DATA 552

Communication and Argumentation

Lecture 5: Middle of a project

Dr. Vikas Menghwani, Assistant Professor of Teaching

Warm-up

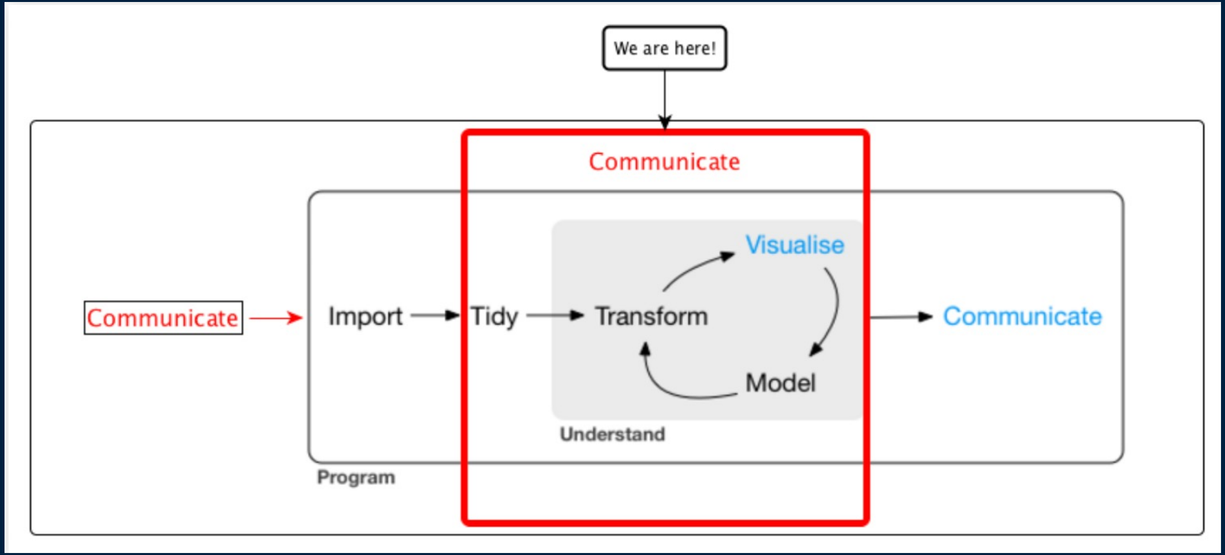
Analyse the following question using the framework we discussed on Tuesday:

- Should I sign up to give a talk at an upcoming data science meetup?

Specifically:

- What is the decision variable?
- What are the alternatives?
- What are the objectives? (if you were making this decision)
- What is the context? (if you were making this decision)

Middle of a data science project



What is happening at this stage of the project

- Making predictions / forming hypotheses
- Identifying relevant data sources, gathering it, organizing it
- Exploratory and in-depth data analysis
- Documenting your observations and processes
- Synthesizing your results

Subjectivity in the structured approach

- Note that all of the above require judgement calls by the data science practitioner, which in turn influence the results of the project!
- In academia, we often refer to this as the ‘researcher degrees of freedom’.

- Your hypotheses in the first step influence the
  - data sources you seek and choose
  - the patterns you look for in that data
  - the insights you focus on
  - and finally, the story you tell

- You are much more likely to dig for reasons as to why your project did not work as planned than you are to question why it did work!!

Subjectivity in the structured approach

- Viewing your own work through a critical lens is hard, but essential.
- Part of this is understanding what you’re good at, and what you’re not.
- Newsflash: many/most of us are not particularly good at the hypothesis/prediction generation stage.
- It is particularly susceptible to our biases.

Degrees of freedom at play

The phenomenon of p-hacking

Subjectivity in the structured approach

- This is concerning — I also just claimed that choice of hypothesis influences the remainder of the project in a cascading fashion!

- This is the reason to be aware of it, and be cautious.

- Play ‘devil’s advocate’ with yourself whenever you can. Team up with people who think in different ways than you. Listen carefully to potential concerns, and ensure you can rationalize and defend your choices.

- “We’re looking to model Y and decided to use data sets A, B, and D. But why not C?” — “C includes participants that fall under category M and are therefore not eligible for Y, their inclusion could seriously skew our ability to find patterns related to Y”

Example

Imagine we're conducting a study to model the relationship between exercise habits (Y) and heart health. We decide to use datasets A (a survey on exercise frequency), B (medical records on heart health), and D (a database tracking gym attendance).

Dataset C, which contains health data from a national health and nutrition examination survey. While this dataset is comprehensive, it includes participants with pre-existing heart conditions (category M) who are on strict exercise regimens prescribed by cardiologists.

