JACOBS UNIVERSITY

CO-560-A
Databases and Web Services

Instructors: Peter Baumann

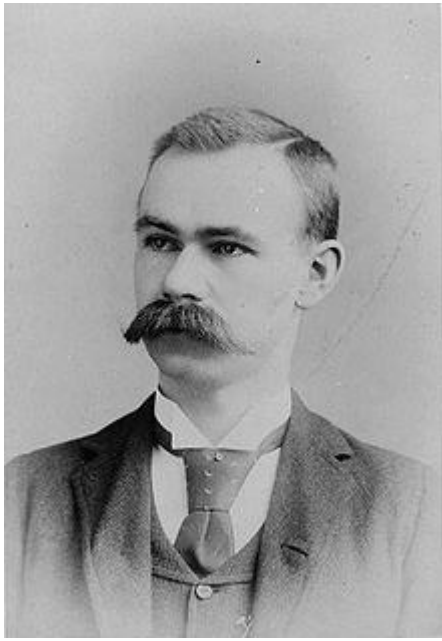
email: p.baumann@jacobs-university.de

office: room 88, Research 1

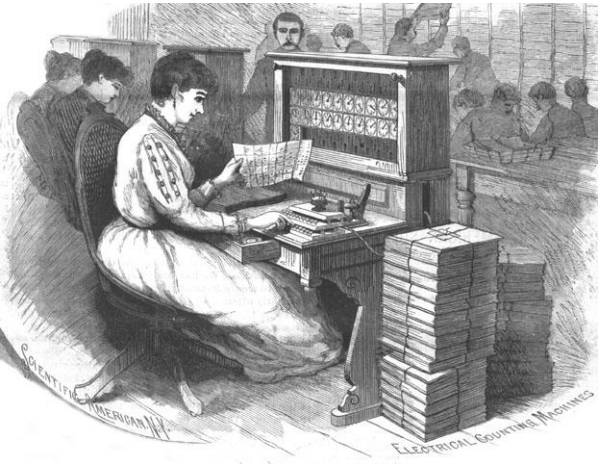
Where It All Started

Source: Wikipedia

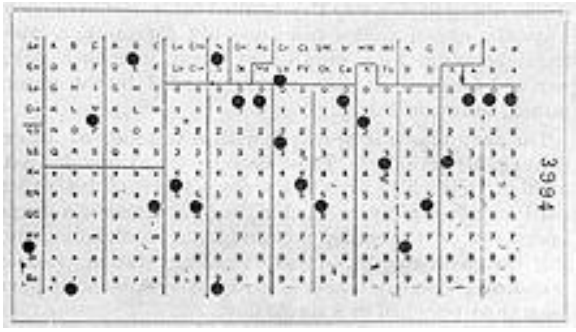
- 1890 census on 62,947,714 US population ← “Big Data”
  - was announced after only six weeks of processing
- Hollerith „tabulating machine and sorter“
- Tabulating Machine Company
  → International Business Machines Corporation



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“?

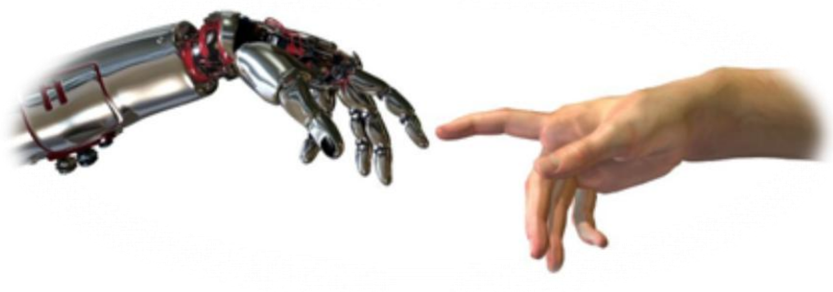
- Internet: the unprecedented information collector
  - 2012: 200m Web servers [Yahoo]
  - estd 50+b static pages [Yahoo]
  - 40 b photos [Facebook]
  - 2012: 31b searches/m [Google]
- 2025: 463 Exabytes / day

- Typical Big Data:
  - Business Intelligence
  - Social networks - Facebook, Twitter, GPS, ...
  - Life Science: patient data, imagery
  - Geo: Satellite imagery, weather data, crowdsourcing, ...

Data = the „new gold“, „new oil“
Petrol industry: „more bytes than barrels“

2012

JACOBS UNIVERSITY



[saubere-zaehne.de] [atlasformen.de]

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

Data Management: The Task

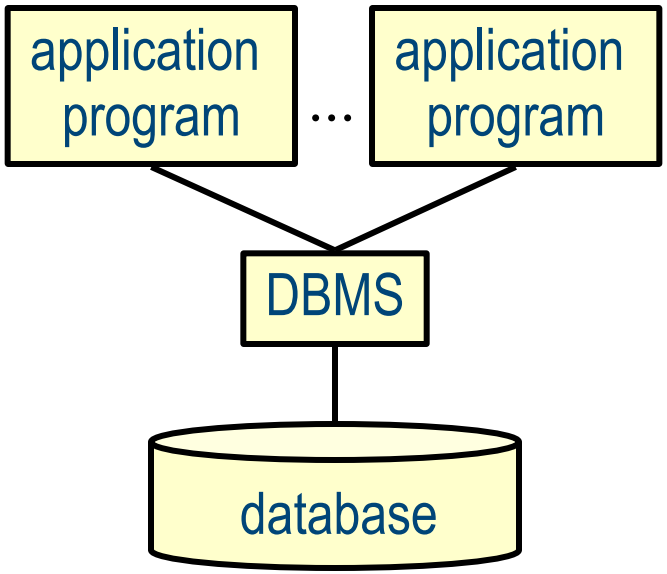
JACOBS UNIVERSITY

- 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, …)

