TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Future Trends in Data Analytics

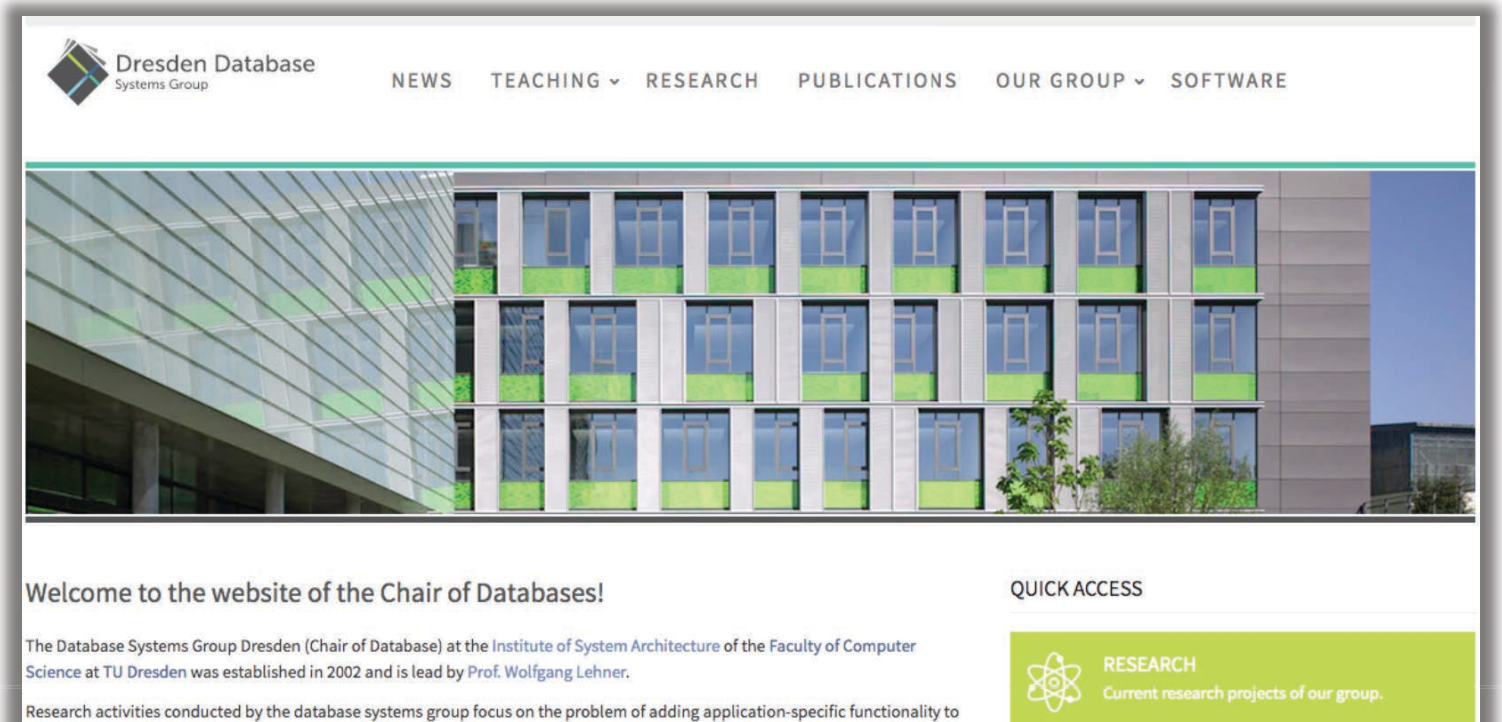
Hannes Voigt

General Information

Dresden Database Systems Group

Lecture Notes

- Further information on our website http://wwwdb.inf.tu-dresden.de

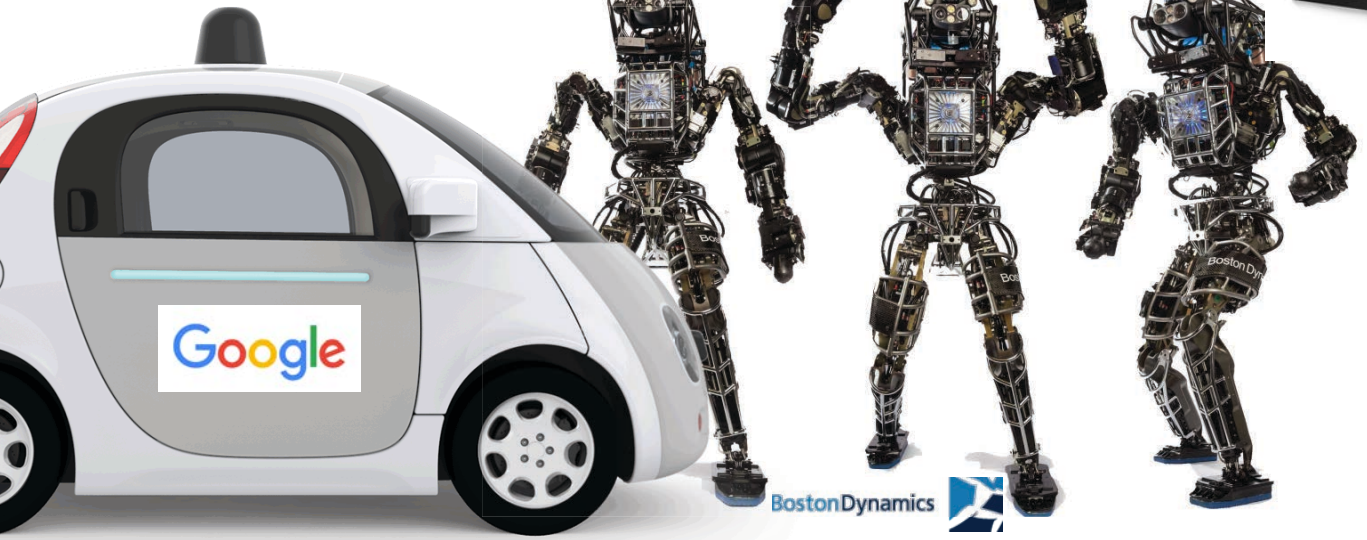


Dresden Database System Group

Dresden Database
Systems Group

Future is Now

- ..., face detection & recogition, autobeam, Siri/Cortant/..., 3D printing, ...



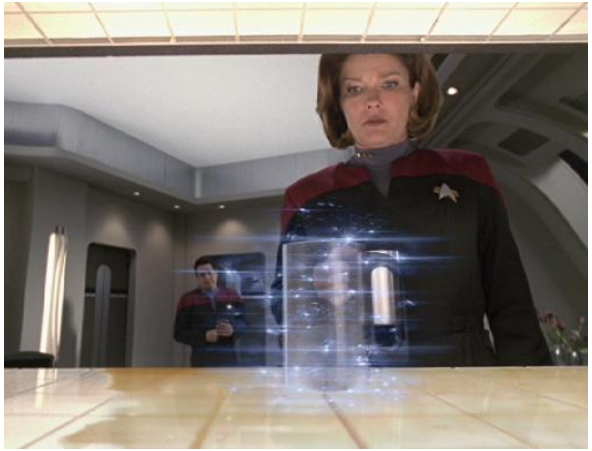
[https://www.google.com/selfdrivingcar/]

[http://www.bostondynamics.com/]

[http://www.ibmwatson.com/]

[http://www.idsc.ethz.ch/research-dandrea/research-projects/cubli.html]

... tomorrow?



Dresden Database Systems Group

Sooner then you think!

When have we entered the future?

What was the threshold we crossed?

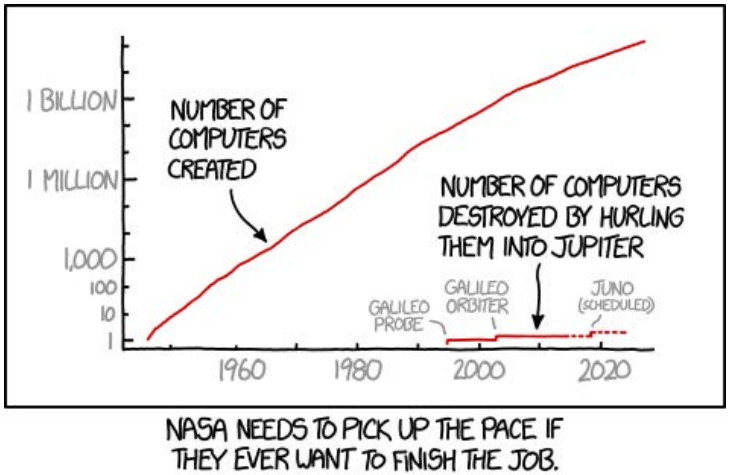
Exponential Growth

Exponential Growth

Dresden Database Systems Group

Outnumbers everything else quickly

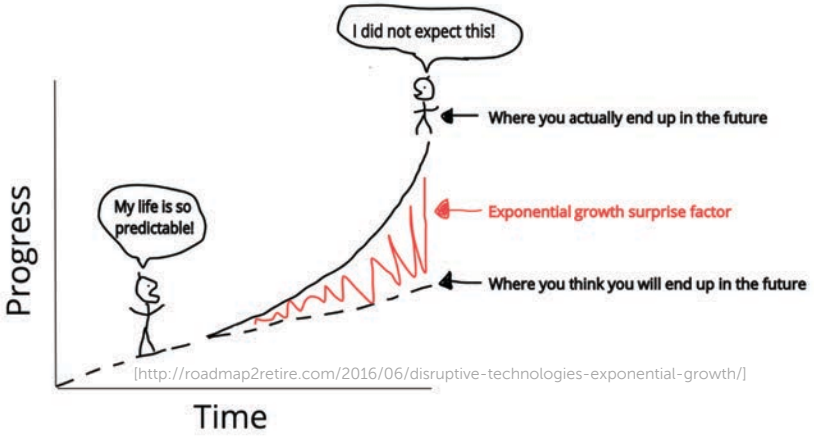
- Asymptotic advantage
- Quickly increasing add-on



[http://xkcd.com/1727/]

Leads to surprising results

- Black swans in economic crisis
- Shooting stars in business, media, sport, etc.



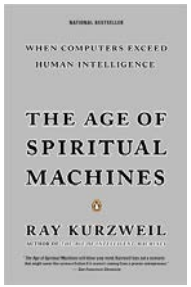
[http://roadmap2retire.com/2016/06/disruptive-technologies-exponential-growth/]

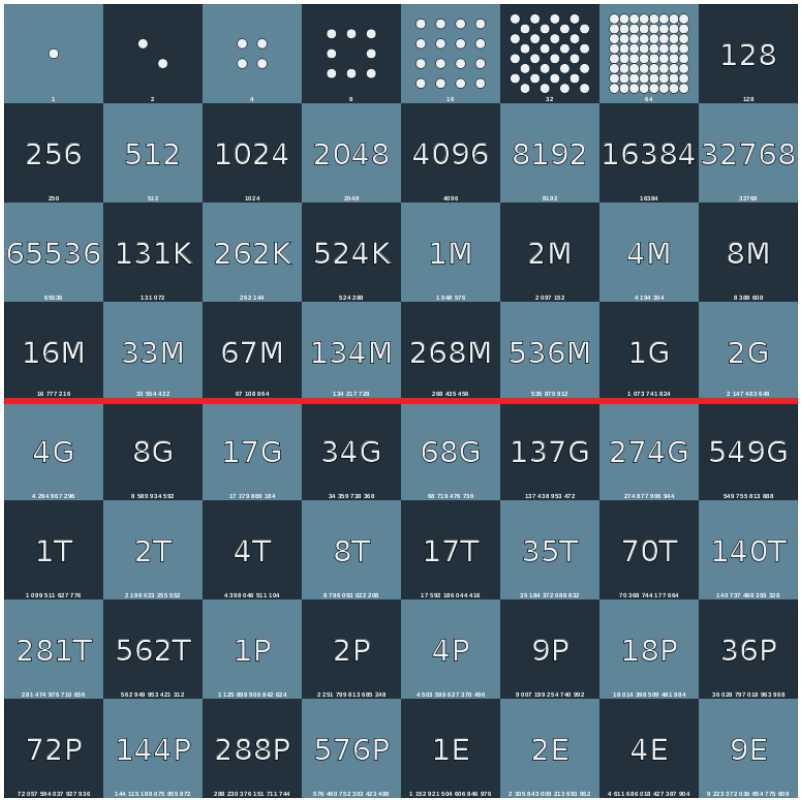
Second Half of the Chessboard

Dresden Database Systems Group

When exponential growth really kicks in

- According to Ray Kurzweil
- Things start to get interesting in the second half of the chess board
- Beyond 4G numbers quickly go beyond human intuition
- What happens in the second half can hardly foreseen





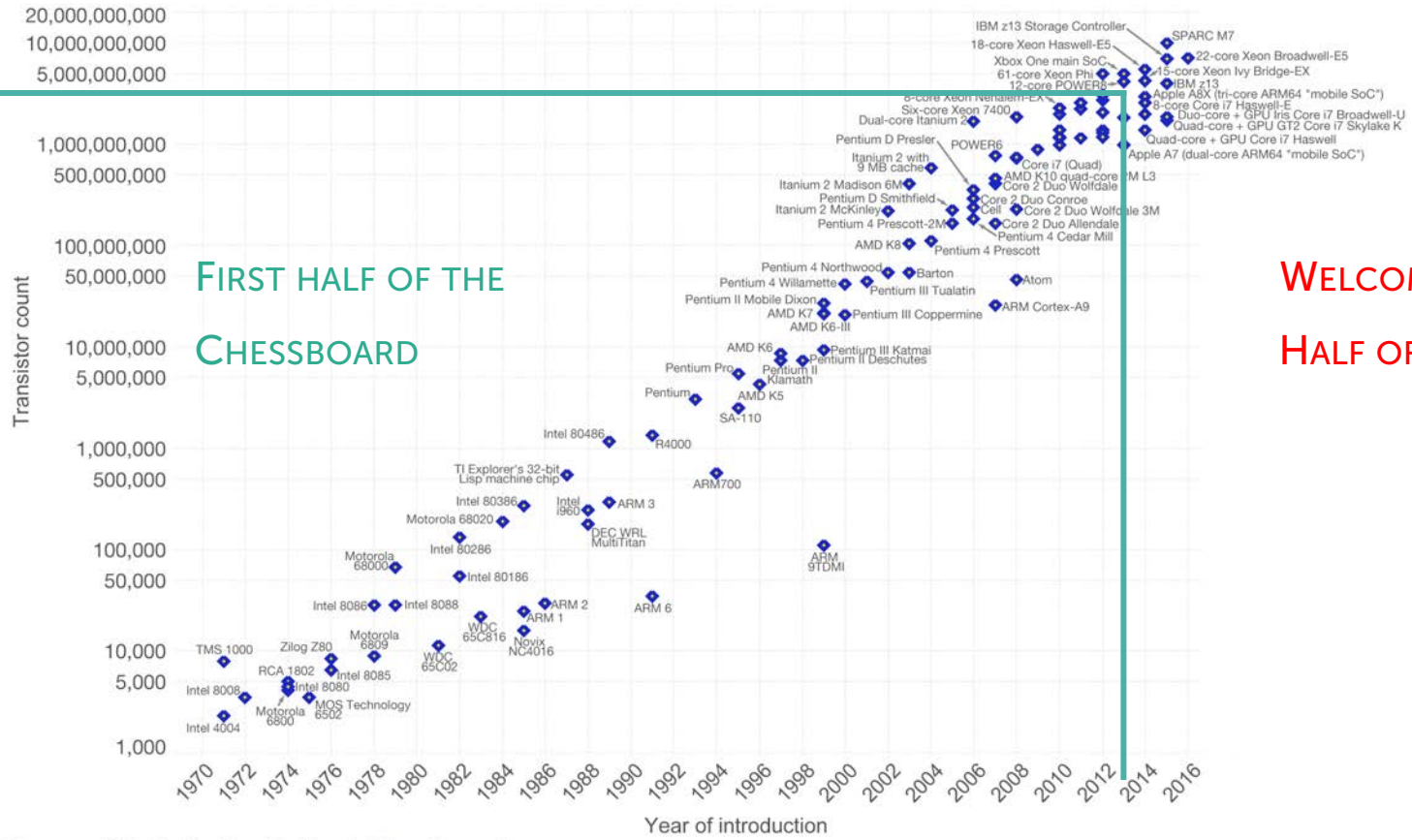
[https://en.wikipedia.org/wiki/File:Wheat_Chessboard_with_line.svg]

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.

