HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

In-Memory Technologie - Treiber spannender Big-Data-Innovationen

Prof. Dr. Christoph Meinel
Hasso Plattner Institute
at the University Potsdam, Germany

Agenda

- **Hasso Plattner Institute**
- In-Memory Technology
- Application in Personalized Medicine
- Application in Security Analytics
- Application in Social Media Analysis

Fact Sheet: Hasso Plattner Institute

... for IT Systems Engineering

Top-Ranking among German speaking computer Sience departments

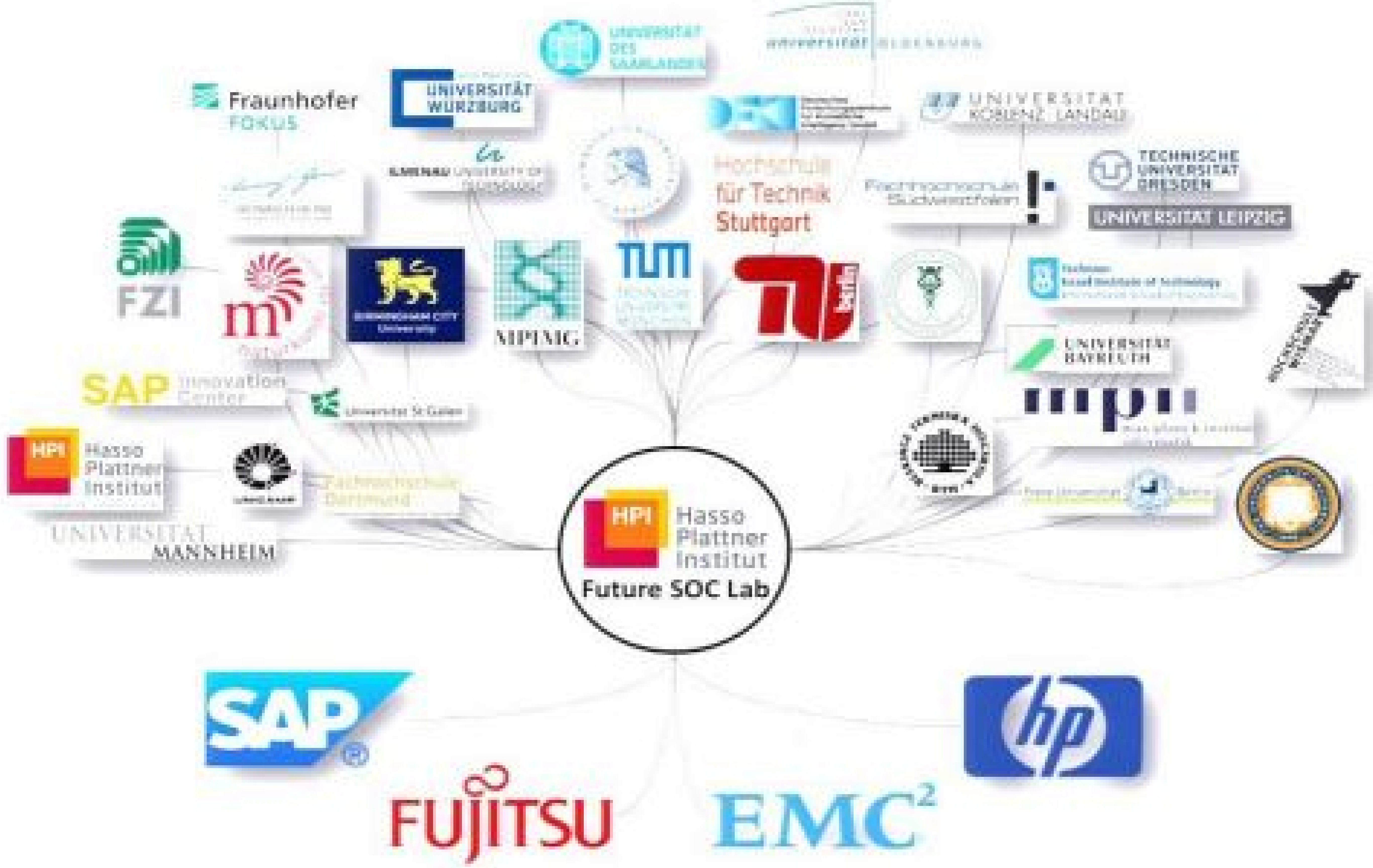
- 10 Professors and departments
- 150 Lecturers, assistant professors, and research assistants
- 120 PhD candidates, 100 internal, 20 external
- 450 Bachelor and master students in IT Systems Engineering
- 160 Students at the HPI D-School
- HPI-Standford Design Thinking Research Program, Future SOC Lab, MOOC platform openHPI and tele-TASK lecture portal, ...

HPI Future SOC Lab – Research in Multicore and In-Memory

Processing Big Data…

- Databases
- Data analysis
- Data management
- Simulation
- Hardware

Future SOC Lab – Vision



HPI Future SOC Lab – Equipment

Highlights

- 1000 Core Cluster with 25TB RAM and 75TB Solid-State Disk
- Hewlett-Packard Converged Cloud
- SAP's In-Memory Computing Appliance HANA and Suite on HANA
- Server with up to 2TB RAM and up to 64 Cores
- State-of-the-art EMC² Storage Systems

Systems

- Fujitsu RX600 S5, RX900 S1, 32 & 64 Cores, 1024 GB RAM
- Hewlett-Packard DL980 G7, 64 Cores, 2048 GB RAM
- EMC² Celerra NS-960 & VNX 5700, 130 TB HDD, 6 TB SSD
- Nvidia Tesla systems with 1792 GPU-Cores
- ...

Agenda

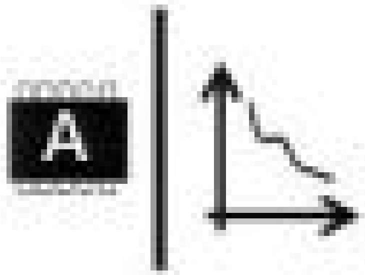
- Hasso Plattner Institute
- **In-Memory Technology**
- Application in Personalized Medicine
- Application in Security Analytics
- Application in Social Media Analysis

Important IT-Innovation from HPI: In-Memory Technology

Important IT-Innovation from HPI: In-Memory Technology

Recent Advances in Hardware

- Multi-core Architectures, e.g. 4 CPUs x 10 Cores on Each Node
- Scaling Across Servers, e.g. 100 Nodes x 40 Cores



- 64 bit Address Space – 4TB in Current Servers
- 25GB/s Data Throughput
- Costs per Enterprise Class Server Node (40 Cores) approx. 29,000 USD

Important IT-Innovation from HPI: In-Memory Technology

Hasso Plattner Institut

Recent Advances in Hardware

- Multi-core Architectures, e.g. 4 CPUs x 10 Cores on Each Node
- Scaling Across Servers, e.g. 100 Nodes x 40 Cores

- 64 bit Address Space – 4TB in Current Servers
- 25GB/s Data Throughput
- Costs per Enterprise Class Server Node (40 Cores) approx. 29,000 USD

Recent Advances in Software

Text Retrieval and Extraction

Insert Only

Compression

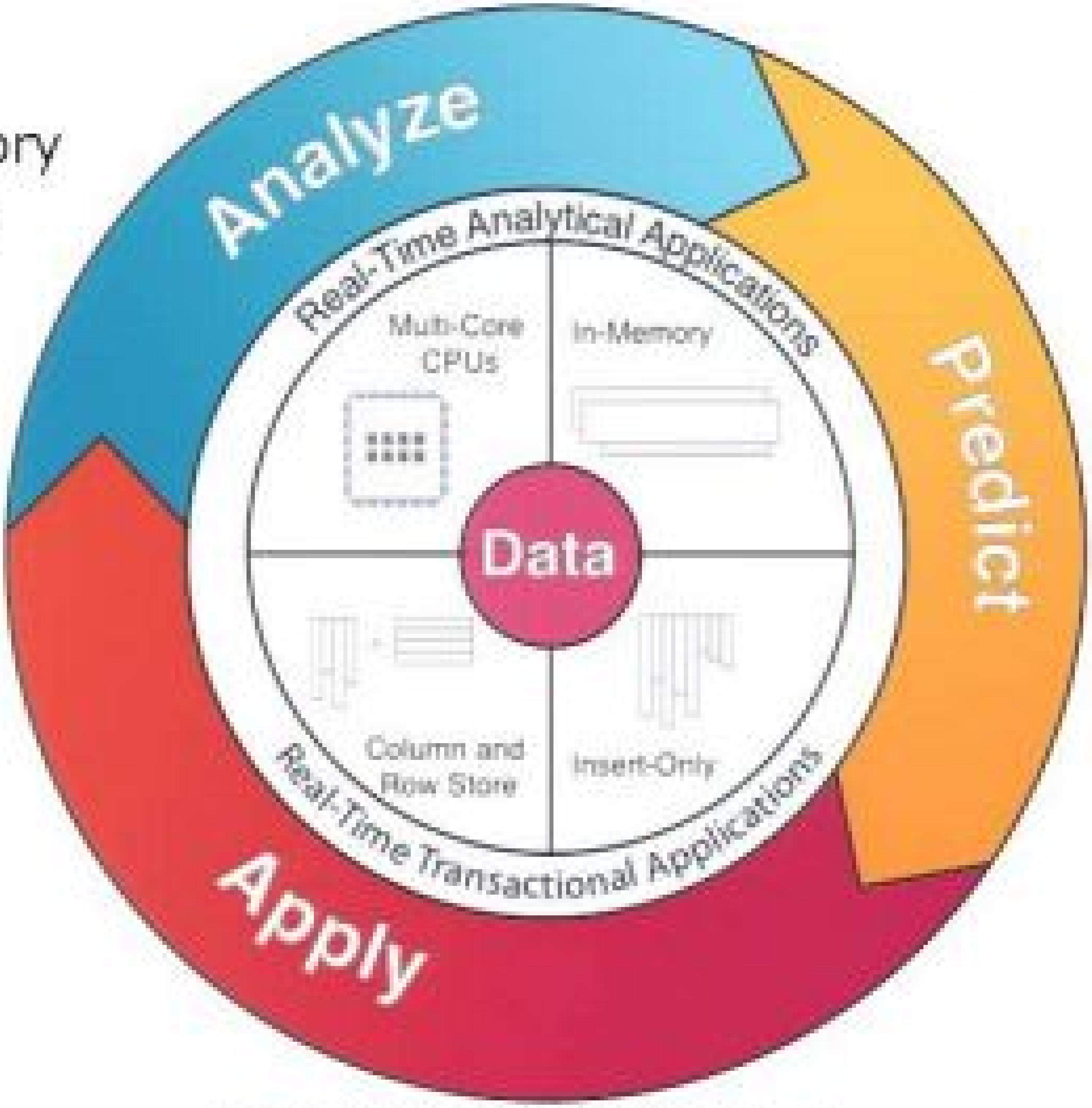
Partitioning

Multi-Core Parallelization

Dynamic Multithreading

Important IT-Innovation from HPI: In-Memory Data Management

- Data-centric architecture: In-Memory database serves as **single source** of truth for all relevant data
- Architecture based on 4 distinct pillars
  - Multi-Core computing
  - In-Memory
  - Column and Row Store
  - Insert-Only
- Enables informed management decisions based on up-to-the-moment data through real-time combination of



Enterprise Performance In-Memory Circle (EPIC)

Breakthrough is Based on Strong Progress in Academic Research …

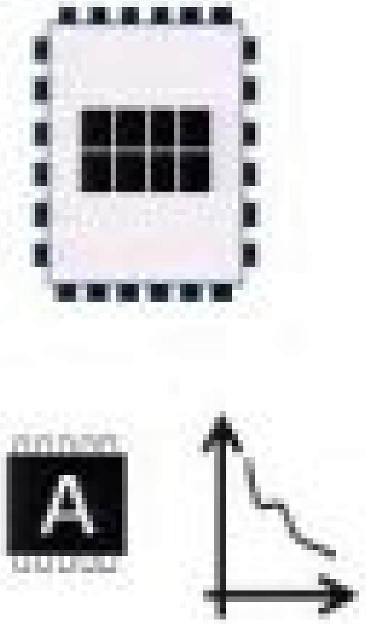
… during the recent years in software for processing data, e.g.

