Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

Announcements

- To get a quicker response from me, please email to my CSE email and not to my IITK email:
  - **My CSE email: soumyad@cse.iitk.ac.in**

Acknowledgements

- Some of the following slides are adapted from the excellent course materials made available by:
  - Prof. Klaus Mueller (State University of New York at Stony Brook)
  - Prof. Tamara Munzner (University of British Columbia)

Visual Design and Visual Variables

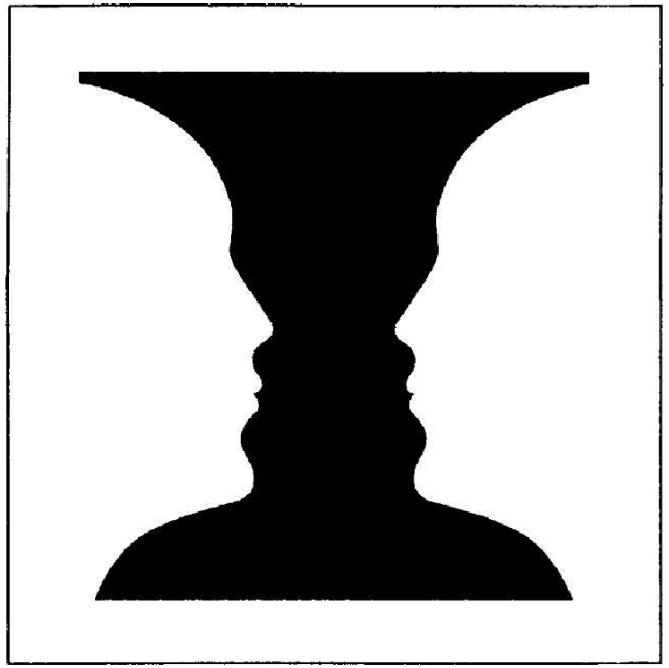
Key Visual Representations

- Gestalt Principles
  - The tendency to perceive elements as belonging to a group, based on certain visual properties
- Pre-attentiveness
  - Certain low level visual aspects are recognized before conscious awareness
- Visual variables
  - The different visual aspects that can be used to encode information

Gestalt Principles

- "Gestalt" is German for "unified whole"
- Grasp the "totality" of something before worrying about the details
- Proximity, similarity, closure, multistability, …

What do you see in this figure?

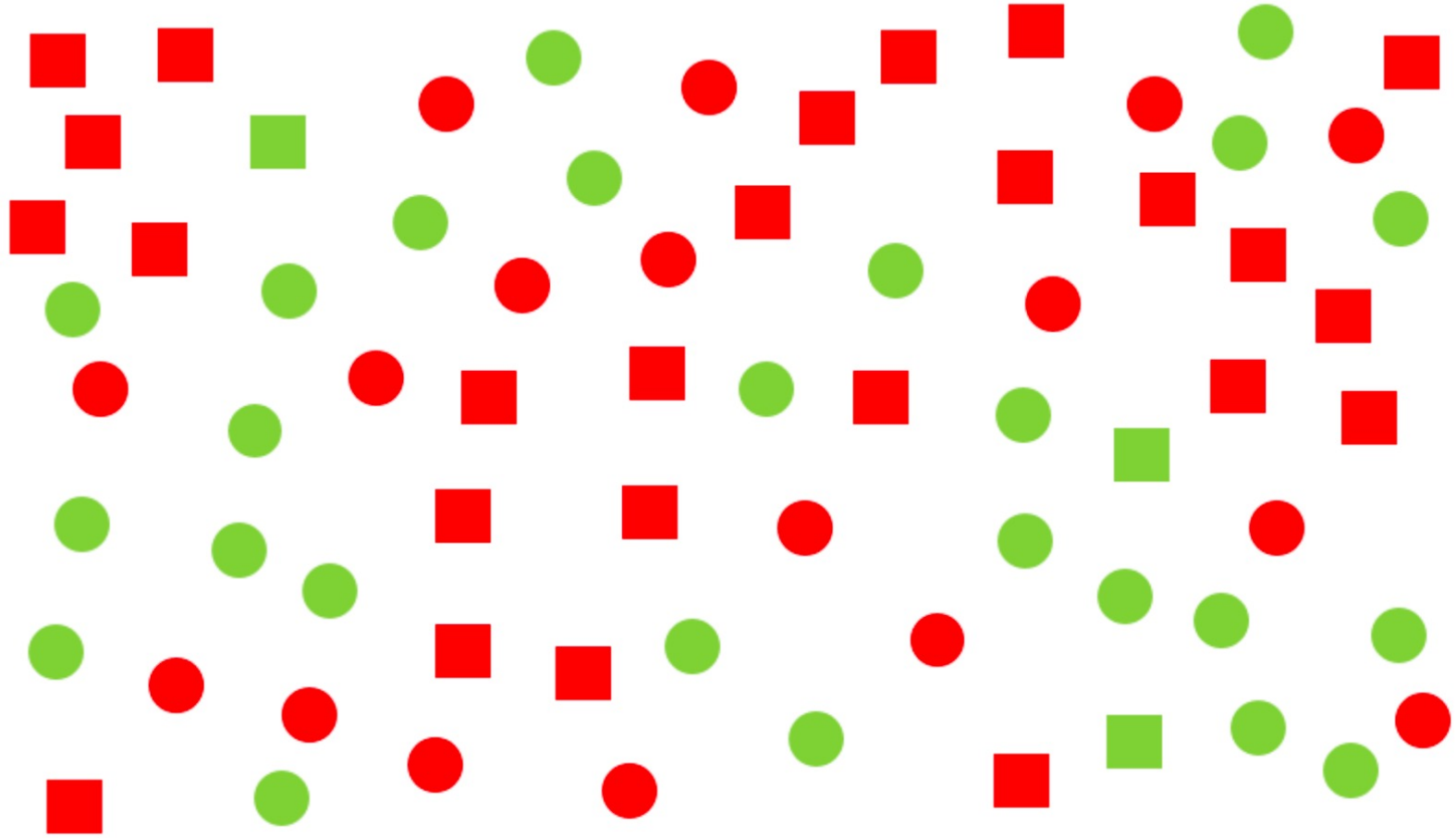


Rubin's vase

What do you see in this figure?

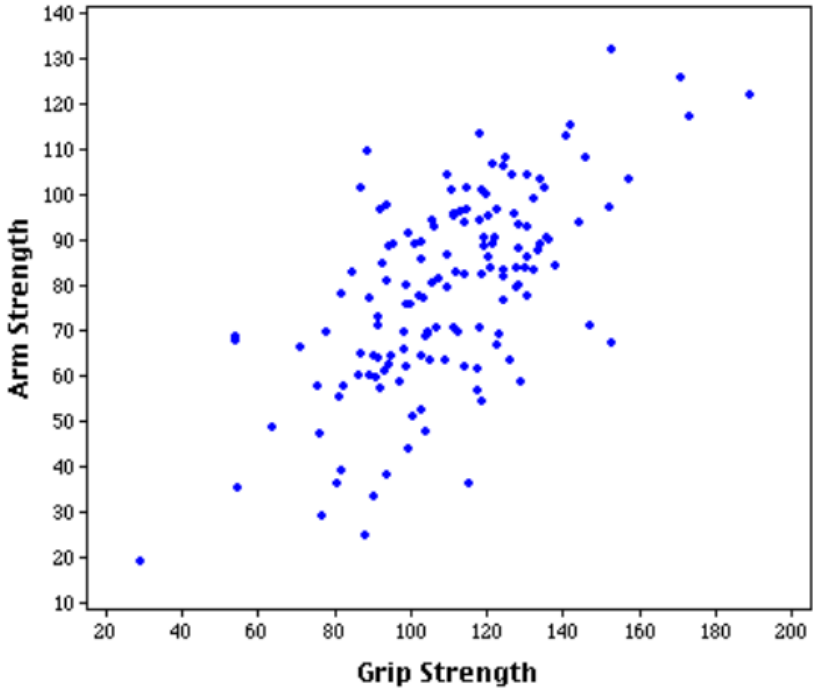
Pre-attentiveness

- Also called pop-out



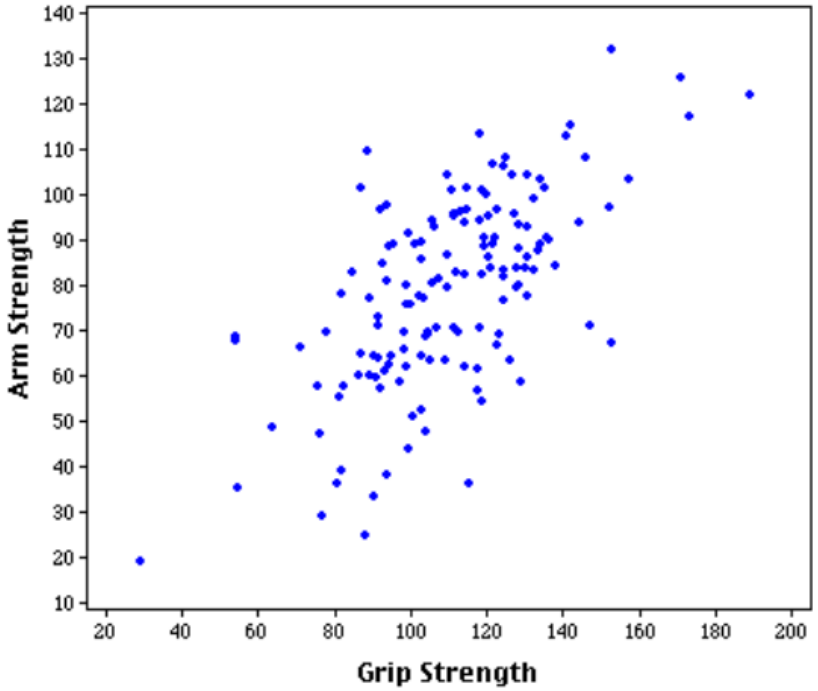
Visual Variables

- Two planar variables
  - Spatial dimensions (X and Y)