Dresden Database Systems Group



First half of the Chessboard

Welcome to the Second Half of the Chessboard

Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Licensed under CC-BY-SA by the author Max Roser.

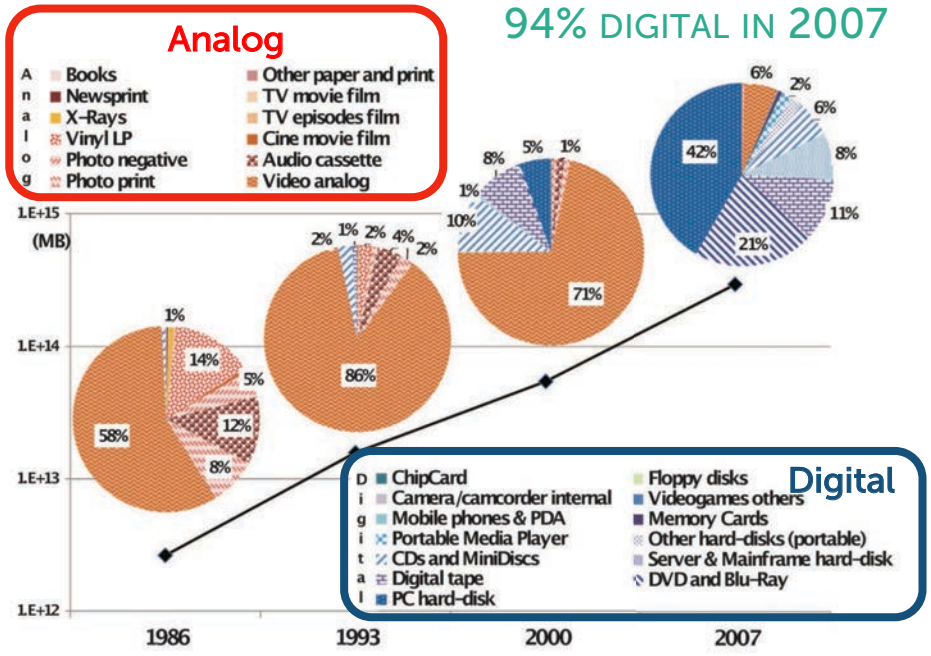
Digitization

Everything is Digital

Dresden Database Systems Group

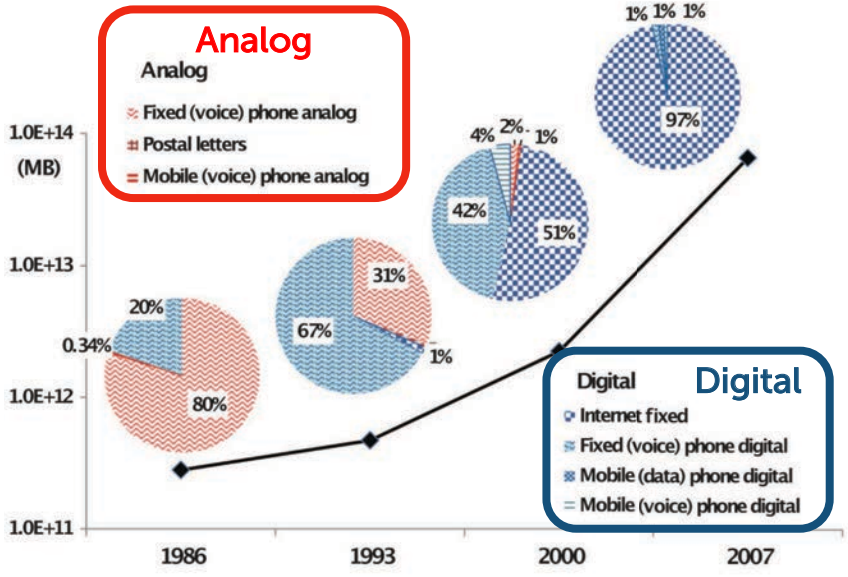
World's capacity to store information

94% digital in 2007



World's capacity to telecommunicate

99% digital in 2007



[M. Hilbert and P. Lopez, The World's Technological Capacity to Store, Communicate, and Compute Information, Science, 332, April 2011, DOI: 10.1126/science.1200970]

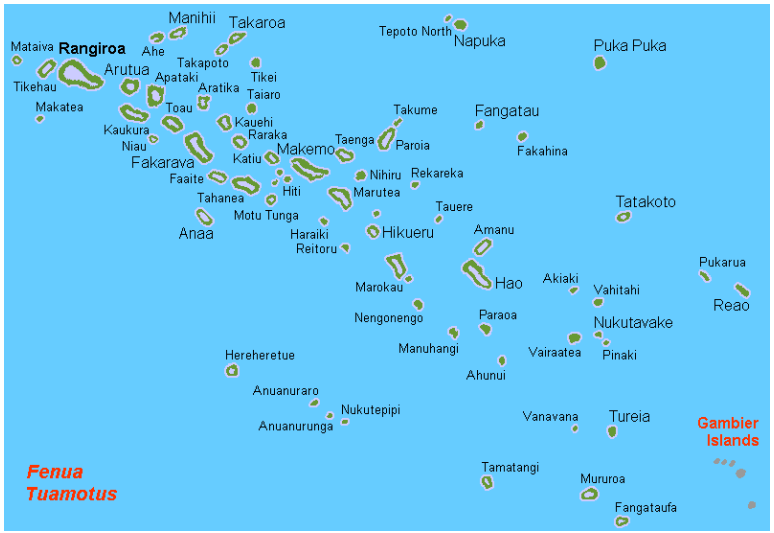
Everything is Digital

Dresden Database Systems Group

Landscape has changed

Dresden Database Systems Group

From Islands of digital data ...



(Tuamotu Archipelago, French Polynesia)

... to ponds of analog signals

(Algonquin Provincial Park, Ontario, Canada)

Everything is Digital

[http://blog.acronis.com/posts/data-everything-8-noble-truths]

Recombinant Innovation

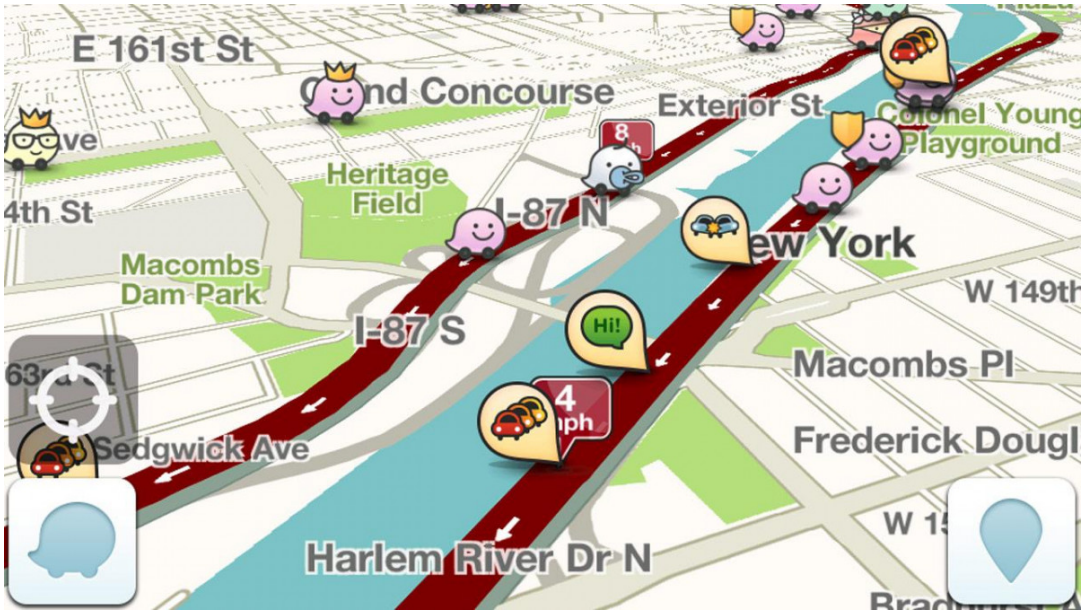
Example: waze

OUTSMARTING TRAFFIC, TOGETHER

Dresden Database Systems Group

CAR ROUTING APP THAT TURNS EVERY USER INTO A SENSOR

- Provides real-time traffic information
- Learns from users about traffic flow
- Crowd-sourcing of traffic information
- User can post alerts about
  - speed traps,
  - accidents,
  - traffic jams,
  - and even gas prices



[http://www.digitaltrends.com/cars/wazes-connected-citizens-program-turn-sunday-driver-ally/]

Example: waze

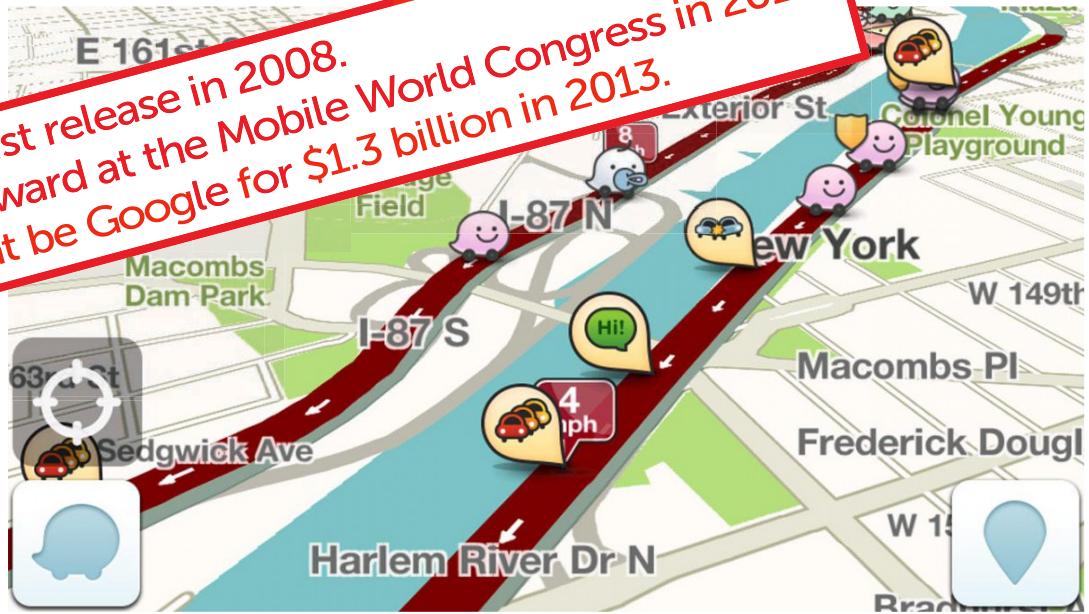
OUTSMARTING TRAFFIC, TOGETHER

Dresden Database Systems Group

CAR ROUTING APP THAT TURNS EVERY USER INTO A SENSOR

- Provides real-time traffic information
- Learns from users about traffic flow
- Crowd-sourcing of traffic information
- User can post alerts about
  - speed traps,
  - accidents,
  - traffic jams
  - and ev

Frist release in 2008.
Best Overall Mobile App award at the Mobile World Congress in 2013.
Company bought be Google for $1.3 billion in 2013.



[http://www.digitaltrends.com/cars/wazes-connected-citizens-program-turn-sunday-driver-ally/]

Waze Ingredients

ALL INGREDIENTS HAVE EXISTED LONG BEFORE

- Digital road maps
- Car navigation and routing
- GPS/GPS car navigation
- Mobile broadband
- PDAs/Smart Phones

TomTom Navigator, 2002

Dresden Database Systems Group

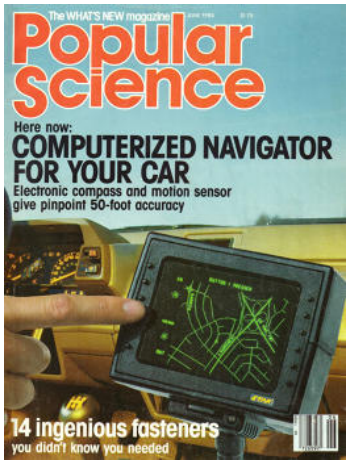
Tele Atlas

founded 1985

NAVTEQ

founded 1985
as Karlin & Collins, Inc.

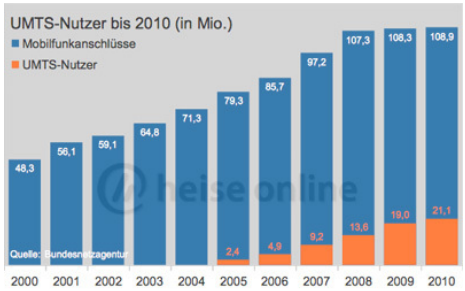
PSN-8 Manpack GPS Receiver
used in Operation Desert Storm
in 1991



Etak Navigator, 1985

Newton, 1993

iPhone 1, 2007



UMTS auction in 2000
UMTS available in 2004

Waze Ingredients

ALL INGREDIENTS HAVE EXISTED LONG BEFORE

- Digital road maps
- Car navigation and routing
- GPS/GPS car navigation
- Mobile broadband
- PDAs/Smart Phones

TomTom Navigator, 2002

Dresden Database Systems Group

Tele Atlas

NAVTEQ

founded 1985
as Karlin & Collins, Inc.

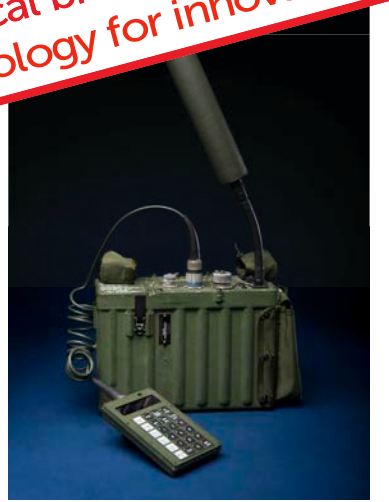
PSN-8 M...

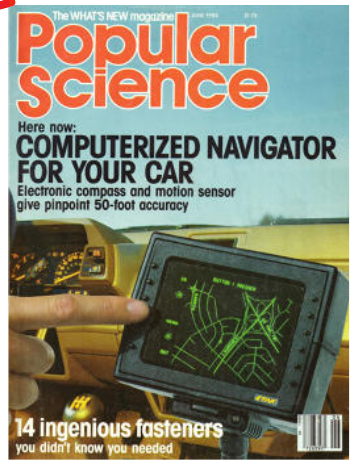
Although major break through in traffic organization,
Waze is no major technological break through.
Waze combines existing technology for innovation.

