TU Delft

Spatial Data Science

Introduction

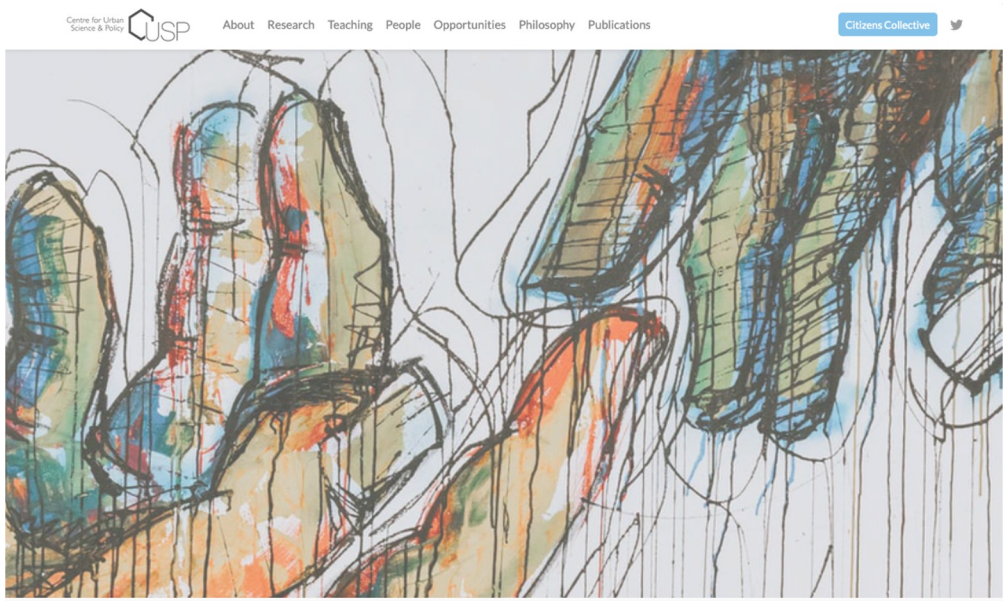
(EPA122A)

Lecture 1

Trivik Verma

Den Haag, NL

Cover Image by Bichi Zhang



TU Delft | TPM

TU Delft BK Bouwkunde

Space and Place

Cities & Social Justice

Understand the complex nature of urban spaces in transformation

Understand how inequalities intersect with space

Propose guidelines and develop tools to support public participation and citizen action

Identify inequalities associated with lack of recognition and legitimacy

Intersectionality

Democratisation of urban design, planning & policy

* Co-led by Dr Juliana Goncalves and Dr Trivik Verma

Teaching Support

Philip Mueller
2nd year EPA

Laura van Geene
2nd year IE

Nachiket Kondhalkar
2nd year EPA

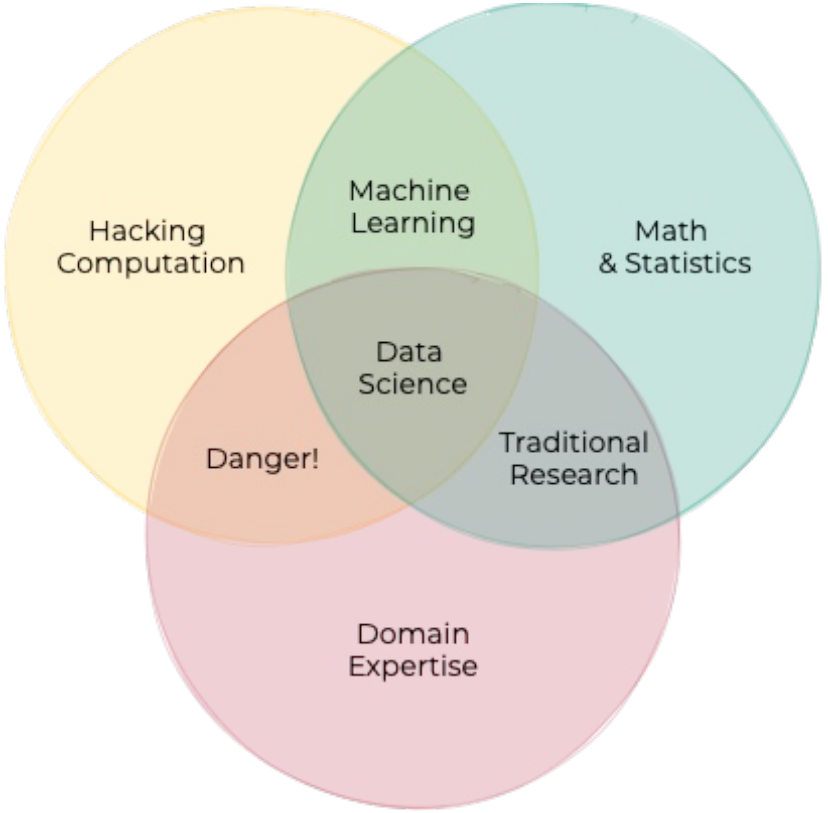
Today

- Introduction to the Course
- Tools - Python and Conda
- Post break, Intro part II

Introduction to the Course

More stats than a GIS course… more GIS than a stats course

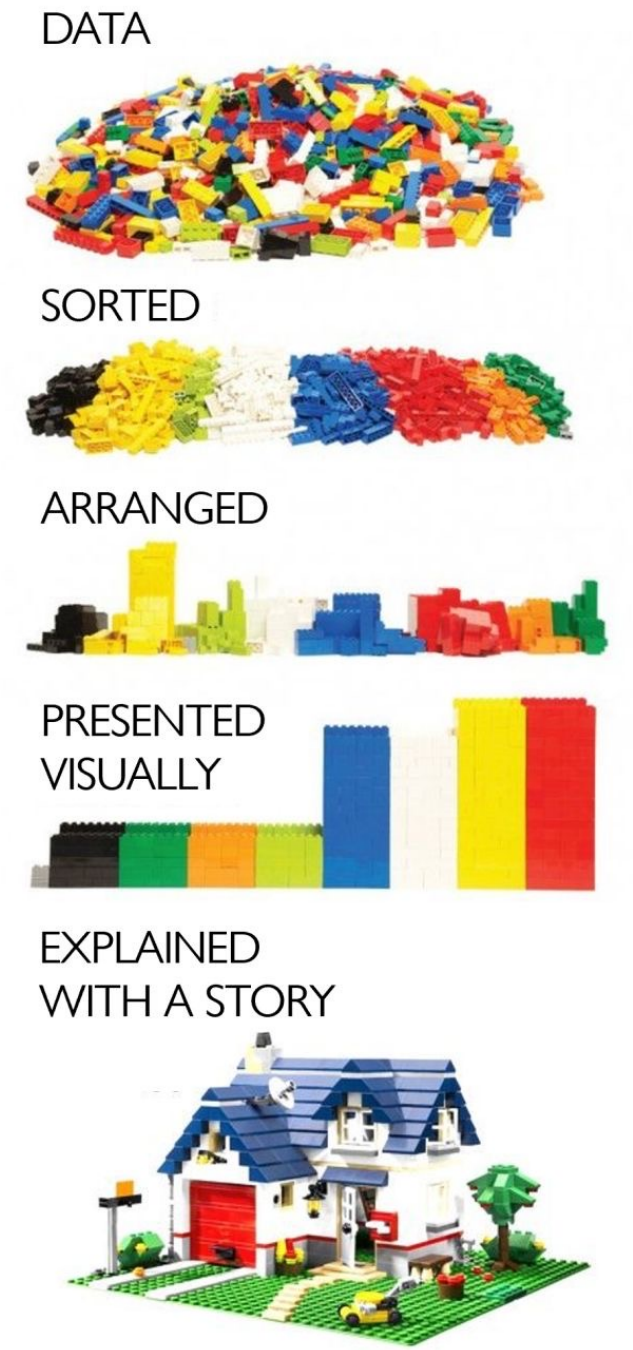
With a few twists!



After this course

You will be able to…

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



Can you find the source for me?

Philosophy

- (Lots of) **methods** and techniques
  - General overview
  - Intuition
  - Very little math
  - Lots of ways to continue learning more
- Emphasis on the **application** and **use**
- Close connection to “**real world**” applications

Format – Ways of Working

Seven weeks of:

- **Prep. Materials**: videos, podcasts, articles… 1h. approx. (most recommended!)
- **2x 1h. Lecture**: interaction, concepts, methods, examples
- **2x 2h. Computer labs**: hands-on, application of concepts, Python (highly *employable*)
- **Extra material**: how to go beyond the minimum*

*(not necessary for the course but useful in life)

Content

- **Weeks 1-4**: “big picture” lectures + introduction to computational tools (learning curve) + lots and lots of data + lots of visualisation
- **Weeks 5-7**: lots of spatial, network and machine learning concepts
- **Weeks 8-10**: wrap up + prepare an awesome final project in groups (opportunities to follow up with internships in the ***CUSP*** lab)

An overview of the course

- **Weeks 1-4**: “big picture” lectures + introduction to computational tools (learning curve) + lots and lots of data + lots of visualisation
- **Weeks 5-7**: lots of spatial, network and machine learning concepts
- **Weeks 8-10**: wrap up + prepare an awesome final project in groups (opportunities to follow up with internships in the ***CUSP*** lab)

Week	Lecture	Topic	Learning Goals	Python Libraries	Labs [1]	Assessment [2][3]
W1	L1	Introduction to Spatial Data Science		Anaconda and Jupyter, Numpy	Lab 0 + 1	
	L2	Spatial and Urban Data			"	
W2	L3	Data Grammar	Obtain, Discuss	Pandas, Seaborn	Lab 2	
	L4	Data Engineering	Manipulate, consolidate	Pandas	"	
W3	L5	EDA and Visualisation	Discuss, manipulate and Consolidate	Geopandas, Matplotlib, Rasterio	Lab 3	Assignment 1
	L6	Geo-Visualisation	Interpret		"	
W4	L7	Networks and Spatial Weights	Describe, Analyse	Networkx, Osmnx, Pysal	Lab 4	
	L8	Exploratory Spatial Data Analysis	Describe, Analyze		"	
W5	L9	Machine Learning for Everyone	Apply	Sklearn, Scipy, Statsmodels	Lab 5	Assignment 2
	L10	Anatomy of a Learning Algorithm	Infer		"	
W6	L11	Clustering	Apply	Pysal, Sklearn-Cluster	Lab 6	
	L12	Dimensionality Reduction	Apply		"	
Winter Break						

Schedule

	Week 46	Week 47	Week 48	Week 49	Week 50	Week 51	Week 52	Week 1	Week 2	Week 3-4	Week 5
	L1: Introduction L2: Spatial and Urban Data 13-17 Nov	L3: Data Grammar L4: Data Engineering 20-24 Nov	L5: EDA and Visualisation L6: Geo-Visualisation 27 Nov- 1 Dec	L7: Networks and Spatial Weights L8: Exploratory Spatial Data Analysis 4-8 Dec	L9: Machine Learning for Everyone L10: Anatomy of a Learning Algorithm 11-15 Dec	L11: Clustering L12: Dimensionality Reduction 18-22 Dec	25-29 Dec	1-5 Jan	L13: Spatial Density Estimation L14: Responsible Data Science 8-12 Jan	15 - 26 Jan	29 Jan - 2 Feb
Mo	Assignment 1 Release Start of the week		Assignment 2 Release Start of the week	Final project release Start of the week	Assignment 3 Release Start of the week						
Tu			Assignment 1 Deadline: 18:00		Assignment 2 Deadline: 18:00				Assignment 3 Deadline: 18:00		
Wed	Lecture 1 13:15 - 15:00 Trivik Lab 1 TAs 15:00 - 17:00	Lecture 3 13:15 - 15:00 Trivik Lab 2 TAs 15:00 - 17:00	Lecture 5 13:15 - 15:00 Trivik Lab 3 TAs 15:00 - 17:00	Lecture 7 13:15 - 15:00 Trivik Lab 4 TAs 15:00 - 17:00	Lecture 9 13:15 - 15:00 Trivik Lab 5 TAs 15:00 - 17:00	Lecture 11 13:15 - 15:00 Trivik Lab 6 TAs 15:00 - 17:00	WINTER BREAK		Lecture 13 13:15 - 15:00 Trivik Lab 7 TAs 15:00 - 17:00		
Th											
Fr	Lecture 2 13:15 - 15:00 Trivik Lab 1 TAs 15:00 - 17:00	Lecture 4 13:15 - 15:00 Trivik Lab 2 TAs 15:00 - 17:00	Lecture 6 13:15 - 15:00 Trivik Lab 3 TAs 15:00 - 17:00	Lecture 8 13:15 - 15:00 Trivik Lab 4 TAs 15:00 - 17:00 Assignment 1 summative feedback Released by: 18:00	Lecture 10 13:15 - 15:00 Trivik Lab 5 TAs 15:00 - 17:00 Group formation Deadline: 23:30	Lecture 12 13:15 - 15:00 Trivik Lab 6 TAs 15:00 - 17:00 Assignment 2 summative feedback Released by: 18:00 Selection of final project Consult and check in with TAs	WINTER BREAK		Lecture 14 13:15 - 15:00 Trivik Lab 7 TAs 15:00 - 17:00 Assignment 3 summative feedback Released by: 18:00 Problem formulation and Data checks Consult and check in with TAs		Final report submission Deadline: 17:00