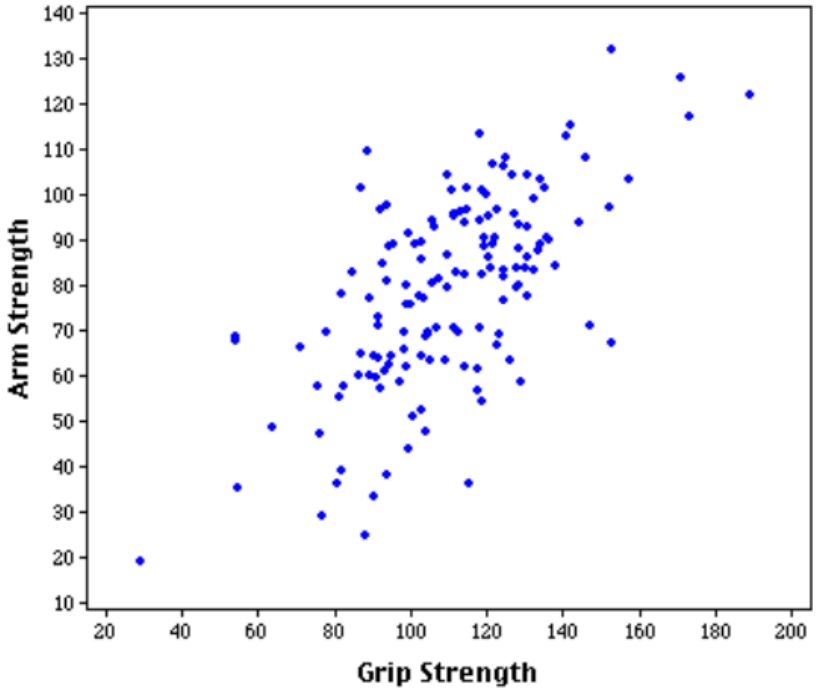
Visual Variables

- Two planar variables
  - Spatial dimensions (X and Y)
- Six Retinal variables
  - Size
  - Color
  - Shape
  - Orientation
  - Texture
  - Brightness

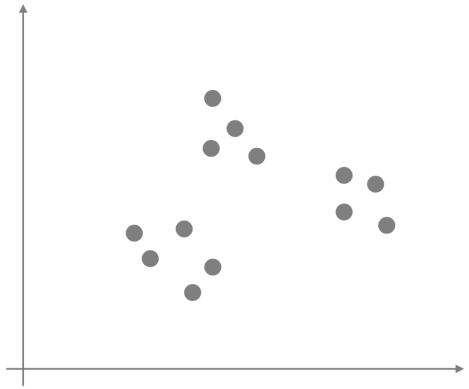


Visual Variables

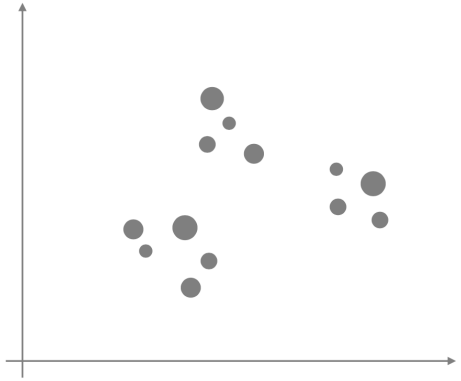
- Two planar variables
  - Spatial dimensions (X and Y)
- Six Retinal variables
  - Size
  - Color
  - Shape
  - Orientation
  - Texture
  - Brightness
- Retinal variables allow for one more variable to be encoded



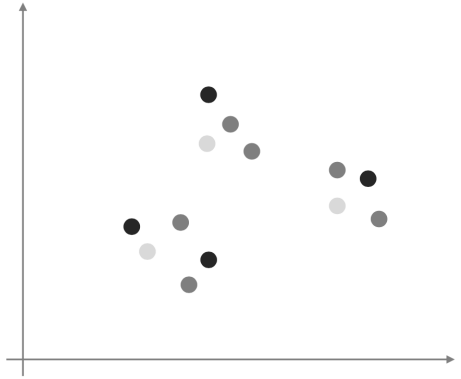
Visual Variables



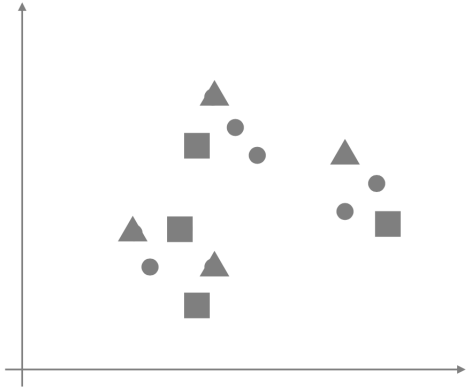
Planar



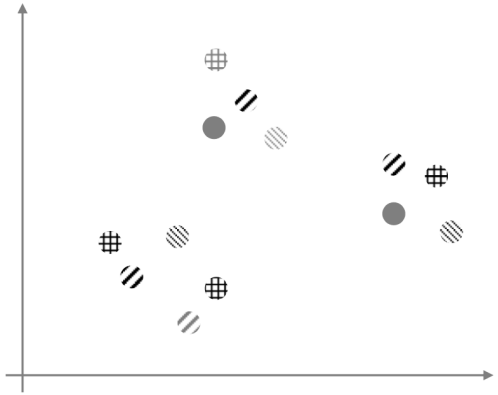
Size



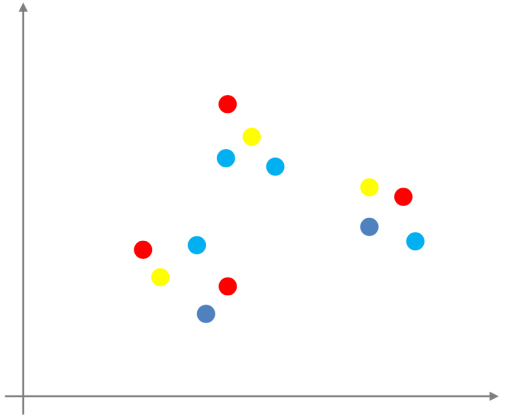
Brightness



Shape



Texture



Color

Orientation

Take Aways

- Planar variable is the strongest visual variable
  - Maps to proximity
  - Provides an intuitive organization of information
  - Things close together are perceptually grouped together (Gestalt)
- Size and brightness are good secondary visual variables to encode relative magnitude
- Color is a good visual variable for labeling
  - Texture can do this as well, but it does not support pop-out much
- Shape provides only limited pop-out

Considerations with Scalability for Big Data

- Must be scalable to
  - Number of data points
  - Number of dimensions
  - Data sources
  - Diversity of data sources (heterogeneity)
  - Number of users

Considerations with Scalability for Big Data

- Must be scalable to
  - Number of data points
  - Number of dimensions
  - Data sources
  - Diversity of data sources (heterogeneity)
  - Number of users

Visual Analytics can help!

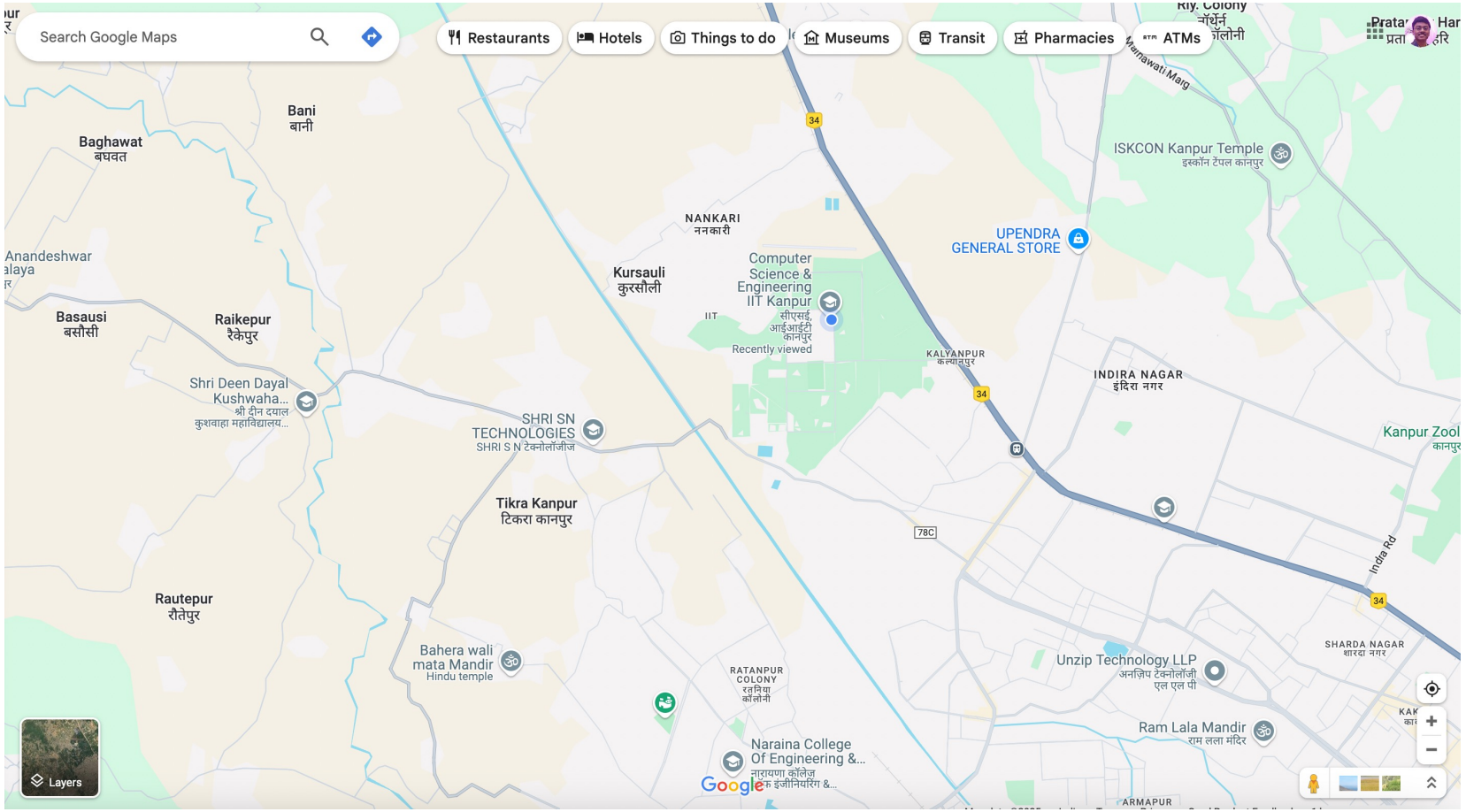
What is Visual Analytics

- Visualization plus...
  - Data processing (analytics)
  - Intelligent computing (AI, machine learning)
  - Interaction (HCI)
  - Pattern discovery
  - Storytelling and sensemaking
  - Behavioral psychology (cognitive science, human factors)

Visual Analytics is the process of analytical reasoning often supported by a highly interactive visual interface/tool

