Analyzing Massive Data Sets
Summer Semester 2019

Prof. Dr. Peter Fischer
Institut für Informatik
Lehrstuhl für Datenbanken und Informationssysteme

Chapter 0: Introduction

About myself

- Since Oktober 2017:
  Chair for Databases and Information Systems
  Before
  – Juniorprofessor for Web Science at University of Freiburg
  – Senior Researcher/Oberassistent at Systems Group, ETH Zürich
  – PhD from ETH Zürich (worked at Uni Heidelberg, TU München)
- Research Interests:
  – Real Time Analytics/Data Streams/Temporal Data
  – Social Media Analytics
  – Databases on modern hardware (Main Memory, Cluster)
  – Analysis and adaptation of Information
  – Assurance of Data Quality: Provenance
- Contact:
  – peter.fischer@informatik.uni-augsburg.de
  – Office hours: Tuesday 14:30 – 15:30 @ 2051 (N)
    or by e-mail appointment

Basic Course Information

- Credits: 4V + 2U (ask examination office for ECTS)
- Language: English
  (feel free to ask/answer in German)
- Lecture: Tuesday/Thursday 10:00-11:30 2045 N
- Exercises (4 groups – 2056N):
  - Tuesday 14:00
  - Wednesday: 10:00
  - Thursday: 12:15 and 14:00
  - Do we need an English-only group?

Workload & Grading

- Exercises
  - Weekly exercise sheets with questions related to the lecture coverage
  - Attendance to exercise sessions is not mandatory, but it is highly recommended to do well in the exam.
- Exam
  - No prerequisites, enroll in STUDIS (punctually!)
  - Written exam, open book
  - July 24 16:30, Mensa

Exercise Sessions

- Two types of exercises:
  - Homework: solve yourself, discuss in session
  - Live exercises: solve together in session
  - Written solutions for homework only, posted in the following week
- No hand-in, no grading (you may ask for feedback on your solution on a best effort basis)
- Enrollment via Digicampus:
  April 23 18:00 – April 26, 17:59
- Sheets will be made available on Friday 14:00
- First sheet: April 20th
- First exercise sessions: Week starting April 29th

Course Material and Literature

- Slides will be uploaded on the evening before
- Recordings will be available a few days after the lecture
- Main book (and basis for many slides):
  Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of Massive Datasets* (2nd ed.). Cambridge University Press, New York, NY, USA.
- Available in the library and on http://www.mmds.org/
- Additional books for selected topics:
  - Jake VanderPlas: Python Data Science Handbook, O'Reilly
  - Chambers/Zaharia: Spark - The definitive guide (or others)
  - C. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval
  - Easly, Kleinberg: Networks, Crowds, and Markets

Prerequisites

- Programming (Info I/II)
  - We will do a short tutorial on Python:
    very popular as data analysis language, many good toolkits
  - C++/Java will also be useful
- Linear Algebra
  - Matrices
- Algorithms (~DS, Info 3)
  - Dynamic programming, basic data structures
  - Graphs and Graph Algorithms
- Basic probability ()
- Let's do a quick poll on the relevant backgrounds

Topics

- Big Data Platforms:
  - Single Node Tools (Python)
  - HDFS/MapReduce/Spark
- Text and High-Dimensional Data:
  - Similarity
  - Clustering
  - Retrieval and Ranking
- Graphs:
  - Link Analysis and Pagerank
  - Social Networks and Community Detection
  - Information Diffusion
- Streams and Temporal Data:
  - Basic Models
  - Sampling, Counting, Trends
- …

Course Motivation: New Analytics

- No longer just structured, “clean” business data:
  - Text data, photos, videos
  - Social media: social networks, social streams
  - Science
  - …
- Much broader range of analytics
  - Information Retrieval
  - Machine Learning: Classification, Mining
  - Statistics
  - Human Interaction: Crowdsourcing, Interactive exploration
- Much larger volumes (think Google, Facebook!)
- Unpredictable workloads
- Results required in real time

Course Motivation: New Platforms

- Increasing CPU/GPU core count: Massive Parallelism
- Increasing RAM, “slower” disks, flash, new storage
- Faster Networks and massive Distribution
  - Racks and Datacenters as new basic building blocks
  - Global Replication, Consistency and Access
- New Processing paradigms
  - Map/Reduce, Distributed In-Memory Computations
  - Graph Computation Systems
  - Event, Data Stream Processing

Data to Knowledge

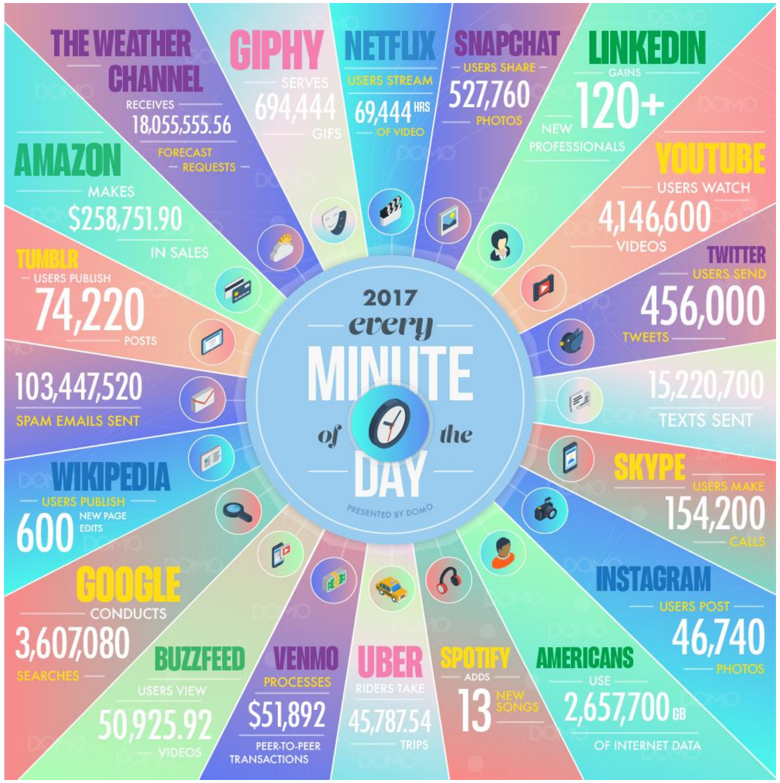
- Gathering insight is main selling point for most organizations when collecting data
- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - And ANALYZED ← this class
- Buzzword Bingo:
  Data Mining ≈ Big Data ≈
  Predictive Analytics ≈ Data Science

Buzzword 1: Big Data

- Shorthand for challenges occurring in current data management and analysis
- No longer just storage and retrieval, but also complex computations
- Often expressed as the 4 -7 V‘s (depending on source)
- Relative Term:
  – Not always Peta/Exabytes
  – My "Big Data" is not Googles, is not CERNs

1st V: Volume

- Scale of Data
  - Scientific applications (CERN: 70MPixel*40M/s, 15PB/year)
  - Genomics: (single genome > 1.5TB)
  - Web Data
  - …
- 90% of all data was created in the last two years!



