

# Introduction to *Urban Data* Science

## Introduction-II

(EPA1316)

Lecture 1

Trivik Verma



# Questions?

- Do you have a working setup of **conda+jupyter** in some form?

# Questions?

- Confusion around labs, homework and assignments...

## Labs

- Practice
- Not graded

## HW (1-8)

- Submit for peer-review on **PEER**
- Not graded

## Assignments (3,5,7)

- Graded
- Submit on BS

## Project (10)

- Graded
- Groups
- Submit on BS.

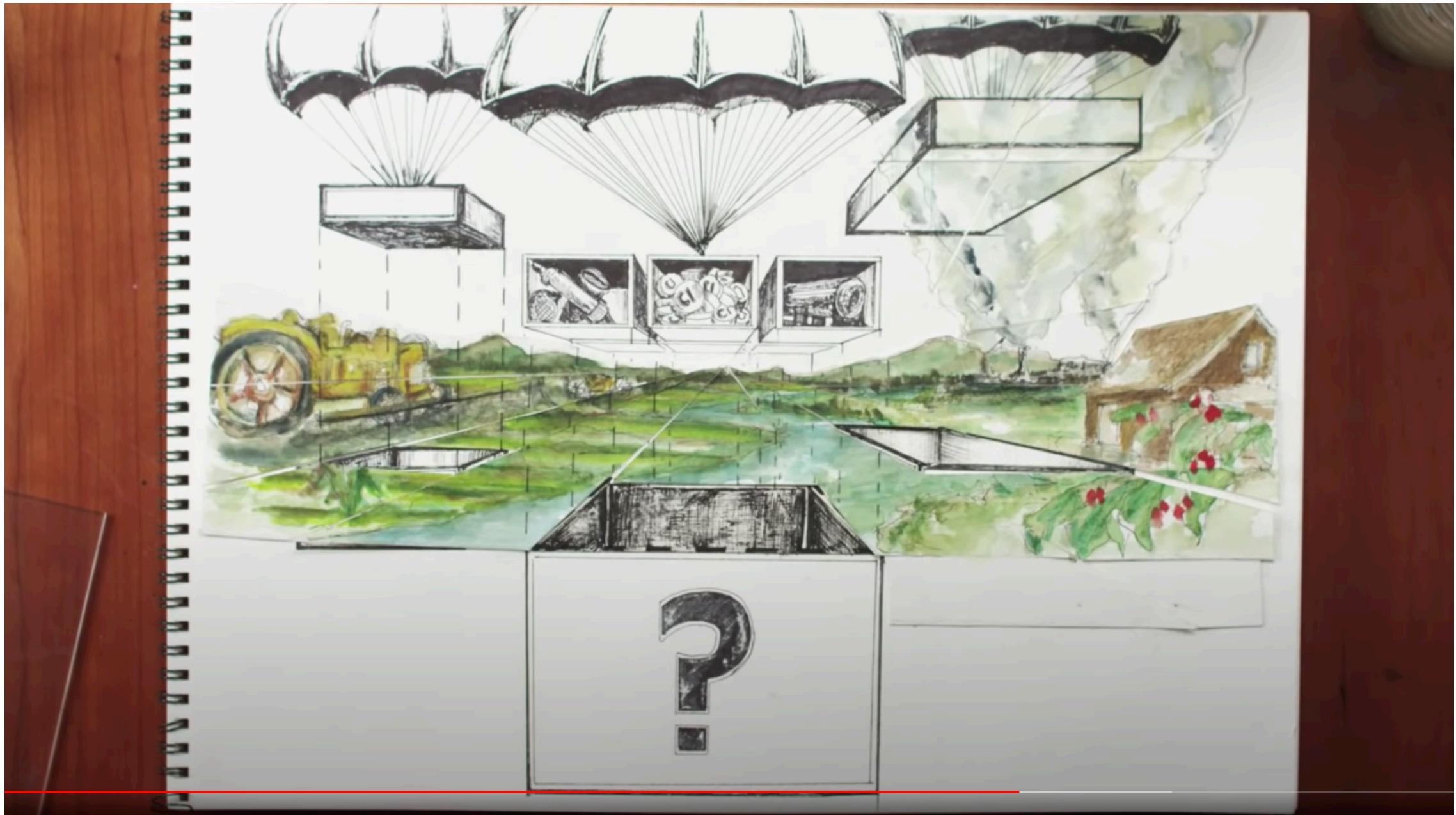
# Last Time

- Introduction to the Course
- The (Geo-)Data Revolution
- (Geo-)Data Science
- Tools - Python and Conda

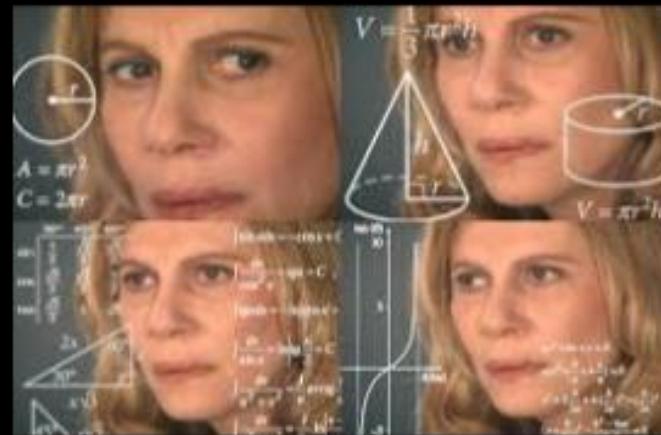
# Today

- Why Data Science?
- What is Data Science?
- Examine the **role** of evidence in policy
- Analyse **data** understanding and preparation **requirements**

# Why Data Science?



what my friends think I do



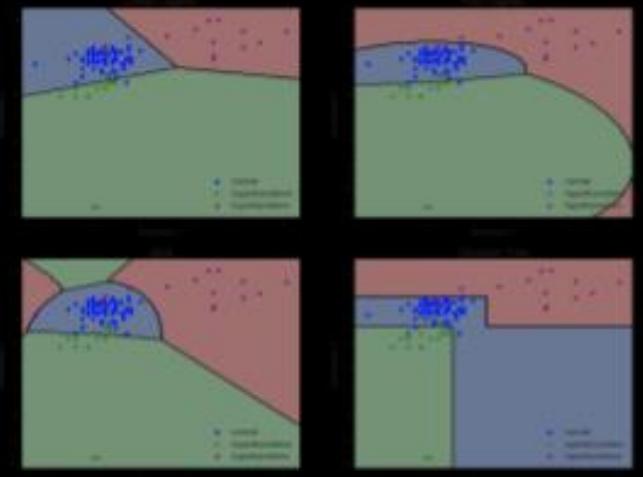
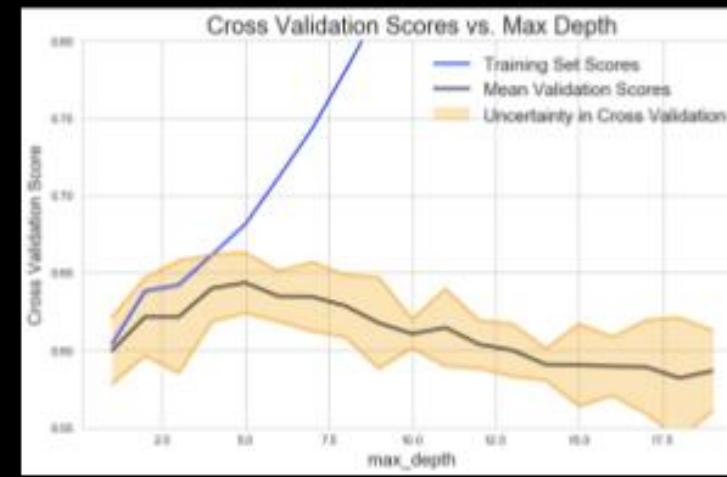
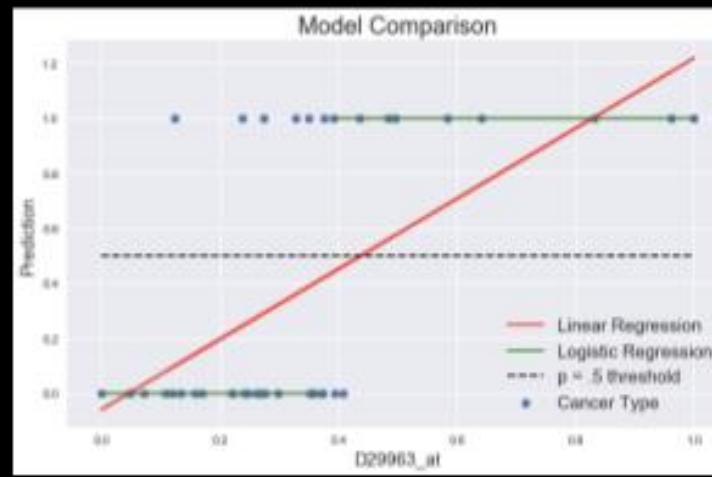
what my family thinks I do



what society thinks I do



what I actually (will) do in Data Science 1



# History

Long time ago (thousands of years) science was only empirical, and people counted stars



# History (cont)

Long time ago (thousands of years) science was only empirical, and people counted stars or crops



# History (cont)

Long time ago (thousands of years) science was only empirical, and people counted stars or crops and used the data to create machines to describe the phenomena



