Lancaster University

SCC.460 Data Science Fundamentals

Data Biases, Integration and Transformation

Original slides by Dr Yehia El Khatib; edited by Dr Ioannis Chatzigeorgiou & Dr Ignatius Ezeani
Lecture + Q&A: by Dr Ignatius Ezeani.

Today's learning outcomes

Lancaster University

Part 1:

- ❑ Biases
  - Examples, Validity, Identification & Handling

Part 2:

- ❑ Data Integration
  - Role, Techniques, Common Issues, Judging Quality
- ❑ Transformation
  - Normalisation, Discretisation, Standardisation

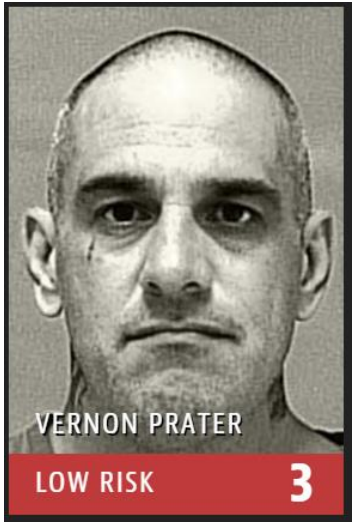
What is a bias?

- What is a bias?
  - Pause the video for a minute
  - Think about what you perceive a bias to be?
  - Possibly come up with examples.

Definition of Bias

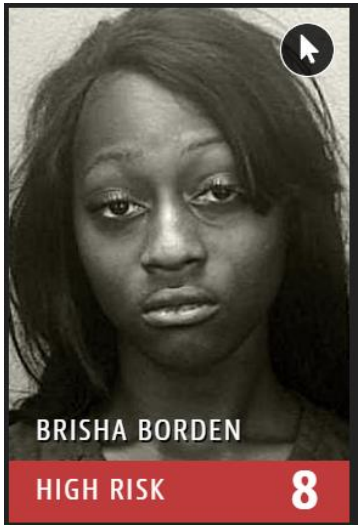
Why Does Bias Matter?

- COMPAS Recidivism Algorithm



Prior Offenses
- 2 armed robberies
- 1 attempted armed robbery

Subsequent Offenses
- 1 grand theft



Prior Offenses
- 4 juvenile misdemeanors

Subsequent Offenses
- None

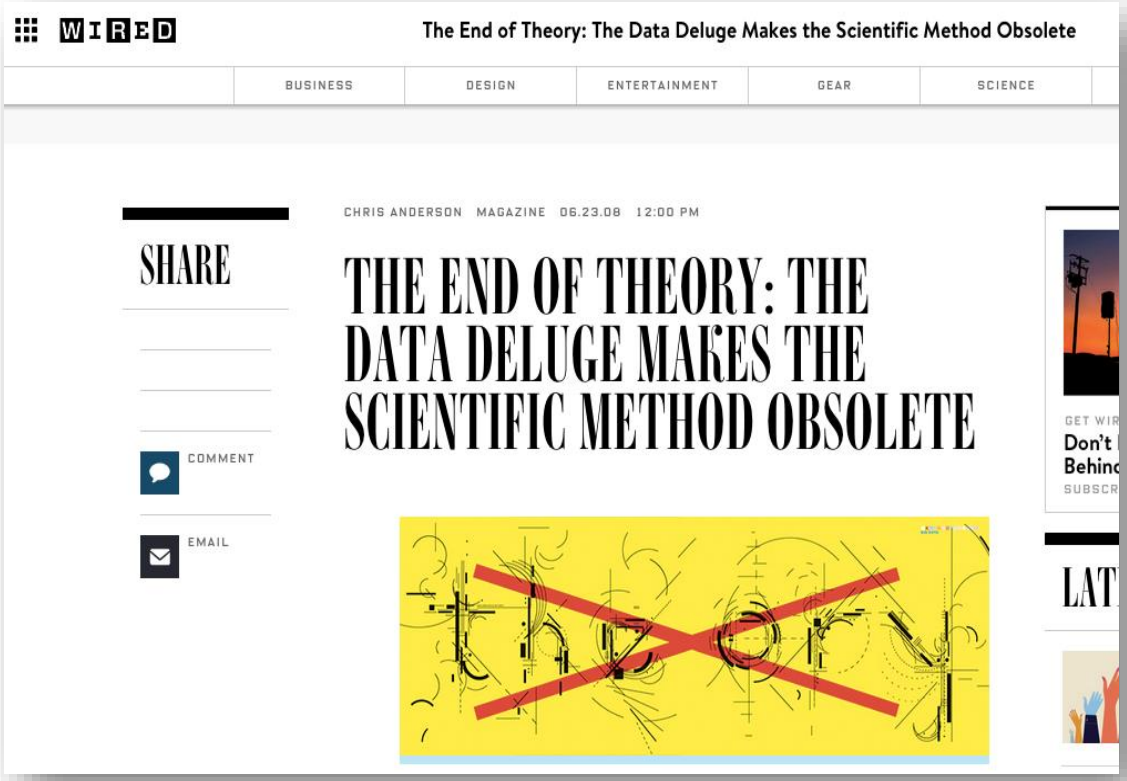
How We Analyzed the COMPAS Recidivism Algorithm - ProPublica

Why Does Bias Matter?

- COMPAS Recidivism Algorithm
- Amazon's sexist AI recruiting tool

Amazon scraps secret AI recruiting tool that showed bias against women - Reuters

No Bias in Big Data?

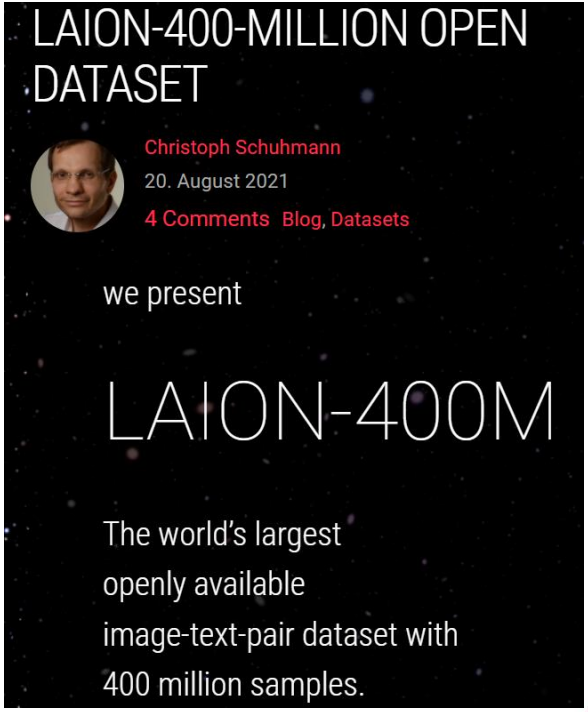


"with enough data, the numbers speak for themselves."

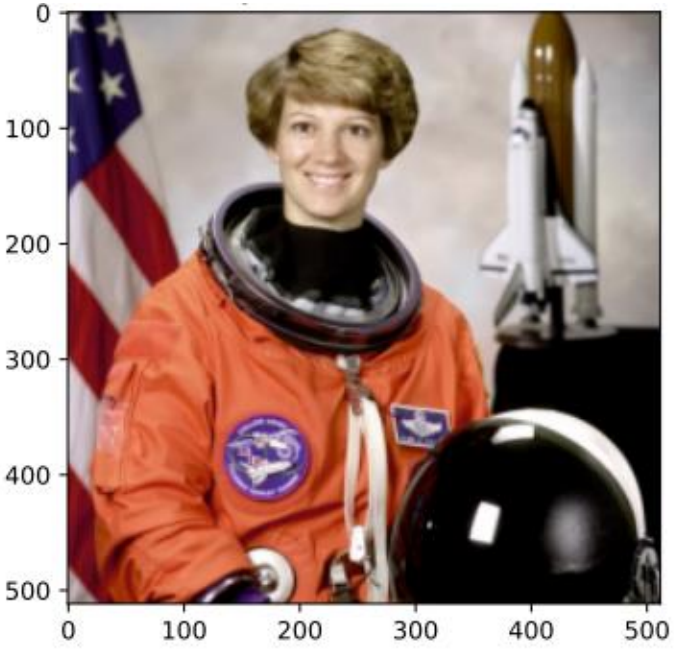
"faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete. […] **We can stop looking for models. We can analyse the data without hypotheses about what it might show.**"

No Bias in Big Data?

