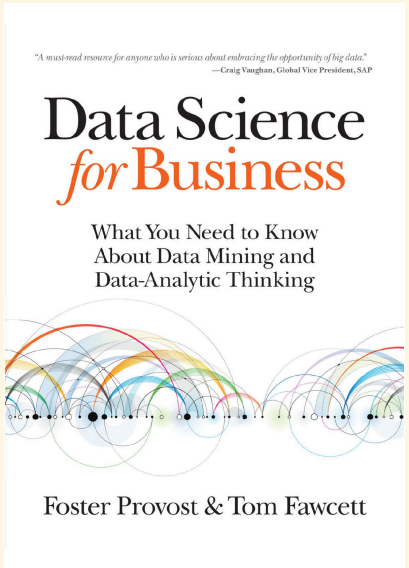
DATA SCIENCE AS 9 PROBLEMS

AP

A GENTLE-YET-FOCUSSED INTRODUCTION

Chapter 2 describes the core computational problems of Data Science



In memoriam: Tom Fawcett

1. Classification and class probability

Instance:

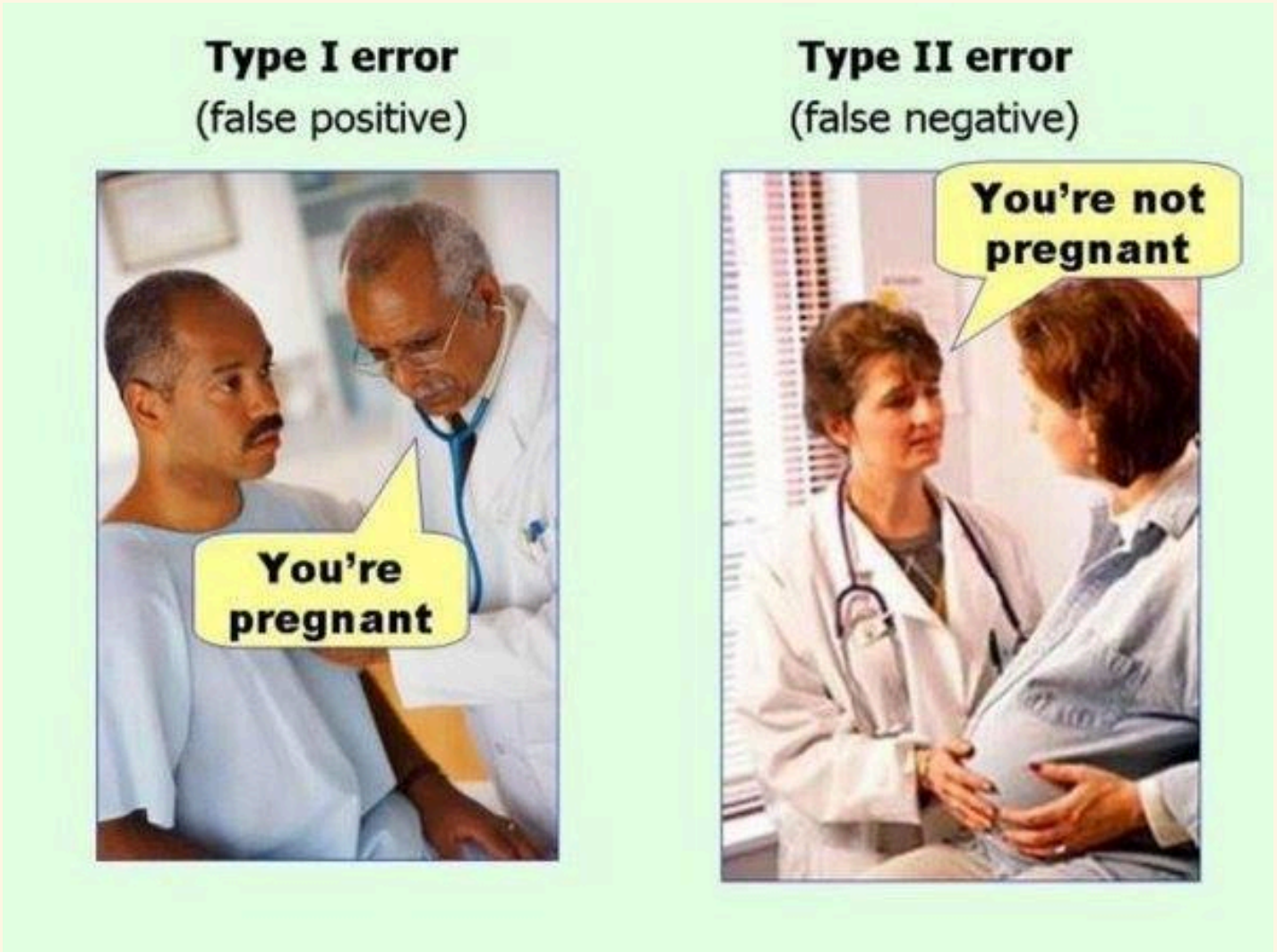
- a collection (dataset) of datapoints from $\mathbf{X}$
- a classification system $C = \{c_1, c_2, \ldots c_k\}$

