René Jäkel
Center for Information Services and High Performance Computing (ZIH)

Overview: challenges and research directions in large scale data analytics and management

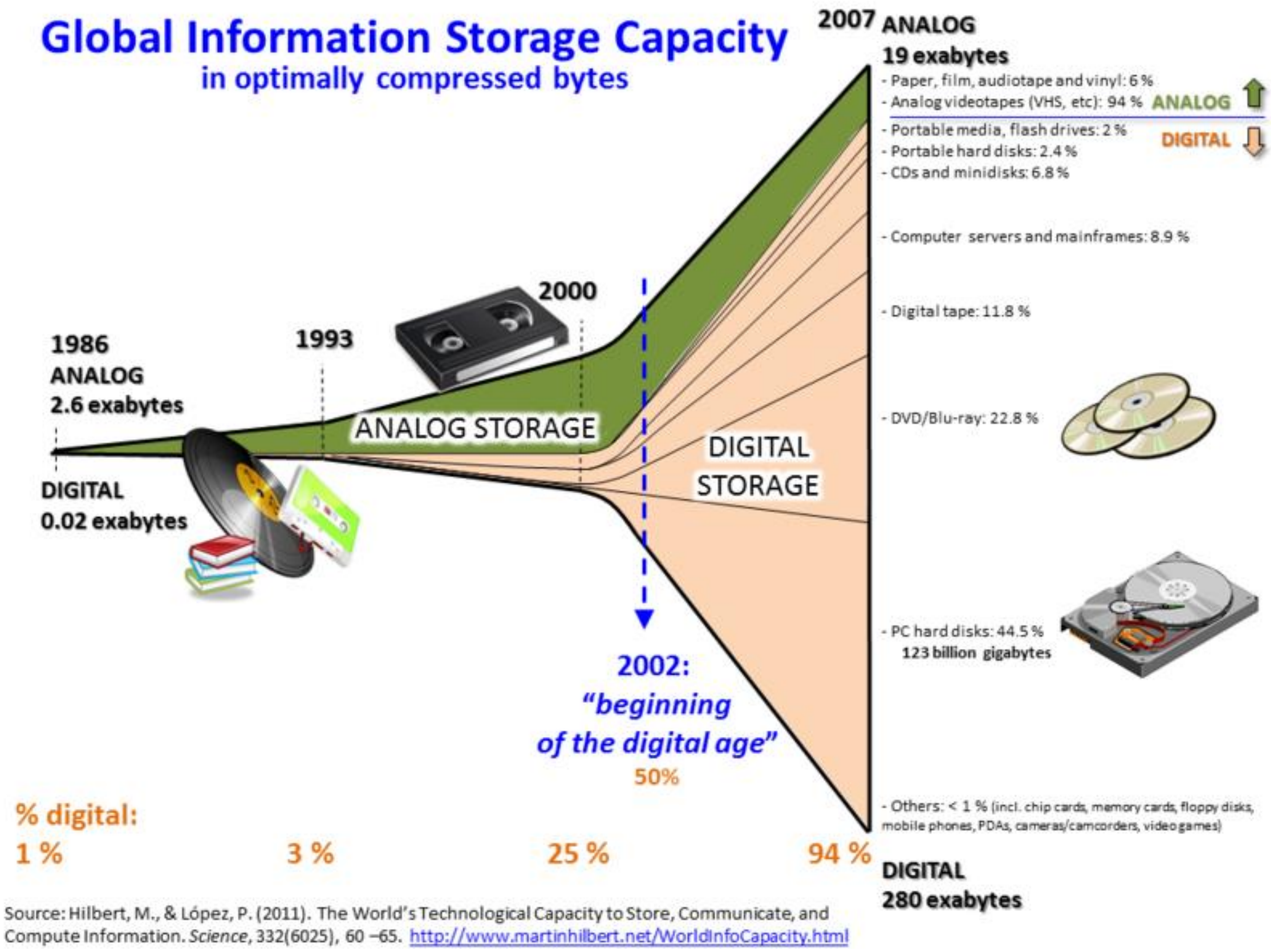
Science at Extreme Scales: Where Big Data Meets Large-Scale Computing Tutorials

Sept. 13 2018, IPAM

Transformation: From scarce to abundant (digital) data (starting early 200x)

4 September 2008

11 February 2011



What means big?

Source: IDC's Digital Universe study, sponsored by EMC, 2014

Study from 2014

„Digital Universe" (IDC – International Data Corporation)

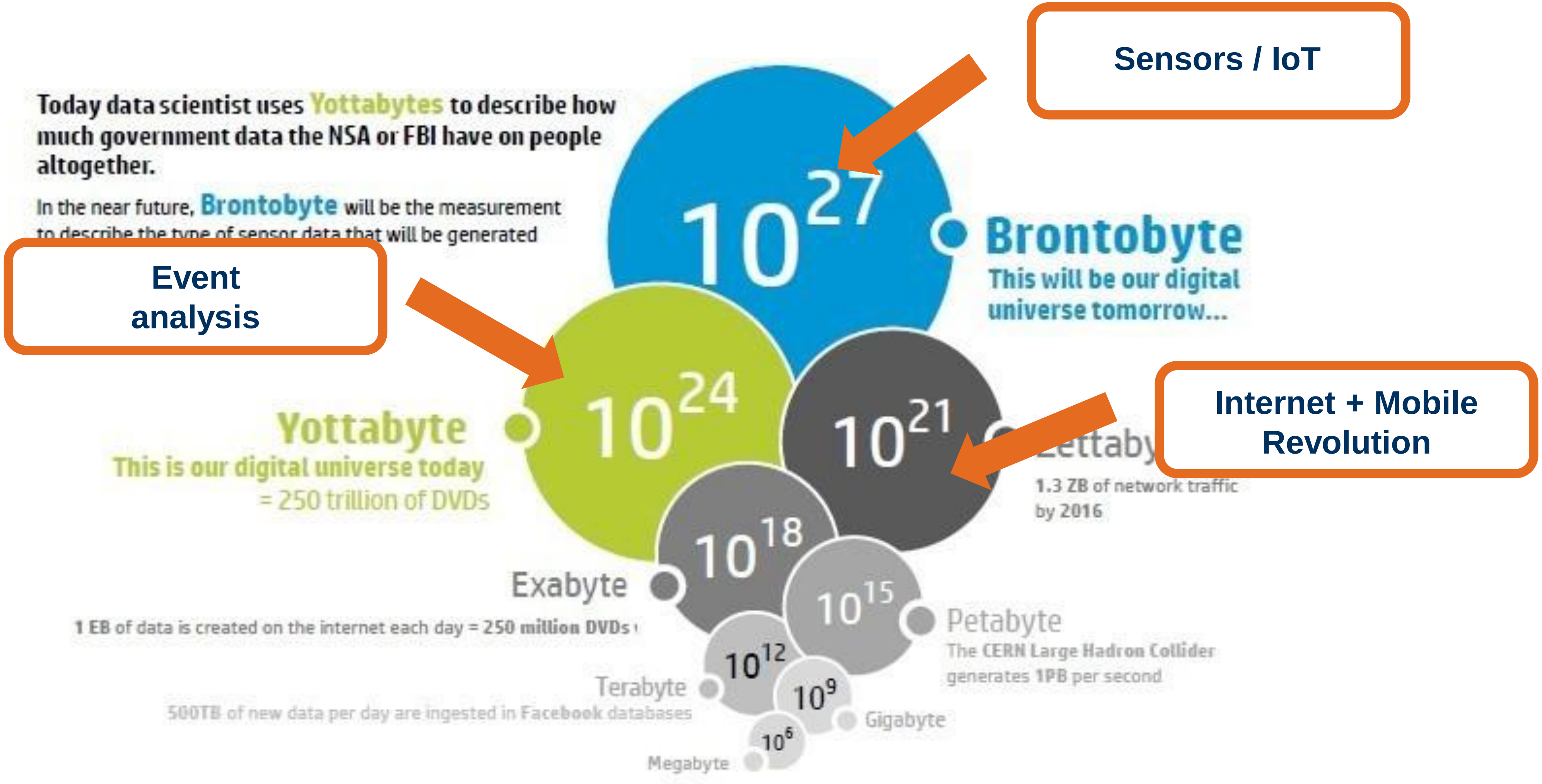
Mostly unstructured data

Big: ZB = 10^{21}B = 10^9 x TB

x1000

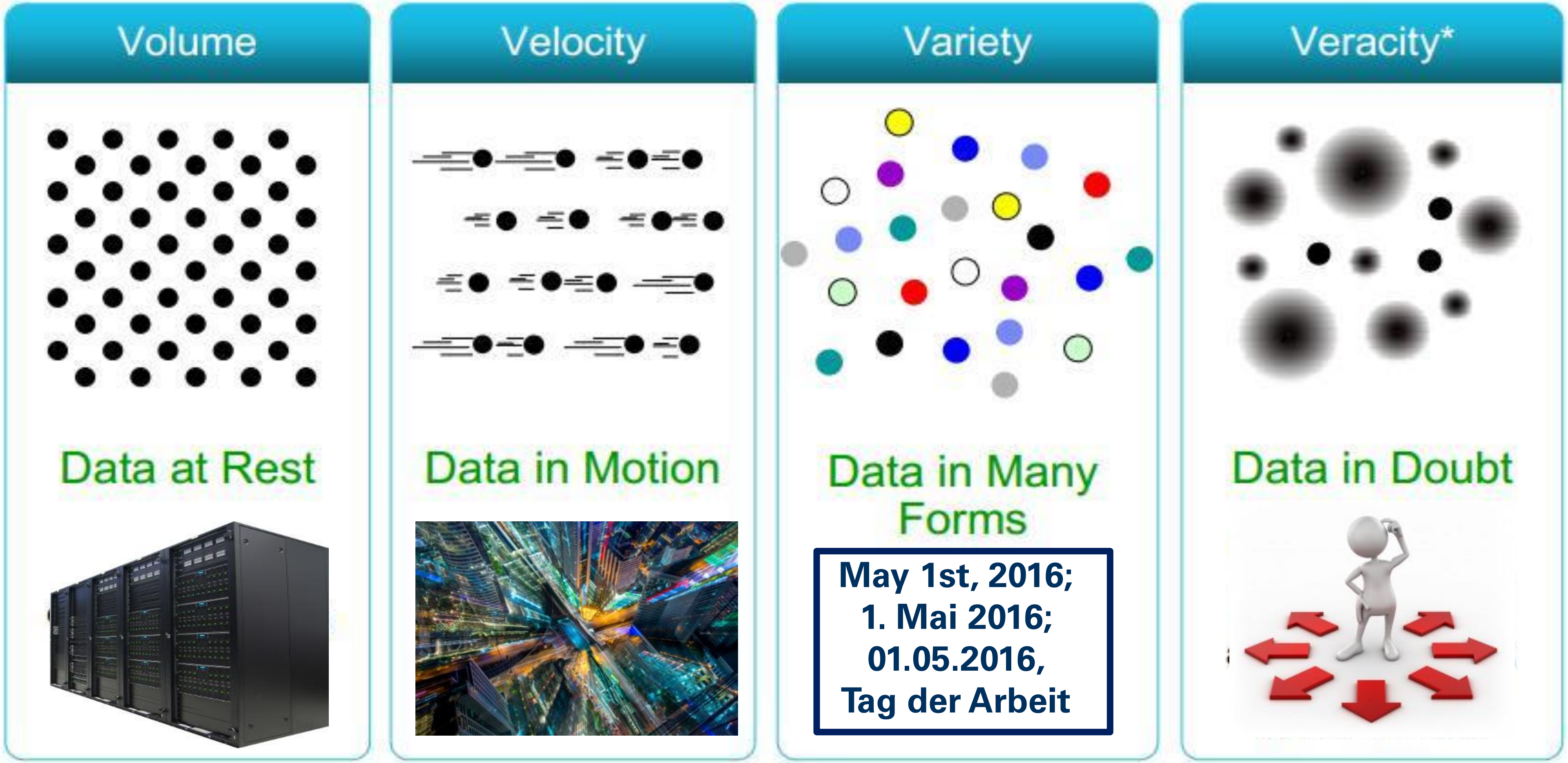
B kB MB GB TB PB EB ZB

What is big?



https://www.digitalistmag.com/cio-knowledge/2016/12/29/brontobyte-of-data-why-care-04773542

Big Data Definition(s)



May 1st, 2016;
1. Mai 2016;
01.05.2016,
Tag der Arbeit

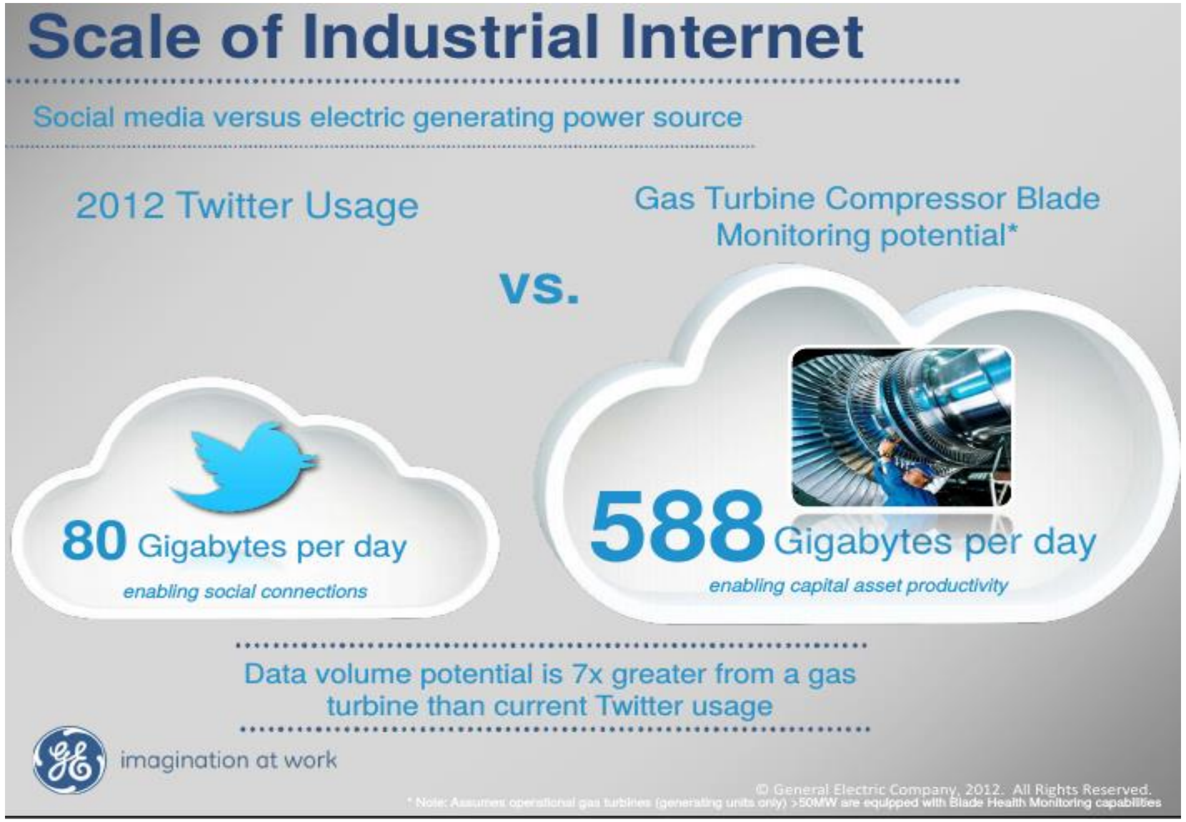
→ more important:

extract new content from database

Driver of Big Data development: Growing data, access to internet and connectivity – is there an commercial benefit?

