TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Future Trends in Data Analytics

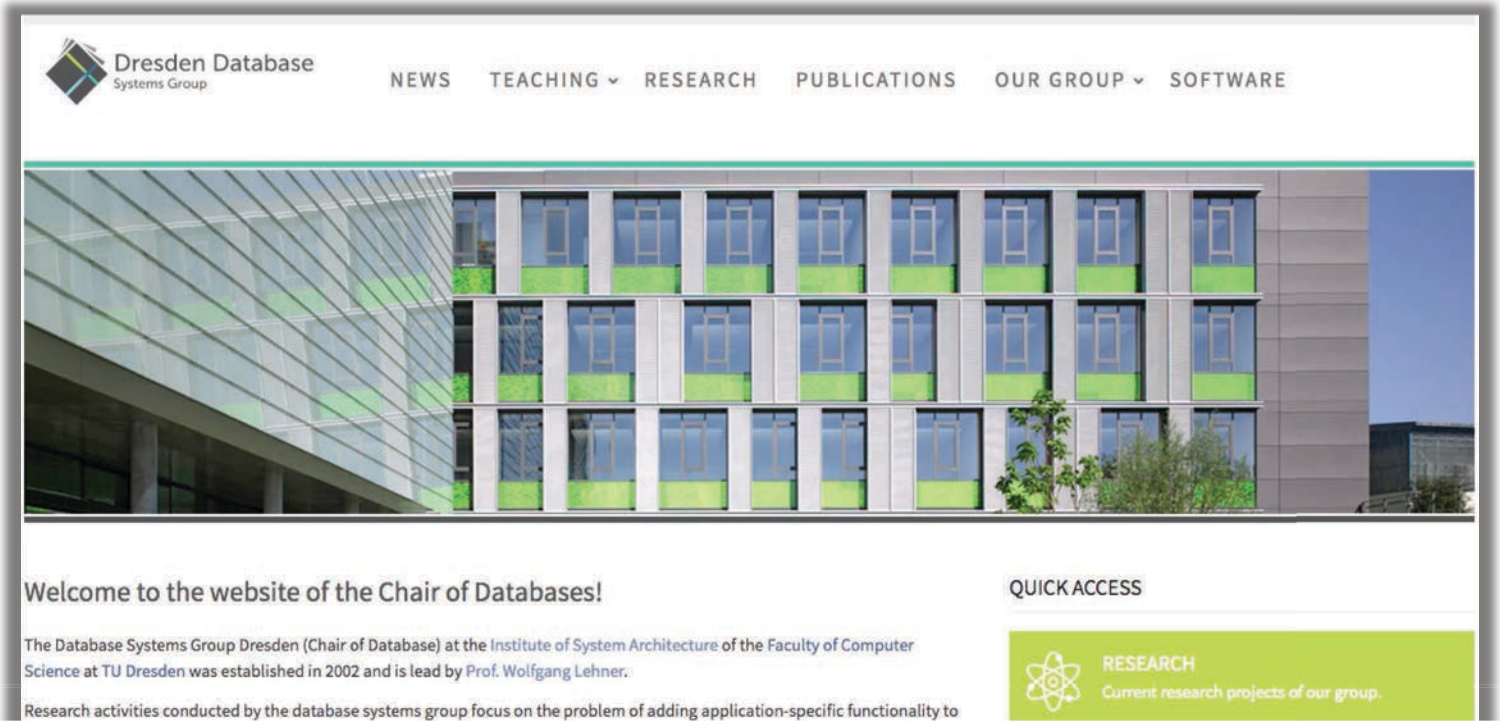
Hannes Voigt

General Information

Dresden Database Systems Group

Lecture Notes

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

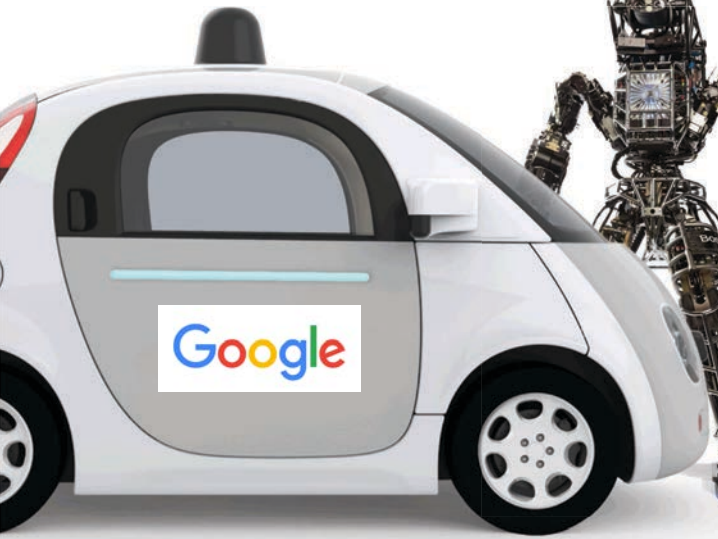


Dresden Database System Group

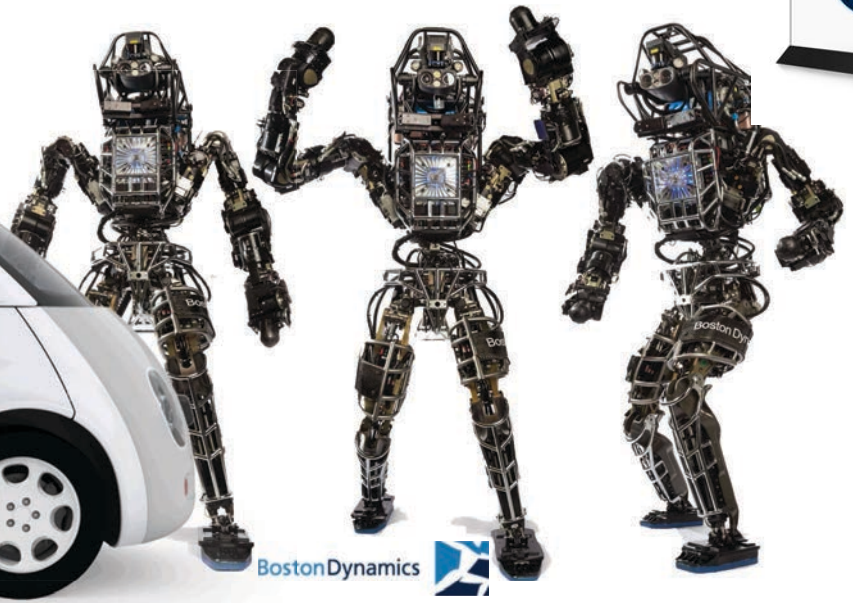
Dresden Database
Systems Group

Future is Now

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



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



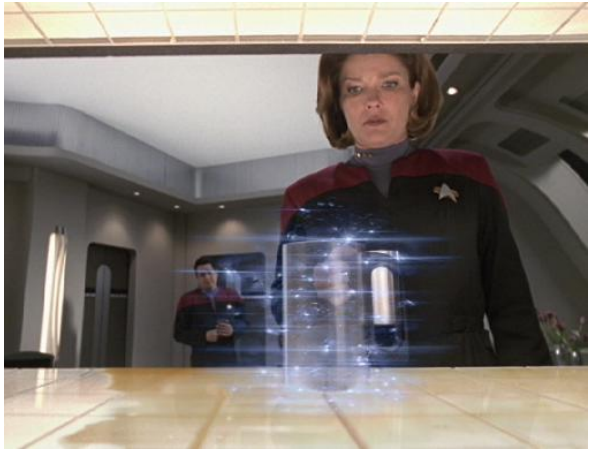
[http://www.bostondynamics.com/]

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

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

… tomorrow?

Sooner then you think!



When have we entered the future?

What was the threshold we crossed?

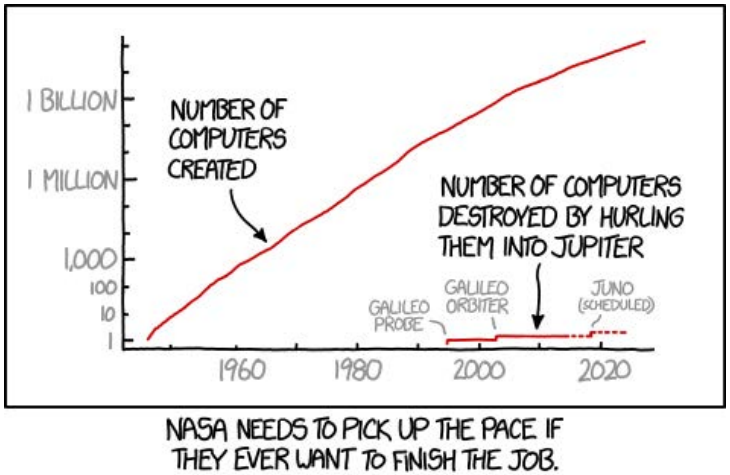
Exponential Growth

Exponential Growth

Dresden Database Systems Group

Outnumbers everything else quickly

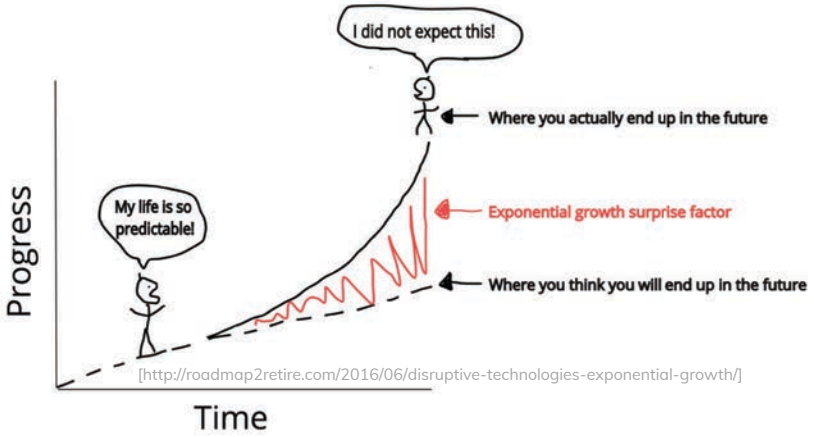
- Asymptotic advantage
- Quickly increasing add-on



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

Leads to surprising results

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



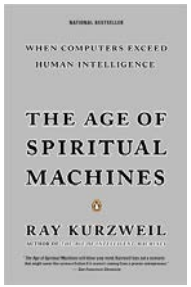
[http://roadmap2retire.com/2016/06/disruptive-technologies-exponential-growth/]

