Databases & Data Mining

Erwin M. Bakker & Stefan Manegold

e.m.bakker s.manegold

@liacs.leidenuniv.nl

https://homepages.cwi.nl/~manegold/DBDM/

http://liacs.leidenuniv.nl/~bakkerem2/dbdm/

DBDM: “Registration”

Please send an email

To: s.manegold@liacs.leidenuniv.nl
Subject: [DBDM-2018] Registration

containing the following information:

- *Your full name*
- *Your email address*
- *Your student ID*
- *Your affiliation (university)*
- *Your program / subject*

By Sunday 16 September 2018, 23:59 CEST.

DBDM: Overview

Period: September 11th - December 4th 2018 (Tuesdays)

Place: Room **312** (LIACS, Snellius building, Niels Bohrweg 1, 2333 CA Leiden)

Time: 15.30 - 17.15

ECTS: 6

Description:

The course Databases & Data Mining consists of a series of lectures in which advanced database and data mining techniques will be discussed, with applications to bioinformatics.

Grading:

There will be 2 database and 2 data mining assignments, i.e., 4 assignments in total, and a final exam (open book). The final grade will be based on a weighted average of the grades obtained for assignments P1, P2, P3, P4 and the Exam (E >5):

Final Grade = (0.5*P1 + P2 + 0.5*P3 + P4 + 3*E)/6.

DBDM: (tentative) Schedule

Date	Room	Subject (tentative)	Topic & Lecturer
11-09	312	Introduction	
18-09	312	Database Techology	Databases and Data Management for Data Mining Stefan Manegold
25-09	312	Database Techology	
02-10	312	Data Preprocessing	
09-10	*312*	*No class*	
16-10	312	Data Warehousing and OLAP	
23-10	312	Data Cube Technology	
			Data Mining Techniques and Applications Erwin Bakker
30-10	312	Basic Data Mining Algorithms I	
06-11	312	Basic Data Mining Algorithms II	
13-11	312	Advanced Data Mining Algorithms	
20-11	312	Mining in Bio-Data	
27-11	312	Graph Mining I	
04-12	312	Graph Mining II	

https://homepages.cwi.nl/~manegold/DBDM/

http://liacs.leidenuniv.nl/~bakkerem2/dbdm/

DBDM: Assignments

- 2 database assignments & 2 data mining assignments
- Will be announced individually during lectures and posted on website

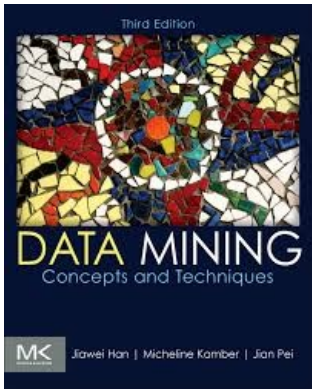
DBDM: Exam

- **open book exam:** you can take with you your book, and printed course notes (slides). *No electronic equipment is allowed, though.*

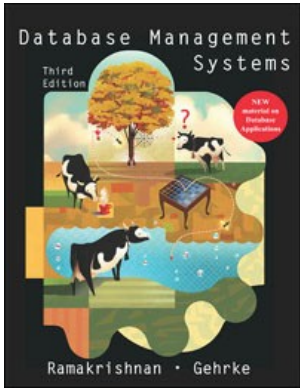
- **Materials to be studied:**
  - All content covered and discussed during lectures
    (slides will be shared).
  - More to be announced.

