In-Memory Data Management for Life Sciences

M. Kraus, M. Schapranow, M. Neves, M. Uflacker

Seminar Kick-Off

Apr 12, 2016

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Agenda

- Seminar Organization
- Introduction Analyze Genomes
- Seminar Topics

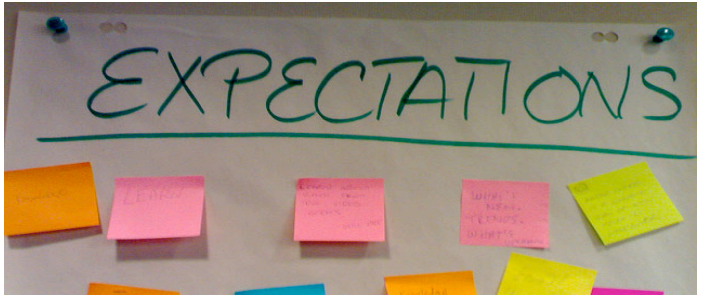
Seminar Organization Setup

- Supervisors: Milena Kraus, Dr. Matthieu-P. Schapranow, Dr. Mariana Neves, Dr. Matthias Uflacker

- Location: HPI Campus II, Room D.E-9/10 (former SNB)
- When: Tuesdays 9:15-10:45 a.m. (s.t.)
- Periods: 4 SWS (6 graded ECTS)
- Enrollment:
  - Prioritized topic wish list via e-mail to milena.kraus@hpi.de
  - Due Fri Apr 22, 2016 (HPI deadline) including
- http://hpi.de/plattner/teaching/summer-term-2016/in-memory-data-management-for-life-sciences.html

Seminar Organization
What you can expect from us

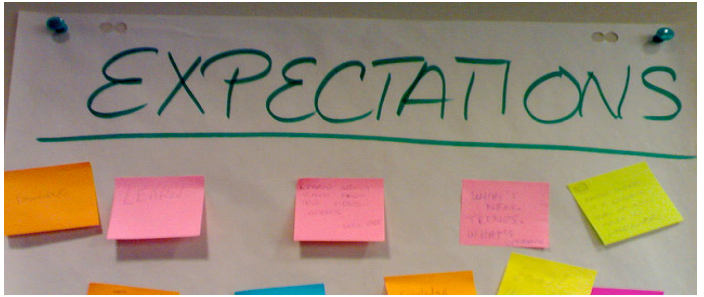
- Broaden your horizon in the fields of
  - In-memory technology,
  - Life sciences, and
  - Your selected seminar topic
- Get in touch and work with real-world data
- Work collaboratively together with experts from industry and research
- Work with latest hard-/software resources, e.g. beta systems in the Future SOC laboratory at HPI
- Get experienced in collaborative project work
- Enhance your skills in English presentation, scientific working, and writing



http://i.kinja-img.com/gawker-media/image/upload/s--cRElB5AZ--/1865smw5hbbt6jpg.jpg

Seminar Organization
What we expect from you

- Commitment on your selected seminar topic
- Perform autonomously research to acquire required knowledge about your selected seminar topic
- Work together in interdisciplinary teams
- Participate in every seminar meeting
- Systematic use of software design and engineering methods
- Contribute with your expertise also to your colleagues / other teams
- Update supervisors regularly on your progress / issues
- Handle sensitive data, e.g. from partners, confidentially



http://i.kinja-img.com/gawker-media/image/upload/s--cRElB5AZ--/1865smw5hbbt6jpg.jpg

Seminar Organization
Grading

- The grading of the seminar works as follows (aka "Leistungserfassungsprozess"):
  - 40% seminar results, i.e.
    - Research prototype
    - Presentations
  - 40% scientific research article
  - 20% individual commitment

- **All individual parts have to be passed** to pass the complete seminar

http://www.hpi.uni-potsdam.de/fileadmin/hpi/presse/Fotos/campus_und_gebaeude/20111017_HPI_Hoersaal.jpg

Next Steps
Enrollment for Seminar Topics

How to apply for a topic?