=> Beyond what a single machine can handle

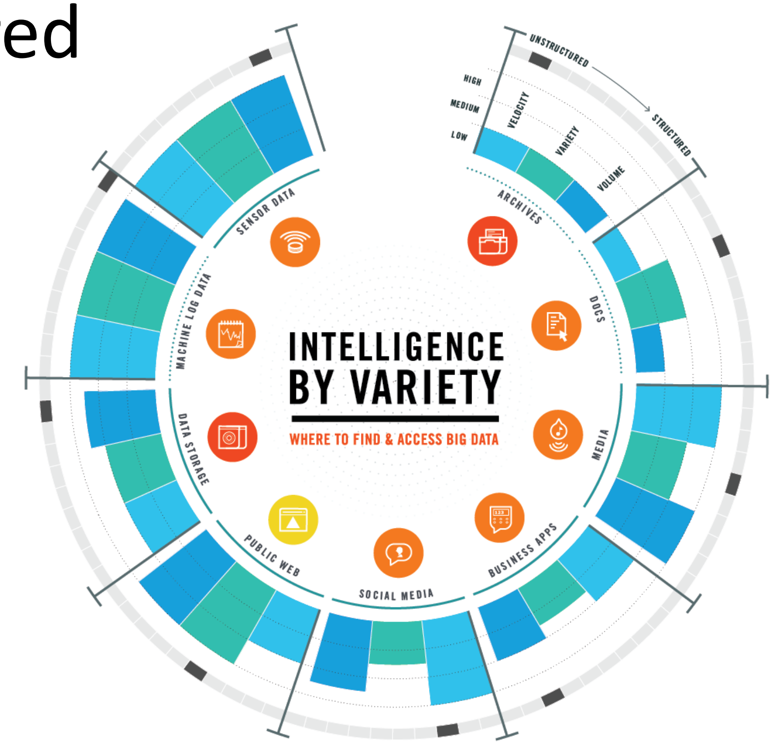
2nd V: Velocity

- Speed of data and expected reactions
- Stock exchanges
  (NASDAQ: >35K msg/s, 1ms for common operations)
- Social Media (>150K msg/s peak on Twitter)
- Environmental Sensors (>100 sensors on a car, ms response time)
- Web indexing (reindex within minutes, queries with less than 0.5 seconds)

- Storing is easy, quick answers are hard
- Data potentially unbounded

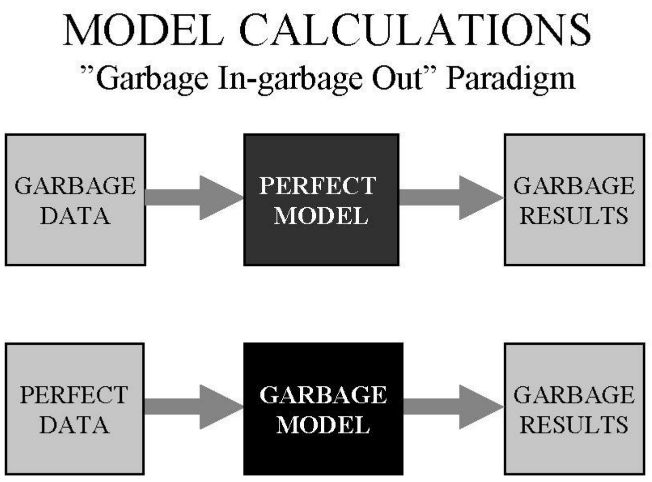
3rd V: Variety

- Form(at) of data not uniform
- Structured vs non-structured (or hidden structure): relations, graphs, text, audio/voice, video, …
- Broad range of sources: customers, transactions, logs, sensors, …
- Skewed data/Power laws dealing with popular data vs long tail



4th V: Veracity

- Uncertainty of Data
- Data Quality and Completeness
  - Sensor readings inconsistent (faults, calibration, …)
  - Social media messages contain slang, abbrevations, colloqualism, …
  - User Profiles faked, duplicated, …
- Interpretation
  - Underlying model unknown
  - Wrong choice of parameters



http://blog.potterzot.com/2007/09/25/garbage-in-garbage-out-and-the-desire-to-cover-our-own-ass-is-ruining-the-world/

5th V: Value

- Data contains value and knowledge
- Specific to application domain
- Ask the right questions, do not blindly apply methods
- Understand usefulness: customer analysis, trading, business/personal/ technology/… improvement)

Additional/Disputed V‘s

- Variability:
  properties of data change over time
- Visualization:
  complex data cannot be understood without appropriate presentation

Buzzword 2: Data Mining

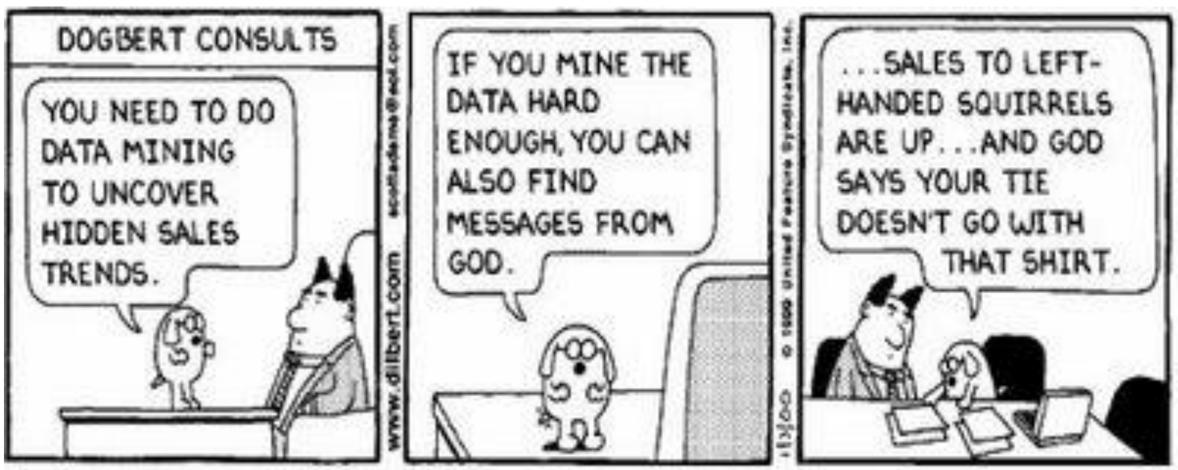
- Given lots of data
- Discover patterns and models that are:
  - **Valid**: hold on new data with some certainty
  - **Useful**: should be possible to act on the item
  - **Unexpected**: non-obvious to the system
  - **Understandable**: humans should be able to interpret the pattern