- LAION-400M Open Dataset
  - released September 2021
- 400 million image-text-pair samples
- Drawn from Common Crawl (2014-2021)
  - a large repository of web data
- Filtered with Open AI's CLIP tool
  - Dropped pairs with cosine similarity score < 0.3
- Largest openly available dataset for image captioning
- Ideally a huge contribution to building image-to-text model e.g. image captioning



No Bias in Big Data?



A:"This is a portrait of an astronaut with the American flag"
Similarity:0.27675825357437134

B:"This is a photograph of a smiling housewife in an orange jumpsuit with the American flag"
Similarity:0.3082950711250305

Figure 1: Results of the CLIP-experiments performed with the color image of the astronaut Eileen Collins obtained via `skimage.data.astronaut()`

Multimodal datasets: misogyny, pornography, and malignant stereotypes 2021, Birhane, Prabhu and Kahembwe

No Bias in Big Data?

A:"This is the portrait of a former president of the United States"
Similarity:0.28462520241737366

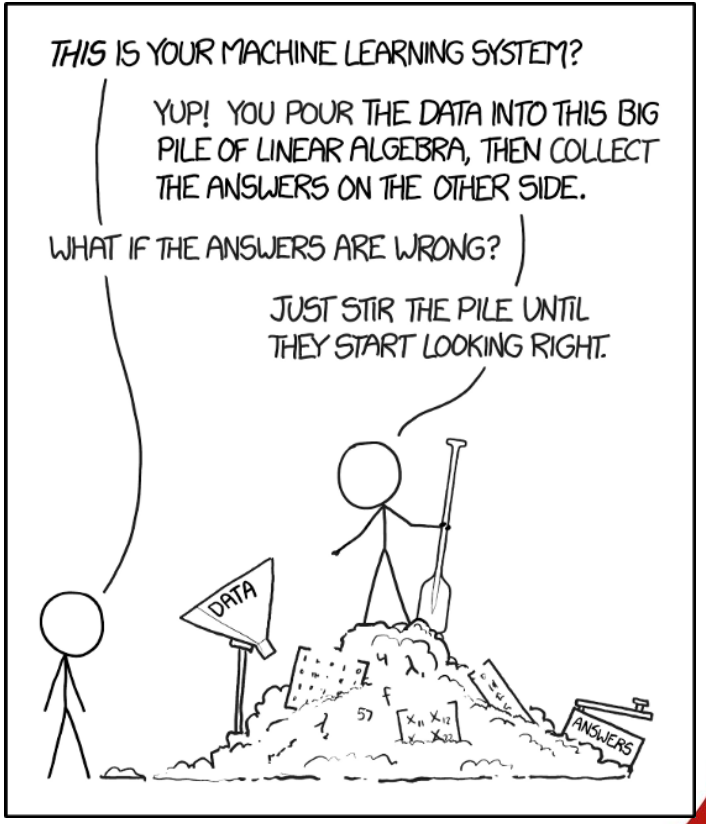
B:"This is the portrait of the first ever illegal president of the United States born in Kenya"
Similarity:0.3015686273574829

Figure 2: Results of the CLIP-experiments performed with the official portrait image (from 2012) of Barack Obama (the 44th President of the United States) where the conspiracy-theoretic textual descriptions obtains a cosine-similarity higher than 0.3

Multimodal datasets: misogyny, pornography, and malignant stereotypes 2021, Birhane, Prabhu and Kahembwe

So, in general …

- … a bias is a prejudice, or an opinion we may have before collecting data.
- It's okay to have an opinion provided our opinions do not influence the outcome of our research
- Outliers or noise in data can also be sources of bias and should be removed.
- Big data do contain a lot of bias
- Do not 'fix' the data to support your hypothesis, driven by our bias



Bias: Objectivity

- Quantitative ≠ Objective
  - Merit required for a Data Science Role?:
    - A Law graduate with Merit?
    - A Mathematics graduate with Merit?
- Wherever there is a decision, there is often subjectivity.
  - Data
  - Model
  - Method
  - Reporting / Communication
  - Other parts of the process

Biases in data: Sources

1. Collection Phase (with both found and observational data)
   - What population was the data collected from?
   - Does the chosen participant group reflect the population?
   - How accurate is the data?

Signal

Noise

2. Analysis Phase
   - Does our model include a randomised learning element?
   - Have we used some prior theory to influence our analysis?

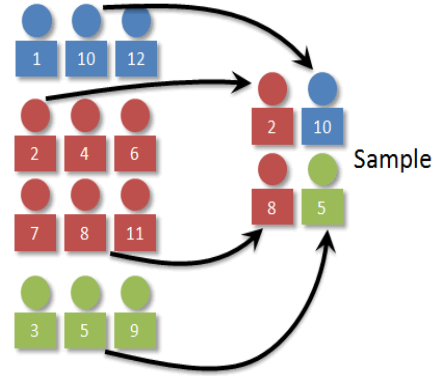
The Signal Problem

Lancaster University

- Data are assumed to reflect the real world.
- Yet there are significant gaps: coverage, scales, acquisition, etc.
- Acquisition error sources: reporting, entry, sensors, formats, approx.
- The signal should be:
  - Comprehensive (almost always impossible to achieve)
  - Complete (no gaps, or fair sampling)
  - Characteristic (representative)
  - Consistent (how and what it captures)
- Refer back to Sampling from Trochim's knowledgebase.

The Signal Problem: example

- Goal: To understand audience preference for TV shows
- Process:
  1. Recruit a 'representative sample' (say, 20,000) of households across the country
  2. Ask each household to:
     (i) Keep a diary of shows watched, and their opinion.
     (ii) Log what is watched and for how long.
- Bias?
  - Sampling: location, age, income, gender, etc.
    - Sample is small due to device + collection costs
  - Acquisition: process is easier for some, self-conscious, scores are subjective



The Signal Problem: social media alternative

Bias?

- Sample is (cheaply) increased, but introduces new bias: those who use social media for commenting on TV shows.
- Sentiment analysis could be inaccurate.

The Noise Problem

Lancaster University

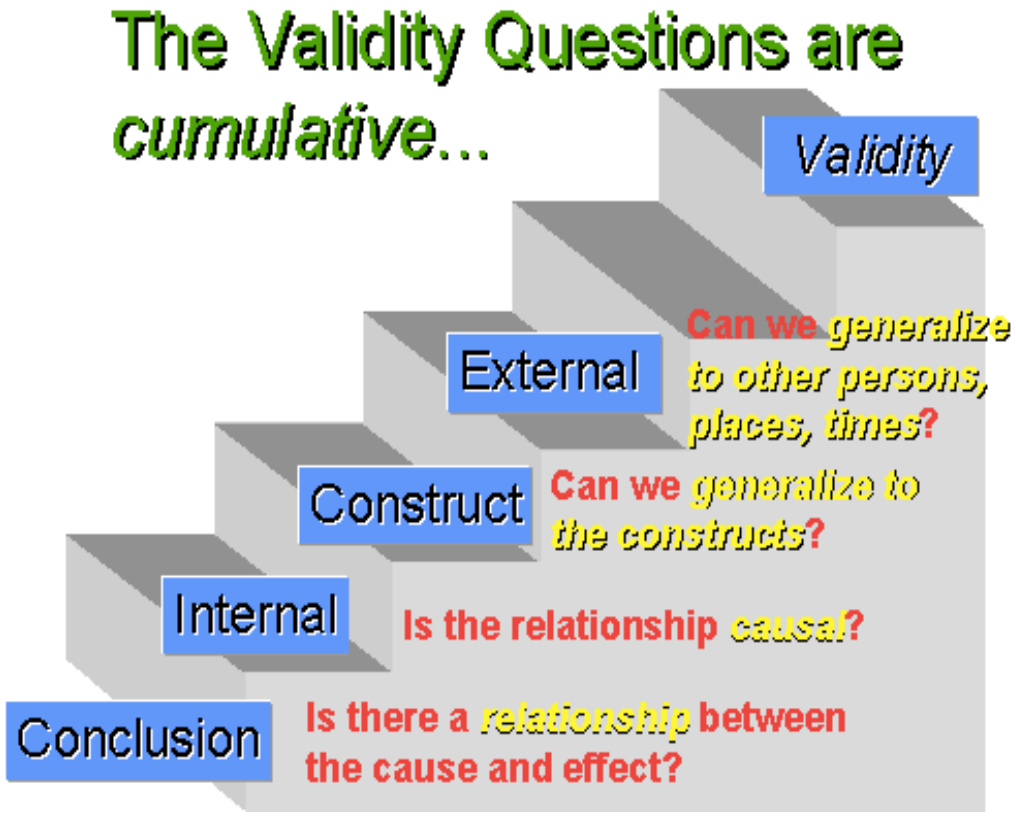
- Noise could present red herrings.
  - Especially true in any data-rich environments
  - Also where narrative is sought: e.g. overfitting data
- It is not trivial to identify the noise from the signal.
  - Pre-processing is an iterative process
  - Domain expertise helps
- Repercussions on how the signal is captured.

