Main points from last time

- The context of our topic is extracting knowledge from data
  - For example, in computer vision we want to derive a high level representation of a scene from image data
- The approach supports connecting data to knowledge through
  - Focussing on the model (inference is separate)
  - Probabilistic calculus is a universal language to combine disparate sources of evidence and previous knowledge
  - Probabilistic methods yields answers (a distribution) which
    - represent your uncertainty
    - naturally support other tasks (as input)
    - provide for prediction

Interdisciplinary Importance

In a collaborative setting, you do not necessarily get to create your own problem (based on what can be done).

In a collaborative setting, theory matters!

Connecting theory to models to data is not well understood by most domain scientists you might collaborate with, so ...

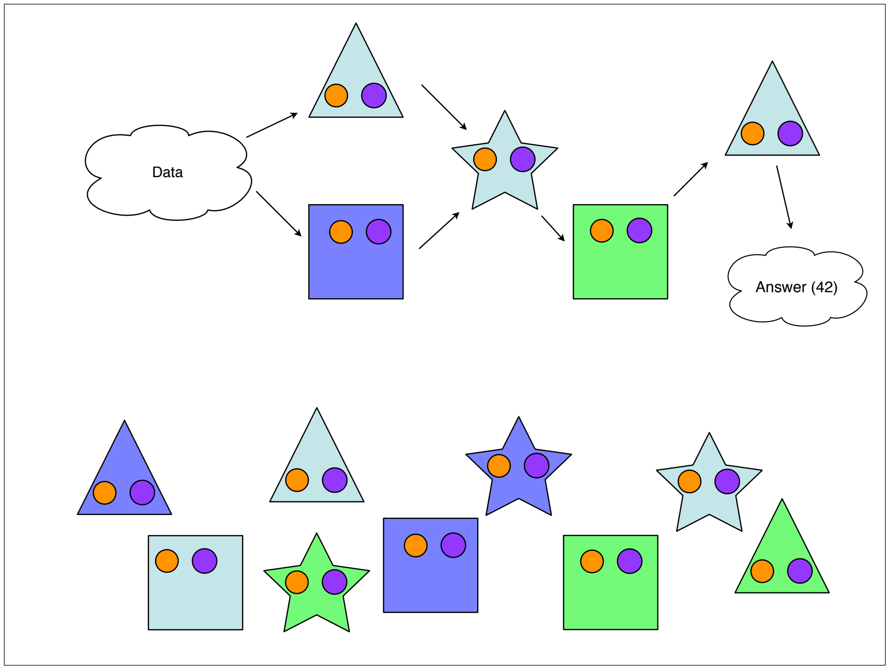
take this course!

