INM430
Principles of Data Science

Week 10
Principles, Issues & Challenges

Aidan Slingsby, giCentre

CITY
UNIVERSITY OF LONDON
EST 1894

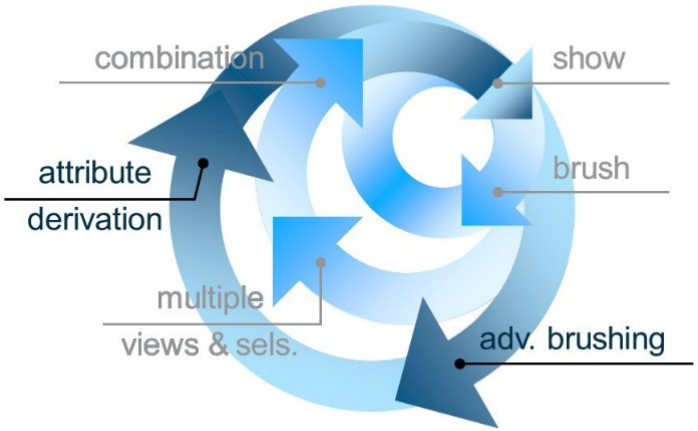
Today

- A gallop through the Data Science process
- Along the way…
  - Tips!
  - Discussion!
  - Examples!
- In the lab, consolidate. Either:
  - Revisit previous labs and lab feedbacks
  - Work on your courses
  - Tackle the option exercise

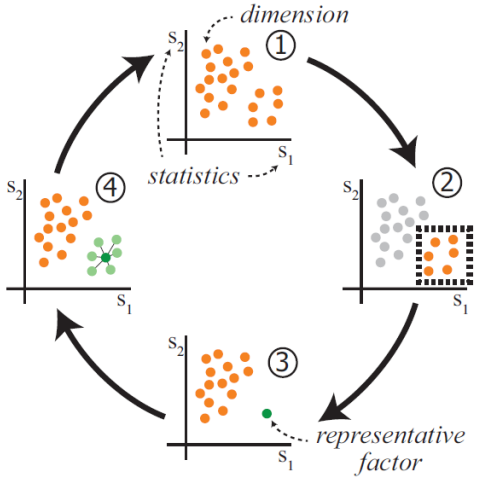
Data Science process (as presented earlier)

- Understand domain needs
- Collect & make data available
- Get the data ready for analysis
- Exploratively (and visually) analyse the data
- Model the phenomena (if needed)
- Evaluate findings
- Communicate findings
- ITERATE (from any stage to any other stage)!
- Useful writeup:
  - https://www.kdnuggets.com/2016/03/data-science-process.html

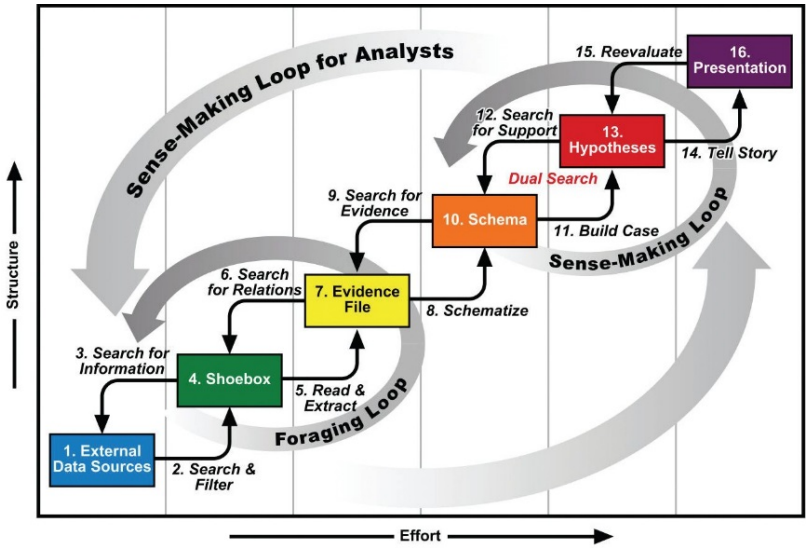
Iterate!