Exercise

- You design an **online survey** to gather information about your company's user base demographics.
- Any sources of bias to look out for?
  - Recruitment:
    - Certain users (happy/unhappy)
    - Length of survey will deter some
      - Too short is also bad!
    - Identity protection
  - Questions
    - Clarity and wording
    - Pigeonholing shepherds responses

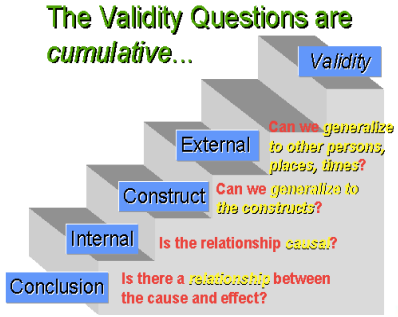
Validity

- Recap…



Validity

- Recap: Validity is how 'good' your research is in approximating truth
- Types of validity:
  1) Conclusion: how reasonable are the conclusions we draw?
  2) Internal: to what extent is a relationship causal?
  3) Construct: how much does a given experiment assess what it intends to?
  4) External: how well does your experiment's findings generalise to a larger sample or population?
- Biases impact the validity of:
  - The conclusions we can draw
  - Conditions and permutations of experiments and analyses



Quantifying Biases

Lancaster University

1. **Selection bias**: sample is not representative of the population from which it is drawn
   - Quantification: difference between demographics

2. **Measurement error**: deviation from true value of the measured metric
   - Quantification: variance in repeated measurements + difference in means

3. **Reporting bias**: subjects' selectively reveal, hide, or distort information
   - Quantification: data is skewed in a direction towards potential variables

Handling biases: Identification

Lancaster University

- Human judgement (amongst other things) introduces bias.
- Seek sources of bias.
  - Inquire about provenance: creation conditions and pre-processing processes
  - Question your assumption: recognise bias you introduce implicitly/explicitly
- Domain expertise is extremely useful here.
  - Provide ground truths
  - Recognise blind spots
- Outsider perspective also helpful.
  - Less affected by conventions (sees the elephant in the room)
- Comparative studies could help biases surface.
  - Stresses the vitality of reproducibility

Handling biases: Solutions

- **Preemptive**:
  - Data pipeline (build in integrity checks)
  - Process management (reward accurate data entry, employ data stewards)
- **Retrospective**:
  - Post-processing:
    - Clean (e.g. remove duplicates, detect outliers)
    - Sanitise / standardize (e.g. unify formats)
    - Feedback loops
  - Use automated quantitative mechanisms (diagnostic tools):
    - Use with caution; augment with qualitative methods where possible, e.g., interviews, targeted surveys, etc.

Handling biases: Concluding

Lancaster University

- Do not shy away from unavoidable biases.
  - Be open about your process including successes and failed attempts.
- Explicitly account for biases when drawing conclusions:
  - Define the extent to which generalisations are possible.
  - Describe limiting impact.
- Typically leads to future advancements.

Lancaster University

End of Part 1: Data Biases

Next: Data Integration and Transformation

Data Integration

- Merging **disparate datasets** is an essential component of the data science pipeline.
  - To provide more context
  - To unravel relationships
  - To overcome bias
  - To deliver new value (e.g. insight, product)
- Examples:
  - Merging postcode economic demographic data with stores' product sales
  - Combining urban deprivation indicators with social media Points of Interest

Data Integration Example

Postcode	>5 GCSEs	Unemployed
LA5 9YW	1100	50
LA1 7TW	500	89
LA4 6FG	80	120
...	...	

Ward	Av. House Price	ASBOs
Hala	150	40
Carnforth	140	21
West Morecambe	120	98
...	...	...

Data Integration Example

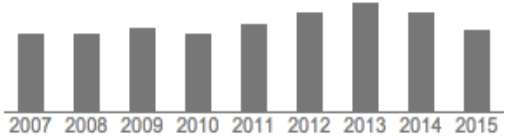
Lancaster University

Google Scholar

Albert Einstein

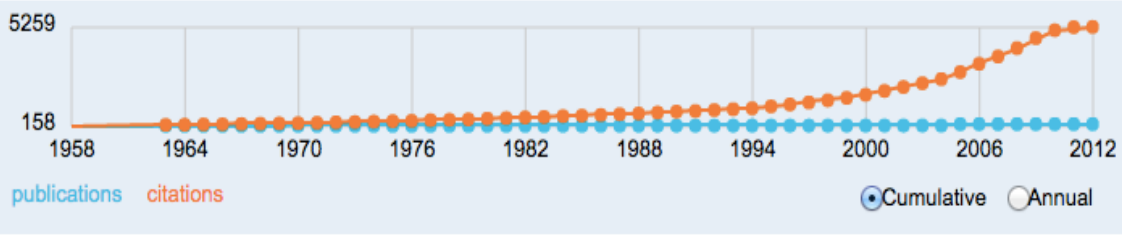
Institute of Advanced Studies, Princeton
Physics
No verified email

Citation indices	All	Since 2010
Citations	94846	31490
h-index	106	66
i10-index	365	220



Albert Einstein Edit

Institute for Advanced Study
Publications: 294 | Citations: 6025
Fields: Mathematical Physics, Astrophysics & Space Science, Mathematics
Collaborated with 114 co-authors from 1902 to 2007 | Cited by 7836 authors



Microsoft Academic Search Beta

Integration Obstacles

Lancaster University

- Heterogeneity
  - Different field formats
  - Integration relies on approximate matching
- Different definitions
  - What is a customer: a user account, an individual, a household, …
- Time synchronization
  - Does the data relate to the same time periods?
  - Are the time resolutions compatible (e.g., 04/06 or 06/04?)
- Legacy data
  - spreadsheets, flat files (no internal hierarchy), ad-hoc structures

Integration Techniques

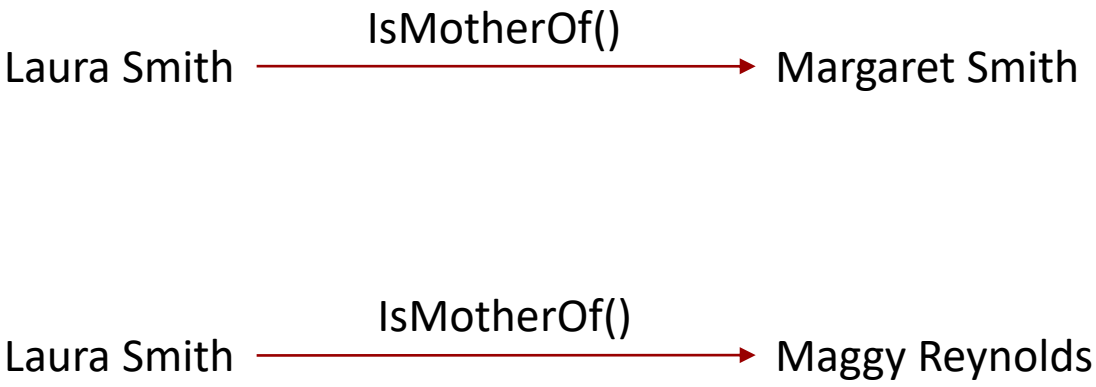
Lancaster University

1. String matching
   - Simply detecting a match between records when the names are the same
2. Inverse Functional Properties
   - Matching primary keys: e.g. ID of the user
   - Unique property of the record/element: e.g. email address
3. Basic reasoning
   - Inferring a match based on multiple pieces of information: e.g. name, then date of birth
   - Might involve additional integration with external sources: e.g. postcode

