CONSTRUCTOR UNIVERSITY

Advanced Databases

instructor: Peter Baumann

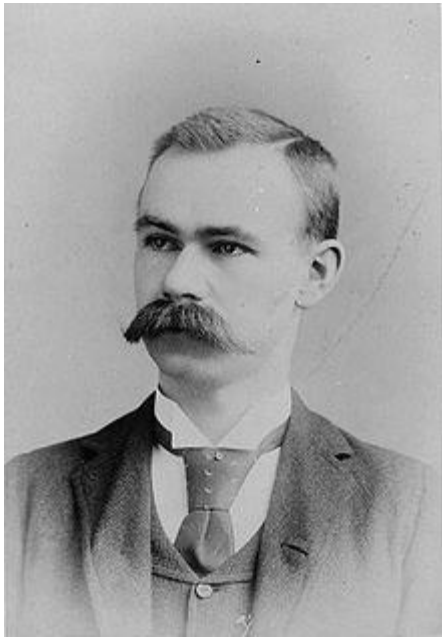
email: pbaumann@constructor.university

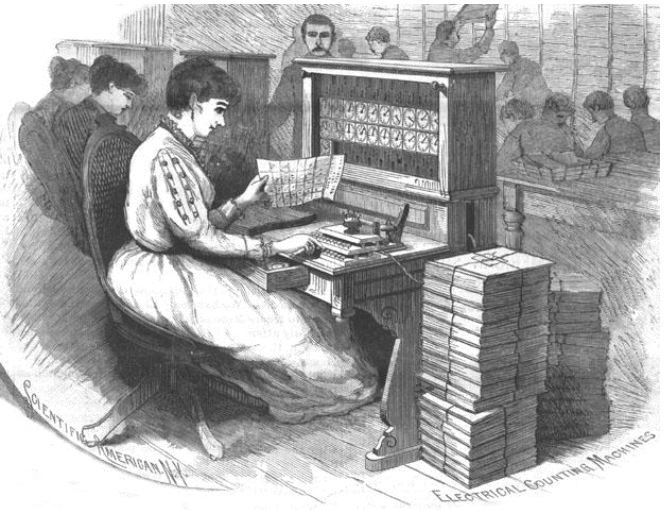
tel: -3178

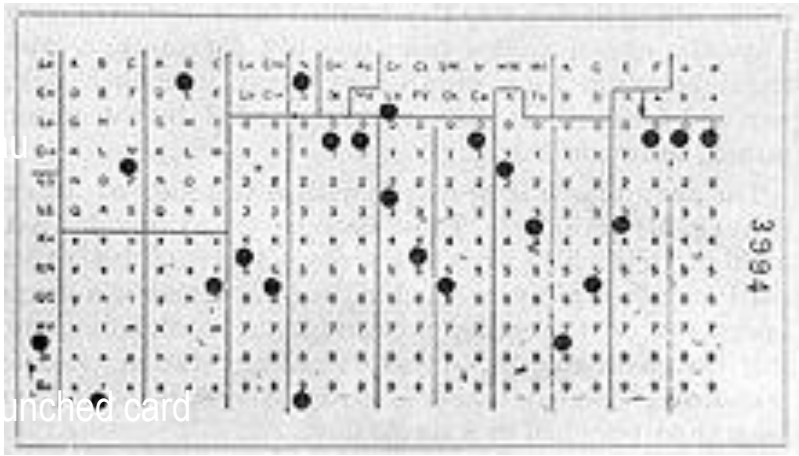
office: Research 1, room 88

Where It All Started

- 1890 census on 62,947,714 US population ← “Big Data”
- Hollerith „tabulating machine and sorter“
  - 2 years faster
- Tabulating Machine Company
  → International Business Machines Corporation







unched card

[Wikipedia]

2012

Big Data

- Internet: the unprecedented information collector
  - 2012: 200m Web servers [Yahoo]
  - estd 50+b static pages [Yahoo]
  - 2012: 31b searches / month [Google]
  - Wayback Machine: 240 billion web pages archived from 1996
- 2025: expected 463 Exabytes / day

- Typical Big Data:
  - Social networks - facebook, twitter, GPS, ...
  - Business: Data Warehousing
  - Geo: Satellite imagery, weather data, ...
  - Petrol industry: „more bytes than barrels“
- ...plus „Deep Web“

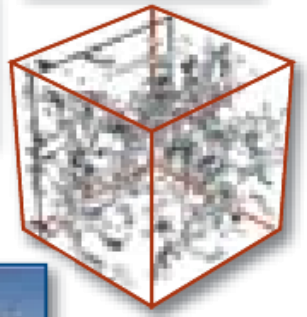
Data = the „new gold“, „new oil“
Petrol industry: „more bytes than barrels“

The 4th Paradigm

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



„Big Data“: The 4+ Vs

- „data too big to transport“,
  but also „too complex to process“

- Volume - ngEO plannings: 10^12 images under ESA custody
- Velocity - NASA EOSDIS: 5 TB/d; LOFAR: 25 TB/h; phones: 1+ PB/d
- Variety - grids; point clouds; general meshes; vectors; text; graphs; ...
- Veracity - Quality, provenance, trust

- ...plus more in blogs: Value, Verisimilitude, Variability, Visualization, ...

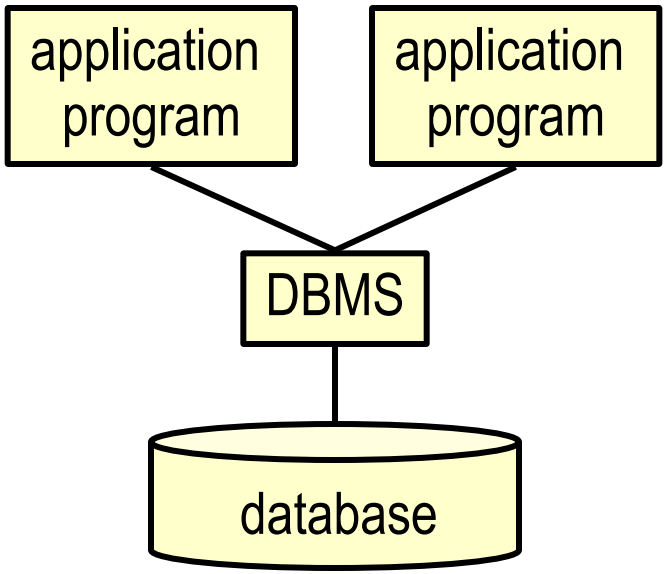
Data Management: The Task

- 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 (Rihanna is taking 320301, …)



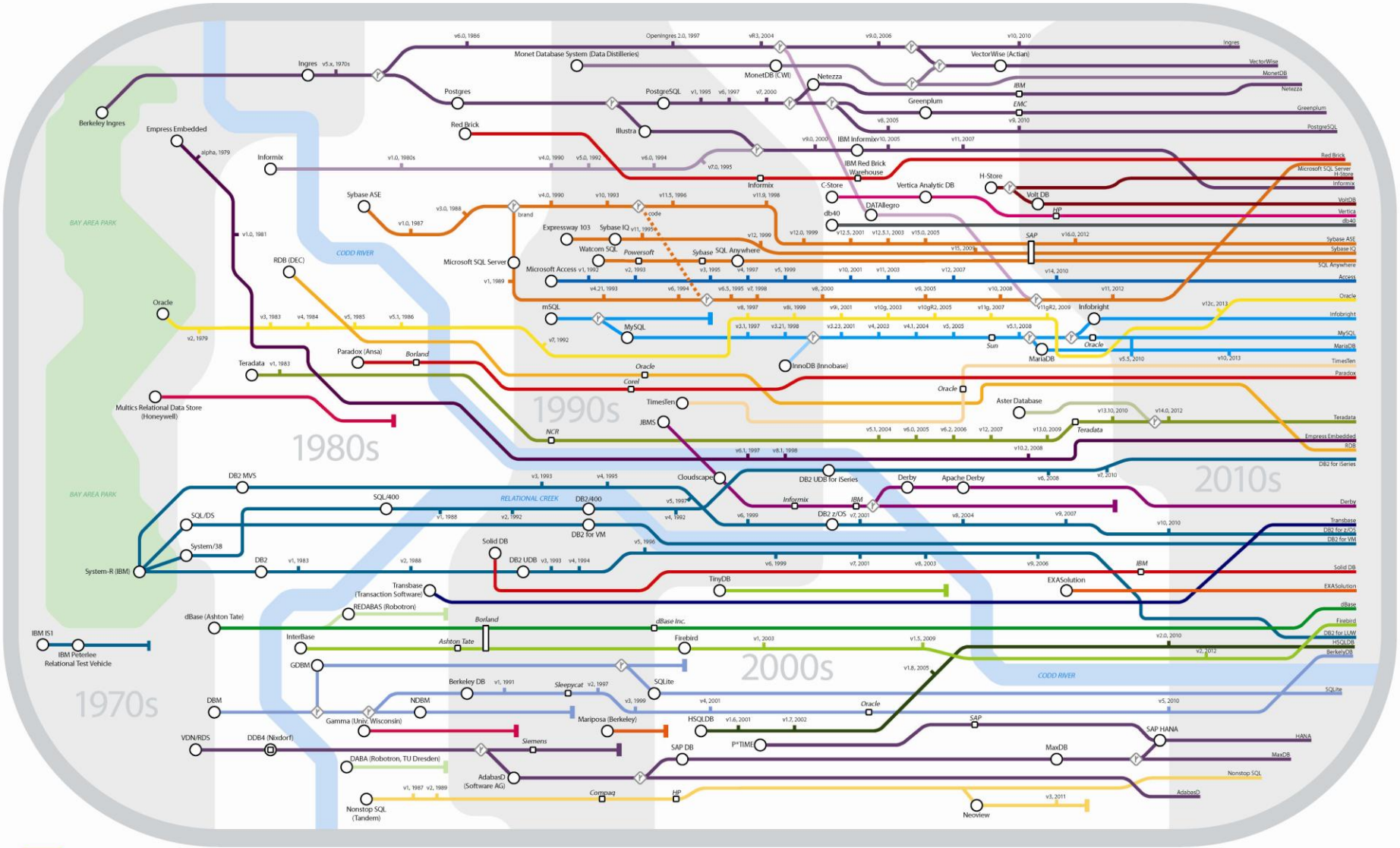
Why Use a DBMS?