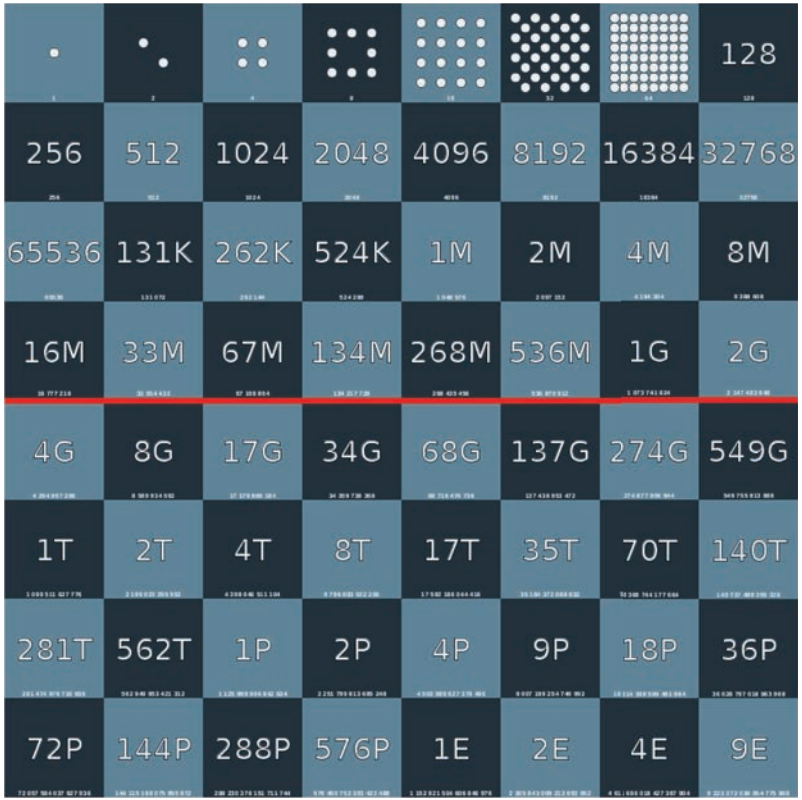
Second Half of the Chessboard

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





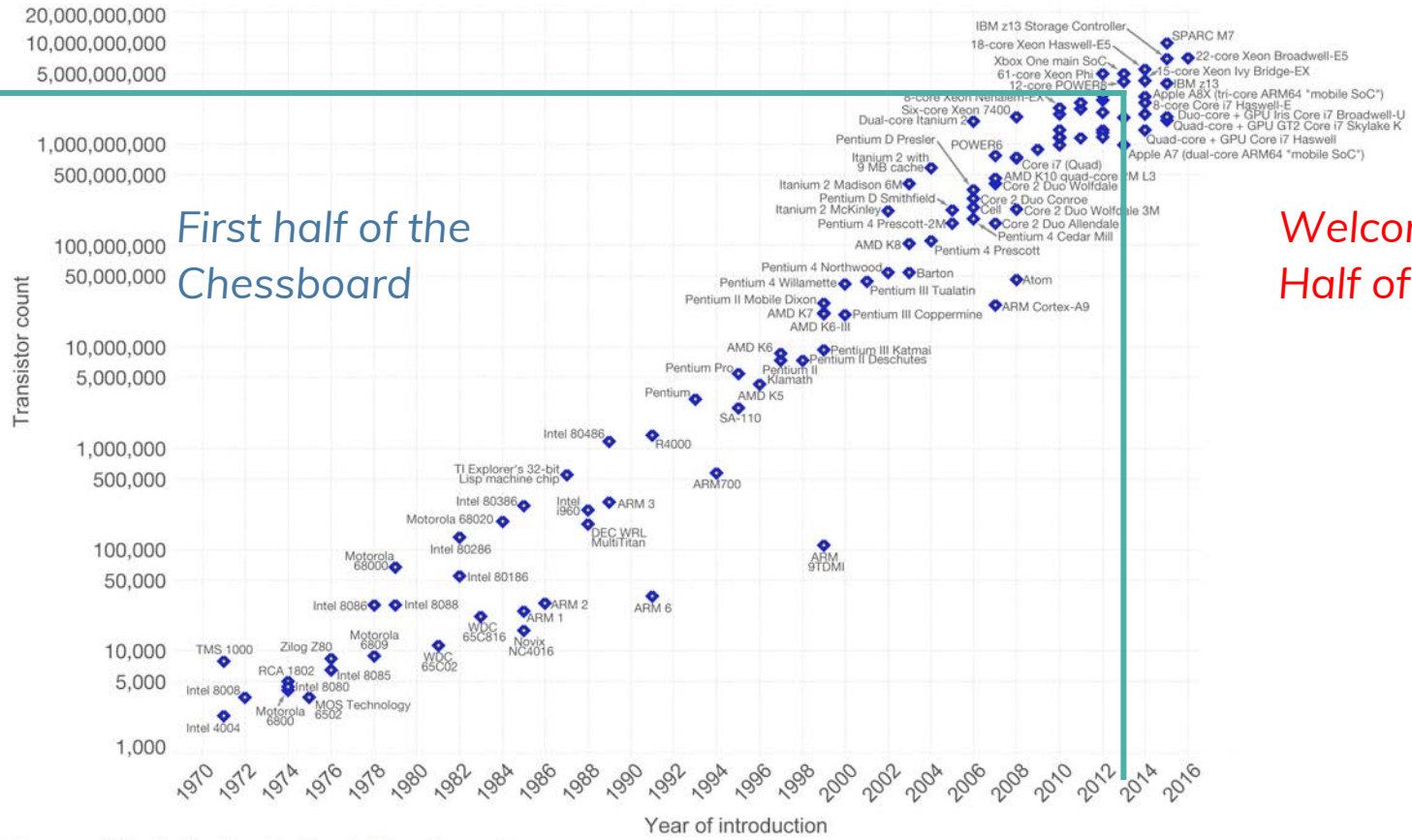
[https://en.wikipedia.org/wiki/File:Wheat_Chessboard_with_line.svg]

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.

Digitization

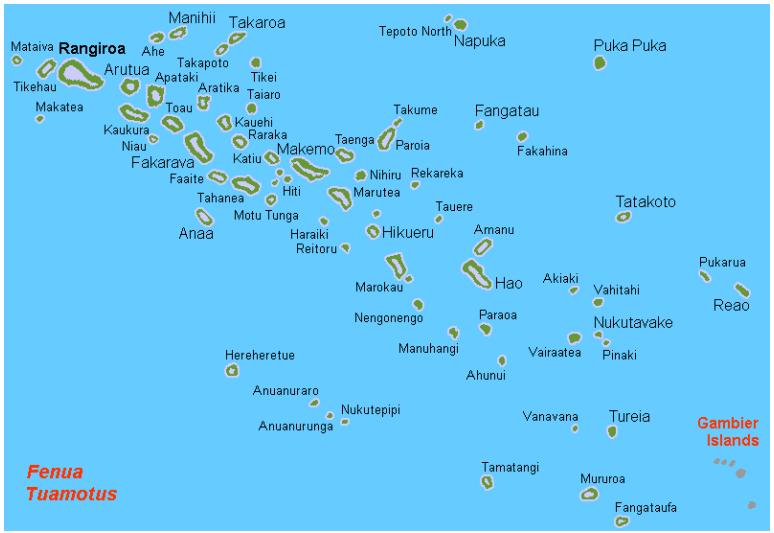
Everything is Digital

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

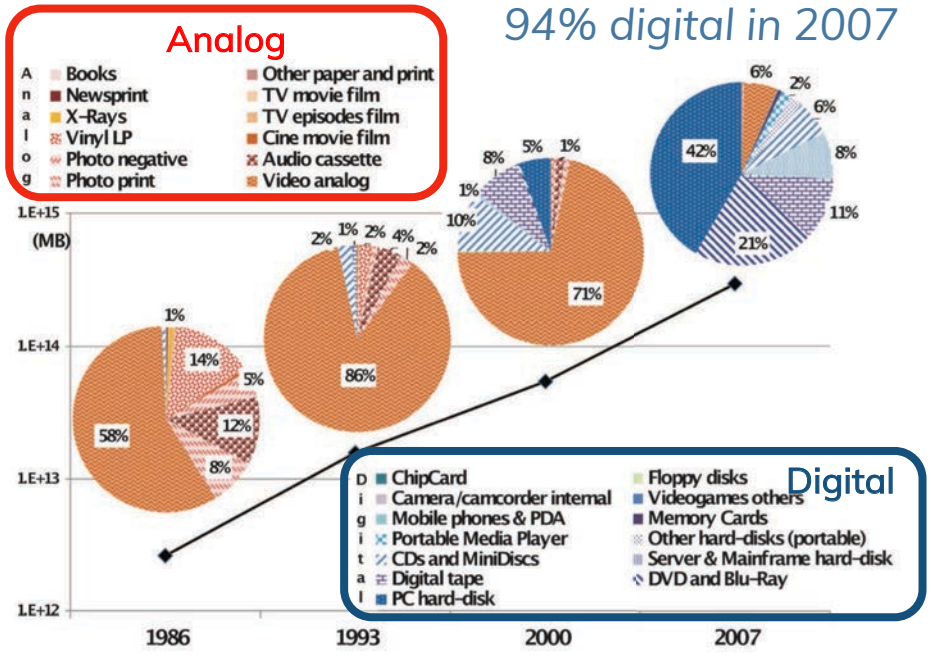
(Algonquin Provincial Park, Ontario, Canada)

Everything is Digital

Dresden Database Systems Group

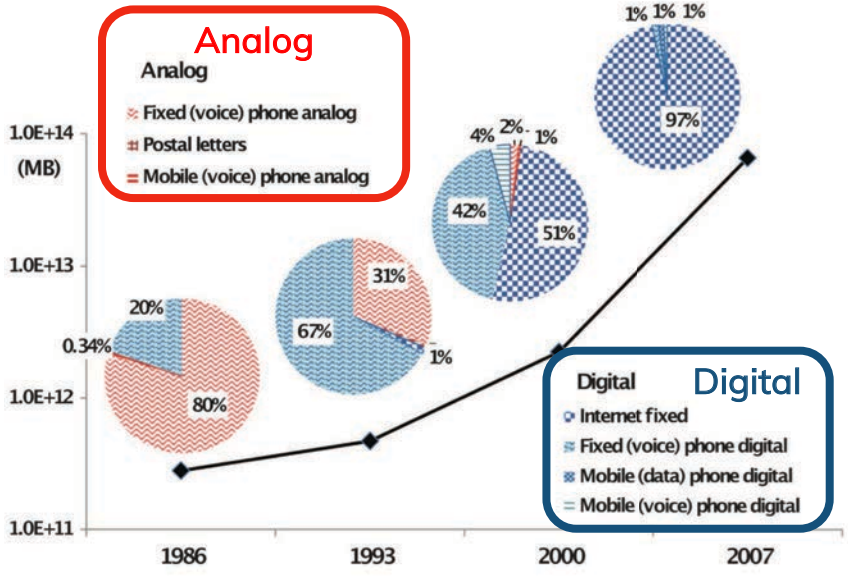
World's capacity to store information

94% digital in 2007



World's capacity to telecommunicate

99% digital in 2007

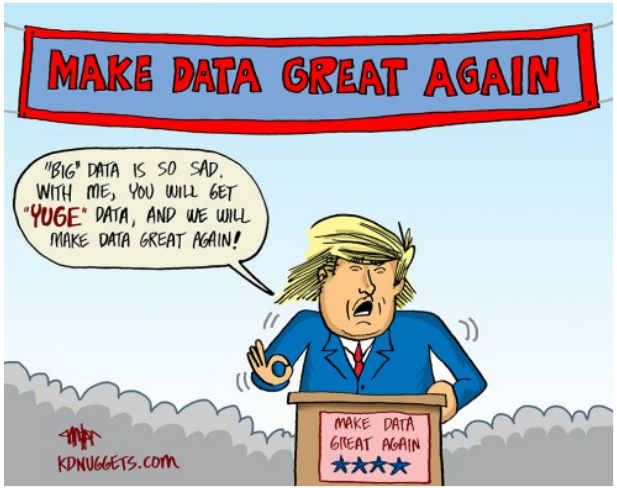


[M. Hilbert and P. Lopez, The World's Technological Capacity to Store, Communicate, and Compute Information, Science, 332, April 2011, DOI: 10.1126/science.1200970]

Everything is Digital

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

Big Data

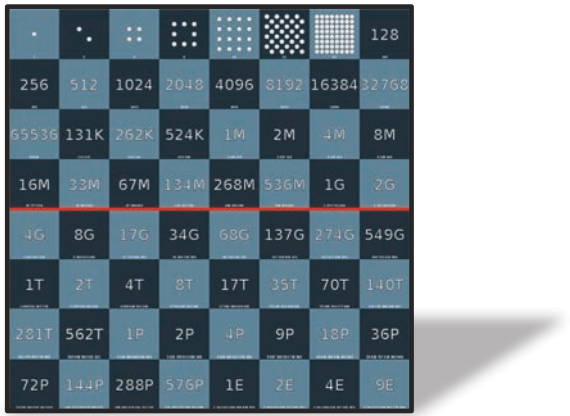

MAKE DATA GREAT AGAIN
"BIG" DATA IS SO SAD. WITH ME, YOU WILL GET "YUGE" DATA, AND WE WILL MAKE DATA GREAT AGAIN!
MAKE DATA GREAT AGAIN
KDNUGGETS.com

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

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.

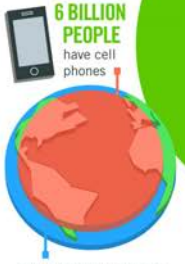
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

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

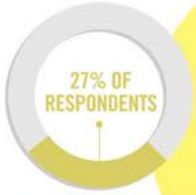
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

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]

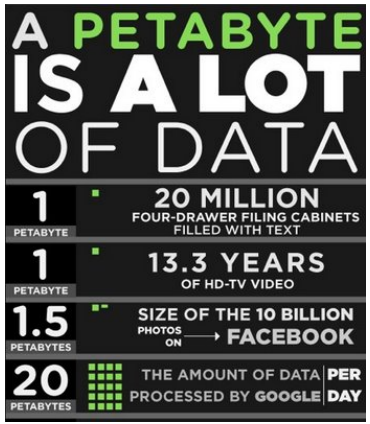
IBM

Data, data, everywhere… Volume

Dresden Database Systems Group

The Petabyte Age

- 2008



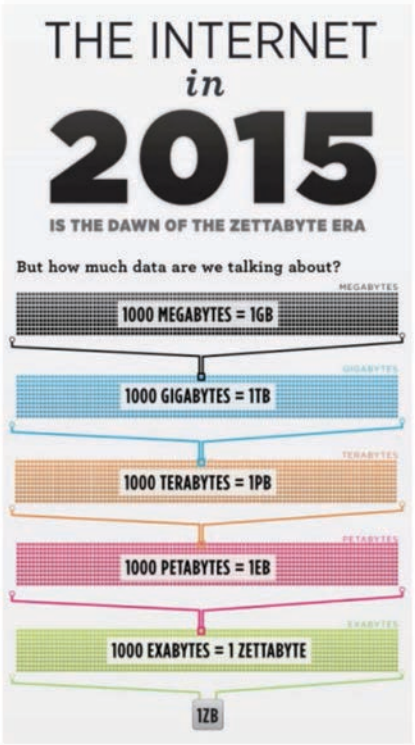
- Eric Schmidt (in 2010): Every 2 Days We Create As Much Information As We Did Up To 2003

