HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

In-Memory Databases: Applications in Healthcare

C. Fähnrich, M. Schapranow, M. Neves, M. Uflacker

Seminar Kick-off

Apr 21, 2015

Agenda

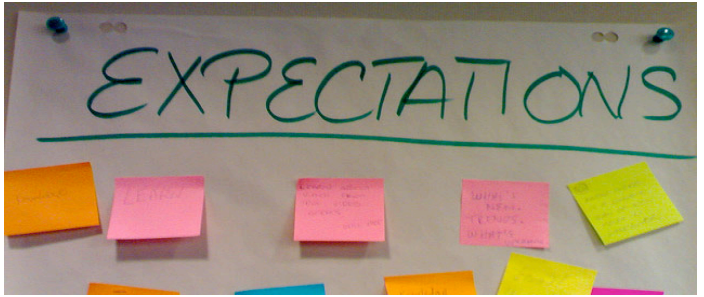
- Seminar Organization
- Seminar Topics
- Introduction Analyze Genomes

Seminar Organization Setup

- Supervisors:
  - Cindy Fähnrich,
  - Dr. Matthieu-P. Schapranow,
  - Dr. Mariana Neves,
  - Dr. Matthias Uflacker
- Location: HPI Campus II, Room D.E-9/10 (former SNB)
- When: Tuesdays and Thursdays 9:15-10:45 a.m. (s.t.)
- Periods: 4 SWS (6 graded ECTS)
- Enrollment: Due Fri April 24, 2015 (HPI deadline)
- http://hpi.de/plattner/teaching/summerterm2015/in-memory-databases-applications-in-healthcare.html

Seminar Organization
What you can expect from us

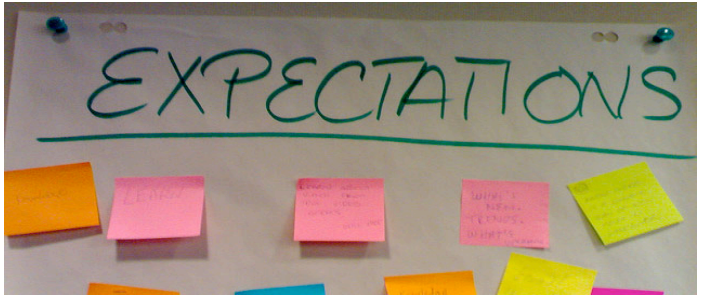
- Broaden your horizon in the fields of
  - In-memory technology,
  - Life sciences, and
  - Your selected seminar topic
- Get in touch and work with real-world data
- Work collaboratively together with experts from industry and research
- Work with latest hard-/software resources, e.g. beta systems 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
    - Intermediate presentations
  - 40% scientific research article
  - 20% individual commitment

- **All individual parts have to be passed** to pass the 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?

- Mail prioritized list of your top 3 topics to Cindy Fähnrich
  - 1 (very high adherence): …
  - 2 (high adherence): …
  - 3 (medium adherence): …
- Deadline: **Wed Apr 22, 2015 12pm (noon)**
- Assignment of seminar topics: **Thu, Apr 23, 2015 12pm (noon)**

Seminar Topics

A. Customized Reference for Genome Data Analysis
B. In-Memory-based Evaluation of Genetic Variants
C. Identification of Genetic Variants within an In-Memory Database
D. Linking Medical Knowledge to Improve Precision Medicine
E. Interactive Data Exploration of Medical Data
F. Integration and Harmonization of Medical Data
G. Analysis of Longitudinal Medical Data
H. Gene-Based Text Summarization
I. Sentiment Analysis for Controversial Topics
J. Text Mining on Gut Microbiota and Human Health

A: Customized Reference For Genome Data Analysis

Issue:
Genome Data Analysis (GDA) does not include population-specific features. By that, population-specific differences that might be relevant for analysis cannot be identified.

```
...
>7 dna:chromosome chromosome:GRCh37:7:1:159138663:1
...
AACATTCAAAAGCTGAGCAGGGCTTTAAAGCTATCTTATTAATAATTATTT
CTGTATTGCGAACTTCAGCATACTTTTTTCTAGTTACATTTGAAATGTTAT
TCTTTTGGGATGTGCTCAAGTGAGTACTGCTTTTTCCTCTGCCTTGCTTCA
TTACTTTTTAGTTTCCTTCATTTGAATCATCATTGTAAGTCTCCCCTTCTC
...
```

