Trends in Bioinformatics
Seminar Kickoff

Cindy Perscheid, Milena Kraus, Harry Freitas da Cruz

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Agenda

- Organization and Schedule
- Topics

Seminar Organization – Setup

- Supervisors: Cindy Perscheid, Milena Kraus, Harry Freitas da Cruz
- Time: Tuesdays 9.15-10.45 AM, and Wednesdays 1.30 – 3.00 PM, 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 24, 11.59 PM**
  - Topic assignment notification by **Thu Oct 25, 1 PM**
  - Sign up for the course until **Fri Oct 26**
  - https://hpi.de//plattner/teaching/winter-term-201819/trends-in-bioinformatics.html

Seminar Organization – Grading

- The grading of the seminar works as follows (aka "Leistungserfassungsprozess"):
  - **40%** intermediate and final presentation
  - **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 24, 11.59 PM**
- Topic Assignments: **Thu Oct 25, 2017 1 PM**
- HPI course registration deadline: **Fri Oct 26, 2017**

Seminar Schedule – Presentations

- ***Nov 26 – 30:*** 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 21 - 25***: Final presentations
  - 30 minutes presentation
  - Slides due at day of presentation, 9 AM
  - Present your approach and planned experimental setup
  - Concrete dates tbd after topic assignment

Seminar Schedule – Paper Writing

- ***Jan 29, 9.15 AM***: Introduction to scientific writing
- ***Mar 10, 11.59 PM***: Paper Submission Deadline
  - One paper per topic
  - 4-6 pages for single students, 6-8 for teams (fixed upper bound!)
  - Iterate with your supervisor
- ***Mar 18:*** Notification of reject or accept w/o (minor) revisions
- ***Mar 30:*** Submission of camera-ready version

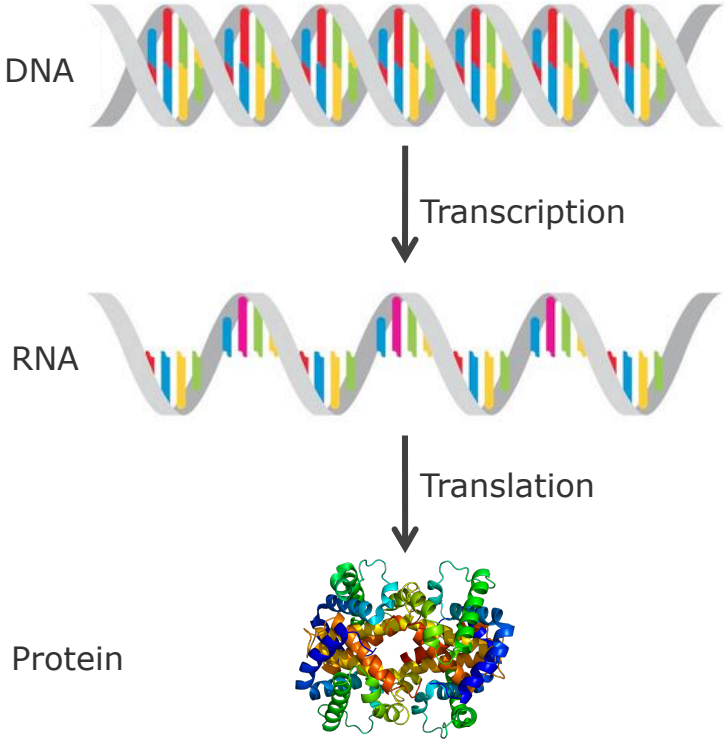
Seminar Topics

A. **Analysis of RNAseq Data (Supervisor: Cindy Perscheid)**
   1. Integrative Gene Selection
   2. Association Rule Mining
   3. Integrative Gene Selection vs. Integrative Clustering
   4. Biological Evaluation of Marker Genes

B. **Analysis of Multi-Omics Data (Supervisor: Milena Kraus)**
   1. Calculate and validate eQTLs in Heart Failure
   2. Calculate and validate pQTLs in Heart Failure
   3. Feasibility of "expQTLs"
   4. Bayesian Clustering of Multi-Omics
   5. Similarity Network Fusion on Multi-Omics

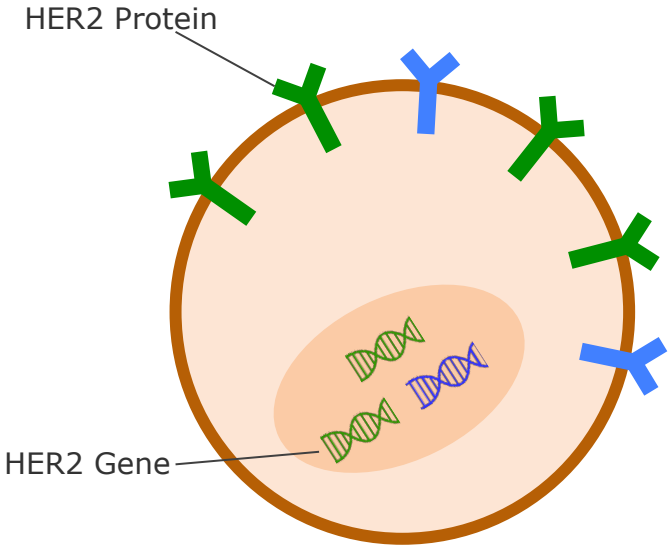
C. **Interpretability Approaches applied to Clinical Predictive Modeling (Supervisor: Harry Freitas da Cruz)**