Newton, 1993

iPhone 1, 2007

UMTS auction in 2000
UMTS available in 2004





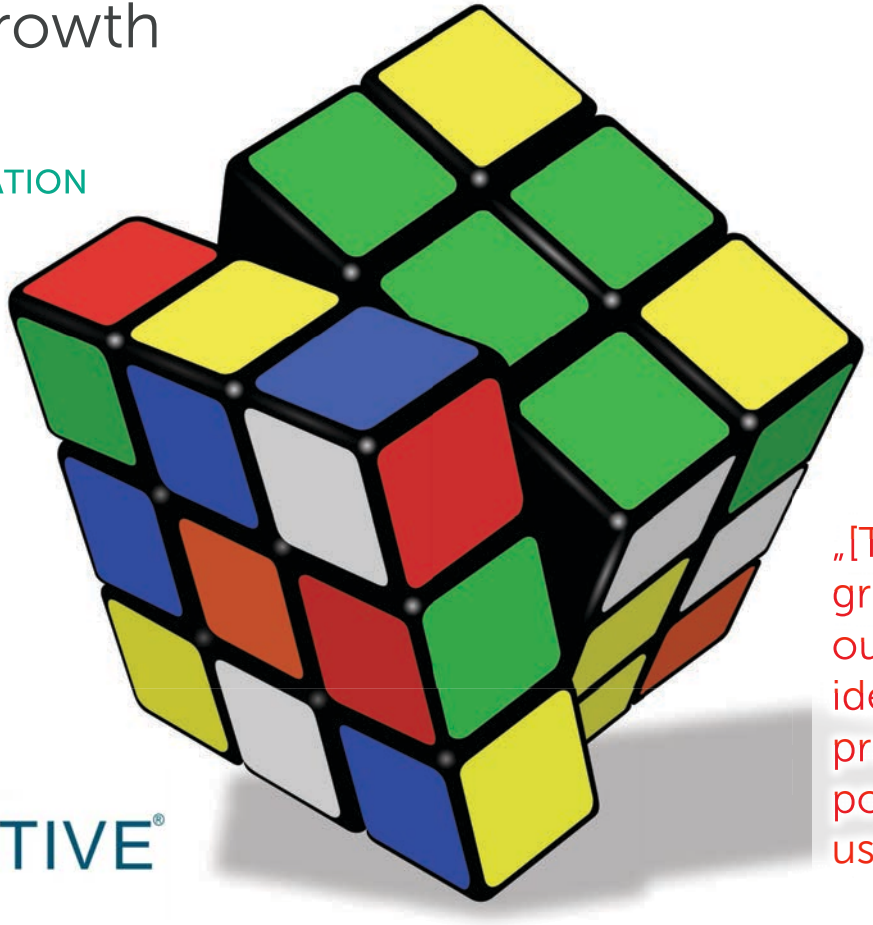
Etak Navigator, 1985

Recombinant Growth

Dresden Database Systems Group

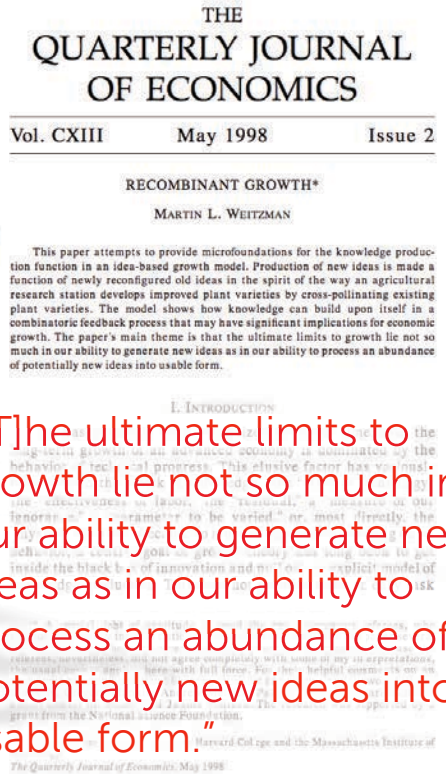
Innovation is recombination of existing ideas

- Possibilities explode quickly
- But we got way better in processing ideas



kaggle

INNOCENTIVE®



THE QUARTERLY JOURNAL OF ECONOMICS

Vol. CXIII May 1998 Issue 2

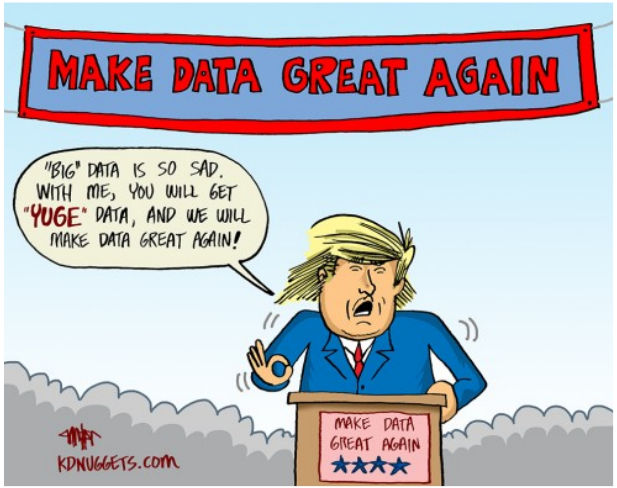
RECOMBINANT GROWTH*

MARTIN L. WEITZMAN

This paper attempts to provide microfoundations for the knowledge production function in an idea-based growth model. Production of new ideas is made a function of newly reconfigured old ideas in the spirit of the way an agricultural research station develops improved plant varieties by cross-pollinating existing plant varieties. The model shows how knowledge can build upon itself in a combinatoric feedback process that may have significant implications for economic growth. The paper's main theme is that the ultimate limits to growth lie not so much in our ability to generate new ideas as in our ability to process an abundance of potentially new ideas into usable form.

„[T]he ultimate limits to growth lie not so much in our ability to generate new ideas as in our ability to process an abundance of potentially new ideas into usable form."

Big Data

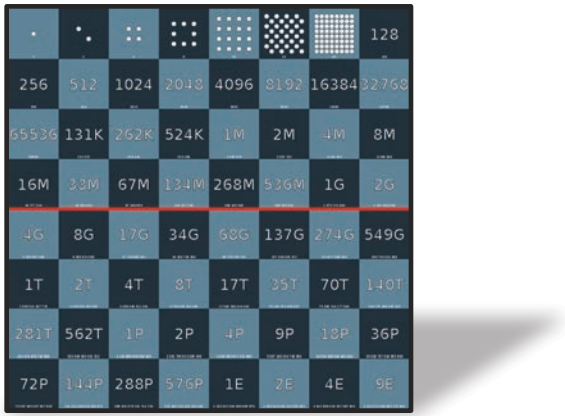

MAKE DATA GREAT AGAIN
"BIG" DATA IS SO SAD. WITH ME, YOU WILL GET "YUGE" DATA, AND WE WILL MAKE DATA GREAT AGAIN!
MAKE DATA GREAT AGAIN
KDNUGGETS.com

Big Data term reflects the three drivers

Dresden Database Systems Group

EXPONENTIAL GROWTH

- beyond intuition
- second half of the chessboard



Big

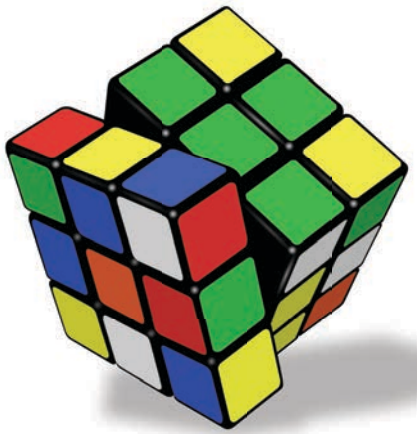
DIGITIZATION

- Everything is digital data
- Analog signals are not part of big data

Data

RECOMBINANT INNOVATION

- explosion of ideas
- all somehow seem to be part of big data



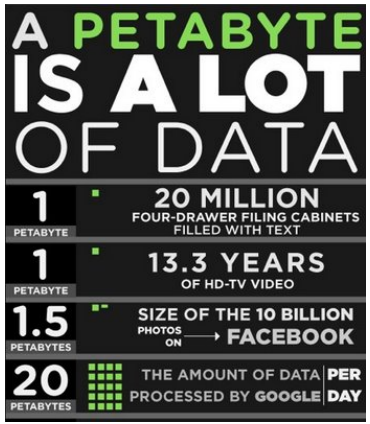
Fuzziness of the term

Data, data, everywhere… ▸ Volume

Dresden Database Systems Group

The Petabyte Age

- 2008



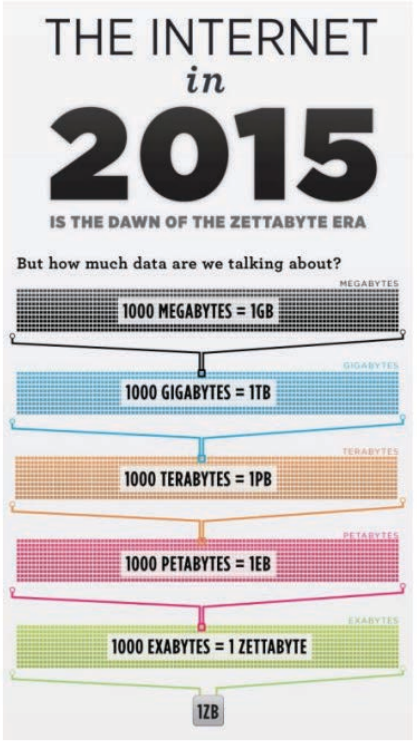
- Eric Schmidt (in 2010): Every 2 Days We Create As Much Information As We Did Up To 2003

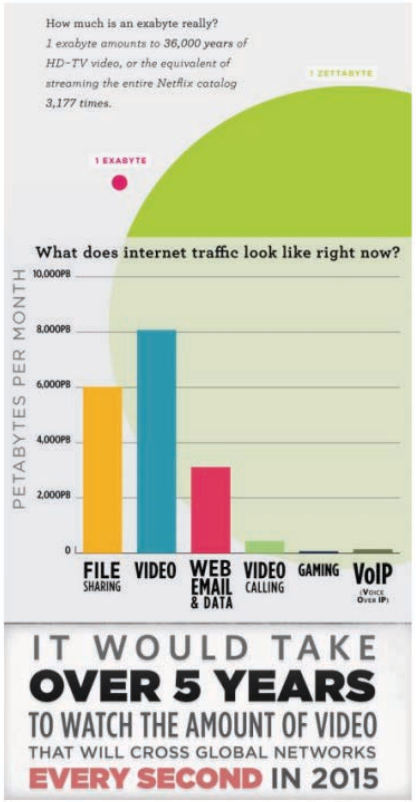
The Zettabyte Age

- 2015
- One Zettabyte = Stack of books from Earth to Pluto 20 times

The internet in 2020

- ~26.3 billion networked devices
- 25.1 GB average traffic per capita per month
- 2.3 Zettabytes annual IP-Traffic





Data is produced continuously ▸ Velocity

Dresden Database Systems Group

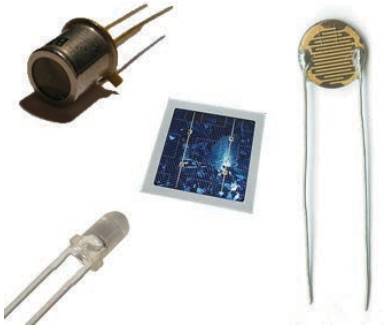
Humane-produced data

- ~300 million email sent/received per minute in 2016
- >0.4 million tweets per minute in 2016
- ~138 million google searches per minute in 2016
- >400 hours of video was uploaded to YouTube per minute in 2016

Machine produced data

- IoT will be boost data velocity greatly
- Sensors become ubiquitous
- Sensors for sound, images, position, motion, temperature, pressure, etc …
- Resolution (in time and space) is continuously increasing

- Assume Waze like cars collecting 20 double values every second
- With one million driving cars that is almost 10 TB every minute



Everything is data ▸ Variety

Dresden Database Systems Group

Data from all kinds of digital sources

- Structure vs. unstructured
- Text vs. image
- Curated vs. automatically collected
- Raw vs. edited vs. refined

Semantic heterogenity

- Decentralized content generation
- Multiple perspectives (conceptualizations) of the reality
- Ambiguity, vagueness, inconsistency

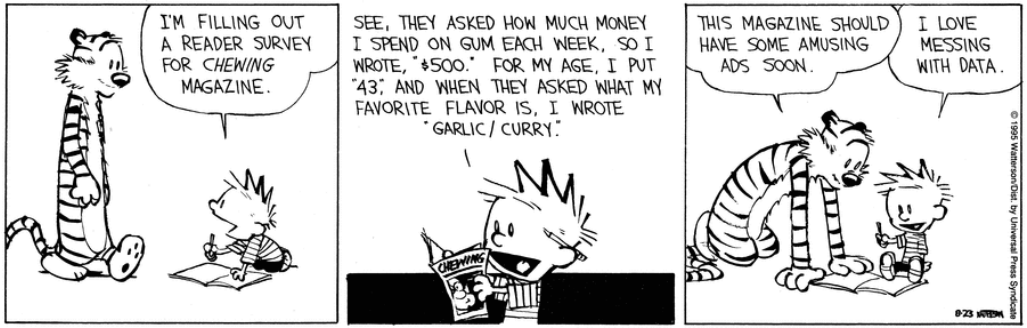
"A lot of Big Data is a lot of small data put together."

Data is messy ▸ Veracity

Dresden Database Systems Group

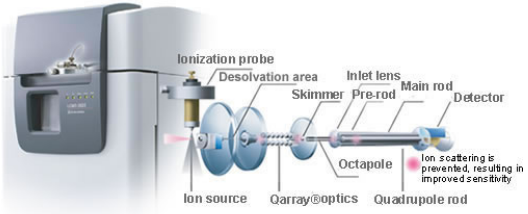
QUALITY OF CAPTURED DATA VARIES GREATLY

- Sensor inaccuracy
- Human mistakes
- Incompleteness
- Untrusted sources
- Deterministic processes
- etc.



"Next to the analytes, we see everything in the results, from the perfume of the lab assistant to the softener in the new machine sitting next."