Early picture (2012)

- Big data applications primarily considered to be of interest in commercial applications
- Influenced by but partly opposite to data base research (MapReduce)
- Outside HPC



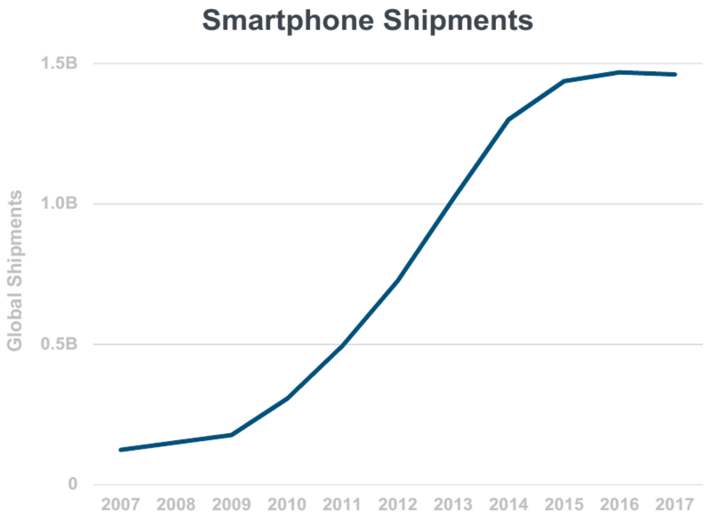
Fisher CIO Leadership Program: "Big Data Analytics: Making Big Data Work"

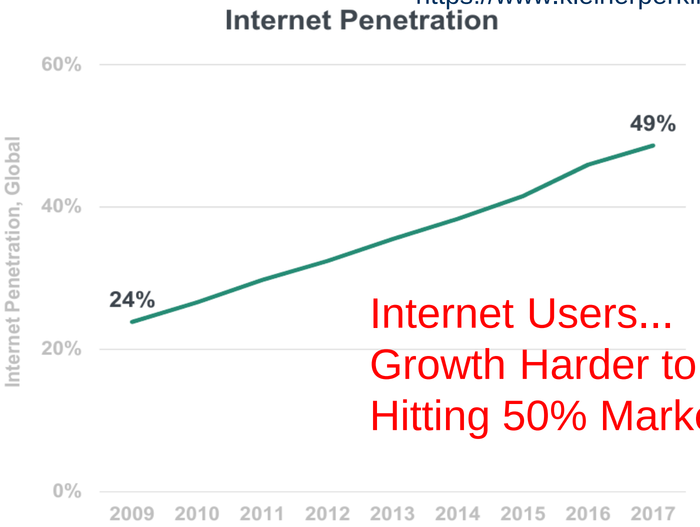
Bill Ruh, Vice President, Global Software Center, GE, November 1, 2012;

https://businessinnovation.berkeley.edu/wp-content/uploads/2018/03/fisher-bigdata-presentations.pdf

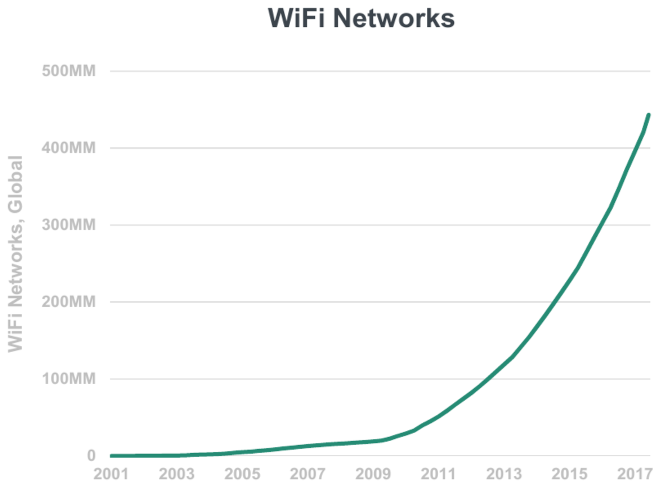
Driver of Big Data development: Data grows / data deluge

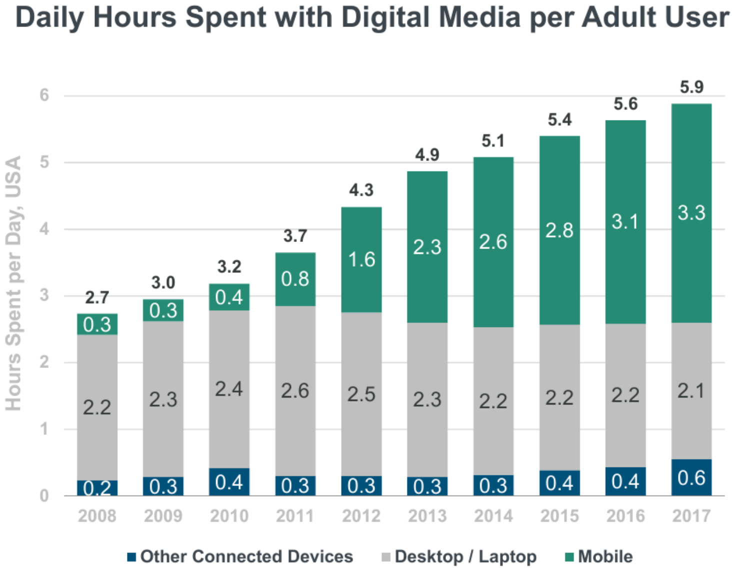
https://www.kleinerperkins.com/perspectives/internet-trends-report-2018





Internet Users...
Growth Harder to Find After
Hitting 50% Market Penetration





Driver of Big Data development: Data grows / data deluge

https://www.kleinerperkins.com/perspectives/internet-trends-report-2018

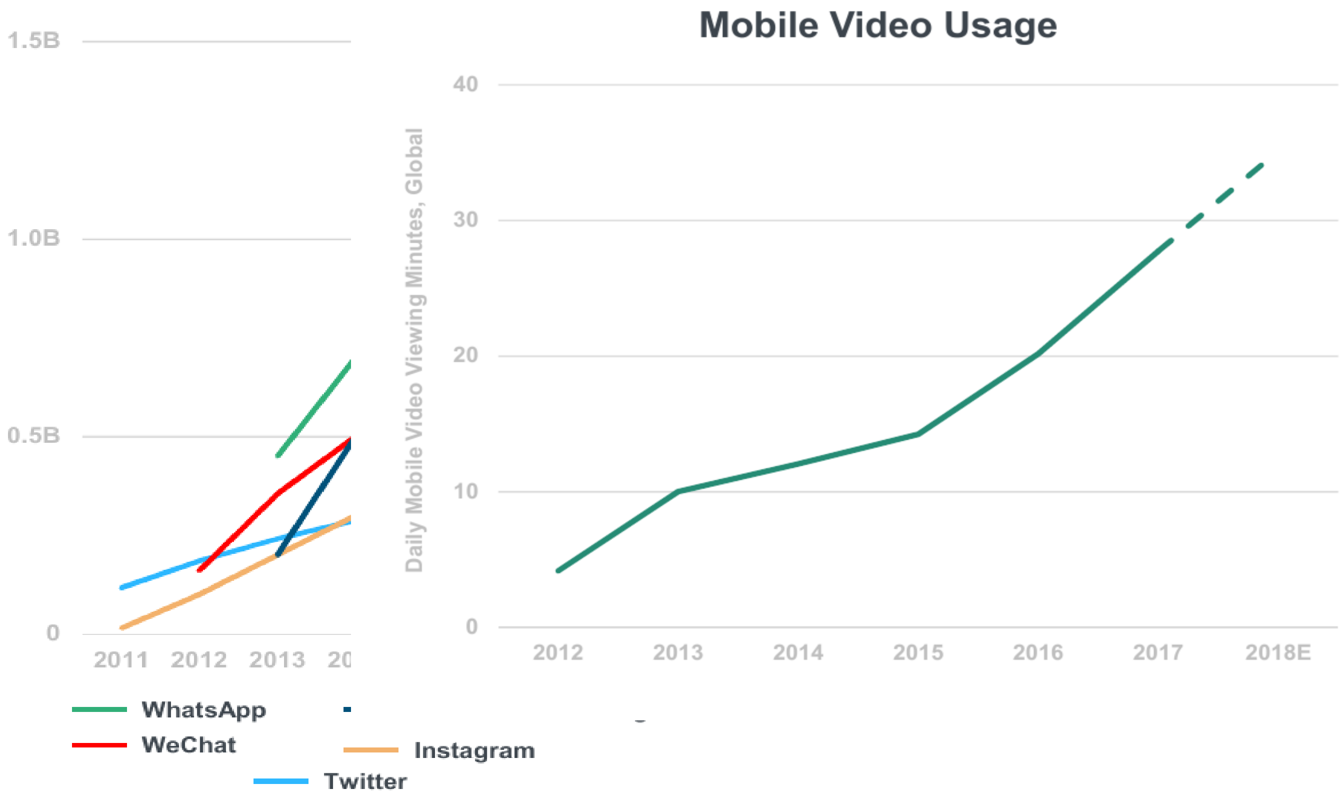
Messaging

Tencent (2000 → 2018)

