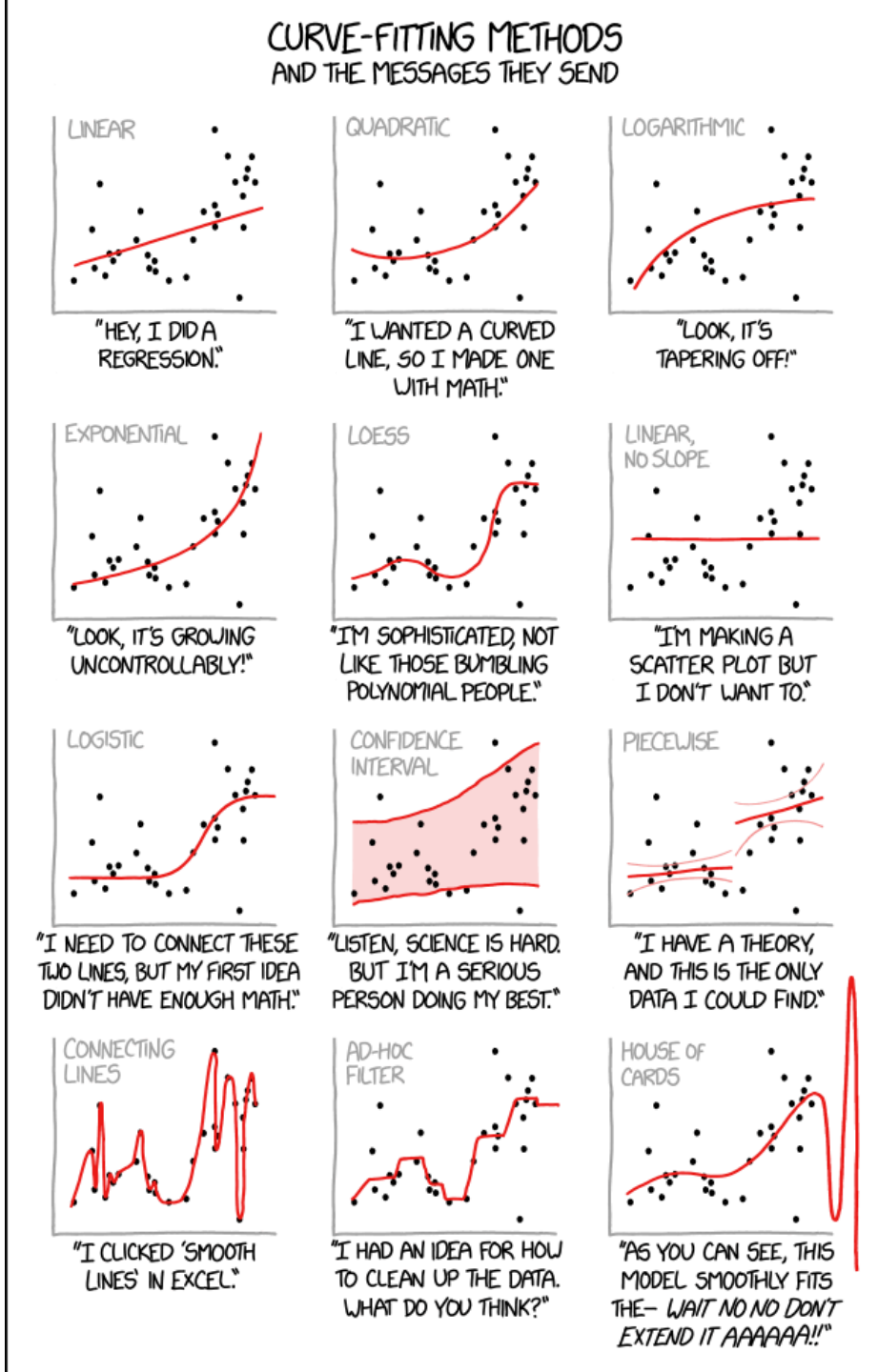
Introduction to *Urban* Data Science

Data Engineering

(EPA1316)