—About Liquid chromatography–mass spectrometry at IPB Halle



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

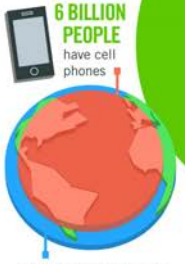
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones



Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

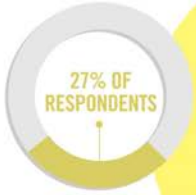
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

[http://www.ibmbigdatahub.com/infographic/four-vs-big-data]

IBM

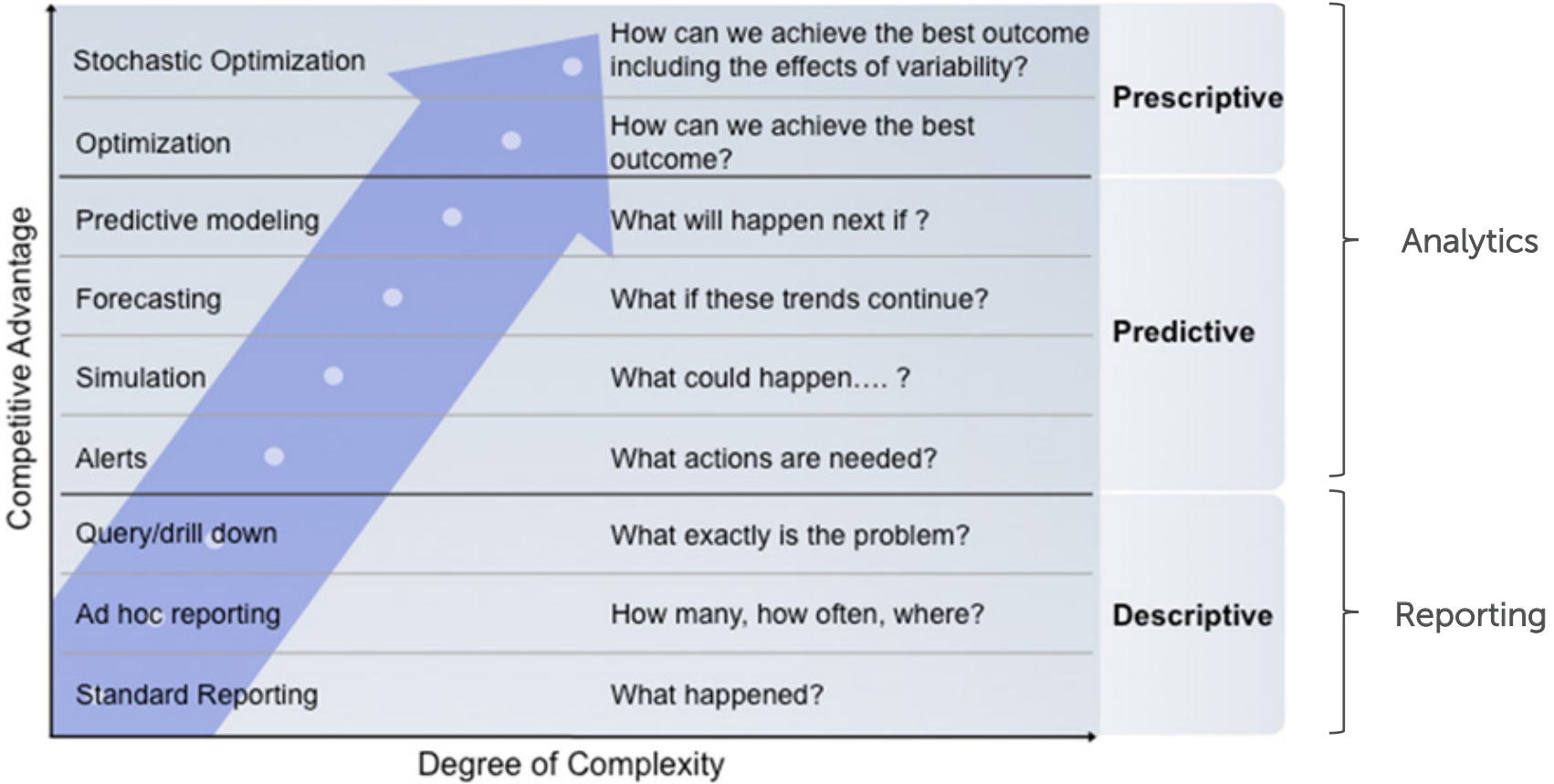
Value – the fifth V of Big Data

Data Science/Data Analysis

… or how to turn raw data into something valuable?

"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; **so must data be broken down, analyzed for it to have value.**" –Clive Humby

Levels of Analysis



Based on: Competing on Analytics, Davenport and Harris, 2007

Descriptive Analytics

Dresden Database Systems Group

How have I done? (and why?)

- Simplest class of analytics
- Condense big data into smaller, more useful bits of information
- Summary of what happened
- 70-80% penetration

Examples

- Database aggregation queries
- Business reporting (e.g. Sales figures)
- Market survey (e.g. GFK)
- (classical) business intelligence, dashboards, scorecards
- Google Analytics

Predictive Analytics

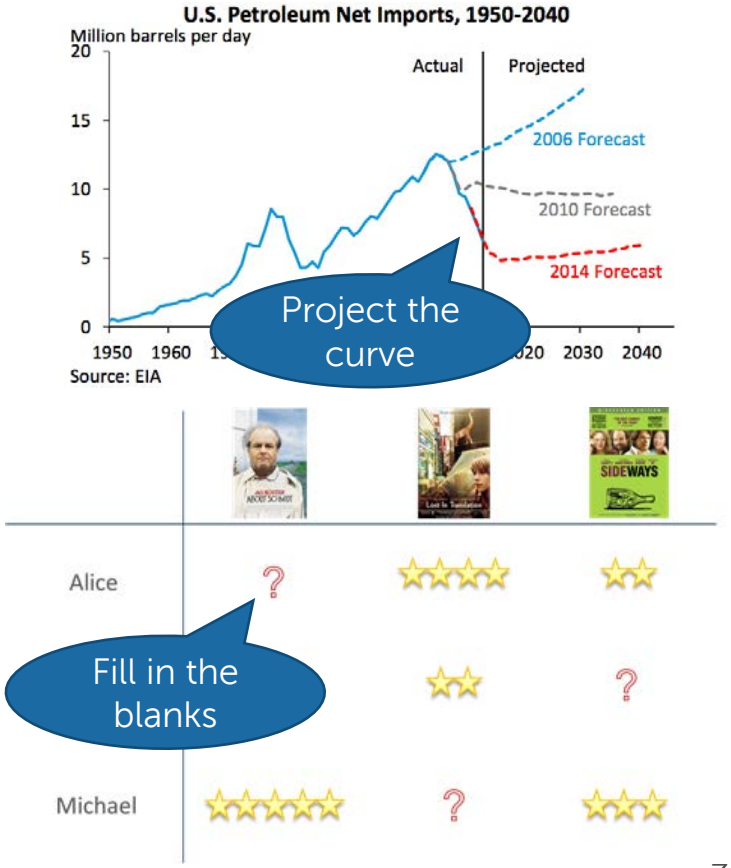
Dresden Database Systems Group

How will I do?

- Next step up in data reduction
- Studies recent and historical data
- Utilizes a variety of statistical, modeling, data mining and machine learning techniques
- Allows (potential inaccurate) predictions about the future
- Use data you have, to create data you do not have
- 15-25% penetration

Examples

- Market developments
- Stock developments
- Movie/product recommendations on netflix/amazon
- Energy demand and supply forecasting
- Preditive policing (e.g. precobs, predpol)



Prescriptive Analytics

ORION –
On-Road Integrated Optimization and Navigation

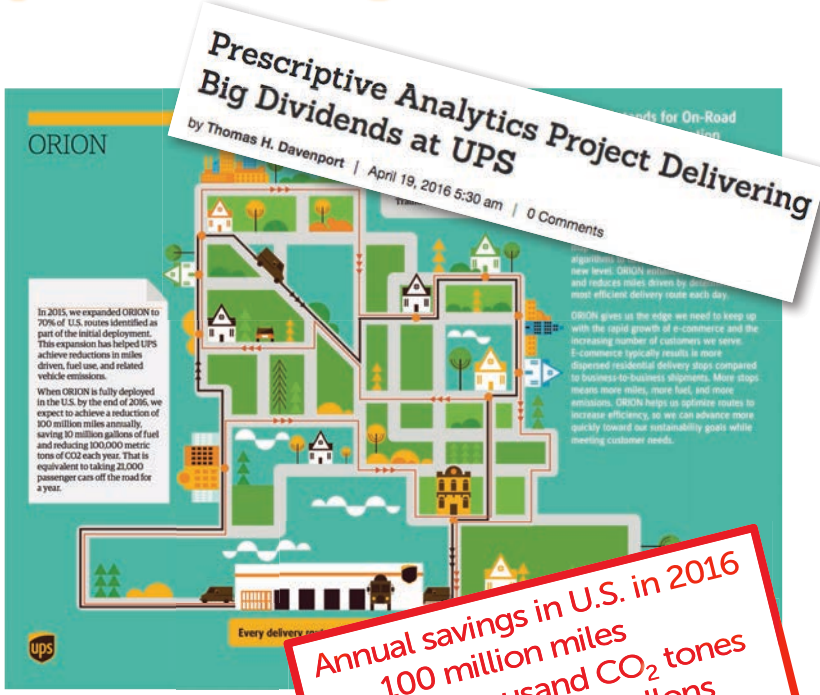
Dresden Database Systems Group

What should I do?

- Predicts "multiple futures" based on the potential actions
- Recommends the best course of action for any pre-specified outcome
- Typically involves a feedback system to track outcome produced by the action taken
- Utilizes predictive methods + optimization techniques
- 1-5% penetration

Examples

- Energy load balancing by flexoffer scheduling
- Inventory optimization in supply chains
- Targeted marketing campaign optimization
- Focus treatment of clinical obesity in health care
- Waze-like car navigation



Annual savings in U.S. in 2016
- 100 million miles
- 100 thousand CO_2 tones
- 10 million fuel gallons
- $300-400 million costs

Same same but different

From Business Intelligence to Data Science

Dresden Database Systems Group

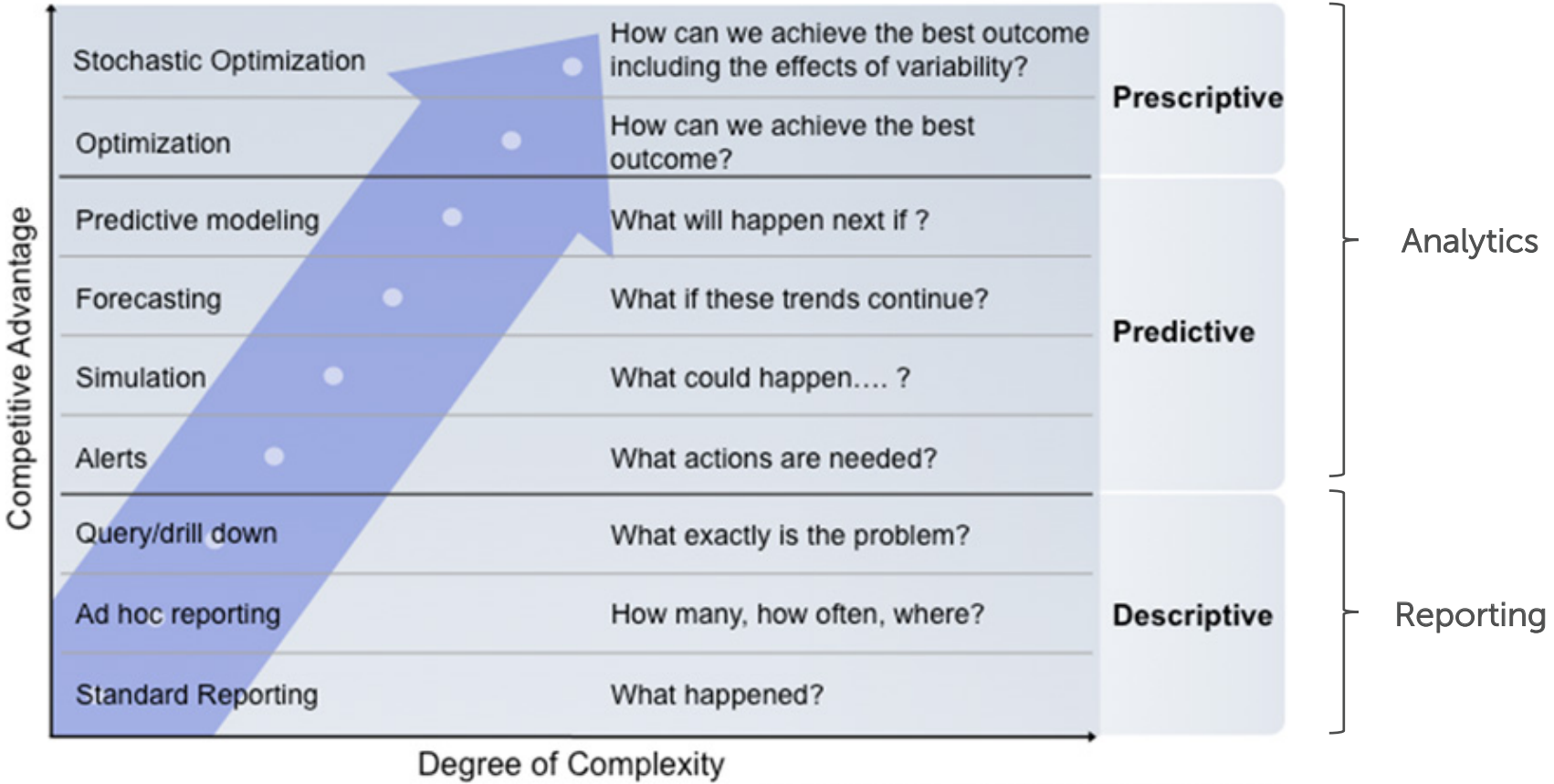
Characteristics of Business Intelligence (BI)

- Provides pre-created dashboards for management
- Repeated visualization of well known analysis steps
- Deals with structured data
- Typically data is generated within the organization
- Central data storage (vs. multiple data silos)
- Handled well by specialized database techniques

Typical types of questions and insight

- Customer service data: “what business causes customer wait times”
- Sales and marketing data: “which marketing is most effective”
- Operational data: “efficiency of the help desk”
- Employee performance data: “who is most/least productive”

Levels of Analysis



Based on: Competing on Analytics, Davenport and Harris, 2007

From Data Warehouse to Data Lake

Dresden Database Systems Group

With cheap storage costs, people promote the concept of the data lake

- Combines data from many sources and of any type
- Allows for conducting future analysis and not miss any opportunity

Collect everything

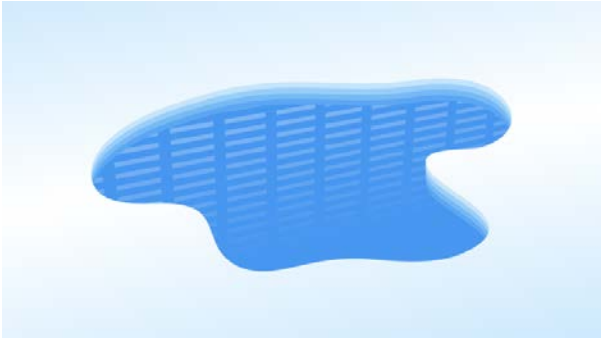
- All data, both raw sources over extended periods of time as well as any processed data
- Decide during analysis which data is important, e.g., no “schema” until read

Dive in anywhere

- Enable users across multiple business units to refine, explore and enrich data on their terms

Flexible access

- Enable multiple data access patterns across a shared infrastructure: batch, interactive, online, search, and others