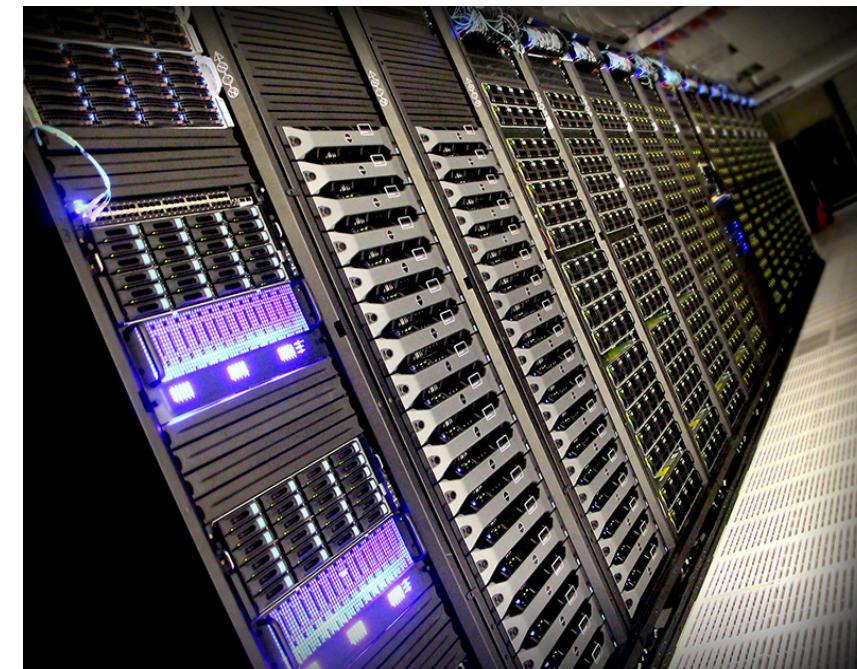
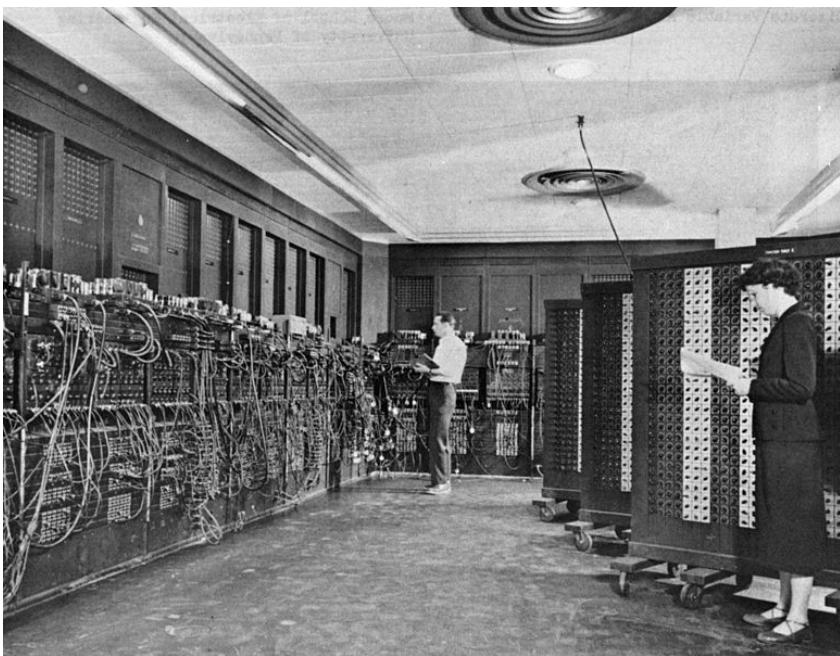
# History (cont)

Few hundred years: theoretical approaches, try to derive equations to describe general phenomena.

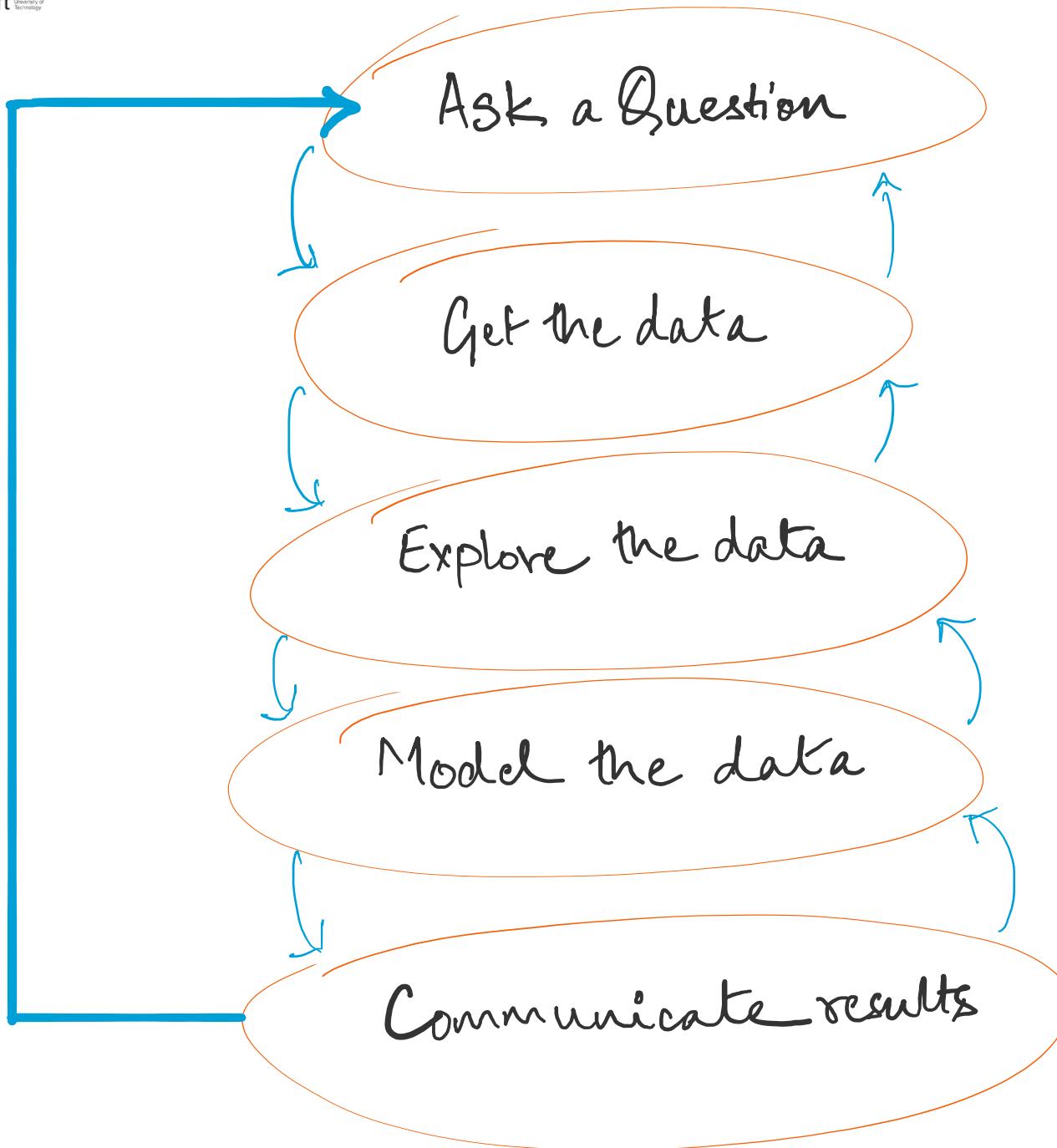
|                                     |  |  |                            |
|-------------------------------------|--|--|----------------------------|
| <b>Maxwell's Equations</b>          | $\nabla \cdot \mathbf{E} = 0$<br>$\nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{H}}{\partial t}$                        | $\nabla \cdot \mathbf{H} = 0$<br>$\nabla \times \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t}$ | J.C. Maxwell, 1865         |
| <b>Second Law of Thermodynamics</b> | $dS \geq 0$  |  | L. Boltzmann, 1874         |
| <b>Relativity</b>                   | $E = mc^2$   |  | Einstein, 1905             |
| <b>Schrodinger's Equation</b>       | $i\hbar \frac{\partial}{\partial t} \Psi = H\Psi$  |  | E. Schrodinger, 1927       |
| <b>Information Theory</b>           | $H = - \sum p(x) \log p(x)$  |  | C. Shannon, 1949           |
| <b>Chaos Theory</b>                 | $x_{t+1} = kx_t(1 - x_t)$  |  | Robert May, 1975           |
| <b>Black-Scholes Equation</b>       | $\frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} - rV = 0$ |  | F. Black, M. Scholes, 1990 |

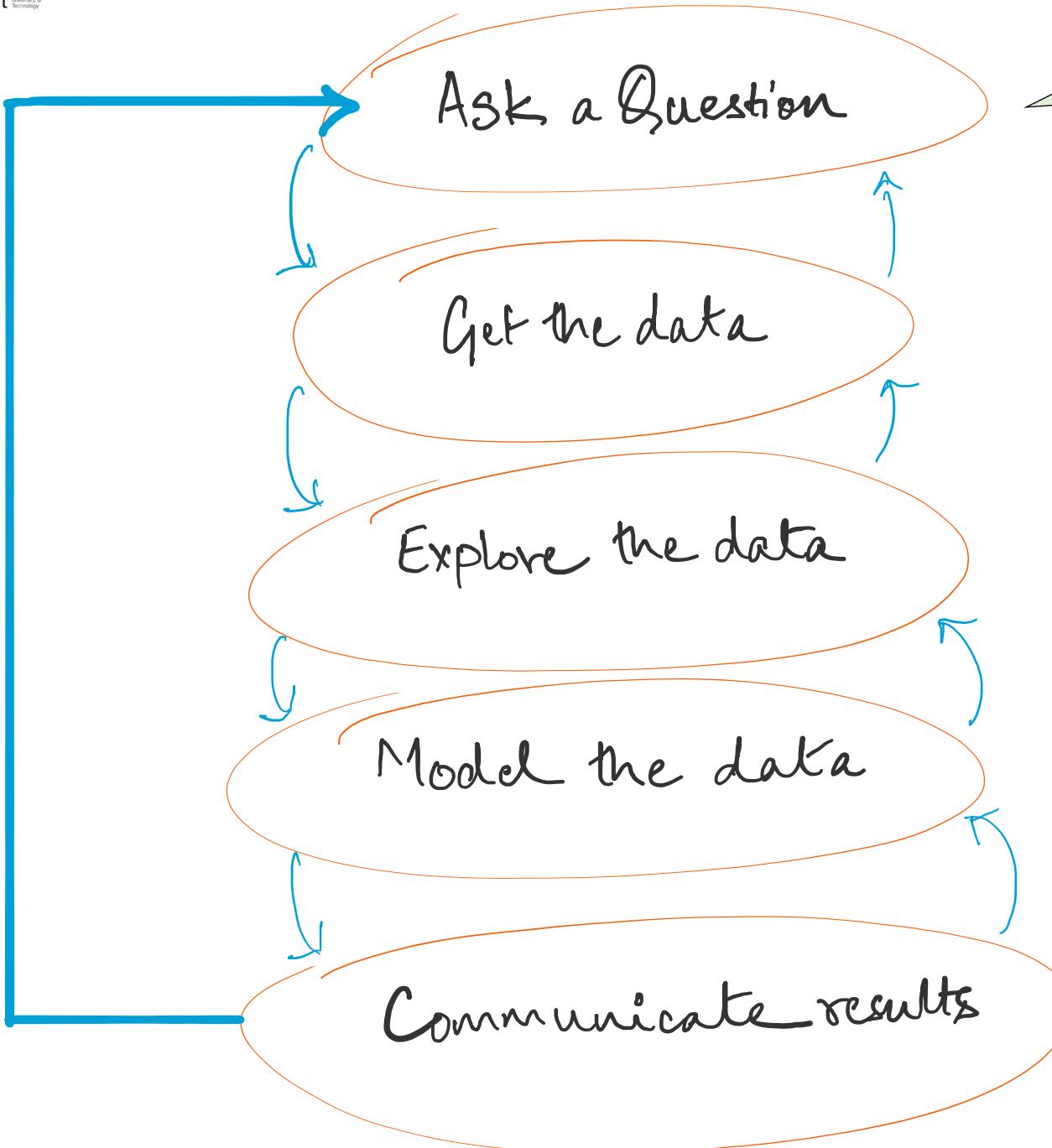
# History (cont)

