

Large Scale Data Mining

Introduction, Logistics

Who are we ?

- **Dr. Avishek Anand**

- post-doc researcher
- Information Retrieval, Text mining, Graph mining
- anand@l3s.de



- **Jaspreet Singh (TA)**

- PhD student
- Information Retrieval
- singh@l3s.de



What is Data Mining?

- Given lots of data
- Discover patterns and models that are:
 - **Valid:** hold on new data with some certainty
 - **Useful:** should be possible to act on the item
 - **Unexpected:** non-obvious to the system
 - **Understandable:** humans should be able to interpret the pattern

Data Mining

- But to extract the knowledge data needs to be
 - Stored
 - Managed
 - And ANALYZED ← this class

**Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Data Science**

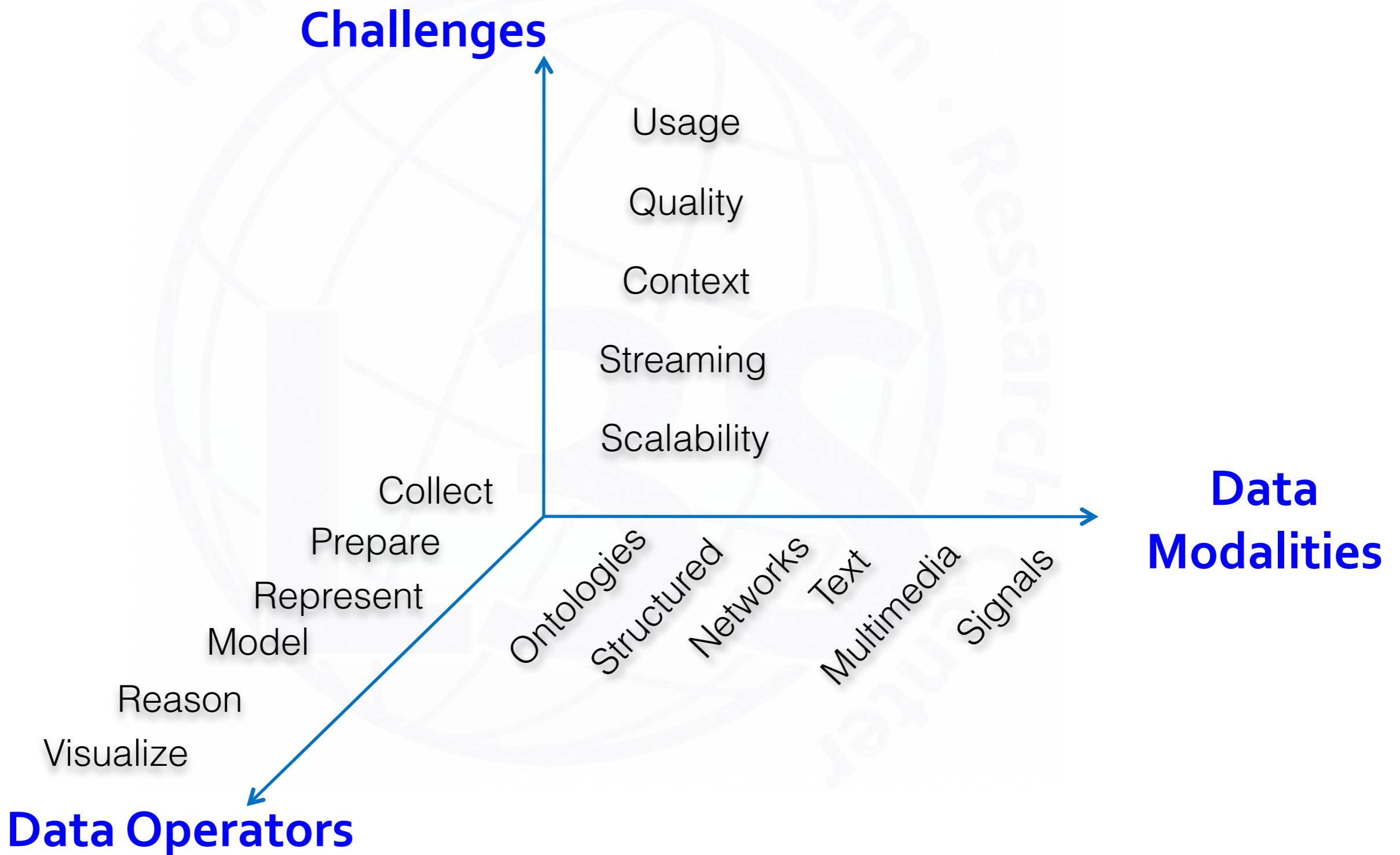
Data Mining Tasks

- **Descriptive methods**
 - Find human-interpretable patterns that describe the data
 - **Example:** Clustering
- **Predictive methods**
 - Use some variables to predict unknown or future values of other variables
 - **Example:** Recommender systems