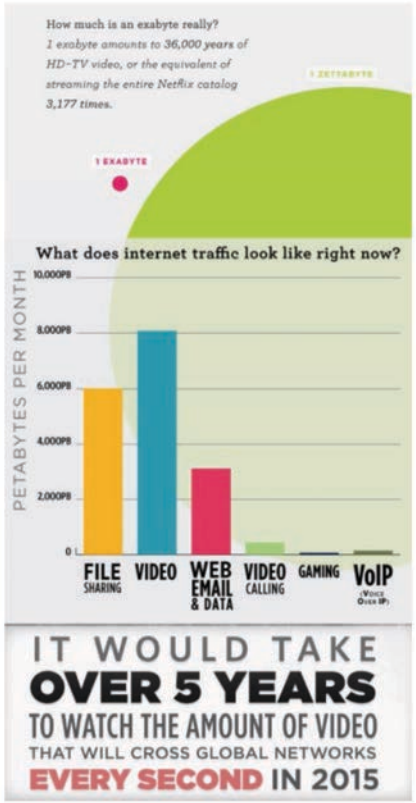
The Zettabyte Age

- 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

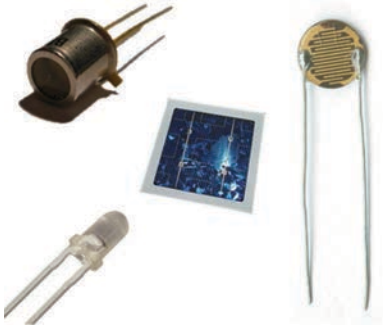
Humane-produced data

- ~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

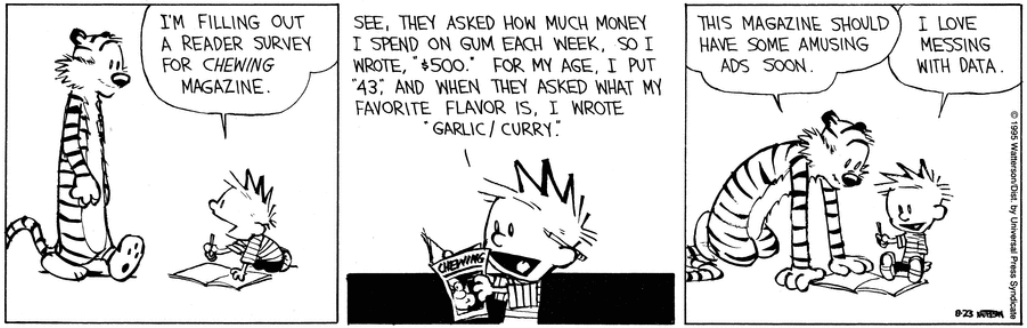
"A lot of Big Data is a lot of small data put together."

Data is messy Veracity

Dresden Database Systems Group

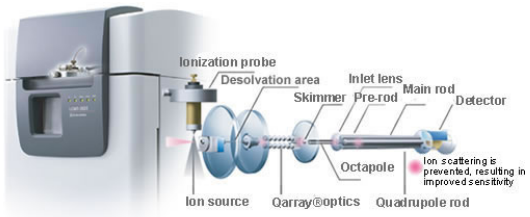
Quality of captured data varies greatly

- 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




information management

Video | Slidesho

BIG DATA & ANALYTICS | DATA MANAGEMENT | MDM & DATA GOVERNANCE | INFRASTRUCTURE | INFO STRATEGY & LEADERSHIP | SECURITY

GET BREAKING NEWS TO YOUR INBOX PLUS MORE EXCLUSIVE BENEFITS! BECOME A REGISTERED ME

NEWS

Messy Big Data Overwhelms Data Scientists

by BOB VIOLINO
FEB 20, 2015 2:00pm ET

Print
Email
Reprints
Comments (2)

Data scientists see messy, disorganized data as a major hurdle preventing them from doing what they find most interesting in their jobs: predictive analysis and data mining for behavioral patterns and future trends, according to a new report from CrowdFlower, a data enrichment platform provider.

A majority of the 153 CrowdFlower online research panel members surveyed (80%) also acknowledged the skills shortage within their field. The respondents work for companies of varied sizes and sectors, mostly in the U.S. All respondents have the term "data scientist" in their job title or job description on LinkedIn, CrowdFlower says.

