



Solve the Puzzle

Dr. Matthieu-P. Schapranow
Code of Life

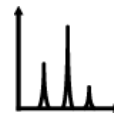
November 2016

IT Challenges

Distributed Heterogeneous Data Sources



Human genome/biological data
600GB per full genome
15PB+ in databases of leading institutes



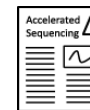
Human proteome
160M data points (2.4GB) per sample
>3TB raw proteome data in ProteomicsDB



Hospital information systems
Often more than 50GB



Cancer patient records
>160k records at NCT



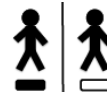
PubMed database
>23M articles



Medical sensor data
Scan of a single organ in 1s
creates 10GB of raw data



Prescription data
1.5B records from 10,000 doctors and
10M Patients (100 GB)



Clinical trials
Currently more than 30k
recruiting on ClinicalTrials.gov

Solve the Puzzle

Schapranow, Nov 2016

Software Requirements in Life Sciences



- Requirements
 - Real-time data analysis
 - Maintained software

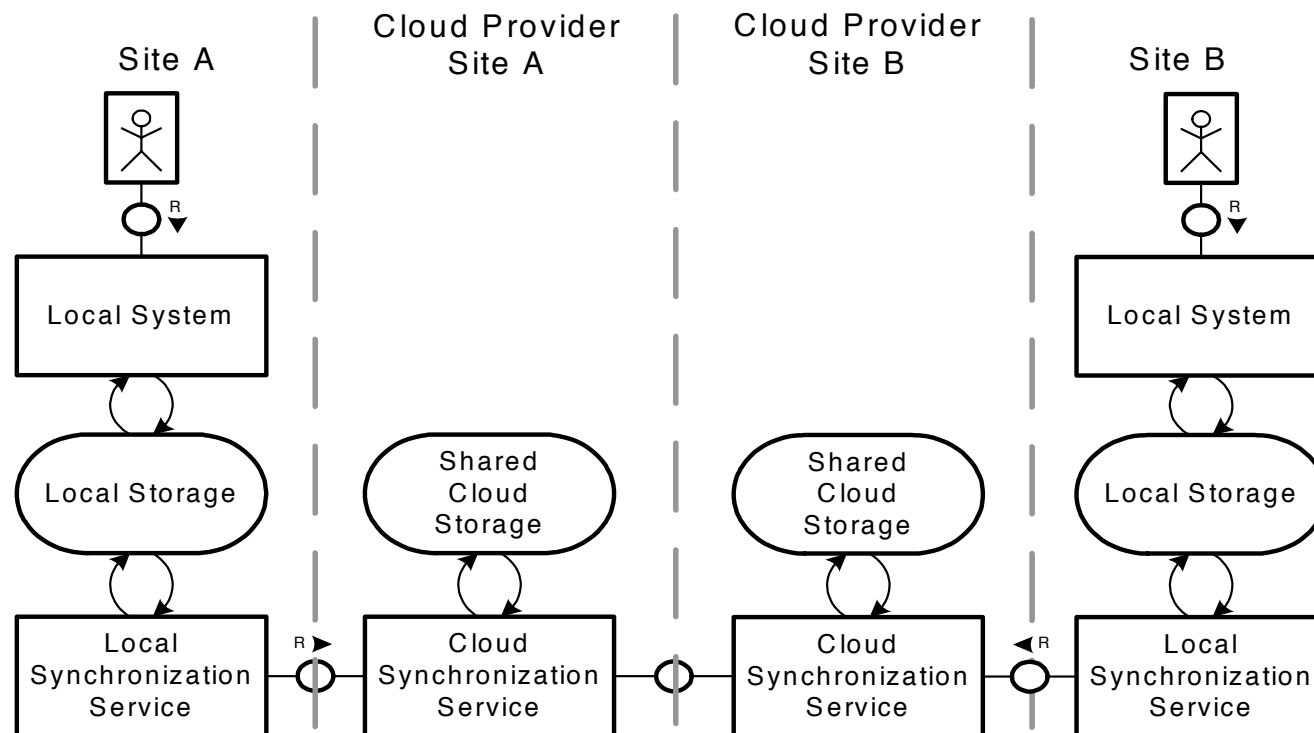
- Restrictions
 - Data privacy
 - Data locality
 - Volume of “big medical data”