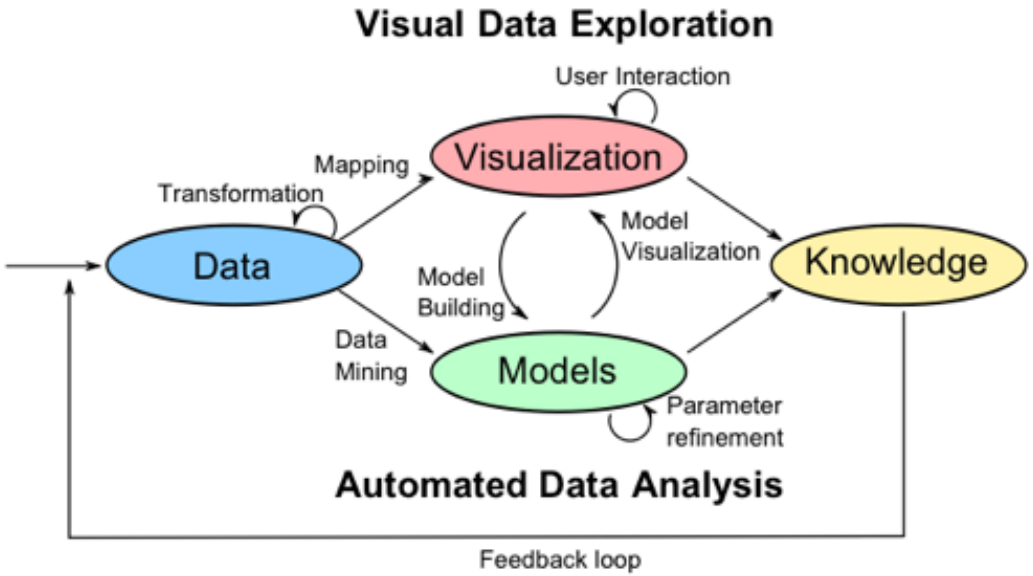
Hauser, 2013



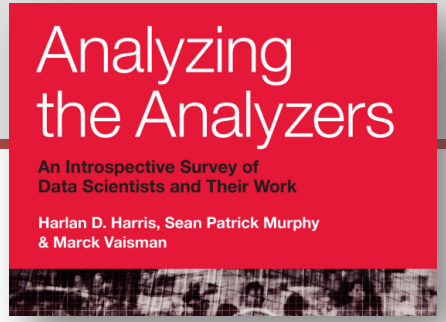
Turkay, 2012



Sense making loop by Pirolli and Card, 2005



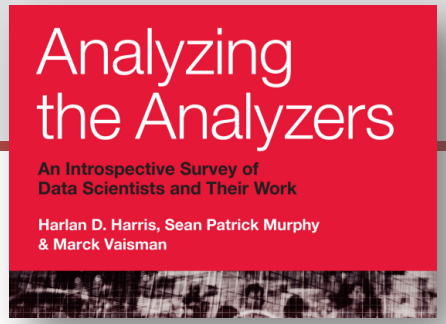
On data analysts – analyst types



Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

https://www.oreilly.com/data/free/files/analyzing-the-analyzers.pdf

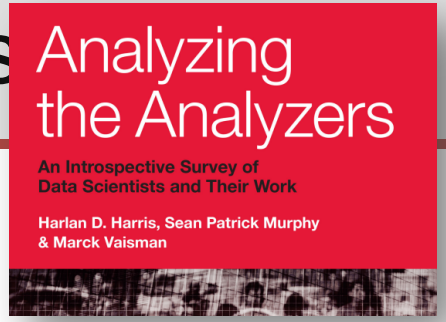
On data analysts – different skills

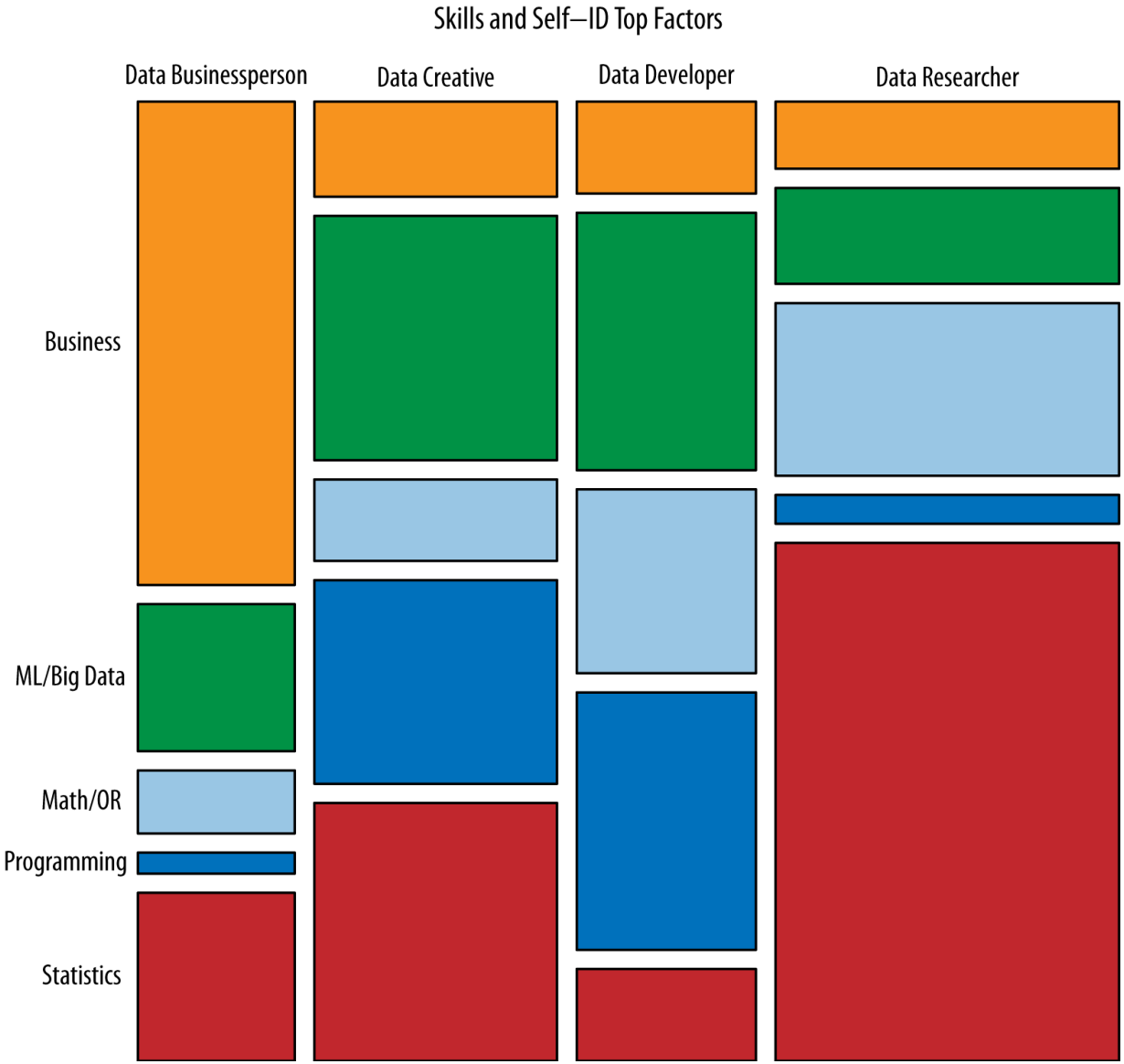


Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

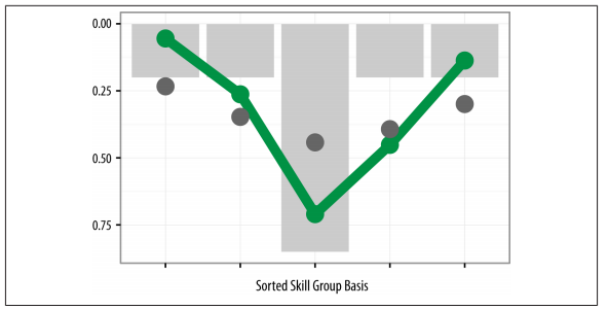
https://www.oreilly.com/data/free/files/analyzing-the-analyzers.pdf

On data analysts – skills vs. types