Visual Information Seeking Mantra

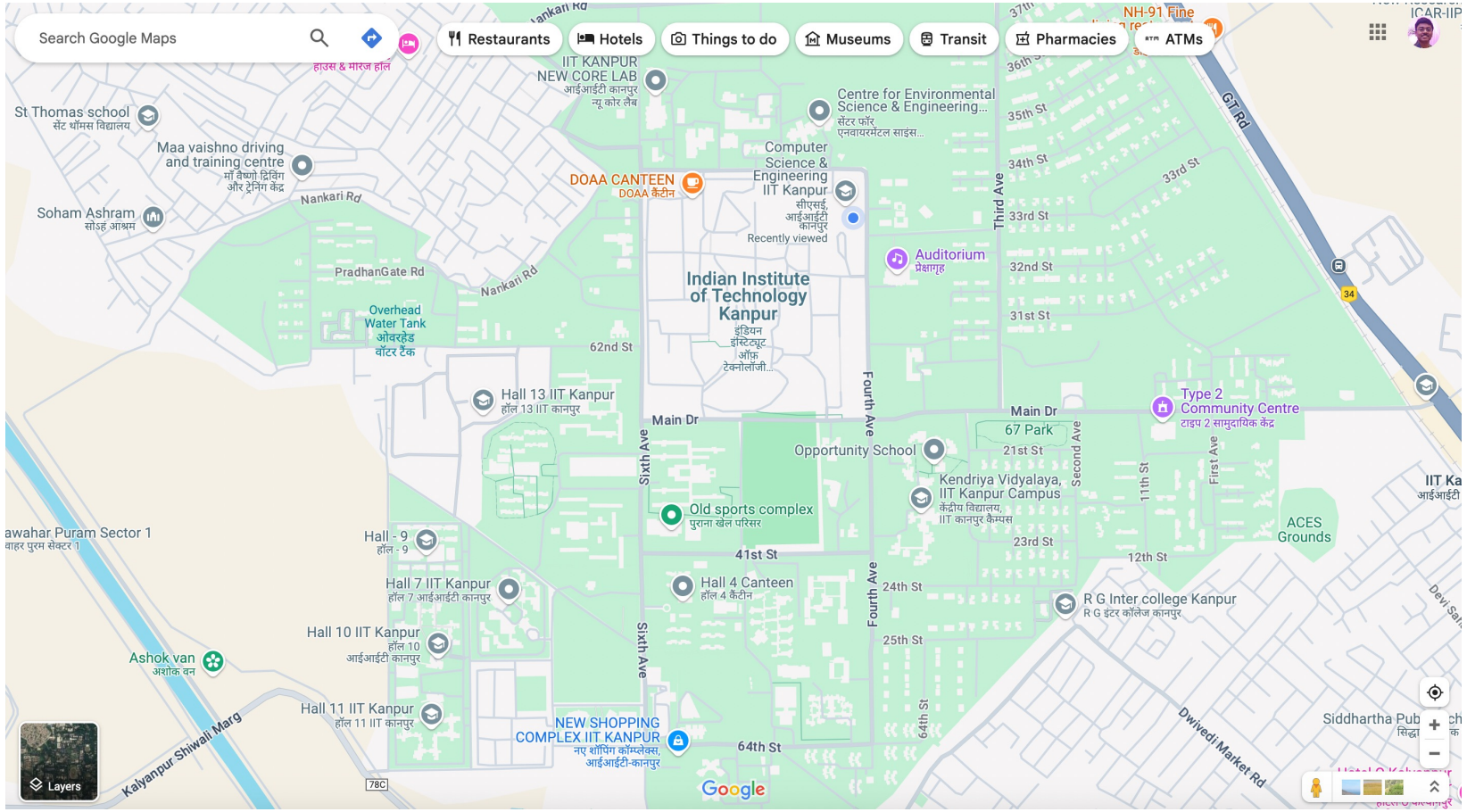
- Ben Shneiderman's Mantra: Overview, zoom and filter, then details-on-demand!



Overview first

Visual Information Seeking Mantra

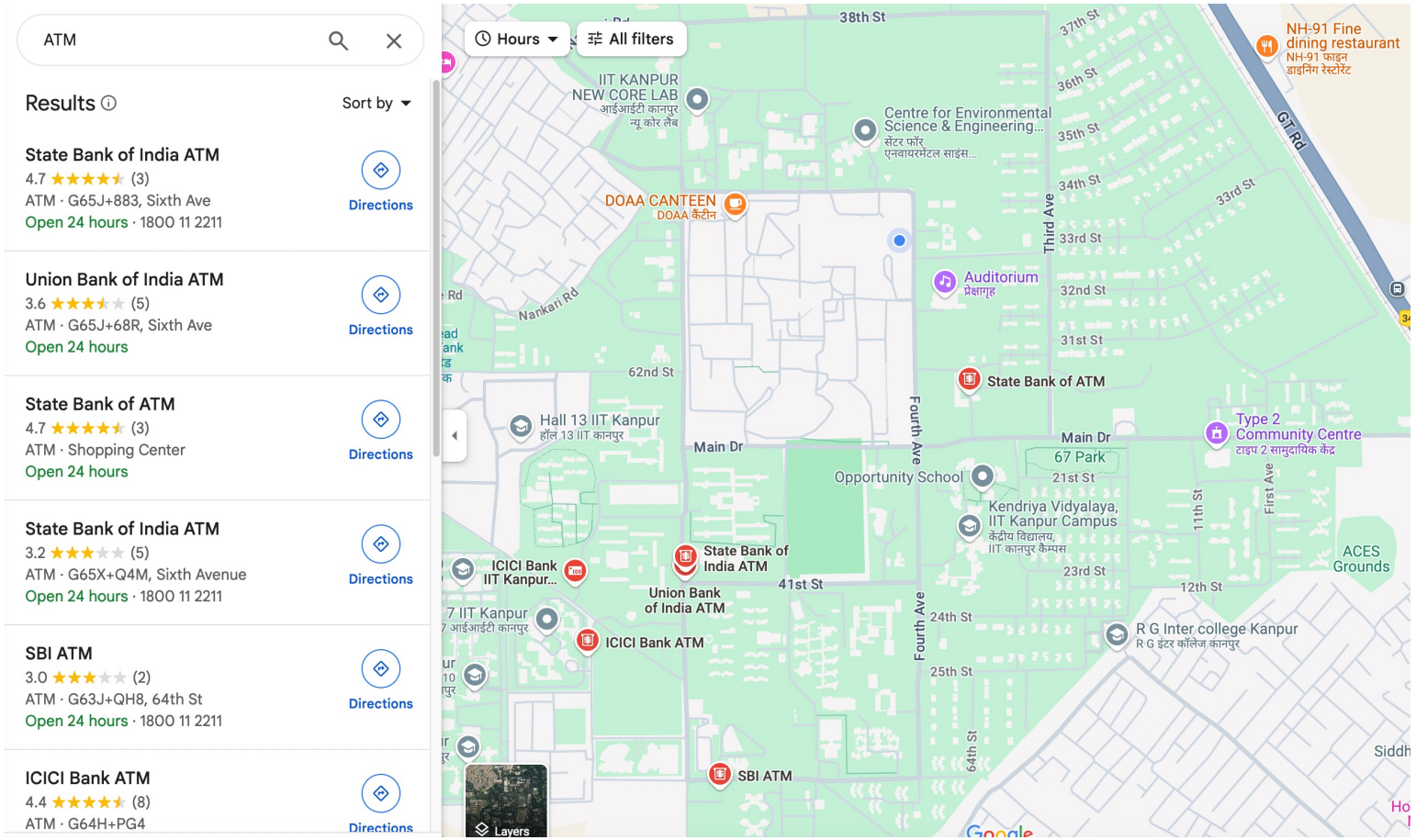
- Ben Shneiderman's Mantra: Overview, zoom and filter, then details-on-demand!



Zoom

Visual Information Seeking Mantra

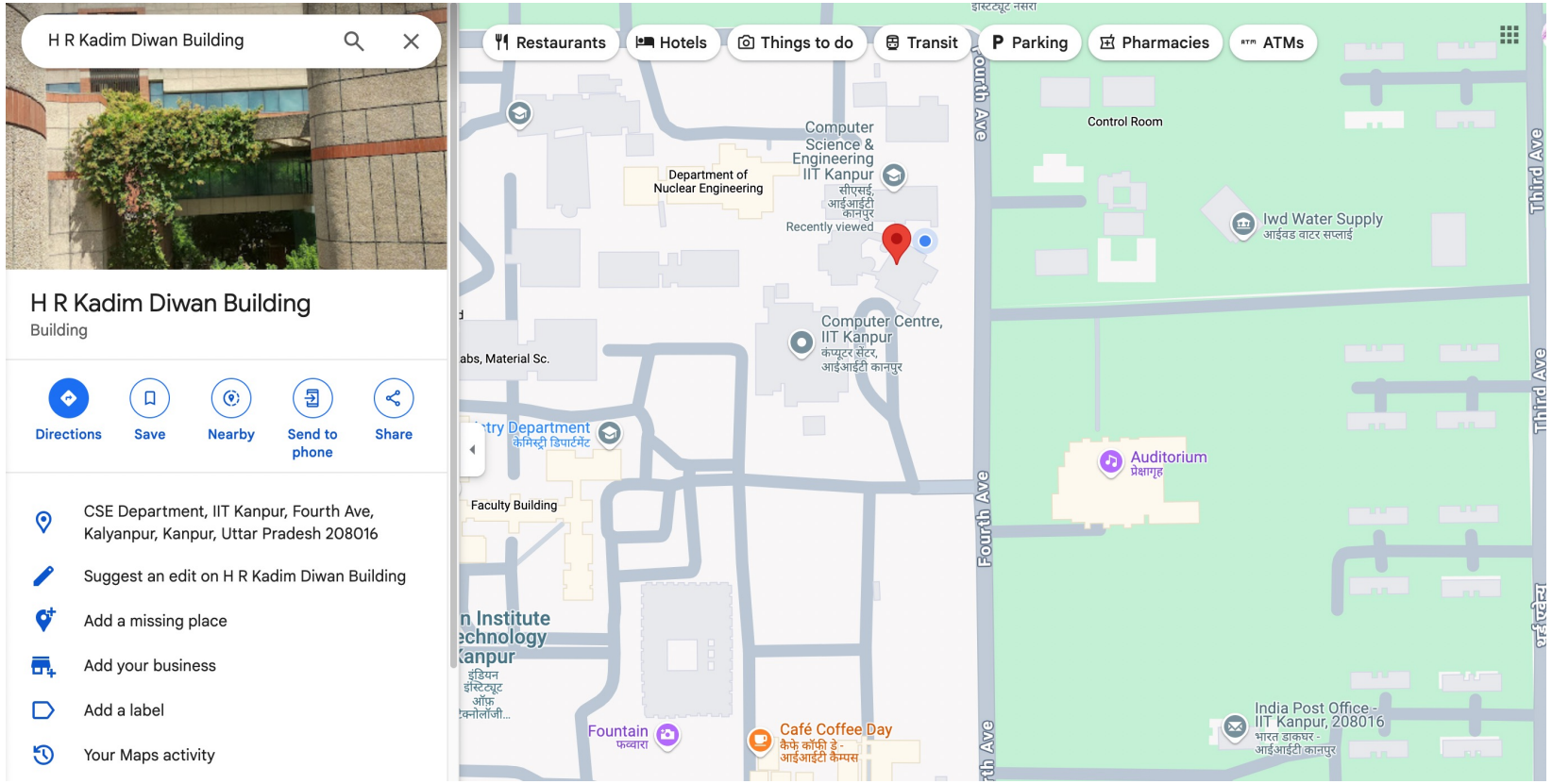
- Ben Shneiderman's Mantra: Overview, zoom and filter, then details-on-demand!