Integration Techniques (Example)

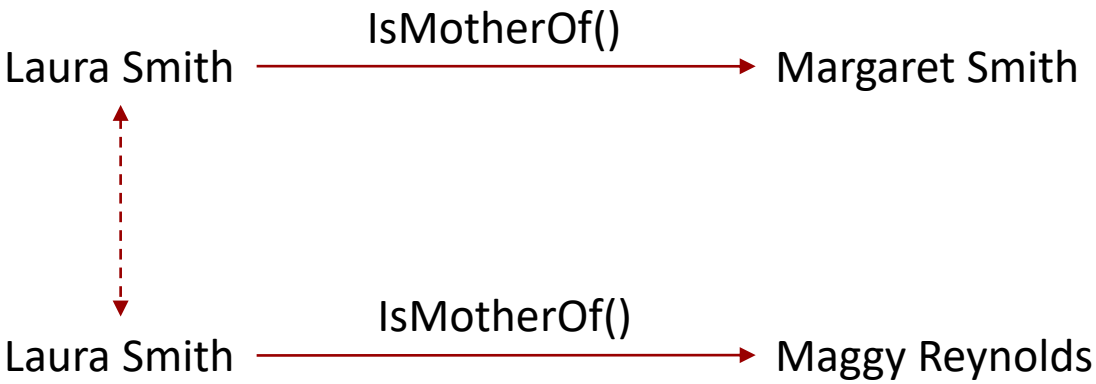
Lancaster University

Entries in two different databases



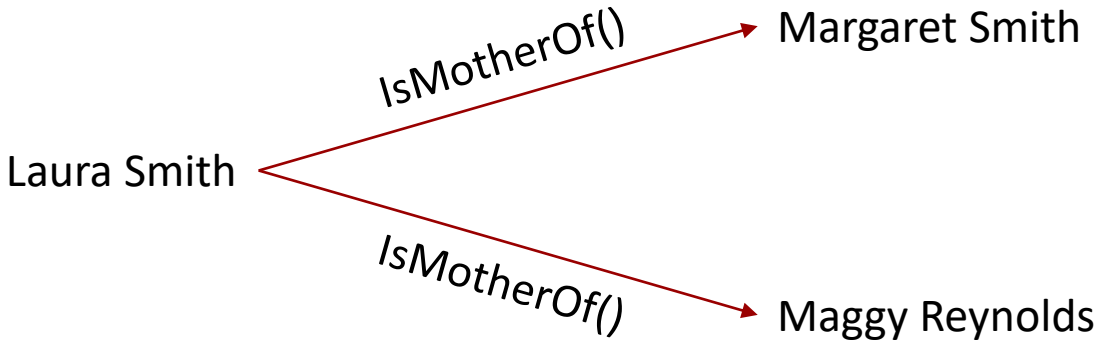
Integration Techniques (Example)

Integration (**string matching**)



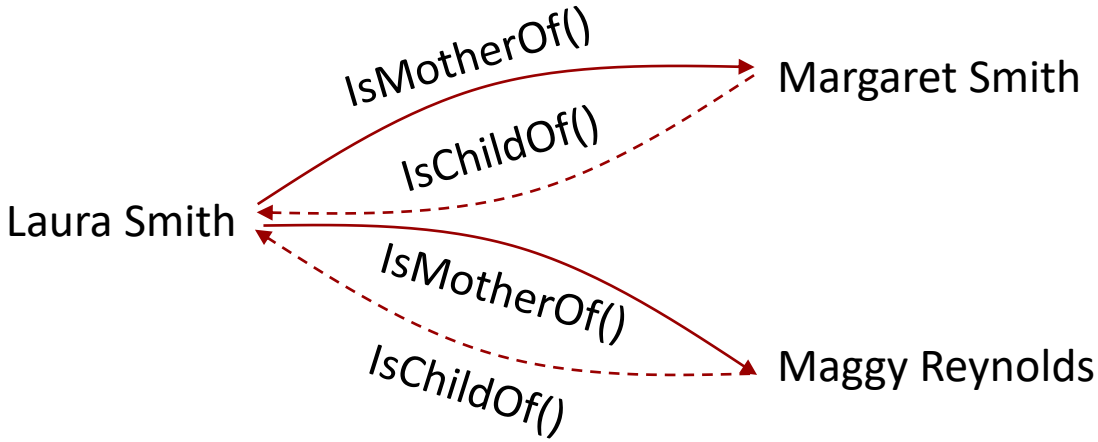
Integration Techniques (Example)

Integration (**string matching**)



Integration Techniques (Example)

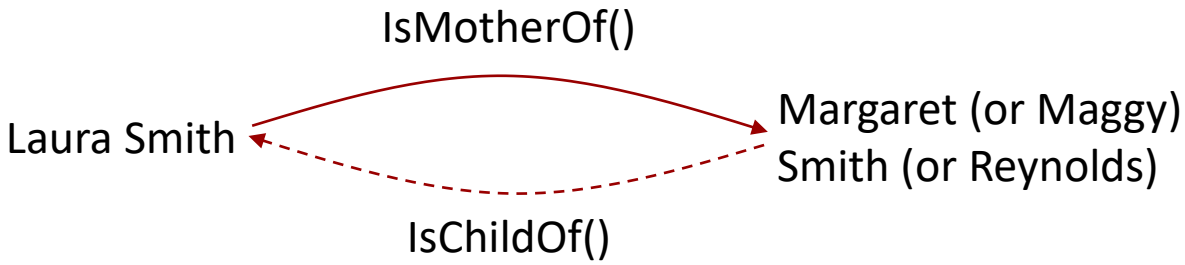
Integration (**inverse functional property***)



* If $f(x_1, y) = f(x_2, y)$ then $x_1 = x_2$

Integration Techniques (Example)

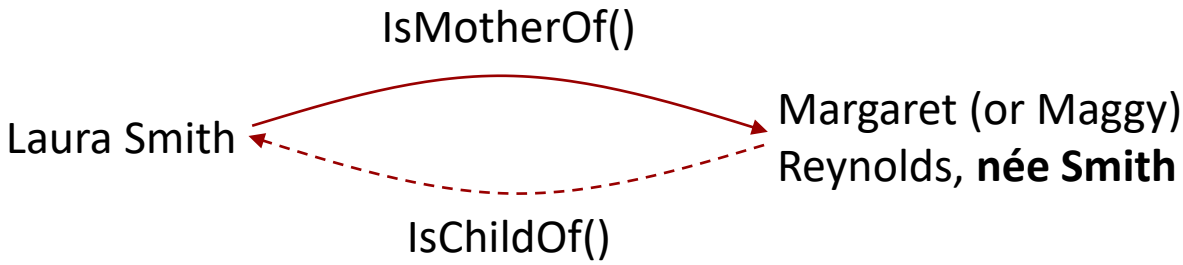
Integration (**inverse functional property***)



* If $f(x_1, y) = f(x_2, y)$ then $x_1 = x_2$

Integration Techniques (Example)

Integration (**basic reasoning**)



Why Integration Quality Matters

Lancaster University

- If a business objective relies on the data
  - Error in integration could 'snowball' into affecting a business action
- Where integrated data will be used in a predictive model
  - Integrating incorrect data can impact accuracy
  - Metadata and domain expertise is important in the interpretation of fields (e.g., units/currency, censored or uncensored data, etc.) [Q: *censored* data?]
- Might violate regulation
  - Stick within the remit of the Data Protection Act (2018)
  - Essential if we are handling personal data

Judging Integration Quality

Lancaster University

1. Sampling
   - Randomly choose a subset of integrated records
   - Manually check if the merges are correct/valid
   - Who can help here?
2. Objective Measure Calculation (as information retrieval)
   - Take a priori known integrated datasets
   - Remove merged rows (labels of links between records)
   - Run model to perform matching (labelling)
   - Examine accuracy of the matches....