QQ

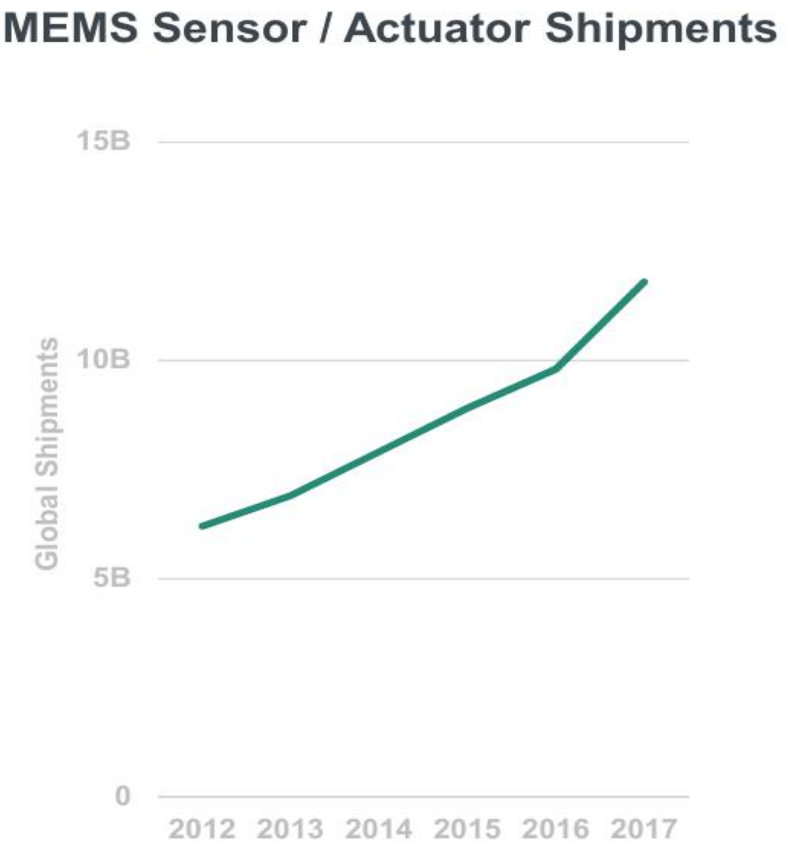
WeChat

Messenger MAUs



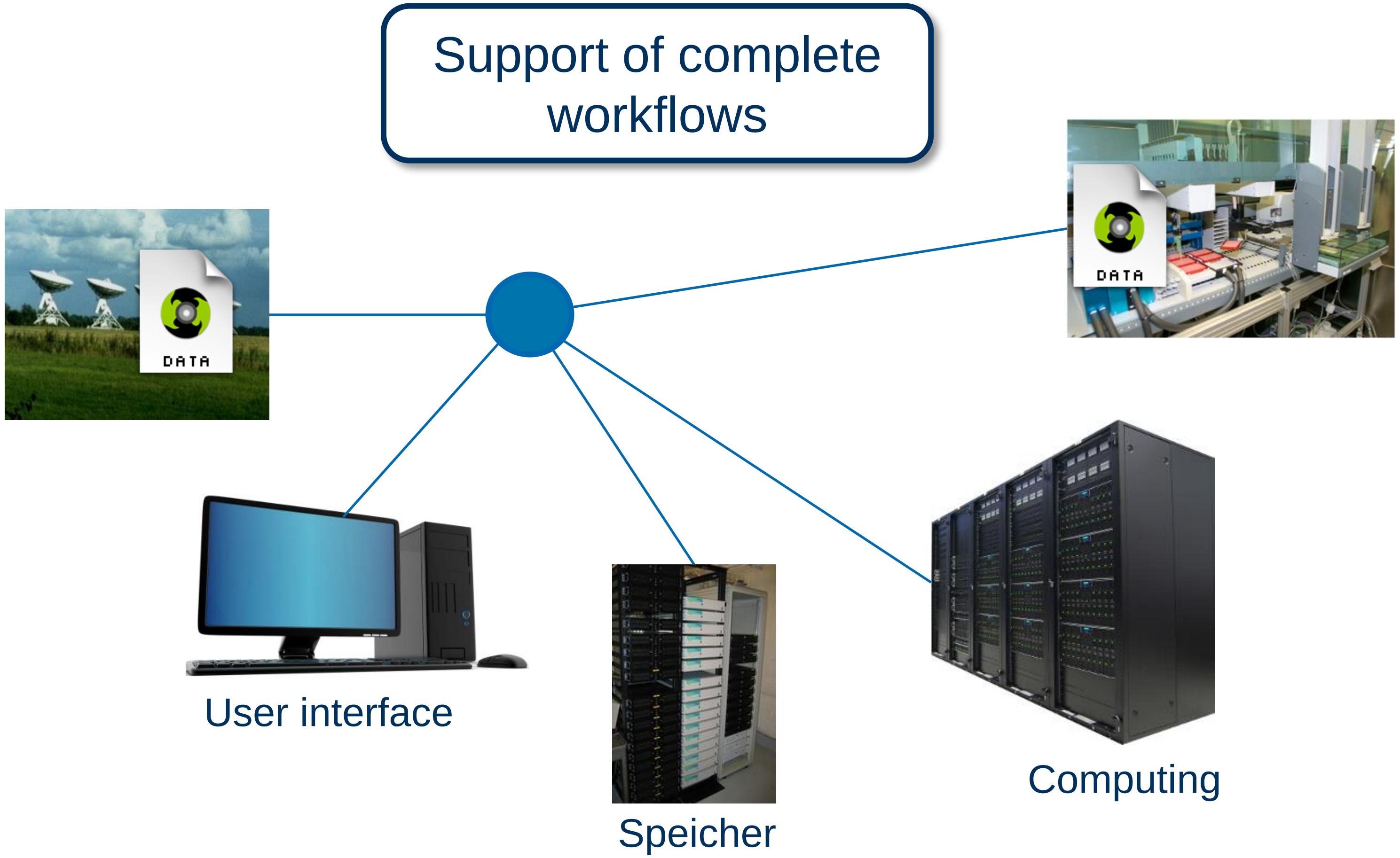
Data grows / data deluge

https://www.kleinerperkins.com/perspectives/internet-trends-report-2018



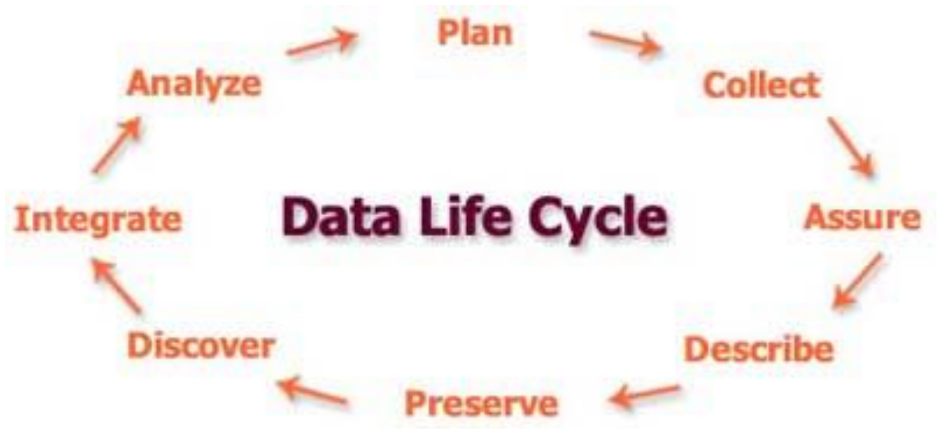
MEMS: Micro-Electro-Mechanical Systems

Users experience in sciences

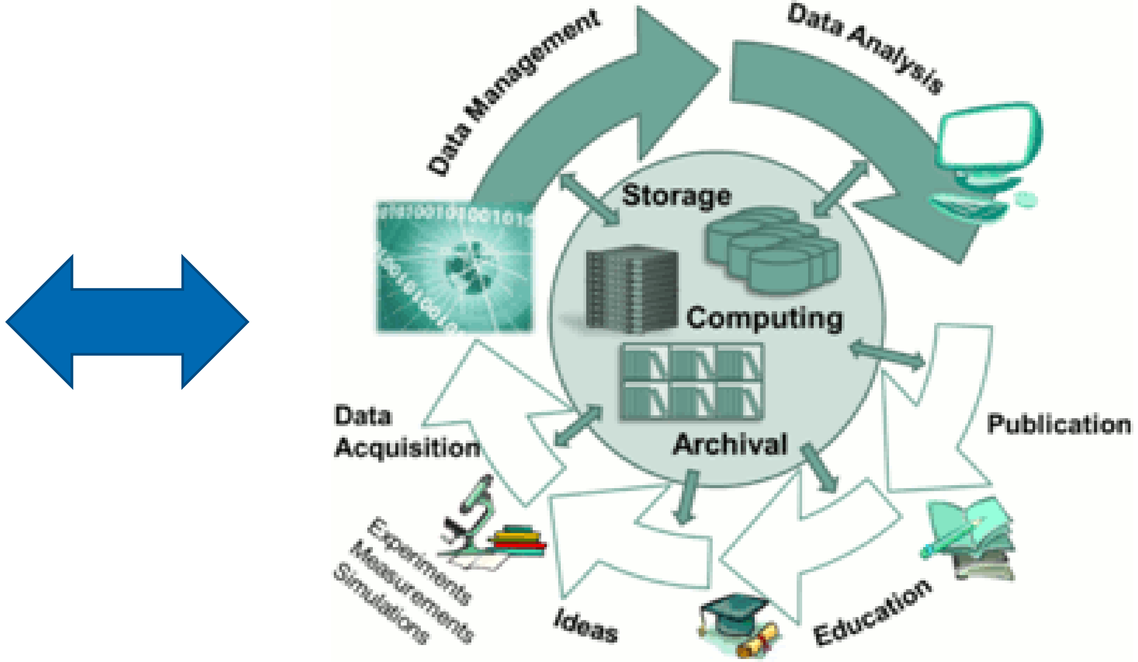


Science point of view: Data life cycle management

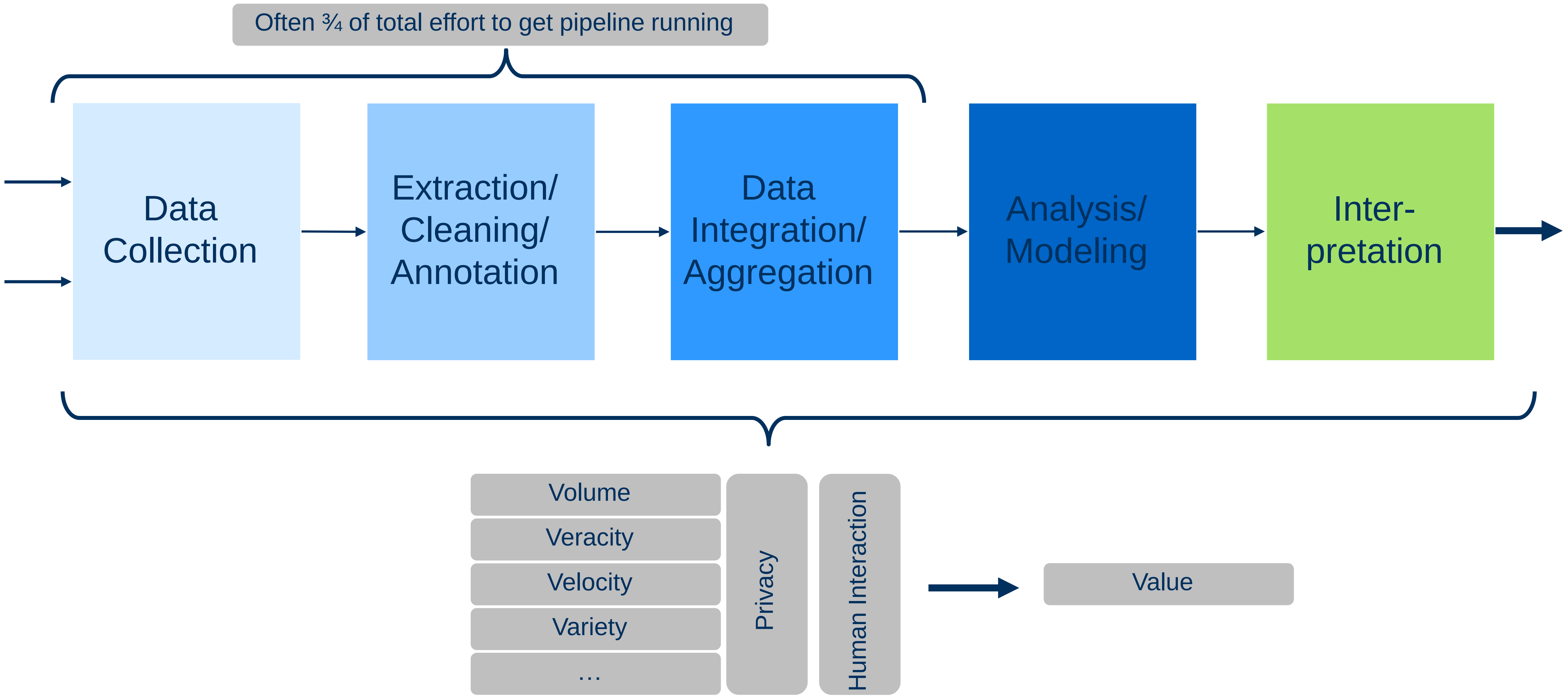
Data(flow) perspective



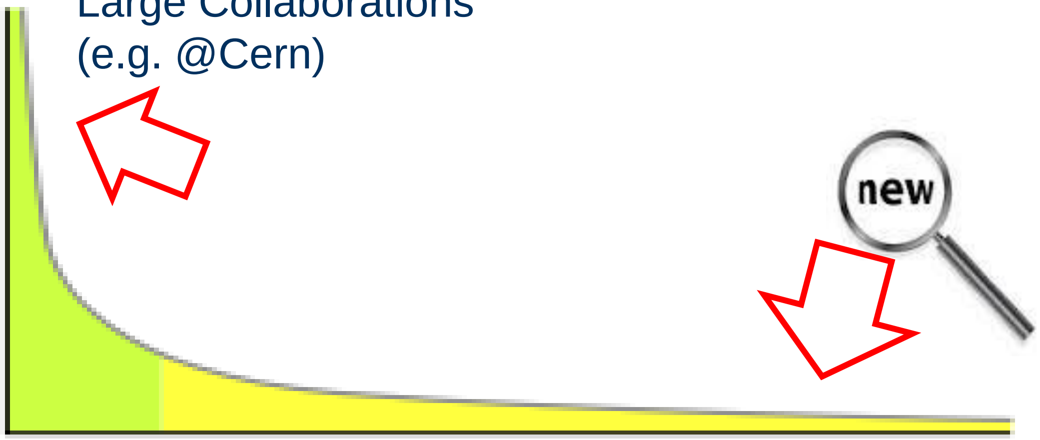
Systems perspective



Data analytics processing pipeline



In science: Not just “big players” – Long Tail of Science



Requirements from the Users perspective

- Data must be managed, annotated and curated to extract their potential
- Many research communities do not have the necessary tools to transform ever-growing data into scientific knowledge

Transportation

Engineering

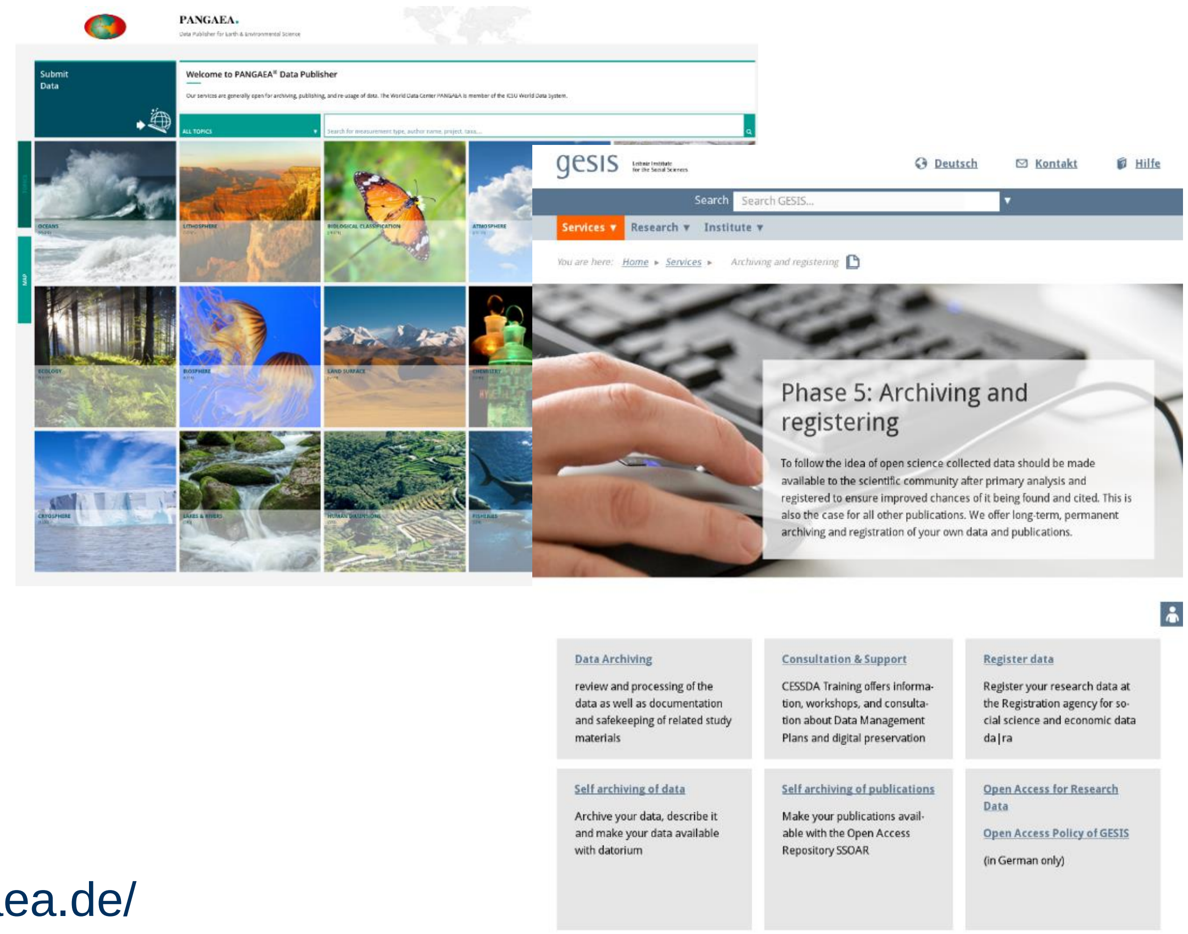
DNA sequencing

And many more!!!

Publishing data - repositories

Communities tend to organize their data themselves:
Disciplinary/domain repositories

- For one scientific discipline/domain
- Usually fixed metadata schema
- A central place to look for data in a domain
- Use it if there is one for you!



Examples

- Pangea – Geodata repository - https://www.pangaea.de/
- GESIS – Social sciences – https://www.gesis.org
- NOMAD Repository - Novel Materials Discovery Repository - http://nomad-repository.eu/cms/

Support the long tail of Science

Open data (open science) initiatives [EU18] [US18]

- Break boundaries to data silos – towards “living” data
- Foster re-usability, reproducibility and exchange among scientists
- Rich source for data science

Science Gateways and Virtual Research Environments

- Access computing infrastructures via standardized interfaces
- Hide (technical) complexity from the users perspective

Why was Big Data not HPC-driven?

Is big data killing HPC?

— Blog post “HPC is dying, and MPI is killing it” by Jonathan Dursi (from April 2015, see [DUR15])

Pictured: The HPC community bravely holds off the incoming tide of new technologies and applications.

[Dur15] http://www.dursi.ca/hpc-is-dying-and-mpi-is-killing-it/

Why is not everything HPC today?

Jonathan Dursi criticism on HPC in general

- "This should be a golden age for High Performance Computing." But it is not. Instead **new technologies are developed by other communities**.
- Analysis of Internet data and DNA sequencing brought **huge amounts of data** in new areas. **Why wasn't HPC the logical solution**?
- Prevailing "Not invented here" or "this is not real HPC" attitudes.
- HPC stayed with traditional concepts largely, both in hardware and software.
  - MPI (Message Passing Interface) was "killer app" in HPC, it was a firm standard for 27 years
  - Very high quality implementations, highest speed, constantly adapting to newest hardware
- **Other communities developed their own solutions, re-inventing several wheels, producing many successful new technologies and software (outside HPC)**.

