JACOBS UNIVERSITY

CO19-320302
Databases and Web Services

Instructors: Peter Baumann, Sebastian Villarroya

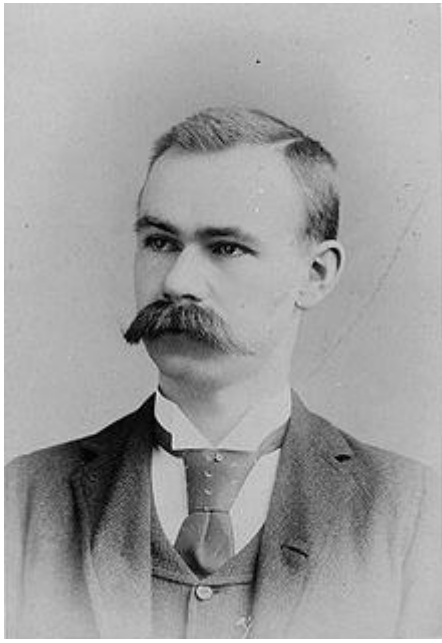
email: p.baumann@jacobs-university.de, s.villarroyafernandez@jacobs-university.de

office: room 88 & 96, Research 1

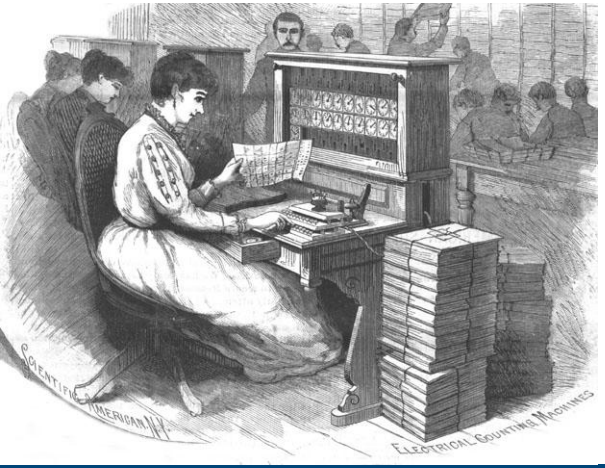
Where It All Started

Source: Wikipedia

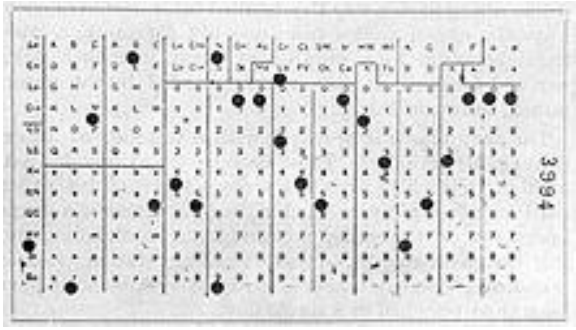
- 1890 census on 62,947,714 US population ← "Big Data"
  - was announced after only six weeks of processing
- Hollerith „tabulating machine and sorter"
- Tabulating Machine Company
  → International Business Machines Corporation



Herman Hollerith in 1888



Hollerith card puncher, used by the United States Census Bureau



Hollerith punched card

What Happens in an **Internet Minute?**

And **Future Growth** is **Staggering**

Today, the number of networked devices = the global population

By **2015**, the number of networked devices = 2x the global population

In **2015**, it would take you **5 years** to view all video crossing IP networks each **second**

intel

What Is „Big Data“?

- Internet: the unprecedented information collector
  - May 2012: 200m Web servers [Yahoo]
  - estd 50+b static pages [Yahoo]
  - 40 b photos [Facebook]
  - 2012: 31b searches/m [Google]
- 2.8 Zettabyte generated in 2012. Adding 2.5 PB every day. [Computerwoche]

- Typical Big Data:
  - Business Intelligence
  - Social networks - Facebook, Twitter, GPS, ...
  - Life Science: patient data, imagery
  - Geo: Satellite imagery, weather data, crowdsourcing, ...
    - *Petrol industry: „more bytes than barrels“*

http://www.sgi.com/go/twitter/#heatmaps

Today: „Data Deluge“

JACOBS UNIVERSITY

- „It is estimated that a week‘s work at the New York Times contains more information than a person in the 18th Century would encounter in their entire lifetime and the thought is that within 10 years the rate of information doubling will occur every 72 hours.“ -- P. „Bud“ Peterson, U Colorado
- “global mobile data traffic 597 petabytes per month in 2011 (8x the size of the entire global Internet in 2000) estimated to grow to 6,254 petabytes per month by 2015” -- Forbes, June 2012
- a typical new car has about 100 million lines of code
  - -- http://www.wired.com/autopia/2012/12/automotive-os-war/

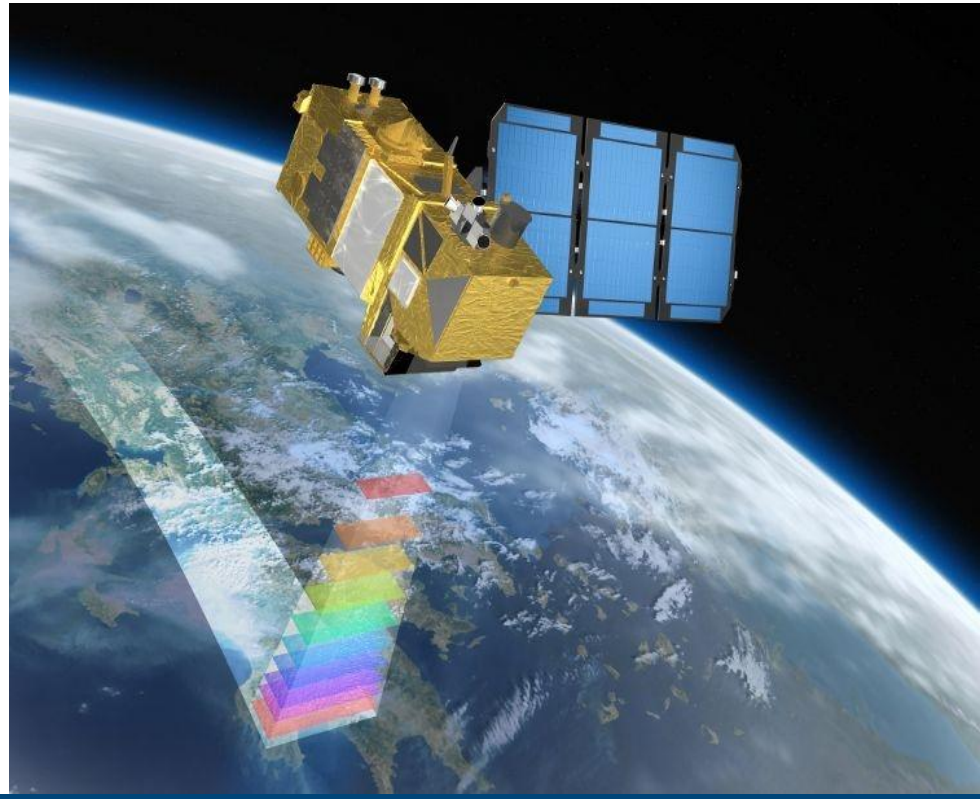
Big Data in Business

[Wikipedia]

