UK Data Service

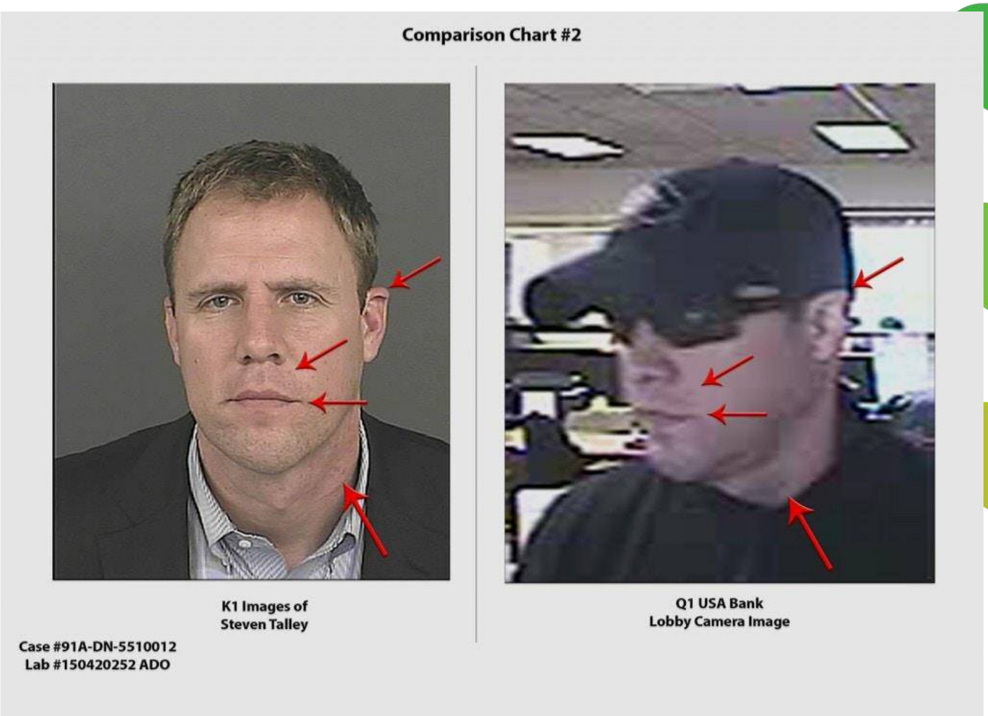
Getting started with secondary data analysis:

Dr Ana Morales-Gomez

Introduction to analysing data about crime using R
Manchester
4-5 February 2020

Data and Crime

- Jack Maple and the **“Charts of the Future”**
- Steve Talley: **How facial recognition can ruin your life**
- Paul Zilly: **Human versus Machine**



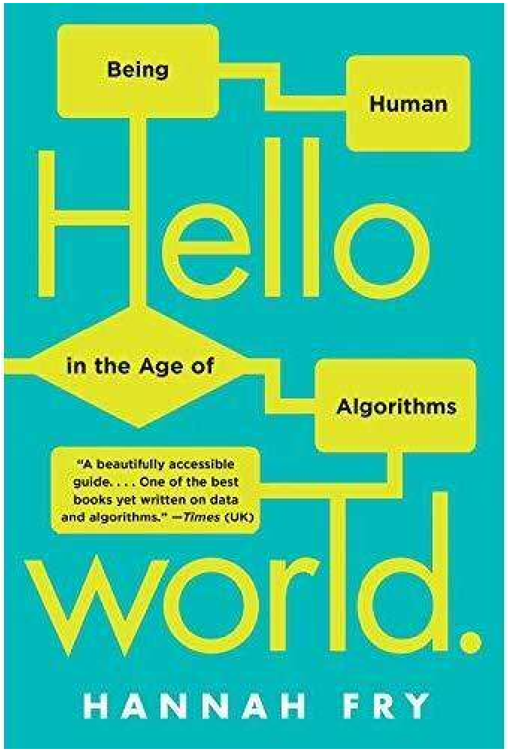
Jack Maple: https://en.wikipedia.org/wiki/Jack_Maple
Steve Talley: https://theintercept.com/2016/10/13/how-a-facial-recognition-mismatch-can-ruin-your-life/
Paul Zilly: https://www.sciencefocus.com/future-technology/can-an-algorithm-deliver-

UK Data Service

More about Data and Crime

Chapter Justice

Chapter Crime



Being

Human

Hello

in the Age of

Algorithms

"A beautifully accessible guide. . . . One of the best books yet written on data and algorithms." —*Times* (UK)

world.