- Column-oriented data organization (the column-store)
  - **Sequential** scans allow best bandwidth utilization between CPU cores and memory
  - **Independence** of tuples within columns allows easy partitioning and therefore parallel processing
- Lightweight Compression
  - Reducing data amount, while..
  - Increasing processing speed through late materialization
- And more, e.g., parallel scan / join / aggregation

Breakthrough is Based on Strong Progress in Hardware ...

... assumptions from yesterday are no more true, new perspectives are possible

- Multi-Core Architecture (96 cores per server)
- One blade ~$50.000 = 1 Enterprise Class Server
- Parallel scaling across blades



- 64 bit address space
- 2TB in current servers
- 25GB/s per core
- Cost-performance ratio rapidly declining
- Memory hierarchies

- Main memory becomes **cheaper and larger**

Breakthrough is Based on Strong Advances in Data Management …

… e.g. Combination of Column and Row Store

- Row stores are designed for operative workload, e.g.
  - Create initial data, e.g. in medical application during first visit, such as name, home address, first contact, etc.
- Column stores are designed for analytical work, e.g.
  - E.g. evaluate the number of patients with the same diagnosis
  - Calculate Kaplan-Meier estimator
- In-Memory approach: Combination of both stores
  - Increased performance for analytical work
  - Without affecting operative performance significantly

Breakthrough is Based on Strong Advances in Data Management …

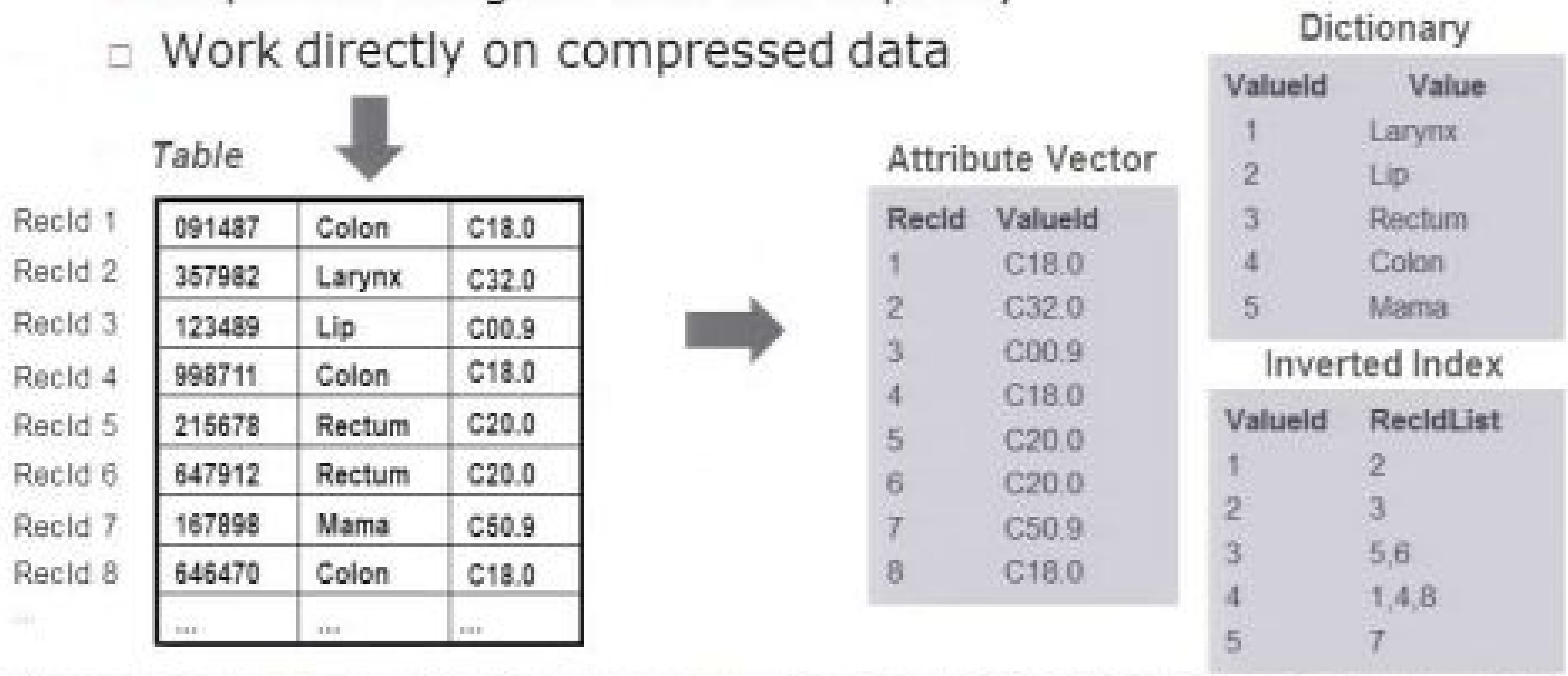
… e.g. Insert-only

- Traditional databases allow four data operations:
  - **INSERT, SELECT,** DELETE, UPDATE
- Last two are destructive since original data is no longer available
- Insert-only requires
  - only first two to store a complete history (bookkeeping systems)
- Insert-only enables travelling through the time, e.g. to
  - To trace changes
  - To document complete history of assessments, therapies, etc.
  - To enable instant Kaplan-Meier estimation on tumor patients

Breakthrough is Based on Strong Advances in Data Management …

… e.g. Lightweighted Compression

- Main memory access is the new bottleneck
- Lightweight compression to reduce bottleneck
  - □ Lossless
  - □ Improved usage of data bus capacity
  - □ Work directly on compressed data



Table

RecId 1	091487	Colon	C18.0
RecId 2	357982	Larynx	C32.0
RecId 3	123489	Lip	C00.9
RecId 4	998711	Colon	C18.0
RecId 5	215678	Rectum	C20.0
RecId 6	647912	Rectum	C20.0
RecId 7	167898	Mama	C50.9
RecId 8	646470	Colon	C18.0
…	…	…	…

Attribute Vector

RecId	ValueId
1	C18.0
2	C32.0
3	C00.9
4	C18.0
5	C20.0
6	C20.0
7	C50.9
8	C18.0

Dictionary

ValueId	Value
1	Larynx
2	Lip
3	Rectum
4	Colon
5	Mama

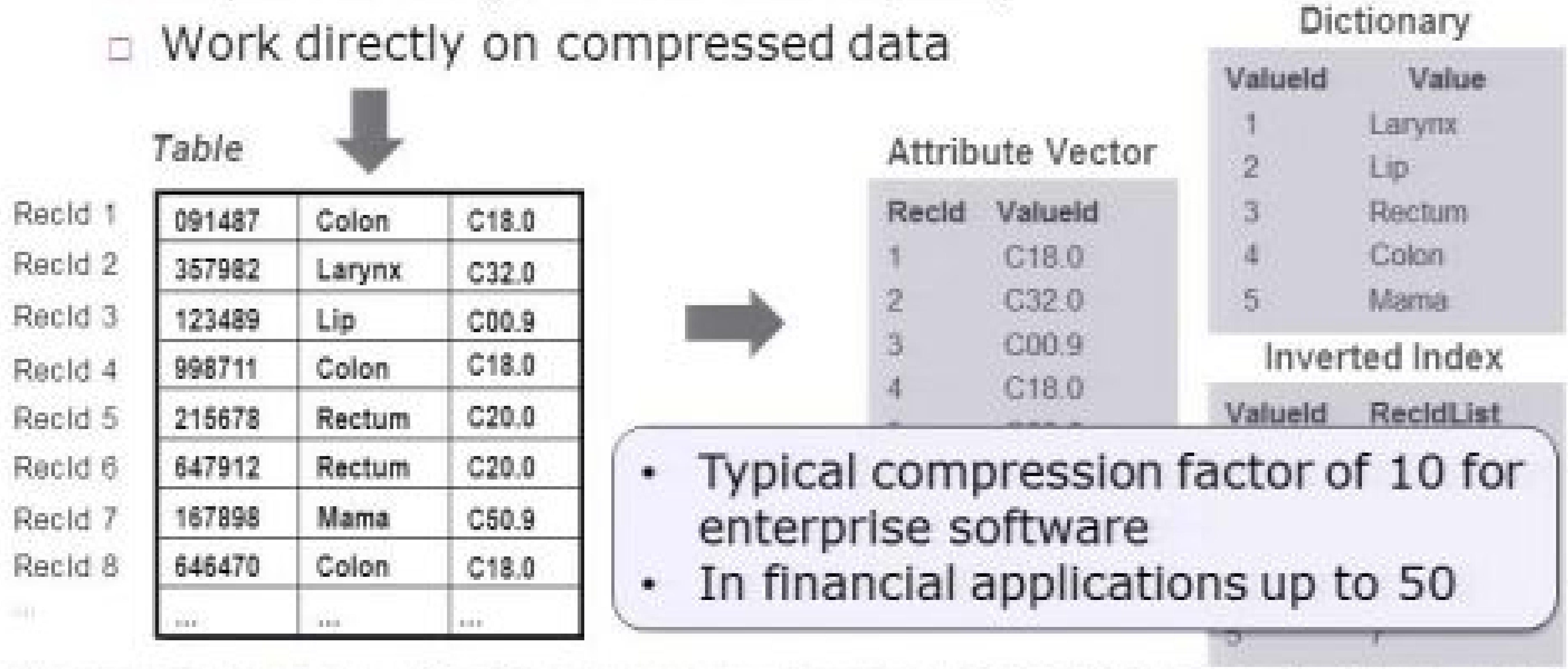
Inverted Index

ValueId	RecIdList
1	2
2	3
3	5,6
4	1,4,8
5	7

Breakthrough is Based on Strong Advances in Data Management …

… e.g. Lightweighted Compression

- Main memory access is the new bottleneck
- Lightweight compression to reduce bottleneck
  - Lossless
  - Improved usage of data bus capacity
  - Work directly on compressed data



Table

RecId 1	091487	Colon	C18.0
RecId 2	357982	Larynx	C32.0
RecId 3	123489	Lip	C00.9
RecId 4	998711	Colon	C18.0
RecId 5	215678	Rectum	C20.0
RecId 6	647912	Rectum	C20.0
RecId 7	167898	Mama	C50.9
RecId 8	646470	Colon	C18.0
…	…	…	…

Attribute Vector

RecId	ValueId
1	C18.0
2	C32.0
3	C00.9
4	C18.0

Dictionary

ValueId	Value
1	Larynx
2	Lip
3	Rectum
4	Colon
5	Mama

Inverted Index

ValueId	RecIdList
5	7

- Typical compression factor of 10 for enterprise software
- In financial applications up to 50

Breakthrough is Based on Strong Advances in Data Management …

… e.g. Partioning

- Horizontal Partitioning
  - □ Cut long tables into shorter segments
  - □ Group patients with same diagnosis
- Vertical Partitioning
  - □ Split off columns to individual resources
  - □ Separate therapy, assessment, diagnosis data
- Partitioning is the basis for
  - □ Parallel execution of database queries
  - □ Data aging
  - □ Data retention management

Breakthrough is Based on Strong Advances in Data Management …

… e.g. Multi-core and Parallelization

- Modern server systems consist of *x* CPUs, e.g. x=4
- Each CPU consists of *y* CPU cores, e.g. y=8
- Consider each of the *x*y* CPU core as individual **worker**
- Each worker can perform one task at the same time in parallel
- Full table scan of database table with 1M entries results in 1/x*1/y search time when traversing in parallel
  - □ reduced response time
  - □ no need for pre-aggregated totals and redundant data
  - □ improved usage of hardware
  - □ instant analysis of data