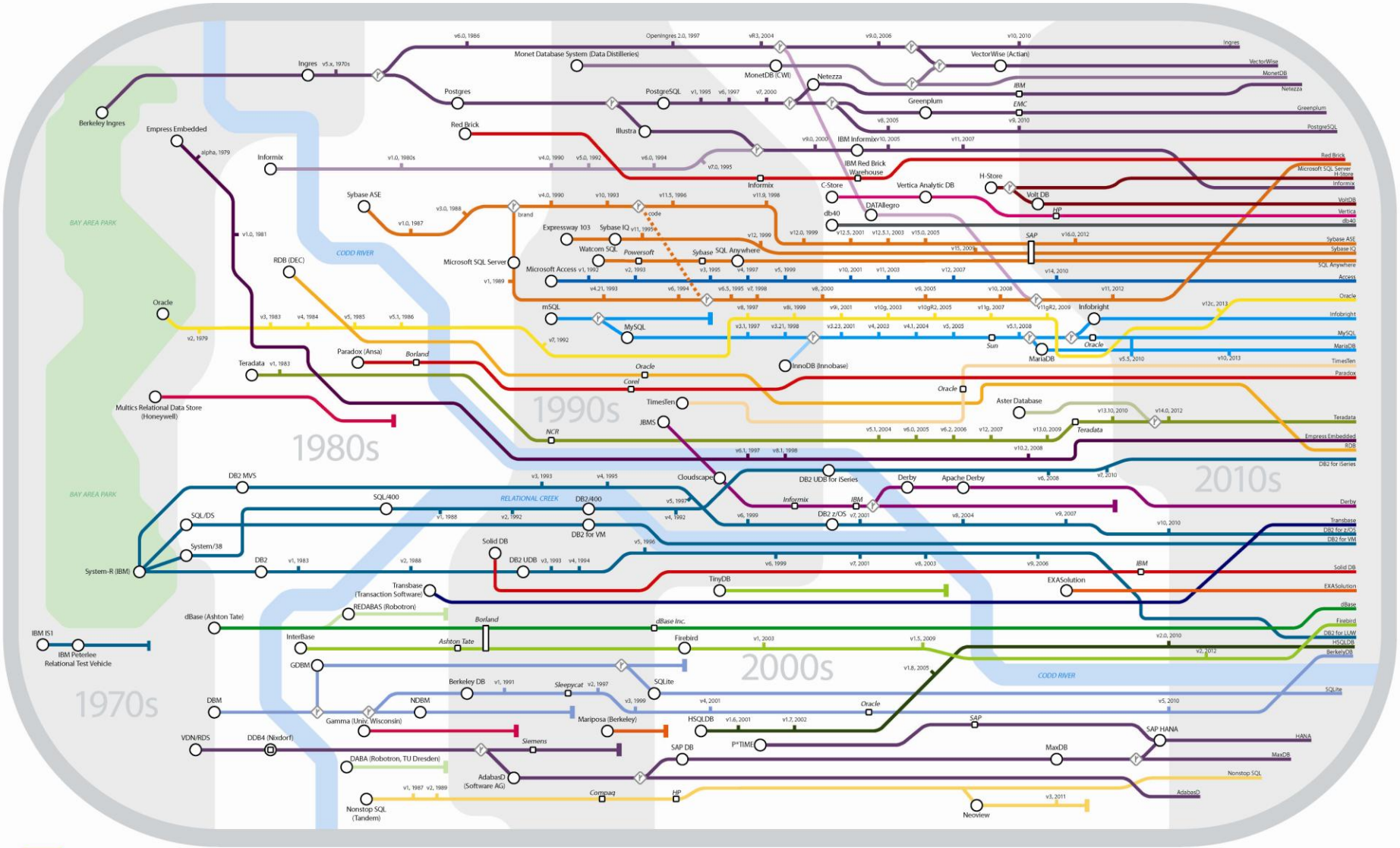


DBMS History

- 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

- Key to success: query language
  - Intuitive (hm…)
  - Yet precise, formalised semantics
  - Declarative = abstracts from internals
  - …hence optimizable

Genealogy of Relational Database Management Systems



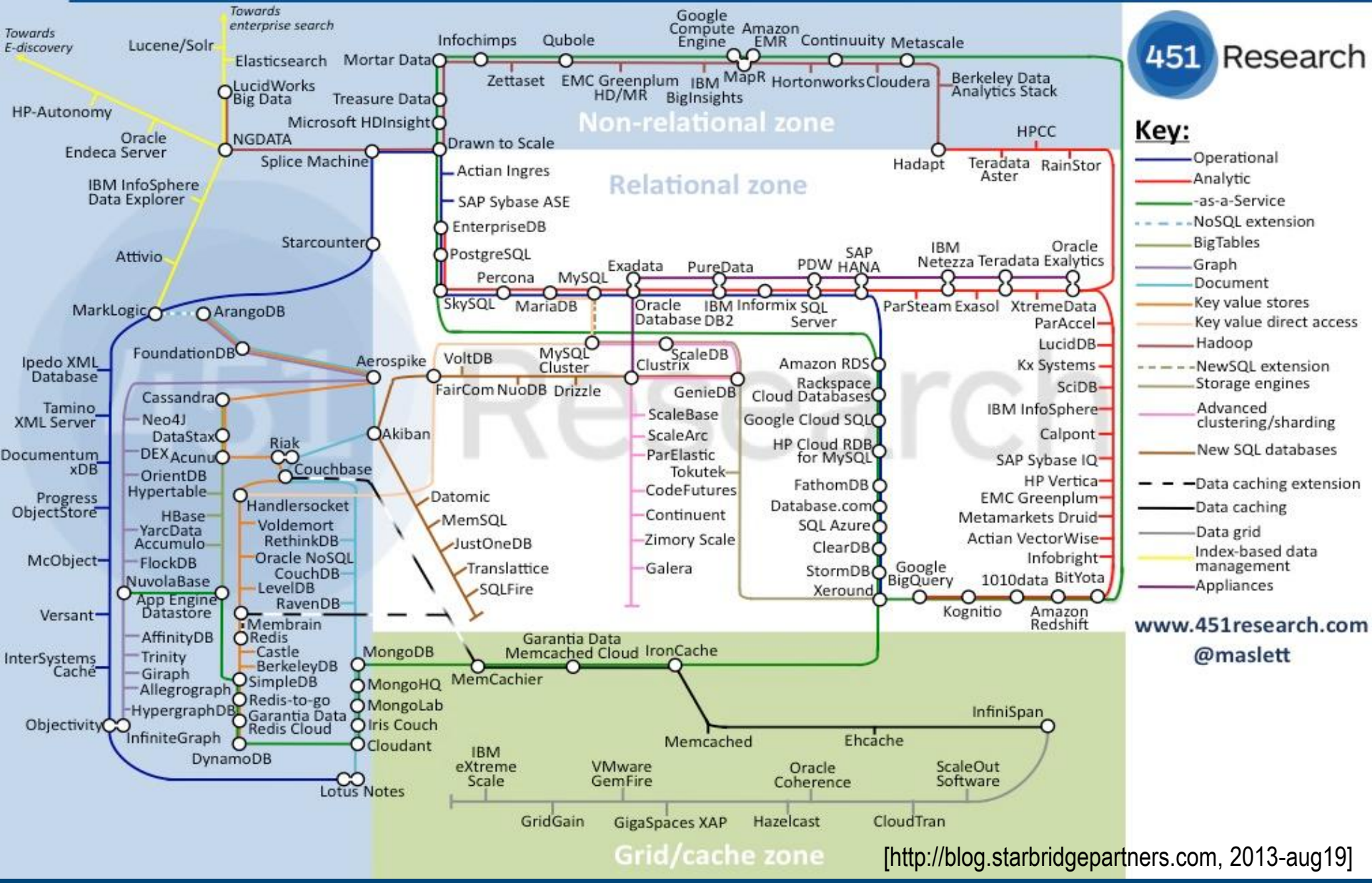
Key to lines and symbols

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]