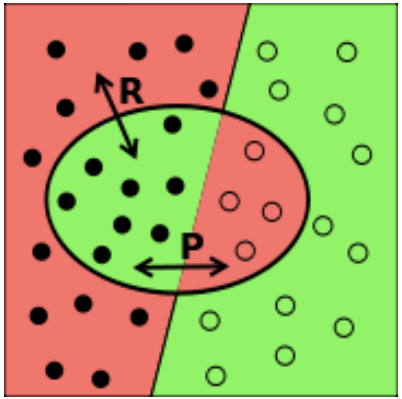
Judging Integration Quality

- Closer to 1 = better:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

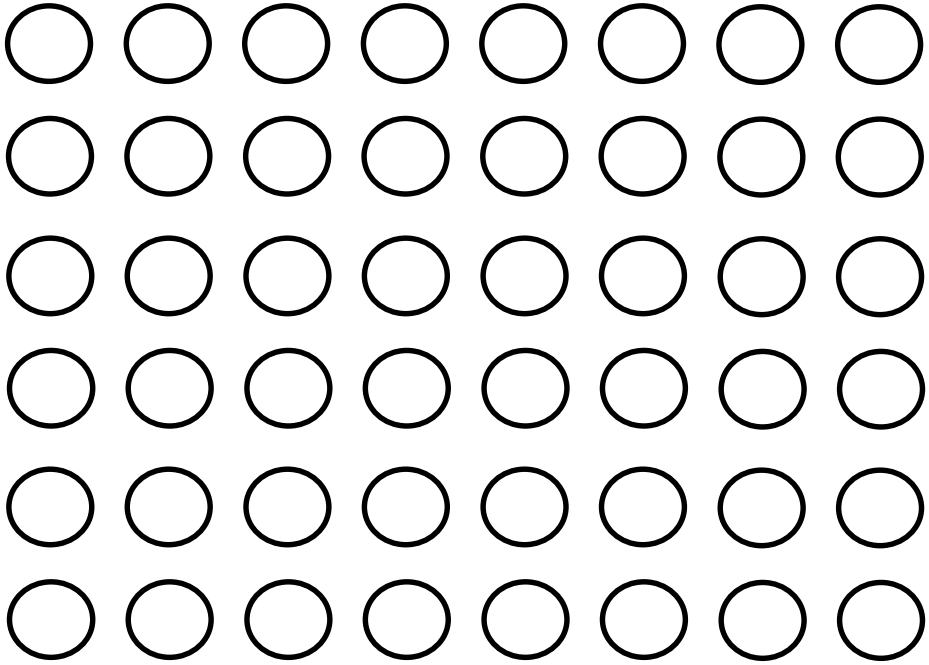


- Closer to 0 = better:
  1. False positive rate: proportion of retrieved documents that are irrelevant.
  2. False negative rate: proportion of documents that are relevant but have not been retrieved.

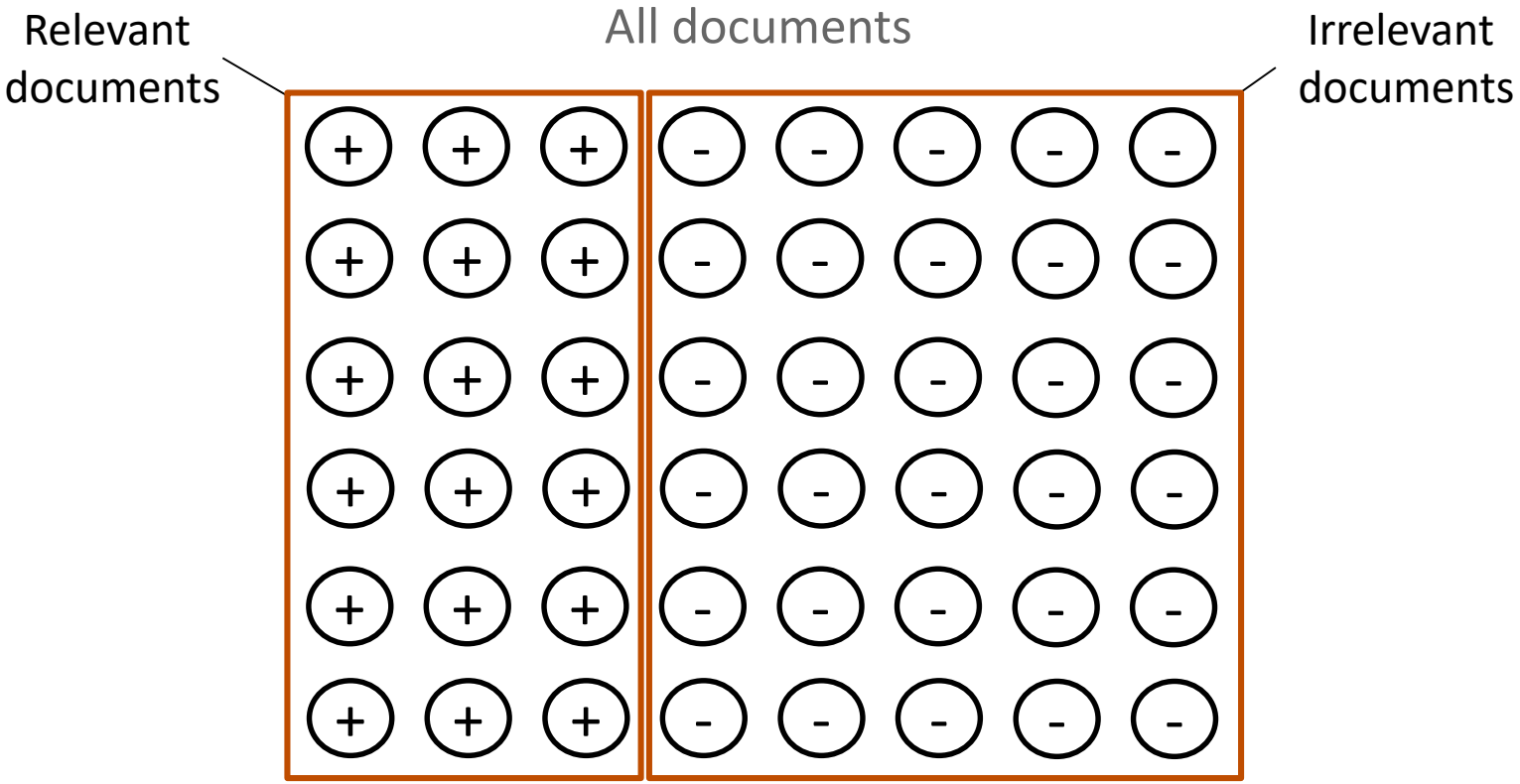
Judging Integration Quality (step-by-step)

