

Introduction to *Urban* Data Science

Introduction-II

(EPA1316A)

Lecture 1

Trivik Verma



Adapted from the work of Sean Perez

Just before the break

- Introduction to the Course
- Tools - Python and Conda

Now..

- The Data Revolution
- (Geo-)Data Science
- Why Data Science?
- What is Data Science?

The data revolution

Exciting times to be a:

- Data Scientist
- Urban Planner
- Policymaker

The world is producing a lot of “**data**”...

Massive Data Revolution

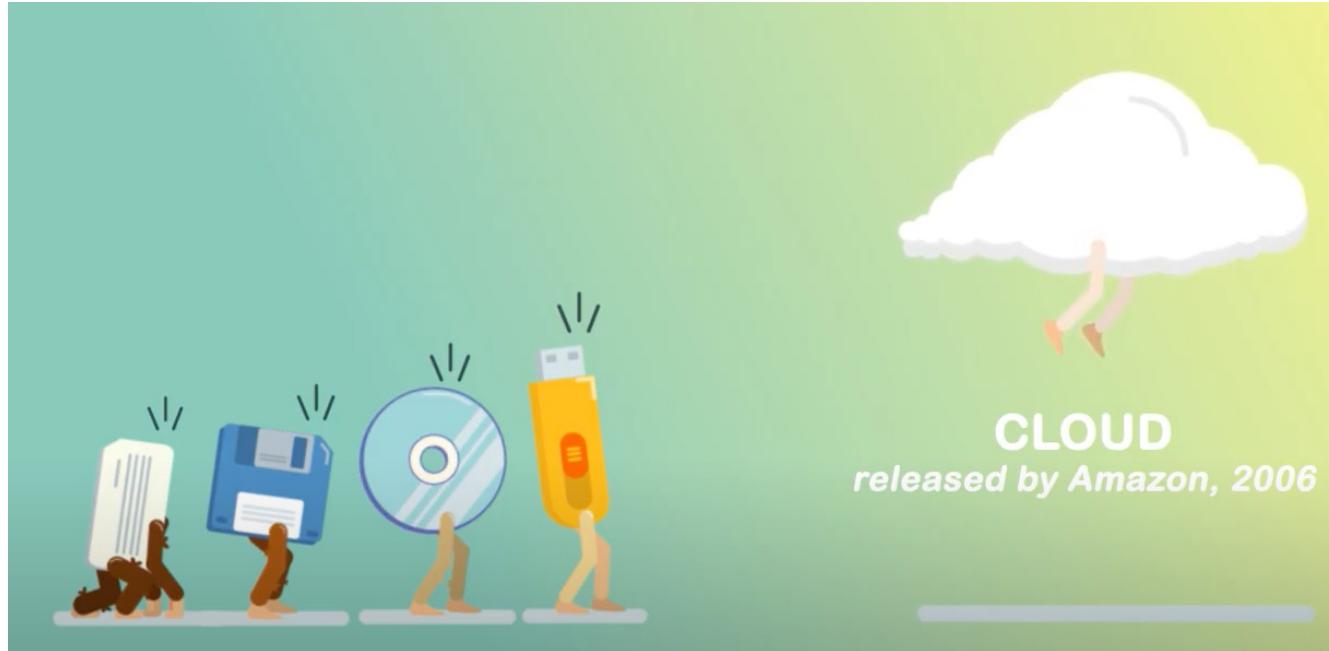
Quantification of phenomena through the systematic recording of data, “taking all aspects of life and turning them into data” ([Cukier & Mayer-Schoenberger](#))

Examples: credit transactions, public transit, tweets, facebook likes, spotify songs, etc.

Implications

- **Window** into human behaviour (this course)
- Opportunities for optimization of systems (Industrial IoT, planning systems...)
- Issues with **representation** and **privacy**
- ...

Why now?



Statistics

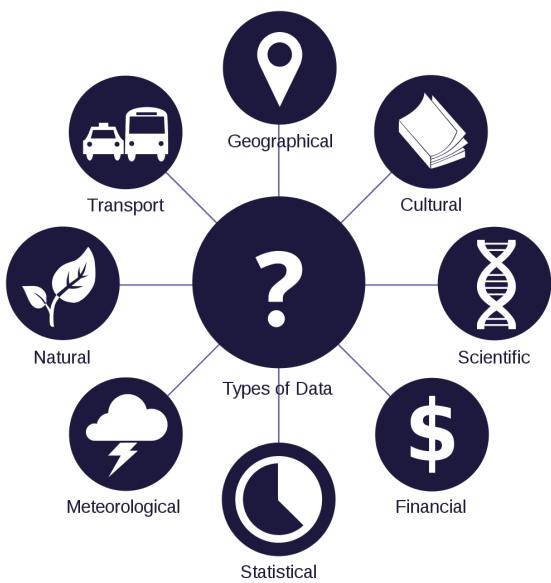
Machine
learning

What's
next?

- Massive Data generation
- Computing power
- R + Python
- Visualisation

“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”

At this point many people have said it..

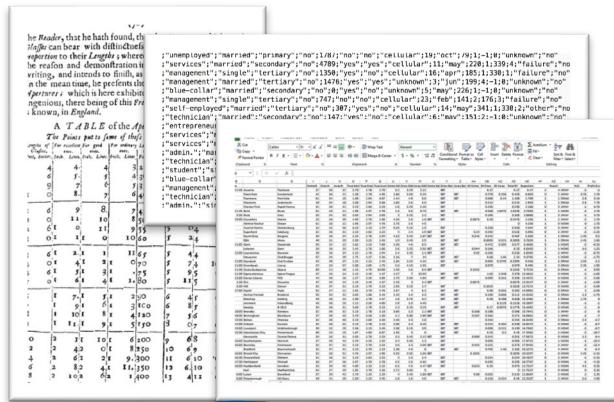


Examples: credit transactions, public transit usage, tweets, census, mobility and migration, etc.

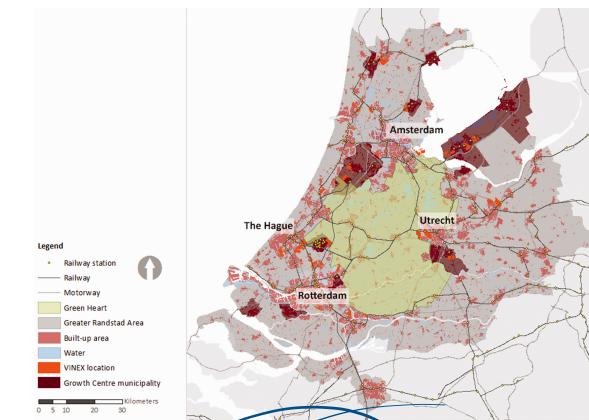
A screenshot of Microsoft Excel displaying a large dataset. The columns are labeled "Length" and "Age". The data includes various entries such as "Geographical", "Cultural", "Scientific", "Financial", "Statistical", "Natural", "Meteorological", and "Transport". The table has approximately 20 rows of data, with some rows containing numerical values and others containing text descriptions.

Formats: CSV, Excel, JSON, Shapefiles

Now, data alone is not very valuable



Data



Action

Methods, tools and techniques to turn data into
actionable knowledge (hence this lesson!)

Class Quiz

Can you think of a real-world context where data and statistics are being used to make a difference? And how?

- Turn to your neighbour and discuss for two minutes
- Then I may ask you to summarise your discussion



Data Science

Statistics + ...

- **Computational** tools → Programming (hence this course's labs and homework!)
- **Communication** skills → “Story telling ” (hence this course's assignments)
- **Domain** expertise → Theories about why the data are the way they are (hence the rest of your degree)

Some examples...

Emmy-winning US TV Shows



Police Detective TV Dramas



Critically Acclaimed Witty TV Shows



Free Online Dating | OkCupid - Mozilla Firefox (Private Browsing)

Free Online D... https://www.okcupid.com

Have an account? Sign in

okcupid

Join the best free dating site on Earth.

I am a

Woman

Continue



Signing up takes two minutes and is totally free.



Our matching algorithm helps you find the right people.



iOS or Android?
You can take us to go.

