Large Scale Data Mining

Conclusion, Exam preparation

credits:www.mmds.org

What we learnt?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending

- **We will learn to use different models of computation:**
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory

What we covered this Lecture

- **Scalability** (big data)
- **Algorithms**
- **Computing architectures**
- Automation for handling **large data**

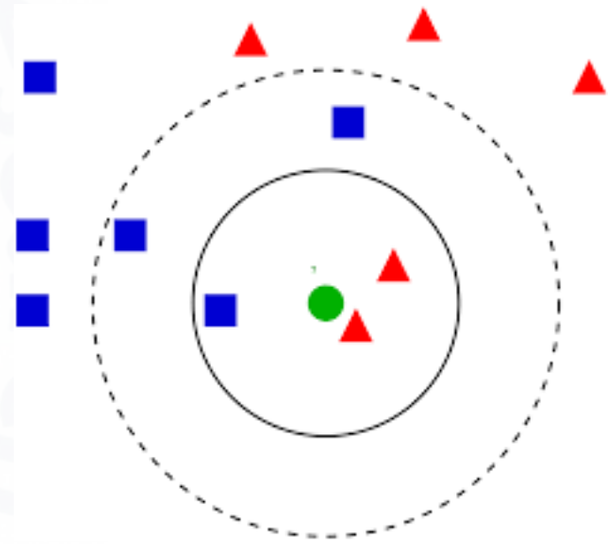
What we learnt ?

- **We learnt to solve real-world problems:**
  - Finding Similar Items
  - Streaming algorithms
  - Clustering and Community detection
  - Graph algorithms

- **We will learn various "tools":**
  - Optimization (stochastic gradient descent)
  - Indexing methods for fast graph computations
  - Hashing (LSH, Bloom filters)
  - Probabilistic analysis

Finding Similar Items

- What are the near duplicates / similar items / nearest neighbours ? How do we find this efficiently for large input ?

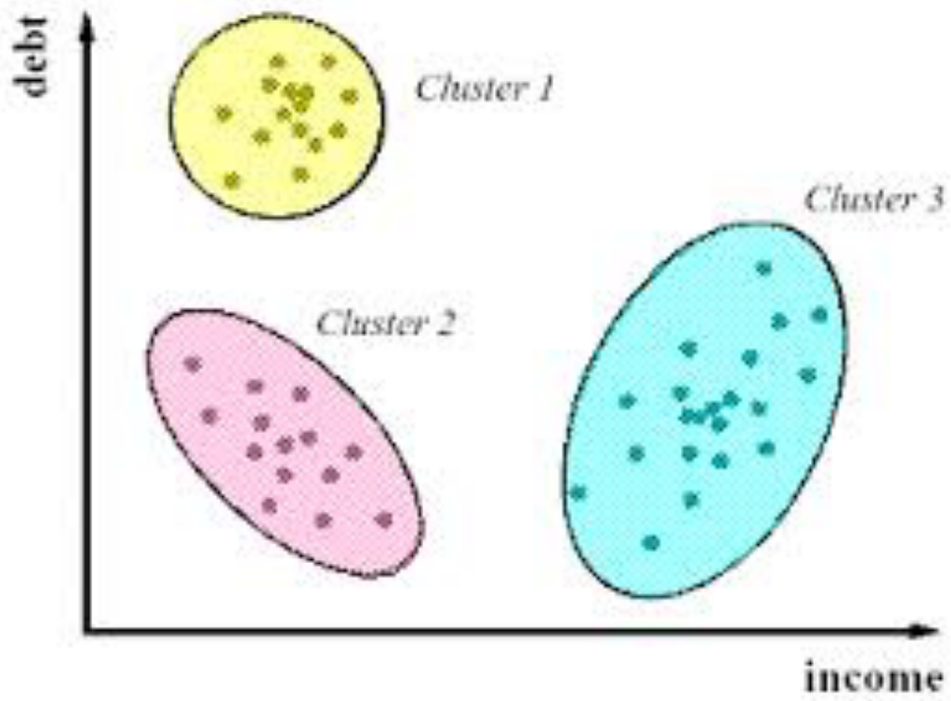


- How can we efficiently find similar items using LSH

- Why LSH works ? Probabilistic analysis + S-Curve
  - Avoid comparing items that are too dissimilar
  - Avoid missing candidate pairs that are actually similar
  - Min-Hashing for Jaccard distances