- Solution?
 - Federated In-Memory Database System vs. Cloud Computing

Solve the Puzzle

Schapranow, Nov 2016

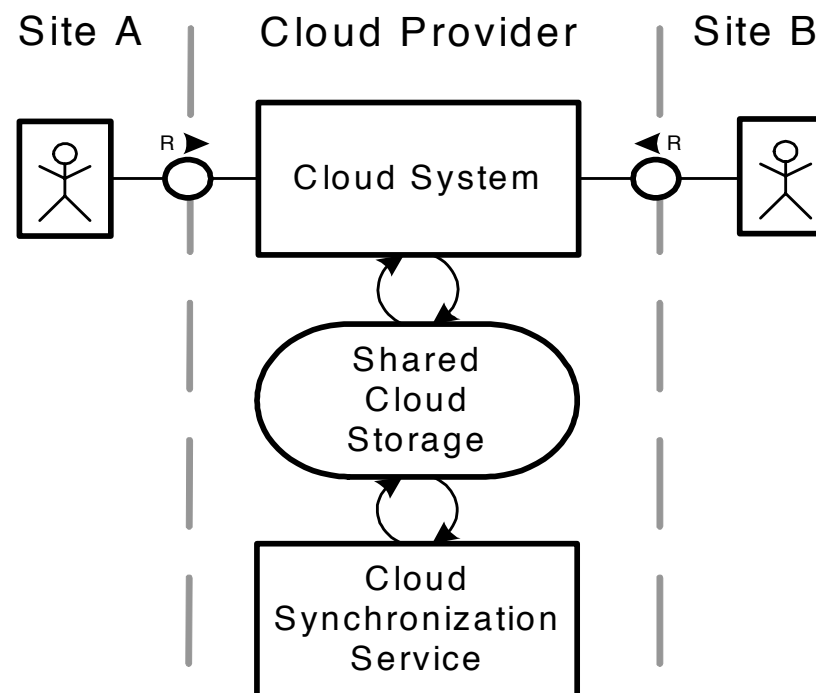
Approach I: Multiple Cloud Service Providers



