Reutlingen University

Panel SEMAPRO/ADVCOMP/DATA ANALYTICS, Porto, 01.10.2013

Cognitive-based Computation, Semantic Understanding, and Web Wisdom

Moderation:

Fritz Laux, Reutlingen University, Germany

Panelists:

Felix Heine, University of Applied Sciences & Arts, Hannover, Germany

Rudolf Berrendorf, Bonn-Rhein-Sieg University, Germany

Alexey Cheptsov, High Performance Computing Center Stuttgart - HLRS, Germany

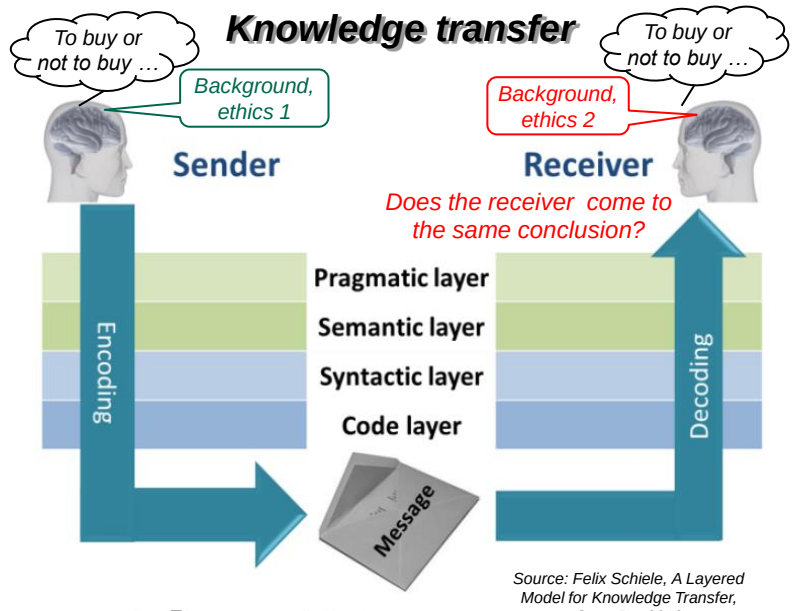
Reutlingen University

Knowledge Stack

- *Data*
  - sequence of symbols from well defined set of symbols
  - information (facts) coded for mechanized processing (DIN)
- *Information*
  - data + metadata
  - recognized (relevant) data
- *Knowledge*
  - information + thinking process
  - Information linked to the person's background knowledge
- *Wisdom*
  - applied knowledge (English, 1999)
  - Understanding in the context of a person's background knowledge and ethics
- *data := facts, data in context, info in context, knowledge in context (applied knowledge) (L. P. English, 1999)*

Reutlingen University

Problem 1: Semantic Loss



Source: Felix Schiele, A Layered Model for Knowledge Transfer, SemaPro 2013

3 /4

Reutlingen University

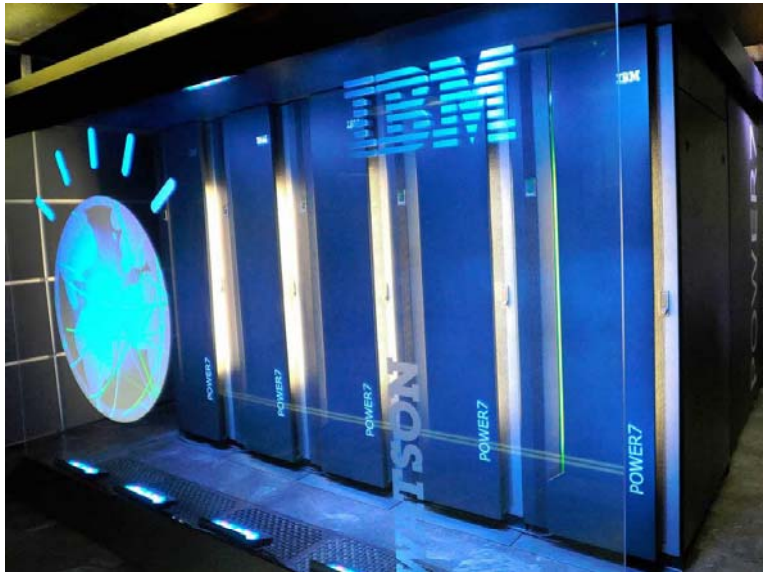
Problem 2: Reliability of Information

Current state of web (Google) query

- ***What is the height of Mt. Everest?***
  - *None answered the question, but the first 4 links including Wikipedia said: 8848 m*
  - *4. link had two different heights (in m)*
  - *5. answer from Encyclopedia Britannica claims 8850 m*
- ***Which one is correct?***
  - *Majority*
  - *Trusted source*

