900
11 Days missing in September 1756
Meaningless in non-western calendars
The year used to start on the 25th of March
Week of the year changes depending on jurisdiction
Excel 2013 for PC starts from 1900
Excel 2013 for Mac starts from 1904
19th March 2018 10:00:00.000 GMT
Unix, Java, Matlab, Excel, R & Postgres all count time differently
Ad-hoc leap seconds
39 different time zones
Countries switch time zones
Daylight savings

```
$ whoami
jakevdp
```

Work...

UNIVERSITY of WASHINGTON eScience Institute

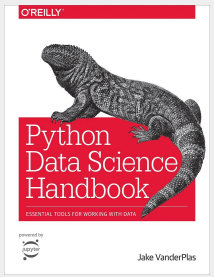
Google

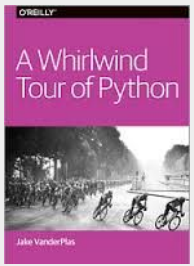
Code...

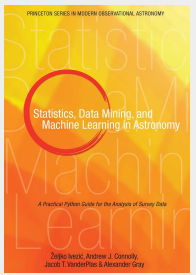
scikit learn

astropy

Writing...

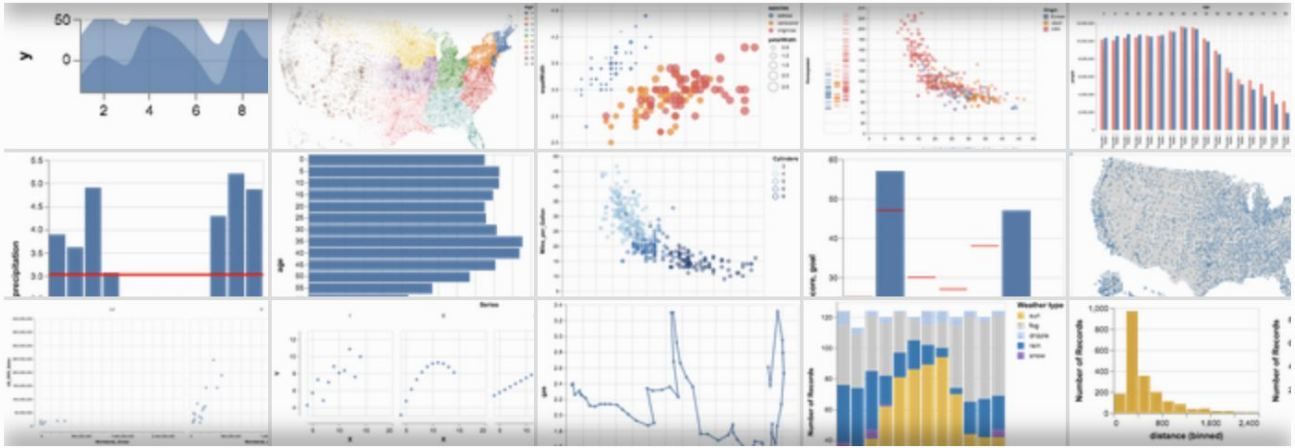






My focus Lately...

Altair: Declarative Visualization in Python



```
alt.Chart(cars).mark_point().encode(
    x='Horsepower',
    y='Miles_per_Gallon',
    color='Origin',
).interactive()
```