# Datenanalyse & Big Data

Harald Schilly, Wifi Wien, 20160616

# Geschichte der Technologien

# Mathematik/Statistik

- Wahrscheinlichkeitstheorie: neue Ansätze und Methoden, philosophische Überlegungen, …
- Wissenschafts&Erkenntnistheorie (Falsifikationismus)
- Bayes Statistik (Rechenaufwändiger, subjektivistischer Wahrscheinlichkeitsbegriff, Vorwissen, Verteilungen, Laplace 1812)
- Bootstrapping (1979)
- Optimierung: LP, MILP, konvex, NLP
- Graphentheorie

# ICT

- Verteiltes Rechnen (Amdahl's Law)
- Funktionales Programmieren / immutable Datenstrukturen
- Approximative Algorithmen
- Relationale Datenbanken und NoSQL
  - guaranteed vs. eventual consistency
- SQL und DataFrames (R)
- Moores Law / Kryders Law
- NUMA Architektur
- Container & Cluster Management (Kubernetes, Mesos, ...)

# Visualisierung

Kommunikation der Ergebnisse, Überblick, sinnliche Erfassung

- John Tukey
- Edward Tufte
- Leland Wilkinson's Grammar of Graphics

# Multidisciplinary: Data Science

**Extraktion von Information aus Daten (information retrival)**

Wissenschaften

- Mathematik
- Statistik/Wahrscheinlichkeitstheorie
- Informationstheorie
- Signalverarbeitung
- Mustererkennung
- Optimierung
- Visualisierung/Design

ICT

- Programmierung
- Datenbanken
- Machine Learning
- Storage
- Cluster (HPC/HTC)
- Streaming

**Computer Science**

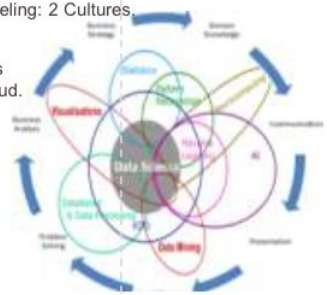
- Liebniz – Binary Logic.
- Babbage, Lovelace
- Boolean Algebra
- Punch cards.
- Turing machines
- Information Theory
- Weiner & Cybernetics
- Von Neumann Architecture.
- First IBM Computers
- DBMS.
- Sort & Search Algorithms – Dijkstra, Kruskal, Shell Sort, …
- Heuristics – Simulated Annealing, …
- Graph Algorithms
- Multigrid methods
- Tree based methods.
- Text/ string search
- 1974 Peter Naur "Concise Survey of Computer Methods", **Data Science, Datalogy**
- Knuth – Art of Computer Programming.
- 1989 First KDD Workshop
- Gregory Piatetsky-Shapiro.
- Database Marketing
- Data Mining, Knowledge Discovery
- "Data science, classification, and related methods."

**Data Technology**

- Catrography
- Astronomical Charts.
- William Playfair
- Charles Minard
- Florence Nightingale.
- Removable Disk drives
- Relational DBMS.
- Desktop, floppy
- SQL, OOP
- High level languages.
- William Cleveland: Data Science
- Leo Breimann: Statistical Modeling: 2 Cultures

**Visualization**

- John Tukey
- Jacques Bertin.
- Edward Tufte.
- Grammar of Graphics
- Word Cloud, Tag Cloud.



**Mathematics/ OR**

- Optimization Methods
- Fourier and other transforms
- Matrix & Generalizations
- Non-euclidean geometries.
- Applications to Military, manufacturing, Communications.
- Networks
- Assignment Problems
- Automation
- Scheduling
- Calculus
- Logarithms
- Newton-Raphson.
- **1962** John W. Tukey, Future of Data Analysis
- 1976 – SAS Institute
- 1977 The International Association for Statistical Computing (IASC).
- Decision Science
- Pattern recognition
- Machine learning.