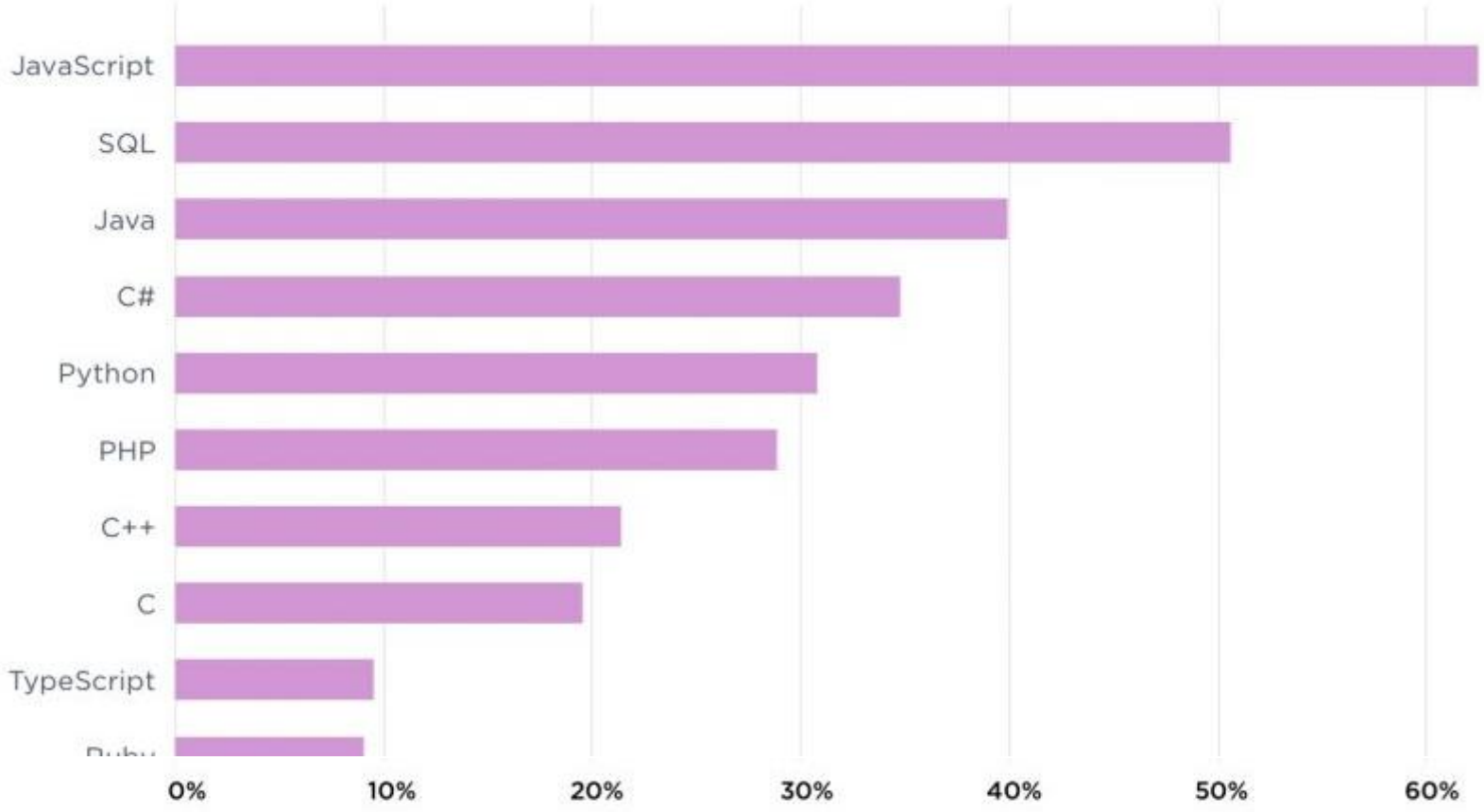
…and Then Came NoSQL

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
  - Cassandra: Writes 0.12 ms avg
- The Empire strikes back: NewSQL

...but still:


JavaScript
SQL
Java
C#
Python
PHP
C++
C
TypeScript
0%
10%
20%
30%
40%
50%
60%
Jelvix
Source: Stack Overflow, Amazon, Statista

JACOBS UNIVERSITY

COURSE & LAB ORGANIZATION

Prerequisites

- Interest, Curiosity, Engagement
- General CS I+II, programming, basic algebra
  - data structures (trees!), object-oriented concepts
  - general programming experience
  - Linux (project!)
- Non-CS majors: contact me!
  - possibly more difficult w/o prerequisites, specifically lab
  - This is an *advanced* CS course!
- *"reading without writing is daydreaming“*
- On any difficulties, contact TAs/me

Resources

- Textbooks Databases:
  - Database Systems: The 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: peter-baumann.org
  → teaching → DBWS
- Instructor: p.baumann@
- DBWS list – will subscribe first batch
  - Latecomers: your responsibility
  - Will NOT use course forum, Moodle!
- Teaching Assistants:
  - Valdrin Smakaj, v.smakaj@
  - Aryans Rathi, a.rathi@
  - Xhersila Olldashi, x.olldashi@
  - Flavia Tasellari, f.tasellari@
  - Alex Tretyakov, atretyakov@
- CLAMV: clabsql
  - a.gelessus@

Lab Project

- Implement core of an individual web service
  - Guided
  - Teams of 2 – 4
- Topics? suggest your own!
  - Earlier examples: cocktail database, stock trade monitoring, hospital drug inventory
- Tech platform: LAMP = Linux, Apache, MYSQL, [PHP | Python | Perl]
- Lab: offline work, submission via repo, discussion in class
  - Weekly lab slots with TA availability: Thu 11:15 – 12:30

Lab Project (contd.)