Solution: classification function $\gamma : \mathbf{X} \to C$

Measure: misclassification

MISCLASSIFICATION: DETECTION

binary, only one class is important



MISCLASSIFICATION: MULTI-CLASS

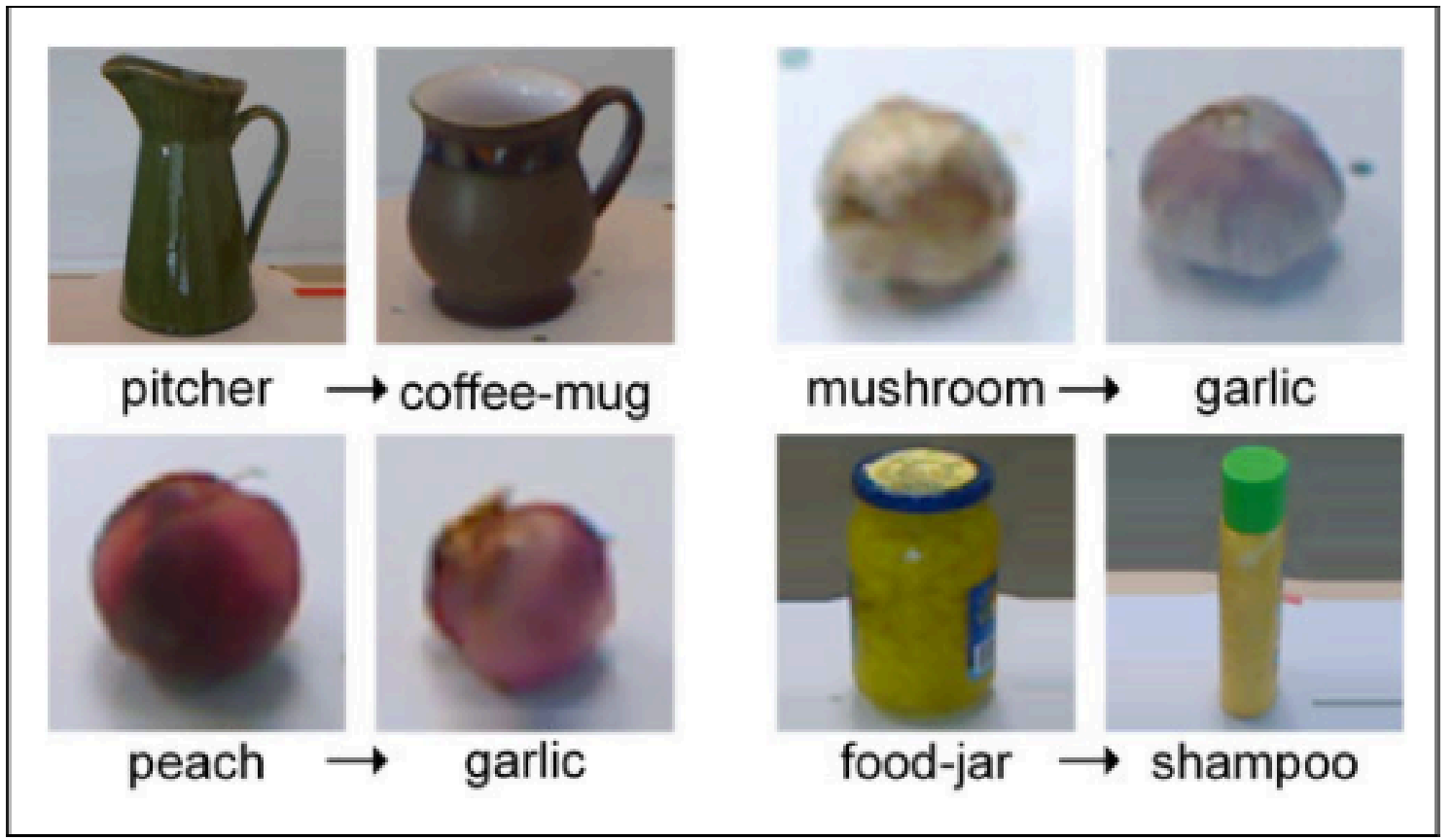
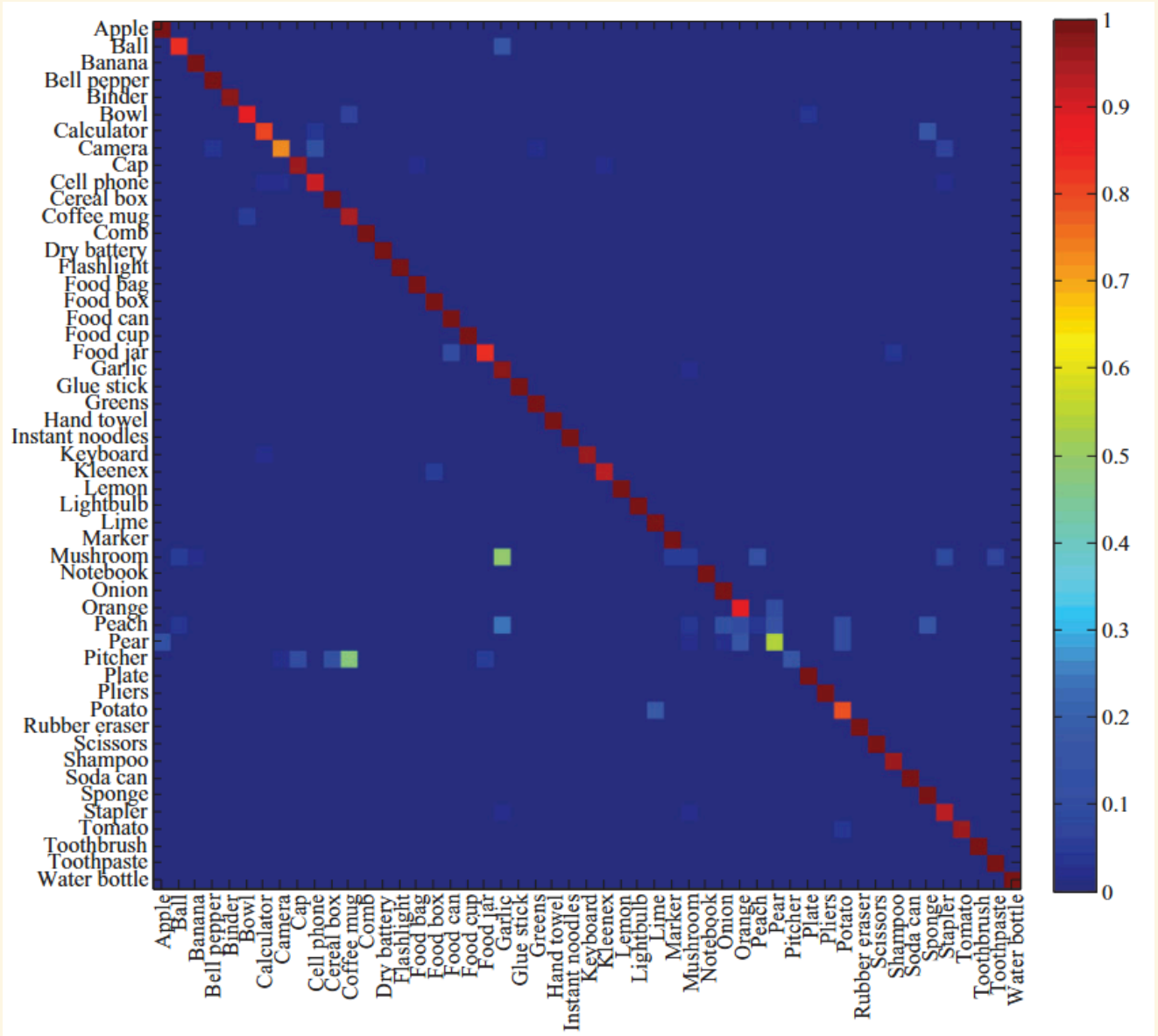