- Walmart: more than 1 million customer transactions every hour; imported into databases estimated to contain more than 2.5 PB of data
  - =167 times all books in the US Library of Congress
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide
- Estd.: business data worldwide x2 every 1.2 years

Big Data in Geo: Satellite Imagery

- 100s of Exabytes expected for 2020
- ngEO: planning for 10^12 satellite images under curation of ESA [ESA]
  - Increased # of instruments flying
    - *A-Train, Landsat, Sentinels, ...*
  - Increased spectral resolution: 5 (Landsat) to 250 (ALI/Hyperion)
  - Increased spatial resolution: meters
- NASA EOSDIS: 5 TB / day

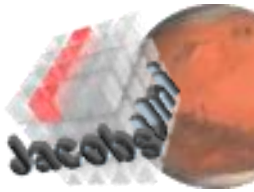


Big Data in Geo: Datacubes

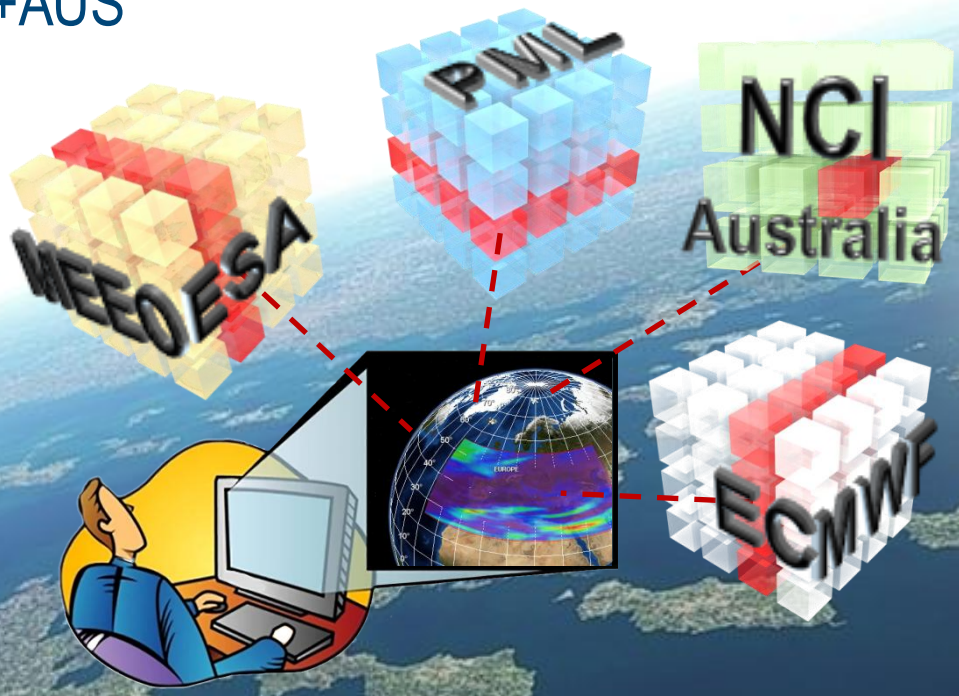
JACOBS UNIVERSITY

EarthServer: Agile Analytics on 2.5+ Petabyte space/time datacubes

- Earth Science (3D sat image timeseries, 4D weather)
- Planetary Science

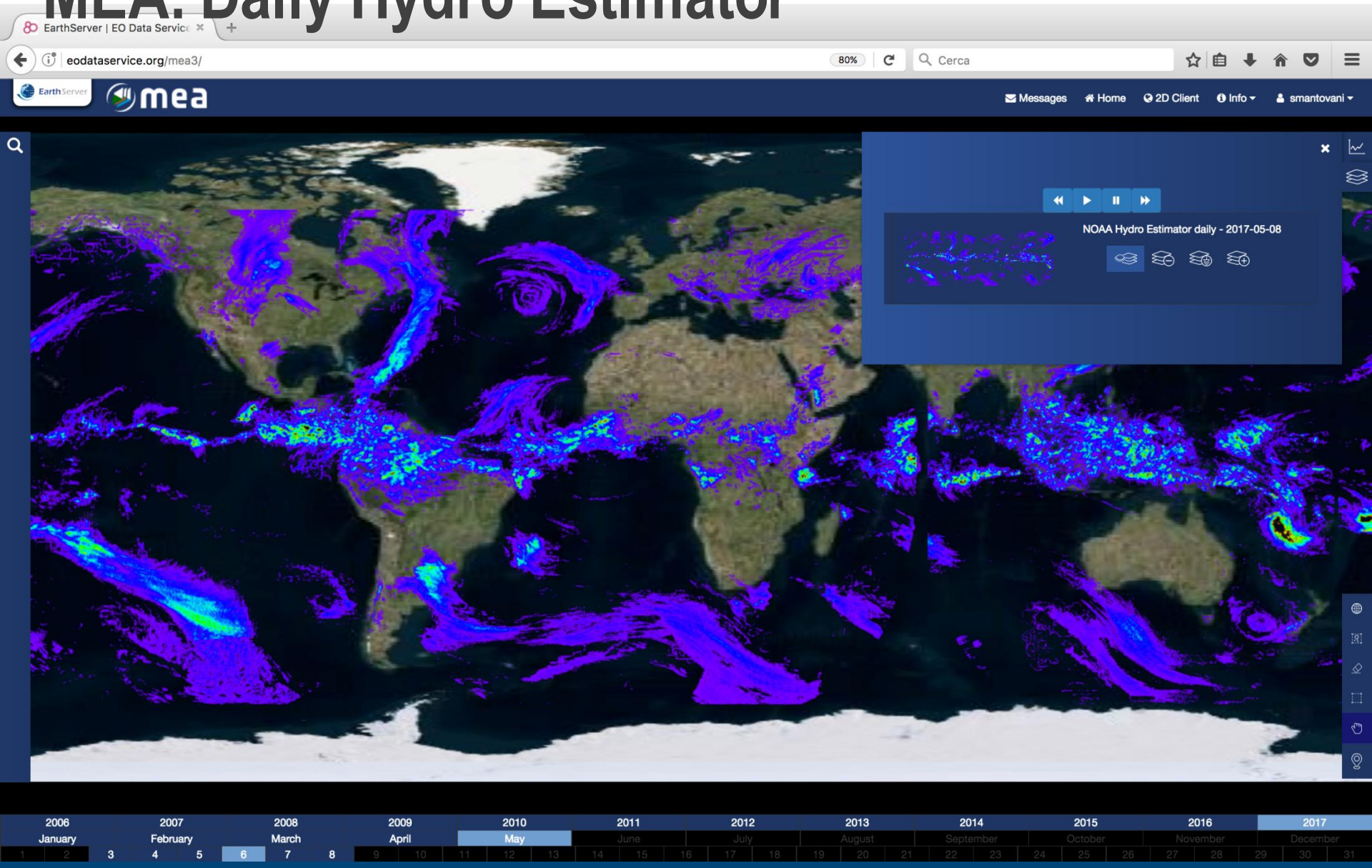


- Intercontinental initiative: EU+US+AUS
- EU rasdaman
  + US NASA WorldWind
- www.earthserver.eu,
  www.planetserver.eu