Lecture 4

Trivik Verma

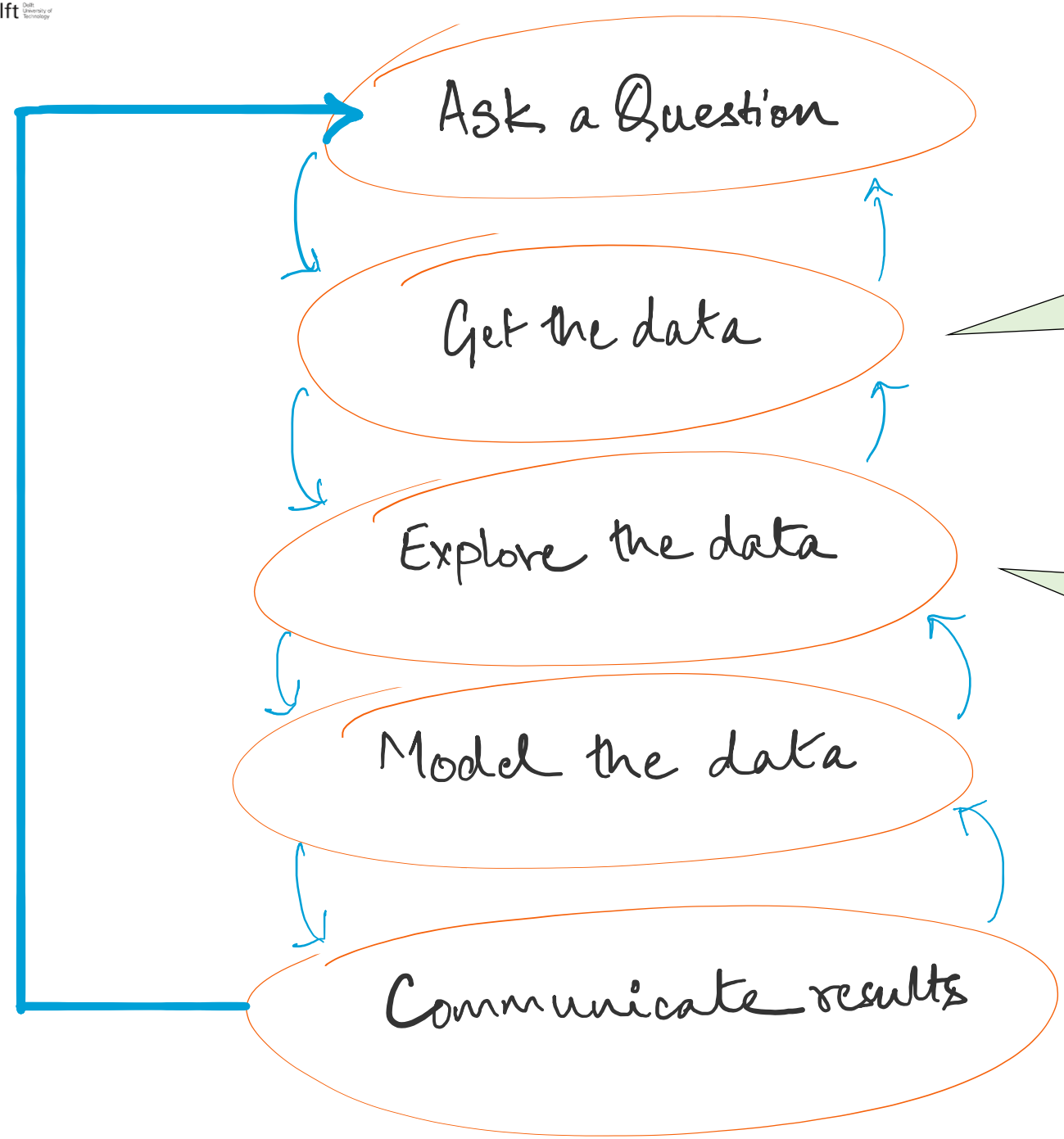


Last Time

- Types of Data
- Grammar
- EDA without Pandas
- EDA with Pandas
- Data Concerns

Today

- Descriptive Statistics
- Data Transformations



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

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

Descriptive Statistics

Basics of Sampling

Population versus sample:

- A **population** is the entire set of objects or events under study. Population can be hypothetical “all students” or all students in this class.
- A **sample** is a “representative” subset of the objects or events under study. Needed because it’s impossible or intractable to obtain or compute with population data.