Filter

Visual Information Seeking Mantra

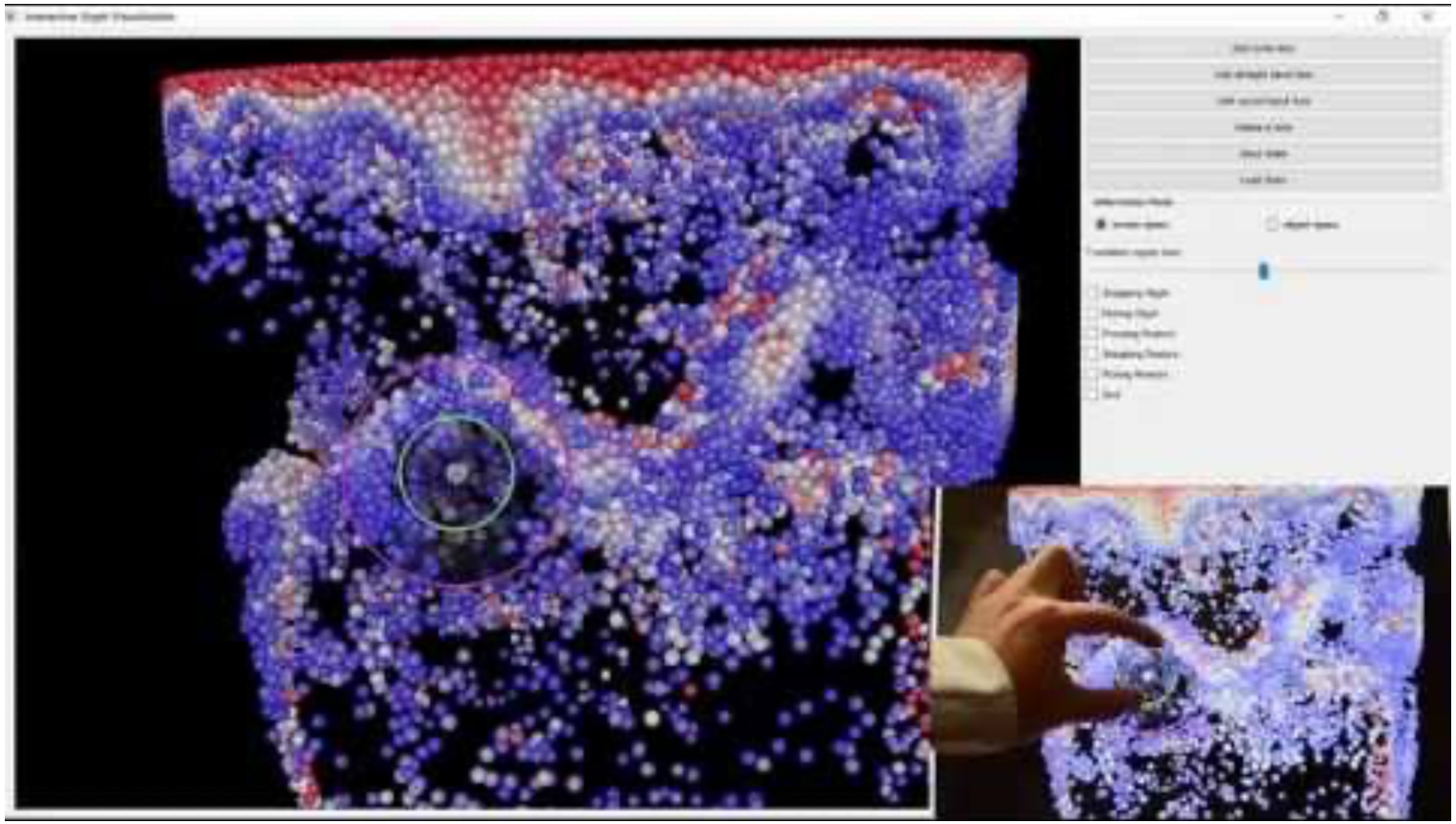
- Ben Shneiderman's Mantra: Overview, zoom and filter, then details-on-demand!



Details on demand

Another Paradigm: Focus + Context

- Focus + Context:
  - One single view which shows information in direct context
  - Maintains continuity and do not require viewer to shift back and forth
  - But: there is distortion!



Use of Visualization

- Visual Perception
  - Fast screening of lot of data
  - Pattern recognition
  - High-level cognition
- Interaction
  - Direct manipulation of data and visualization (Human in the loop)
  - Two-way communication

Humans are important!

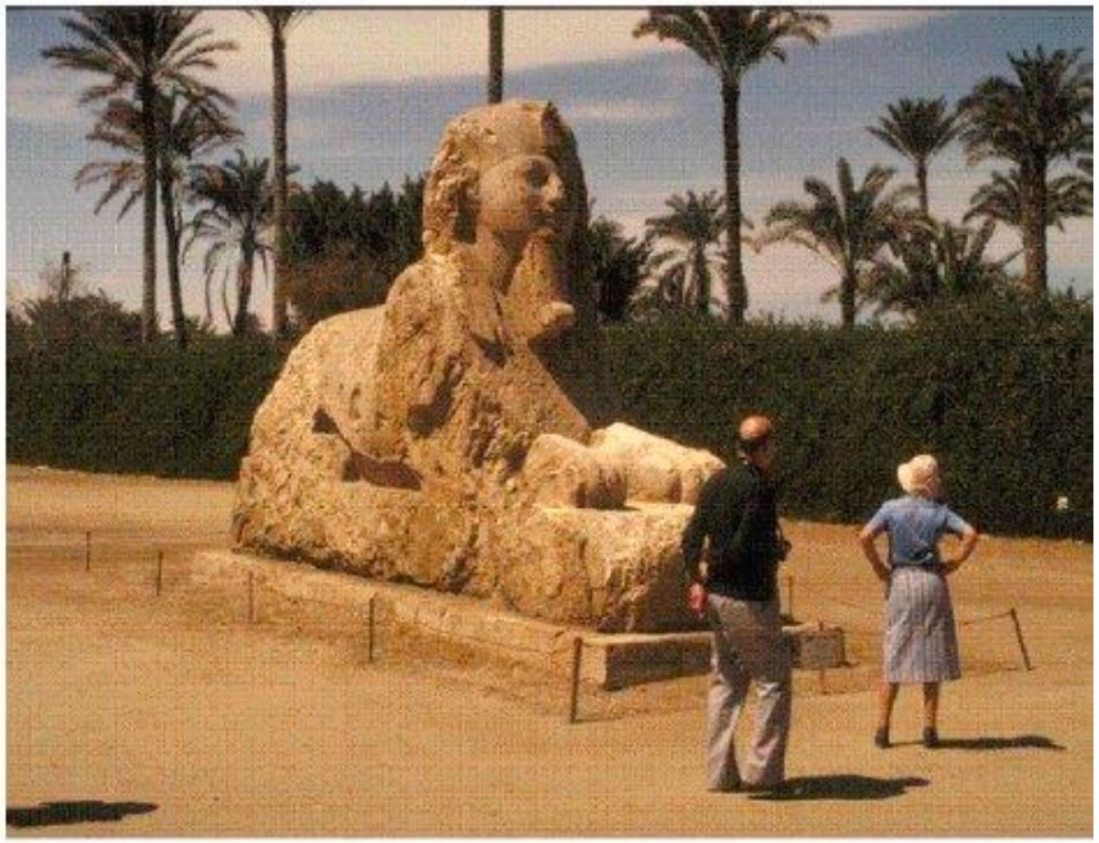
But Humans are imperfect too!!

Humans Are Imperfect

- Humans tend to overlook/ignore non-focused (and unexpected) objects even when they are very close and obvious
- Humans also have limited working memory
  - Fine details are quickly forgotten when focus changes
  - Need to preserve temporal context

Humans Are Imperfect

- Spot the difference: Change blindness



Source: Google

Humans Are Imperfect

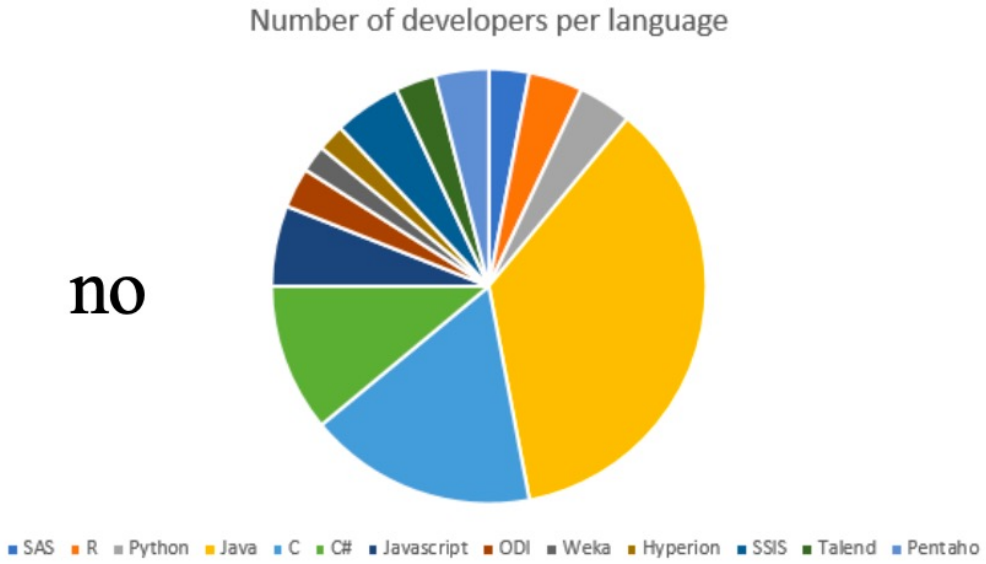
- Spot the difference: Change blindness

Source: Wikipedia

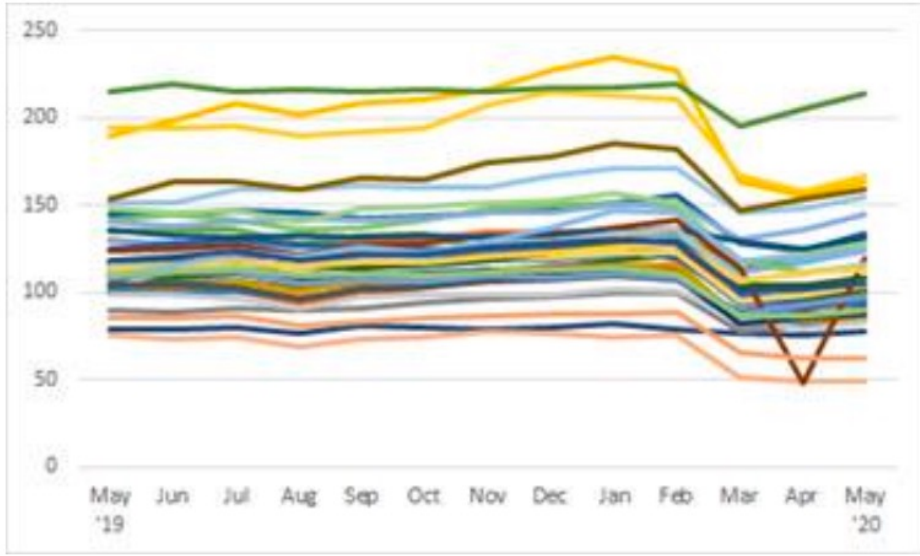
Human Limitations for Visualization

- The Magic Number Seven (7 ± 2) for visualization
  - Not more than 7 ± 2 segments in a pie chart
  - Not more than 7 ± 2 colors in a line chart
  - and so on …..