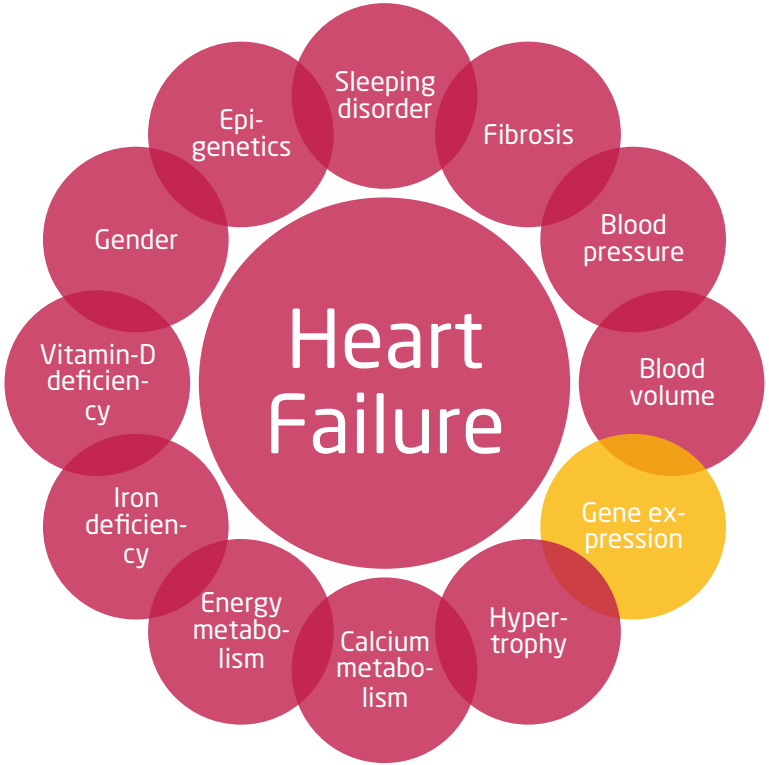
- Send prioritized list of top 3 topics to Milena Kraus (milena.kraus@hpi.de)
  - 1st choice: …
  - 2nd choice: …
  - 3rd choice: …
- Deadline: **Thu Apr 21, 2016 12pm (noon)**
- Assignment of seminar topics: **Fri Apr 22, 2016 12pm (noon)**

Seminar Topics

1. Processing of large RNAseq data sets to elucidate causes of heart failure
2. Natural Language Processing for Biomedicine

SMART – Systems Medicine Approach of Heart Failure

- SMART elucidates influences on the onset, progression, and treatment of heart failure
- Every aspect is evaluated by a different partner of a research consortium
- Combined analysis is facilitated by the HPI

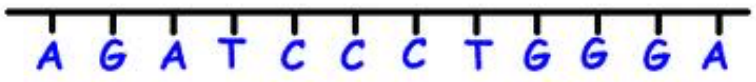


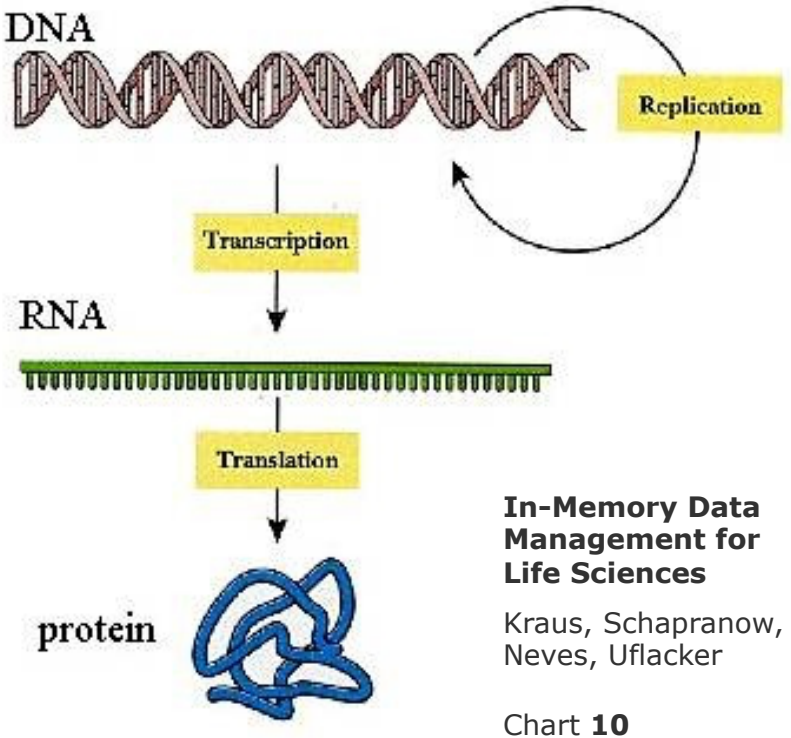
What is gene expression?