Data Mining Tasks

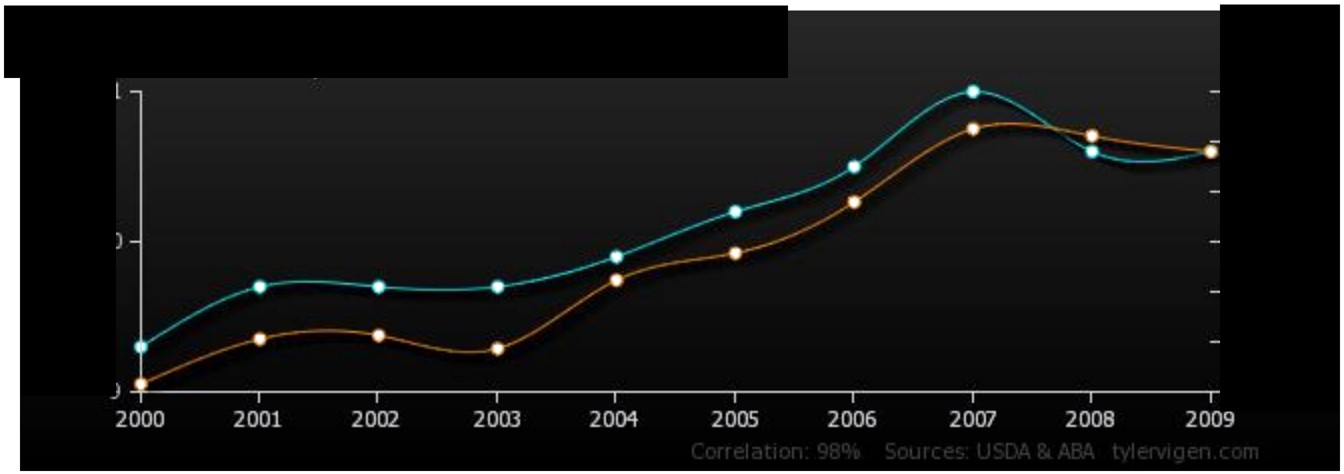
- Descriptive methods
  - Find human-interpretable patterns that describe the data
    - Example: Clustering
- Predictive methods
  - Use some variables to predict unknown or future values of other variables
    - Example: Recommender systems

Meaningfulness of Analytic Answers

- 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

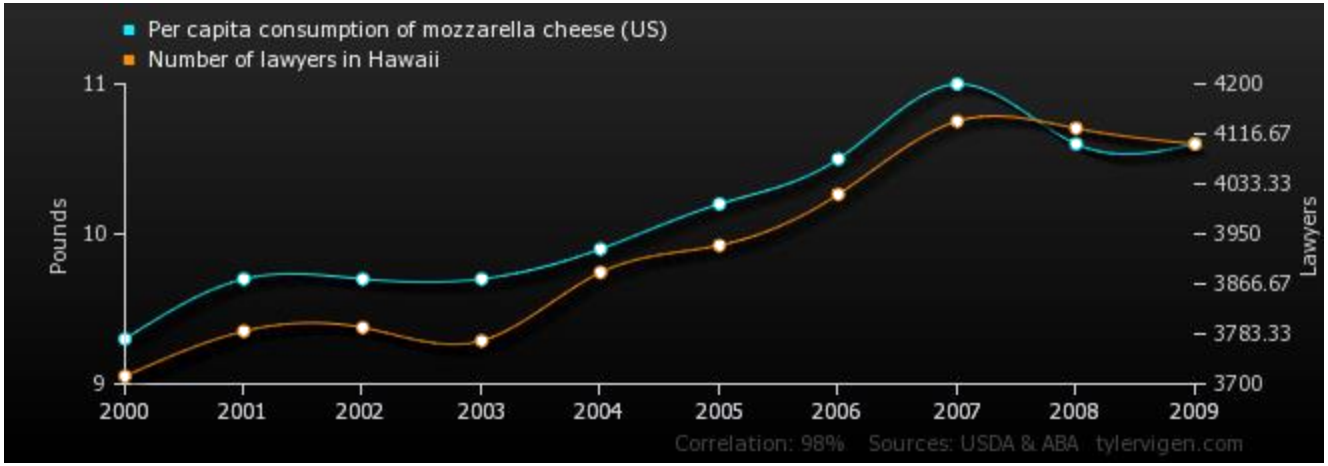


Interesting Correlations



- Looks legitimate
- Correlation Coefficient: 0.98
- What could it be?

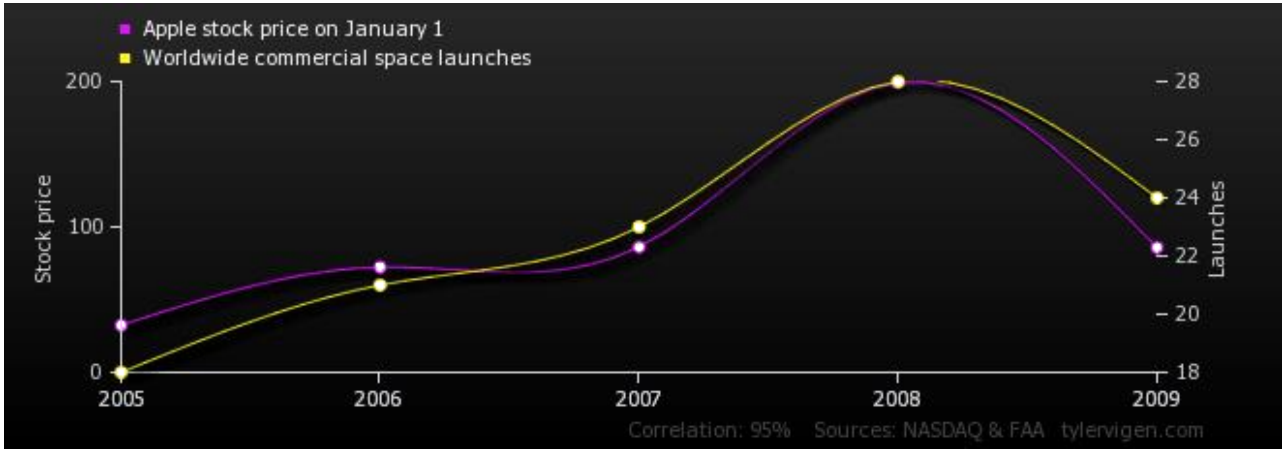
Interesting Correlations



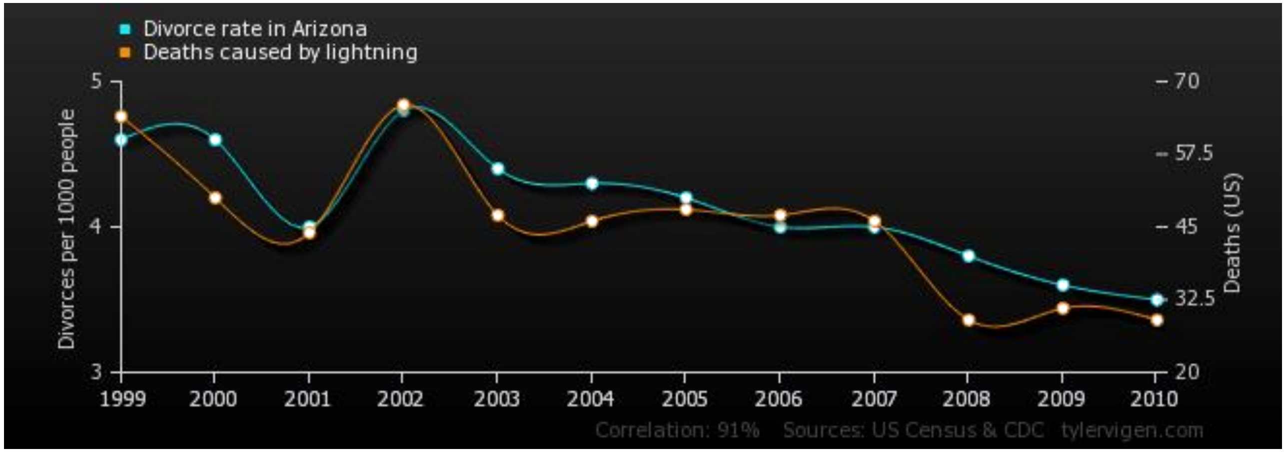
- Per capita consumption of mozarella cheese
- Number of laywers in Hawaii