no

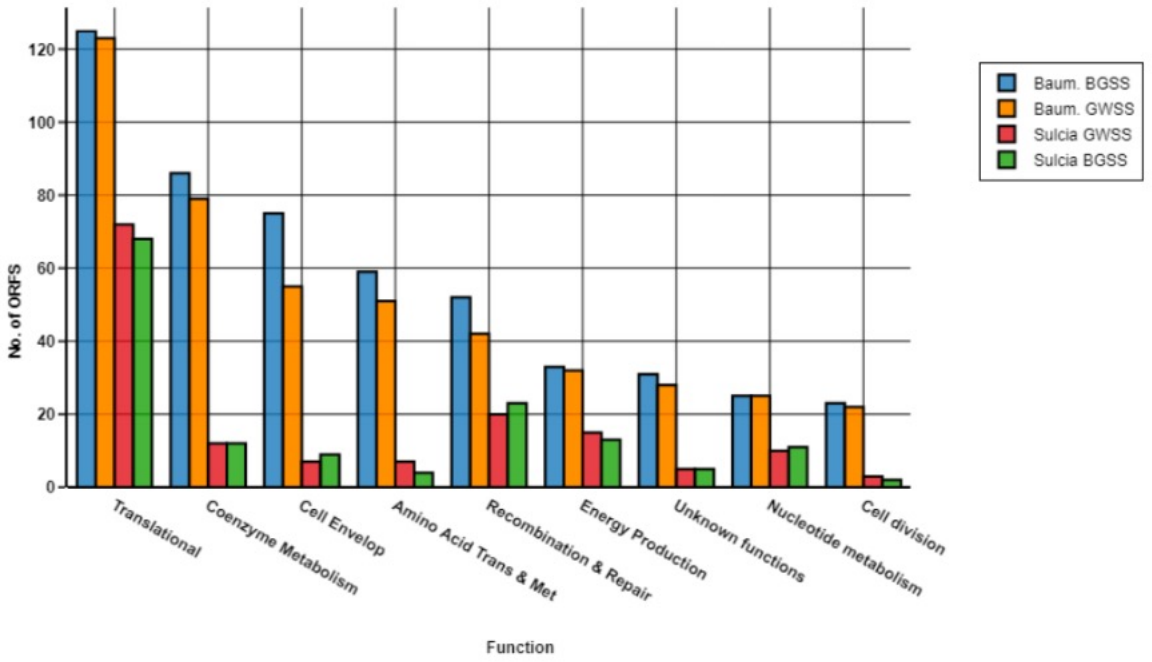


no

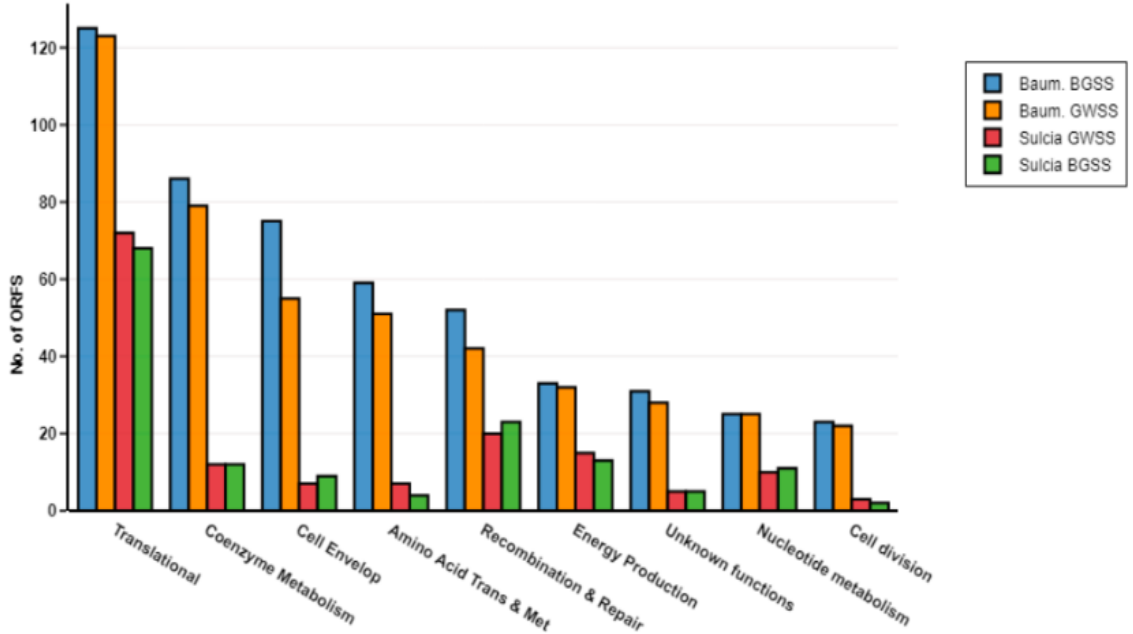


Miller, G.. (1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information".

Example of Visual Complexity



Do we really need the background grid?



Maybe not!

Handling Data

What Do We Do After Getting the Raw Data?

- Real world data can be dirty!

- Data cleaning (Wrangling)
  - Missing values
  - Noisy data
  - Deal with outliers
  - Standardize/normalize
  - Resolve inconsistency
  - Fuse/merge

Data Cleaning Cycle

Missing Data: Why?

- Data may not be always available/complete!

- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Many more other reasons

Missing Data: How to Handle?

- How would you estimate the missing value for a dataset?
  - Ignore or put in a default value
  - Manually fill in (can be tedious or infeasible for large data)
  - Use the available value of the nearest neighbor
  - Average over all the values
  - Use a probabilistic methods (regression, Bayesian, decision tree)
  - Use AI/ML models to predict missing data

Data Normalization and Standardization

- Sometimes we like to have all variables on the same scale
  - Min-max normalization

$$v' = \frac{v - min}{max - min}$$

  - Standardization

$$v' = \frac{v - \overline{v}}{\sigma_v}$$

Data Normalization and Standardization

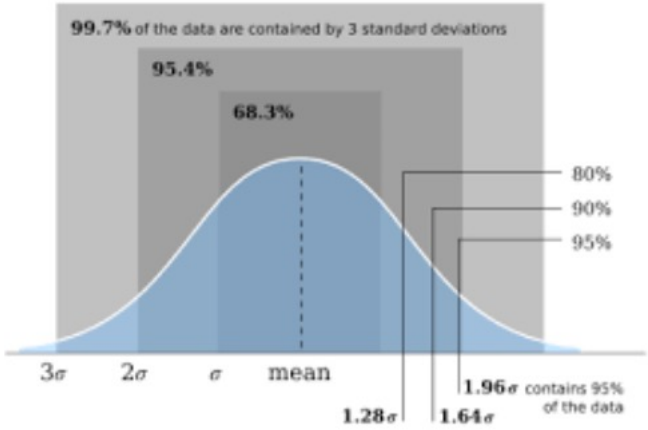
- Sometimes we like to have all variables on the same scale
  - Min-max normalization