Figure 8. Examples of some easily misclassified categories. Misclassification occurs due to the strong similarities in the objects' colours and shapes.


Apple
Ball
Banana
Bell pepper
Binder
Bowl
Calculator
Camera
Cap
Cell phone
Cereal box
Coffee mug
Comb
Dry battery
Flashlight
Food bag
Food box
Food can
Food cup
Food jar
Garlic
Glue stick
Greens
Hand towel
Instant noodles
Keyboard
Kleenex
Lemon
Lightbulb
Lime
Marker
Mushroom
Notebook
Onion
Orange
Peach
Pear
Pitcher
Plate
Pliers
Potato
Rubber eraser
Scissors
Shampoo
Soda can
Sponge
Stapler
Tomato
Toothbrush
Toothpaste
Water bottle
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Source

2. Regression/value estimation

Instance:

- a collection (dataset) of numerical $< \mathbf{x}, y >$ datapoints
- a regressor (independent) value $\mathbf{x}$

Solution: a regressand (dependent) value y

that complements $\mathbf{x}$

Measure: error over the collection

[PF] “classification predicts whether something will happen, whereas regr. predicts how much something will happen.”

3. Similarity

Identify similar individuals based on data known about them.

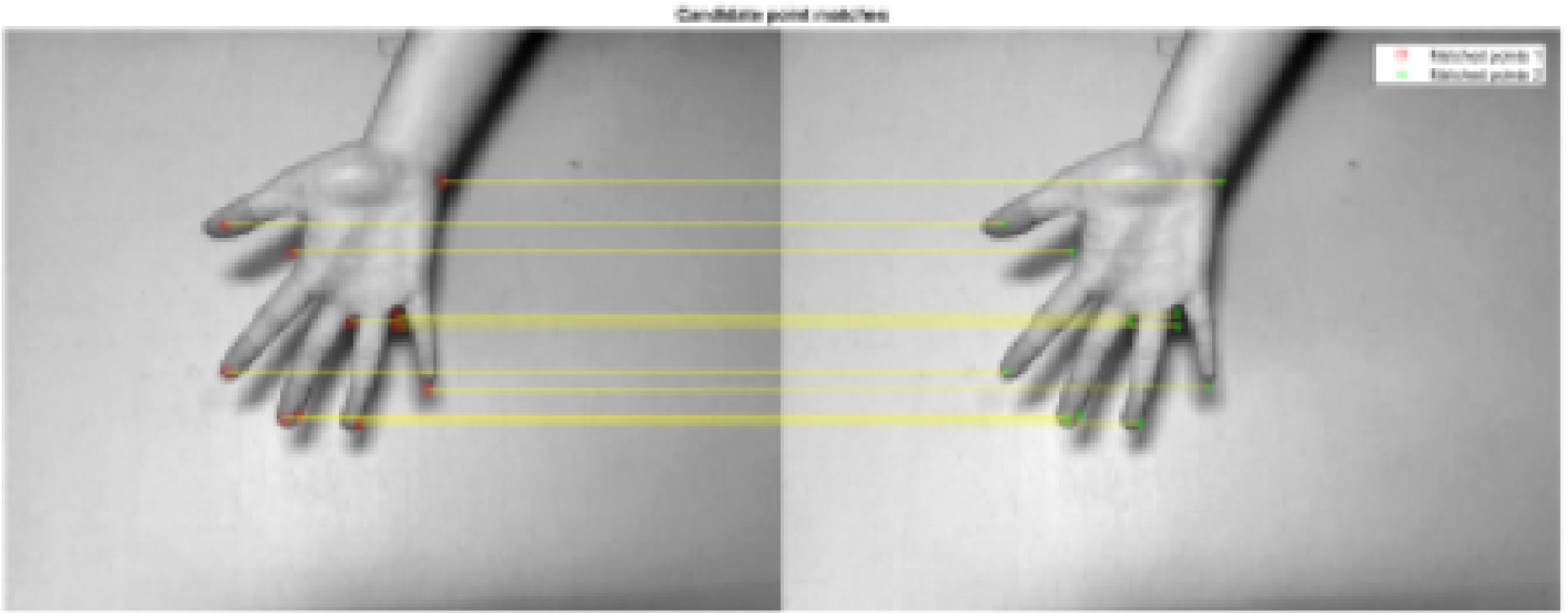
Instance:

- a collection (dataset) of datapoints from $\mathbf{X}$, e.g., $\mathbb{R}^n$
- (distance functions for some of the dimensions)

Solution: similarity function $\sigma : \mathbf{X} \to \mathbb{R}$

[**Measure:** error]

Good similarity measures are the key to accurate detection/classification



4. Clustering (segmentation)

group individuals in a population together by their similarity (but not driven by any specific purpose)

Instance:

- a collection (dataset) $\mathbf{D}$ of datapoints from $\mathbf{X}$, e.g., $\mathbb{R}^n$
- a relational structure on $\mathbf{X}$ (a graph)
- a small integer k

Solution: a partition of $\mathbf{D}$ into $\mathcal{C}_1, \ldots \mathcal{C}_k$

Measure: network modularity Q: proportion of the relational structure that *respects* the clusters.

Detection version: k is part of the output.

See an example research work (from yours truly)

5. Co-occurence (frequent itemset mining)

similarity of objects based on their appearing together in transactions.

Instance:

- a collection (dataset) $\mathbf{T}$ of itemsets (subsets of $\mathbf{X}$) or sequences
- a theshold τ

Solution: All *frequent patterns:* subsets that appear in $\mathbf{T}$ above τ

Detection version: τ is part of the output.

Market-basket analysis, (some) recommendation systems

6. Profiling (behaviour description)

Instance:

- a user description $\mathbf{u}$ drawn from a $\mathbf{D}$ collection
- a stimulus $a \in \mathbf{A}$
- a set of possible responses $\mathbf{R}$

Solution: a functional reaction of **u** to **a**, i.e., $\rho : \mathbf{U} \times \mathbf{A} \rightarrow \mathbf{R}$

Application: anomaly/fraud detection.