Central Dogma of Molecular Biology – From DNA to RNA

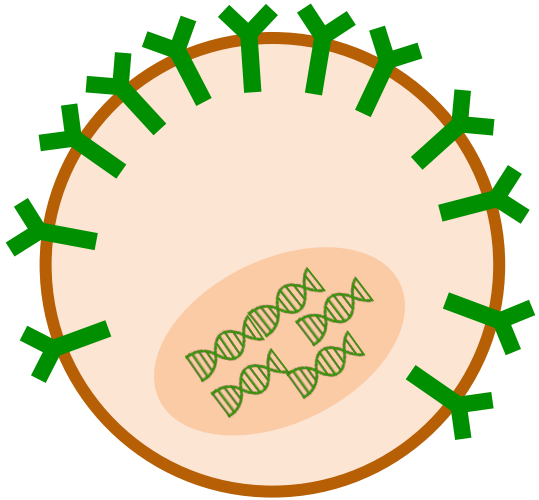


- **Protein**: Gene product controlling cell metabolisms
- **Gene Expression**: Cell process where protein is built from gene information encoded in DNA
- **Expression Level**: Production rate of protein

Gene Expression Rates – What Differentiates Cells

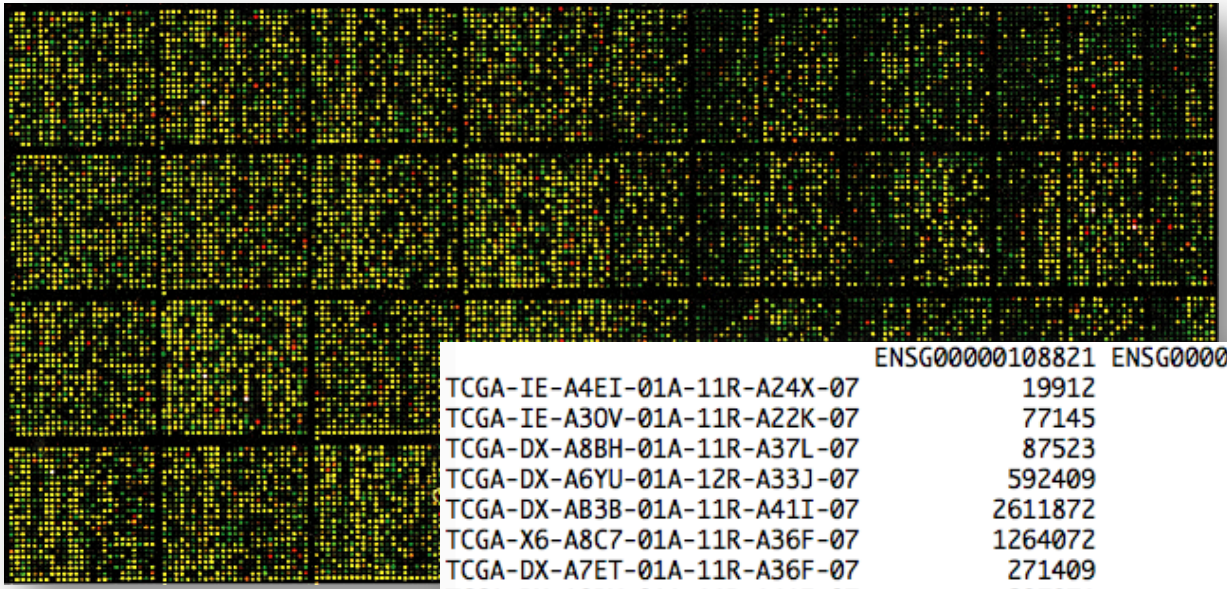


Breast - Healthy



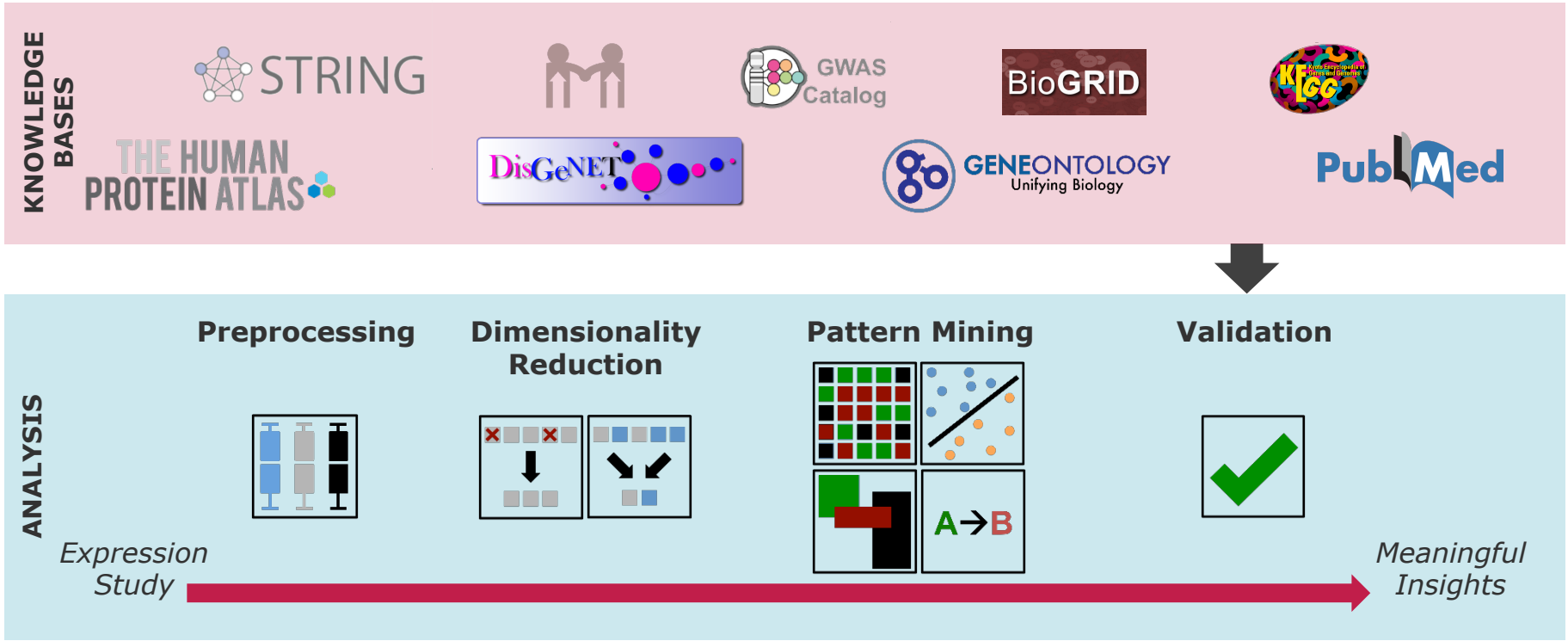
Breast - Cancer

RNAseq –
A Complete Snapshot of a Cell's Gene Activity

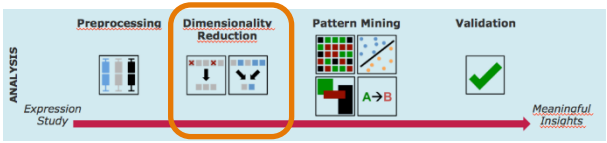


	ENSG00000108821	ENSG00000168542	ENSG00000164692	ENSG00000115414
TCGA-IE-A4EI-01A-11R-A24X-07	19912	90836	182664	138454
TCGA-IE-A3OV-01A-11R-A22K-07	77145	277426	232004	770781
TCGA-DX-A8BH-01A-11R-A37L-07	87523	51317	127552	569322
TCGA-DX-A6YU-01A-12R-A33J-07	592409	734140	284342	316996
TCGA-DX-AB3B-01A-11R-A41I-07	2611872	188255	2042859	198150
TCGA-X6-A8C7-01A-11R-A36F-07	1264072	115894	190062	162531
TCGA-DX-A7ET-01A-11R-A36F-07	271409	77395	173086	48850
TCGA-DX-A8BM-01A-11R-A41I-07	207871	387120	83256	572023
TCGA-DX-A6BE-01A-41R-A32Q-07	2675793	1114789	1038408	165573
TCGA-WK-A8XS-01A-11R-A37L-07	595470	580555	545340	174225

Integrating Biological Context into the Analysis of Gene Expression Data



A1. Integrative Gene Selection

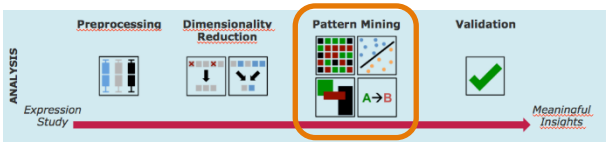


- Integrative approaches have shown to improve gene selection
  - Higher accuracy