Developments in Big Data

Apache Project Ecosystem

Big Data (36):

- Airavata, Ambari, Apex, Avro, Bigtop, BookKeeper, Calcite, CouchDB, Crunch, DataFu (Incubating), DirectMemory (in the Attic), Drill, Falcon, Flink, Flume, Giraph, Hama, Helix, Ignite, Kafka, Knox, MetaModel, Oozie, ORC, Parquet, Phoenix, Quarks (Incubating), REEF, Samza, Spark, Sqoop, Storm, Tajo, Tez, VXQuery, Zeppelin

Database (25):

- Accumulo; Cassandra, Cayenne, Cocoon, CouchDB, Curator, Derby, Empire-db, Forrest, Gora, **Hadoop**, Hbase, Hive, Jackrabbit, Lucene Core, Lucene.Net, Lucy, MetaModel, OFBiz, OpenJPA, ORC, Phoenix, Pig, Torque, ZooKeeper
- Many more to explore: [Apa17] https://projects.apache.org/projects.html?category

Extension of simplified Map/Reduce approach

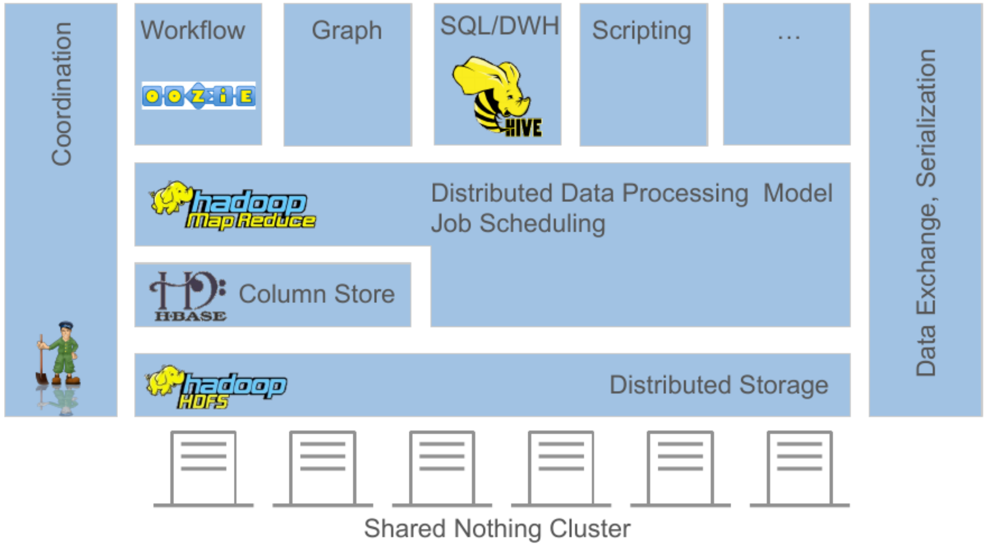
General Purpose

Big SQL

Big Graph

Big Stream

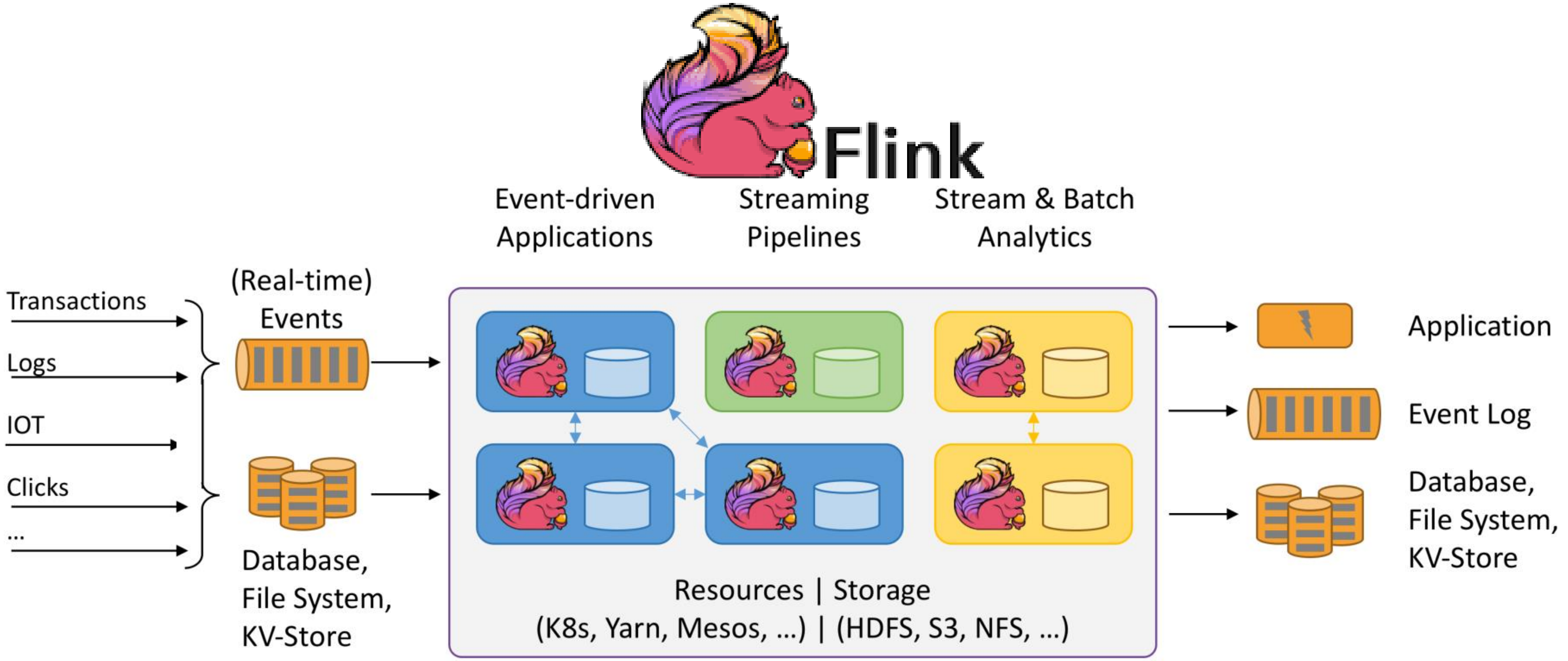
Hadoop 1.0 ecosystem



Synergies between Big Data and ML software stack

There are already several sophisticated frameworks supporting Machine Learning methods extending general Big Data frameworks

- Spark ML
- Flink ML
- Mahout
- H2O
- …



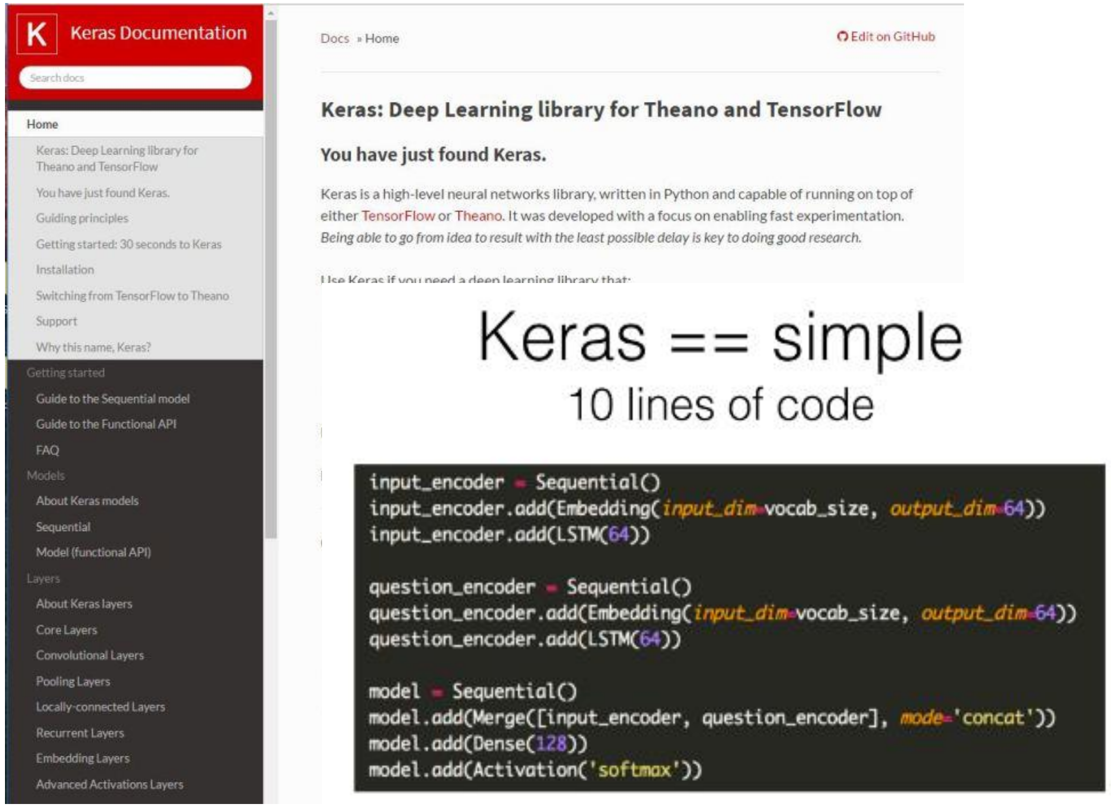
[ALEX14] A. Alexandrov, R. Bergmann, S. Ewen, J. C. Freytag, F. Hueske, A. Heise, O Kao, M. Leich, U. Leser, V. Markl and e. al, “The Stratosphere platform for big data analytics," , VLDB J. vol. 23, no. 6, 2014.

[CARB15] P. Carbone, A. Katsifodimos, A. Ewen, V. Markl and e. al:, "Apache Flink™: Stream and Batch Processing in a Single Engine.," IEEE Data Eng. Bull., vol. 38, no. 4, pp. 28-38, 2015.

Synergies between Big Data and ML software stack

There are already several sophisticated frameworks supporting Deep Learning applications like

- Keras
- Caffee
- CuDNN
- DeepLearning4j
- Tensorflow

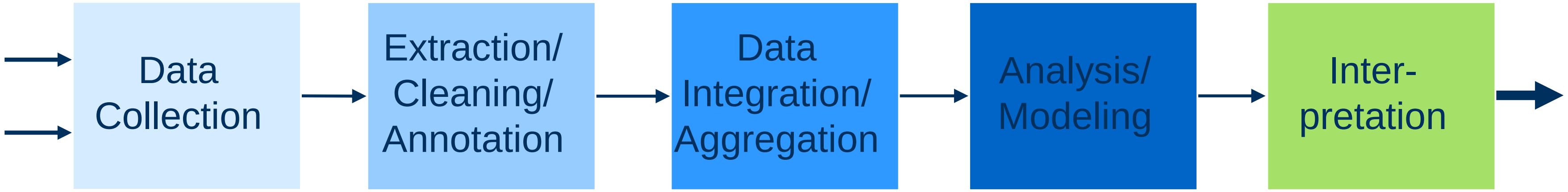


Data Analytics workflow requirements

If you start from the very beginning:

Goal: find the best match for complex analytical workflow – is there any?

- Abstraction: workflow consist of different stages and tools
- Every step has different resource requirement and data utilization



Data Integration

Data Aggregation/ Transformations

Classification/Interpretation

- How to realize? -> efficient data management, scheduling
- Is my implementation correct and performant?

From the systems perspective:
How to support users with infrastructures

HPC vs. Data Analytics

- Bring computing to data, or data to computing (data mover)?
- HPC: traditional rather monolithic usage, e.g. simulations
- Big Data analytics: more data centric, but not all and every analysis is embarrassingly parallel, iterative models still induce large data movements
- There is no unique big data blueprint! – which would fit all requirements
- Which way to follow – more HPC like approach or dynamic possibilities of Big Data frameworks?
  - Systems and infrastructure should support users, not forcing them to follow rigid regiments
  - Let user pick up approach, which is best for individual use case

Convergence patterns of HPC and Big Data

E.g discussed by BDEC [BDEC18]

- Experts group from BD and HPC researchers
- Recently published report discussing ideas about architectural changes and requirements for large-scale data analytics applications in “BDEC Pathways to Convergence Report” [BDEC17]
- Are there common convergence patterns?
  - From the application/users perspective
  - From the software ecosystems perspective
  - From the architectural perspective: centralized vs. decentralized

Convergence patterns of HPC and Big Data

From the application/users perspective:

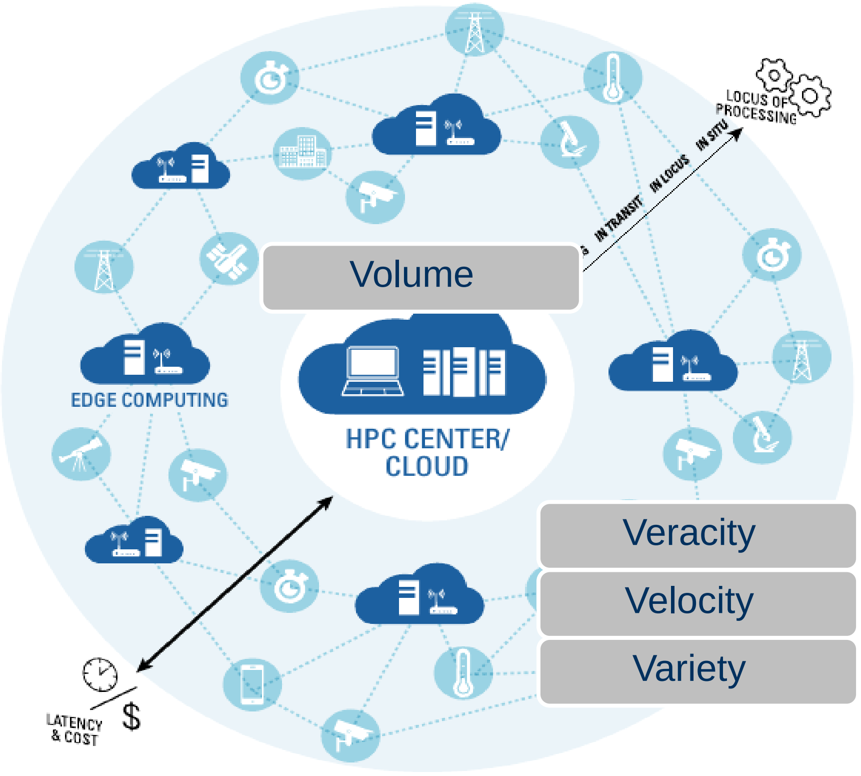
— Traditionally, data and compute is in rather close proximity, e.g. at the data center; not the case in cloud settings

— “Data concentration” gets reduced towards edges

— Developments driven by big data aspects “velocity”, “veracity” and “variety” (heterogeneous data)

— Dealing with the data “deluge” is harder at the edges of the data space

— Edge devices may produce high data rates: e.g. microscopes, gene sequencers, sensor networks, …



Convergence patterns of HPC and Big Data

From the application/users perspective:

— HPC + HDA (High-end Data Analytics) workflows

— Analytics scenarios are not singular applications, but rather complex workflows

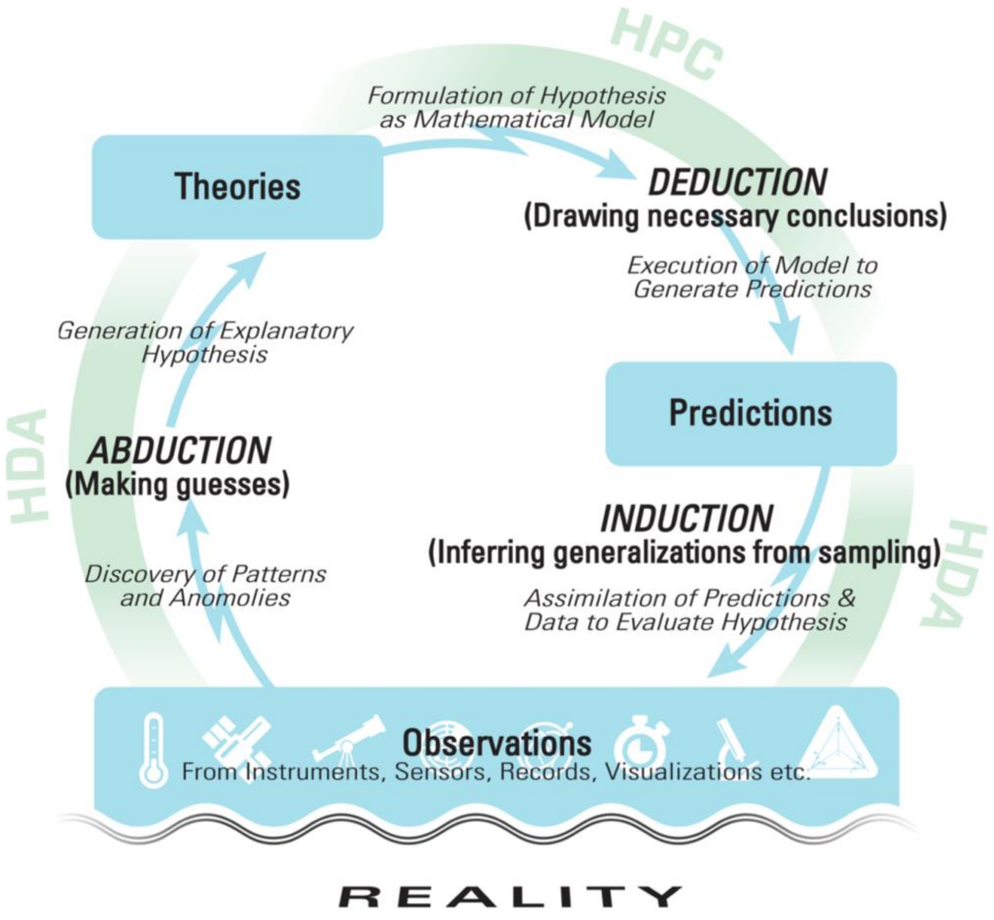
— Need to cover full data life cycle

— Need „user in the loop“: confirm model iterations or outcome of predictive analytics

