Trends in Bioinformatics
Seminar Kickoff

Cindy Perscheid, Milena Kraus, Harry Freitas da Cruz, Mariana Neves

Agenda

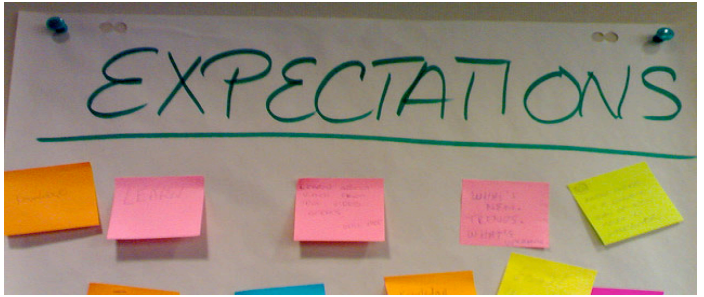
- Seminar Organization
- Seminar Topics

Seminar Organization Setup

- Supervisors: Cindy Perscheid, Milena Kraus, Harry Freitas da Cruz
- Time: Tuesdays 9.15-10.45 AM, individual appointments with your supervisor
- Location: D.E-9/10, HPI Campus II
- Periods: 4 SWS (6 graded ECTS)
- Enrollment:
  - Prioritized topic wish list via e-mail to *cindy.perscheid (at) hpi.de*
  - Due **Wed Oct 25, 11.59 PM**
  - Topic assignment notification by **Thu Oct 26, 1 PM**
  - Sign up for the course until **Fri Oct 27**
  - https://hpi.de//plattner/teaching/winter-term-201718/trends-in-bioinformatics.html

Seminar Organization
What you can expect from us

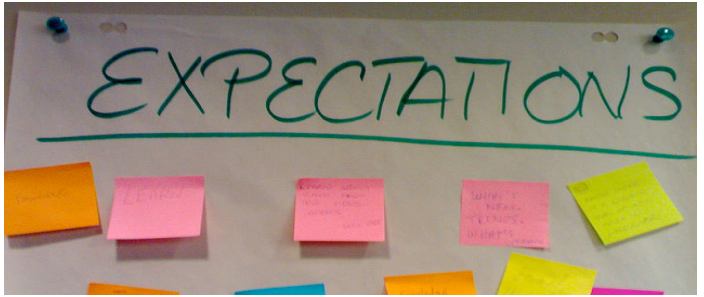
- Broaden your horizon in the fields of
  - Bioinformatics,
  - Life sciences, and
  - Your selected seminar topic
- Get in touch and work with real-world data
- Enhance your skills in English presentation, scientific working, and writing



http://i.kinja-img.com/gawker-media/image/upload/s--cREIB5AZ--/1865smw5hbbt6jpg.jpg

Seminar Organization
What we expect from you

- Commitment on your selected seminar topic
- Perform autonomous research to acquire knowledge about your selected seminar topic
- Hands-on experiments of selected tools on benchmarking data
- Participate in every seminar meeting
- Contribute with your expertise also to your colleagues / other teams
- Update supervisors regularly on your progress / issues



http://i.kinja-img.com/gawker-media/image/upload/s--cREIB5AZ--/1865smw5hbbt6jpg.jpg

Seminar Organization
Grading

- The grading of the seminar works as follows (aka "Leistungserfassungsprozess"):
  - **40%** intermediate presentation, final presentation, and abstract
  - **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

Seminar Organization
Enrollment for Seminar Topics

How to apply for a topic?

- Send prioritized list of top 3 topics to Cindy Perscheid (*cindy.perscheid (at) hpi.de*) until: **Wed Oct 25, 11.59 PM**
- Topic Assignments: **Thu Oct 26, 2017 1 PM**
- HPI course registration deadline: **Fri Oct 27, 2017**

Seminar Organization
Schedule (I/II)

- ***Dec 11 – 15:*** Intermediate presentations
  - 10 minutes presentation
  - Introduce your topic, problem/motivation, how you want to solve it
  - Slides due at day of presentation, 9 AM
  - Concrete dates tbd after topic assignment
- ***Jan 23, 9.15 AM***: Introduction to scientific writing
- ***Feb 5 – 9***: Final presentations
  - 30 minutes presentation
  - One-page abstract due one week prior to the presentation
  - Slides due at day of presentation, 9 AM
  - Present your approach and planned experimental setup
  - Concrete dates tbd after topic assignment

Seminar Organization
Schedule (II/II)

- ***Mar 31, 11.59 PM***: Scientific report
  - One report per topic
  - 4-6 pages for single students, 6-8 for teams (fixed upper bound!)

- ***TBD***: Excursions (optional)
  - Gläsernes Labor: Hands-on wet lab session; only if at least 5 students sign up
  - Max-Planck-Institute for Molecular Plant Physiology: Lab visit
  - Max-Planck-Institute for Molecular Genetics: Sequencing machines
  - We will schedule excursions once you have registered for the course

Seminar Topics

A. Data Mining on Gene Expression Data

1. An Interestingness Measure for Gene Expression Associations
2. Bi-Clustering with Biological Context Information
3. Causal Inference of Gene Expression Data
4. Verification of Gene Expression Patterns in Public Knowledge Bases
5. Optimize Calling of Genetic Variants from RNAseq Data
6. Clinical Interpretation of Omics Clustering Results
7. Statistical Basis of Differential Gene Expression (DGE) Analysis