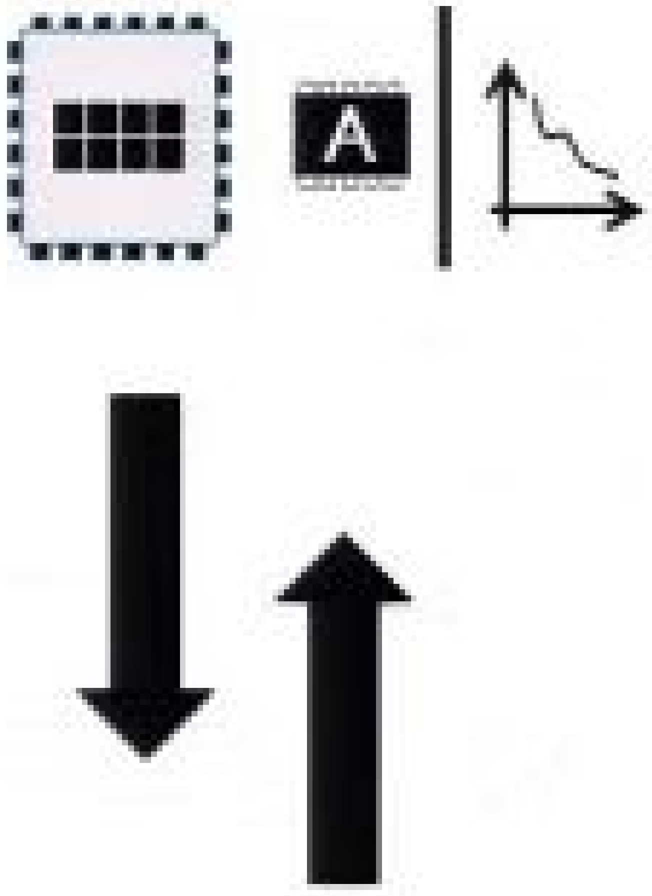
In-Memory Data Management: Summary

Advances in Hardware

Multi-Core Architecture (4 x 8core CPU per blade)

Parallel scaling across blades

One blade ~$50.000 = 1 Enterprise Class Server



64 bit address space – 2TB in current server boards

25GB/s data throughput

Cost-performance ratio rapidly declining

Advances in Software

Row and Column Store

Insert Only

Compression

Partitioning

Parallelization

Active & Passive Data Stores

Agenda

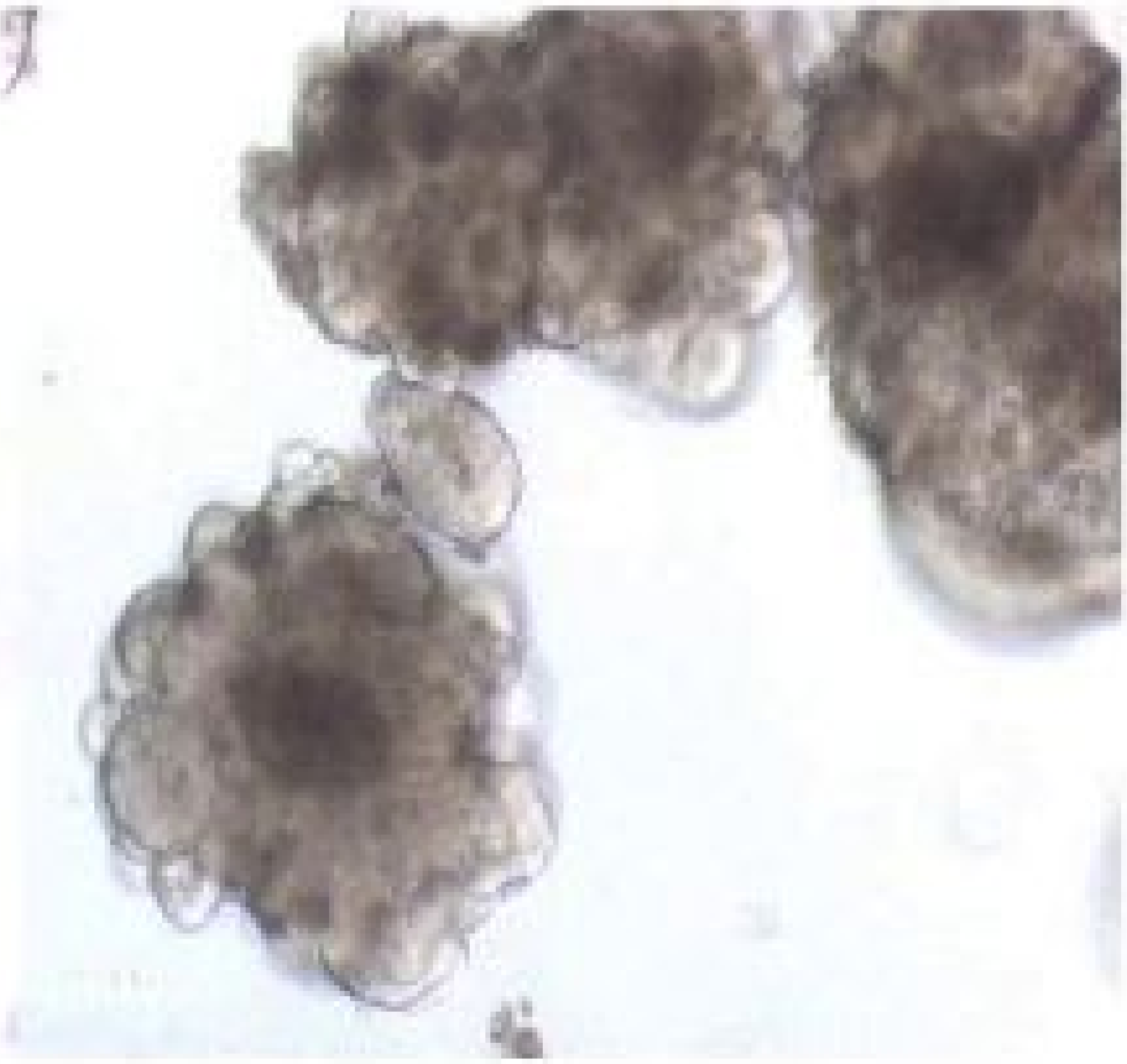
- Hasso Plattner Institute
- In-Memory Technology
- **Application in Personalized Medicine**
- Application in Security Analytics
- Application in Social Media Analysis
- Outlook

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

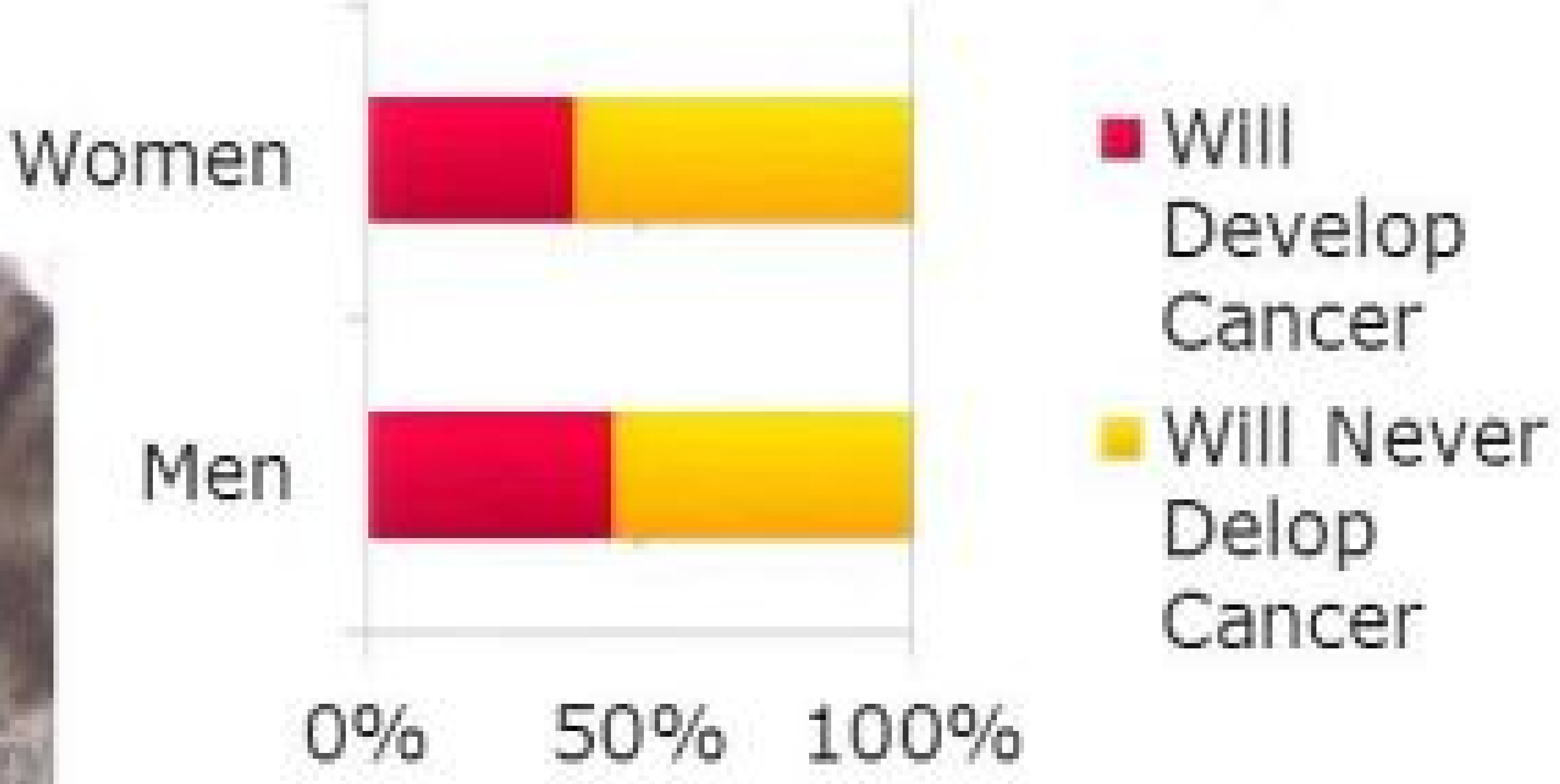
In-Memory Technology – Enabler for Personalized Medicine

Conventional Medicine – Facts one Should Know ...



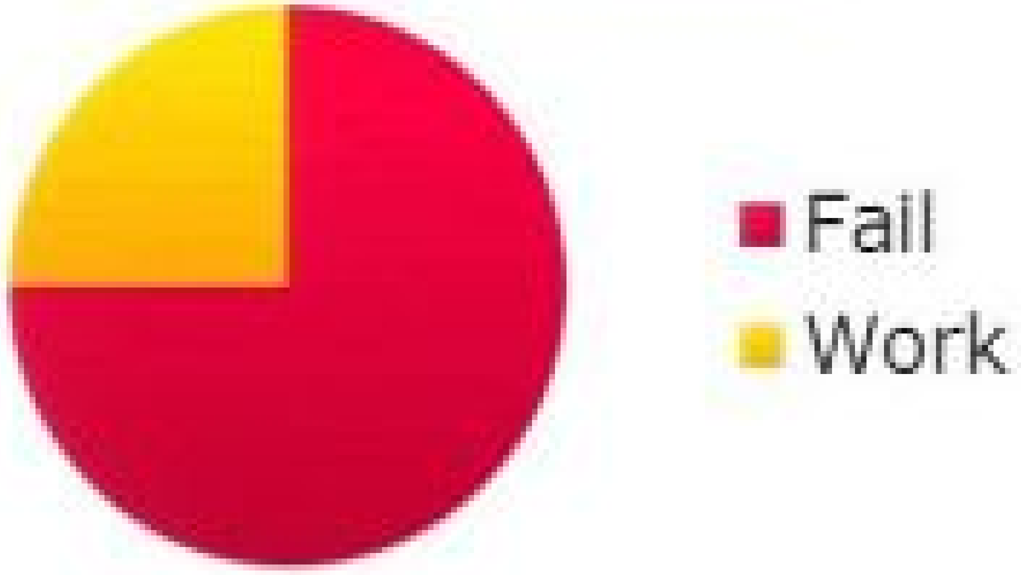
Conventional Medicine – Facts one Should Know …

HPI Hasso Plattner Institut



American Cancer Society, Surveillance Research, 2012

Chemotherapies



Personalized Medicine – Challenges

> "Personalized medicine aims at treating patients specifically based on their individual dispositions, e.g. genetic or environmental factors"

(K. Jain, *Textbook of Personalized Medicine*. Springer, 2009)

Personalized Medicine – Challenges

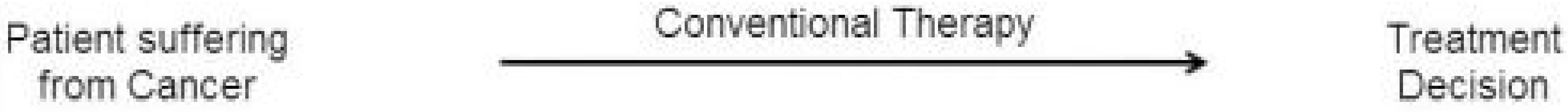
"Personalized medicine aims at treating patients specifically based on their individual dispositions, e.g. genetic or environmental factors"