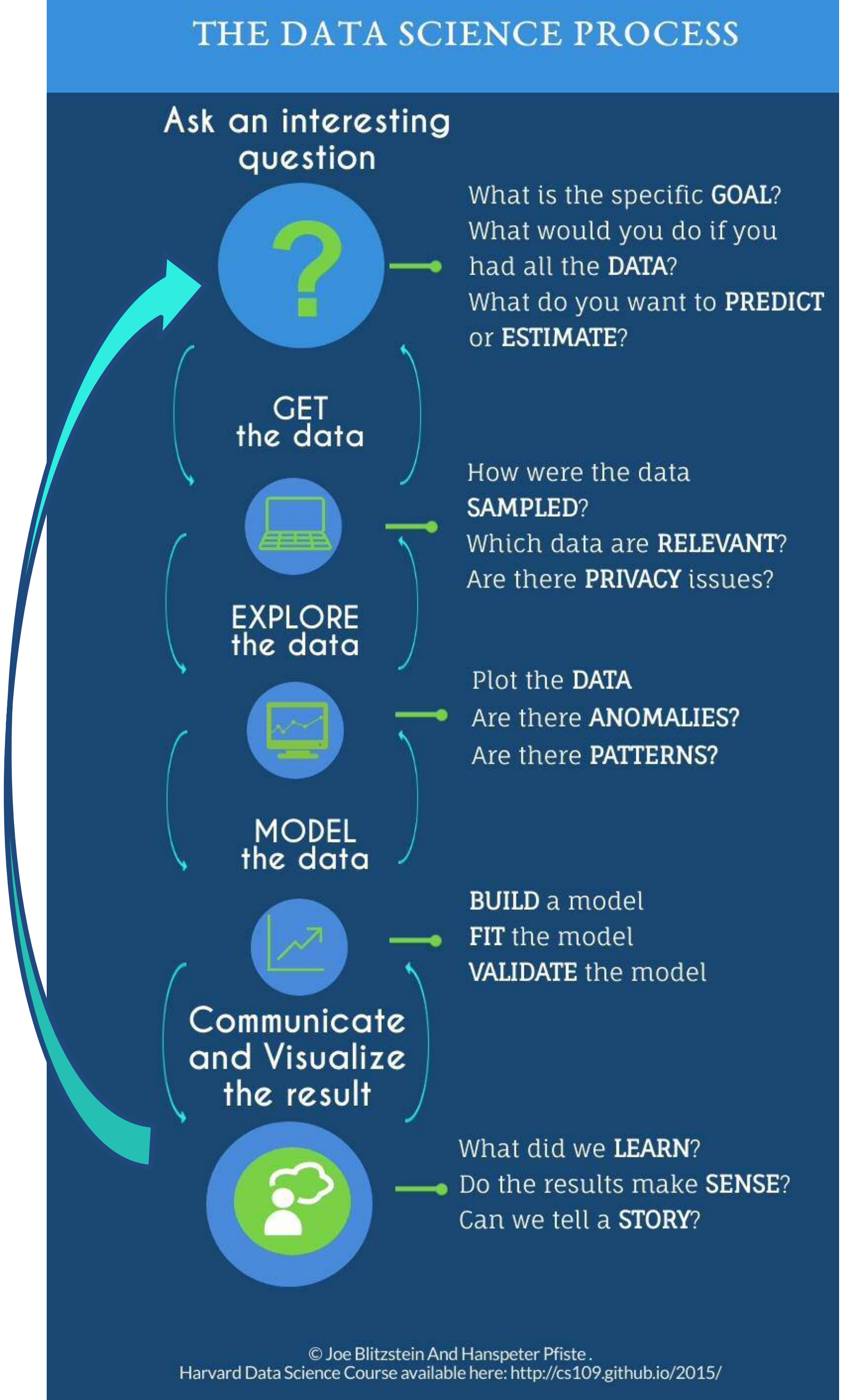
HANNAH FRY

Twitter: @FryRsquared

UK Data Service

What is Research?

Adding a contribution to an existing body of knowledge



THE DATA SCIENCE PROCESS

Ask an interesting question

What is the specific **GOAL**?
What would you do if you had all the **DATA**?
What do you want to **PREDICT** or **ESTIMATE**?

GET the data

How were the data **SAMPLED**?
Which data are **RELEVANT**?
Are there **PRIVACY** issues?

EXPLORE the data

Plot the **DATA**
Are there **ANOMALIES?**
Are there **PATTERNS?**

MODEL the data

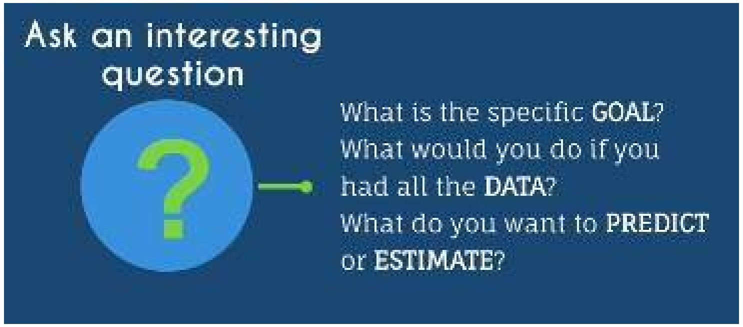
BUILD a model
FIT the model
VALIDATE the model

Communicate and Visualize the result

What did we **LEARN**?
Do the results make **SENSE**?
Can we tell a **STORY**?

© Joe Blitzstein And Hanspeter Pfiste .
Harvard Data Science Course available here: http://cs109.github.io/2015/

The data Science process (1): Ask an interesting question



Ask an interesting question

?

What is the specific GOAL?
What would you do if you had all the DATA?
What do you want to PREDICT or ESTIMATE?

- ✓ **A topic of interest:**
  - ➢ Crime
  - ➢ Health inequalities
  - ➢ Pollution

- ✓ **Specific goal**
  1. Confidence in the Criminal Justice System in England
  2. Antisocial behaviour in Manchester

- ✓ **What do you want to predict or estimate?**
  - ➢ National level estimates
  - ➢ Local level indicators
  - ➢ CJS as a whole or concentrate on Police, prisons, Sentencing?

UK Data Service

The data Science process (2): Get the data

GET the data

How were the data **SAMPLED**?
Which data are **RELEVANT**?
Are there **PRIVACY** issues?

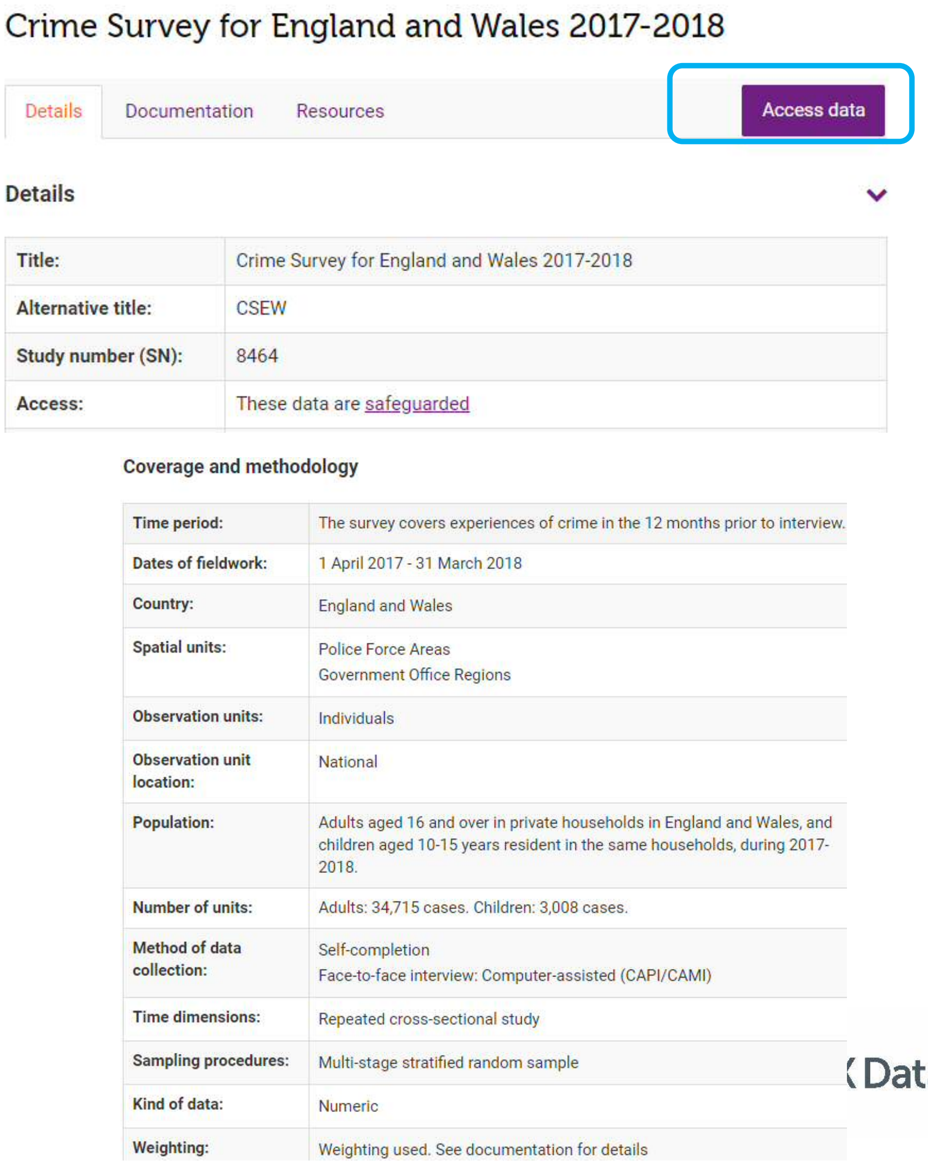
- ✓ **Police recorded crime data**
- ✓ **CSEW: For England and Wales**
- ✓ **Scottish Crime Survey**
- ✓ **European Social Survey**
- ✓ **Others:**
  - ➢ Administrative data of prisons
  - ➢ Administrative data sentencing council