About a hundred years ago: computational approaches

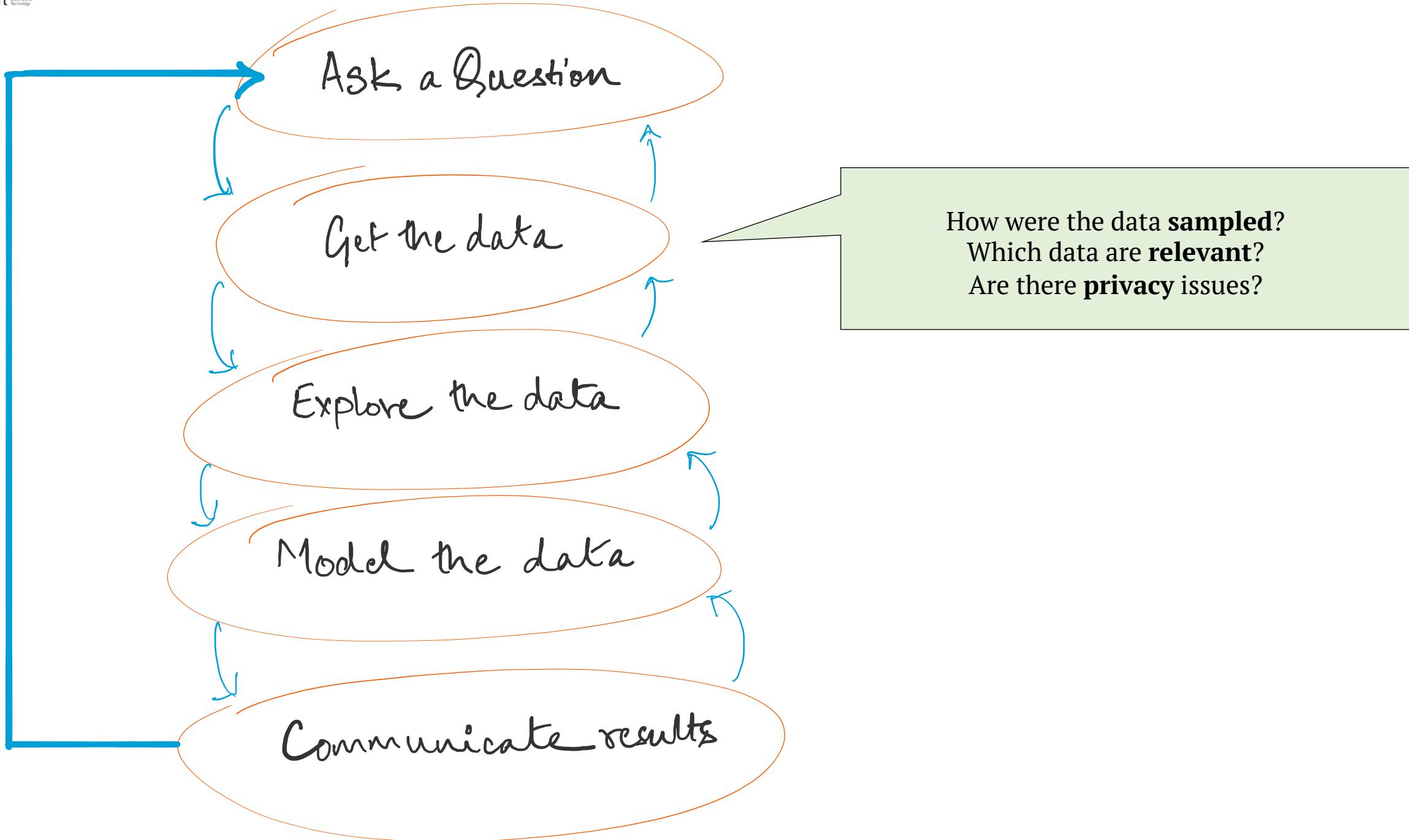


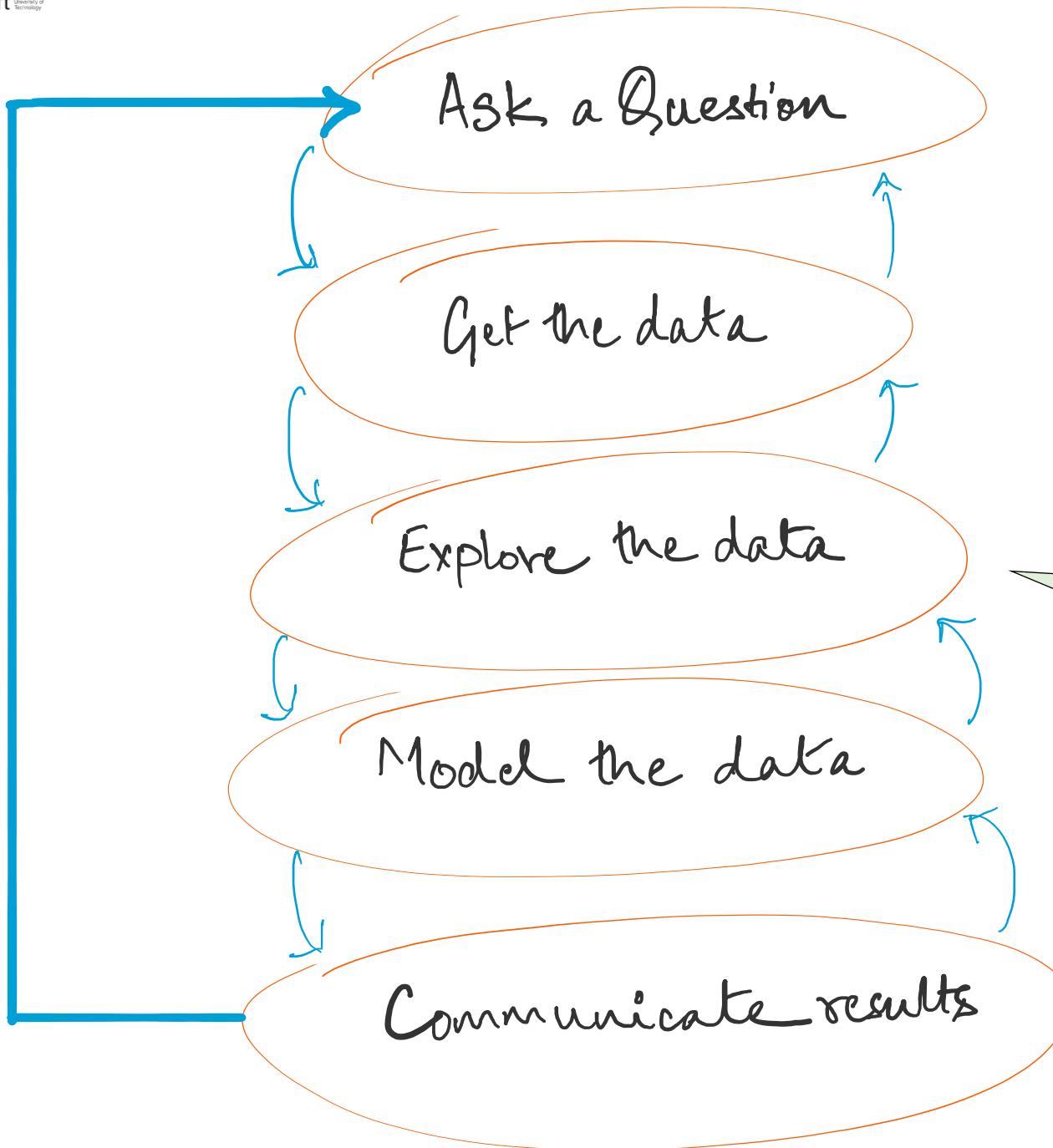
# What is Data Science?