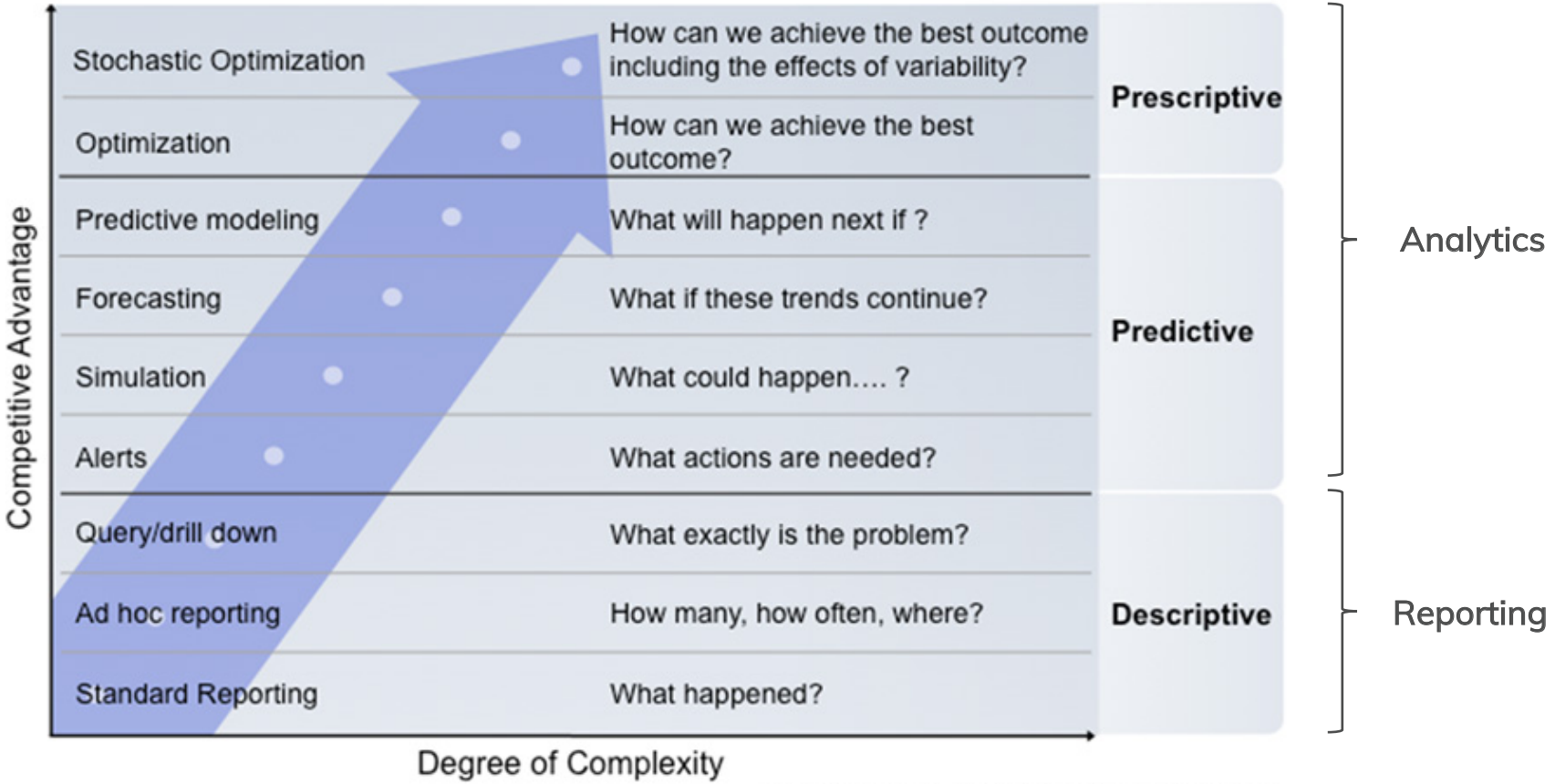
Value – the fifth
V of Big Data

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

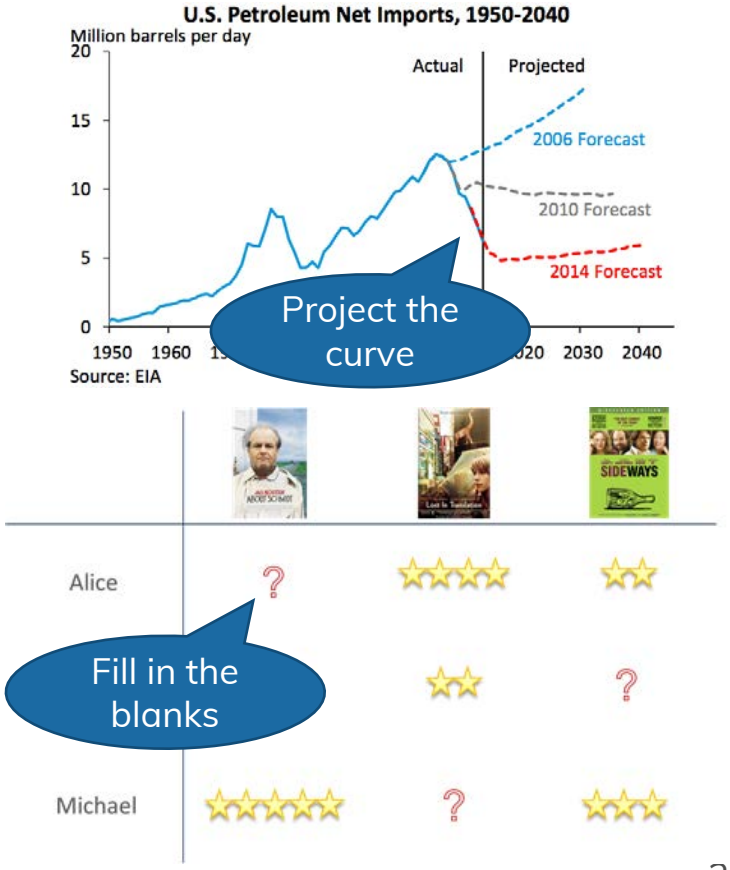
Dresden Database Systems Group

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

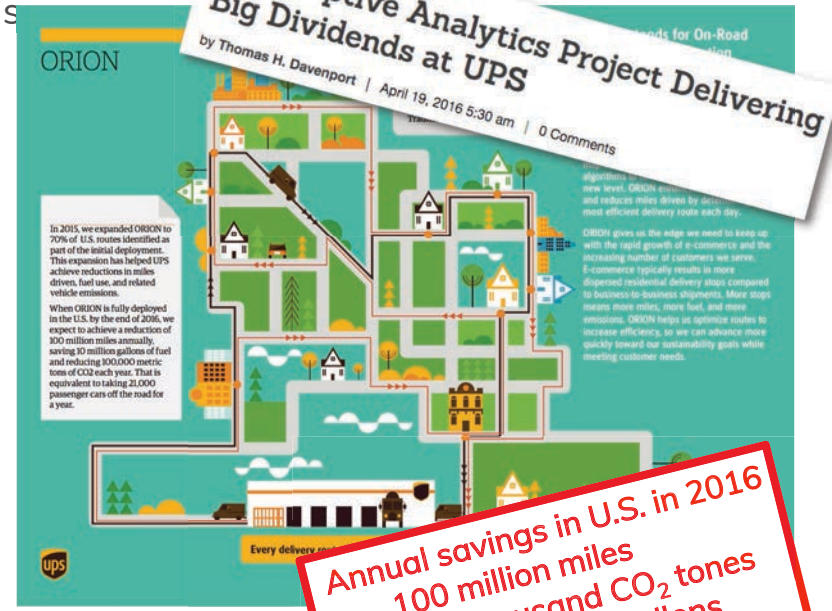
Dresden Database Systems Group

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

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

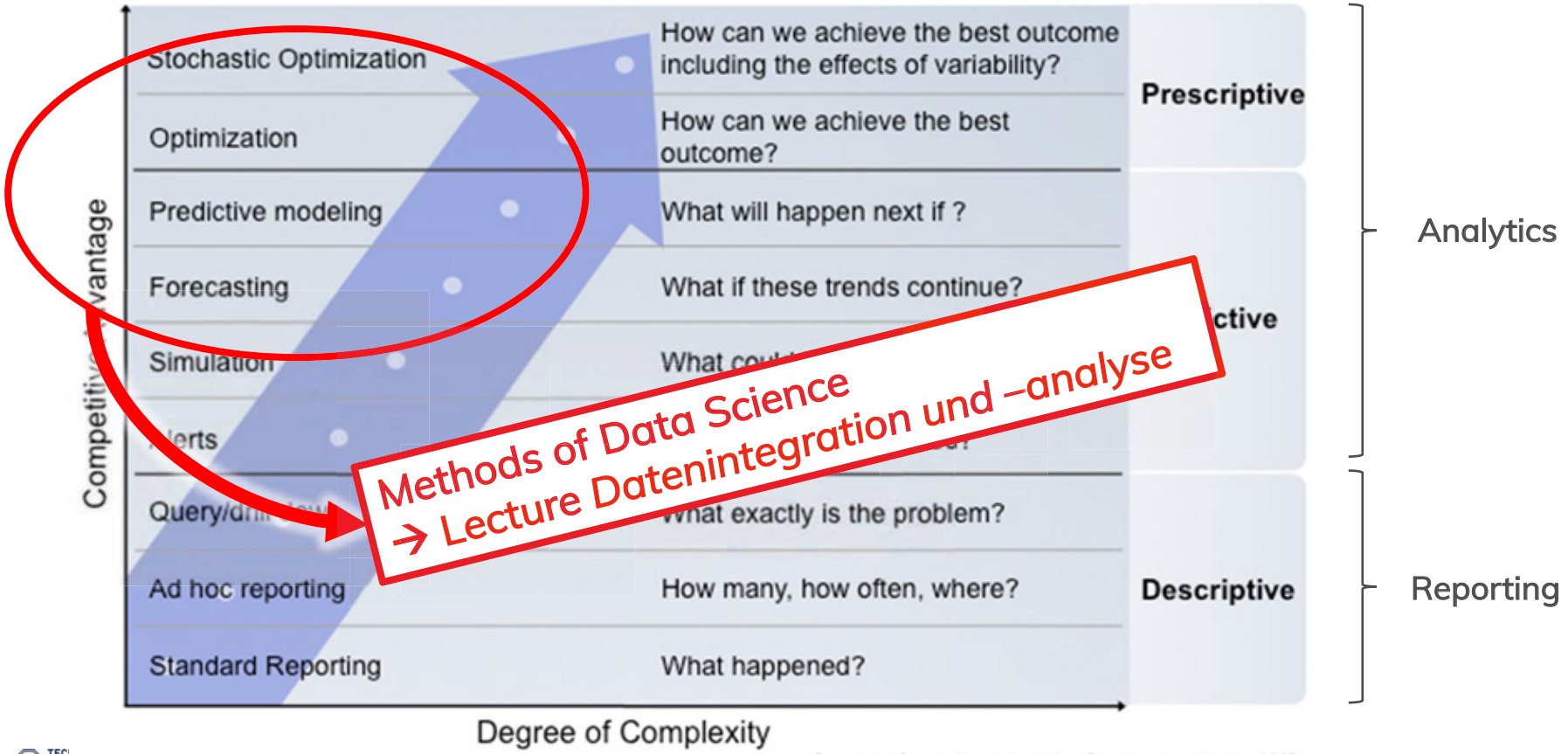
Data engineer

- 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 and Research

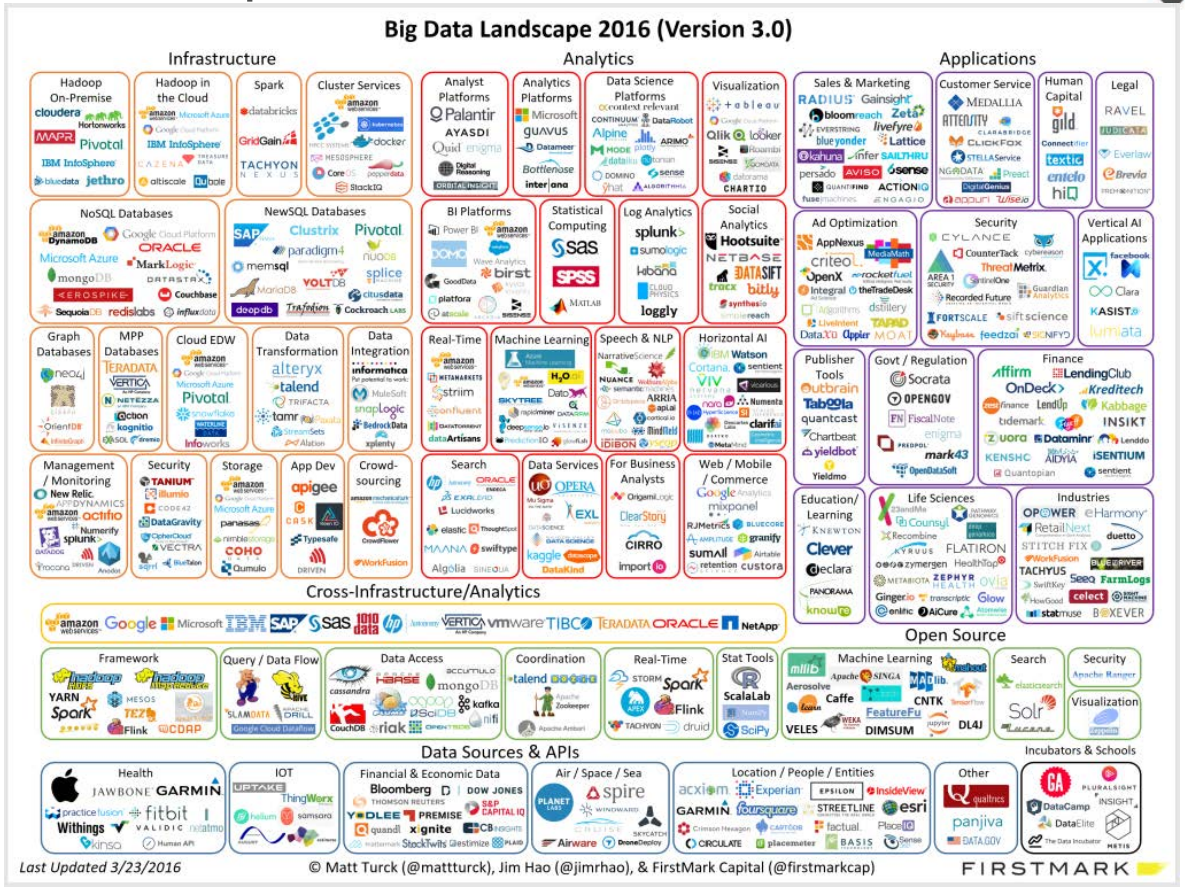
Levels of Analysis

Dresden Database Systems Group



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

Big Data Landscape(s)



Data Systems

Dresden Database Systems Group

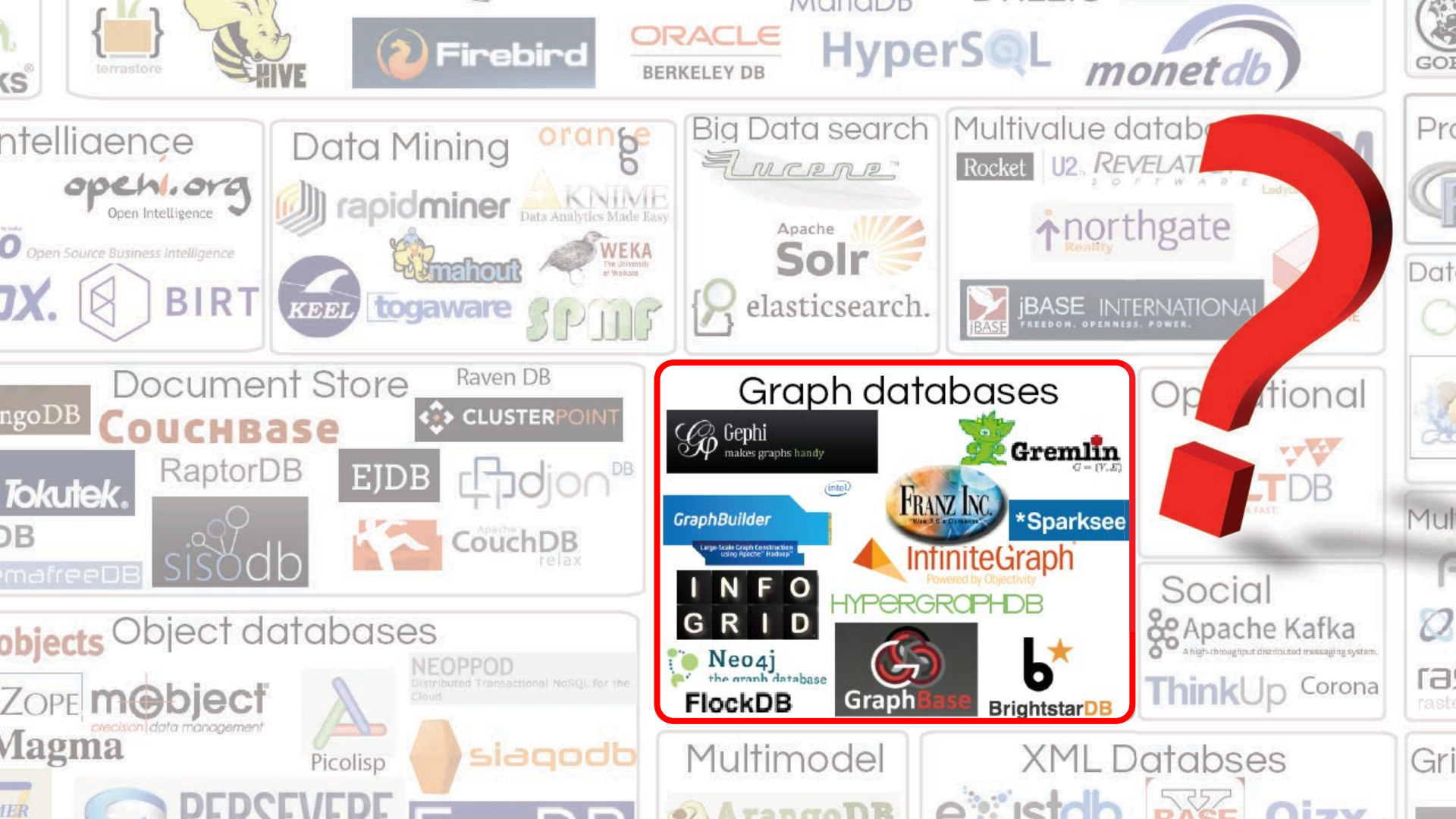
data systems are in the middle of all this

a data system…

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

Different data systems use different data models

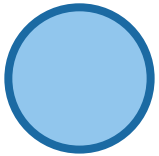
(Big) Data System Landscape(s)


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 databases
Rocket
U2
REVELATION SOFTWARE
northgate
Reality
jBASE
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
Large-Scale Graph Construction using Apache Hadoop
FRANZ INC.
*Sparksee
InfiniteGraph
Powered by Objectivity
INFO GRID
HYPERGRAPHDB
Neo4j
the graph database
GraphBase
BrightstarDB
FlockDB
Operational
LTDB
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
existdb
BASEX

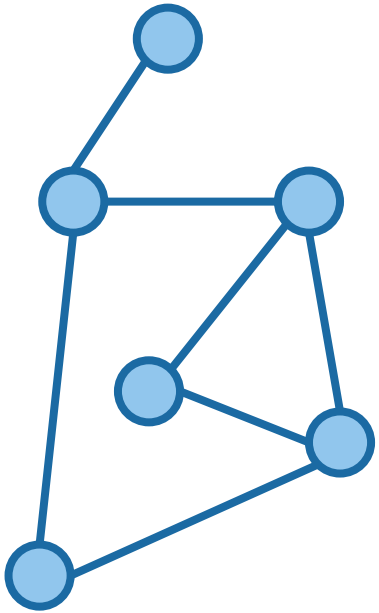
Graph Building Blocks

Dresden Database Systems Group

Nodes (Dots)



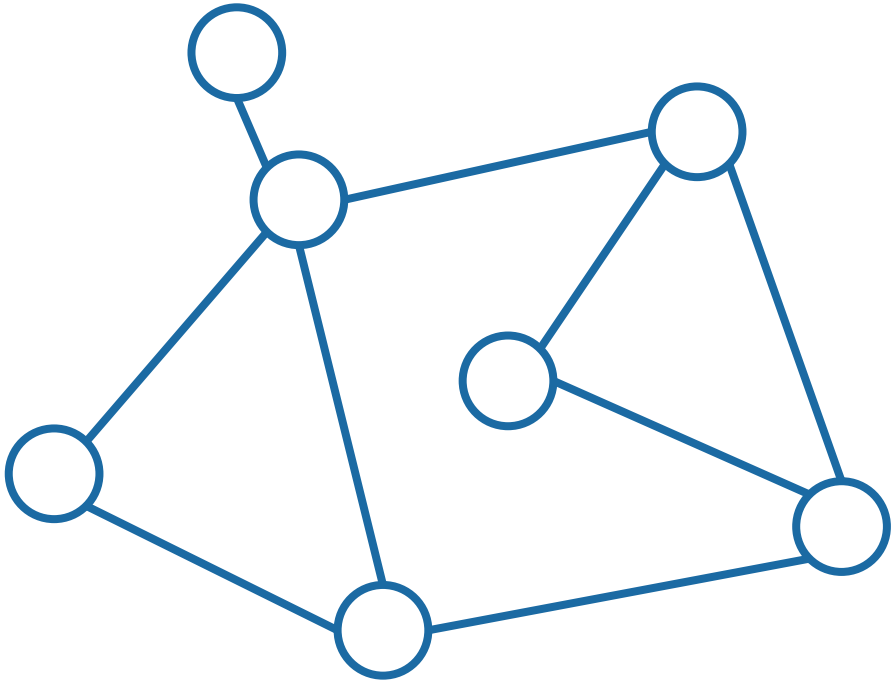
- Like an entity in ER
- Exist on their own
- Have object identity