- 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 with you – *better understand what you submit*
- main evaluation criteria (no particular order):
  - complete wrt. requirements
  - engineering (bug-free, project & code documentation, coding quality, ...)
  - user-friendliness, professional look & feel
  - complexity (in absolute terms & in comparison to other teams' work)
  - own understanding

Where to Work

- CLAMV has reserved clabsql machine
- Connect with:
  - ssh <CampusNet Name>@clabsql.clamv.jacobs-university.de
  - ssh <CampusNet Name>@10.72.1.14
  - Password as distributed on paper
    - *ssh <CampusNet Name>@ 10.17.2.8*
- Assistance:
  - TAs
  - Dr Geleßus, A.Gelessus@jacobs-university.de (only CLAMV topics!)

Interactive SQL Access

- Login to *clabsql*
- Launch mysql client: `mysql -u user -p`
- Pick database: `use dbws;`
- List tables: `show tables;`
- List table definition: `describe Sailors;`
- Send SQL query: `select * from Sailors;`

Web Pages

- On *clabsql*, files sitting in your home directory -> public_html/ are accessible via web server
- Example:
  - User pbaumann
  - File public_html/index.html
  - Accessible via http://clabsql.clamv.jacobs-university.de/~pbaumann/index.html
- Caveat: web server must have permissions to access, minimum:
  - Files: permissions 644
  - Home directory & public_html & subdirectories: permissions 755

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

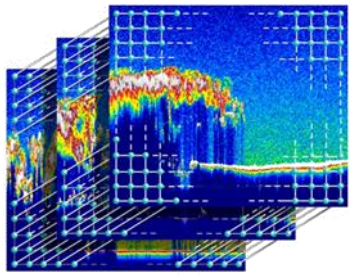
- Database design
  - Entity-Relationship Model; UML
- The relational database model
  - Relations; SQL intro;
    ER mapping; views
  - SQL:
    queries, constraints, triggers
- Database application development
- Internet service architectures
  - HTTP, XML, JSON
- Database services revisited
  - Logical/Physical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

JACOBS UNIVERSITY

OUR RESEARCH

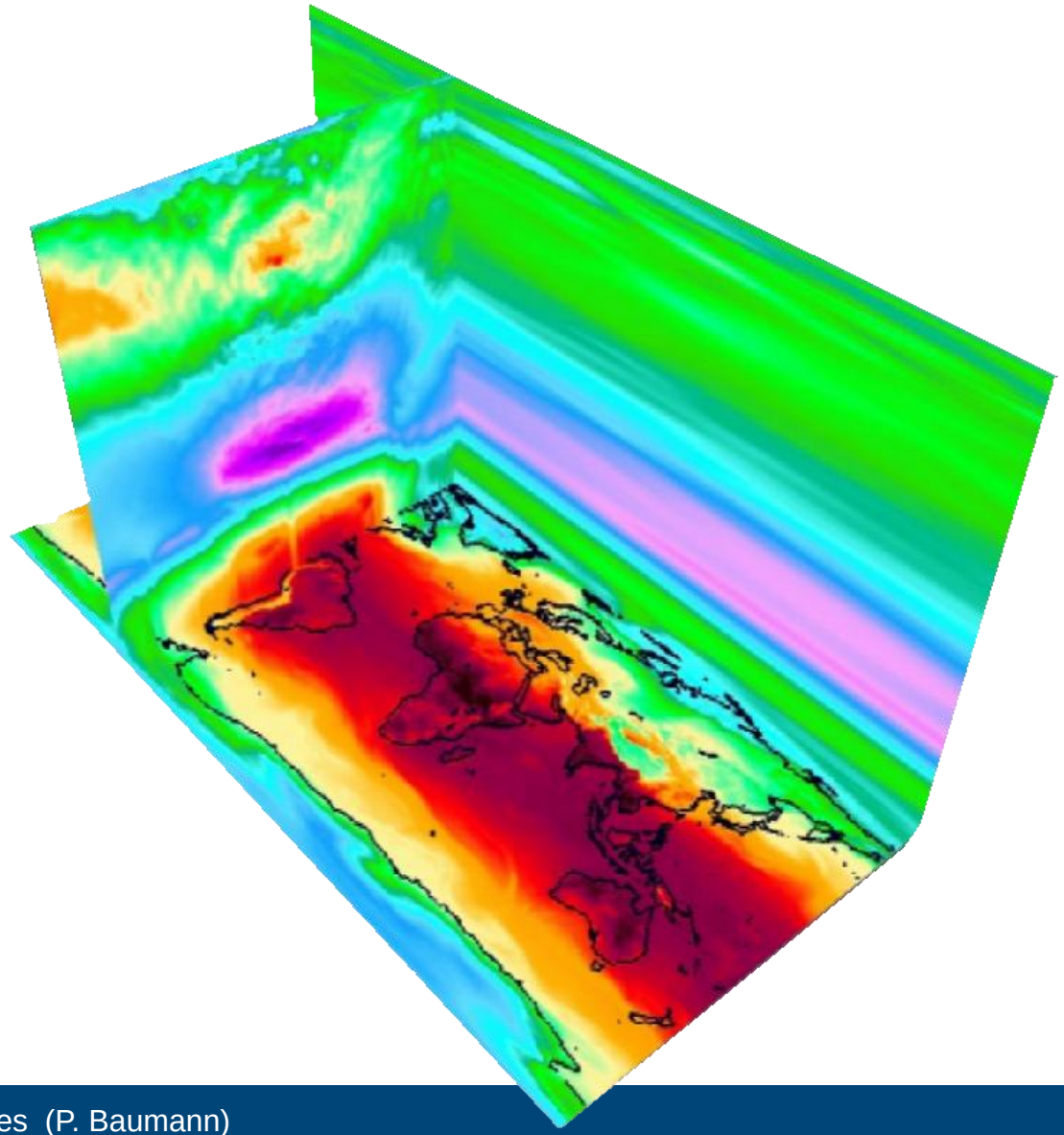
Our Research: Array Databases

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



- Main visible results:
  - rasdaman Array DBMS - worldwide in operational use
  - Datacube standards in OGC, ISO, INSPIRE – eg, SQL/MDA
- *Got rock-solid coding skills? Join us!*
  - *C++, Java, JavScript*