- DBMS to maintain & query large datasets
- Quality of service
  - Flexible, efficient (=fast) access to large data assets
  - Concurrent access
  - Data independence
- Efficiency
  - Uniform data administration
  - Reduced application development time
- Safety
  - Data integrity & security
  - Crash recovery

The Real Life

- 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; Sybase; 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 extensions
  - XML databases

Genealogy of Relational Database Management Systems



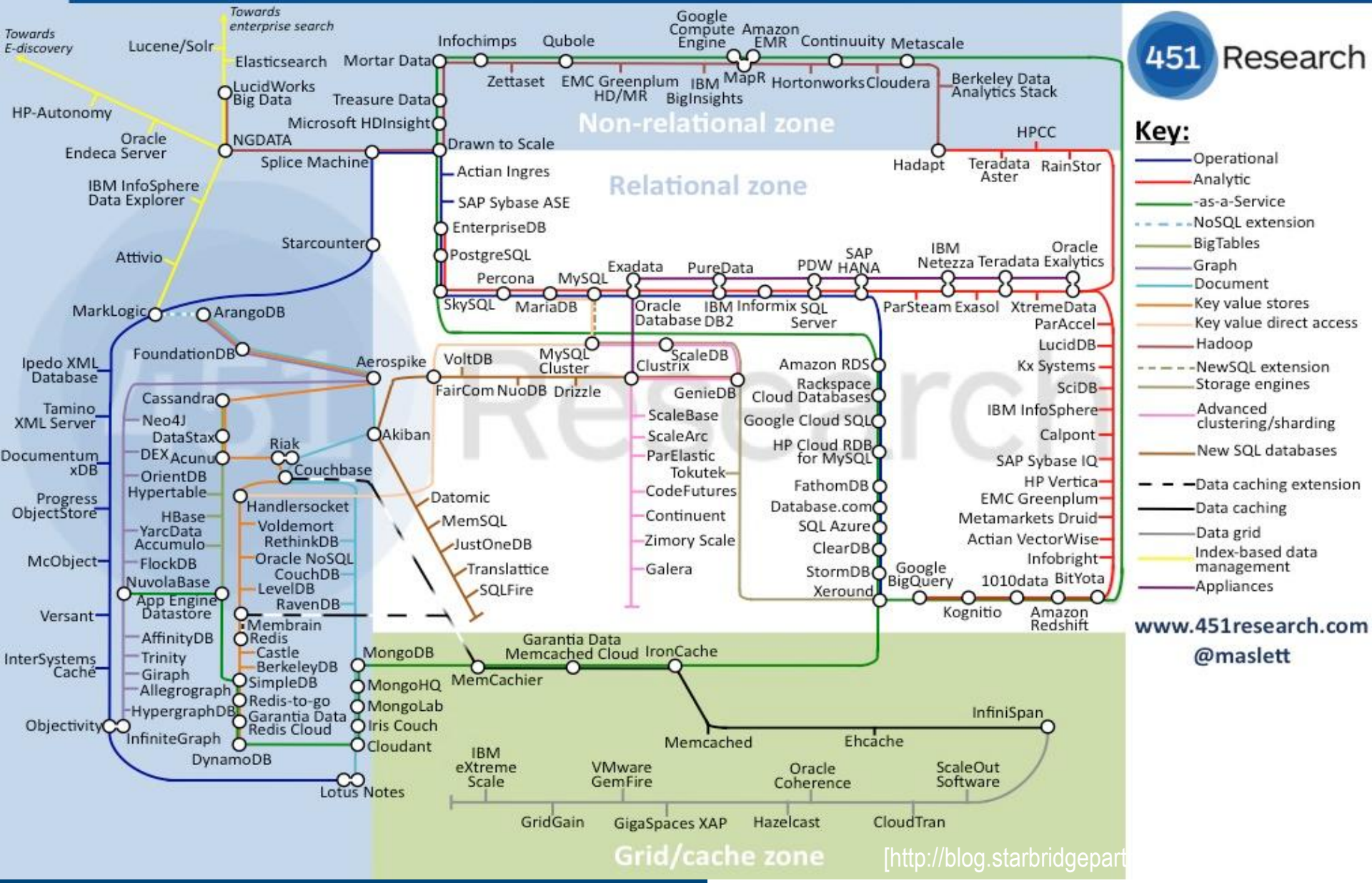
Key to lines and symbols

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

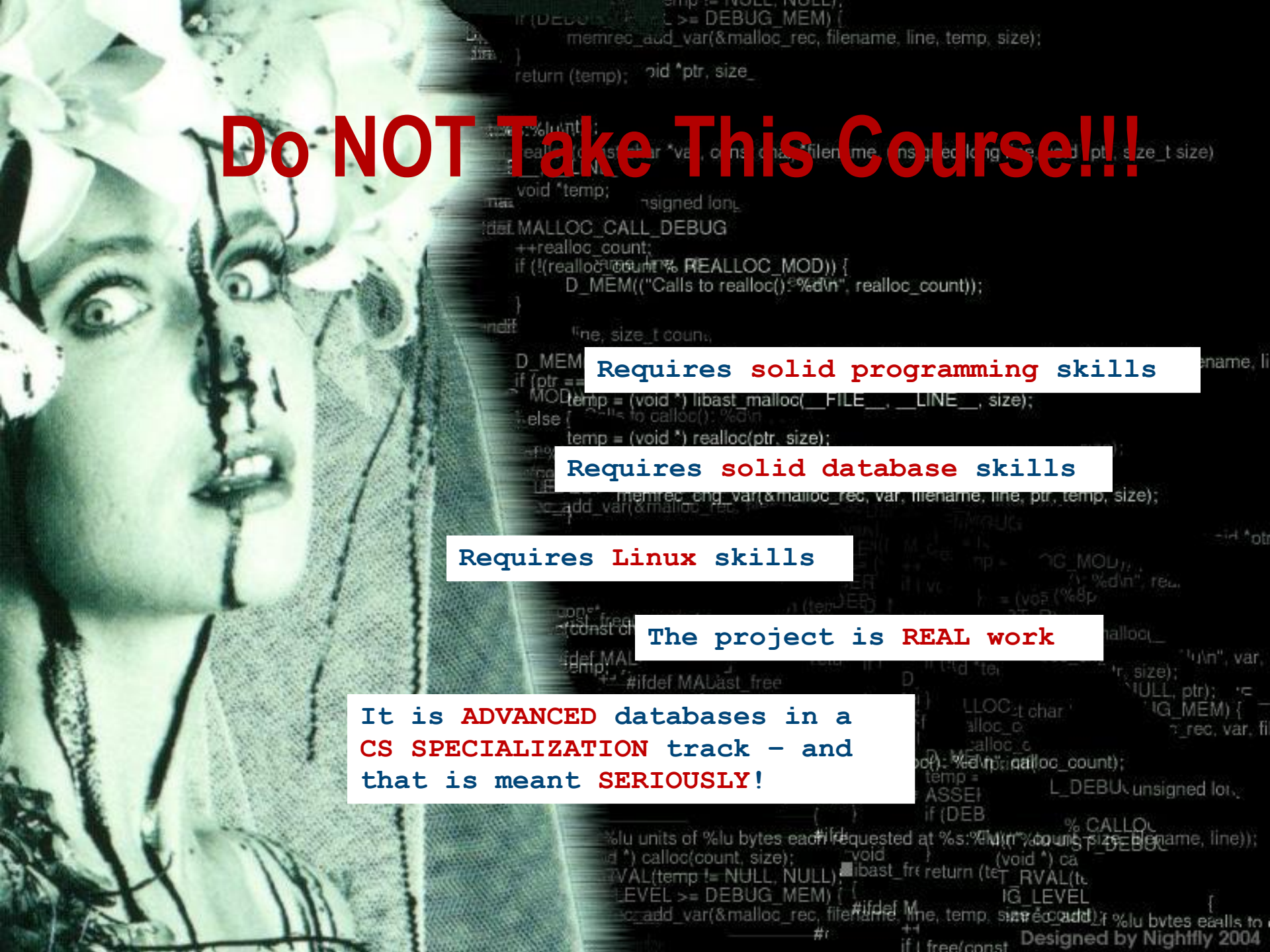
Felix Naumann, Jana Bauckmann, Claudia Exeler, Jan-Peer Rudolph, Fabian Tschirschnitz
Contact - Hasso Plattner Institut, Potsdam, Felix.Naumann@hpi.uni-potsdam.de
Design - Alexander Sandt Grafik-Design, Hamburg

