Essentials of quantitative research

Shannon Mickelson Campbell, MPP
Principal Data Scientist, St James's Hospital
shcampbell@stjames.ie

Winter Clinical Research Seminar, 24 Feb 2026

OSPIDÉAL SAN SÉAMAS
ST JAMES'S HOSPITAL

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

ECI COMPREHENSIVE Cancer Centre

Quantitative data foundations

- Framing the study & selecting variables
- Data structure
- Data quality
- Statistics & model selection
- Communicating results

Quantitative data foundations

How you think you'll spend your time on a quantitative project

Quantitative data foundations

How you think you'll spend your time on a quantitative project

How you'll actually spend your time on a quantitative project

Framing the study

- What are you trying to do?
  - Explore or describe: e.g., learn more about the characteristics of a particular patient population
  - Explain or evaluate: e.g., how does the presence of gene A influence the progression of disease B, with or without considering other factors; does this intervention impact patient outcomes?
- What is it you're truly interested in, the core of what you want to learn and why?
  - Operationalising that – from both a maths/technical perspective and an ethics/philosophical perspective

Framing the study

- What type of data are you going to use?
  - Collected for this purpose – clinical trials, surveys, etc.
  - "Real world data" (RWD)
    - Gathered in normal course of medical practice, not necessarily for further evaluation or research purposes – e.g., EPR data, insurance claims, registries
  - Pros/cons of each

Framing the study

- Selecting and understanding your variables (both before and after dataset selection)
  - Subject matter expertise
  - Selectiveness – an art and a science
    - Multicollinearity (multiple variables measuring similar concept)
    - Realism of gathering (fewer high quality data points vs. literally everything with extensive missing values or inaccuracies)
    - Tight definitions – mutually exclusive categories; all possible answers incorporated; consistency in measurements
  - Timing

"On the Origins of Data"

- Operational understanding
  - Data provenance – ongoing learning throughout each stage – where is your data coming from?
    - How is it collected? What's the population? What decisions are made during collection?
    - Example: ED visit analysis

"On the Origins of Data"

- Sample questions to ask and document:
  - Who is included? Who is *excluded*? Does it include all potential interventions or only some?
  - How is it generated? (Patient self-reports? Clinical measurements? Consultant opinions?)
  - How "raw" is the data? What has been recoded? Are recodings consistent? Do they match *your* understanding of the concepts?
- Take the time to consult with those who understand the operations best and understand – don't make assumptions

Data structure

- Organisation of data is critical to successful analysis
- Terminology
  - Rows = observations
  - Columns = fields/variables
- Data structure can and will vary for different models
  - A "safe" base
    - 1 row per observation (whether that's per person, per event, etc.), with separate and clearly labeled fields for each variable
    - NO fancy formatting; don't skip rows or columns, don't add headers or images, etc.

Data structure

OSPIDÉAL SAN SÉAMAS
ST JAMES'S HOSPITAL

Appointment Records

Patient ID	Age	Sex	Consultant	Specialty	Appointment Date	Status
AKS79809	52	F	Eoin O Brien	Cardiology	13/01/2025	Attended
MES65300	67	M	Eoin O Brien	Cardiology	13/01/2025	DNA
LWP09123	39	M	Susan Murphy	Cardiology	13/01/2025	DNA
UYS24356	41	F	Susan Murphy	Cardiology	14/01/2025	Attended
HHJ73758	25	F	Fiona Kelly	Urology	10/01/2025	Attended
KEW83948	81	M	James Smith	Dermatology	10/01/2025	DNA