— Potentially highly iterative

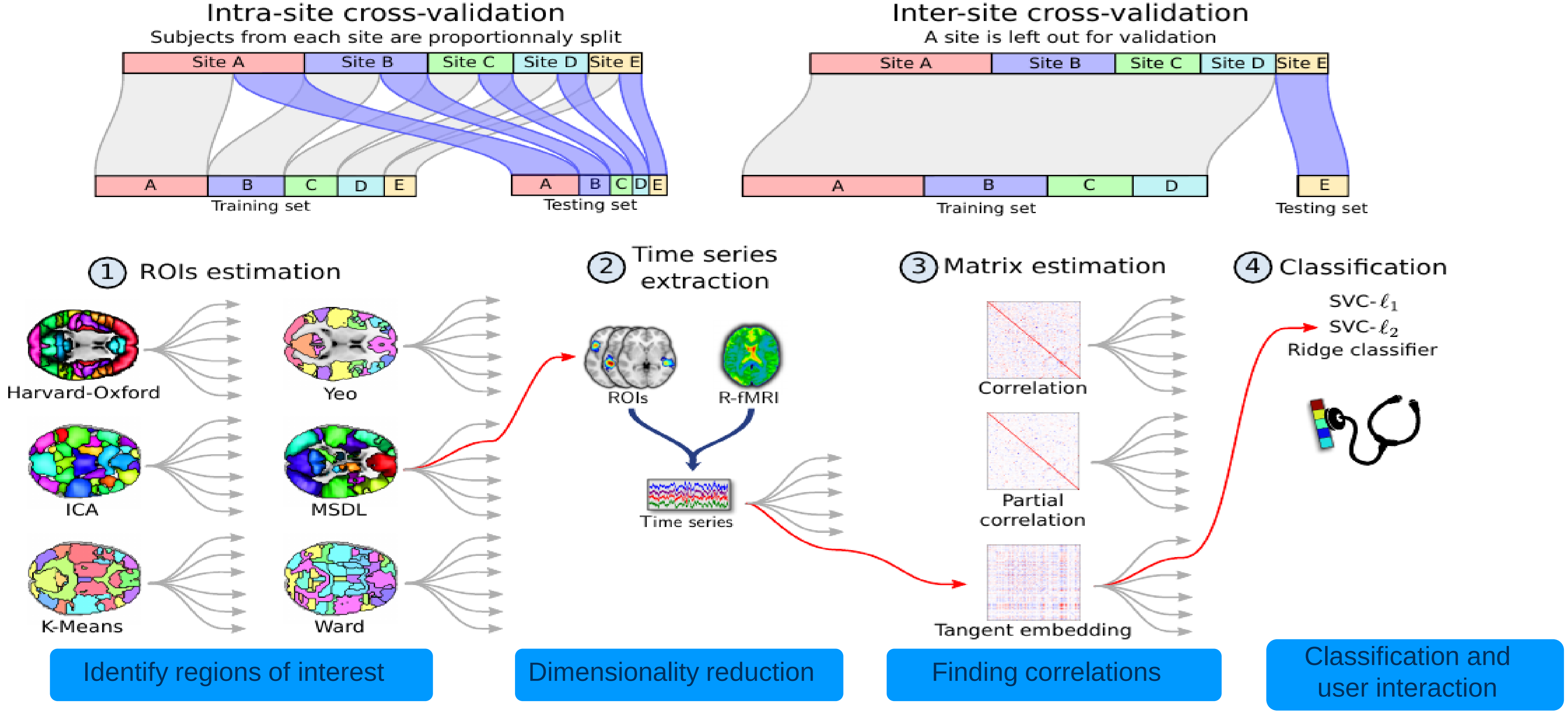
— Flexible enough to react on changes in software stack

- Interoperability of tools
- Reproducibility of results
- Adaptability to systems



Convergence patterns of HPC and Big Data

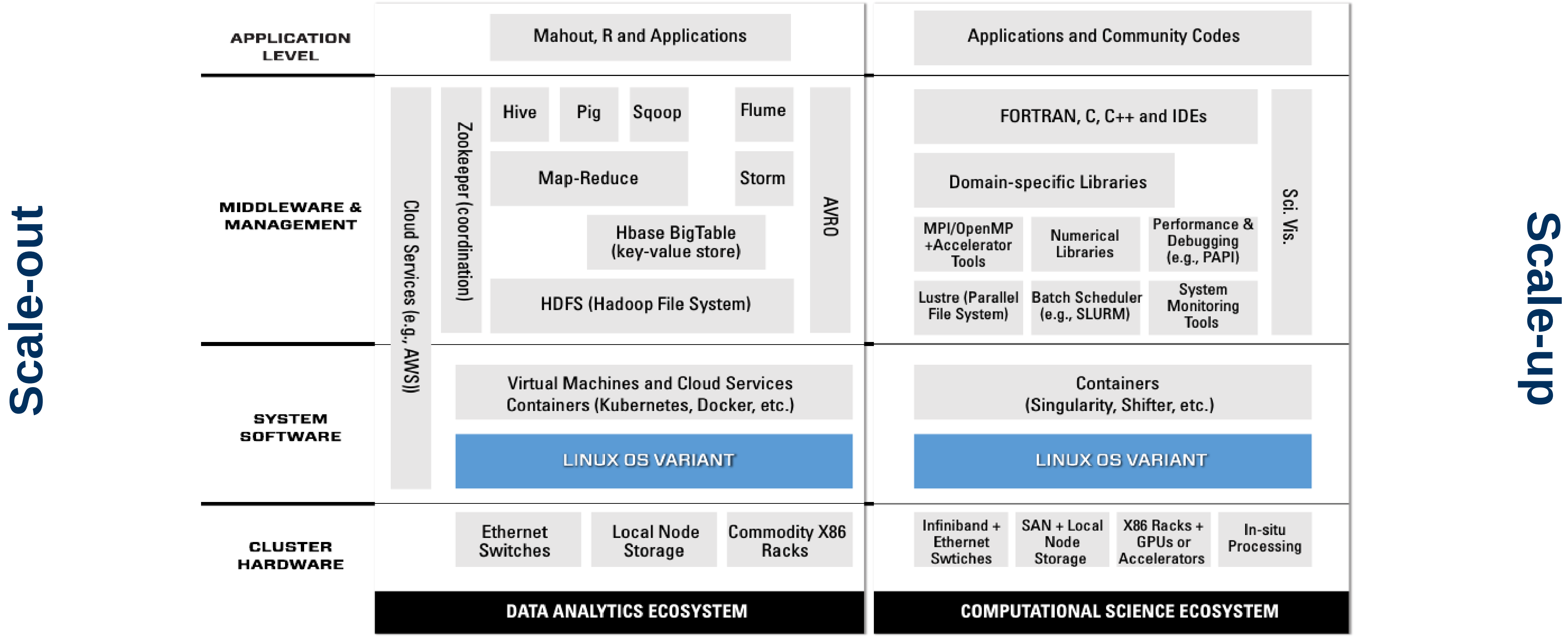
From the application/users perspective: Complex analytical MRI- workflows



[ABR16] Alexandre Abraham, Michael P. Milham, Adriana Di Martino, R. Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage, 147:736 – 745, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2016.10.045. URL http://www.sciencedirect.com/science/article/pii/S1053811916305924.

Software stacks of Big Data and HPC

From the software ecosystems perspective: The past and current status

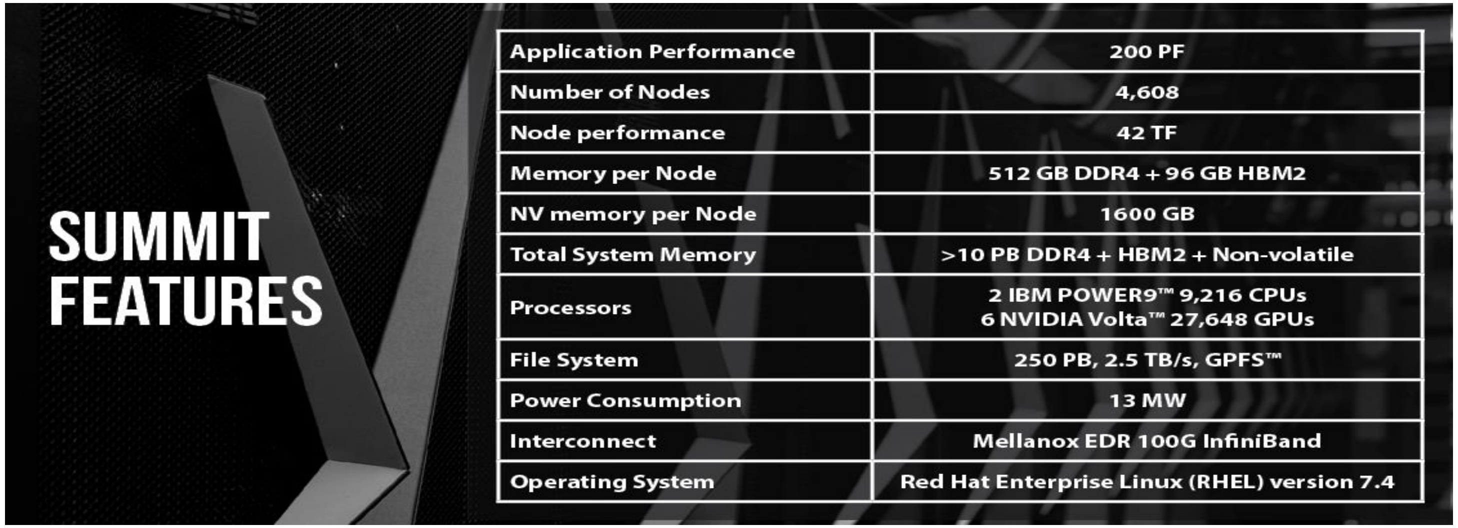


How to fuse the two worlds?

Current trend: Specialization
Top 1 in HPC 500 – June 2018

Summit - IBM Power System AC922

- IBM POWER9 22C 3.07GHz
- NVIDIA Volta GV100
- Dual-rail Mellanox EDR Infiniband
- Majority of cores provided by GPU-units, not classical by CPU cores



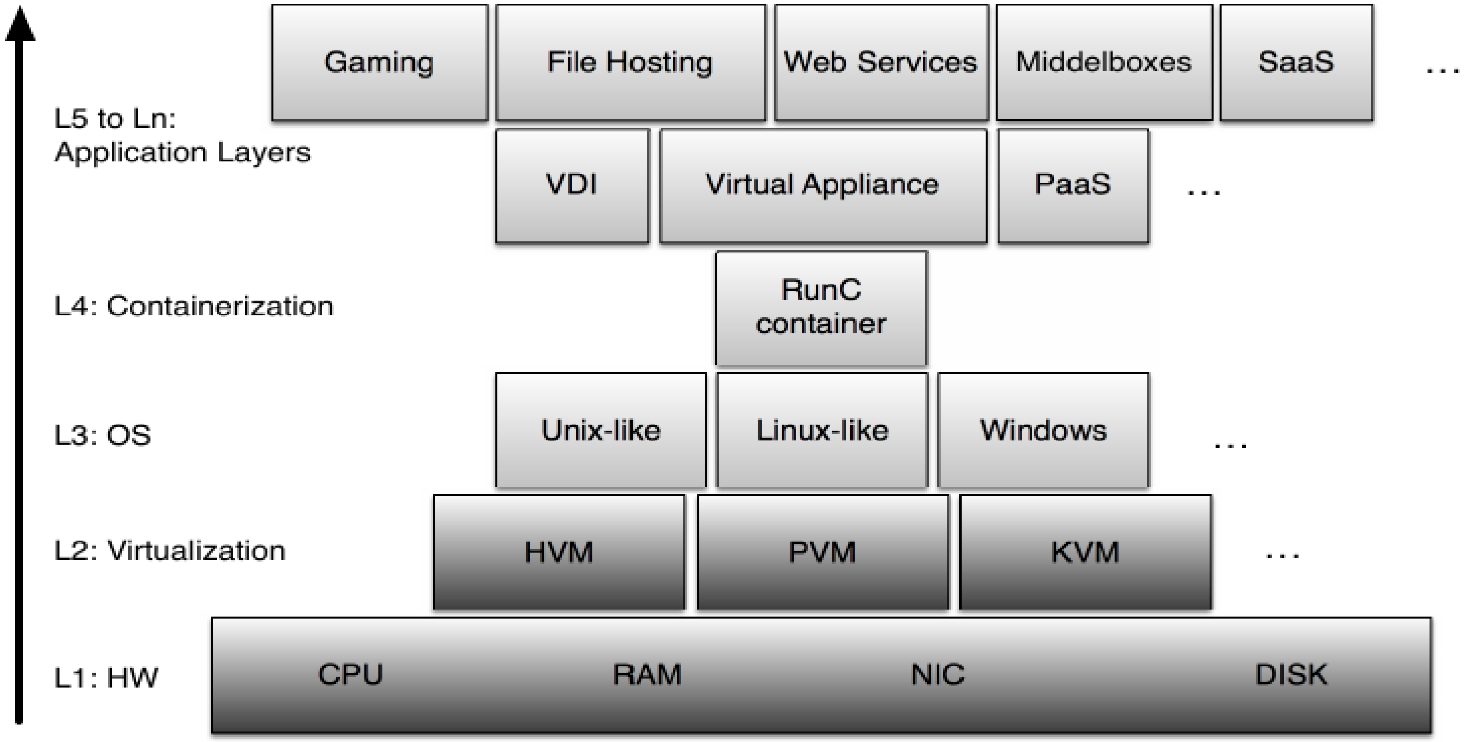
Application Performance	200 PF
Number of Nodes	4,608
Node performance	42 TF
Memory per Node	512 GB DDR4 + 96 GB HBM2
NV memory per Node	1600 GB
Total System Memory	>10 PB DDR4 + HBM2 + Non-volatile
Processors	2 IBM POWER9™ 9,216 CPUs 6 NVIDIA Volta™ 27,648 GPUs
File System	250 PB, 2.5 TB/s, GPFS™
Power Consumption	13 MW
Interconnect	Mellanox EDR 100G InfiniBand
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4