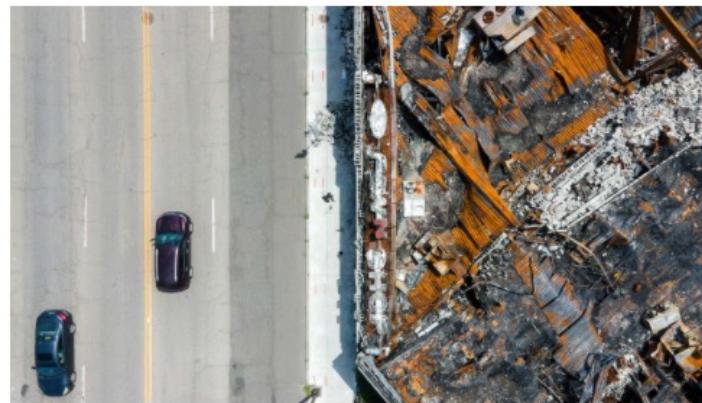
The (Geo-)Data Revolution

The Global Picture: Urban Inequalities

Rising Seas



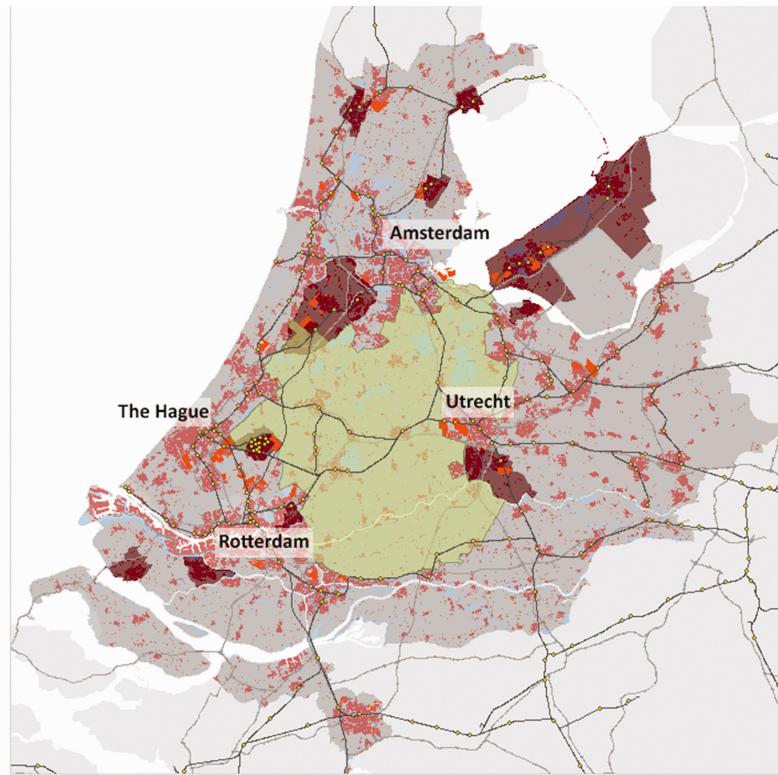
Economic
Inefficiencies



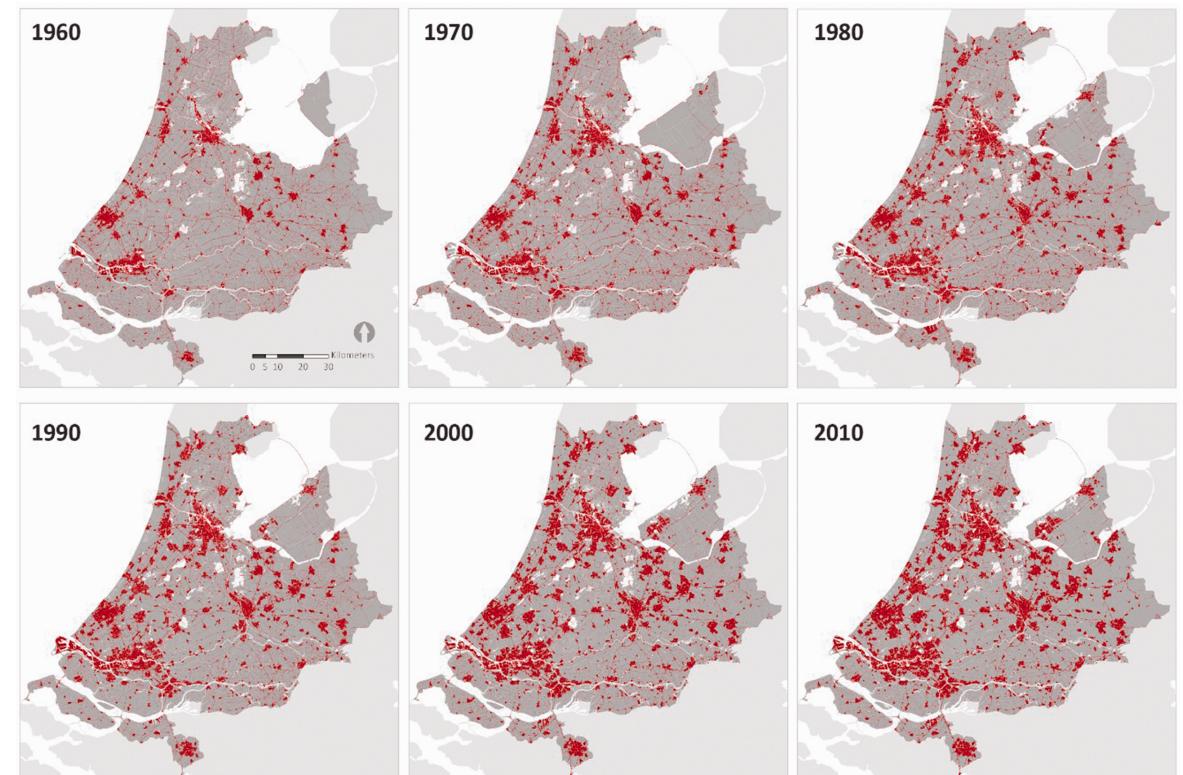
Growing Energy
Demand



The Local Picture: Randstad



Urbanisation of the Greater Randstad Area 1960-2010



Space is important!

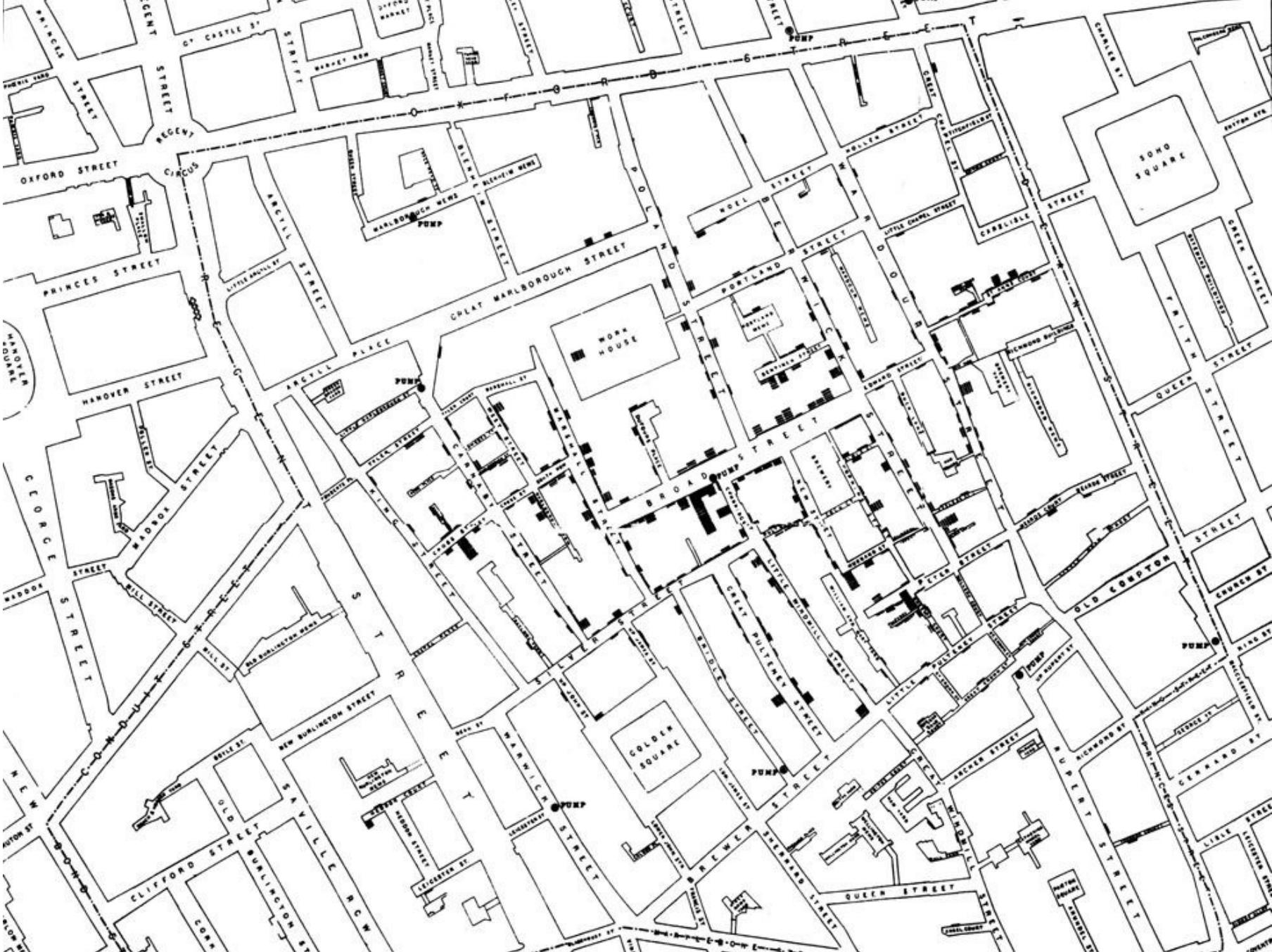
- The space around us
- Thus, the geography of our location
- Geolocated data

The world is producing a lot of geo-located “**data**”...