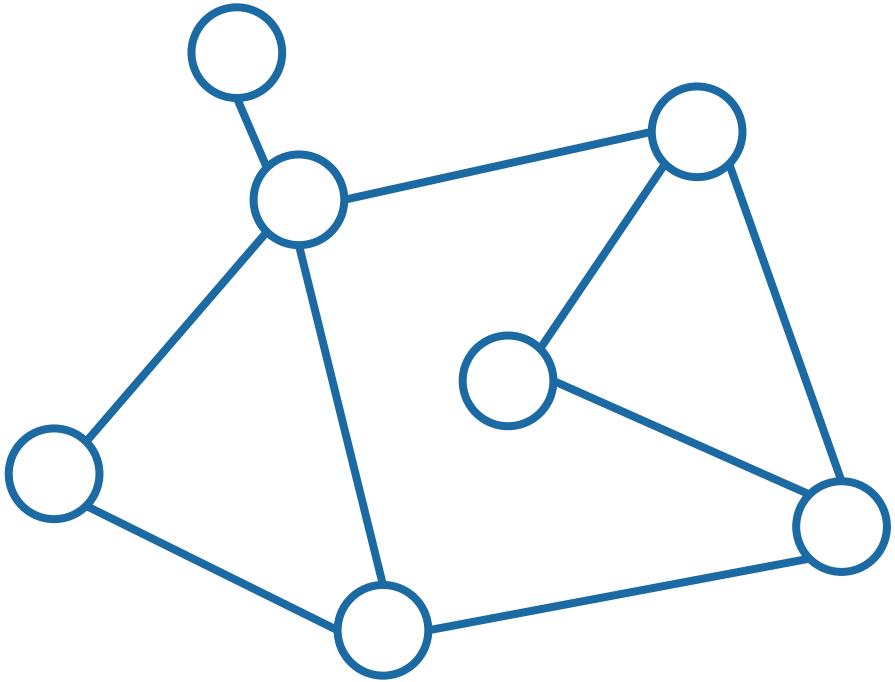
Edges (Lines)

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

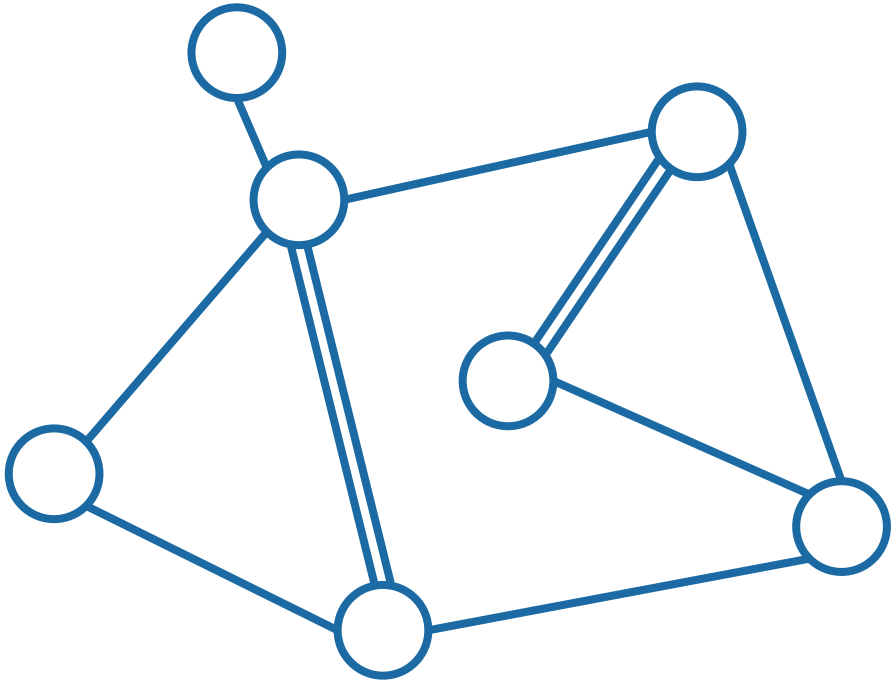
Graph Data – Social Network



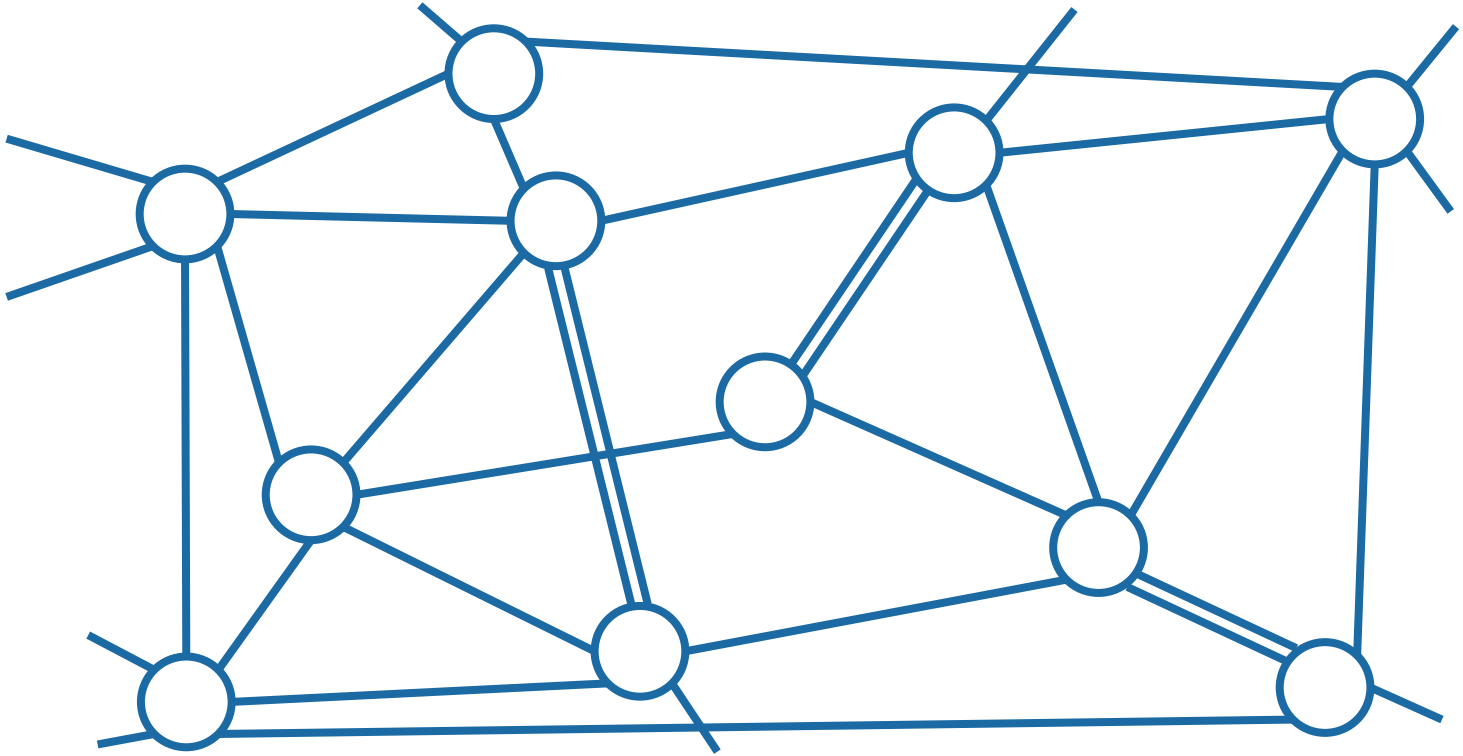
Graph Data – Social Network



Graph Data – Social Network



Graph Data



Social Graphs

Dresden Database Systems Group

Facebook

- May 2013

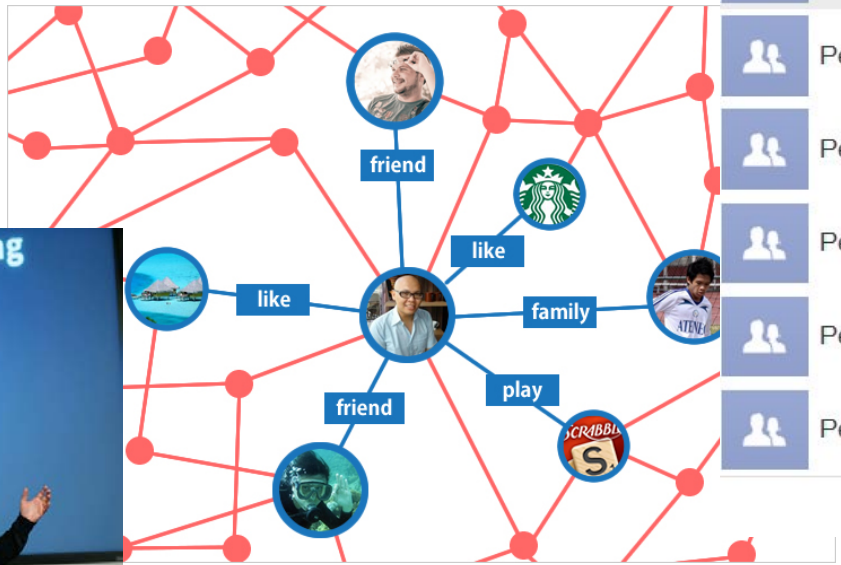
As of 12/2014: "1.39 billion active users with more than 400 billion edges"
[Ching et al. One Trillion Edges: Graph Processing at Facebook-Scale. PVLDB (12)2015]

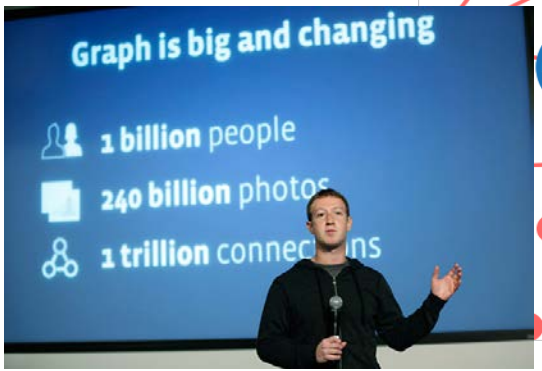
Structured Search

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

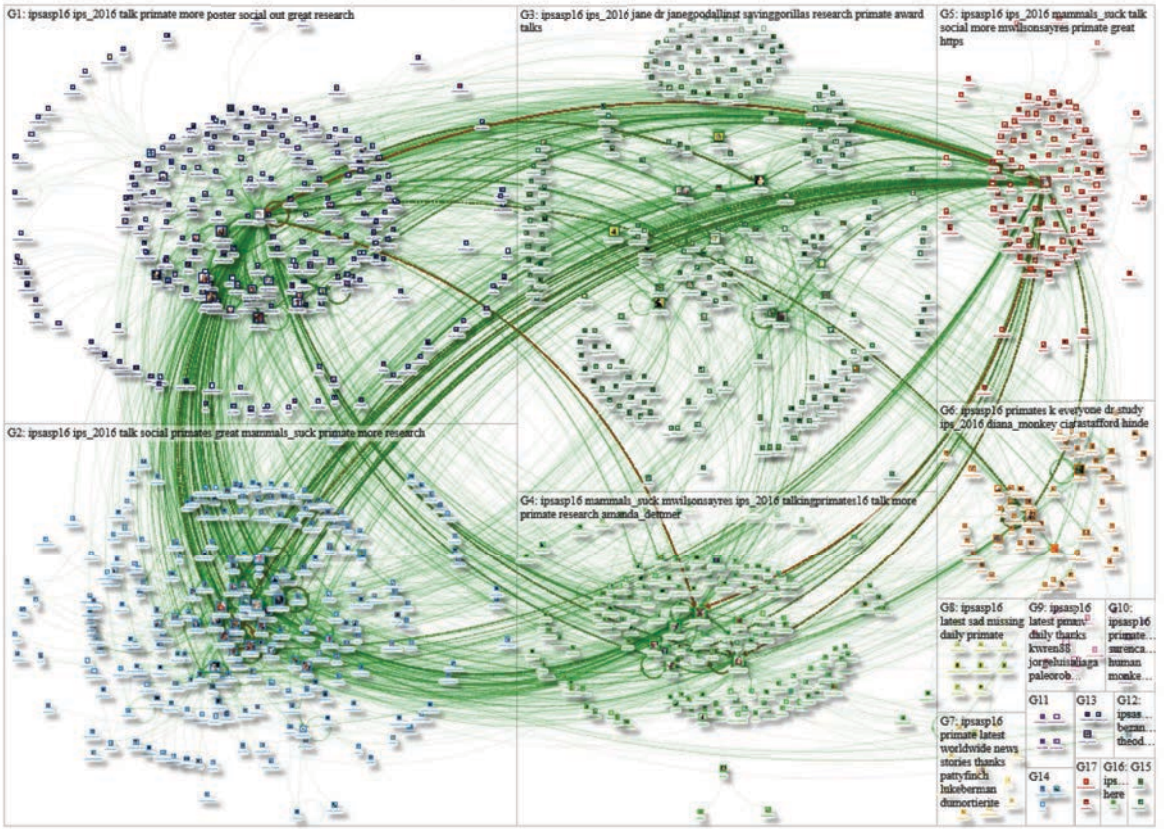
IPS ASP CHICAGO 2016

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



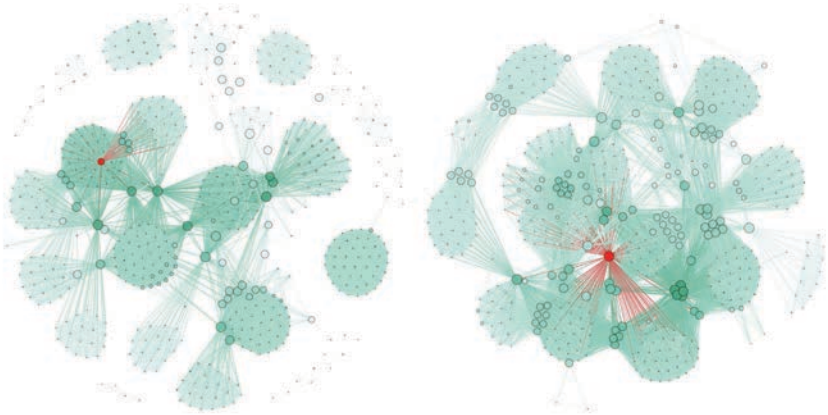
[https://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=76277]