Biases in samples:

- **Selection bias**: some subjects or records are more likely to be selected
- Volunteer/**nonresponse bias**: subjects or records who are not easily available are not represented

Examples?

Sample mean

- The **mean** of a set of *n* observations of a variable is denoted $\bar{x}$ and is defined as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The mean describes what a “typical” sample value looks like, or where is the “center” of the distribution of the data.
- **Important** : there is always uncertainty involved when calculating a sample mean to estimate a population mean.

Sample median

- The **median** of a set of n number of observations in a sample, ordered by value, of a variable is defined by

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2} + x_{(n+1)/2}}{2} & \text{if } n \text{ is even} \end{cases}$$

- Example (already in order):

  Ages: 17, 19, 21, 22, 23, 23, 23, 38

  Median = (22+23)/2 = 22.5

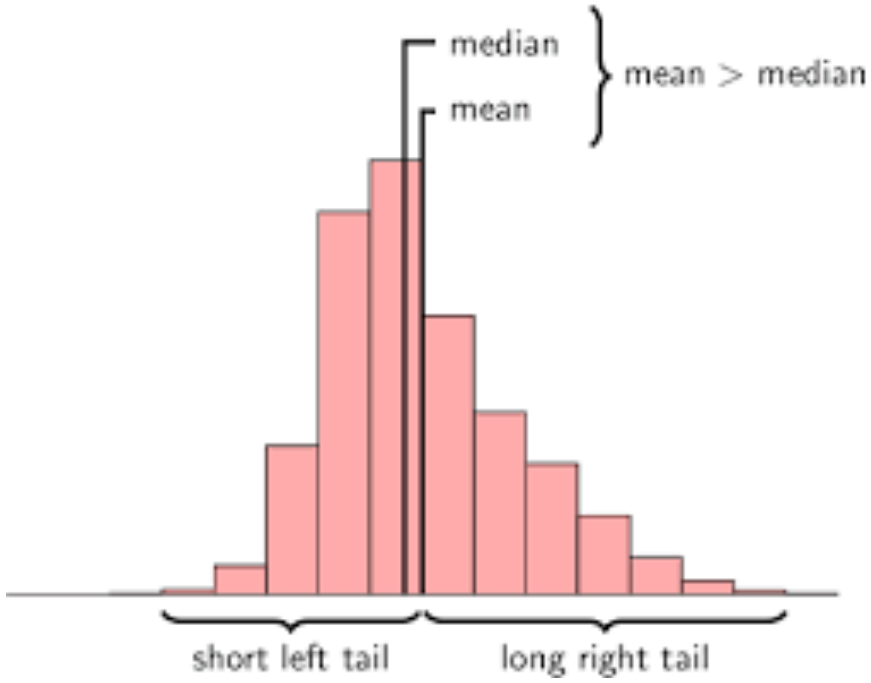
- The median also describes what a typical observation looks like, or where is the center of the distribution of the sample of observations.