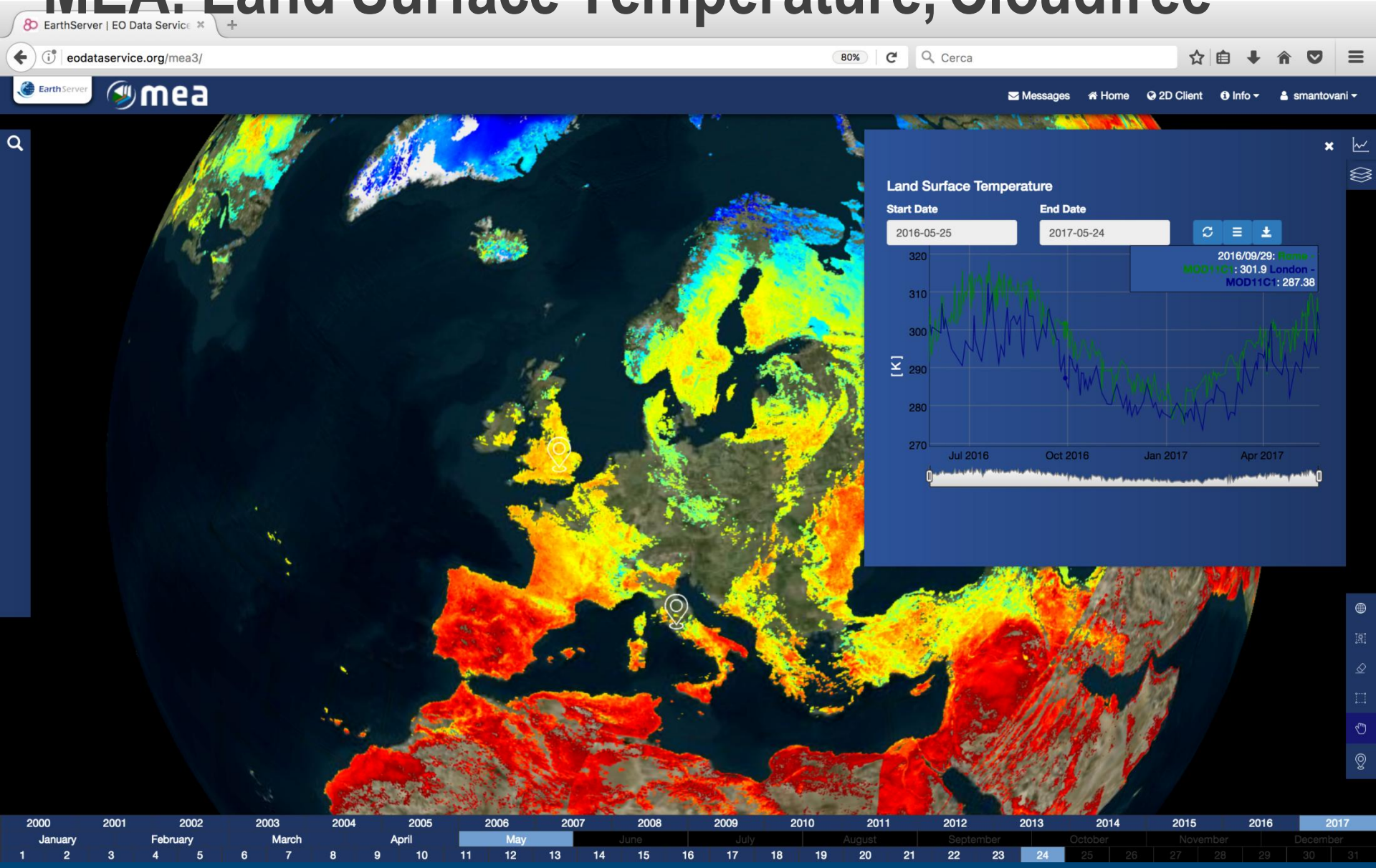


Co-funded by
the European Union

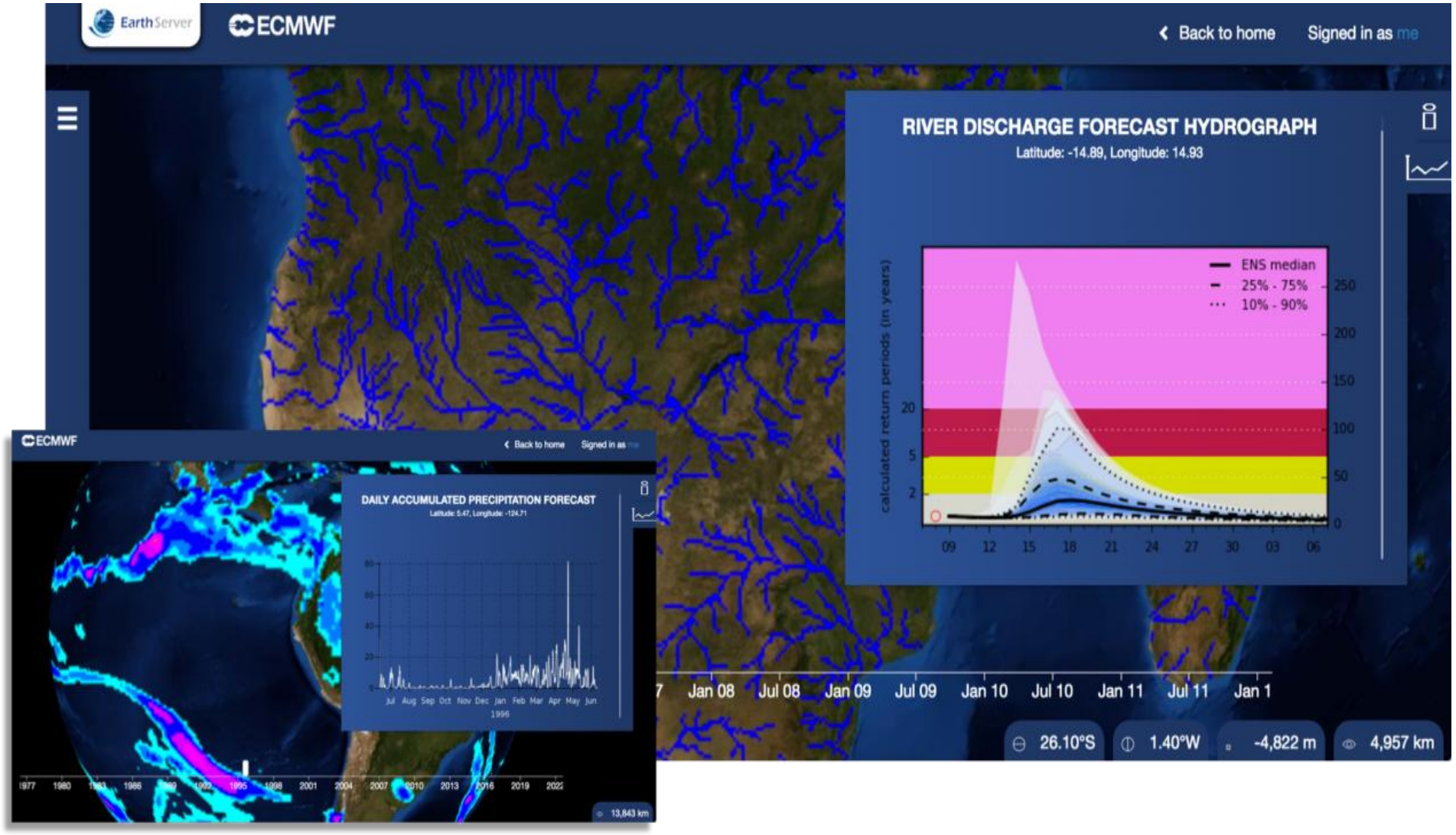
MEA: Daily Hydro Estimator



MEA: Land Surface Temperature, Cloudfree



ECMWF: River Discharge

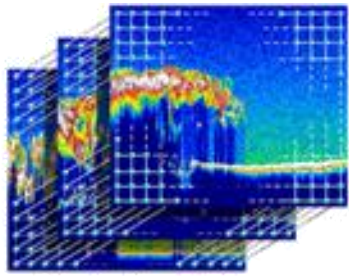


Our Research: Big Datacubes

[gamingfeeds.com]

JACOBS UNIVERSITY

- Large-Scale Scientific Information Services (L-SIS) Research Group
  - flexible, scalable services on massive multi-dimensional arrays