Other examples

Apple Stock price (Jan 1st)
Vs.
Commercial space launches



Divorce rate in Arizona
Vs.
Deaths caused by lightning

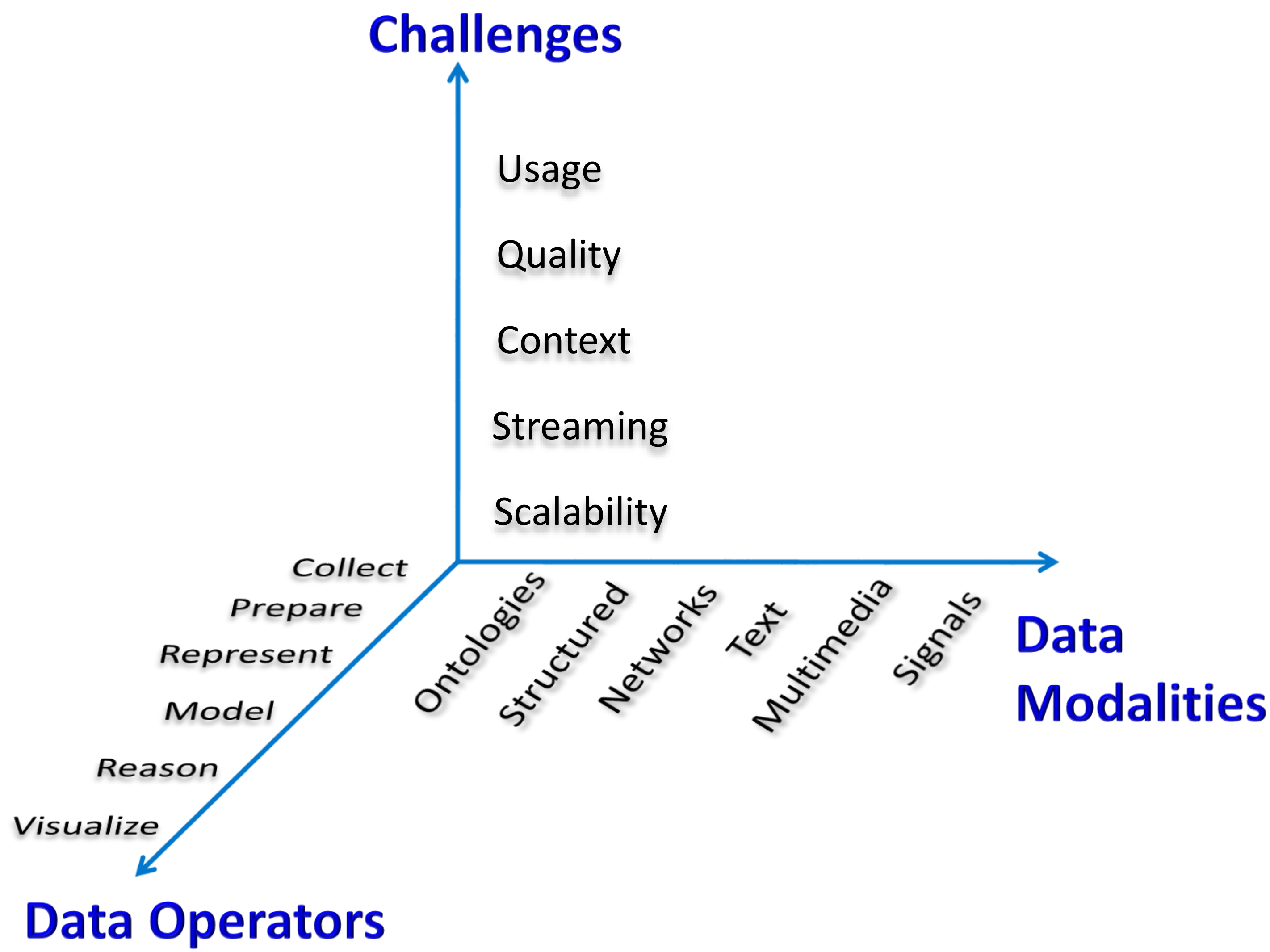


http://www.tylervigen.com/spurious-correlations

More serious: suspect detection

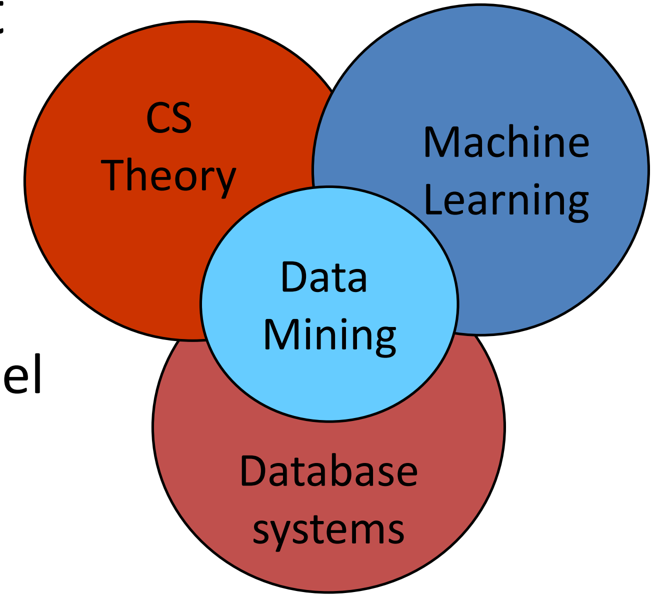
- We want to find (unrelated) people who at least twice have stayed at the same hotel on the same day
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
  - If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?
- Expected number of "suspicious" pairs of people:
  - 250,000
  - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

What matters when dealing with data?



Data Mining: Cultures

- Data mining overlaps with:
  - Databases: Large-scale data, simple queries
  - Machine learning: Small data, Complex models
  - CS Theory: (Randomized) Algorithms
- Different cultures:
  - To a DB person, data mining is an extreme form of analytic processing – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data-mining is the inference of models
    - Result is the parameters of the model
- In this class we will do both!