Pitfalls and Risks

- A risk with “Data mining” is that an analyst can “discover” patterns that are meaningless
- Statisticians call it **Bonferroni’s principle**:
 - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

What matters when dealing with data?



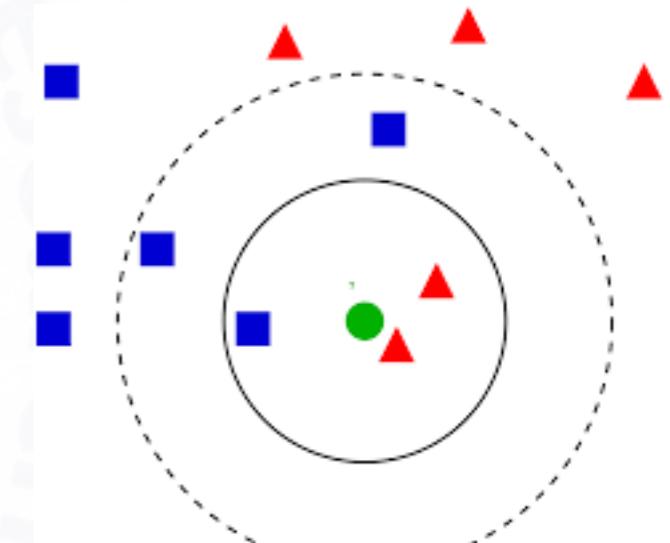
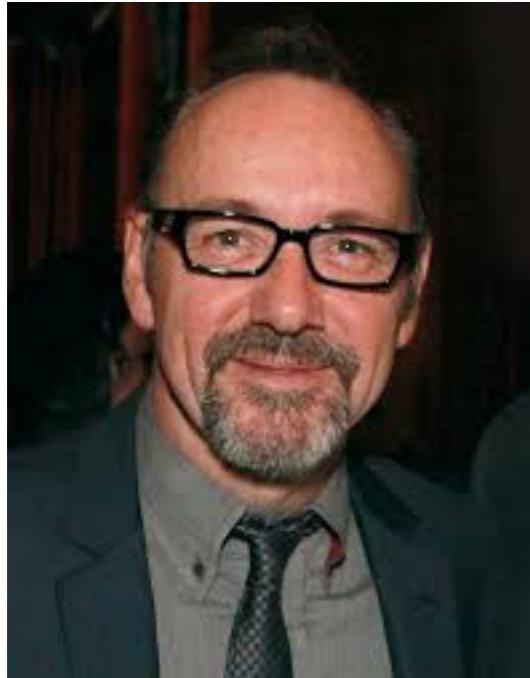
This Lecture

- This class overlaps with machine learning, statistics, artificial intelligence, databases but more stress on
 - **Scalability** (big data)
 - **Algorithms**
 - **Computing architectures**
 - Automation for handling
large data

What will we learn?