Logistics - Website

Course Material

- https://trivikverma.github.io/spatial-data-science/_index.html#

Lectures and Labs

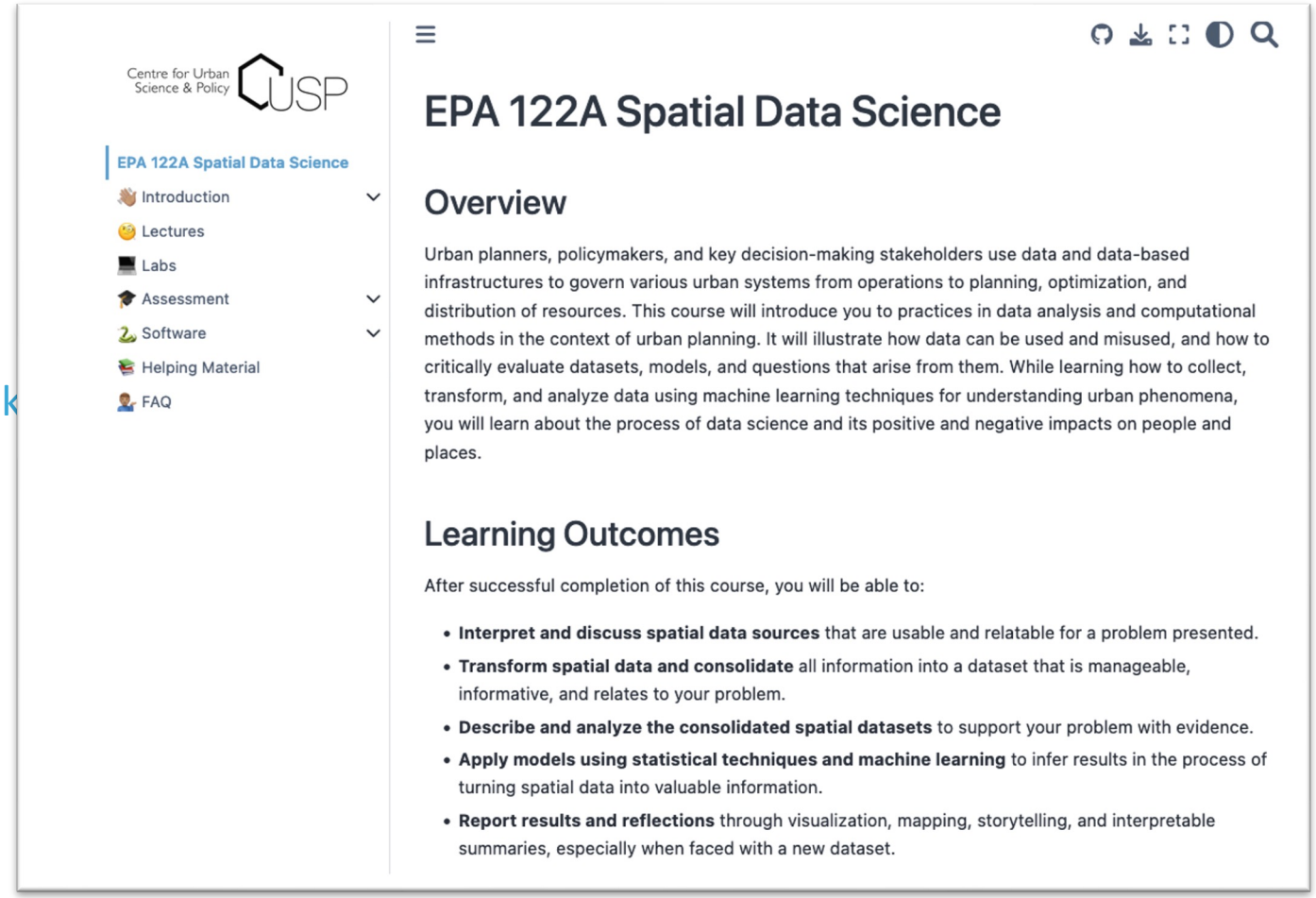
- Physical Space: Check Timetable every week

Communication Channels

- With instructors and TA: Email
- Announcements on Brightspace

Submissions, Groups, Grades and Feedback

- Brightspace



Self-directed learning

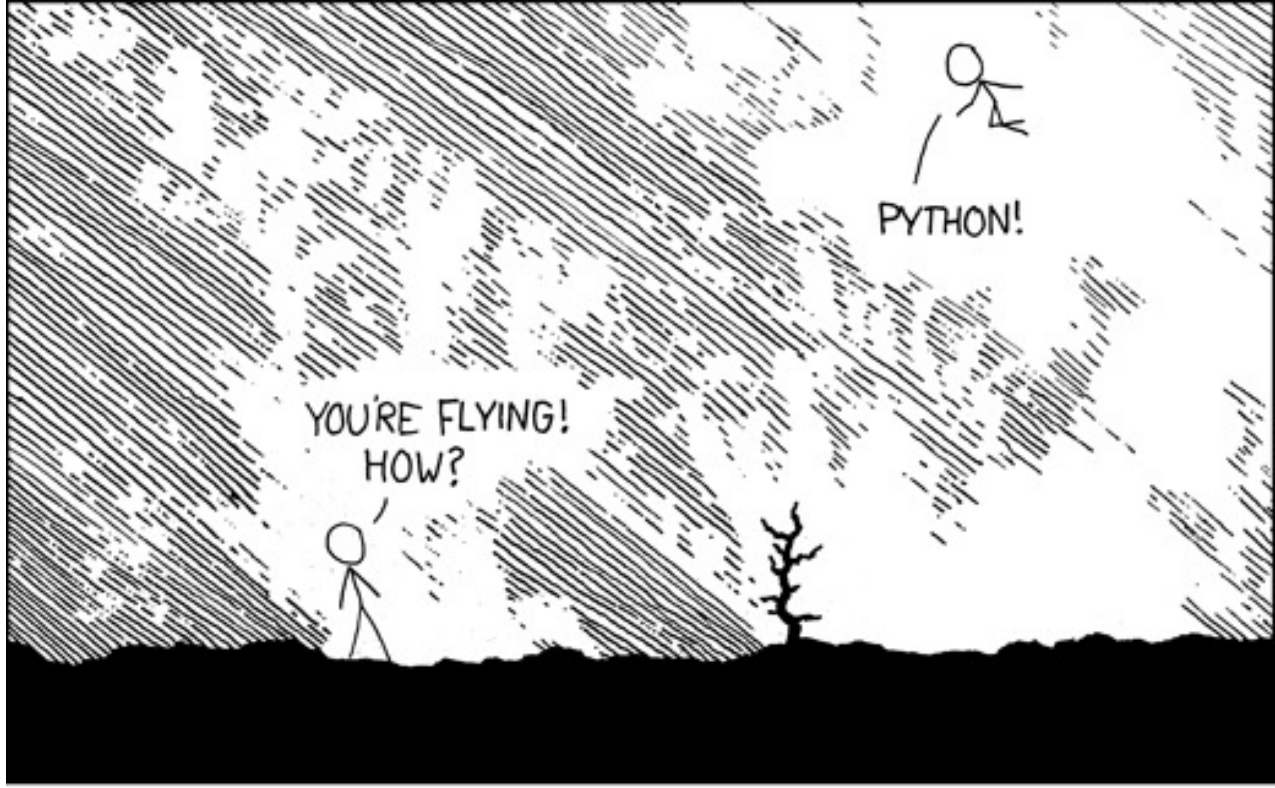
This course is much more about "**learning to learn**" and problem solving rather than acquiring specific programming tricks or stats wizardry

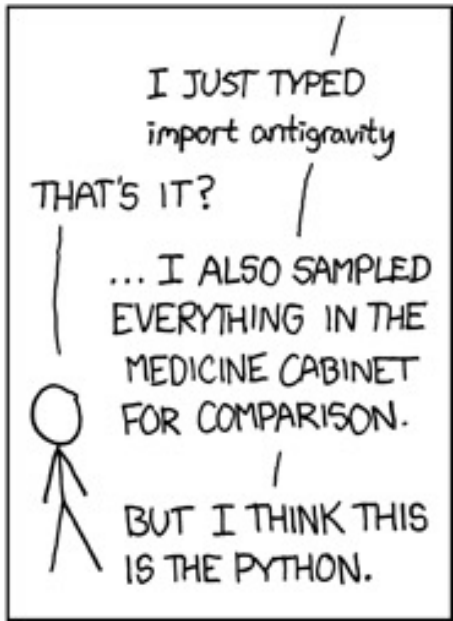
- **Prepare** for the labs
- **I won't** be leading/lecturing at the computer labs. TAs will be present for abundant help and feedback.
- **Go over the notebooks** before the lecture and the computer lab
- If the first time you see a notebook is at the lab, you may find it difficult to follow on. The best thing to do is to prepare a set of questions to ask us.
- **Bring** questions, comments, feedback, (informed) rants to class/labs. The more you bring, the more we all learn.
- **Collaborate** (it's **NOT** a zero-sum win!)

Python

- **General purpose** programming language
- "Sweet spot" between *"proof-of-concept"* and *"production-ready"*
- Industry standard: **GIS** (Esri, QGIS) and **Data Science** (510, World Bank, OECD, The Atlantic, Gemeente Den Haag…)

How many of you have written a line of computer code before?





Assessments

Formative : These are ungraded

- 7 In-Class Labs
- 7 Homework Labs

Discuss the completed homework with your peers using a list of Do’s and Don’ts to evaluate each other's work

Assessments

Summative : There are 4 graded components that contribute to the final mark for the course as follows:

50% Individual assessment include (very easy and meant to keep up with the course)

- Assignment 1: Data Collection and Wrangling (15%)
- Assignment 2: Geographic Visualisation (15%):
- Assignment 3: Prediction/Inference (20%)

50% Group assessment include

- Final Project

Rubric for Assignments

- Assignments are graded based on **four criteria**
- The criteria have different weights that add up to a **total of 10**
- The score for each criterion range from **0 to 2 points**
- Only English

Criteria	Indicative weight	**Needs Improvement (0.5-0.9)**	**Meets Some Expectations (1.0-1.9)**	**Exceeds Expectation (2)**
Output	0.5			
Formatting	0.5			
Methods (e.g. Tidy Data, EDA, Graphical Excellence, Spatial Autocorrelation, Network Weights, Regression, etc.)	2			
Documentation (Markdown/Comments)	2			

Rubric for Assignments

Assignments are about **Methods and Documentation,** not your ability to code

- Output
- Formatting
- Methods
- Documentation

Criteria	Indicative weight	**Needs Improvement (0.5-0.9)**	**Meets Some Expectations (1.0-1.9)**	**Exceeds Expectation (2)**
Output	0.5			
Formatting	0.5			
Methods (e.g. Tidy Data, EDA, Graphical Excellence, Spatial Autocorrelation, Network Weights, Regression, etc.)	2			
Documentation (Markdown/Comments)	2			there is extensive documentation explaining your code's logic and your choices in the analysis. - hypothesis clearly stated - explanation of model choices clearly presented (ex. variables, number of neighbours or clusters, etc.) - interpretation of model results, weakness and strengths (ex. errors in regression) - well substantiated argumentation of results and choices - plots are clearly described and interpreted

Do's

- Finish the corresponding labs before starting an assignment
- Think about the objective of the assignment: data exploration vs data analysis
- Think about the method and its limitations
- Make sure that your code runs and produces the expected output
- Use clear and interpretable variable names: 'SP.DYN.TFRT.IN' vs 'average_fertility_rate'
- Think about the data quality: are there missing values? How will you deal with them? What are the implications?
- Use headers to structure your notebook
- Use markdown to explain the code and interpret the findings