- Current “Exaflop application”: Machine Learning application on Summit using reduced precision
  - comparative genomics code at 1.88 exaflops using the Tensor Core capability

[TOP500] https://www.top500.org/news/summit-up-and-running-at-oak-ridge-claims-first-exascale-application/

Convergence patterns of HPC and Big Data

From the software ecosystems perspective: container and virtualization as bridging element



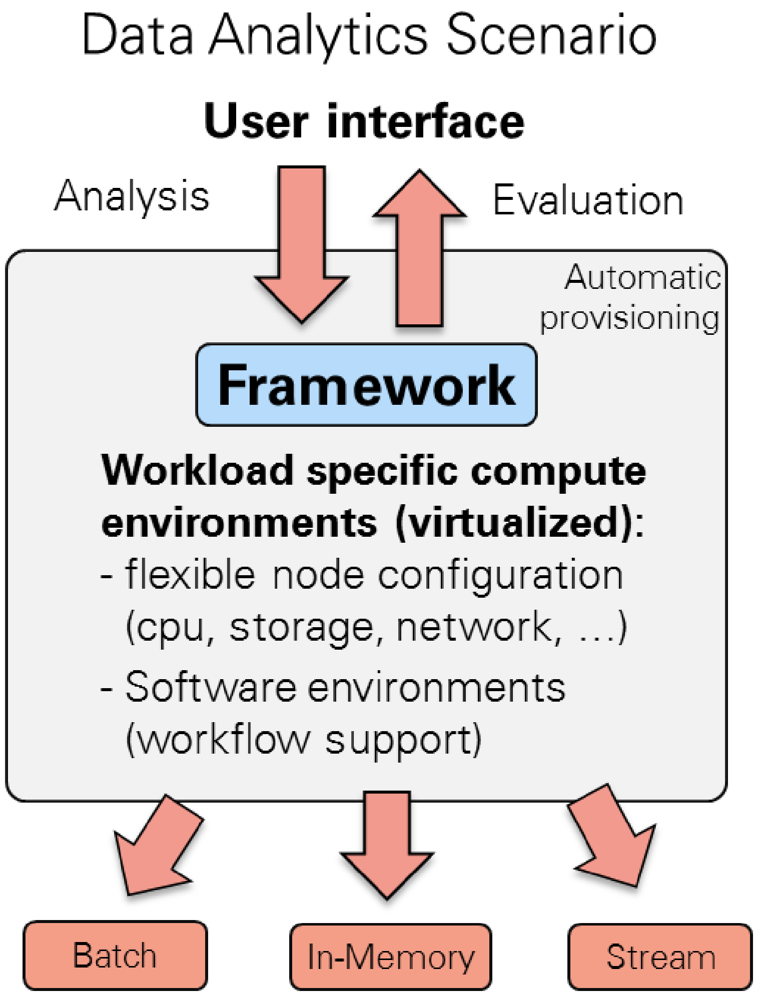
kubernetes

[FU16] Fu, Jiangchuan Liu, Xiaowen Chu, and Yueming Hu. Toward a standard interface for cloud providers: The container as the narrow waist. IEEE Internet Computing, 20(2):66–71, 2016.

Big Data and HPC – convergence patterns

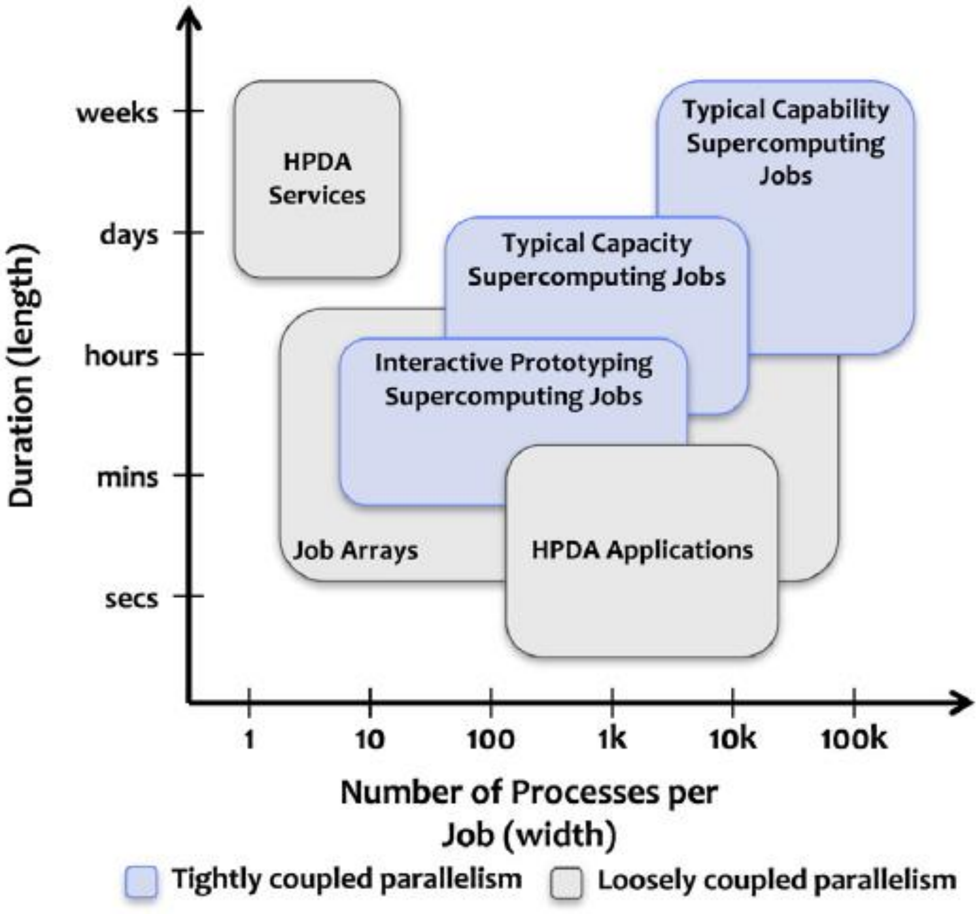
Requirements to support Big Data workloads on HPC

- Support frameworks/environments: more versatile software stacks
- Fast access to data: not just self-production of data (simulation), but also use 3rd-party data (open data, domain repositories)
- Support different data processing paradigms on the same system
  - Streaming / Batch / Iteration
- Better support of evaluation of (temporary) results, e.g. visualization frontends
- Service orientation (working environments)



Convergence patterns of HPC and Big Data: Improve HPC schedulers

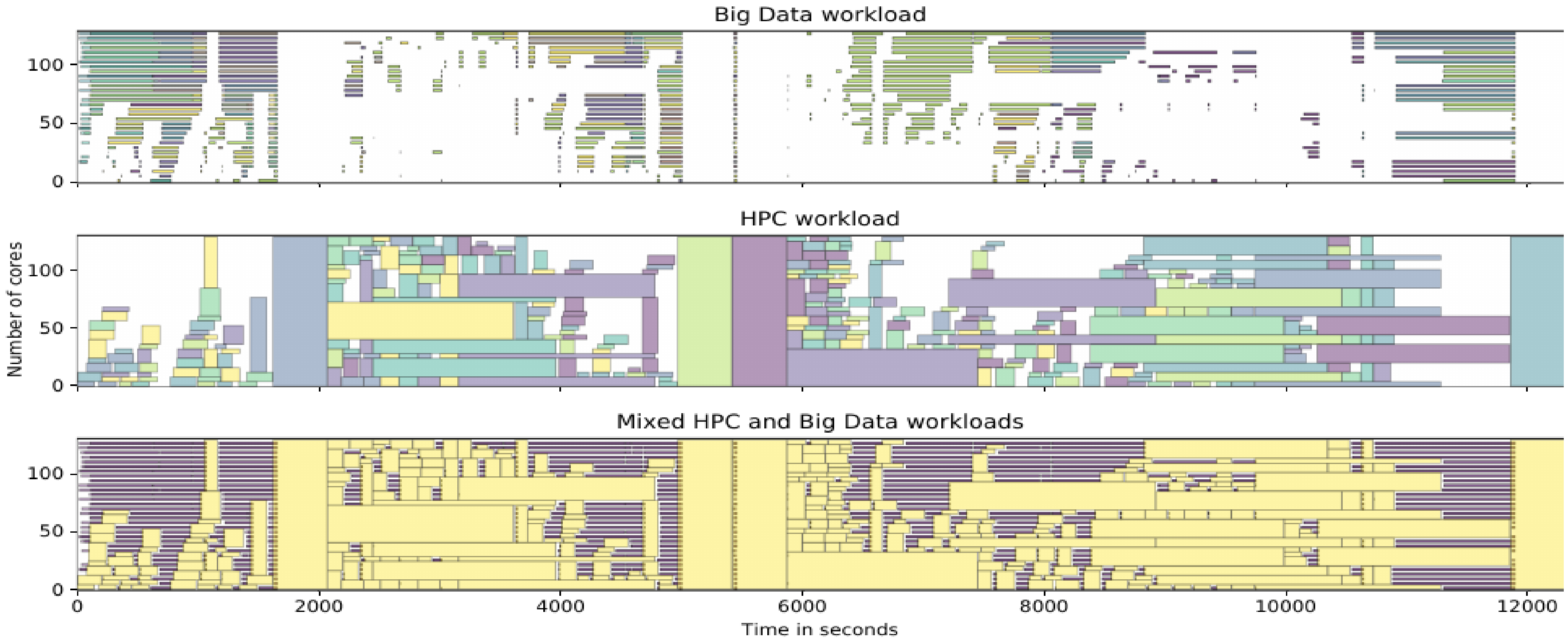
- HPC and Big Data schedulers have many common features.
- HPC schedulers handle parallel jobs better, while Big Data ones have better API.
- Schedulers display little overhead for jobs longer than 30 s.
- Some schedulers have significant overhead for jobs shorter than 10 s.



[REU18] A. Reuther et al. J. Parallel Distrib. Comput. 111 (2018) 76–92

Big Data Analytics @HPC

Ongoing research to integrate different workload on HPC architectures



[MER17] M. Mercier, D. Glesser, Y. Georgiou and O. Richard, "Big data and HPC collocation: Using HPC idle resources for Big Data analytics," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 347-352. doi: 10.1109/BigData.2017.8257944

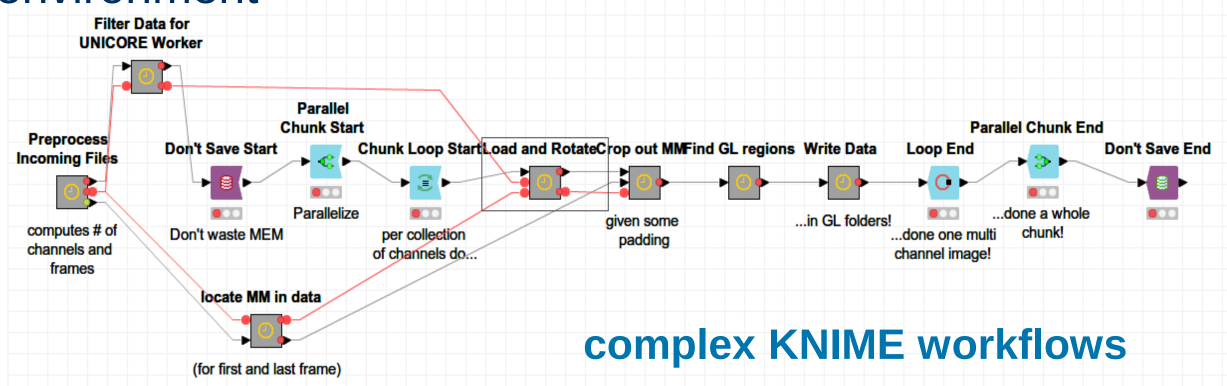
Some own experiences….

Execution of large data-driven workflows

Execution of data-intensive user workflows in HPC environment

— No prior HPC-knowledge required on user side

— Formulation of workload directly in workflow environment



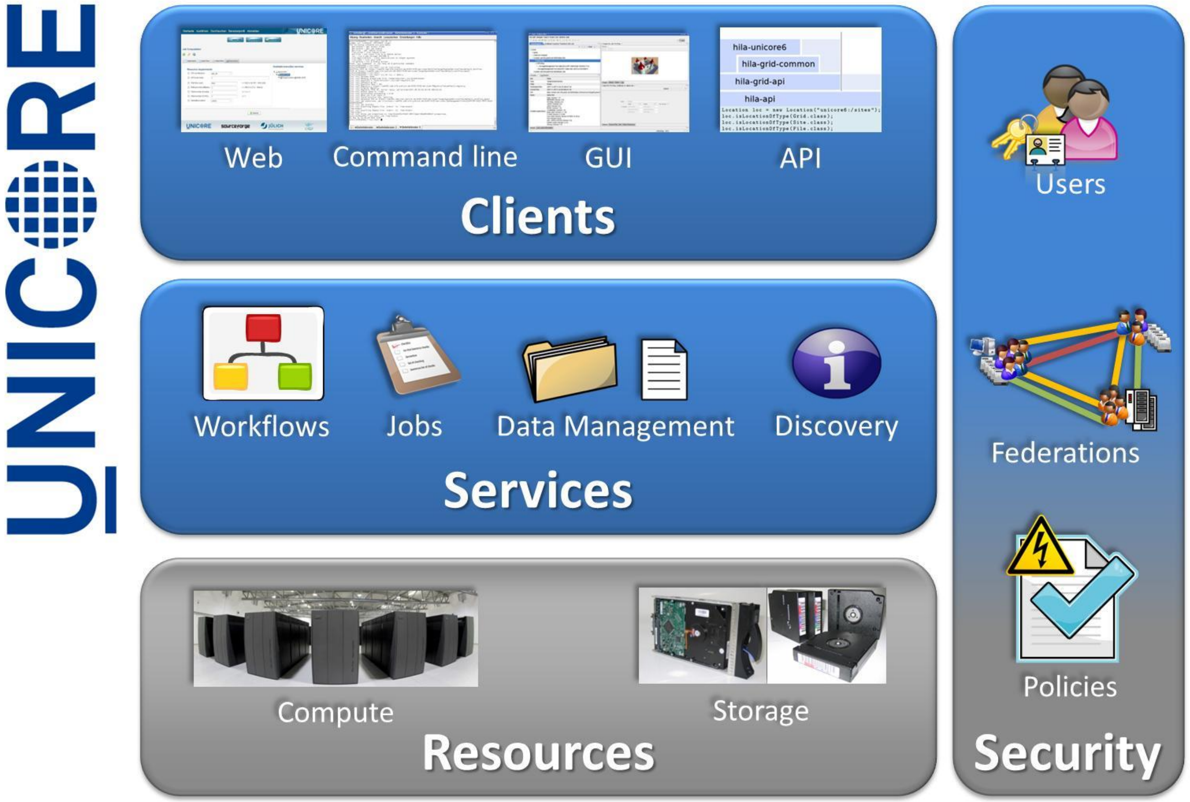
complex KNIME workflows

— Integration of HPC scheduling mechanism into workflow modelling framework to offload tasks or complete workflows on HPC cluster