Solve the Puzzle

Schapranow, Nov 2016

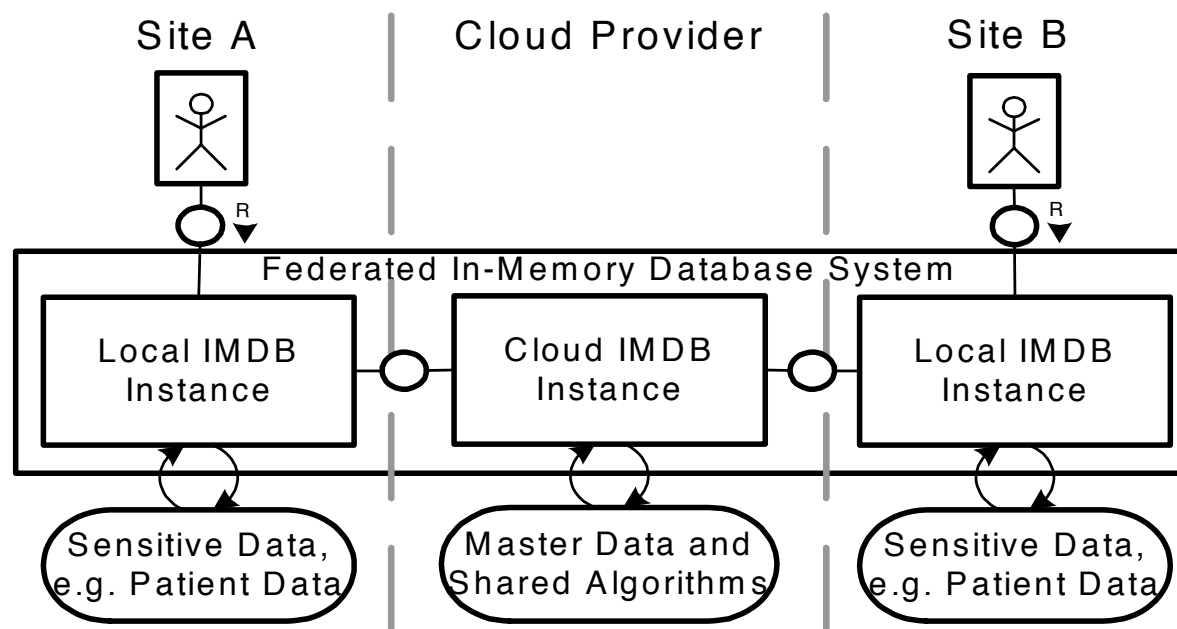
Approach II: A Single Service Provider



Solve the Puzzle

Schapranow, Nov 2016

Multiple Sites Forming the Federated In-Memory Database System (FIMDB)