Mean vs Median

The mean is sensitive to extreme values **(outliers)**

Mean, median, and skewness

The mean is sensitive to outliers:

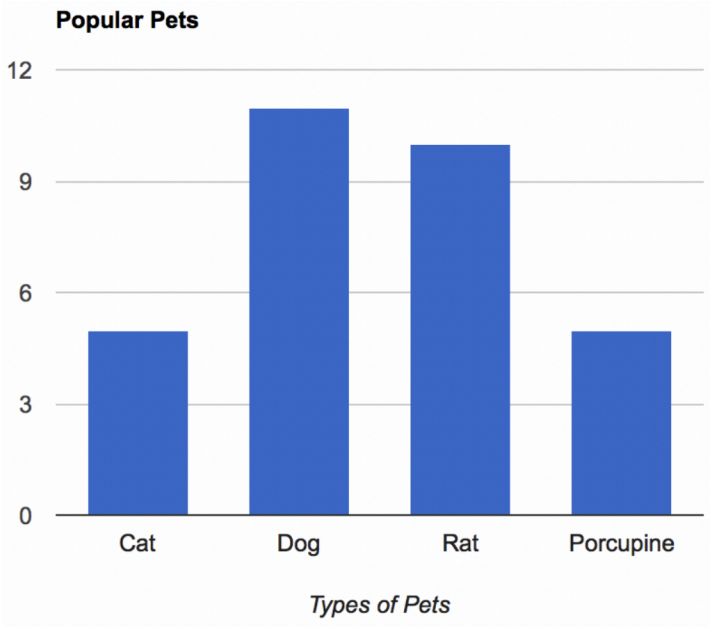


The above distribution is called **right-skewed** since the mean is greater than the median.

Note: **skewness** often “follows the longer tail”.

Regarding Categorical Variables…

For categorical variables, neither mean or median make sense. Why?



The mode might be a better way to find the most “representative” value.

Measures of Spread: Range

The spread of a sample of observations measures how well the mean or median describes the sample.

One way to measure spread of a sample of observations is via the **range**.

Range (R) = (Max)imum Value - (Min)imum Value

Measures of Spread: Variance

- The (sample) **variance**, denoted s^2, measures how much on average the sample values deviate from the mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2$$

- Note: the term $|x_i - \bar{x}|$ measures the amount by which each x_i deviates from the mean $\bar{x}$. Squaring these deviations means that s^2 is sensitive to extreme values (outliers).
- Note: s^2 doesn't have the same units as the x_i :(
- What does a variance of 1,008 mean? Or 0.0001?

Measures of Spread: Standard Deviation

The (sample) **standard deviation**, denoted *s (or sigma)*, is the square root of the variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2}$$

Note: *s* does have the same units as the x_i. Phew!

