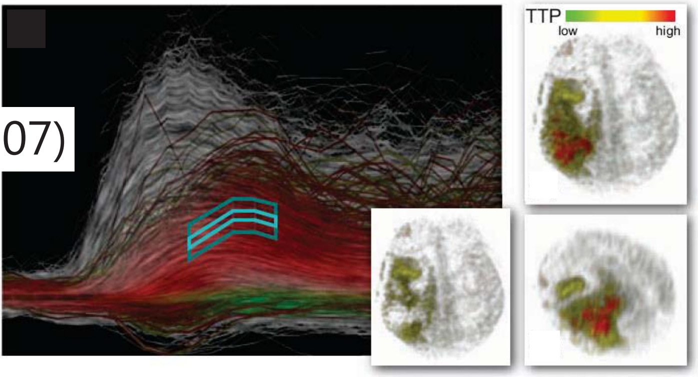
Big Data – visualization and visual analytics

Helwig Hauser

University of Bergen, Dept. of Informatics

Introduction

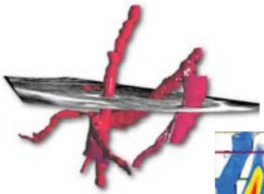


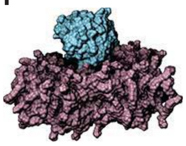
HH:

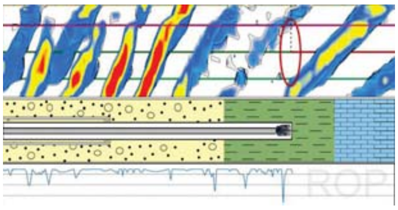
- prof. in visualization (UiB: 2007)
- about 20 years in the field
- interactive visual analysis

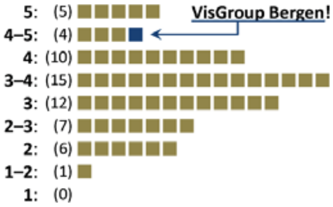
VisGroup @ UiB.no:

- research group @ UiB Dept. of Inf.
- appl.-oriented basic research
  - medicine
  - oil & gas
  - biology
  - fluid mechanics
  - …









Big Data

What is “Big Data”?

- well, lots of data, right? *… we come back to this in a moment.*
- certainly, a buzz-word… *… but a relevant one!*

Examples

- big data from numerous **sensors** (Internet of Things, …)
- bid data in large **social networks** (Facebook, Twitter, …)

Broadly used definition

- 3V-def.: “Big data” is **high-volume**, **-velocity** & **-variety** information assets that demand cost-effective, innovative forms of **information processing** for **enhanced insight** and **decision making**. [Doug Laney, 2001 / Gartner]

Big Data, V#1: Volume

Certainly, *Big Data* (usually) **refers to lots of data!**

> "Big data" refers to datasets
> whose size is **beyond the ability of typical database** software tools to **capture**, **store**, **manage**, and **analyze**.

[McKinsey Global Institute 2011]

Available data grows exponentially

– Exabytes of data available world-wide
  - 1 EB = 1000 PB = 1 million TB = 1 billion GB
  - hundreds of EB transferred via the Internet, annually
  - EB of new information stored, annually

Big Data, V#2: Variety

Big Data beyond numbers

– text, images & sound, relational data, …
unstructured data

– 30 billion pieces of information on Facebook per month!
400 million tweets per day
4 billion hours of videos are watched on YouTube / month
>400 million wearable, wireless health monitors

– Daniel Keim, 2007: 100 million FedEx transactions per day, 150 million VISA credit card transactionen per day, 300 million long distance calls in ATT's network per day, 50 billion e-mails worldwide per day, 600 billion IP packets per day DE-CIX backbone

Dark Data: available, but unused data

Big Data, V#3: Velocity

Real-time Big Data / Streaming Data Analysis, but also

- rapidly changing data
- data at different speeds and uneven rates (bursts)

Big Data – a moving target!

- lots of generated information cannot be stored!
  - 90% of health care data is discarded (videos, *etc.*)

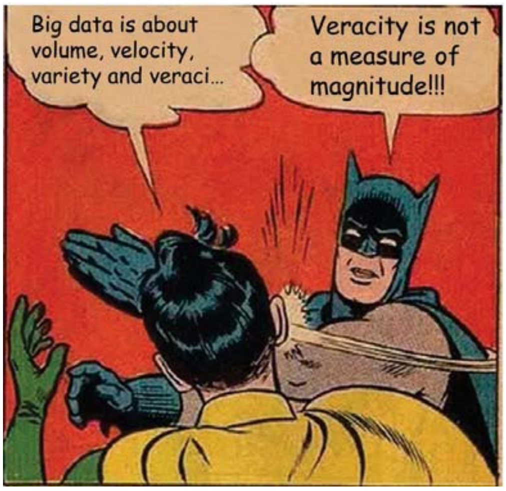
Big Data, V#4(?): Veracity [IBM, …]

Uncertain / low-quality data

- >$3 trillion loss to US economy due to bad data quality
- high degree of uncertainty

D. Laney blogs:

- Batman on Big Data:


Big data is about volume, velocity, variety and veraci…
Veracity is not a measure of magnitude!!!