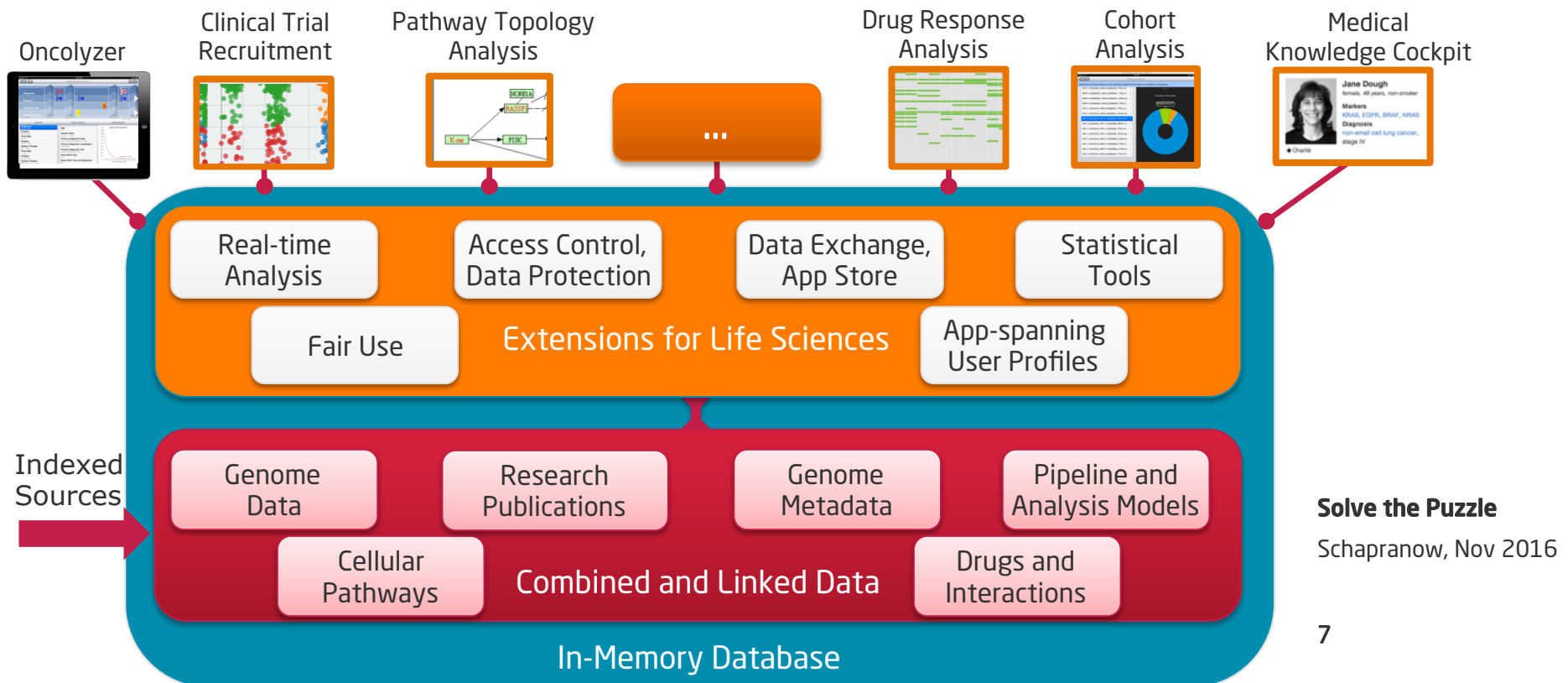


Solve the Puzzle

Schapranow, Nov 2016

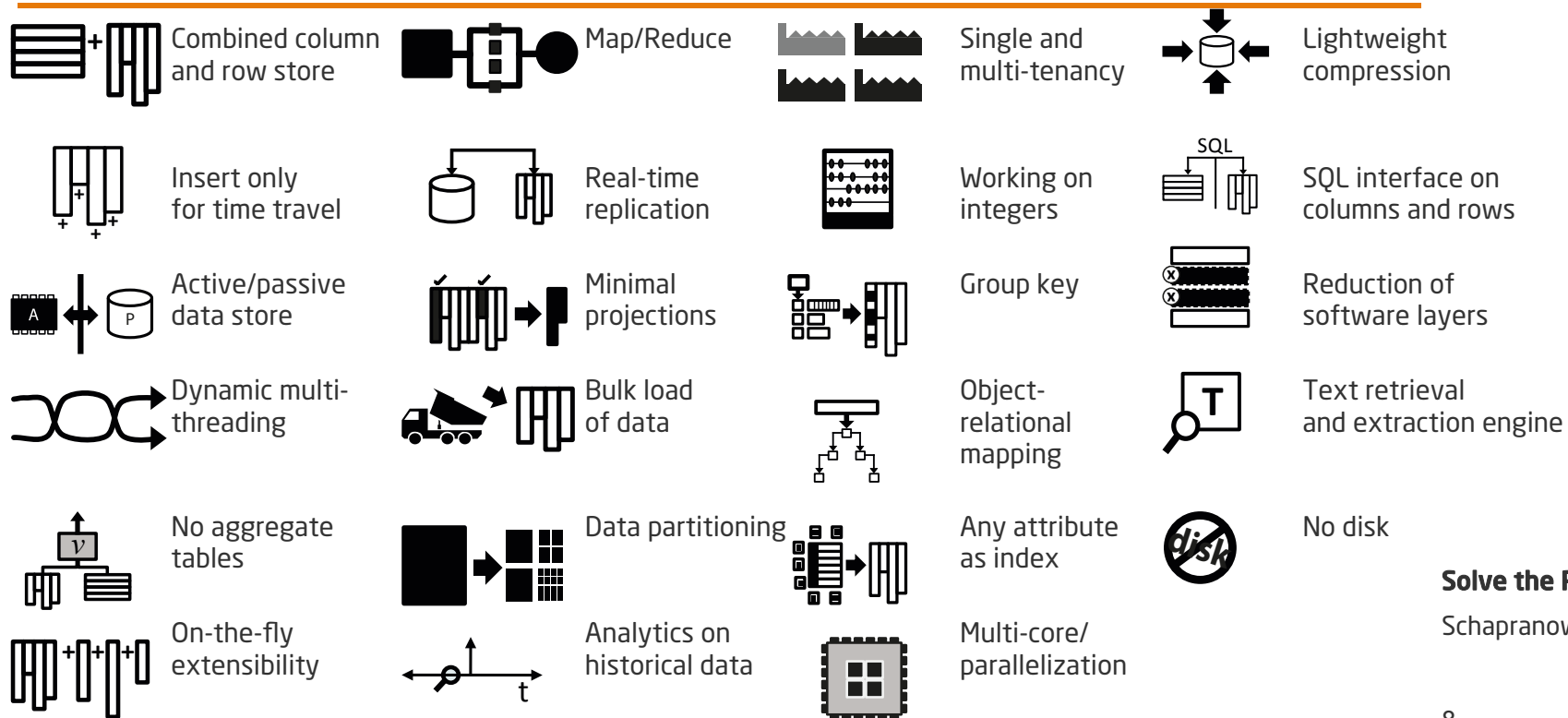
Our Approach

Analyze Genomes: Real-time Analysis of Big Medical Data



Our Technology

In-Memory Database Technology



Solve the Puzzle

Schapranow, Nov 2016

App Example: From Raw DNA to Variants



Hasso Plattner-Institut Design IT. Create Knowledge.