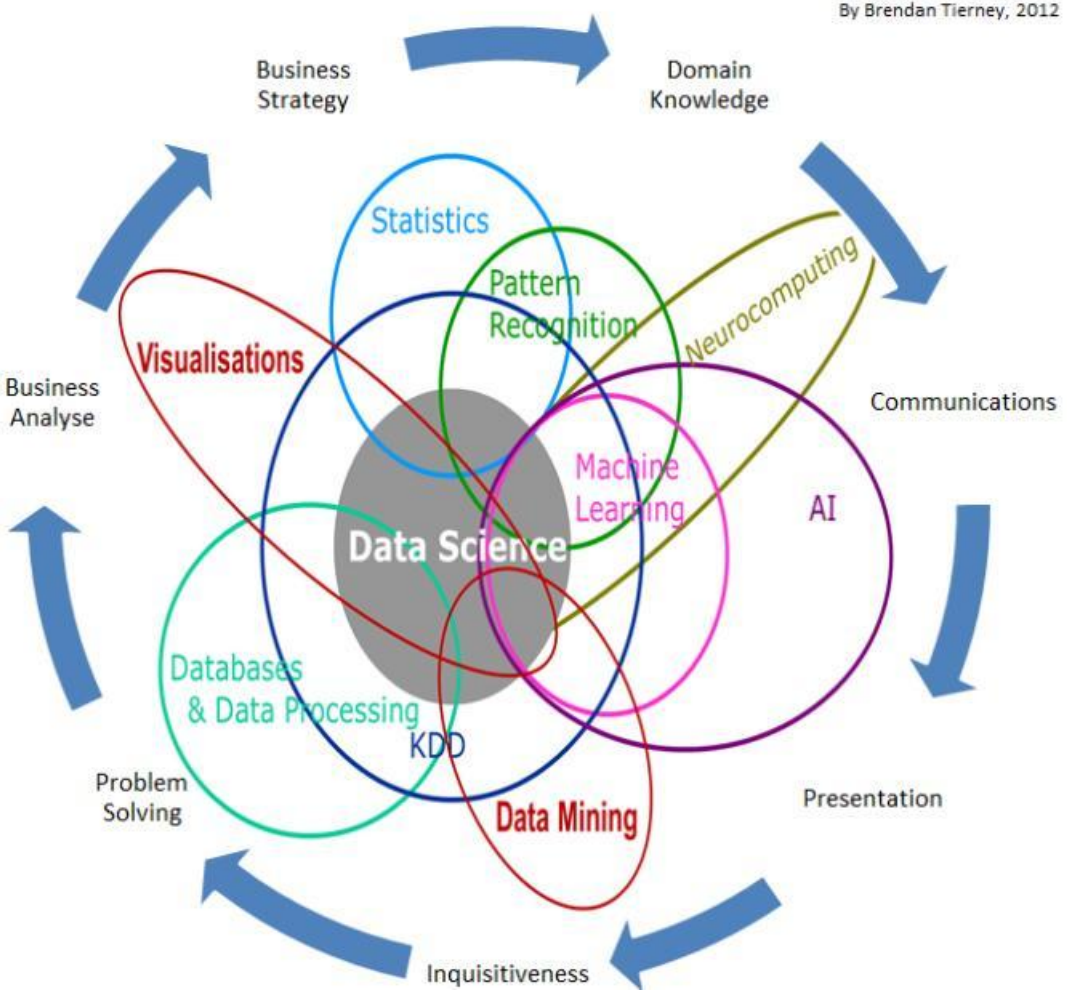
**Statistics**

- Theoretical Foundations of Modern Stats
- Hypothesis, DOE
- Mathematical Statistics.
- Probability
- Correlation
- Bayes Theorem.
- Regression, Least Squares
- Time Series.
- Bayesian Methods
- Time Series Methods (Box Cox, Survival, etc.)
- Stochastic Methods.
- Simulation, Markov
- Computational Statistics.

Pre 1800s | 1800-1900 | 1900-1940 | 1940-1960 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010