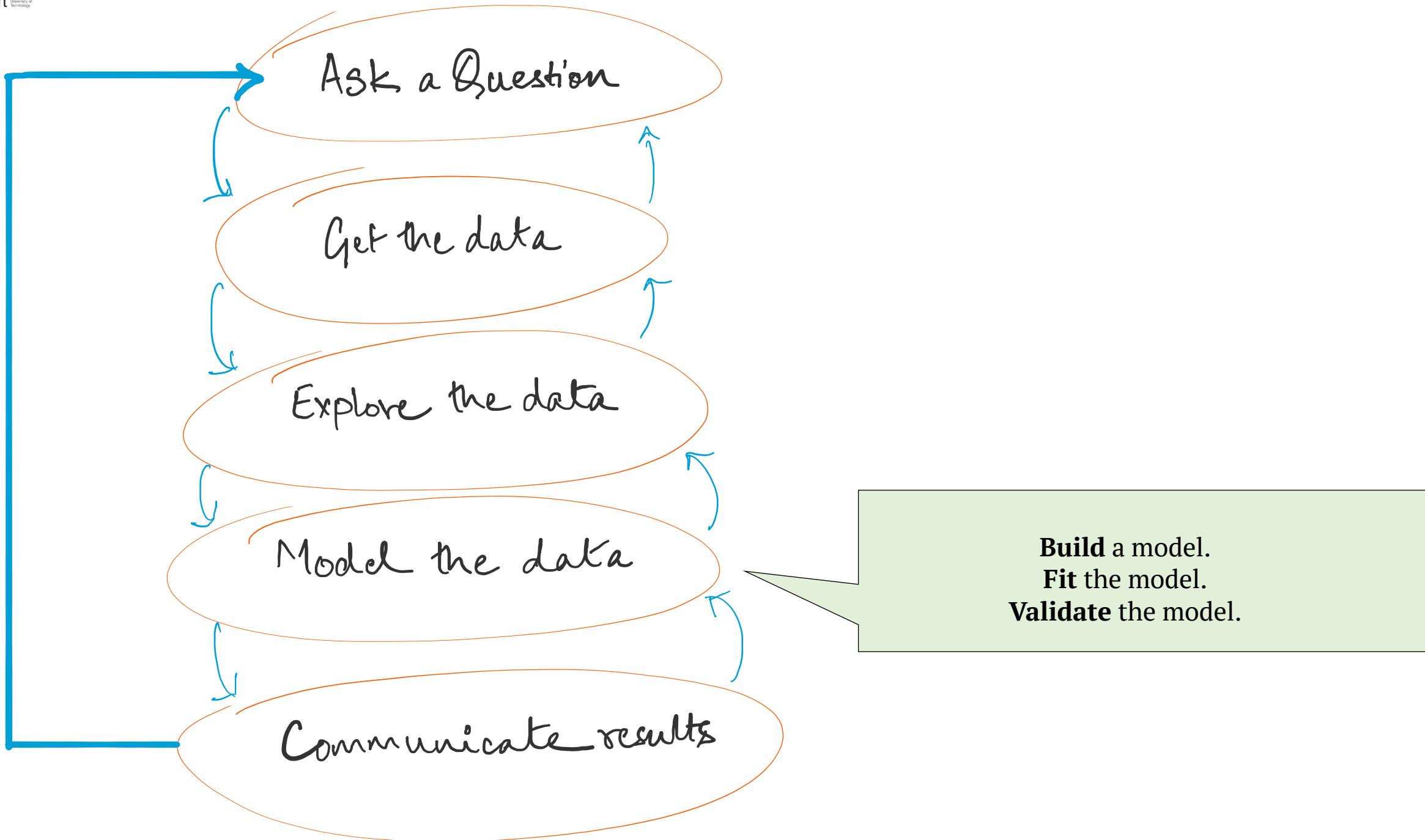


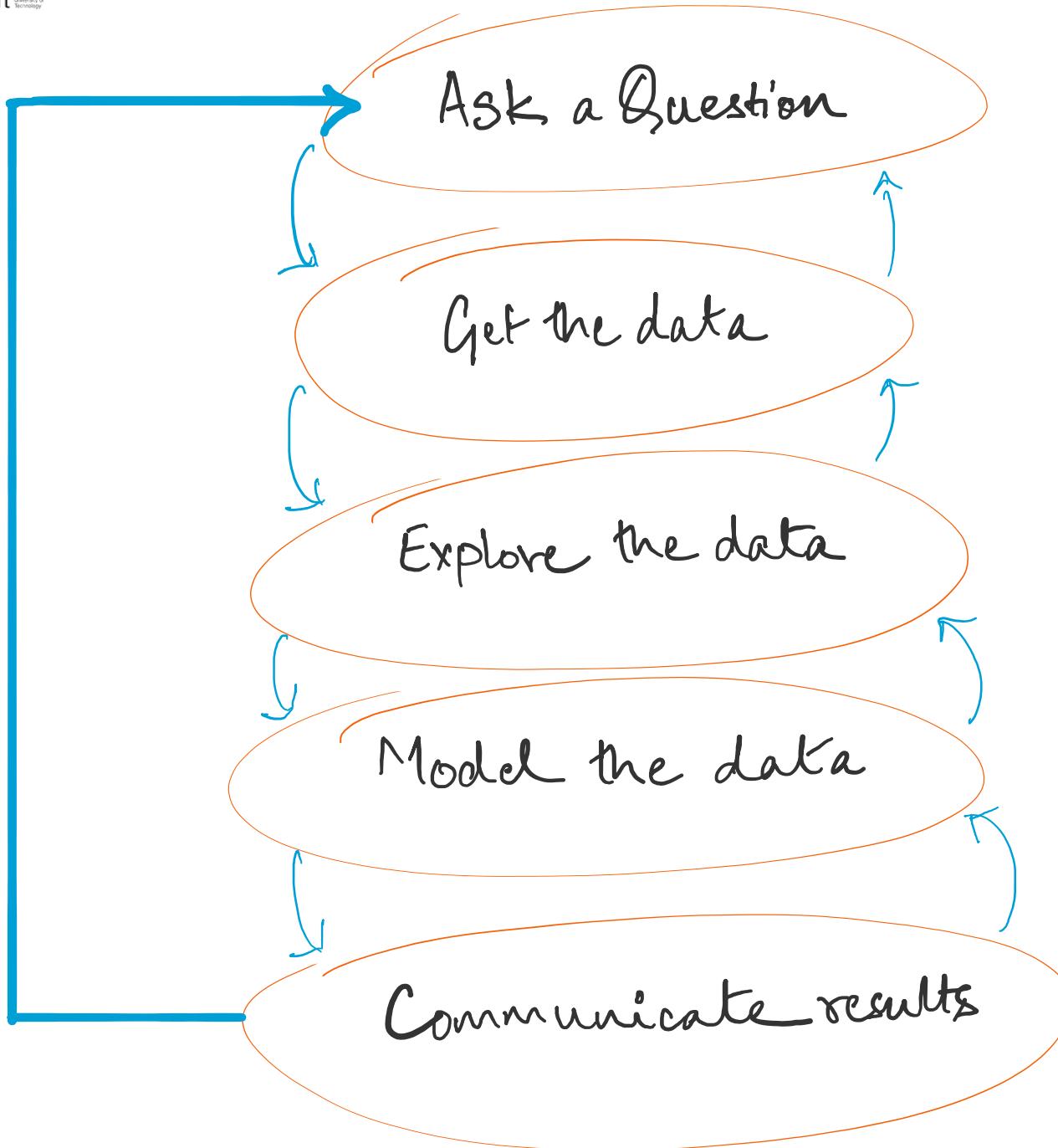


What is the scientific **goal**?  
What would you do if you had all the **data**?  
What do you want to **predict** or **estimate**?