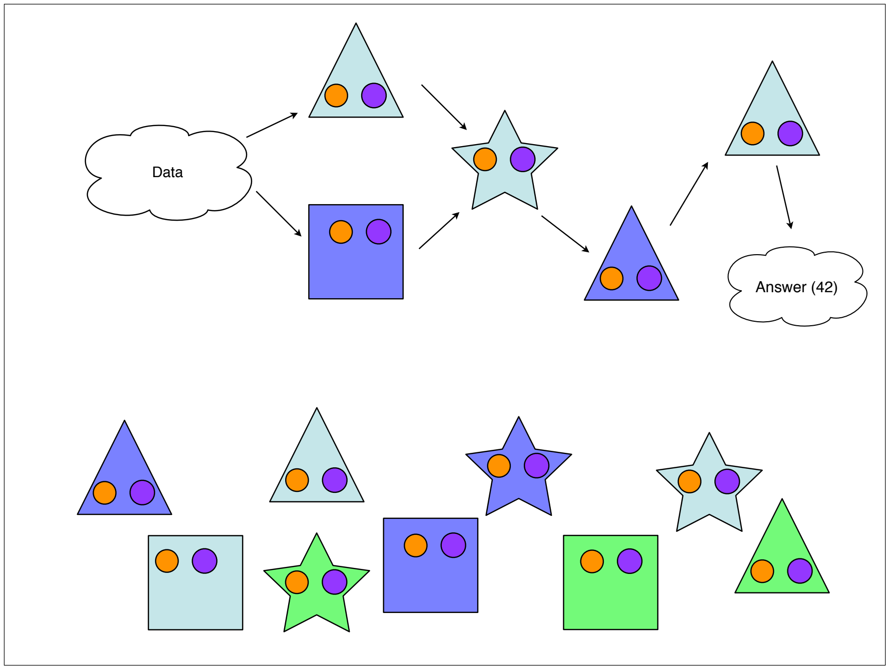
Social Sciences Example

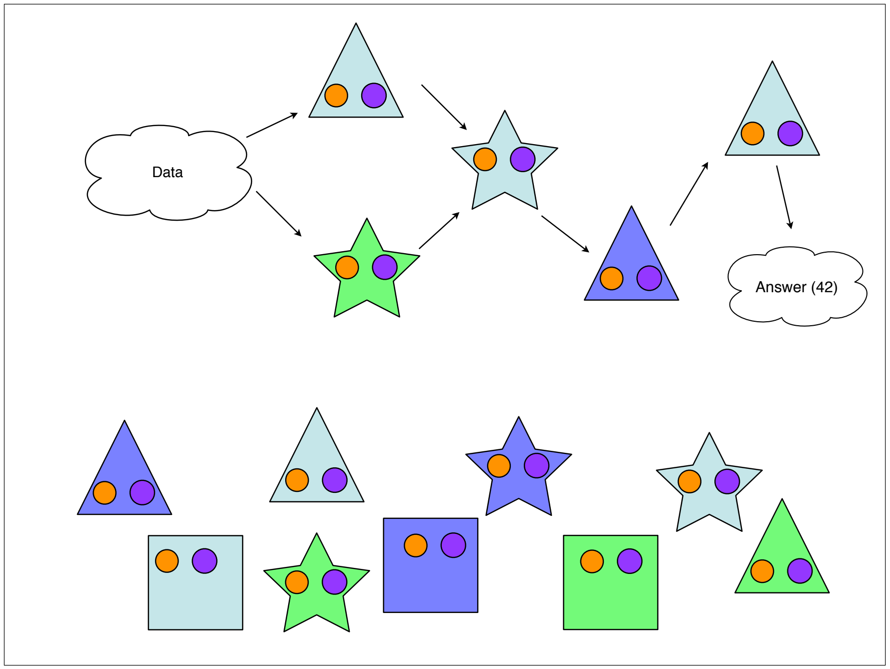
- Your collaborator says
  - I want to understand emotional processes when people in close relationships (e.g., couples) interact and how that influences health outcomes.
- Roughly
  - I have cool data, and I want to do something interesting with it
    - Find patterns, form theories, test them
- You hear
  - Cool data ... do something

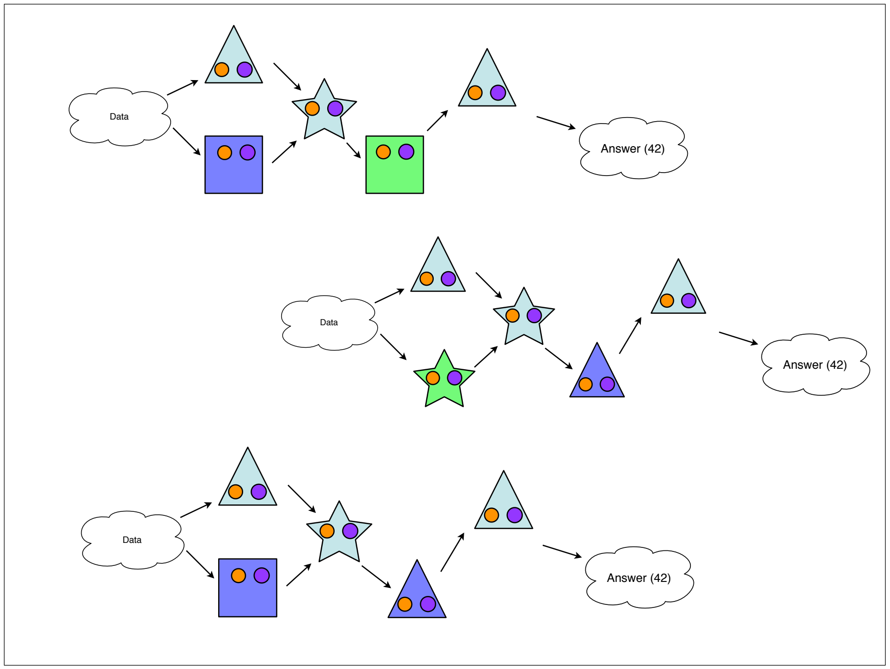
Social Sciences Example

- Your collaborator says
  - This data is really complex. Our analytic tools don’t seem sufficient.
- You think
  - I know lots of tools and I like playing with them


Data
Answer (42)


Data
Answer (42)

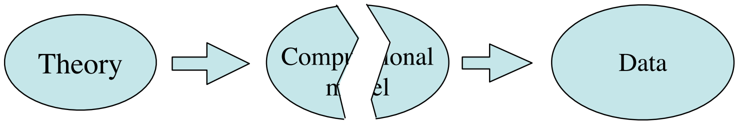

Data
Answer (42)


Data
Answer (42)
Data
Answer (42)
Data
Answer (42)

Cross Validation to the Rescue?

