Big Data with Big Spoon
Veliko podatkov z veliko žlico

Dunja Mladenić
Jožef Stefan Institute and
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

BIG-DATA DEFINITION

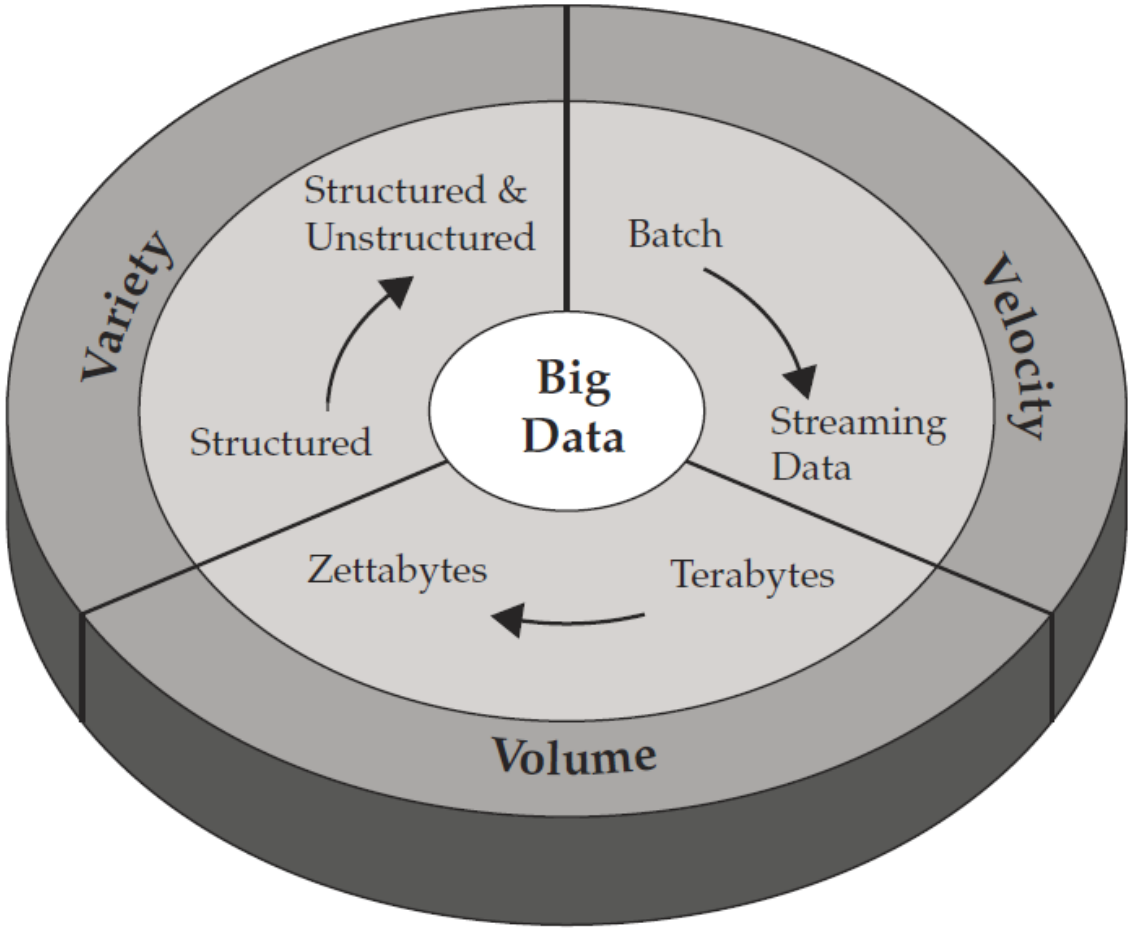
…so, what is Big-Data?

- 'Big-data' is similar to 'Small-data', but bigger
  - Recently getting popular expression "Midsize data"

- Having data bigger requires somewhat different approaches
  - techniques, tools, architectures
- …with an aim to solve new problems
  - …or old problems in a better way.

Characterization of Big Data: volume, velocity, variety (V3)

- **Volume** – challenging to load and process (how to index, retrieve)
- **Variety** – different data types and degree of structure (how to query semi-structured data)
- **Velocity** – real-time processing influenced by rate of data arrival



From “Understanding Big Data” by IBM

The extended 3+n Vs of Big Data

- 1. **Volume** (lots of data = “Tonnabytes”)
- 2. **Variety** (complexity, curse of dimensionality)
- 3. **Velocity** (rate of data and information flow)

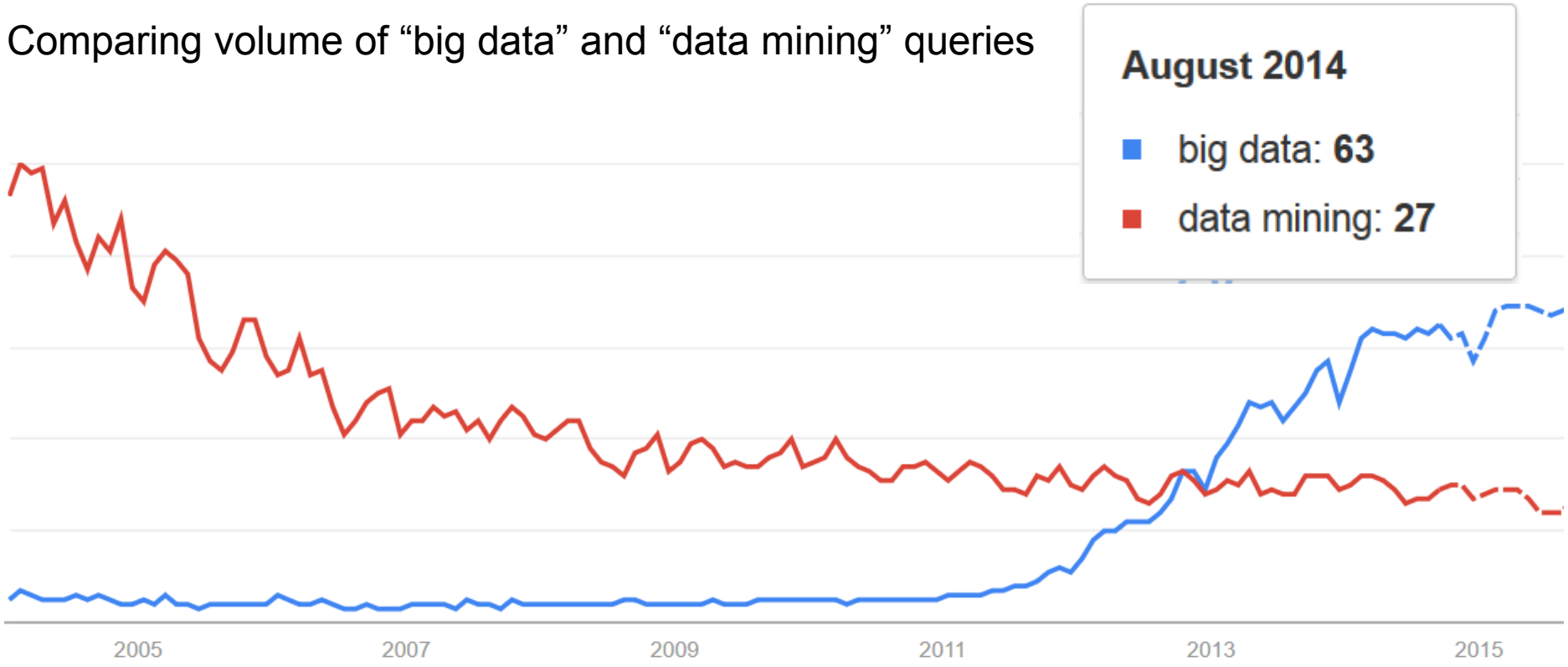
- 4. **Veracity** (verifying inference-based models from comprehensive data collections)
- 5. **Variability**
- 6. **Venue** (location)
- 7. **Vocabulary** (semantics)

MOTIVATION FOR BIG-DATA

Big-Data popularity on the Web

(through the eyes of "Google Trends")

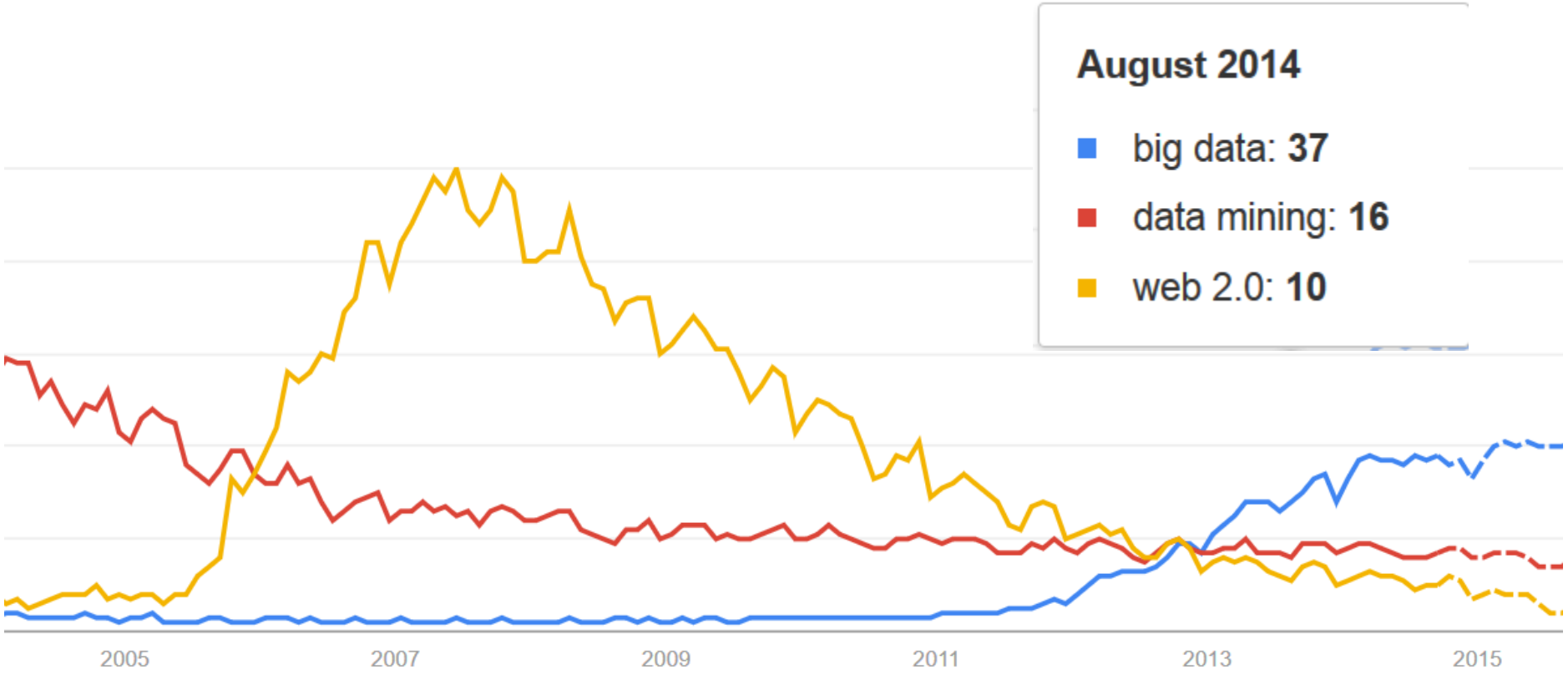
Comparing volume of "big data" and "data mining" queries