- **Date:** *Monday, January 7, 2019*
- **Time:** *14:00 - 17:00*
- **Place:** *Room F104, Van Steenisgebouw, Einsteinweg 2, 2333 CC Leiden*

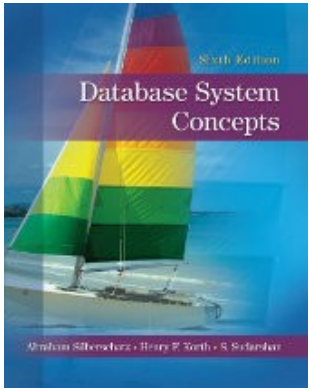
DBDM: Recommended Books



- **Data Mining:**
  - *J. Han*, *M. Kamber*, *J. Pei*. **Data Mining Concepts and Techniques (3rd Edition)**, Morgan Kaufman Publishers, July 2011 (ISBN 978-0123814791)



- **Database systems (e.g.):**
  - *Ramakrishnan*, *Gehrke*: **Database Management Systems (3rd International Edition)**, McGraw-Hill, 2003 (ISBN 0-07-246563-8)
  - *A. Silberschatz*, *H. F. Korth*, *S. Sudarshan:* **Database System Concepts (6th Edition),** McGraw-Hill, 2010 (ISBN 0-07-352332-1)



DBDM: “Registration”

Please send an email

To: s.manegold@liacs.leidenuniv.nl
Subject: [DBDM-2018] Registration

containing the following information:

- *Your full name*
- *Your email address*
- *Your student ID*
- *Your affiliation (university)*
- *Your program / subject*

By Sunday 16 September 2018, 23:59 CEST.