Roles in Big Data Projects

Dresden Database Systems Group

Data scientist

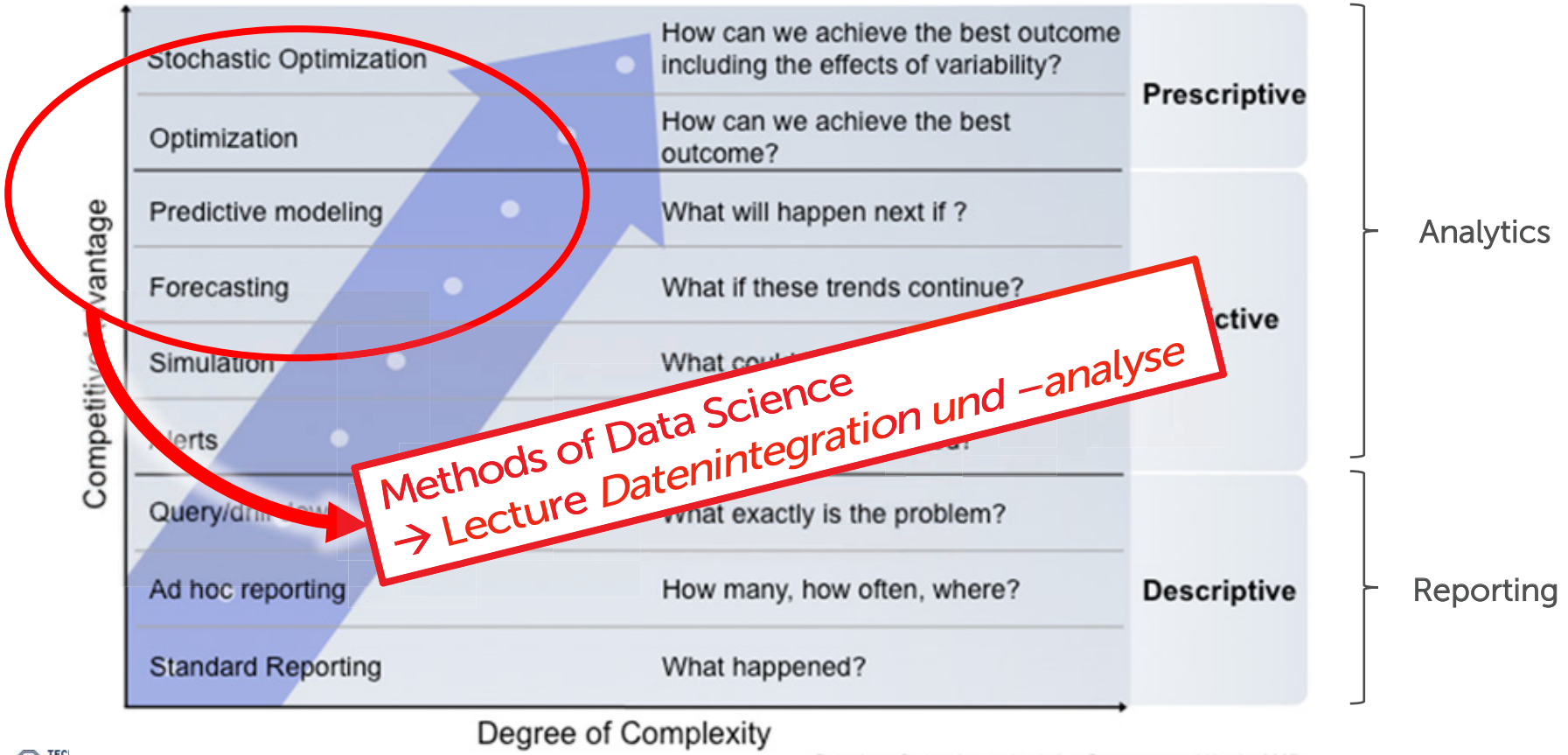
- Data science is a systematic method dedicated to knowledge discovery via data analysis
- In business, optimize organizational processes for efficiency
- In science, analyze experimental/observational data to derive results
- Typical skills
  - Statistics + (mathematics) background
  - Computer science: Programming, e.g.: R, (SAS,) Java, Scala, Python; Machine learning
  - Some domain knowledge for the problem to solve

Data engineer

- Data engineering is the domain that develops and provides systems for managing and analyzing big data
- Build modular and scalable data platforms for data scientists
- Deploy big data solutions
- Typical skills
  - Computer science background
  - Databases
  - Software engineering
  - Massively parallel processing
  - Real-time processing
  - Languages: C++, Java, (Scala,) Python
  - Understand performance factors and limitations of systems

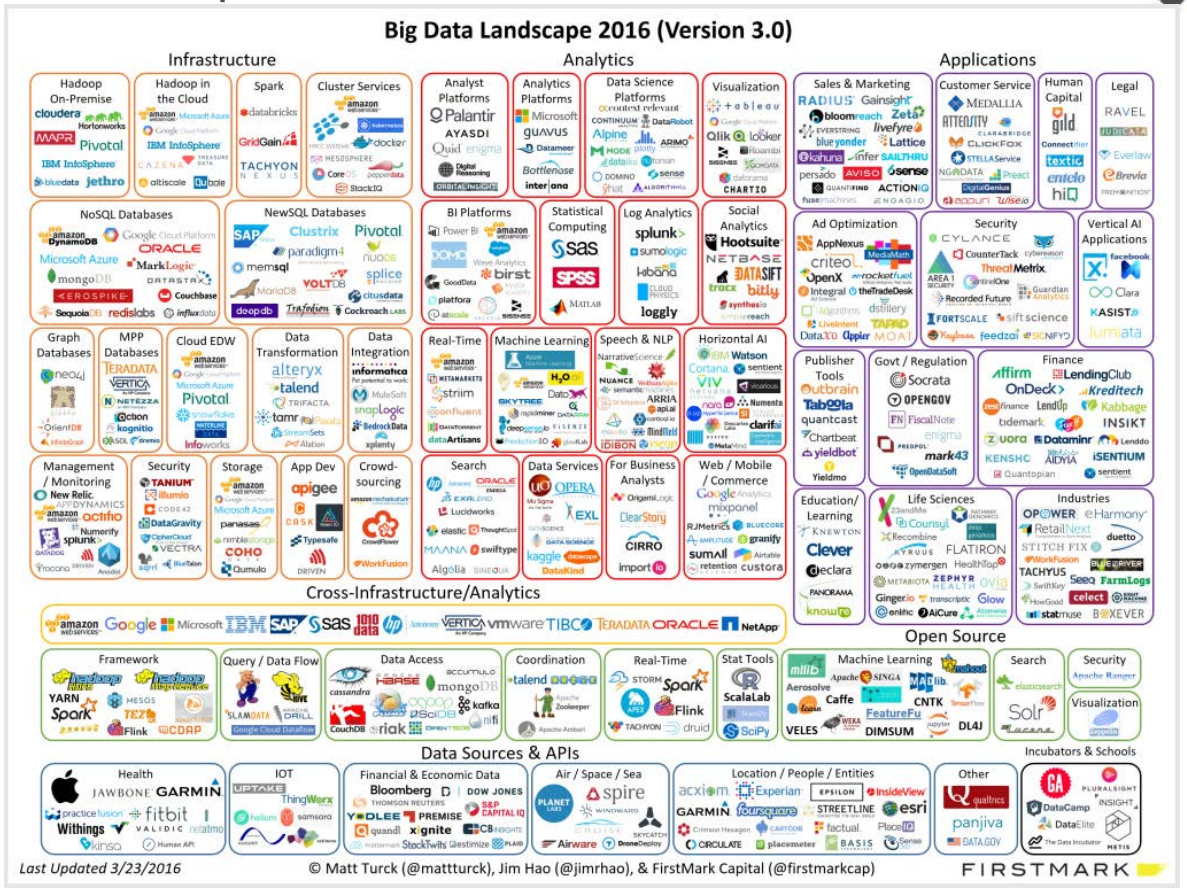
Our Focus in Teaching in Research

Levels of Analysis



Based on: Competing on Analytics, Davenport and Harris, 2007

Big Data Landscape(s)



Data Systems

Dresden Database Systems Group

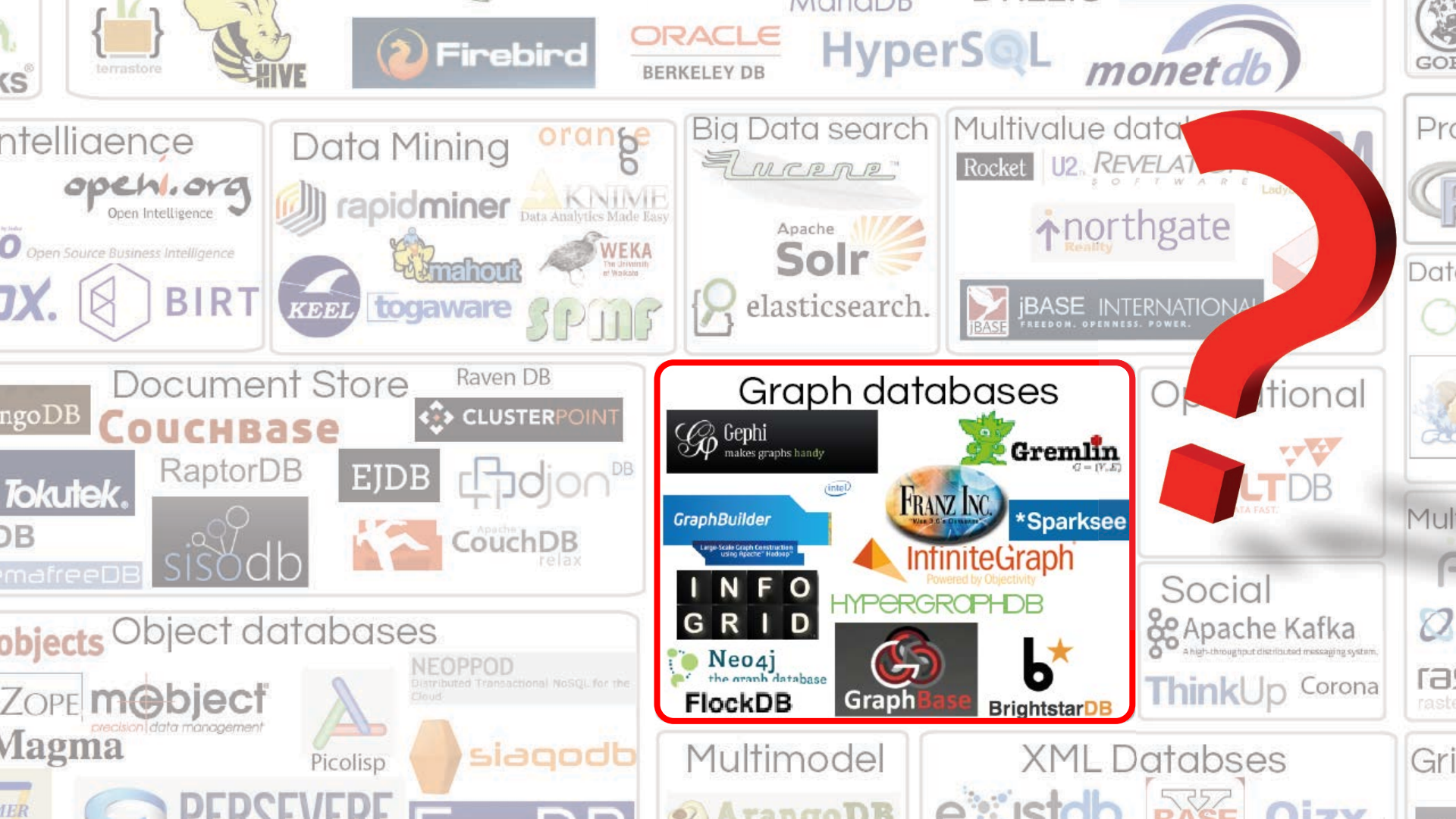
DATA SYSTEMS ARE IN THE MIDDLE OF ALL THIS

A DATA SYSTEM...

- ...stores data...
- ...provides access to data...
- ...and (ideally) makes data analysis easy

DIFFERENT DATA SYSTEMS USE DIFFERENT DATA MODELS

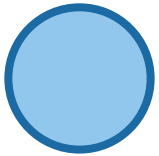
(Big) Data System Landscape(s)


terrastore
HIVE
Firebird
ORACLE
BERKELEY DB
HyperSQL
monetdb
Intelligence
openi.org
Open Intelligence
Open Source Business Intelligence
BIRT
Data Mining
orange
rapidminer
KNIME
Data Analytics Made Easy
mahout
WEKA
KEEL
togaware
SPMF
Big Data search
Lucene
Apache
Solr
elasticsearch.
Multivalue data
Rocket
U2
REVELATION SOFTWARE
northgate
Reality
jBASE INTERNATIONAL
FREEDOM. OPENNESS. POWER.
Document Store
Raven DB
CouchBase
CLUSTERPOINT
Tokutek
RaptorDB
EJDB
djon DB
sisodb
Apache CouchDB
relax
Graph databases
Gephi
makes graphs handy
Gremlin
intel
GraphBuilder
FRANZ INC.
*Sparksee
InfiniteGraph
Powered by Objectivity
INFO GRID
HYPERGRAPHDB
Neo4j
the graph database
GraphBase
BrightstarDB
FlockDB
Social
Apache Kafka
A high-throughput distributed messaging system.
ThinkUp
Corona
Object databases
objects
ZOPE
mobject
precision | data management
Magma
Picolisp
NEOPPOD
Distributed Transactional NoSQL for the Cloud
siaqodb
PERSEVERE
Multimodel
ArangoDB
XML Databases
exist-db
BASEX

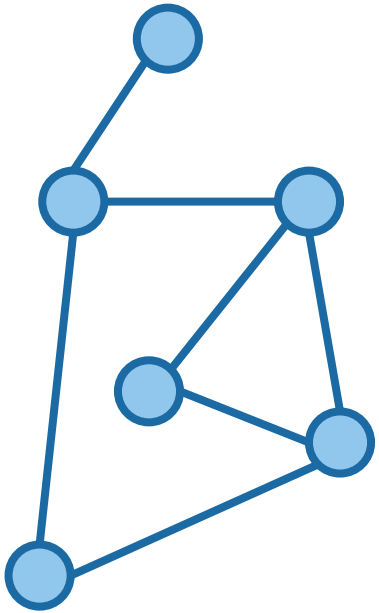
Graph Building Blocks

Dresden Database Systems Group

Nodes (Dots)



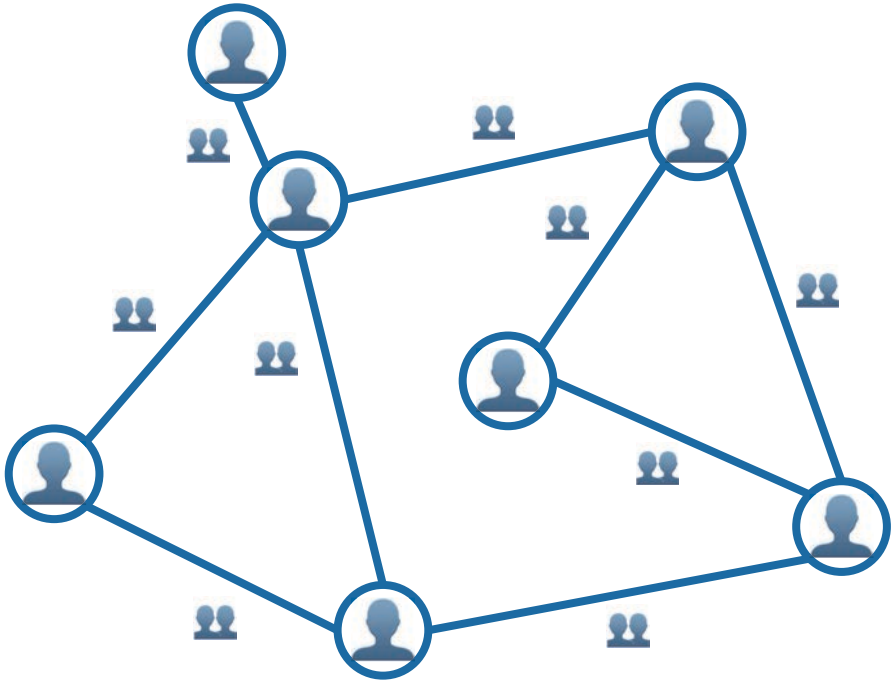
- Like an entity in ER
- Exist on their own
- Have object identity