ANALYZE GENOMES by Hasso Plattner Institut

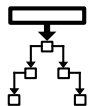
Log in | Results Genome Browser

Genome file: Upload or choose a file Pipeline category: Select pipeline category Pipeline: Select pipeline Process

Recent alignment tasks

Pages: 1 2 3 4 5 6 ... 29 (30 rows per page)

ID	Progress	Pipeline	Results	Cost
1458	Completed in 00h06m00s	BENCH_128 fastq Isaac_25	50132	Free
1457	Completed in 00h08m36s	BENCH_64 fastq Isaac_25	82271	Free
1456	Completed in 00h12m03s	BENCH_32 fastq Isaac_25	124164	Free
1455	Completed in 00h17m04s	BENCH_16 fastq Isaac_25	196269	Free
1454	Completed in 00h29m16s	BENCH_8 fastq Isaac_25	286552	Free
1453	Completed in 00h00m00s	BENCH_4 fastq Isaac_25	440512	Free
1452	Completed in 00h07m09s	BENCH_128 fastq Isaac_10	50132	Free
1451	Completed in 00h07m40s	BENCH_64 fastq Isaac_10	82271	Free
1450	Completed in 00h12m08s	BENCH_32 fastq Isaac_10	124163	Free
1449	Completed in 00h19m02s	BENCH_16 fastq Isaac_10	196301	Free
1448	Completed in 00h27m37s	BENCH_8 fastq Isaac_10	286500	Free
1447	Completed in 00h02m08s	BENCH_4 fastq Isaac_10	440512	Free
1446	Completed in 00h05m15s	BENCH_128 fastq Isaac_5	50132	Free
1445	Completed in 00h07m31s	BENCH_64 fastq Isaac_5	82272	Free
1444	Completed in 00h10m38s	BENCH_32 fastq Isaac_5	124163	Free
1443	Completed in 00h17m37s	BENCH_16 fastq Isaac_5	196303	Free
1442	Completed in 00h32m05s	BENCH_8 fastq Isaac_5	286554	Free



Standardized Modeling and runtime environment for analysis pipelines

- Control center for processing of raw DNA data, such as FASTQ, SAM, and VCF
- Personal user profile guarantees privacy of uploaded and processed data
- Supports reproducible research process by storing all relevant process parameters
- Implements prioritized data processing and fair use, e.g. per department or per institute
- Supports additional service, such as data annotations, billing, and sharing for all Analyze Genomes services
- **Honored by the 2014 European Life Science Award**