Databases & Data Mining

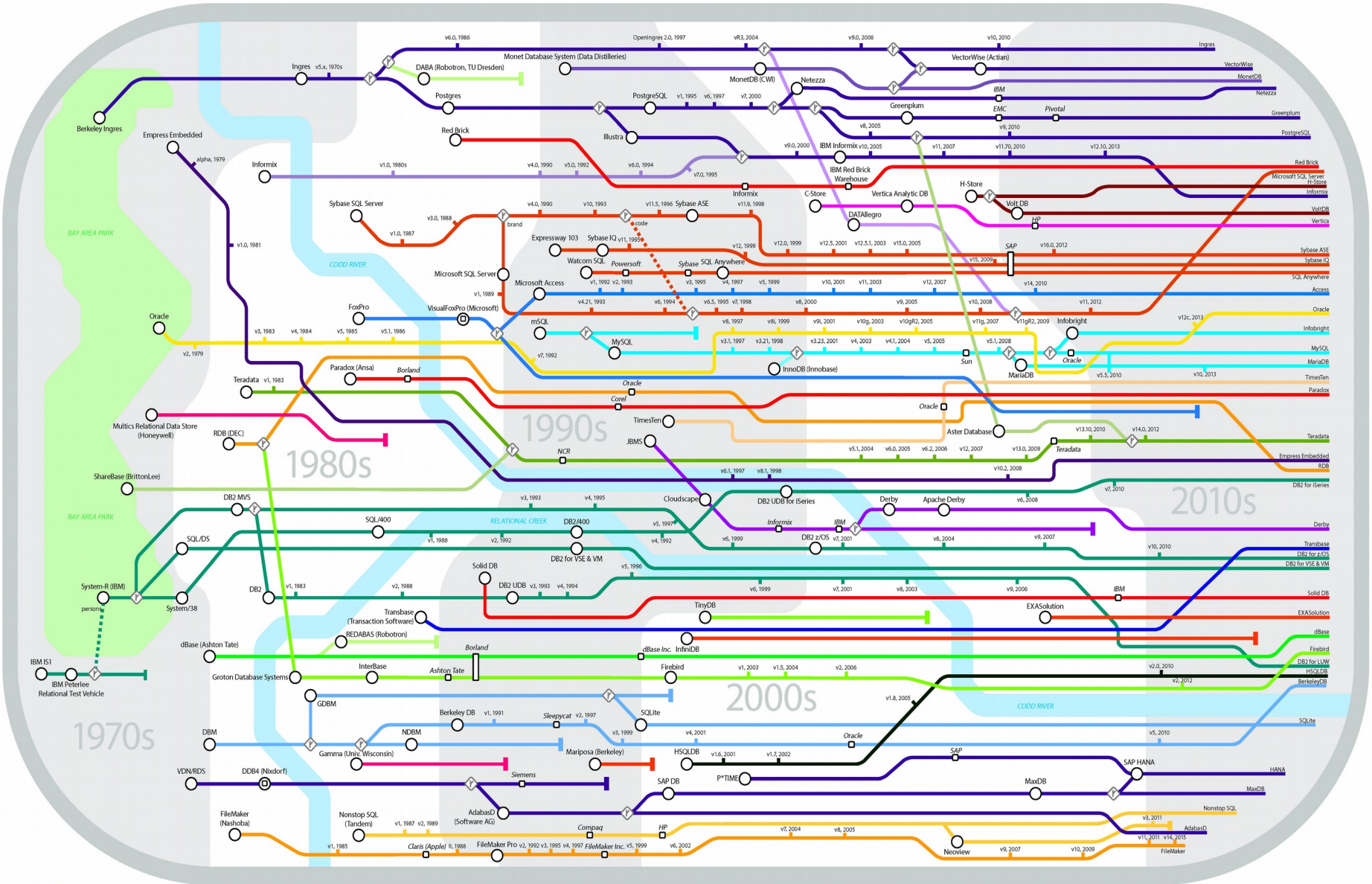
Stefan Manegold

Group leader Database Architectures
Centrum Wiskunde & Informatica (CWI)
Amsterdam
http://homepages.cwi.nl/~manegold/

http://www.monetdb.org/

Prof. Data Management (0.2 fte)
LIACS & LCDS
Faculty of Science, Leiden University

Genealogy of Relational Database Management Systems



Key to lines and symbols

Publishing Date | Acquisition | v9, 2006CC 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 5.0 - October 2015
http://www.hpi.uni-potsdam.de/naumann/projekte/rdbms_genealogy.html