- Main visible results:
  - rasdaman Array DBMS - worldwide in operational use
  - „Big Earth Data“ standards in OGC, ISO, INSPIRE – eg, SQL/MDA
- *If you have rock-solid coding skills, why not join us?*

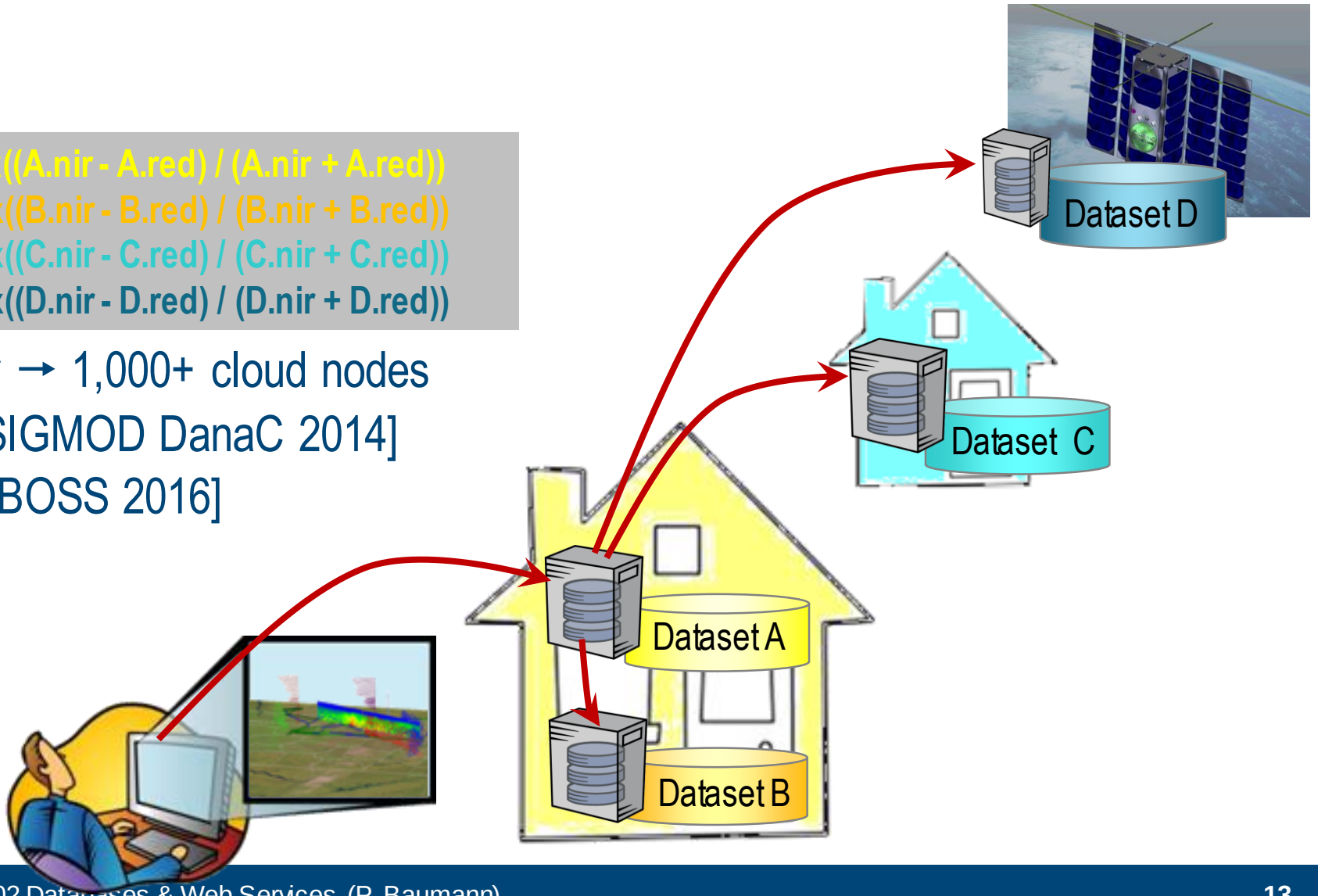
rasdaman Distributed Processing

```
max((A.nir - A.red) / (A.nir + A.red))
- max((B.nir - B.red) / (B.nir + B.red))
- max((C.nir - C.red) / (C.nir + C.red))
- max((D.nir - D.red) / (D.nir + D.red))
```

