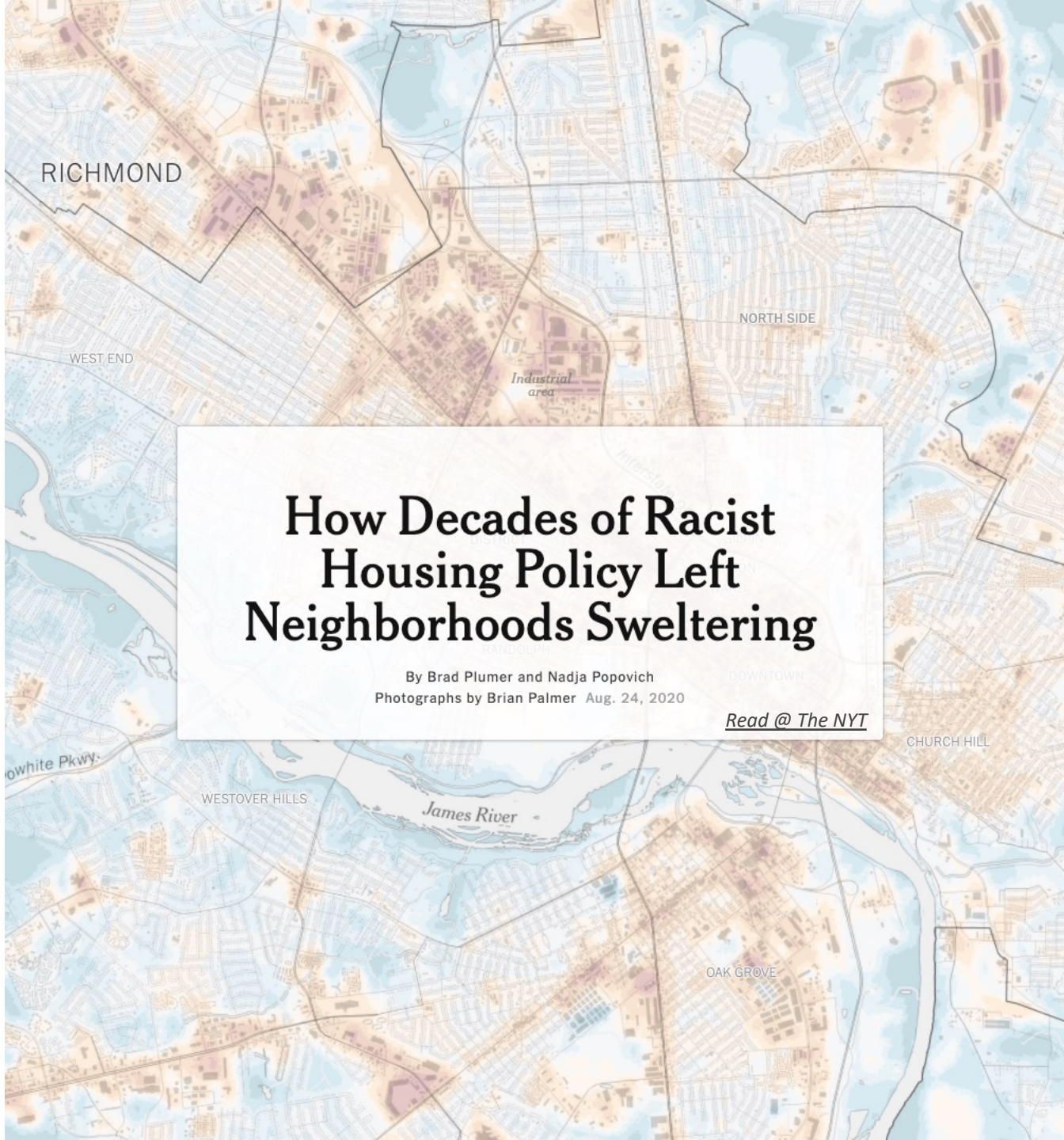
Introduction to *Urban* Data Science

Spatial and Urban Data

(EPA1316A)

Lecture 2

Trivik Verma



Questions?

- Confusion around labs, homework and assignments…

Labs

- Practice
- Not graded

Homework

- Practice
- Peer-Review
- Not graded

Assignments

- Graded
- Submit on Brightspace

Project (10)

- Graded
- Groups
- Submit on BS

More help…?

- First troubleshoot yourself
- Then write on the discussion forum
- If that doesn’t work, contact a TA for help
- Finally, if nothing works, write to me

Last Time

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

Today

- The Data Science Process
- Examine the **role** of evidence in policy
- Analyse **data** understanding and preparation **requirements**
- What are data?
- Types of (geo-)data
- Traditional and new sources of spatial data
- Opportunities and Challenges
- New ways for traditional approaches

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

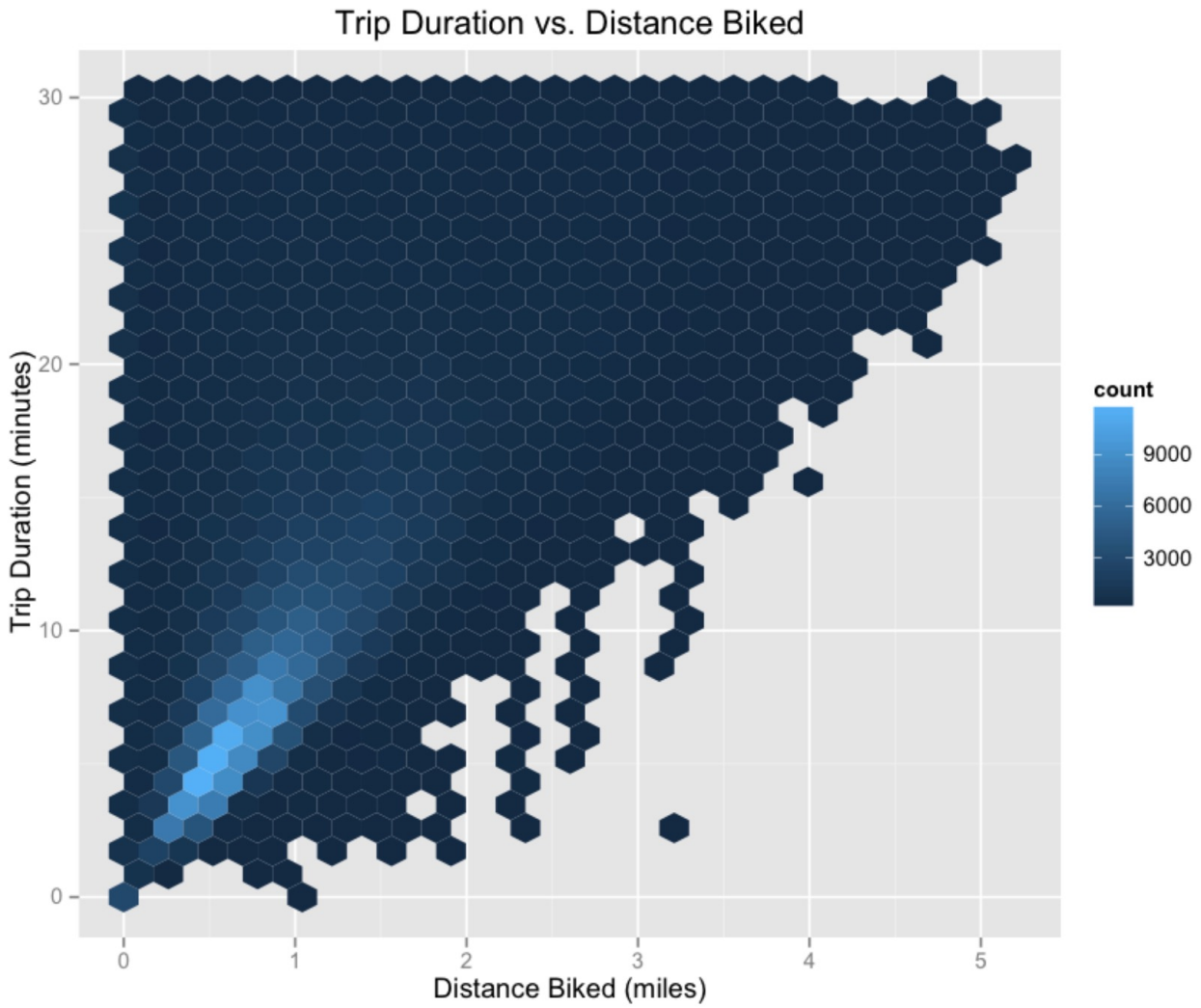
How? Questions that combine variables.