  - Lower computational complexity

- Network-based approaches are most promising
  - Map genes to protein-protein networks or pathways
  - Identify densely coupled subnetworks

- Your task: Implement an integrative approach for gene selection
  - Review existing literature for integrative approaches for gene selection
  - Integrate approach into existing framework
  - Evaluate against existing approaches

A2. Association Rule Mining on RNAseq Data

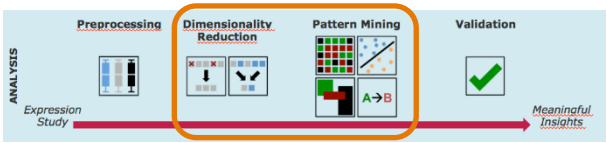


- Association rule mining can help to identify correlations between expression profiles and genes, e.g.

GeneA ↑ → *GeneB* ↑

- Your Task: Apply association rule mining on RNAseq data
  - Benchmark overall feasibility
  - Identify limitations and address one selected limitation
  - Integrate into existing framework

A3. Integrative Gene Selection vs. Integrative Clustering

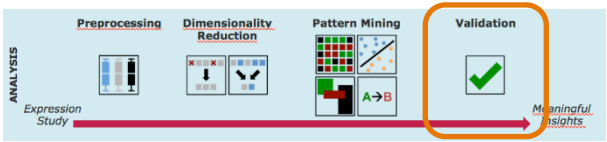


- Integrating external resources into the analysis can…
  - … reduce computational complexity
  - … deliver biologically relevant results

- External resources can be incorporated at multiple points
  - Gene selection
  - Pattern mining

- Your task: Evaluate the effect of integrating external information at different steps in the analysis pipeline
  - Integrate external information into clustering
  - Integrate approach into existing framework
  - Evaluate integrative gene selection vs. integrative clustering