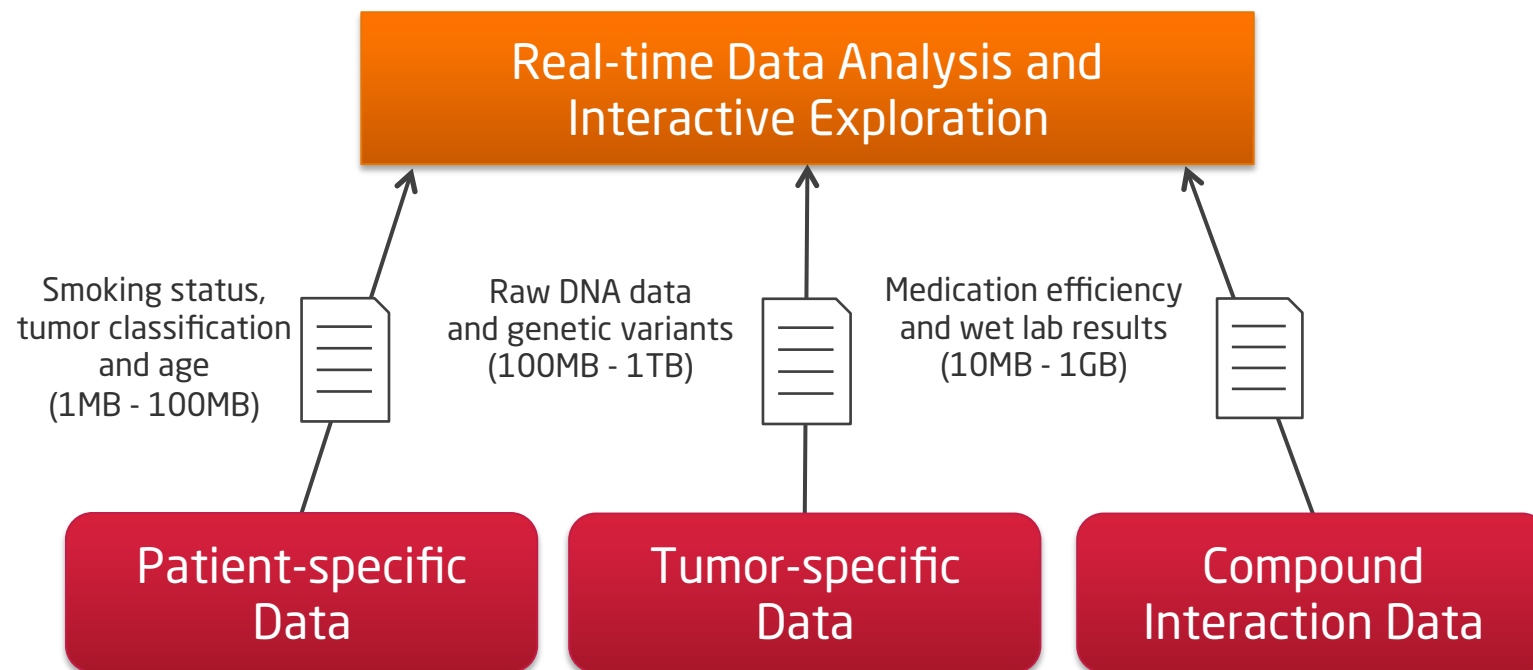
Solve the Puzzle
Schapranow, Nov 2016

App Example: Identification of Optimal Chemotherapy



Solve the Puzzle
Schapranow, Nov 2016

10



■ Honored by the 2015 PerMediCon Award

App Example: Latest Medical Knowledge for Patients and Clinicians

Medical Knowledge Cockpit

How about [A2M](#), [AADAT](#) or [PIK3CA](#)?

NRAS x

BRAF x

EGFR x

KRAS x

non-small cell lung cancer x



Jane Dough
female, 48 years, non-smoker

Markers
[KRAS](#), [EGFR](#), [BRAF](#), [NRAS](#)

Diagnosis
[non-small cell lung cancer](#),
stage IV

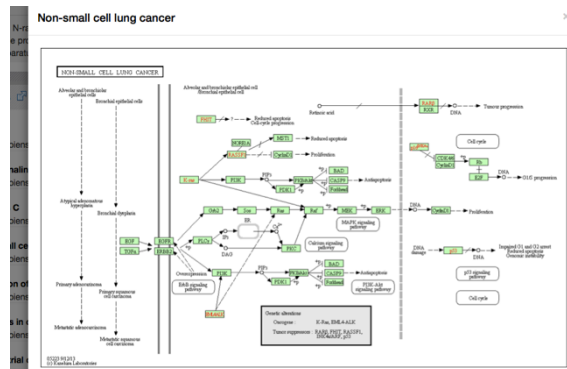
+ Charité

- Query-oriented search interface
- Seamless integration of patient specifics, e.g. from EMR
- Parallel search in international knowledge bases, e.g. for biomarkers, literature, cellular pathway, and clinical trials

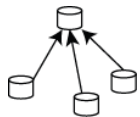
Solve the Puzzle

Schapranow, Nov 2016

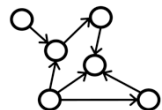
App Example: Pathway Topology Analysis



- Search in pathways is limited to “is a certain element contained” today
- Integrated >1,5k pathways from international sources, e.g. KEGG, HumanCyc, and WikiPathways, into HANA
- Implemented graph-based topology exploration and ranking based on patient specifics
- Enables interactive identification of possible dysfunctions affecting the course of a therapy before its start