(K. Jain, *Textbook of Personalized Medicine*. Springer, 2009)

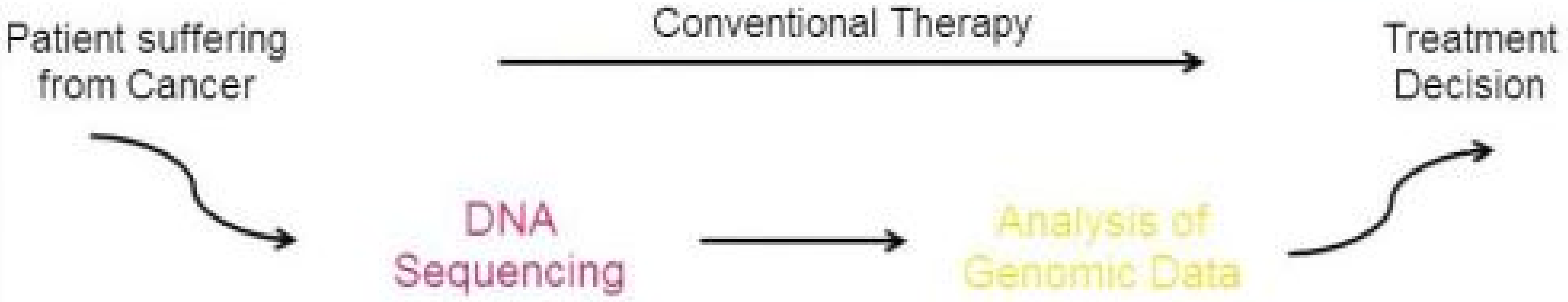
Enhanced by	Limiting Factor
World-wide medical research activities	Research results heterogeneously formatted in distributed databases

NCBI UCSC PubMed.gov OMIM Online Mendelian Inheritance in Man COSMIC KEGG

Personalized Medicine – Using Multicore and In-Memory



Personalized Medicine – Using Multicore and In-Memory



Patient suffering from Cancer

Conventional Therapy

Treatment Decision

DNA Sequencing

Analysis of Genomic Data

Personalized Medicine: Our Motivation

- Today analysis of genome data necessary for personalized treatment takes 4-6 weeks
- Huge data size of human genome:
  3.2 GB * 2 DNA strands * 30 = 192 GB
- A study with only 300 patients ➔ already 57.6 TB of genomic data

Personalized Medicine: Our Motivation

- Today analysis of genome data necessary for personalized treatment takes 4-6 weeks
- Huge data size of human genome:
  3.2 GB * 2 DNA strands * 30 = 192 GB
- A study with only 300 patients → already 57.6 TB of genomic data
- In-memory technology is suitable to accelerate genome analysis
  - □ Fast analysis of large amounts of data
  - □ Pattern recognition
  - □ Combined search in structured and unstructured data

Challenge:

- How to analyze and interpret the entire data of a patient including his genome during a doctor's visit?

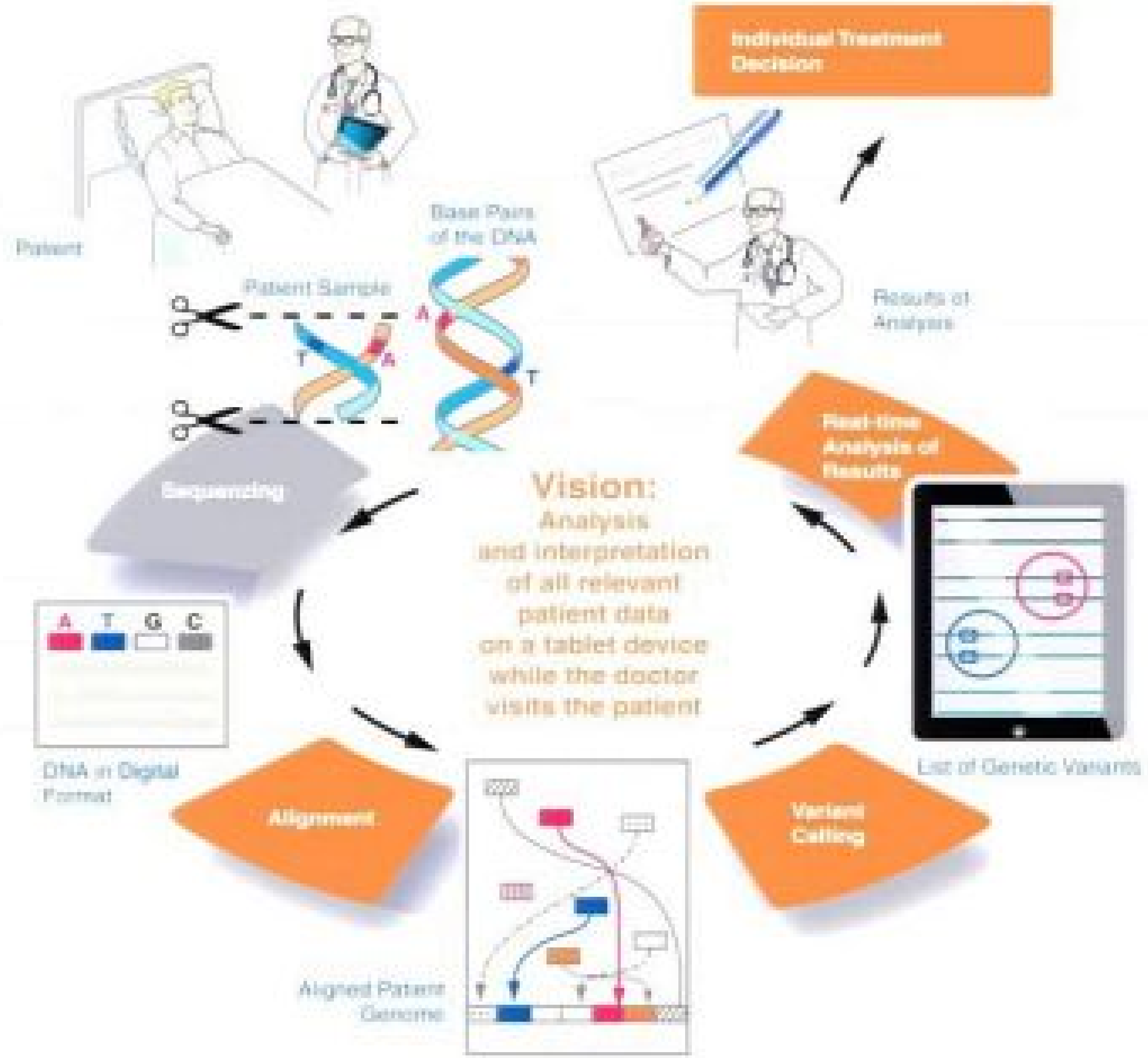
Personalized Medicine: Our Vision

To provide a combined IT-platform for clinician, physicians, researchers and patients that

- Includes all the data from latest research for sophisticated analyses
- Delivers in comprehensive information in real-time about
  - potential sources of a disease
  - cures for a disease
  - related cases
  - relevant literature and annotations
  - clinical trials and proprietary knowledge

HPI High-Performance In-Memory Genome Project
Integration of Genomic Data

HPI Hasso Plattner Institut



Individual Treatment Decision

Sequencing

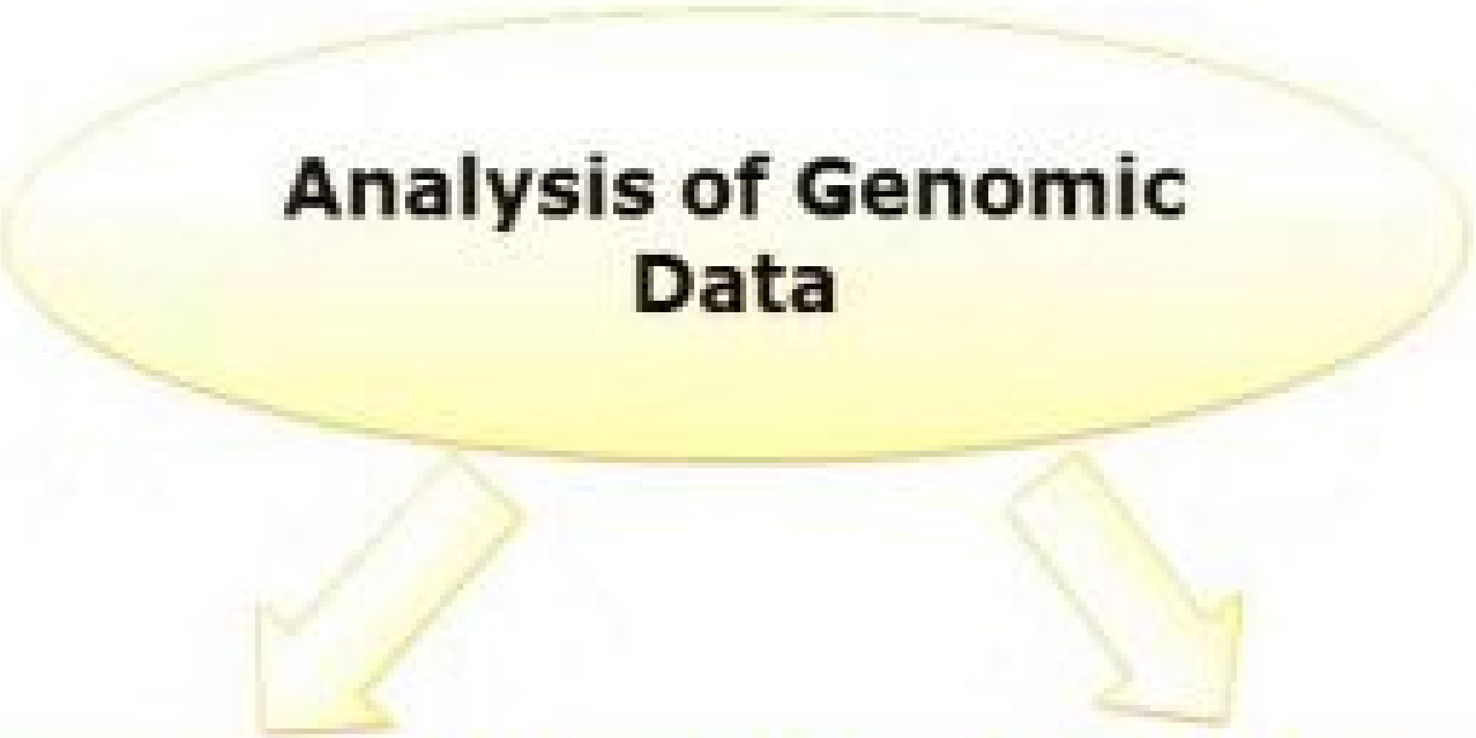
Alignment

Variant Calling

Real-time Analysis of Results

Vision: Analysis and interpretation of all relevant patient data on a tablet device while the doctor visits the patient