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

Database Landscape Map – December 2012



Back to the Course


Do NOT Take This Course!!!
Requires solid programming skills
Requires solid database skills
Requires Linux skills
The project is REAL work
It is ADVANCED databases in a CS SPECIALIZATION track – and that is meant SERIOUSLY!
Designed by Nightfly 2004

Course Plot

- Databases
  - RDBMS recap & engine deep dive
- Database application development
- NoSQL, NewSQL, MapReduce
- OLAP
- Virtualization & Cloud
- Security

Advanced Databases Project

„The way to your goal starts the day
you take over 100% responsibility for your actions.“
– Dante Alighieri

- Establish core of your own Web service
  - Full stack: database backend, business logic, Web frontend
- Team of optimally 3-4
- Assignments guide through steps
- Final presentation in class

Prerequisites

- General database concepts, SQL, some SQL API binding
  - *This course is not about SQL!*
- Some general CS / IT knowledge
  - Algorithms & data structures, object-oriented concepts, programming
- Motivation, Interest, Curiosity

"*reading without writing is daydreaming*"

Resources

- "Database Management Complete Book"
  Ullman & Garcia Molina & Widom, Prentice Hall
- www.peter-baumann.org
  → teaching
  → Advanced Databases
- peer group
- mailing list *course-bdcs*
- TA + me

Grading

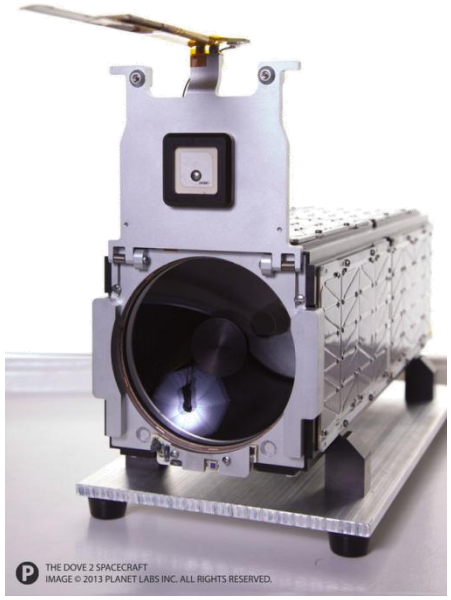
- Exam
  - written, @ end of semester
- Lab
  - Semester project: build your own Web service
  - Sequence of tasks, individually evaluated
  - Lab grade is sum of task grades

Big Data: a Kaleidoscope

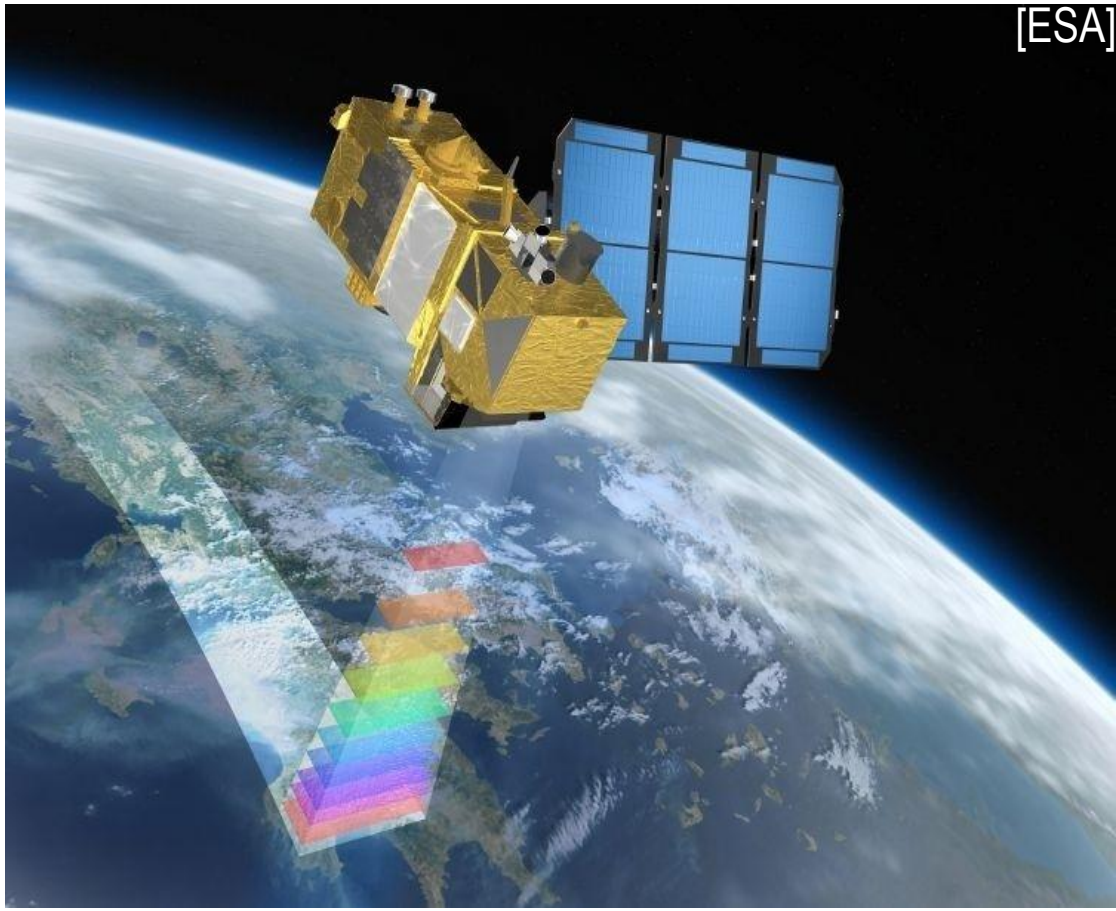

BIG EARTH DATA
The Digitized Planet
→ Youtube

Big Data in the Earth Sciences

- „Exaflood“: ~100s of Exabytes in 2020
  - Spectral bands
  - resolution: km → ~20cm
- CubeSats are coming!



[Planet.com]



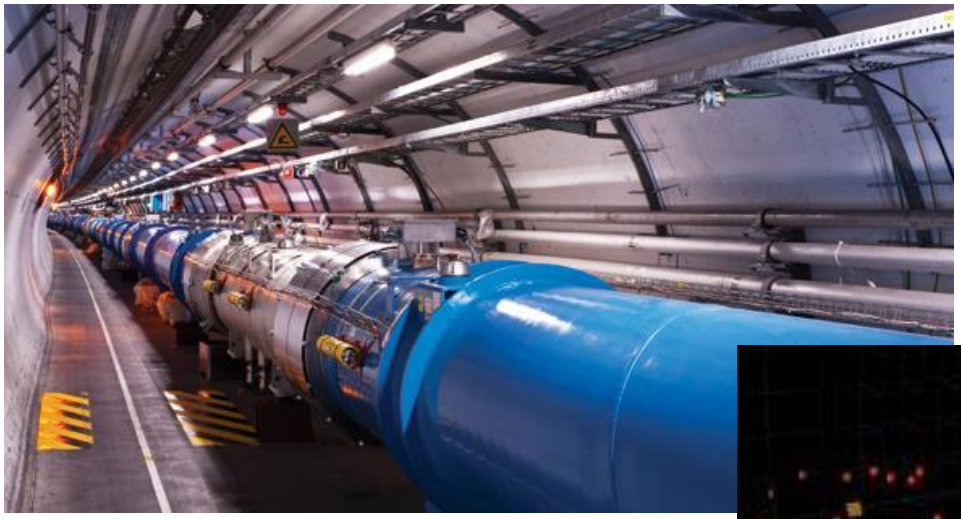
[ESA]

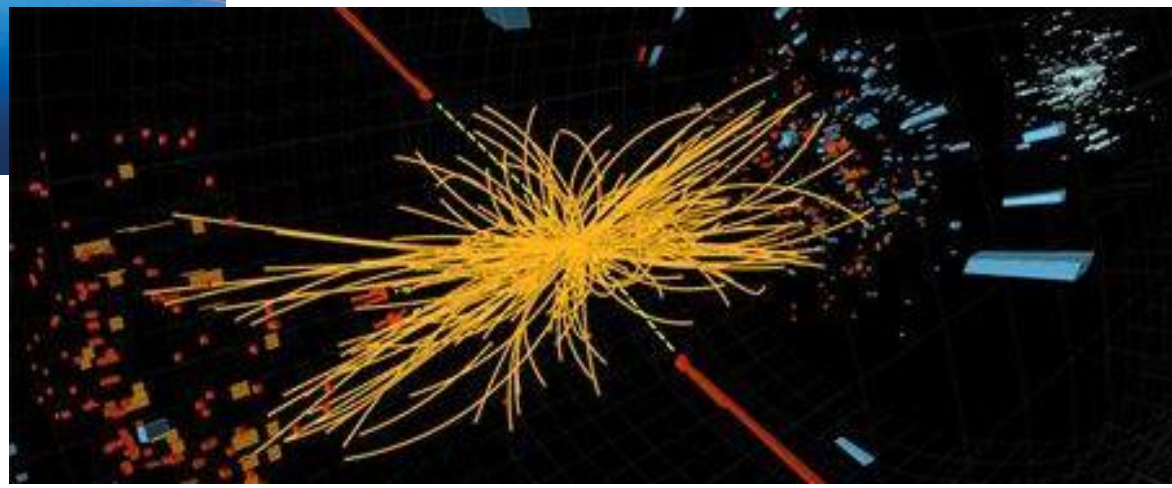
Variety in Oceanography

[OGC Ocean Interoperability Experiment]

Big Data in High Energy Physics

- CERN, Large Hadron Collider:
  13 PB in 2010





[CERN]

Big Data in the Life Sciences

- Neuro Sciences: Human Brain Project (EU, ~1b €), BRAIN (US)
  - Multi-scale models of the human brain (molecular - behavioral)
- Data aggregation integration → cost saving, improved care
  - Personalised patient care
    - *Real-time observation & agent adjustment*
- Genome medicine
  - 23andme.com: personalised analysis of your DNS
  - „Microsoft is an equal opportunity employer. All qualified applicants will receive consideration for employment without regard to race, color, gender, sexual orientation, gender identity or expression, religion, national origin, marital status, age, disability, veteran status, genetic information, or any other protected status."