- Please do not use any language other than **Python and English**

Don’ts

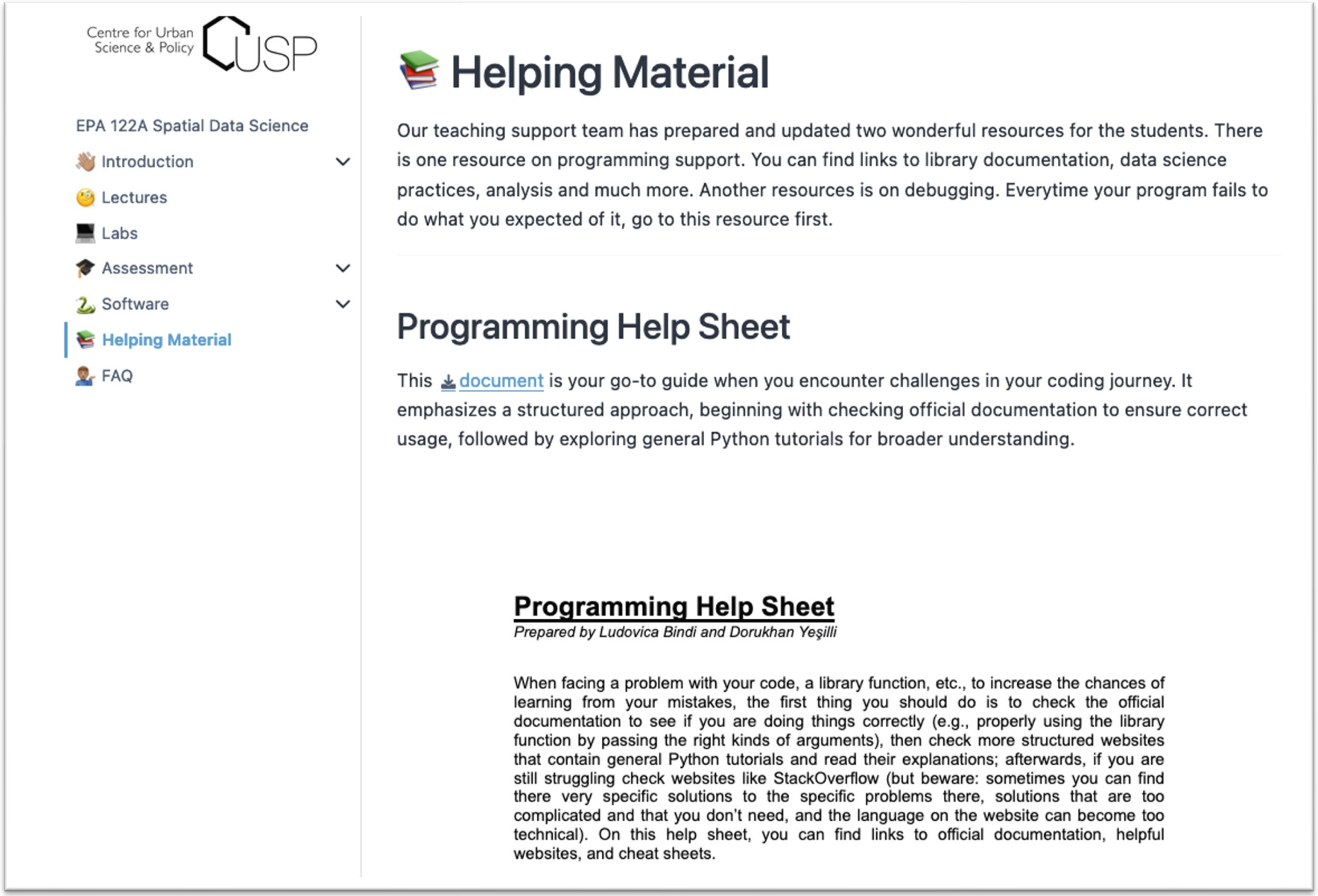
- Don’t copy and paste the code from the labs without understanding what the added value is
- Don’t print huge matrices or use enormous font. Make sure the notebook can be read as a report
- Don’t print empty cells, import useless packages or import packages multiple times
- Don’t print figures without labels or titles
- Don’t use colours that are difficult for colour-blind people to distinguish
- Don’t hardcode, try instead to use functions to make the code easily reusable later
- Don’t overwork, focus on what is asked from you in the assignment

Rubric for Final Project

- Projects are graded based on seven criteria
- The criteria have different weights and the score for each criterion range from 1 to 10 points in increments of .5

Criteria (increments of .5)	Indicative %	(1-4) Unacceptable	(4.5-5.5) Insufficient	(6-6.5) Minimally acceptable	(7-8.5) Meets Expectations	(9-10) Exceeds Expectations
		Defines work that is generally incomplete or substandard	An incomplete attempt to address the task.	An adequate accomplishment of the task.	An above average accomplishment of the task.	An exemplary accomplishment of the task.
Formatting and Legibility (use Grammarly.com for grammar and spelling if unsure)	5%					
Problem Structuring	15%					
Related Work	5%					
Data Collection and Processing	20%					
Methodology (reproducibility means that I understand what you did and how you did it and can repeat the analysis if I had access to the same data)	20%					
Results and Visualisation	20%					
Discussion/Conclusion	15%					

Support!



Centre for Urban Science & Policy USP

EPA 122A Spatial Data Science

- 👋🏽 Introduction
- 🧐 Lectures
- 💻 Labs
- 🎓 Assessment
- 🐍 Software
- 📚 Helping Material
- 🤦🏾 FAQ

📚 Helping Material

Our teaching support team has prepared and updated two wonderful resources for the students. There is one resource on programming support. You can find links to library documentation, data science practices, analysis and much more. Another resources is on debugging. Everytime your program fails to do what you expected of it, go to this resource first.

Programming Help Sheet

This document is your go-to guide when you encounter challenges in your coding journey. It emphasizes a structured approach, beginning with checking official documentation to ensure correct usage, followed by exploring general Python tutorials for broader understanding.

Programming Help Sheet

Prepared by Ludovica Bindi and Dorukhan Yeşilli

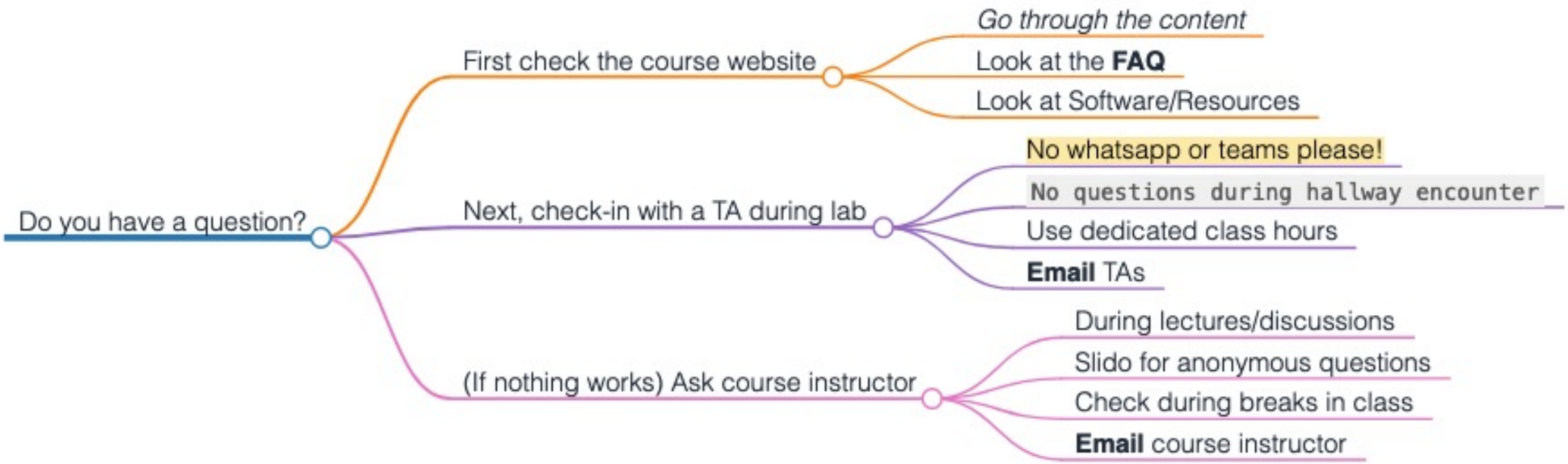
When facing a problem with your code, a library function, etc., to increase the chances of learning from your mistakes, the first thing you should do is to check the official documentation to see if you are doing things correctly (e.g., properly using the library function by passing the right kinds of arguments), then check more structured websites that contain general Python tutorials and read their explanations; afterwards, if you are still struggling check websites like StackOverflow (but beware: sometimes you can find there very specific solutions to the specific problems there, solutions that are too complicated and that you don't need, and the language on the website can become too technical). On this help sheet, you can find links to official documentation, helpful websites, and cheat sheets.

More Support!!!

- **Ask** questions
- **Help** others as much as you can (the best way to learn is to share perspectives)
- **Search** heavily on browser + Stack Overflow (learn to troubleshoot)
- Bring questions, comments, feedback, (informed) rants to class
- Collaborate with each other

Figure adapted from https://github.com/thepracticaldev/orly-full-res under Non-Commercial 2.0 Generic License

General Support Structure



Installing Python

🐍 Software

This course is best followed if you can reproduce the examples and tutorials provided with it. To do so, you will need to install in your machine a series of software packages. These are all open-source and available for free to download.

There are three main pathways to install required Python libraries on your machine.

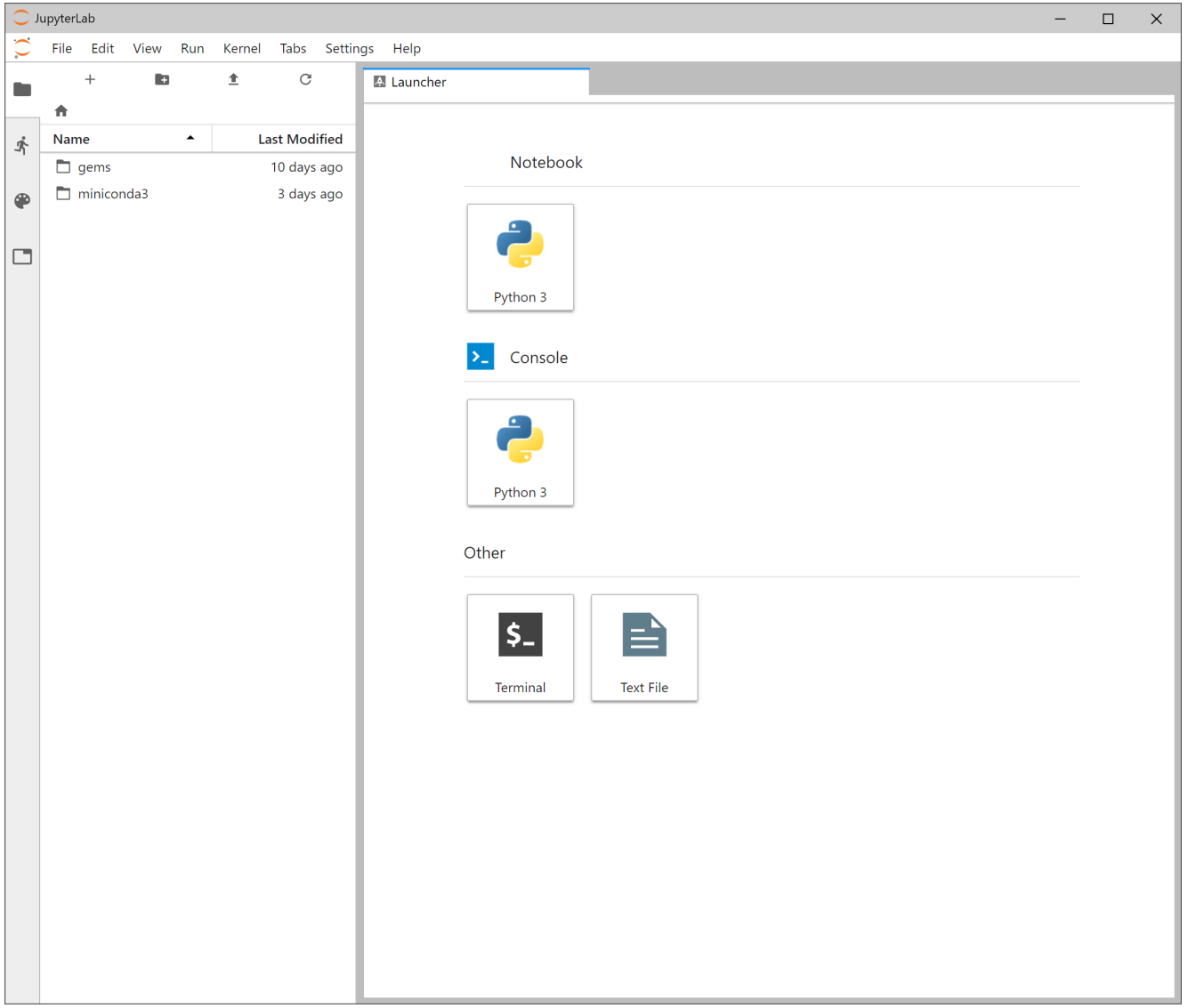
- A **(1) standard** one is installing the software on your operating system without using the command line interface.
- A **(2) minimalist** one will provide basic Python resources and the ability to expand them.
- A **(3) comprehensive** one will install not only a Python stack but also several useful libraries (including some from the programming language, R).

If you want to learn to explore Python and its capabilities, while going beyond this course, I **recommend option 2**. If all else fails, option 3 is the last resort for this course. It is guaranteed to work and very powerful, so you will not be limited in any way. But it does not allow you to install new libraries, which means you are limited by what it offers.