Unified access to multiple formerly disjoint data sources

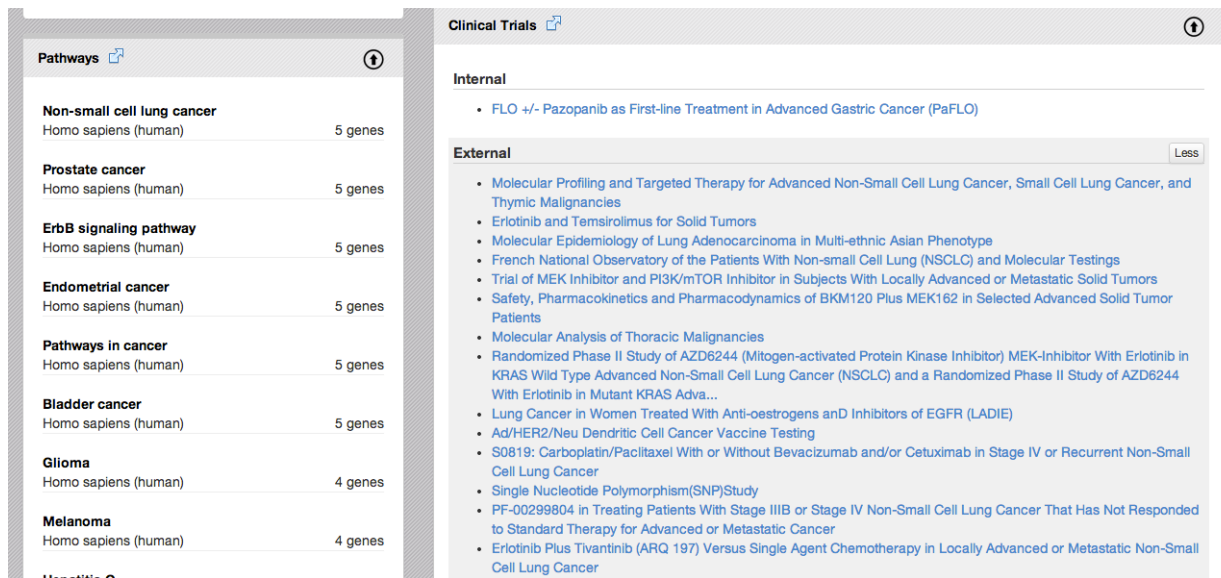


Pathway analysis of genetic variants with graph engine

Solve the Puzzle

Schapranow, Nov 2016

App Example: Latest International Clinical Trials



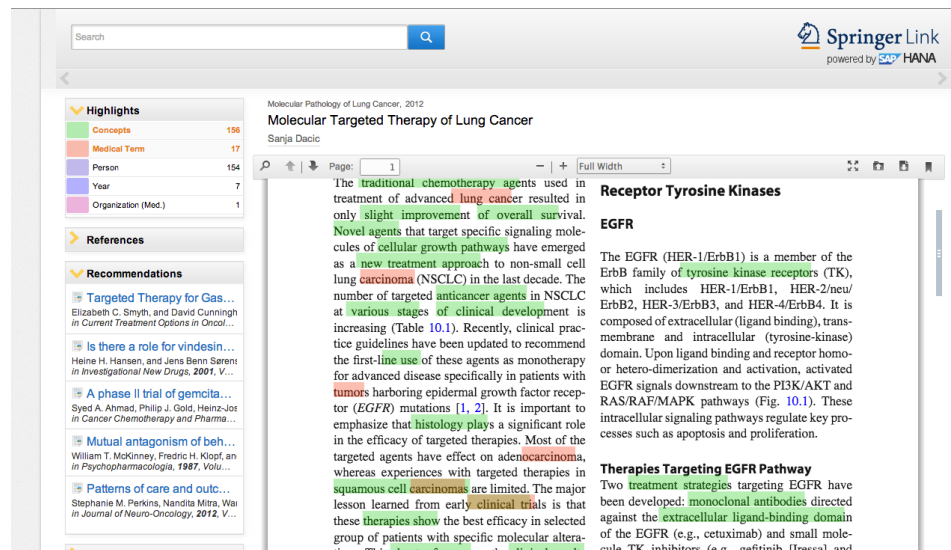
The screenshot shows a web application interface for clinical trials. On the left, a sidebar titled 'Pathways' lists various cancer types and their associated gene counts: Non-small cell lung cancer (5 genes), Prostate cancer (5 genes), ErbB signaling pathway (5 genes), Endometrial cancer (5 genes), Pathways in cancer (5 genes), Bladder cancer (5 genes), Glioma (4 genes), and Melanoma (4 genes). The main content area is titled 'Clinical Trials' and is divided into 'Internal' and 'External' sections. The 'Internal' section lists a single trial: 'FLO +/- Pazopanib as First-line Treatment in Advanced Gastric Cancer (PaFLO)'. The 'External' section, which is expanded, lists multiple trials related to Non-Small Cell Lung Cancer (NSCLC), including studies on Erlotinib, MEK inhibitors, and various clinical trial phases.

