UNIVERSITY OF HELSINKI

Data Exploration

Ziye Zhou

Course: Seminar on big data management

OVERVIEW

- Motivation
- Data exploration and problems
- Solution
- Conclusion

Why Explore Data?

Motivation

1. Data in the real world

- **Incomplete**

Lacking attribute values, or containing only aggregate data. e.g., Salary =" "

- **Noisy**

Containing errors or outliers.

e,g., Age = "-20"

- **Inconsistent**

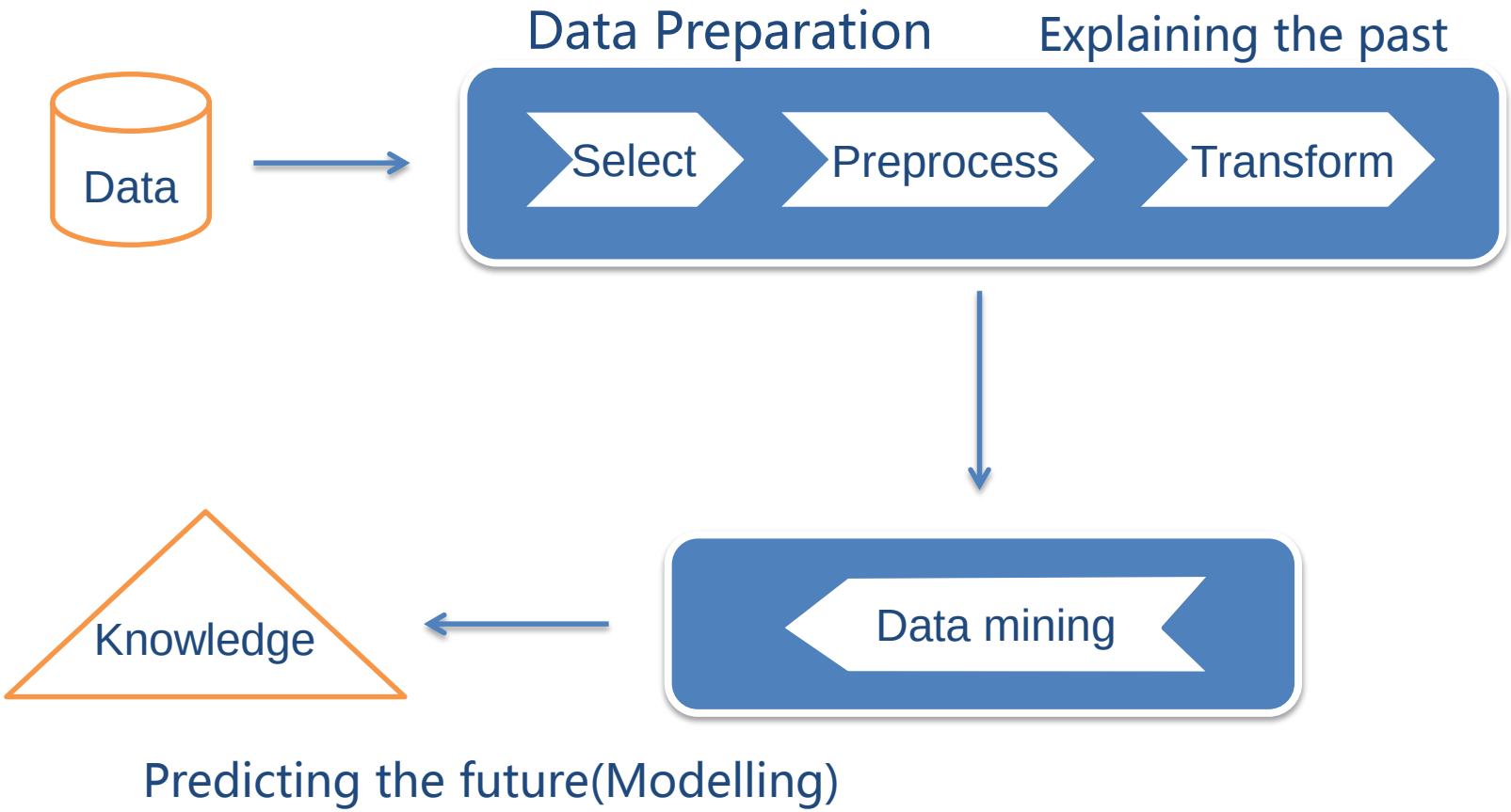
Discrepancies in names.

e,g., rating" 1,2,3" and "A,B,C"

2. Users do not what they are looking for

They will know that something is interesting only after they find it.