Break

WATER

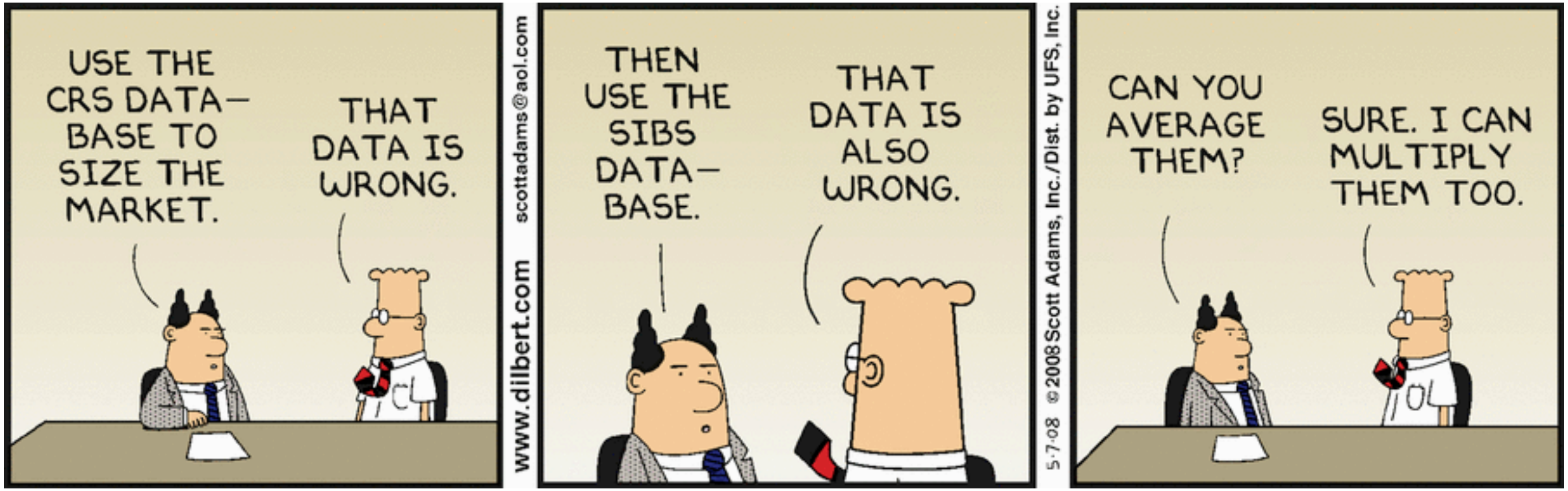
WALK

COFFEE OR TEA

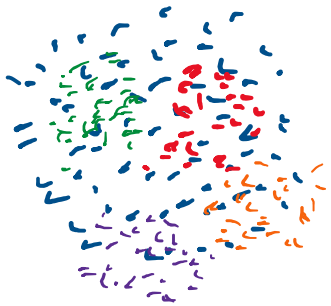
MAKE FRIENDS

Data Transformations

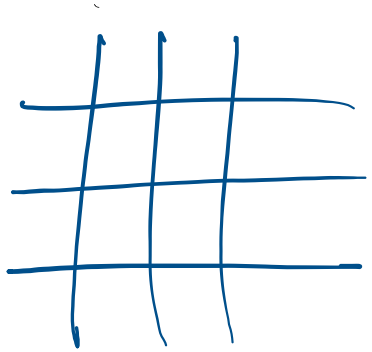
Why Transform Data



feature engineering



RAW



TABULAR

features

objects

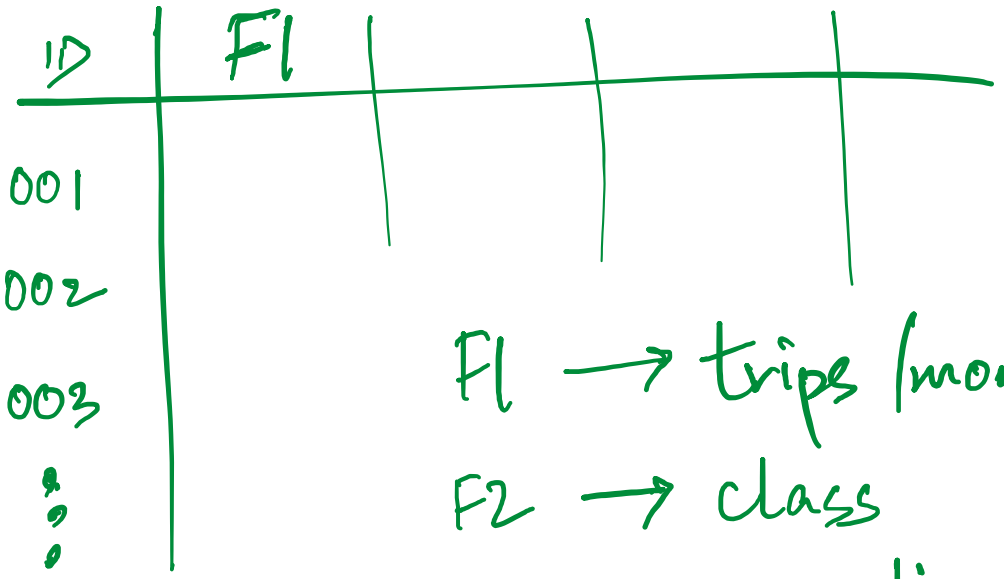
	f_1	f_2	...
O1			
O2			
⋮			

Based upon Domain knowledge

Eg. SMART-CARD CHECK-IN LOGS

Features (measurable)