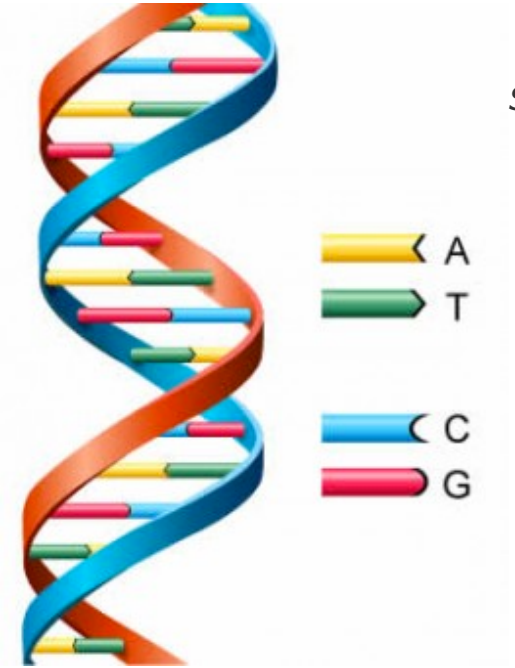
B. Text Mining for Biomedicine

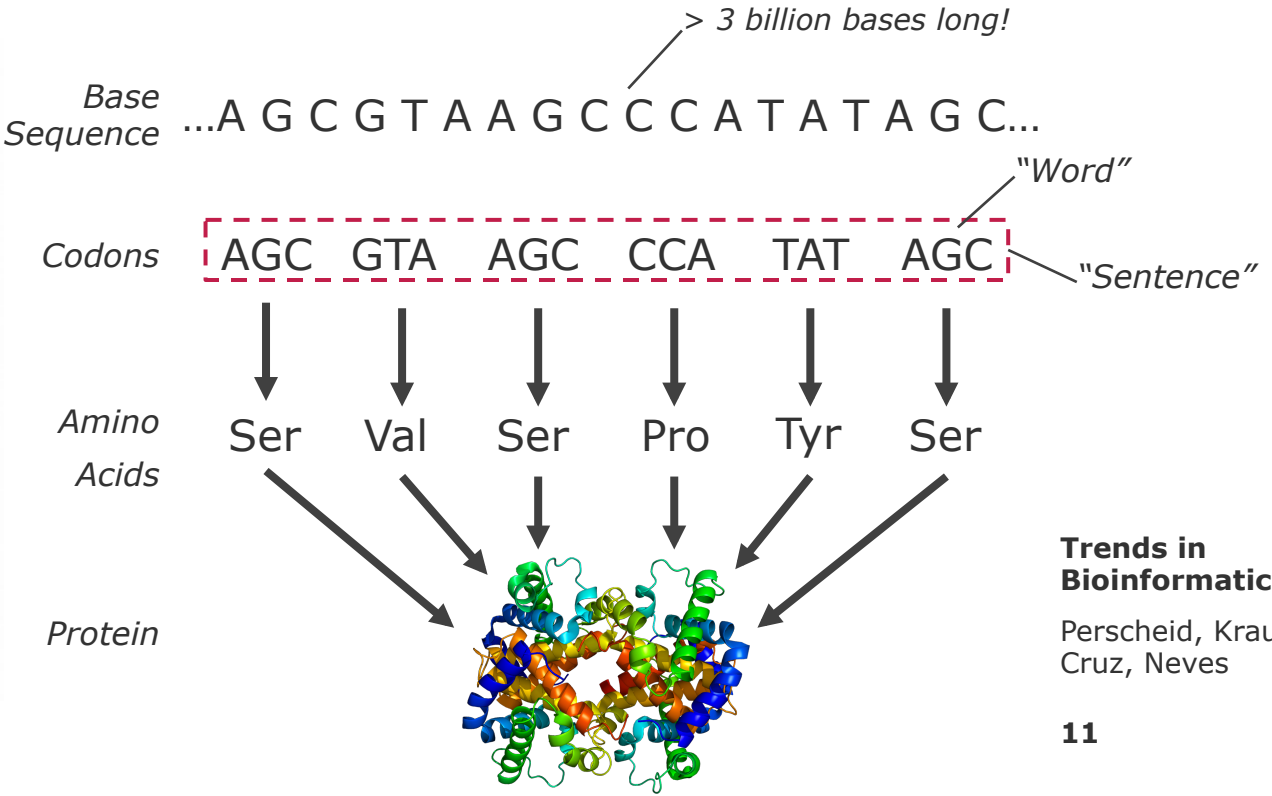
1. Extracting Scientific Entities and Relations from Publications to Support Searching for Alternative Methods to Animal Experiments

C. Prediction of Patient-Level Outcomes

1. Prediction of Patient Outcomes after Renal Replacement Therapy (RRT) in the ICU
2. Prediction of Incidence of Acute Kidney Injury in Cardiac Surgery