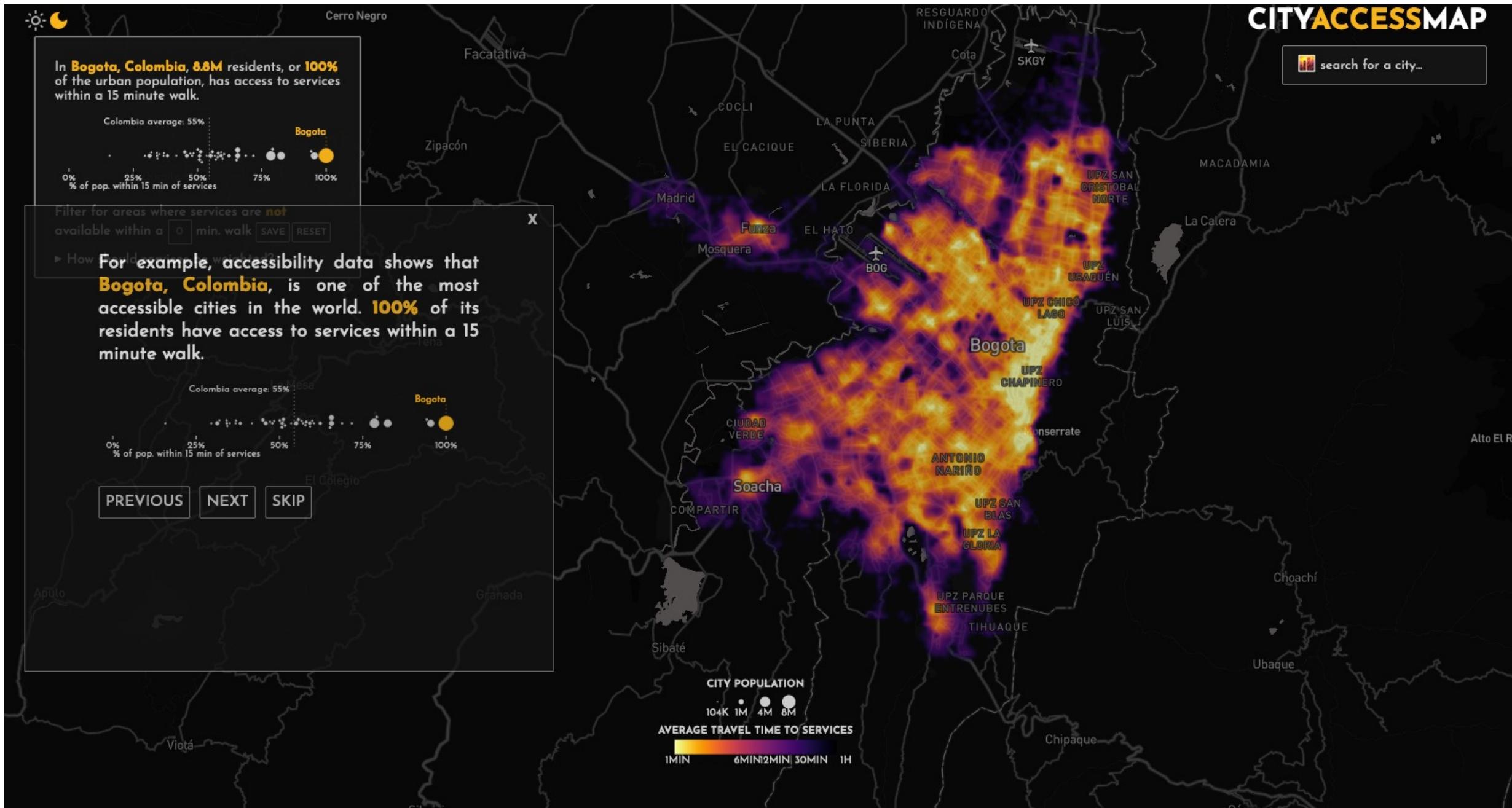
(Geo-)Data Science

(Geo-)Data Science

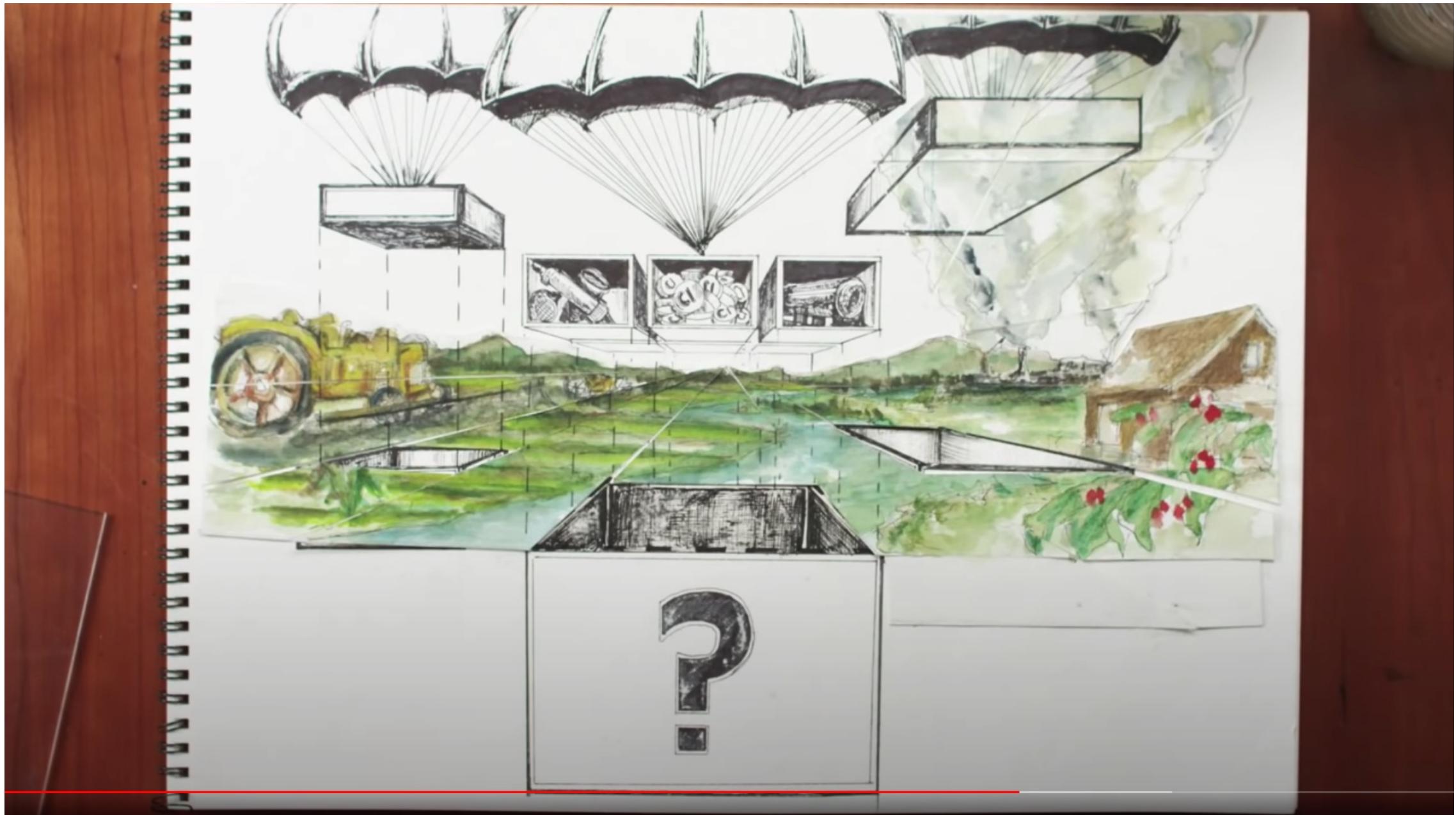
- A (very) large portion of all these new data are inherently **geographic** or can be traced back to some location over space.
- Spatial is special.
- Some of the methods require an explicitly spatial treatment -> (Geo-)Data Science
- Some examples...