http://www.google.com/trends/explore#q=big%20data%2C%20data%20mining

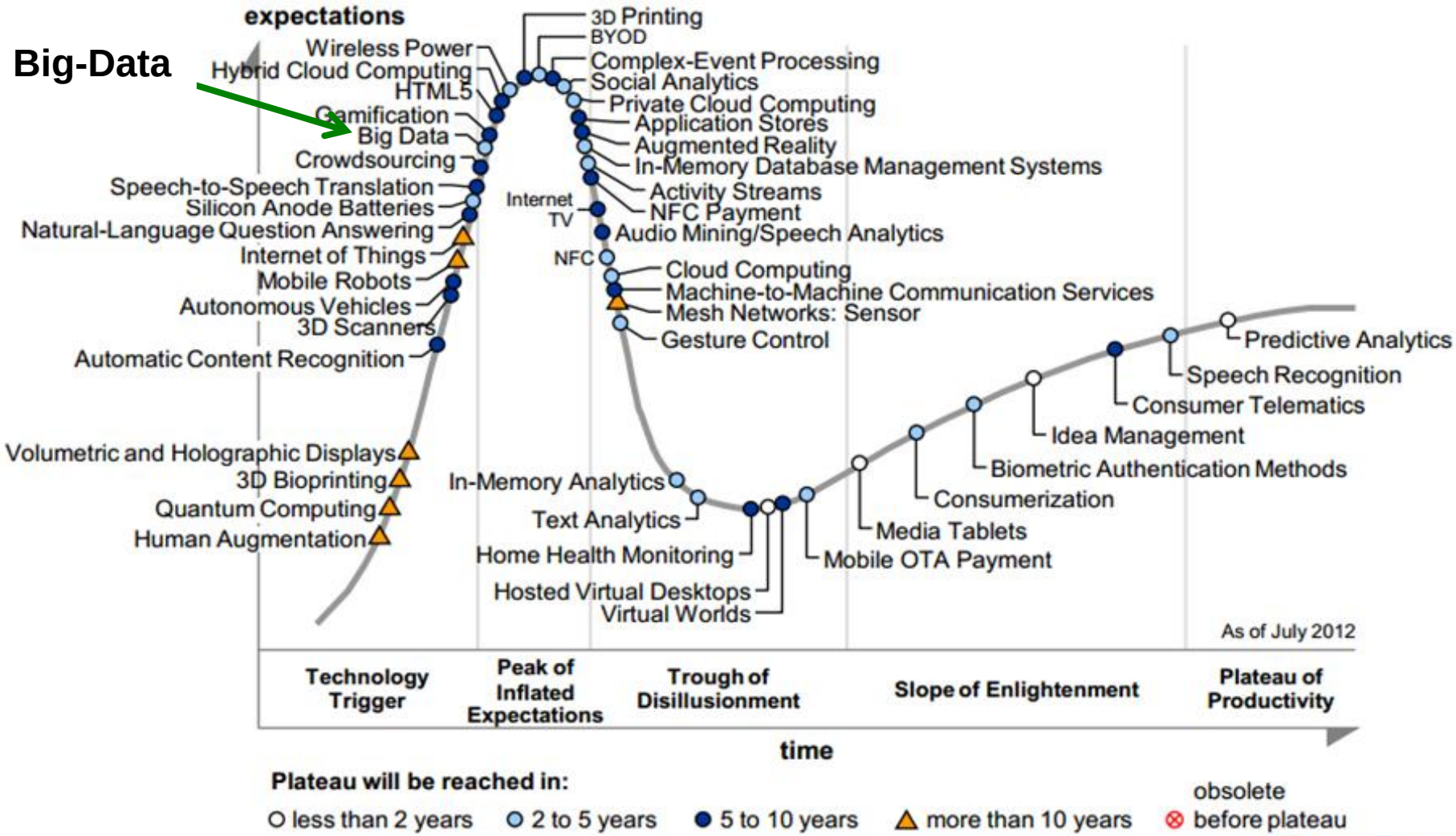
…but what can happen with "hypes"

…adding "web 2.0" to "big data" and "data mining" queries volume



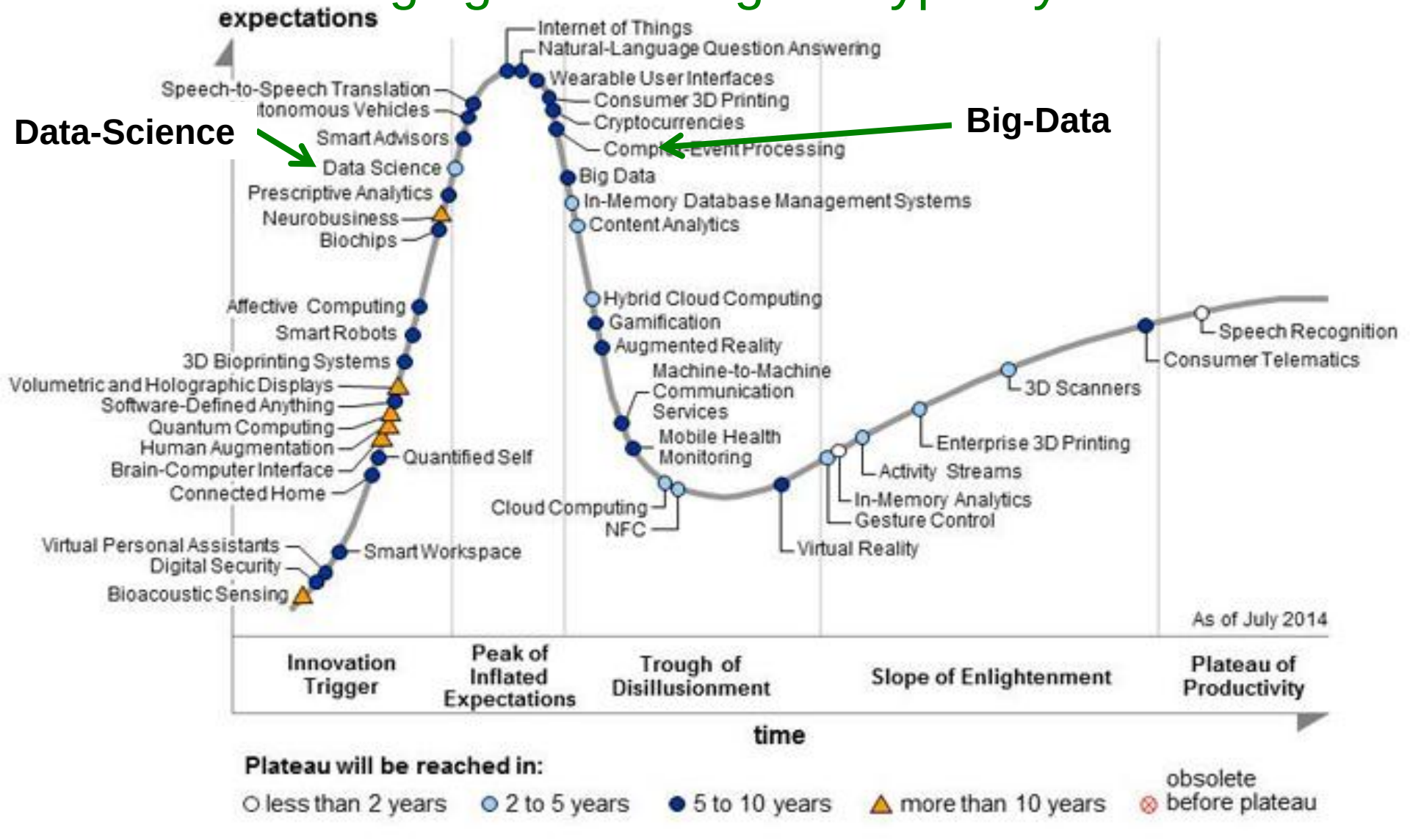
http://www.google.com/trends/explore#q=big%20data%2C%20data%20mining%2C%20web%202.0

Gartner: Emerging Technologies Hype Cycle 2012



Gartner

Gartner: Emerging Technologies Hype Cycle 2014



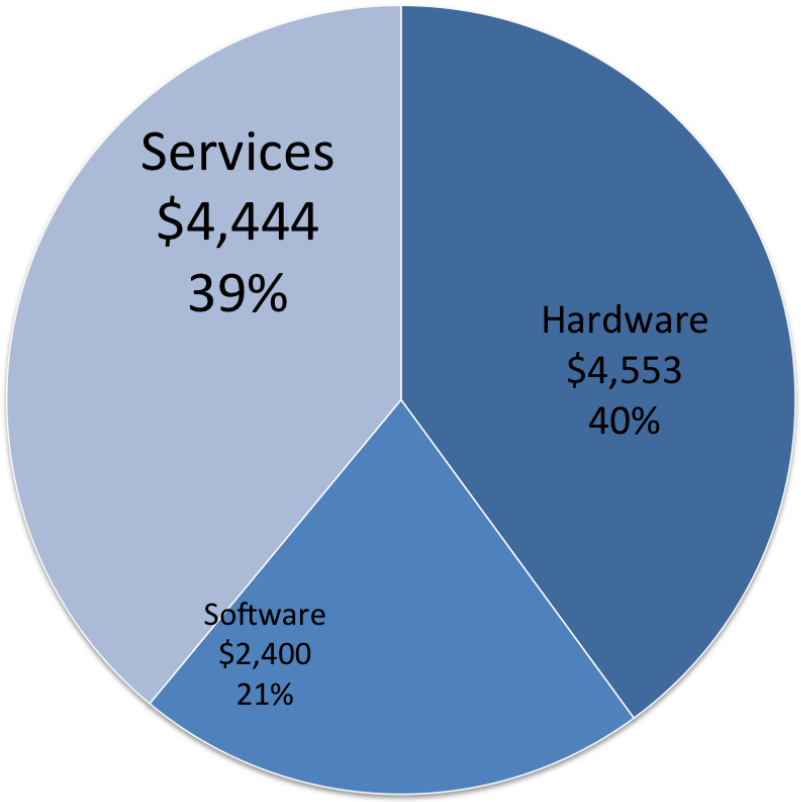
Why Big-Data Now?