- 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

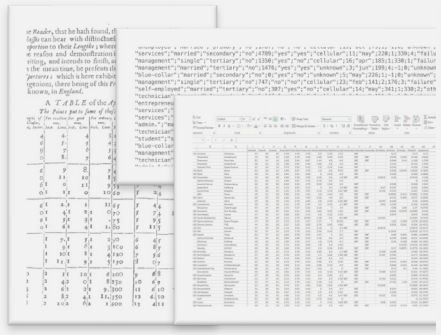
So how well did we do in formulating creative hypotheses and manipulating the data for answers?



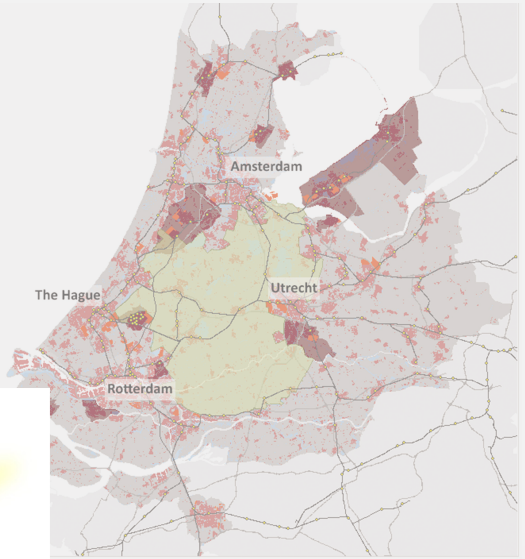
Role of Evidence in Policy

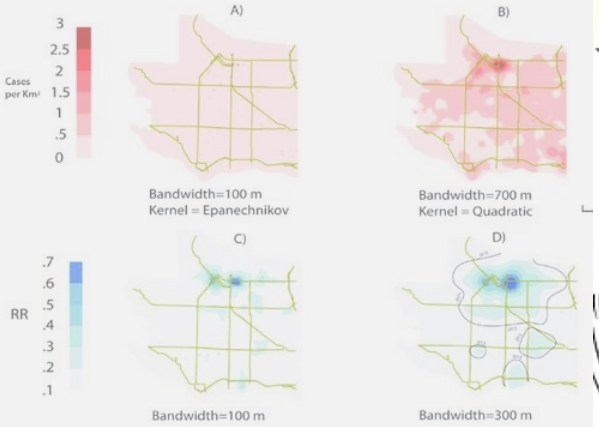
Example: Resilient Transportation

Workshop Report: Designing for Resilient Transportation, by Resilience Accelerator The Hague




LET'S SOLVE THIS PROBLEM BY
USING THE BIG DATA NONE
OF US HAVE THE SLIGHTEST
IDEA WHAT TO DO WITH


Amsterdam
The Hague
Utrecht
Rotterdam


A)
B)
C)
D)
Cases per Km²
RR
Bandwidth=100 m
Kernel = Epanechnikov
Bandwidth=700 m
Kernel = Quadratic
Bandwidth=100 m
Bandwidth=300 m


? UM ?!!


DATA


SORTED


ARRANGED


PRESENTED
VISUALLY


Geographical
Cultural
Transport
Natural
Scientific
Types of Data
Meteorological
Financial
Statistical


CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

Problem Understanding (Vision)

SUSTAINABLE

RESILIENT

INCLUSIVE

SMART

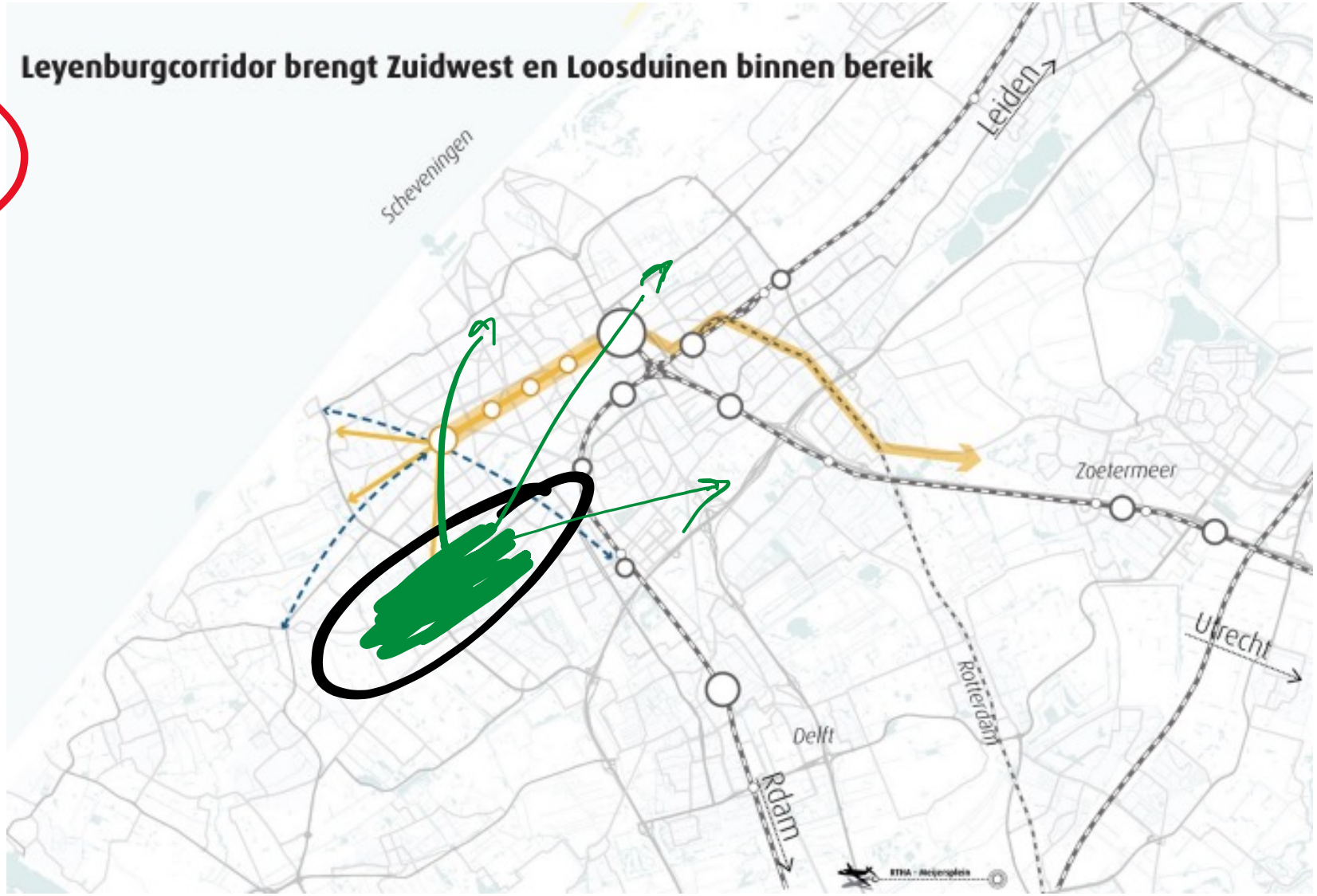
EQUITABLE

Workshop Report: Designing for Resilient Transportation, by Resilience Accelerator The Hague