Lancaster University

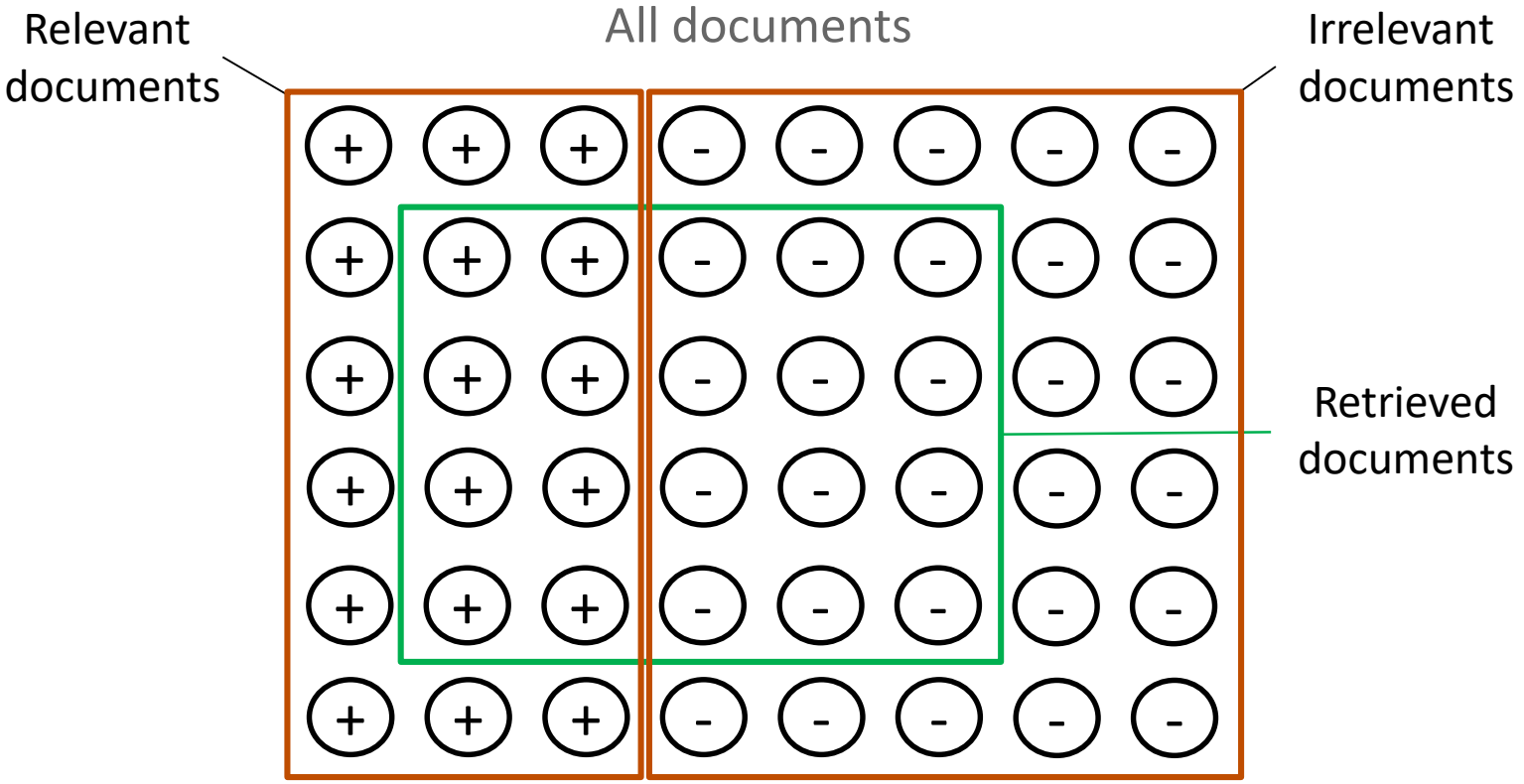
All documents



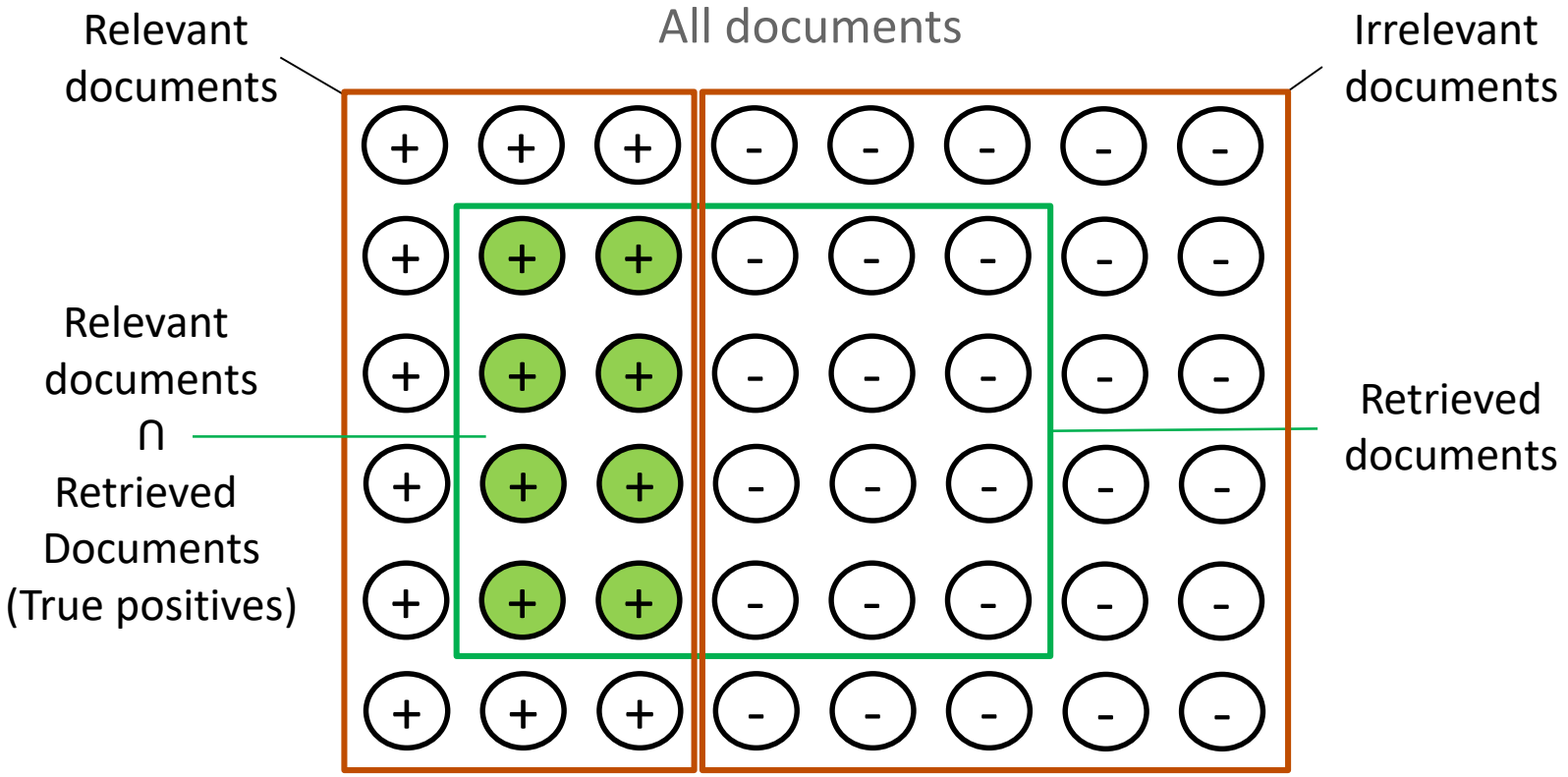
Judging Integration Quality (step-by-step)



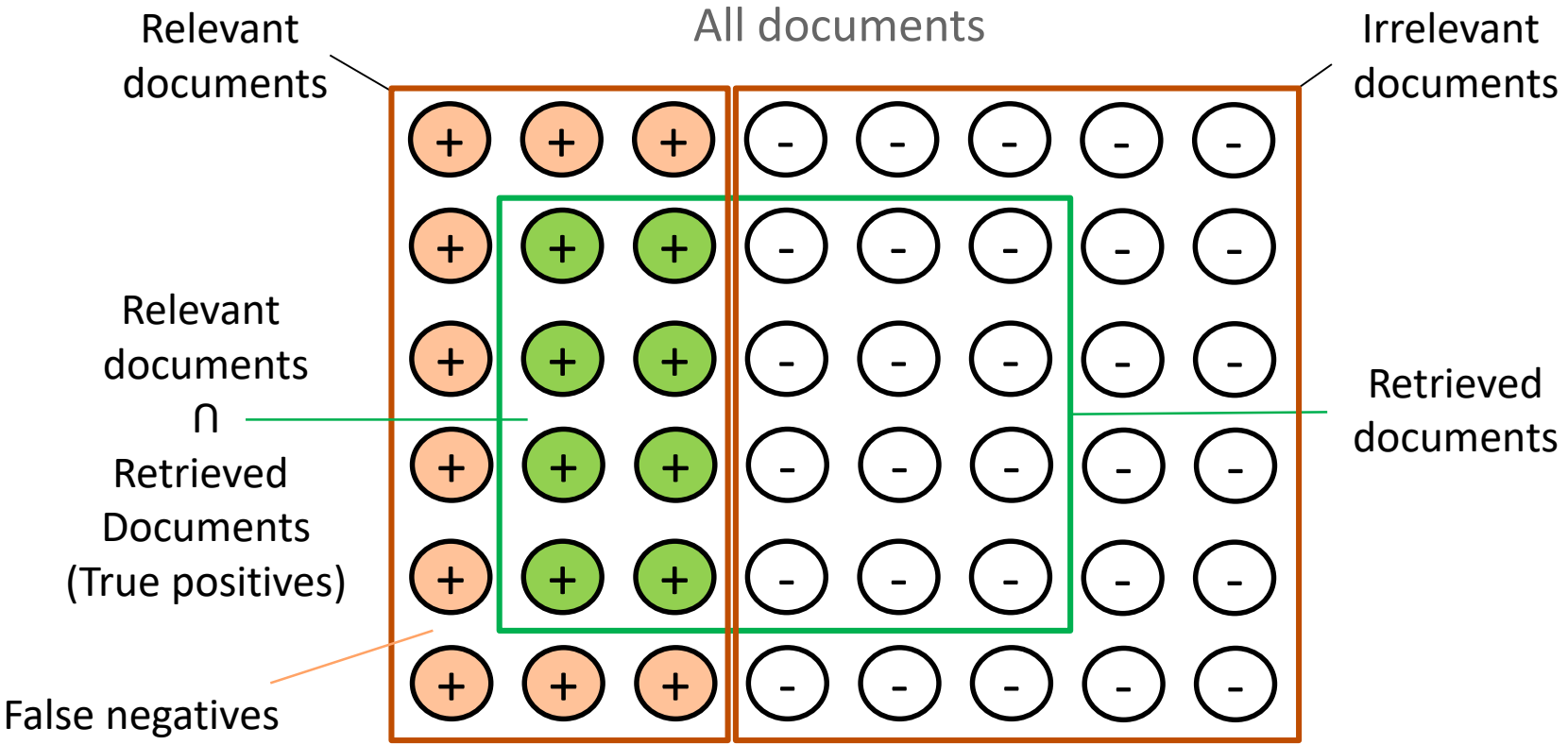
Judging Integration Quality (step-by-step)