Clustering

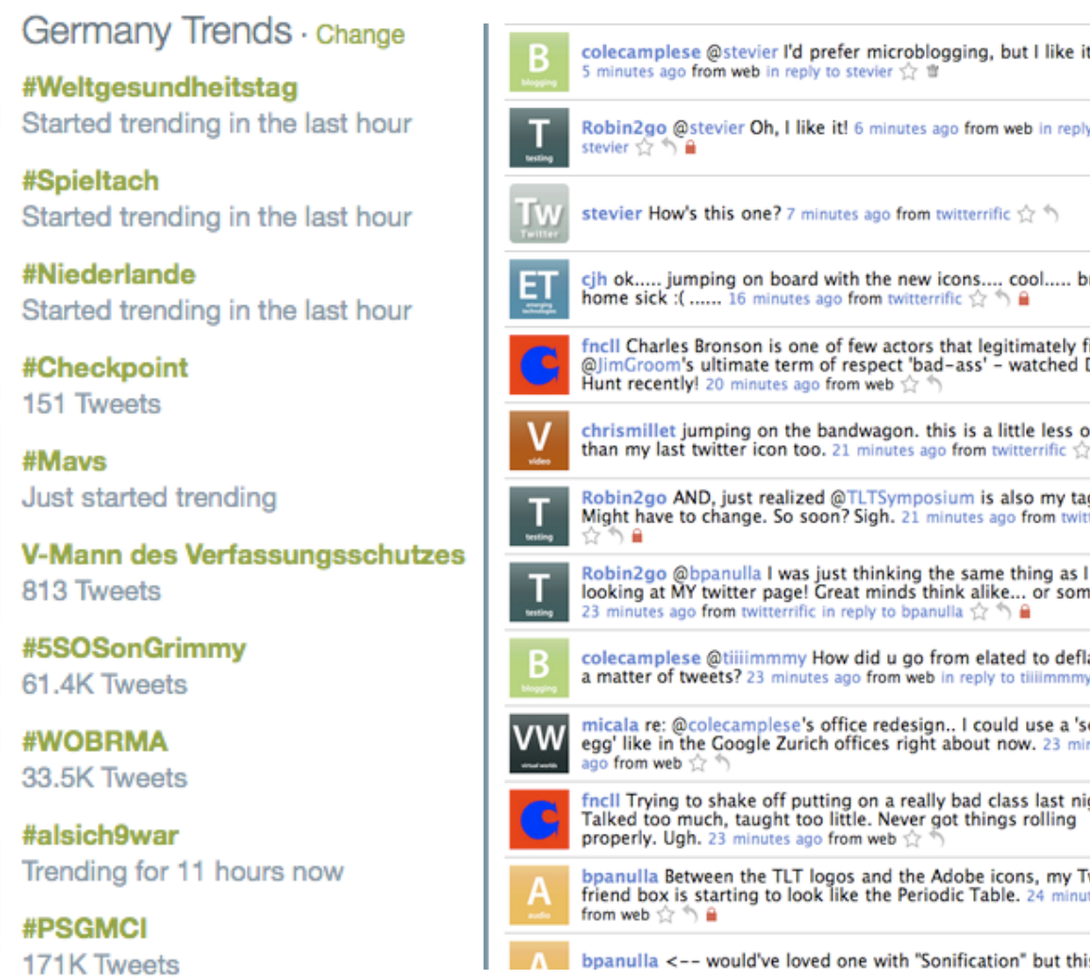
- What are the clusters given large input collections ?
- Agglomerative
- K-Means



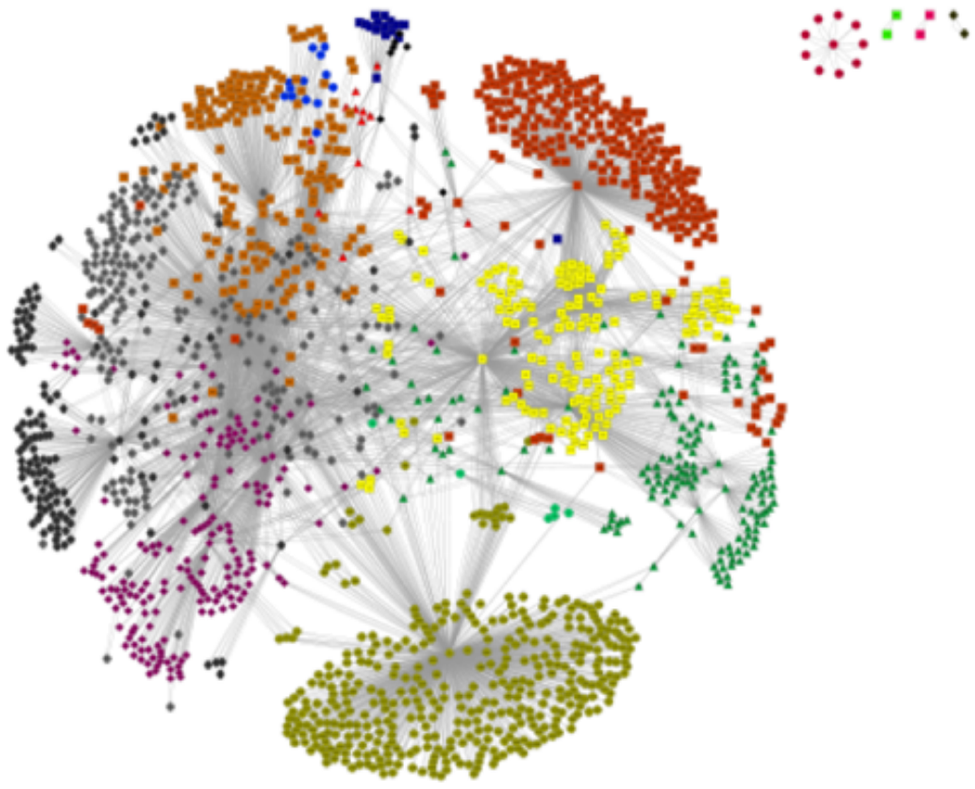
- How do we scale up using BFR framework