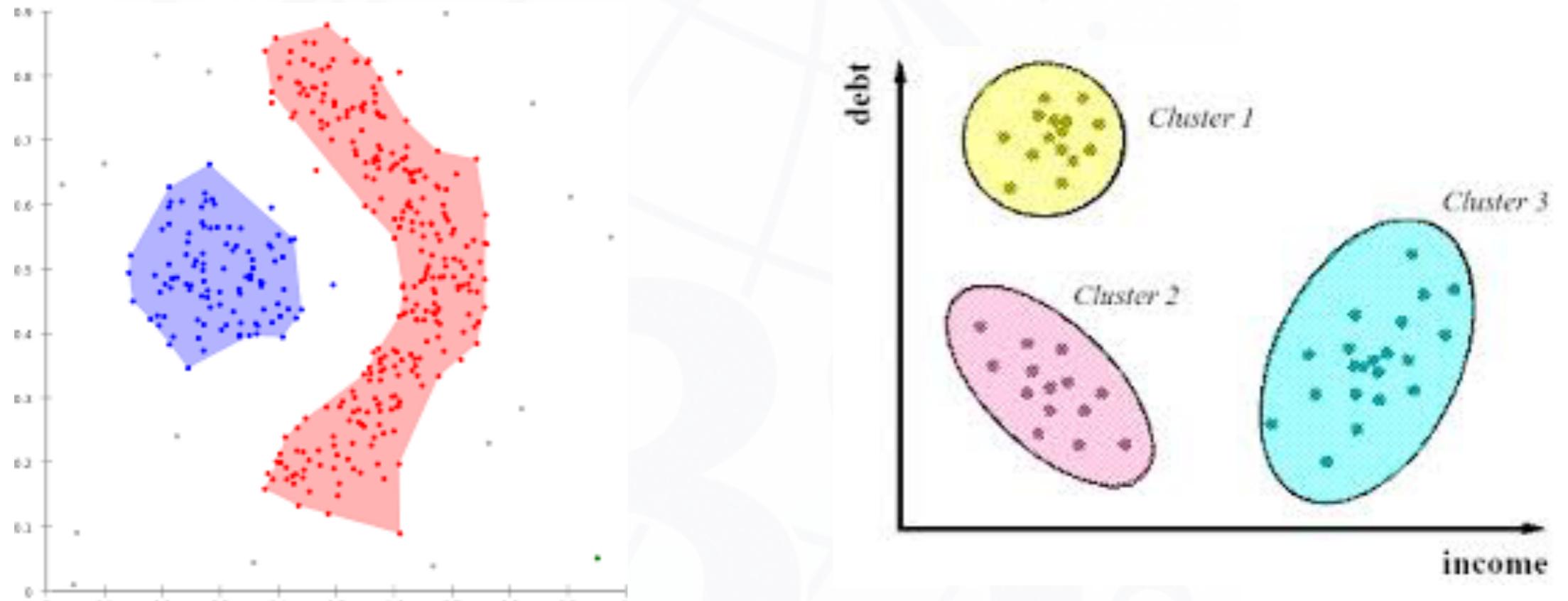
- We will learn to **solve real-world problems**:
 - Finding Similar Items
 - Market Basket Analysis
 - Spam detection
 - Duplicate document detection
- We will learn **various “tools”**:
 - Optimization (stochastic gradient descent)
 - Dynamic programming (frequent itemsets)
 - Hashing (LSH, Bloom filters)
 - Probabilistic analysis

Finding Similar Items



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

Clustering



- What are the **clusters** given large input collections ?

Mining Data Streams

B blogging	colecamples @stevier I'd prefer microblogging, but I like it a *lot* 5 minutes ago from web in reply to stevier ☆ ☺
T testing	Robin2go @stevier Oh, I like it! 6 minutes ago from web in reply to stevier ☆ ☺
Tw Twitter	stevier How's this one? 7 minutes ago from twitterrific ☆ ☺
ET engineering	cjh ok..... jumping on board with the new icons.... cool..... but, still home sick :(..... 16 minutes ago from twitterrific ☆ ☺
C Charles Bronson	fncll Charles Bronson is one of few actors that legitimately fits @JimGroom's ultimate term of respect 'bad-ass' - watched Death Hunt recently! 20 minutes ago from web ☆ ☺
V video	chrismillet jumping on the bandwagon. this is a little less ominous than my last twitter icon too. 21 minutes ago from twitterrific ☆ ☺
T testing	Robin2go AND, just realized @TLTSymposium is also my tag. Might have to change. So soon? Sigh. 21 minutes ago from twitterrific ☆ ☺
T testing	Robin2go @bpanulla I was just thinking the same thing as I was looking at MY twitter page! Great minds think alike... or something. 23 minutes ago from twitterrific in reply to bpanulla ☆ ☺
B blogging	colecamples @tiiimmyy How did u go from elated to deflated in a matter of tweets? 23 minutes ago from web in reply to tiiimmyy ☆ ☺
VW virtual worlds	micala re: @colecamples's office redesign.. I could use a 'serenity egg' like in the Google Zurich offices right about now. 23 minutes ago from web ☆ ☺
C Charles Bronson	fncll Trying to shake off putting on a really bad class last night. Talked too much, taught too little. Never got things rolling properly. Ugh. 23 minutes ago from web ☆ ☺
A audio	bpanulla Between the TLT logos and the Adobe icons, my Twitter friend box is starting to look like the Periodic Table. 24 minutes ago from web ☆ ☺
A audio	bpanulla <-- would've loved one with "Sonification" but this'll do!