- Gene expression = synthesis of a protein with the help of genetic information

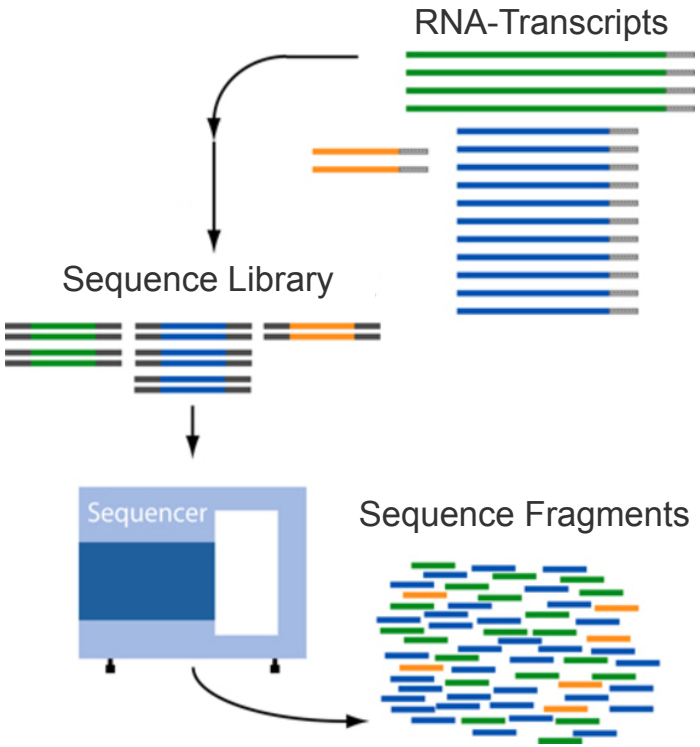
Most important facts for your task:

- A cell of a failing heart expresses other genes than a healthy heart cell → expression profile
- The number of found RNAs of one gene gives you the quantity of the corresponding protein
- RNA consists of the letters A, T(U), C, and G





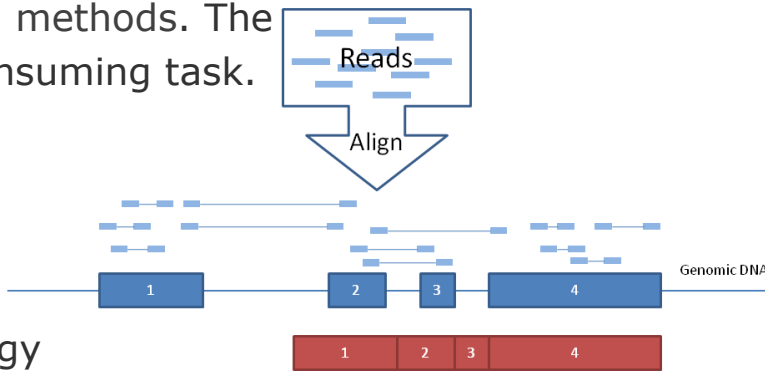
Creating the transcriptome out of raw experimental sequencing data



- RNA transcripts are broken into smaller (puzzle) pieces of short sequence reads
- Reads need to be "sorted" and aligned to a reference genome
- Aligned Reads are counted to give the respective RNA quantity
- Differences between conditions (ill, healthy) are computed through statistical methods and visualized accordingly

A: Processing of large RNAseq data sets to elucidate causes of heart failure

Issue: The transcriptome of a patient provides rich information for the elucidation of causes of heart failure. It needs to be build from raw RNAseq data with computationally and algorithmically challenging methods. The recreation of the transcriptome is a complex and time-consuming task.