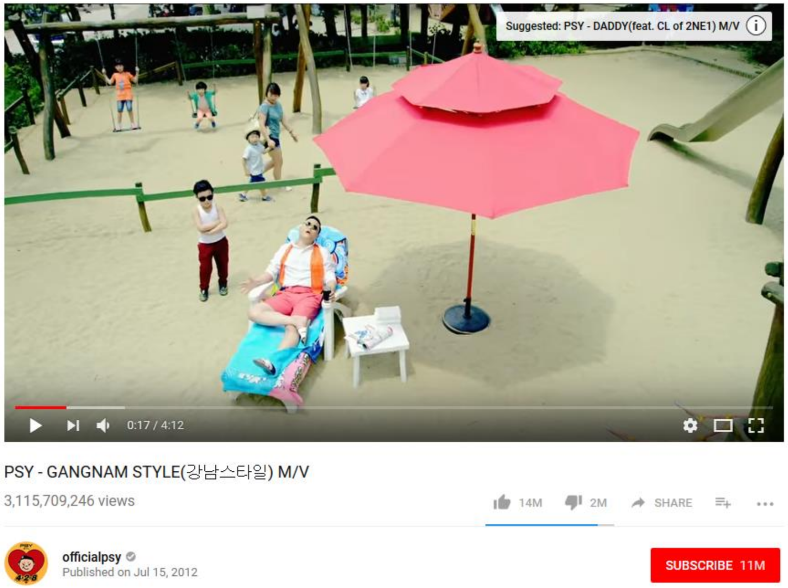
What will we learn?

- We will learn to mine different types of data:
  - Data is high dimensional
  - Data is labeled
  - Data is a graph
  - Data is infinite/never-ending
- We will learn to use different models of computation:
  - Single machine in-memory
  - MapReduce/Spark
  - Streams and online algorithms

Example Application: Tracing Information Diffusion

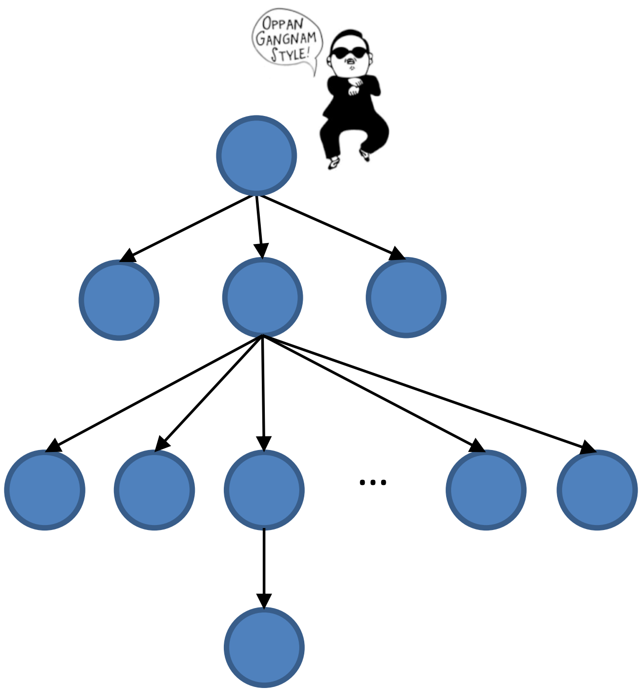
- Understanding how a information spreads
  - Who was the source of information?
  - Who forwarded it at what time and why?
  - …
- Conceptual similarity to epidemiology (also shared vocabulary)
- Applies techniques shown in this lecture
- Part of my ongoing research

Motivation



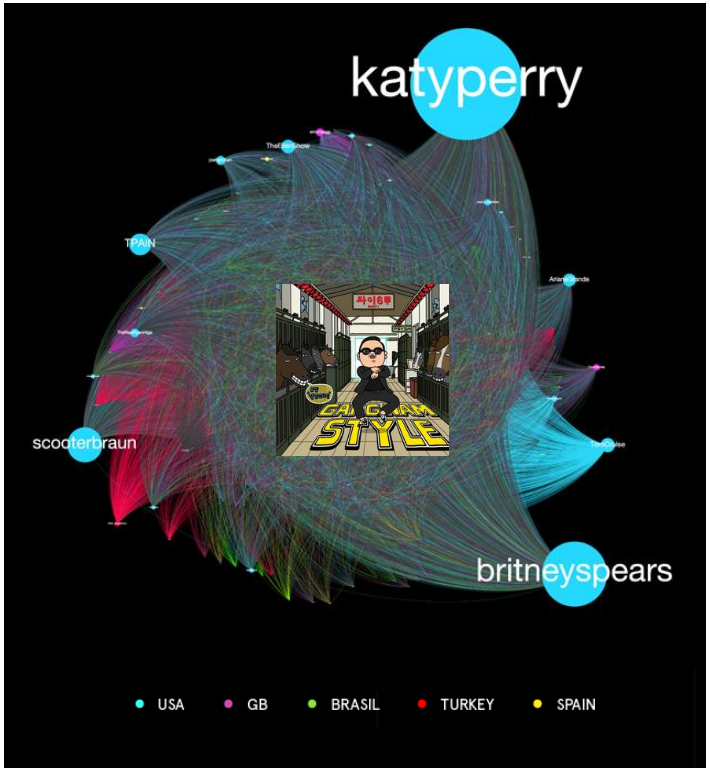
Suggested: PSY - DADDY(feat. CL of 2NE1) M/V
0:17 / 4:12
PSY - GANGNAM STYLE(강남스타일) M/V
3,115,709,246 views
14M
2M
SHARE
officialpsy
Published on Jul 15, 2012
SUBSCRIBE 11M

Motivation


OPPAN GANGNAM STYLE!
...


Suggested: PSY - DADDY(feat. CL of 2NE1) M/V
0:17 / 4:12
PSY - GANGNAM STYLE(강남스타일) M/V
3,115,709,246 views
14M
2M
SHARE
officialpsy
Published on Jul 15, 2012
SUBSCRIBE 11M

Motivation



Source:
http://blog.datasift.com/2013/05/08/gangnam-style-vs-harlem-shake-the-take-by-face/

Motivation

- 2016 US precedential election: orchestrated bots supporting certain candidates
- Trump received disproportionally positive messages
  from bots compared to Clinton.
- Biases his public perception & endangers democratic processes

Motivation

- **Large share of population** participates in social media
- **Large audiences**:
  - Information can be **easily spread** and consumed:
  - No information **verification**/ provenance
- Identify:
  - who spreads information and influences others?
  - how information is diffused?
  - what are the sources
    (indications for trust and relevance assessment)?
- Analyze in a **prompt way** to mitigate the negative effects of diffusion

Research Question Q1

How to model and trace information by unraveling user-to-user influence?

Challenges:

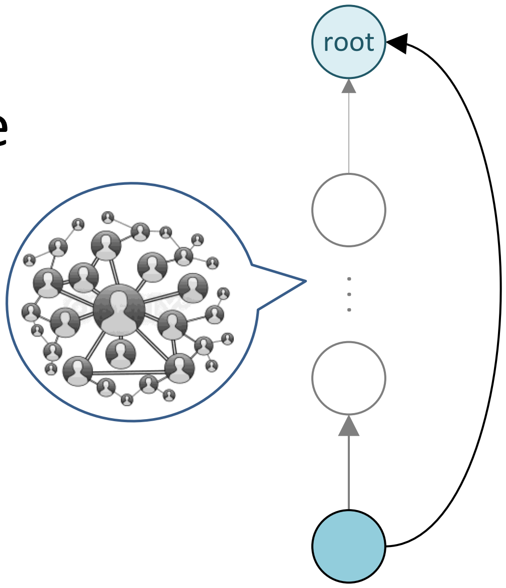
- Incomplete/ non existing social media provenance
- Latent/ external influence
- Lack of ground truth & high uncertainty
- Incomplete datasets/ API restrictions
- Lack of consistent models

Research Question Q1

How to model and trace information by unraveling user-to-user influence?