UK Data Service

The data Science process (2): Get the data

How were the data **SAMPLED**?
Which data are **RELEVANT**?
Are there **PRIVACY** issues?

- ✓ **Police recorded crime data:**
  - ✓ **[CSEW UK Data Service]**
  - ➢ **Coverage**
    - Date range
    - Spatial units
  - ➢ **What data**
    - Available for surveys
    - Open data may not have any
  - ➢ **Format**
    - Depending on the source. UKDS: Stata, SPSS,
    - Excel
    - Text



Crime Survey for England and Wales 2017-2018

Details

Title:	Crime Survey for England and Wales 2017-2018
Alternative title:	CSEW
Study number (SN):	8464
Access:	These data are safeguarded

Coverage and methodology

Time period:	The survey covers experiences of crime in the 12 months prior to interview.
Dates of fieldwork:	1 April 2017 - 31 March 2018
Country:	England and Wales
Spatial units:	Police Force Areas Government Office Regions
Observation units:	Individuals
Observation unit location:	National
Population:	Adults aged 16 and over in private households in England and Wales, and children aged 10-15 years resident in the same households, during 2017-2018.
Number of units:	Adults: 34,715 cases. Children: 3,008 cases.
Method of data collection:	Self-completion Face-to-face interview: Computer-assisted (CAPI/CAMI)
Time dimensions:	Repeated cross-sectional study
Sampling procedures:	Multi-stage stratified random sample
Kind of data:	Numeric
Weighting:	Weighting used. See documentation for details

The data Science process (3a): Explore the data

EXPLORE the data

Plot the DATA

Are there ANOMALIES?

Are there PATTERNS?

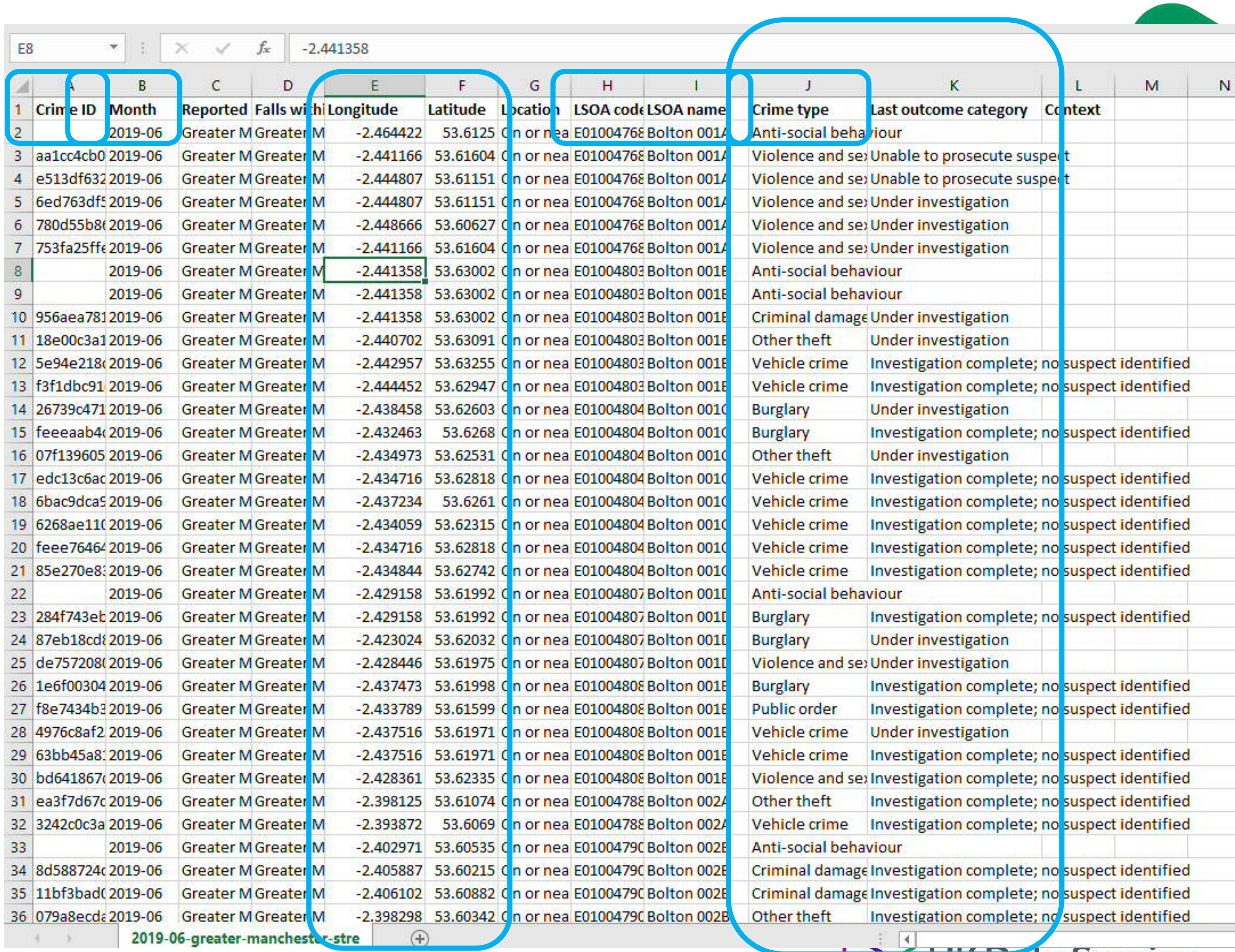
✓ **What data do we have?**

- **Variables**

  (name some variables)