Short Biology Crash Course: The Human Genome

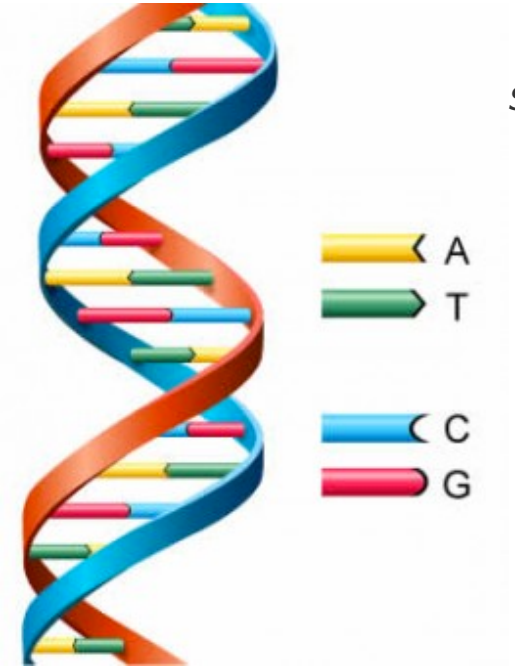


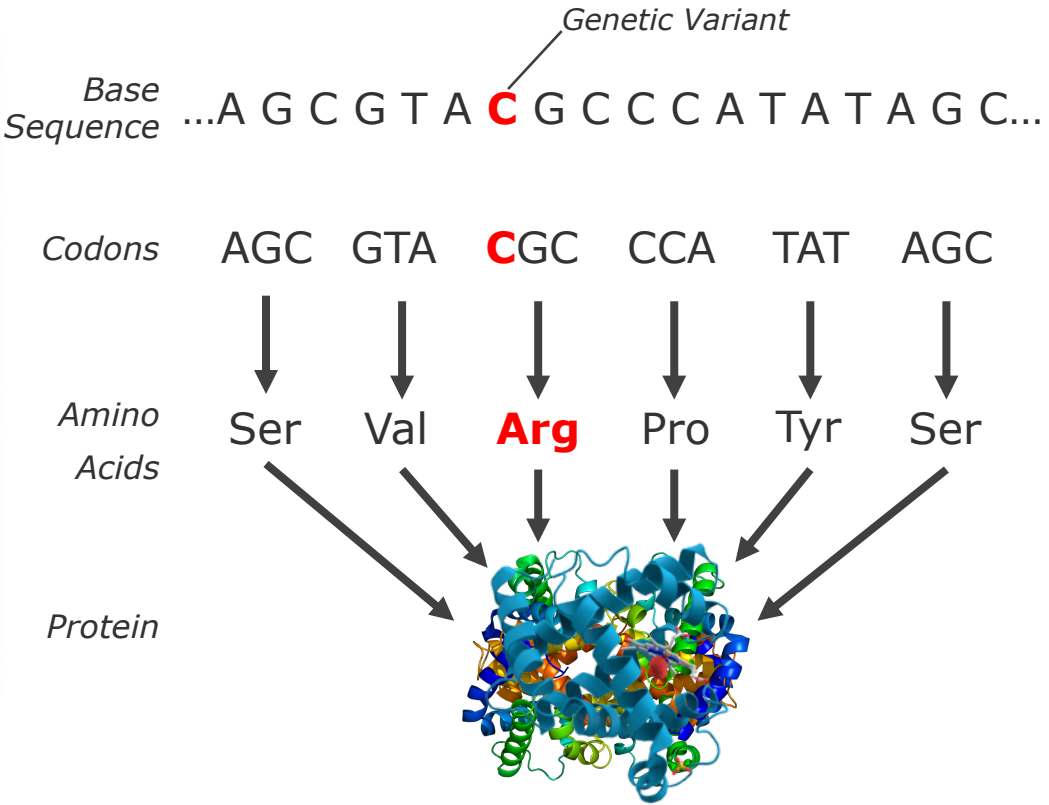


Trends in Bioinformatics

Perscheid, Kraus, Cruz, Neves

Short Biology Crash Course: Genetic Variants



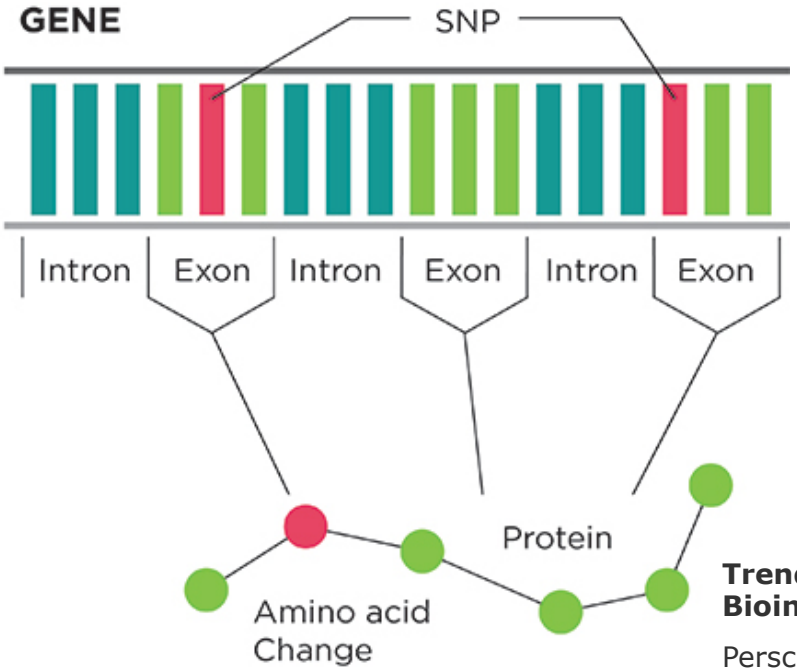


Trends in Bioinformatics

Perscheid, Kraus, Cruz, Neves

Crash Course: Genome vs. Transcriptome

- DNA provides more information
- RNA is cheaper to sequence
- Both contain redundant information



Trends in Bioinformatics

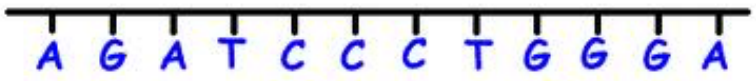
Perscheid, Kraus, Cruz, Neves

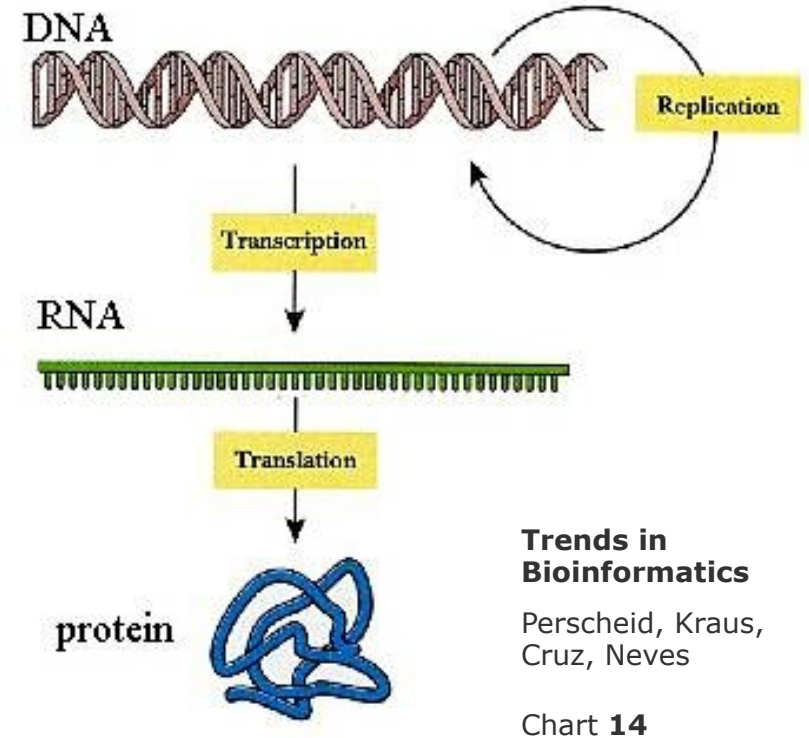
Short Biology Crash Course: What is Gene Expression?