# Data Science Is Multidisciplinary

By Brendan Tierney, 2012



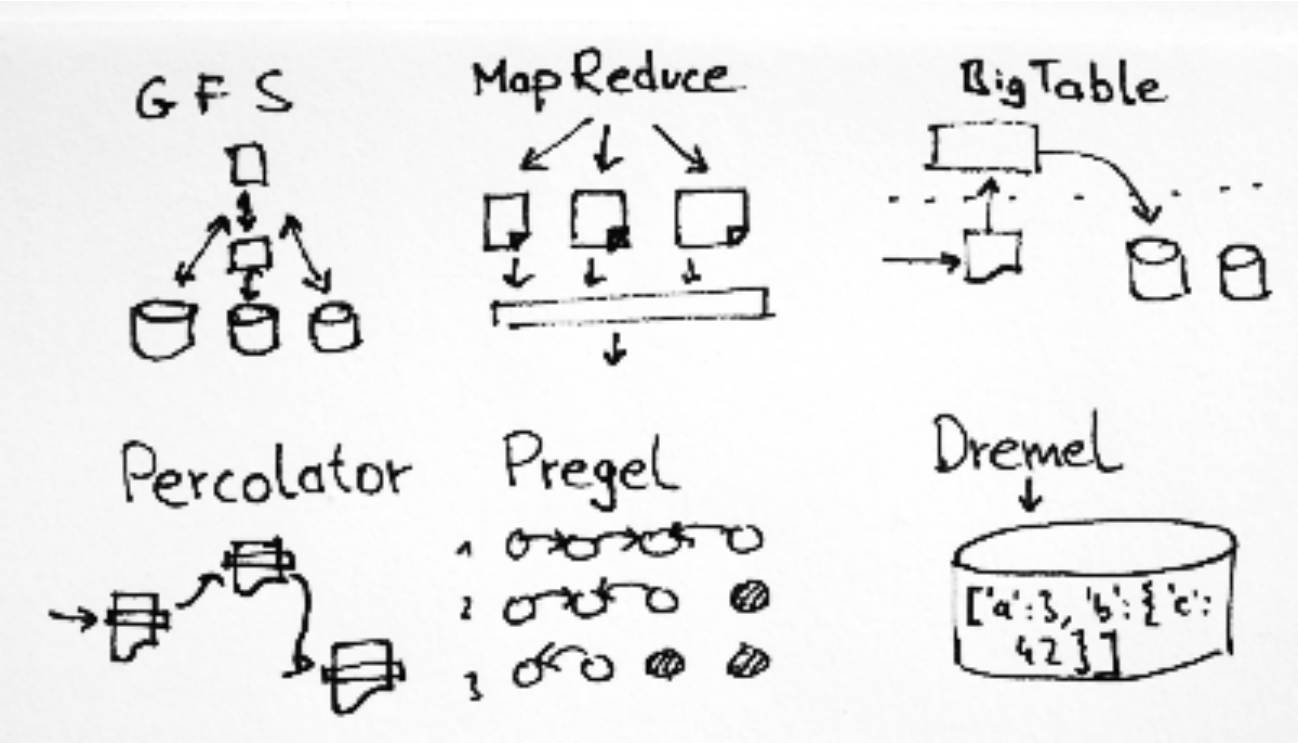
# Industrielle Forschung

(zuvor militärische)

# Industry Research / Google & Co.

- Google Filesystem 2003: dx.doi.org/10.1145%2F945445.945450
- MapReduce 2004: research.google.com/archive/mapreduce.html
  - `Map(k1,v1) → list(k2,v2)` und `Reduce(k2, list (v2)) → list(v3)`
- BigTable 2006: research.google.com/archive/bigtable.html (u.a. columnar storage)
- Pregel 2010: dl.acm.org/citation.cfm?id=1807184
- Dremel 2010: Interactive Analysis of Web-Scale Datasets (BigQuery)
- Stream Processing Frameworks:
  Perculator (Google Index), S4 (Yahoo!), Storm (Twitter)
- TensorFlow 2015: tensorflow.org (whitepaper)

# Industry Research / Google & co




http://blog.mikiobraun.de/2013/02/big-data-beyond-map-reduce-googles-pap

# Software

unvollständiger Überblick

# Software Lösungen / Distributed Computing

- MPI - Klassisch, Rechenintensiv
  - mpich
- Cassandra, HBase, Hive, ...
  - inspiriert von BigTable
- Hadoop: hadoop.apache.org
  - HDFS (GFS, ...)
  - (mit Oozie) von Yahoo!
- **Spark**: spark.apache.org
  - Verallgemeinertes Modell
  - umfasst ETL, Computation, GraphX, Machine-Learning und Stream Processing.
- Giraph: giraph.apache.org
  - Pregel, bei Facebook für Graph der Freundschaften

# Software Lösungen: Spark (2011)

- **Idee**: Hadoop ist langsam wegen Festplatte → nützen wir den Speicher!
- Handling von Datenverarbeitungsschritten und Fehlern? → **RDDs**!
- High-Level API zur Beschreibung der Schritte, Caching, und Realisierung.
- Framework zum verteilten Ausführen, mit Management, etc. (aufgesetzt auf Zookeeper, Mesos)
- Gewann Daytona GraySort contest

  (3x schneller, 10x weniger Nodes als Hadoop)
- Machine Learning, Graph Analyse, Streaming, etc.