T-shaped data scientists



https://www.oreilly.com/data/free/files/analyzing-the-analyzers.pdf

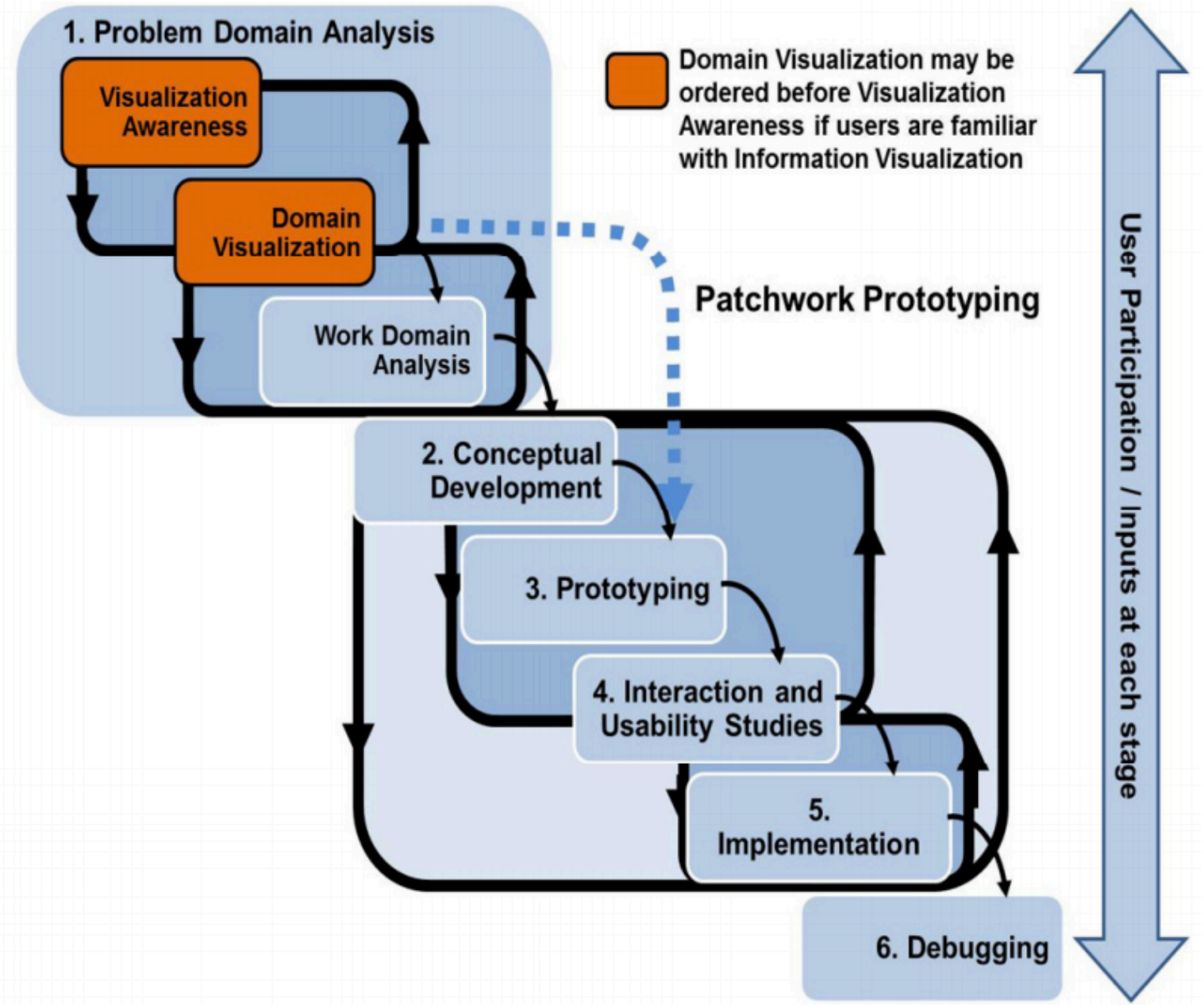
UNDERSTAND DOMAIN NEEDS

Often overlooked

- If we are helping someone else answer a question, need to make sure
  - We understand **what they want to know**
  - We understand **what the data mean**
  - We **verify patterns and anomalies found**, as they will know whether important and erroneous
- Academic working on analytical/visualisation techniques also guilty of this!
- Agree on (initial) **research questions** (or initial **tentative investigation**). Acknowledge that this may (and probably should) **evolve/change**.