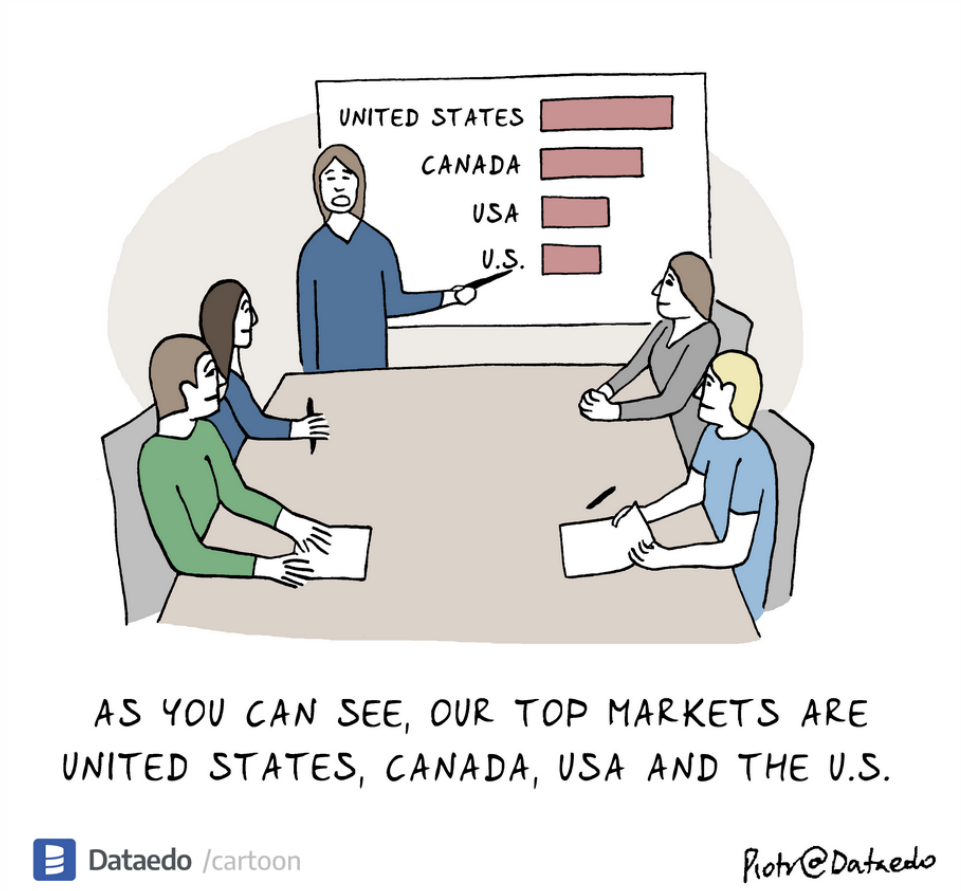
Data structure

	A	B	C	D	E	F	G
1	**Patient ID**	**Age**	**Sex**	**Consultant**	**Specialty**	**Appointment Date**	**Status**
2	AKS79809	52	F	Eoin O Brien	Cardiology	13/01/2025	Attended
3	MES65300	67	M	Eoin O Brien	Cardiology	13/01/2025	DNA
4	LWP09123	39	M	Susan Murphy	Cardiology	13/01/2025	DNA
5	UYS24356	41	F	Susan Murphy	Cardiology	14/01/2025	Attended
6	HHJ73758	25	F	Fiona Kelly	Urology	10/01/2025	Attended
7	KEW83948	81	M	James Smith	Dermatology	10/01/2025	DNA

Data structure

- Structuring our variables: clean, clear, and standardised
  - Clear and accurate in raw or literary form may not be clear and accurate in statistical analysis
    - E.g., how often does patient do X? 2-3 times/week. 10x/month. A few times a week. Etc.
  - Categorical variables should be discrete and exclusive
    - For text (string) variables, be aware of case sensitivity, potential for spelling errors, etc. in review
  - Number format should be consistent (not blending text and numbers, using same units of measurement)
- **Regardless of tool/model demands, a well-crafted, consistent, clean dataset at baseline can be quickly retooled for different purposes**

Examples



On average, how often does patient do X activity?
Never
1-3 times/week
4-7 times/week
Daily
Monthly
Never

Examples

Patient ID	Age	Sex	Consultant	Specialty	Appointment Date	Status
AKS79809	52	F	Eoin O Brien; Susan Murphy	Cardiology	13/01/2025	Attended
MES65300	67	M	Susan Murphy; Eoin O'Brien	Cardiology	13/01/2025	Didn't attend
LWP09123	39	M	Susan Murphy	Cardiology	13/01/2025	DNA
UYS24356	41	F	Susan Murphy	Cardiology	14/01/2025	Atttended
HHJ73758	25	F	Fiona Kelly MCRN 12345	Urology	10/01/2025	Attended
KEW83948	81	M	F Kelly MCRN 12345	Urology	17/01/2025	Attended
KEW83948	81	M	Jim Smith	Dermatology	10/01/2025	Did not attend
AKS79809	52	F	James Smith	Dermatology	15/01/2025	dna