* Users



F1 → trips/month

F2 → class

F3 → Avg. time of trip

F4 → total price

⋮

* Alternative: trips/station

Scale Data

F1 F2 F3 F4

Dimensions

1 – 500 trips/month

200 – 2000 CHF/month

Why Scaling

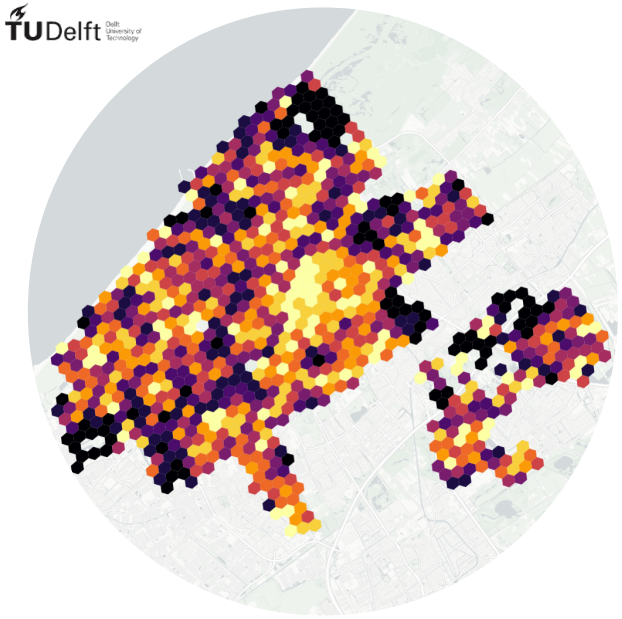
- Comparison of groups of Object

Example: Access to infrastructure in Cities

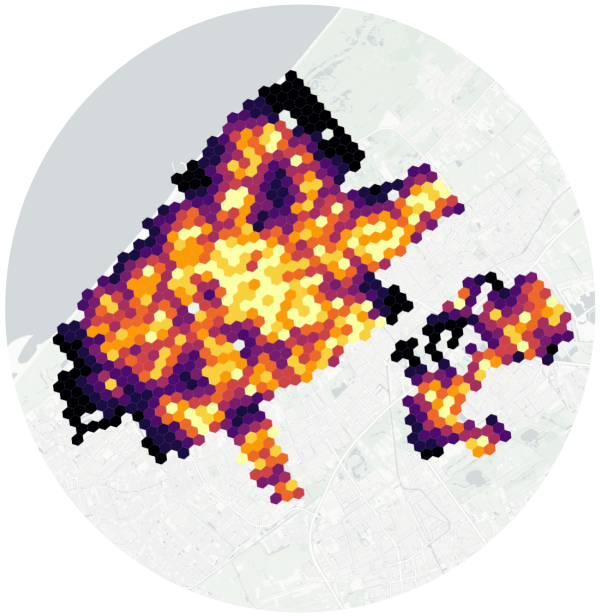
- ML algorithms use Euclidean distance (higher magnitude will weigh more) –