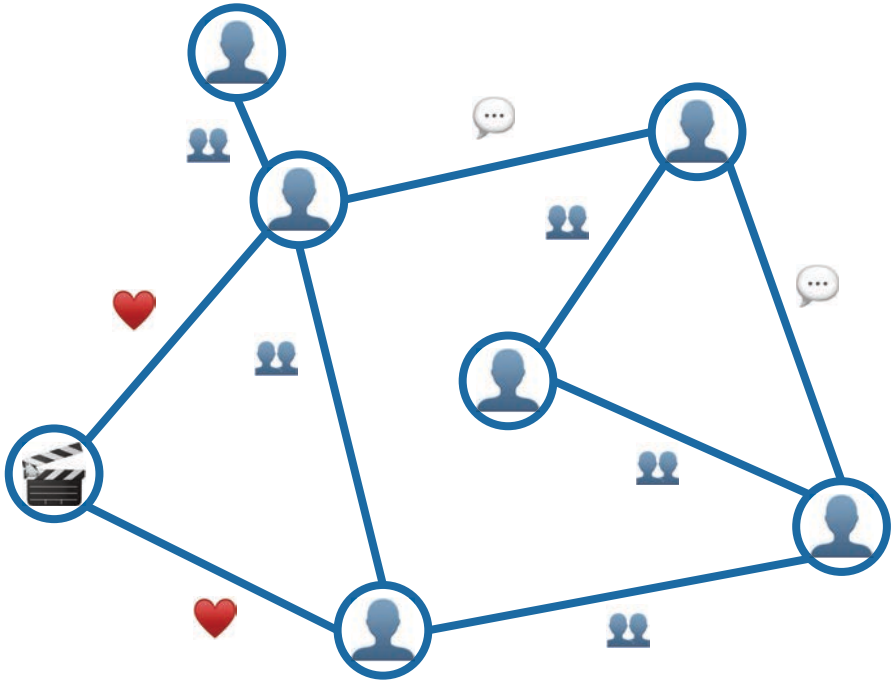
Edges (Lines)

- Like a relationship in ER
- Exist only between nodes
- Identity depends on the nodes they connect

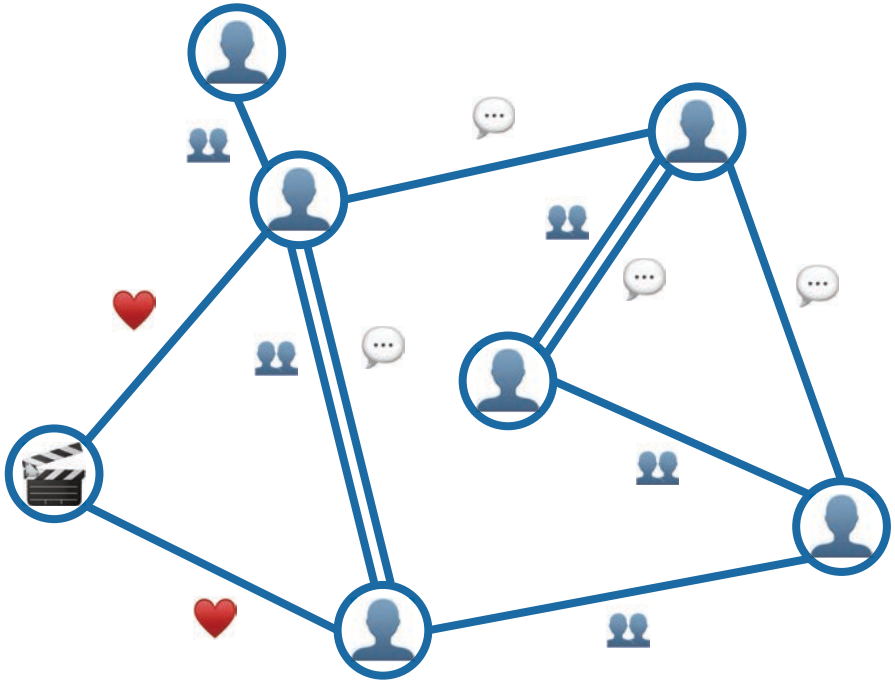
Graph Data – Social Network



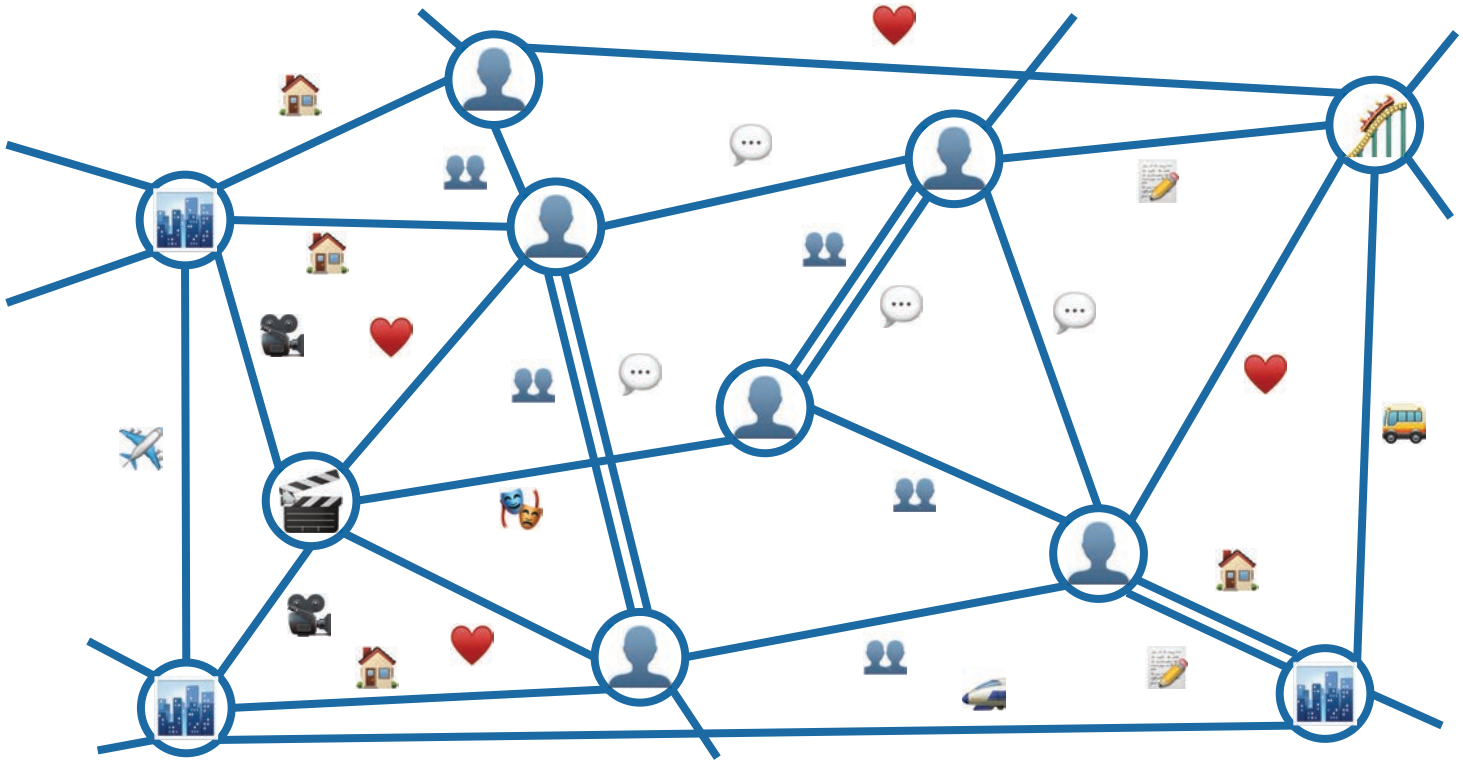
Graph Data – Social Network



Graph Data – Social Network



Graph Data



Social Graphs

Dresden Database Systems Group

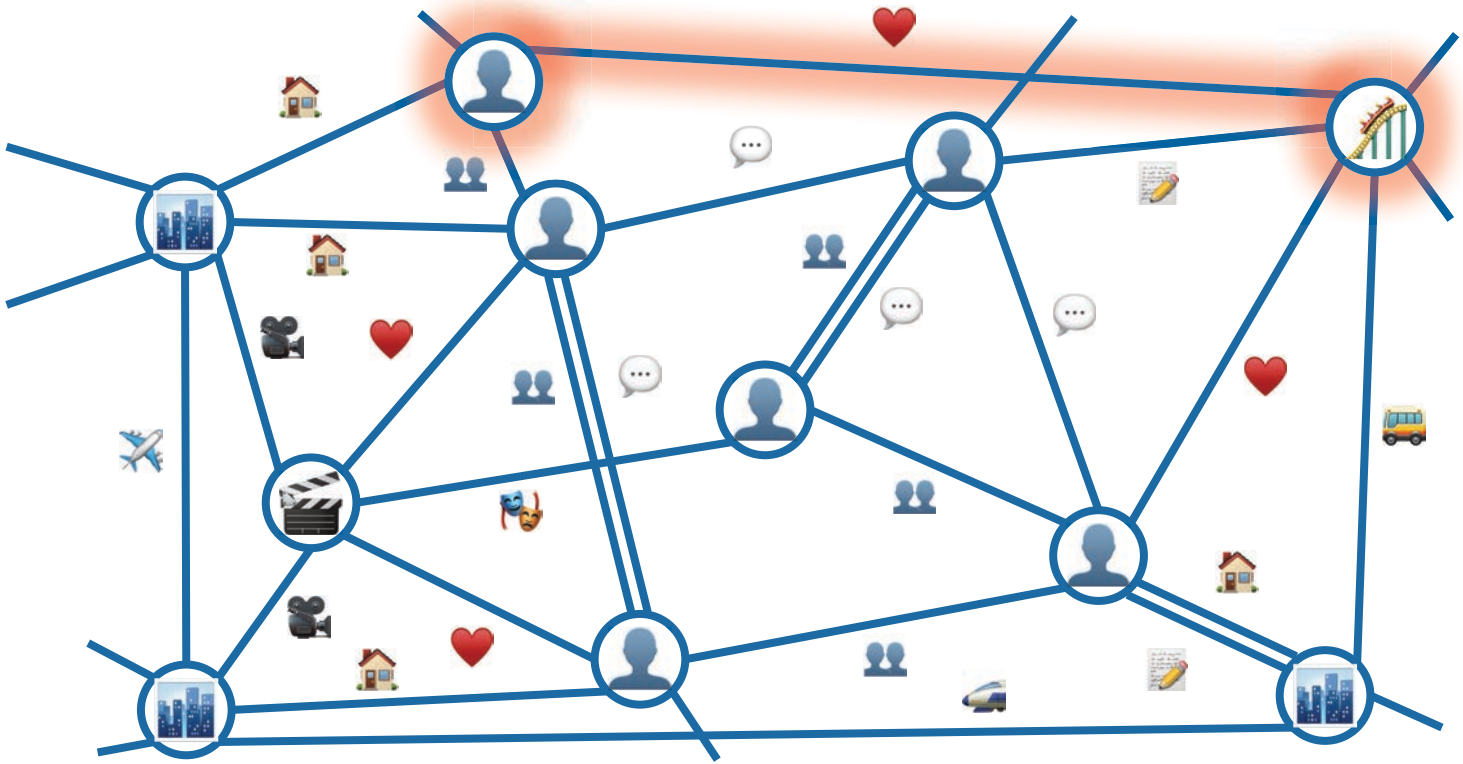
Facebook

- May 2013

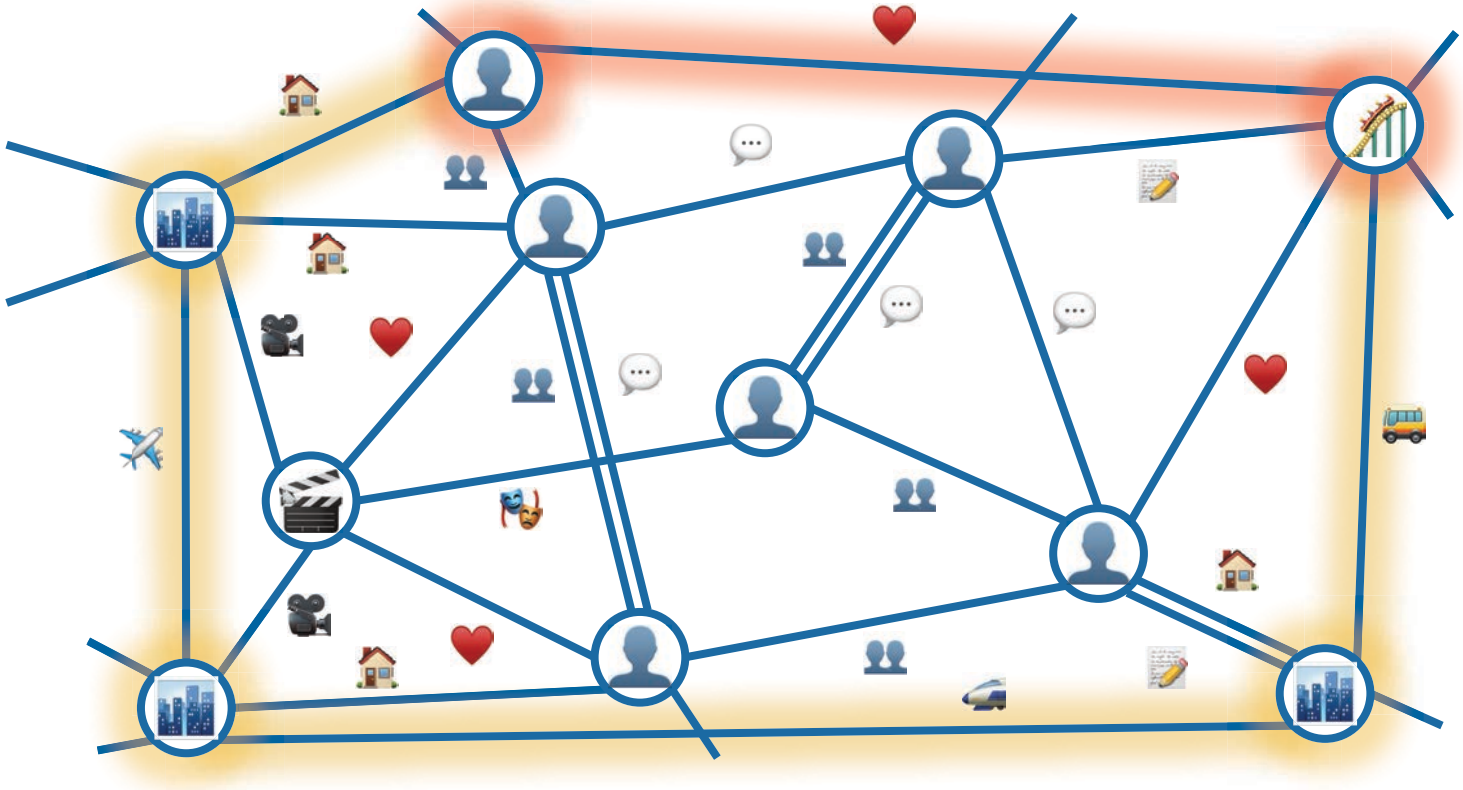
As of 12/2014: "1.39 billion active users with more than 400 billion edges"

[Ching et al. One Trillion Edges: Graph Processing at Facebook-Scale. PVLDB (12)2015]