Examples

Patient ID	Age	Sex	Consultant	Specialty	Appointment D
AKS79809	52	F	Eoin O Brien; Susan Murphy	Cardiology	13/01/2
MES65300	67	M	Susan Murphy; Eoin O'Brien	Cardiology	13/01/2
LWP09123	39	M	Susan Murphy	Cardiology	13/01/2
UYS24356	41	F	Susan Murphy	Cardiology	14/01/2
HHJ73758	25	F	Fiona Kelly MCRN 12345	Urology	10/01/2
KEW83948	81	M	F Kelly MCRN 12345	Urology	17/01/2
KEW83948	81	M	Jim Smith	Dermatology	10/01/2
AKS79809	52	F	James Smith	Dermatology	15/01/2

Row Labels	Count of Patient ID
Eoin O Brien; Susan Murphy	1
F Kelly MCRN 12345	1
Fiona Kelly MCRN 12345	1
James Smith	1
Jim Smith	1
Susan Murphy	2
Susan Murphy; Eoin O'Brien	1
Grand Total	**8**

Row Labels	Count of Patient ID
Attended	3
Atttended	1
Did not attend	1
Didn't attend	1
DNA	2
Grand Total	**8**

Examples

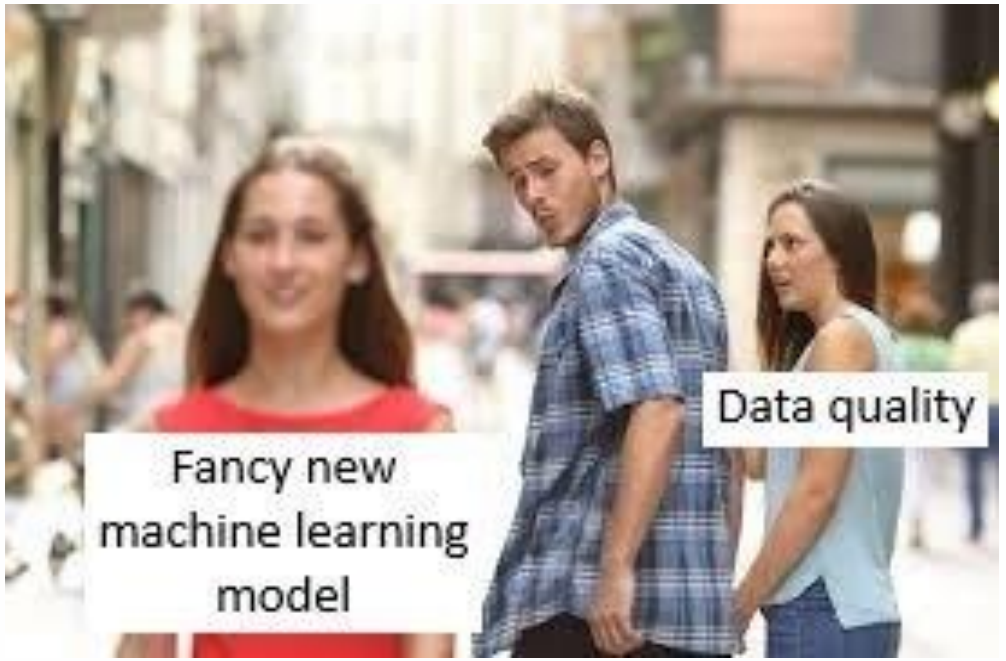
Patient ID	Age	Sex	Consultant	Specialty	Appointment Date	Status	Primary Consultant	Status New	Attended	Cardiology	Urology	Dermatology
AKS79809	52	F	Eoin O Brien; Susan Murphy	Cardiology	13/01/2025	Attended	Eoin O Brien	Attended	1	1	0	0
MES65300	67	M	Susan Murphy; Eoin O'Brien	Cardiology	13/01/2025	DNA	Susan Murphy	DNA	0	1	0	0
LWP09123	39	M	Susan Murphy	Cardiology	13/01/2025	DNA	Susan Murphy	DNA	0	1	0	0
UYS24356	41	F	Susan Murphy	Cardiology	14/01/2025	Atttended	Susan Murphy	Attended	1	1	0	0
HHJ73758	25	F	Fiona Kelly MCRN 12345	Urology	10/01/2025	Attended	Fiona Kelly	Attended	1	0	1	0
KEW83948	81	M	F Kelly MCRN 12345	Urology	17/01/2025	Attended	Fiona Kelly	Attended	1	0	1	0
KEW83948	81	M	Jim Smith	Dermatology	10/01/2025	Did not attend	James Smith	DNA	0	0	0	1
AKS79809	52	F	James Smith	Dermatology	15/01/2025	dna	James Smith	DNA	0	0	0	1

Data structure

- Two key transformation tips
  - Retain your original data within the dataset
  - When in doubt, use flags so you can easily revisit

Data quality

- The best study design, the most sophisticated model...
  ...it can all be destroyed by bad data.