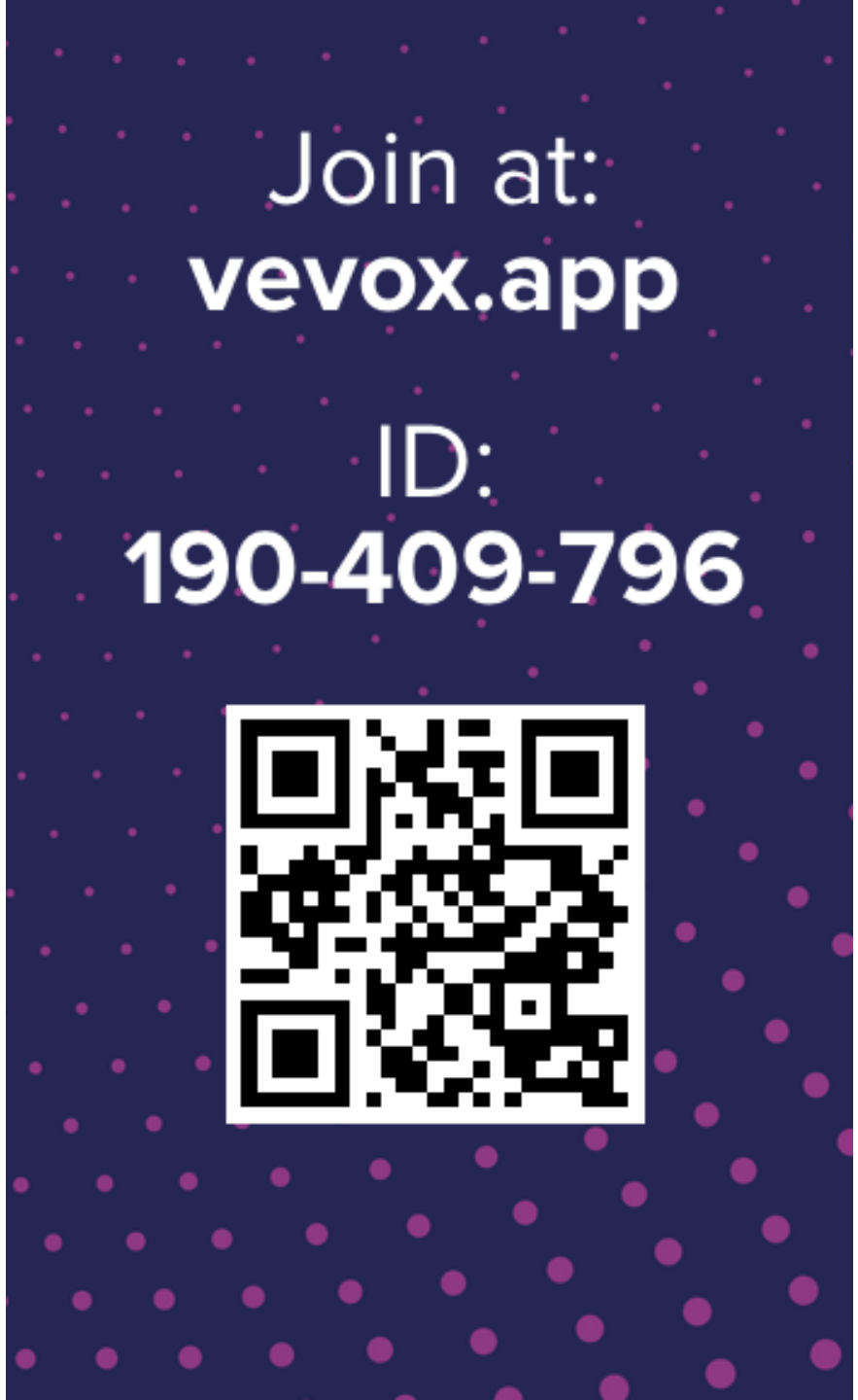
Hint

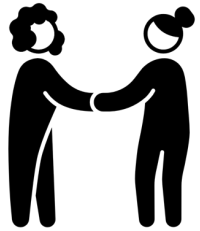
The difference in these options can be explained through the illustration of a living place. If you own a house, you might be able to expand it, paint the walls, add new furniture, even keep a dog. This is akin to the **minimalist** approach which gives you everything you need and the freedom to build upon it. Instead, if you rent a house, in most cases you will not be allowed to make any changes. A **comprehensive** approach gives you everything too, but no freedom to experiment with new python libraries. The **standard** option is like visiting a hotel where others service you for a bit without you having to do the heavy lifting. You choose what works for you!

Installing Python



You?





Meet your peers

Turn to your neighbors and chat for three minutes about,

- Name
- What did you study before coming here?
- What do you expect from this course?
- How do you see this course helping you with your future ambitions?

Community

EPA + Others

- We are from different parts of the world
- We have different educational backgrounds
- We have different experiences in life

Adapted from the work of Sean Perez

Discussions

Only by listening to others can I become aware of the conceptual shackles imposed by my own identity and experiences.

- David Takacs

Experience
Identity
Theories and Evidence

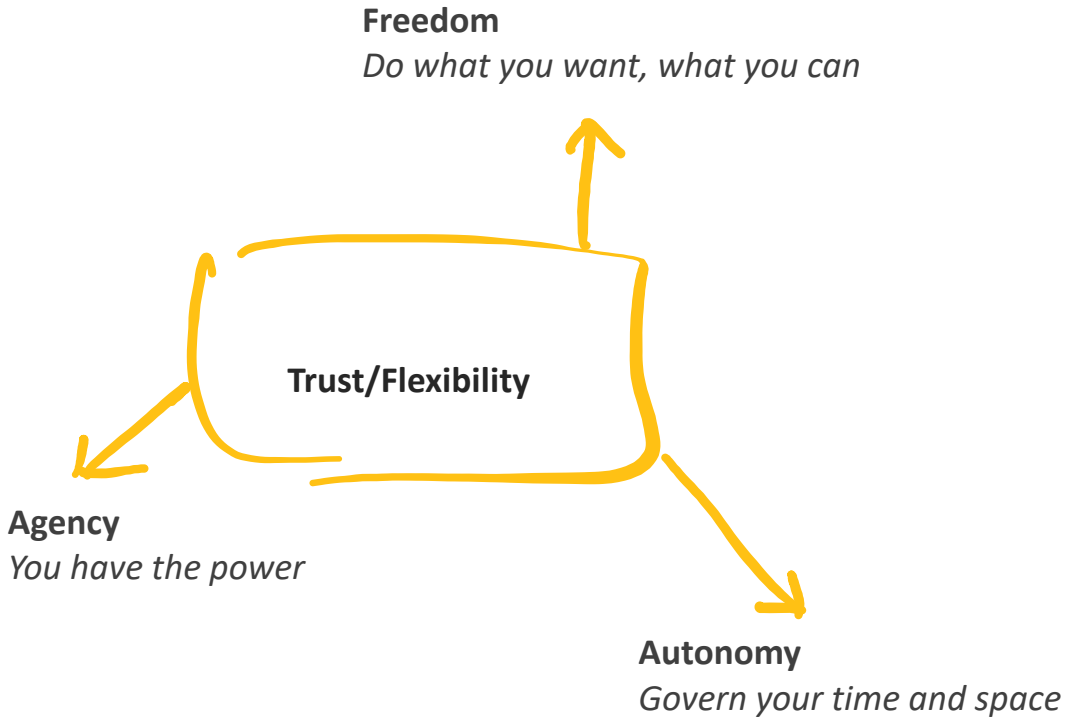
David Takacs, How Does Your Positionality Bias Your Epistemology?, Thought & Action Thought & Action 27 (2003).
Available at: http://repository.uchastings.edu/faculty_scholarship/1264

Flexibility

in learning

Flexibility

in everything



You do what fits your lives!

Flexibility

- You do not owe me “productivity” or “efficiency”. I just want your participation, so we can all learn something new.
- You do not owe me any information concerning your personal situation or mental or physical health condition.
- If you want to, you are welcome to talk to me about anything you are going through. **Just drop me an email**, and we will figure it out from there.
- This course is just one small part of life. If you have to work around it to figure life out, go ahead and do it. I trust you will reach out to me if you need support, and I will be here to offer it.
- Exams only focus on your ability to regurgitate knowledge in a 3-hour window. There is no exam in this course. The final project will provide you an opportunity to learn from each other and create something awesome.

Deadline Philosophy

Deadlines are there for a reason!

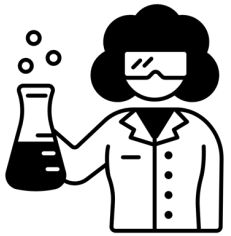
But if you need more time because… well life, you need to send me an **email** explaining **why** you need more time and send me that email in good time.

Slothilda.com

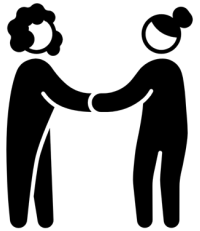
Break

CHILL

WALK



COFFEE OR TEA



MAKE FRIENDS

Icons by Iconathon on The Noun Project

Spatial Data Science

Introduction-II

(EPA122A)

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?

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

Statistics

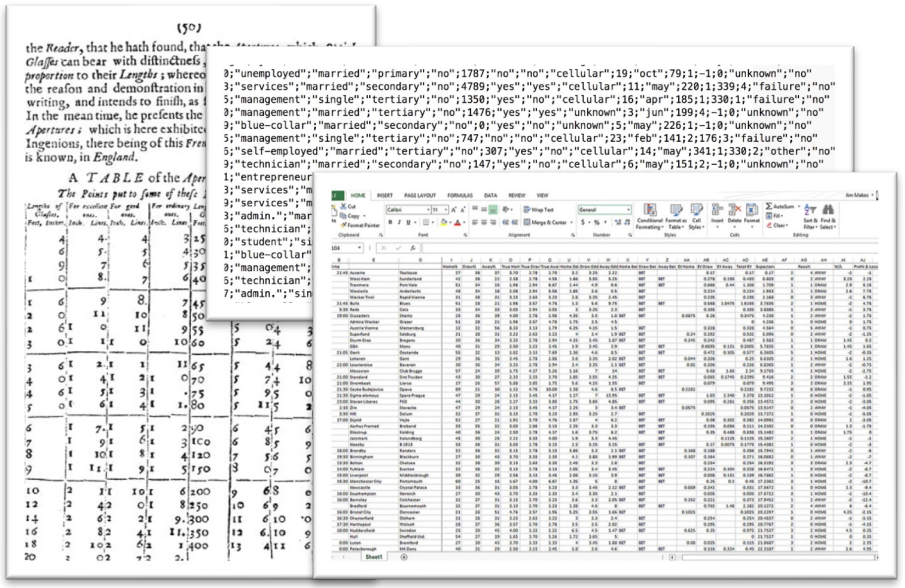
Machine Learning

What's next?

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

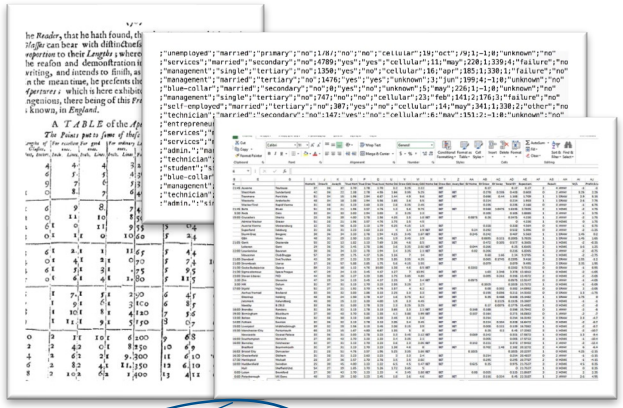
At this point many people have said it..

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

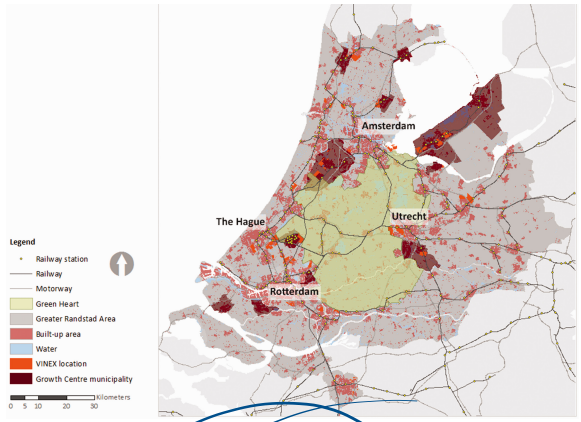


Formats: CSV, Excel, JSON, Shapefiles

Now, **data alone is not very valuable**



Data



Action

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

Kasraian, Dena, et al. "The impact of urban proximity, transport accessibility and policy on urban growth: A longitudinal analysis over five decades." *Environment and Planning B: Urban analytics and City Science* (2017)