- 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





Trends in Bioinformatics

Perscheid, Kraus, Cruz, Neves

A1. An Interestingness Measure for Gene Expression Associations

- Association rule mining can help to identify correlations between expression profiles

GeneA ↑ → *GeneB*↑

- Challenge: Filter rules to identify relevant assocations

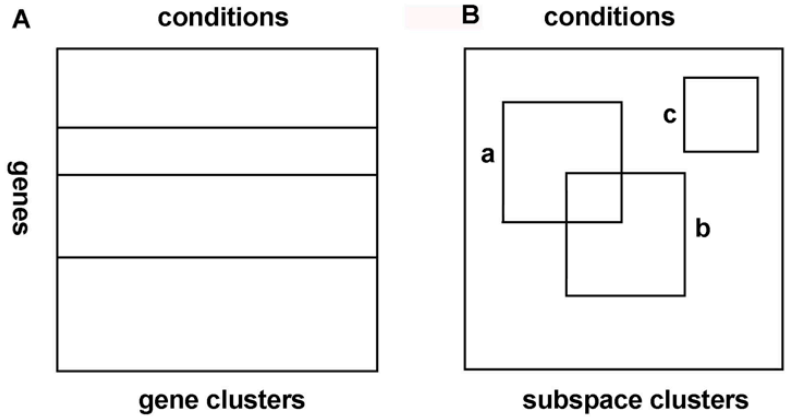
→ How do we know if a rule is *biologically* relevant?

- Your task: Define a subjective interestingness measure that takes into account the biological relevance of a rule
  - Build on support/confidence/lift
  - Integrate knowledge from external resources as rating criteria

A2. Bi-Clustering with Biological Context Information

- Traditional clusterings do not accurately reflect cell processes
  - Genes participate in multiple processes
  - Clustering result highly depends on selected genes
  - Clusters cannot necessarily be mapped to separate cell processes

- Your task: Use subspace/biclustering on gene expression data
  - Generate overlapping clusters for a more fine-grained distinction
  - Integrate existing knowledge on cell processes into clustering



A3. Causal Inference of Gene Expression Data

A simplified world

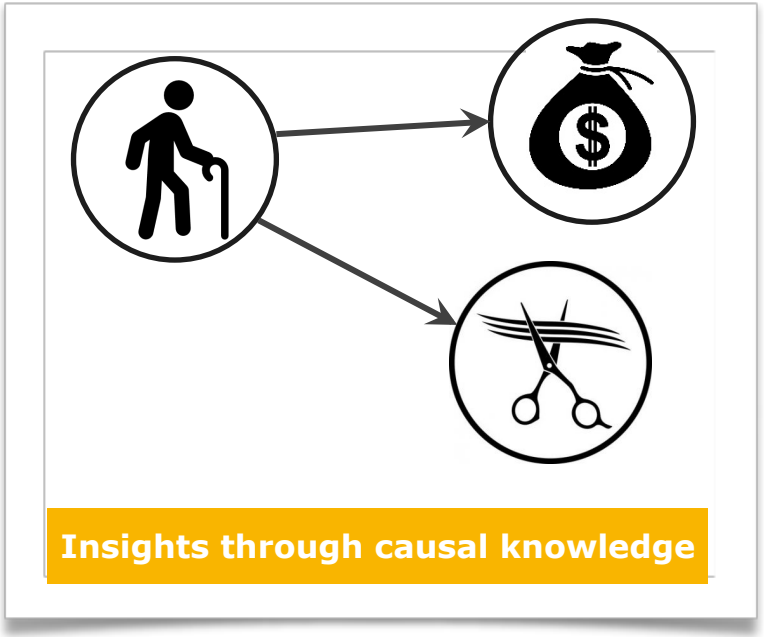
- Your life in three variables:
  - Age
  - Salary
  - Hair
  - → Age determines all

Challenge

- *Predict* and *optimize* salary!

Solution

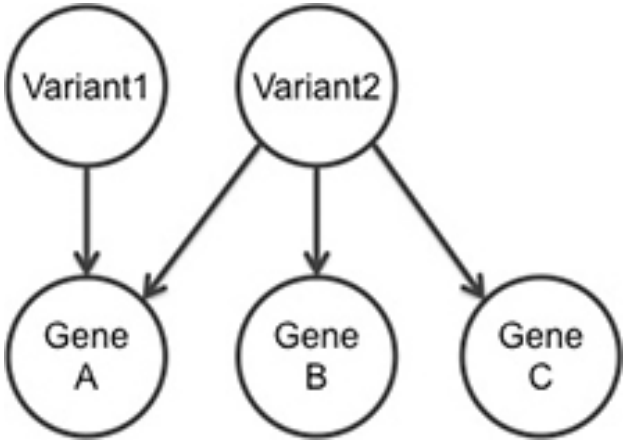
- Deep Learning can not help…
- …nor can standard Machine Learning!
- → But **Causal Inference can**!



A3. Causal Inference of Gene Expression Data

- Only 5% of the human genome encodes proteins
  - The rest is involved in regulatory processes
  - Large parts are still unexplored
  - → Can we use causal inference to identify causation between specific genetic variants and expression levels?

- Your task: Combine genetic variants and gene expression profiles for a causal inference analysis
  - Adapt the algorithm to cope with different data types
  - Implement a statistical model for each data type
  - Find efficient computing strategies for the algorithm