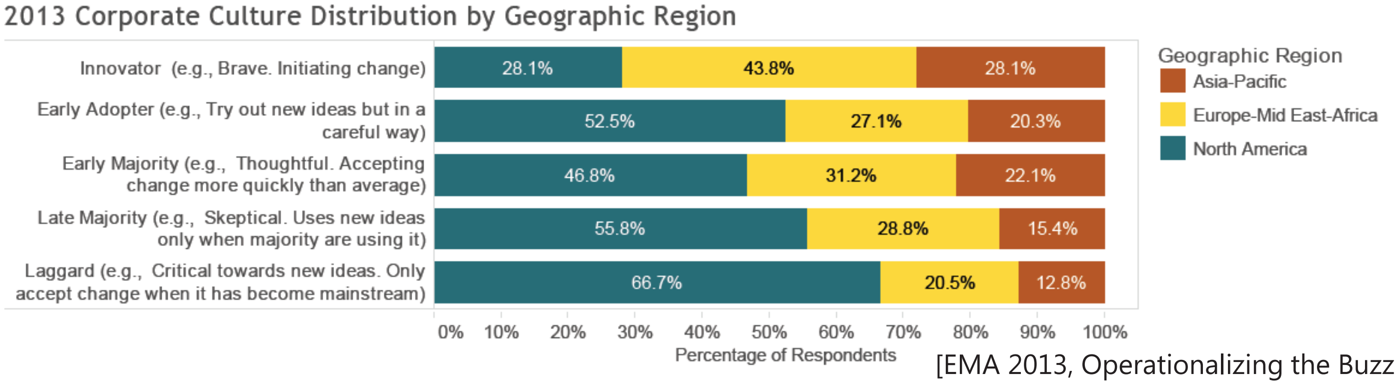
Even more Vs: [K. Normandeau]

- validity: the right data for the right decisions?
- volatility: when valid, storing for how long, etc.?

Big Data in Practice

Big data is

- generated, aggregated, analyzed, and consumed
- **sensed, collected** (networks), **stored** (cloud), **and analyzed** (machine learning, visualization, …)
- **process-mediated** ("nicer" data),
  machine-generated (Internet of Things),
  human-sourced (from messages to videos)



[EMA 2013, Operationalizing the Buzz]

Big Data – Challenges & Opportunities

Selected Challenges

- shortage of *Big Data* talent (up to 200.000 needed in the US plus 1.5 million «data-savvy» managers)
- contextualization of Big Data – Big Data needs to be complimented by Big Judgment [Harvard Business Review]
- prediction difficult without theory

Selected Opportunities

- annually $300 billion to the US health care system, incl. cost savings up to 8%
- annually $250 billion to the European public sector adm.
- job opportunity (analysts, managers, *et al.*)!

[McKinsey GI, 2011]

Big Data in Business

Five opportunities according to McKinsey GI, 2011:

- **reduced searching & processing time**, e.g., in the public administration sector, as well as **concurrent engineering** in manufacturing due to **accessible Big Data**
- enabling **experimentation** to **discover needs, expose variability,** and **improve performance**
- **segmenting populations** to **customize actions**
- **replacing/supporting human decision making** with **automated algorithms** based on **Big Data Analytics**
- innovating **new business models**, **products**, and **services**

Active enterprises include:

- **eBay, Amazon, Walmart, Facebook,** in *finance*, *real estate*, …

Big Data Technology – selection

Conceptual

- MapReduce [Google, 2004]
  - **map**: distribution of queries to many nodes
  - **reduce**: gathering of results and delivery
- NoSQL (“not only SQL”), for ex. Cassandra (key-value)

Software

- Hadoop [Apache], MongoDB

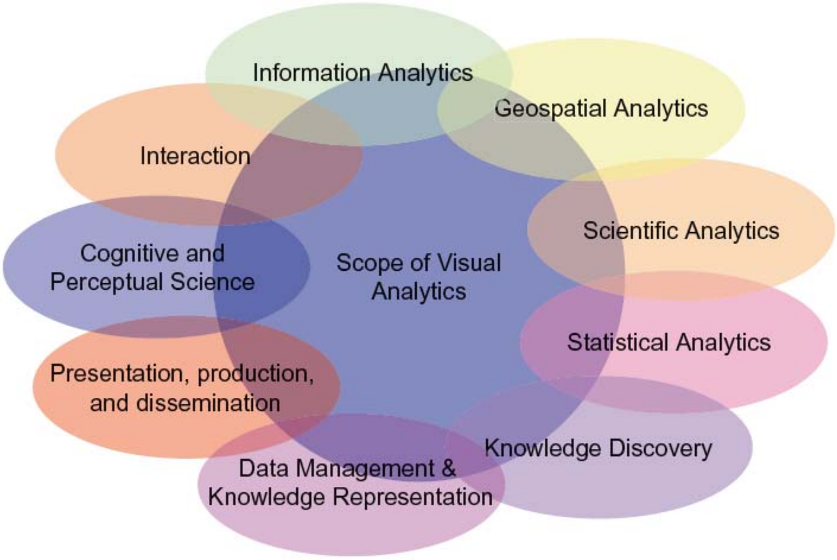
Analytics Technologies

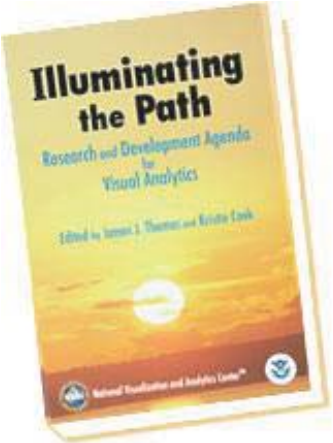
- A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, natural language processing, signal processing, simulation, time series analysis and visualization [McKinsey, 2011]

Big Data and Visual Analytics

Visual Analytics

- Illuminating the Path book: 2005
- VisMaster book: 2010


Information Analytics
Geospatial Analytics
Interaction
Scientific Analytics
Cognitive and Perceptual Science
Scope of Visual Analytics
Statistical Analytics
Presentation, production, and dissemination
Knowledge Discovery
Data Management & Knowledge Representation