Idea:

- Familiarize with processing of RNAseq data
- Evaluate means of optimization through IMDB technology
- Implement different processing pipelines in an IMDB
- Benchmark and evaluate your pipeline(s) with real patient data and compare to existing solutions

B: Statistical analysis of the transcriptome and differentially expressed genes

Issue: Methods for statistical analysis and visual exploration of RNAseq processing pipeline outputs exist and need to be implemented in our system. Inherent capabilities of the IMDB (PAL, Lumira) and R can be used to meet the requirements of our partner researchers.

Idea:

- Familiarize with the output RNAseq preprocessing
- Explore possibilities of statistical analysis in our IMDB and in R and also the IMDB-Rserv interface in particular
- Explore visualization capabilities of Lumira and R
- Choose and implement the best options for statistical analysis

C: Integration and Harmonization of Medical Data

Issue:

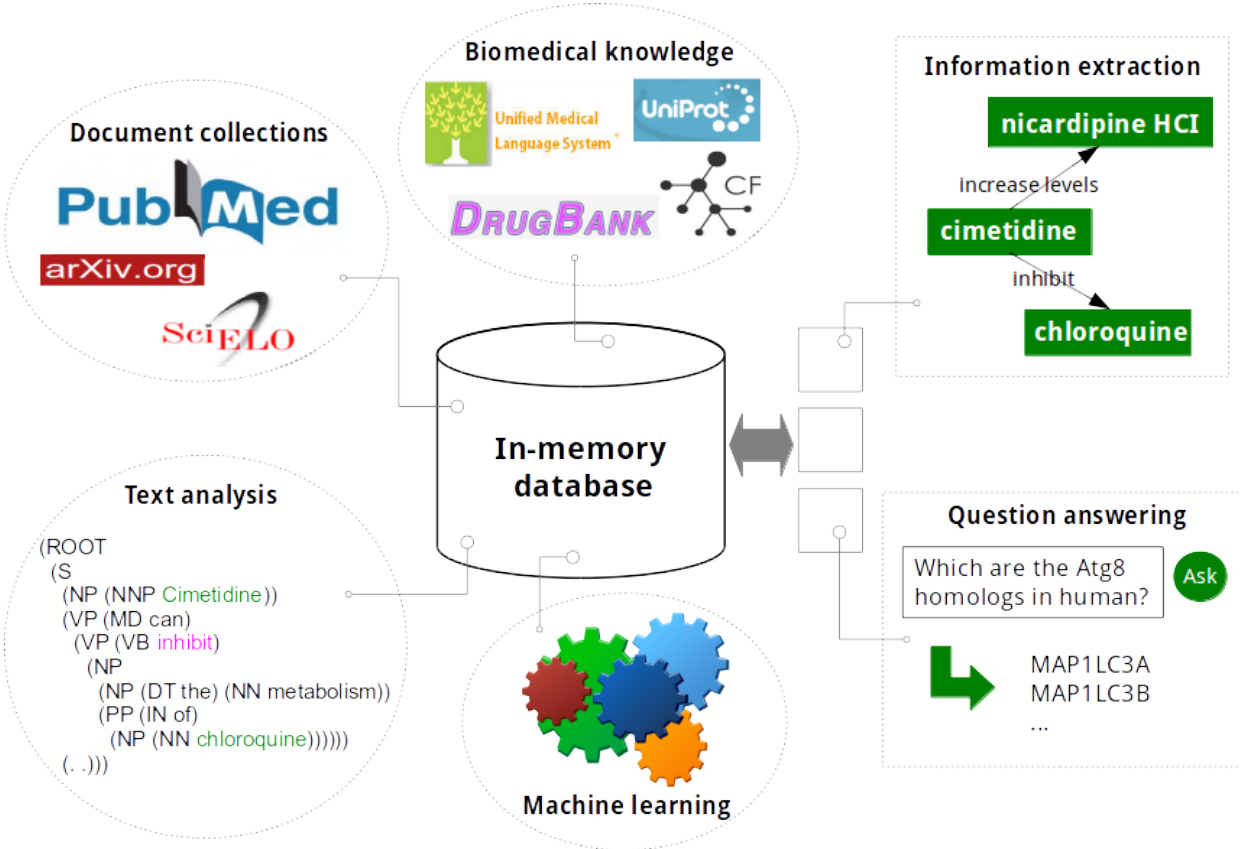
Clinical data is acquired in heterogeneous data formats in distributed data silos. Combining existing data sets for analysis is a manual task, which prevents efficient exploration of existing knowledge.