Challenges of Genome Data Analysis

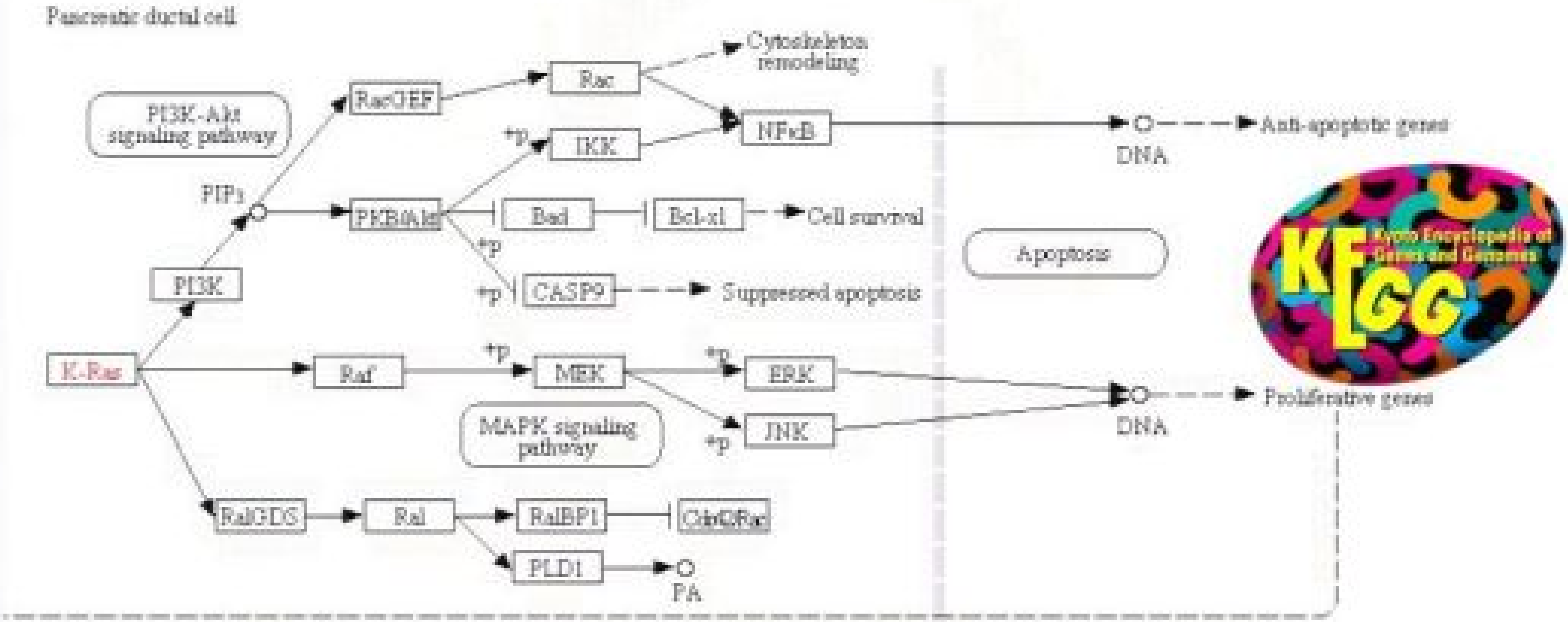


Analysis of Genomic Data

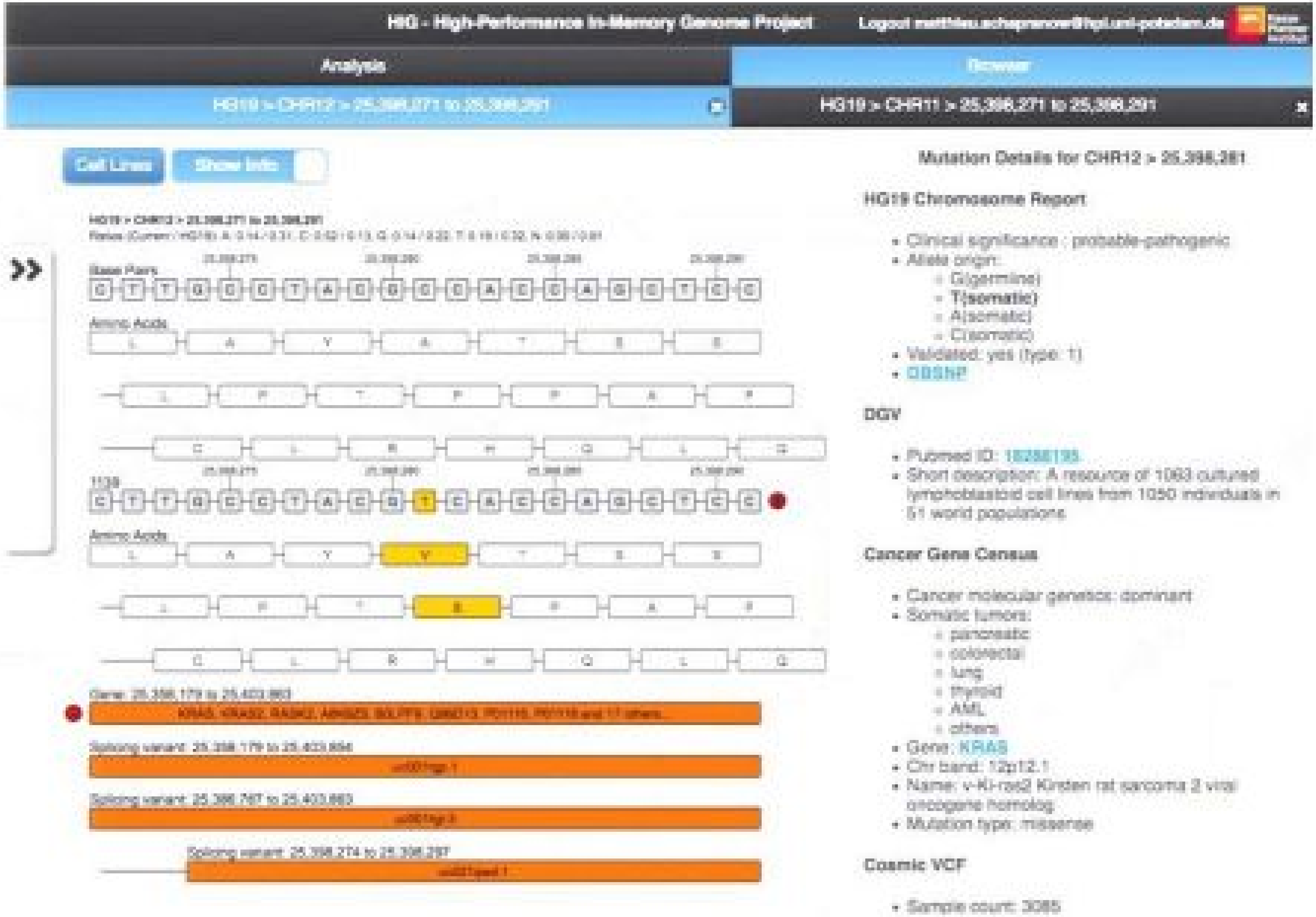
	Alignment and Variant Calling	Analysis of Annotations in World-wide DBs
Bound To	CPU Performance	Memory Capacity
Duration	Hours – Days	Weeks
HPI	Minutes	Real-time
In-Memory Technology	Multi-Core	Partitioning & Compression

HPI High-Performance In-Memory Genome Project
Integration of Genetic Pathways

- Integration of research findings about cause-and-effect relationships in genetic networks into genome analysis
- Use in-memory technology to persist and search in genetic pathways in real-time for causes/effects of certain mutations



HPI High-Performance In-Memory Genome Project
Real-time Analysis of Genome Data



HIG - High-Performance In-Memory Genome Project | Logout matthieu.schapranow@hpi.uni-potsdam.de

Analysis | Browser

HG19 > CHR12 > 25,398,271 to 25,398,291 | HG19 > CHR11 > 25,398,271 to 25,398,291

Cell Lines
Show Info
HG19 > CHR12 > 25,398,271 to 25,398,291
Ratios (Current / HG19): A: 0.14 / 0.31, C: 0.52 / 0.13, G: 0.14 / 0.22, T: 0.19 / 0.32, N: 0.00 / 0.01
25,398,271
25,398,280
25,398,285
25,398,290
Base Pairs
C T T G C C T A C G C C A C C A G C T C C
Amino Acids
L A Y A T S S
L P T P P A P
C L R H Q L Q
1139
C T T G C C T A C G T C A C C A G C T C C
Amino Acids
L A Y V T S S
L P T S P A P
C L R H Q L Q
Gene: 25,358,179 to 25,403,863
KRAS, KRAS2, RASK2, ARRSD5, SOLPF8, Q8BD13, P01116, P01118 and 17 others...
Splicing variant: 25,358,179 to 25,403,854
uc001rgp.1
Splicing variant: 25,386,767 to 25,403,863
uc001rgq.3
Splicing variant: 25,398,274 to 25,398,297
uc021qwd.1

Mutation Details for CHR12 > 25,398,281

HG19 Chromosome Report

- Clinical significance : probable-pathogenic
- Allele origin:
  - G(germline)
  - **T(somatic)**
  - A(somatic)
  - C(somatic)
- Validated: yes (type: 1)
- DBSNP

DGV

- Pubmed ID: 18288195
- Short description: A resource of 1063 cultured lymphoblastoid cell lines from 1050 individuals in 51 world populations

Cancer Gene Census

- Cancer molecular genetics: dominant
- Somatic tumors:
  - pancreatic
  - colorectal
  - lung
  - thyroid
  - AML
  - others
- Gene: KRAS
- Chr band: 12p12.1
- Name: v-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog
- Mutation type: missense

Cosmic VCF

- Sample count: 3085

HPIHigh-Performance In-Memory Genome Project
Combination of Latest Research Results

HPIHigh-Performance In-Memory Genome Project
Combination of Latest Research Results

Genome Browser

- Comparison of multiple mapped genomes with reference
- Exploration of individual genome locations combined with latest relevant annotations and literature e.g. NCBI, dbSNP, UCSC, Sanger

HPIHigh-Performance In-Memory Genome Project
Combination of Latest Research Results

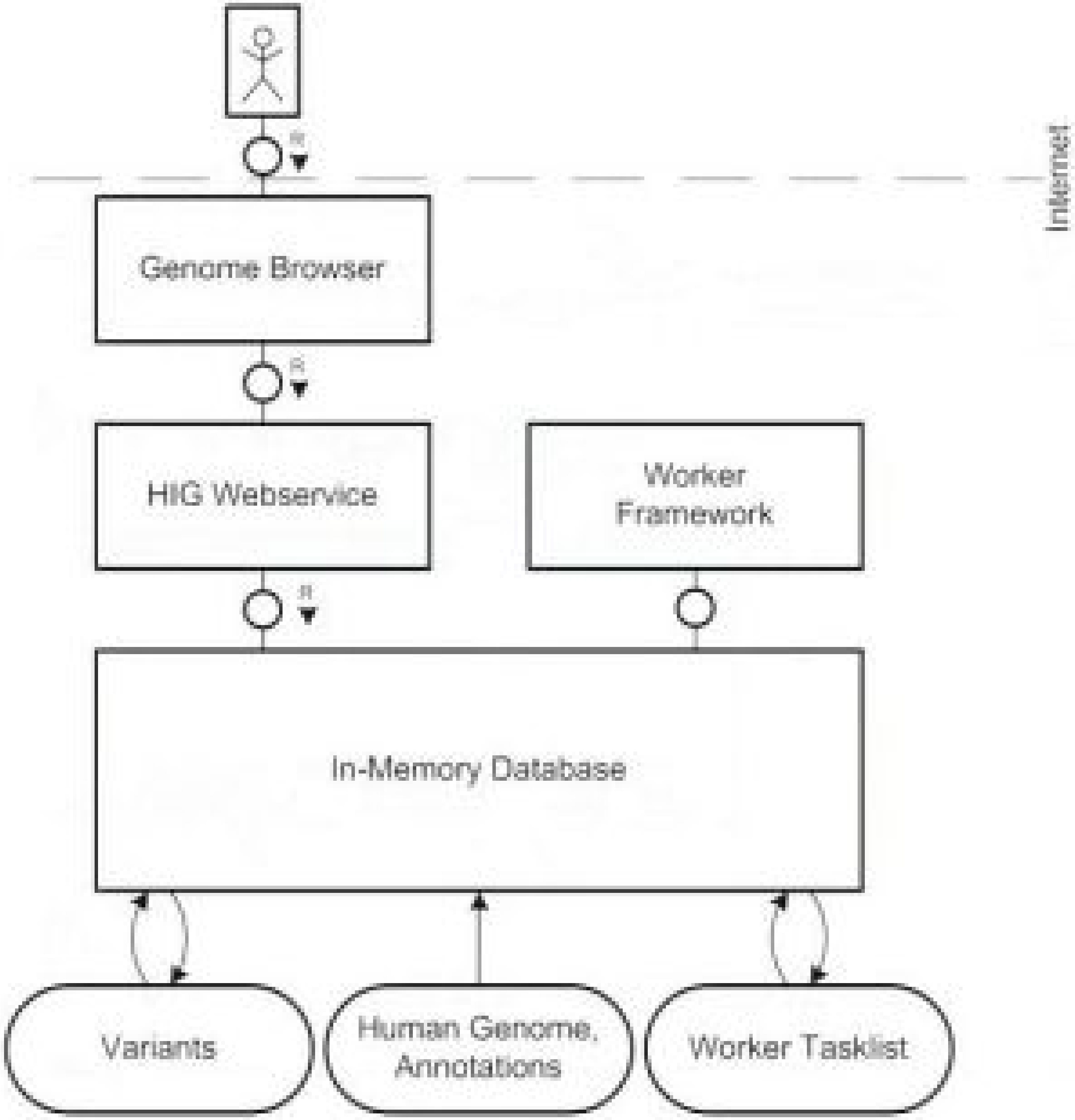
Genome Browser

- Comparison of multiple mapped genomes with reference
- Exploration of individual genome locations combined with latest relevant annotations and literature e.g. NCBI, dbSNP, UCSC, Sanger

Interpretation of Variants

- Variants are sorted, e.g. accordingly to known associated diseases
- All variants are linked to genome browser
- Multiple patients can be compared to identify individual dispositions

High-Performance In-Memory Genome Project
Pilot Installation at HPI Future SOC Lab



1,000 core cluster

- 25 identical nodes
- 40 cores
- 1 TB main memory
- 2.40 GHz, 30 MB Cache

Agenda

- Hasso Plattner Institute
- In-Memory Technology
- Application in Personalized Medicine
- **Application in Security Analytics**
- Application in Social Media Analysis

Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

HPI – Security Analytics Lab and Software Surveillance

→ www.hpi-vdb.de

HPI-SAL – Security Analytics Lab ...