Problem Understanding

Poor Neighbourhood

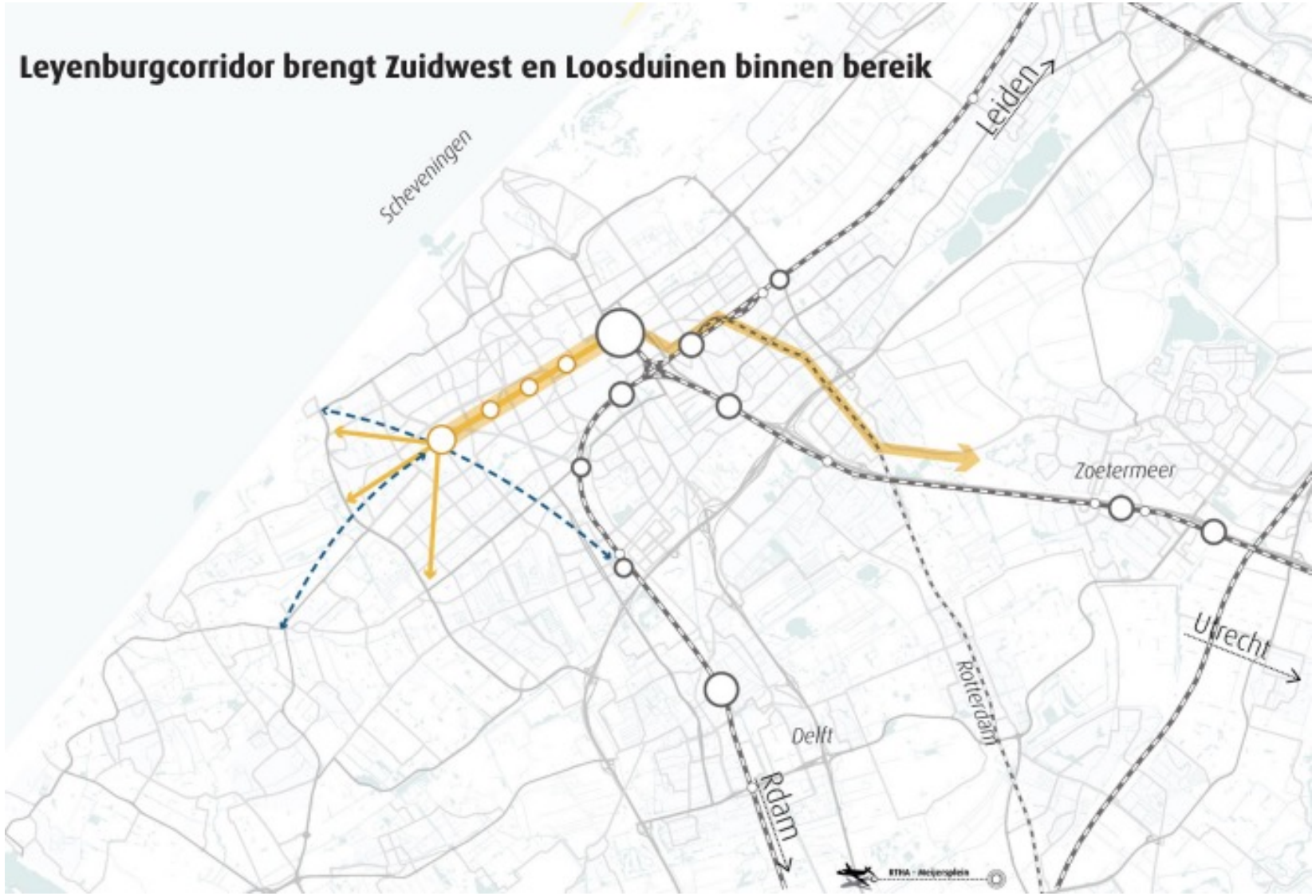
Promote
- → Development
- → Social Cohesion
- → Access to jobs + infrastructure
- → Reduce car trips + increase PT use



Workshop Report: Designing for Resilient Transportation, by Resilience Accelerator The Hague

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



Workshop Report: Designing for Resilient Transportation, by Resilience Accelerator The Hague