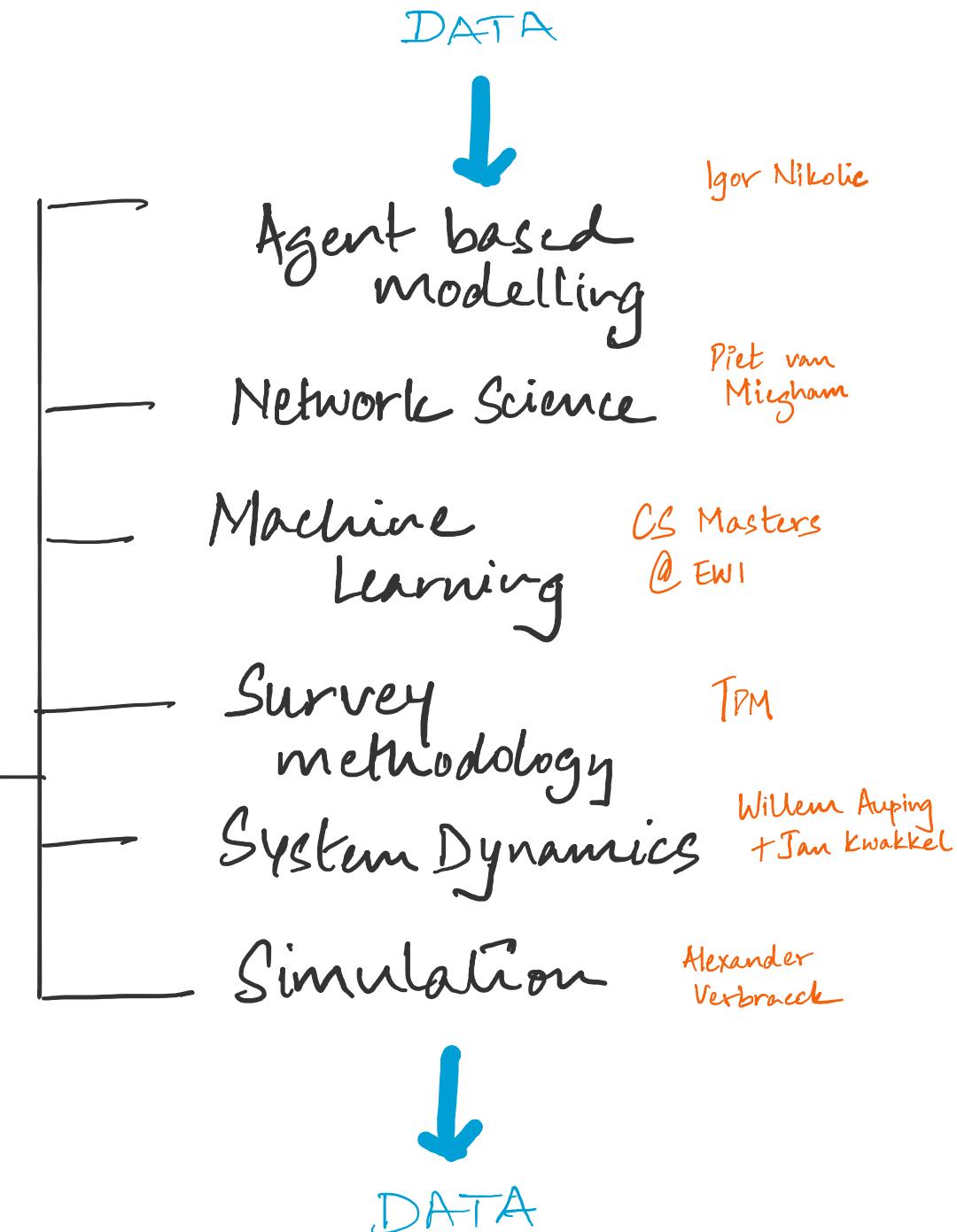
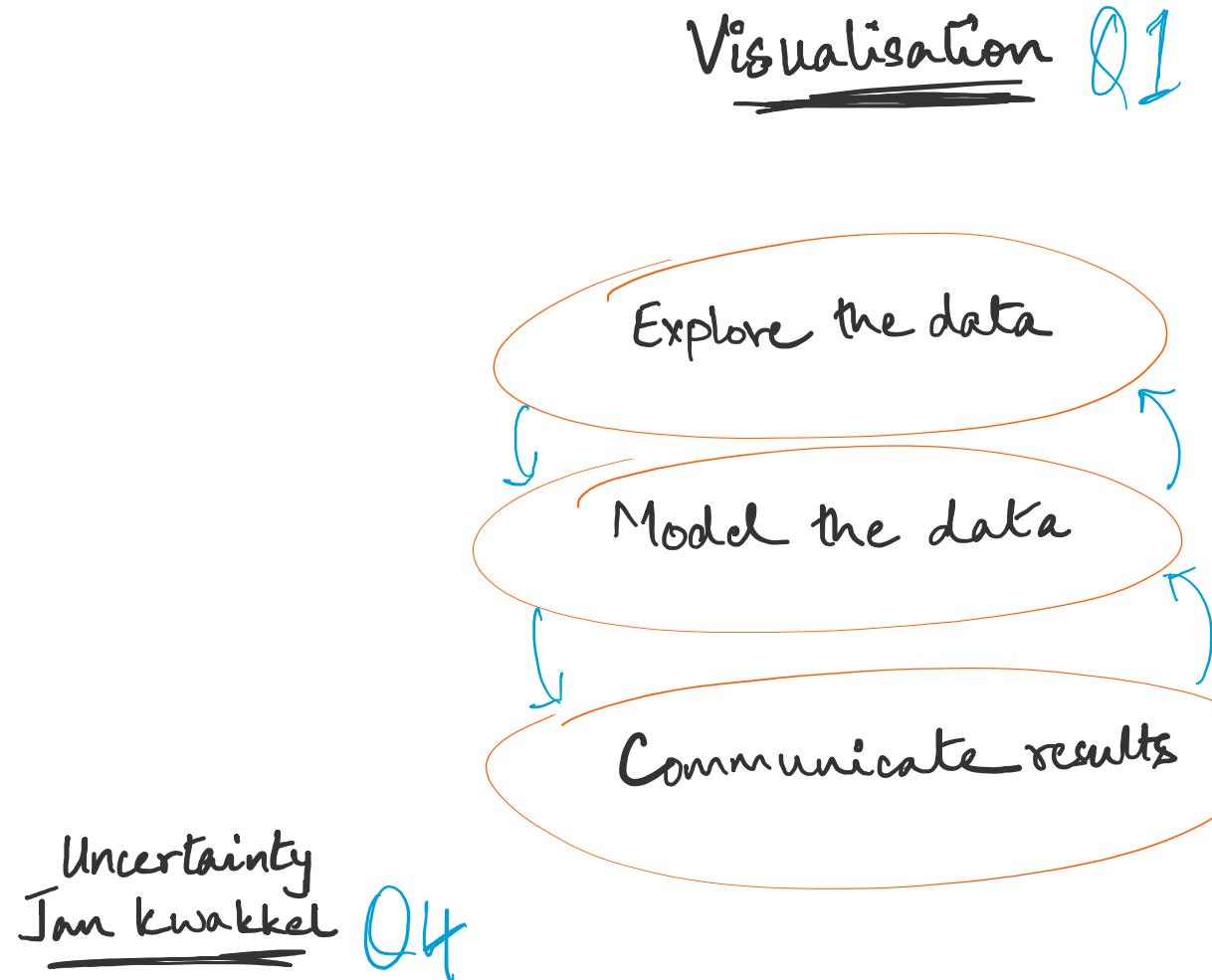






What did we **learn**?  
Do the results make **sense**?  
Can we tell a **story**?

# EPA Programme



# What are we going to do?

The material of the course will integrate the five key facets of an investigation using data:

1. Obtain: Obtaining data from multiple open data sources.
2. Scrub: Data cleaning, *wrangling*, sampling to consolidate all information into a dataset that is manageable, informative and relates to your problem.
3. Explore: Exploratory data analysis to make sense of what your data is trying to say (build intuition).
4. Model: Estimation and modelling based on statistical tools such as regression and clustering.
5. Interpret: Communicating results and reflections through visualisation, storytelling and interpretable summaries.

# The Data Science Process

The Data Science Process is like the scientific process - one of observation, model building, analysis and conclusion:

- Ask questions
- Data Collection
- Data Exploration
- Data Modeling
- Data Analysis
- Visualization and Presentation of Results

**Note:** This process is by no means linear!

# Analysing Hubway Data

- **Introduction:** Hubway (now called BlueBikes) is metro-Boston's public bike share program, with more than 1600 bikes at 160+ stations across the Greater Boston area. Hubway is owned by four municipalities in the area.
- By 2016, Hubway operated 185 stations and 1750 bicycles, with 5 million rides since launching in 2011.
- **The Data:** In April 2017, Hubway held a Data Visualization Challenge at the Microsoft NERD Center in Cambridge, releasing 5 years of trip data.
- **The Question:** What does the data tell us about the ride share program?

# The Data Exploration Cycle

Our original question: '**What does the data tell us about the ride share program?**' is a reasonable slogan to promote a hackathon. It is not good for guiding scientific investigation.

Before we can refine the question, we must look at the data!

|   | seq_id | hubway_id | status | duration | start_date         | strt_statn | end_date           | end_statn | bike_nr | subsc_type | zip_code | birth_date | gender |
|---|--------|-----------|--------|----------|--------------------|------------|--------------------|-----------|---------|------------|----------|------------|--------|
| 0 | 1      | 8         | Closed | 9        | 7/28/2011 10:12:00 | 23.0       | 7/28/2011 10:12:00 | 23.0      | B00468  | Registered | '97217   | 1976.0     | Male   |
| 1 | 2      | 9         | Closed | 220      | 7/28/2011 10:21:00 | 23.0       | 7/28/2011 10:25:00 | 23.0      | B00554  | Registered | '02215   | 1966.0     | Male   |
| 2 | 3      | 10        | Closed | 56       | 7/28/2011 10:33:00 | 23.0       | 7/28/2011 10:34:00 | 23.0      | B00456  | Registered | '02108   | 1943.0     | Male   |
| 3 | 4      | 11        | Closed | 64       | 7/28/2011 10:35:00 | 23.0       | 7/28/2011 10:36:00 | 23.0      | B00554  | Registered | '02116   | 1981.0     | Female |
| 4 | 5      | 12        | Closed | 12       | 7/28/2011 10:37:00 | 23.0       | 7/28/2011 10:37:00 | 23.0      | B00554  | Registered | '97214   | 1983.0     | Female |

Based on the data, what kind of questions can we ask?

# The Data Exploration Cycle

**Who?** Who's using the bikes?

Refine into specific hypotheses:

- More men or more women?
- Older or younger people?
- Subscribers or one-time users?

# The Data Exploration Cycle

**Where?** Where are bikes being checked out?

Refine into specific hypotheses:

- More in Boston than Cambridge?
- More in commercial or residential?
- More around tourist attractions?

**Sometimes the data is given to you in pieces and must be merged!**

# The Data Exploration Cycle

**When?** When are the bikes being checked out?

Refine into specific hypotheses:

- More during the weekend than on the weekdays?
- More during rush hour?
- More during the summer than the fall?

**Sometimes the feature you want to explore doesn't exist in the data and must be engineered!**

# The Data Exploration Cycle

**Why?** For what reasons/activities are people checking out bikes?

Refine into specific hypotheses:

- More bikes are used for recreation than commute?
- More bikes are used for touristic purposes?
- Bikes are used to bypass traffic?