Trends in Bioinformatics

Perscheid, Kraus, Cruz, Neves

A4. Validation of Gene Expression Patterns in Public Knowledge Bases

- Analysis results must be validated by researchers

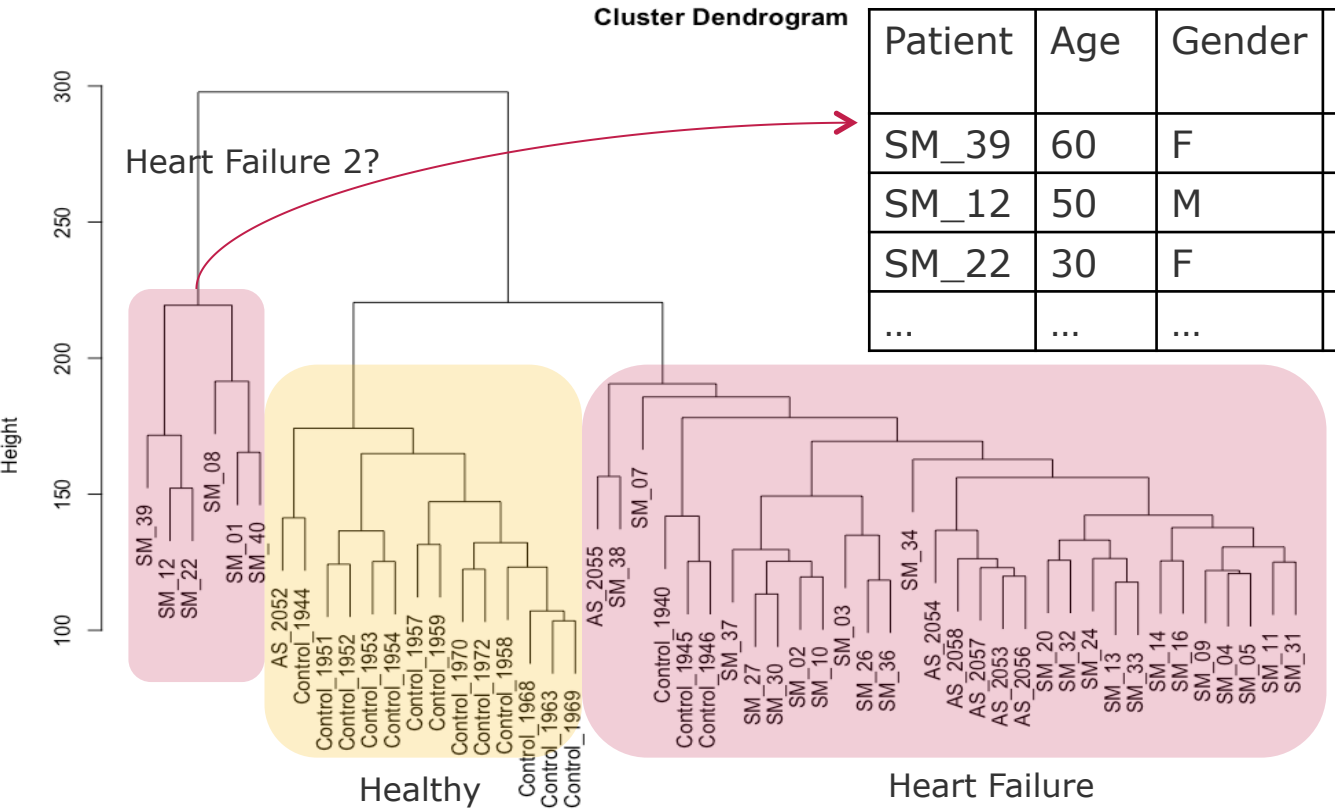
- Many associations can be validated in existing knowledge bases
  - Literature review
  - Keyword search
  - ...
  - → Data is hidden in heterogenous silos (20+ gene ids...)

- Your task: Implement a framework for the automatic validation of given gene-gene/gene-disease correlations
  - Identify suitable resources
  - Define ranking criteria

A5. Optimize Calling of Genetic Variants from RNAseq Data

- Understand:
  - Why current variant callers do not generally behave well with RNA-seq data
- Try out on new benchmark data set:
  - GATK Best practices
  - Optimized preprocessing (Opossum approach)
  - Optimized performance of the pipeline (HalvadeRNA)
  - Optimized subsequent filtering steps (e.g. from SNiPR).
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Discuss benefits and drawbacks of the approaches

A6. Clinical Interpretation of Omics Clustering Results



Patient	Age	Gender	Blood Pressure	Ethnicity	Weight
SM_39	60	F	120	Caucasian	60
SM_12	50	M	100	African	60
SM_22	30	F	110	Caucasian	60
…	…	…	…	…	…

- SMART set: > 150 variables that may explain a clustering result

Trends in Bioinformatics

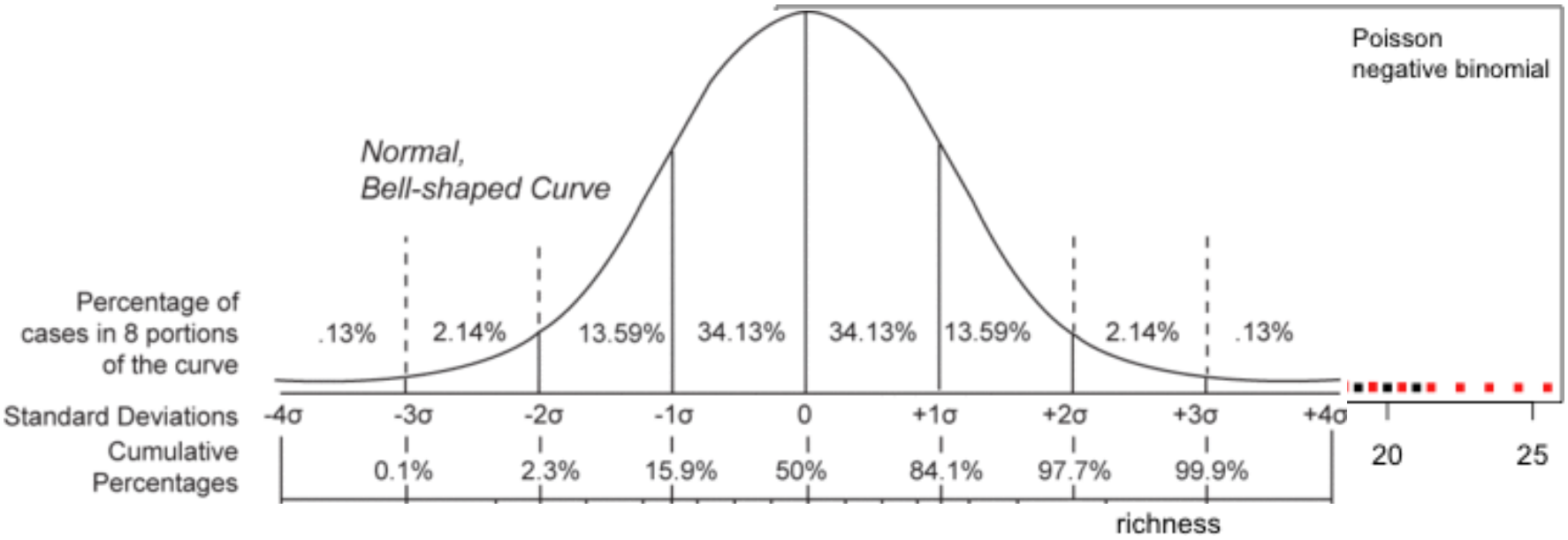
Perscheid, Kraus, Cruz, Neves

Chart **21**

A6. Clinical Interpretation of Omics Clustering Results

- Understand:
  - Decision trees
  - Clusters derived from omics data sets
  - Medical and biological background of heart failure
- Try out:
  - Train a decision tree on clinical data
  - Find parameters that may explain the given omics clustering result
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Explain computational and biological meaning of found parameters with the help of SMART cardiologist