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 + …

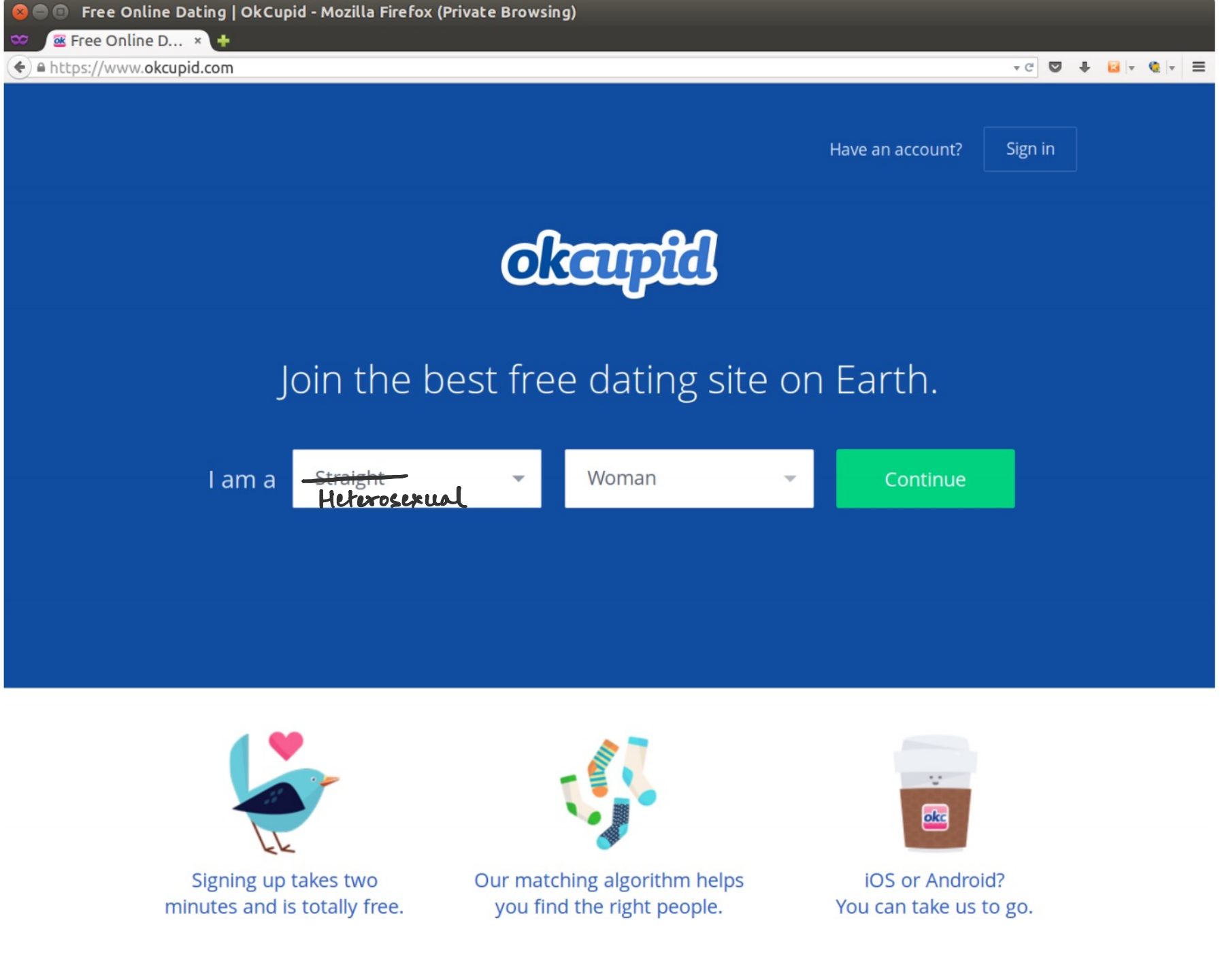
- **Computational** tools –> Programming (hence this course's labs and homework!)
- **Communication** skills –> "Story telling " (hence this course's assignments)
- **Domain** expertise –> Theories about why the data are the way they are (hence the rest of your degree)

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
Straight
Heterosexual
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

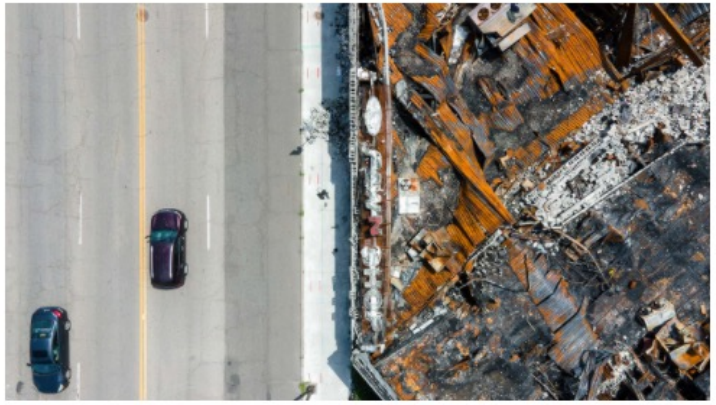
Rising Seas

Forced Displacement

Economic Inefficiencies

Growing Energy Demand

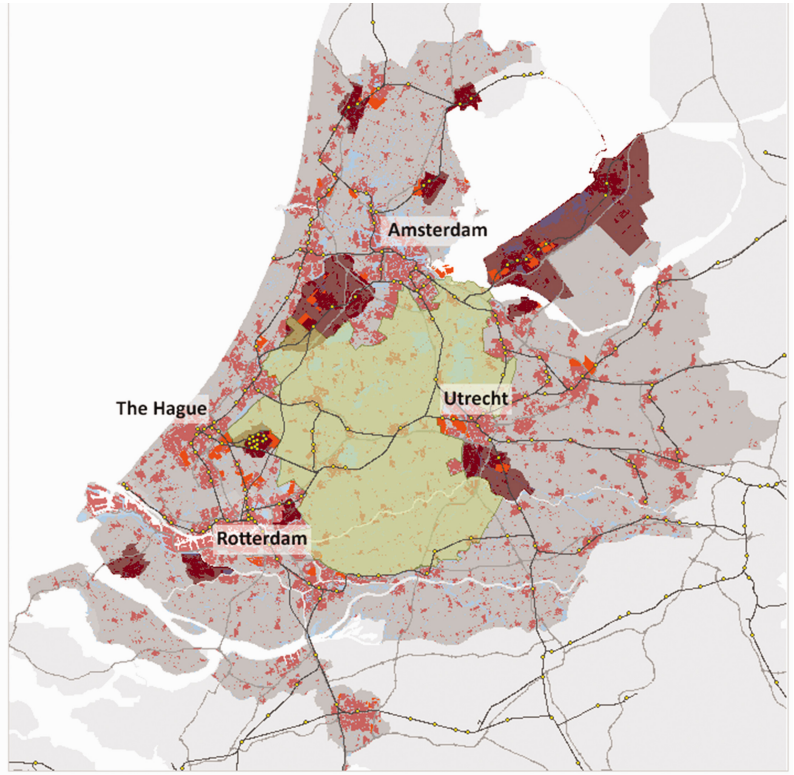




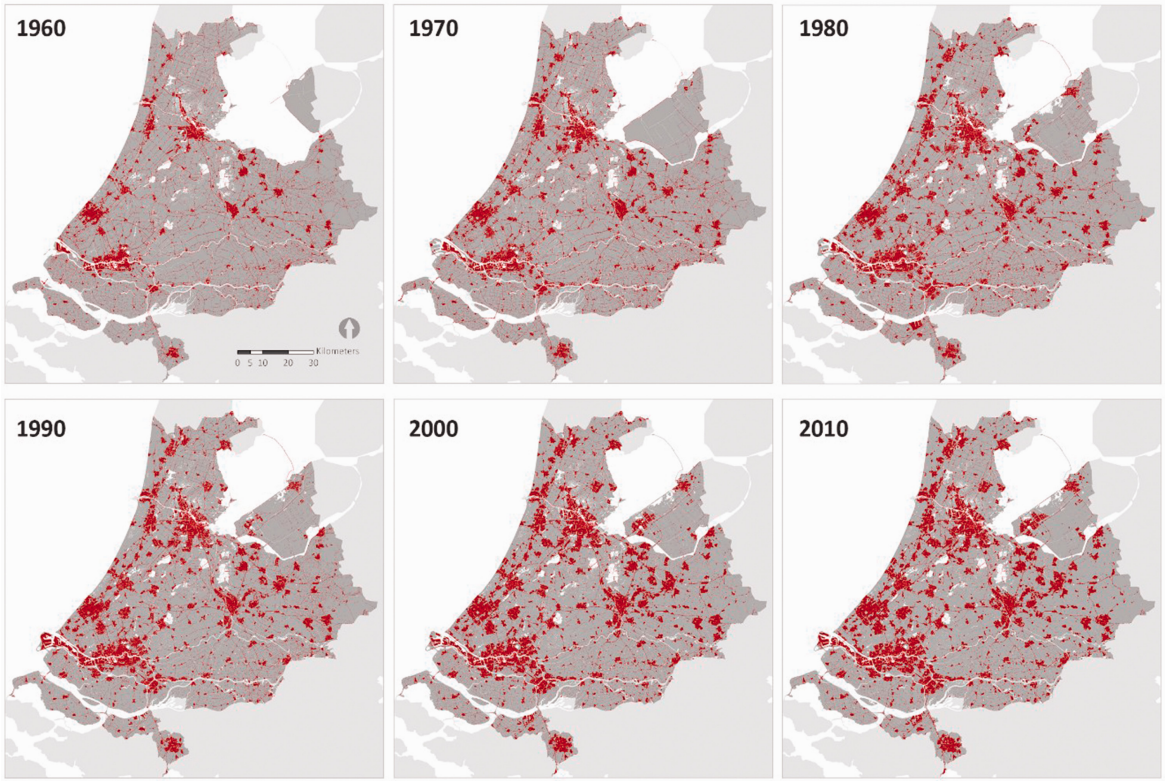

Photographs by Johnny Miller https://unequalscenes.com/

Kasraian, Dena, et al. "The impact of urban proximity, transport accessibility and policy on urban growth: A longitudinal analysis over five decades." *Environment and Planning B: Urban analytics and City Science* (2017)

The Local Picture: Randstad



Urbanisation of the Greater Randstad Area 1960-2010



Space is important!

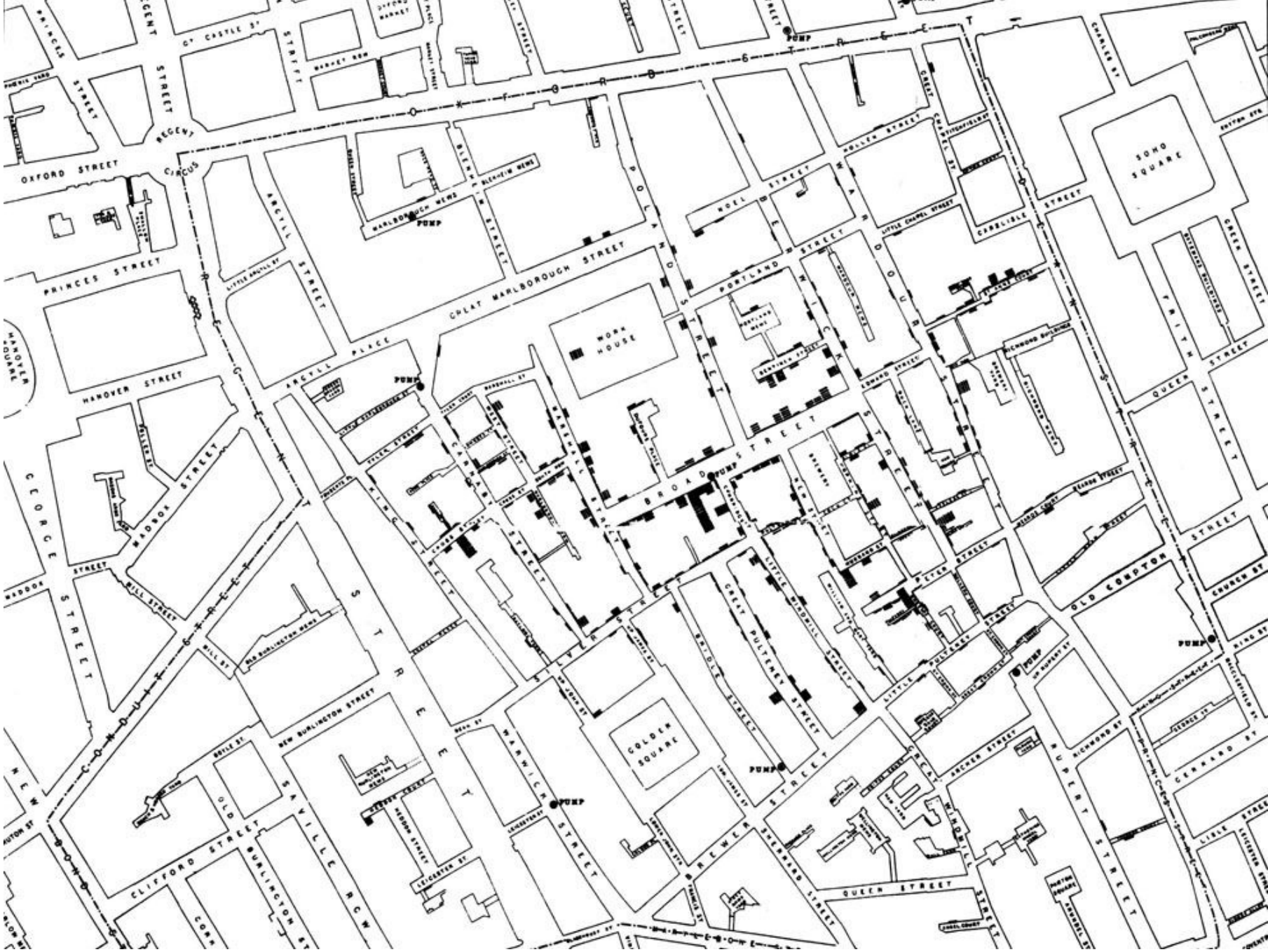
- The space around us
- Thus, the geography of our location
- Geolocated data