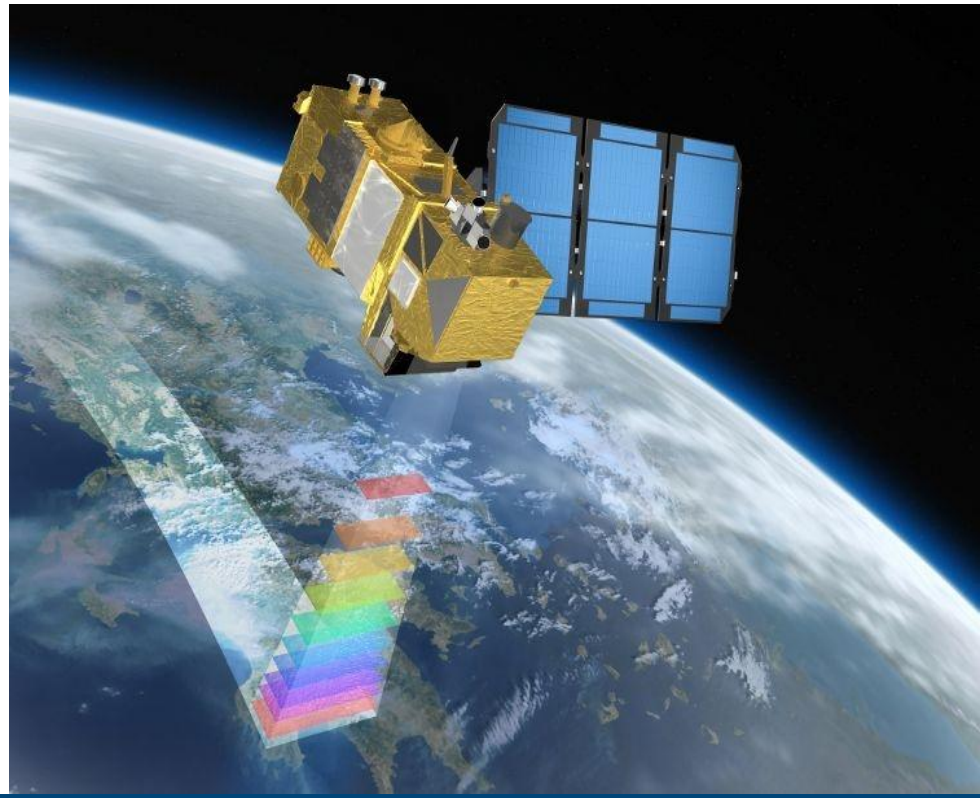
Arrays, aka Datacubes



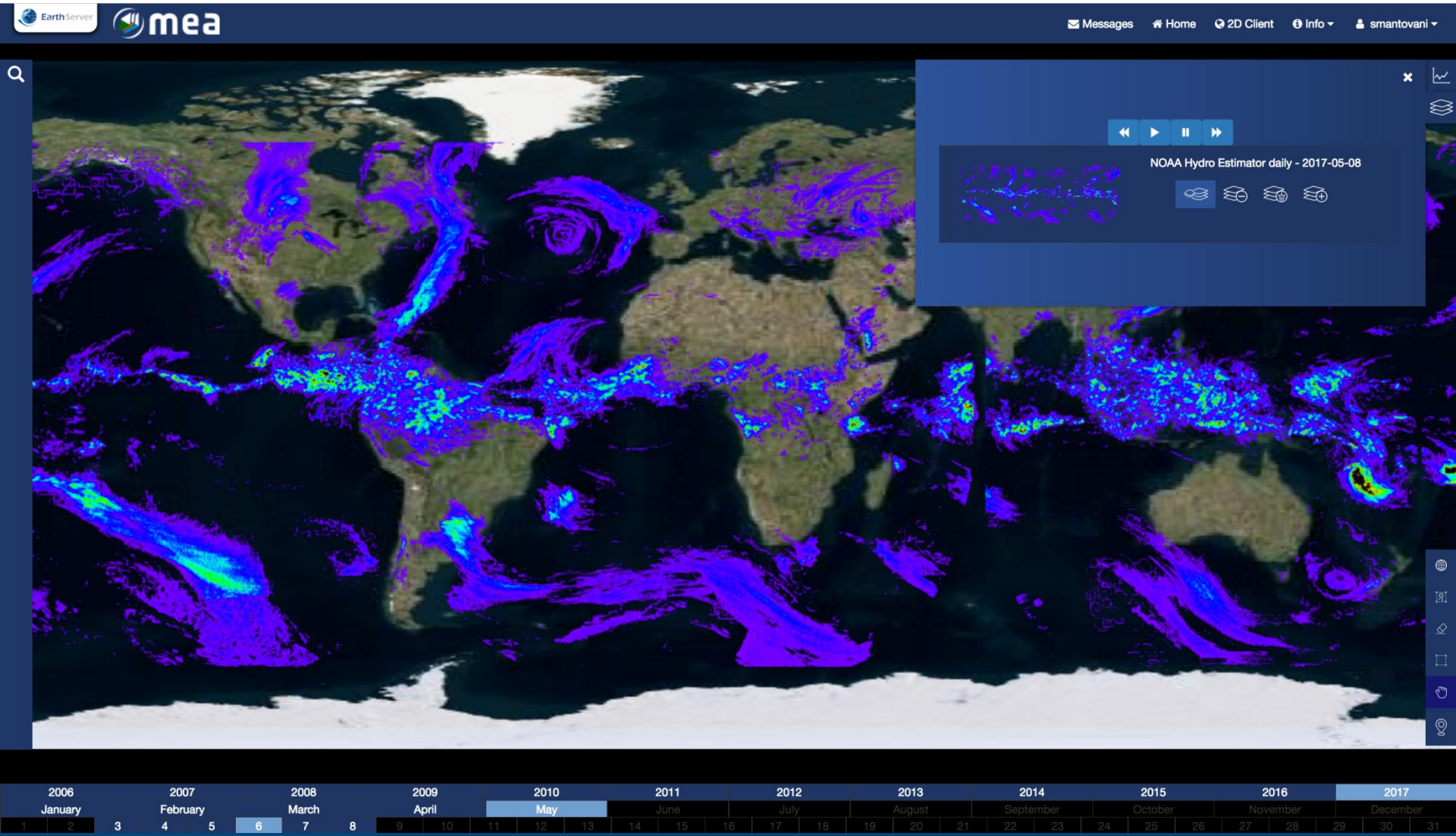
[DKRZ]