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

Problem Understanding

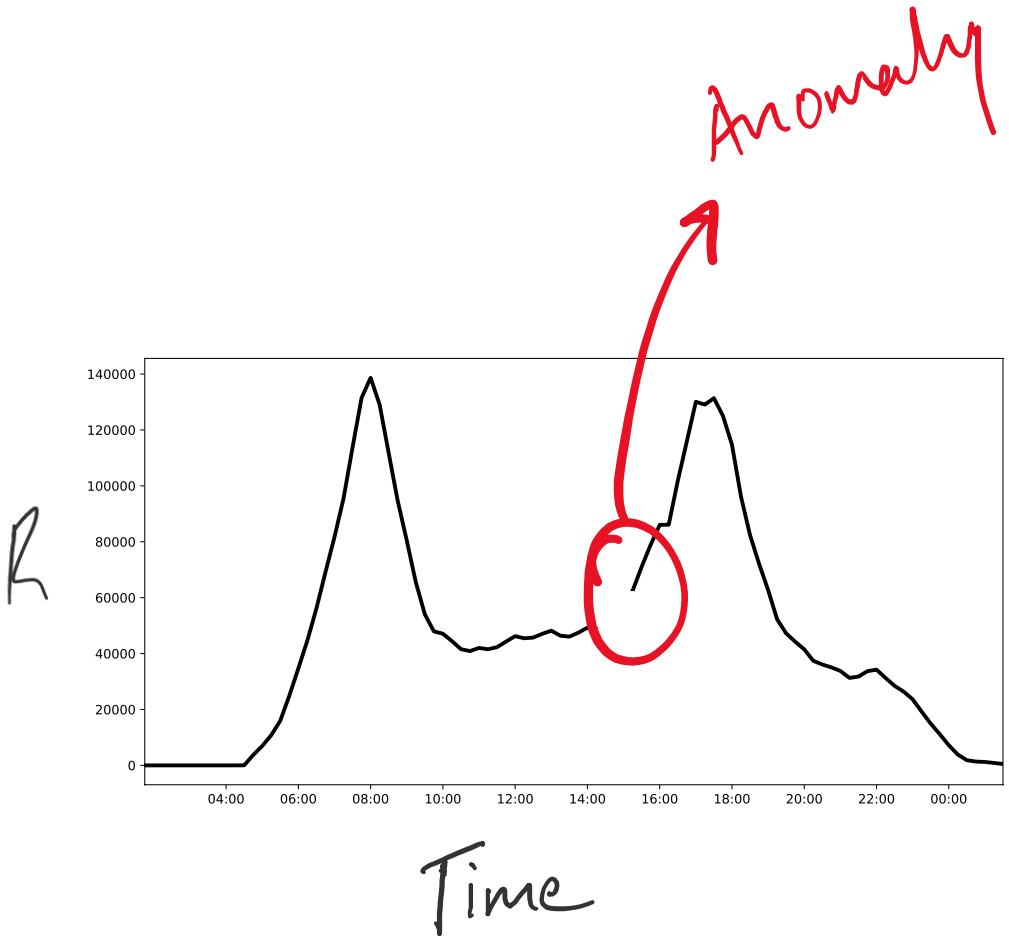
My DOs and DON'TS

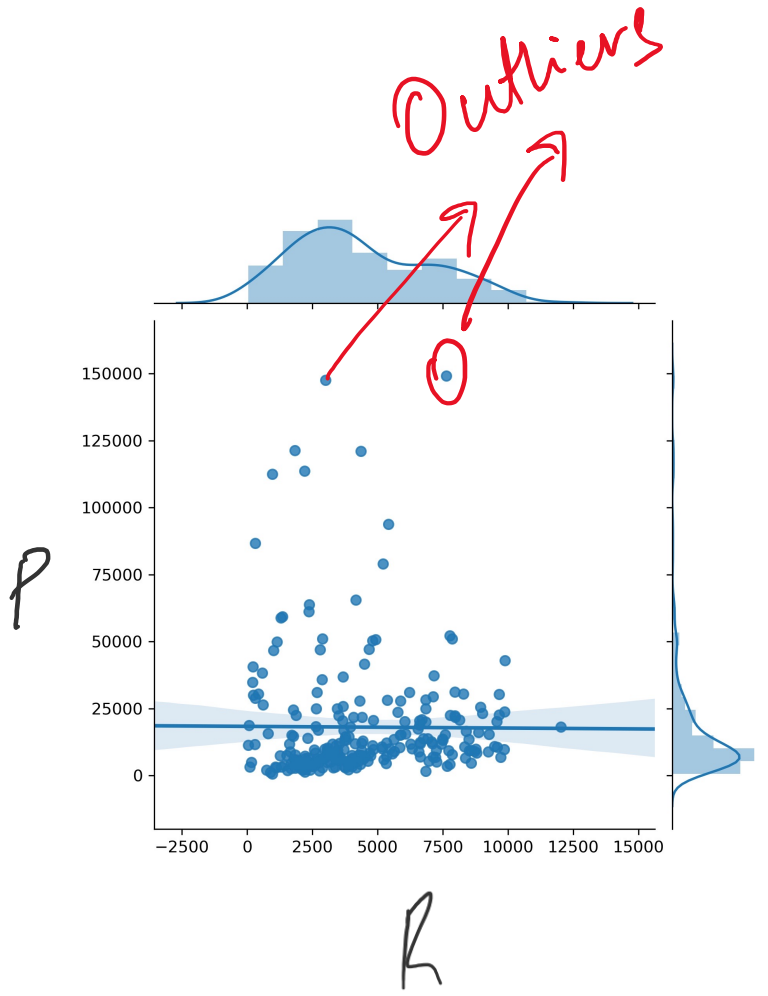
- Be extremely patient for vaguely defined problems
- Concretely reduce the scope of the idea
  - Data Samples are essential
  - Real-life case studies
  - Measures for success
- Ill-defined and unrealistic? Go to the beach

Data Understanding & Preparation

Data Understanding

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

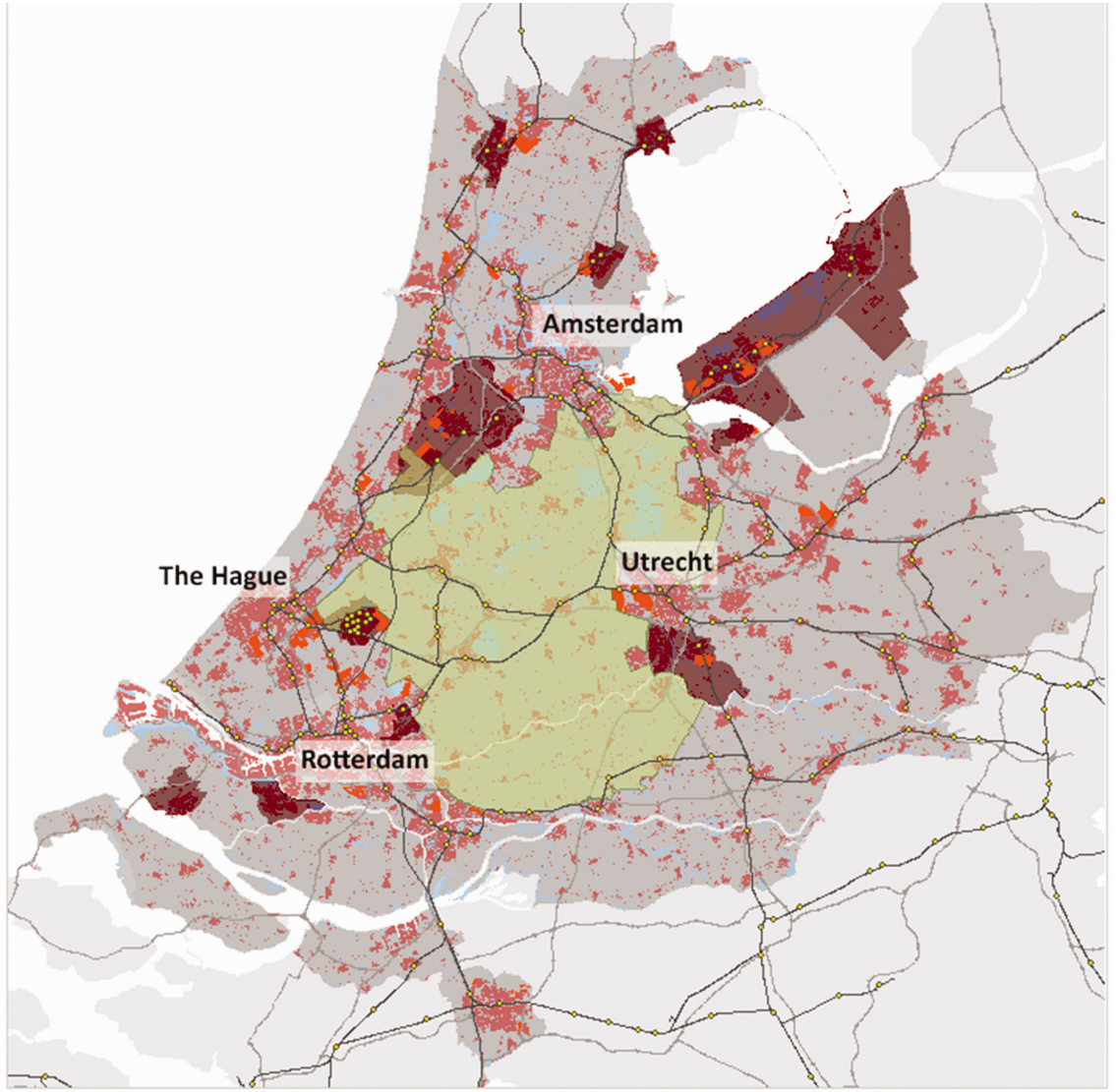




Data Understanding

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



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)