Map of the book "On the Mode of Communication of Cholera" by John Snow, originally published in 1854 by C.F. Cheffins, Lith., Southampton Buildings, London, England.



Why Data Science?



History

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



© Trivik Verma. All rights reserved.

History (cont)

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

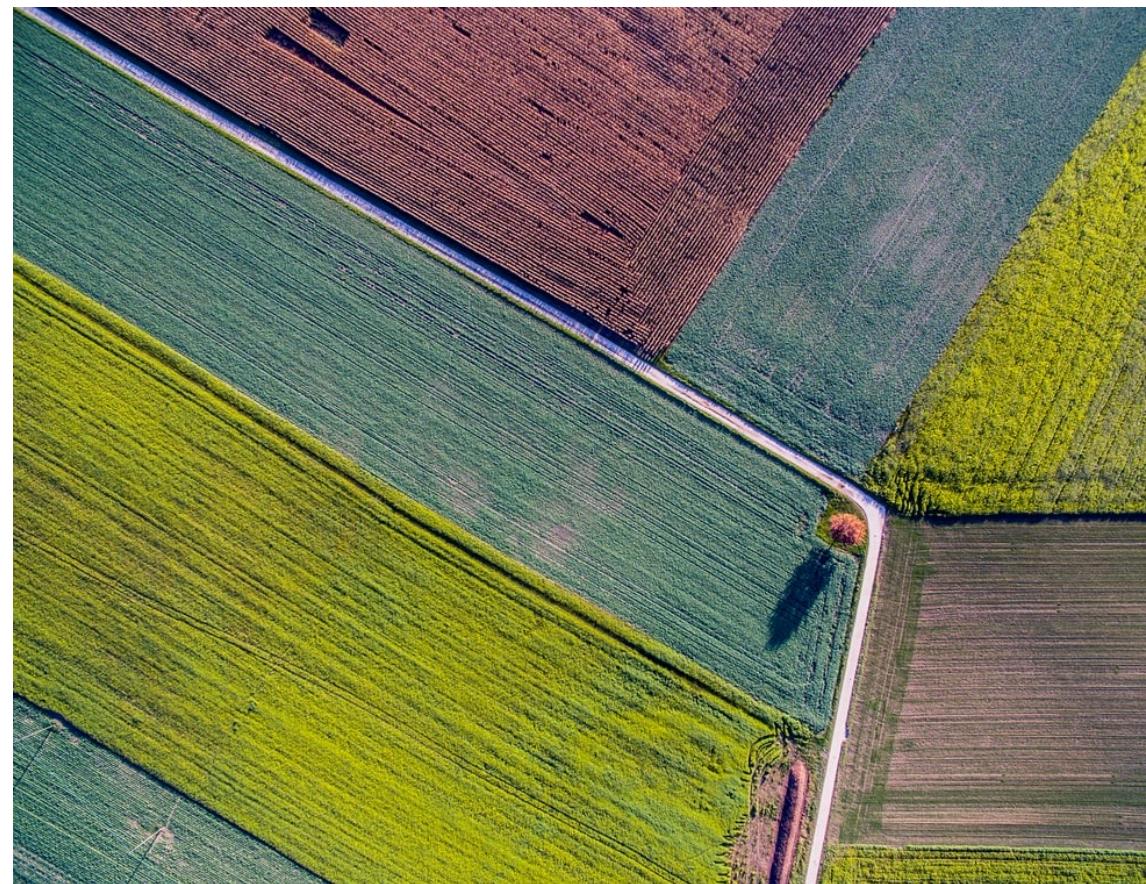


Photo by [jean wimmerlin](#) on [Unsplash](#)

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



Photo by [Frank Chou](#) on [Unsplash](#)

History (cont)

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

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

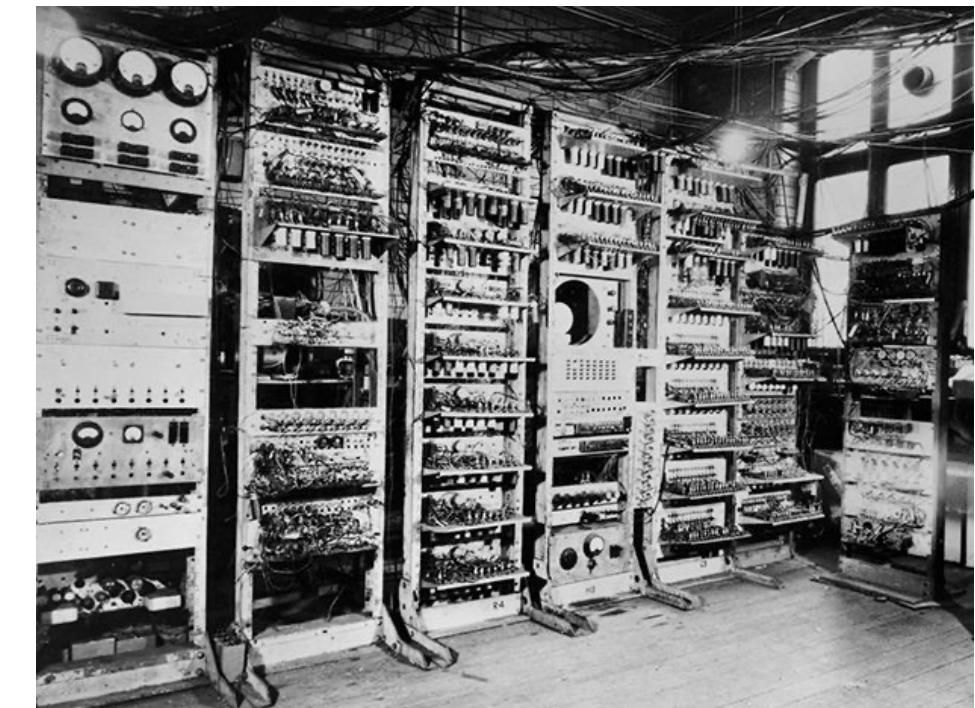
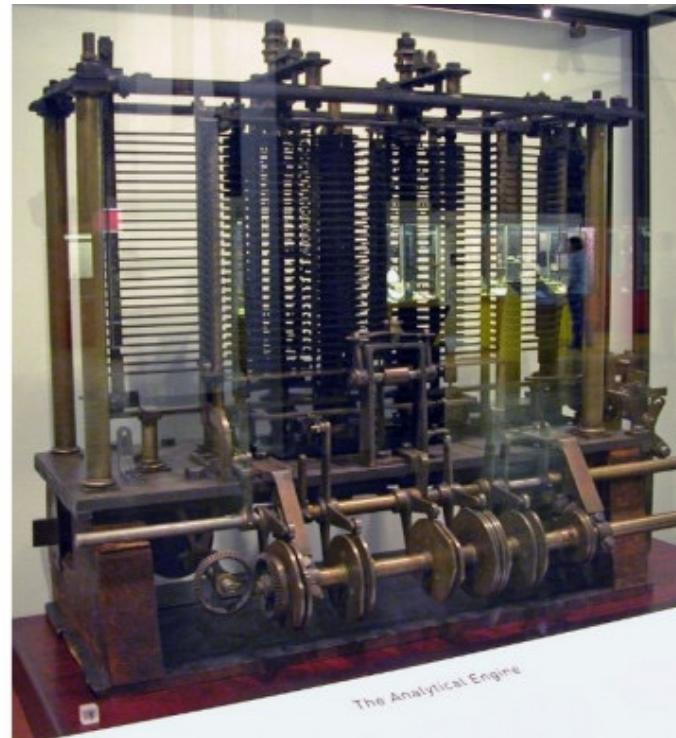
Stewart, I. (2012). *In pursuit of the unknown: 17 equations that changed the world*. Basic Books.

History (cont)

About a hundred years ago: computational approaches



Scanned from *The Calculating Passion of Ada Byron* by Joan Baum.
Analytical Machine [Wikimedia Commons](#)



SSPL/Getty Images The Manchester Mark I at Manchester University's Computer Machine Laboratory.