4 /4

New Opportunities
with Clever Techniques **and** Big Iron



courtesy: IBM

courtesy: LLNL

Traditional Way

- Use a Mobile Phone / Tablet / PC / server / mainframe and run a program / an algorithm to fulfill some task
- All allgorithms variants in use: deterministic, randomized, …
- Most times, you need to know in advance (when you program) what you want to do (e.g., signal processing)

Another Way

- You have already Giga/Tera/Peta/Exa/Zeta/Yotta bytes of information (in some way structured, unstructured,…)

- Learn from this existing databases
  - to answer questions (e.g., Watson)
  - to solve problems
  - to detect problems
  - to make projections into the future
  - …

- Lots of techniques known around this idea for quite some time (AI, Machine Learning, Neural Networks, Data Mining, …)

Clever Algorithms Are Sometimes Not Enough

- As data becomes **really large** and/or algorithms need to be **more clever** (need much more time to compute), **usual** mobile phone / tablet / PC /… do not suffice any more

- **Limitations** are at any single point in the usual hardware: raw compute power, available memory, I/O bandwidth, network bandwidth,…

- Some people stop here!

Here Comes the Sun …

	PC	LLNL Sequoia
cores	<10	≈ 1.6 million
FP performance	< 100 GFlops	≈ 20 PetaFlops
main memory	4-8 GB	1.6 PetaBytes
network bandwidth	1 GigaBits/s	≈ 30 PetaBytes/s (internal network)

courtesy: LLNL

Use It!

- Use the raw power you need somewhere in the spectrum from smaller up to big, big machines
  - to process / learn from big, big data
  - to find better solutions
  - to answer additional questions that could not be answered before
  - …

- For example:
  - run many, many filters / mining algorithms in parallel and combine intermediate results
  - for optimization problems, start processing with many different seeds in parallel

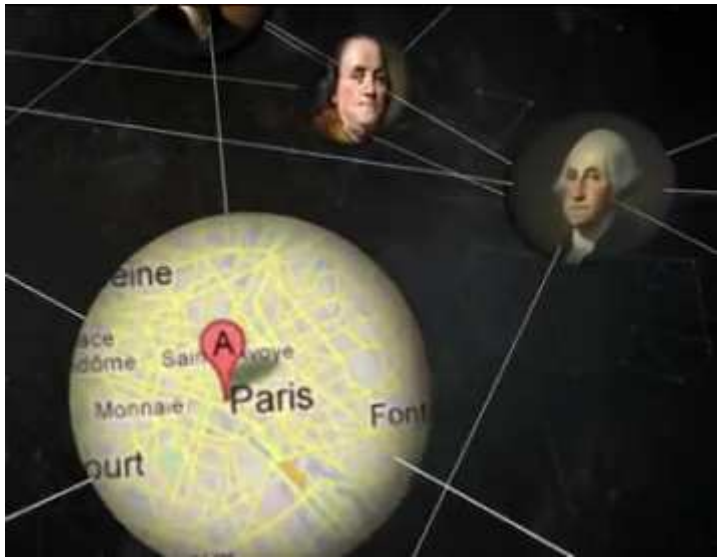
- Start thinking about **the opportunities with tomorrow‘s compute capacities**