A7. Statistical basis of differential gene expression (DGE) analysis

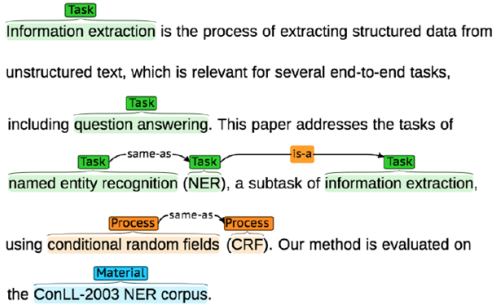


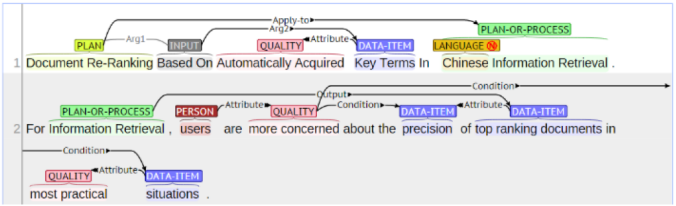
Perscheid, Kraus, Cruz, Neves

A7. Statistical basis of differential gene expression (DGE) analysis

- Understand:
  - Statistical methods in DGE analysis
  - Method used in kallisto/sleuth or sailfish
- Try out:
  - Kallisto/sleuth
  - Any other method(s) of your choice
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Discuss differences, benefits and drawbacks of methods

B1. Extracting Scientific Entities and Relations from Publications to Support Searching for Alternative Methods to Animal Experiments





- Researchers are required to **carefully search the biomedical literature for alternative methods to animal experiments**, e.g., in vitro instead of in vivo methods.
- Relevant publications should address the **same research goal** as proposed in the in vivo publication but should **describe an in vitro method**.
- Your task: Identify the **elements in a scientific abstract** and/or classifying the relationships between these.
- It will involve **supervised learning algorithms** for **named-entity recognition** and/or **relation extraction**.

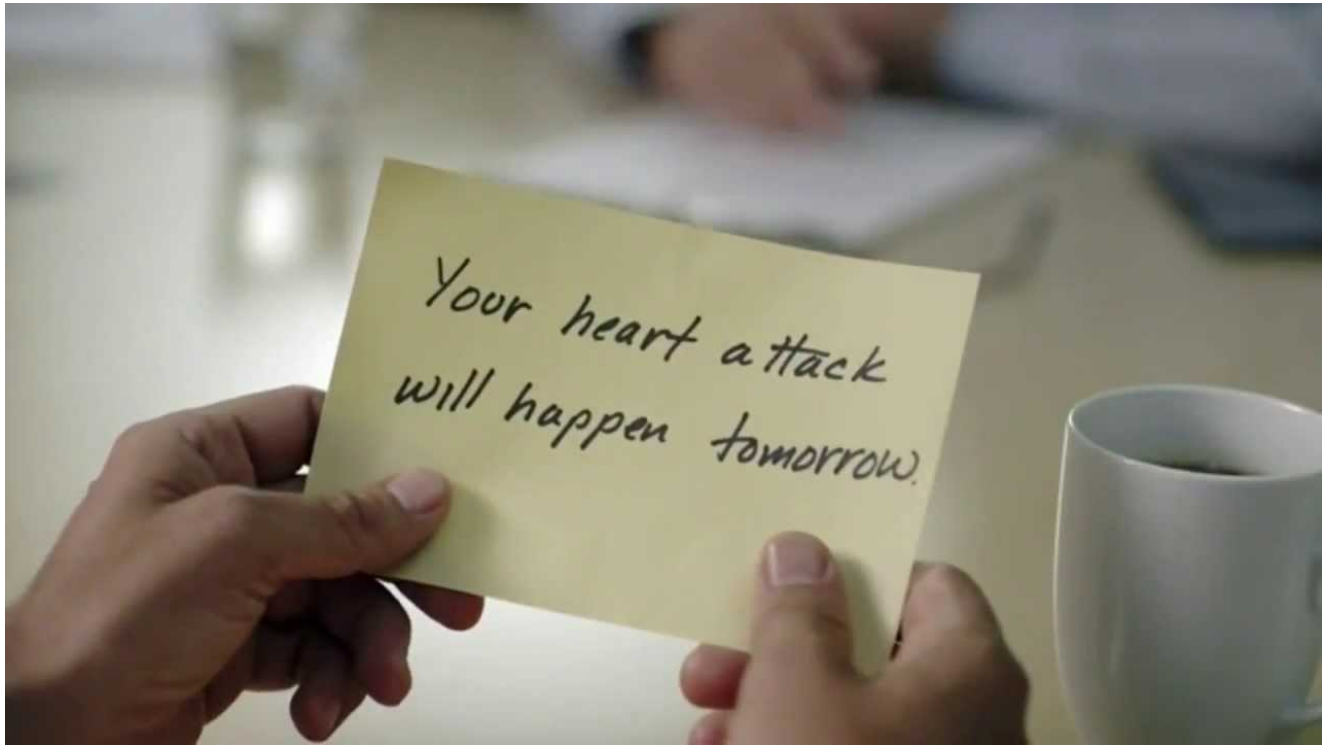
Trends in Bioinformatics

Perscheid, Kraus, Cruz, Neves

B1. Extracting Scientific Entities and Relations from Publications to Support Searching for Alternative Methods to Animal Experiments

- Understand:
  - Named-entity recognition (NER) and relation extraction (RE) methods
  - Machine learning (Supervised learning)
- Try out:
  - Train existing NER and RE tools for the task
  - Experiment with different schemata and corpora
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Discuss differences, benefits and drawbacks of the schemata