- Key enablers for the appearance and growth of “Big Data”:
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data

Big Data Revenue by Type, 2012

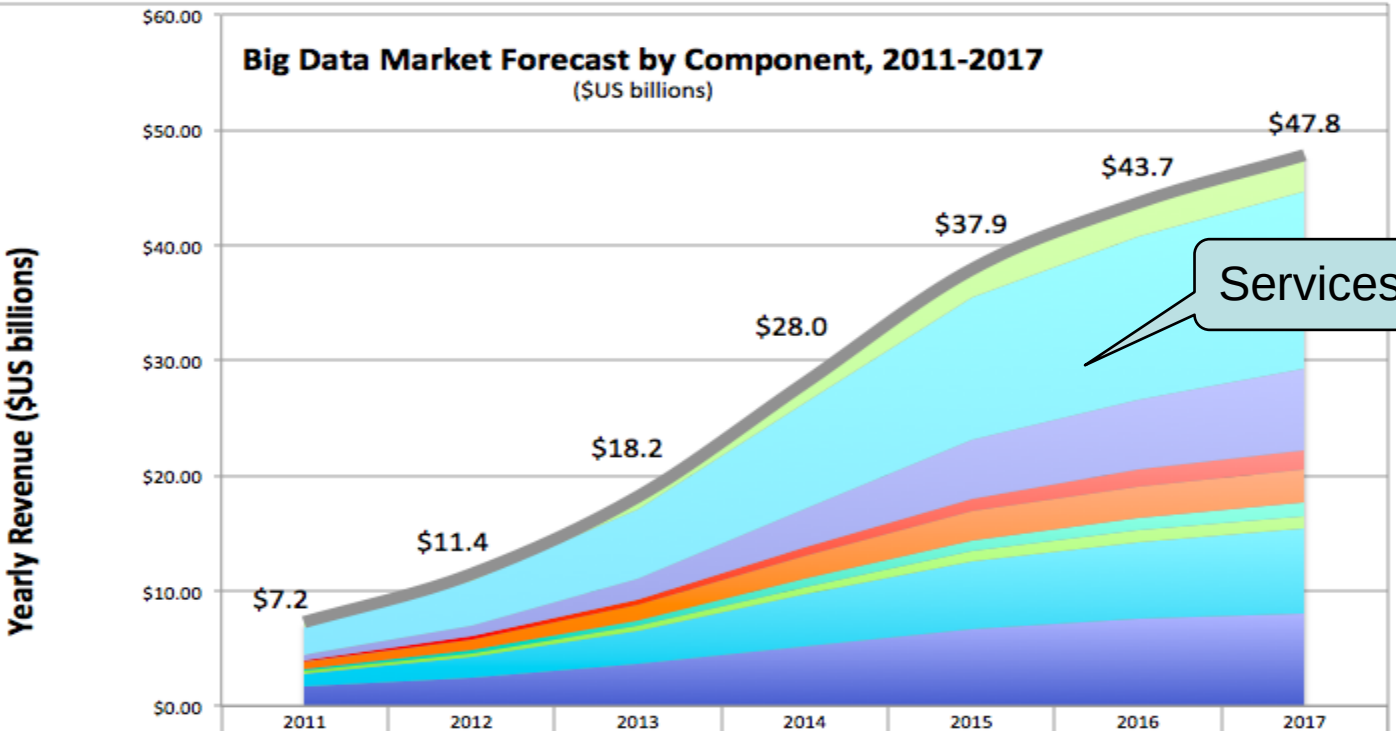
(http://wikibon.org/w/images/f/f9/Segment_-_BDMSVR2012.png)

Wikibon



Big Data Market Forecast (2011-2017)

(http://wikibon.org/w/images/b/bb/Forecast-BDMSVR2012.png)



	2011	2012	2013	2014	2015	2016	2017
Big Data XaaS Revenue	$0.35	$0.61	$1.05	$1.74	$2.47	$2.91	$3.24
Big Data Professional Services Revenue	$2.45	$3.87	$6.10	$9.29	$12.37	$14.14	$15.38
Big Data Application (Analytic and Transactional) Software	$0.49	$0.94	$1.80	$3.29	$5.02	$6.15	$7.00
Big Data NoSQL Database Software	$0.10	$0.19	$0.39	$0.73	$1.14	$1.41	$1.62
Big Data SQL Database Software	$0.72	$1.02	$1.45	$1.99	$2.47	$2.73	$2.90
Big Data Infrastructure Software	$0.16	$0.26	$0.43	$0.70	$0.96	$1.12	$1.24
Big Data Networking Revenue	$0.18	$0.28	$0.44	$0.67	$0.89	$1.02	$1.11
Big Data Storage Revenue	$1.16	$1.83	$2.89	$4.40	$5.86	$6.70	$7.28
Big Data Compute Revenue	$1.64	$2.45	$3.64	$5.23	$6.70	$7.50	$8.06
Total Big Data Revenue	$7.2	$11.4	$18.2	$28.0	$37.9	$43.7	$47.8

TECHNIQUES & RESEARCH

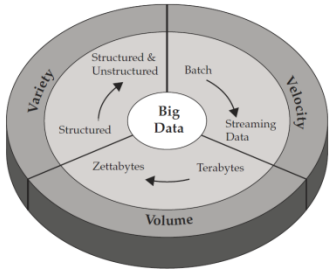
When Big-Data is really a hard problem?

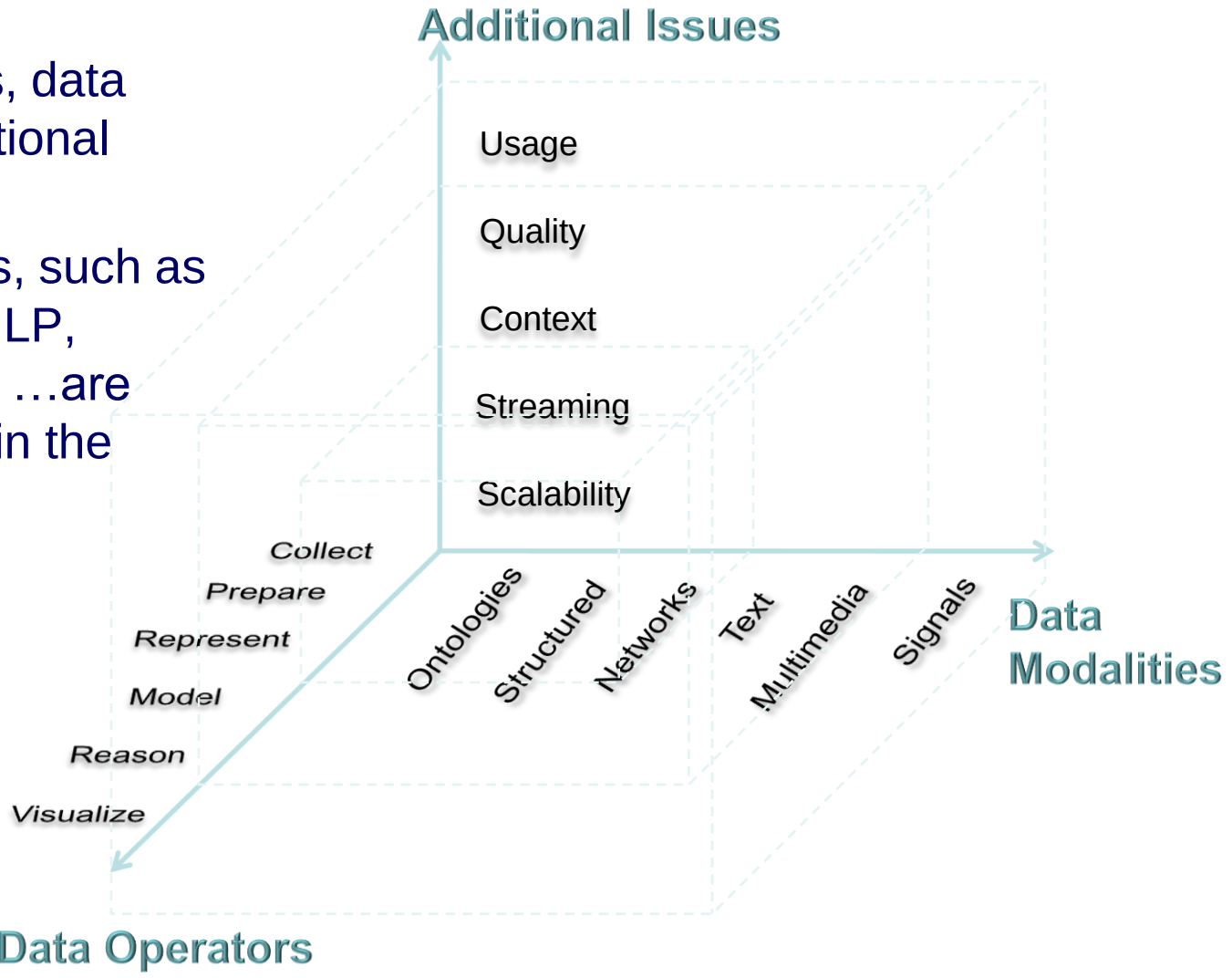
- when the operations on data are complex:
  - eg., simple counting is not a complex problem
  - Modeling and reasoning with data of different kinds can get extremely complex

- good news about big-data:
  - often, because of vast amount of data, modeling techniques can get simpler (e.g. smart counting can replace complex model-based analytics)…
  - …as long as we deal with the scale

What matters when dealing with data?

- Data modalities, data operators, additional issues
- Research areas, such as IR, KDD, ML, NLP, Semantic Web, …are sub-cubes within the data cube





Truth or a random phenomenon?