A4. Biological Evaluation of Marker Genes



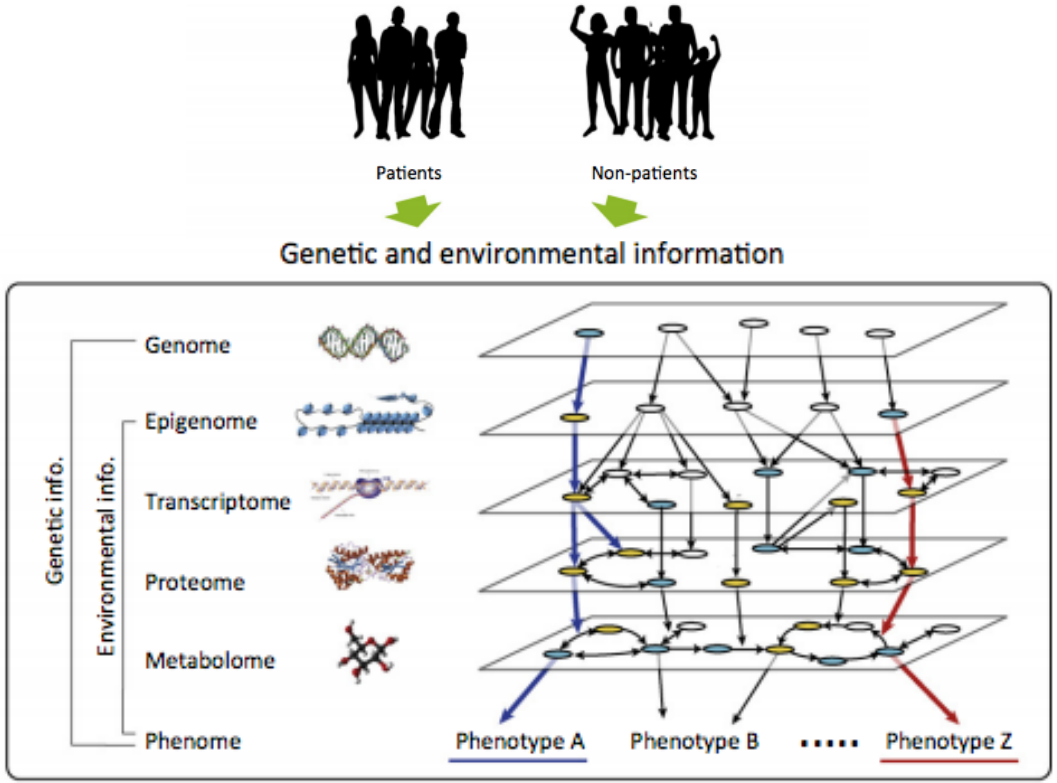
- Analysis results must be validated for their biological relevance
  - State of the art: Gene Set Enrichment Analysis (GSEA)
  - Literature review
  - Keyword search

- Your task: Implement an automatic evaluation for marker genes
  - Identify suitable resources
  - Decide on evaluation strategy, e.g. GSEA
  - Integrate approach into existing framework

TiB Seminar Kickoff

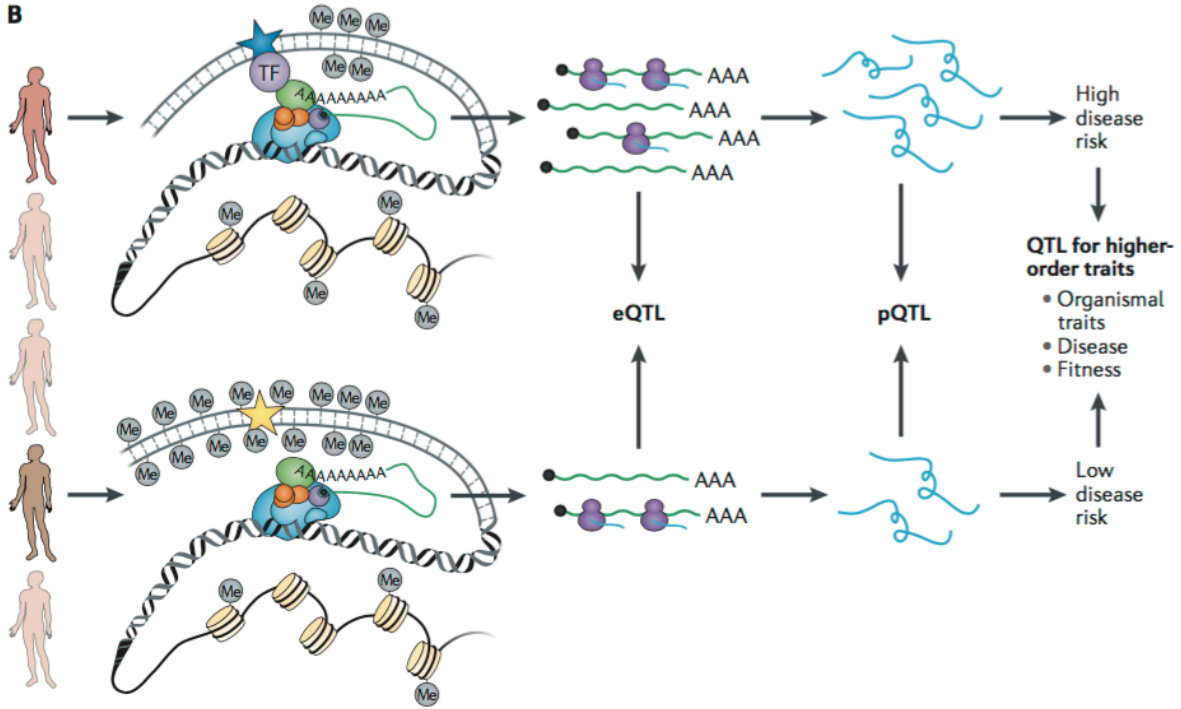
Perscheid, Kraus, Cruz

B. Multi-level Data Integration in Systems Medicine of Heart Failure



Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers, *Yugi, K.,* (2016), Cell Press