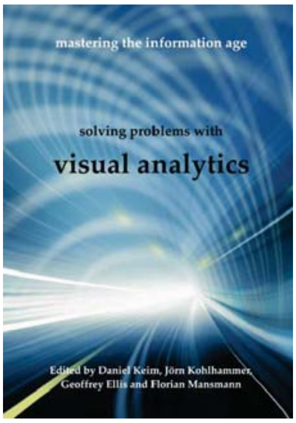


Illuminating the Path

Research and Development Agenda for Visual Analytics

Edited by James J. Thomas and Kristin Cook

National Visualization and Analytics Center™



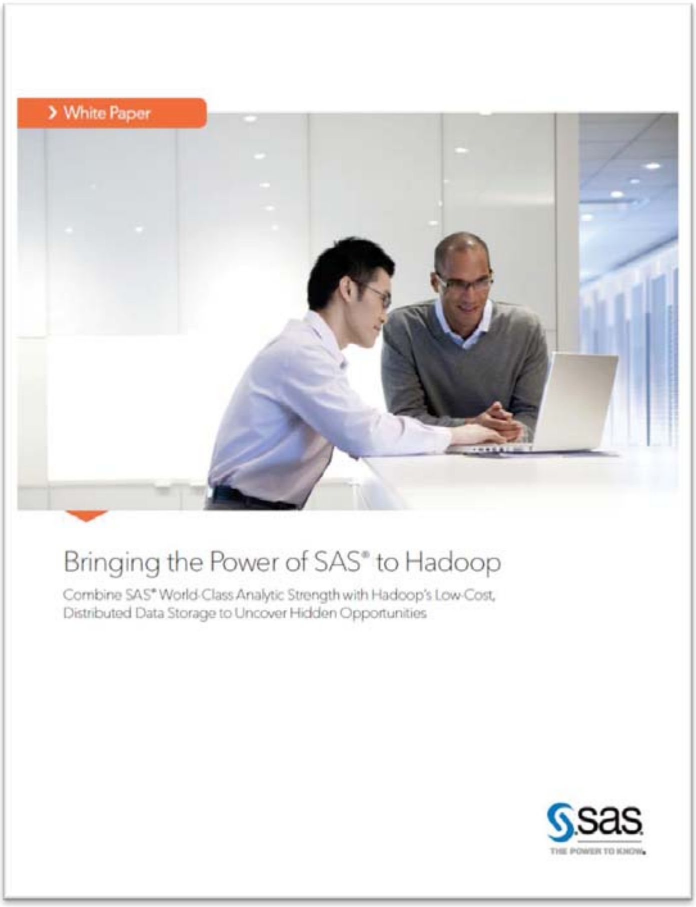
mastering the information age

solving problems with

visual analytics

Edited by Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis and Florian Mansmann

SAS' Approach



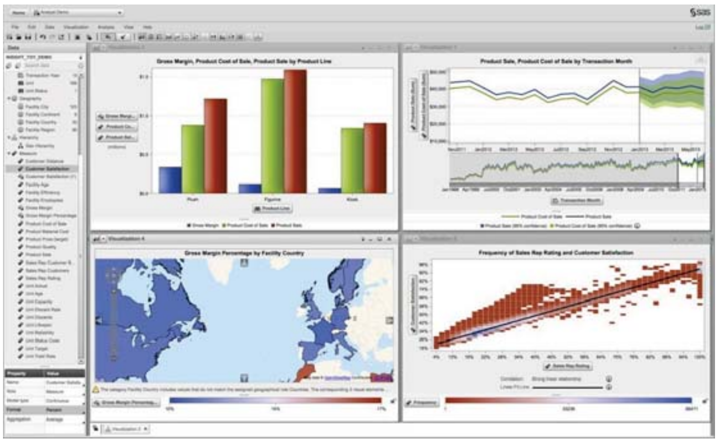
› White Paper

Bringing the Power of SAS® to Hadoop

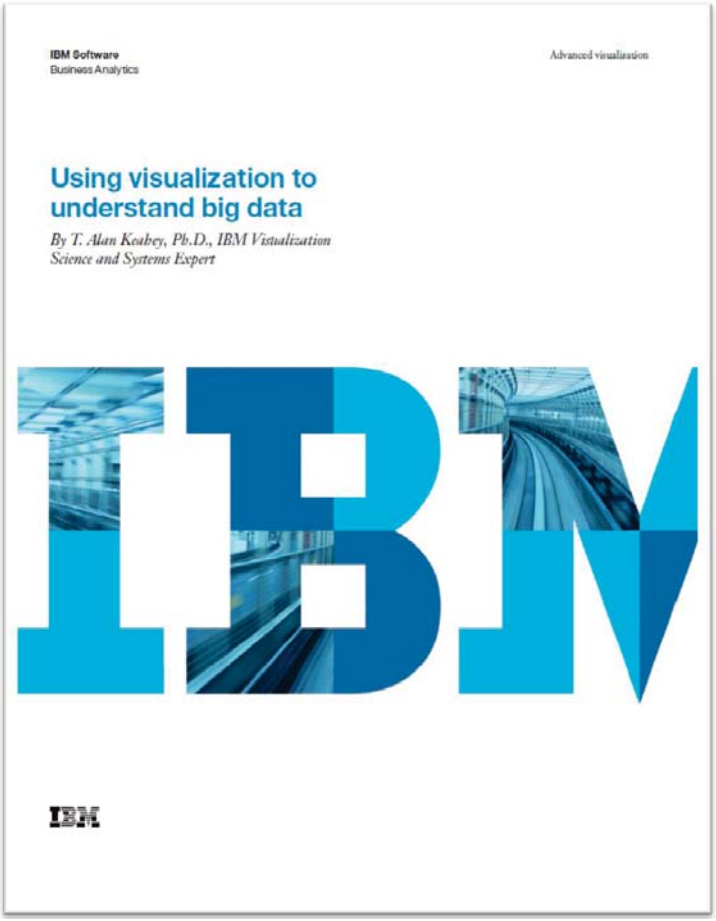
Combine SAS® World-Class Analytic Strength with Hadoop's Low-Cost, Distributed Data Storage to Uncover Hidden Opportunities

sas THE POWER TO KNOW.