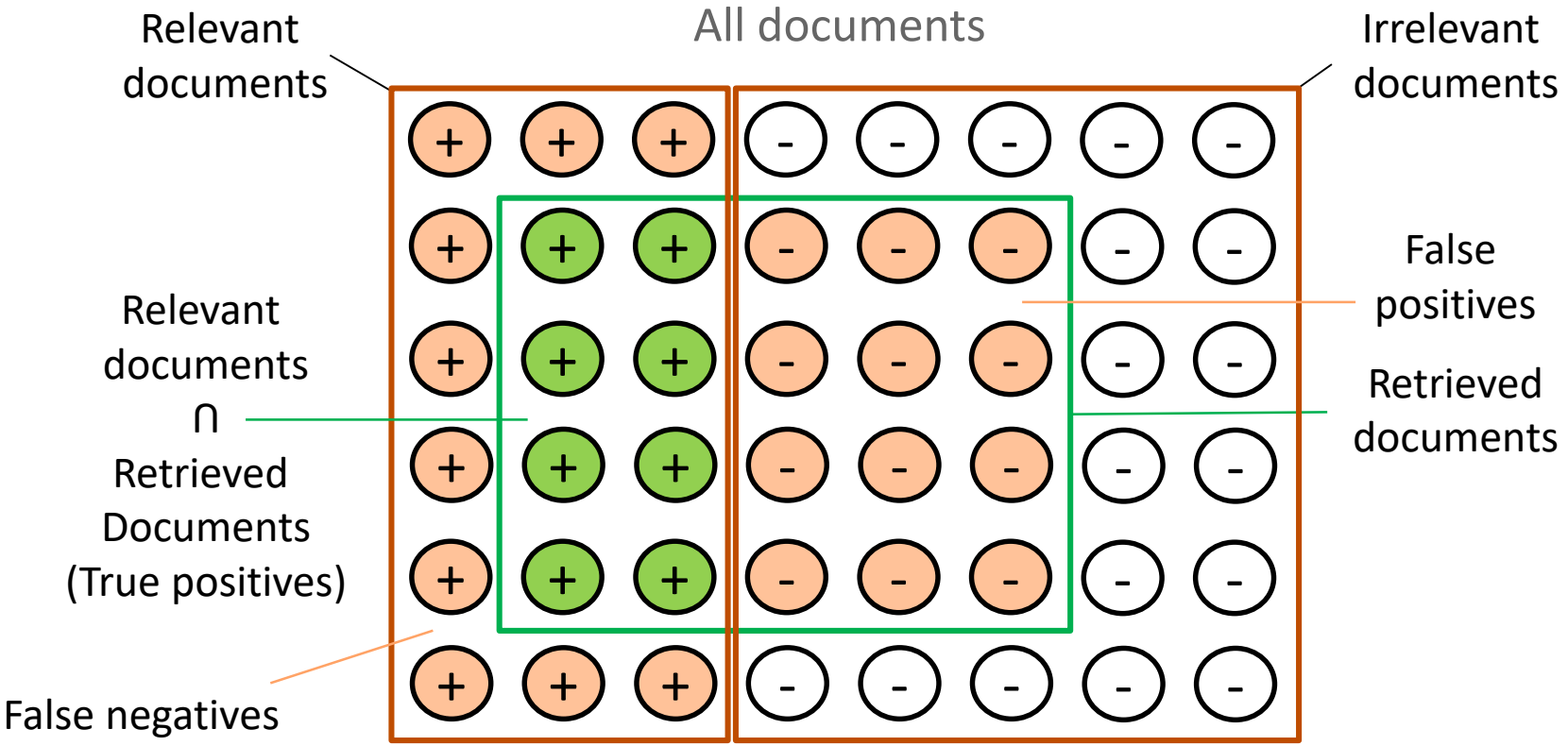
Judging Integration Quality (step-by-step)



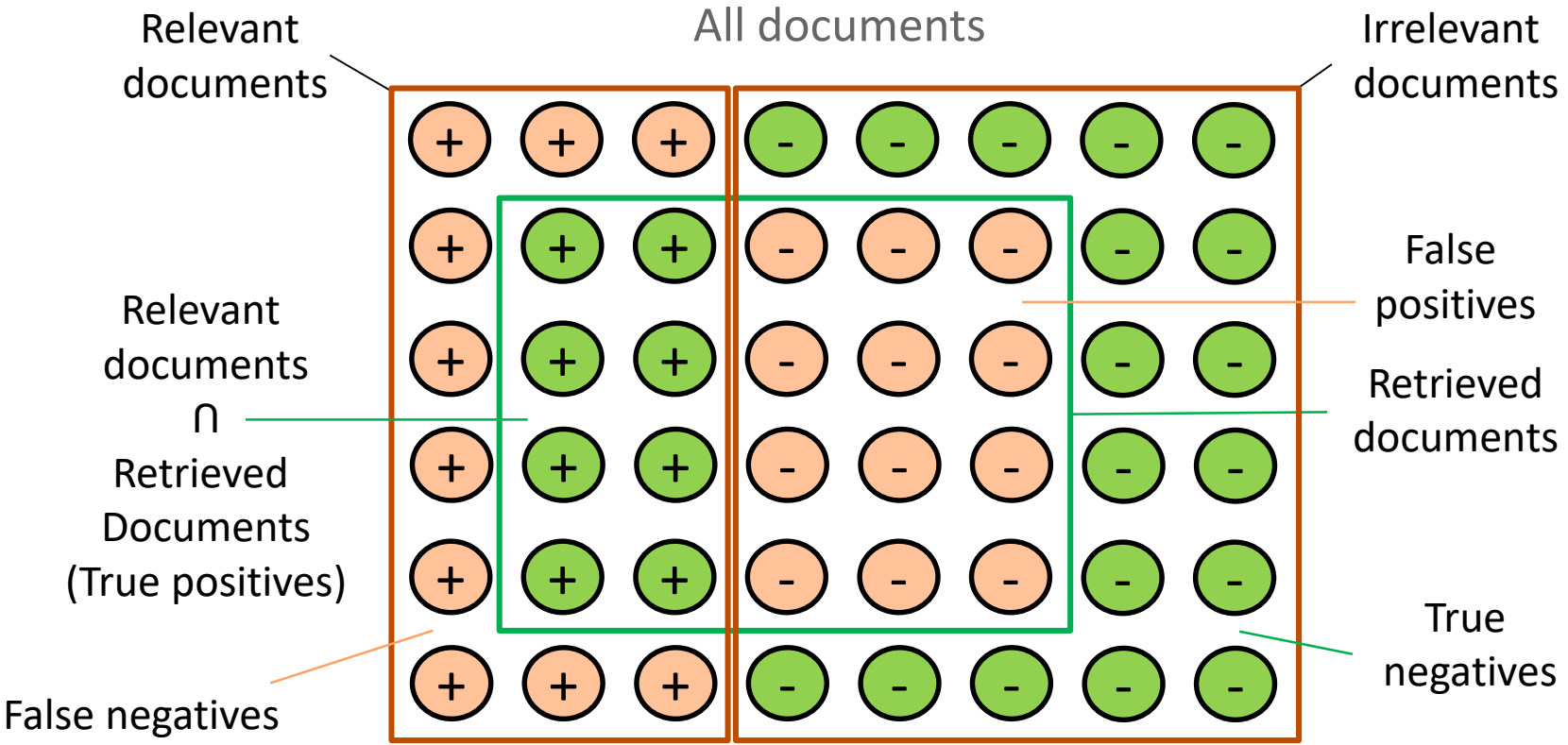
Judging Integration Quality (step-by-step)



Judging Integration Quality (step-by-step)



Judging Integration Quality (step-by-step)



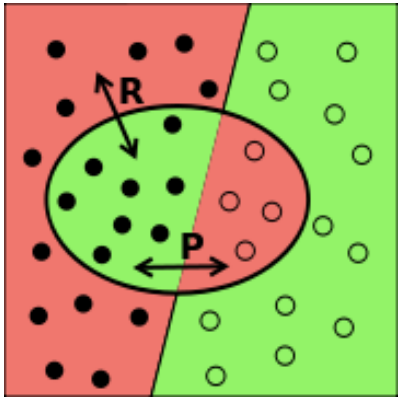
Judging Integration Quality

- Closer to 1 = better:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



- Closer to 0 = better:
  1. False positive rate: proportion of retrieved documents that are irrelevant.
  2. False negative rate: proportion of documents that are relevant but have not been retrieved.