Citation Networks

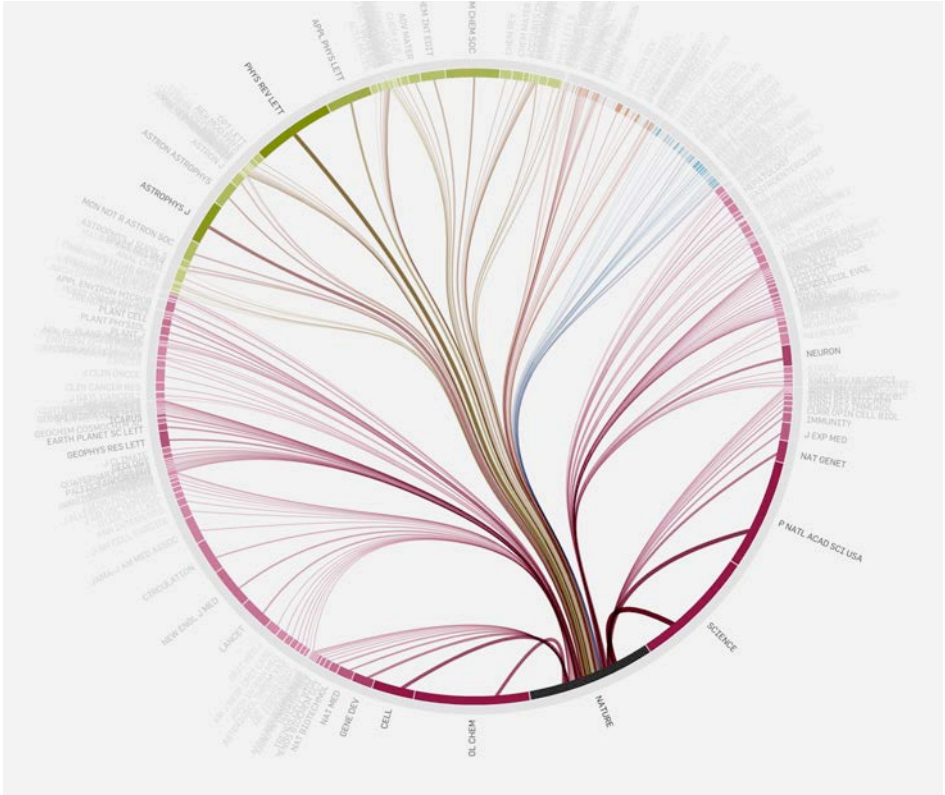
Dresden Database Systems Group

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]



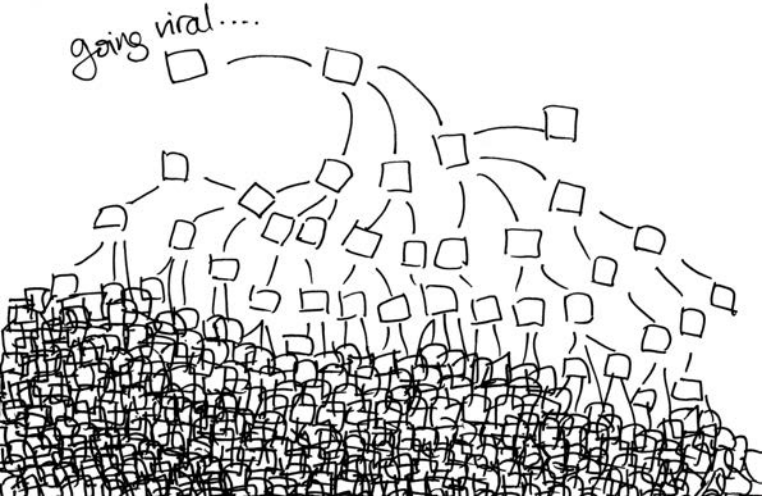
[http://well-formed.eigenfactor.org/projects/well-formed/radial.html#/]

Viral Marketing

Dresden Database Systems Group

Viral Marketing

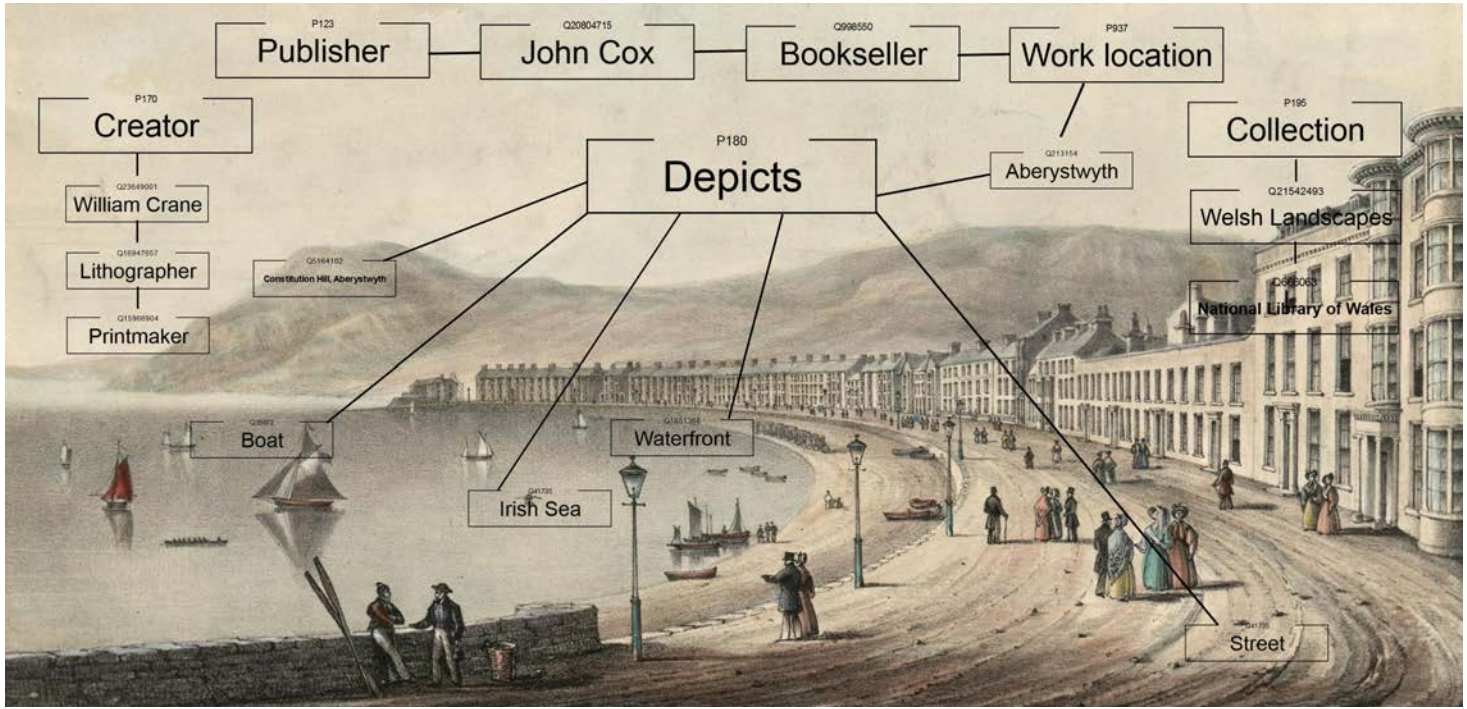
- spreading content to one person so that more than one person engaging with the content
- Techniques
  - Influence estimation
  - Influence maximization



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

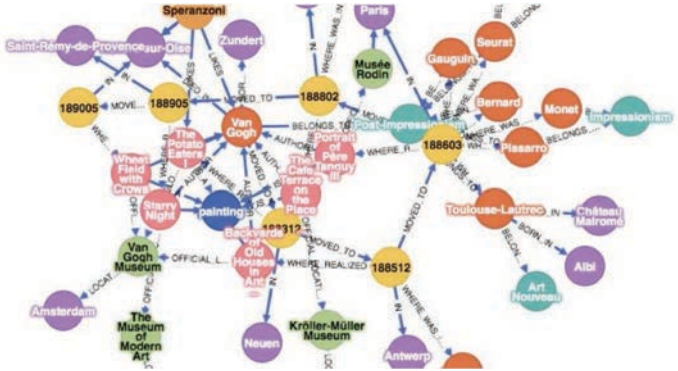
SLUB
Wir führen Wissen.

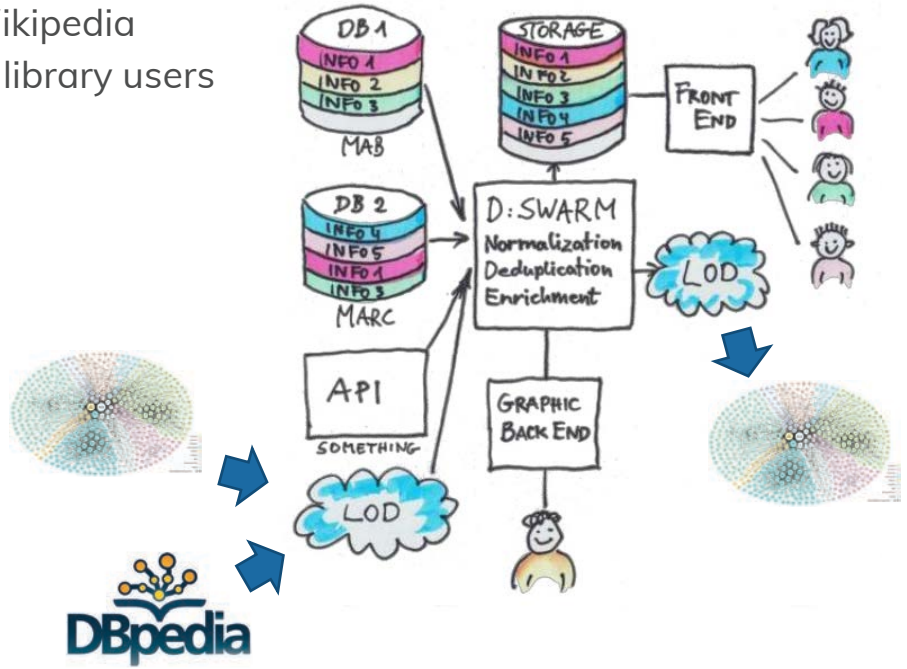
Avantgarde Labs

Dresden Database Systems Group

Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB)

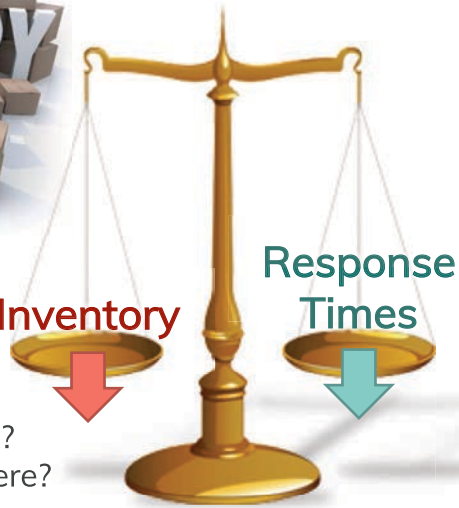
- Adds semantics search to library online catalog
- Utilizing multi-lingual knowledge data from Wikipedia
- Significant improvements in search quality for library users





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

Supply Chain Management

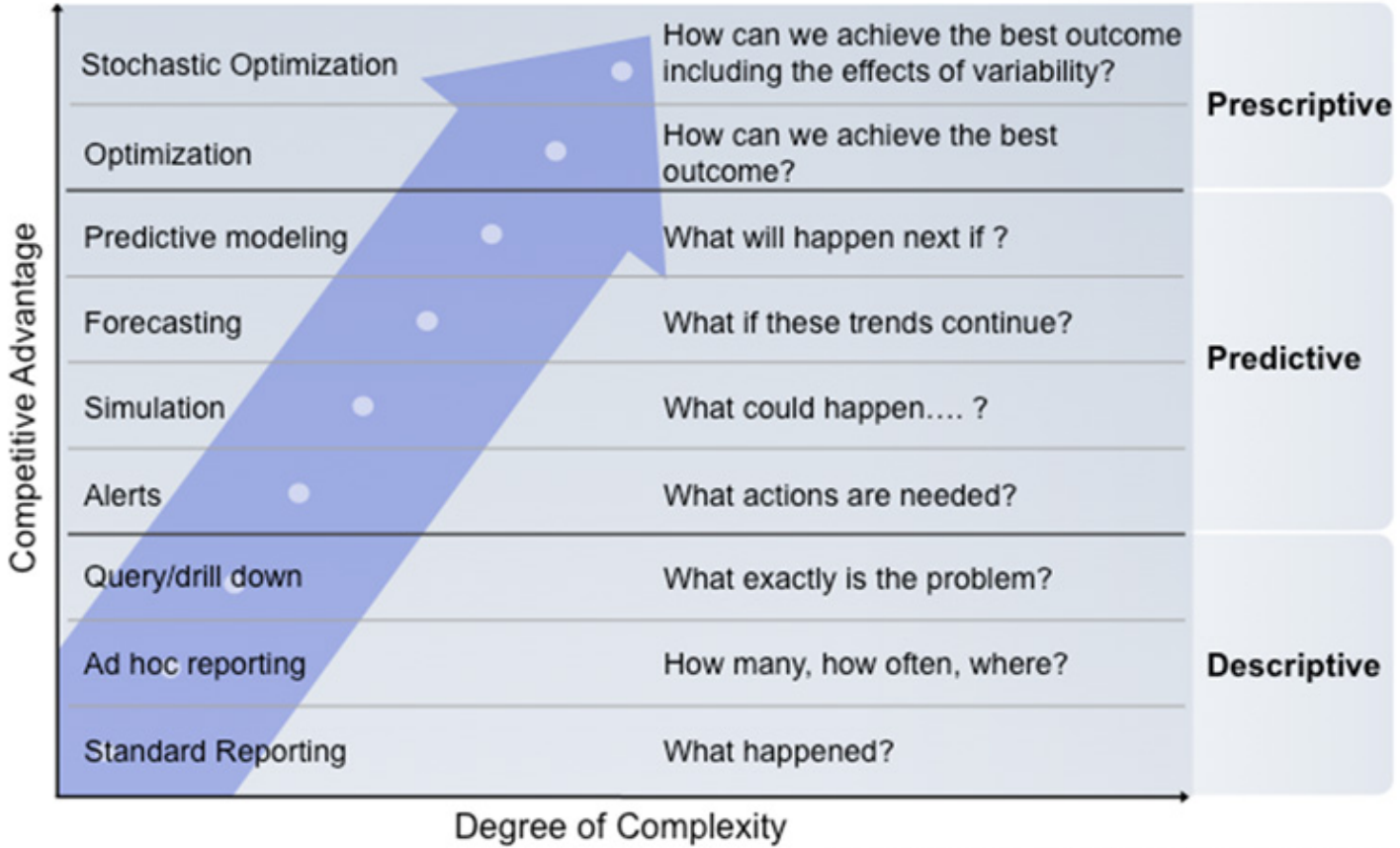


When to send?
How much to send?
From where to where?