- Personalized clinical trials, e.g. by incorporating patient specifics
- Classification of internal/external trials based on treating institute

Solve the Puzzle

Schapranow, Nov 2016

App Example: Latest International Publications



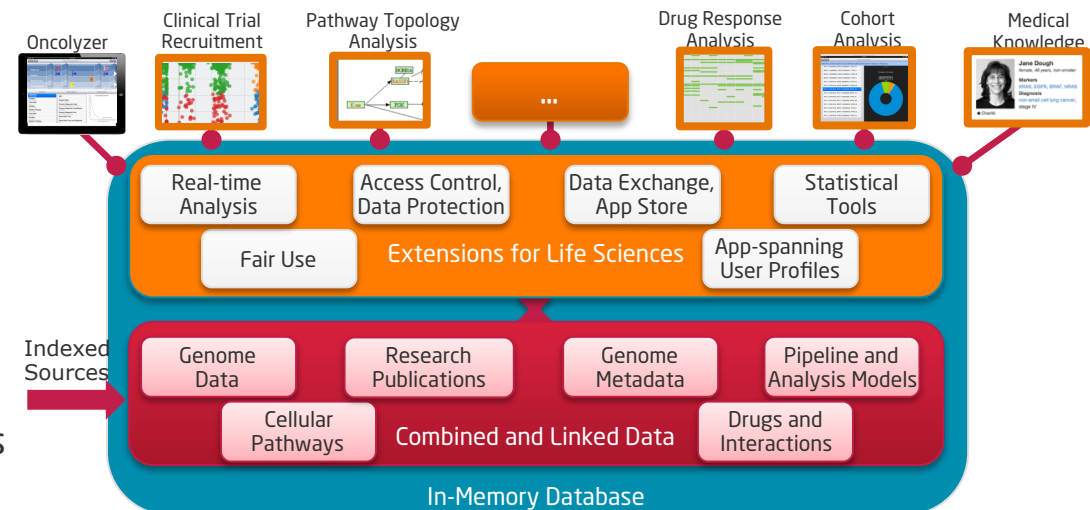
- Interactively explore relevant publications, e.g. PDFs
- Improved ease of exploration, e.g. by highlighted medical terms and relevant concepts

Solve the Puzzle

Schapranow, Nov 2016

What to take Home?

- Medical data is diverse
- Genetic data is a subset of all medical data
- Linkage and real-time analysis of data is crucial for doctors
- Specific software tools are available to support researchers, clinicians, and patients
- Test-drive selected apps at we.analyzegenomes.com



Solve the Puzzle

Schapranow, Nov 2016

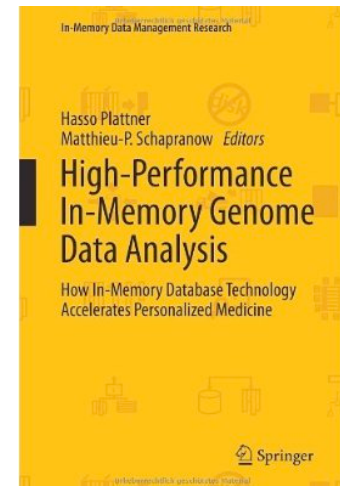
Keep in contact with us!



Dr. Matthieu-P. Schapranow
Program Manager E-Health & Life Sciences

Hasso Plattner Institute
August-Bebel-Str. 88
14482 Potsdam, Germany

schapranow@hpi.de
<http://we.analyzegenomes.com/>



Solve the Puzzle

Schapranow, Nov 2016