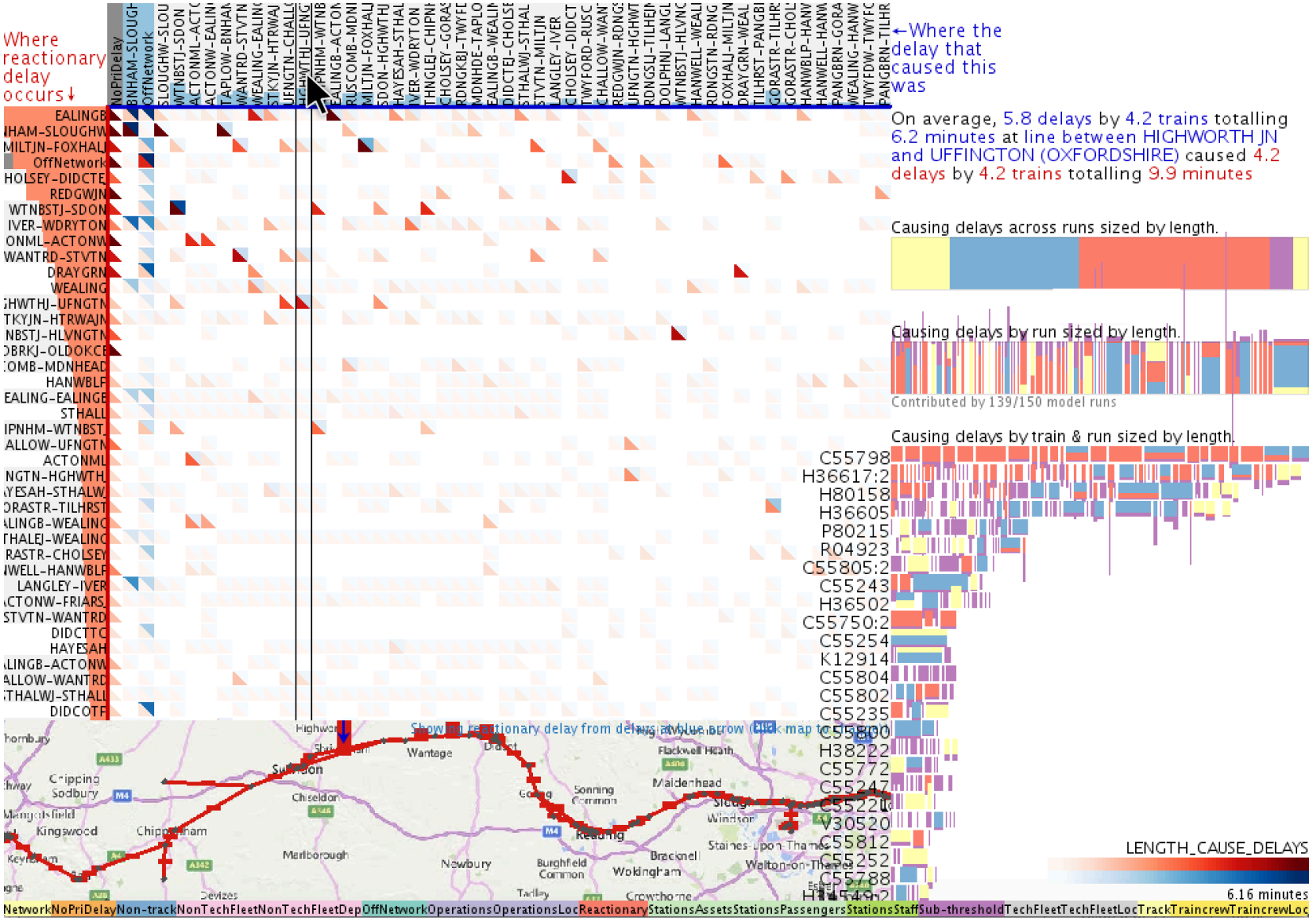
Beware separation between “user” & “analyst”

- Academics not immune to this!
- Impossible to separate:
  - “users” may not know what’s possible (beyond “magic”)
  - both sides need to engage in collaborate design and brainstorming
  - both sides need to monitor/discuss progress
  - (Some) data scientists ideally placed!
- An iterative Agile-like approach
- Some users such want “magic”, but we need to ensure that it’s not misguided magic…



http://openaccess.city.ac.uk/id/eprint/555/

Example: My train project



COLLECT DATA

A confession (often the other way round)

- We often start with the data.
- A new source of data becomes available:
  - “What can we do with it?”
- In this case, the previous step comes second
  - But we often will present the business case or report the other way round.