Panel SEMAPRO/ADVCOMP/DATA ANALYTICS

Dr.-Ing. Alexey Cheptsov

Large-scale Graph Computing: Some Examples

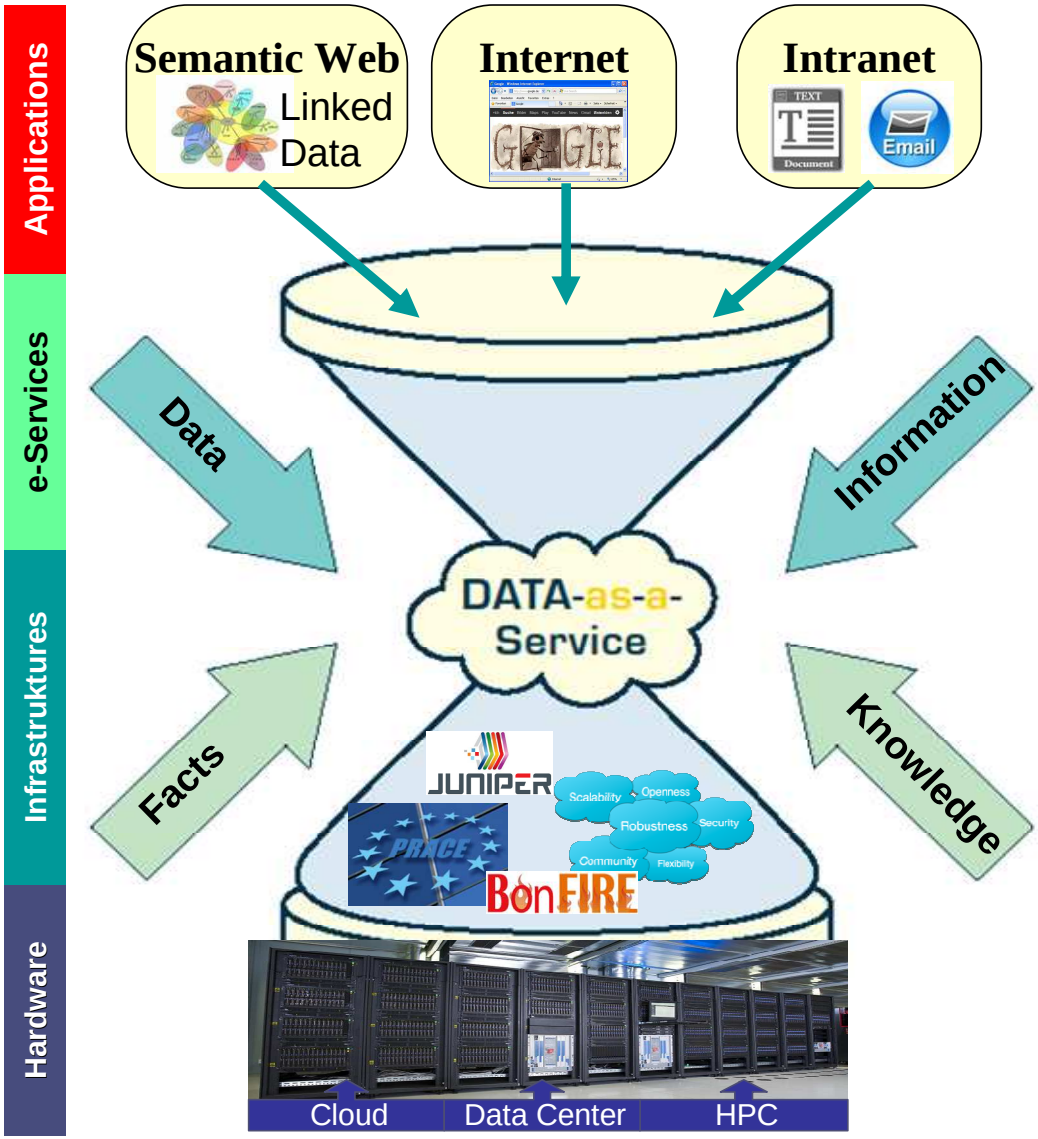
- Google Knowledge Graph
  - **700 million** nodes
  - **20 billion** facts
  - **several terabytes** of files



- Facebook's Social Graph
  - 60 PB of graph structured data
- Twitter's Interest Graph
- NoSQL database solutions

Credit: Google,
http://www.stateofsearch.com/search-in-the-knowledge-graph-era/

Cognitive-based Computation, Semantic Understanding, and Web Wisdom



Use of Supercomputers

Hermit – the HLRS mainstream system

- **Cray XE6 architecture**
- **Performance of 1,2 PetaFLOP (**10^15 floating point operations per second)
- **3552** compute nodes
- **64GB** RAM per node
- **2,7 PB** disc space

What are the challenges: Semantic Web View

- Infrastructure „on demand“
  - shared/distributed memory parallel clusters
  - multicore machines
  - GPGPU devices
  - FPGA
  - alltogether?

- Programming models to achieve high performance
  - MapReduce
  - MPI
  - „New“ programming languages

JUNIPER – Java platform for high performance and real-time large scale data management

Main Results

- HPC is going to face new challenges related to data-centric application expansion.
- Parallel programming models (mainly MapReduce and MPI) are the key enablers of HPC to data-centric applications
- Reaching near-peak performance is going to be the major challenge