Problem Understanding

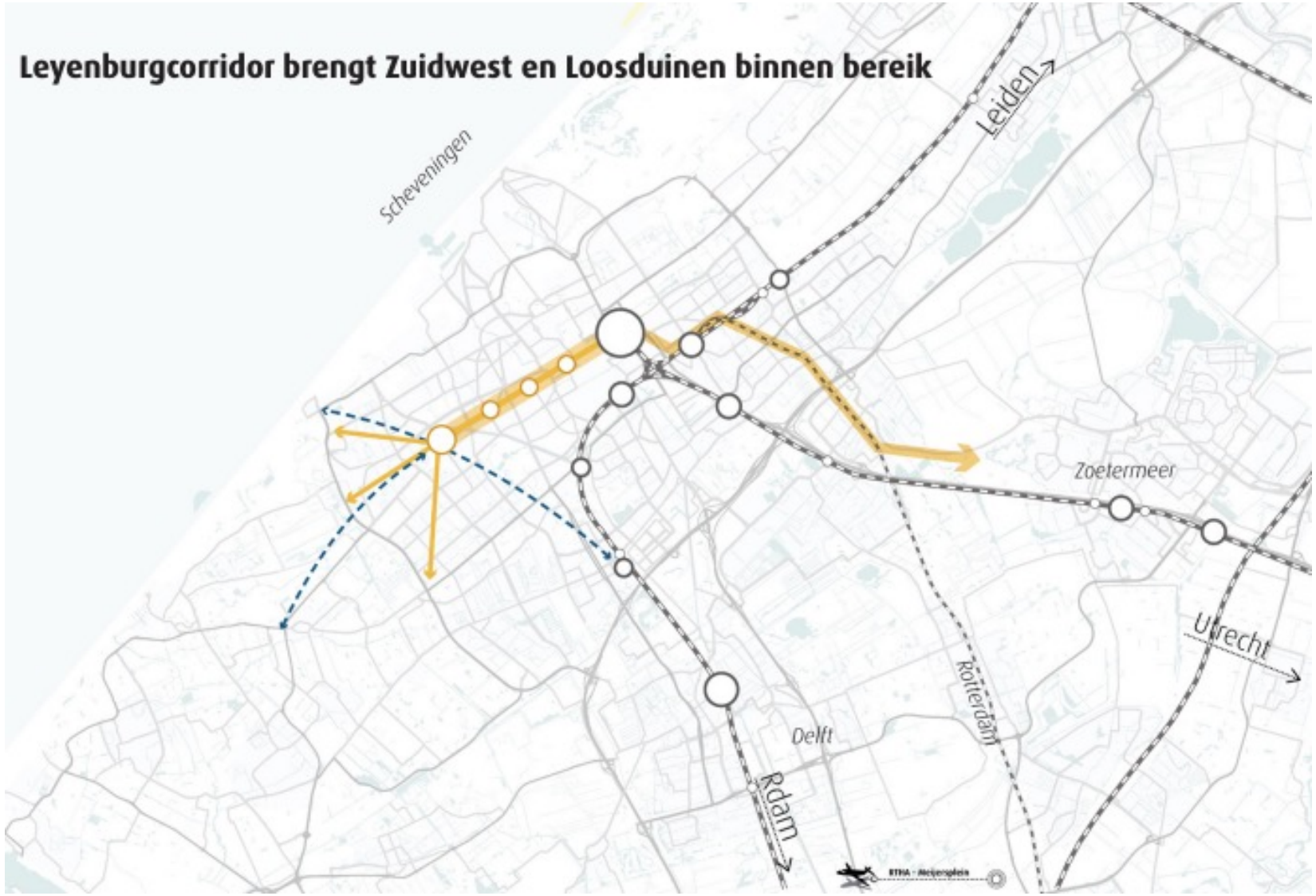
Data Understanding

Data Preparation

Data Preparation

My DOs and DON'TS

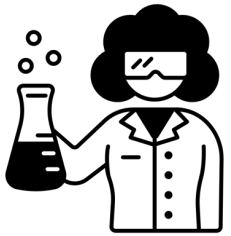
- Automate this step as much as possible – new data / new case
- When merging sources, track data origin
- **Document everything!** Create a workflow
- Manage the stakeholders' expectations that these tasks will take roughly 50% of your time



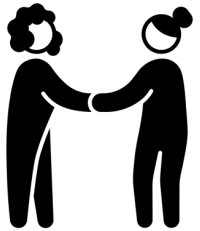
Break

CHILL

WALK



COFFEE OR TEA



MAKE FRIENDS

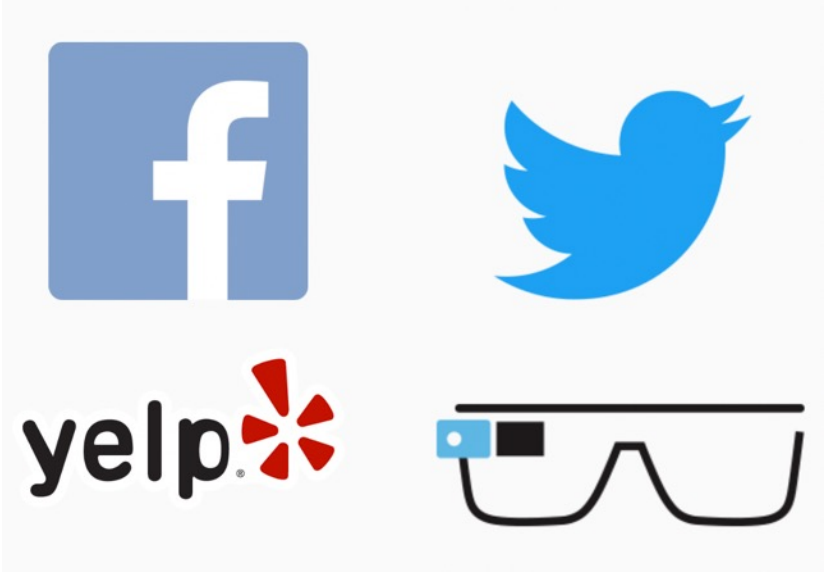
Icons by Iconathon on The Noun Project

What are Data?

What are Data?

“A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data are multiple such measurements.”

Claim: everything is (can be) data!



Where do data come from?

- **Internal sources**: already collected by or is part of the overall data collection of your organization.
  For example: business-centric data that is available in the organization data base to record day to day operations; scientific or experimental data.
- **Existing External Sources**: available in ready to read format from an outside source for free or for a fee.
  For example: public government databases, stock market data, Yelp reviews, [your favorite sport]-reference.
- **External Sources Requiring Collection Efforts**: available from external source but acquisition requires special processing.
  For example: data appearing only in print form, or data on websites.

Ways to gather Online data

How to get data generated, published or hosted online:

- **API (Application Programming Interface):** using a prebuilt set of functions developed by a company to access their services. Often pay to use. For example: Google Map API, Facebook API, Twitter API
- **RSS (Rich Site Summary)**: summarizes frequently updated online content in standard format. Free to read if the site has one. For example: news-related sites, blogs
- **Web scraping**: using software, scripts or by-hand extracting data from what is displayed on a page or what is contained in the HTML file (often in tables).