Data quality and sample size

- Often impossible to get perfect data. Does that mean we shouldn't use it?
- Not necessarily!
  - (Discussion on use of Twitter in VA module)

Understand the limitations of the data

- Know how the data was collected, therefore how representative it is
- If it's likely to be biased, simply ensure that this is taken into account
  - e.g. explicit caveats on interpretation at the end

Geographical data

- Covered in Visual Analytics
- Where records have geographical references
  - Named/coded (location) places or (area) places, coordinates (need to understand its precision/accuracy)
  - **Enables** geographical visualisation and analysis, but it's **not necessary** to do geographical visualisation/analysis
- To georeference named/coded places, need list of geometries and a lookup
  - Nice free source: https://www.geonames.org/
  - Nice source for UK: https://borders.ukdataservice.ac.uk/, including UK postal codes

https://twitter.com/jwoLondon/status/957653599194173440

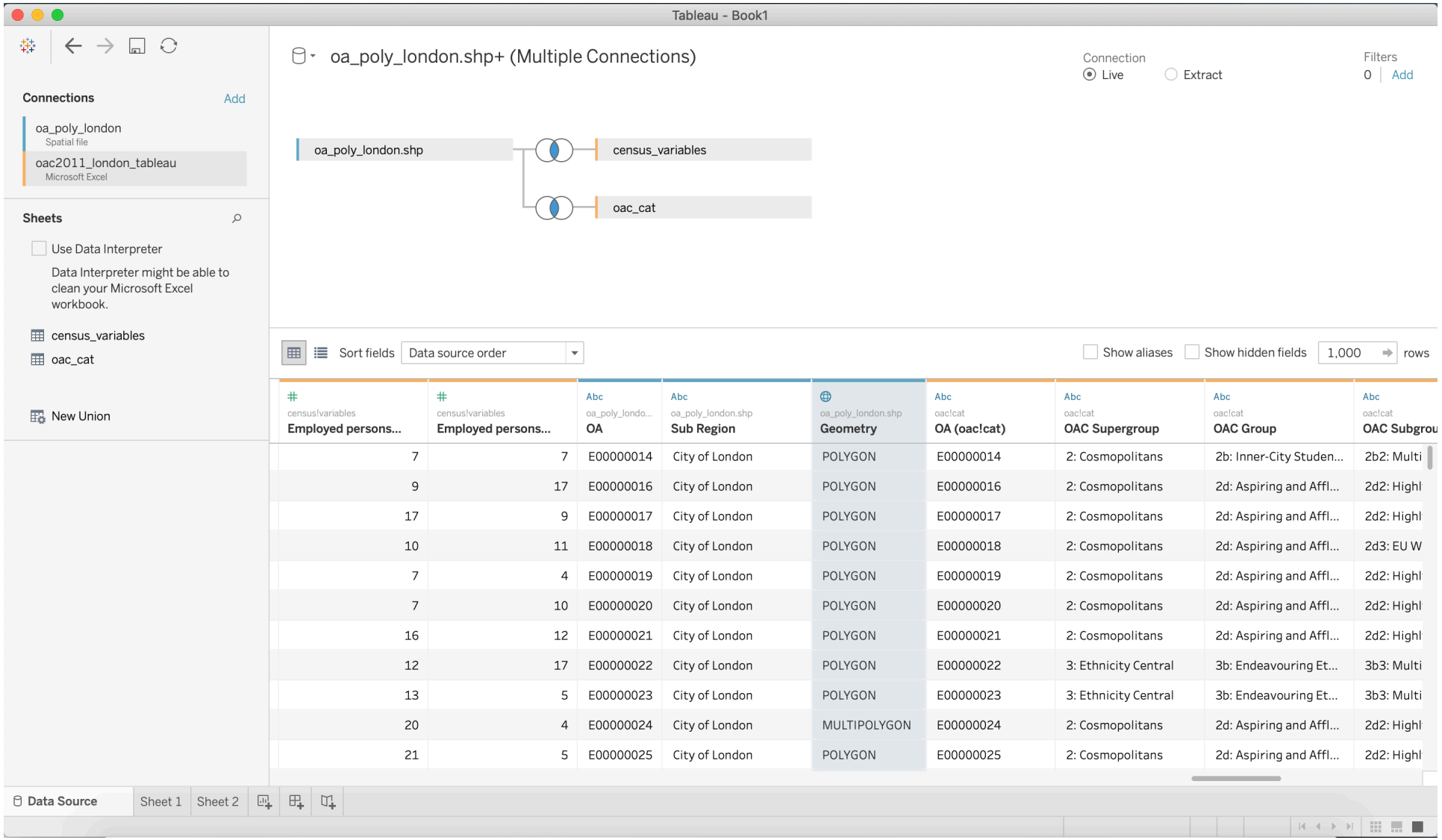


Tableau - Book1
oa_poly_london.shp+ (Multiple Connections)
Connection
Live
Extract
Filters
0
Add
Connections
Add
oa_poly_london
Spatial file
oac2011_london_tableau
Microsoft Excel
Sheets
Use Data Interpreter
Data Interpreter might be able to clean your Microsoft Excel workbook.
census_variables
oac_cat
New Union
oa_poly_london.shp
census_variables
oac_cat
Sort fields
Data source order
Show aliases
Show hidden fields
1,000
rows
census!variables
Employed persons...
census!variables
Employed persons...
oa_poly_londo...
OA
oa_poly_london.shp
Sub Region
oa_poly_london.shp
Geometry
oac!cat
OA (oac!cat)
oac!cat
OAC Supergroup
oac!cat
OAC Group
oac!cat
OAC Subgrou
7 7 E00000014 City of London POLYGON E00000014 2: Cosmopolitans 2b: Inner-City Studen... 2b2: Multi
9 17 E00000016 City of London POLYGON E00000016 2: Cosmopolitans 2d: Aspiring and Affl... 2d2: Highl
17 9 E00000017 City of London POLYGON E00000017 2: Cosmopolitans 2d: Aspiring and Affl... 2d2: Highl
10 11 E00000018 City of London POLYGON E00000018 2: Cosmopolitans 2d: Aspiring and Affl... 2d3: EU W
7 4 E00000019 City of London POLYGON E00000019 2: Cosmopolitans 2d: Aspiring and Affl... 2d2: Highl
7 10 E00000020 City of London POLYGON E00000020 2: Cosmopolitans 2d: Aspiring and Affl... 2d2: Highl
16 12 E00000021 City of London POLYGON E00000021 2: Cosmopolitans 2d: Aspiring and Affl... 2d2: Highl
12 17 E00000022 City of London POLYGON E00000022 3: Ethnicity Central 3b: Endeavouring Et... 3b3: Multi
13 5 E00000023 City of London POLYGON E00000023 3: Ethnicity Central 3b: Endeavouring Et... 3b3: Multi
20 4 E00000024 City of London MULTIPOLYGON E00000024 2: Cosmopolitans 2d: Aspiring and Affl... 2d2: Highl
21 5 E00000025 City of London POLYGON E00000025 2: Cosmopolitans 2d: Aspiring and Affl... 2d2: Highl
Data Source
Sheet 1
Sheet 2