Big Data in Business Intelligence

- Business data worldwide 2x every 1.2 years [estd]
- Walmart: 1+ million customer transactions / h
  - estd. 2.5+ PB databases =167x US LoC
- FICO Falcon CC Fraud Detection System
  - 2.1b CC accounts worldwide
- Equifax: multi-million customers' key data compromised

Possible Early Warning Sign for Market Crashes

[www.wired.com]

Big Data in Industry

- Industry 4.0: integration of production & IT
  - Optimising value chain & life cycle
- Automobiles
  - Typically, ~100m LoC
  - Networked with co-traffic, traffic lights, ...
    - *2.8 ZB in 2012, plus 2.5 PB / Tag [Computerwoche]*
- Airplanes:
  - A380: 1b LoC
  - Per engine: 1 TB / 3 min
    - *LHR → JFK = 640 TB*

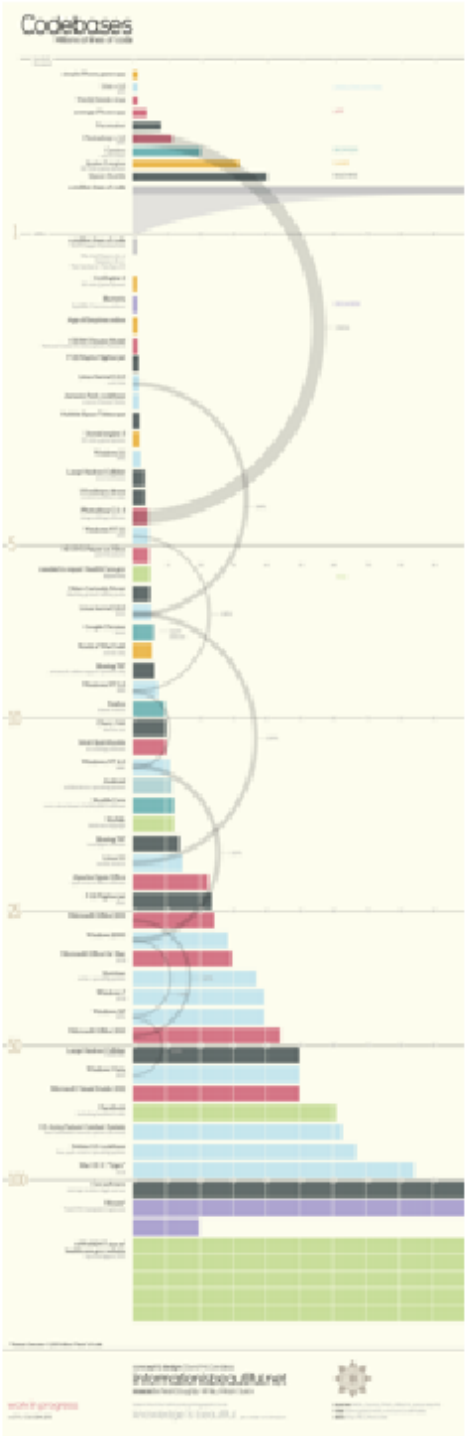
[Kristen Nicole]

[Airbus]

Big Code – Lines of Code

Average iPhone app	= 50.000 lines
Hubble Space Telescope	= 2 million lines
Windows 3.1 (1992)	= 2.5 million lines
Control software for US military drone	= 3.5 million lines
Windows NT 3.1 (1993)	= 4.5 million lines
HD DVD Player Xbox	= 4.5 million lines
World of Warcraft Server	= 5.5 million lines
Google Chrome	= 6.5 million lines
Windows NT 4 (1996)	= 11 million lines
MySQL	= 12 million lines
Boing 787 Flight Software	= 14 million lines
F35 Fighter jet	= 23 million lines
Microsoft Office 2013	= 44 million lines
Large Hadron Collider	= 50 million lines
Facebook	= 61 million lines
US Army Future Combat System	= 63 million lines
MacOS X 4.1 Tiger	= 85 million lines
Average high-end car	= 100 million lines
1.3+ million iPhone apps, 1.3+ million Android apps	= 170billion lines

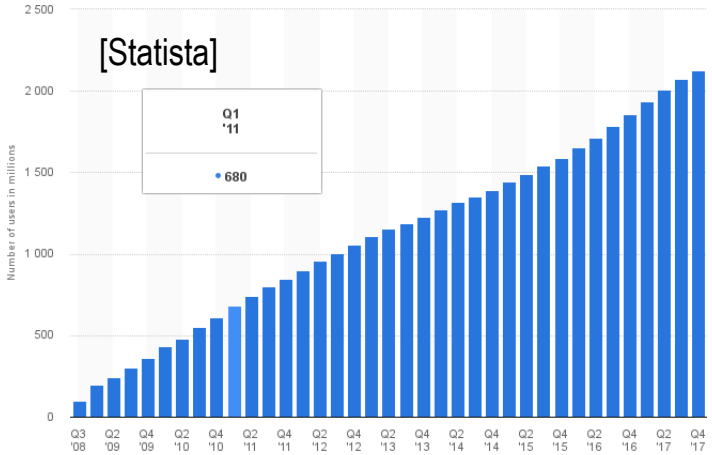
source: http://www.informationisbeautiful.net/visualizations/million-lines-of-code/



Big Data in Social Networks

- Facebook

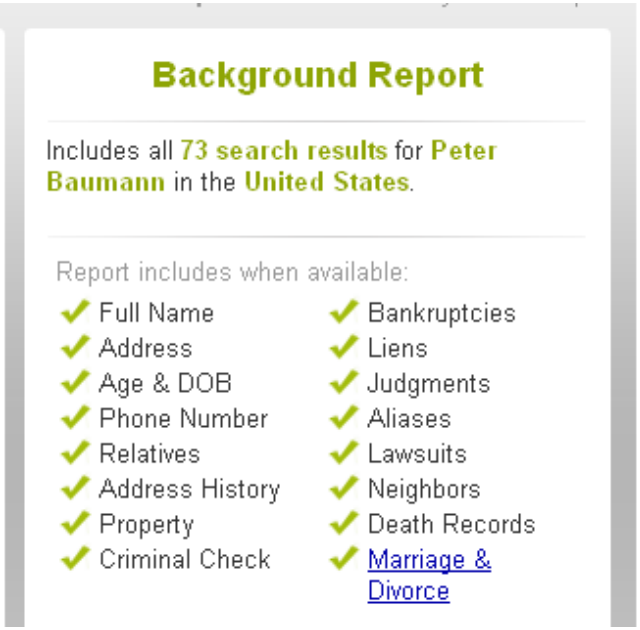
[M. Rodriguez, Aurelius]

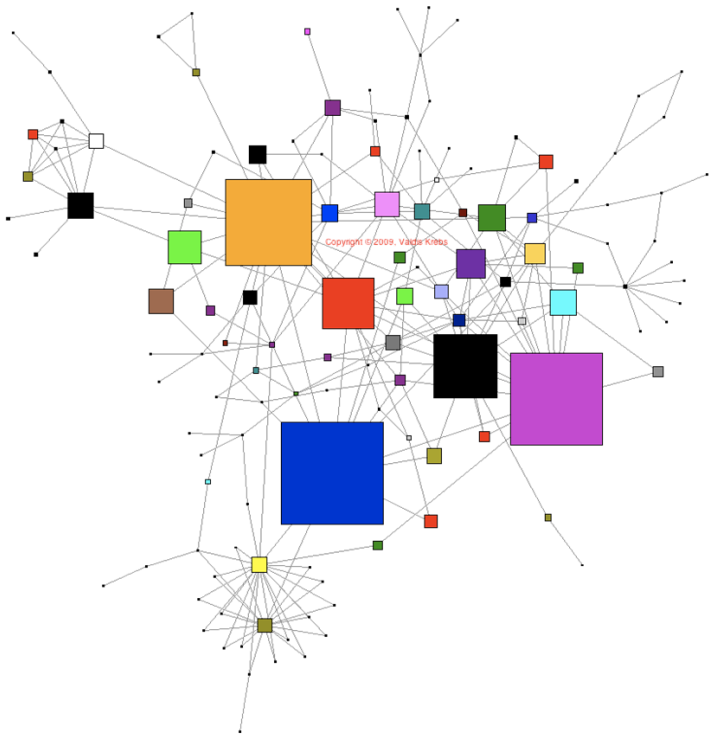


- MS Messenger: 30b chats, 240m participants [Leskovec, 2008]
- Global mobile phone traffic: 80,000 PB in 2016 [Gartner]