Risk with “Big Data mining”

- we can “discover” patterns that occur by chance
- …if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

Bonferroni’s principle

find a statistical artifa e looking for

– a patter is there by a he phenomenon

“*...truth is simple, straight and with a smile. You don't have to remember it. You have to say it. You know it and then you have to live it. It is so simple.* ”

[Y.Bhajan]

Why is Big Data *BIG?*

Mostly due to repeated observations over time and/or space

- Examples
  - Web logs with millions of visits per day
  - Supermarket transactions log - thousands of retail stores with tens of thousands of products and millions of customers
  - Satellites regularly sending images

Big data – “data whose size forces us to look beyond the tried-and-true methods that are prevalent at the time”

[A. Jacobs, CACM-2009]

Big Data from Data Stream

Data stream is a common source of **big data**

- web logs, social media, stock market, sensor networks,…

- Data stream **management**
  - Problematic are blocking query operators – need the entire input to produce any result (eg, sort, sum, max)
  - Use approximations, sampling, window of data
- Data stream **processing**
  - Maintain simple statistics on stream (mean, standard deviation)
  - Use time window:
    - sliding (fixed size – eg. the last 100 values),
    - landmark (fixed start – eg. from the start of the day)
    - tilted (recent data in more details – eg, last hour in 15 mins, last day in 24 hours, last month in days, last year in months)

Big Spoon for Big Data

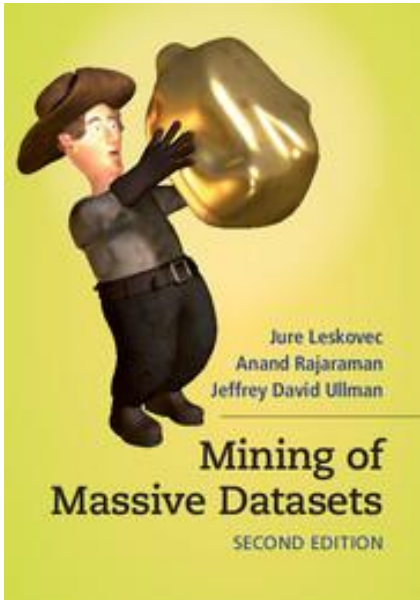
- **Smart sampling** of data
  - …reducing the original data while not losing the statistical properties of data
- **Finding similar items**
  - …efficient multidimensional indexing
- **Incremental updating** of the models
  - (vs. building models from scratch)
  - …crucial for streaming data
- **Distributed linear algebra**
  - …dealing with large sparse matrices

Analytical operators on Big-Data

- On the top of the previous ops we perform usual data mining/machine learning/statistics operators:
  - Supervised learning (classification, regression, …)
  - Non-supervised learning (clustering, different types of decompositions, …)
  - …
- …we are just more careful abut the algorithms that we choose
  - typically linear or sub-linear versions of the algorithms

…guide to Big-Data algorithms

- An excellent overview of the "Big Data" algorithms is the book "**Leskovec, Rajaraman, Ullman: Mining of Massive Datasets**"
  - Downloadable from: http://www.mmds.org/
  - Associated MOOC (from Oct 2014): https://www.coursera.org/course/mmds



"Big Data Research" Journal

- In August 2014 Elsevier started new "Big Data Research" journal
  - http://www.journals.elsevier.com/big-data-research/

- Articles from the first issue (Special Issue on Scalable Computing for Big Data):
  - FlexAnalytics: A Flexible Data Analytics Framework for Big Data Applications with I/O Performance Improvement
  - A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments
  - GDPS: An Efficient Approach for Skyline Queries over Distributed Uncertain Data
  - A Near-Linear Time Subspace Search Scheme for Unsupervised Selection of Correlated Features
  - Efficient Indexing and Query Processing of Model-View Sensor Data in the Cloud

Sampling on Big-Data

Sampling

- Deals with velocity and volume
- Enables off-line data analysis
- Enables performing expensive operations (eg, join of two streams via join of two samples)

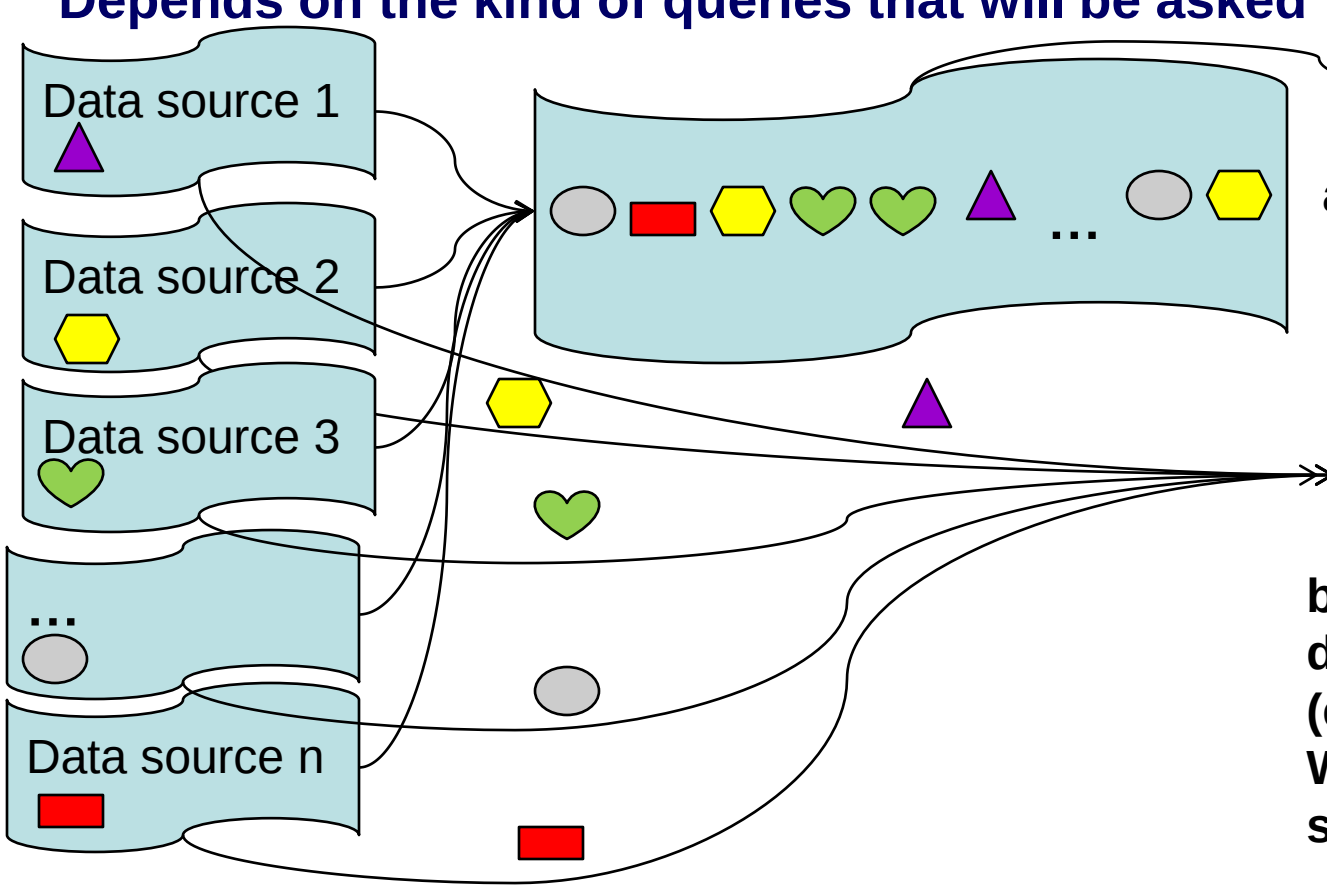
Reservoir sampling – maintaining a sample of fixed size by probabilistically replacing an old element by a new one

Sampling from different data sources – depending on the kind of queries to be asked decide whether to consider info. about the data source

Sampling *a/b* fraction of the data – hash data into *b* buckets and decide whether to store the data point based on the calculated value of the hash function

Sampling different data sources

Depends on the kind of queries that will be asked



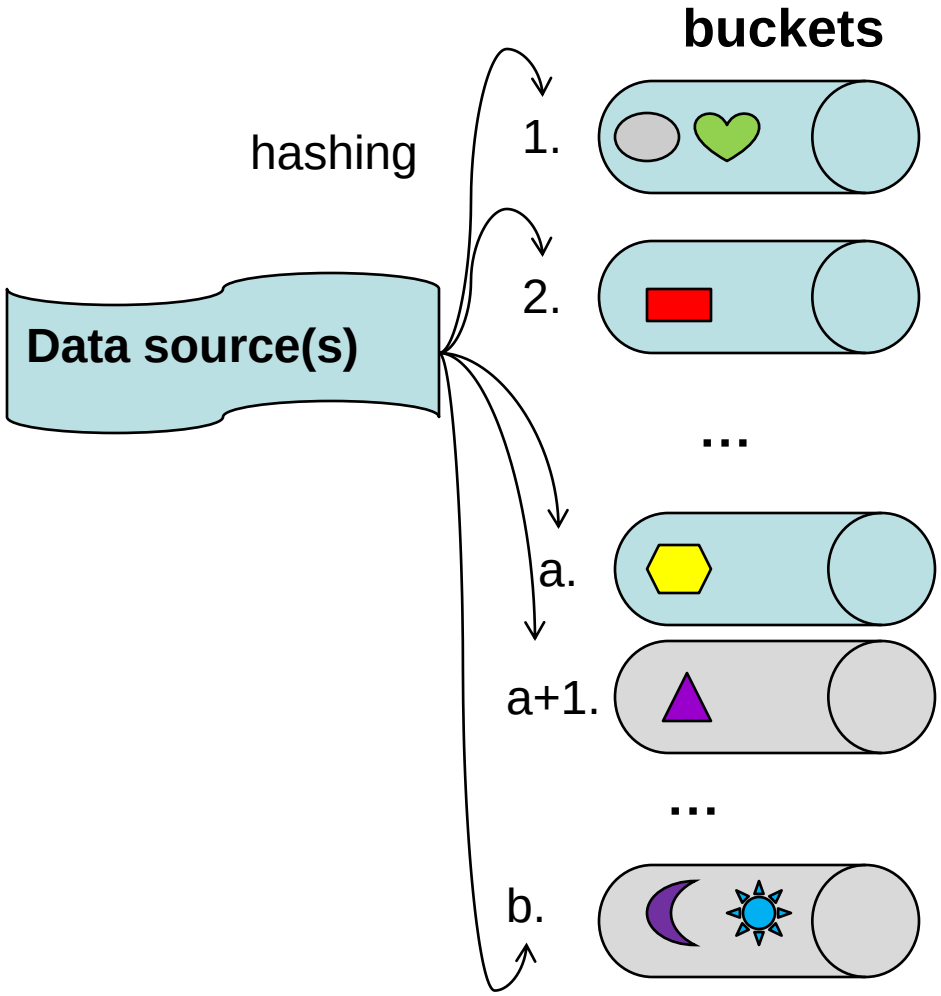
a) sampling x% of all the data points regardless the source

b) sampling x% of the data sources (eg., users accessing Web site or sensors sending measurements)

Example: average number of Web pages revisited by the same user?