What is Data Science?

what my friends think I do



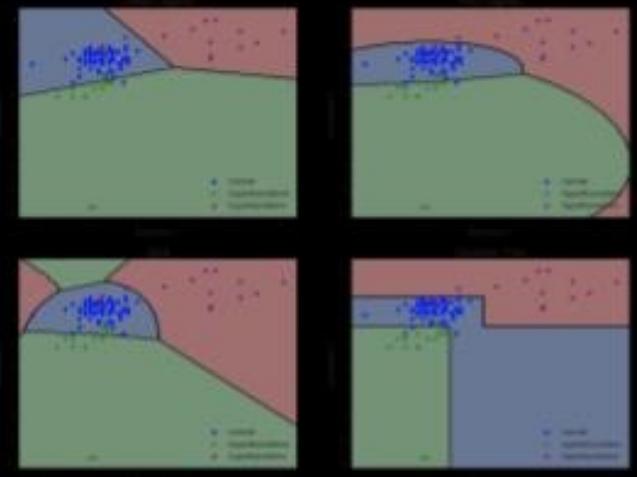
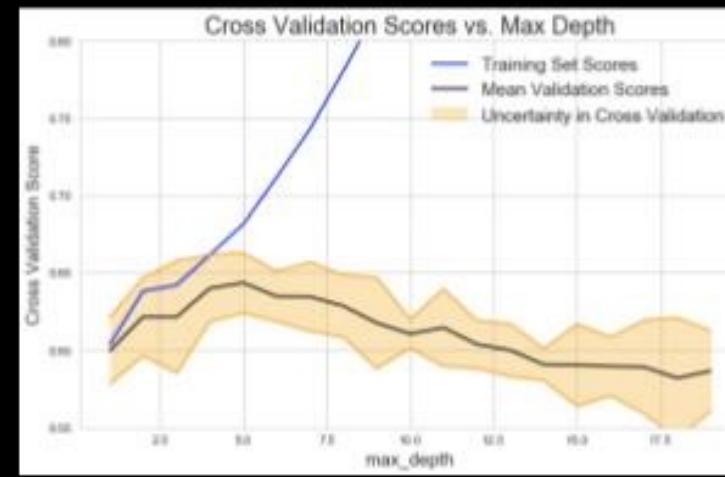
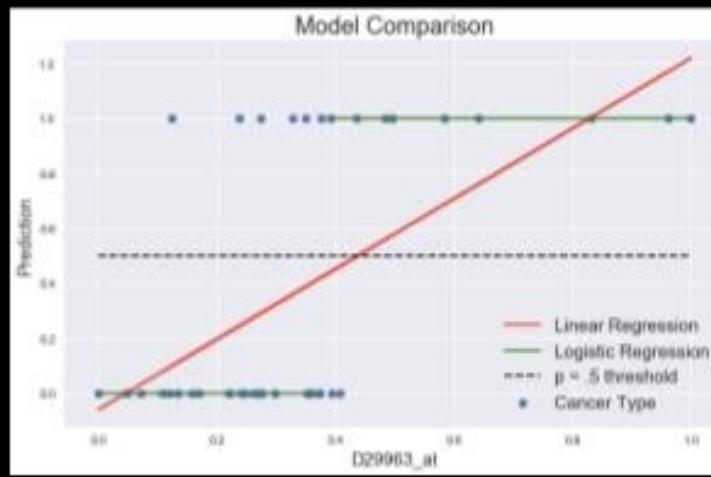
what my family thinks I do