— Middleware UNICORE used for HPC interaction

Execution of large data-driven workflows

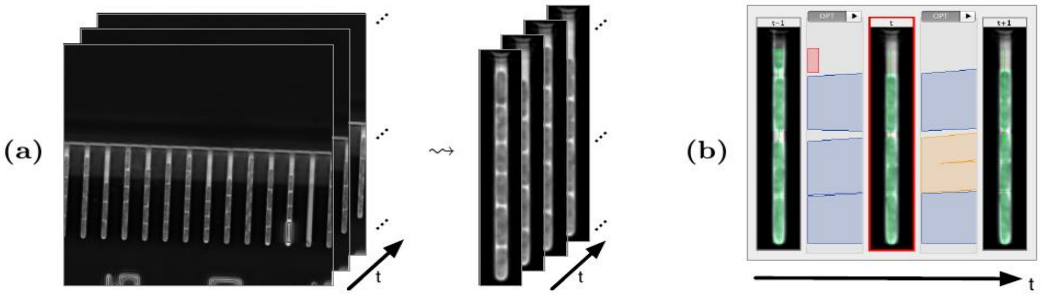
http://www.kdnuggets.com/2016/02/gartner-2016-mq-analytics-platforms-gainers-losers.html

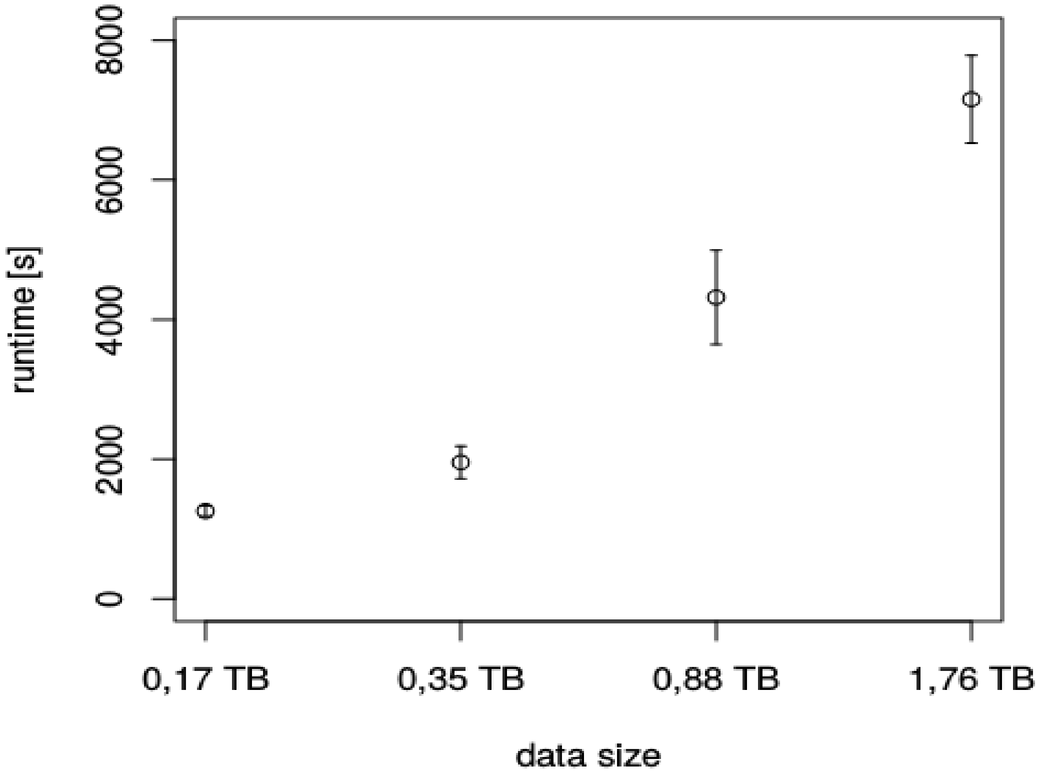


Execution of large data-driven workflows

Use Case: processing pipeline for cell tracking (bacteria E.coli) over time in collaboration with MPI-CBG + KNIME + UNICORE + TUD

- Experiment data set: 1,8 TB in ~7,5 M files coming from fast microscopes
- Reconstruction: follow cell movement over time as preprocessing step for development analysis
- Runtime improvement:
  previously: 17d on 4 cores
  now: 2h on 800 cores

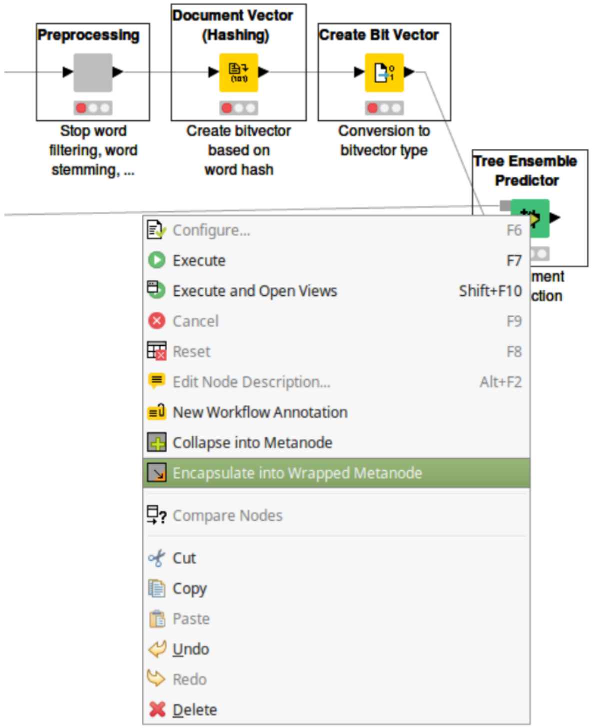


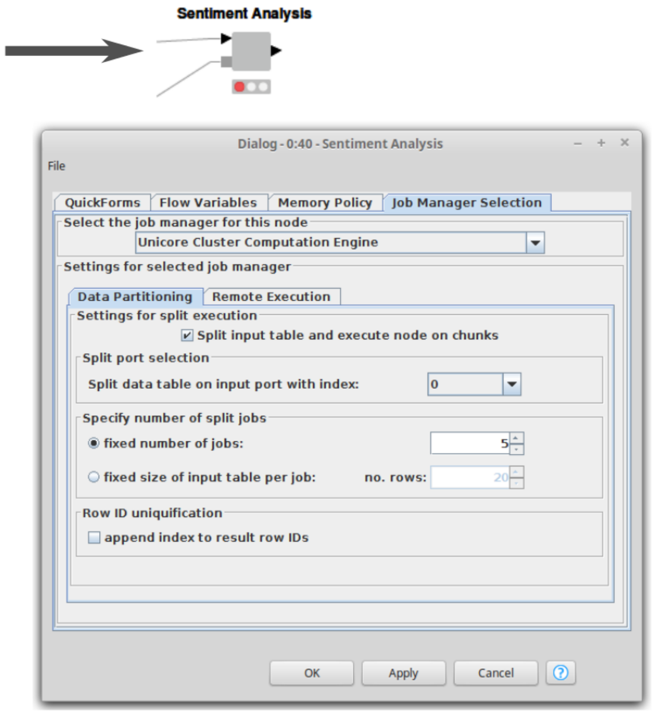


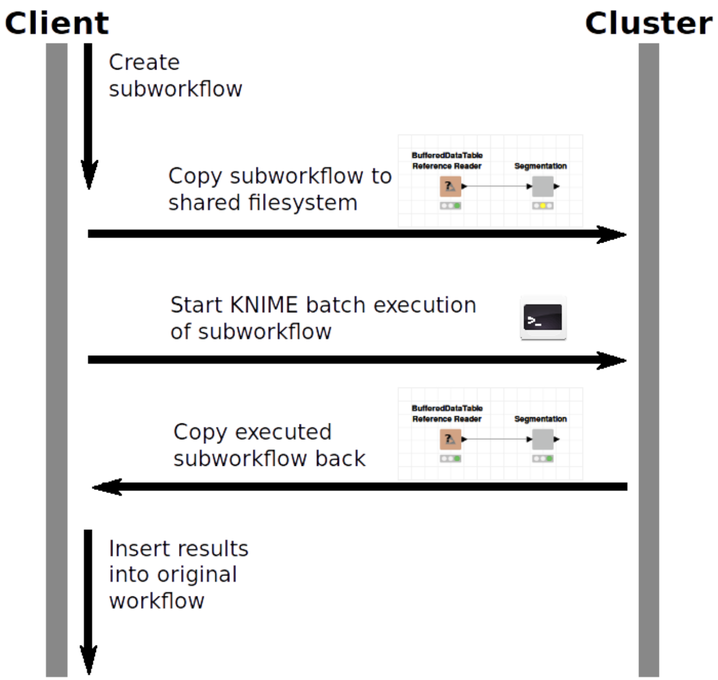
Execution of large data-driven workflows

Extension: selection of relevant sub-tasks (nodes) for remote processing

Specification of parameters for cluster interaction and splitting of data into bunches





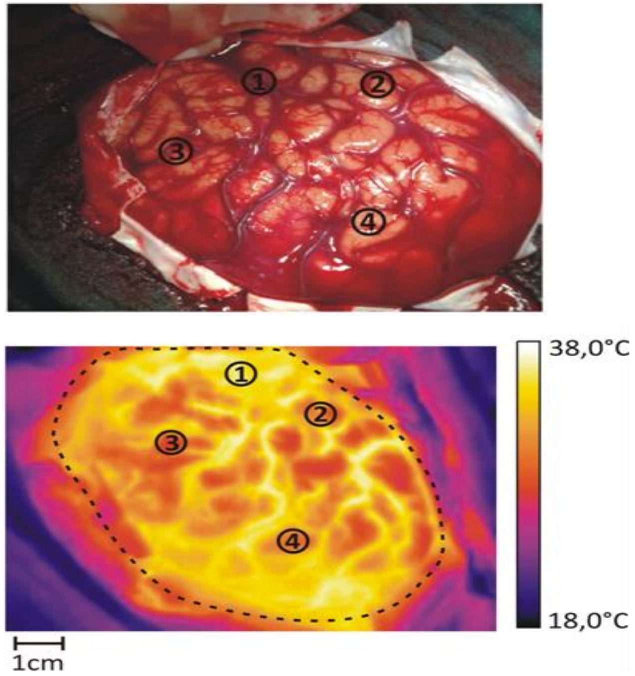


Data Analytics for Intraoperative Thermal Imaging

Application area: low delay operation support using thermal imaging processing

Quasi real-time data processing required in decision support during surgery – University Hospital Dresden (UKD)

- Neural activity monitoring require long-term intraoperative measurements (~10 minutes)
- Fast preprocessing required to decrease delay for subsequent analysis workflows and result presentations
  => minimize overall OP delay
- Iterative process: 3000 frames (5.4 GB) have to be processed every minute (50 Hz sampling rate)



Thermal image of acute subdural hematoma

[HOF17] N. Hoffmann et al. Learning Thermal Process Representations for Intraoperative Analysis of Cortical Perfusion During Ischemic Strokes. Springer International Publishing, 2016.

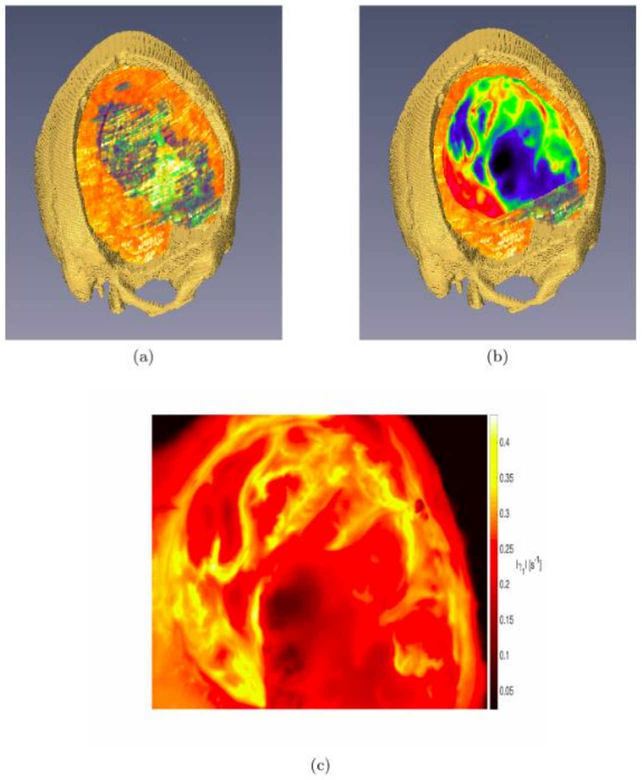
Data Analytics for Intraoperative Thermal Imaging

Application area: low delay operation support using thermal imaging processing

Quasi real-time data processing required in decision support during surgery – University Hospital Dresden (UKD)

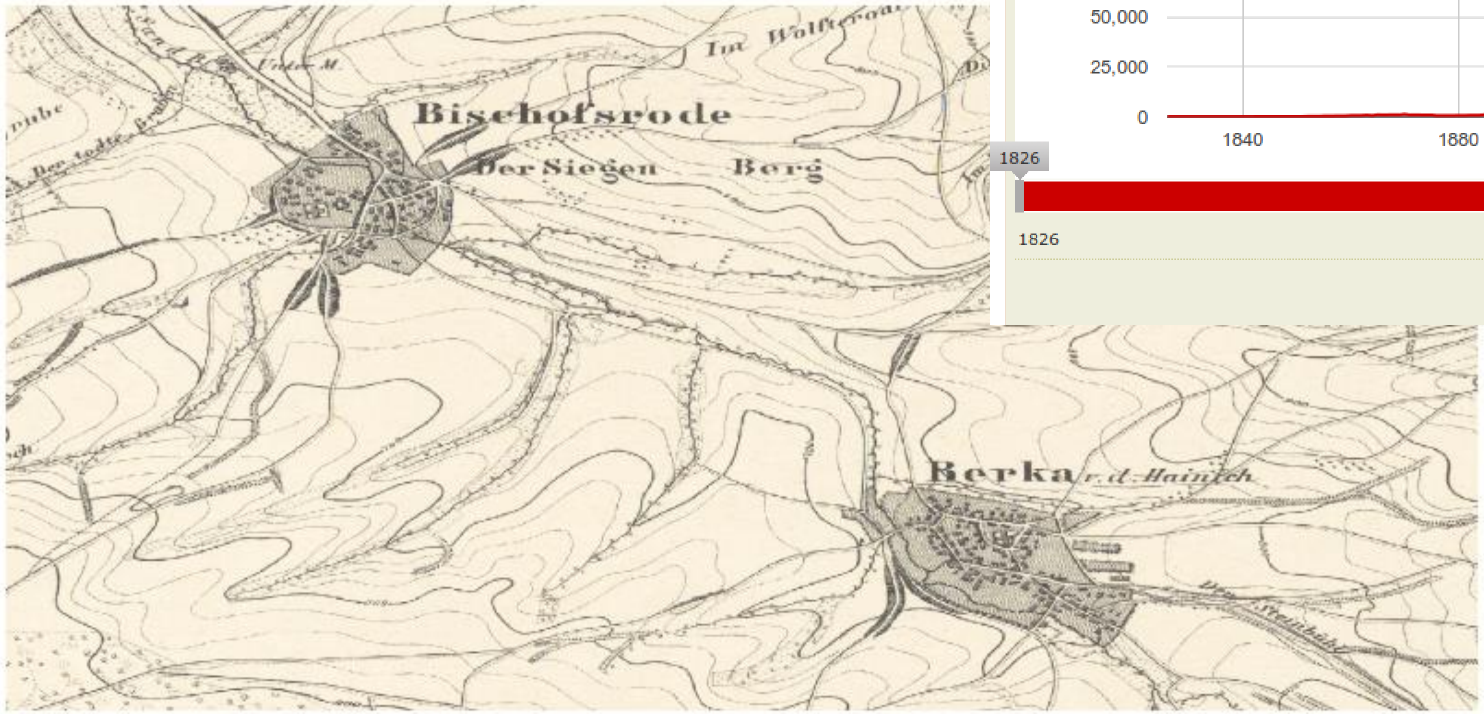
- Solution: Provision of Spark-Cluster @HPC
  - Fast SSD-backend to speed-up IO; fail-safe storage of imaging data
- Runtime improvement:
  - UKD-workstation: ~7000s/30.000 images
  - Spark cluster @Taurus: ~32s/30.000 images