- based on MapReduce
- SAS Visual Analytics
- SAS Visual Statistics


Gross Margin, Product Cost of Sale, Product Sale by Product Line
Product Sale, Product Cost of Sale by Transaction Month
Gross Margin Percentage by Facility Country
Frequency of Sales Rep Rating and Customer Satisfaction

IBM's Approach



IBM Software
Business Analytics

Advanced visualization

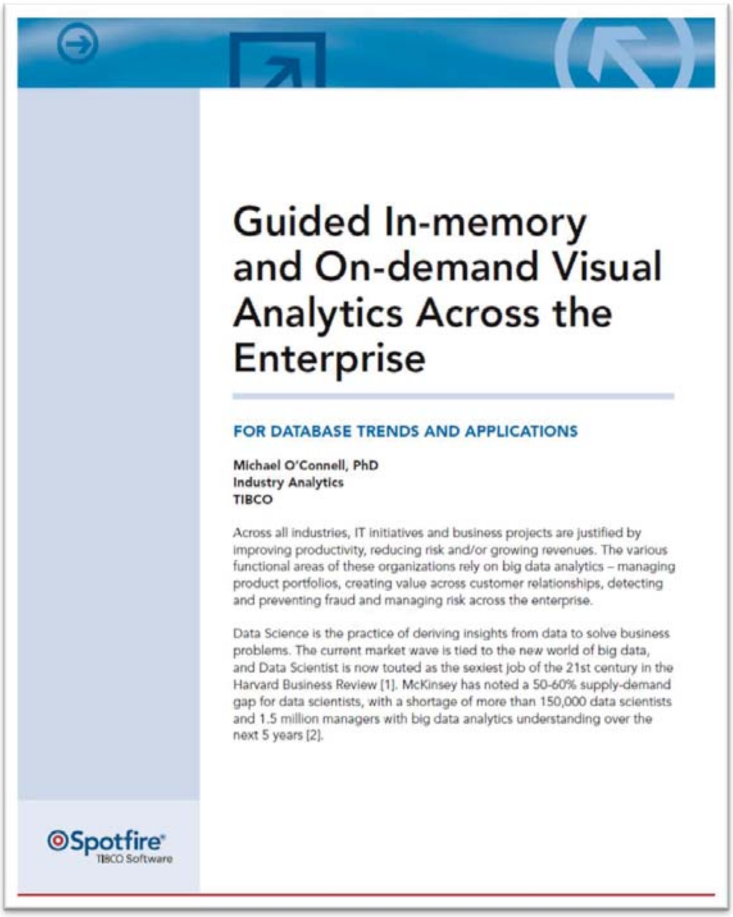
Using visualization to understand big data

By T. Alan Keahey, Ph.D., IBM Visualization Science and Systems Expert

IBM

- visualization
  → externalization
- standard visualization
  - up to 10^6 data items; pies < bars < lines < scatterplots < treemaps
  - up to $10^{9–12}$, when combined with analytics
- IBM Rapidly Adaptive Visualization Engine (RAVE)
- analyticszone.com/visualization

TIBCO's Approach (Spotfire)



Guided In-memory and On-demand Visual Analytics Across the Enterprise

FOR DATABASE TRENDS AND APPLICATIONS

Michael O'Connell, PhD
Industry Analytics
TIBCO

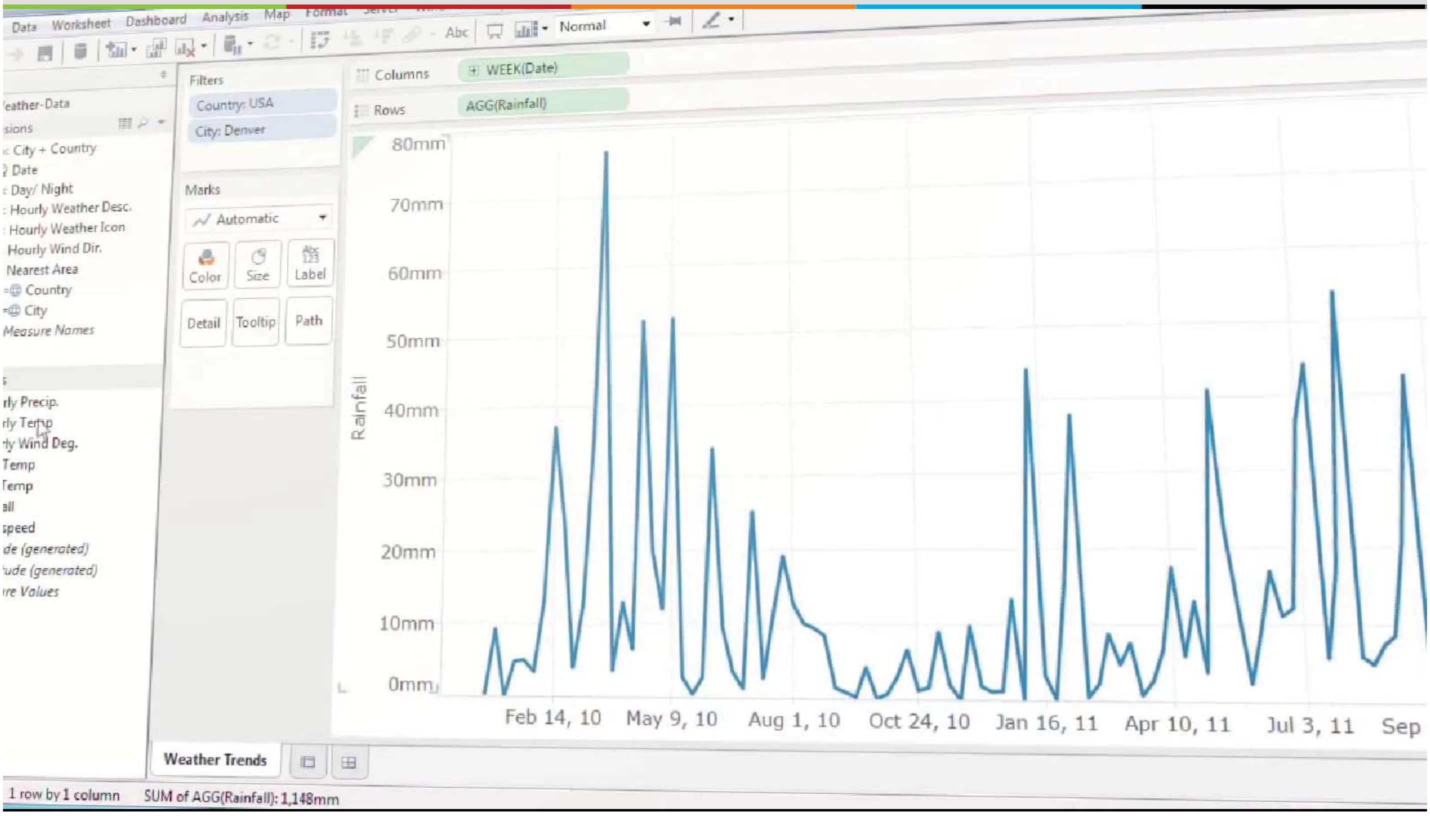
Across all industries, IT initiatives and business projects are justified by improving productivity, reducing risk and/or growing revenues. The various functional areas of these organizations rely on big data analytics – managing product portfolios, creating value across customer relationships, detecting and preventing fraud and managing risk across the enterprise.