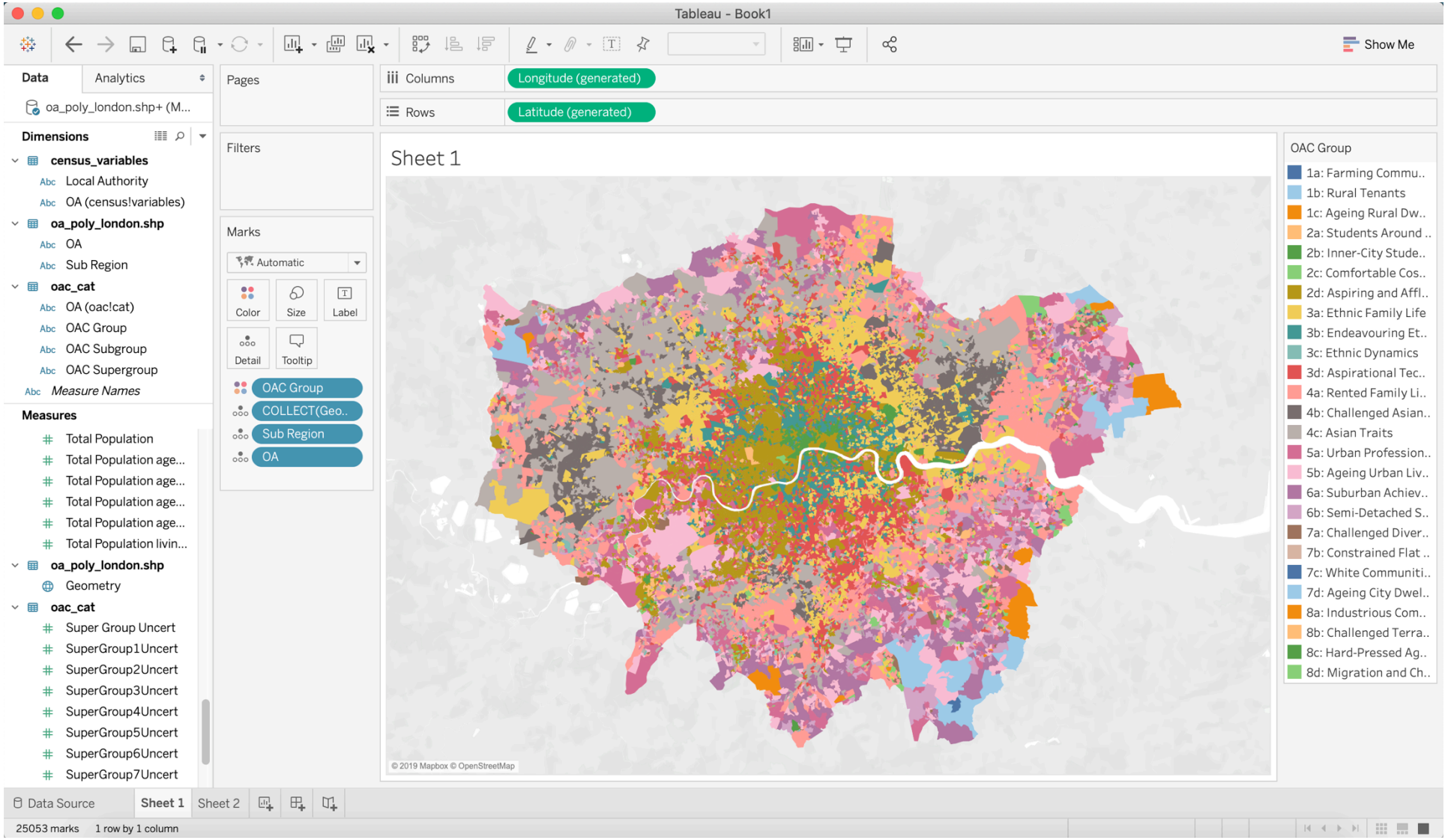


Tableau - Book1
Show Me
Data
Analytics
oa_poly_london.shp+ (M...
Dimensions
census_variables
Local Authority
OA (census!variables)
oa_poly_london.shp
OA
Sub Region
oac_cat
OA (oac!cat)
OAC Group
OAC Subgroup
OAC Supergroup
Measure Names
Measures
Total Population
Total Population age...
Total Population age...
Total Population age...
Total Population age...
Total Population livin...
oa_poly_london.shp
Geometry
oac_cat
Super Group Uncert
SuperGroup1Uncert
SuperGroup2Uncert
SuperGroup3Uncert
SuperGroup4Uncert
SuperGroup5Uncert
SuperGroup6Uncert
SuperGroup7Uncert
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
OAC Group
COLLECT(Geo..
Sub Region
OA
Columns
Longitude (generated)
Rows
Latitude (generated)
Sheet 1
© 2019 Mapbox © OpenStreetMap
OAC Group
1a: Farming Commu..
1b: Rural Tenants
1c: Ageing Rural Dw..
2a: Students Around ..
2b: Inner-City Stude..
2c: Comfortable Cos..
2d: Aspiring and Affl..
3a: Ethnic Family Life
3b: Endeavouring Et..
3c: Ethnic Dynamics
3d: Aspirational Tec..
4a: Rented Family Li..
4b: Challenged Asian..
4c: Asian Traits
5a: Urban Profession..
5b: Ageing Urban Liv..
6a: Suburban Achiev..
6b: Semi-Detached S..
7a: Challenged Diver..
7b: Constrained Flat ..
7c: White Communiti..
7d: Ageing City Dwel..
8a: Industrious Com..
8b: Challenged Terra..
8c: Hard-Pressed Ag..
8d: Migration and Ch..
Data Source
Sheet 1
Sheet 2
25053 marks
1 row by 1 column

Ethics

- Scraping (including Twitter)
  - Can be a bit of a grey area
  - Observe “terms of use”
  - If grey area, always ask permission before making public
  - Be VERY careful about revealing personal information, even if already “public”
- If you acquire data, make sure that it can be used in a new context and my new analytics
- If you collect personal information, ensure you comply with research ethics policies.

Cambridge Analytica scandal

WRANGLE

Wrangle

- Restructuring, linking and recoding data