1 query → 1,000+ cloud nodes

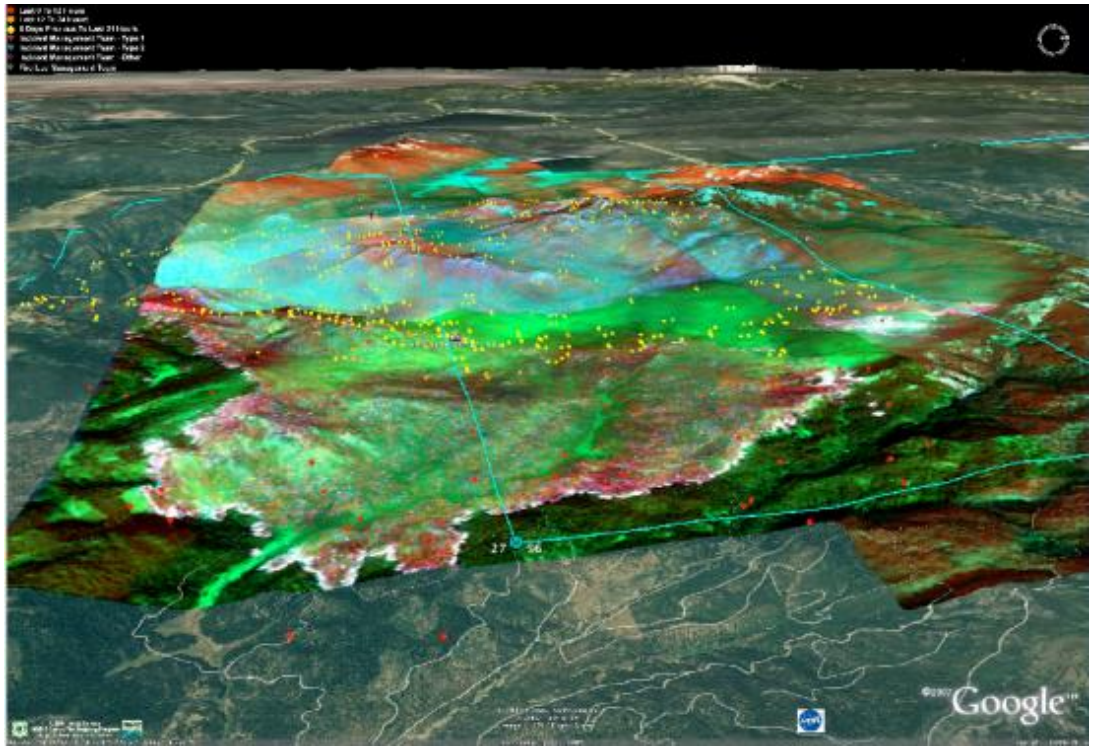
[ACM SIGMOD DanaC 2014]

[VLDB BOSS 2016]



Next: On-Board Query Intelligence

ORBiDANse:
Orbital Big Data Analytics Service



[images: ESA, NASA]

JACOBS UNIVERSITY

...BACK TO THE COURSE

Data Management: The Task

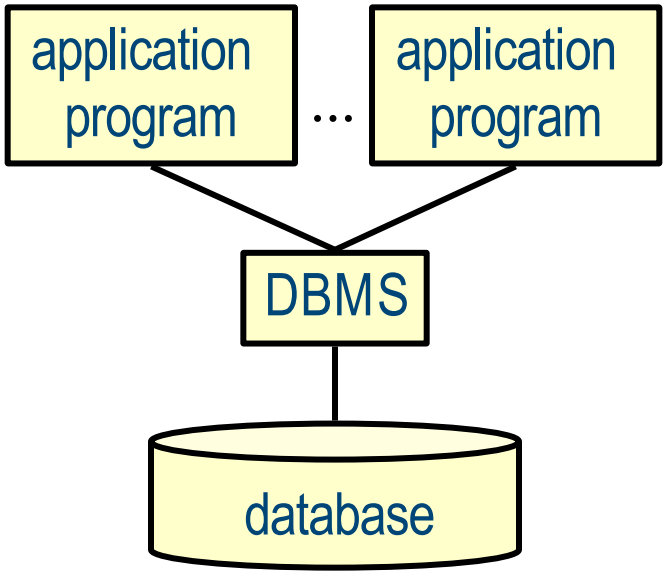
JACOBS UNIVERSITY

- Manifold information,
  accessed by users in manifold (often unanticipated) ways
  - Standard task
  - Many variations
- Solution: individually configurable standard tool

- *...is this marketing speak???*

What Is a Database [System]?

- Database = DB = an integrated collection of data
  - With a well-described structure = schema
- Database [Management] System = DBMS
  = software to store and manage databases
  - …and no one else!
- describes excerpt of real-world enterprise
  - "Universe of Discourse" (UoD), "mini world"
- Example:
  - Entities (students, courses, …)
  - Relationships (Madonna is taking 320301, …)