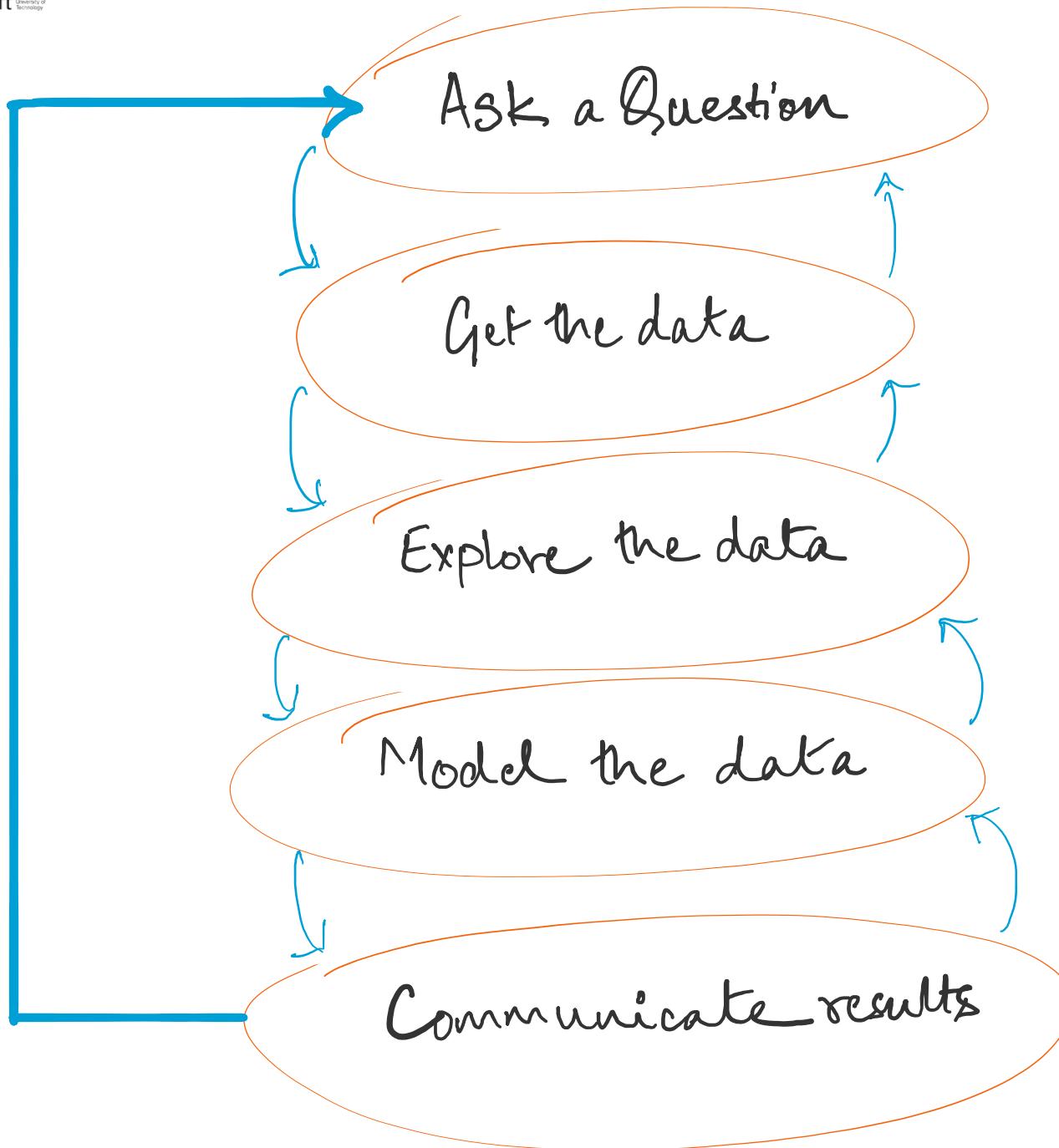


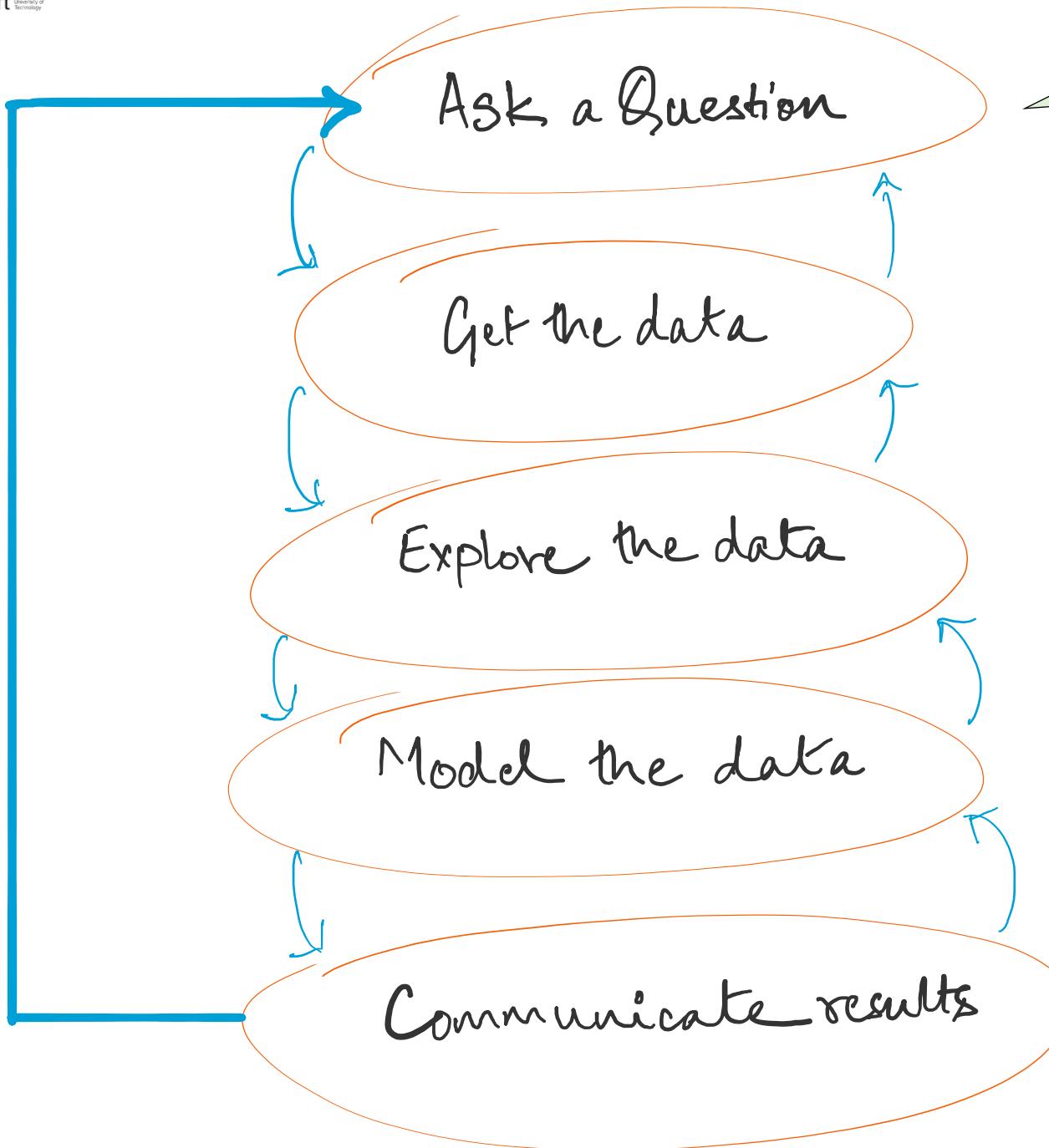
what society thinks I do



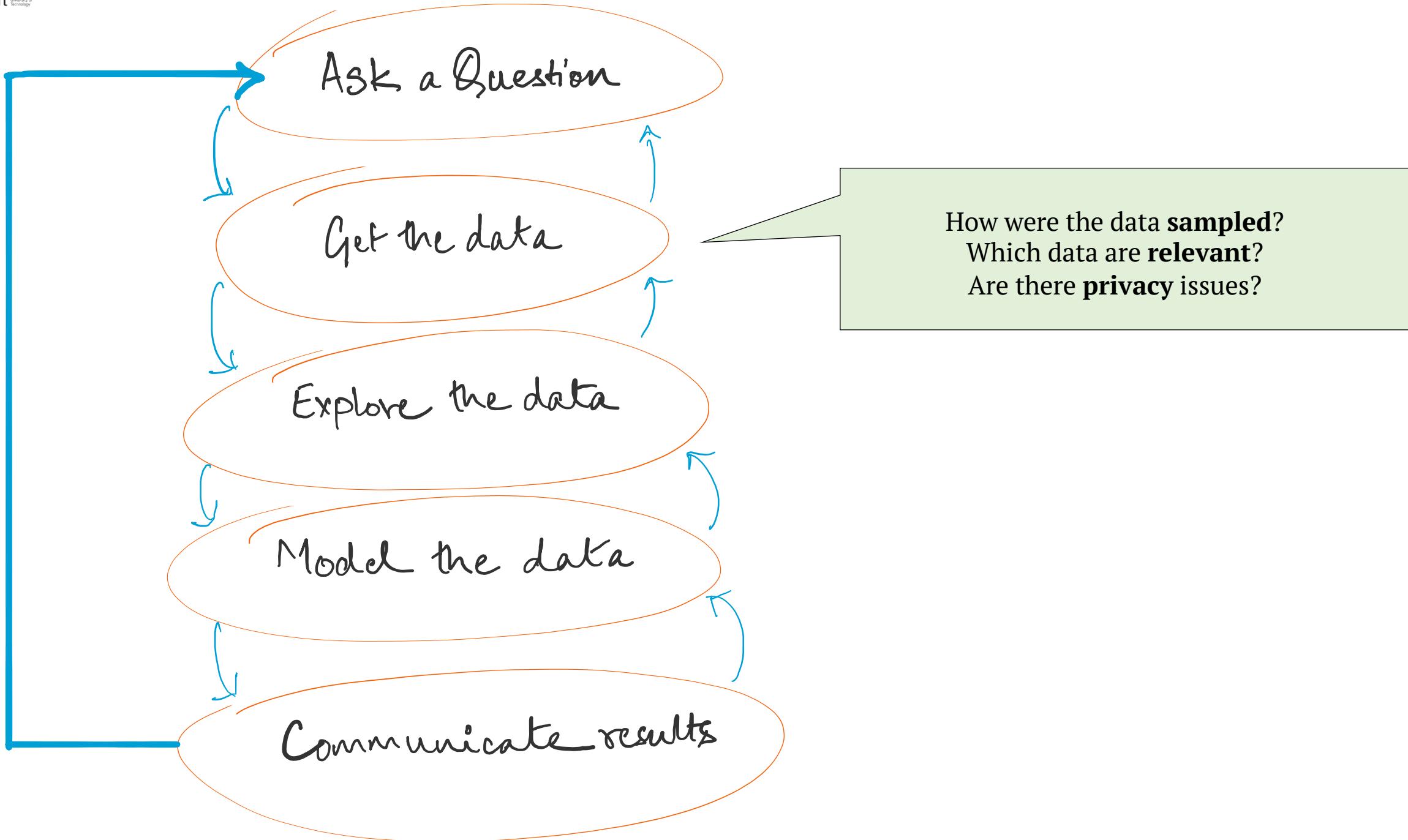
what I actually (will) do in Data Science 1