Idea:

- Explore existing data silos
- Define an integrated database model for harmonization
- Use existing analysis tools to test analysis capabilities of your data model
- Work in interdisciplinary teams with our cooperation partner

http://www.programmableweb.com/wp-content/FirstGiving_2.jpg

Natural Language Processing for Biomedicine



In-Memory Data Management for Life Sciences

Kraus, Schapranow, Neves, Uflacker

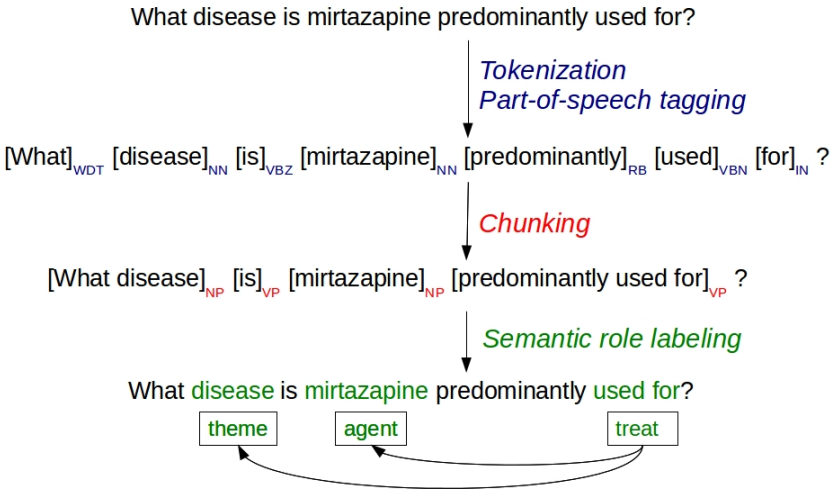
D: Semantic Role Labeling to Support Question Answering

Issue:

Finding the exact answer for a question requires the use of advanced natural language techniques.

Idea:

- Improve the current implementation of our semantic role labeling algorithm.
- Explore new methodologies and/or resources.
- Evaluate SRL in the scope of our question answering system.



E: Natural Language Processing to Support Clinical Decision

Issue:

Physicians frequently need to screen many publications to search for answers for clinical cases.

What is the patient's diagnosis?

What tests should the patient receive?

How should the patient be treated?

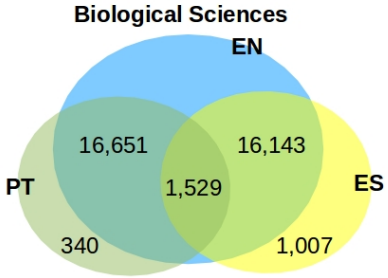
Idea:

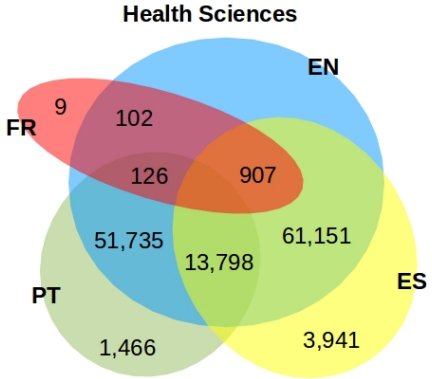
- Process and understand given textual clinical cases.
- Retrieve relevant full text publications to support answering the above questions.
- Integrate it into our text mining application.
- Take part in the TREC'16 Clinical Decision

F: Automatic Translation of Scientific Publications

Issue:

Most scientific publications are available only in English. On the other hand, publications from regional journals are only available in their local language.