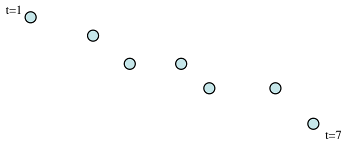
- Cross validation is extremely important but there are too many choices to try
- Even if you can try them all, you are limited to testing on a small subset of biased data sets.
- Even if you deal with all that, what have you really learned?
- You need something to help you choose (theory)

Collaborative Metaphor



Simple example

We observe a sequence of points ((x,y) coordinates) from an unknown physical process or sensor



What are the statistical dependencies? The points do not seem independent!

We might declare a plausible model that the points are independent, conditioned on a line model.

Why does modeling work at all?

Complex example:
Video feed from a camera watching the world

Summary points:
There is structure in the world

Real world **high** dimensional data is **not**!

Brute force representation of a high dimensional distribution is a bad idea for two reasons
It is completely impractical
It misses the forest for the trees

More summary

- Our models should capture what is important
- Mechanistically (i.e., explain)
- Statistically (simplify the joint density)
  - Two extremes, neither are useful
    - Everything is independent
    - Everything is dependent on everything

Clustering in high dimensions

Consider observed multidimensional data points (e.g., animal attributes such as color (3 attributes), height, mass normalized for height, texture (multiple numbers, ear length by height, ...)

$$\left(x_1, x_2,\ \ldots,\ x_n\right) \qquad \text{(data is for a particular animal)}$$

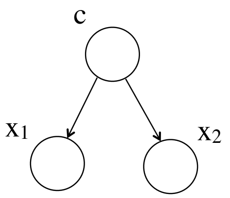
An explanation of our data is that the data is from species, c. Consider the joint probability

$$p\left(x_1, x_2,\ \ldots,\ x_n, c\right) = p(c)\, p\left(x_1, x_2,\ \ldots,\ x_n \middle| c\right)$$
$$= p(c) \prod_i p\left(x_i \middle| c\right)$$

Generative models

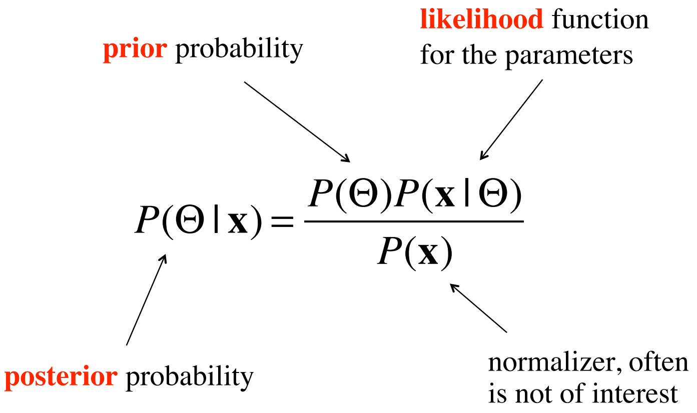
Informally, tells a story about how data comes to be

Illustrated using ancestral sampling



$$p\left(x_1, x_2, c\right) = p(c)\, p\left(x_1 \middle| c\right) p\left(x_2 \middle| c\right)$$

Bayesian inference



prior probability

likelihood function for the parameters

$$P(\Theta \mid \mathbf{x}) = \frac{P(\Theta) P(\mathbf{x} \mid \Theta)}{P(\mathbf{x})}$$

posterior probability

normalizer, often is not of interest

Simple example*

- What you know
  - John is coughing
- What do you conclude?
  - John has a cold
  - John has lung cancer
  - John has stomach problems

*Adopted from Josh Tenenbaum

Why this approach

Read first chapter of K&F linked from the lecture notes page.

Separates representation, modeling, and inference

Model is separated into prior and likelihood

Encourages being precise about the relationship between models and observed data

Priors are often key to handling complex models with enough variables to over-fit the data (need “regularization”)

Handles fitting and learning similarly

What is known is always represented as a distribution which is versatile for whatever the task is. Note that we evaluate models by their ability to predict.

Preparing for next time

If you have had minimal exposure to probability (or it was a long time ago), you need to take the up coming review seriously

Suggested reading is Kholar/Friedman chapter two (posted on lecture page)