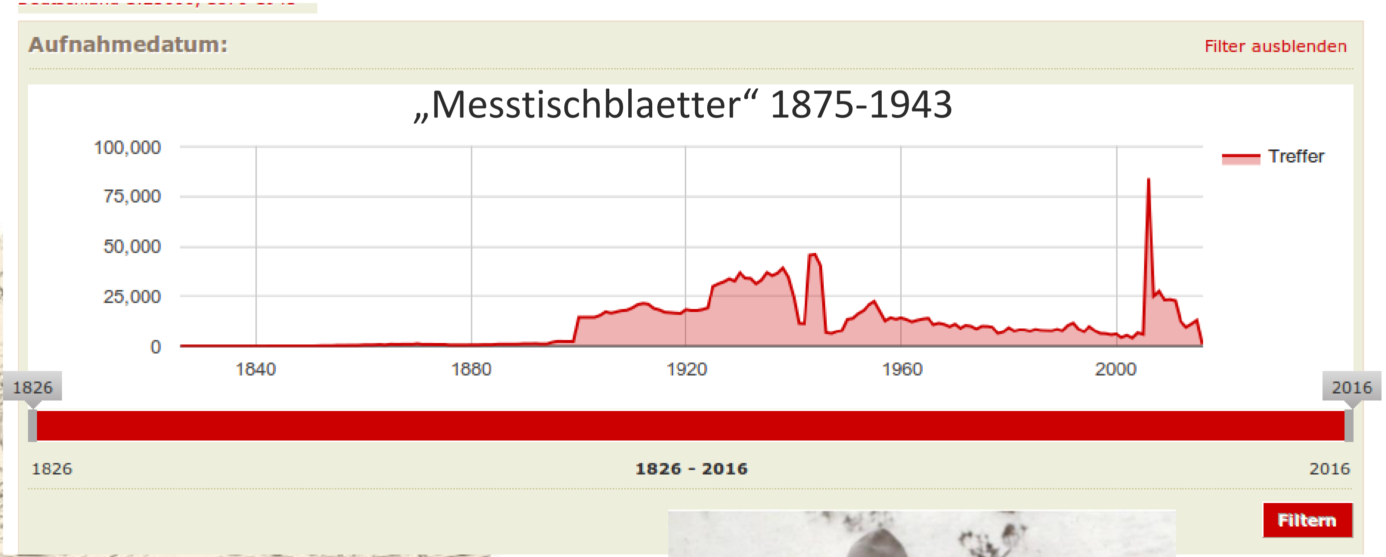
→ >200x faster



[HOF17] N. Hoffmann et al. Learning Thermal Process Representations for Intraoperative Analysis of Cortical Perfusion During Ischemic Strokes. Springer International Publishing, 2016.

Use Case: land-use detection and monitoring settlement detection using binary segmentation

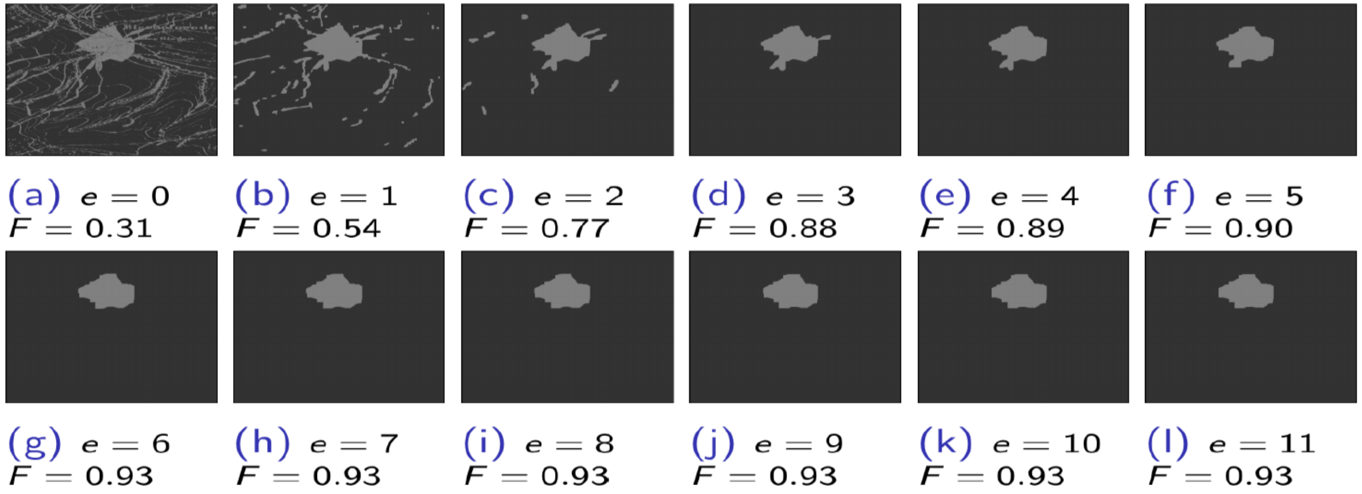


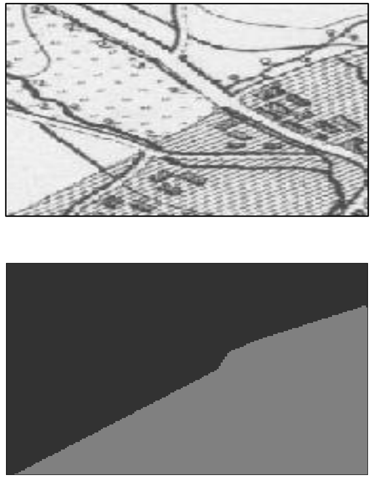


Use Case: land-use detection and monitoring settlement detection using binary segmentation

Binary segmentation: standard tool in Computer Vision to distinguish between regions of interest

- Settlement/background in maps
- Collaboration between Computer Vision experts, environmental scientists and data center (HPC)





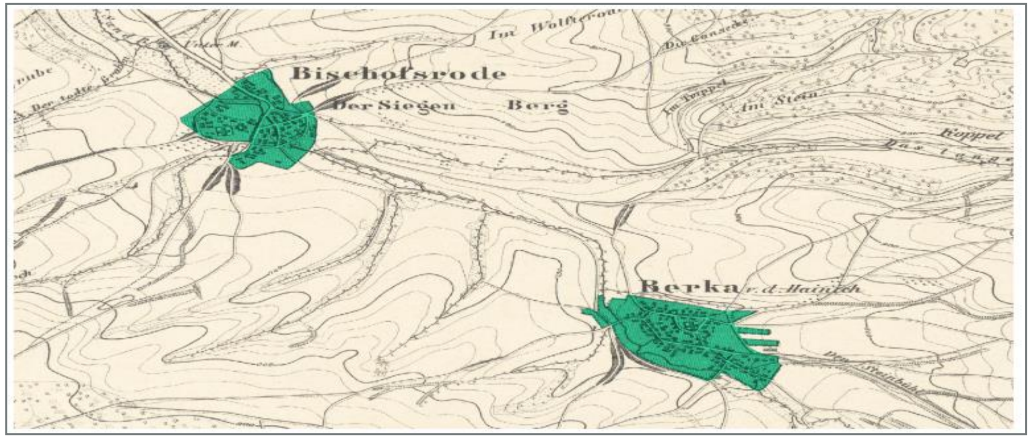
Use Case: land-use detection and monitoring settlement detection using binary segmentation

Results:

— Train model in minutes on HPC (instead of hours on workstations)

— Random forest (RF) + conditional random field (CRF), now evaluating DL approaches

— Required to re-iterate model on different input types of ma

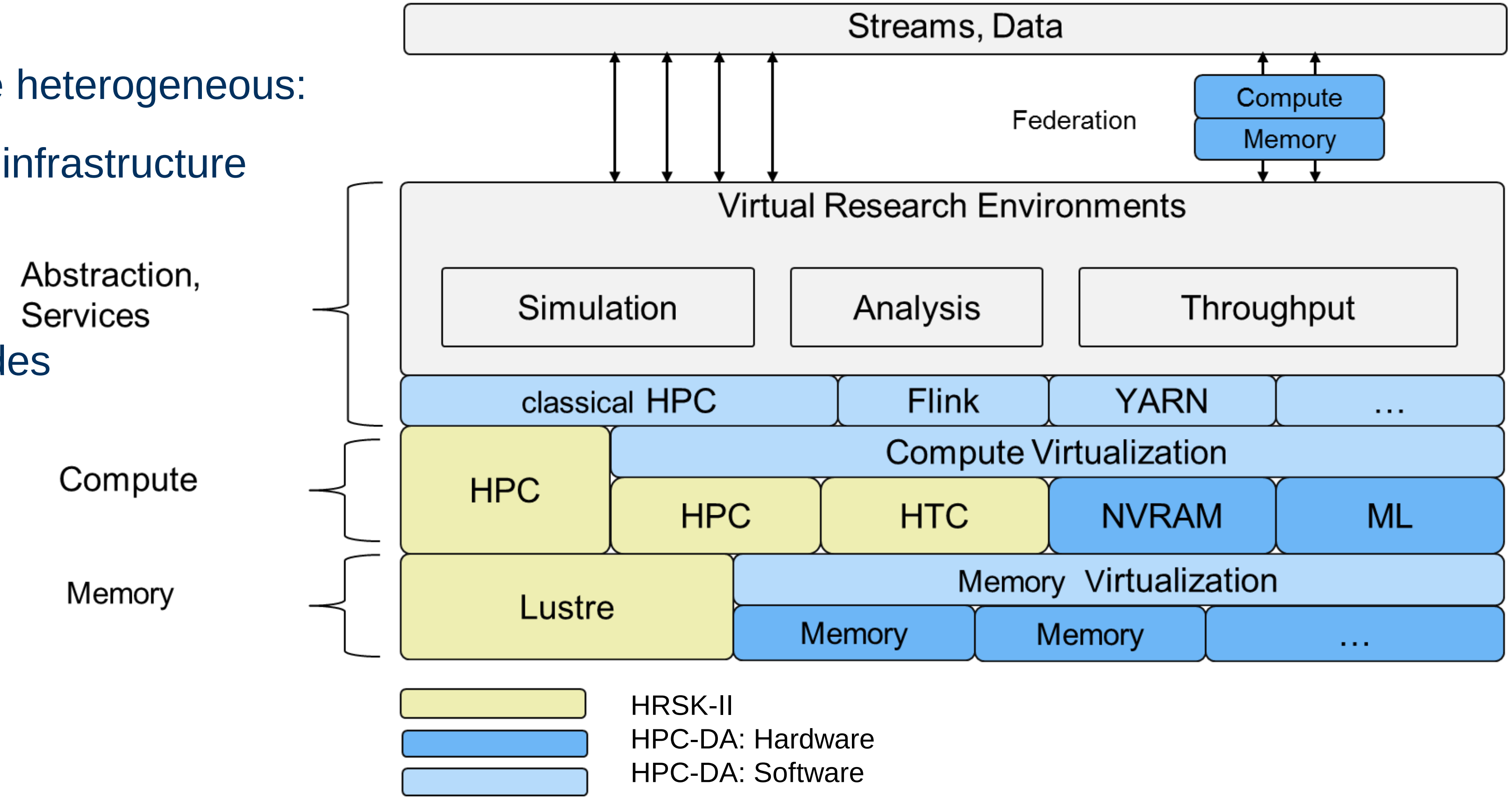


Extension of TUD HRSK-II for HPC Data Analytics (HPC-DA)

HPC systems tend to be more heterogeneous:

Extension of our current HPC infrastructure with new components

- NVMe island
- Power 9/NVIDIA Volta nodes
- Warm archive



Summary

Big Data and HPC have developed independently in the past

- Scale-up vs. Scale-out
- Big Data methods might be helpful for managing/transforming data and fast access to analytics functionality
- Not able to compete specific and tailored implementations
- Big Data concepts could transform HPC architectures: virtualization/containerization and flexible analytics software stacks needs to be involved
- From the application perspective: bring people together! Domain scientists + methods experts

Thanks!

Acknowledgements:
Ralph Müller-Pfefferkorn, Jan Frenzel, Peter Winkler, Nico Hoffmann, Richard Grunzke, Dima Schlesinger, Michael Kluge, Sunna Torge, Bernd Schuller

References:

[ABR16] Alexandre Abraham, Michael P. Milham, Adriana Di Martino, R. Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage, 147:736 – 745, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2016.10.045. URL http://www.sciencedirect.com/science/article/pii/S1053811916305924.

[ALEX14] A. Alexandrov, R. Bergmann, S. Ewen, J. C. Freytag, F. Hueske, A. Heise, O Kao, M. Leich, U. Leser, V. Markl and e. al, “The Stratosphere platform for big data analytics," , VLDB J. vol. 23, no. 6, 2014.

[APA18] https://projects.apache.org/projects.html?category

[BDEC17] BDEC Pathways to Convergence Report “Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry” https://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/bdec_pathways.pdf

[BDEC18] Big Data and Extreme scale Computing, http://www.exascale.org/bdec

[CARB15] P. Carbone, A. Katsifodimos, A. Ewen, V. Markl and e. al:, "Apache Flink™: Stream and Batch Processing in a Single Engine.," IEEE Data Eng. Bull., vol. 38, no. 4, pp. 28-38, 2015.

[Dur15] http://www.dursi.ca/hpc-is-dying-and-mpi-is-killing-it

[EU18] https://data.europa.eu/euodp/en/home

References:

[FOX] Geoffrey Fox: Lecture: Building a Library at the Nexus of High Performance Computing and Big Data, Indiana University

[FISH12] Fisher CIO Leadership Program: “Big Data Analytics: Making Big Data Work”, Bill Ruh, Vice President, Global Software Center, GE, November 1, 2012; https://businessinnovation.berkeley.edu/wp-content/uploads/2018/03/fisher-bigdata-presentations.pdf

[FU16] Fu, Jiangchuan Liu, Xiaowen Chu, and Yueming Hu. Toward a standard interface for cloud providers: The container as the narrow waist. IEEE Internet Computing, 20(2):66–71, 2016

[HOF17] N. Hoffmann et al. Learning Thermal Process Representations for Intraoperative Analysis of Cortical Perfusion During Ischemic Strokes. Springer International Publishing, 2016.

[KLEI18] Kleiner Perkins Internet Trends Report: https://www.kleinerperkins.com/perspectives/internet-trends-report-2018

[REU18] A. Reuther et al. J. Parallel Distrib. Comput. 111 (2018) 76–92

[SAPU] Supun Kamburugamuve, Pulasthi Wickramasinghe, Saliya Ekanayake, and Geoffrey C Fox, Anatomy of machine learning algorithm implementations in MPI, Spark, and Flink , The International Journal of High Performance Computing Applications, Vol 32, Issue 1, pp. 61 – 73

[TOP500] https://www.top500.org/news/summit-up-and-running-at-oak-ridge-claims-first-exascale-application/

[US18] U.S. Government’s open data: http://data.gov