- requires b) sampling - all the data for x% of the users

Sampling data using hashing



- use a hash function to hash the key components of the data stream (eg., username),
- based on the value of the function decide whether to store the current data or not

Example:

- hash the username to *b* buckets,
- if the user falls into one of the first *a* buckets store the data

Finding similar items

- Approach as a problem of finding sets with large intersections
  - Jaccard similarity: set_intersection/set_union
- Focus on similarity between the promising pairs of items
  - Eg., usernames with the same hash value, documents of the same length

Example problem

- Similarity of documents (plagiarism, mirror Web pages, news articles from the same source)
- Collaborative filtering for movie/book/… recommendation

Storing Big Data

- Data arriving in streams, rapidly so it is not feasible to store all the data
  - Eg., measurements of sensors at different locations – even if one stream is not of high speed, there is multitude of streams
- **What to store** depends on the queries that will be asked
  - **Standing query** (event pattern)
    - trigger an alarm, perform an operation on each arrival of a data point (eg., average the last 100 readings of sensor), report max. temperature so far
  - **Ad-hoc query**
    - Store **sliding window** of the last *n* data points
      - eg., the last 10 values of wind speed
    - Store **the last *t* time units** readings
      - eg., wind speed during the last hour,
      - eg., the number of unique users on the Web site in the past month – store the complete stream for the last month with the time stamp, remove the old data as new arrives

Storing Data on more Machines

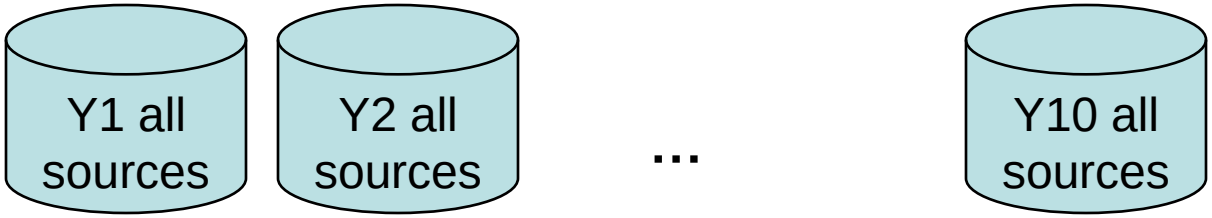
- Most big data have inherent temporal and/or spatial dimension
  - Data with time dimension should be stored and processed at least in a partial temporal ordering
  - Distributed storing of the data should consider the kind of queries that will be asked
    - if we want different type of queries i.e. over time and over location the data can be replicated to improve efficiency (and provide redundancy over potential hardware failure)
- A cluster of 10 machines is 10 times more likely to require a service than one machine

Storing data on more Machines

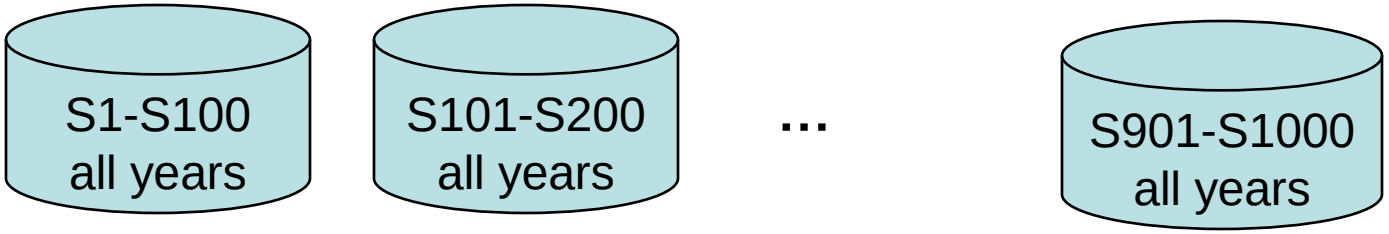
Example:10 years of observations collected at 15s intervals from 1000 sensor sites can be stored on 10 machines:

a) All observations for each year on one machine (eg., to return average value for the last year of all sensors)

b) All observations for 100 sensors on one machine (eg., to make analysis for one sensor over 10 years)



a) Time –centric



b) Source-centric

The Era of Big Data

- In science available massive streams of data
  - astronomy, high-energy physics, ecology, genetics and molecular biology
- In technology, personalization
  - data on fine-grained aspects of human behavior permitting the development of new services that are tailored to individuals

Big Data requires consideration of

- systems issues
  - how to store, index and transport data at massive scales; how to exploit parallel and distributed platforms,
- statistical issues
  - how to cope with errors and biases of all kinds; how to develop models and procedures that work big data,
- algorithmic issues
  - how to perform computations using resources that scale as linear or sub-linear functions
- legal, commercial and social issues

[M. Jordan, 2011]

Big Data for Business

Be smart when using Big Data, combine different aspects of the mind to achieve efficient utilization of:

- data and input (analytical mind)
- people and time (administrative mind)
- funds (financial mind)
- taking all into account in making executive decisions (executive mind)

[Sadhana Singh et al., 2015]

Data is a valuable asset in business, but before going for using (big) data (executive), check:

- What is the business problem or goal?
- Is the available data suitable? (analytical)
- What is the expected return on investment? (financial)
- Can we do it with the available resources timely?

Big Data for Business (cont.)

- Volume, Velocity, Variety of Big Data requires tradeoff on data freshness, query response time, data quality and answer quality
- Research challenges:
  - Support feature engineering and selection (eg., scoring individual features)
  - Learning from partially labeled data (eg., active learning)
  - Managing missing data across heterogeneous stream
  - Combining offline and online learning
  - Interactive and collaborative mining
  - Visualization of Big Data
  - Privacy and transparency
- Approach Big Data in scientific, practical and economic fashion

[Gopalkrishnan et al., BigMine-2012]

APPLICATION: RECOMMENDATION

Data

- User visit logs
  - Track each visit using embedded JavaScript
- Content
  - The content and metadata of visited pages
- Demographics
  - Metadata about (registered) users

Why news recommendation?

- "Increase in engagement"
  - Good recommendations can make a difference when keeping a user on a web site
  - Measured in number of articles read in a session
- "User experience"
  - Users return to the site
  - Harder to measure and attribute to recommendation module

- Predominant success metric is the attention span of a user expressed in terms of time spent on site and number of page views.

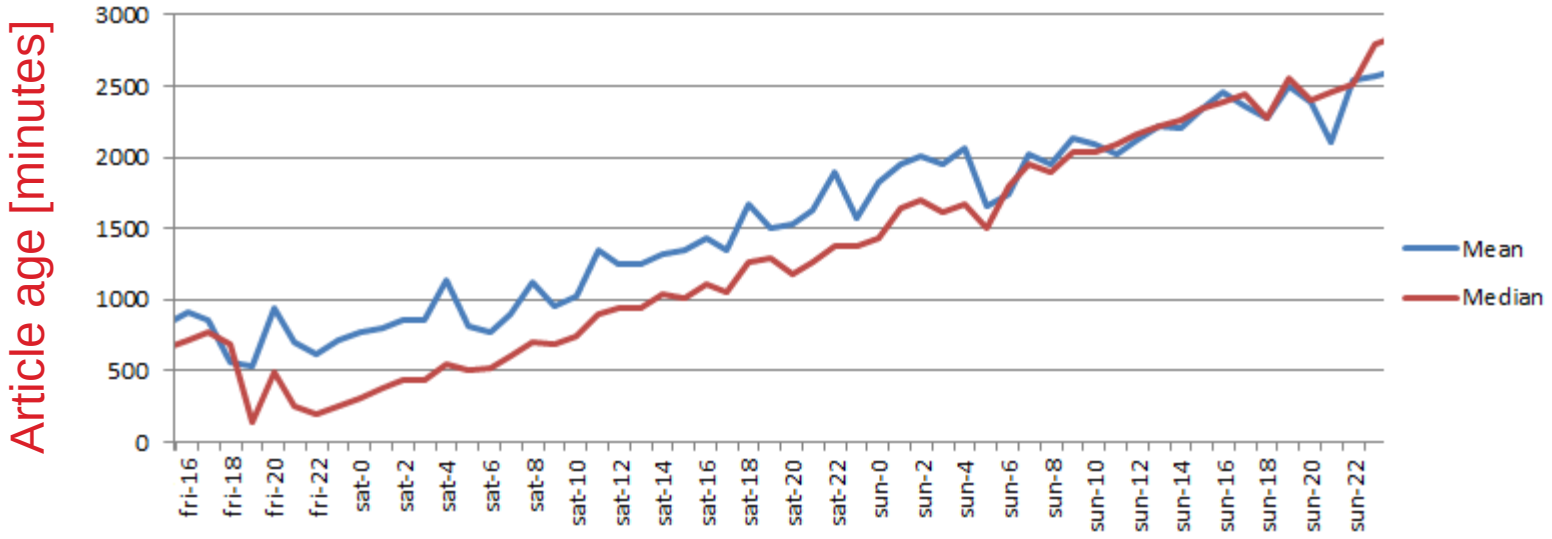
Why is it hard?

- Cold start
  - Recent news articles have little usage history
  - More sever for articles that did not hit homepage or section front, but are still relevant for particular user segment

- Recommendation model must be able to generalize well to new articles.