- Necessary, time-consuming step and often don't get any credit (which is a bit of a tragedy), but…
- …often the first stage of exploratory data analysis
  - insights into **how suitable data** are (previous step)
  - will need to return to it **during exploratory analysis** (next step)
  - will need to return to it **during analysis**/modelling (the step after next) e.g. getting data out of model results
- Python is very powerful, but lots of GUI tools that incorporate visual summaries
  - https://www.trifacta.com/ or Tableau Prep

Coding

- Don’t worry about code efficiently in early stages
  - Only when it becomes a problem (slow)
  - Only when you’re working with other (next week)

EXPLORATORY ANALYSIS

Investigating the data

- Generally, descriptive statistics
  - Types of variables
  - Counts and distributions: mean, standard deviation, percentiles
  - Relationships between variables: correlation
  - Representativeness, balance, bias, missingness, time series gaps…
- Informs
  - Assessment of suitability, dropping/imputing, etc
  - Transforming or creating new features

Creating new features

- Creating new features based on research question
  - Are you **studying change in something**? Create a feature to describe that.
  - Are you interested in difference between **above/below a known value**? Create a feature to describe that.
  - Are you interested in **geographical/temporal patterns** of point data? Create a feature to describe that (by doing spatio-temporal binning; see VA module)
  - Are you interested in **sentiment or topics** discussed in text? Create a feature to describe that (by using a appropriate library)
  - Are you interested in whether **users' profile picture have cats in**? Create a feature to describe that.

Use visualisation!

- Plot
  - Graphs
  - difference graphs
  - normalised graphs
- Use them
  - inform feature transformation/creation
  - inform what things need to be investigated
  - inform the use of subsequent analytical means
  - inform interpretation of the data