Web Scraping

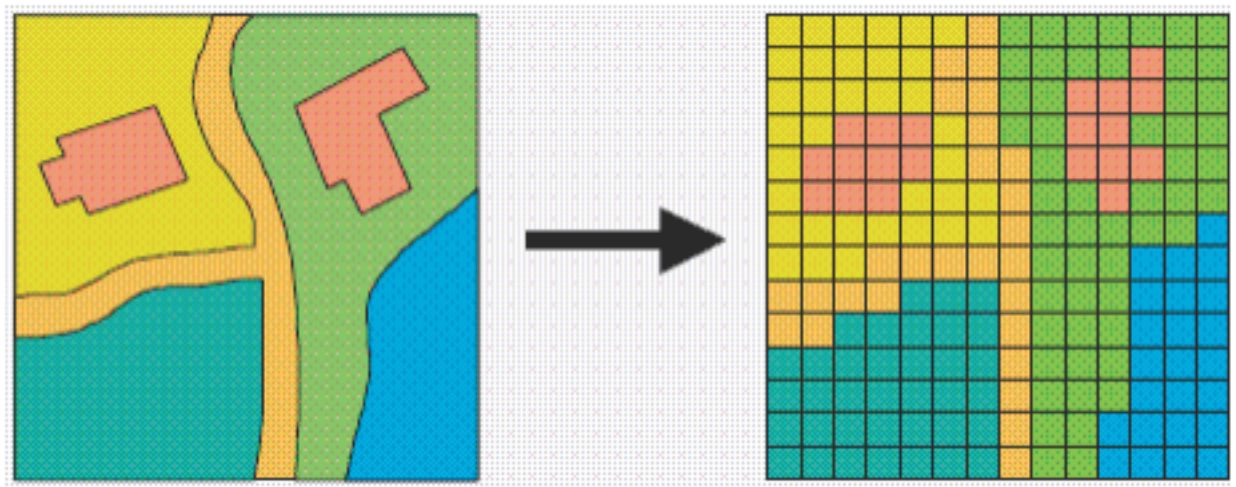
- **Why do it?** Older government or smaller news sites might not have APIs for accessing data or publish RSS feeds or have databases for download. Or, you don't want to pay to use the API or the database.
- **How do you do it?** (beautifulsoup / python package) – some material in lab 1
- **Should you do it?**
  - You just want to explore: Are you violating their terms of service? Privacy concerns for website and their clients?
  - You want to publish your analysis or product: Do they have an API or fee that you are bypassing? Are they willing to share this data? Are you violating their terms of service? Are there privacy concerns?

Representing the world digitally

GIS Data

Traditionally, geographic information is represented as:

- Vector finite set of entities (shapes/geometries)
- Raster images encoding surfaces (values, colours, etc.)



One of the most common types of raster data is land cover derived from satellite imagery. Land-cover data is produced by assigning each pixel in a Landsat thematic mapper image to one of 16 land-cover classes using a procedure known as unsupervised classification.

Raster Basics provided as a tutorial by ArcGIS

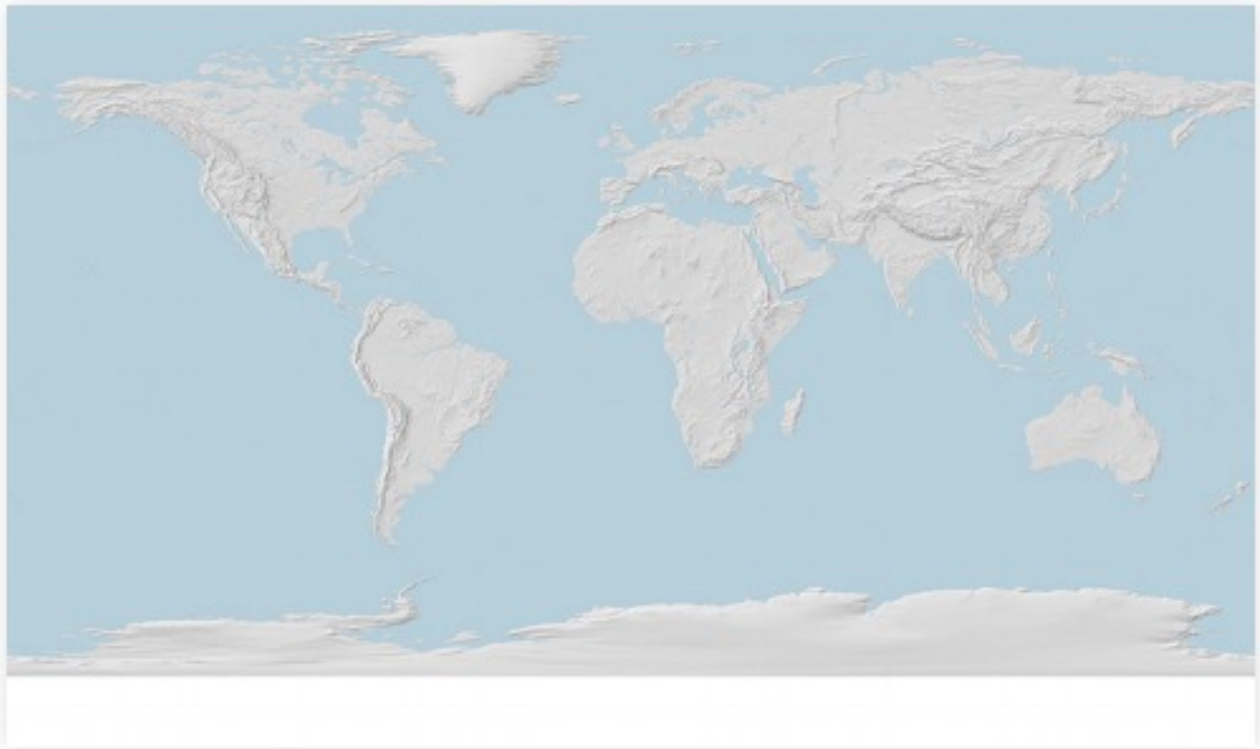
Example

NLCD 2016 Land Cover for the conterminous United States represented as 16 land cover classes.

Vector

Images taken from: Arribas-Bel, D. (2019). A course on geographic data science. *Journal of Open Source Education*, *2*(16), 42.

Raster



Images taken from: Arribas-Bel, D. (2019). A course on geographic data science. *Journal of Open Source Education*, *2*(16), 42.

Good old Spatial Data

[source]

Good old Data (+)

Traditionally, datasets used in the (social) sciences are:

- Collected for the purpose –> carefully **designed**
- Detailed in information (*"…rich profiles and portraits of the country…"*)
- **High quality**

Good old Data (-)

But also:

- Massive enterprises (*"…every single person…"*) –> **costly**
- **Coarse** in resolution (to preserve privacy they need to be aggregated)
- **Slow**: the more detailed, the less frequent they are available

Examples

- Decenial census (and census geographies)
- Longitudinal surveys
- Customly collected surveys, interviews, etc.
- Economic indicators
- …

New sources of *Spatial* Data

New Sources of *Spatial* Data

New sources are appearing that are:

- **ACCIDENTAL** –> created for different purposes but available for analysis as a side effect
- Very **diverse** in nature, resolution, and detail but, potentially, much more detailed in both space and time
- **Quality** also varies greatly

Different ways to categorise them…

Lazer and Radford (2017)

- Digital life: digital actions (Twitter, Facebook, WikiPedia…)
- Digital traces: record of digital actions (CDRs, metadata…)
- Digitalised life: nonintrinsically digital life in digital form (Government records, web…)

Arribas-Bel (2014)

Three levels, based on how they originate:

- Bottom up: “Citizens as sensors”
- Intermediate: Digital businesses/businesses going digital
- Top down: Open Government Data (The Hague Cijfers)

Class Quiz

Class Quiz

What is the origin of the following sources of (geo-)data:

- Geo-referenced tweets –>
- Land-registry house transaction values –>
- Google maps restaurant listing –>
- ONS Deprivation Indices –>
- Liverpool bikeshare service station status –>

Citizens as Sensors

- Technology has allowed widespread adoption of sensors (bands, smartphones, tablets…)
- (Almost) every aspect of human life is subject to leave a digital trace that can be collected, stored and analyzed
- Individuals become content/data creators (sensors, Goodchild, 2007)
- Why relevant for geographers? –> Most of it (80%?) has some form of spatial dimension

Example: Free To Be

GIRLS' SAFETY WALKS

As part of the Unsafe in the City report launch in Sydney, we facilitated a young women-led Girls' Safety Walk on International Day of the Girl 2018.