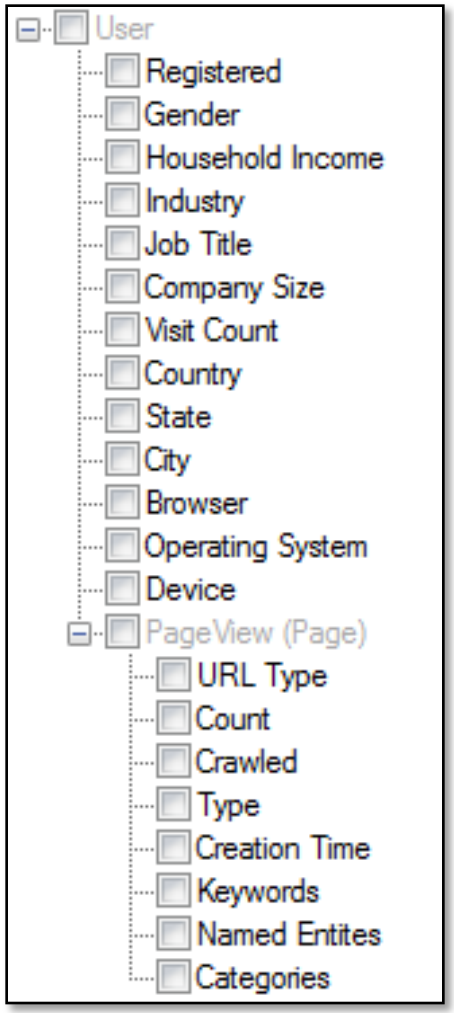
Example: Bloomberg.com

- Access logs analysis shows, that half of the articles read are less then ~8 hours old
- Weekends are exception



User Modeling

- Feature space
  - Extracted from subset of fields
  - Using vector space model
  - Vector elements for each field are normalized
- Training set
  - One visit = one vector
  - One user = a centroid of all his/her visits
  - Users from the segment form positive class
  - Sample of other users form negative class
- Classification algorithm
  - Support Vector Machine
  - Good for dealing with high dimensional data
  - Linear kernel
  - Stochastic gradient descent
    - Good for sampling



Experimental setting

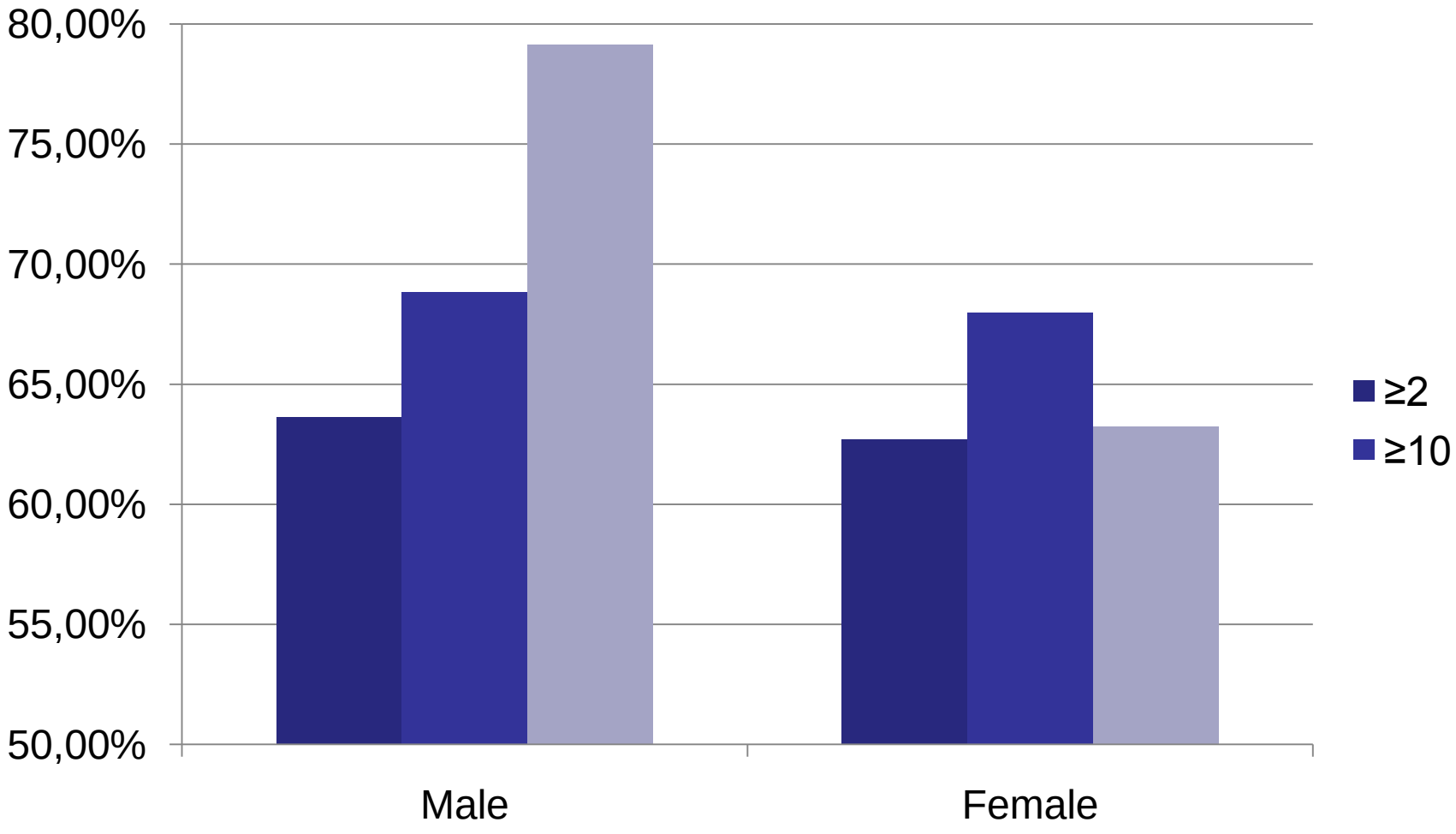
- Real-world dataset from a major news publishing website
  - 5 million daily users, 1 million registered
- Tested prediction of three demographic dimensions:
  - Gender, Age, Income
- Three user groups based on the number of visits:
  - ≥2, ≥10, ≥50
- Evaluation:
  - Break Even Point (BEP)
  - 10-fold cross validation

Category	Size
Male	250,000
Female	250,000

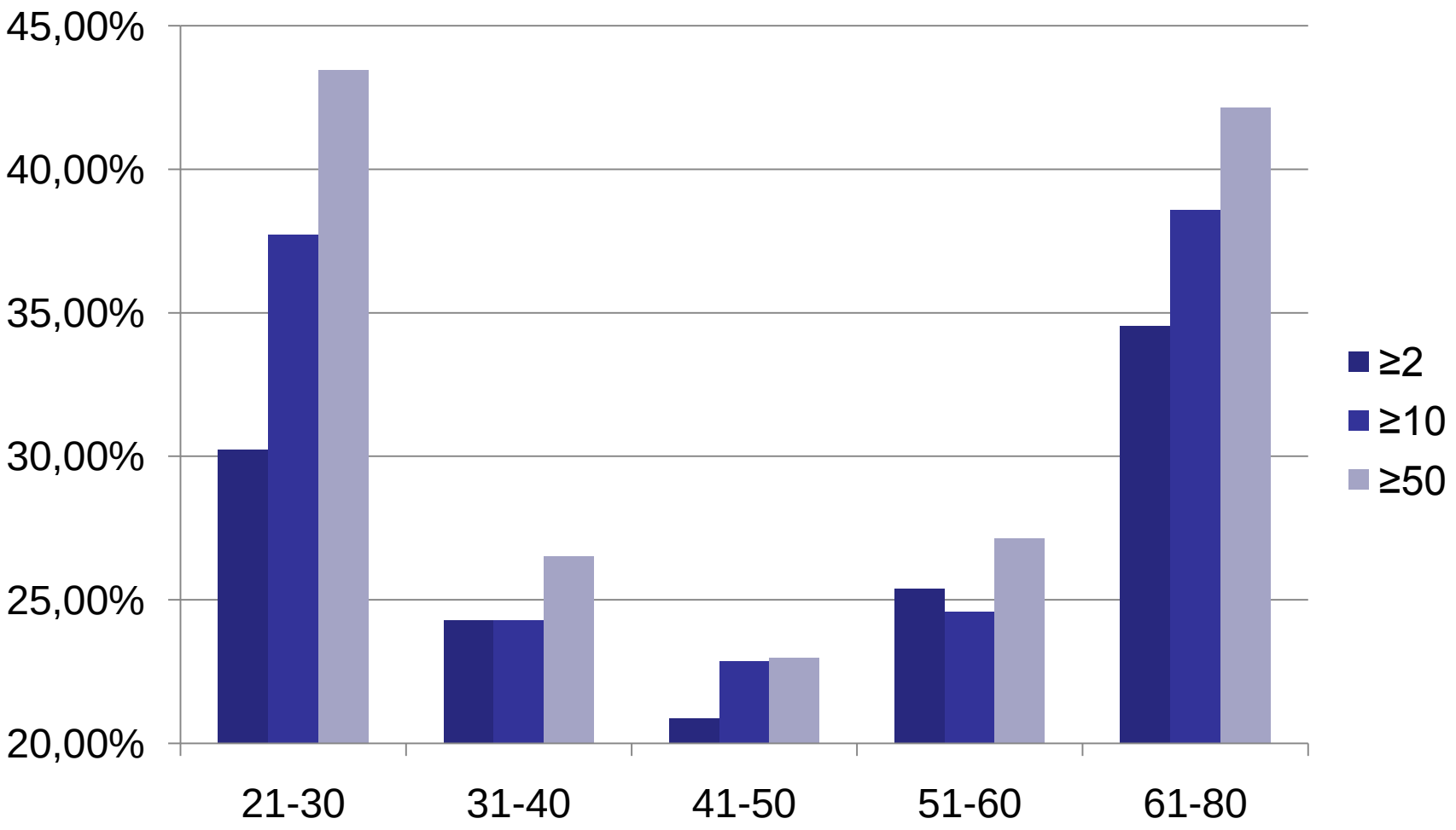
Category	Size
21-30	100,000
31-40	100,000
41-50	100,000
51-60	100,000
61-80	100,000

Category	Size
0-24k	50,000
25k-49k	50,000
50k-74k	50,000
75k-99k	50,000
100k-149k	50,000
150k-254k	50,000