Data

Data Management

Database

Data Mining

The age of Big Data

1500TB/min =
1000 full drives
per minute

= a stack of
20meter high

4000 million
TeraBytes
=
3 billion
full disk drives


SAP


Microsoft


LABS hp


Sun
ORACLE

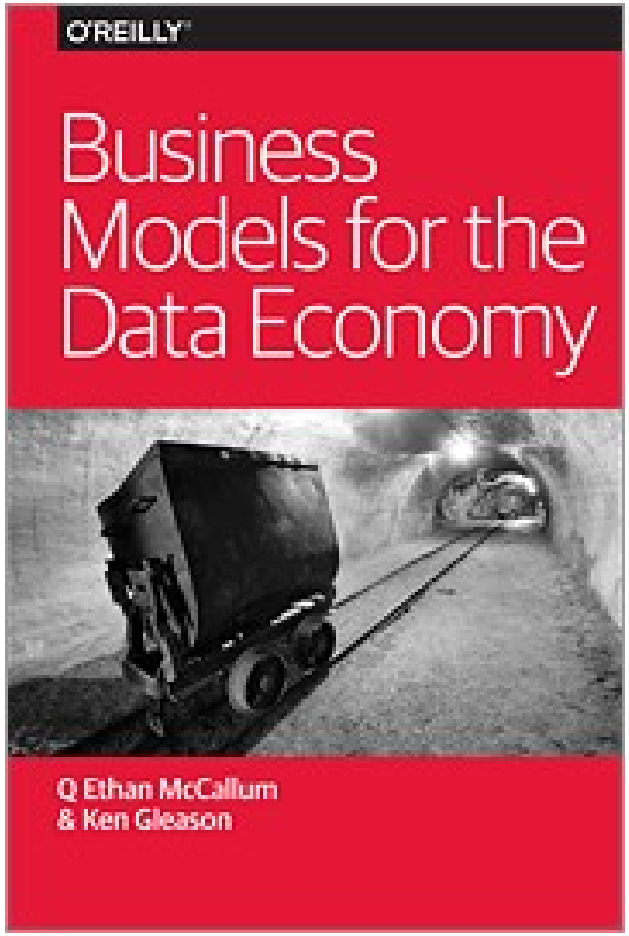

IBM


twitter


Google


facebook

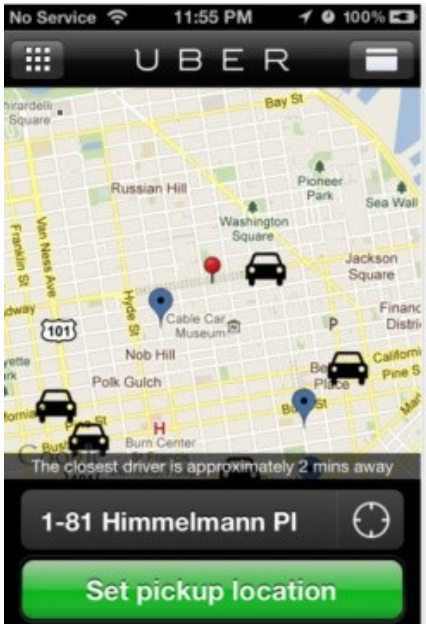
The Data Economy

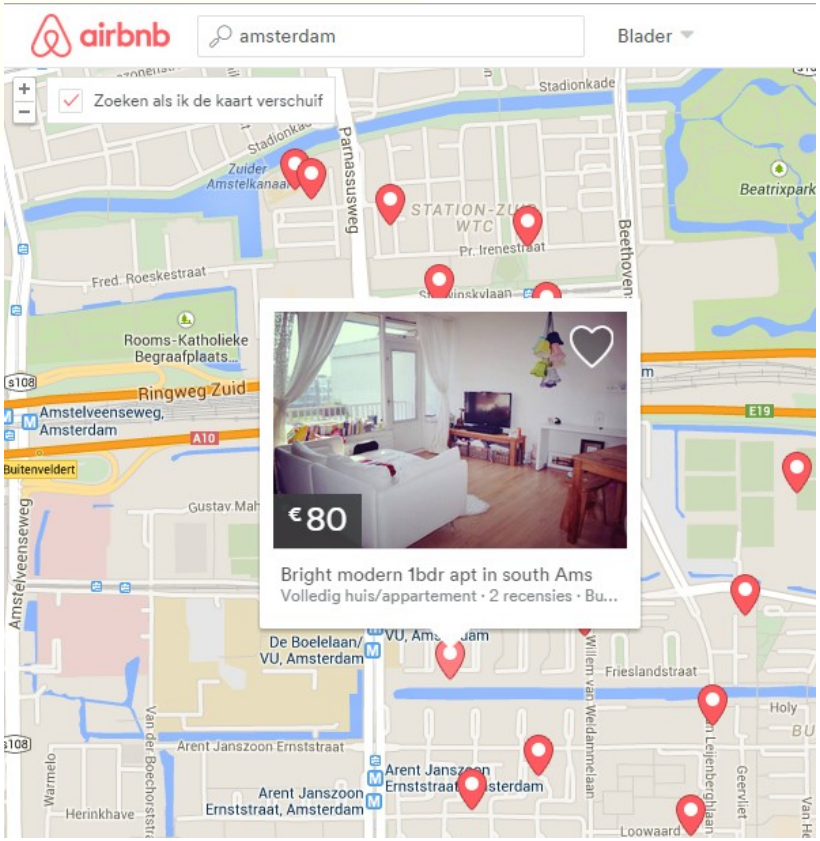


Disruptions by the Data Economy

UBER

EVERYONE'S PRIVATE DRIVER™





DBDM: Selected Challenges

GIS (LIDAR):

Massive point clouds: 640 Billion (x,y,z) points / 15 TB

=> spatial joins between point cloud and polygons

netherlands eScience center

Logistics:

> 5 trillion (10^12) GPS points (grows with >60k points/sec)

COMMIT/

Seismology:

~ 4 M files, ~ 500 GB (10x compressed)

=> Transparent data ingestion: *Data Vault*

Remote sensing:

~2 PB satellite image data

=> Array data processing: *SciQL*

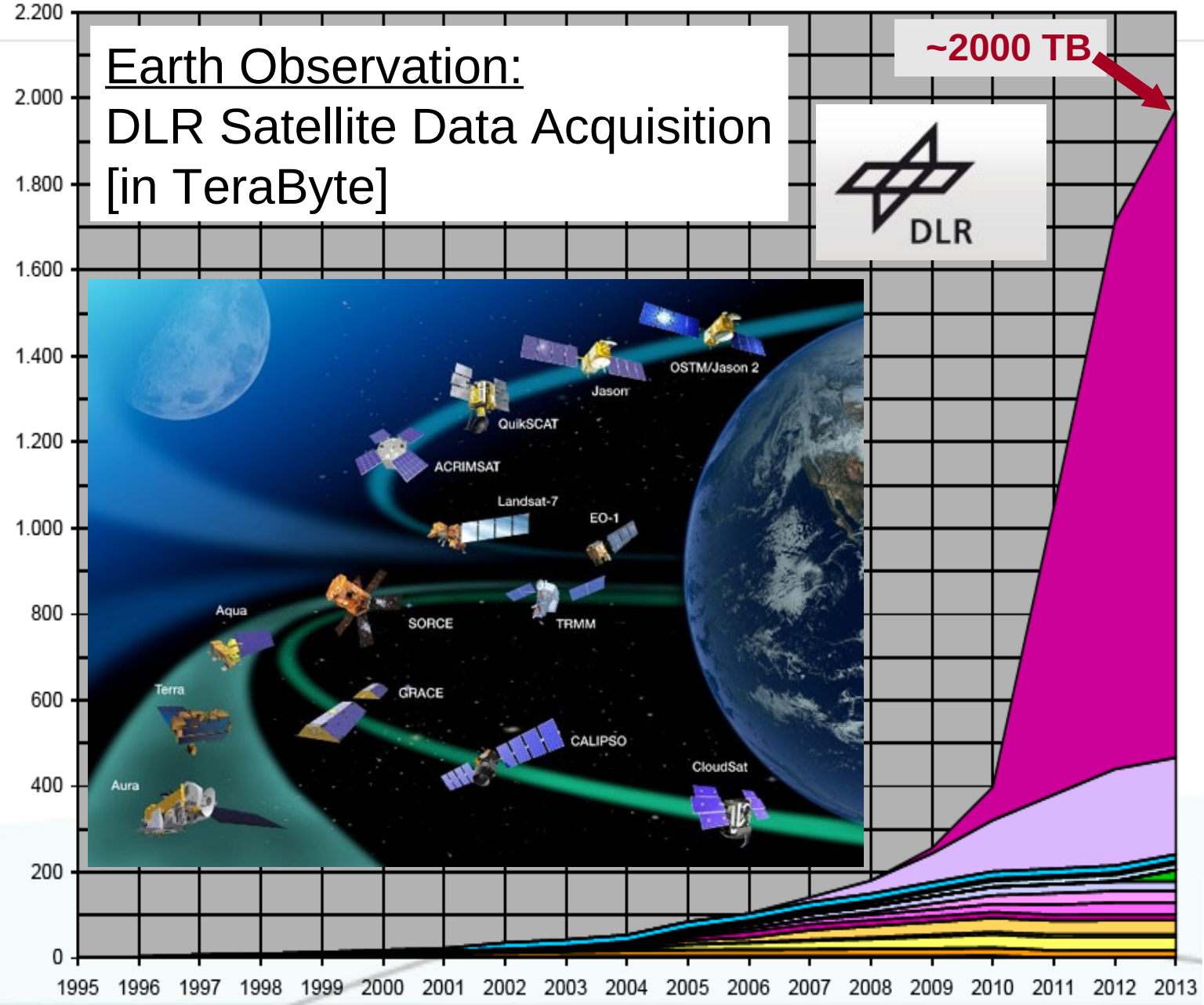
TELEIOS

Astronomy:

Raw data: *25 TB / hour*; derived data: *100 TB / year*

=> Transient detection inside DBMS

DBDM: Earth Observation



TELEIOS

LOFAR Low Frequency Array for Radio Asronomy

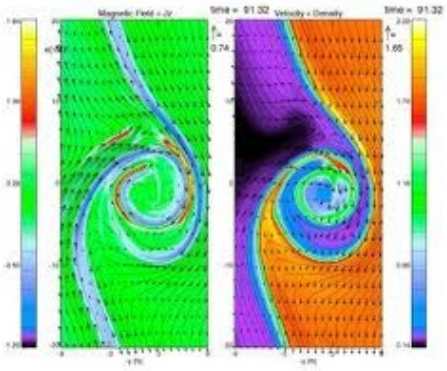
Raw data:
~25 TB / *hour*

Derived data:
~100 TB / year

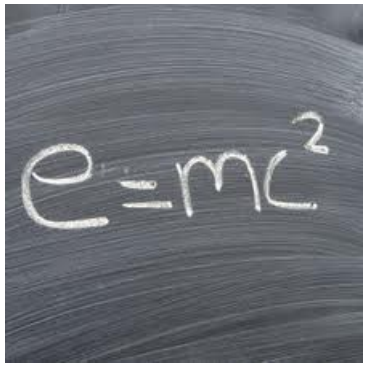
monetdb

Data Disrupting Science: Paradigm Shift in Scientific Research

collecting & analyzing data *data exploration* (eScience)