Quantitative Trait Loci (QTL)

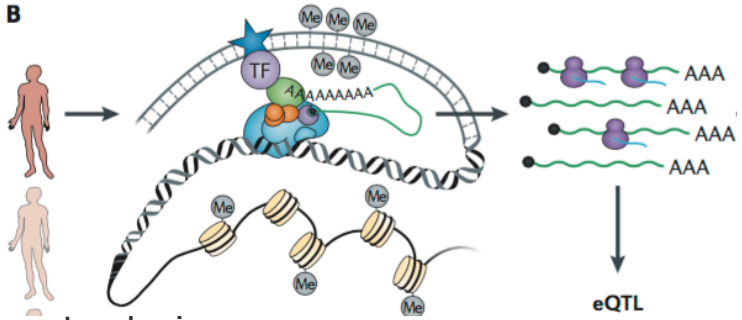


The role of regulatory variation in complex traits and disease, *Frank W. Albert1,2 and Leonid Kruglyak* (2005), Nature Reviews Genetics

TiB Seminar Kickoff

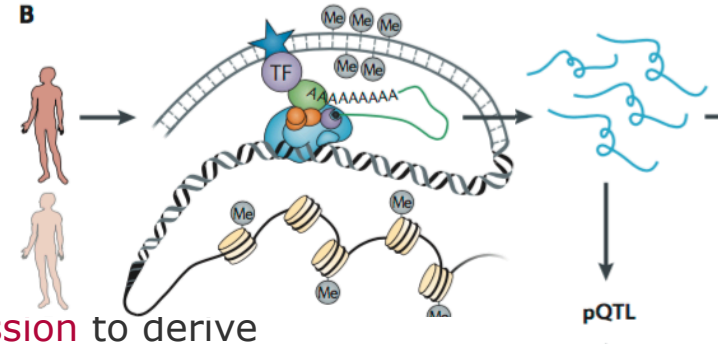
Perscheid, Kraus, Cruz

B1. Calculate and Validate eQTLs in Heart Failure



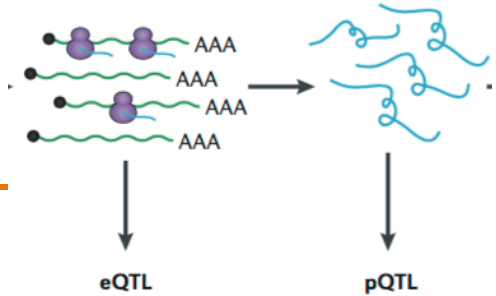
- Understand:
  - How to combine genomic variation and RNA expression to derive eQTLs
- Try out:
  - PEERs normalization for confounding variation in expression data and known confounders
  - Matrix QTL package to infer genomic regions that alter RNA expression
  - Compare found eQTLs to GWAS and know HF variants
- Write:
  - Describe the algorithms and experiments in a **scientific** paper
  - Discuss results in a technical and biological manner
  - (Optional: Compare your results with B2)

B2. Calculate and Validate pQTLs in Heart Failure



- Understand:
  - How to combine genomic variation and protein expression to derive pQTLs
- Try out:
  - PEERs normalization for confounding variation in expression data and known confounders
  - Matrix QTL to infer genomic regions that alter protein expression
  - Compare found pQTLs to GWAS and known HF variants
- Write:
  - Describe the algorithms and experiments in a **scientific** paper
  - Discuss results in a technical and biological manner
  - (Optional: Compare your results with B1)

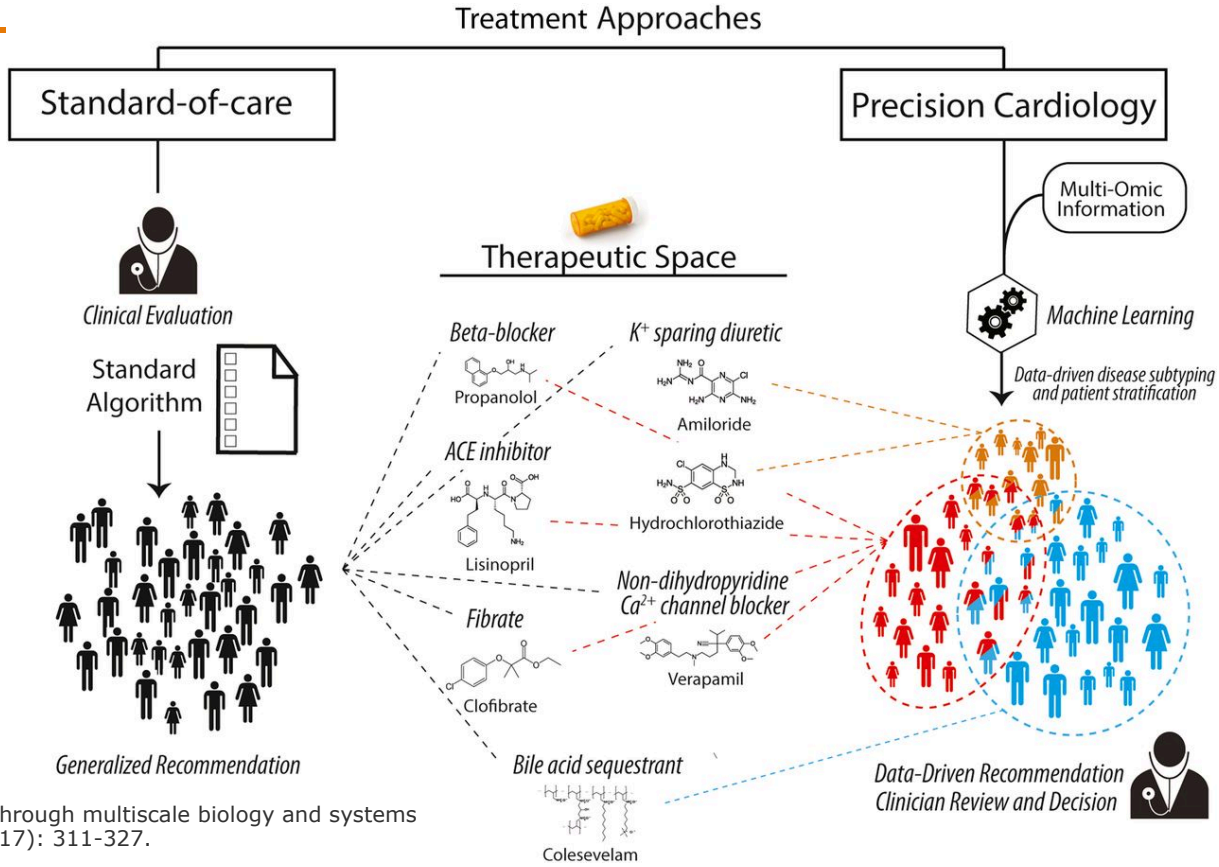
B3. Assess the Feasibility of Expressed QTLs