Contributions:

- Identification, classification & computation of user interactions
  - Explicit: partial social media provenance
    - Direct linkage based
    - Source based
      → Inference: social graph



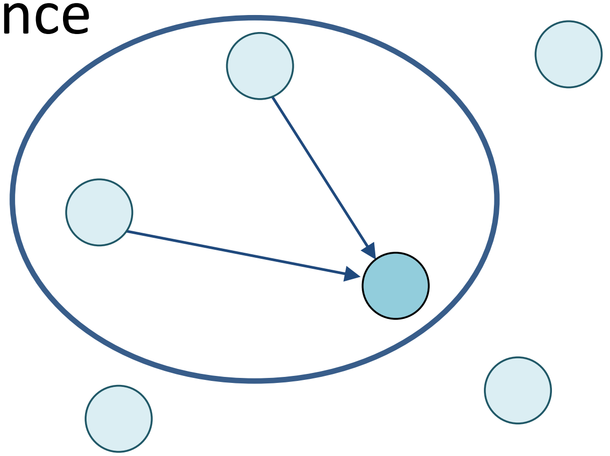
Source based

Research Question Q1

How to model and trace information by unraveling user-to-user influence?

Contributions:

- Identification, classification & computation of user interactions
  - Explicit: partial social media provenance
    - Direct linkage based
    - Source based
      → Inference: social graph
  - Implicit: latent influence
    - Similarity based
    - Additional influence indicators
      → user conventions to reveal influence



Research Question Q2

Is it feasible to trace information diffusion in an online fashion?

Challenges:

- Fast social media rates & huge social graphs
- Very large search space
- Limited support for systems that traces information diffusion in a online fashion
- Scalability, short response times

Research Question Q2

Is it feasible to trace information diffusion in an online fashion?

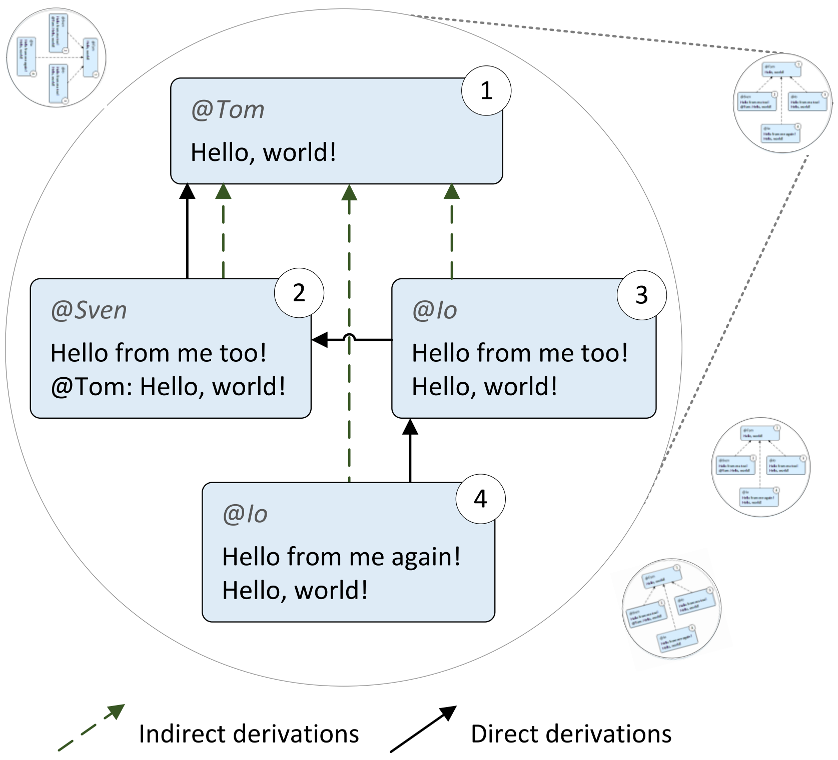
Contributions

- Explicit interactions:
  - Streaming reconstruction of information diffusion graphs
    → stream iterative problem combined with large social graphs
- Implicit interactions:
  - Streaming computation of latent influece
    → similarity computations and clustering over infinite streams
-

Implicit Interactions

How to model and trace information by unraveling user-to-user influence?

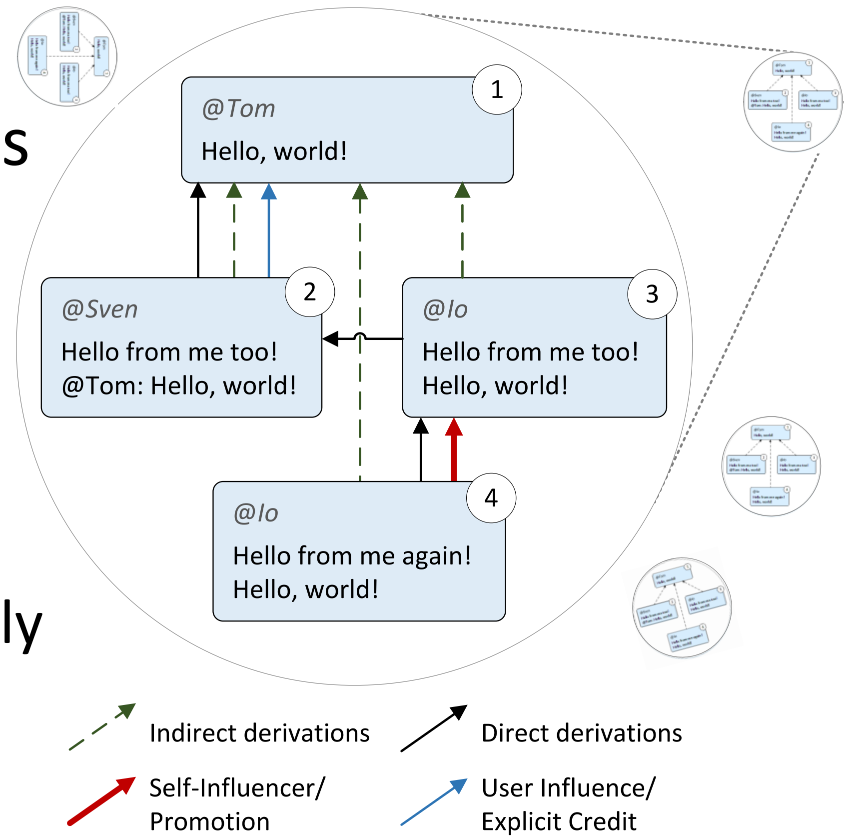
- Clustering of similar messages with SimClus, lower similarity threshold
  - Tf-idf, cosine similarity
- Within each cluster:
  - Coarse grained provenance: Indirect derivations → oldest message
  - Fine grained provenance: Direct derivations → most similar message



Implicit Interactions – Influence indicators

How to model and trace information by unraveling user-to-user influence?

- Empirical methods to identify user activity patterns
- In this example:
  – User Influence/Explicit credit: mention of the influencer → @Tom
  – Self Influence/ Promotion: user promotes some previously written content



Web-scale implicit provenance reconstruction - Challenges

Is it feasible to trace information diffusion in an online fashion?