# Software Lösungen:

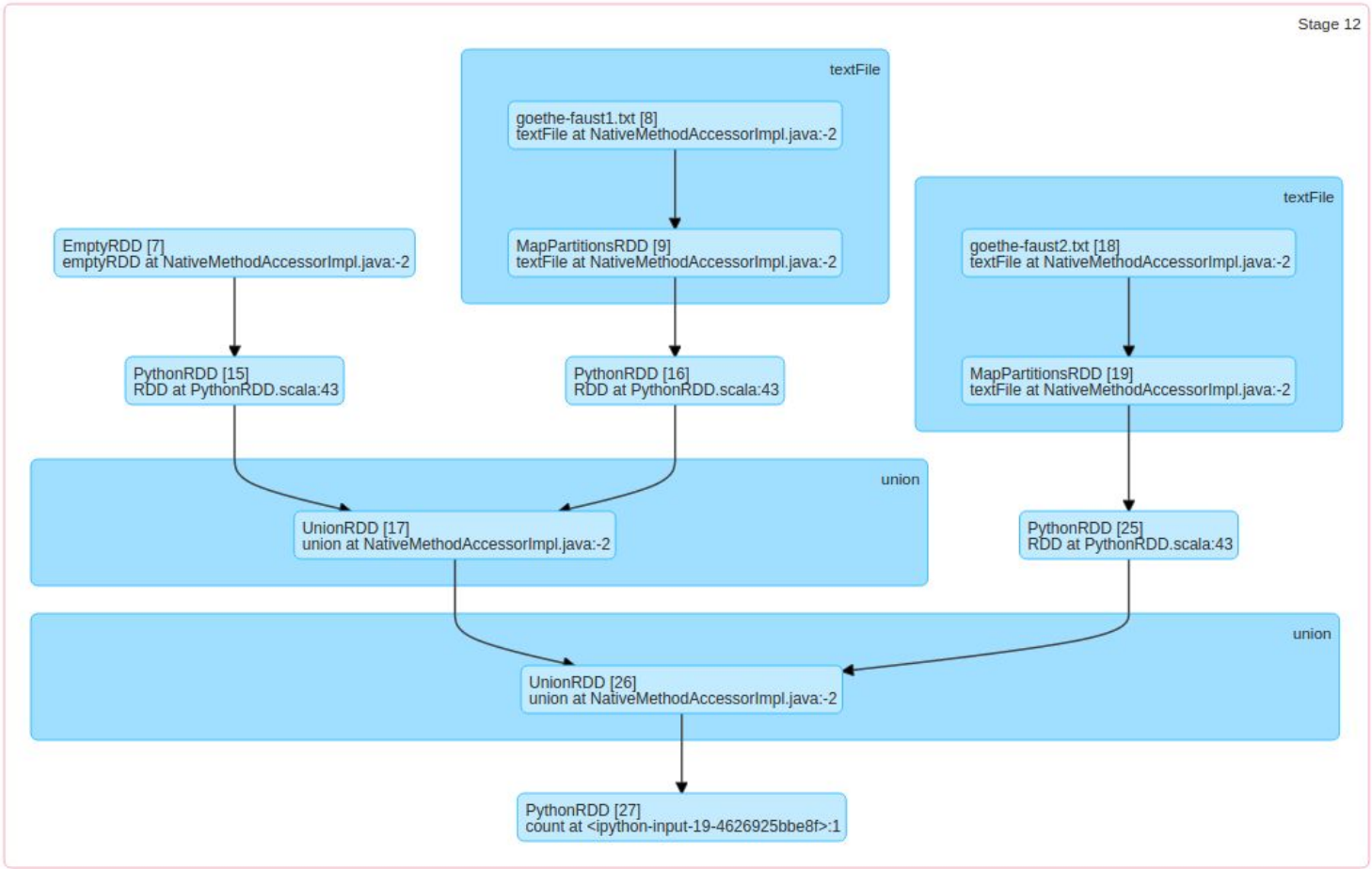


- Start 2011
- RDDs Initialisierung: `range`, `parallelize`, oder `textFile` (auch `hdfs://`)
- Zwei funktionale Grundbausteine:
  - **Transformationen**: RDD → RDD