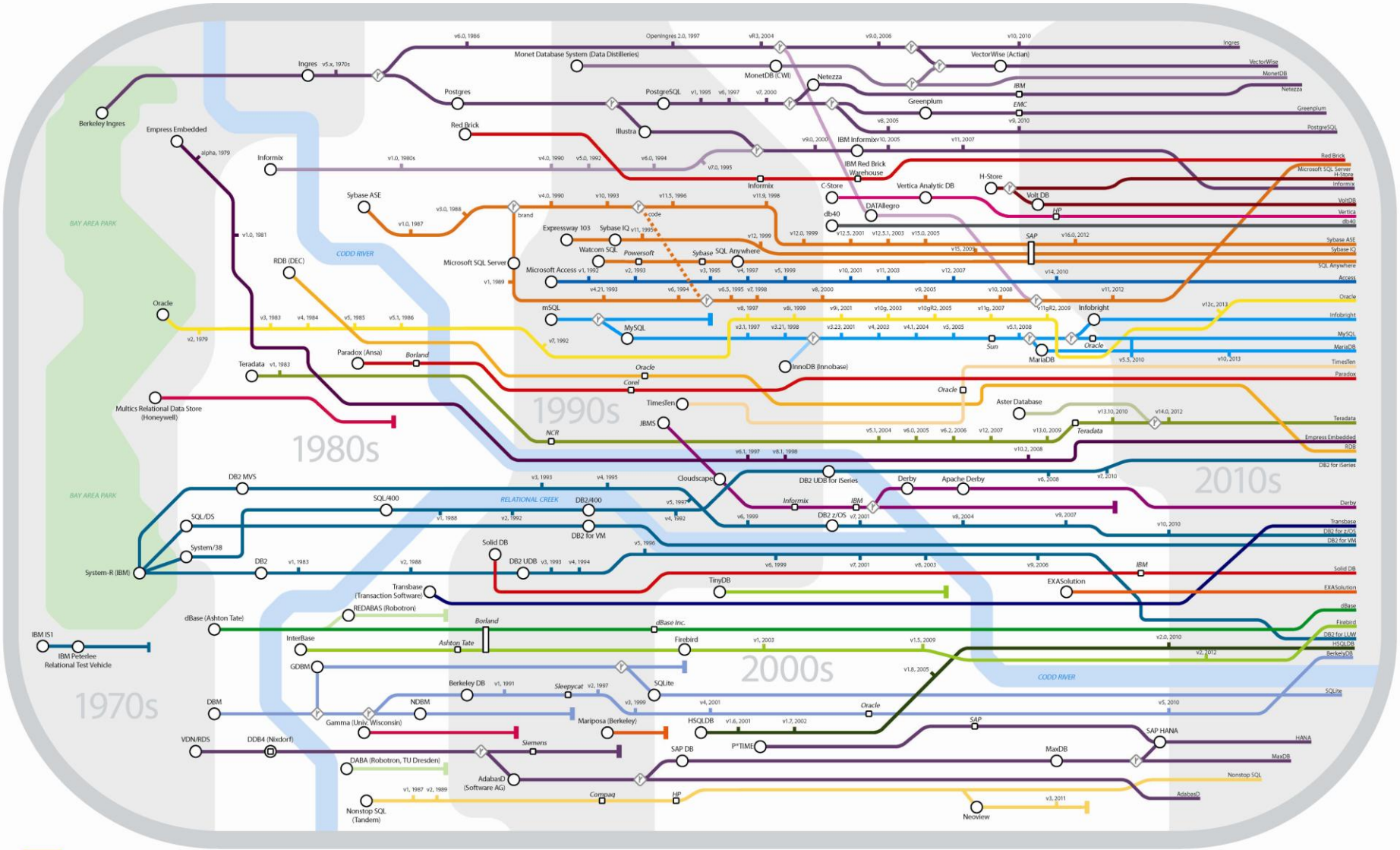


DBMS History

- History:
  - 60s… IMS (hierachical model, for tapes), CODASYL (network model, still tapes)
  - 1974 SEQUEL defined (Chamberlain et al.)
  - 1977 IBM prototype System R; Oracle starts implementation
  - 1979 first Oracle SQL DBMS shipped
  - 1981 IBM ships SQL/DS
  - 1983 IBM introduces DB2
  - 1985 Ingres, Informix switch to SQL
  - 1987 ISO 9075 Database Language SQL
  - 1988 dBASE IV with SQL
  - 1989 ISO SQL-89
  - 1992 ISO SQL-92
  - 1999 SQL:1999 (SQL3): extensibility
  - 2003 SQL:2003

- SQL / relational DBMS dominate
  - Oracle, IBM DB2, Informix, MS SQL Server; MySQL; Postgres; ...
- Key to success: query language
  - Intuitive (hm…)
  - Yet precise, formalised semantics
  - Declarative = abstracts from internals
  - …hence optimizable
- Some Trends
  - Information retrieval = full text databases *Silently integrated*
  - (Object-oriented DBMSs), Object-relational
  - XML databases
  - NoSQL, Array Databases, Graph Databases

Genealogy of Relational Database Management Systems



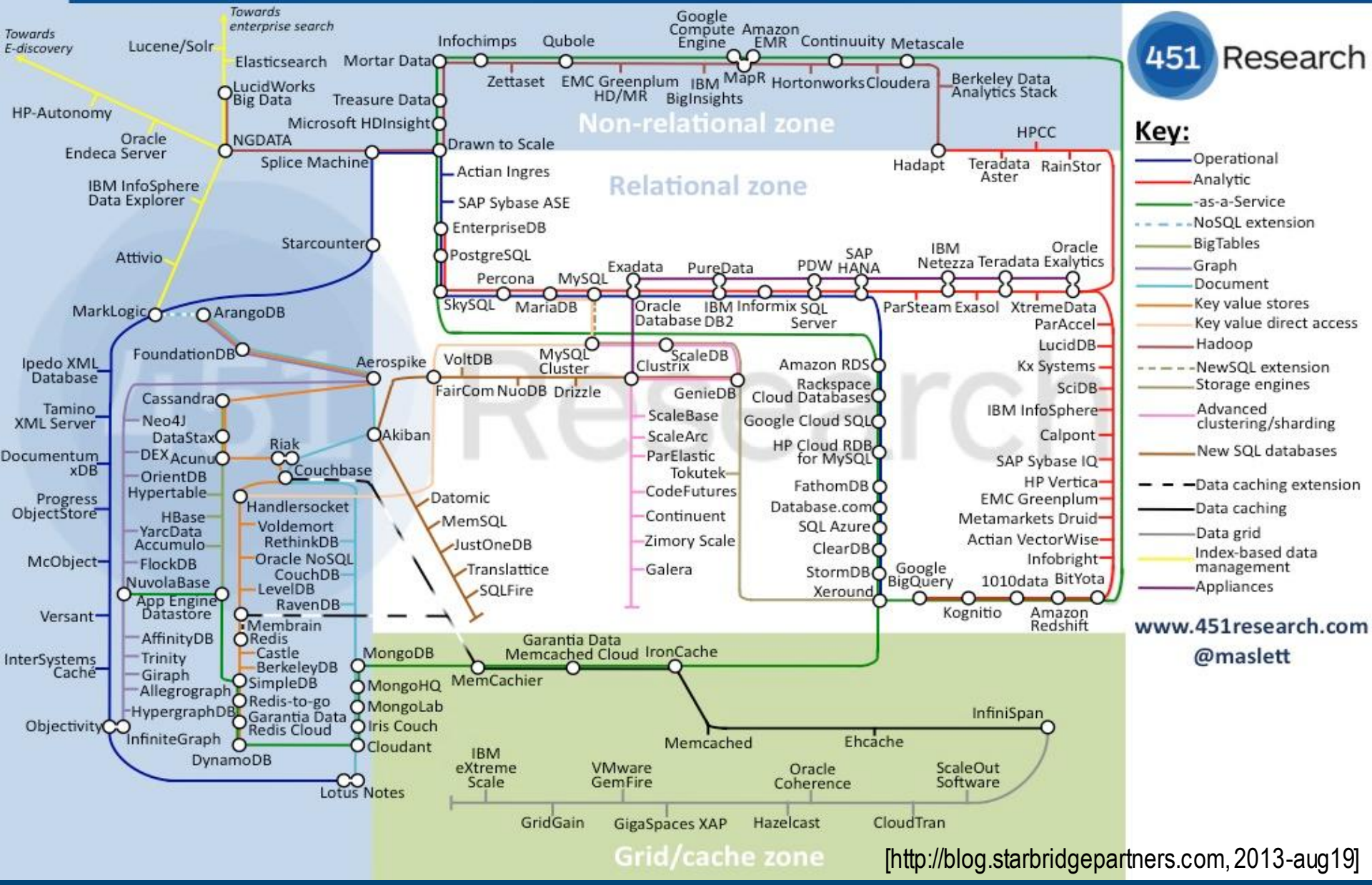
Key to lines and symbols

Publishing Date | Acquisition | Versions | Discontinued | Branch (intellectual and/or code) | Crossing lines have no special semantics

Felix Naumann, Jana Bauckmann, Claudia Exeler, Jan-Peer Rudolph, Fabian Tschirschnitz
Contact - Hasso Plattner Institut, Potsdam, Felix.Naumann@hpi.uni-potsdam.de
Design - Alexander Sandt Grafik-Design, Hamburg

Version 4.0 - August 2013
http://www.hpi.uni-potsdam.de/naumann/projekte/rdbms_genealogy.html