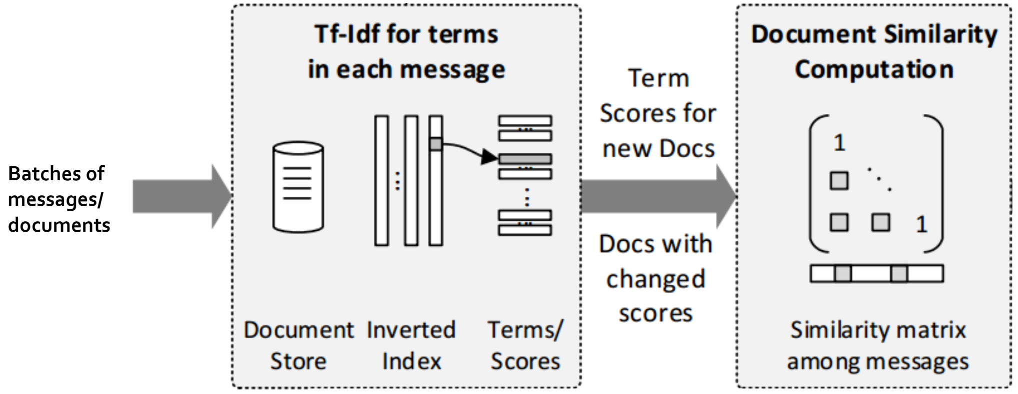
- All previous messages might be relevant (in the range of millions)
- Constant changes in TF-IDF model, clustering and provenance
- Similarity matrix → quadratic complexity (# of documents)

Web-scale implicit provenance reconstruction - Architecture

Is it feasible to trace information diffusion in an online fashion?

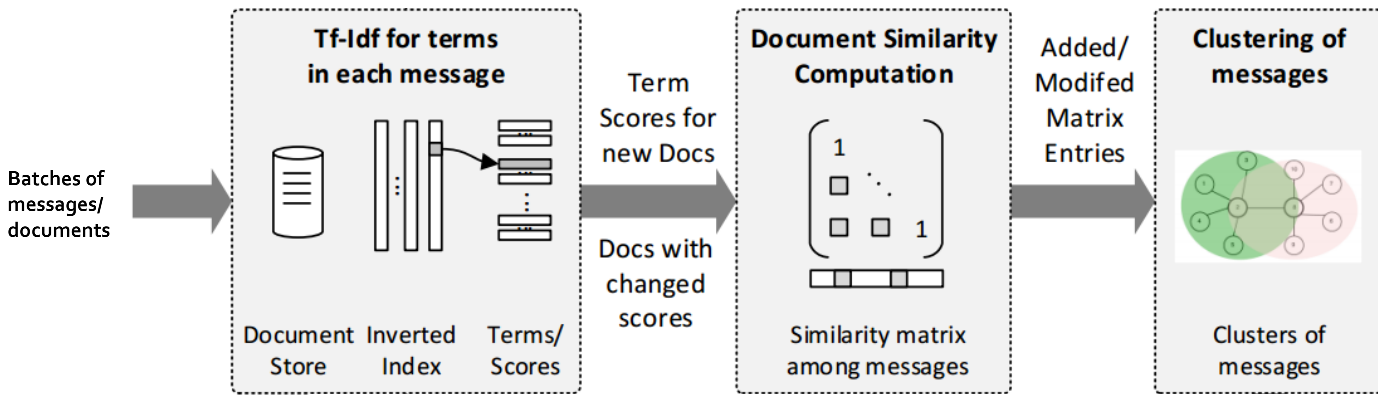
Web-scale implicit provenance reconstruction - Architecture

Is it feasible to trace information diffusion in an online fashion?



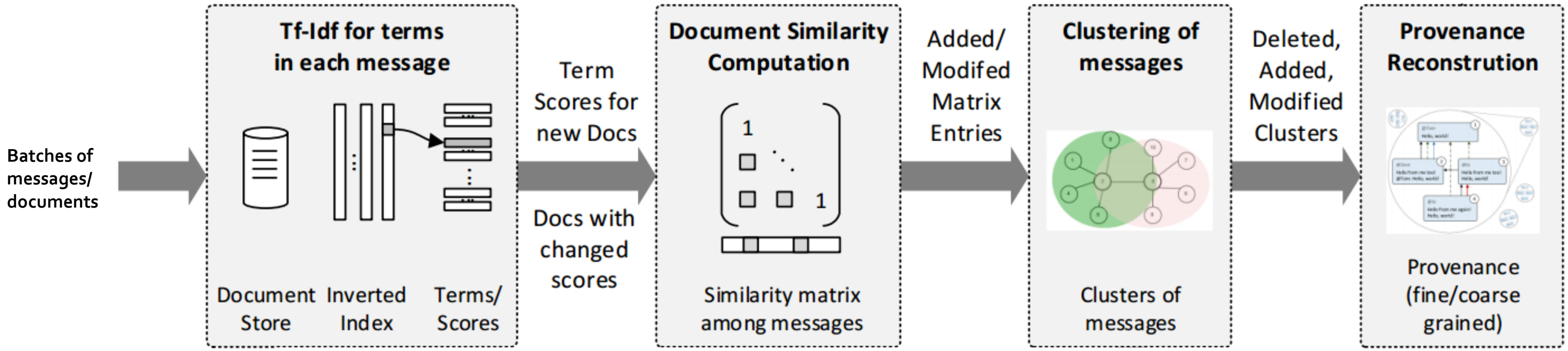
Web-scale implicit provenance reconstruction - Architecture

Is it feasible to trace information diffusion in an online fashion?



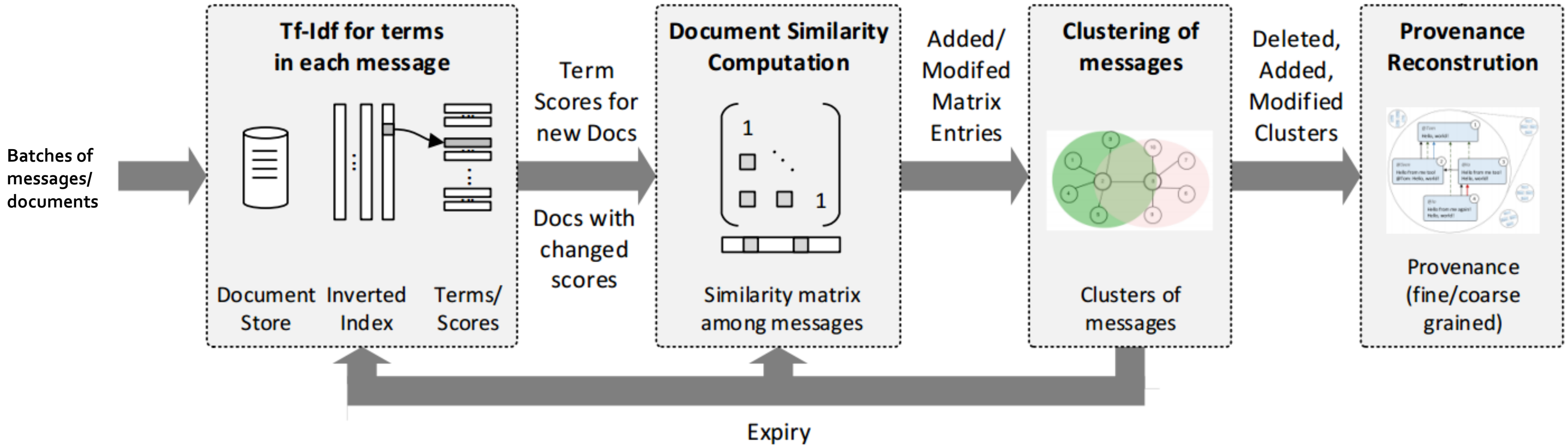
Web-scale implicit provenance reconstruction - Architecture

Is it feasible to trace information diffusion in an online fashion?