Our youth activists took key stakeholders on an immersive walk around the city, looking at 'hot spots' and leading activities based on the data and young women's stories.

These walks give planners, decision makers and local leaders a glimpse into the experiences of girls and young women as they move around our cities. By drawing attention to the themes from the Free to Be data in a practical way, it helps participants identify how their work can influence the experiences of women and girls.

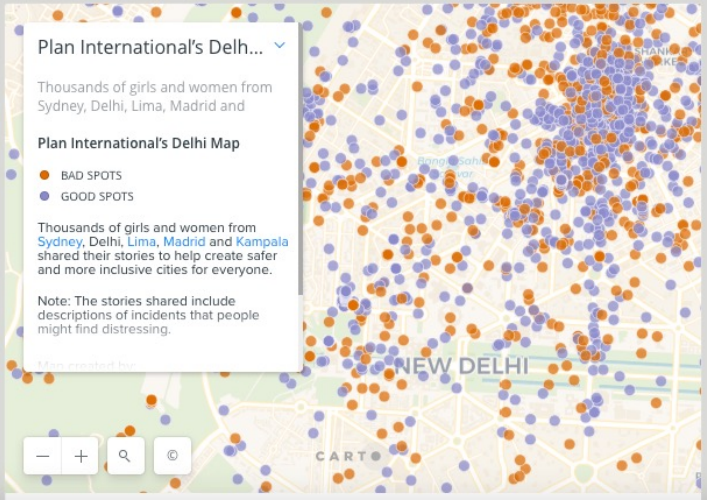
FREE TO BE MAPS

What girls from Sydney, Australia have to say about their city

Explore the map below

Explore the map or pick another city below

Delhi | Kampala | Lima | Madrid | Sydney



OUR REPORTS: UNSAFE IN THE CITY

Through this ground-breaking research, thousands of girls and young women have shared their stories of harassment and violence for the first time, providing a never-before seen glimpse of what they experience in their cities and the impact this has on their lives. Based on research in Delhi, Kampala, Lima, Madrid and Sydney, Unsafe in the city reveals relentless sexist and sexual harassment and abuse – and calls for specific actions to allow girls and young women to live without experiencing fear or discrimination on our streets.

READ THE REPORT

REPORTING TO AUTHORITIES

Utilising data collected via our Safer Cities Free to Be project in 2018 with Monash XYX Lab and Crowdspot, the 'Reporting to Authorities' report is a major study on street harassment, and urges authorities to improve reporting outcomes for girls and young women. Across five cities, the report found that young women's reports of harassment were largely trivialised by authorities, with responses ranging from belittling, disbelief and dismissal, to further harassment from authorities themselves and a complete lack of justice, resulting in frustration and a lack of trust in the system.

READ THE REPORT

https://www.plan.org.au/you-can-help/join-the-movement-for-girls-rights/free-to-be/

Businesses moving online

- Many of the elements and parts of business activities have been **computerized** in the last decades
- This implies, without any change in the final product or activity per se, a lot more digital data is “available” about their operations
- In addition, entirely new business activities have been created based on the new technologies (“**internet natives**”)
- Much of these data can help researchers better understand how cities work

Example: Picnic

https://www.walkscore.com/

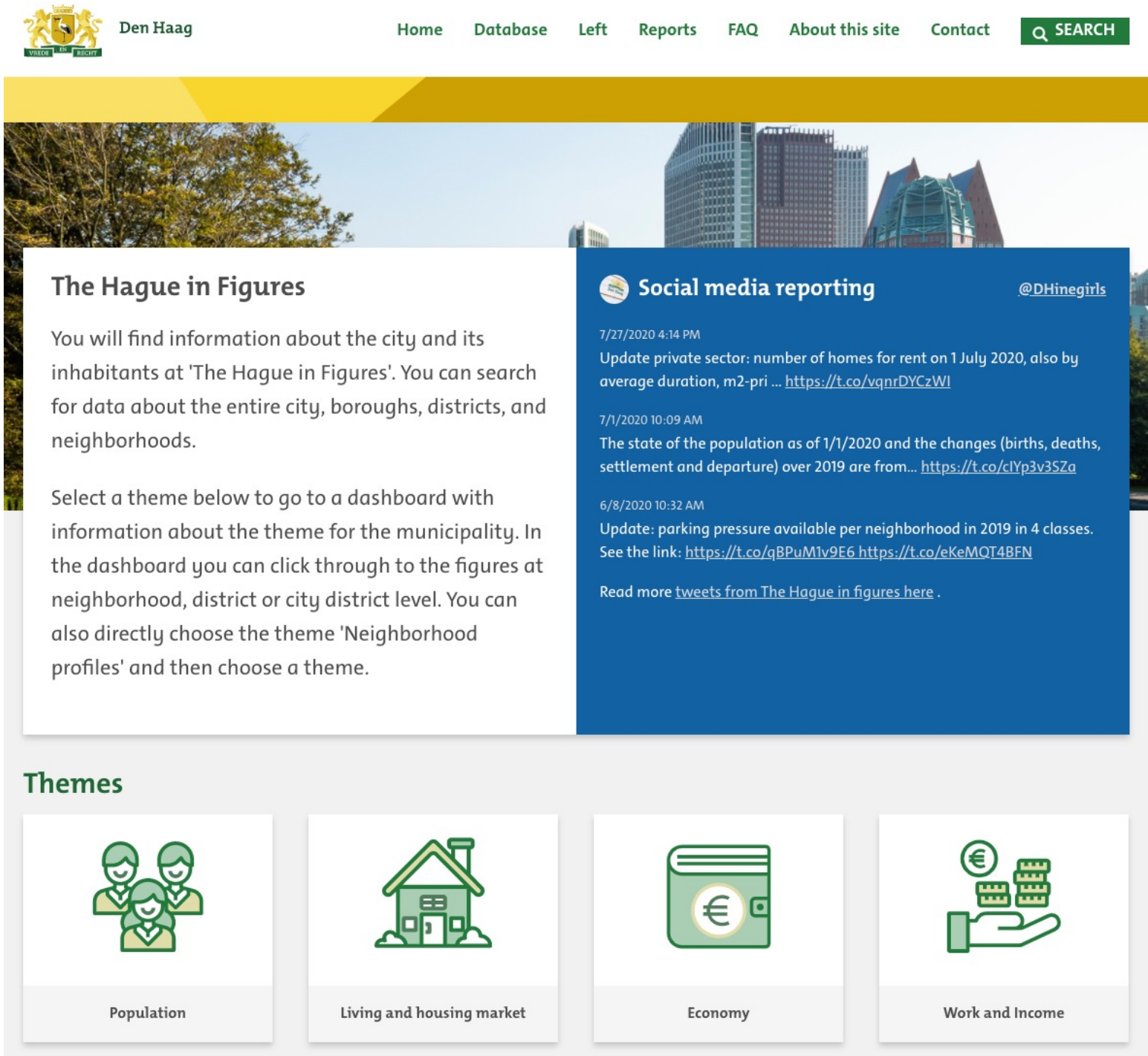
Open Data for Open Governments

Government institutions release (part of) their internal data in open format. Motivations (Shadbolt, 2010):