Data scientist as a cartographer

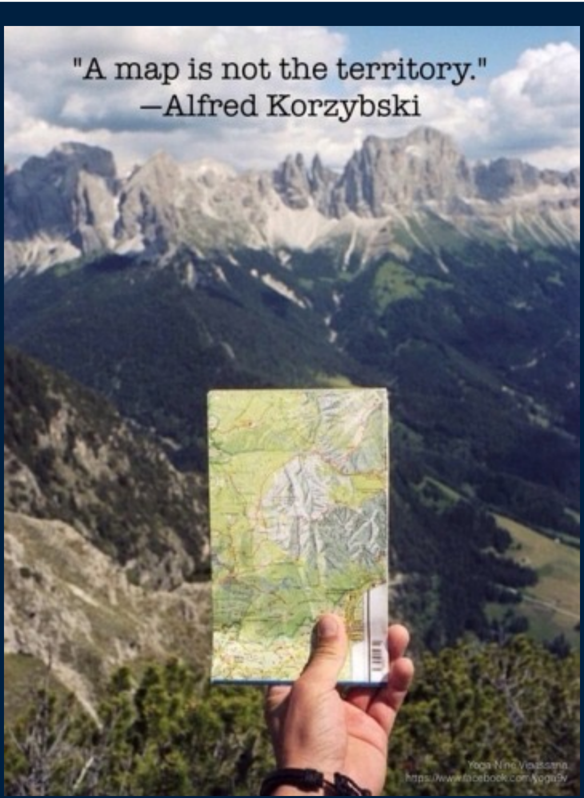
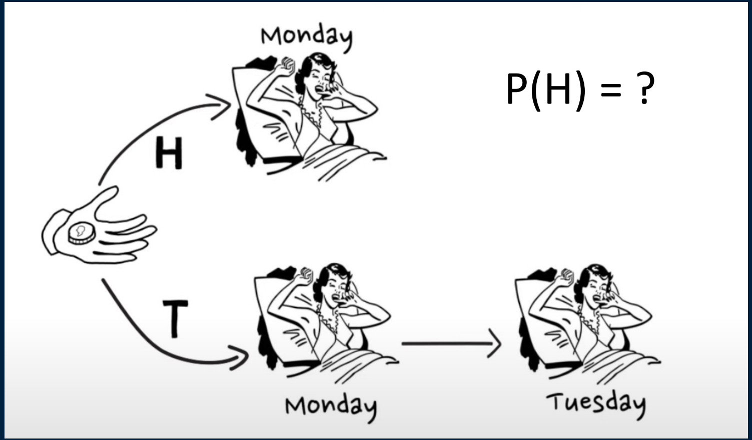

"A map is not the territory."
—Alfred Korzybski
Yoga Nino Viegsanna
https://www.facebook.com/yoga9v

image from quoteswave.com

A map is basically making specific claims about the territory. Depending on how detailed and accurate the map is (resolution / details) it may be closer to or far from a true depiction of the territory.

Subjective probability



The Sleeping Beauty Problem

Subjective probability is a type of probability derived from an individual's personal judgment or own experience about whether a specific outcome is likely to occur.

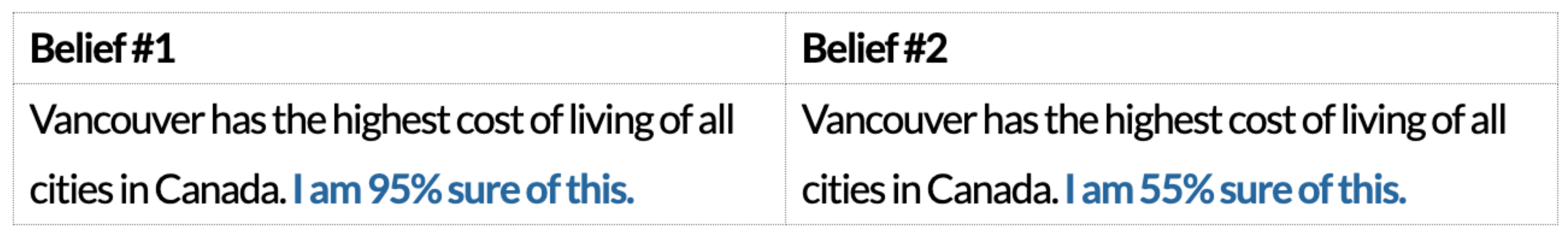
- investopedia

What does it mean, when we are modeling data

Let's imagine that the following claim is true:

- Vancouver has the highest cost of living of all cities in Canada.

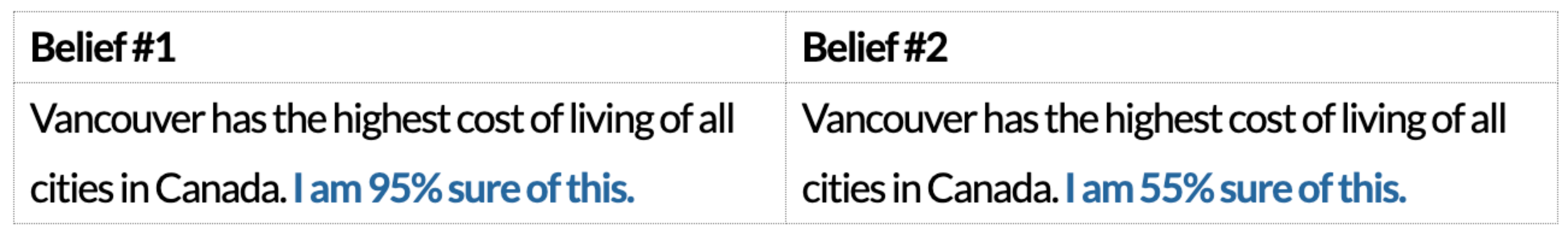
Now let's consider a few beliefs we could hold (these are potential sections of the map):



Belief #1	Belief #2
Vancouver has the highest cost of living of all cities in Canada. I am 95% sure of this.	Vancouver has the highest cost of living of all cities in Canada. I am 55% sure of this.