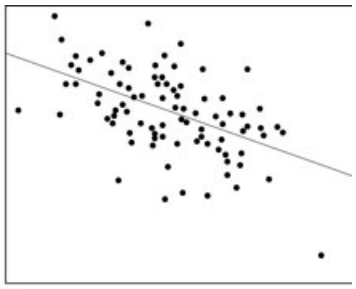
Main steps in statistical data analysis



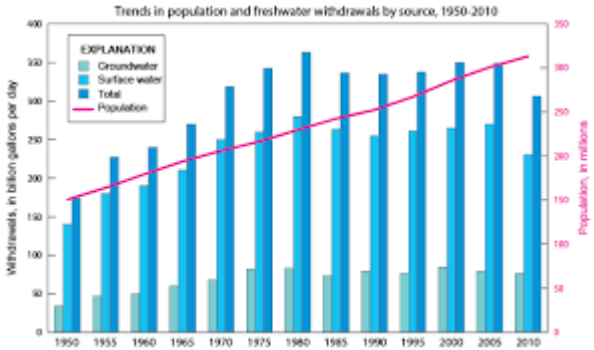
Ref:https://www.slideshare.net/21_venkat/data-exploration-validation-and-sanitization

Data Exploration

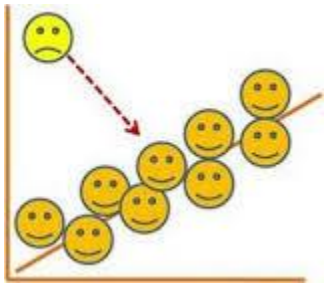
Definition: It is the first step in data analysis and typically involves summarizing the main characteristics of a dataset even if we do not know exactly what we are looking for.