Germany Trends · Change

#Weltgesundheitstag

Started trending in the last hour

#Spieltach

Started trending in the last hour

#Niederlande

Started trending in the last hour

#Checkpoint

151 Tweets

#Mavs

Just started trending

V-Mann des Verfassungsschutzes

813 Tweets

#5SOSONGRIMMY

61.4K Tweets

#WOBRMA

33.5K Tweets

#alsich9war

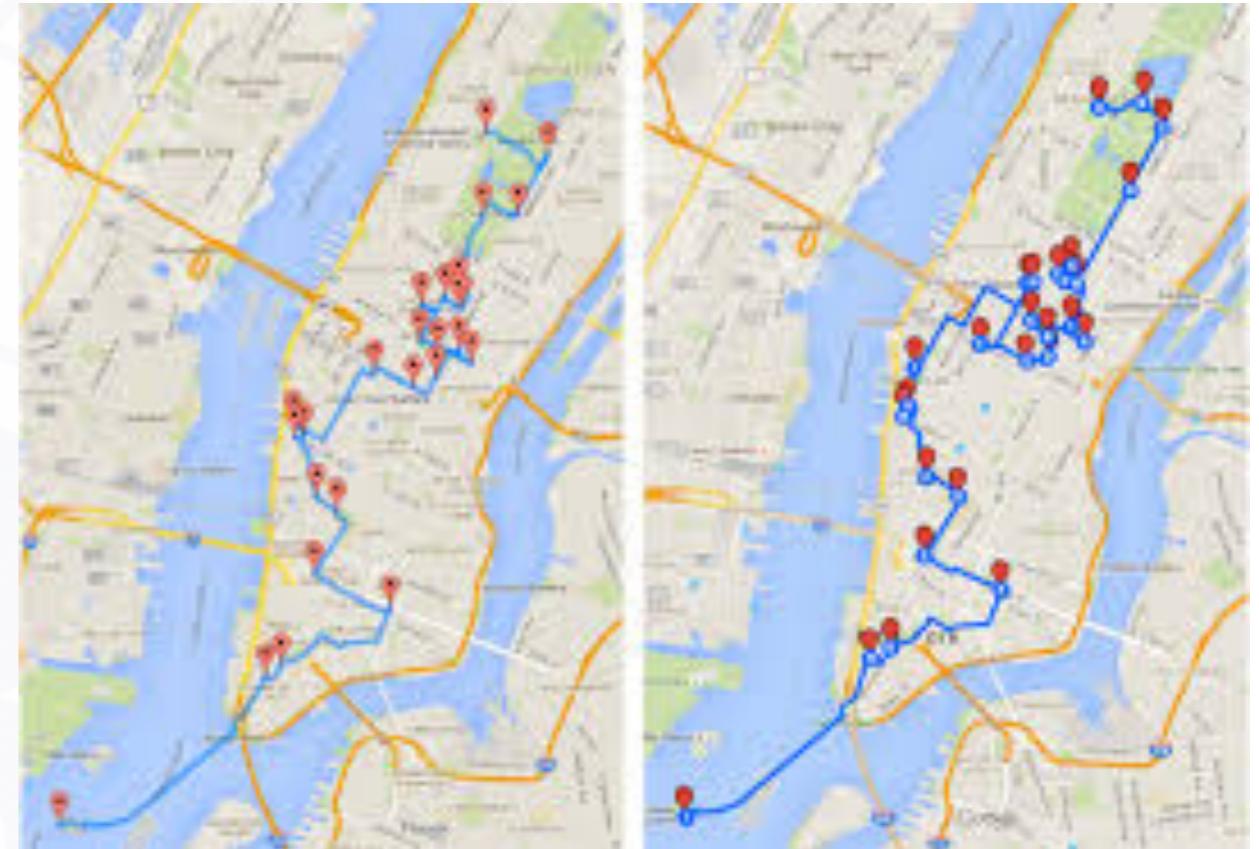
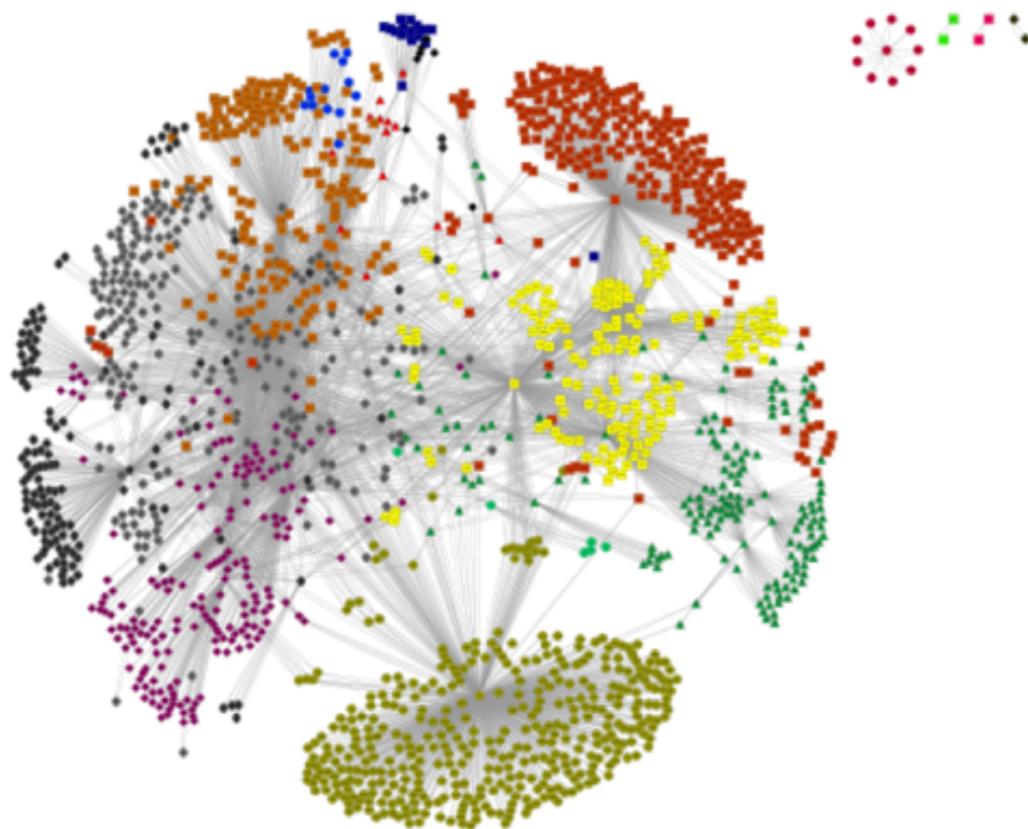
Trending for 11 hours now