Big Data in Geo: Satellite Imagery

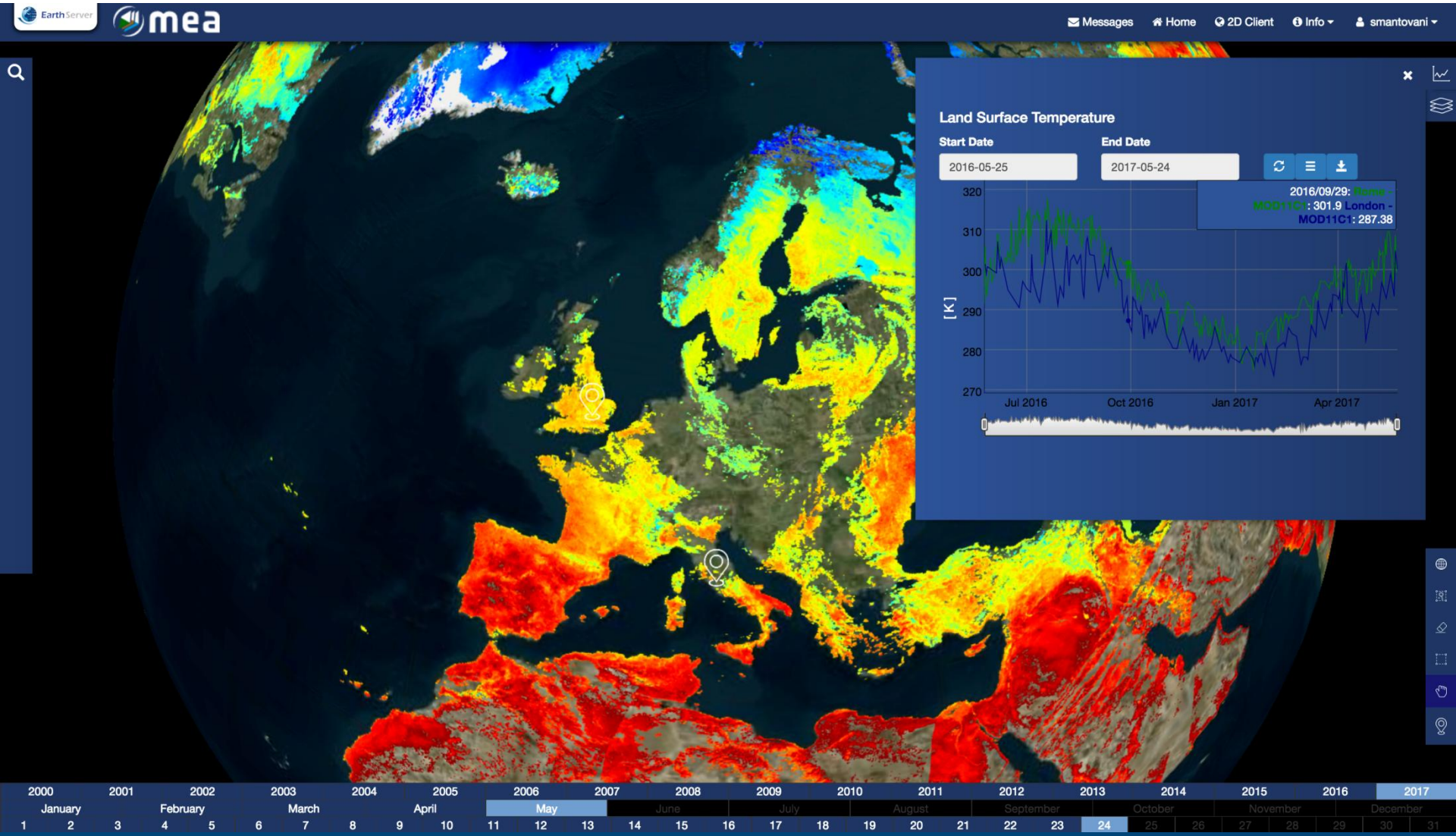
- 100s of Exabytes expected for 2020
- ngEO: planning for 10^12 satellite images under curation of ESA [ESA]
  - Increased # of instruments flying
    - *A-Train, Landsat, Sentinels, ...*
  - Increased spectral resolution:
    5 (Landsat) to 250 (ALI/Hyperion)
  - Increased spatial resolution:
    few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