$$v' = \frac{v - min}{max - min}$$

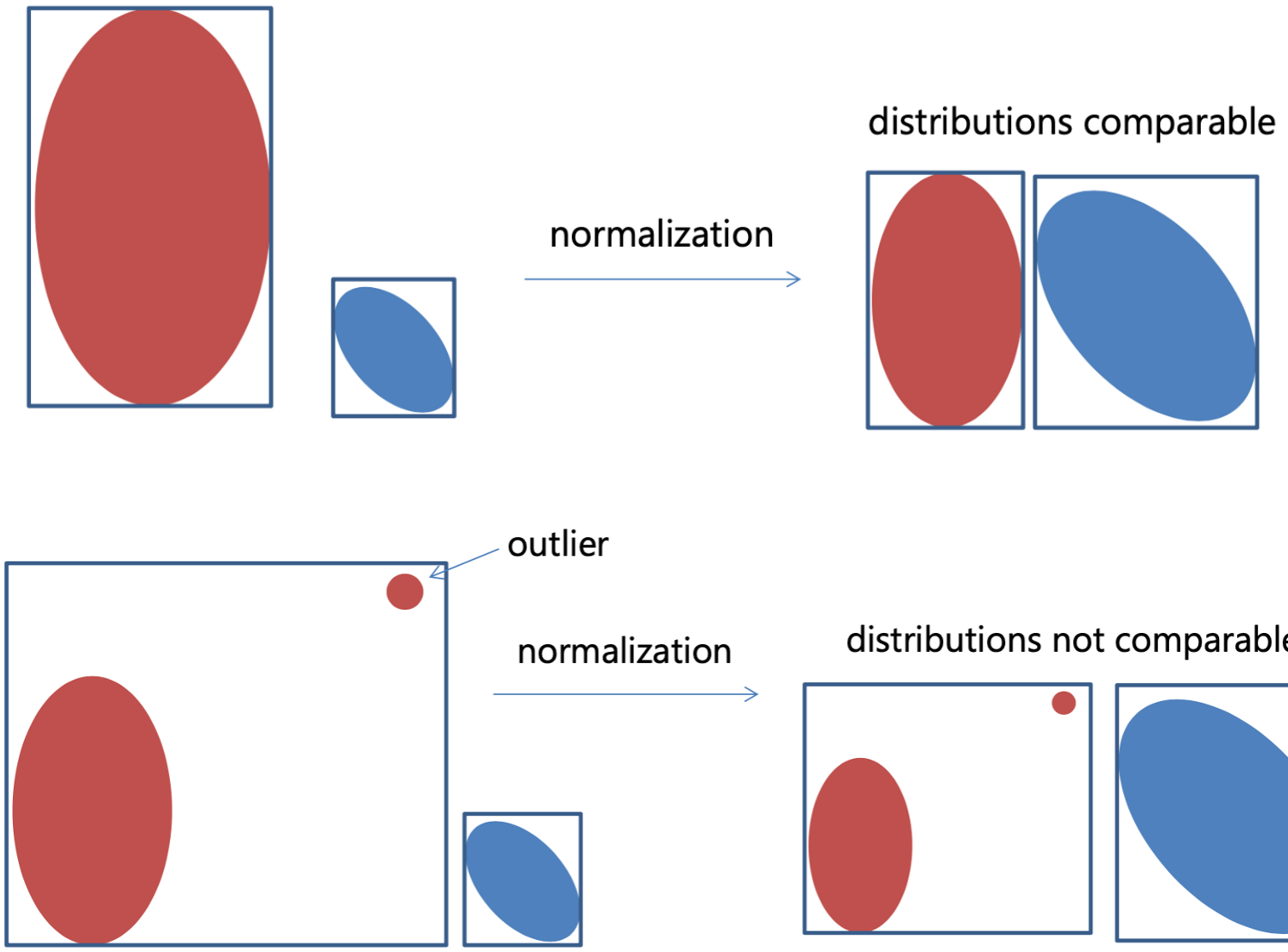
  - Standardization

$$v' = \frac{v - \overline{v}}{\sigma_v}$$

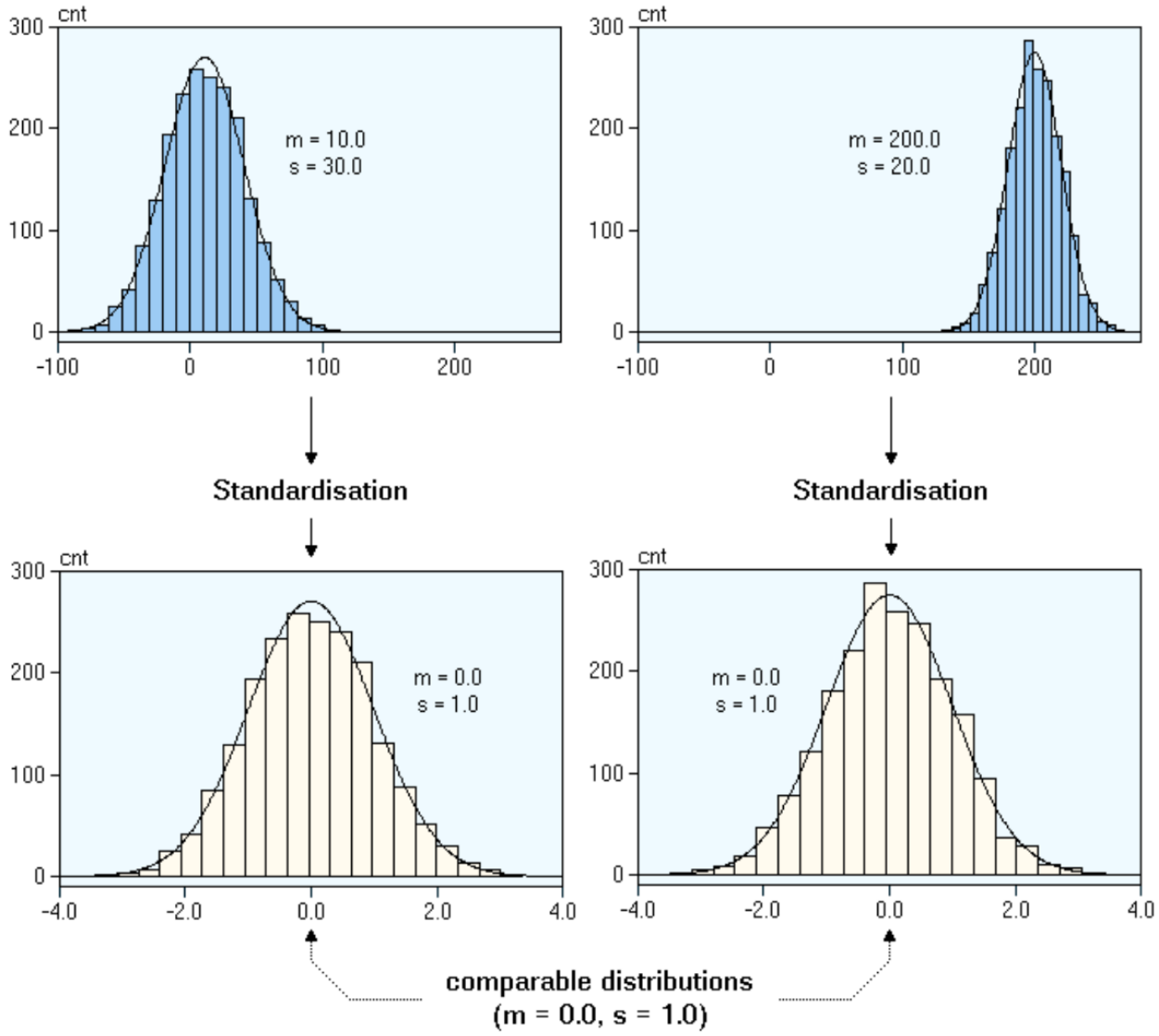
  - Clipping tails and outliers
    - set all values beyond ± 3s to value at 3s



Normalization



Standardization

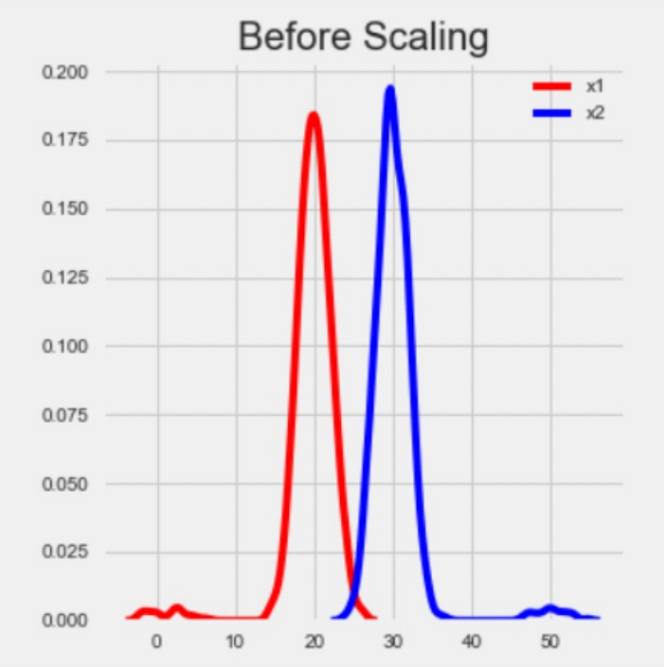


Robust Scaling

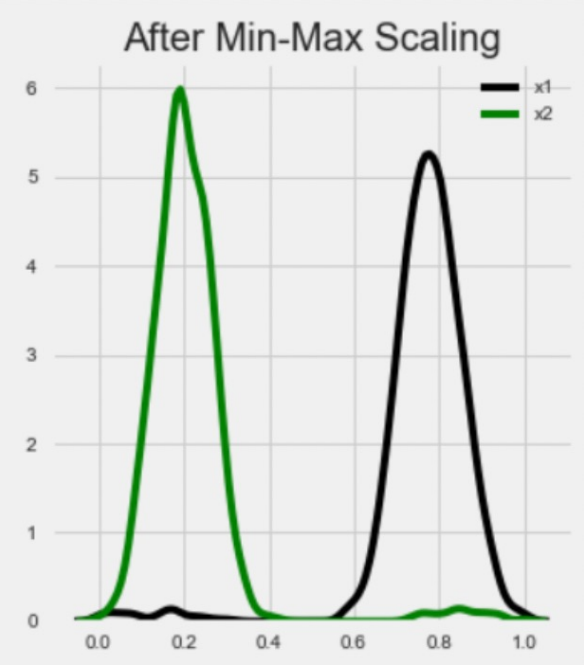
$$x_{\text{scaled}} = \frac{x - \text{Median}(X)}{\text{IQR}}$$

- IQR = Q3 – Q1
  - Difference between the 75th percentile and the 25th percentile data
- Immune to outliers
  - Relies on the median and IQR, which are robust to extreme values
  - Ensures that most of the data falls within a consistent range after scaling

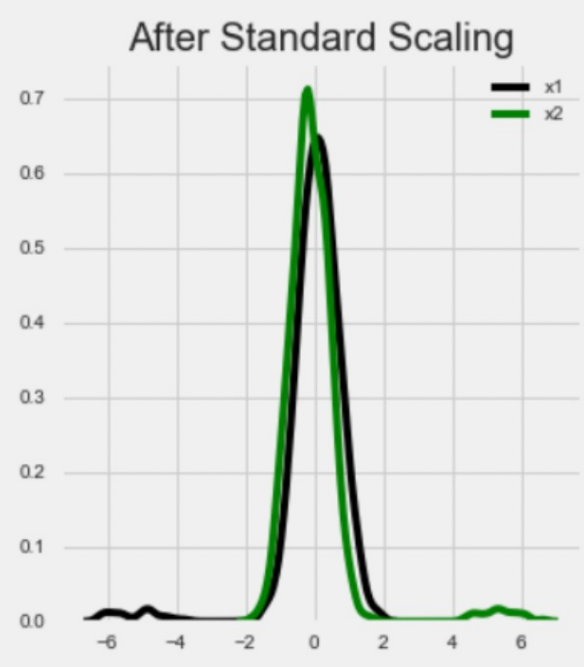
Comparison Among Diff. Methods of Scaling