- **Type of data**
  - Numeric?
  - Attribute (character)

- **Is it ready to analyse?**
  - Data cleaning
  - Manipulation



E8 -2.441358

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Crime ID	Month	Reported	Falls withi	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type	Last outcome category	Context		
2		2019-06	Greater M	Greater M	-2.464422	53.6125	On or nea	E01004768	Bolton 001A	Anti-social behaviour				
3	aa1cc4cb0	2019-06	Greater M	Greater M	-2.441166	53.61604	On or nea	E01004768	Bolton 001A	Violence and se	Unable to prosecute suspect			
4	e513df632	2019-06	Greater M	Greater M	-2.444807	53.61151	On or nea	E01004768	Bolton 001A	Violence and se	Unable to prosecute suspect			
5	6ed763df5	2019-06	Greater M	Greater M	-2.444807	53.61151	On or nea	E01004768	Bolton 001A	Violence and se	Under investigation			
6	780d55b86	2019-06	Greater M	Greater M	-2.448666	53.60627	On or nea	E01004768	Bolton 001A	Violence and se	Under investigation			
7	753fa25ffe	2019-06	Greater M	Greater M	-2.441166	53.61604	On or nea	E01004768	Bolton 001A	Violence and se	Under investigation			
8		2019-06	Greater M	Greater M	-2.441358	53.63002	On or nea	E01004803	Bolton 001B	Anti-social behaviour				
9		2019-06	Greater M	Greater M	-2.441358	53.63002	On or nea	E01004803	Bolton 001B	Anti-social behaviour				
10	956aea781	2019-06	Greater M	Greater M	-2.441358	53.63002	On or nea	E01004803	Bolton 001B	Criminal damage	Under investigation			
11	18e00c3a1	2019-06	Greater M	Greater M	-2.440702	53.63091	On or nea	E01004803	Bolton 001B	Other theft	Under investigation			
12	5e94e218c	2019-06	Greater M	Greater M	-2.442957	53.63255	On or nea	E01004803	Bolton 001B	Vehicle crime	Investigation complete; no suspect identified			
13	f3f1dbc91	2019-06	Greater M	Greater M	-2.444452	53.62947	On or nea	E01004803	Bolton 001B	Vehicle crime	Investigation complete; no suspect identified			
14	26739c471	2019-06	Greater M	Greater M	-2.438458	53.62603	On or nea	E01004804	Bolton 001C	Burglary	Under investigation			
15	feeeaab4c	2019-06	Greater M	Greater M	-2.432463	53.6268	On or nea	E01004804	Bolton 001C	Burglary	Investigation complete; no suspect identified			
16	07f139605	2019-06	Greater M	Greater M	-2.434973	53.62531	On or nea	E01004804	Bolton 001C	Other theft	Under investigation			
17	edc13c6ac	2019-06	Greater M	Greater M	-2.434716	53.62818	On or nea	E01004804	Bolton 001C	Vehicle crime	Investigation complete; no suspect identified			
18	6bac9dca9	2019-06	Greater M	Greater M	-2.437234	53.6261	On or nea	E01004804	Bolton 001C	Vehicle crime	Investigation complete; no suspect identified			
19	6268ae110	2019-06	Greater M	Greater M	-2.434059	53.62315	On or nea	E01004804	Bolton 001C	Vehicle crime	Investigation complete; no suspect identified			
20	feee76464	2019-06	Greater M	Greater M	-2.434716	53.62818	On or nea	E01004804	Bolton 001C	Vehicle crime	Investigation complete; no suspect identified			
21	85e270e83	2019-06	Greater M	Greater M	-2.434844	53.62742	On or nea	E01004804	Bolton 001C	Vehicle crime	Investigation complete; no suspect identified			
22		2019-06	Greater M	Greater M	-2.429158	53.61992	On or nea	E01004807	Bolton 001D	Anti-social behaviour				
23	284f743eb	2019-06	Greater M	Greater M	-2.429158	53.61992	On or nea	E01004807	Bolton 001D	Burglary	Investigation complete; no suspect identified			
24	87eb18cd8	2019-06	Greater M	Greater M	-2.423024	53.62032	On or nea	E01004807	Bolton 001D	Burglary	Under investigation			
25	de7572080	2019-06	Greater M	Greater M	-2.428446	53.61975	On or nea	E01004807	Bolton 001D	Violence and se	Under investigation			
26	1e6f00304	2019-06	Greater M	Greater M	-2.437473	53.61998	On or nea	E01004808	Bolton 001E	Burglary	Investigation complete; no suspect identified			
27	f8e7434b3	2019-06	Greater M	Greater M	-2.433789	53.61599	On or nea	E01004808	Bolton 001E	Public order	Investigation complete; no suspect identified			
28	4976c8af2	2019-06	Greater M	Greater M	-2.437516	53.61971	On or nea	E01004808	Bolton 001E	Vehicle crime	Under investigation			
29	63bb45a81	2019-06	Greater M	Greater M	-2.437516	53.61971	On or nea	E01004808	Bolton 001E	Vehicle crime	Investigation complete; no suspect identified			
30	bd641867c	2019-06	Greater M	Greater M	-2.428361	53.62335	On or nea	E01004808	Bolton 001E	Violence and se	Investigation complete; no suspect identified			
31	ea3f7d67c	2019-06	Greater M	Greater M	-2.398125	53.61074	On or nea	E01004788	Bolton 002A	Other theft	Investigation complete; no suspect identified			
32	3242c0c3a	2019-06	Greater M	Greater M	-2.393872	53.6069	On or nea	E01004788	Bolton 002A	Vehicle crime	Investigation complete; no suspect identified			
33		2019-06	Greater M	Greater M	-2.402971	53.60535	On or nea	E01004790	Bolton 002B	Anti-social behaviour				
34	8d588724c	2019-06	Greater M	Greater M	-2.405887	53.60215	On or nea	E01004790	Bolton 002B	Criminal damage	Investigation complete; no suspect identified			
35	11bf3bad0	2019-06	Greater M	Greater M	-2.406102	53.60882	On or nea	E01004790	Bolton 002B	Criminal damage	Investigation complete; no suspect identified			
36	079a8ecda	2019-06	Greater M	Greater M	-2.398298	53.60342	On or nea	E01004790	Bolton 002B	Other theft	Investigation complete; no suspect identified			

2019-06-greater-manchester-stre

The data Science process (3b): Explore the data

EXPLORE the data

Plot the DATA

Are there ANOMALIES?