Idea:

- Implement machine translation methods using in-memory technology.
- Evaluate on the Scielo dataset for English, Spanish, French and Portuguese.
- Create a prototype for automatic translation of biomedical publications.

G: Relation Extraction based on Distant Supervision

Issue:

The amount of annotated documents to support algorithms based on supervised learning is limited.

Idea:

- Extend our relation extraction system currently based only on supervised learning.
- Evaluate the algorithm for information extraction of biological data.
- Integrate it into our intelligent annotation tool.

General information	
Organism	Human herpesvirus 6
Tissue	-
EC Class	3.4.21
SABIO reaction id	11741
Variant	wildtype
Recombinant	expressed in Escherichia coli M15
Experiment Type	in vitro
Event Description	-

Substrates		
name	**location**	**comment**
Succinyl-RRYIKASEPPV-NH2	-	-
H2O	-	-

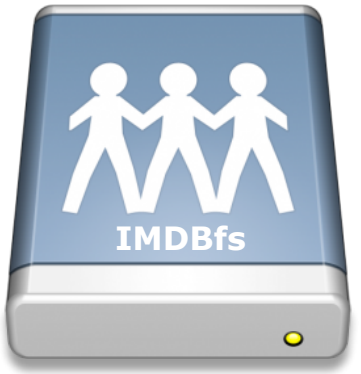
Products		
name	**location**	**comment**
SEPPV-NH2	-	-
Succinyl-RRYIKA	-	-

Y: IMDBfs - Adaption and evaluation of a shared high-performance file system built on in-memory technology

Hasso Plattner Institut

- IMDBfs := User space Distributed File System (DFS) extension built on libfuse, which incorporaes an In-Memory Database (IMDB) as persistency

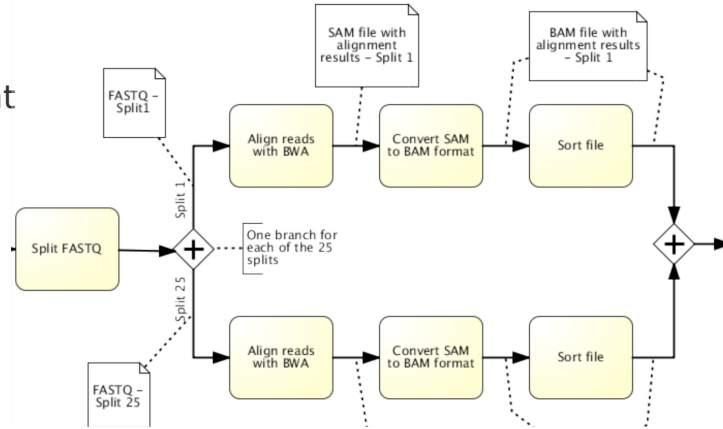
```
mount.imdbfs -ouser=IMDBFS,password=private,DSN=DNP,database=FUSE ../fs
```



- **Research tasks**
  - Setup benchmarks, perform measurements, compare with existing DFS alternatives
  - Identify performance bottlenecks, improve selected artifacts
- **Ideas**
  - Adapt file system for a concrete use case, e.g. from life science
  - Accelerate existing data processing tools by replacing long-running data import/ export into/from IMDB

Z: Distributed Execution: Adaption and evaluation of a distributed IMDB execution engine

- "Analyze Genomes" allows distributed execution of arbitrary BPMN processes using an IMDB for scheduling and job management



- **Research tasks**
  - Setup large scaled benchmark (individual physical locations)
  - Perform measurements and compare to existing frameworks
- **Ideas**
  - Derive robustness of software architecture
  - Identify latency when perform distributed processing, e.g. minimize data transfers
  - Improve selected artifacts