Database Landscape Map – December 2012



[http://blog.starbridgepartners.com, 2013-aug19]

…and Then Came NoSQL

JACOBS UNIVERSITY

www.nosql-database.org

- original intention: modern web-scale databases
  - began early 2009, has grown rapidly
  - Broadened into "Next Generation Databases"
- Fast: On >50 GB data:
  - MySQL: Writes 300 ms avg, Reads 350 ms avg
  - Cassandra: Writes 0.12 ms avg Reads 15 ms avg
- Prime characteristics:
  - non-relational
  - distributed
  - horizontally scalable

Prerequisites

- Motivation, Interest, Curiosity
- General CS I+II, some programming, basic algebra
  - data structures (trees!), object-oriented concepts
  - HTML, Linux (project!)
  - Consider Java course, HTML course
- If something's missing: contact me!
  - Non-CS-majors like Bioinformatics etc.
- *"reading without writing is daydreaming"*

Resources

- Textbooks Databases:
  - Database Management Complete Book
    Ullman & Garcia Molina & Widom, Prentice Hall
  - Database Management Systems
    Ramakrishnan & Gehrke, McGraw Hill
- Textbook Web services:
  - Open Source Web Development with LAMP
    Lee & Brent, Addison Wesley
  - The Web – manifold tutorials, find your favourite
- Course material:

  www.faculty.jacobs-university.de/pbaumann

  → teaching → DBWS
  - Not all slides presented in class! (why that?)
- DBWS mailing list: **eecs-dbwa@...**
  - Subscribe now!
  - Not listed on CampusNet for spam
  - Will NOT use course forum!
- Instructor:
  - p.baumann@..., s.villarroyafernandez@...
- Teaching Assistant:
  - Tbd
- CLAMV help:
  - a.gelessus@..., f.neu@...

How to Handle This Course

JACOBS UNIVERSITY

- Slide sets available on my university page
  - Accessible only from campus networks
  - Caveat: not all slides published (cf intro slide set)
- Strategy:
  - Download slides before lecture
  - Bring paper + pen
  - Take notes
  - Look into book
  - Missed class? Ask colleagues!

Homework / Web Service Project

JACOBS UNIVERSITY

- Implement core of an individual web service
  - Guided, part of homework assignments
- Topics? suggest your own!
  - Teams of 2 – 4
  - Form team & claim topic until 2 weeks from now – contact me for discussion!
  - Earlier examples: cocktail database, stock trade monitoring, hospital drug inventory
- Tech platform: LAMP = Linux, Apache, MYSQL, [PHP | Python | Perl]

Web Service Project (contd.)

JACOBS UNIVERSITY

- Choose a project title, and always use it as email subject
- Develop wherever you want, but final handover on a ClamV Linux box!
  - Support only for ClamV – you will want to do it there
  - Will inspect & discuss source code - zero grade otherwise!
- main evaluation criteria (no particular order):
  - complete wrt. requirements
  - engineering (bug-free, project and code documentation, coding quality, ...)
  - user-friendliness and appealing look&feel
  - complexity (in absolute terms and in comparison to other teams' work)
  - own understanding (assessed through final review)

Grading Scheme

- Final Exam: graded, 100%
- Homework: prerequisite for sitting exam: 80% overall achievement

Course Plot – or: why should I take it?

JACOBS UNIVERSITY

- How to design databases, and how to search them
- How to design (Internet) services

What industry expects a CS graduate to know

- Database services revisited
- Practice: set up a Web service

Your entry point to the DB [dev/admin] world

Course Plot, Refined

- Database design
  - Entity-Relationship Model; UML
- The relational database model
  - Relations; SQL intro; ER mapping; views
  - SQL: queries, constraints, triggers
- Database application development
- Internet service architectures
  - HTTP, XML, JSON
- Database services revisited
  - Logical/Physical Design, Transaction Management, Security, Authorization
- Big Data
- Outlook