Are there PATTERNS?

✓ What data do we have?

- **Variables**

  (name some variables)

- **Type of data**
  - Numeric?
  - Attribute (character)

- **Is it ready to analyse?**
  - Data cleaning
  - Manipulation

✓ Descriptive statistics

- **Central tendency measures**
  - Any correlations?
  - Anomalies?

- **Plot the data**
  - Anomalies?
  - Patterns?

- **More questions**
  - Are the data enough for my RQ?
  - Do we need more data?
  - Is there more data?
  - Change RQs?

The data Science process (4): Model the data

MODEL the data
BUILD a model
FIT the model
VALIDATE the model

✓ **What is the best approach to understand the data we have?**

- **Depends on…**
  - Our research questions
  - Our data available

- **Example:**
  - **Correlation to look for association of two variables**
  - **Generalised Linear models /Regression based models for**
    - Multiple linear regression (continuous outcome)
    - Logistic/Probit regression (binary outcome)
    - Ordinal regressions
    - Multilevel models (clusters and hierarchy dependence)
    - Longitudinal models (samples at different time points)

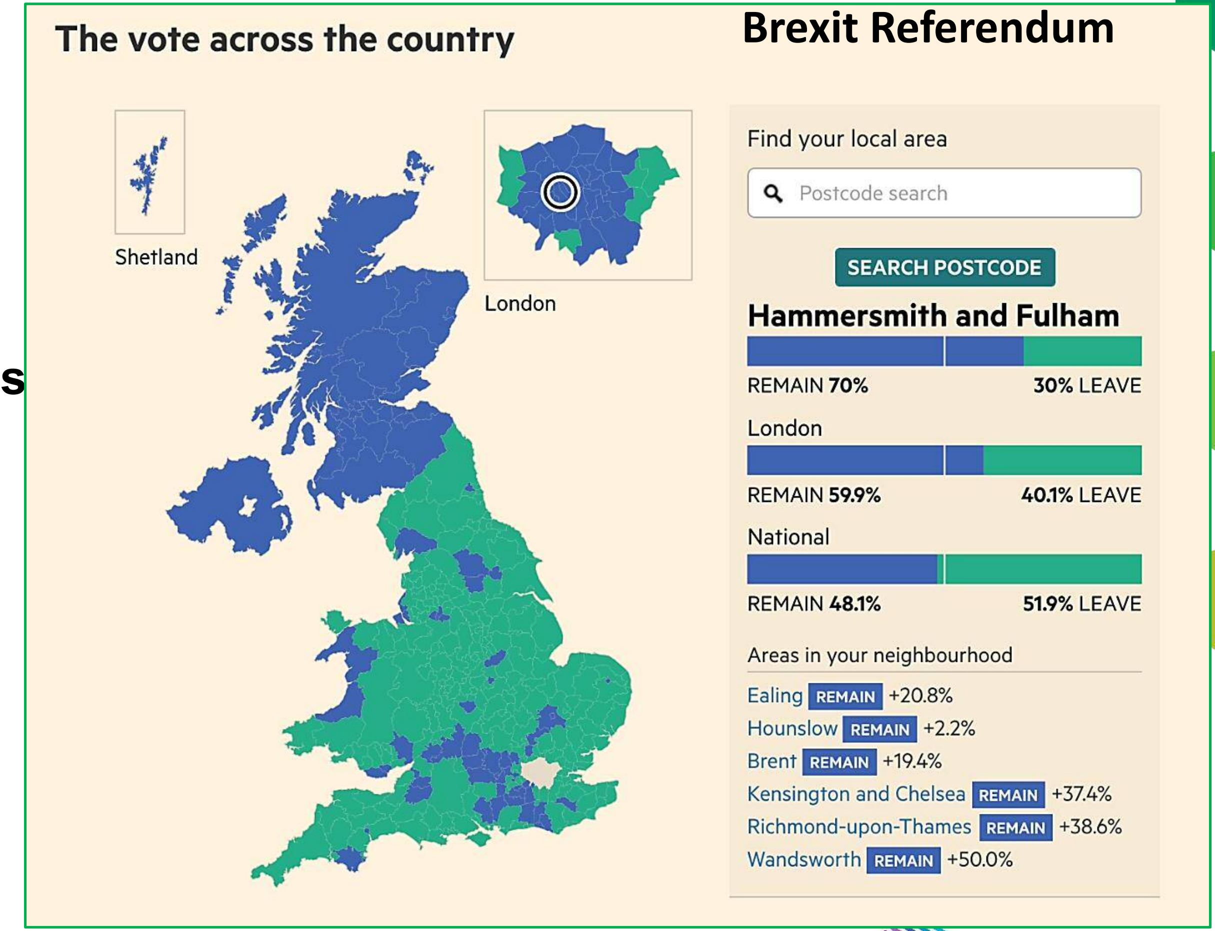
The data Science process (5): Communicate and visualise the results

Communicate and Visualize the result

What did we LEARN?
Do the results make SENSE?
Can we tell a STORY?

- ✓ **Visualise the results**
  - ➢ Tables
  - ➢ Figures
  - ➢ Plots
  - ➢ Maps

- ✓ **Communicate the results**
  - ➢ **Know your audience**
    - Effective
    - The right details for each audience
    - Academic ≠ Local Government officers


Brexit Referendum

The vote across the country

Find your local area

Postcode search

SEARCH POSTCODE

Hammersmith and Fulham

REMAIN **70%** | **30%** LEAVE

London

REMAIN **59.9%** | **40.1%** LEAVE

National

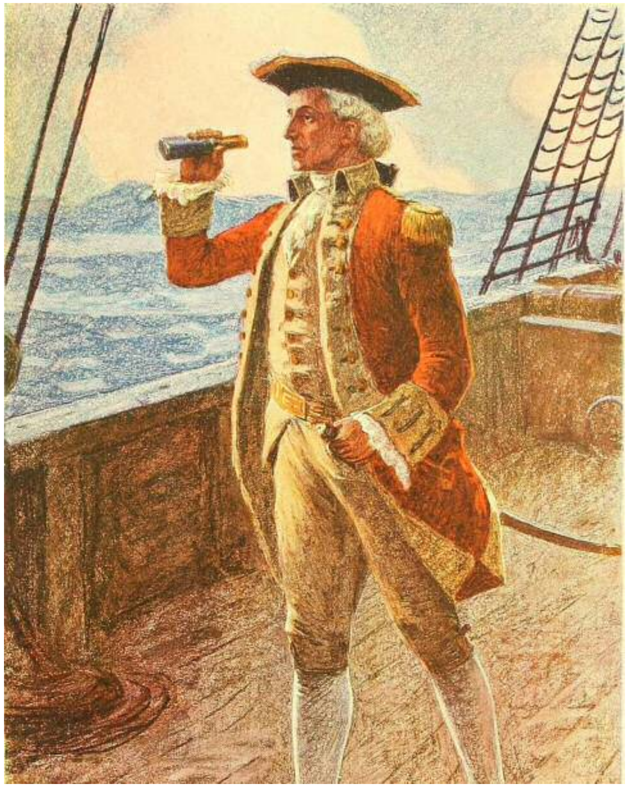
REMAIN **48.1%** | **51.9%** LEAVE

Areas in your neighbourhood

Ealing	REMAIN	+20.8%
Hounslow	REMAIN	+2.2%
Brent	REMAIN	+19.4%
Kensington and Chelsea	REMAIN	+37.4%
Richmond-upon-Thames	REMAIN	+38.6%
Wandsworth	REMAIN	+50.0%