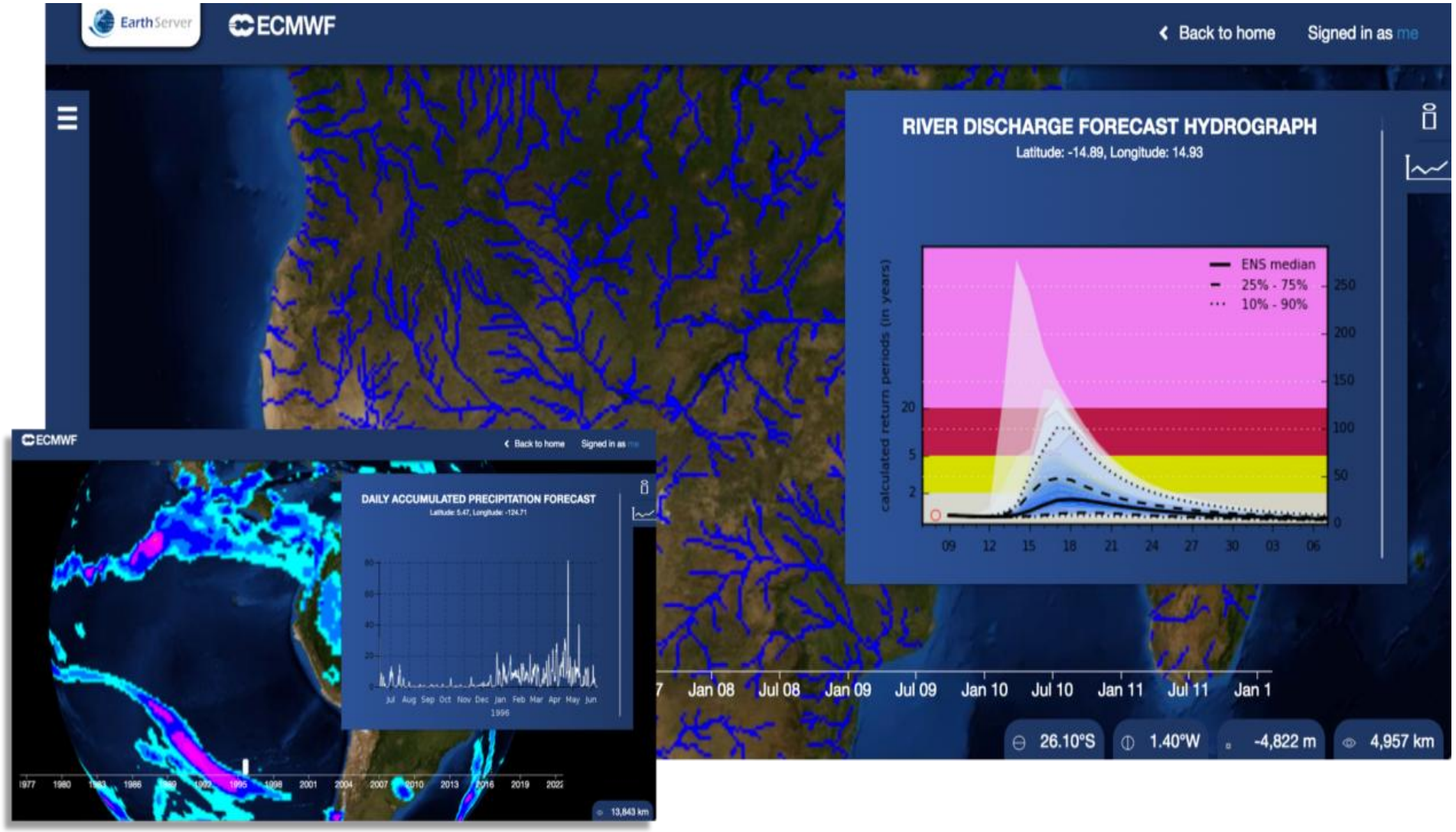
Daily Hydro Estimator



Land Surface Temperature, Cloudfree



ECMWF: River Discharge

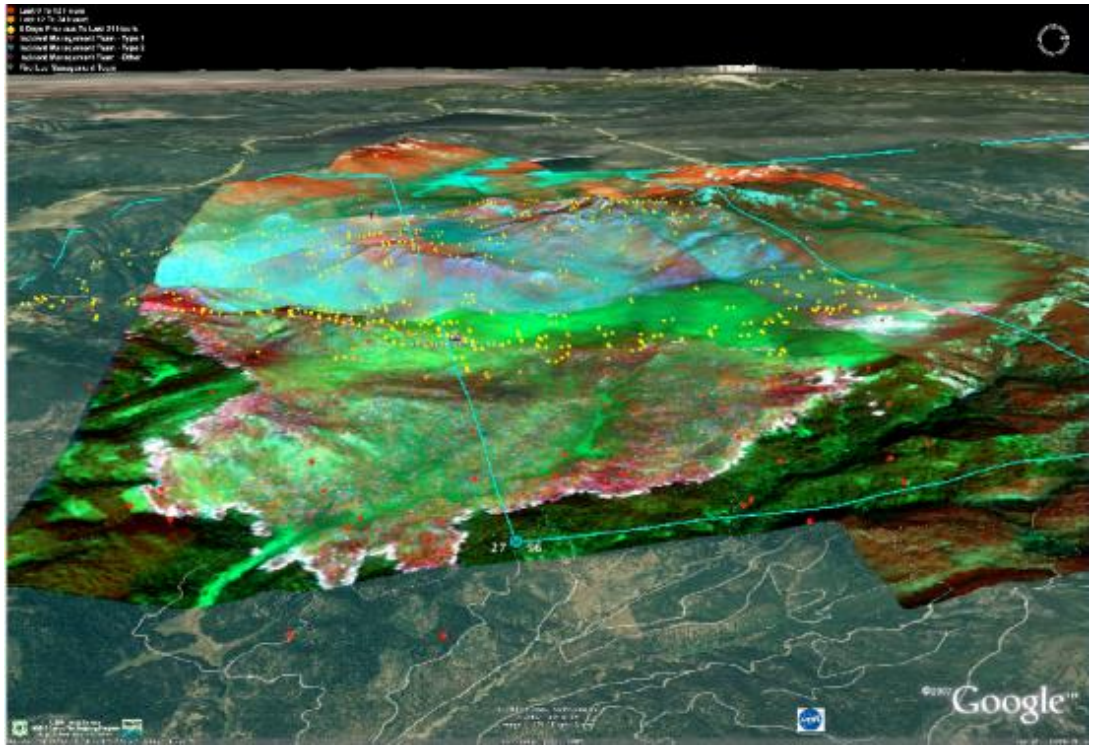


BIG EARTH DATA

The Digitized Planet

On-Board Datacube Intelligence

ORBiDANse:
Orbital Big Data Analytics Service



[images: ESA, NASA]

JACOBS UNIVERSITY

CAREER RELEVANCE

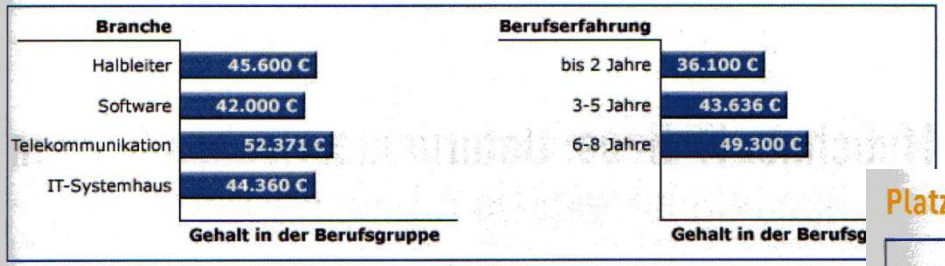
Job Opportunities with DB Knowledge

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

IT Salaries in Germany

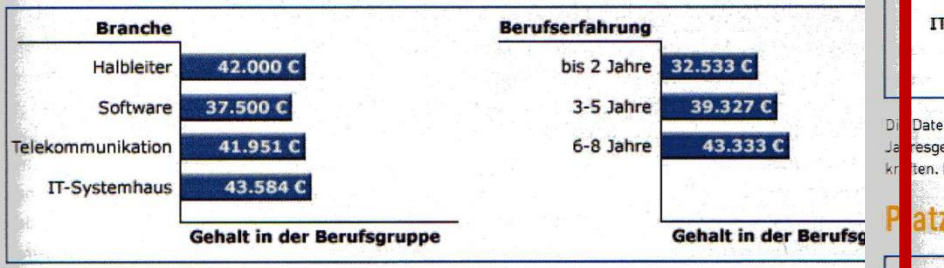
Samples, outdated – check current figures

Platz 3: Softwareentwicklung Durchschnittliches Gehalt: 43.862 Euro



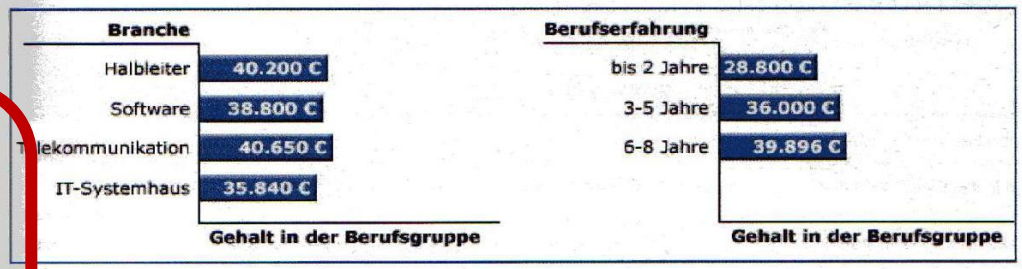
Die Gehälter der Softwareentwickler landen mit einem Durchschnitt von fast 44.000 Euro auf Platz drei. Für diesen Bereich: Die Teleko tions-Branche zahlt Softwareentwicklern am besten!