Data quality

- Already discussed understanding the data, the processes that generated it, restructuring and cleaning – all these are also important to data quality!
- But the job isn't finished yet – time to be a 'data detective'

Data detective: the “gut check”

- What do you already know to be true?
- Research should be allowed to challenge our assumptions and surprise us – but also, use your expertise to critically evaluate what it shows you
- Also, back to collaboration with those involved in this clinical/operational area – consider involving them

True or false: There were 30,000 ED visits by 50,000 unique people at SJH last year.

Data detective: the “gut check”

- What do you already know to be true?
- Research should be allowed to challenge our assumptions and surprise us – but also, use your expertise to critically evaluate what it shows you
- Also, back to collaboration with those involved in this clinical/operational area – consider involving them

True or false: There were 24 million ED visits by 10 million unique people at SJH last year.

Data detective: cross-tabs & distributions

- Actually *look* at your data
- Examples of things to look out for:
  - Missing data, and to what degree
  - Date formatting – e.g., 27/10/2025 and 10/27/2025
  - Are unique IDs truly unique
  - Values that make no sense/can't coincide
    - Discharge dates before admission dates; dead and alive simultaneously; conflicting demographic details; etc.
  - Outliers
  - Accuracy in recodings (again, keeping the original data!)

Data detective: from beginning to end and back again

- How many patients/events in each stage of analysis? Are individual patients translating properly from beginning to end?
  - Example: you have a dataset derived from the EPR, and you have done some work transforming it to work for a survival analysis.
    - Looking up several patient records in the EPR that should be included – did they make it into the dataset? Do their various data points translate accurately? Are your transformations reliable?
    - And in reverse – select several cases from the dataset and verify within the EPR.

Examples

Patient ID	Age	Sex	Consultant	Specialty	Appointment Date	Status	Primary Consultant	Status New	Attended	Surgery Scheduled	Deceased
AKS79809	52	F	Eoin O Brien; Susan Murphy	Cardiology	13/01/2025	Attended	Eoin O Brien	Attended	1		Y
MES65300	67	M	Susan Murphy; Eoin O'Brien	Urology	13/01/2025	DNA	Susan Murphy	DNA	0	01/06/2000	N
LWP09123	39	M	Susan Murphy	Cardiology	13/01/2025	DNA	Susan Murphy	DNA	0	15/05/2025	N
UYS24356	41	F	Susan Murphy	Cardiology	14/01/2025	Atttended	Susan Murphy	Attended	1		Y
HHJ73758	25	F	Fiona Kelly MCRN 12345	Urology	10/01/2025	Attended	Fiona Kelly	Attended	1	05/06/2025	N
KEW83948	81	M	F Kelly MCRN 12345	Urology	17/01/2025	Attended	Fiona Kelly	Attended	1	02/06/2025	N
KEW83948	71	M	Jim Smith	Dermatology	10/01/2025	Did not attend	James Smith	DNA	0	30/01/2025	N
AKS79809	52	F	James Smith	Dermatology	15/01/2025	DNA	James Smith	DNA	0	30/01/2025	N

Examples

Patient ID	Age	Sex	Consultant	Specialty	Appointment Date	Status	Primary Consultant	Status New	Attended	Surgery Scheduled	Deceased	Requires further review
AKS79809	52	F	Eoin O Brien; Susan Murphy	Cardiology	13/01/2025	Attended	Eoin O Brien	Attended	1		Y	0
MES65300	67	M	Susan Murphy; Eoin O'Brien	Urology	13/01/2025	DNA	Susan Murphy	DNA	0	01/06/2000	N	1
LWP09123	39	M	Susan Murphy	Cardiology	13/01/2025	DNA	Susan Murphy	DNA	0	15/05/2025	N	1
UYS24356	41	F	Susan Murphy	Cardiology	14/01/2025	Atttended	Susan Murphy	Attended	1		Y	0
HHJ73758	25	F	Fiona Kelly MCRN 12345	Urology	10/01/2025	Attended	Fiona Kelly	Attended	1	05/06/2025	N	0
KEW83948	81	M	F Kelly MCRN 12345	Urology	17/01/2025	Attended	Fiona Kelly	Attended	1	02/06/2025	N	1
KEW83948	71	M	Jim Smith	Dermatology	10/01/2025	Did not attend	James Smith	DNA	0	30/01/2025	N	1
AKS79809	52	F	James Smith	Dermatology	15/01/2025	DNA	James Smith	DNA	0	30/01/2025	N	0