#PSGMCI

171K Tweets

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

What will we learn?

- 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

About the Course

Course Logistics

- Course website:

<http://www.l3s.de/~anand/lsm16/>

- Lecture slides (at least 30min before the lecture)
- Homeworks, solutions

- **Readings:** Book **Mining of Massive Datasets** with A. Rajaraman and J. Ullman
- ITIS Students: part of the **theoretical foundations**

Free online:

<http://www.mmds.org>

Logistics: Communication

- **For e-mailing us, always use:**
 - anand@l3s.de,singh@l3s.de
 - Use this prefix in the subject field [**lsdm16**]
 - You can message via Stud.ip as well
- **We will post course announcements to the course webpage - via twitter (make sure you check it regularly)**

Work for the Course

- **7-8 Assignments:**
 - Theoretical and programming questions
 - **Assignments take time. Start early!!**
 - **5-6 questions per assignment, online on the lecture day**
 - **Due date Tuesday after the lecture 11:59 am**
- **How to submit?**
 - **Homework write-up:**
 - **Submit @ Jaspreet's office or (11:59 am),**
 - **Scan and send to singh@l3s.de (11:59 pm)**

Tutorials

- After the lecture (with Jaspreet)
- Successfully evaluated assignments will be offered for solution presentation
- Each presentation (whiteboard) should be ~10 minutes..max 15 minutes
- 3 successful presentations give you 1 bonus point
- 1 bonus point = 0.3 grade improvement in your final exam

Final Exam

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

Prerequisites

- **Algorithms and Data Structures**
 - Dynamic programming, basic data structures
- **Basic probability**
 - Moments, typical distributions, MLE, ...
- **KDD1**
 - Foundation lecture running this semester

Quiz

Quiz

- **Anonymous**
- **Tear off your Quiz Number on the top right**
 - Save it with you if you want to know your scores
- **Duration : 20 minutes**
- After the test: Shuffle the deck and redistribute
- Crowdsourcing