... research in a in-memory security information and event management (**SIEM**)

Goal: Continuously real-time analysis of security sensor data

- complex system information, diverse vulnerabilities, giant range of attacks

Source: Huge multi-types and heterogeneous real-time security sensor data

- Log files (OS/App), scanning reports, IDS Alerts, Virus/Firewall warnings, monitoring logs (e.g., third-party SIEMs, e.g., Splunk, Graylog2, etc.) from different sources, e.g., files, DBs, registries,...

HPI-SAL – Security Analytics Lab ...

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

HPI – Security Analytics Lab and Software Surveillance

→ www.hpi-vdb.de

HPI-SAL – Security Analytics Lab ...

... research in a in-memory security information and event management (**SIEM**)

Goal: Continuously real-time analysis of security sensor data

- complex system information, diverse vulnerabilities, giant range of attacks

Source: Huge multi-types and heterogeneous real-time security sensor data

- Log files (OS/App), scanning reports, IDS Alerts, Virus/Firewall warnings, monitoring logs (e.g., third-party SIEMs, e.g., Splunk, Graylog2, etc.) from different sources, e.g., files, DBs, registries,...

HPI-SAL – Security Analytics Lab …

… research in a in-memory security information and event management (**SIEM**)

Goal: Continuously real-time analysis of security sensor data

- complex system information, diverse vulnerabilities, giant range of attacks

Source: Huge multi-types and heterogeneous real-time security sensor data

- Log files (OS/App), scanning reports, IDS Alerts, Virus/Firewall warnings, monitoring logs (e.g., third-party SIEMs, e.g., Splunk, Graylog2, etc.) from different sources, e.g., files, DBs, registries,…

Continuous live analysis:

- Post-Processing (filtering, compressing, ..)
- Aggregation/Clustering/Correlation
- Visualization
- Correlation of interesting events
- Detection of complex attack scenarios
- Rapid decision and response

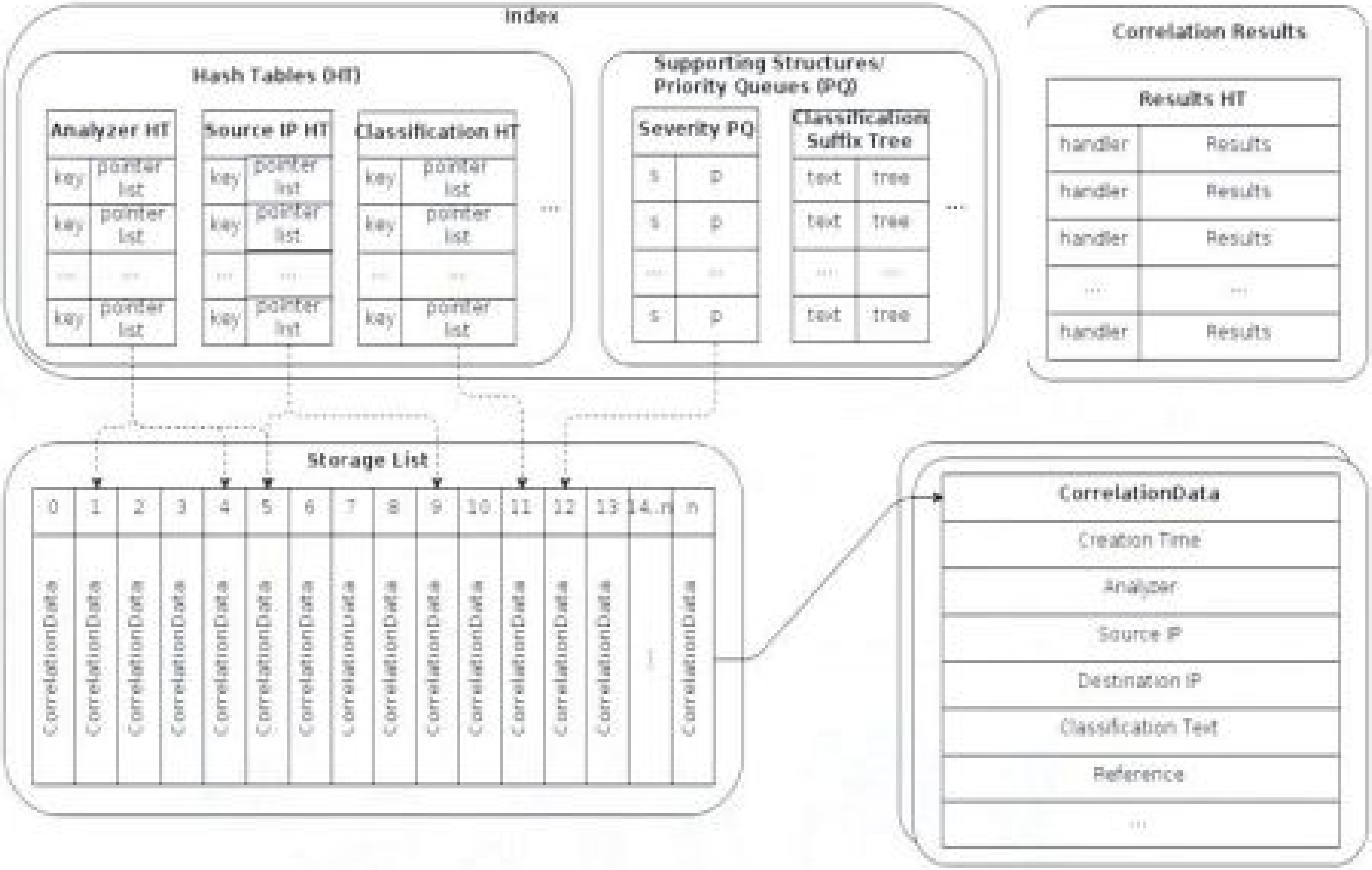
SecurityAnalyticsLab

ID	Analyzer	Classification	Creation Time	Source Address	Target Address	Reference	Message Id
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

HPI-SAL – Security Analytics Lab Feature List

- **Normalization**: Unified representation of real-time Event Information
- **Centralization** using **In-Memory** Data Storage (e.g., **SAP HANA**)
  - Organizing data in index tables
  - Storing results for multi-step correlation
- **Analytics**: **correlation** and **visualization**, for example,
  - Context-based Event correlation: event informatin are analyzed in the context of enviromental Information and known vulnerability information (both are reasoned into the Attack Graph)
- **Multi-core** supported correlation
  - Known parallel programming approaches, e.g., scala, CUDA, Hadoop…
  - Computation-intensive algorithms, e.g., k-means clustering, etc.

SAL – Security Analytics Lab
In-Memory Data Storage



Index

Hash Tables (HT)

Analyzer HT	
key	pointer list
key	pointer list
...	...
key	pointer list

Source IP HT	
key	pointer list
key	pointer list
...	...
key	pointer list

Classification HT	
key	pointer list
key	pointer list
...	...
key	pointer list

...

Supporting Structures/ Priority Queues (PQ)

Severity PQ	
s	p
s	p
...	...
s	p

Classification Suffix Tree	
text	tree
text	tree
...	...
text	tree

...

Correlation Results

Results HT	
handler	Results
handler	Results
handler	Results
...	...
handler	Results

Storage List

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14..n	n
CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	...	CorrelationData

CorrelationData
Creation Time
Analyzer
Source IP
Destination IP
Classification Text
Reference
...

HPI-SAL – Security Analytics Lab Feature List

- **Normalization**: Unified representation of real-time Event Information
- **Centralization** using **In-Memory** Data Storage (e.g., **SAP HANA**)
  - Organizing data in index tables
  - Storing results for multi-step correlation
- **Analytics**: **correlation** and **visualization**, for example,
  - Context-based Event correlation: event informatin are analyzed in the context of enviromental Information and known vulnerability information (both are reasoned into the Attack Graph)
- **Multi-core** supported correlation
  - Known parallel programming approaches, e.g., scala, CUDA, Hadoop…
  - Computation-intensive algorithms, e.g., k-means clustering, etc.

HPI-SAL – Security Analytics Lab ...

... research in a in-memory security information and event management (**SIEM**)

Goal: Continuously real-time analysis of security sensor data

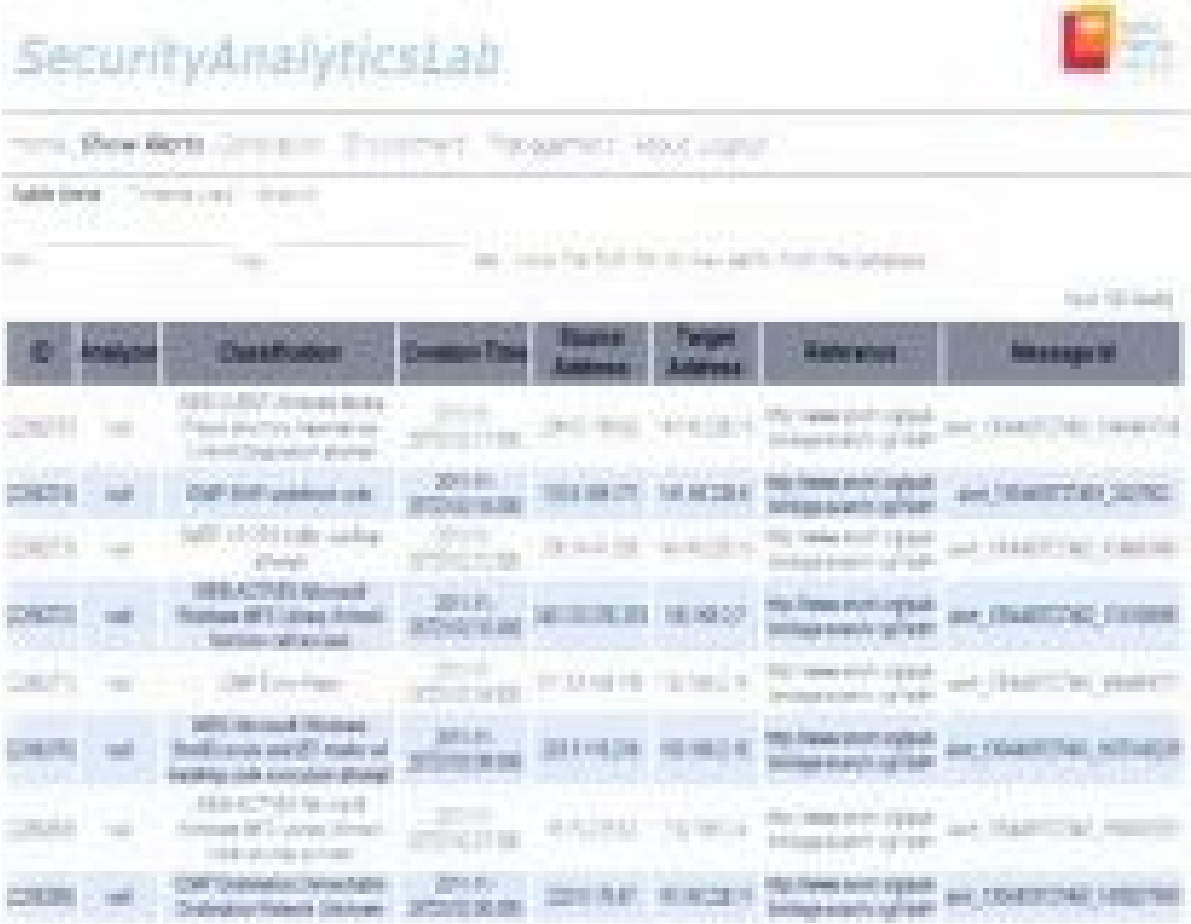
- complex system information, diverse vulnerabilities, giant range of attacks

Source: Huge multi-types and heterogeneous real-time security sensor data

- Log files (OS/App), scanning reports, IDS Alerts, Virus/Firewall warnings, monitoring logs (e.g., third-party SIEMs, e.g., Splunk, Graylog2, etc.) from different sources, e.g., files, DBs, registries,...