**Do we have the data to answer these questions with reasonable certainty?**

**What data do we need to collect in order to answer these questions?**

# The Data Exploration Cycle

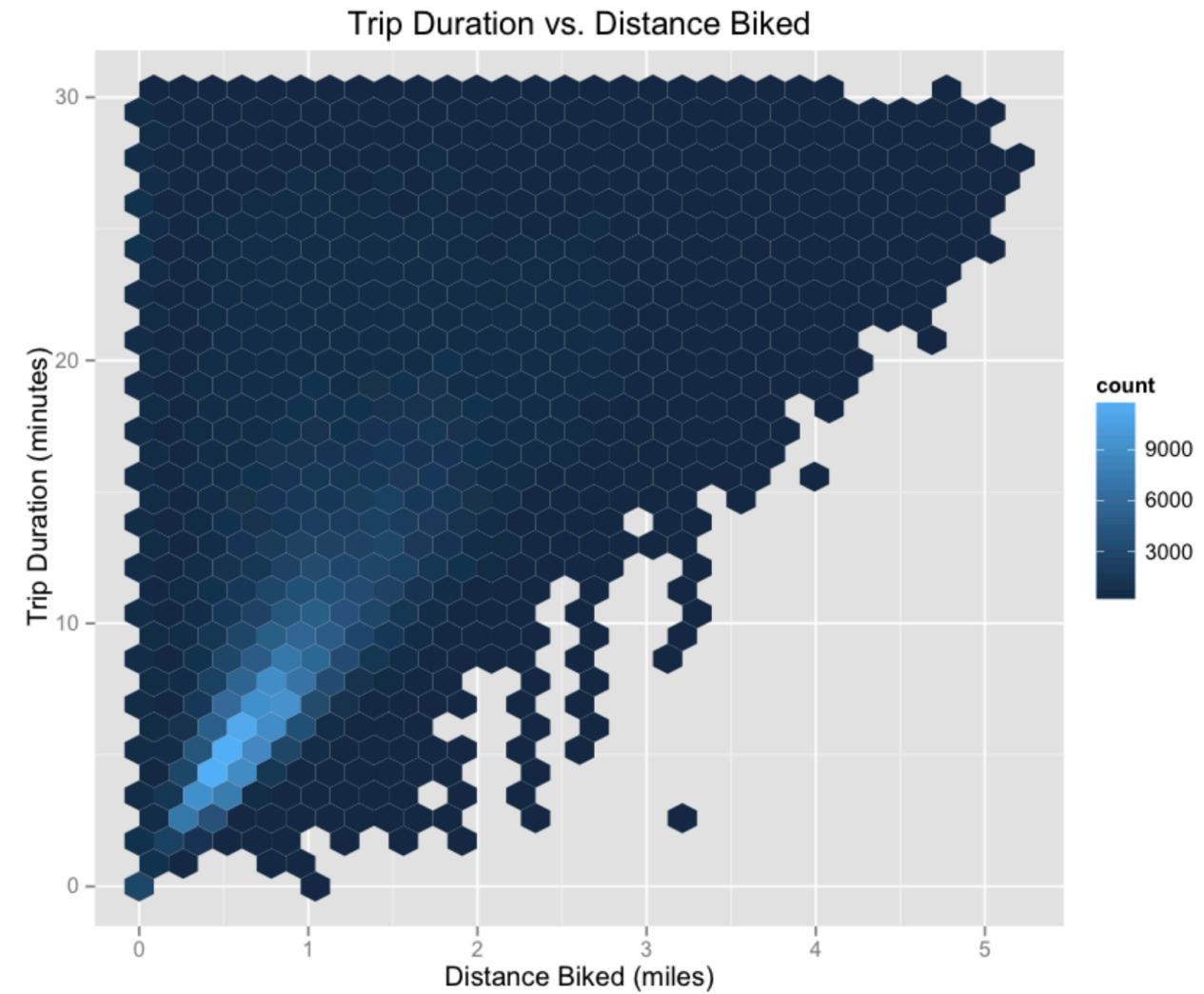
**How?** Questions that combine variables.

- How does user demographics impact the duration the bikes are being used? Or where they are being checked out?
- How does weather or traffic conditions impact bike usage?
- How do the characteristics of the station location affect the number of bikes being checked out?

**How questions are about modeling relationships between different variables.**

# Inspiration for Exploring

So how well did we do in formulating creative hypotheses and manipulating the data for answers?



# Break



WATER



WALK



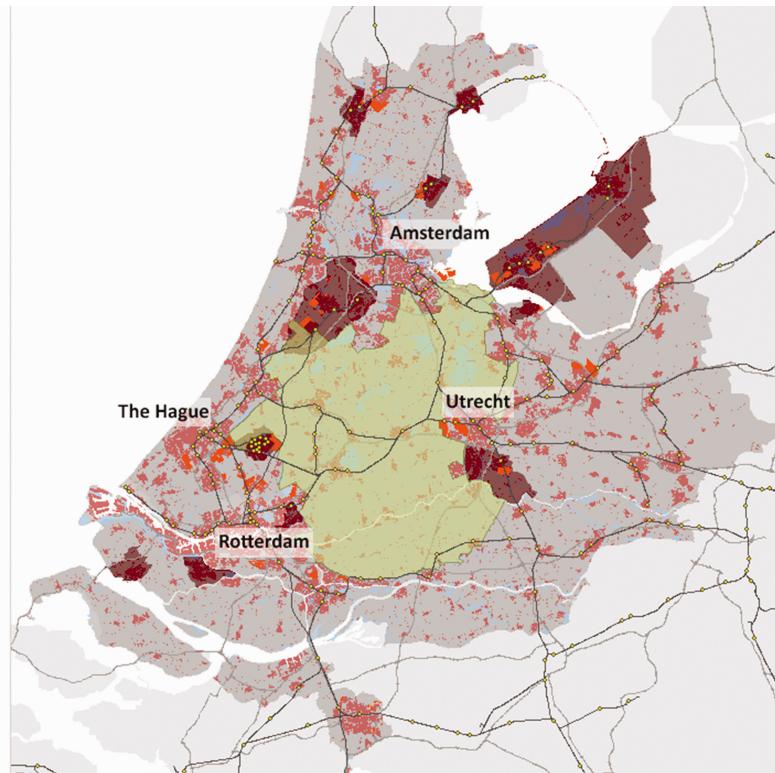
COFFEE OR TEA



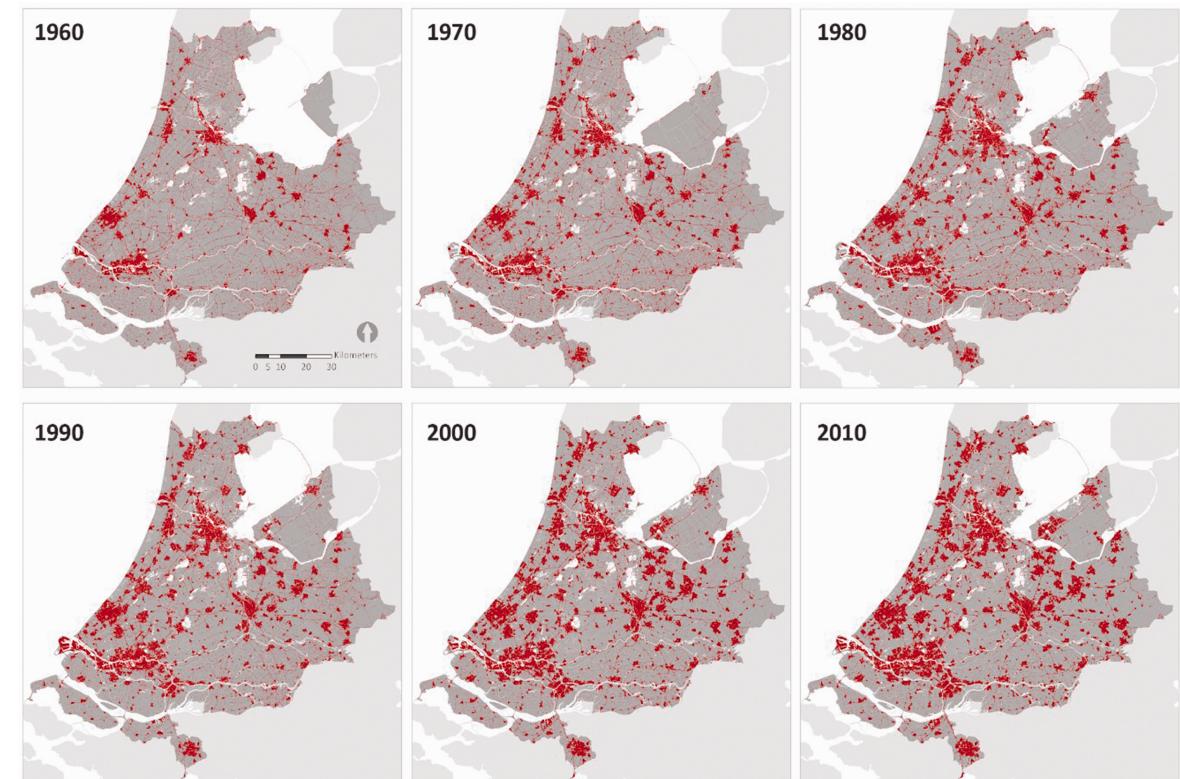
MAKE FRIENDS

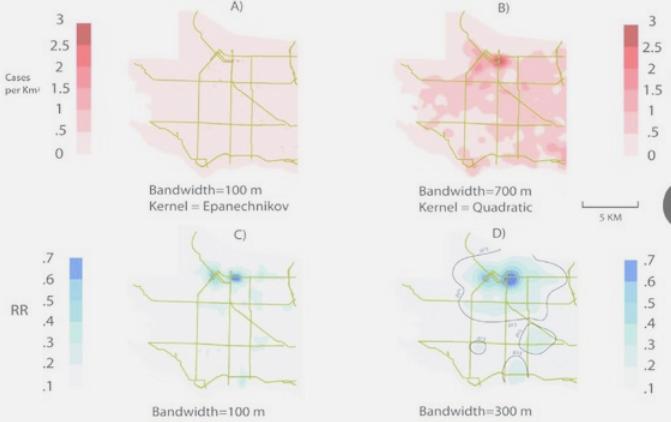
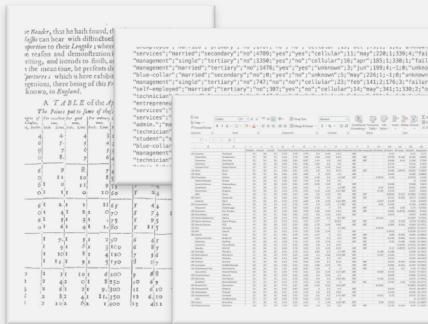
# Role of Evidence in Policy