- Transparency and accountability
- Economic and social value
- Public service improvement
- Creation of new industries and jobs

Example: The Hague Cijfers

https://denhaag.incijfers.nl/



Opportunities and Challenges

Opportunities (Lazer & Radford, 2017)

- Massive, passive
- Nowcasting
- Data on social systems
- Natural and field experiments (“ always-on” observatory of human behaviour)
- Making big data small

Challenges (Arribas-Bel, 2014)

- Bias
- Technical Barriers
- Methodological “mismatch”

Bias

- Traditional data meet some quality standards (representativity, accuracy…)
- Because they're *accidental*, new data sources might not
- Researchers need to have extra care and put more thought into what conclusions they can reach from analyses with new sources of data
- In some cases, bias can run in favour of researchers, but this should never be taken for granted

Technical barriers to access

- Much of these data are available
- However, their accidental nature makes them *difficult* to access
- Usually, **a different set of skills** is required to tap into their power
  - Basic programming
  - Computing literacy (understanding of the internet, APIs, databases…)
  - Software savvy-ness (a.k.a. “go beyond Word and Excel”)

New Methods

The nature of these data is not the same as that of more traditional datasets. For example:

- Spatial aggregation: Polygons Vs. Points
- Temporal aggregation(frequency): Decadal Vs. Real-time

Some of this does not “play well” with techniques employed traditionally to analyse data in Geography or any other discipline –> borrow techniques from other disciplines, or even create new ones

New + Old

Traditional data:

- High quality, detailed, and reliable
- Costly, coarse, and slow

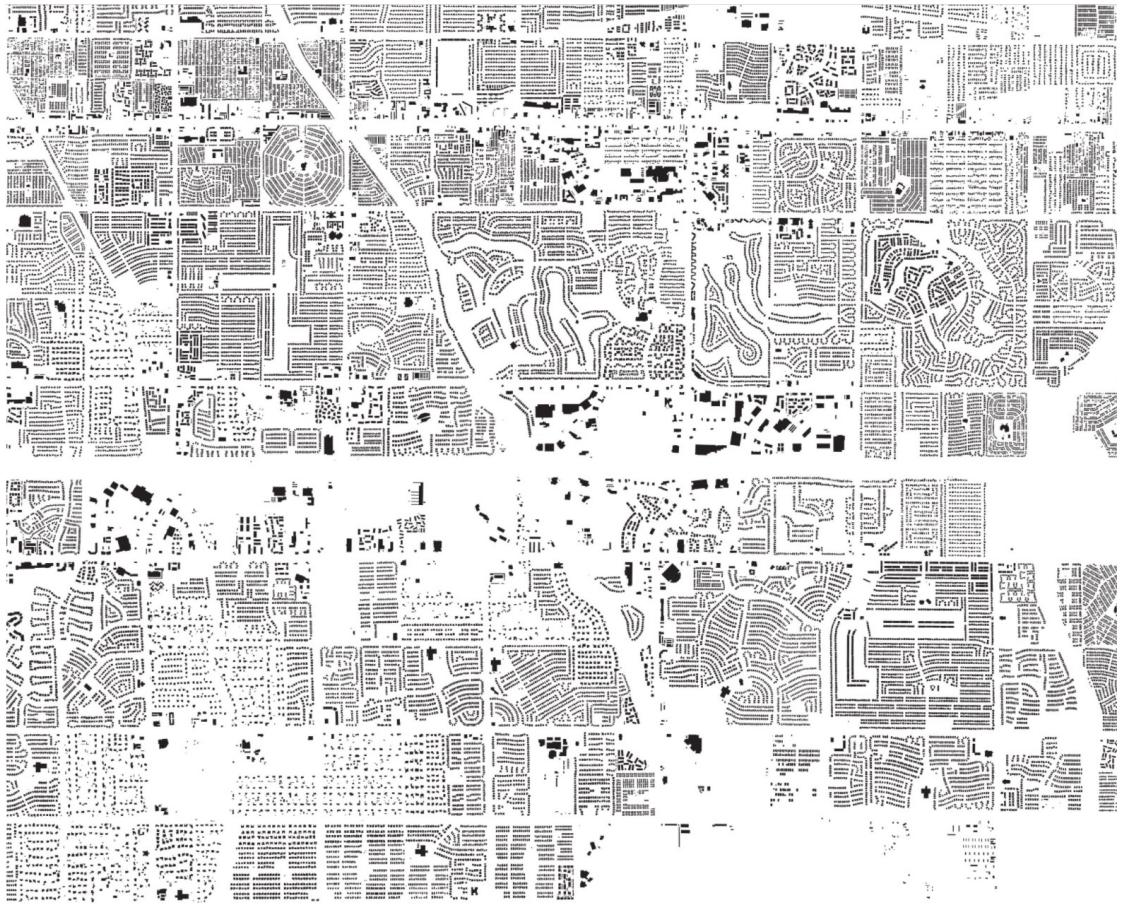
Accidental data:

- Cheap, fine-grained, and fast
- Less reliable, harder to access, and potentially uninteresting

Old/New, raster/vector . . .

Traditional approaches to represent the world in a computer are blending thanks to new forms of data

Keep an open mind to tools, approaches, and methods



A NATION OF SUBURBS

Mesa, Ariz. America's suburban streets twist and flow, with their wild involutions and curving cul-de-sacs. Mesa's suburbs are especially imaginative, particularly from above. The feeling of meandering through a place whose layout is designed to thwart speed and comprehension is familiar to anyone who, in the days before GPS, needed to pick up a friend or deliver a pizza in an unfamiliar neighborhood.

An interactive NYT article showing a map of every building in America

EUROPEAN COMMISSION

Global Human Settlement

European Commission > EU Science Hub > GHSL

Home | About | Copernicus | GEO | Documents | Atlases | Applications | Degree of Urbanisation | Data | Tools | Visualisation | News

GHSL - Global Human Settlement Layer

A new open and free tool for assessing the human presence on the planet

- Produces new global spatial information, evidence-based analytics and knowledge describing the human presence on the planet
- Operates in an open and free data and methods access policy (open input, open method, open output)
- Supported by the Joint Research Centre (JRC) and the DG for Regional and Urban Policy (DG REGIO) of the European Commission, together with the international partnership GEO Human Planet Initiative GEO GROUP ON EARTH OBSERVATIONS

News ***26/03/2020*** ***Call for Contribution to the JRC Atlas of the Human Planet 2020*** *it will showcase applications of the GHSL data. Go to our* ***news page*** *for the details*

Global Human Settlement Layer:
https://ghsl.jrc.ec.europa.eu/index.php

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