- Understand:
  - The impact of genomic variants in coding regions
- Try out:
  - Mapping of tissue-specific eQTLs to expressed genomic regions from GTEX
  - Matrix QTL on expressed genomic regions that alter RNA and/or protein expression for at least one GTEX tissue → expQTLs
  - Compare expQTLs and eQTLs
- Write:
  - Describe your algorithm and experiments in a **scientific** paper
  - Quantify expQTLs
  - Evaluate the feasibility to extract expQTLs from eQTL data

Multi-omics Clustering

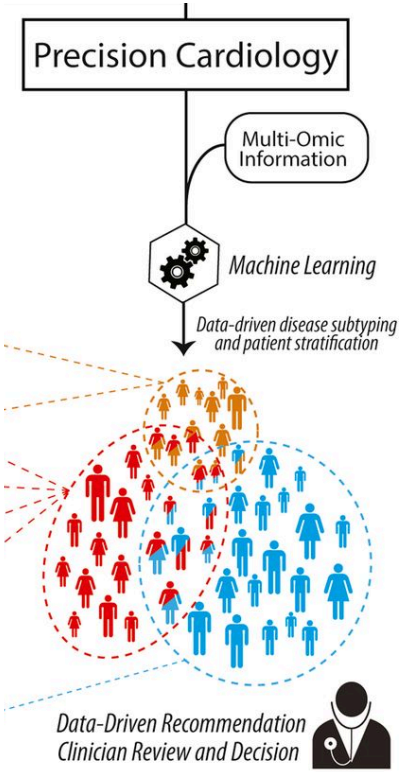
- High-throughput multi-omic information is available
- Unsupervised classification is used to classify molecular profiles on a single omic basis
- Patient subgroup detection may help to find a personalized therapy based on molecular data



Johnson, Kipp W., et al. "Enabling precision cardiology through multiscale biology and systems medicine." *JACC: Basic to Translational Science* 2.3 (2017): 311-327.

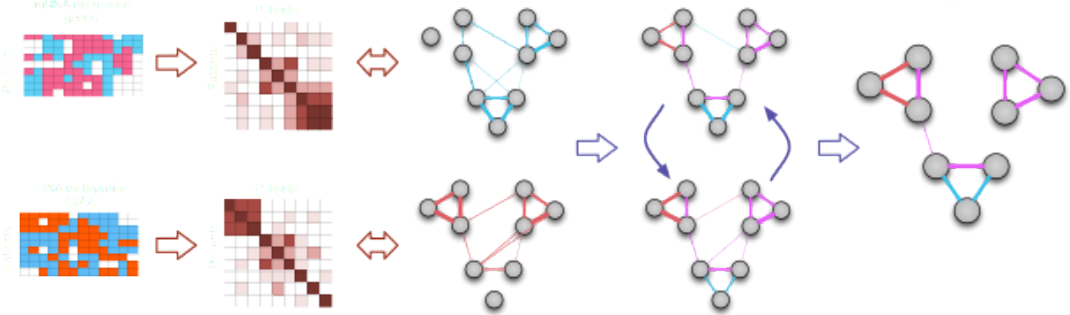
B4. Bayesian Clustering of Multi-Omics

- Understand:
  - How to perform a model-based multi-omics clustering
- Try out:
  - iClusterBayes unsupervised subgroup detection for multiple omics data sets
  - Infer molecular subgroups of heart failure patients
  - Link subgroups to clinical features (e.g., obesity or HF regression)
- Write:
  - Describe the algorithms and experiments in a **scientific** paper
  - Discuss results in a technical and biological manner



B5. Similarity Network Fusion on Multi-Omics

- Understand:
  - The (dis-) advantages of late integration for omics clustering
- Try out:
  - SNF for unsupervised subgroup detection in multiple omics data sets
  - Infer molecular subgroups of heart failure patients
  - Link subgroups to clinical features (e.g., obesity or HF regression)
- Write:
  - Describe the algorithms and experiments in a **scientific** paper



B6. Acceptance of the DEAME Application for Clinical Research

- Understand:
  - How differential gene expression analysis is implemented in our DEAME application
- Try out:
  - Create a user questionnaire
  - Conduct user interviews with our clinical partners
  - Find strengths and weaknesses of our current application
- Write:
  - Describe the user research metodology and interviews in a **scientific** paper
  - Discuss if DEAME is a valuable tool for clinical research