Platz 4: Datenbank-Administration Durchschnittl. Gehalt: 39.650 Euro



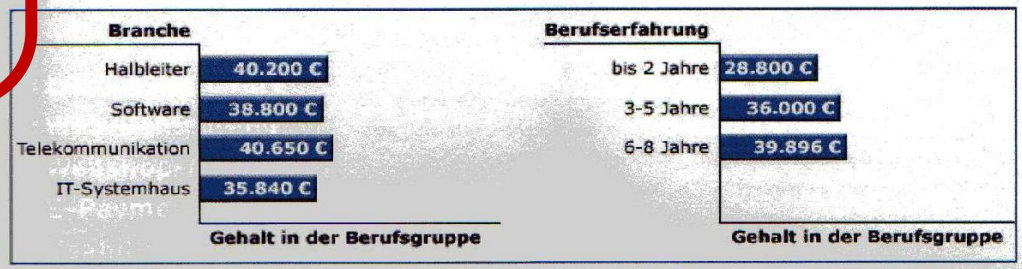
In der Datenbank-Administration beschäftigte ITler belegen mit einem Gehalt von fast 40.000 Euro jährlich Platz vier der Gehaltsskala. S der Branche IT-Systemhaus die besten Verdienstmöglichkeiten.

Platz 5: System- und Netzwerkadministration Durchschnittl. Gehalt: 39.650 Euro



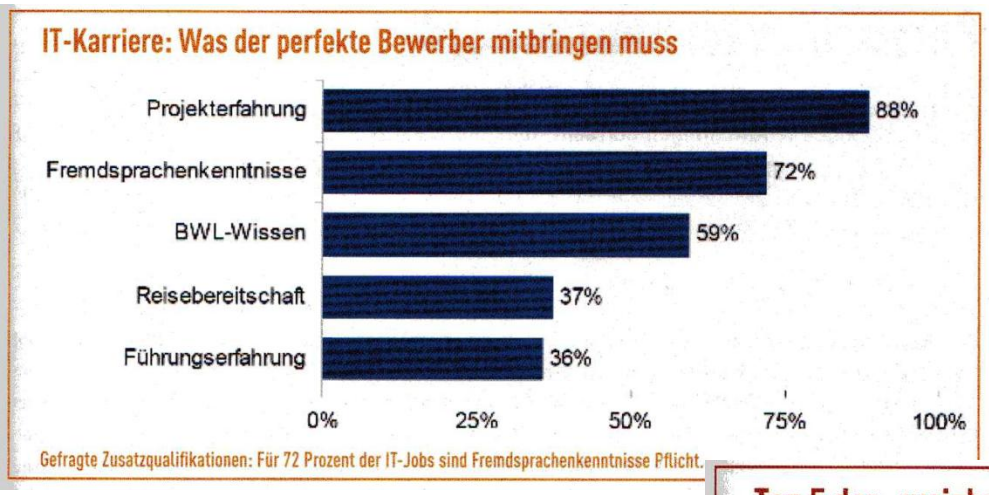
Die Datenbank-Administratoren werden dicht gefolgt von den Kollegen in der System- und Netzwerkadministration: Mit einem durchschnittlichen Jahresgehalt von 37.500 Euro liegen die Verdienstmöglichkeiten nur knapp hinter den in der Datenbank-Administration beschäftigten IT-Fachkräften. Die Branchen Telekommunikation und Halbleiter zahlen System- und Netzwerkadministratoren am besten.

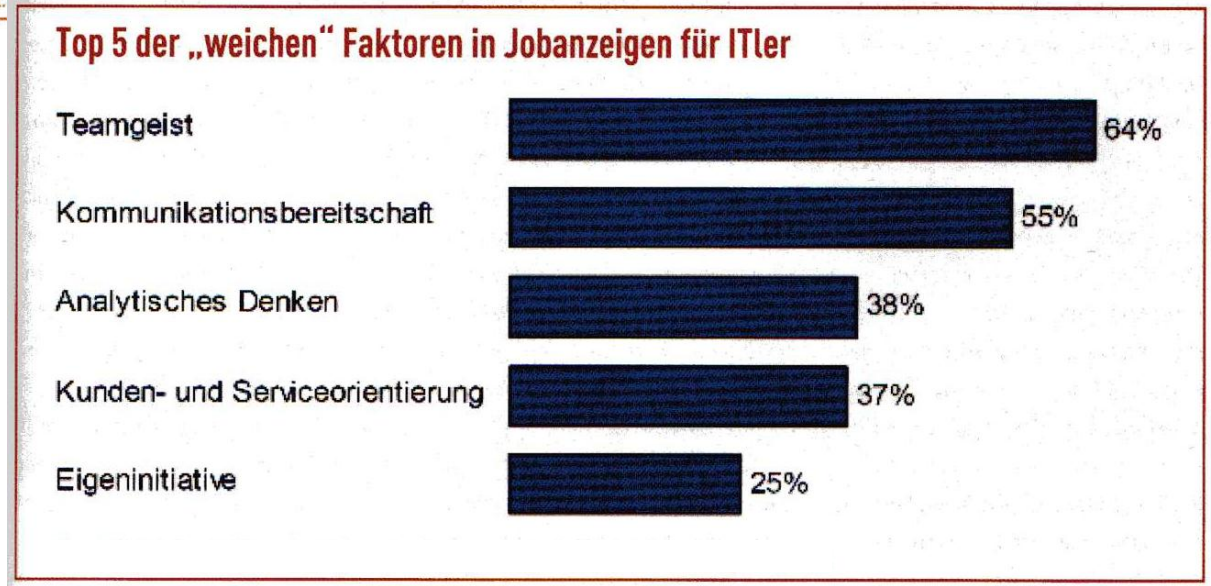
Platz 6: Anwender Support Durchschnittliches Gehalt: 33.571 Euro



Der Bereich Anwender Support bildet das Schlusslicht der Gehaltsskala für IT-Fachkräfte. Etwa 33.500 Euro verdient man in diesem Bereich. Einsteiger fangen mit einem Jahresgehalt von deutlich unter 30.000 Euro an. Die Halbleiter-Branche liegt bei der Höhe der Jahresgehälter klar an der Spitze.

Skills Expected





Summary: Why Learn 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!