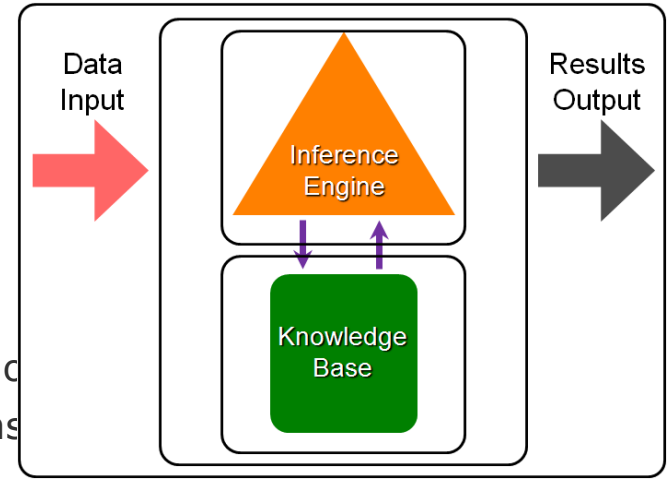
Prediction of Patient-Level Outcomes



https://i.ytimg.com/vi/xEemneA_qwE/maxresdefault.jpg

Prediction of Patient-Level Outcomes
Clinical Decision Support Systems

- Usual scenarios
  - Diagnostic support
  - Preventive care
  - Treatment planning / recommendations
- How can this be achieved?
  - Contextual retrieval of highly relevant informatic
  - Patient-specific reminders and recommendations
  - Organization and presentation of information
- Information logistics / 5 „rights“
  - Information, person, format, channel, time



Architecture components of CDSS (Kola, n.d.)

Berner, E. (2009). Clinical decision support systems: state of the art. AHRQ Publication, (9). Retrieved from https://healthit.ahrq.gov/sites/default/files/docs/page/09-0069-EF_1.pdf

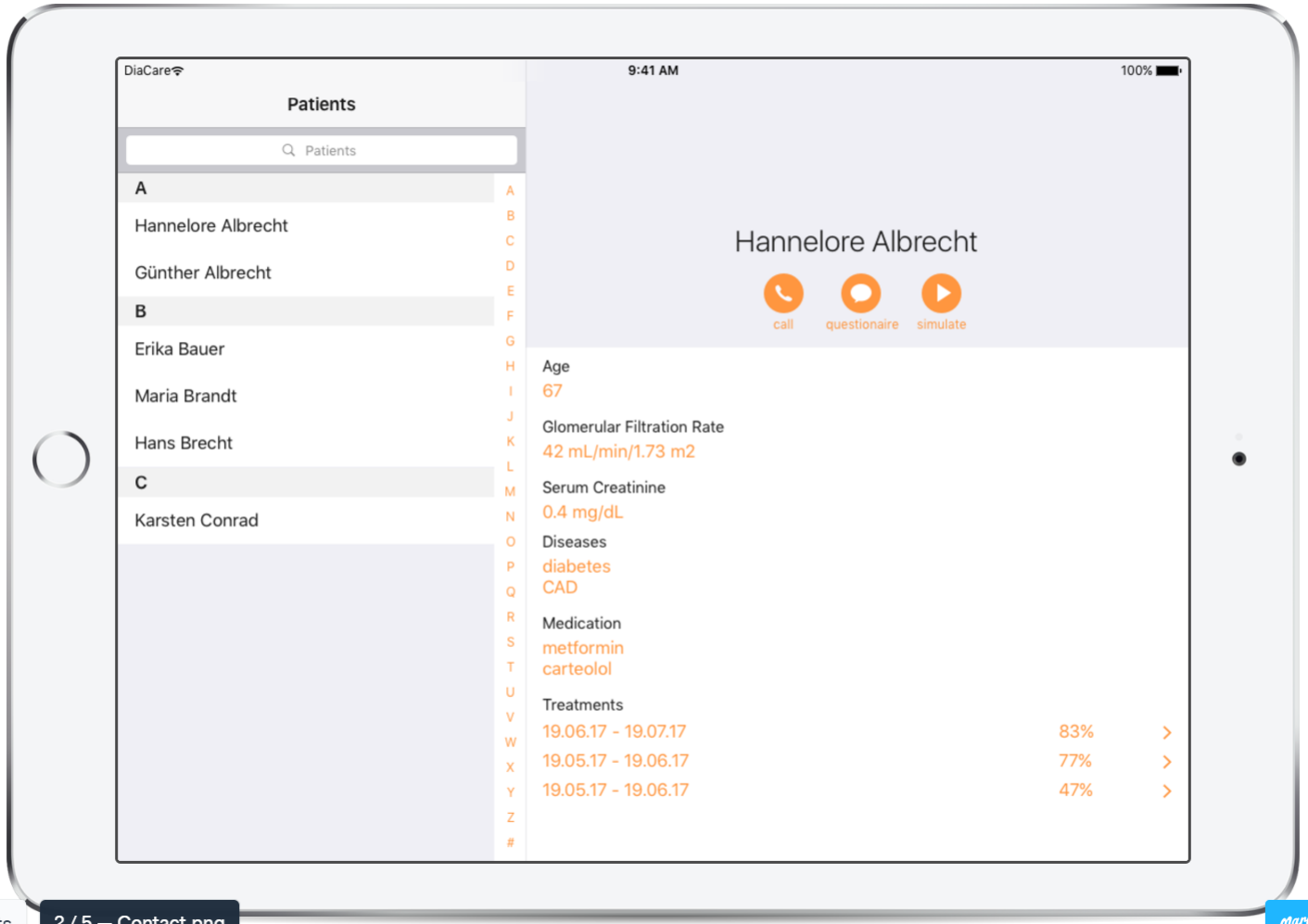
C1. Prediction of Patient Outcomes after Renal Replacement Therapy in the ICU

- **Dialysis in Germany[1]**
  - 70,000 patients / 2.5 Mio. EUR p.a.
  - 100,000 patients by 2020
  - High risk of mortality / high costs
- **Tasks:**
  - Predict patient outcomes using:
    - Gradient-boosted Decision Trees
    - Bayesian Networks
  - Dissect the respective algorithms
  - Write up a research paper