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

(Geo-)Data Science

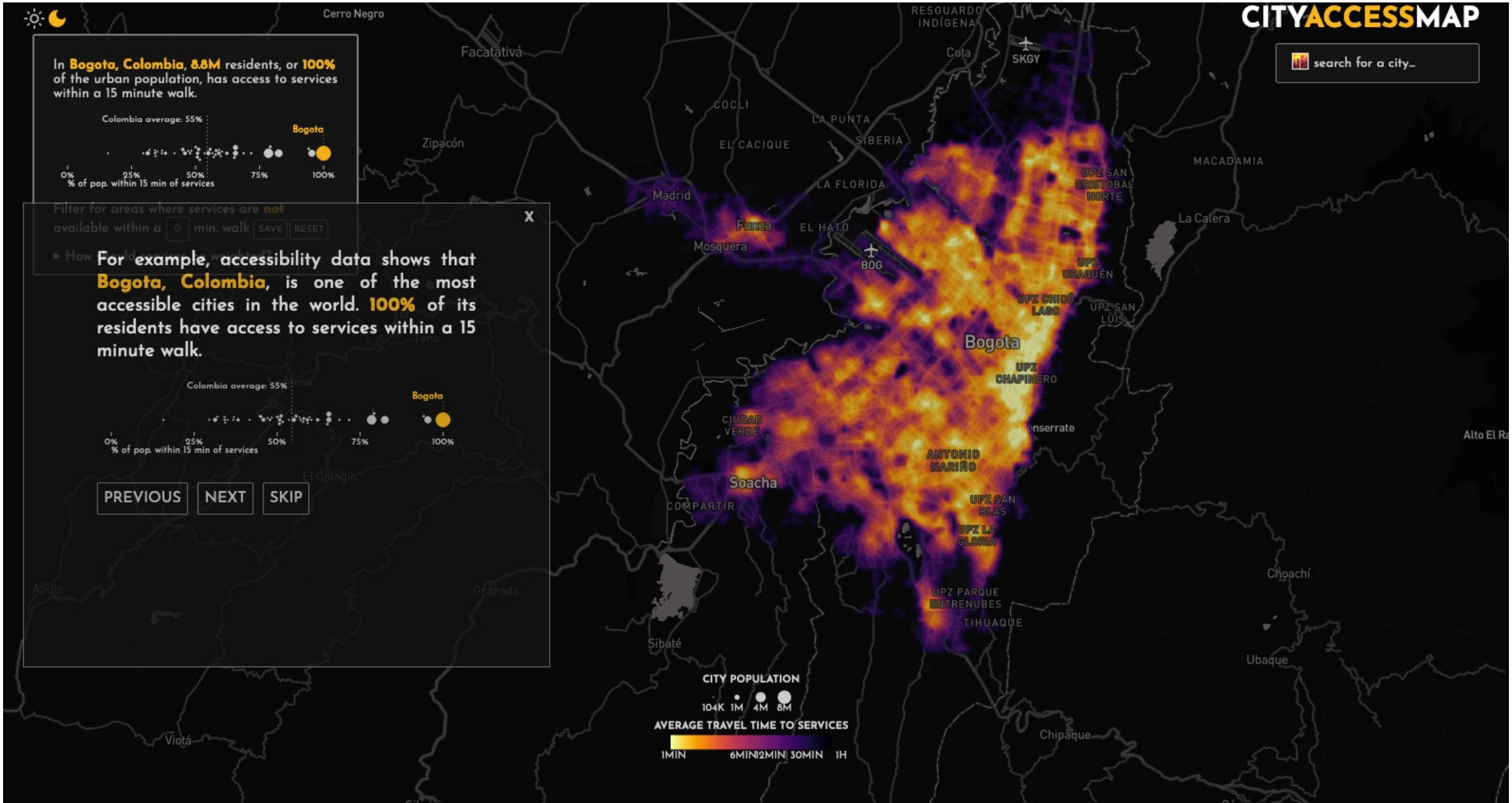
*(Geo-)*Data Science

- 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.

Nicoletti, L., Sireno, M., & Verma, T. (2021). Unequal Access to Urban Infrastructure in Cities across the World. In Preparation.



To do before class [Takes about 1 hour of prep at home]

As a way to whet your appetite about the content of the first class, I recommend you:

- Listen to this interview with Hilary Mason, Max Shron, and Alex Pentland about the power of data.
- Watch this video by Mike Flowers, Chief Analytics Officer, at the City of New York about how data is used to influence policy decisions.
- Read What New Yorkers are complaining about and reflect on if the cost of running such data systems worth the price of knowing?

ARCHIVE

Is the Cost of 311 Systems Worth the Price of Knowing?

311 systems have revolutionized the way cities gather information, allowing them to tackle small problems before they get too big. But running them can be extremely costly.

February 24, 2014 • **Tod Newcombe**

Minneapolis, Minn. FlickrCC/Photo Phiend

Why Data Science?

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

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)

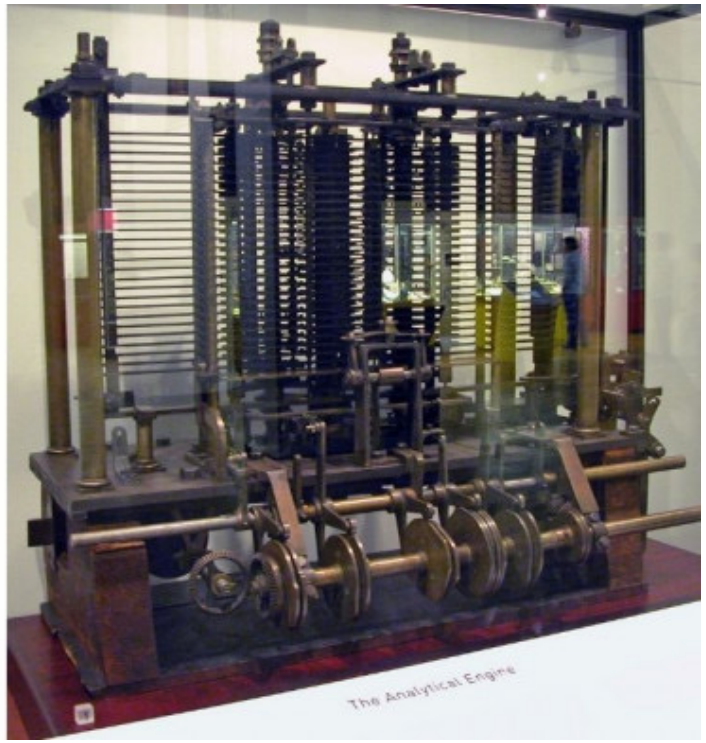
Few hundred years: theoretical approaches, try to derive equations to describe general phenomena.

Maxwell's Equations	$\nabla \cdot \mathbf{E} = 0 \quad \nabla \cdot \mathbf{H} = 0$ $\nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{H}}{\partial t} \quad \nabla \times \mathbf{H} = \frac{1}{c}\frac{\partial E}{\partial t}$	J.C. Maxwell, 1865
Second Law of Thermodynamics	$\mathrm{d}S \geq 0$	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

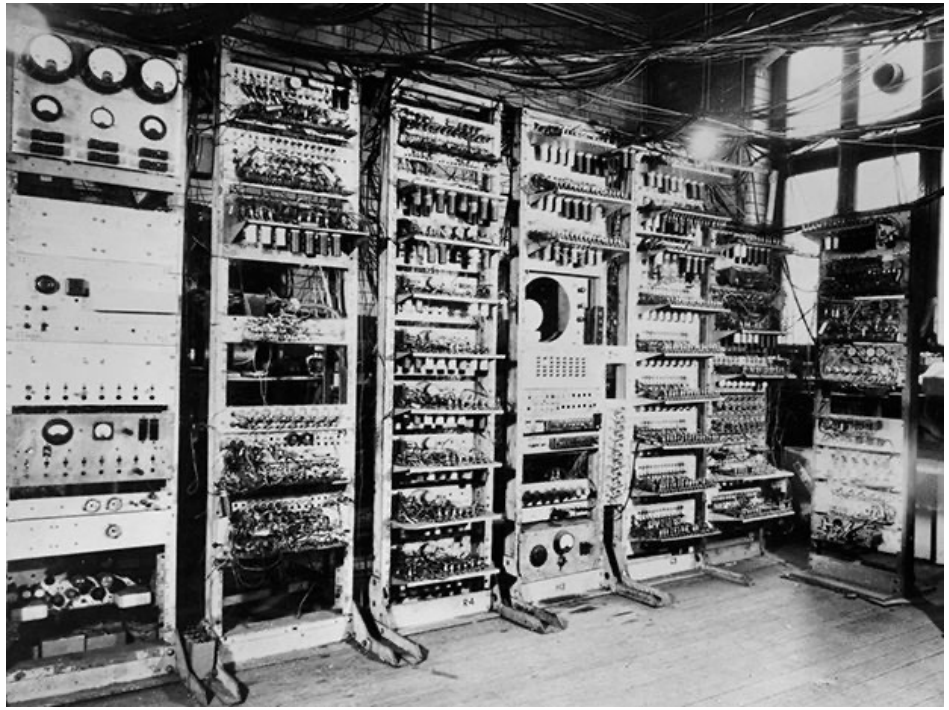
Stewart, I. (2012). *In pursuit of the unknown: 17 equations that changed the world*. Basic Books.

History (cont)

About a hundred years ago: computational approaches



Scanned from *The Calculating Passion of Ada Byron* by Joan Baum.
Analytical Machine Wikimedia Commons

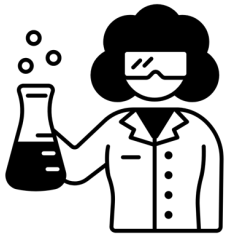


SSPL/Getty Images The Manchester Mark I at Manchester University's Computer Machine Laboratory.

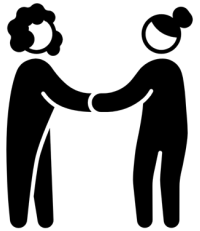
Break

CHILL

WALK



COFFEE OR TEA



MAKE FRIENDS

Icons by Iconathon on The Noun Project

Created by The UCLA Luskin Institute on Inequality and Democracy

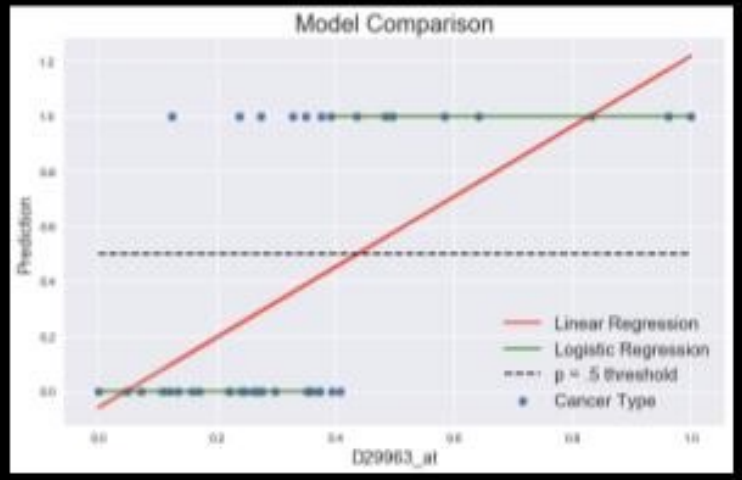
What is Data Science?