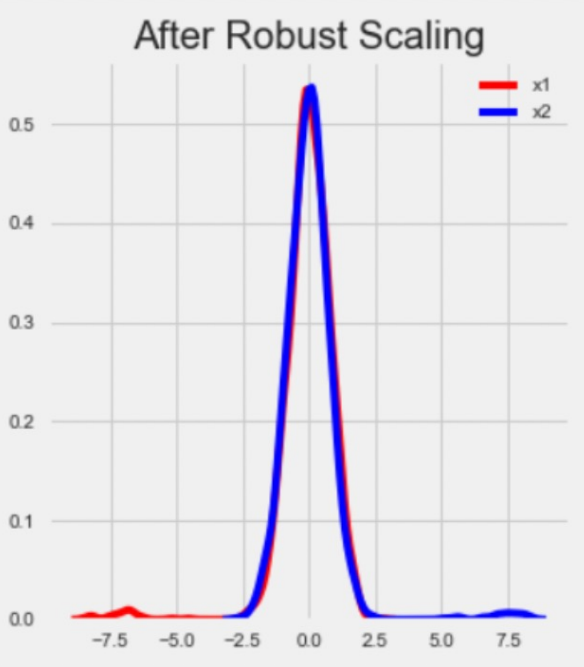
Raw Data



Min-max normalization



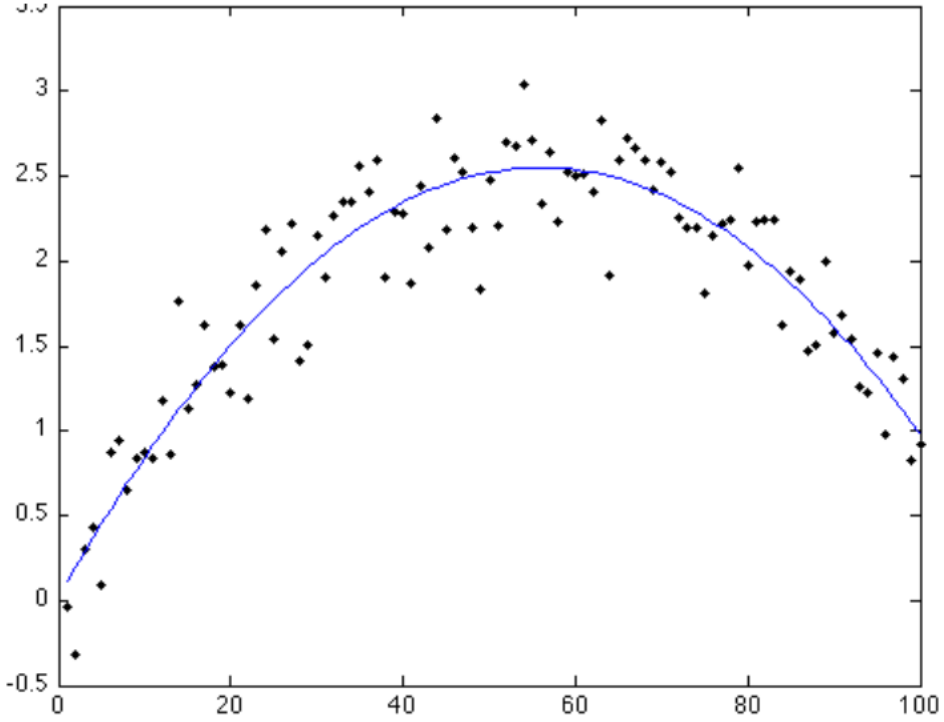
Standardization



Robust Scaling

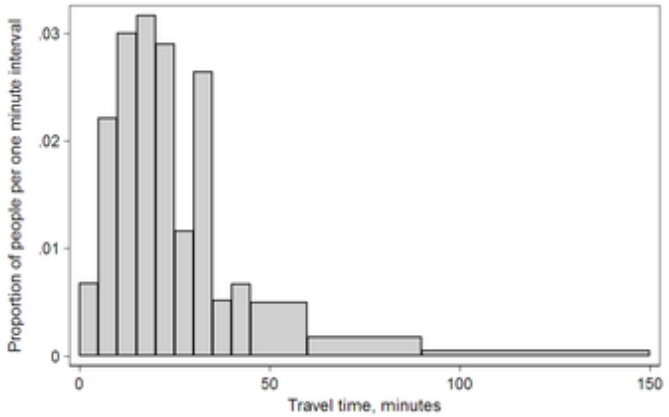
Noisy Data

- Noise = Random error in a measured variable
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention



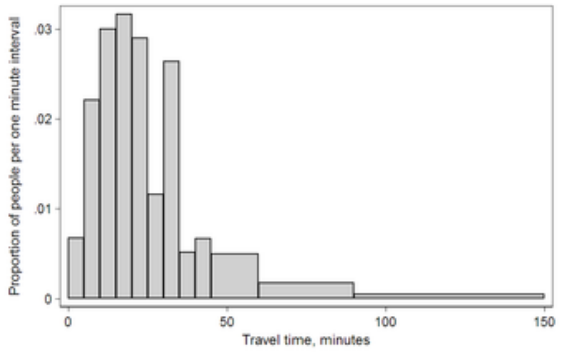
Noisy Data: What to Do?

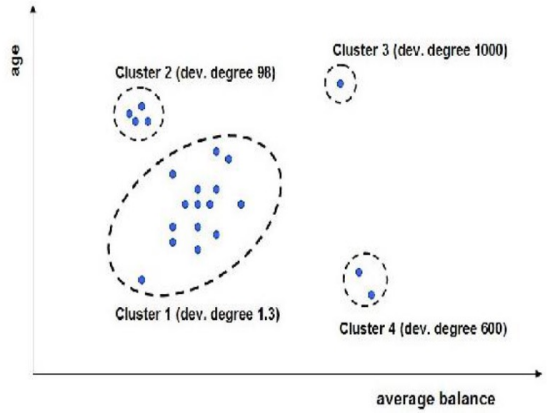
- Binning
  - Replace data with bin centers



Noisy Data: What to Do?

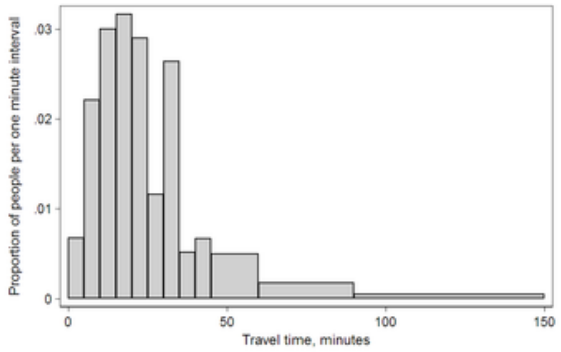
- Binning
  - Replace data with bin centers
- Clustering
  - Detect and remove outliers

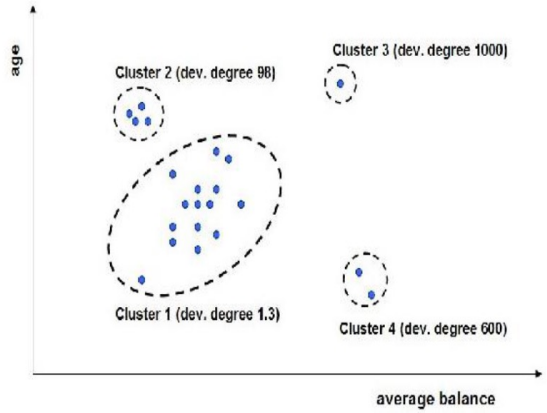




Noisy Data: What to Do?

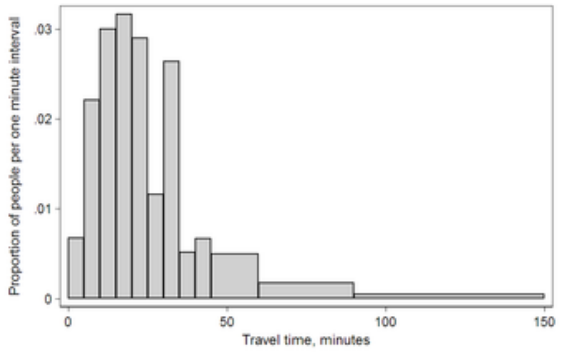
- Binning
  - Replace data with bin centers
- Clustering
  - Detect and remove outliers
- Semi-automated method
  - Combined human and computer inspection
  - Detect suspicious value and check manually

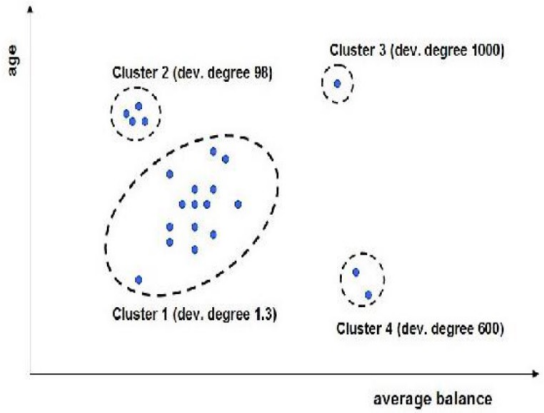


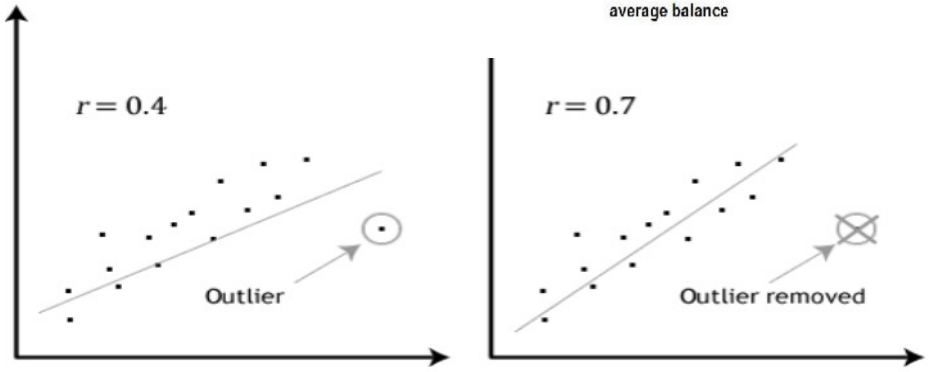


Noisy Data: What to Do?

- Binning
  - Replace data with bin centers
- Clustering
  - Detect and remove outliers
- Semi-automated method
  - Combined human and computer inspection
  - Detect suspicious value and check manually
- Regression
  - Smooth data by fitting to a regression function

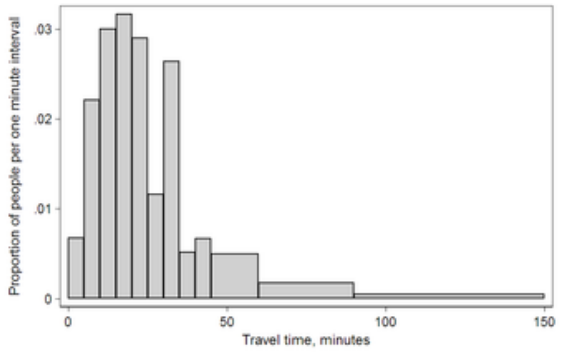


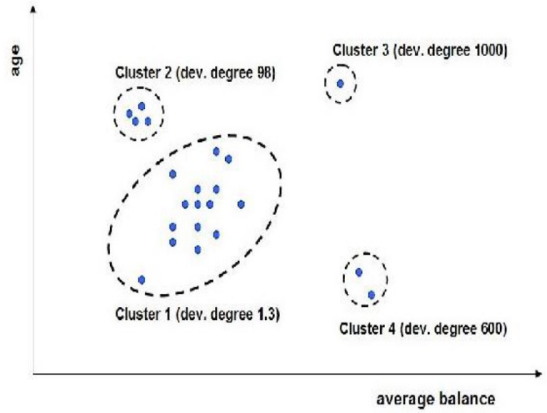


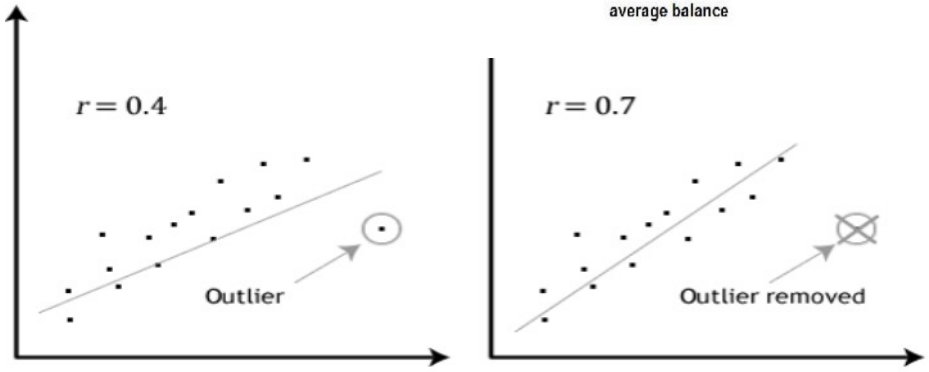


Noisy Data: What to Do?