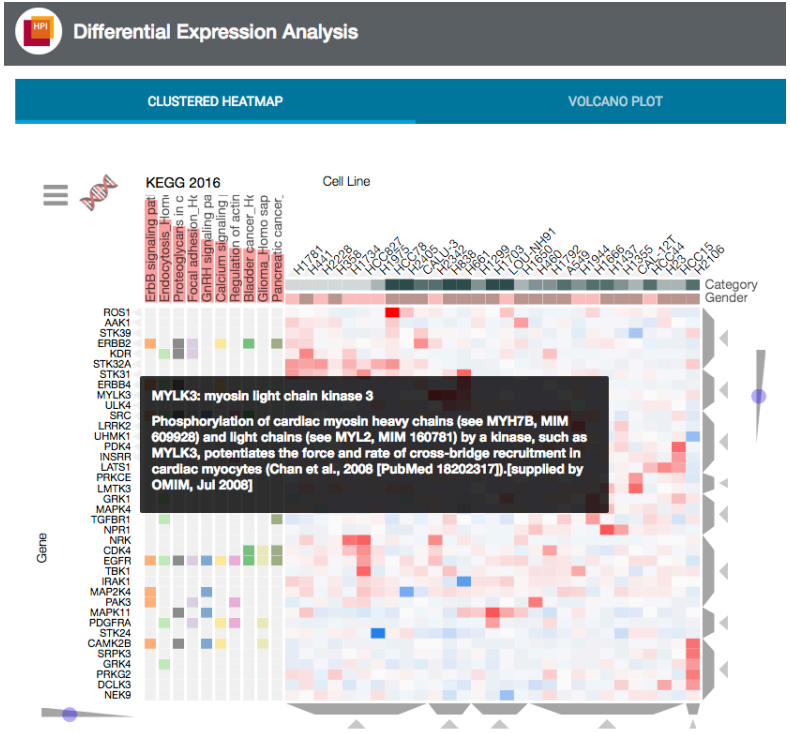


Chart **25**

C. Interpretability Approaches applied to Clinical Predictive Modeling

- Modeling of patient-level outcomes:
  - Hospital mortality
  - Length of ICU stay
  - Onset of complications
  - Disease recovery, etc.

- It can help doctors answer questions like:
  - Will patient develop disease ‘x’?
  - Should this patient be treated with ‘y’?
  - Should testing be done?
  - Is this patient likely to recover?

http://www.mii.ucla.edu/images/research/areas/clinical_decision.png

It’s Elementary, my Dear IBM Watson!

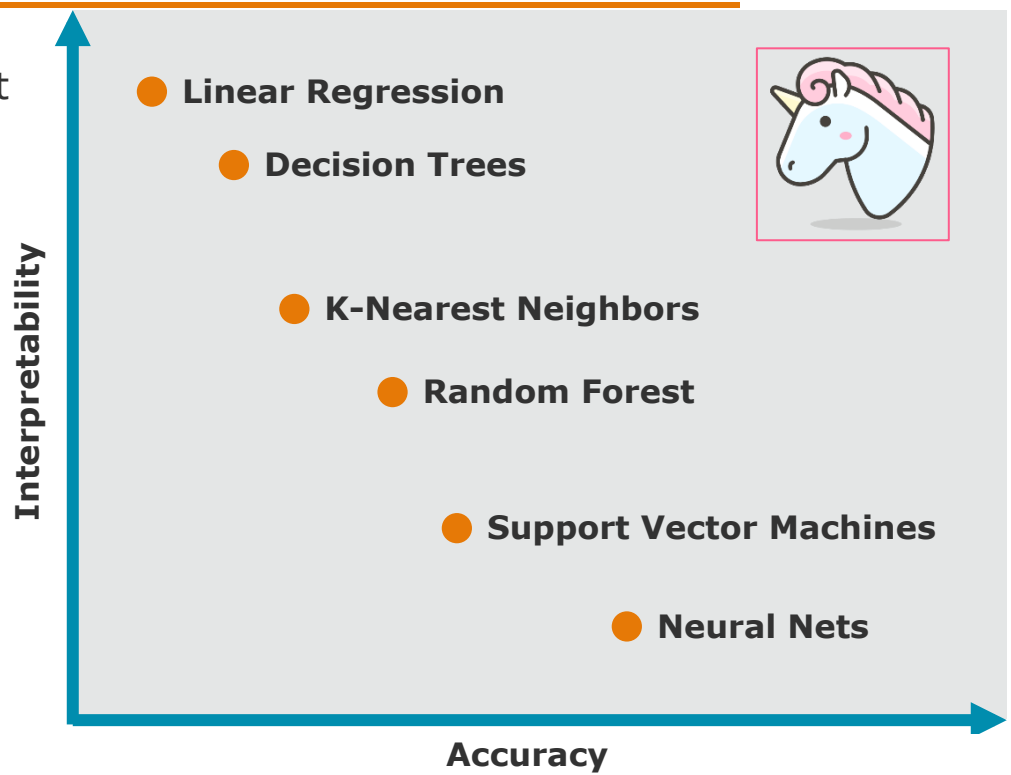
Who among you have already met Dr. Watson in ‘silico’?