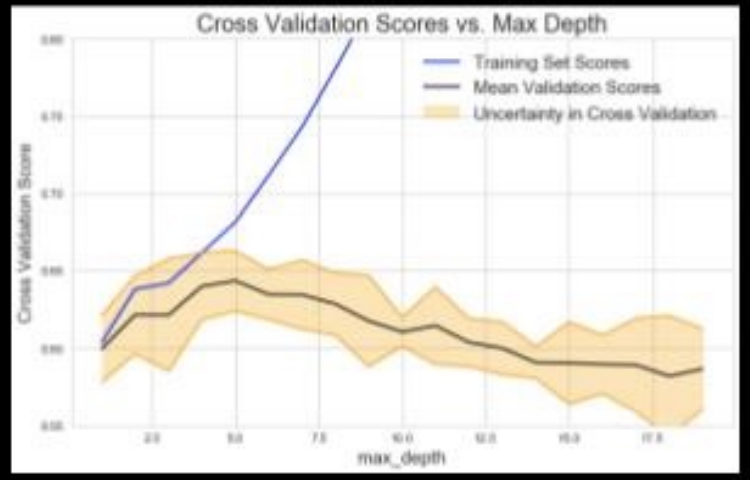
what my friends think I do

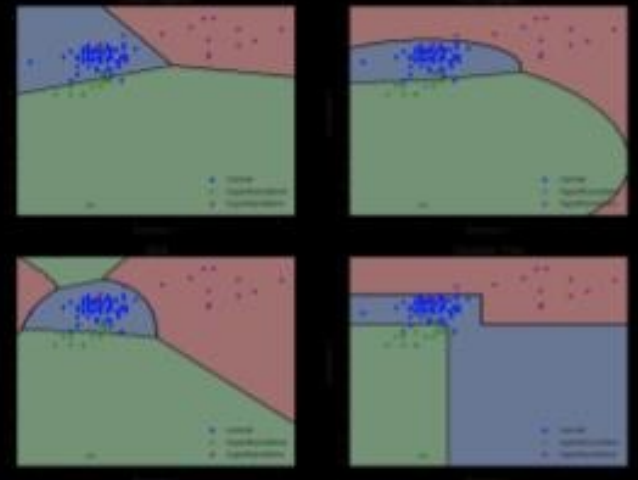
what my family thinks I do

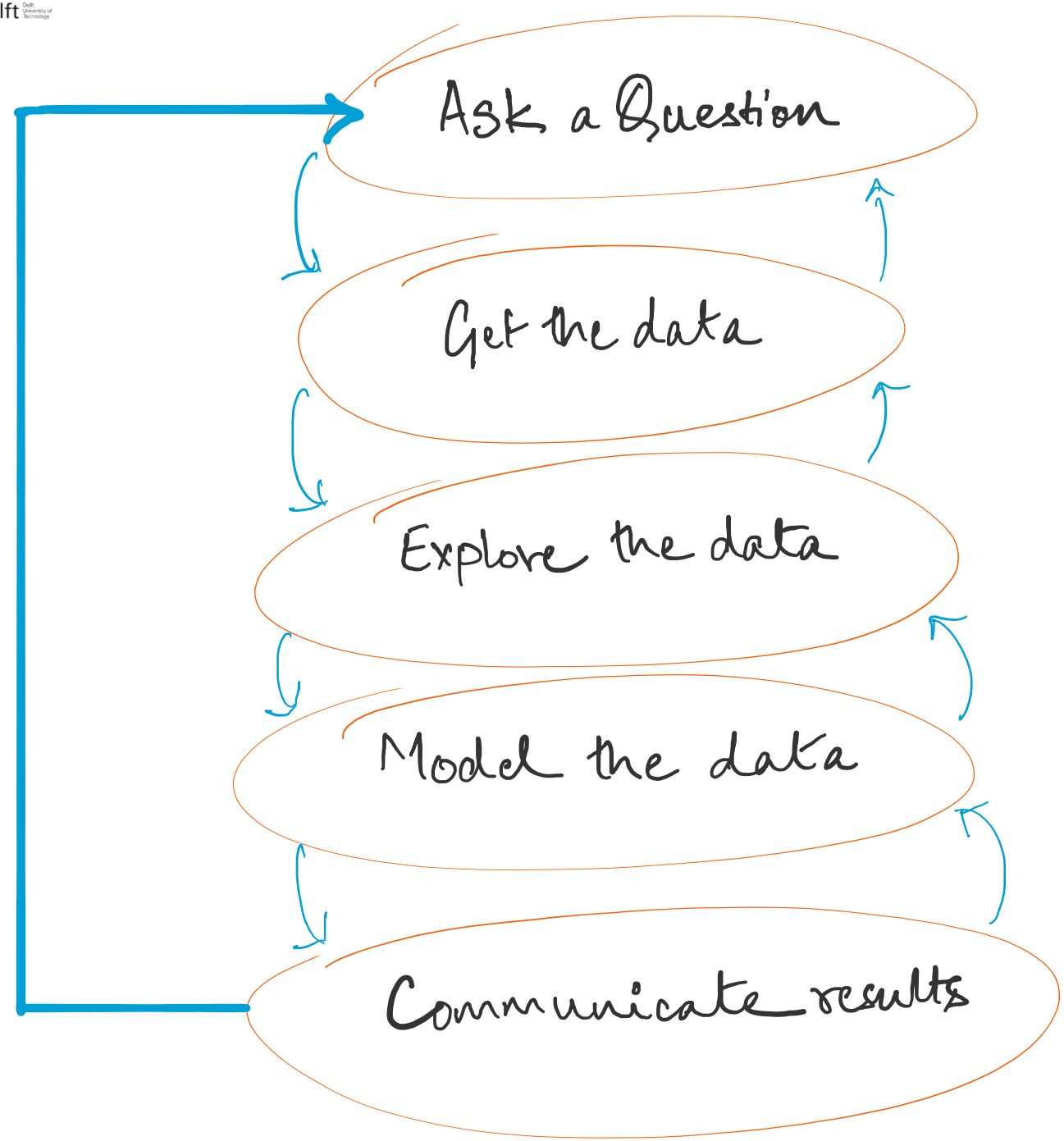
what society thinks I do

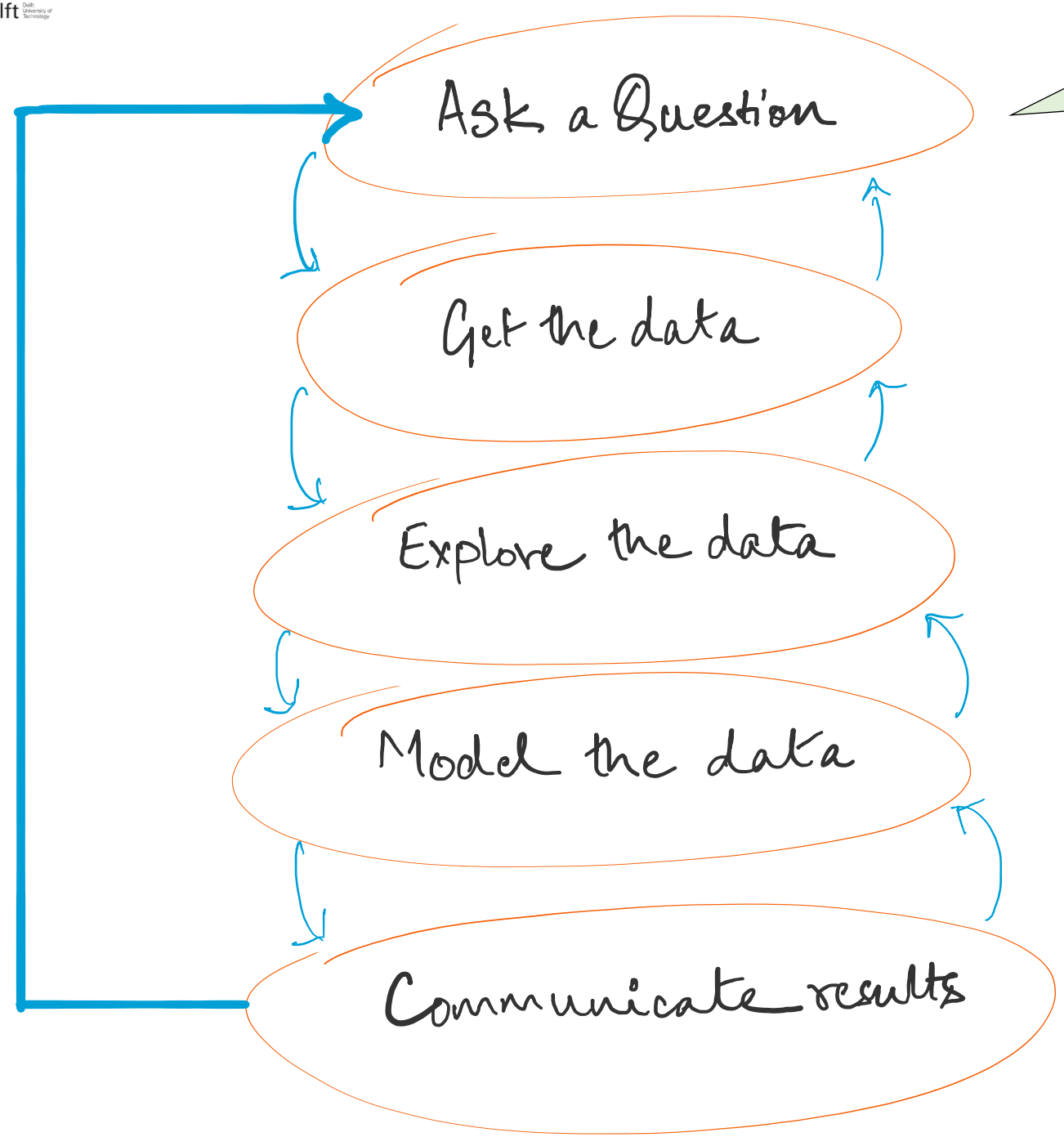
what I actually (will) do in Data Science 1