Source: Financial Times: https://ig.ft.com/sites/elections/2016/uk/eu-referendum/

UK Data Service

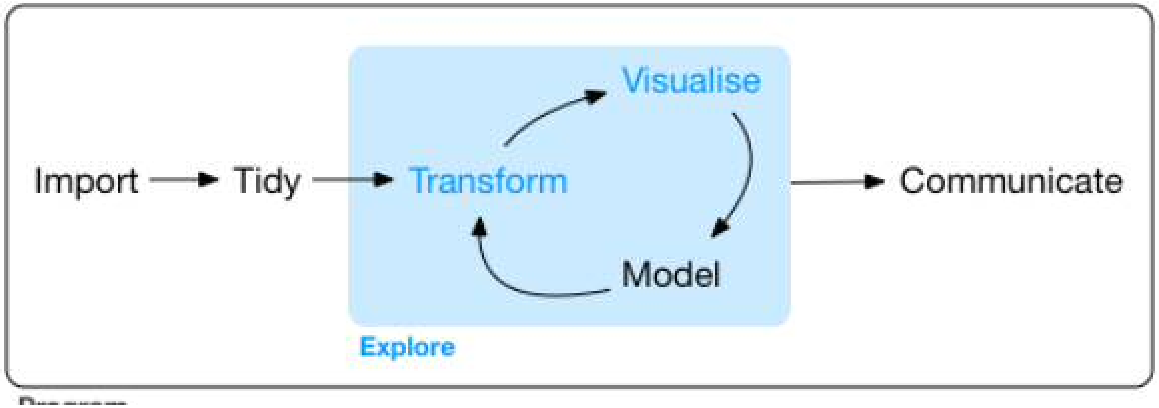


Exploratory Data Analysis

UK Data Service

Exploratory data analysis

Flowchart for data preparation



From R for data science

https://r4ds.had.co.nz/explore-intro.html

What is data?

- **Information**, especially **facts or numbers**, collected to be examined and considered and used **to help decision-making**, or information in an electronic form that can be stored and used by a computer (Cambridge dictionary)
    - Numeric
    - Images
    - Attributes (characters)

NOMINAL
unordered descriptions
I'm a TURTLE!
I'm a snail!
I'm a butterfly!

ORDINAL
ordered descriptions
I am unhappy.
I am OK.
I am AWESOME!!!

BINARY
ONLY 2 MUTUALLY EXCLUSIVE OUTCOMES
I am EXTINCT!
HA.
@allison_horst

CONTINUOUS
measured data, can have ∞ values within possible range.
I AM 3.1" TALL
I WEIGH 34.16 grams

DISCRETE
observations can only exist at limited values, often counts.
I HAVE 8 LEGS and 4 SPOTS!
@allison_horst

Figures: © Allison Horst

Describe the data

- ✓ **To Understand:**
  - ➢ data availability,
  - ➢ Types,
  - ➢ quality,
  - ➢ data complexity (i.e. nonlinearity, requires transformation, etc)

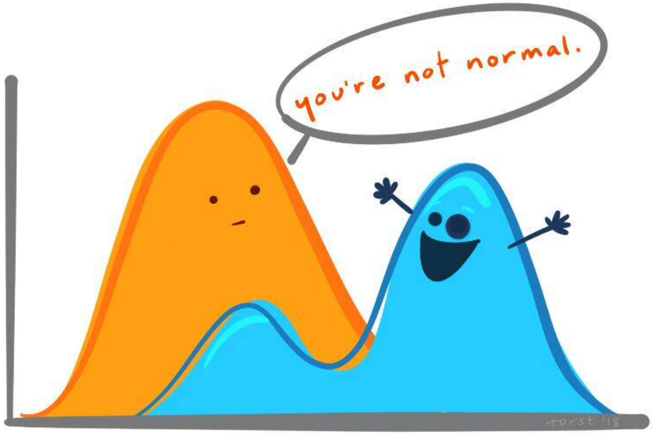
- ✓ **Guided by two types of questions (Grolemund and Wickham, 2016):**
  - ➢ What type of covariation occurs between my variables?
  - ➢ What type of variation occurs within my variables?

UK Data Service

How to describe the data (1)

✓ **Distribution of numerical variables:**

- ➢ Extreme values (outliers)
- ➢ Shape of the distribution
- ➢ Missing cases
- ➢ Unusual patterns



✓ **Distribution of categorical variables**

- ➢ Missing cases
- ➢ Odd values
- ➢ Unusual patterns
- ➢ Most common values

Figure: © Allison Horst

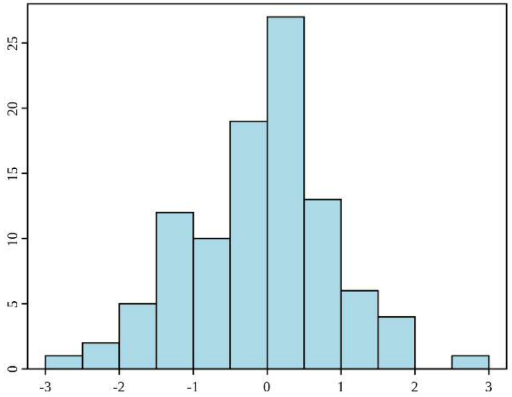
UK Data Service

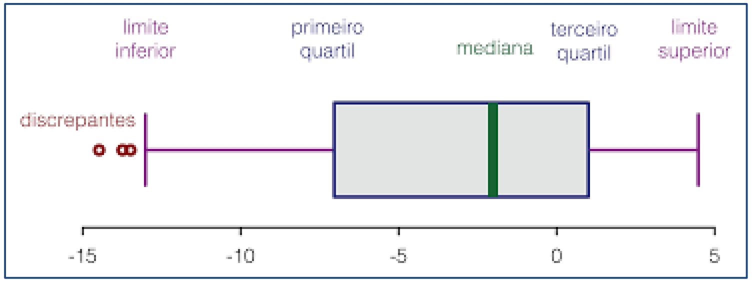
How to describe the data (2)