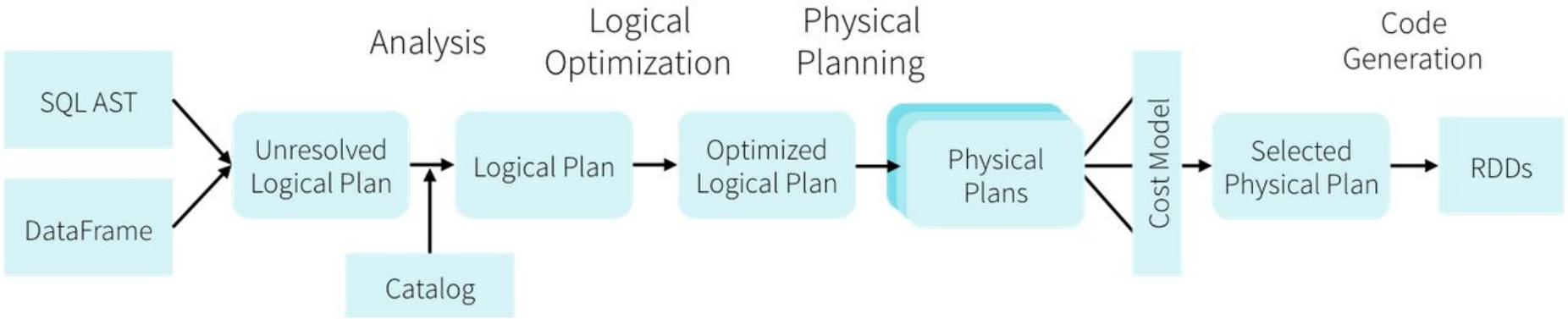
    **`map`**`, flatMap, `**`filter`**`, sample, union, intersection, distinct, groupByKey/ reduceByKey/ aggregateByKey/ sortByKey, join, cartesian, pipe, ...`
  - **Aktionen**: RDD → Value

    **collect**, **count**, **reduce**, first / take / takeOrdered, foreach, …
- **Extras**: `cache, saveAs..., sc.accumulator, sc.broadcast, ...`

# Software Lösungen: Spark - Lineage Graph



# Software Lösungen: Spark - Query Execution



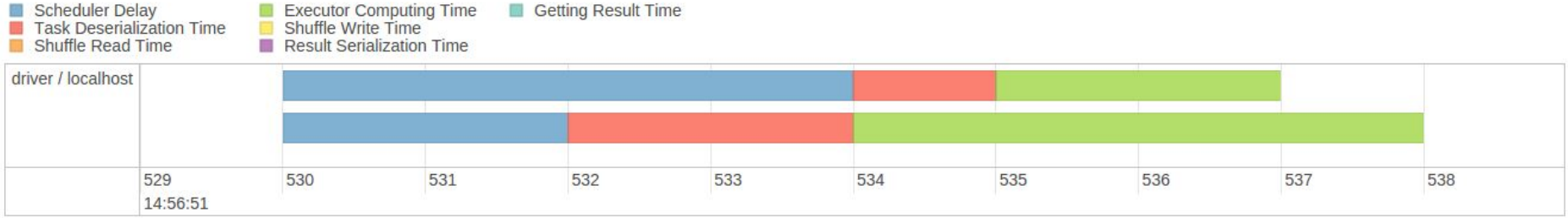

https://databricks.com/blog/

# Details for Stage 30 (Attempt 0)

**Total Time Across All Tasks:** 6 ms
**Locality Level Summary:** Process local: 2
**Input Size / Records:** 2.7 KB / 40

- DAG Visualization
- Show Additional Metrics
- Event Timeline

Enable zooming




## Summary Metrics for 2 Completed Tasks

| Metric | Min | 25th percentile | Median | 75th percentile | Max |
|---|---|---|---|---|---|
| Duration | 2 ms | 2 ms | 4 ms | 4 ms | 4 ms |
| GC Time | 0 ms | 0 ms | 0 ms | 0 ms | 0 ms |
| Input Size / Records | 1376.0 B / 20 | 1376.0 B / 20 | 1376.0 B / 20 | 1376.0 B / 20 | 1376.0 B / 20 |