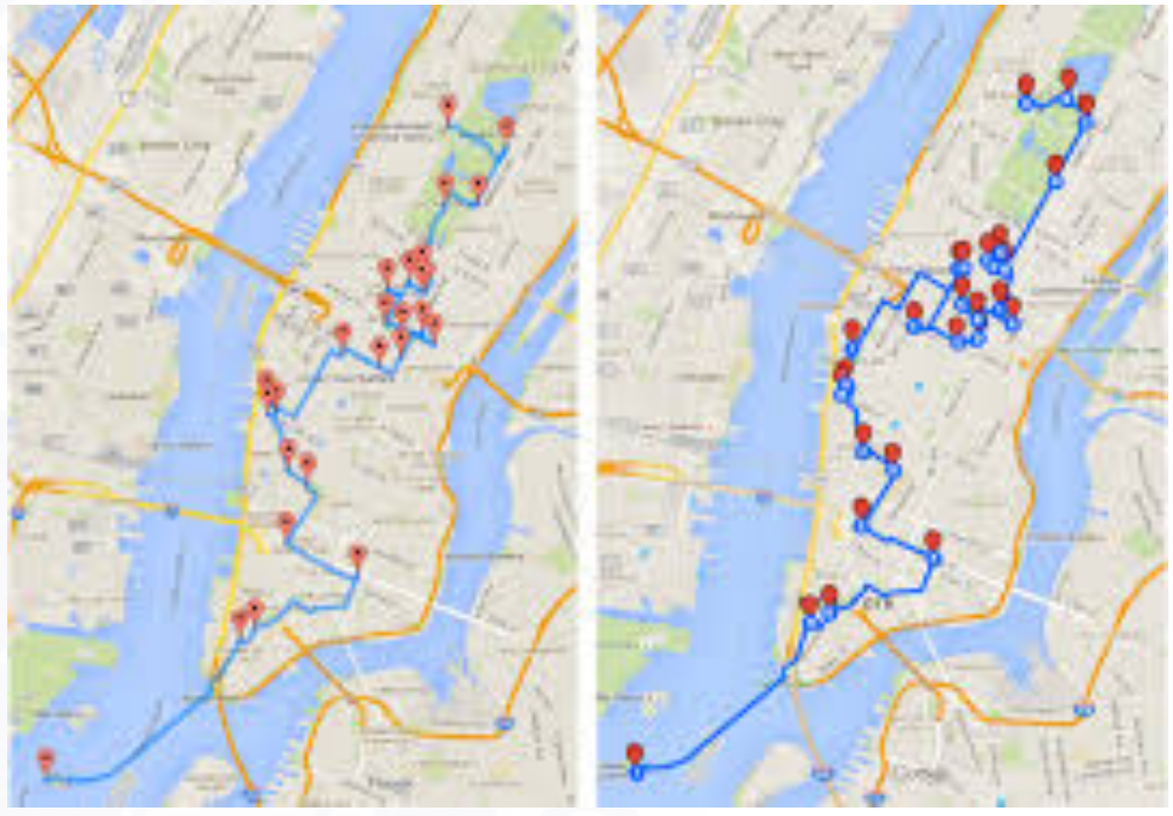
Mining Data Streams

- How do we sample data in a stream ?
- How do we count ones ?
- How do we count distinct elements in a stream ?
- How do we estimate moments in a stream ?



Mining Graphs





- How do we find communities in large graphs ?
- How do we find shortest paths in massive graphs in reasonable time ?

Final Exam

- Written Exam -
- Duration : 2 hours
- 1 bonus point = 0.3 grade improvement in your final exam
  - 1.3 + 1 bonus point = 1.0
  - 5.0 + 1 bonus point = 5.0
- Modelled on Assignments
- More applied and algorithmic aspects rather than memorising

July 29, 14:00 - 16:00

Exam correction review: August 4, 10:00 AM

Venue: to be announced

Exam Layout

- **Six questions (120 points)**
  - **Finding Similar Items (20 points)**
  - **Streaming Mining - I (20 points)**
  - **Streaming Mining - II (20 points)**
  - **Clustering (20 points)**
  - **Graph Mining - I (20 points)**
  - **Graph Mining - II (20 points)**

Graded from 100 points, Best of 5 questions

Programming Assignments

- Make a **Github/bitbucket** account
- Solve the assigned problems using the data provided and algorithm discussed in the lecture (optimized implementation) - NOT NAIIVE
- Can solve upto **three** problems
- Submit, link to github/bitbucket account
  - Code + runnable master script
  - REAMDE file with usage info
- Submit until **10 July,** Come to class on **14th July** for looking up your results