Data Science is the practice of deriving insights from data to solve business problems. The current market wave is tied to the new world of big data, and Data Scientist is now touted as the sexiest job of the 21st century in the Harvard Business Review [1]. McKinsey has noted a 50-60% supply-demand gap for data scientists, with a shortage of more than 150,000 data scientists and 1.5 million managers with big data analytics understanding over the next 5 years [2].

Spotfire
TIBCO Software

- The comb. of in-memory, in-database on-demand, predictive, interactive and visual analytics;
- with self-service guided and collaborative workflows for the masses,
- and in-line deployment in real-time event systems,
- is the future. This is Data Science 2.0

Tableau



Filters
Country: USA
City: Denver

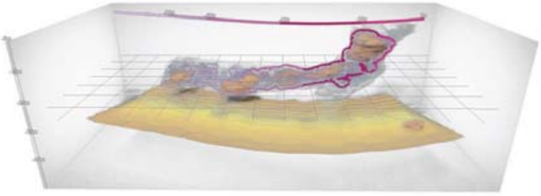
Columns: WEEK(Date)

Rows: AGG(Rainfall)

1 row by 1 column SUM of AGG(Rainfall): 1,148mm

Visualizes Big Data through Google BigQuery

Visualization



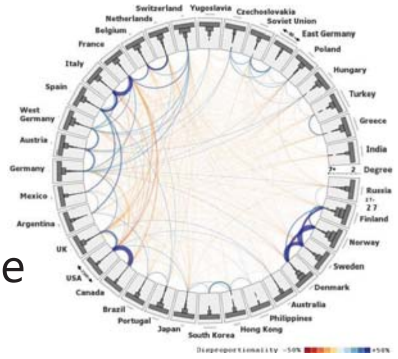
Bridge between user and data

- large-scale measurements (medical tomography, seismic data, etc.)
- computational simulation (computational fluid dynamics, etc.)
- mathematical modeling

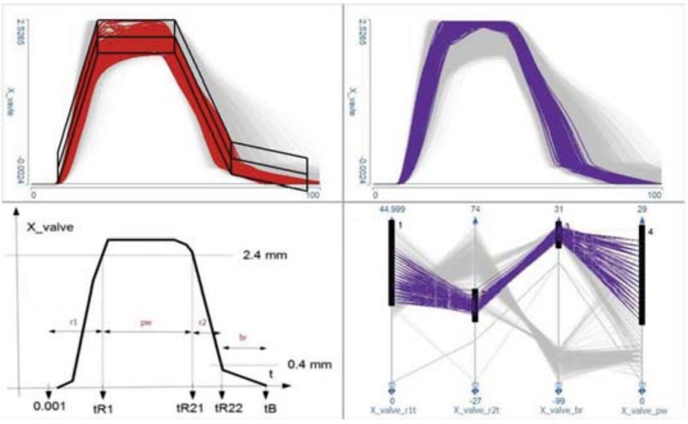
Enabling insight, decision support, improving communication

Good for:

- data exploration – finding the unknown
- data analysis – checking hypotheses
- presentation – communicate & disseminate



Information Visualization

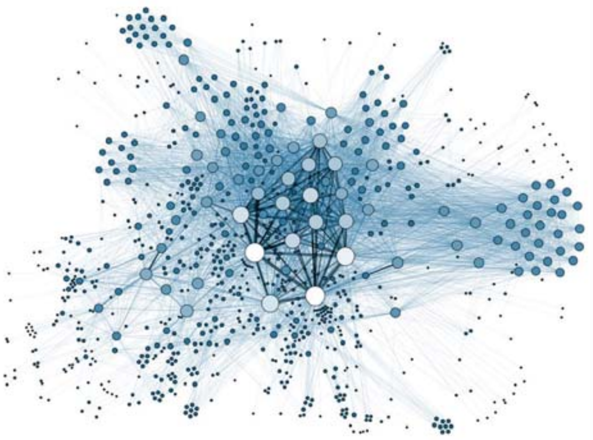


Abstract data visualization:

- tables, databases
- networks, graphs
- texts, collections

Interactive Visualization

- filtering, drill-down
- linking & brushing
- multiple perspectives
- analyze, produce, search, query

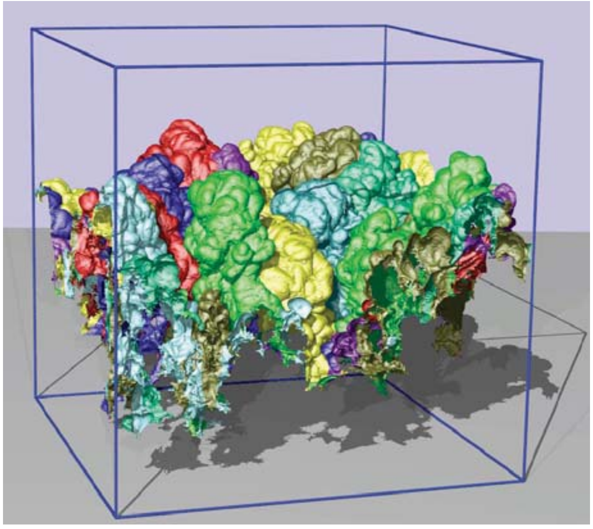


Large-scale Visualization

Scientific Visualization

(spatiotemporal data fields, etc.)

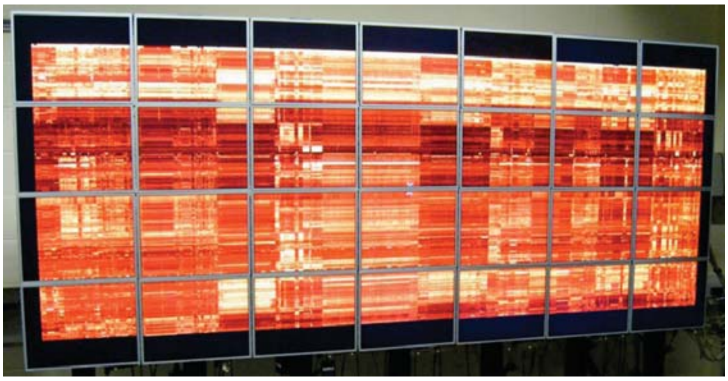
- often from computational simulation (fluids, for ex.)
- up to TB, PB



Information Visualization

(abstract data, etc.)

- data more heterogeneous
- up to MB, GB

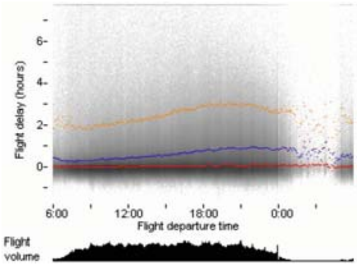


Visualization (Research) & Big Data

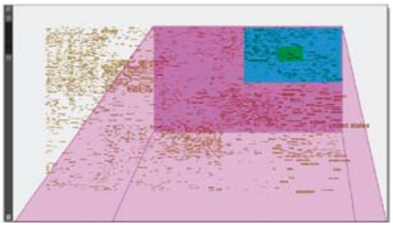
THE FIRST WORKSHOP ON BIG DATA VISUALIZATION

First Attempts,
including

- VisReduce: … (Im et al.);
  150M records, >100dims.



- Visualizing Big SPH Sim. (Reichl et al.);
  10 billion points
- Typograph: …
  (Endert et al.);
  all of Wikipedia
- Egocentric Storylines (Muelder et al.);
  >10k nodes

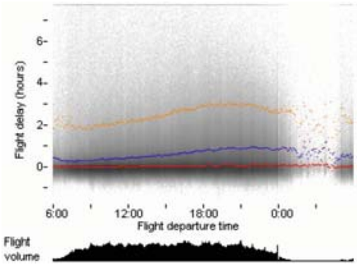


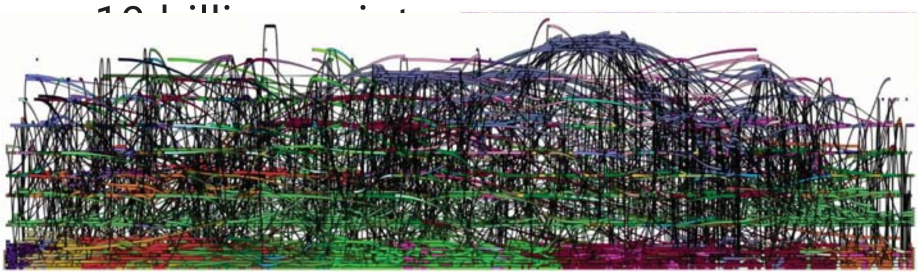
Visualization (Research) & Big Data

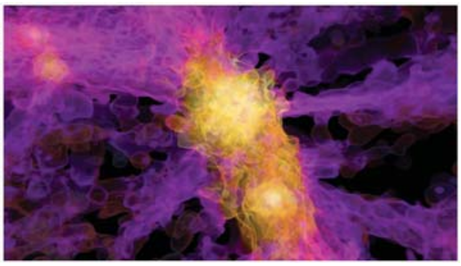
THE FIRST WORKSHOP ON BIG DATA VISUALIZATION

First Attempts,
including

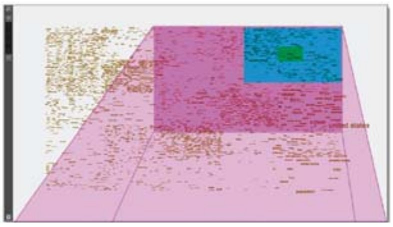
- VisReduce: … (Im et al.);
  150M records, >100dims.
- Visualizing Big SPH Sim. (Reichl et al.);