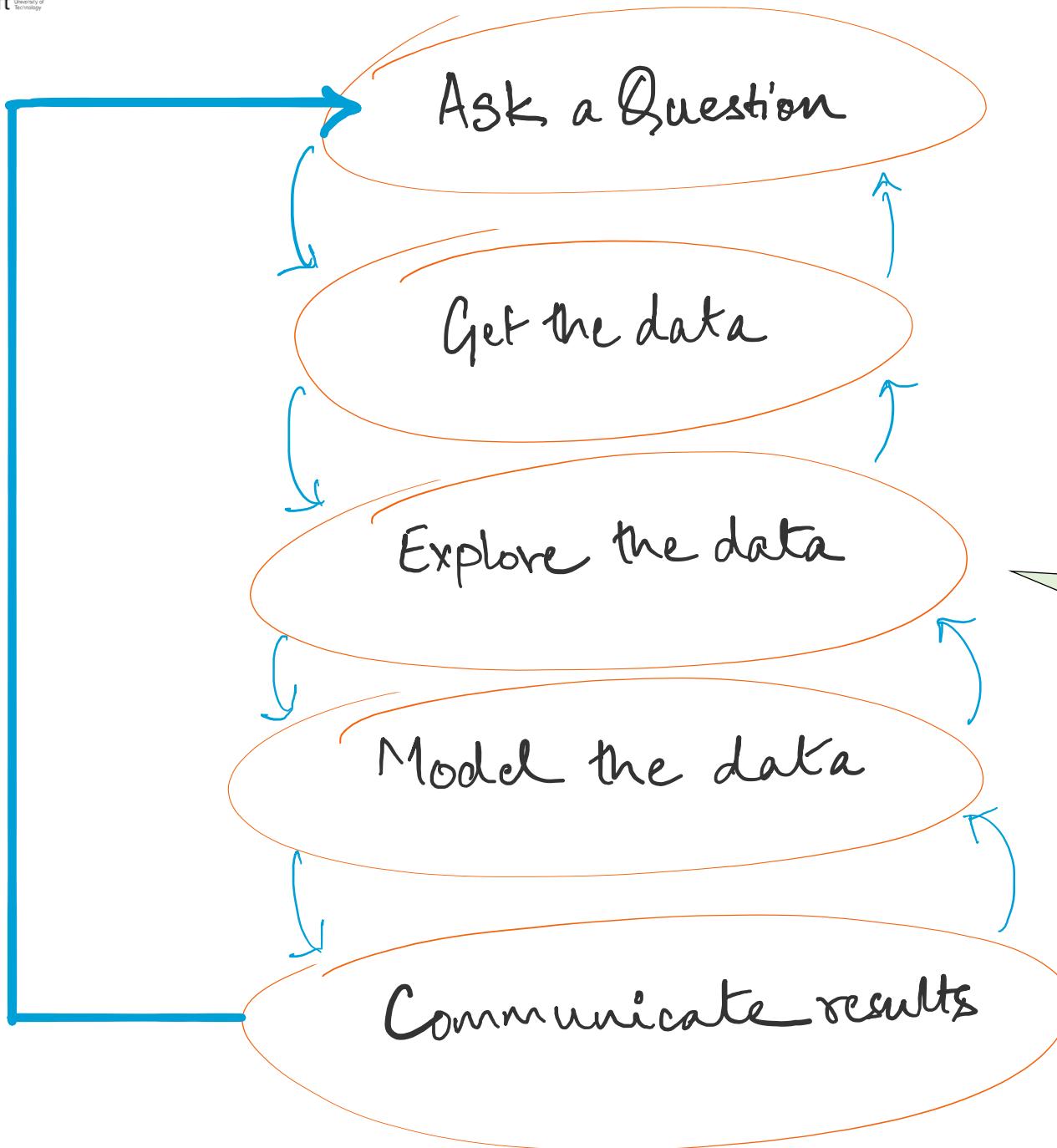


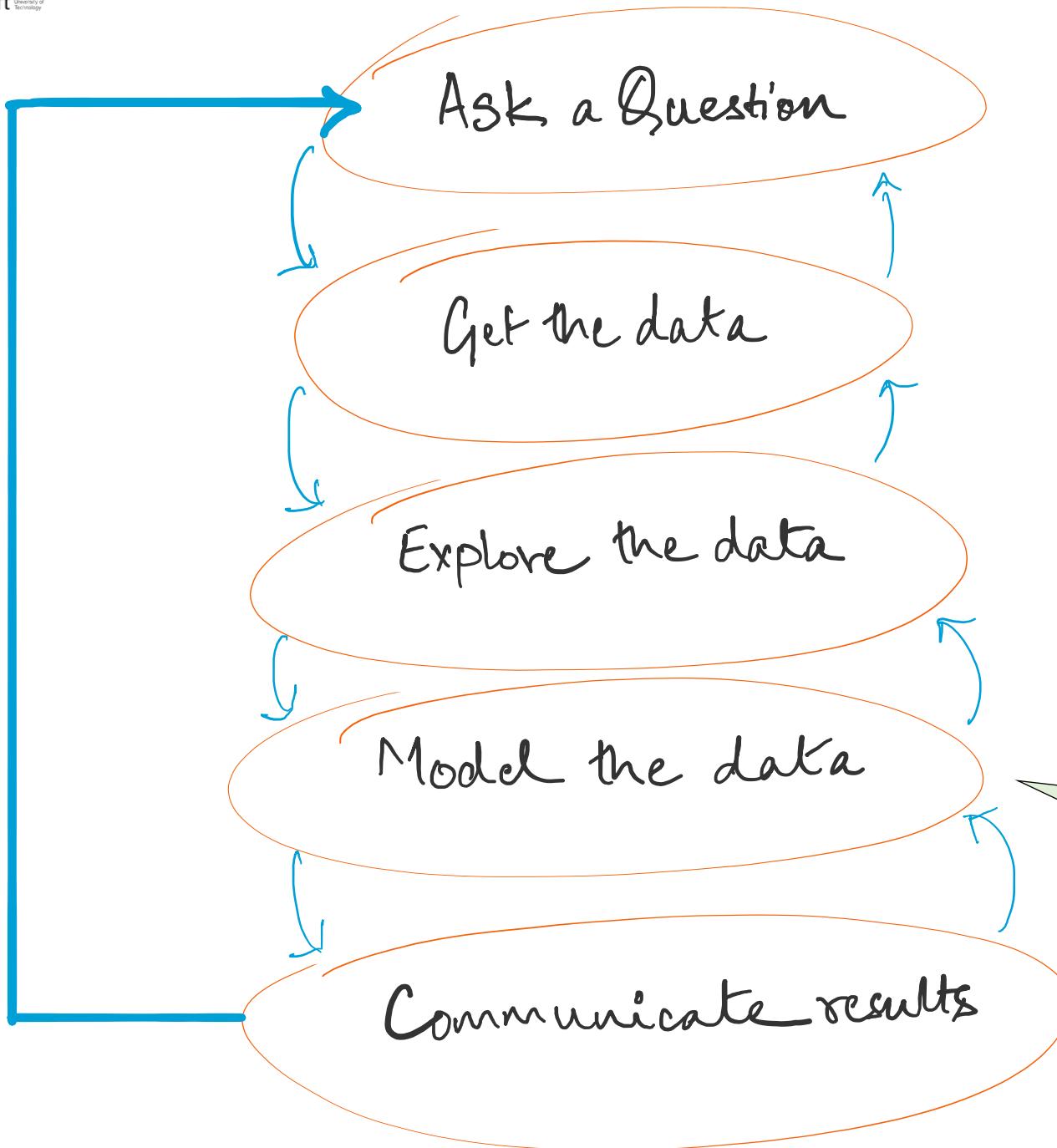


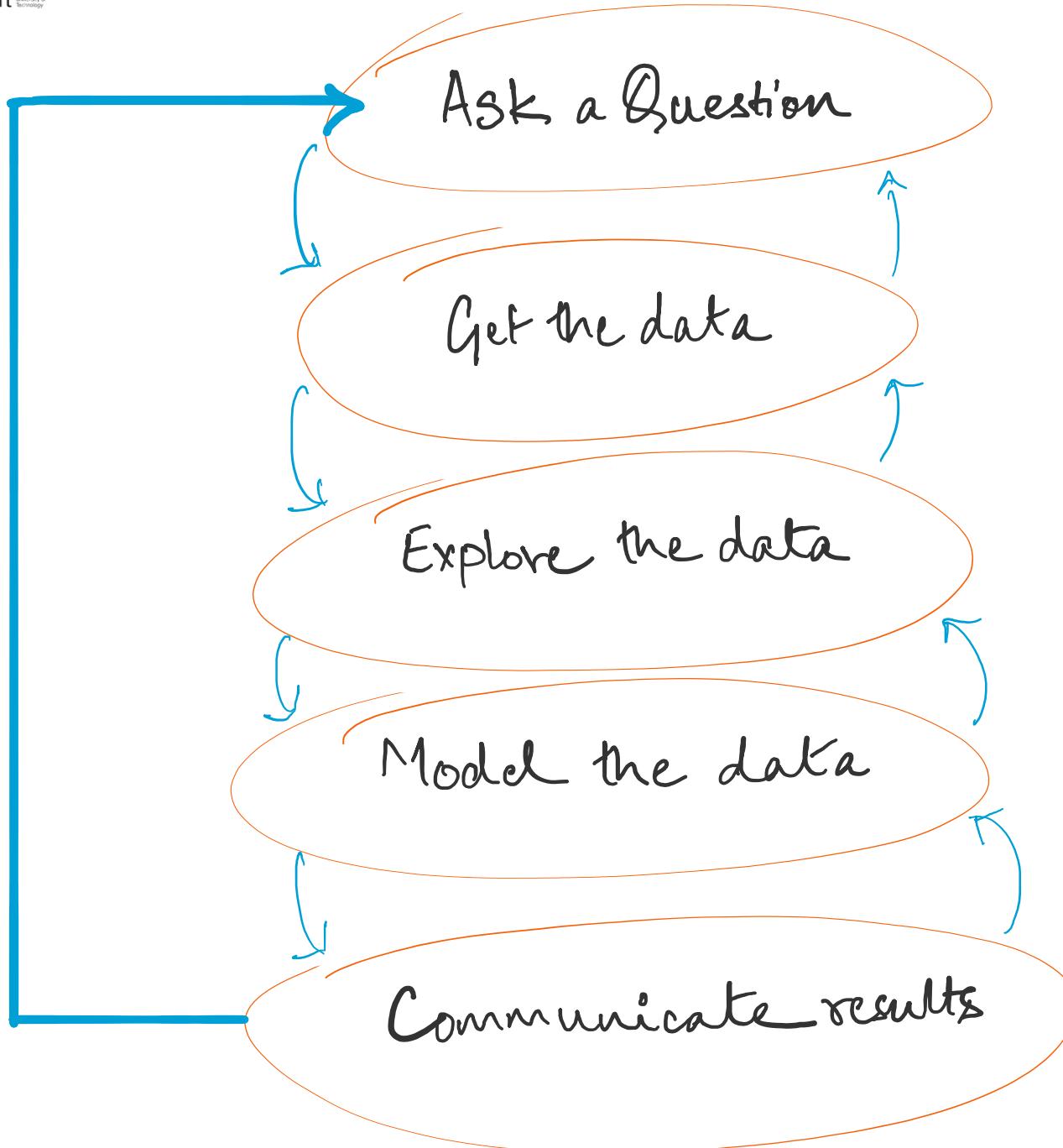


What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?