Big Data in Social Networks

- Social Network Analysis, Sentiment Analysis, Human Analytics:
  - How isolated / connected / central / important is a person?
  - How / where from / where to does information flow? opinions?
- Intelius.com: „live in the know“





Pizza ordering in the future

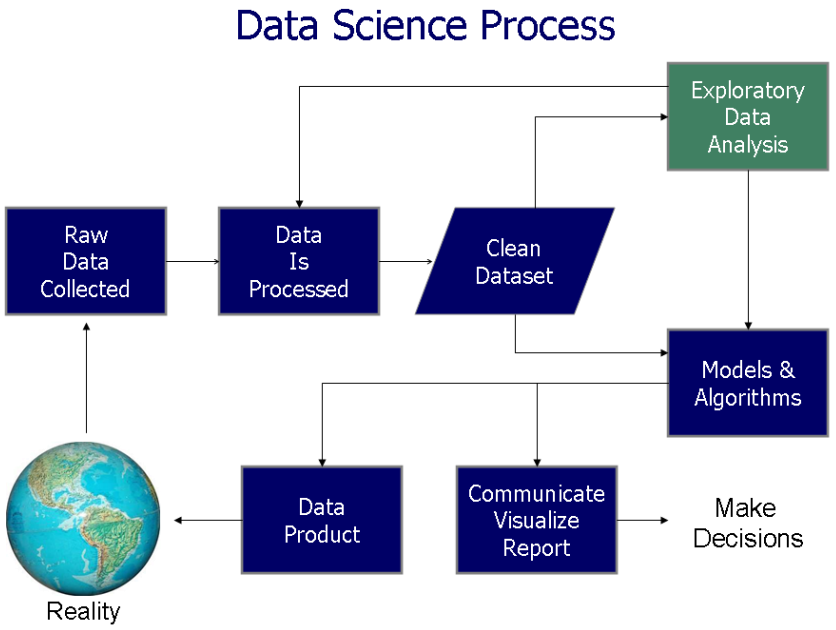
Internet of Things (IoT)

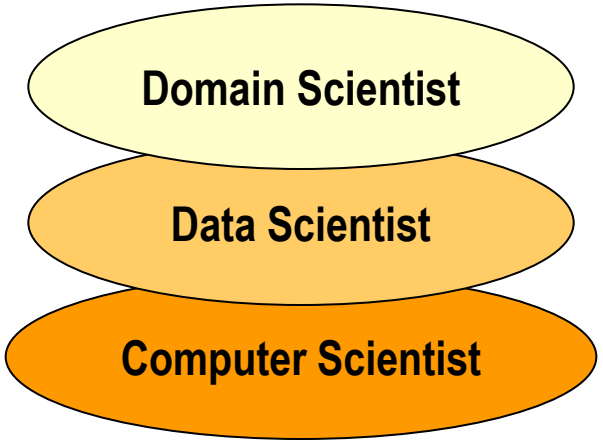
- Every thing is in the Internet
  - „the Internet“ knows state of *physical* world
- Not really new
  - ABS, emergency stop via light sensor, RFID, ...
- New: comprehensiveness, data fusion, AI...in realtime
  - T-Shirt, refrigerator, beer bottle, Fitbit, car, family, neighbours, boss, insurance, ...
- Data privacy? security?
  - Known issues, new dimensions

[Shutterstock, Forbes]

Data Scientist / Engineer

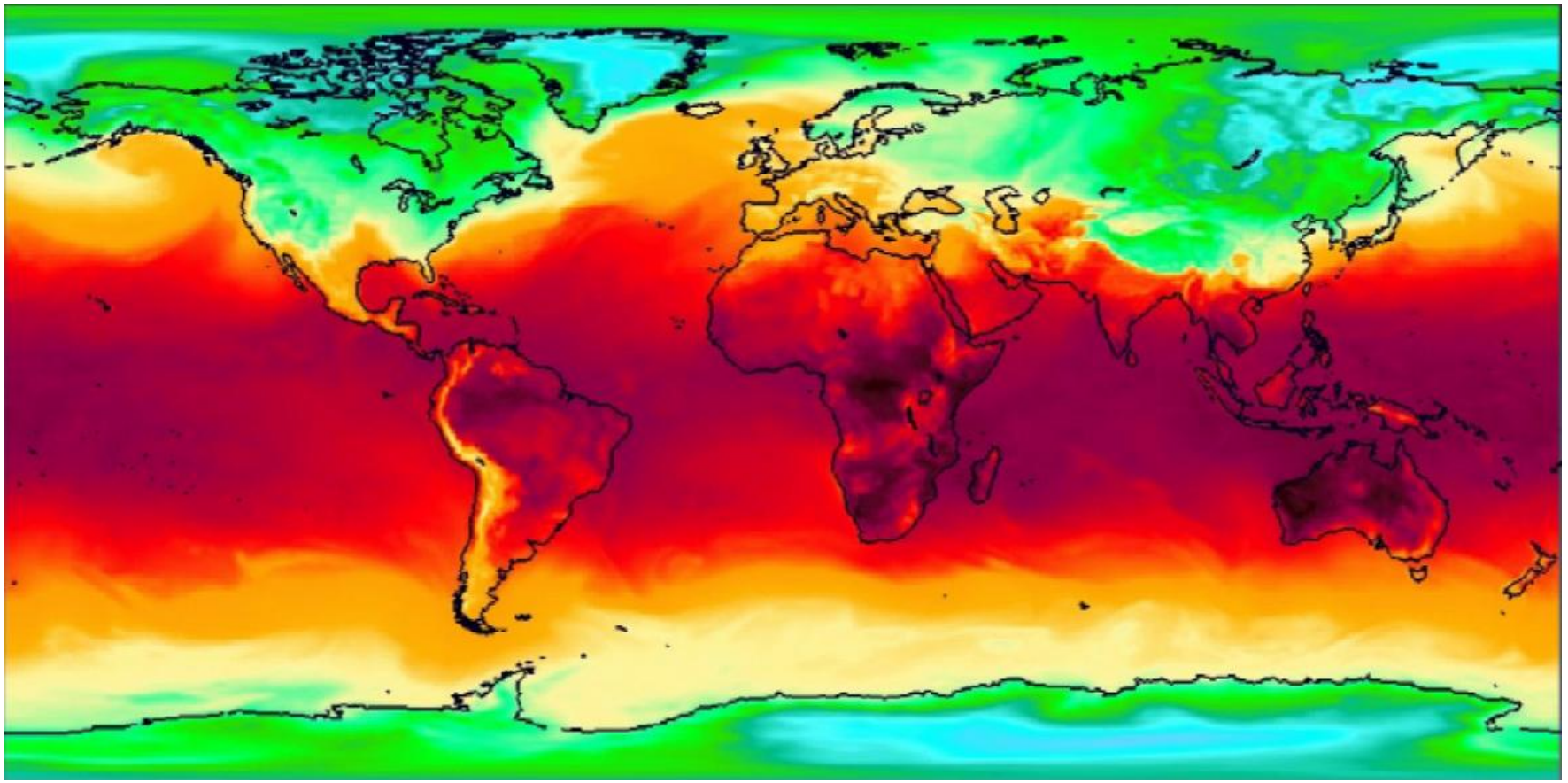
- "Data Scientist:
  The Sexiest Job of the 21st Century"
  - Harvard Business Review, 2012
- Data Scientist =
  Statistics + tool skills
  + domain expertise
  + communication
- Data Scientist ≠ Computer Scientist !