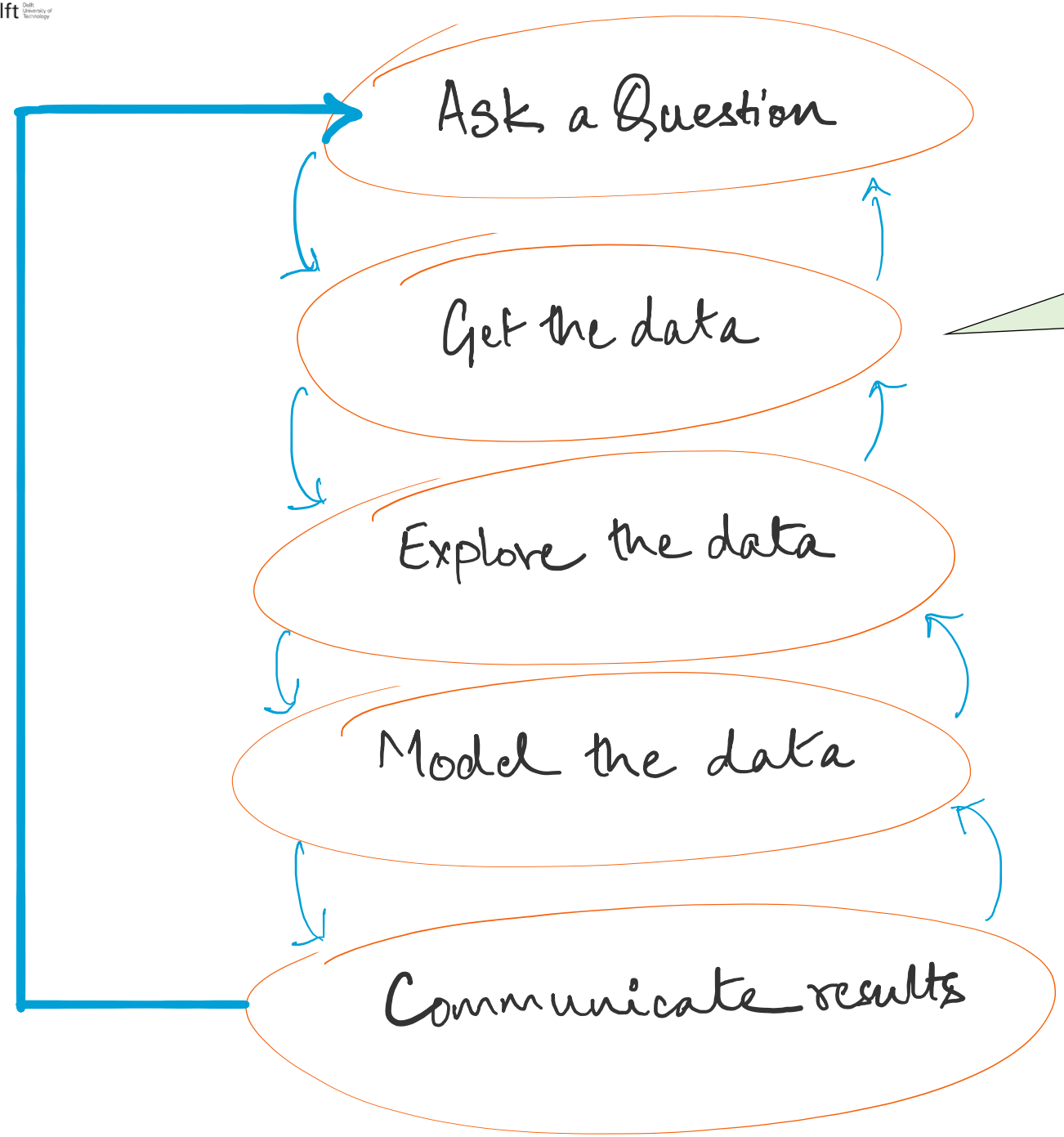




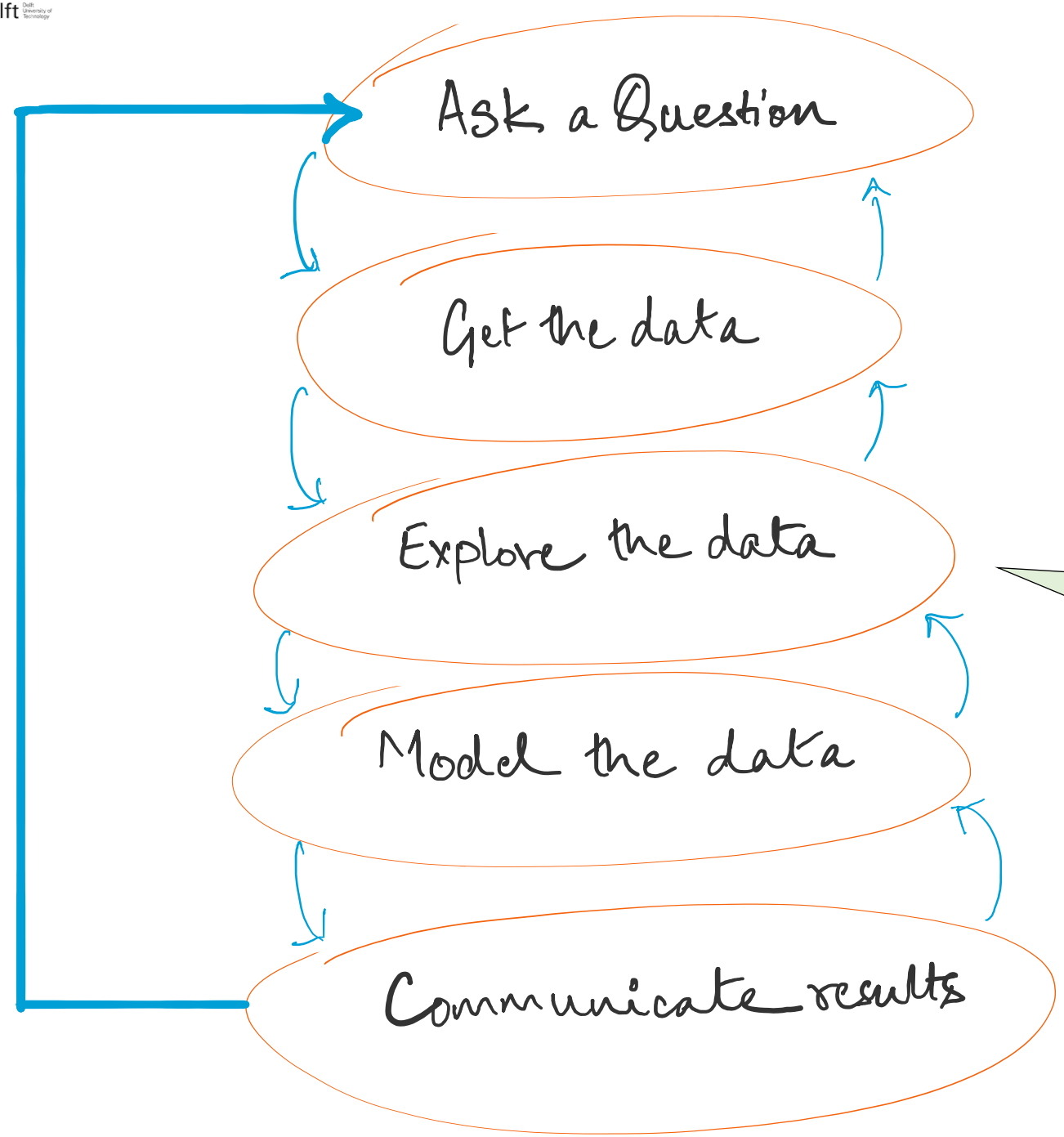

Ask a Question
Get the data
Explore the data
Model the data
Communicate results



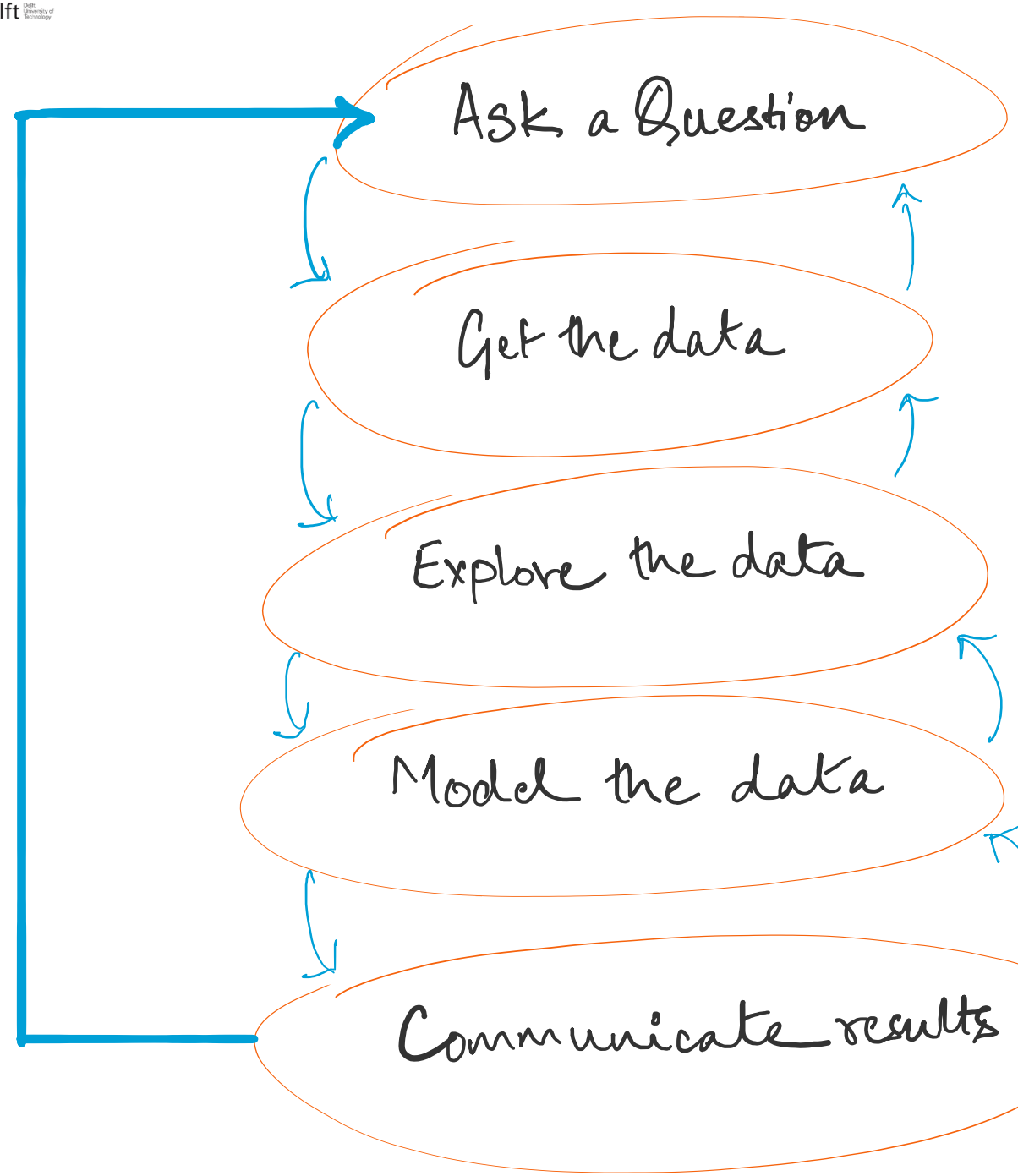
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?



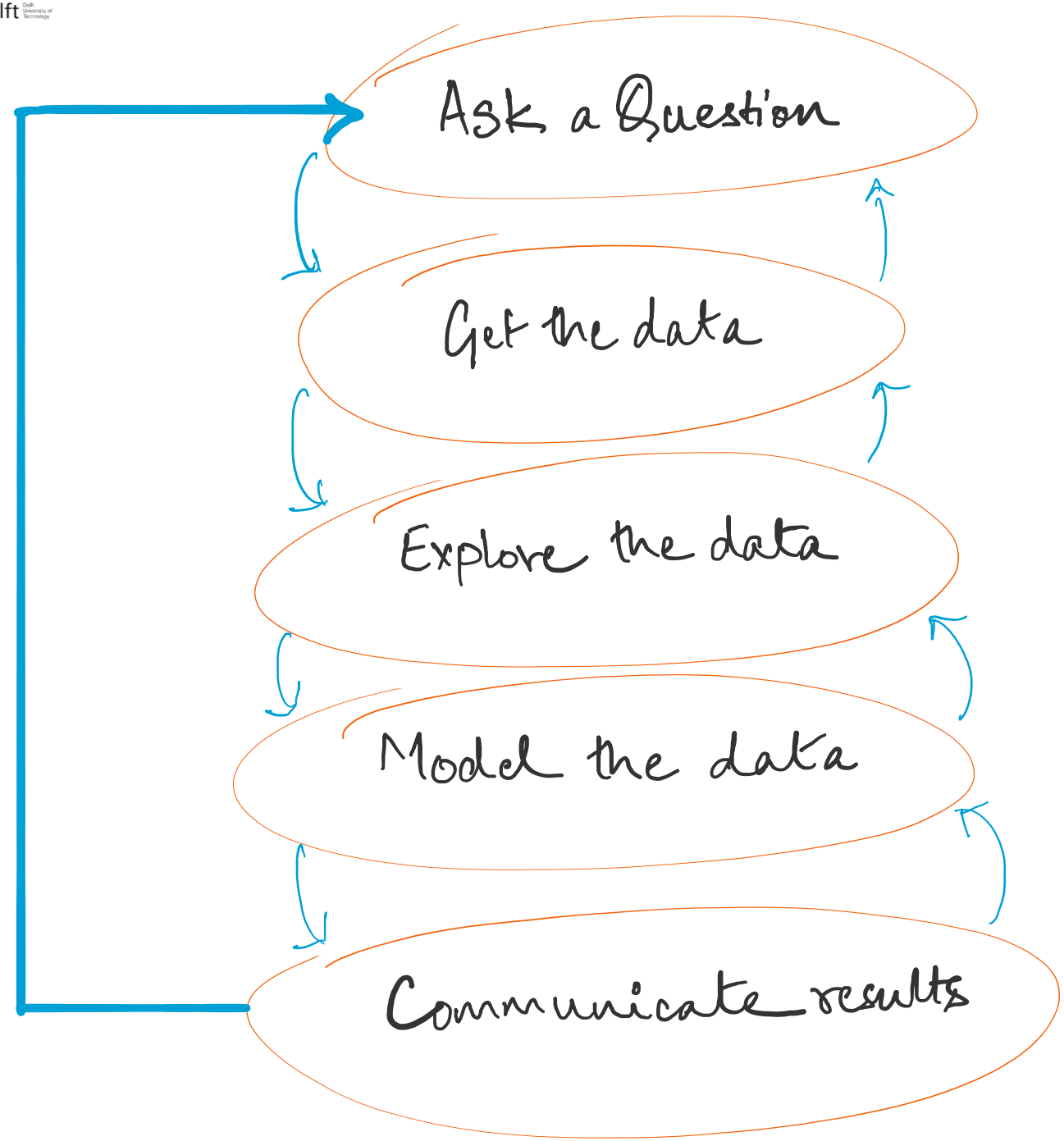
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?



Plot the data.
Are there **anomalies**?
Are there **patterns**?



Build a model.
Fit the model.
Validate the model.



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

EPA Programme

DATA

Agent based modelling — Igor Nikolic

Network Science — Piet van Mieghem

Machine Learning — CS Masters @ EWI

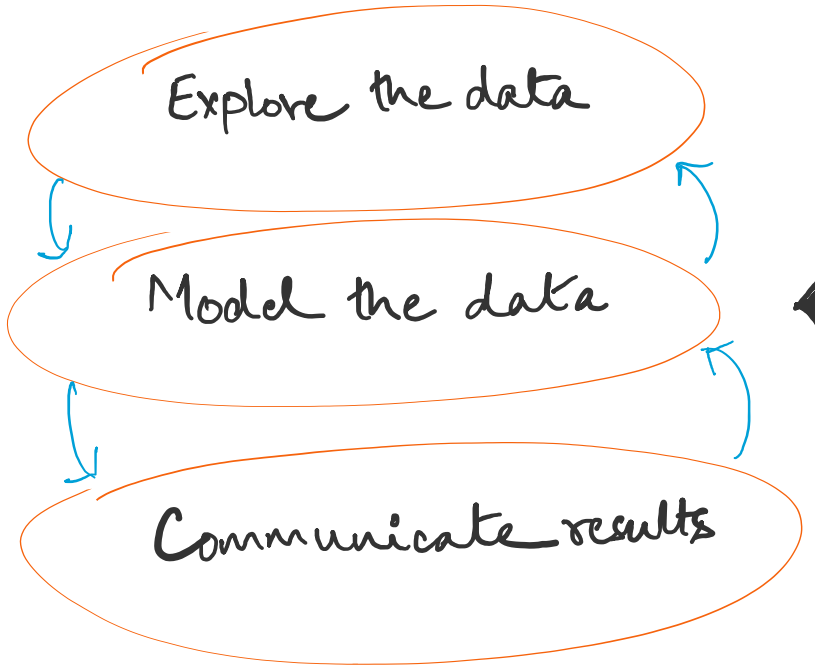
Survey methodology — TPM

System Dynamics — Willem Auping + Jan Kwakkel

Simulation — Alexander Verbraeck

DATA

Visualisation Q1



Uncertainty
Jan Kwakkel Q4

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!

Critical Data Science

- A student handbook led by Laura van Geene

Data Science Process	**Inclusion** *Who is (not) included in the data?*	**Inequality** *What role does inequality play in data science methods?*	**Participation** *Who is (not) involved in the data science process?*	**Power** *How does the data reflect existing power dynamics?*	**Positionality** *What is your own positionality with the research?*
Focus of Analysis *Theories, processes & stakeholders that drive the analysis*	Investigation of exclusive practices of past and present relating to the research focus, and how these affect the diversity of the people represented in the data (Boyd, 2021a; Lee et al., 2022).	What tools can be reliably used to explore the research topic? Research the limitations of the methods, particularly their influence to structural inequalities (Boyd, 2021a).	Use a participatory modelling approach and include stakeholders that may not have otherwise been involved in the design of the research process and discuss how to include topics/perspectives that are not commonly researched (Lee et al., 2022). Discuss the possibility of multiple framings of the research topic (Delbosc, 2023).	Investigation of where and with whom power was distributed in the situations referenced with the data (Lee et al., 2022). Also investigate potential histories of injustices and oppression of the sampling population (Harrington et al., 2021).	Critically reflect on your own position to the research (Boyd, 2021a): 1. Why are your doing research about this specific topic? Why are you specifically involved in this research? What makes you suitable for this research? 2. What is the story that you are trying to tell with this research? Consider biases: do you already have ideas about how this story should go? 3. Is there potential that you cause harm or erasure with your research about this topic?

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 for more practice

Check Assignment contents and due date

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