advanced topics will be explored in week 6

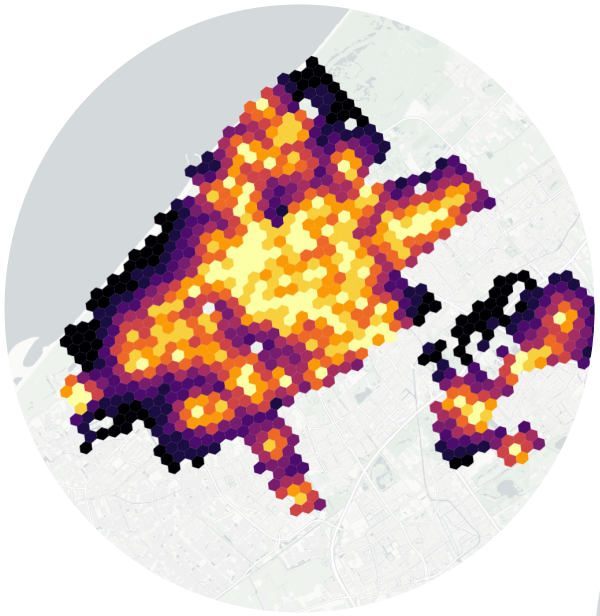
Measure of Access



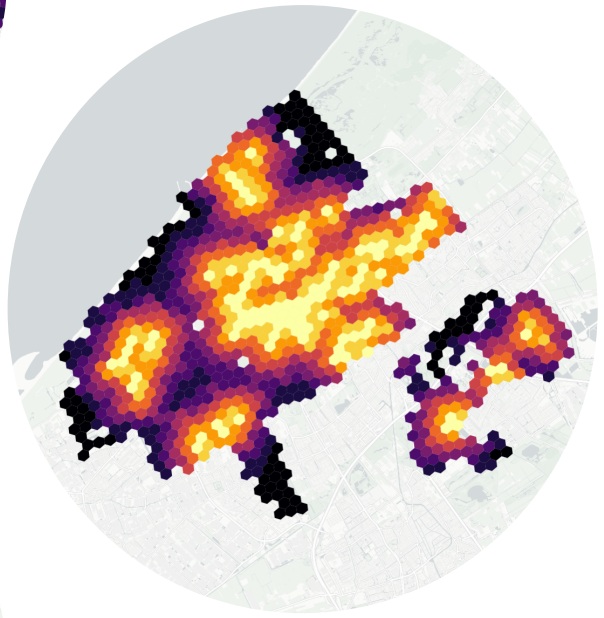
Active Living



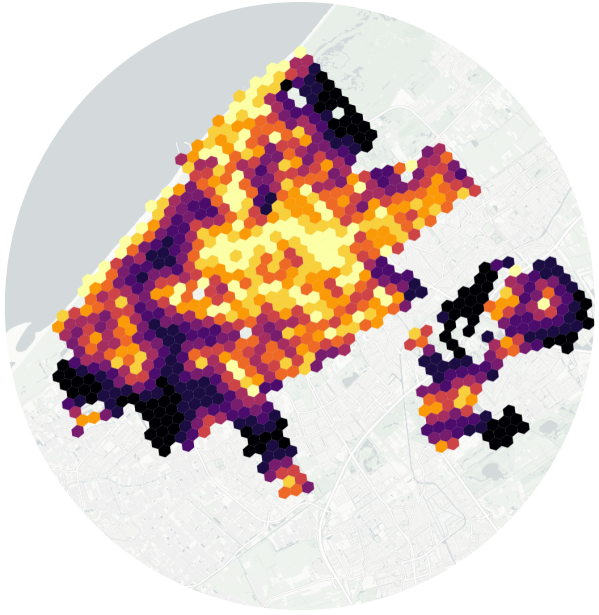
Education



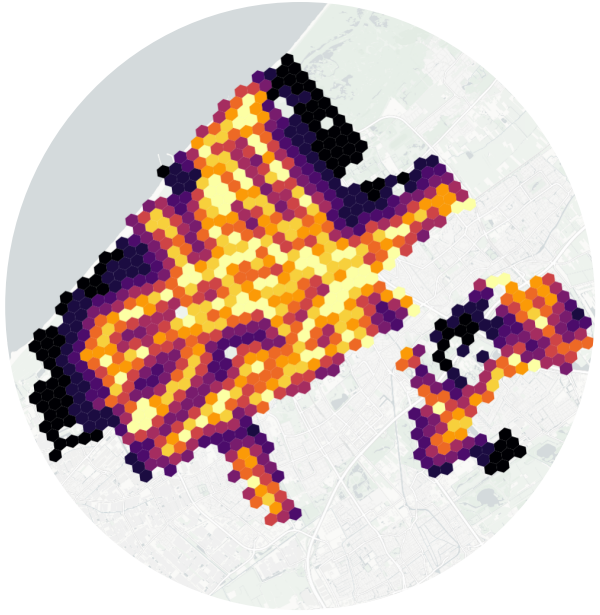
Health and Well Being



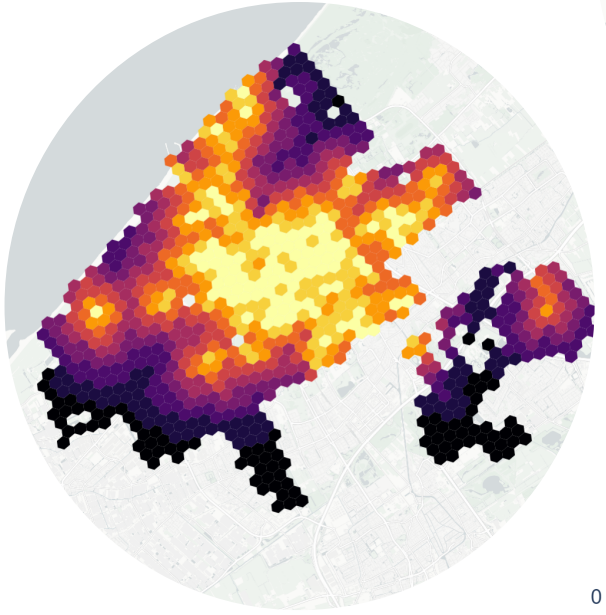
Nightlife



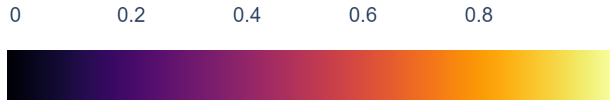
Food Choices



Mobility



Community Space



Dealing with Missing Data

- If your data is big, sacrifice examples with missing features
- Data Imputation techniques
  - Use average of the feature for replacing a missing value $x_i^0 \leftarrow \bar{x}$
  - **Advanced**: regression modelling to estimate missing values

Normalisation

- Transformation of data to a different range [a - b]
- Normally [0-1]
- Create new variables from the transformations.