Our Research: Array Databases

Spatio-Temporal Datacubes



[DKRZ]

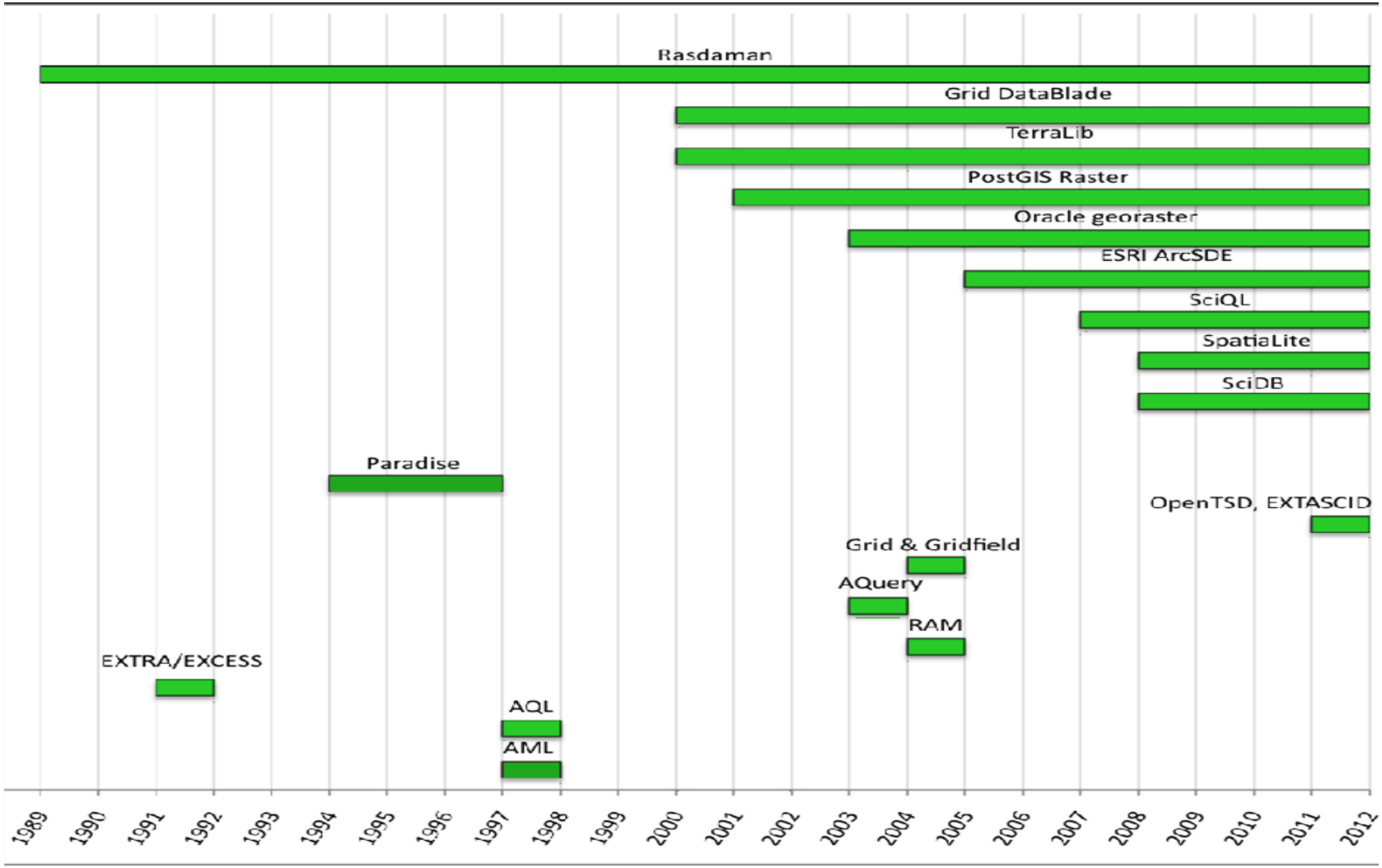
CONSTRUCTOR UNIVERSITY

rasdaman: Agile Datacube Analytics

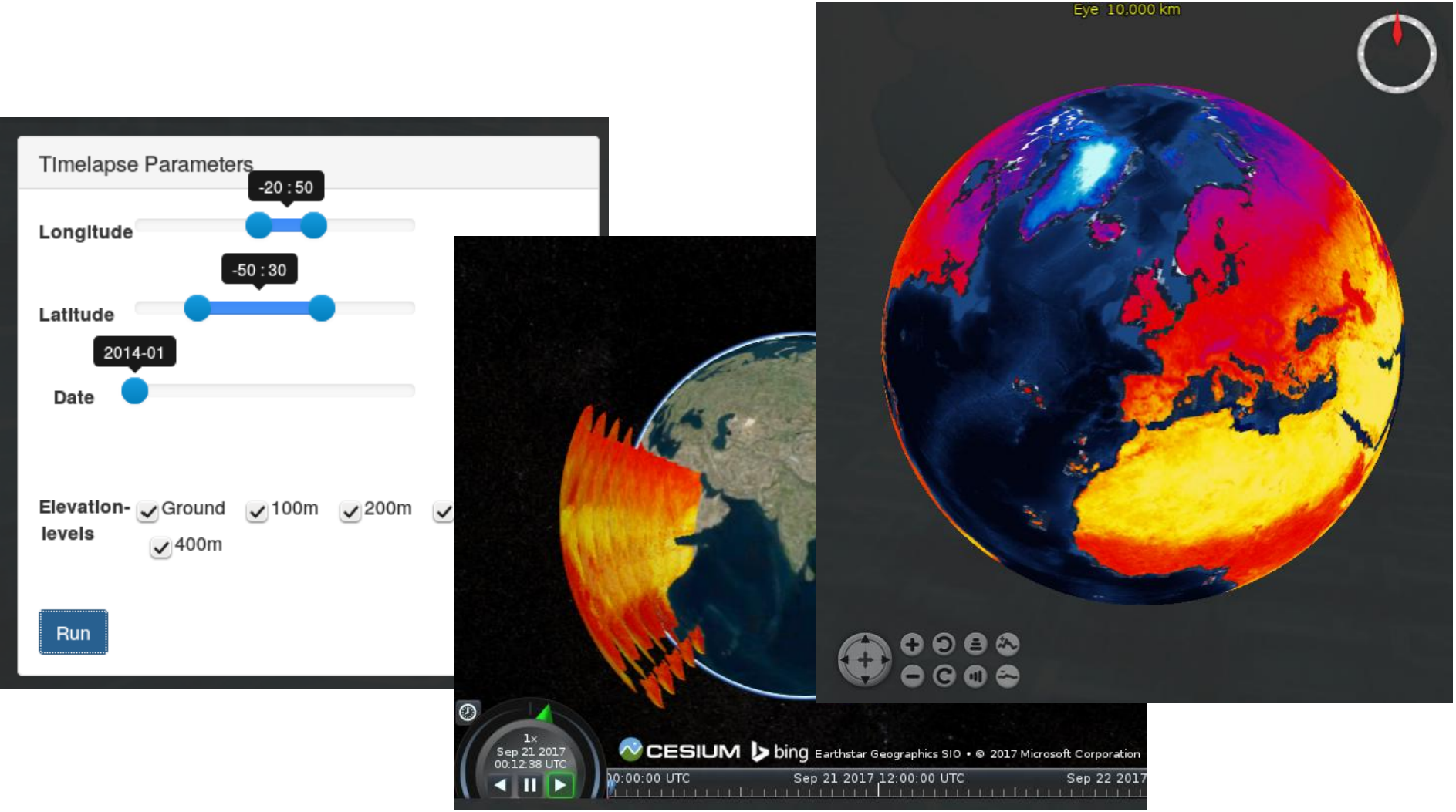
= „raster data manager“: SQL + n-D arrays

- Mature, operational, on OSGeo Live www.rasdaman.org
  - 2.5+ PB databases, 1000x parallelization, federation
- OGC, ISO, INSPIRE datacube standards crafted by rasdaman team
  - Reference Implementation