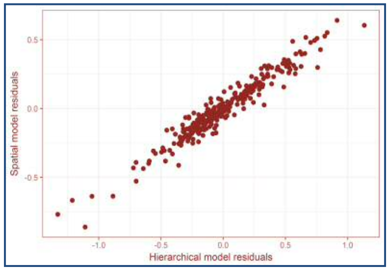
- ✓ **Central tendency measures**
  - ➢ Mean
  - ➢ median
  - ➢ mode

- ✓ **Measures of spread**
  - ➢ Variance and standard deviation
  - ➢ Range:
    - ➢ Interquartile range (IQR)

- ✓ **Visualisations**
  - ➢ Histograms, boxplots, bar plots, scatterplots







How to describe the data (3)

✓ Mean (sample vs. population):

$$\mu = \frac{\sum x}{N}$$

- The "average" number; found by adding all data points and dividing by the number of data points

✓ Median

- Middle value if odd number of values, or average of the middle two values otherwise

✓ Mode

- Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

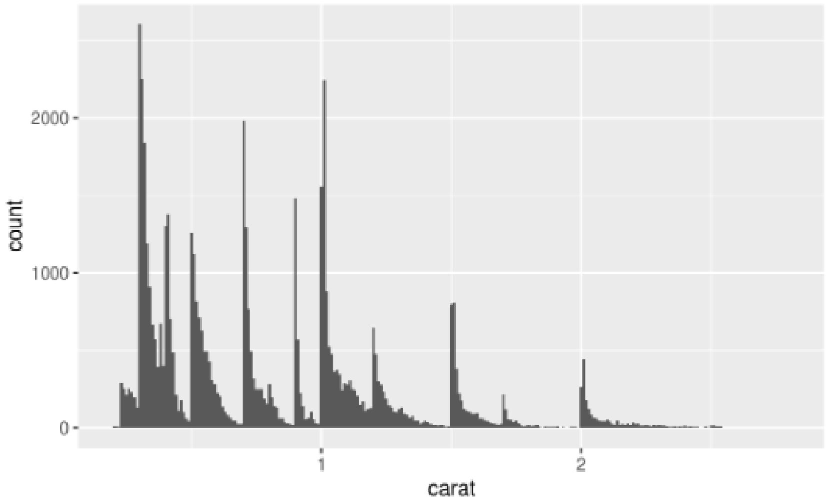
Is the mean always the best central tendency measure?

The problem with the mean

“There are two pieces of bread. You eat two. I eat none. Average consumption: one bread per person.“

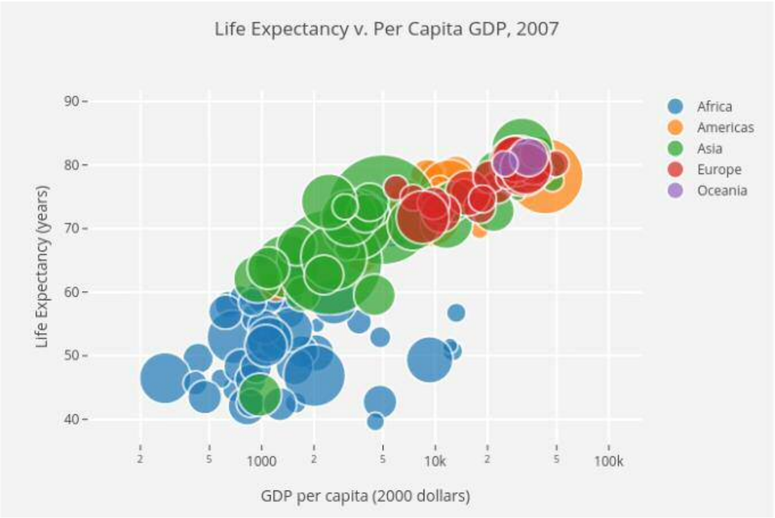
Nicanor Parra, (Anti)Poet, Mathematician and Physicist

More about visualisations



Two types:

1. **Exploring and getting to know the data**
   1. Assess the data: decide what to do next
   2. Accurate
   3. Internal, never reach the wider audience



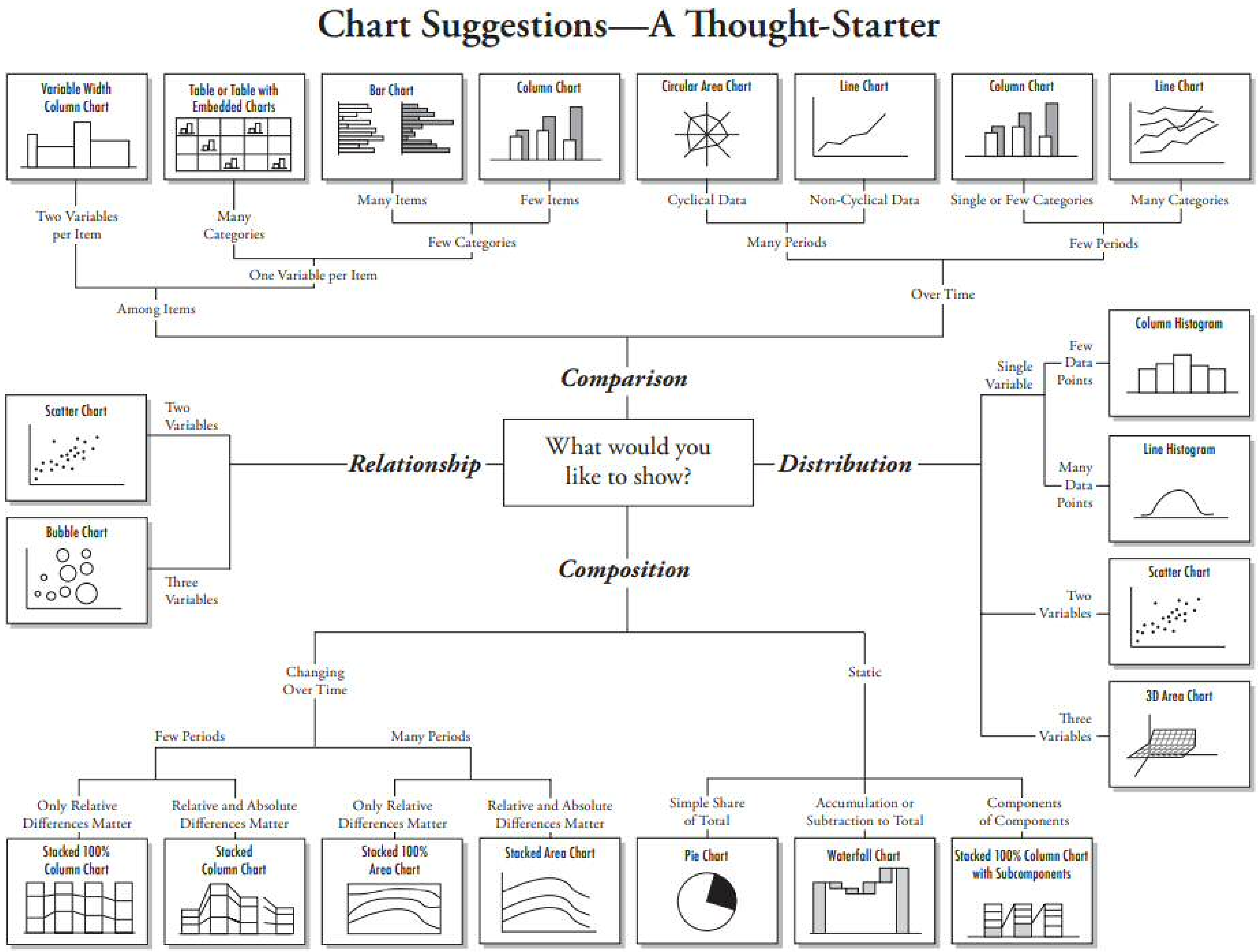
2. **Communication**
   1. Present data and ideas
   2. Accurate: provide evidence
   3. Easy to understand
   4. Effective
   5. It would depend on the audience