Correlations



General trends



Outliers

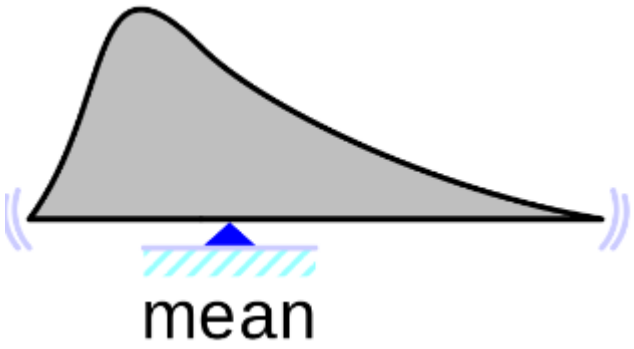
Ways to Explore Data

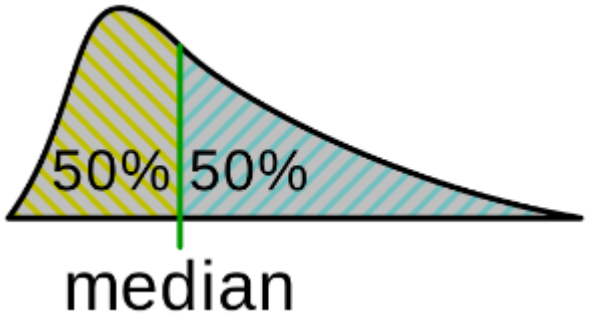
-Summary Statistics

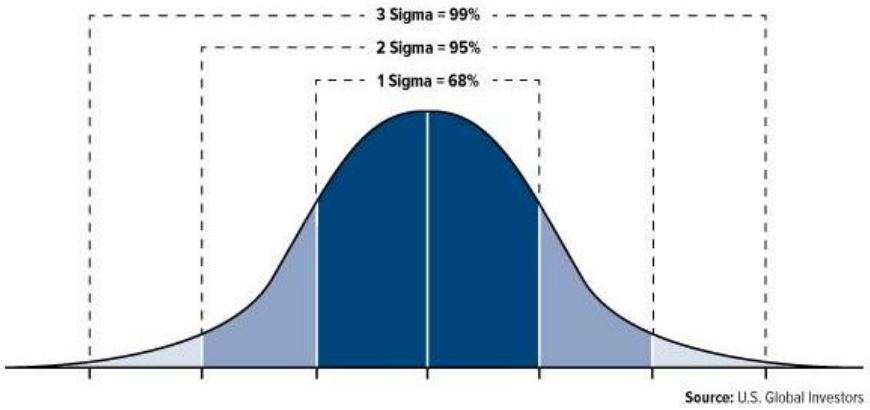
-Visualization

Summary Statistics

-Information that summarizes dataset



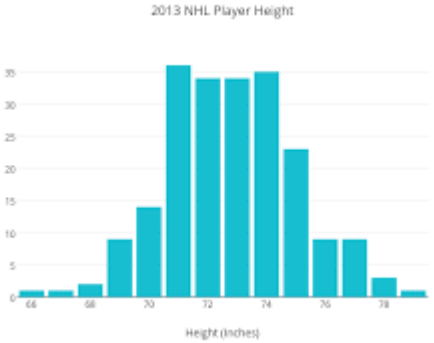




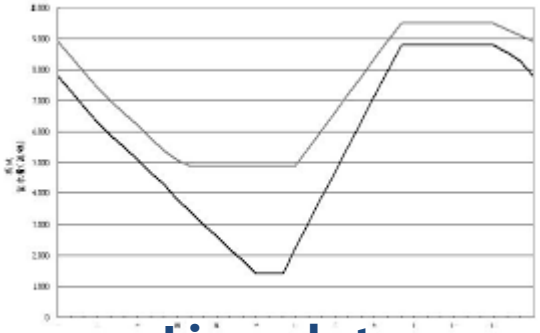
Standard Deviation

Data visualization

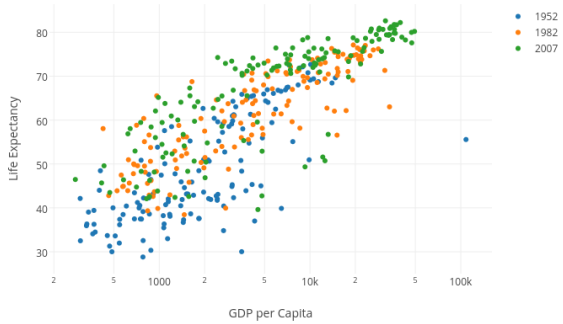
-Look at data graphically



Histogram



Line plot



Scatter plot

Problem?

Goal: Help users to make sense of very big datasets.

Available Tools

For professional data scientists, requires a deep knowledge of mathematics, statistics or computer science.

Database Challenges for Exploratory Computing

Marcello Buoncristiano[1], Giansalvatore Mecca[1], Elisa Quintarelli[2]
Manuel Roveri[2], Donatello Santoro[1], Letizia Tanca[2]

[1]Università della Basilicata – Potenza – Italy
[2]Politecnico di Milano – Milano – Italy

(Ref: ACM SIGMOD Record, 2015, 44(2): 17-22.)