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

The Power of Networks

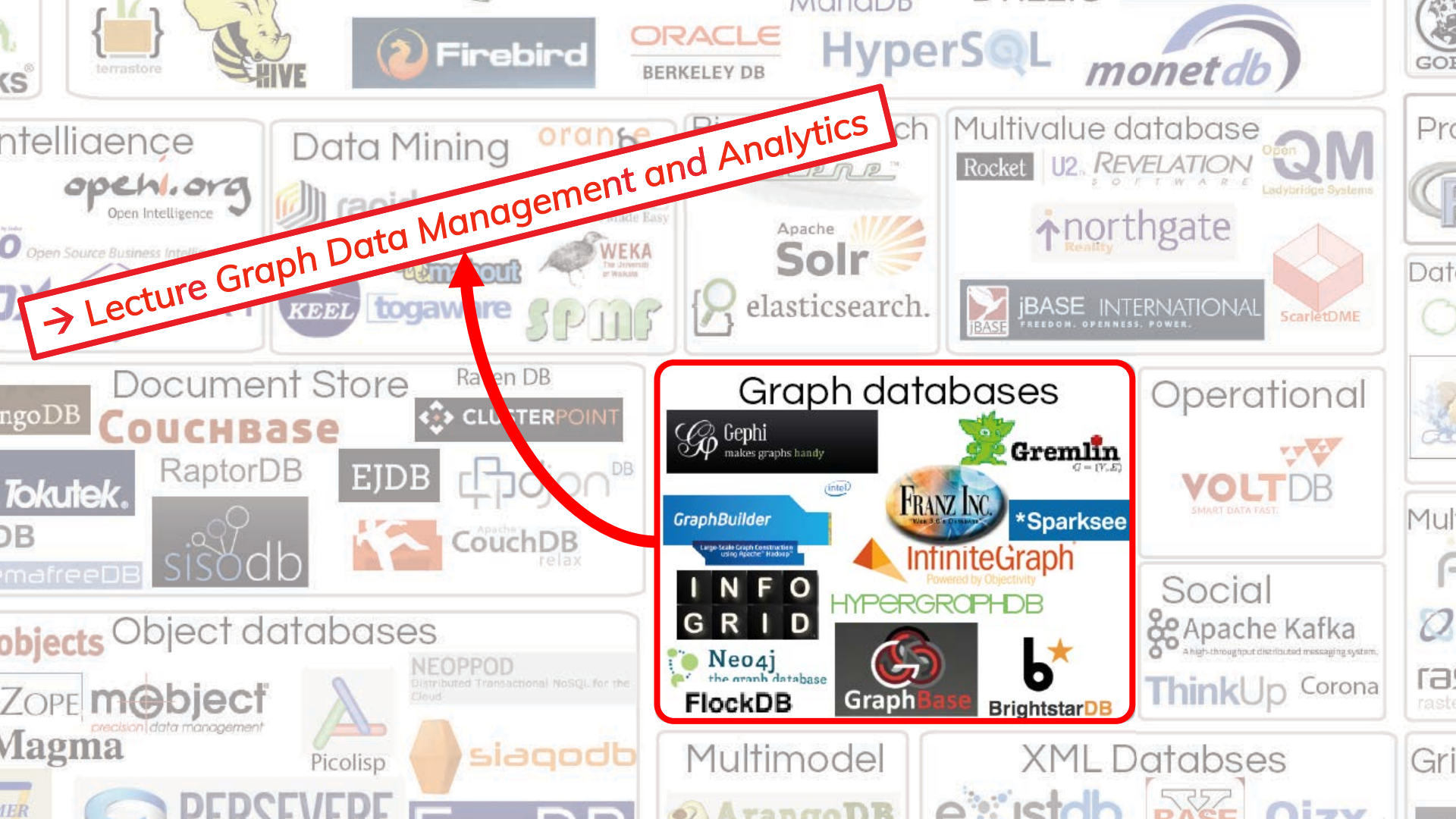
Dresden Database Systems Group

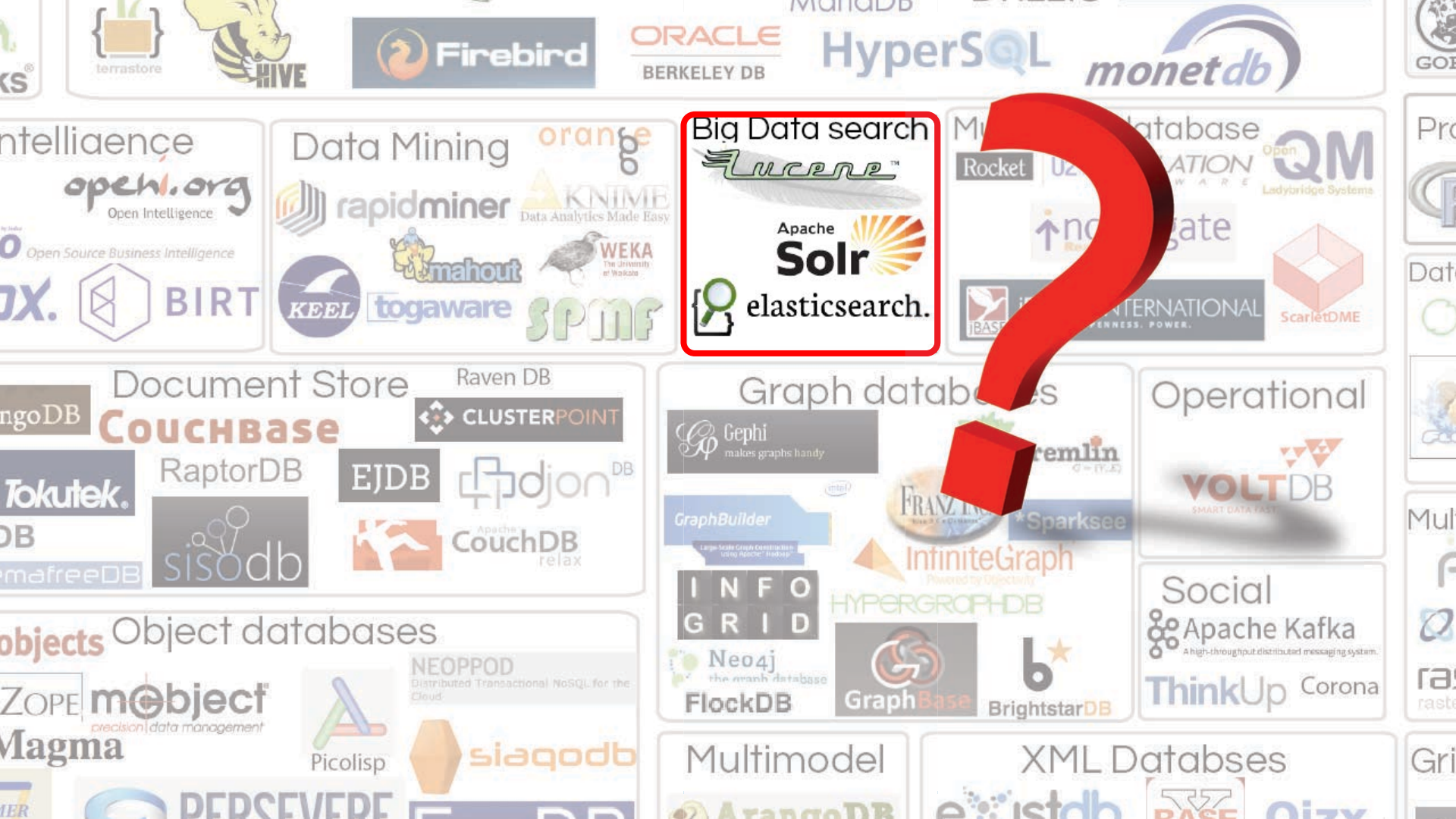
CognitiveMedia **Manuel Lima, The Power of Networks,** London 8.12.2011

Lecture given at the RSA, London on the 8th December, 2011. Animated May, 2012

Manuel Lima explores the way in which we are all connected to the world around us, and how this notion has adapted over the years. From trees of knowledge to the Web of Life to modern social networking, Manuel unearths all the ways life is linked in this RSA Animate.

Watch on YouTube: https://youtu.be/nJmGrNdJ5Gw


→ Lecture Graph Data Management and Analytics
Graph databases
Gephi
makes graphs handy
Gremlin
FRANZ INC.
GraphBuilder
*Sparksee
InfiniteGraph
Powered by Objectivity
INFO GRID
HYPERGRAPHDB
Neo4j
the graph database
FlockDB
GraphBase
BrightstarDB
Document Store
CouchBase
RaptorDB
EJDB
CouchDB
relax
sisodb
Object databases
ZOPE
Magma
Picolisp
siaqodb
Multivalue database
Rocket
U2
REVELATION
QM
northgate
jBASE INTERNATIONAL
ScarletDME
Data Mining
WEKA
KEEL
togaware
spmf
Intelligence
openi.org
Open Intelligence
Apache Solr
elasticsearch.
Operational
VOLTDB
Social
Apache Kafka
ThinkUp
Corona
Multimodel
XML Databases
Raven DB
CLUSTERPOINT
Tokutek
siaqodb


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.
Rocket
QM
Ladybridge Systems
ScarletDME
Document Store
Raven DB
CouchBase
CLUSTERPOINT
Tokutek
RaptorDB
EJDB
djon DB
sisodb
Apache CouchDB relax
Graph databases
Gephi
makes graphs handy
GraphBuilder
FRANZ
Sparksee
InfiniteGraph
INFO GRID
HYPERGRAPHDB
Neo4j
the graph database
FlockDB
GraphBase
BrightstarDB
Operational
VOLTDB
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
PERSEVERE
Multimodel
ArangoDB
XML Databases
exist db
BASE X
Qizx