# Example: Randstad



Urbanisation of the Greater Randstad Area 1960-2010





# Problem Understanding (Vision)

SUSTAINABLE

RESILIENT

INCLUSIVE

SMART

EQUITABLE

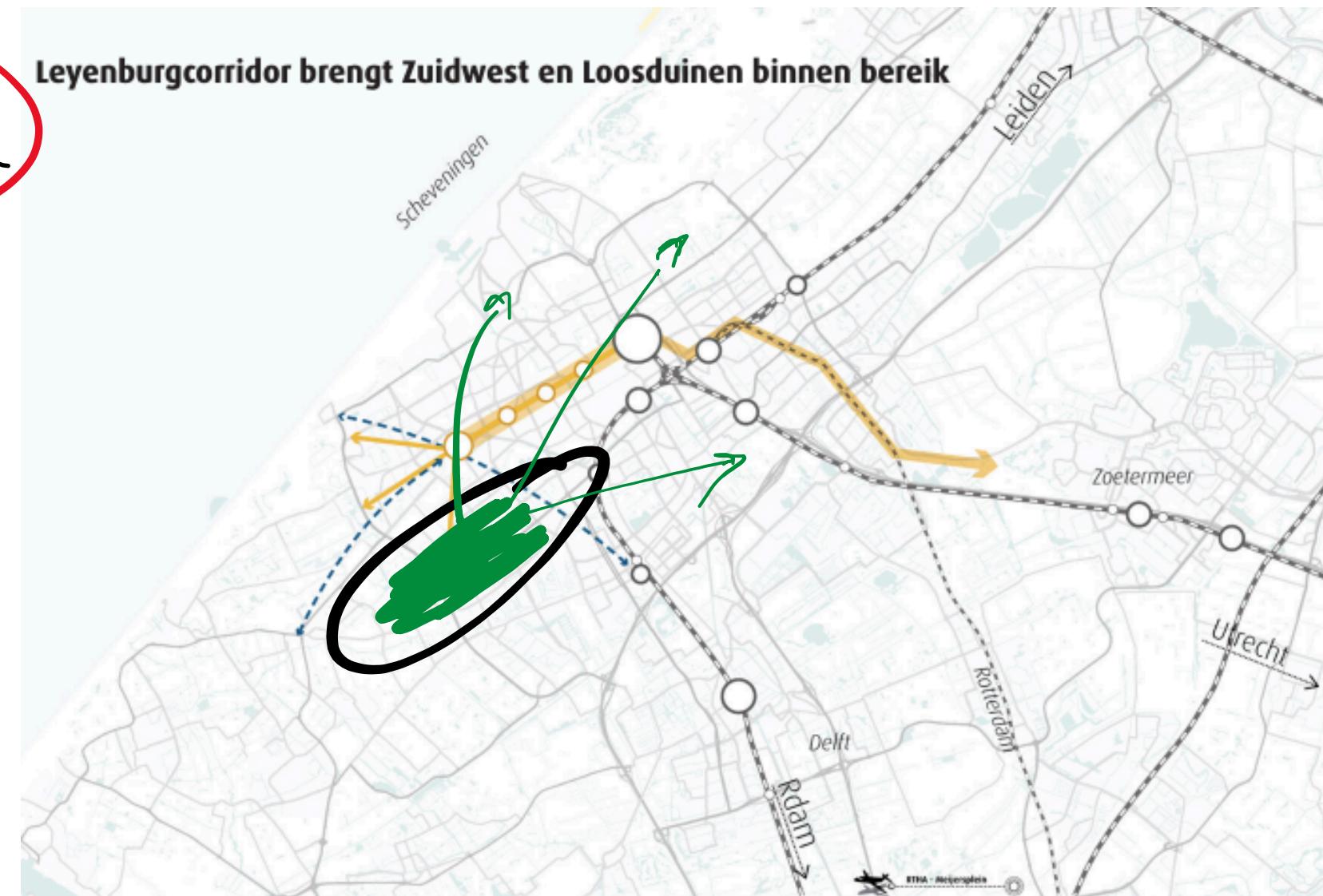


# Problem Understanding

Poor Neighbourhood

Promote

- Development
- Social Cohesion
- Access to jobs + infrastructure
- Reduce car trips + increase PT use



# Problem Understanding

- Determine what the objectives are
- Assess the situation **resources, risks, costs and benefits**
- Determine data mining goals
- Develop a project plan **estimate timeline, budget, but also tools and techniques**



# Problem Understanding

- Difficult!
- Often, new knowledge required
- Explain limitations to non-experts
  - Do you have data? “No”
  - Accuracy will be 0.5%
  - Not next month, maybe next year



IN CS, IT CAN BE HARD TO EXPLAIN  
THE DIFFERENCE BETWEEN THE EASY  
AND THE VIRTUALLY IMPOSSIBLE.

# Problem Understanding

My DOs and DON'TS

- Be extremely patient for vaguely defined problems
- Concretely reduce the scope of the idea
  - Data Samples are essential
  - Real-life case studies
  - Measures for success
- Ill-defined and unrealistic? Go to the beach



IN CS, IT CAN BE HARD TO EXPLAIN  
THE DIFFERENCE BETWEEN THE EASY  
AND THE VIRTUALLY IMPOSSIBLE.



Cesar A. Hidalgo  
@cesifoti

When you hear people define their academic field, among highly overlapping disciplines, the field names give you as much information about people's allegiance to a group, than about a set of concepts.

Learn to *look past* the keywords to basic *theory* and *concepts*

Ex. Design a sustainable food production system

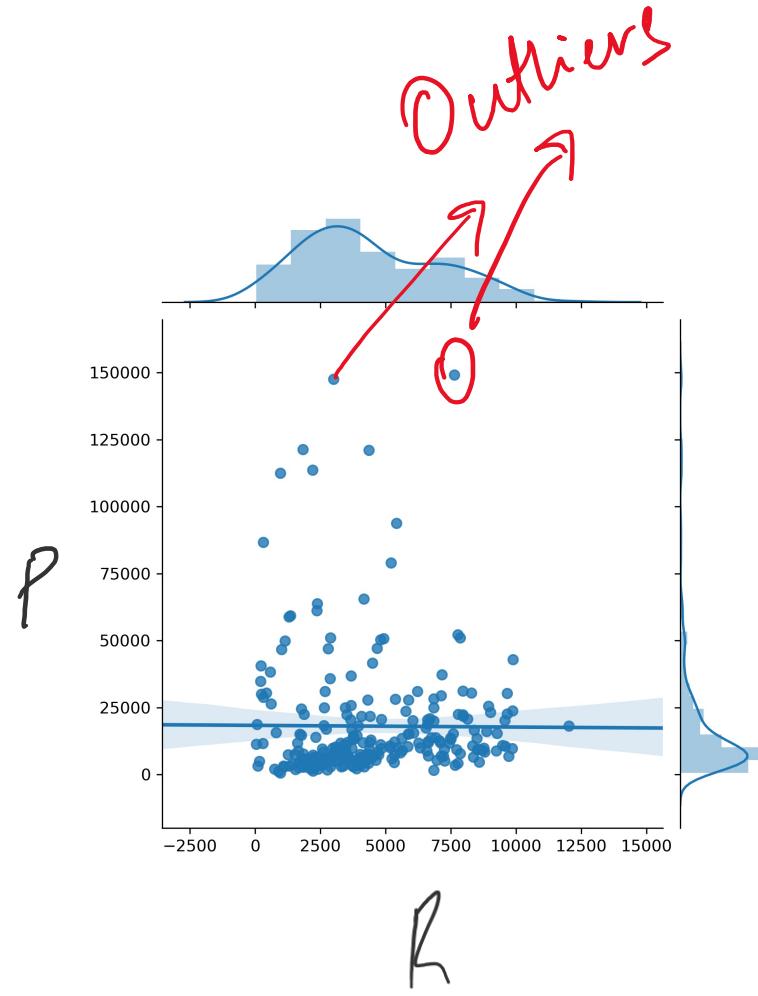
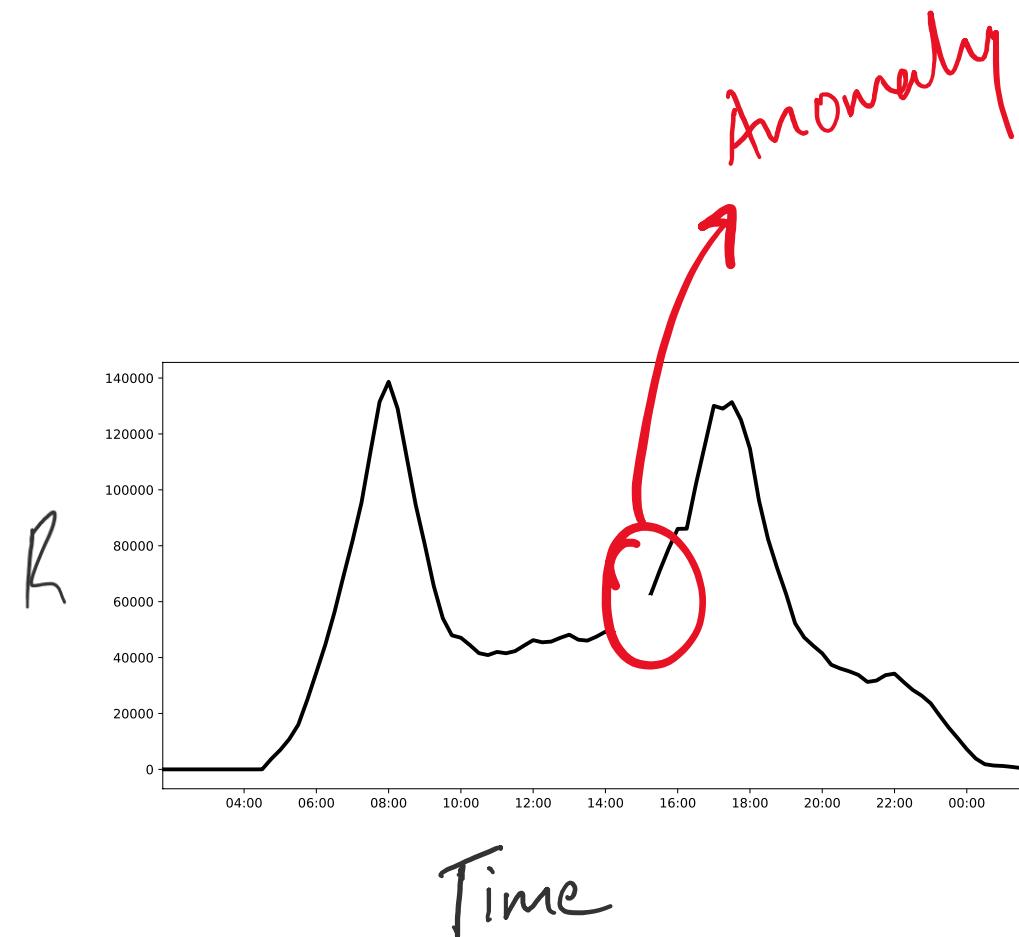
DS: What is a **sustainable** system?