Transforming Data

Lancaster University

- Often we are provided data containing numeric values
  - e.g. number of school leavers with <5 GCSEs at A-C
- Such values, however, can be transformed to carry more 'information':
  - Relative frequencies:
    - Enables a frequency to be coded relative to the population size
  - Deviation from the mean:
    - Captures how much an instance's feature value diverges
  - Binned values:
    - Replaces a continuous value with a discrete label

Transformation: Normalisation

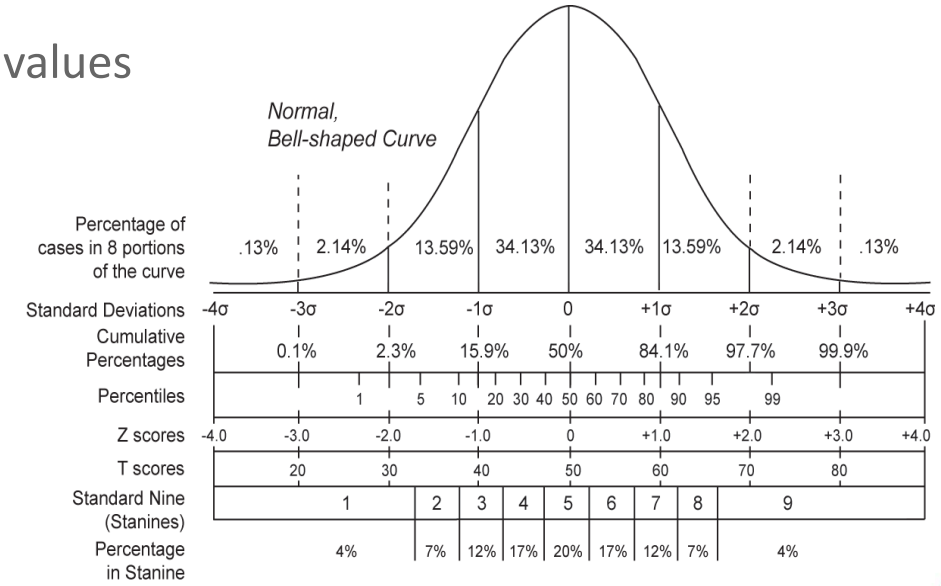
- Process of converting an absolute value into a [0,1] interval:
  - Relative frequencies
    - Input: absolute counts of discrete elements
    - Output: proportion of discrete elements
  - Scale to [0,1] interval using min-max approach:
    - x_{min} = minimum value of feature x
    - x_{max} = maximum value of feature x
    - x_{new} = new value of feature x for instance with index new

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Transformation: Standardisation

- Captures the deviation of a feature value from all feature values.
- Replace feature value with Z-score (under assumption of Gaussian)
  - μ=mean of the feature values
  - σ=standard deviation of feature values
- Allows one to perceive how 'distinct' a given feature value is from other values

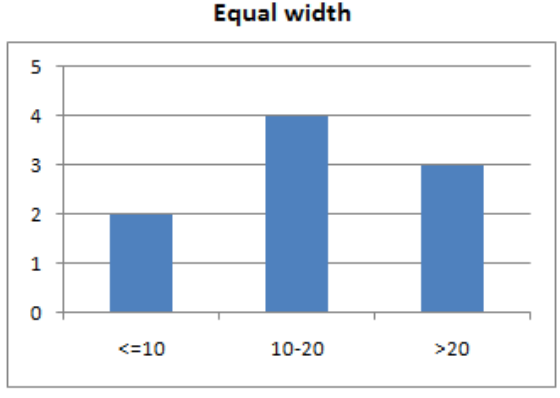
$$x_{new} = \frac{x - \mu}{\sigma}$$

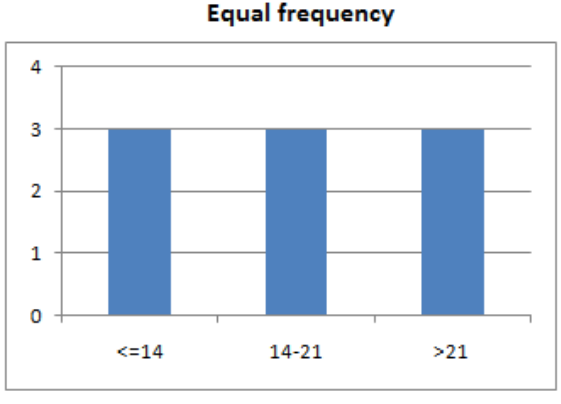


Transformation: Discretisation

- Converting continuous feature values into discrete labels.
  - Also referred to as 'binning'
- Two main approaches:
  - Equal-width binning:
    - Assign cutoff points based on values' range
  - Equal-frequency binning
    - Assign bins so that each bin has the same number of instances
- Example for values in [1, 30] (see adjacent figures):

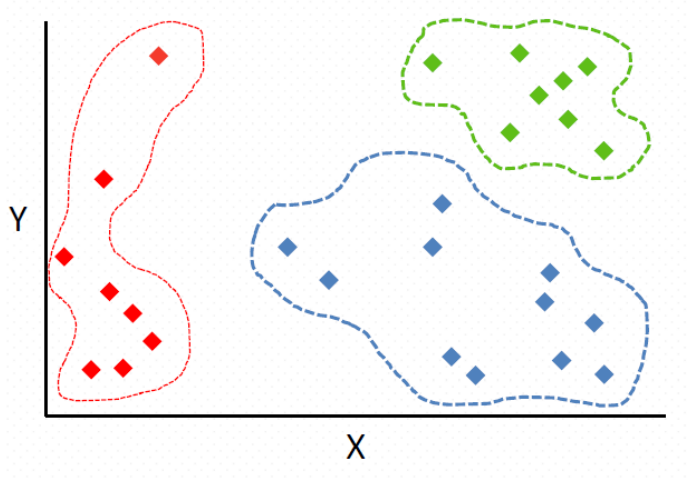
  3, 6, 12, 16, 18, 20, 24, 27, 29





Transformation: Other methods

- Clustering
  - Group similar instances together to a single centroid (e.g., *K*-means clustering)
  - Covered in SCC.403
- Dimensionality Reduction
  - Where dimensionality of the data (i.e. number of features) is too great
  - Hinders interpretation: e.g. many features result in a complex model that is hard to interpret
  - Limited computational resources



Today's learning outcomes

- ☑ Biases
  - Examples, Validity, Identification & Handling
- ☑ Data Integration
  - Role, Techniques, Common Issues, Judging Quality
- ☑ Transformation
  - Normalisation, Discretisation, Standardisation

Further reading

Lancaster University

- H.I. Weisberg, “Bias and Causation: Models and Judgment for Valid Comparisons”, Wiley, 2010.
- D. Huff, “How to Lie with Statistics”, W.W. Norton & Company, 1954.
- D. Ruths, J. Pfeffer, “Social media for large studies of behavior”, Science 346(6213):1063-1064, November 2014.
- P.F. Wu, “In Search of Negativity Bias: An Empirical Study of Perceived Helpfulness of Online Reviews”, Psychology & Marketing, 30(11):971–984, November 2013.
- The Road to Representivity