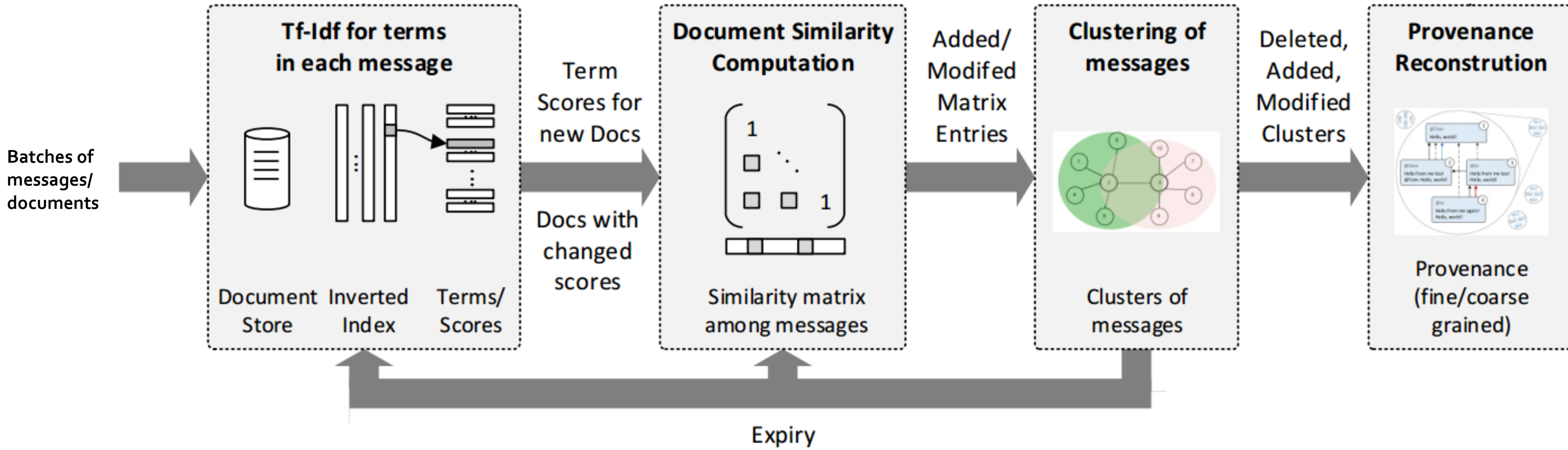
Web-scale implicit provenance reconstruction - Architecture

Is it feasible to trace information diffusion in an online fashion?



Web-scale implicit provenance reconstruction - Architecture

Is it feasible to trace information diffusion in an online fashion?

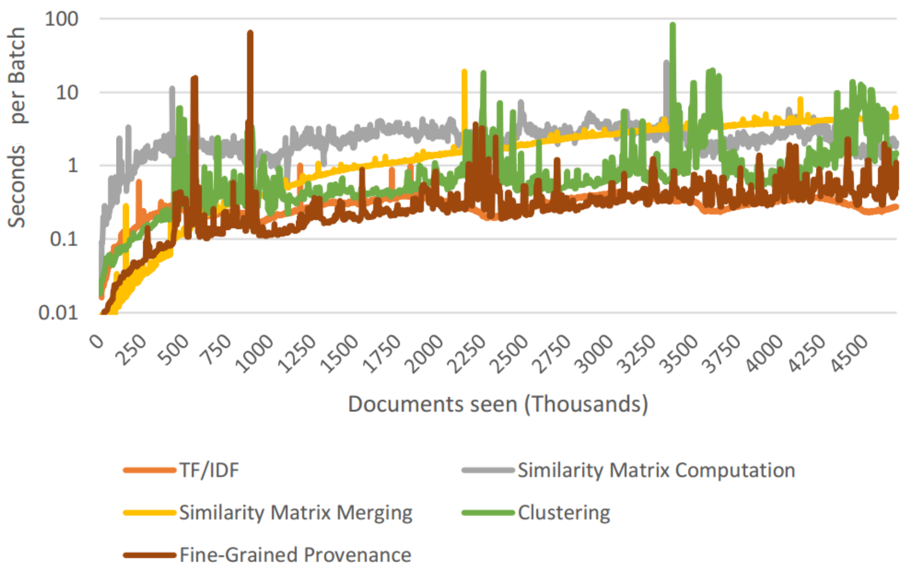


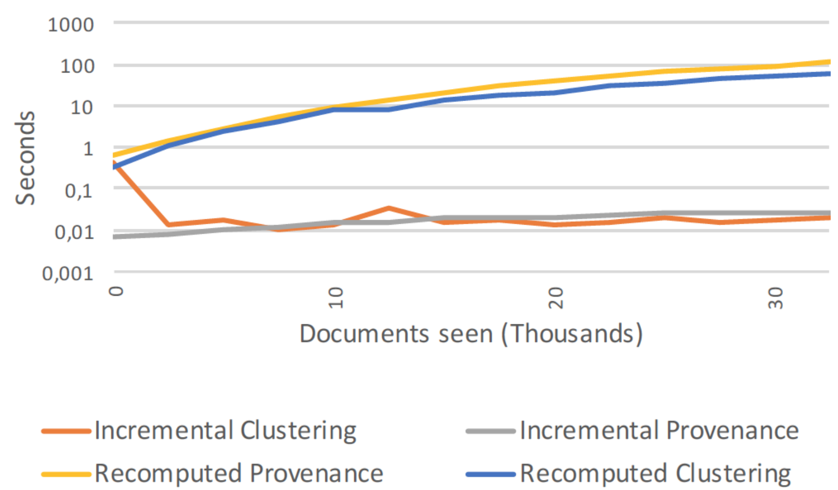
Optimizations at every stage to

1. Reduce unnecessary computations
2. Update the model on demand

Large scale Evaluation – Computational Times

- Dataset: 2012 Olympics in London, terms:“olympics”& “london2012, August 3 to 7th, 2012, ~4.6 M messages, similarity threshold 0.75, batch size: 2500
- Results:
  - Stable computational costs over time, scalability for the twitter streams
  - Incremental vs re-computed: up tp 4 orders of magnitude speed-up





Wrap up

- Data has little value on itself, extracting knowledge is crucial
- Fallacy of mining "fake" results
- Complexity and Volume make "mining" hard
- Wide variety of models and methods
- Mindset+Methods from multiple directions needed