# Data Understanding & Preparation

# Data Understanding



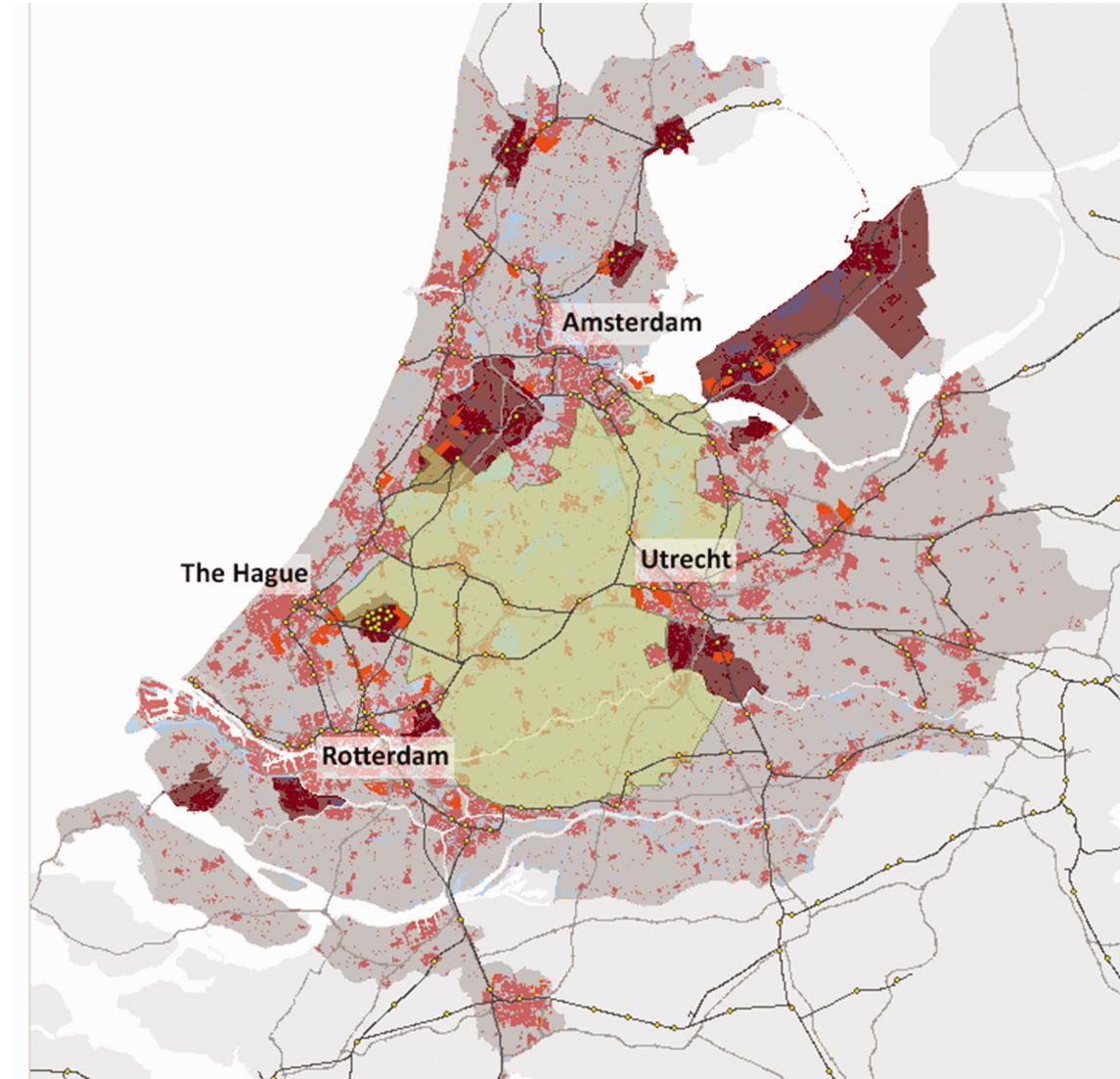
- Collect initial data
- Describe data
- Explore data
- Verify data quality  
carefully document  
problems and issues

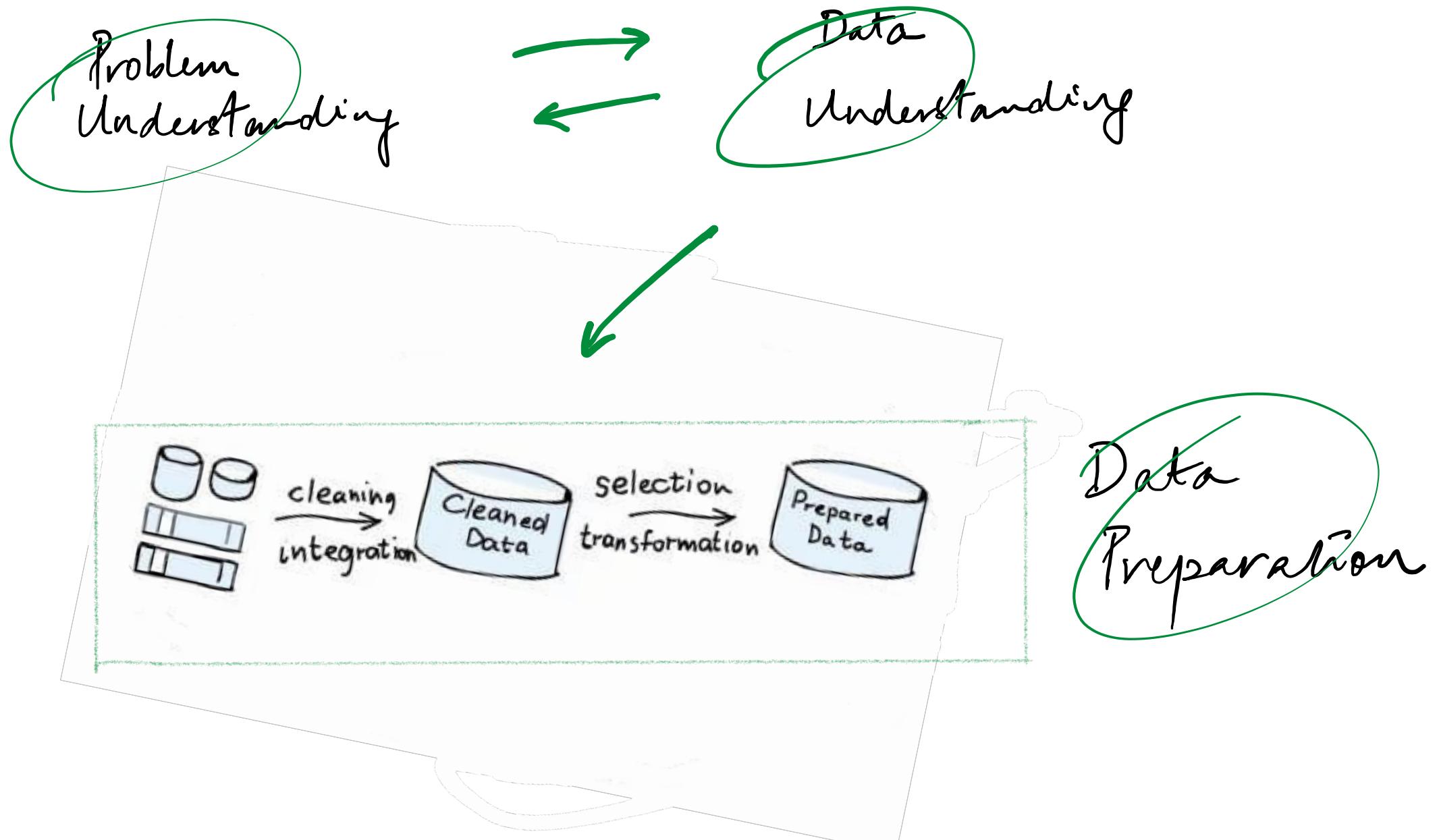


# Data Understanding

My DOs and DON'TS

- Do not economise in this step
  - Data has issues
  - Understand data to understand the domain, very important for modelling later
- Do not trust the stakeholders supplying data for quality
- Verify data is **correct, complete, coherent, unique, representative, independent, up-to-date and stationary**
- Was the data processed? Anonymised?  
Still useful?
- Understand anomalies and outliers





# Data Preparation

My DOs and DON'TS

- Automate this step as much as possible – new data / new case
- When merging sources, track data origin
- **Document everything!**  
Create a workflow
- Manage the stakeholders' expectations that these tasks will take roughly 50% of your time



Before you go for discussion..

# Why do we use Functional programming

- **Organization** -- As programs grow in complexity, having all the code live inside the main() function becomes increasingly complicated. A function is almost like a mini-program that we can write separately from the main program, without having to think about the rest of the program while we write it. This allows us to reduce a complicated program into smaller, more manageable chunks, which reduces the overall complexity of our program.
- **Reusability** -- Once a function is written, it can be called multiple times from within the program. This avoids duplicated code (“Don’t Repeat Yourself”) and minimizes the probability of copy/paste errors. Functions can also be shared with other programs, reducing the amount of code that must be written from scratch (and retested) each time.
- **Testing** -- Because functions reduce code redundancy, there’s less code to test in the first place. Also because functions are self-contained, once we’ve tested a function to ensure it works, we don’t need to test it again unless we change it. This reduces the amount of code we must test at one time, making it much easier to find bugs (or avoid them in the first place).
- **Extensibility** -- When we need to extend our program to handle a case it didn’t handle before; functions allow us to make the change in one place and have that change take effect every time the function is called.
- **Abstraction** -- In order to use a function, you only need to know its name, inputs, outputs, and where it lives. You don’t need to know how it works, or what other code it’s dependent upon to use it. This lowers the amount of knowledge required to use other people’s code (including everything in the standard library).

# For next class..



**Finish** Lab 01 to practice programming



**Submit** Homework 01 for peer review on Brightspace



**Check** Assignment 1 – due in **Week 3** on Friday at **2330**



**See** “To do before class” for every lecture (~ 1 hour of self study)



**Read** paper for **Discussion** session before every Friday



**Post** questions on the **Discussion** forum on Brightspace (especially on **Anaconda/Jupyter** section for this week)