Images: https://r4ds.had.co.nz/exploratory-data-analysis.html
https://towardsdatascience.com/5-quick-and-easy-data-visualizations-in-python-with-code-a2284bae952f

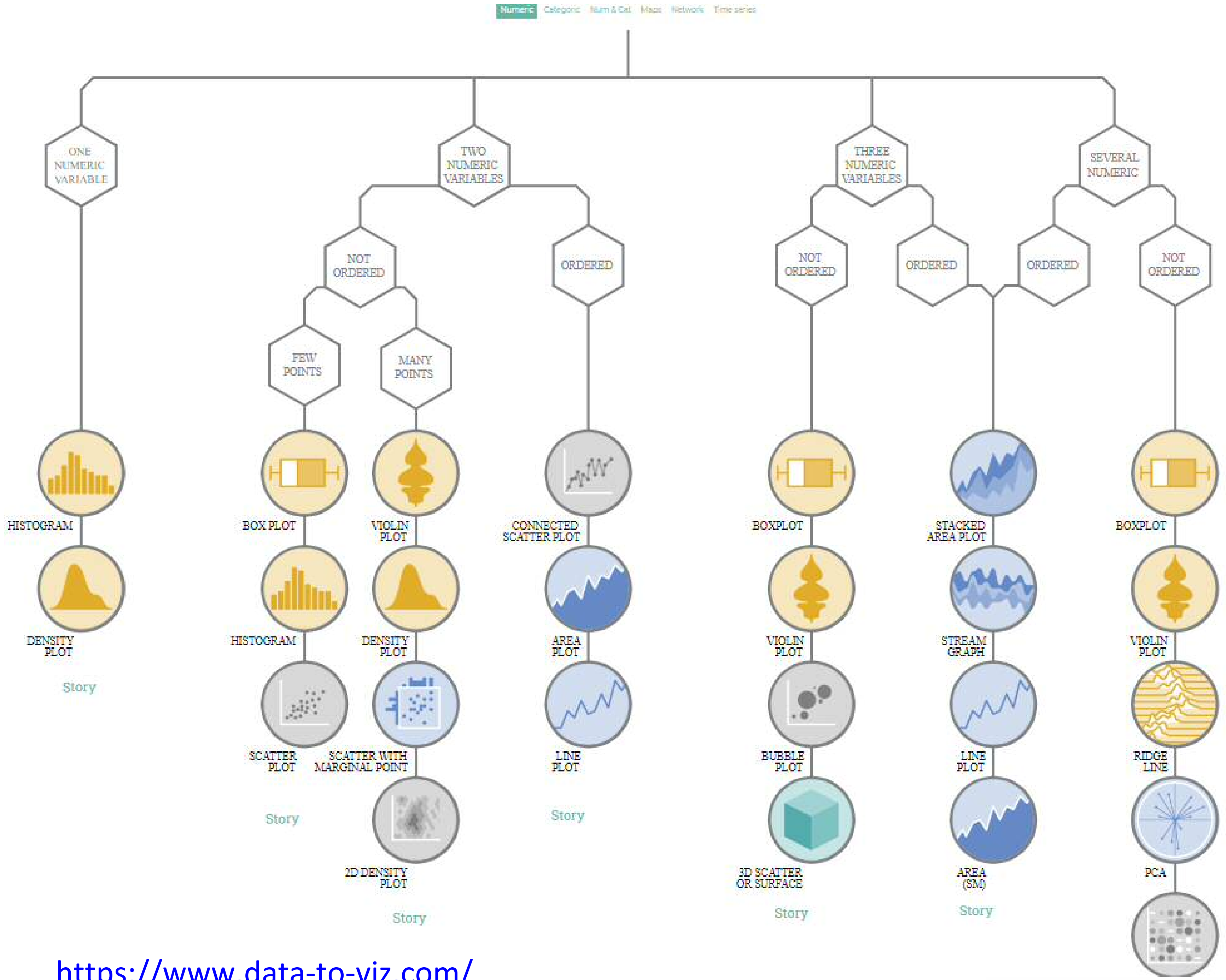
Effective visualisations for communication

- ✓ **Simple but effective (don’t over do it!)**
  - ✓ Easy to understand
- ✓ **Use the right type of graph of figure**
  - ✓ Not a fit them all purpose graph
- ✓ **Appropriate use of colours (colour blind people)**
- ✓ **Know your audience**

Effective visualisations for communication



https://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf



https://www.data-to-viz.com/

Effective visualisations for communication: Use the right display

- ✓ Comparisons:
  - ➢ Bars
  - ➢ lines
- ✓ Proportions
  - ➢ Pie charts
  - ➢ Stacked charts
- ✓ Trends over time
  - ➢ Lines
  - ➢ Scatterplots
- ✓ Distributions
  - ➢ Density plots
  - ➢ Histograms
- ✓ Correlations
  - ➢ Scatterplots

Your turn

Questions

Ana Morales-Gomez

ana.morales@Manchester.ac.uk