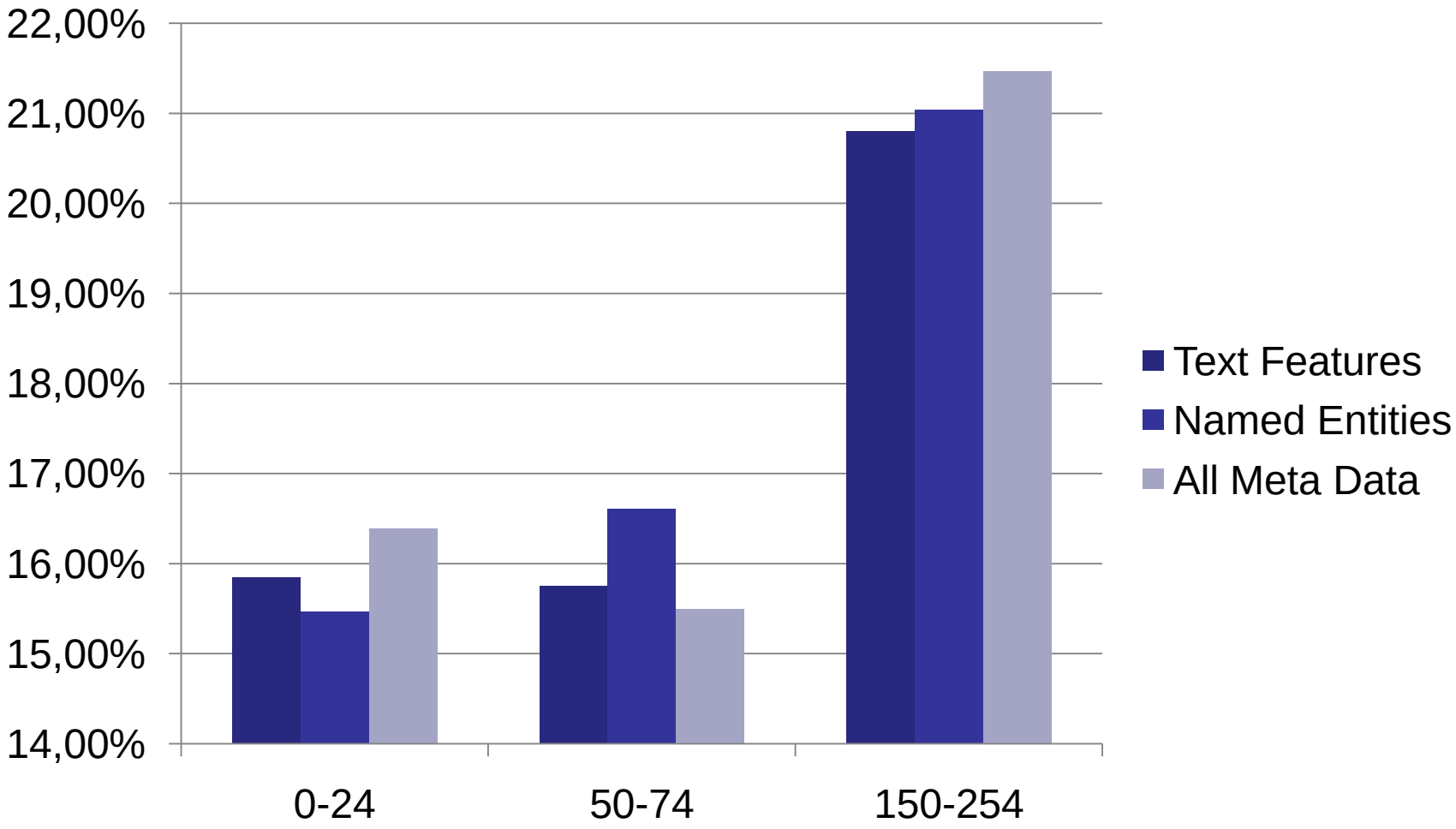
Gender



Age



Income (≥10 visits)



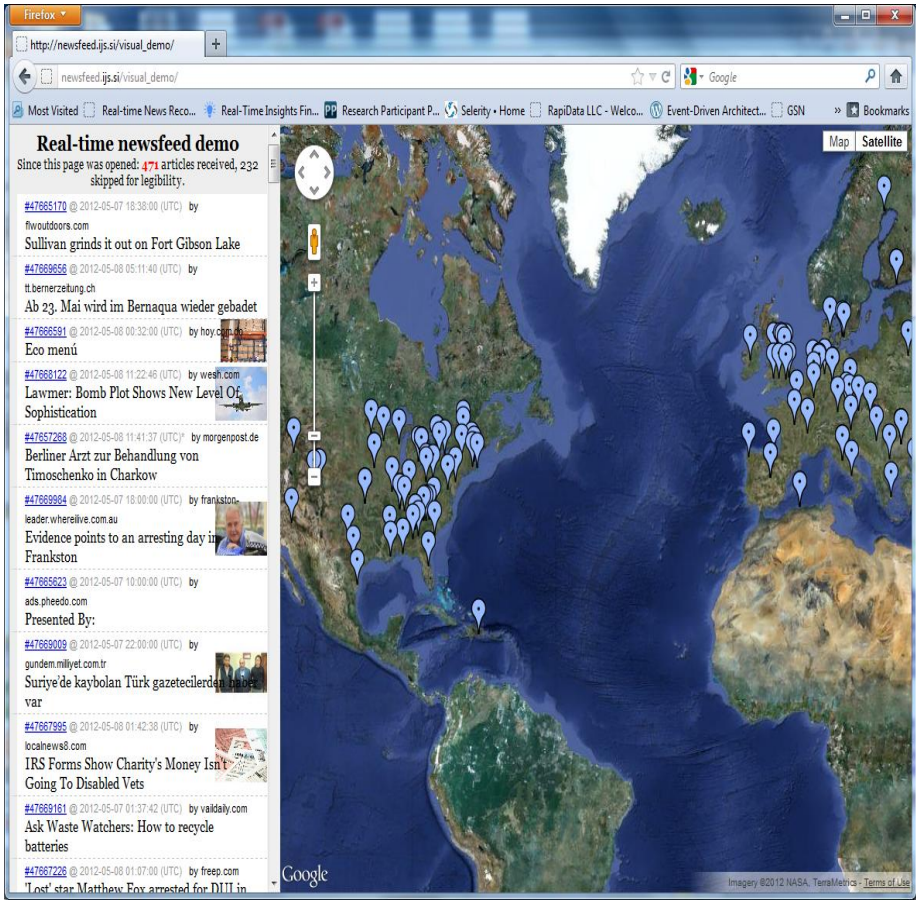
APPLICATION:
GLOBAL MEDIA MONITORING

Application: Monitoring global media

- The aim is to collect and analyze global main-stream and social media
  - documents are crawled from 100 thousands of sources
  - each crawled document gets cleaned, linguistically and semantically enriched
  - connect documents across languages (cross-lingual technology)
  - identify and connect events

Collecting global media in near-real-time (http://newsfeed.ijs.si)

- The NewsFeed.ijs.si system collects
  - 40.000 main-stream news sources
  - 250.000 blog sources
  - Twitter stream
- …resulting in ~500.000 documents + #N of twits per day
- Each document gets cleaned, linguistically and semantically annotated

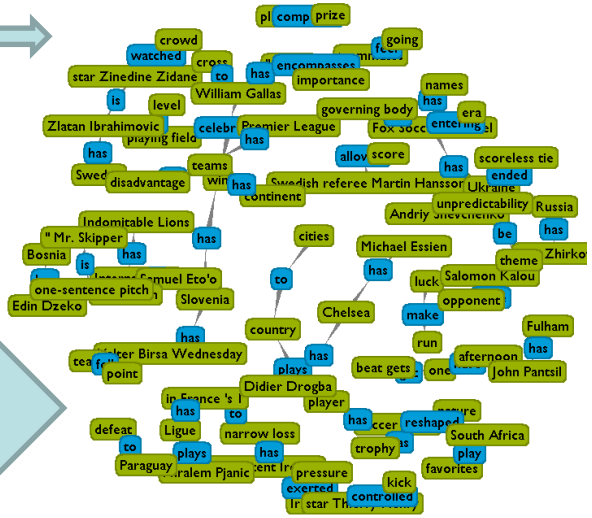


Semantic text enrichment (DBpedia, OpenCyc, …) with **Enrycher** (http://enrycher.ijs.si/)

Slovenia's dramatic w… Russia Wednesday, and to a lesser extent Ireland's narrow loss to France, capped off a grueling two-year qualifying period that saw some of the smallest countries in the world kick some of soccer's biggest names in the teeth. After a century of near-domination from the likes of Brazil, Italy and Germany, intern… entering the era of the Cinderella. It may not hap… but given the increasing flow of talent, training a… borders, it's almost certain that a small upstart nat… athletes and better luck will make a legitimate ru… coveted trophy.

Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's Valter Birsa Wednesday.

Text Enrichment



entities

- Brazil
- Italy
- Germany
- Cinderella
- Paris
- John O'Shea
- Manchester United
- Robbie Keane
- Shay Given
- Greece
- Portugal
- Bosnia-Herzegovina
- Cristiano Ronaldo
- Uruguay

keywords

Sports, Soccer, CONCACAF, Competitions, United States, Sports and Hobbies, Kids and Teens, World Cup, Women,

categories

- Top/Kids_and_Teens/Sports_and_Hobbies/Sports/Soccer
- Top/Sports/Soccer/Competitions
- Top/Sports/Soccer/Competitions/World_Cup
- Top/Sports/Soccer/CONCACAF

Diego Maradona Semantics:
owl:sameAs: http://dbpedia.org/resource/Diego_Maradona
owl:sameAs: http://sw.opencyc.org/concept/Mx4rvofERZwpEbGdrcN5Y29ycA
rdf:type: http://dbpedia.org/class/yago/ArgentinaInternationalFootballers
rdf:type: http://dbpedia.org/class/yago/ArgentineExpatriatesInItaly
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballManagers
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballers

Robbie Keane Semantics:
owl:sameAs: http://dbpedia.org/resource/Robbie_Keane
rdf:type: http://dbpedia.org/class/yago/CoventryCityF.C.Players
rdf:type: http://dbpedia.org/class/yago/ExpatriateFootballPlayersInItaly
rdf:type: http://dbpedia.org/class/yago/F.C.InternazionaleMilanoPlayers

"Enrycher" is available as as a web-service generating Semantic Graph, LOD links, Entities, Keywords, Categories, Text Summarization

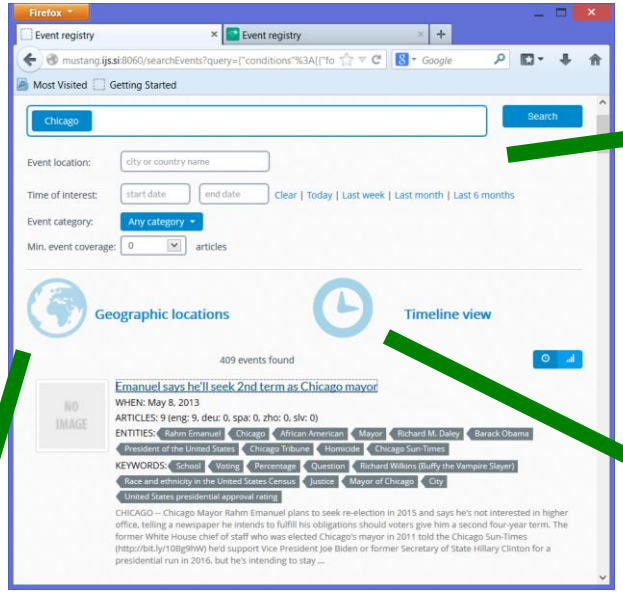
DiversiNews – exploring news diversity (http://aidemo.ijs.si/diversinews/)