Example research work on Social media profiling

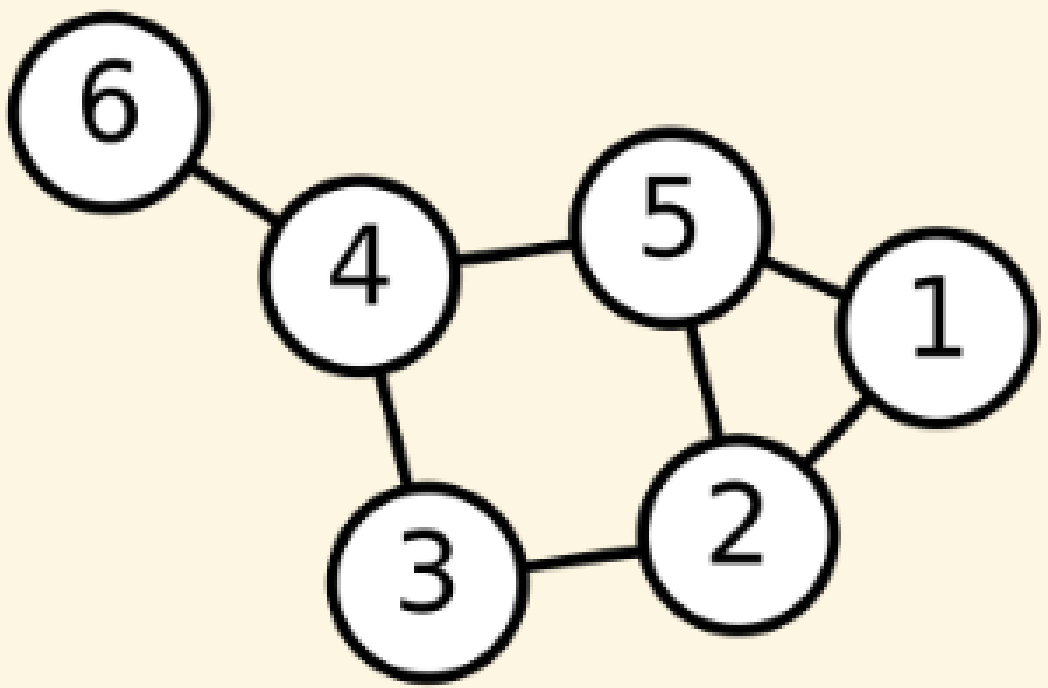
7. Link prediction

Instance: a dynamical graph (network) $\mathbf{G}$, i.e., a sequence

$< V, E >,$

$< V, E' = E + \{(u, v)\} >,$

$< V, E'' = E' + \{(r, s)\} > \ldots$



Question: what is the next link to be created?

What YouTube video will you watch next?

Alternatives: predict the **strength** of the new link; link deletion.

8. Data reduction

Instance:

- a collection (dataset) $\mathbf{D}$ of datapoints from $\mathbf{X}$, e.g., $\mathbb{R}^m$
- [a distinct independent variable x_i]

Solution: a projection of $\mathbf{D}$ onto $\mathbb{R}^n, n < m$

Measure: error in the estimation of x_i

Example: genre identification in consumer behaviour analysis

9. Causal modelling

Instance:

- a collection (dataset) $\mathbf{D}$ of datapoints from $\mathbf{X}$, e.g., $\mathbb{R}^m$
- a distinct dependent variable x_i

Solution: a variable x_j of $\mathbf{D}$ that controls x_i

Measure: effectiveness of x_j *tuning* to *tune* x_i in turn.

Example: Exactly What food causes you to put on weight?

Controlled clinical trials, A/B testing.

FROM PROBLEMS TO ALGORITHMS

Computer Science is often metaphoric: it uses abstractions that allow us to

- focus our problem-solving
- communicate with the computer

Examples:

- interpreted languages (Python)
- data models (SQL and relational tables)
- TCP/IP computer networks

Problem $\longrightarrow$ Algorithm $\longrightarrow$ Implementation (code)

For a given problem, more than one algorithm may be available

For a given algorithm, more than one implementation is possible

Only with clarity about the problem we can look for the algorithms.

SUPERVISED ALGORITHMS

Previous instance/solution pairs are available and fed to the a.

A. may ‘memorise’ past solutions and re-apply them, via some similarity criterion

A. may also ‘learn’ a model and apply it to future inputs

METHOD

- obtain a dataset of examples, inc. the “target” dimension, called *label*
- split it in training and test data
- run a. on the test data, find a putative solution
- test the quality/pred. power against test data

Regression has a numeric target while classification has a categorical/binary one

P. WITH GOOD SUPERVISED ALGORITHMS

1: Regression

2: Classification

9: Causal Modelling

P. WITH MOSTLY UNSUPERVISED A.

4: Clustering

5: co-occurrence grouping

6: profiling

P. WITH A MIX OF SUPERVISED AND UNSUPERVISED

3: Similarity matching,

7: link prediction,

8: data reduction