Source: Anna Frodesiak, CC0

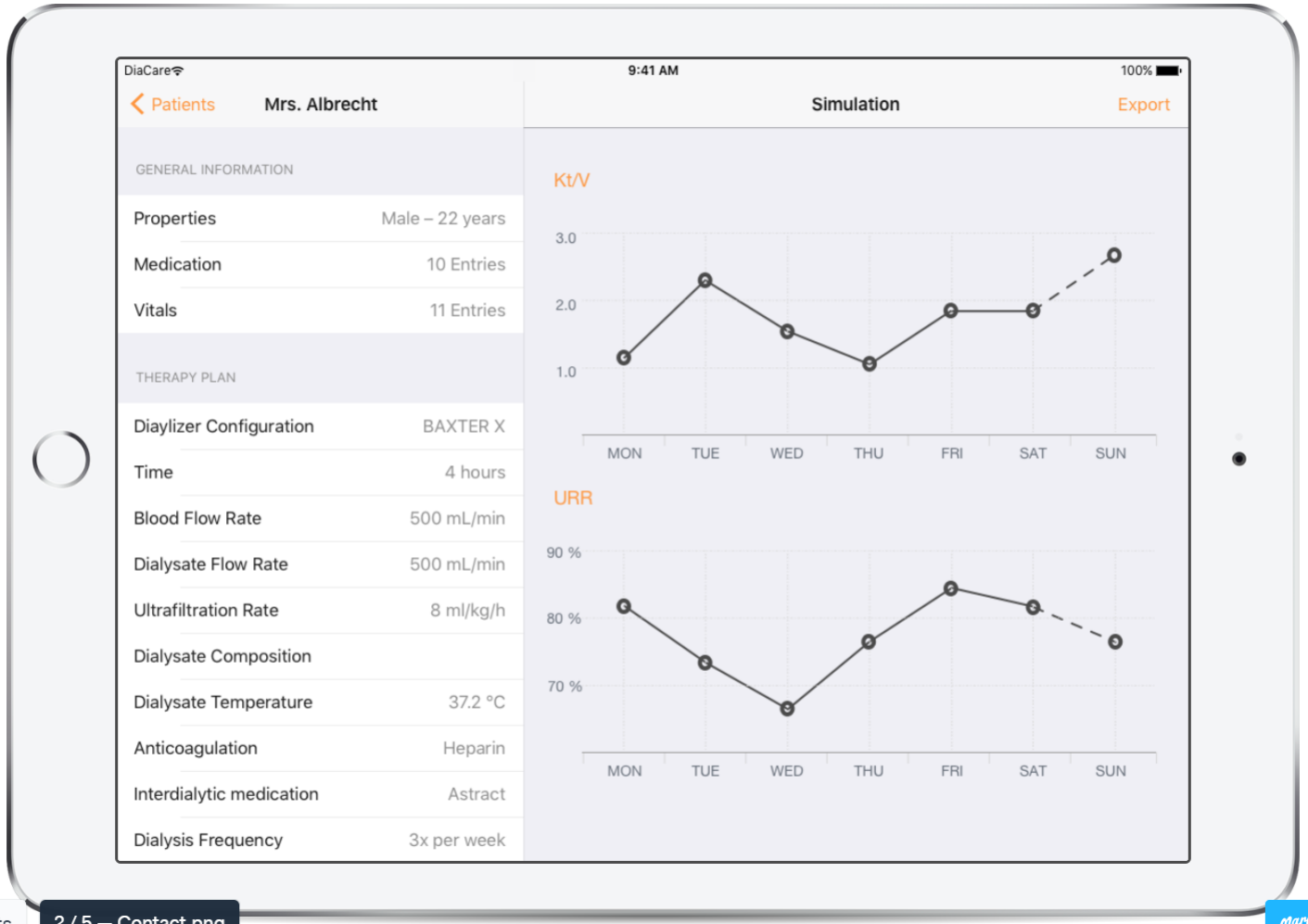
[1] http://www.aerzteblatt.de/nachrichten/41258/Zahl-der-Dialysepatienten-steigt

C1. Prediction of Patient Outcomes after RTT in the ICU



Optimizer Mock-up

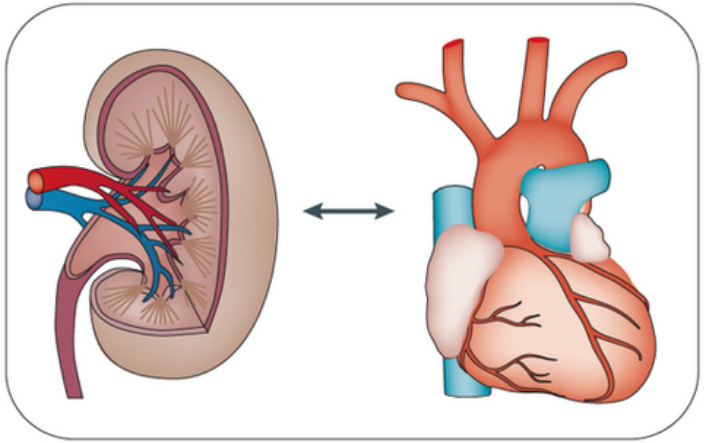
C1. Prediction of Patient Outcomes after RTT in the ICU



Optimizer Mock-up

C2. Prediction of Incidence of Acute Kidney Injury in Cardiac Surgery

- **Heart and kidneys are deeply connected**
  - AKI after cardiac surgery is relatively common (3 tc
  - Associated with complications and mortality
  - Patients under risk must be carefully monitored
- **Tasks**
  - Predict likelyhood of AKI using:
    - Gradient-boosted Decision Trees
    - Bayesian Networks
  - Dissect the respective algorithms
  - Write up a research paper



http://www.nature.com/nrneph/journal/v12/n10/abs/nrneph.2016.113.html

Trends in Bioinformatics

Perscheid, Kraus, Cruz, Neves

O'Neal, J. B., Shaw, A. D., & Billings, F. T. (2016). Acute kidney injury following cardiac surgery: current understanding and future directions. Critical Care, 20(1), 187. http://doi.org/10.1186/s13054-016-1352-z

Thanks for your attention!

- Choose your favorite topics by **Wed Oct 25, 11.59 PM**
- Come by at office V-0.01 for questions