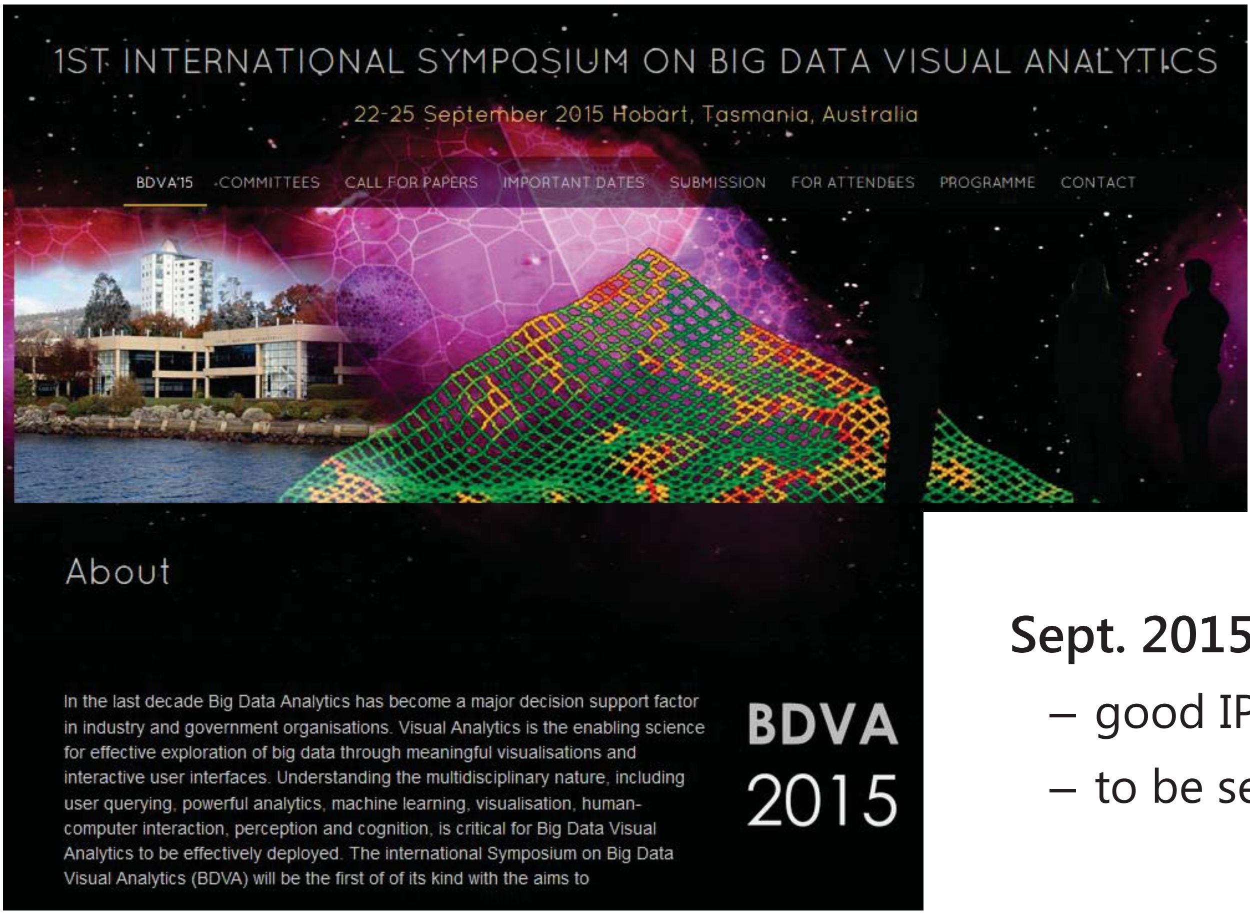




- Egocentric Storylines (Muelder et al.);
  >10k nodes



Visual Analytics (Research) & Big Data



1ST INTERNATIONAL SYMPOSIUM ON BIG DATA VISUAL ANALYTICS

22-25 September 2015 Hobart, Tasmania, Australia

About

In the last decade Big Data Analytics has become a major decision support factor in industry and government organisations. Visual Analytics is the enabling science for effective exploration of big data through meaningful visualisations and interactive user interfaces. Understanding the multidisciplinary nature, including user querying, powerful analytics, machine learning, visualisation, human-computer interaction, perception and cognition, is critical for Big Data Visual Analytics to be effectively deployed. The international Symposium on Big Data Visual Analytics (BDVA) will be the first of of its kind with the aims to

BDVA 2015

Sept. 2015…

- good IPC
- to be seen

Conclusions

Big Data is maturing, it's unavoidable

EMA 2013: the next Big Data challenge: Ethics!

Big Data is transforming Science (4th paradigm, etc.)

- Chris Anderson, Wired, 2008: The End of Theory

New opportunities, new challenges

- big business, P4 medicine
- "the other" Vs, dark data
- how to turn data into knowledge?
- technological challenges, new ways of thinking
- it's – not at the least – also an educational challenge!

Acknowledgements

You! ☺

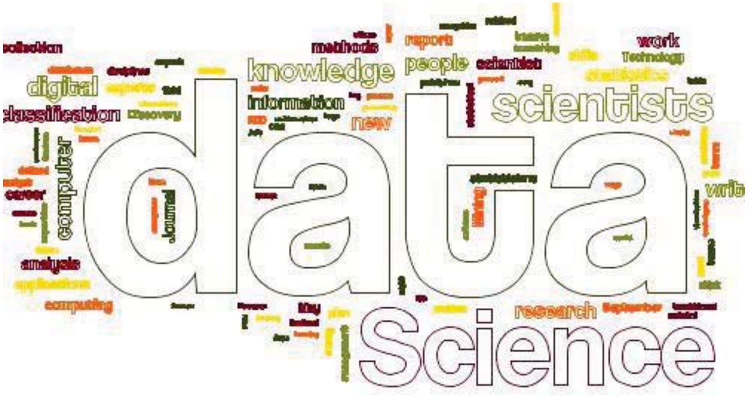
Questions?