IBM Watson™

Source: https://www.ibm.com/watson/

C. Interpretability Approaches applied to Clinical Predictive Modeling

- Progress in machine learning has yet to deliver on its promises
- There is often a trade-off between accuracy and model complexity
- Specially in sensitive domains such as medicine, interpretability is key
- New GDPR 2018: establishes the ”right to explanation”

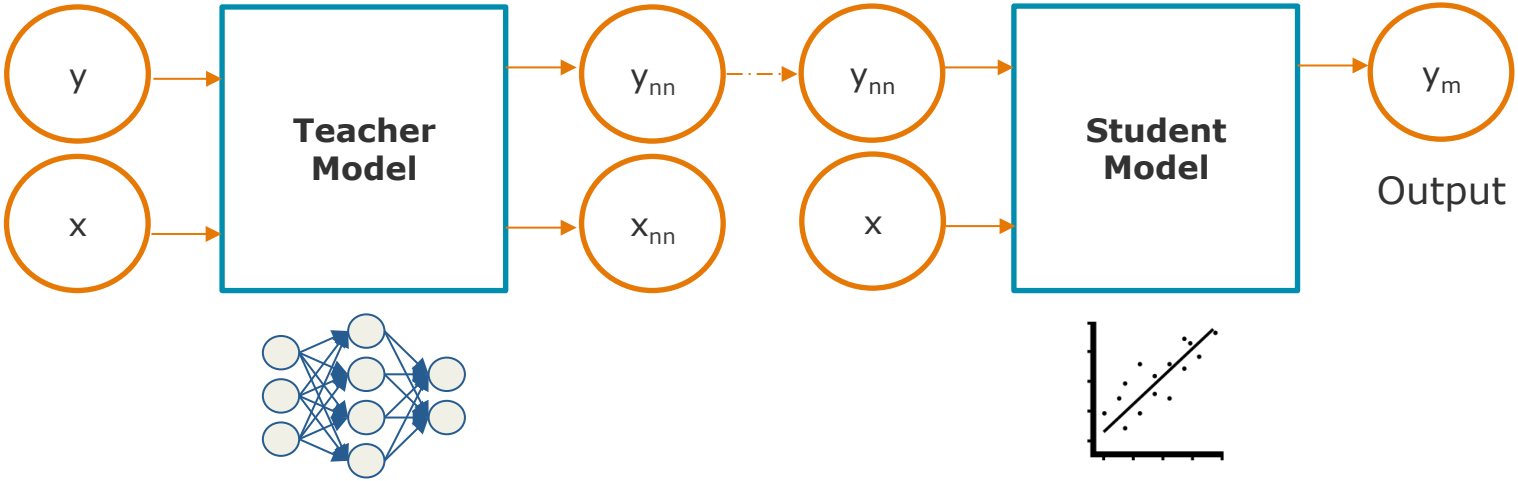


Source: https://blog.fastforwardlabs.com/2017/09/01/LIME-for-couples.html

Hall, P., & Gill, N. *An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability,Transparency, and Explainable AI*. O'Reilly M (2018). Retrieved from http://www.oreilly.com/data/free/an-introduction-to-machine-learning-interpretability.csp

C. Interpretability Approaches applied to Clinical Predictive Modeling

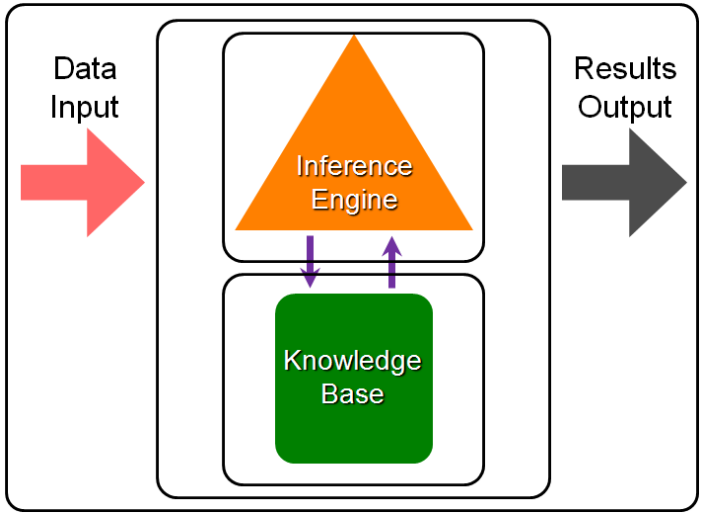
- Interpretability approaches are needed, e.g. mimic learning
- Use a complex model in combination with a more intelligible one



Zhengping Che, Sanjay Purushotham, Robinder Khemani and Yan Liu: *Interpretable Deep Models for ICU Outcome Prediction* (2017)
Doshi-Velez, F., & Kim, B. *Towards A Rigorous Science of Interpretable Machine Learning* (2017).

C. Interpretability Approaches applied to Clinical Predictive Modeling

- Your tasks:
  - Develop a clinical prediction model (CPM) together with clinical experts
  - Perform literature research on state-of-the-art interpretability approaches
  - Implement, evaluate and compare selected methods
  - Identify key areas for improvement



Source: Bonney (2011)

- The tools you will need:
  - Python + SQL
  - ML tookit scikit-learn

Wilfred Bonney (2011). Impacts and Risks of Adopting Clinical Decision Support Systems, Efficient Decision Support Systems - Practice and Challenges in Biomedical Related Domain, Prof. Chiang Jao (Ed.), ISBN: 978- 953-307-258-6, InTech

Thanks for your attention!

- Choose your favorite topics by **Wed Oct 24, 11.59 PM**
- Come by at our offices for questions:
  - V-1.19, Campus II
  - G-2.2.16, Campus III