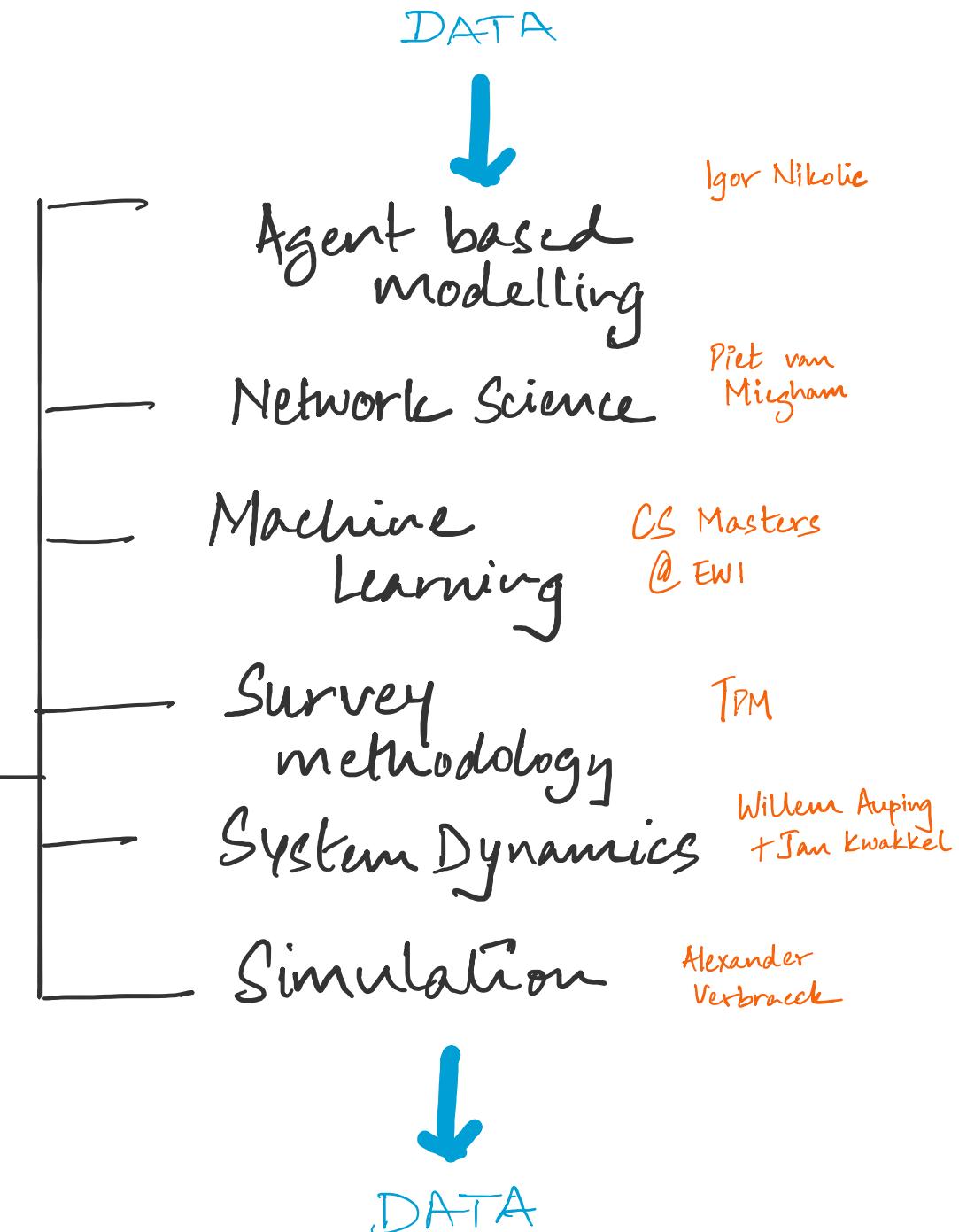
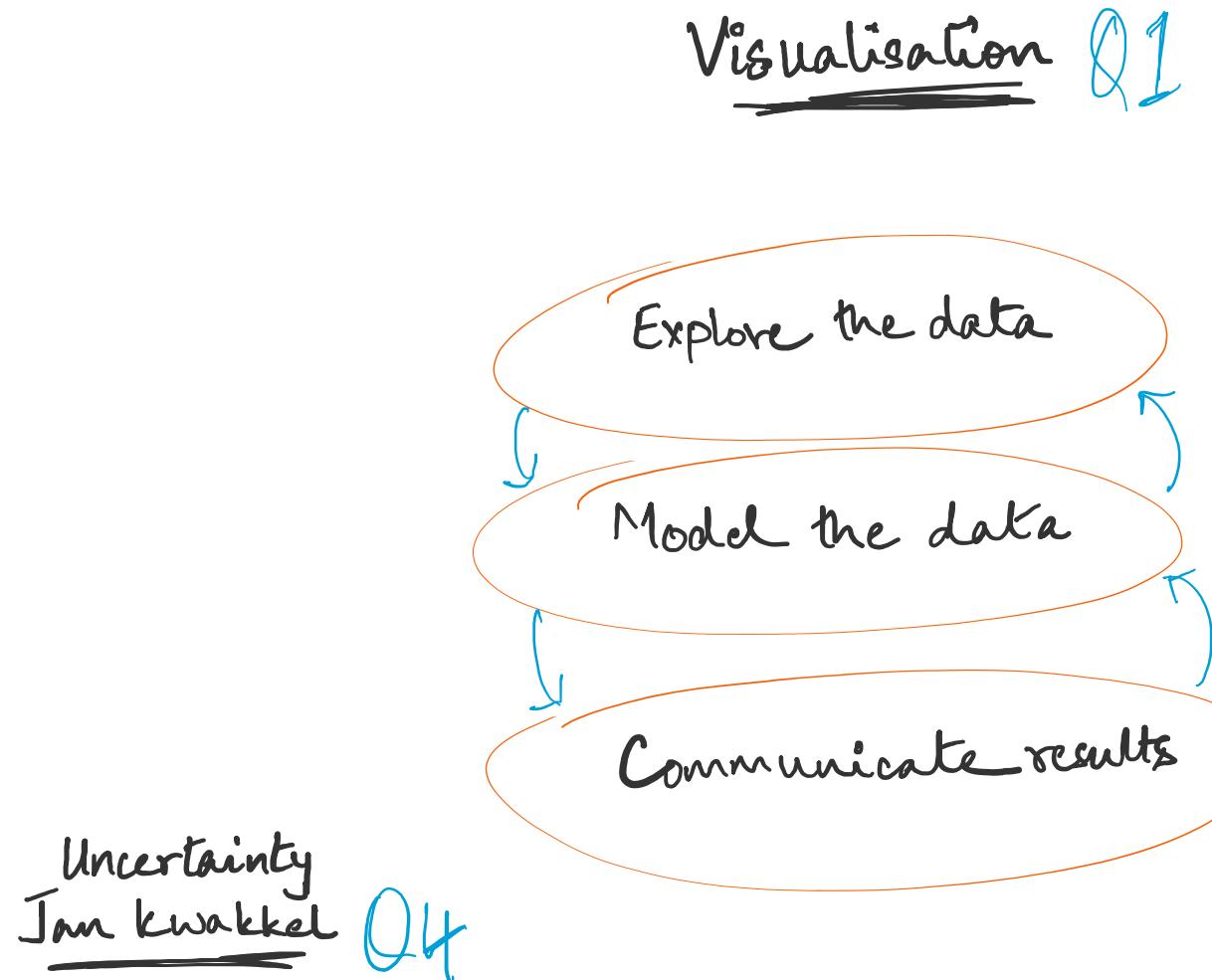






What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

EPA Programme



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!

Before you start Lab 0..

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 Labs to practice programming



Complete Homework and review your peers' work



Check Assignment contents and due date



See "To do before class" for next lecture (~ 1 hour of self study)



Read paper for **Discussion** session before next week (~ 1 hour)



Post questions on the **Discussion** forum on Brightspace