Stefan Bruckner

Arvid Lundervold

Lots of references…

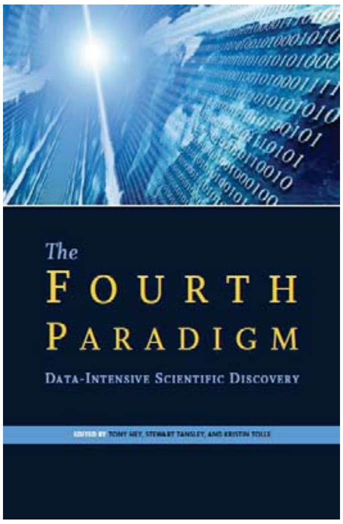
But let's talk about science a little..



Big Data and the Fourth Paradigm

2009, Microsoft: the 4th paradigm: data-intensive scientific discovery

- refers to the last talk by Jim Gray, 2007, "A Transformed Scientific Method"
- from **empirical** (initially), via **theoretical** (modern times), and **computational science** (last decades) to **data-intense science** (now)
- eScience: **capture**, **curation**, **analysis**, **vis**.
- needle-in-a-haystack problems comparably "easy" (Higgs)
- more difficult: trends, clusters, patterns (N^2, or more)


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Big Data in Science

Sources of Big Data

- **meteorology**, **genomics**, **connectomics**, **complex physics simulations**, and **biological** and **environmental research**
- **mobile phones**, **remote sensing**, **logs**, **cameras & microphones**, **RFID sensors** & **sensor networks**

Big Science Examples

- The Large Hadron Collider experiments:
  - about 150 million sensors
  - delivering about 40 millions times per second (!!)
- Sloan Digital Sky Survey (since 2000)
  - more data in a few weeks than all of astronomy so far
  - about 200 GB per night, now >140TB of data

Big Data in Medicine

P4 medicine [Leroy Hood]

- predictive, preventive, personalized, and participatory

Computational Medicine [Arvid Lundervold, 2014]

- embracing IT, bioinformatics, etc., for "systems medicine"

Examples:

- predictive medicine
- large-scale cohort studies

Our Vision:
Enabling Real-Time Personalized Medicine
Interpret all patient data during a patient's visit
SAP HANA
Real-time Big Data Platform
Real-time Big Data Convergence
Clinical Data (EMRs)
+
Biological data
+
Lifestyle Data

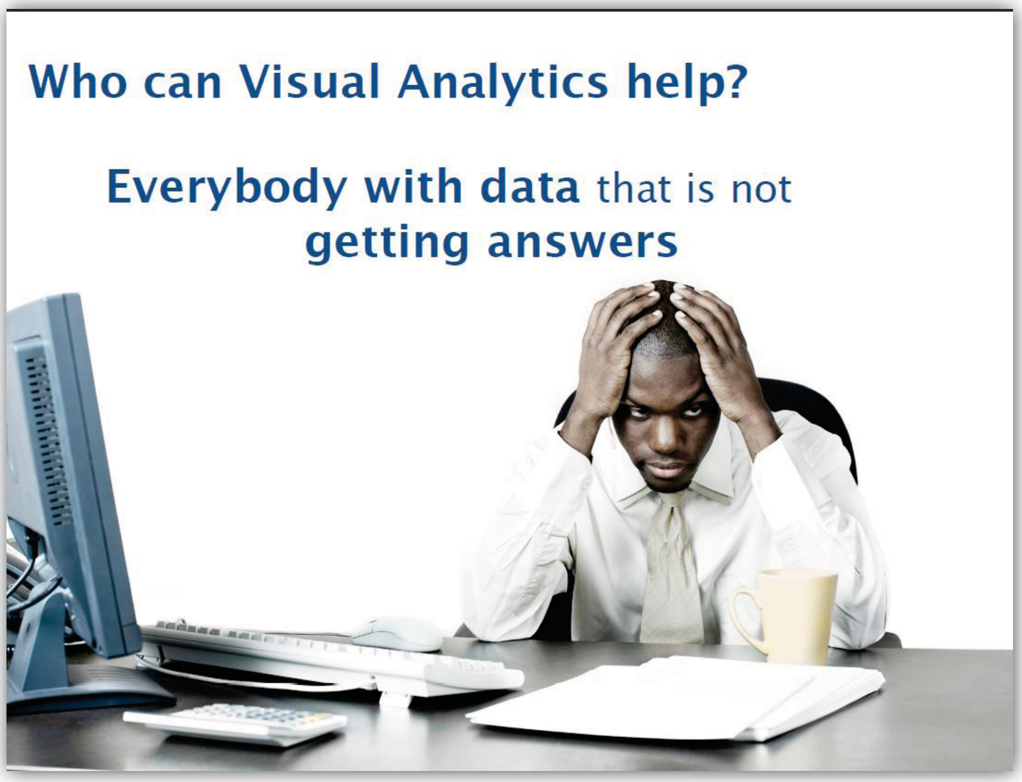
[SAP]

Case: [EMA 2013 Operationalizing the Buzz]

- Brigham and Women's Hospital: improved drug risk awareness due to Big Data (much fast results)

Big Problems with Small Data

Christian Chabot (CEO of Tableau), 2008:



Who can Visual Analytics help?

Everybody with data that is not getting answers

VAST Keynote

Big Data and Privacy Concerns

Snowden informed about NSA...

As data get large, networked, reside in the cloud, we fear

- unauthorized access
- data misuse
- identity theft

Examples:

- leaked health data
- credit card fraud
- monitored privacy