Use Cases of Search

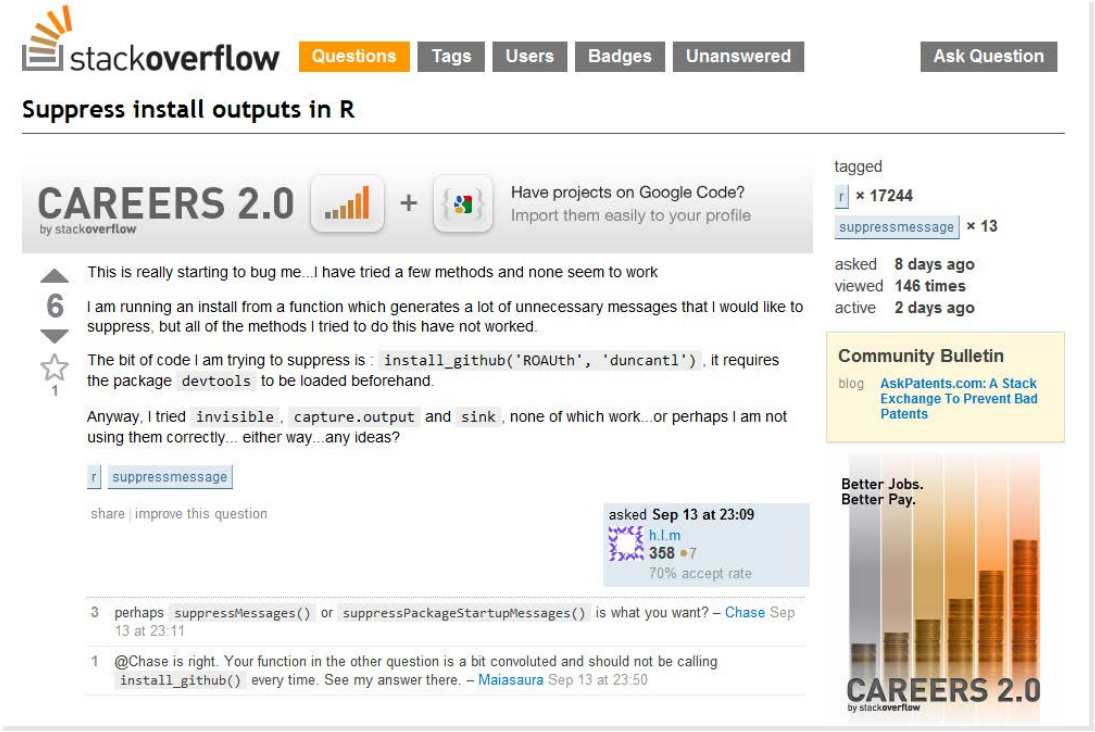
Dresden Database Systems Group

Searching Textual Contents

Use Cases of Search

Dresden Database Systems Group

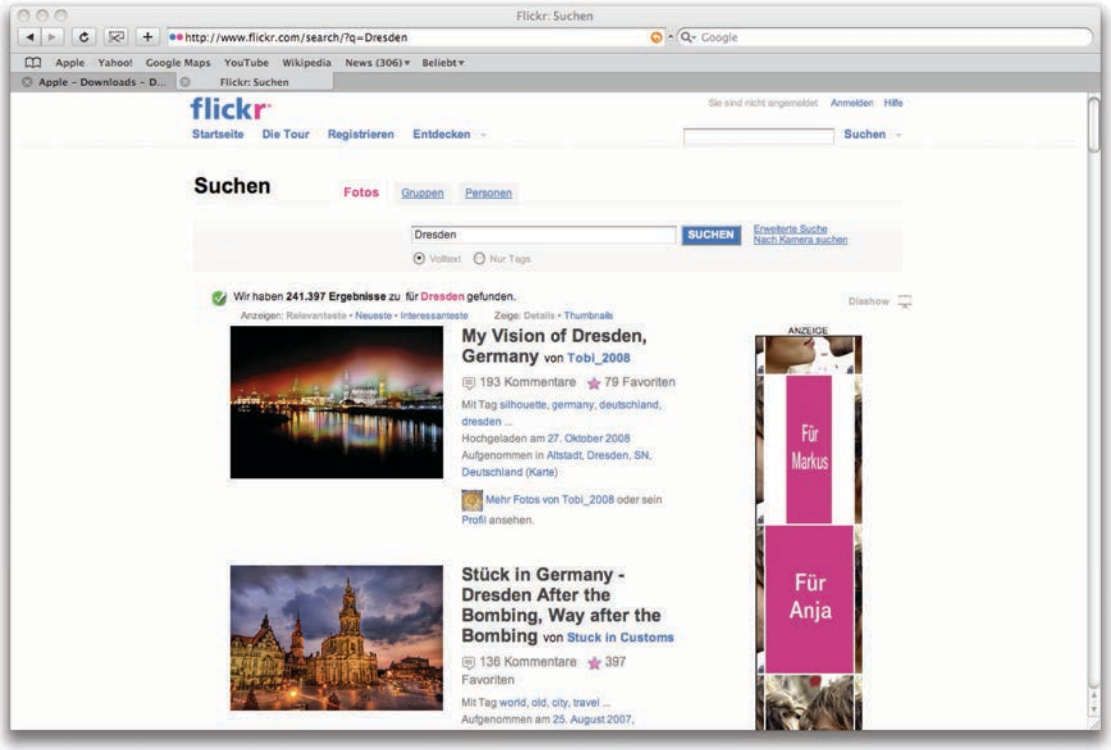
Stack Overflow → Programming Q & A site



Use Cases of Search

Dresden Database Systems Group

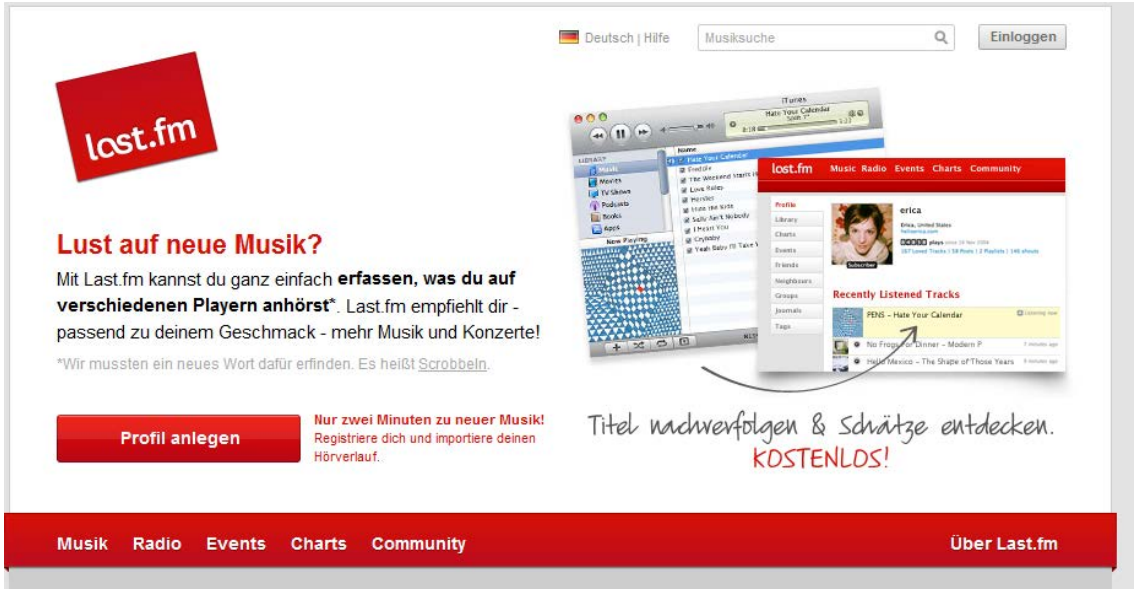
Image Search



Use Cases of Search

Dresden Database Systems Group

Last.fm → Music Search



Example: Library of Congress

Dresden Database Systems Group

FACTS AT A GLANCE

In fiscal year 2008, the Library of Congress:

- Welcomed more than 1.6 million onsite visitors
- Provided reference services to 545,084 individuals in person, by telephone and through written and electronic correspondence
- Recorded a total of 141,847,810 items in the collections:
  - 21,218,408 cataloged books in the Library of Congress classification system
  - 11,599,606 books in large type and raised characters, incunabula (books printed before 1501), monographs and serials, music, bound newspapers, pamphlets, technical reports and other printed material
  - 109,029,796 items in the nonclassified (special) collections, including:
    - 3,005,028 audio materials, such as discs, tapes, talking books and other recorded formats
    - 62,778,118 manuscripts
    - 5,357,385 maps
    - 16,086,572 microforms
    - 5,674,956 pieces of printed sheet music
    - 14,388,175 visual materials, as follows:
      - 1,207,776 moving images
      - 12,536,764 photographs
      - 98,288 posters
      - 545,347 prints and drawings
- Circulated more than 22 million disc, cassette and braille items to more than 500,000 blind and physically handicapped patrons
- Registered 232,907 claims to copyright
- Completed 871,287 research assignments for Congress through the Congressional Research Service
- Prepared 1,529 legal research reports for Congress and other federal agencies through the Law Library
- Recorded more than 85 million visits and 610 million page views on the Library's website. At year's end, the Library's online historical collections contained 15.3 million digital files
- Employed a permanent staff of 3,637 employees
- Operated with a total fiscal 2008 appropriation of $613,496,414, including the authority to spend $50,447,565 in receipts

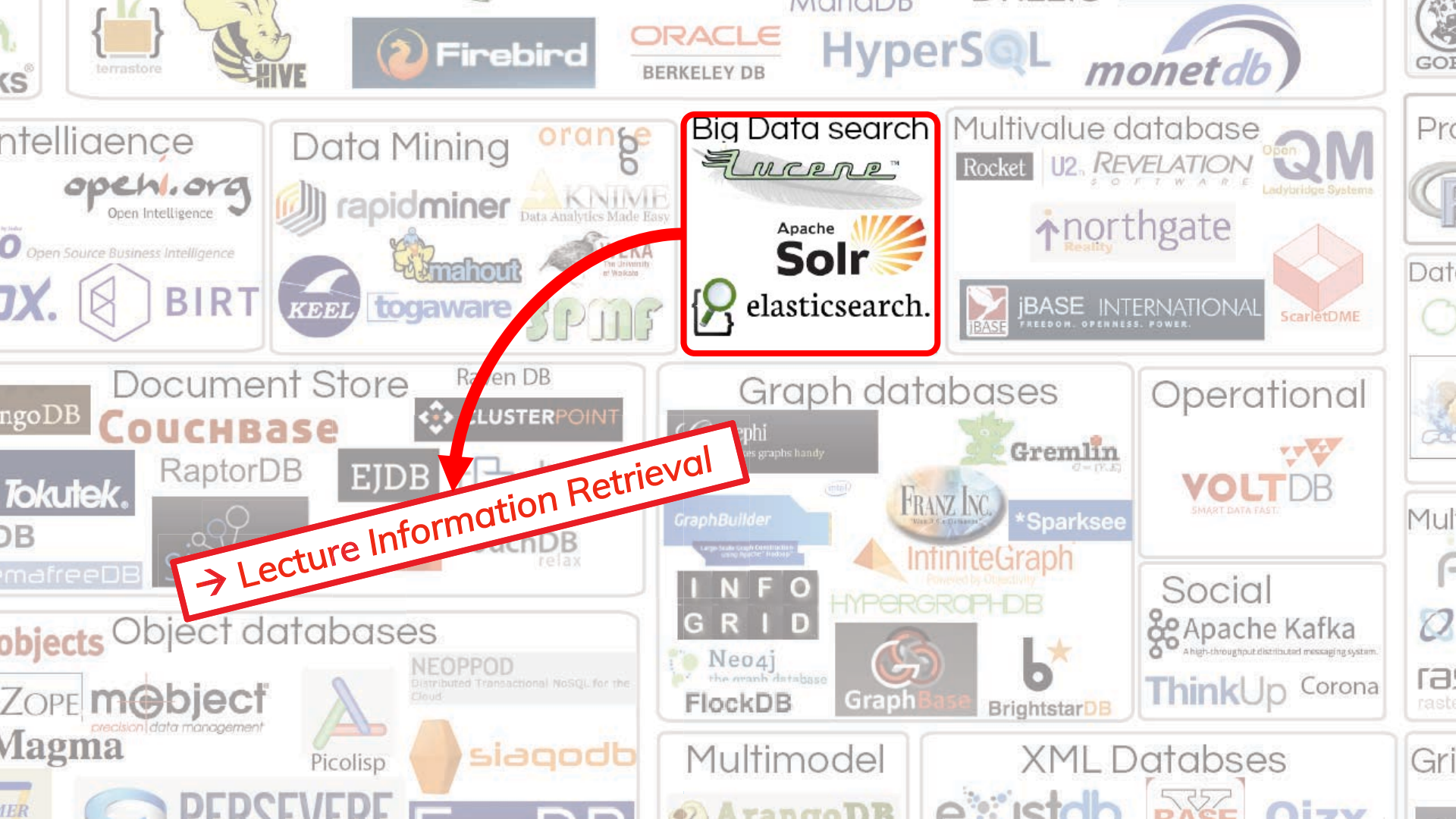
The Great Hall of the Thomas Jefferson Building, Library of Congress

Databases versus Information Retrieval

Dresden Database Systems Group

	Database	Information Retrieval
Matching	Exact Match	Partial Match, Best Match
Model	Deterministic	Probabilistic
Query Language	Structured / Formal	Natural
Query Specification	Complete	Incomplete
Queried Objects	Matching	Relevant
Error Sensitivity	Sensitive	Insensitive

- Hard to formulate Queries
- Iterative workflow base on reponses
- Tons of results, but only a few are relevant
- Ranking of results (instead of set of results)
- Representation of document content often inadequate / inacurate


Big Data search
Lucene
Apache Solr
elasticsearch.
Data Mining
rapidminer
KNIME
Data Analytics Made Easy
mahout
KEEL
togaware
SPMF
Multivalue database
Rocket
U2
REVELATION
QM
northgate
jBASE INTERNATIONAL
ScarletDME
Document Store
CouchBase
CLUSTERPOINT
RaptorDB
EJDB
Graph databases
Gremlin
FRANZ INC.
*Sparksee
GraphBuilder
InfiniteGraph
INFO GRID
HYPERGRAPHDB
Neo4j
FlockDB
GraphBase
BrightstarDB
Operational
VOLTDB
Social
Apache Kafka
ThinkUp
Corona
Object databases
ZOPE
mobject
Magma
Picolisp
NEOPPOD
siaqodb
Multimodel
XML Databases
→ Lecture Information Retrieval

Summary

Dresden Database Systems Group

Big Data

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

Related research and lectures

- Data Science → Datenintegration und -analyse
- Data Systems → Architektur von Datenbanksystemen
  → Big Data Platforms
- Graph Data → Graph Data Management and Analytics
- Search → Information Retrieval