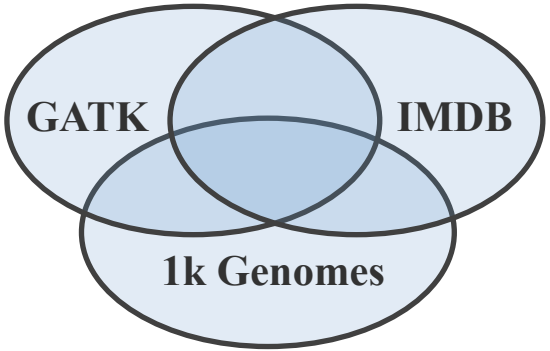
Idea:

- Analyze a sample set of individuals from the 1,000 genomes project for common genetic variants
- Customize the existing reference genome with those variants
- Run existing GDA tools with the customized reference genome and examine the impact of your changes made to the reference genome on the result set

B: In-Memory-based Evaluation of Genetic Variants

Issue:

There exist several tools for identifying genetic variants, which differ in their result sets and are suitable for a different use case. Estimating what tool delivers the best results for your use case, you need to analyze their result sets for sample data.



Idea:

- Get familiar with the data format for genetic variants and find a suitable data mapping for storing it in an in-memory database
- Identify evaluation criteria for comparing variant sets, e.g. include external data sources of well-known genetic variants
- Implement functionality to compare two or more variant sets within an in-memory database

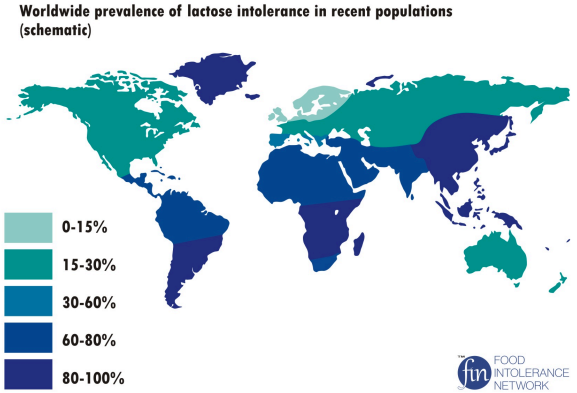
C: Identification of Genetic Variants within an In-Memory Database

Issue:

Currently, there are very simple computation models applied when identifying genetic variants. However, the procedures need to consider data quality and knowledge about genetic variation from research to deliver accurate results.

Idea:

- Research on currently existing statistical models in variant calling
- Identify and discuss relevant aspects for more accurate variant calling
- Adapt an existing in-memory-based variant calling algorithm to include selected aspects from your research, e.g. external data sources



http://upload.wikimedia.org/wikipedia/commons/2/27/Worldwide_prevalence_of_lactose_intolerance_in_recent_populations.jpg

D: Linking Medical Knowledge to Improve Precision Medicine

Issue:

Doctors need to find latest relevant knowledge, e.g. similar patient cases or publications, even if they do not know about it.

Idea:

- Similar patients also suffers from …
- Apply similarity metrics to improve ranking of results
- Work collaboratively with our cooperation partner
- Define concrete requirements for oncology use case
- Extend an existing system, e.g. the Medical Knowledge Cockpit, with required functionality

http://static.ddmcdn.com/gif/hippa-4.jpg

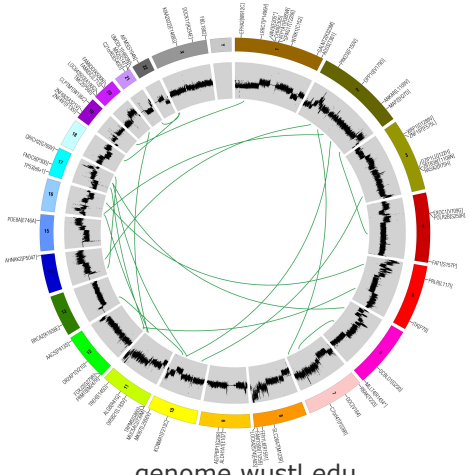
E: Interactive Data Exploration of Medical Data

Issue:

The Cancer Genome Atlas (TCGA) provides access to real patient data but its exploration and analysis is a time-consuming task

Idea:

- Explore existing patient data sets, e.g. RNA or TCGA data
- Derive concrete requirements for their interactive analysis
- Apply existing tools to implement research prototype enabling real-time analysis using in-memory technology
- Work in interdisciplinary teams with our cooperation partner
- Build upon and enhance exiting Analyze Genomes functionality



genome.wustl.edu

F: 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 prevent efficient exploration of existing knowledge.