$$X_i' = \frac{X_i - \min(x)}{\max(x) - \min(x)} \times [b - a] + a$$

Rescaled value

Original value

Min value in feature

New range

Standardisation

or, Z-score normalisation

- Transformation of data to a different range that is normally distributed with mean 0 and standard deviation 1.

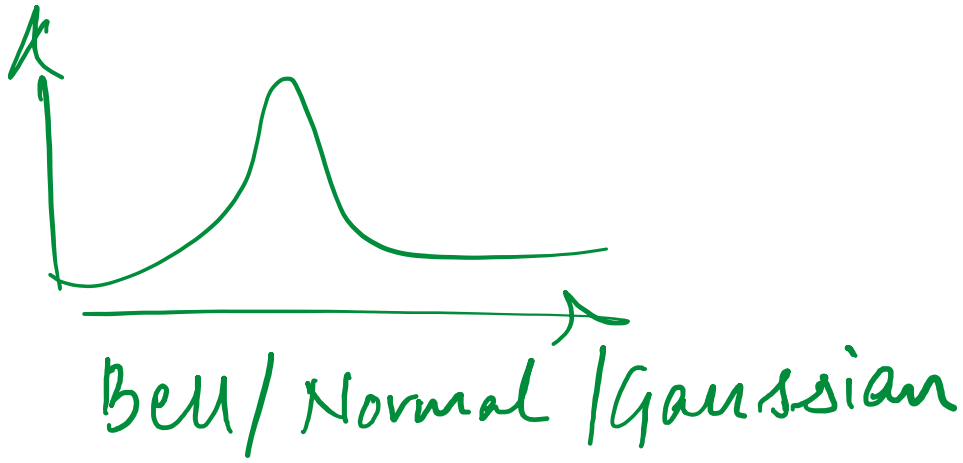
$$N(\mu=0, \sigma=1)$$

Rescaled value

$$X_i' = \frac{X_i - \mu_i}{\sigma_i}$$

Use S (All others N)

- Features are normally distributed (**not normalisation**)



- Many outliers (normalization squashes them in a limited range)
- All unsupervised learning algorithms, like clustering or dimensionality reduction

For next class..

Finish Lab 03 to practice programming

Submit Homework 03 for peer review on Brightspace

Submit 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 **Pandas and Data Features**)