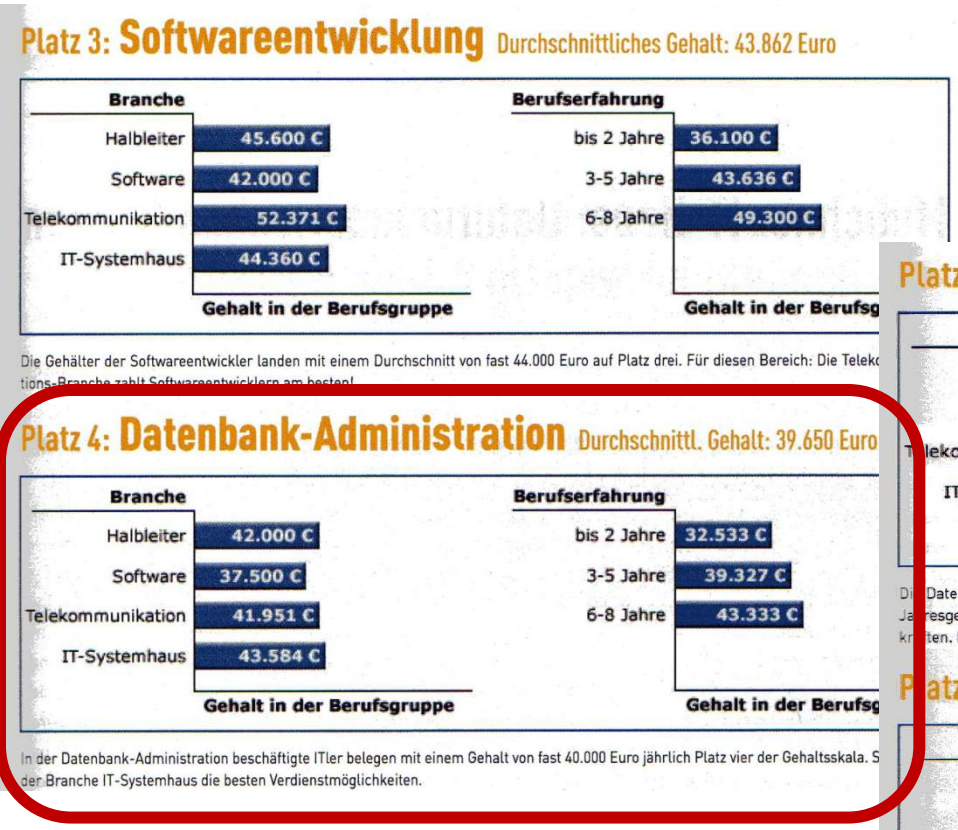
Job Opportunities with DB Knowledge

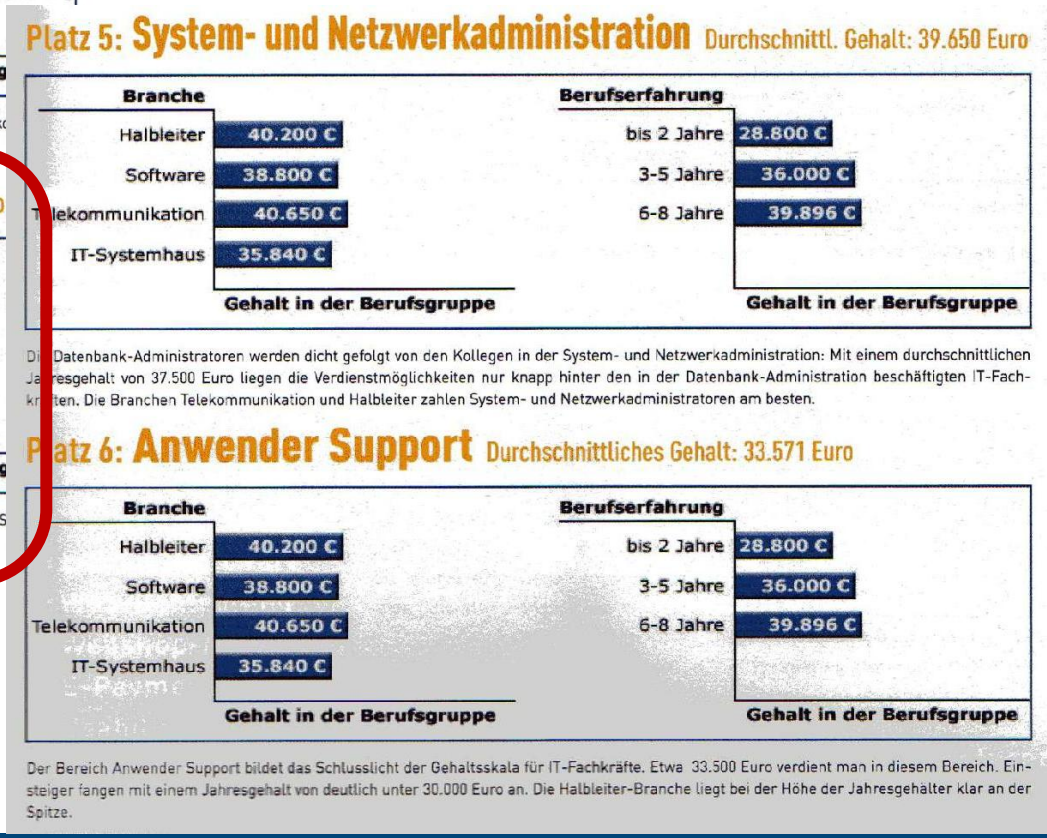
JACOBS UNIVERSITY

- DBMS implementor (with DBMS vendor)
- DB administrator (DBA)
- Database consultants
- Software developer
  - …without basic DB knowledge? No way!

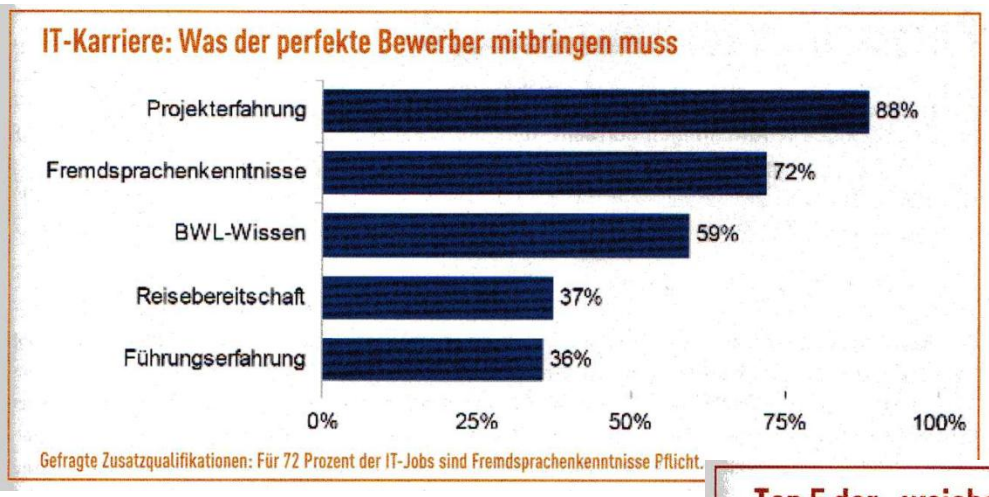
IT Salaries in Germany

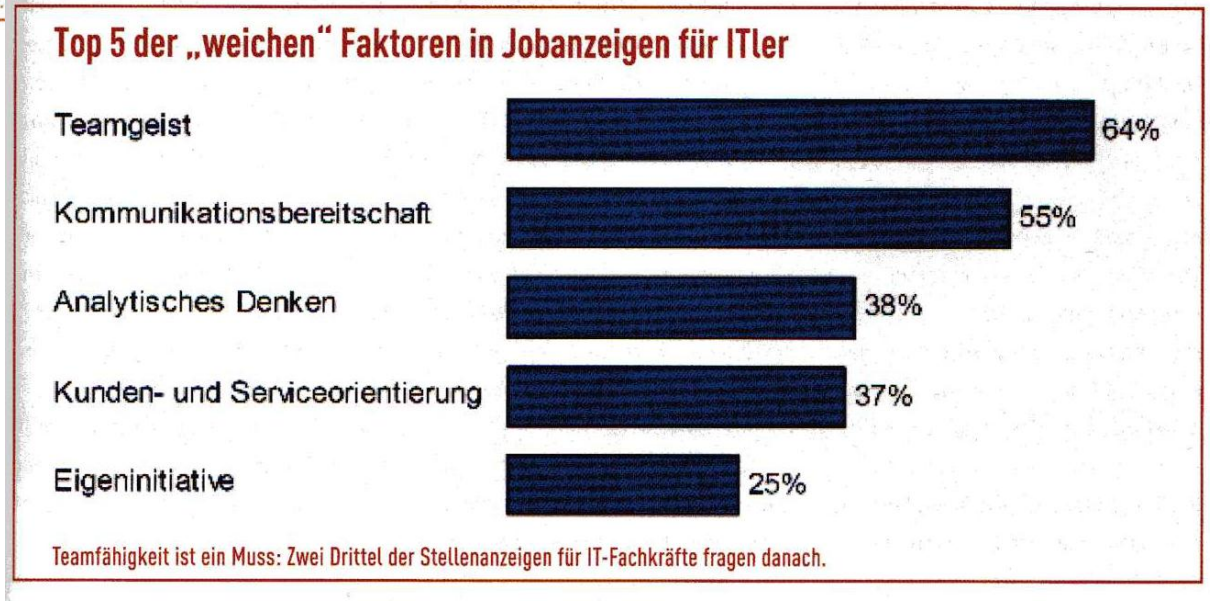
Sample data – check current salary levels





Skills Expected





Summary: Why Study Databases?

JACOBS UNIVERSITY

- Fun & challenge
  - DBMS unique mix of most of CS:
    OS, programming languages, complexity theory, AI, logic, statistics, hardware, …
- Money
  - Computer experts *with database knowledge*
    hold responsible jobs…and are well-paid!