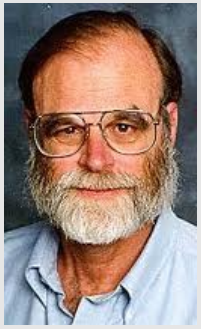


simulating
computational
3rd

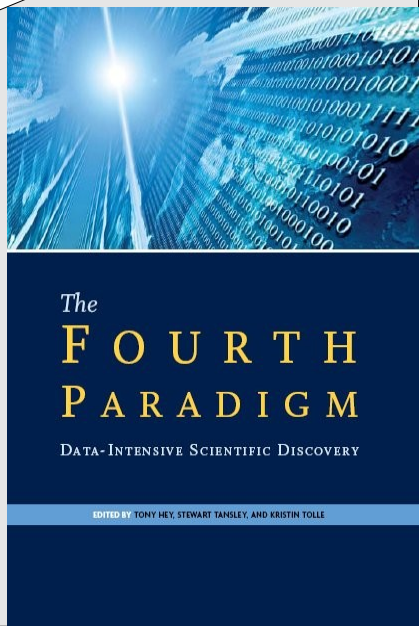


modeling
theoretical
2nd

observing
empirical
1st



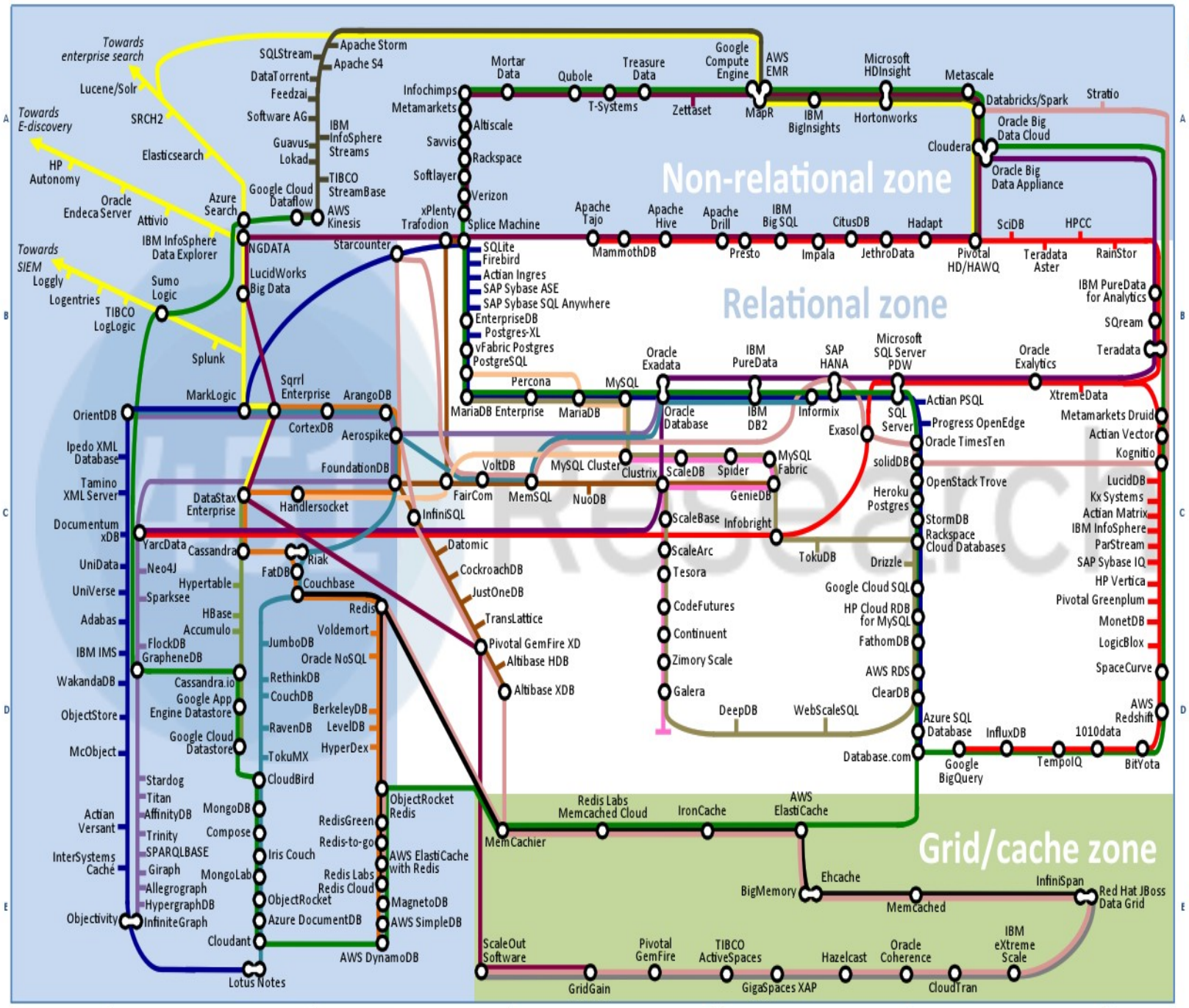
Jim Gray
(1944 - 2007)