## Aggregated Metrics by Executor

| Executor ID ▲ | Address | Task Time | Total Tasks | Failed Tasks | Succeeded Tasks | Input Size / Records |
|---|---|---|---|---|---|---|
| driver | localhost:55233 | 15 ms | 2 | 0 | 2 | 2.7 KB / 40 |

## Tasks

| Index ▲ | ID | Attempt | Status | Locality Level | Executor ID / Host | Launch Time | Duration | GC Time | Input Size / Records | Errors |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 51 | 0 | SUCCESS | PROCESS_LOCAL | driver / localhost | 2016/01/13 14:56:51 | 2 ms | | 1376.0 B (memory) / 20 | |

# Machine Learning in Spark

Alle Funktionen in Scala; Schnittstellen für Python hat großen Umfang.

Overview: DataFrame→Transformation (Feature Extraction, ...) → Estimation → Model&Parameter

**MLLib: http://spark.apache.org/docs/latest/mllib-guide.html**

- Klassisch: Statistiken, Sampling, Optimierung (L-BFGS, ...), ...
- Clustering: KMeans, gauss mixture, LDA, ...
- Regression/Classification: Linear (auch SVMs, Logistisch, ...), NaiveBayes, DecisionTree, Gradient-Boosted Trees, RandomForest, Isotonic,
- Recommendation: ALS
- Accelerated failure time model: AFT
- Python API: http://spark.apache.org/docs/latest/api/python/index.html
- oft auch support für “streaming data”...

# Stream Processing in Spark

- Quellen: Kafka, Flume, Twitter, ZeroMQ, Kinesis, or TCP sockets
- Datenpakete in Zeitfenstern als “DStream” (=RDD für Streaming)
- Windowing (sliding interval), Verarbeitung in Transformationen und ML-Modellen, ...
- Ausgabe
- http://spark.apache.org/docs/latest/streaming-programming-guide.html


# Software Lösungen: Drill



Drill: drill.apache.org - schema-free SQL engine

- “Treat your data like a table even when it's not”:
  mischen von Daten über Grenzen der Datenbank/Files hinweg, etc.
- Ideen von Dremel
- Durch SQL kompatibel zu BI Toolchain (Tableau, etc.)

# Software Lösungen / weitere

- Parquet: parquet.apache.org - columnar storage
- Hive: hive.apache.org - DataWarehouse, query large datasets
- HBase: hbase.apache.org - Facebook, Yahoo! , ...
- Accumulo: accumulo.apache.org - BigTable + sorting
- Presto: prestodb.io - Facebook (300PB warehouse), Airbnb, Dropbox
- Impala: impala.io - SQL query engine für HDFS, HBase
- **BlinkDB**: blinkdb.org - schneller als in-Memory? → Ja!

  approximative Abfragen (stratified/poissonized sampling)

# Chancen

# Wissenschaftliche Methodik

- Hypothese, Empirie, ...
- Datenbasierte Entscheidungen
- Nachvollziehbarkeit
- Automatisierung
- Szenarienanalyse
- bessere Validierung
- Statistik
- Unsupervised Learning
- Predictive Modelling

# Datenverarbeitung

- ETL Pipeline
- weniger Mehrgleisigkeit
- schnellere Resultate
- Flexibilität
- weniger vendor lock-in
- ad-hoc
- Skalierbarkeit
- Mashup

# Visualisierung

- bessere Kommunikation von Effekten/Resultaten
- interaktive Webseiten statt statischer Dokumente (Bokeh, Shiny, Spyre, …)

2012


HBR.ORG
Harvard Business Review
OCTOBER 2012
46 The Big Idea
The True Measures Of Success
Michael J. Mauboussin
84 International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz
93 Leadership
What Ever Happened To Accountability?
Thomas E. Ricks
GETTING CONTROL OF
BIG DATA
How vast new streams of information are changing the art of management
PAGE 59

McKinsey Global Institute

May 2011

# Big data: The next frontier for innovation, competition, and productivity

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation **(2011)**


# *FIN*