Idea:

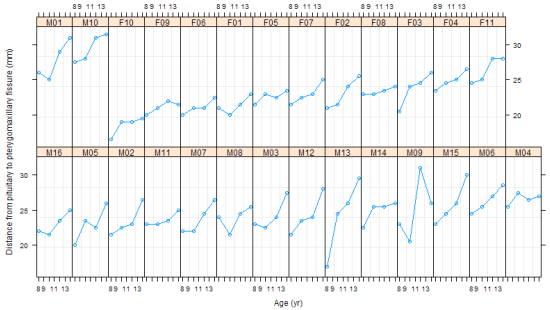
- Explore existing data silos
- Define an integrated database model for harmonization
- Use existing analysis tools to test analysis capabilities of your data model
- Work in interdisciplinary teams with our cooperation partner

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

G: Analysis of Longitudinal Medical Data

Issue:

Identification of historic patient cases can be used to improve treatment guidelines.



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

Idea:

- Analyze selected healthcare insurance data
- Define proper alternatives to detect pattern recognition
- Provide a prototypical implementation to identify patterns in longitudinal data and assess similar patient cases
- Work in interdisciplinary teams with our cooperation partner

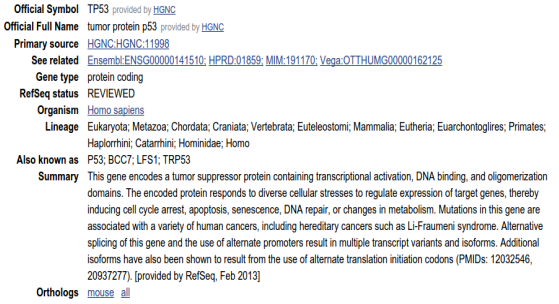
H: Gene-Based Text Summarization

Issue:

Biologists usually deal with long list of genes derived from microarrays experiments and they frequently need to search the vast scientific literature to learn more information.

Idea:

- Automatically generate summaries to help biologists to better translate their findings to clinical benefits
- Summaries should provide short and useful descriptions of each gene, e.g., functions of their proteins, interactions to other genes and associations to diseases
- Retrieve sentences that potentially cite the information needed
- Build summaries by merging phrases from different sentences and documents in a coherent manner and order



Official Symbol	TP53 provided by HGNC
Official Full Name	tumor protein p53 provided by HGNC
Primary source	HGNC:HGNC:11998
See related	Ensembl:ENSG00000141510; HPRD:01859; MIM:191170; Vega:OTTHUMG00000162125
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	P53; BCC7; LFS1; TRP53
Summary	This gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. The encoded protein responds to diverse cellular stresses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Mutations in this gene are associated with a variety of human cancers, including hereditary cancers such as Li-Fraumeni syndrome. Alternative splicing of this gene and the use of alternate promoters result in multiple transcript variants and isoforms. Additional isoforms have also been shown to result from the use of alternate translation initiation codons (PMIDs: 12032546, 20937277). [provided by RefSeq, Feb 2013]
Orthologs	mouse all

http://www.ncbi.nlm.nih.gov/gene/2768677

I: Sentiment Analysis for Controversial Topics

Issue:

Different studies frequently come to conflicting conclusions on the effect of food, medicaments and treatments on the human health.

Idea:

- Use text mining and sentiment analysis to automatically analyze the scientific literature
- Understand how these controversial topics have varied over the last years and the reasons for the changes
- Integrate domain terminologies for finding citations about the topics
- Apply information extraction for finding supporting findings
- Use sentiment analysis for figuring out the opinions about the topics
- Analyze how opinions and facts have changes over the years

J: Text Mining on Gut Microbiota and Human Health

Issue:

Recent scientific findings have been showing that many human disorders are somehow influenced by the human gut microbiota, the groups of bacteria and microbes that lives in our gut.

Idea:

- Use text mining for automatically identify existing findings on this topic
- Retrieve citations which describes association between bacteria and microbes of the human gut and human diseases
- Identify existing drugs and treatments based on this findings
- Identify differences across different countries and ethnic groups
- Facilitate access of scientists to this information

http://dchealthybites.com/tag/gut-microbiota-2/page/2/