A paradigm:
Step-by-step " conversation"
of a user and a system

Step 1

Starting the conversation

Activity

Initial hints about his interests

It might be interesting to explore the type of activities

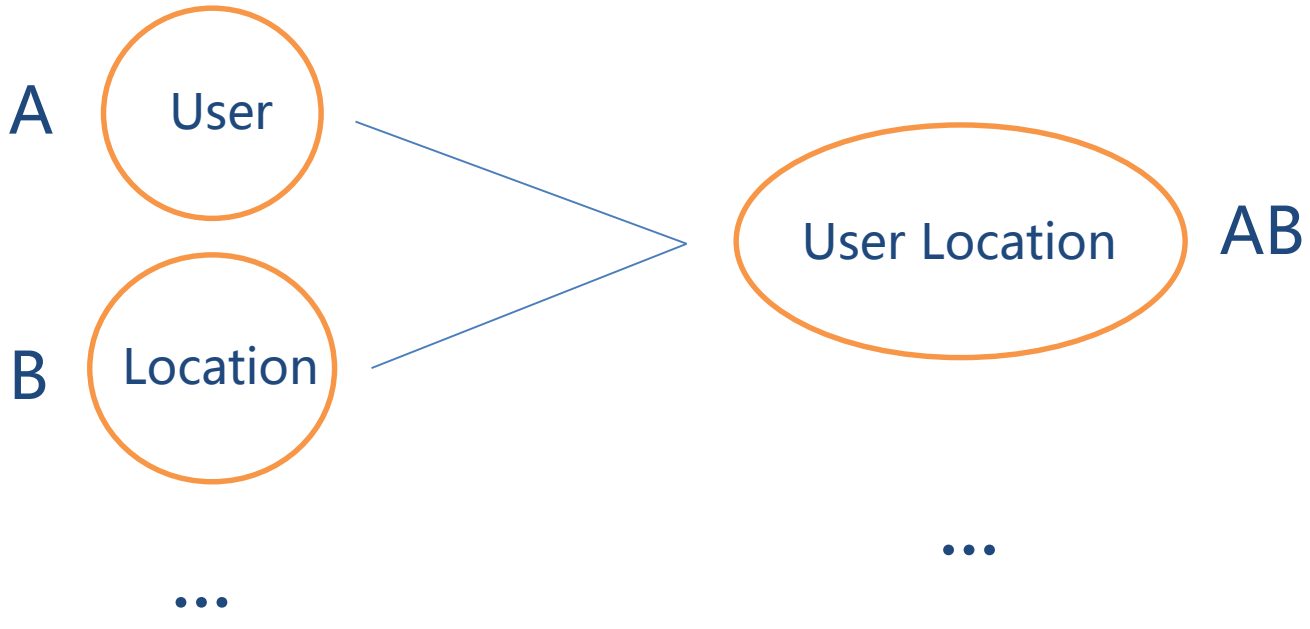
Potentially interesting perspectives

In fact:
running over 50%, while
cycling< 20%

A Model of Relevance

How to extract relevant feature?

Relevance —Based on→ Frequency distribution of attributes in the view lattice



How the Conversation goes on?

It might be interesting to explore the type of activity

User selected view Activity, and relevant feature type

shows a subset of values for the type attribute

Running

Add new view to the lattice(Activity, type, running)

Reseach challenges

- Responsiveness
- Summarization
- CPUs vs GPUs
- Fast Statistical Operators
- User Involvement

A Step Forward

Solve the technical problem of implementing a database-exploration system.

◆ Preliminary

- The critical step is the development of a statistical algorithm to measure the difference between two tuple-sets T^{Q1}, T^{Q2} with a common target feature, in order to compute the relative relevance.

A Step Forward

- Sample tuple sets T^{Q1}andT^{Q2} to extract subsets q_1 and q_2 of cardinality much lower than the one of T^{Q1}, T^{Q2},i.e., $|q_i| \ll |T^{Qi}|$.
- This can be done using different sampling strategies: sequential, random or hybrid

A Step Forward

- The method relies on an ensemble of hypothesis tests operating on randomly-extracted subsets of the original tuple-sets.
- The main intuition is that hypothesis tests should be conducted incrementally, in order to increase scalability, while at the same time keeping the emergence of false positives under control.

Step1: Sampling

- Sample tuple sets T^{Q1}andT^{Q2} to extract subsets q_1 and q_2 of cardinality much lower than the one of T^{Q1}, T^{Q2},i.e., $|q_i| \ll |T^{Qi}|$.
- This can be done using different sampling strategies: sequential, random or hybrid

Step2: Comparison

- Let X_1 and X_2 be the projections of q_1 and q_2 over a specific attribute(feature). The data in X_1 and X_2 can be either numerical or categorical.
- The comparison step aims at assessing the discrepancy between X_1 and X_2 through theoretically-grounded statistical hypothesis tests of the form $test_i(X_1,X_2)$
- Examples of these tests:
  - one-sample Chi-square test: assessing the distribution of a subset with discrete values.
  - two-sample Kolmogorov-Smirnov test: whether two subsets have been generated by two continuous different probability density functions.

Step3: Iteration

- Repeat the extraction and comparison steps M times.
- At the j-th iteration, a new pair of subsets X_1andX_2 are extracted and $test_i(X_1,X_2)$ is computed.
  - If the test rejects the null hypothesis, we stop the incremental procedure since we have enough statistical confidence that there is a difference in the data distributions of T^{Q1}andT^{Q2} .
  - Otherwise, the procedure proceeds o the next iteration.

Step4: Query ranking

- The procedure described above can be applied to different pairs of tuple sets.
- The difference between their empirical distributions is computed using the Hellinger distance.
- Based on this, we can rank the tuple sets to find out those exhibiting the largest differences.

Conclusion

- Briefy introduction to data exploration

Effciently extracting knowledge from data even if we do not know exactly what we are looking for.

- “Conversation” model

User and System that help each other to refine the data exploration process, ultimately gathering new knowledge that concretely fullfils the user needs.

- Technical challenges and solutions


THANKS!