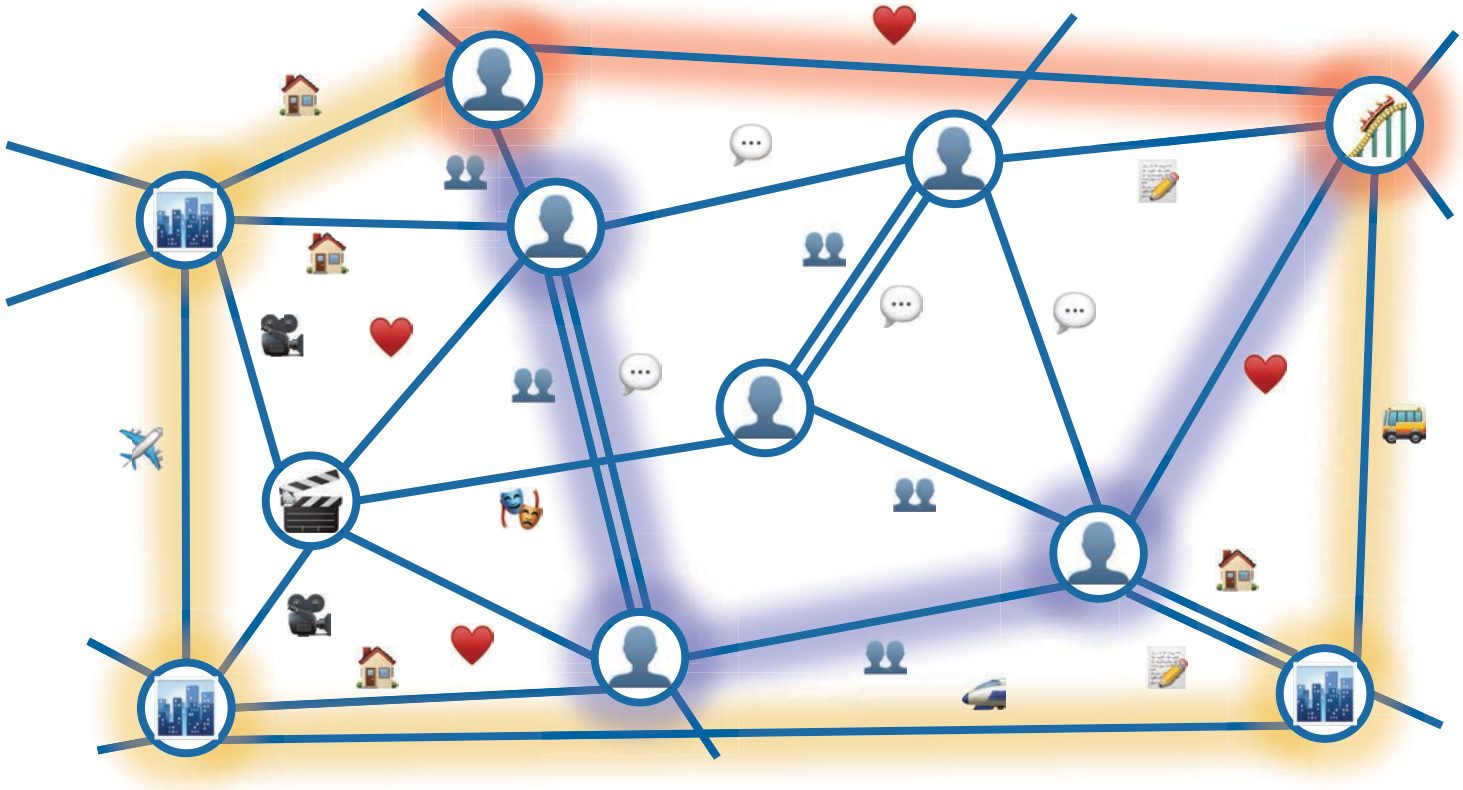
Graph Data



Graph Data



Graph Data

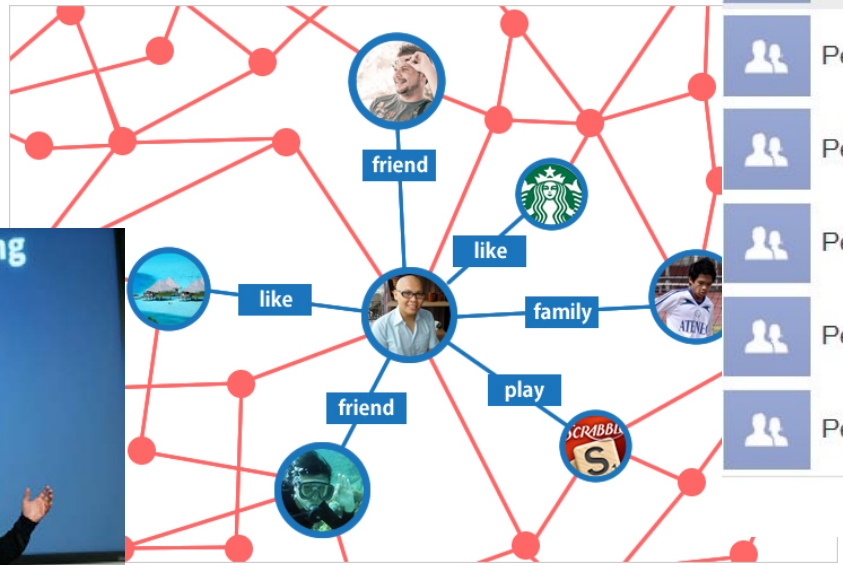


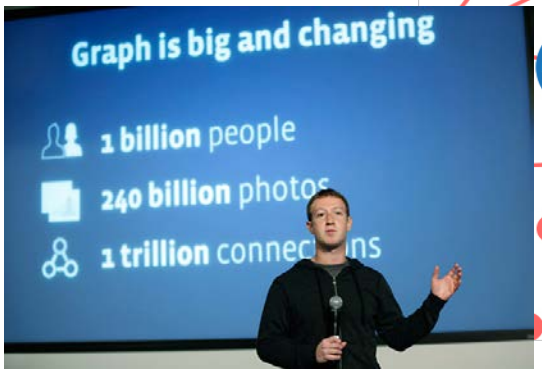
Structured Search

Dresden Database Systems Group

Example: Facebook Graph Search

- Finding subgraph structures
- Very natural way of formulating queries





[http://socialnewsdaily.com/15865/facebook-social-graph-search-a-great-way-to-find-working-professionals-in-your-network/]

Social Network Analysis

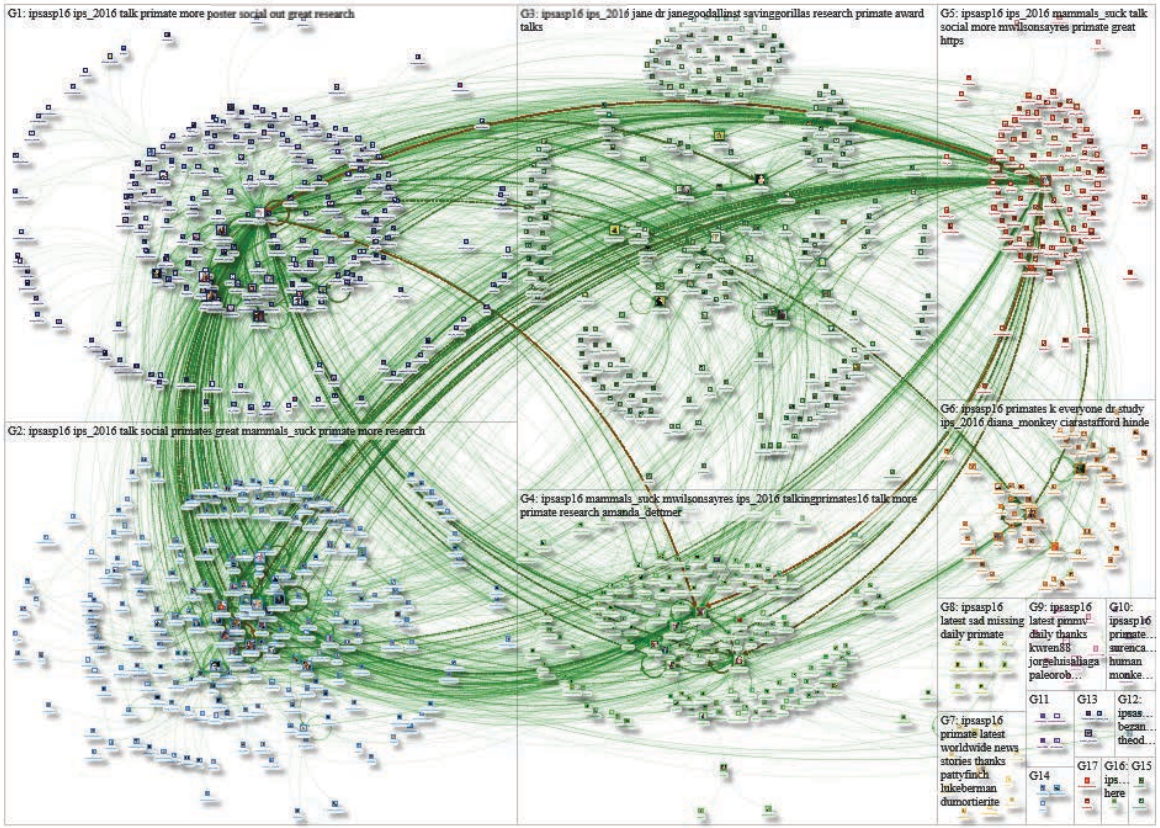
IPS ASP CHICAGO 2016

Dresden Database Systems Group

Example: Twitter communication

- Users tweeting on a specific topic
- Others reply of retweet
- Users can be grouped based on communication topology (-> graph clustering)

- Analysis reveals user groups and dominant communication patterns



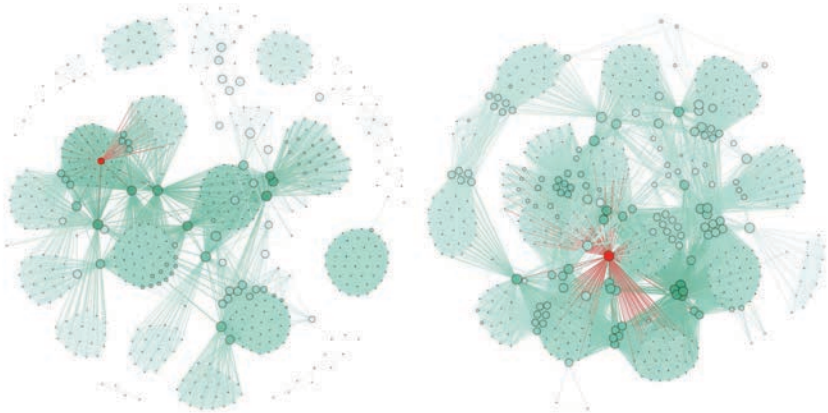
[https://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=76277]

Citation Networks

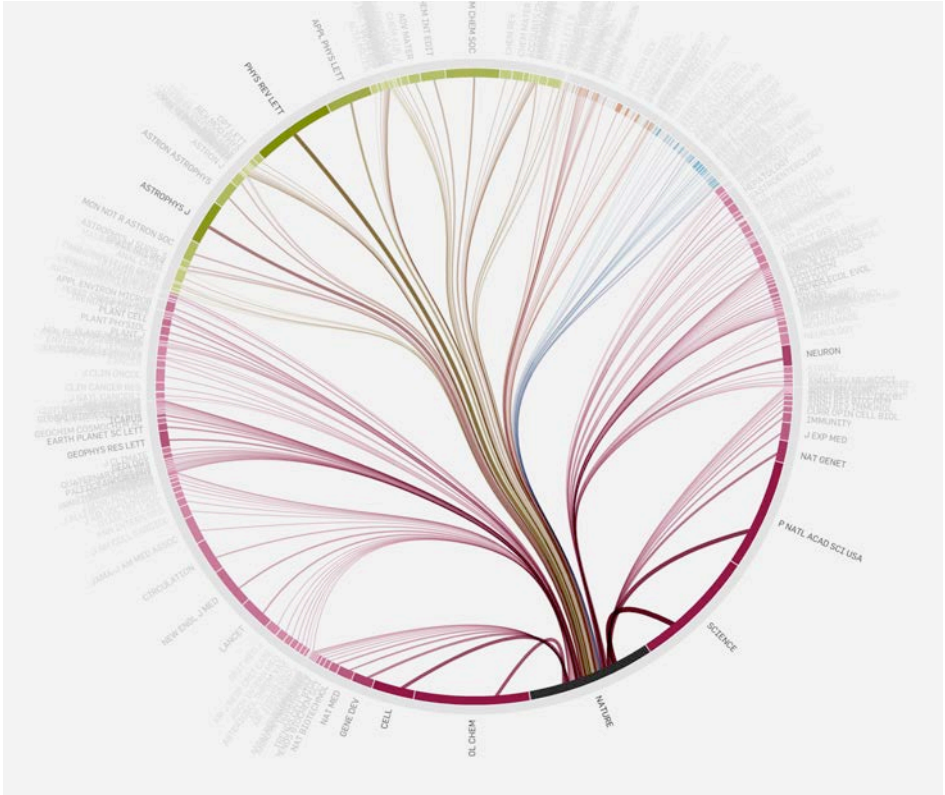
Dresden Database Systems Group

Example: DBLP

- Open bibliographic information on major computer science journals and proceedings
- >3.4 million publication
- >7000 new publication per month
- >1.7 million authors



[Emre Sarigöl et al. Predicting Scientific Success Based On Coauthorship Networks. EPJ Data Science, 2014]



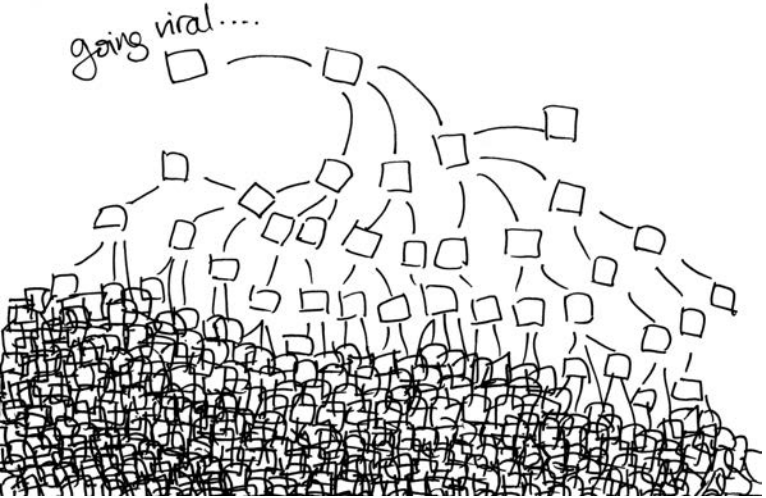
[http://well-formed.eigenfactor.org/projects/well-formed/radial.html#/]

Viral Marketing

Dresden Database Systems Group

Viral Marketing

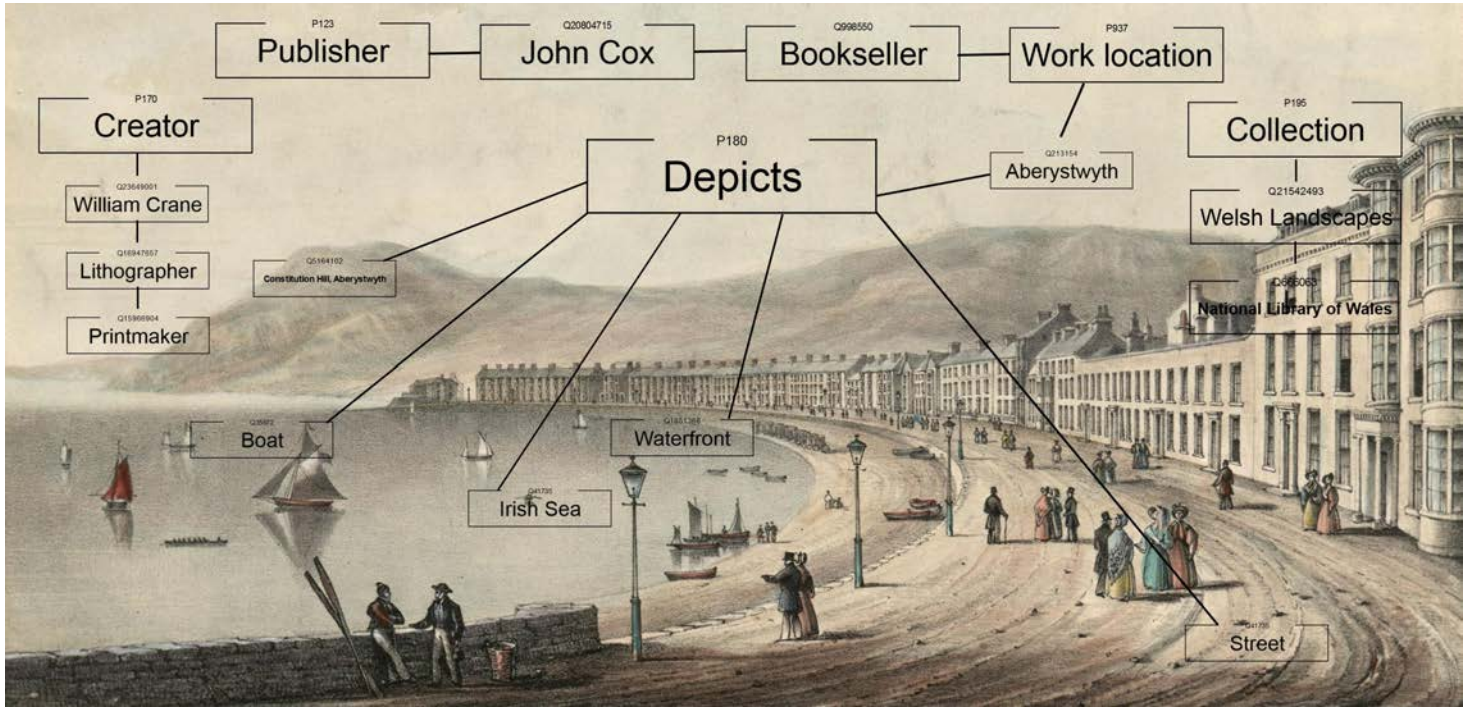
- spreading content to one person so that more than one person engaging with the content
- Techniques
  - Influence estimation
  - Influence maximization