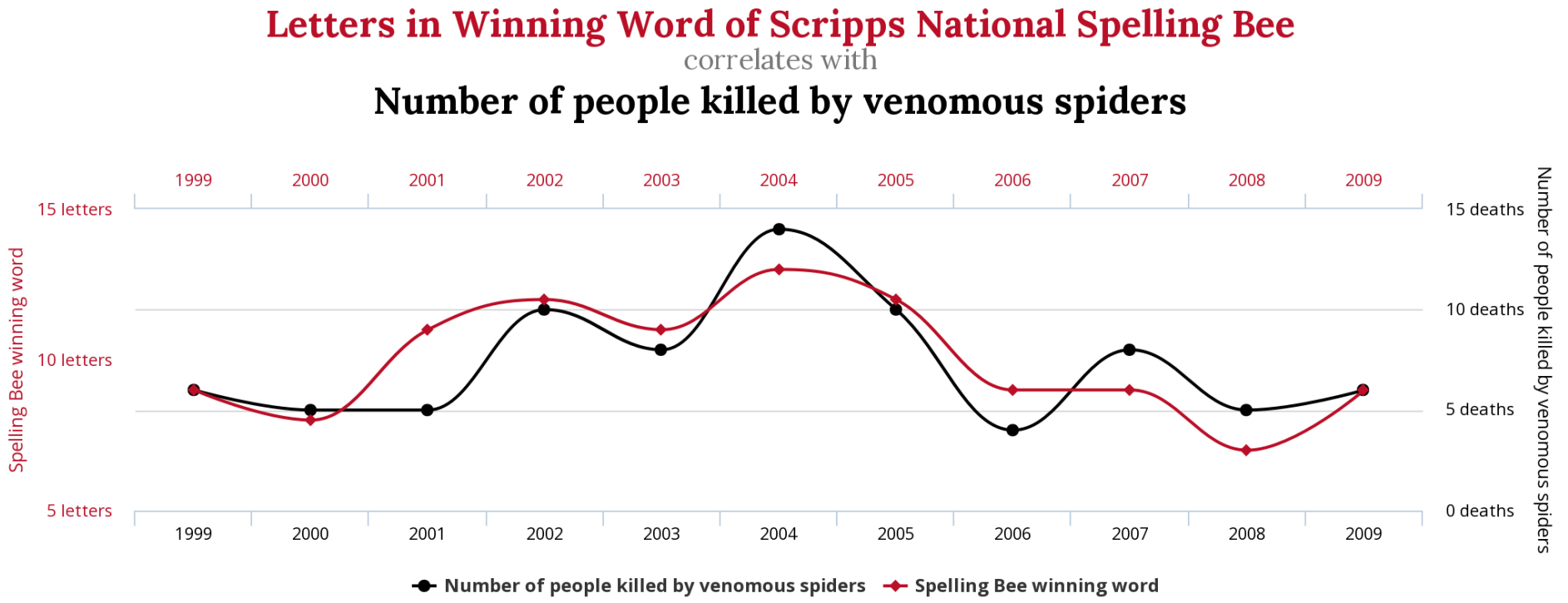
ANALYSE (AND MODEL IF NECESSARY)

Analysing/modelling

- Let’s not look just at the data, let’s start looking at what the data can tell us

Spurious correlations

- Examples
  - http://www.tylervigen.com/spurious-correlations
- Particularly for explanatory models, choose your independent variables carefully



Many analytical techniques at our disposal

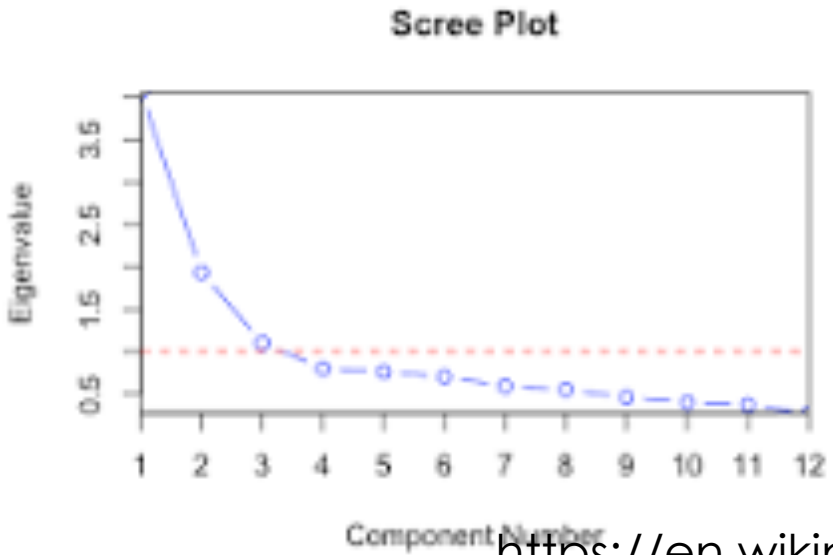
- Dimensional reduction
  - Collapse multiple variables/column/dimensions into a fewer
- Clustering
  - allocate records to groups
- Classification
  - allocate labels/categories to records
- Regression
  - establish mathematical a relationship between a dependent and 1 or more independent variables

Dimensional reduction (used for visualisation)

- Multidimensional scaling (MDS)
- t-SNE/UMAP
  - Non-linear - emphases local structures over global ones
- See demo
  - https://projector.tensorflow.org/

Dimensional reduction

- Principal component analysis (PCA)
  - Generates new variables (components) as weighted combinations of original variables.
  - Represent different (orthogonal) axes of variation
  - The first few usually describe most variation
  - Make sure variables are normalised



https://en.wikipedia.org/wiki/Scree_plot

Dimensional reduction

- Linear Discriminant Analysis (LDA)
  - Supervised
  - Finds the (linear, synthetic) dimension that best discriminates existing categories

Clustering

Classification

Regression

Feature generation

- Some feature generation may involve a lot of analysis and a lot of investigation.

Analytical methods

Does our sample describe the phenomenon?

- Are patterns in the sample are likely to also exist in the wider population?
- Null hypothesis testing and p-values (of the frequentists) are designed to do that. But as discussed, some concerns
  - null hypothesis is not the subject of the research
  - often a reliance on critical values; open to p-hacking
  - often makes assumptions about data distribution
  - very formulaic (good; established process) but may discourage critical thinking.
- These negatives often arise from improper use of these techniques

Bayesian statistics

- **Inferential statistics** and **Bayesian statistics** have different emphases
  - **Inferential statistics**: probabilities based on repeatable random events
  - **Bayesian statistics**: broader definition that can include initial beliefs, past experience. You update your view based on new evidence
- There's more to it than that (and I don't fully understand the theory)…
- …but for our purposes, the Bayesian approach is increasingly being adopted
  - more compatible with repurposing data and taking data from multiple sources

REFLECT ON FINDINGS (AND MAKE RECOMMENDATIONS)

COMMUNICATE FINDINGS