Which belief is better?

But what if it's actually Toronto that has the highest cost of living in Canada?



Belief #1	Belief #2
Vancouver has the highest cost of living of all cities in Canada. I am 95% sure of this.	Vancouver has the highest cost of living of all cities in Canada. I am 55% sure of this.

Which belief is better now?

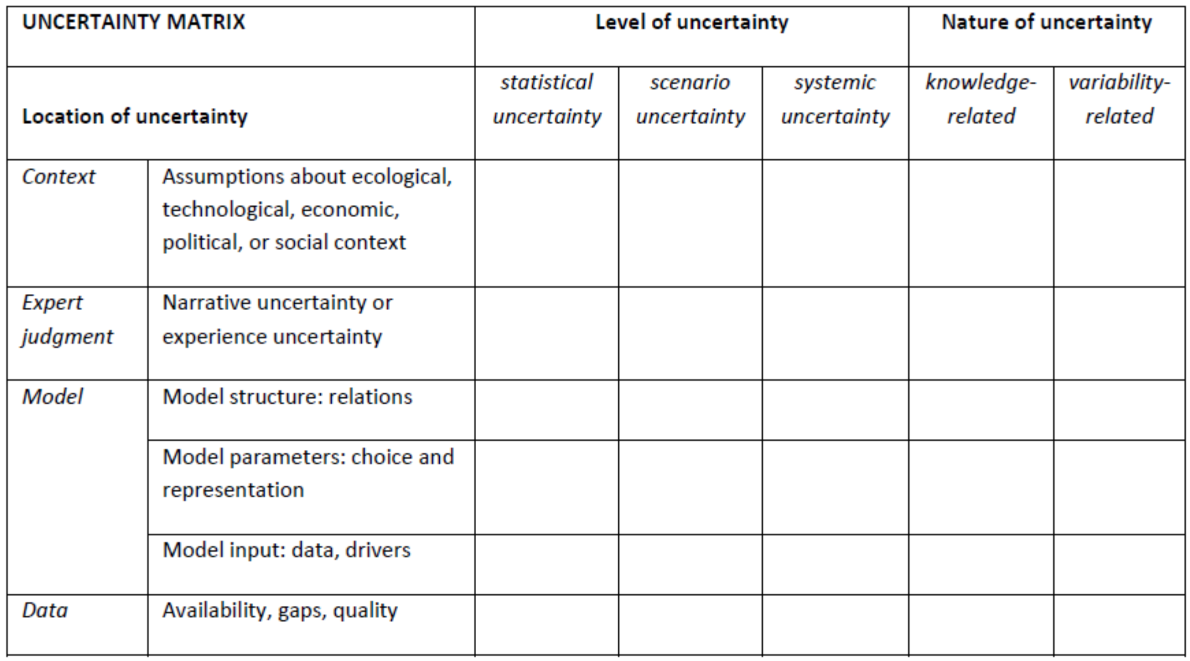
- We don't just want to be right. We want to be confident when we're right and hesitant when we're wrong.

Certainty

- Certainty in your results relies on your confidence in EVERY stage of the data science process.
- Did you design the experiment/survey?
- If yes...do you trust the calibration of your machines? Or, do you
- trust the neutrality of your polling questions?
- If not...how much do you trust the original source of the data?
- How much do you trust your cleaning of the data? Did you skew the results by removing observations with missing values? Are the missing values reasonably assumed to be missing at random? If not, why are they missing and is that important to my question?
- How much do you trust your teammates’ work?
- In short: you’ll never be 100% certain that every decision along the way was the right one. Ask questions, double check (time permitting).

Types of uncertainties

From the field of environmental management



UNCERTAINTY MATRIX		Level of uncertainty			Nature of uncertainty	
Location of uncertainty		*statistical uncertainty*	*scenario uncertainty*	*systemic uncertainty*	*knowledge-related*	*variability-related*
Context	Assumptions about ecological, technological, economic, political, or social context					
Expert judgment	Narrative uncertainty or experience uncertainty					
Model	Model structure: relations					
	Model parameters: choice and representation					
	Model input: data, drivers					
Data	Availability, gaps, quality					

- **statistical uncertainty:** calculated error and probabilities are known, and the decision risk can be calculated
- **scenario uncertainty:** how an impacted system might change is fairly understood, but the likelihood and extent of change are not known
- **systemic uncertainty:** uncertainties that cannot be estimated by any current method or technique—we simply don’t know

★

Portfolio

Example

Scott Alexander

THE UNIVERSITY OF BRITISH COLUMBIA