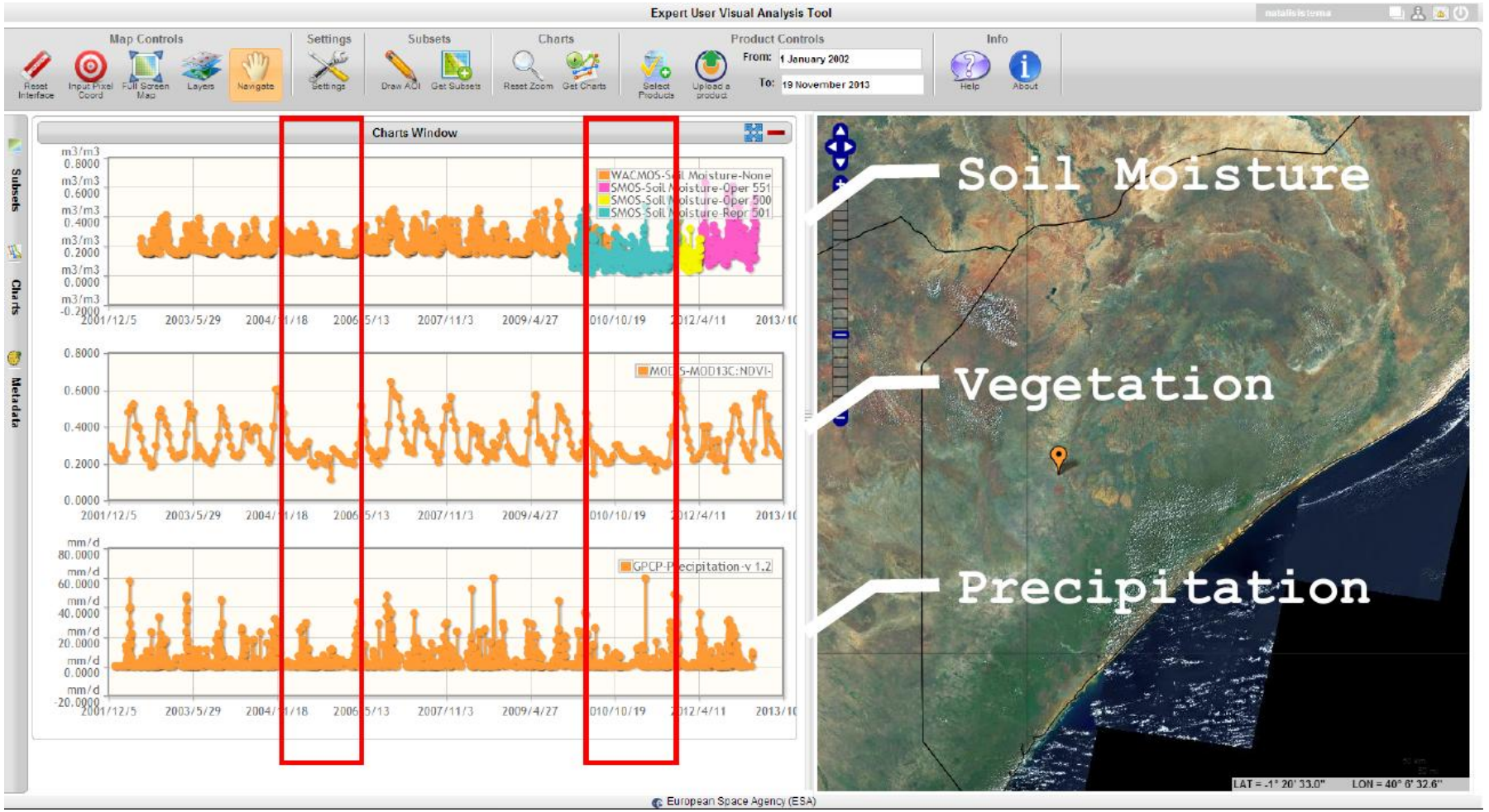
A Brief History of Array Databases



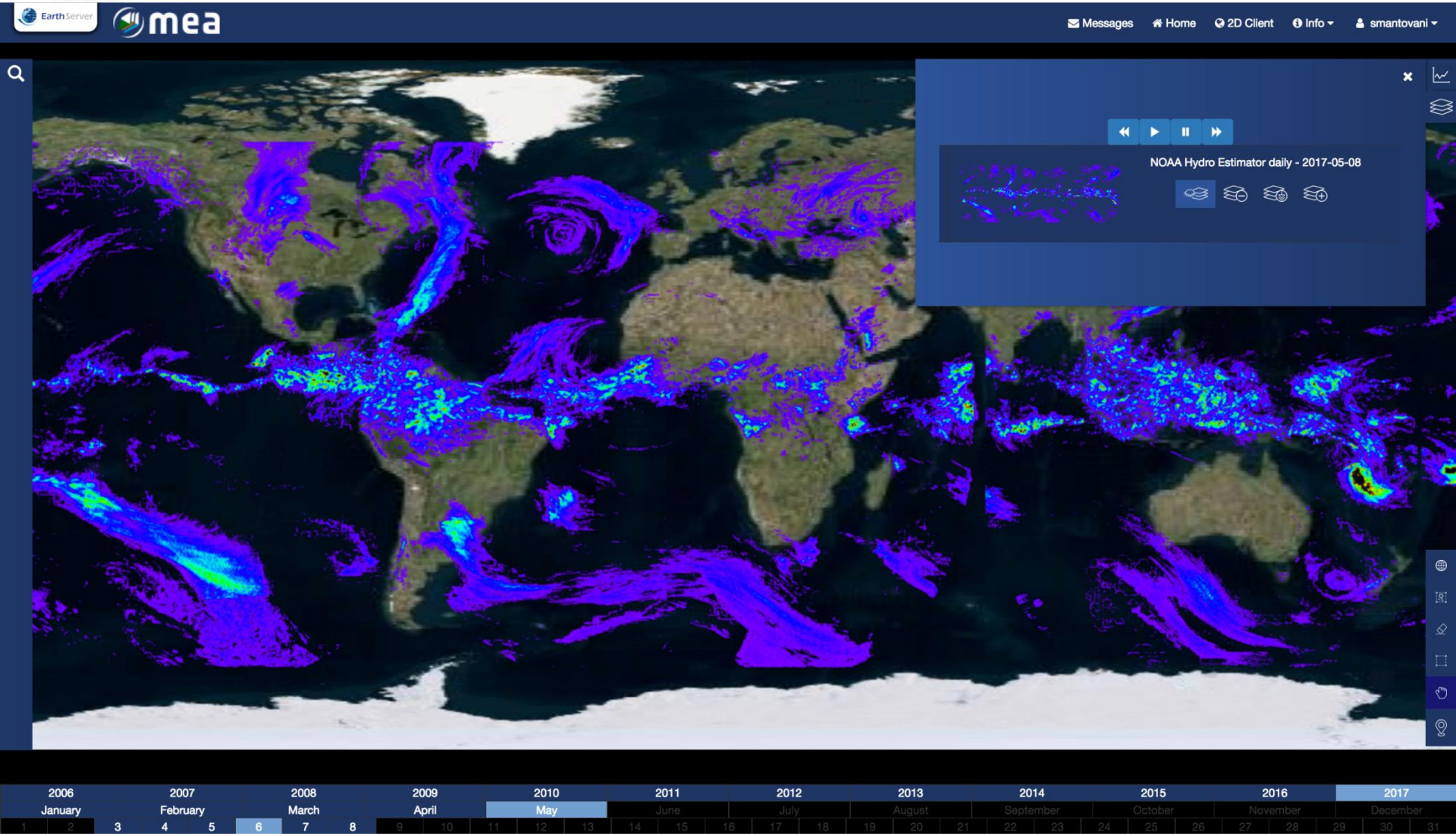
Spatio-Temporal Datacubes on Virtual Globes