- Binning
  - Replace data with bin centers
- Clustering
  - Detect and remove outliers
- Semi-automated method
  - Combined human and computer inspection
  - Detect suspicious value and check manually
- Regression
  - Smooth data by fitting to a regression function
- Outliers are not always noise! Be careful!







Deal with Small Data

- Can you invent meaningful new data?

Deal with Small Data → Data Augmentation

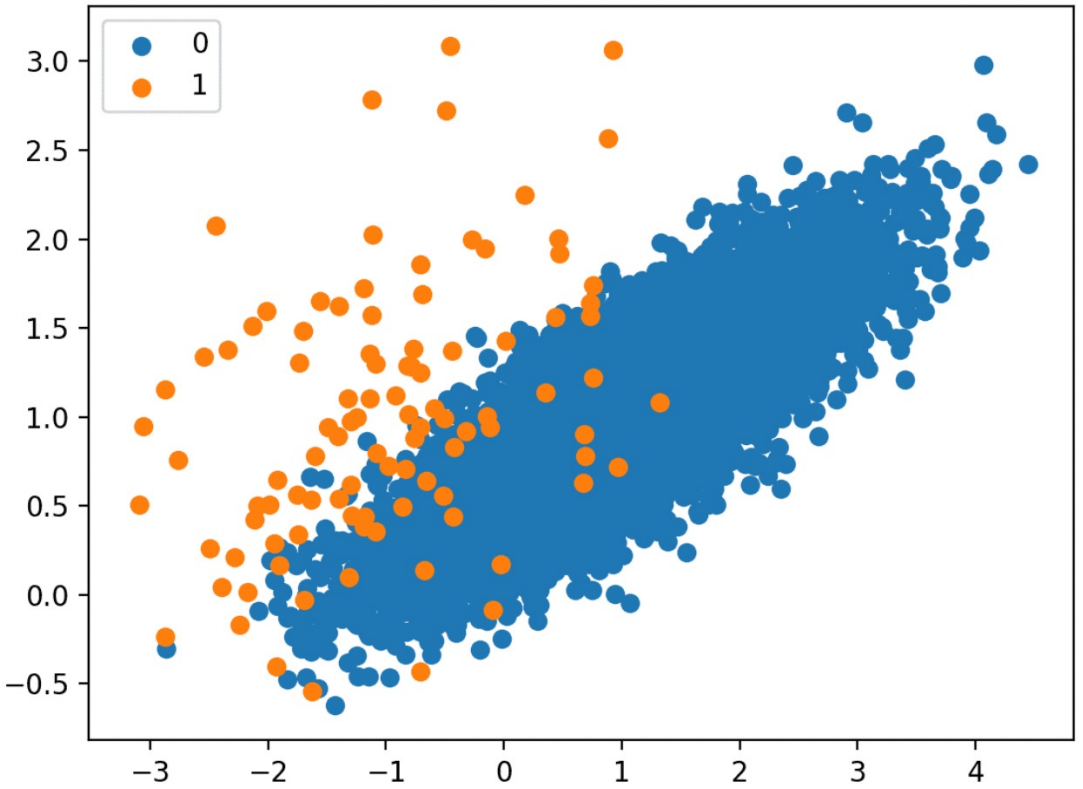
- Can you invent meaningful new data?
- Data Augmentation
  - Strategy to artificially synthesize new data from existing data

Deal with Small Data → Data Augmentation

- Can you invent meaningful new data?
- Data Augmentation
  - Strategy to artificially synthesize new data from existing data
- Common techniques are (for images)
  - rotations
  - Translations
  - Zooms
  - Flips
  - color perturbations
  - crops
  - add noise by jittering

Synthetic Data Generation for Imbalanced Classification

- When data has severe imbalance in the class representation
- If you use such data for ML model training, it will perform poorly for the minority class
- SMOTE (Synthetic Minority Oversampling Technique) can help
  - A data augmentation method



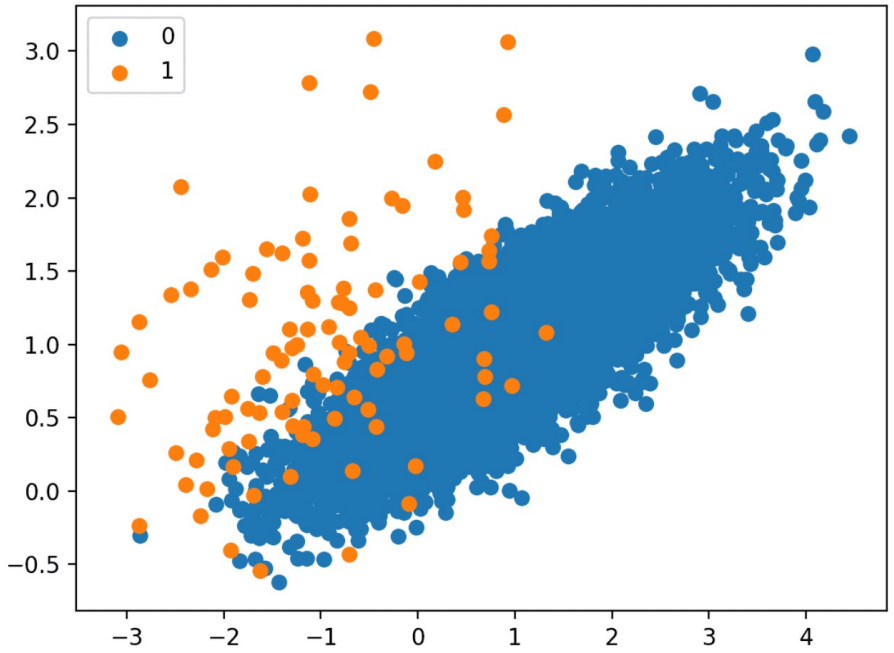
Imbalanced Data

SMOTE: Synthetic Data Generation for Imbalanced Classification

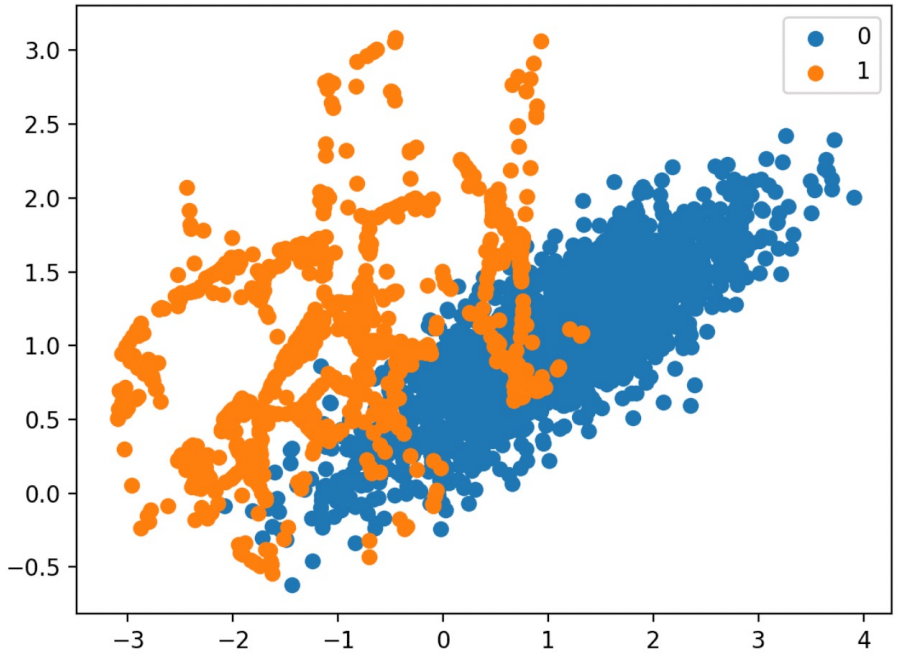
- How do we generate samples for minority class?
  1. Randomly under-sample the majority class
  2. Select a minority class instance (x) at random and find its k-nearest minority class neighbors
  3. Select one of the k neighbors at random, say (y)
  4. The synthetic instances are generated as a convex combination of the two chosen instances x and y

SMOTE: Synthetic Data Generation for Imbalanced Classification

- Example:



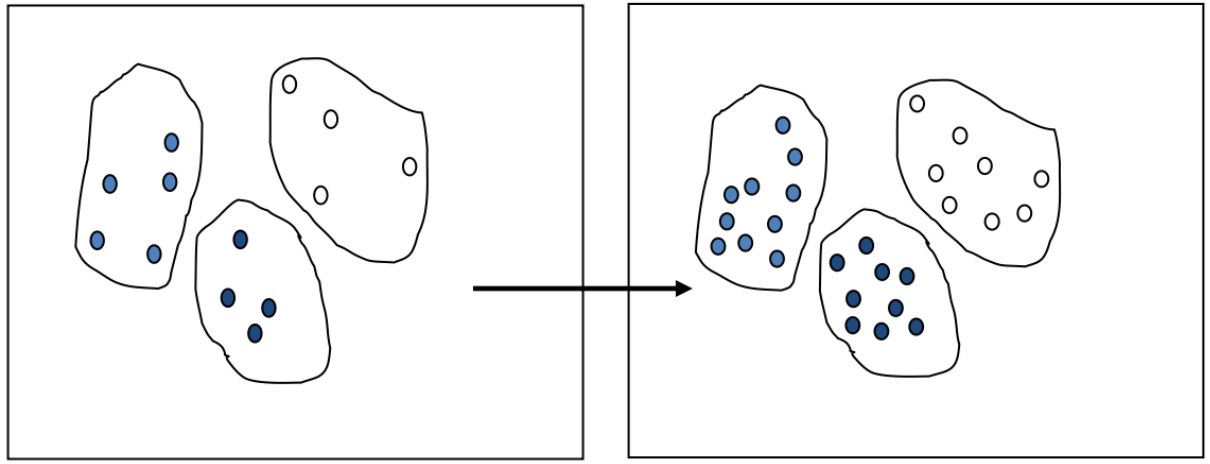
Imbalanced Data



SMOTE + random under-sampling

Data Augmentation for Visualization

- Generate new samples according to the data distributions
  - Cluster the data (outliers may form clusters!)
  - The size of each cluster represents its percentage in the population
  - Randomize new samples – bigger clusters get more samples



Augmentation rate ~ Cluster size

Deal with Big Data → Data Reduction!

- Purpose
  - Reduce the data to a size that can be feasibly stored without missing on important information
  - Reduce the data so a mining algorithm can be feasibly run

- Alternatives
  - Buy more storage
  - Buy more computers or faster ones
  - Develop more efficient algorithms

- In practice, all of this is happening at the same time
  - But the growth of data and complexities is faster
  - So, data reduction is important!

Data Reduction: How?

- Summarization (Later in the course)
  - Binning
  - Distribution-based
  - Clustering
- Sampling (Later in the course)
  - Systematic/Regular
  - Random
  - Stratified
  - Adaptive/Data-driven
  - Importance-driven
  - Cluster-based
- Dimension Reduction (Later in the course)
- AI/ML techniques (Later in the course)