Continuous live analysis:

- Post-Processing (filtering, compressing, ..)
- Aggregation/Clustering/Correlation
- Visualization
- Correlation of interesting events
- Detection of complex attack scenarios
- Rapid decision and response



SecurityAnalyticsLab

ID	Analyzer	Classification	Creation Time	Source Address	Target Address	Reference	Message ID
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

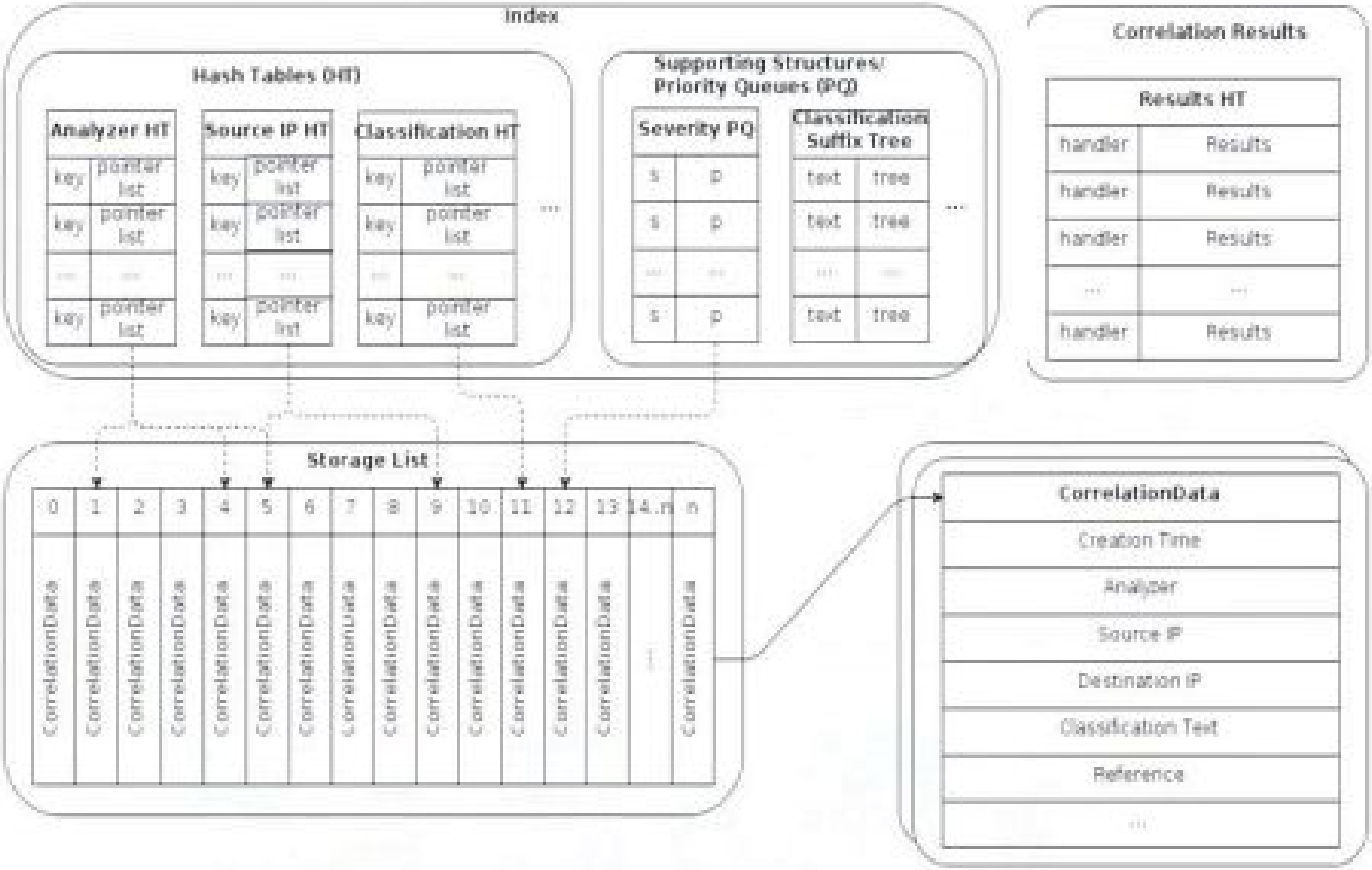
HPI-SAL – Security Analytics Lab Feature List

- **Normalization**: Unified representation of real-time Event Information
- **Centralization** using **In-Memory** Data Storage (e.g., **SAP HANA**)
  - Organizing data in index tables
  - Storing results for multi-step correlation
- **Analytics**: **correlation** and **visualization**, for example,
  - Context-based Event correlation: event informatin are analyzed in the context of enviromental Information and known vulnerability information (both are reasoned into the Attack Graph)
- **Multi-core** supported correlation
  - Known parallel programming approaches, e.g., scala, CUDA, Hadoop…
  - Computation-intensive algorithms, e.g., k-means clustering, etc.

HPI-SAL – Security Analytics Lab Feature List

- **Normalization**: Unified representation of real-time Event Information
- **Centralization** using **In-Memory** Data Storage (e.g., **SAP HANA**)
    - Organizing data in index tables
    - Storing results for multi-step correlation
- **Analytics**: **correlation** and **visualization**, for example,
    - Context-based Event correlation: event informatin are analyzed in the context of enviromental Information and known vulnerability information (both are reasoned into the Attack Graph)
- **Multi-core** supported correlation
    - Known parallel programming approaches, e.g., scala, CUDA, Hadoop…
    - Computation-intensive algorithms, e.g., k-means clustering, etc.

SAL – Security Analytics Lab
In-Memory Data Storage



Analyzer HT	
key	pointer list
key	pointer list
...	...
key	pointer list

Source IP HT	
key	pointer list
key	pointer list
...	...
key	pointer list

Classification HT	
key	pointer list
key	pointer list
...	...
key	pointer list

Severity PQ	
s	p
s	p
...	...
s	p

Classification Suffix Tree	
text	tree
text	tree
...	...
text	tree

Results HT	
handler	Results
handler	Results
handler	Results
...	...
handler	Results

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14..n	n
CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	CorrelationData	...	CorrelationData

CorrelationData
Creation Time
Analyzer
Source IP
Destination IP
Classification Text
Reference
...

HPI-SAL – Security Analytics Lab Performance

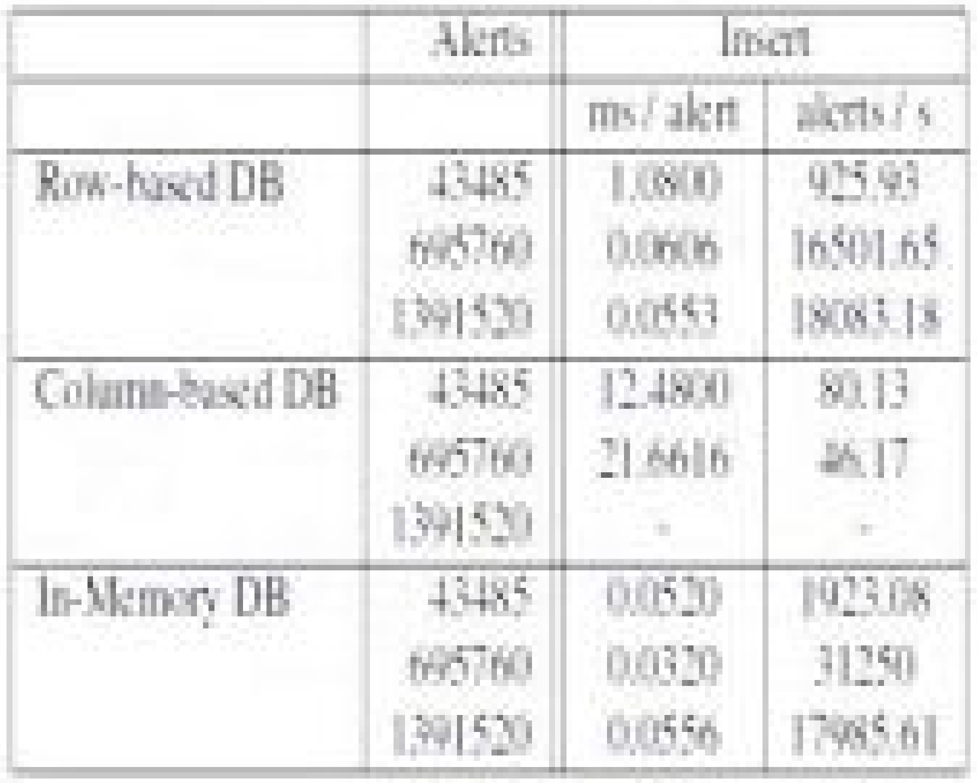
Preliminary Remarks:

- Typical memory allocation: 152 Bytes/Security Item
- According to our tests: 1,391,520 event items can be generated from one host in six months, i.e., 7481 items/day
- Our testing platform is an HP DL980 G7 with **2 TB of main memory**, i.e., 2,199,023,255,552 Bytes, which can hold $1.44672583*10^{10}$ event items
- This means, this mentioned "normal" server can host and process within in-memory an amount of event items generated from 1 hosts in **5298.26** years, or 1000 hosts in about **5 years 3 months**

Security Analytics Lab (SAL): Performance Comparison

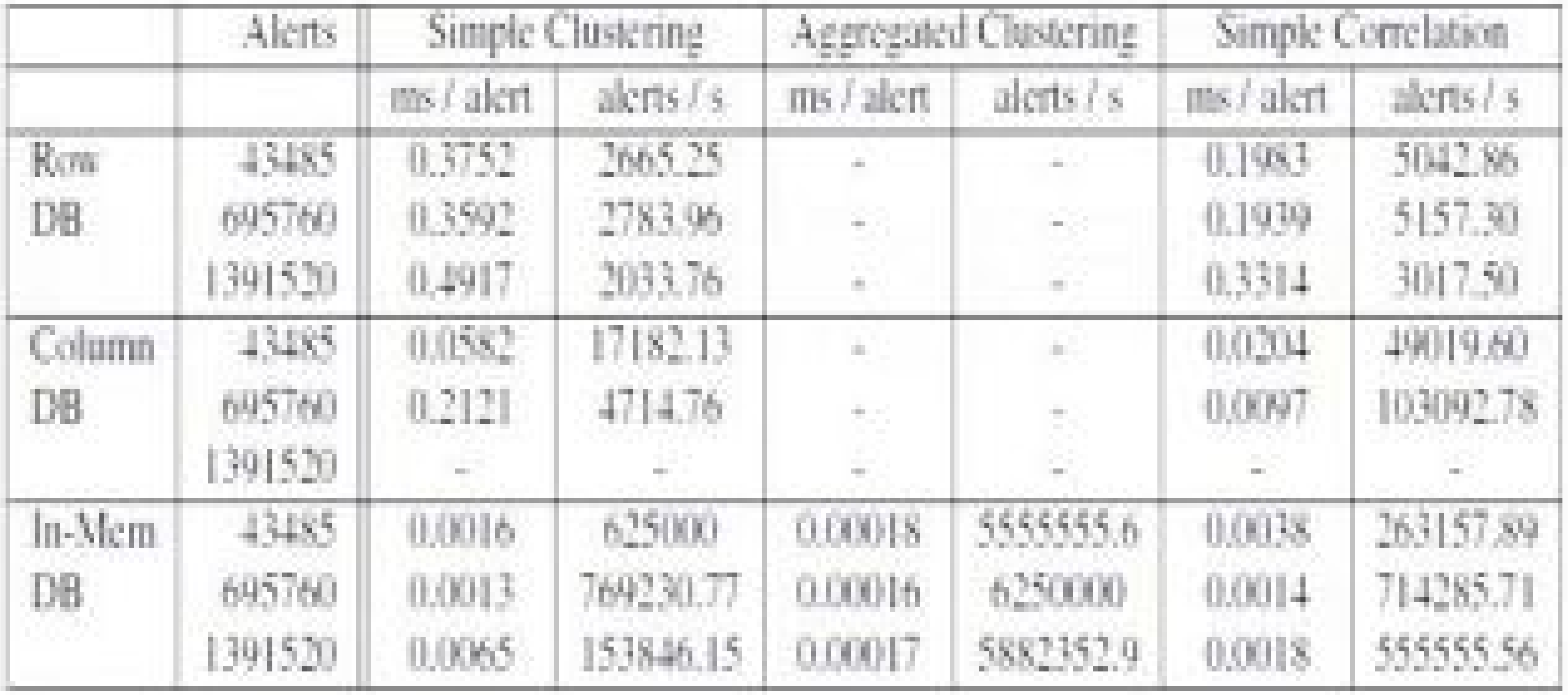
Hasso Plattner Institut

Insert Operation:



	Alerts	Insert	
		ms / alert	alerts / s
Row-based DB	43485	1.0800	925.93
	695760	0.0606	16501.65
	1391520	0.0553	18083.18
Column-based DB	43485	12.4800	80.13
	695760	21.6616	46.17
	1391520	-	-
In-Memory DB	43485	0.0520	1923.08
	695760	0.0320	31250
	1391520	0.0556	17985.61

Some Simple Clustering, Aggregation, and Correlation Algorithms:



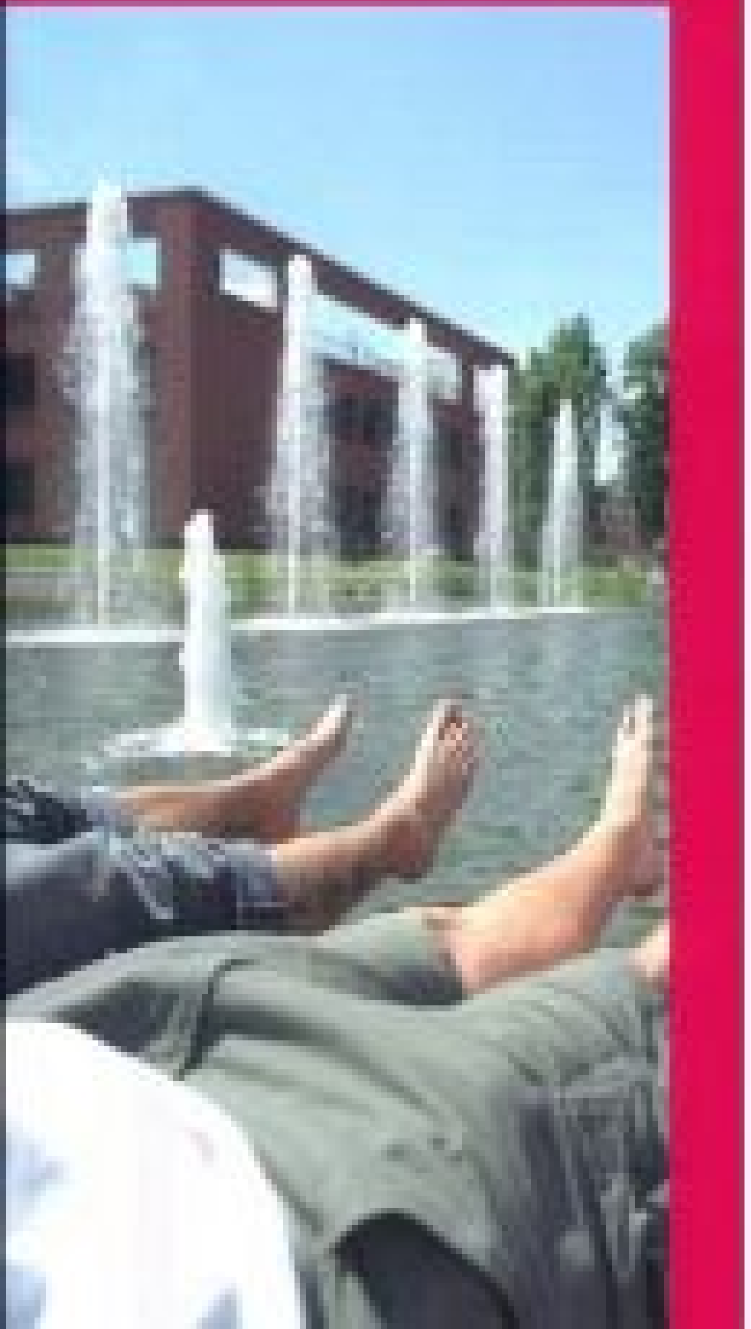
	Alerts	Simple Clustering		Aggregated Clustering		Simple Correlation	
		ms / alert	alerts / s	ms / alert	alerts / s	ms / alert	alerts / s
Row DB	43485	0.3752	2665.25	-	-	0.1983	5042.86
	695760	0.3592	2783.96	-	-	0.1939	5157.30
	1391520	0.4917	2033.76	-	-	0.3314	3017.50
Column DB	43485	0.0582	17182.13	-	-	0.0204	49019.60
	695760	0.2121	4714.76	-	-	0.0097	103092.78
	1391520	-	-	-	-	-	-
In-Mem DB	43485	0.0016	625000	0.00018	5555555.6	0.0038	263157.89
	695760	0.0013	769230.77	0.00016	6250000	0.0014	714285.71
	1391520	0.0065	153846.15	0.00017	5882352.9	0.0018	555555.56

Agenda

- Hasso Plattner Institute
- In-Memory Technology
- Application in Personalized Medicine
- Application in Security Analytics
- **Application in Social Media Analysis**

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam



blogintelligence®

explore the blogosphere.

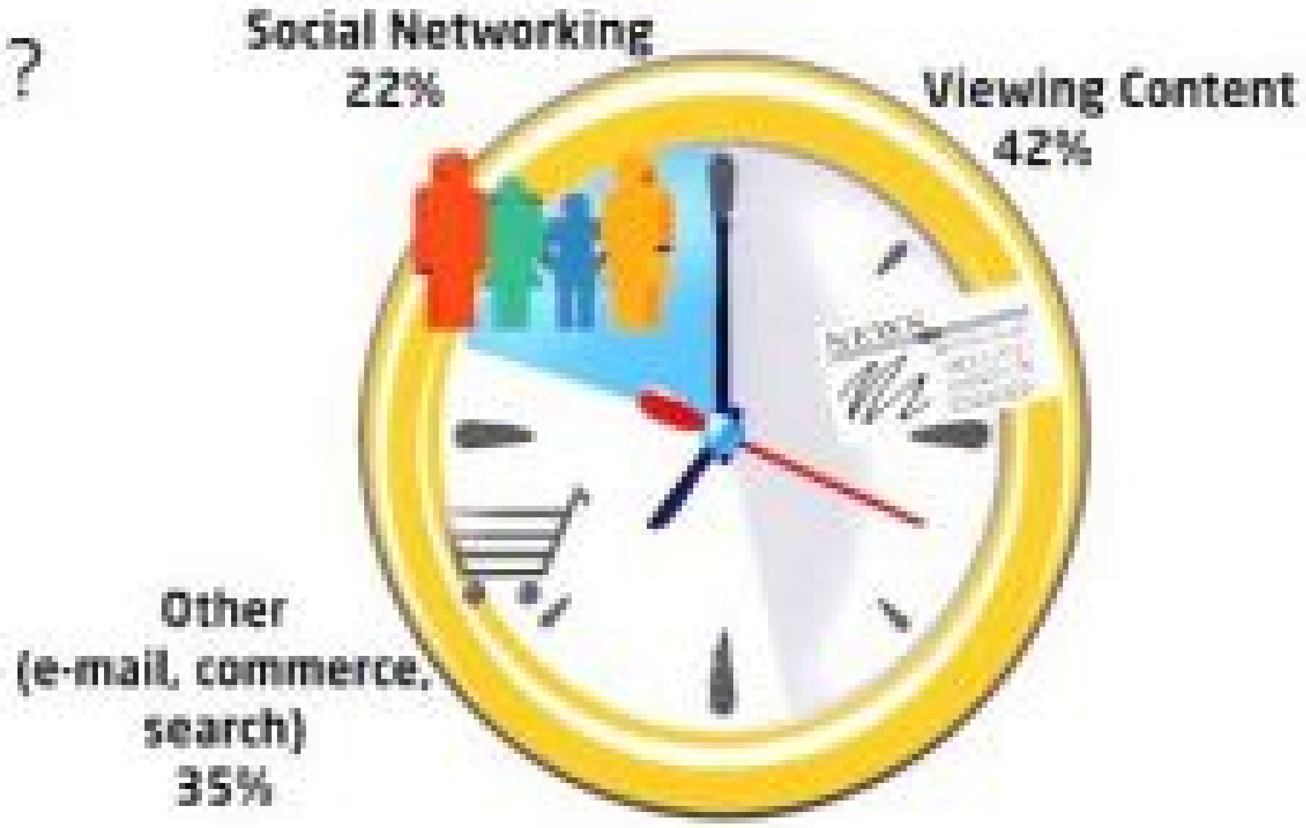
The Blogosphere – Social Media Analysis

Mass Data of Social Media Networks

What happens within anIinternet minute?

- 100 000 tweets[1]
- 300 000 Facebook updates[2]
- 80 000 blog posts[3]

Time Spent on the World Wide Web



Social network postings evolve to a every changing and exponentially growing mass communication network

Tons of Terabytes of structured and unstructured data are produced

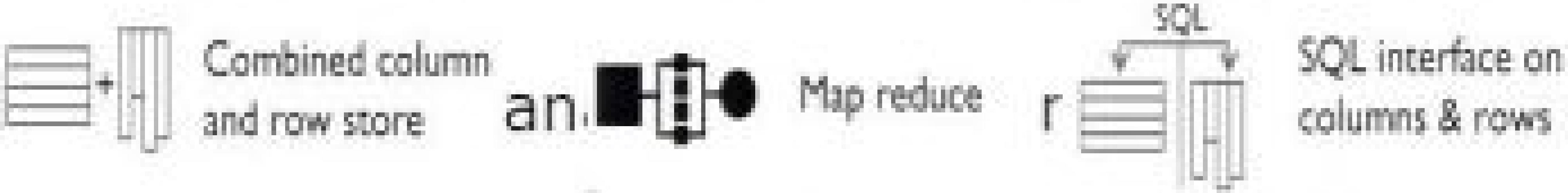
Main Memory as Big Data Storage

Thanks to the continuous development of memory, RAM gets extremely affordable

Our single node database scales up to **6 TB of RAM**

An in-memory database delivers high performance by combining various concepts

SAP HAN an r ıred data


Combined column and row store
Map reduce
SQL
SQL interface on columns & rows

Performance Improvements Compaired to Convential Data Basis

Hasso Plattner Institut

Rank pages based on incoming links:
1000 times faster

- SELECT COUNT(*) as incomingLinks , toHost FROM link GROUP BY toHost ORDER BY COUNT(*) DES

Join link ranking and retrieve blog information:
>18.000 times faster

- SELECT POSTTITLE , POSTAUTHOR FROM WEBPAGE INNER JOIN (SELECT toUrl , COUNT(*) as links FROM link GROUPBY toUrl) as link ON link.toUrl = webpage.ID WHERE type='POST' ORDERBY links desc

Complex selection of new urls to crawl:
60 times faster

- SELECT ID FROM WEBPAGE WHERE FETCHTIME <1358765520000 ORDER BY SCORE DESC LIMIT 10000

Heavy insert performance: still two times faster ...

In-Memory Technology – Big Data Innovation Enabler

Prof. Dr. Christoph Meinel
Hasso-Plattner-Institute
at University of Potsdam
Campus Griebnitzsee
14482 Potsdam, Germany

www.hpi-web.de/meinel