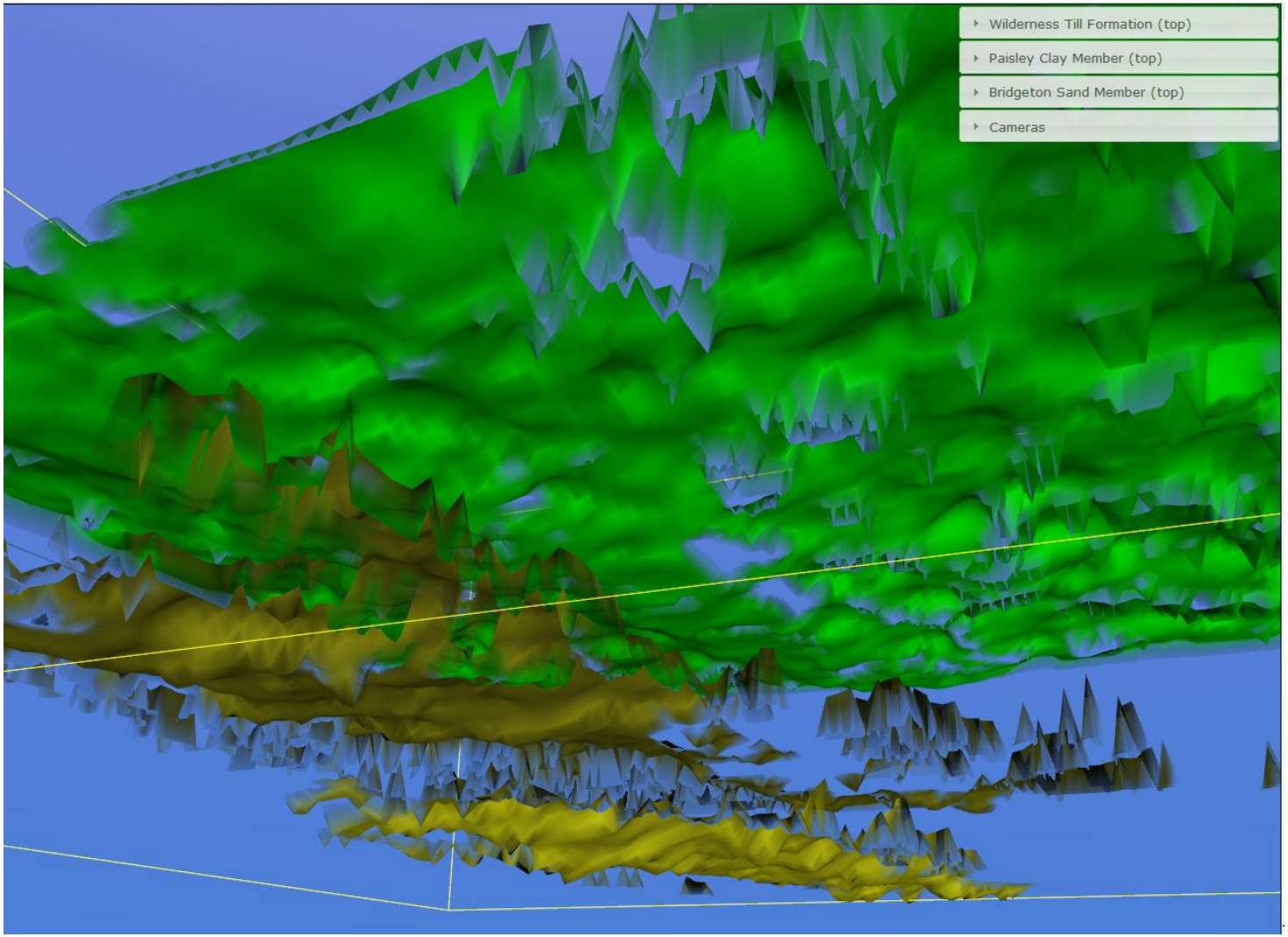
Agriculture



Daily Hydro Estimator

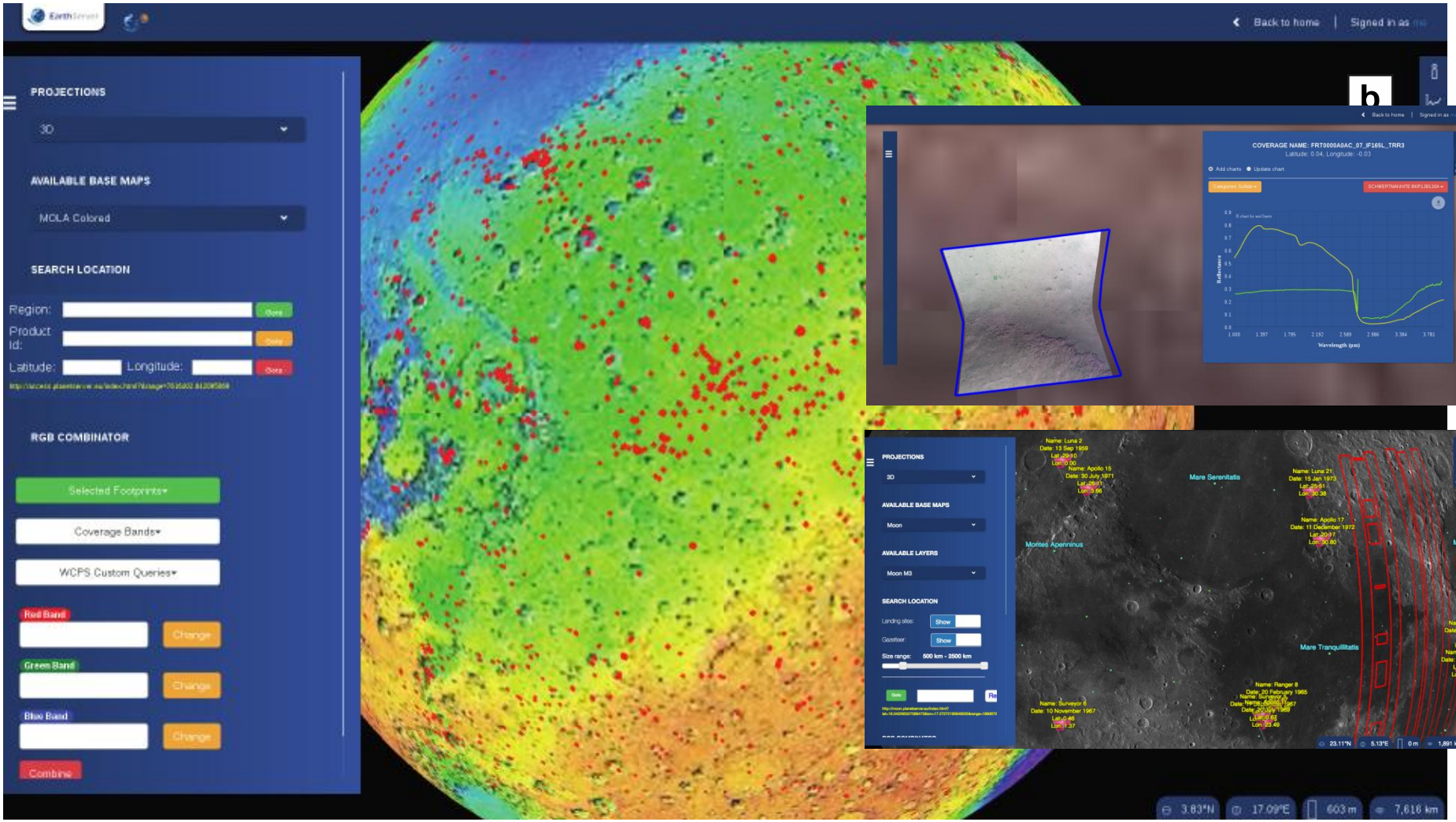


British Geological Service



[BGS 2013]

PlanetServer

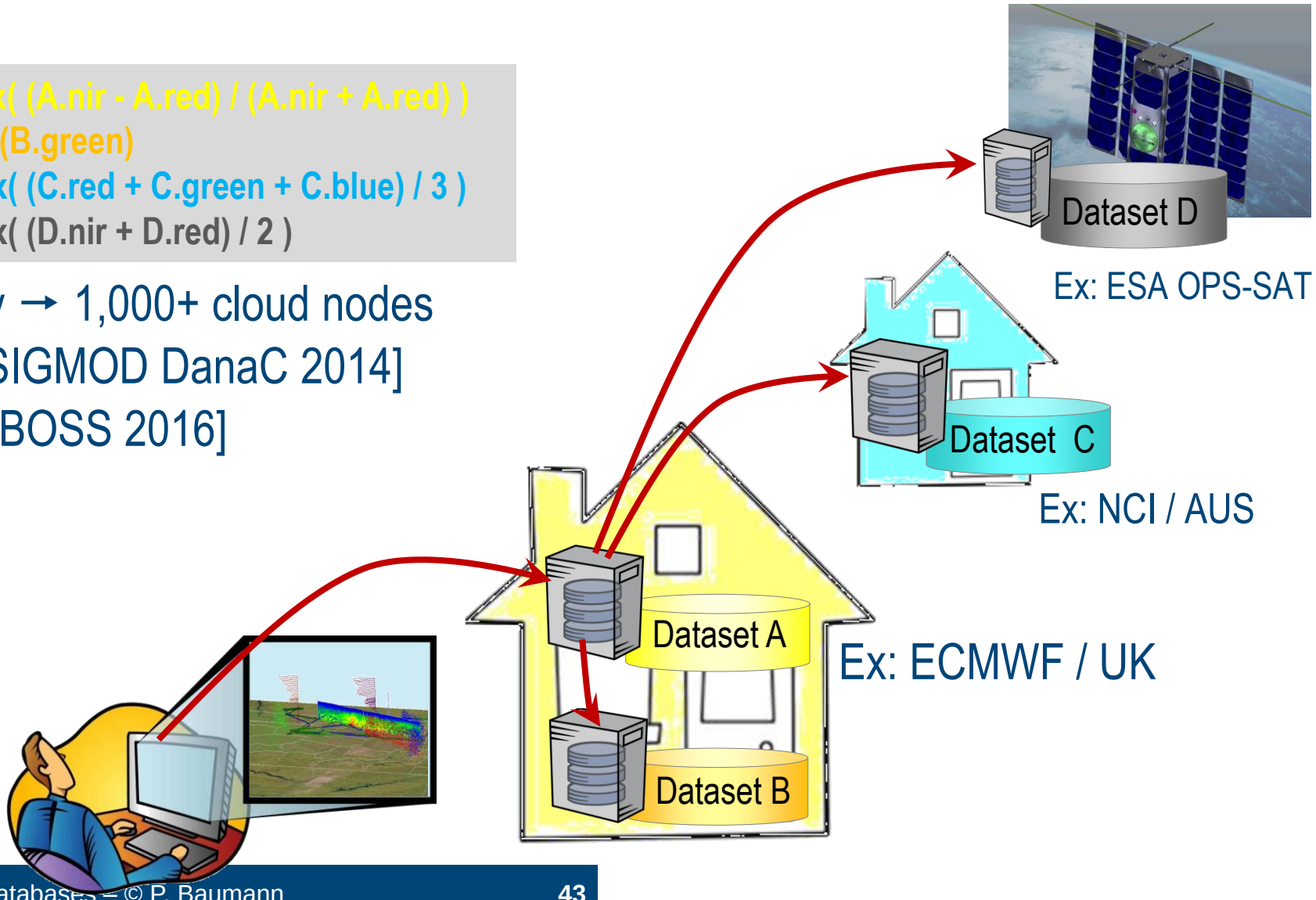


[rasdaman backend]

Parallel, Distributed Processing

```
    max( (A.nir - A.red) / (A.nir + A.red) )
  + avg(B.green)
  + max( (C.red + C.green + C.blue) / 3 )
  + max( (D.nir + D.red) / 2 )
```

1 query → 1,000+ cloud nodes
[ACM SIGMOD DanaC 2014]
[VLDB BOSS 2016]



Demo

Standards: ISO Array SQL [SSDBM 2014]

Information technology — Database languages — SQL —

Part 15:
Multi-Dimensional Arrays (SQL/MDA)

Technologies de l'information — Langages de base de données — SQL —
Partie 15: Tableaux multi-dimensionnels (SQL/MDA)

```
create table LandsatScenes(
    id: integer not null, acquired: date,
    scene: row( band1: integer, ..., band7: integer ) mdarray [ 0:4999,0:4999] )
```

```
select  id, encode(scene.band1-scene.band2)/(scene.band1+scene.band2)), „image/tiff“ )
from    LandsatScenes
where acquired between „1990-06-01“ and „1990-06-30“ and
          avg( scene.band3-scene.band4)/(scene.band3+scene.band4)) > 0
```

Big Datacube Standards (By Us)

- Open Geospatial Consortium (OGC) :
  - Spatio-Temporal „Big Geo Data“ standards suite
  - http://myogc.org/go/coveragesDWG
- ISO:
  - TC211: Spatio-Temporal „Big Geo Data“ standards suite
  - SC32: SQL/MDA („Multi-Dimensional Arrays“)
- INSPIRE:
  - Co-shaping harmonized European Spatial Data Infrastructure


Let‘s Roll!