Data quality

- And back to interrogating your data from a contextual and inclusion/exclusion standpoint:
  - If you exclude cases for DQ (e.g. those missing a certain field), who and why? Is impact proportionate/randomised across populations, or concentrated among certain types of patients or events?
  - Not always perfect answers, but must be able to account for and explain

Statistics & modeling

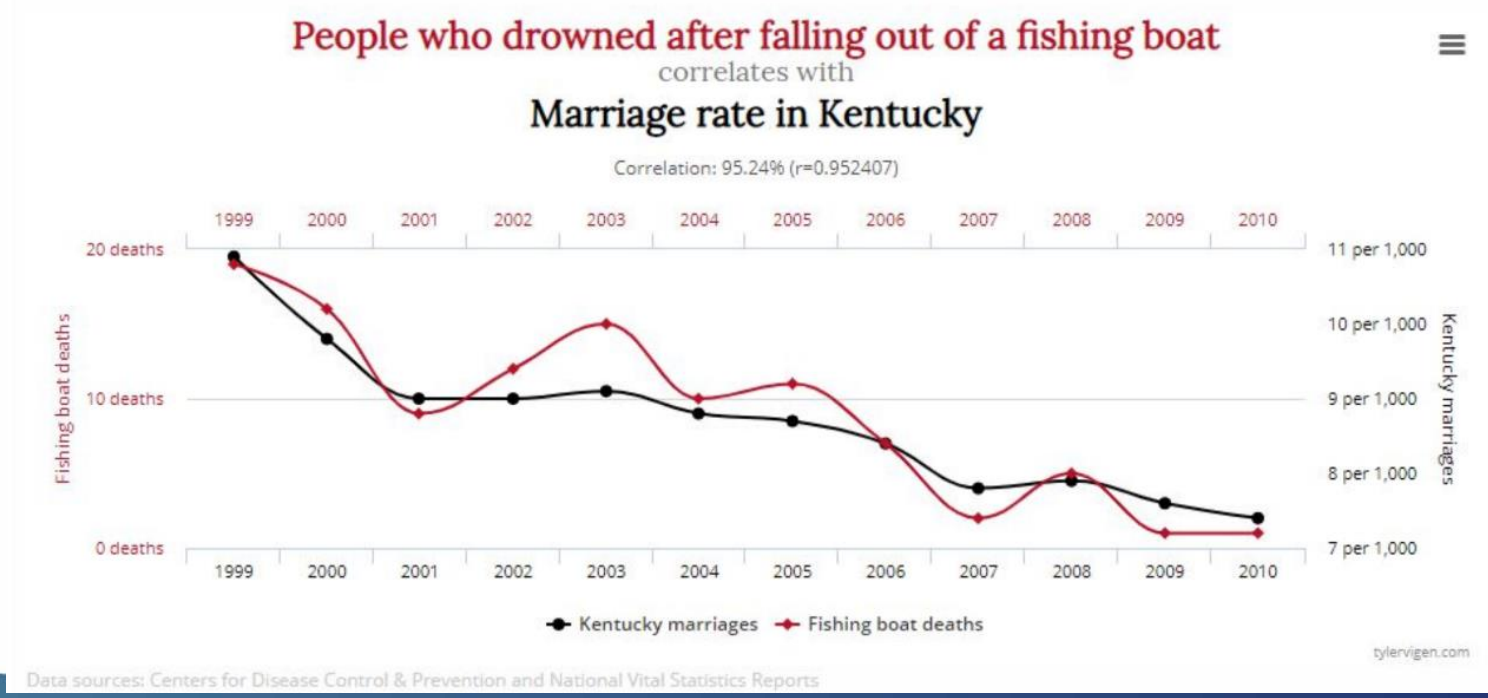
Two things can be true:

1. Better to have very simple descriptive stats derived from excellent data than a highly complex model based on terrible data
2. Multivariate/more nuanced models also help guide you away from oversimplistic correlations

Statistics & modeling

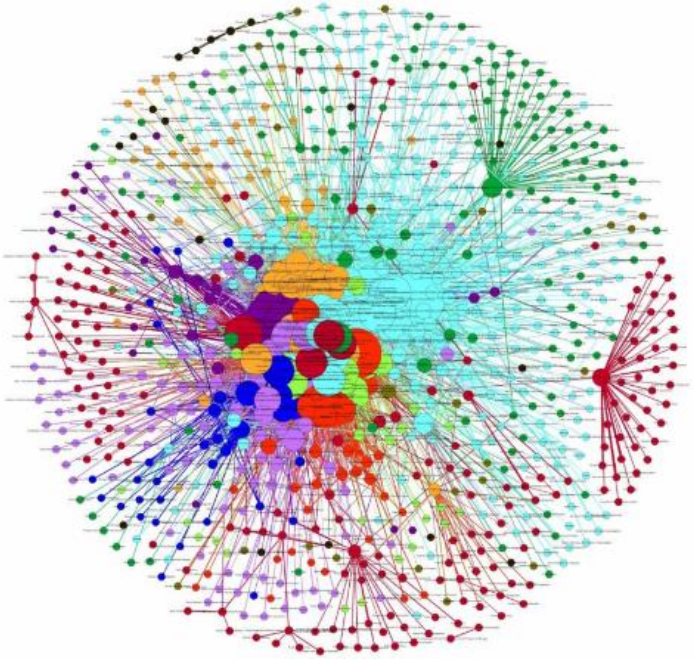
Two things can be true:

1. Better to have very simple descriptive stats derived from excellent data than a highly complex model based on terrible data
2. Multivariate/more nuanced models also help guide you away from oversimplistic correlations



Statistics & modeling

- Descriptive statistics
  - *X% of patients who took Drug Y experienced Side Effect Z.*
    - This can be "sliced and diced" many different ways
- Multivariate analysis
  - *Taking Drug Y increased the odds of Side Effect Z occurring by X% when controlling for A, B, and C demographics and comorbidities.*
  - Getting away from said "slicing and dicing" for a more holistic picture: controlling for other potential influences (while also being aware of and addressing how those other potential influences may influence each other)



Statistics & modeling

- Choosing the best model will depend on what your dependent variable is, your study's relationship to time, your comparison group (to name a few)
  - E.g., dependent variable is a...
    - Binary outcome (did this event occur, yes or no)?
      - Repeatable event?
    - Counts outcome (how many events occurred)?
    - Linear outcome (changes in a linear lab measurement)?
    - Scaled outcome (low, medium, high groups)?
    - *Etc.*

Statistics & modeling

- Relationship to time; e.g.:
  - Measuring the same individuals repeatedly over time?
  - One point in time?
  - Is time to event important?
- Comparators via...
  - Randomised control trial?
  - Propensity score matching?
  - Same cohort, pre/post?

Statistics & modeling

- Predictive analytics:
  - *These patients have the highest risk of experiencing Side Effect Z in the next 30 days.*

Statistics & modeling

- Statistical significance, p-values, and effect size/explanatory power
  - What is meaningful, what is actionable, what is publishable...these may not always coincide
  - What is the risk of being wrong? (patient harms, costs, etc.) - ethical and operational questions
- Post-estimation tests and model assumptions
- There is no perfect model; always tradeoffs
  - Pursuing due diligence and transparency
  - "Now you can see what a fraud I am!"

Communicating results

- Correlation=/=causation
- Communicating caveats – who isn't represented in the data, alternative interpretations, further areas to explore
- Recognise the potential harm of over- or understatement of findings, and of how results may be taken and interpreted
- What new areas this has opened for further exploration
- In other words: *know the limits and be transparent with your audience*

Takeaways

- Data quality is the foundation of everything – don't neglect a strong foundation
  - Better to have simple descriptive stats on excellent data than the most impressive AI model on terrible data
- Quantitative data isn't actually objective – it's the result of a thousand tiny subjective decisions along the way – understand and own those
- Know and understand your data
  - Context of data is as important as the data itself
  - Your responsibility to interrogate its accuracy and if fit for purpose – even if coming to you third party
- Know and understand your results
  - The caveats, the limitations – operate in transparency