Future Work

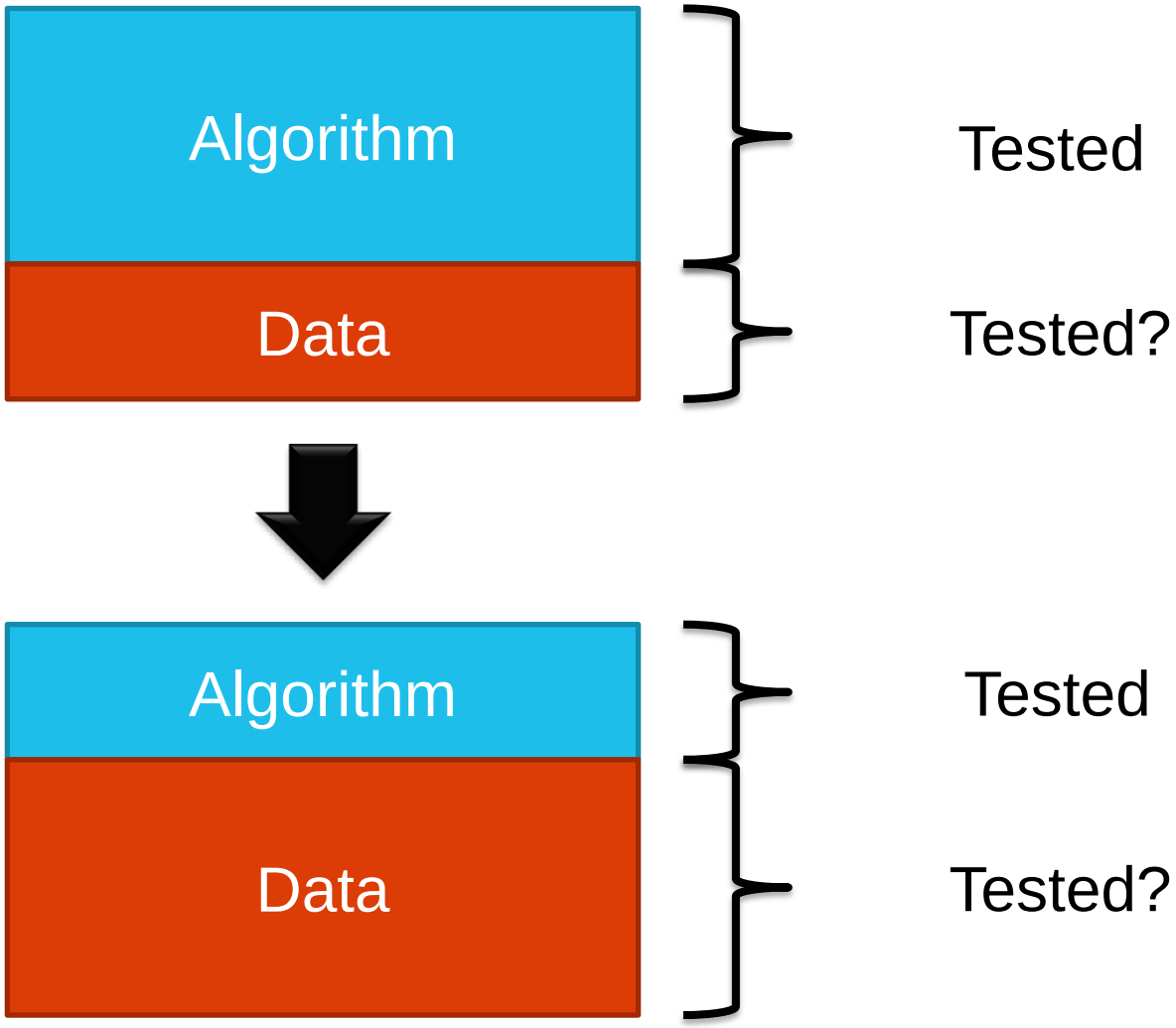
- Promote existing technologies, such as MPI, to solving new challenges, such as Big Data.
- Making existing framework more data-centric.

HOCHSCHULE
HANNOVER
UNIVERSITY OF
APPLIED SCIENCES
AND ARTS
–
Fakultät IV
Wirtschaft und

Panel: Data Analytics 2013 Data Quality and Big Data

Felix Heine

Data driven decisions



Data Quality

Data gets more and more important.

However, Data Quality is **underestimated**:

> "My data does not have any errors"
>
> "Yes, I know data quality is important,
> however, I will spend my budget first for new features"

Compare with Algorithms:

> "My program does not have any error" ???
>
> "Yes, I know, testing is important,
> however, I will first develop new features" ???

Data Quality and Big Data

- Big Data: 3 V‘s
  - **Volume**: Large amounts of data
  - **Velocity**: Stream data with very high data rates
  - **Variety**: Not only relational data: text, binary, XML, ...
  - Sometimes also: **Veracity**
    - Not a property of the data!
- Collect first, analyse later
- Monitor your data quality!

Quality of Big Data: What can we do?

- Understand your data!
  - Use data profiling tools
  - Research challenge: more sophisticated profiling
    - Statistics, machine learning, time series, ...
    - Good visualization of the results
    - Scalability

- Keep the knowledge for constant monitoring
  - In which language?
  - Research challenge:
    - What will be the SQL for Big Data?
    - Declarative language for data analytics