Knowledge Graphs

Dresden Database Systems Group

KNOWLEDGE GRAPH OF A PICTURE



[The National Library of Wales, https://www.llgc.org.uk/blog/?p=11246]

Knowledge Graphs

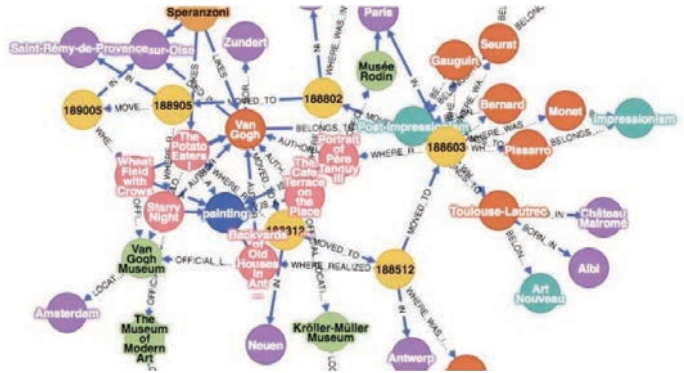
SLUB
Wir führen Wissen.

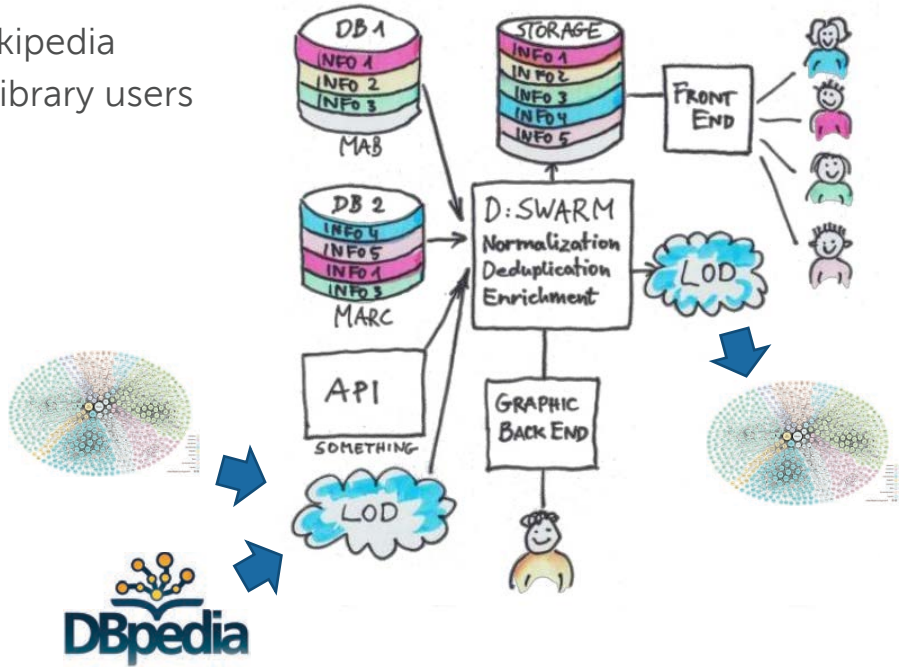
Avantgarde Labs

Dresden Database Systems Group

Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB)

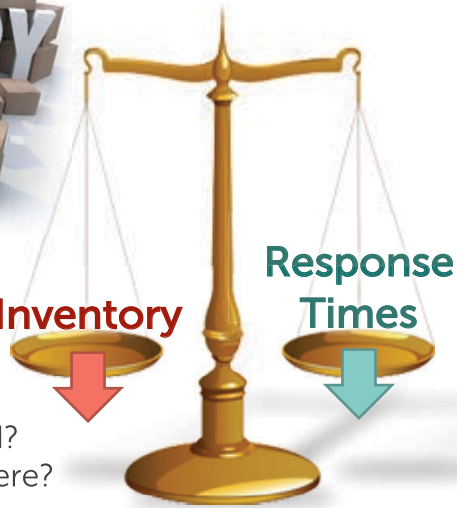
- Adds semantics search to library online catalog
- Utilizing multi-lingual knowledge data from Wikipedia
- Significant improvements in search quality for library users





[http://www.slideshare.net/JensMittelbach/dswarm-a-library-data-management-platform-based-on-a-linked-open-data-approach]

Supply Chain Management

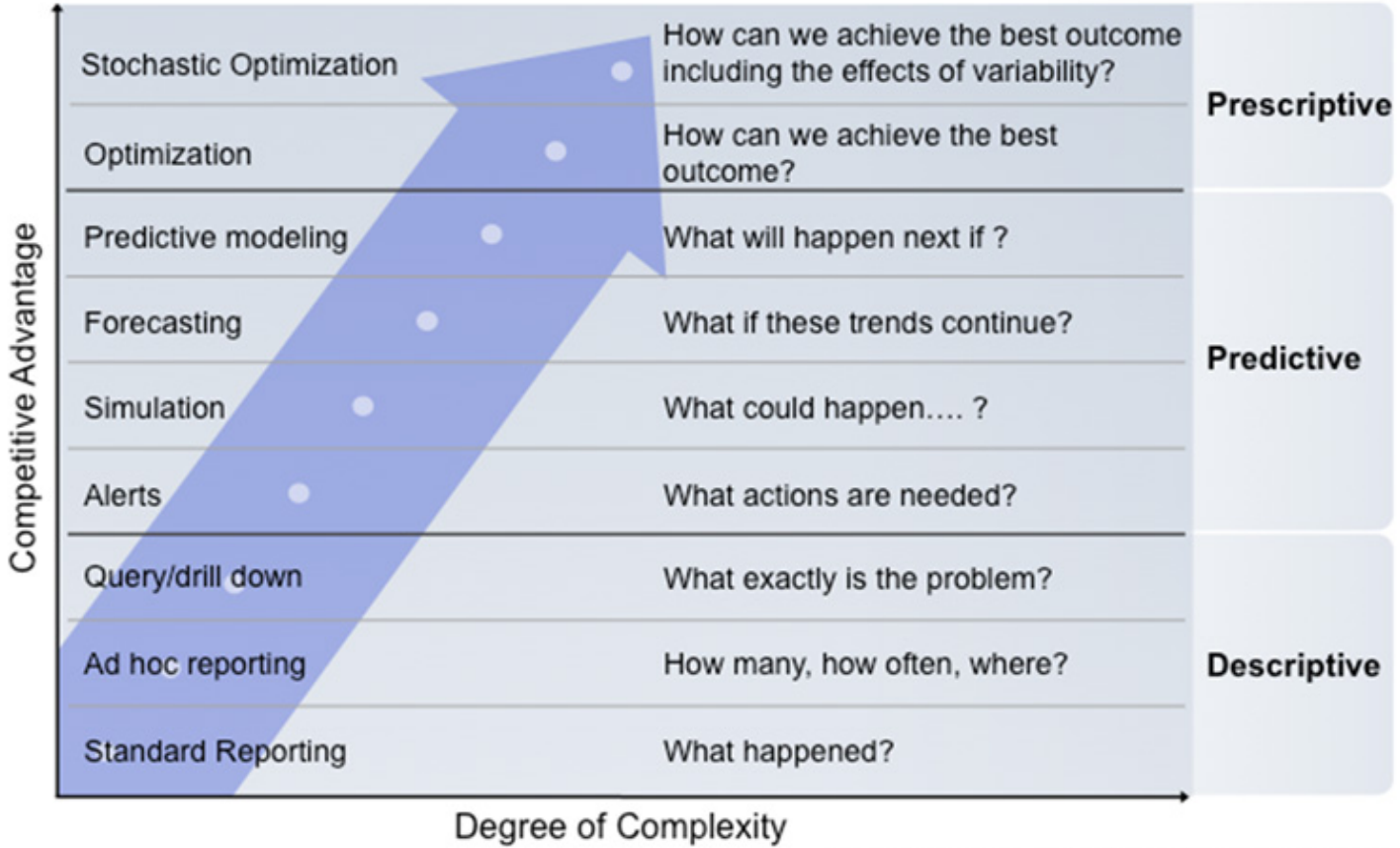


When to send?
How much to send?
From where to where?

Supply Chain Optimization
- Customer A: 10-30% reduction in inventory
- Customer B: 8% reduction in transportation costs

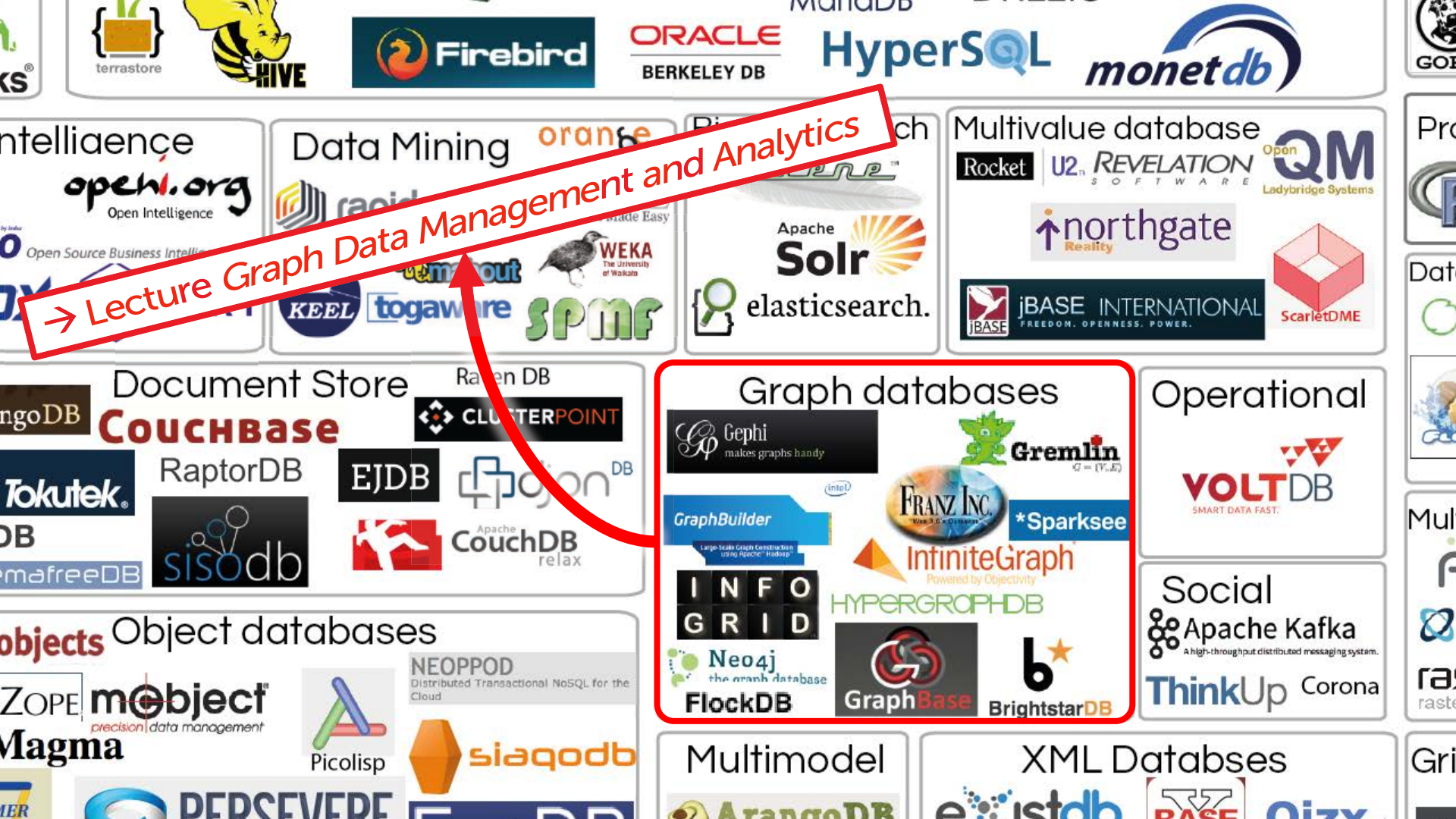
[http://www.logicblox.com/solutions/supply-chain-optimization/]

Level of Analytics



Graph Data Management Applications

Based on: Competing on Analytics, Davenport and Harris, 2007


→ Lecture Graph Data Management and Analytics
terrastore
HIVE
Firebird
ORACLE
BERKELEY DB
HyperSQL
monetdb
Intelligence
openi.org
Open Intelligence
Data Mining
orange
rapid
WEKA
The University of Waikato
KEEL
togaware
SPMF
Apache
Solr
elasticsearch.
Multivalue database
Rocket
U2
REVELATION SOFTWARE
Open QM
Ladybridge Systems
northgate
Reality
jBASE INTERNATIONAL
FREEDOM. OPENNESS. POWER.
ScarletDME
Document Store
CouchBase
RaptorDB
EJDB
sisodb
Apache CouchDB
relax
Graph databases
Gephi
makes graphs handy
Gremlin
FRANZ INC.
GraphBuilder
*Sparksee
InfiniteGraph
Powered by Objectivity
INFO GRID
HYPERGRAPHDB
Neo4j
the graph database
GraphBase
BrightstarDB
FlockDB
Operational
VOLTDB
SMART DATA FAST.
Social
Apache Kafka
A high-throughput distributed messaging system.
ThinkUp
Corona
Object databases
ZOPE
mobject
precision | data management
Magma
Picolisp
NEOPPOD
Distributed Transactional NoSQL for the Cloud
siaqodb
Multimodel
XML Databses

Summary

Dresden Database Systems Group

Big Data

- Crossing thresholds in exponential growth, digitization, recombinant innovation
- Technical challenges in volume, velocity, variety, veracity, value

Related research and lectures

▪ Data Science methods	→ Datenintegration und -analyse
▪ Data Systems	→ Architektur von Datenbanksystemen → Big Data Platforms
▪ Graph Data	→ Graph Data Management and Analytics
▪ Search in large documents set	→ Information Retrieval