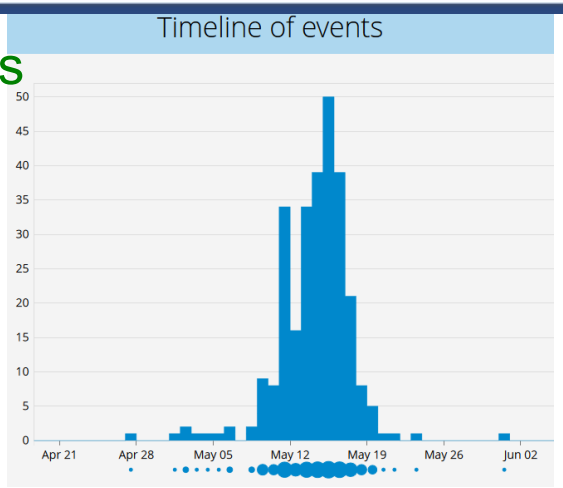
- Reporting has bias – same information is being reported in different ways
- DiversiNews system allows exploring news diversity along:
  - Topicality
  - Geography
  - Sentiment

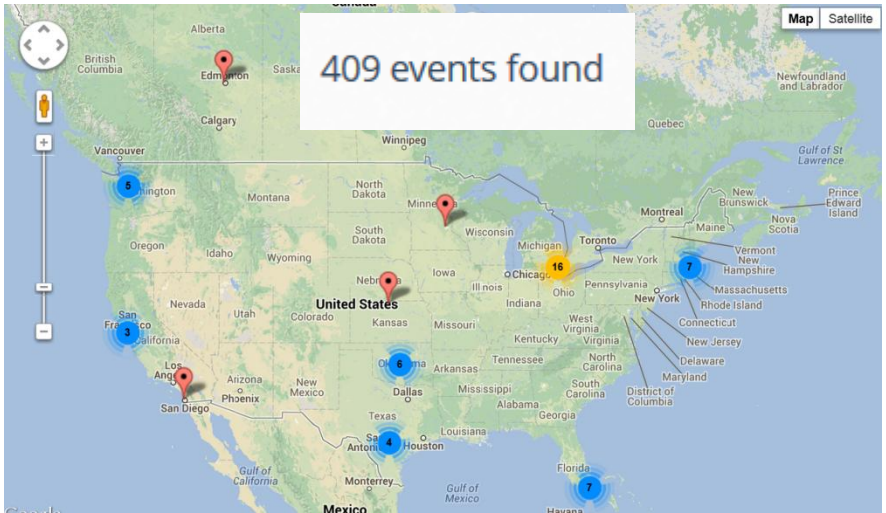
"Event Registry" system for global media monitoring (http://eventregistry.org)

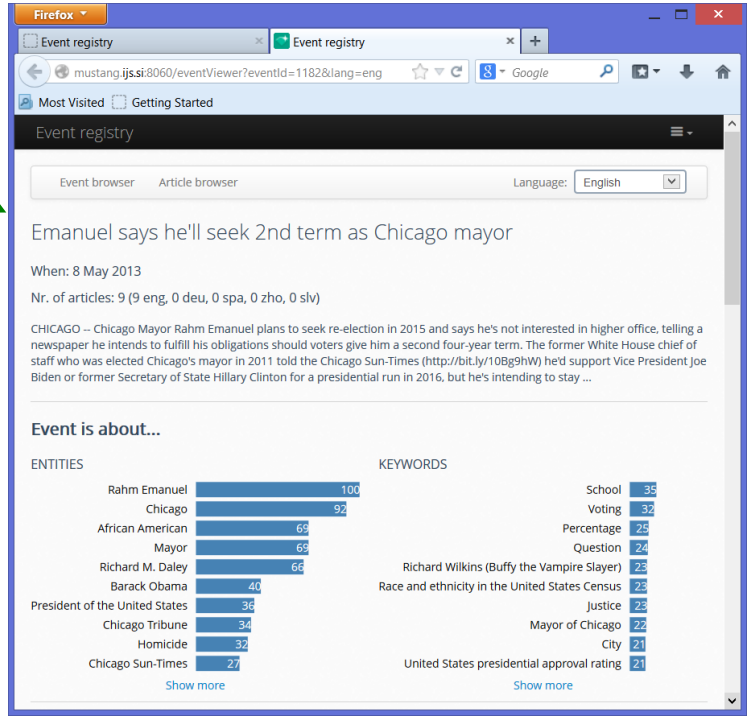
- Having a stream of news & social media, the task is to structure documents into events
- "Event Registry" system allows for:
  - Identification of events from documents
  - Connecting documents across many languages
  - Tracking events and constructing story-lines
  - Describing events in a (semi)structured way
  - UI for exploration through Search & Visualization
  - Export into JSON/RDF (Storyline ontology)

"Event Registry" example on "Chicago" related events



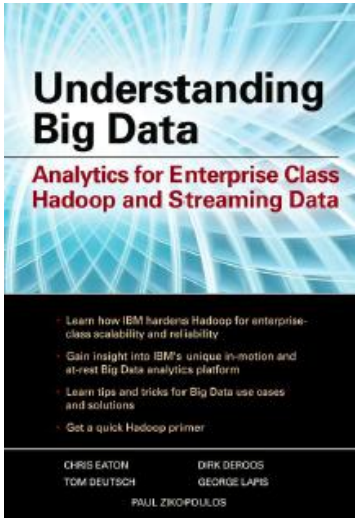


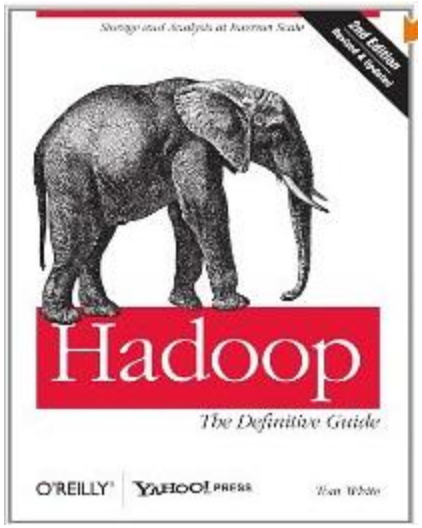


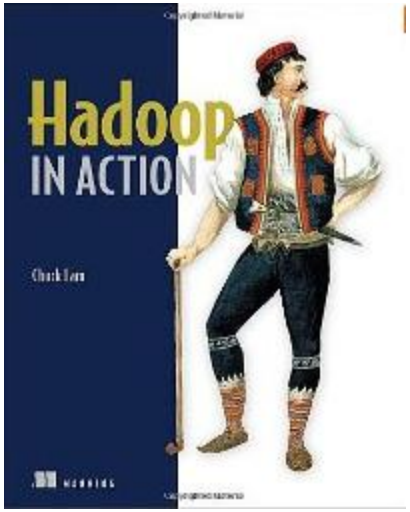


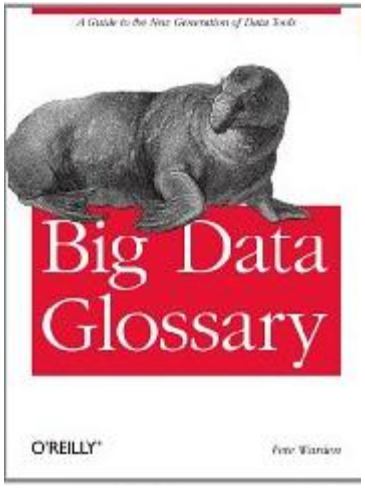
FINAL THOUGHTS

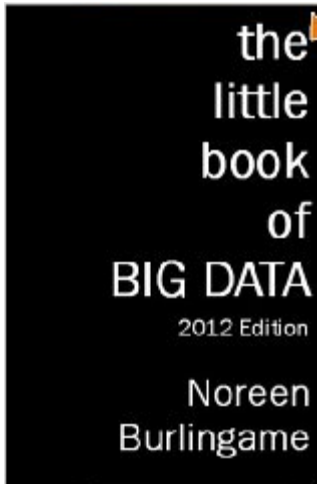
Literature on Big-Data

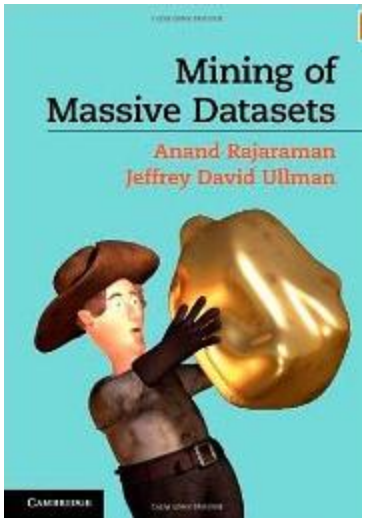


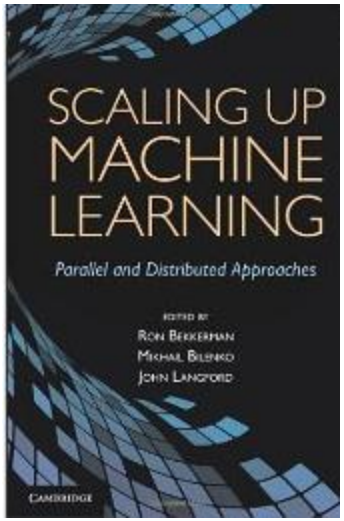












…to conclude

- Big-Data is everywhere, we are just not used to deal with it
- The "Big-Data" hype is very recent
  - …growth seems to be going up
  - …evident lack of experts to build Big-Data apps
- Can we do "Big-Data" without big investment?
  - …yes – many open source tools, computing machinery is cheap (to buy or to rent)
  - …the key is knowledge on how to deal with data
  - …data is either free (e.g. Wikipedia) or to buy (e.g. twitter)

"*This is the Information Age — everybody can be **informed about anything and everything**. There is no secret, therefore there is no sacredness. Life is going to become an open book. When your computer is more loyal, truthful, informed and excellent than you, you will be challenged. You do not have to compete with anybody. You have to compete with yourself.*" [Y. Bhajan]