CWI
Database Architectures

Universiteit Leiden

Data Management & Data Mining


Towards enterprise search
Towards E-discovery
Towards SIEM
Lucene/Solr
SRCH2
Elasticsearch
HP Autonomy
Oracle Endeca Server
Attivio
IBM InfoSphere Data Explorer
Loggly
Logentries
Sumo Logic
TIBCO LogLogic
Splunk
Azure Search
NGDATA
LucidWorks Big Data
MarkLogic
SQLStream
Apache Storm
Apache S4
DataTorrent
Feedzai
Software AG
IBM InfoSphere Streams
Guavus
Lokad
TIBCO StreamBase
AWS Kinesis
Google Cloud Dataflow
Sqrrl Enterprise
ArangoDB
CortexDB
Aerospike
FoundationDB
Handlersocket
DataStax Enterprise
InfiniSQL
Starcounter
Infochimps
Metamarkets
Savvis
Softlayer
Verizon
xPlenty
Trafodion
Splice Machine
Mortar Data
Qubole
Treasure Data
Google Compute Engine
AWS EMR
MapR
Zettaset
T-Systems
IBM BigInsights
Microsoft HDInsight
Hortonworks
Metascale
Databricks/Spark
Stratio
Cloudera
Oracle Big Data Cloud
Oracle Big Data Appliance
Altiscale
Rackspace
Non-relational zone
Apache Tajo
Apache Hive
Apache Drill
IBM Big SQL
CitusDB
Hadapt
MammothDB
Presto
Impala
JethroData
Pivotal HD/HAWQ
SciDB
HPCC
Teradata Aster
RainStor
Relational zone
SQLite
Firebird
Actian Ingres
SAP Sybase ASE
SAP Sybase SQL Anywhere
EnterpriseDB
Postgres-XL
vFabric Postgres
PostgreSQL
MariaDB Enterprise
Percona
MariaDB
MySQL
Oracle Exadata
Oracle Database
IBM PureData
IBM DB2
SAP HANA
Informix
Exasol
Microsoft SQL Server PDW
SQL Server
Oracle Exalytics
Teradata
IBM PureData for Analytics
SQream
XtremeData
Actian PSQL
Progress OpenEdge
Oracle TimesTen
solidDB
OpenStack Trove
Heroku Postgres
StormDB
Rackspace Cloud Databases
Drizzle
Google Cloud SQL
HP Cloud RDB for MySQL
FathomDB
AWS RDS
ClearDB
Azure SQL Database
Database.com
InfluxDB
Google BigQuery
1010data
TempoIQ
BitYota
AWS Redshift
SpaceCurve
LogicBlox
MonetDB
Pivotal Greenplum
HP Vertica
SAP Sybase IQ
ParStream
IBM InfoSphere
Actian Matrix
Kx Systems
LucidDB
Kognitio
Actian Vector
Metamarkets Druid
MySQL Cluster
Clustrix
ScaleDB
Spider
MySQL Fabric
GenieDB
Infobright
ScaleBase
ScaleArc
Tesora
CodeFutures
Continuent
Zimory Scale
Galera
DeepDB
WebScaleSQL
TokuDB
VoltDB
FairCom
MemSQL
NuoDB
Datomic
CockroachDB
JustOneDB
TransLattice
Pivotal GemFire XD
Altibase HDB
Altibase XDB
OrientDB
Ipedo XML Database
Tamino XML Server
Documentum xDB
UniData
UniVerse
Adabas
IBM IMS
WakandaDB
ObjectStore
McObject
Actian Versant
InterSystems Caché
Objectivity
YarcData
Cassandra
Riak
FatDB
Couchbase
Neo4J
Hypertable
Sparksee
HBase
Accumulo
FlockDB
GrapheneDB
Cassandra.io
Google App Engine Datastore
Google Cloud Datastore
Redis
Voldemort
JumboDB
Oracle NoSQL
RethinkDB
CouchDB
RavenDB
TokuMX
CloudBird
BerkeleyDB
LevelDB
HyperDex
Stardog
Titan
AffinityDB
Trinity
SPARQLBASE
Giraph
Allegrograph
HypergraphDB
InfiniteGraph
MongoDB
Compose
Iris Couch
MongoLab
ObjectRocket
Azure DocumentDB
Cloudant
Lotus Notes
RedisGreen
Redis-to-go
Redis Labs Redis Cloud
AWS DynamoDB
ObjectRocket Redis
AWS ElastiCache with Redis
MagnetoDB
AWS SimpleDB
Grid/cache zone
Redis Labs Memcached Cloud
IronCache
AWS ElastiCache
MemCachier
BigMemory
Ehcache
Memcached
InfiniSpan
Red Hat JBoss Data Grid
ScaleOut Software
GridGain
Pivotal GemFire
TIBCO ActiveSpaces
GigaSpaces XAP
Hazelcast
Oracle Coherence
CloudTran
IBM eXtreme Scale
451 Research
Data Platforms Map
October 2014
Key:
General purpose
Specialist analytic
-as-a-Service
BigTables
Graph
Document
Key value stores
Key value direct access
Hadoop
MySQL ecosystem
Advanced clustering/sharding
New SQL databases
Data caching
Data grid
Search
Appliances
In-memory
Stream processing
https://451research.com/dashboard/dpa
© 2014 by 451 Research LLC. All rights reserved

BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO