ETH zürich

Machine Learning for Personalized Medicine

Damian Roqueiro
Machine Learning & Computational Biology Lab
D-BSSE, ETH Zürich

WORLDWEBFORUM 2017, Machine Learning

Part I. Motivation

What is machine learning?
What is personalized medicine?

Machine learning

A definition by Tom Mitchell The Discipline of Machine Learning. (2006)

…"[A] machine learns with respect to a particular task $\mathcal{T}$,
performance metric $\mathcal{P}$,
and type of experience $\mathcal{E}$,
if the system reliably improves its performance $\mathcal{P}$
at task $\mathcal{T}$,
following experience $\mathcal{E}$…"

- Recommender systems, e.g. Netflix®
- Tailoring of ads and newsfeeds in social networks, e.g. Facebook®
- Web searches and ranking of pages, e.g. Google® search
- Spam filtering of e-mails, and many others

Machine learning

A definition by Tom Mitchell The Discipline of Machine Learning. (2006)

…"[A] machine learns with respect to a particular task $\mathcal{T}$,
performance metric $\mathcal{P}$,
and type of experience $\mathcal{E}$,
if the system reliably improves its performance $\mathcal{P}$
at task $\mathcal{T}$,
following experience $\mathcal{E}$…"

- Recommender systems, e.g. Netflix®
- Tailoring of ads and newsfeeds in social networks, e.g. Facebook®
- Web searches and ranking of pages, e.g. Google® search
- Spam filtering of e-mails, and many others

Personalized Medicine

What is personalized medicine?

- Doctors recognize that every patient is unique → tailor treatment as best they can
- Important discovery: Matching a blood transfusion to a blood type
  - "...What if matching a cancer cure to our genetic code was just as easy, just as standard?"

Barack Obama. obamawhitehouse.archives.gov/precision-medicine

- Definition of personalized medicine: "...delivering the right treatments, at the right time, every time to the right person."
- Launch of the Precision Medicine Initiative (January, 2015)
  - +1,000,000 participants
  - biological samples, genetic data, lifestyle information **linked** to electronic health records

Science Magazine. doi:10.1126/science.aaa6436; "All of Us". www.nih.gov/allofus-research-program

Personalized Medicine

What is personalized medicine?

- Doctors recognize that every patient is unique → tailor treatment as best they can
- Important discovery: Matching a blood transfusion to a blood type
  - "...What if matching a cancer cure to our genetic code was just as easy, just as standard?"

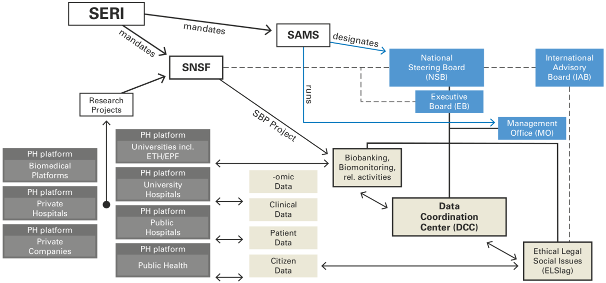
Barack Obama. obamawhitehouse.archives.gov/precision-medicine

- Definition of personalized medicine: "...delivering the right treatments, at the right time, every time to the right person."
- Launch of the Precision Medicine Initiative (January, 2015)
  - +1,000,000 participants
  - biological samples, genetic data, lifestyle information **linked** to electronic health records

Science Magazine. doi:10.1126/science.aaa6436; "All of Us". www.nih.gov/allofus-research-program

Initiatives in Switzerland

Swiss Personalized Health Network (SPHN)


SERI
mandates
mandates
SAMS
designates
SNSF
runs
National Steering Board (NSB)
International Advisory Board (IAB)
Executive Board (EB)
Management Office (MO)
Research Projects
SBP Project
PH platform Biomedical Platforms
PH platform Private Hospitals
PH platform Private Companies
PH platform Universities incl. ETH/EPF
PH platform University Hospitals
PH platform Public Hospitals
PH platform Public Health
-omic Data
Clinical Data
Patient Data
Citizen Data
Biobanking, Biomonitoring, rel. activities
Data Coordination Center (DCC)
Ethical Legal Social Issues (ELSIag)

SAMS Bulletin 01.2016. `www.samw.ch/en/Projects/Personalized-Health.html`

- Two phases:
  - 2017-2018: Funding of infrastructure
  - 2019-2020+: Funding of research projects
- Two clusters: Lausanne-Geneva and Zurich-Basel
- Major goal: Medical informatics systems in Swiss University Hospitals must be interoperable

SIB Swiss Institute of Bioinformatics

→ Establishment of standards for data production and storage

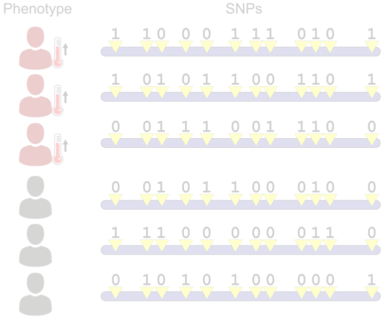
Part II. Applications

Significant pattern mining
Other ongoing projects

Significant pattern mining

Definition F. Llinares-López et al. KDD 2015

- The goal of *significant pattern mining* is to identify sets of items that occur statistically significantly more often in one class than in the other.



1 1 0 0 0 1 1 1 0 1 0 1

1 0 1 0 1 1 0 0 1 1 0 1

0 0 1 1 1 0 0 1 1 1 0 0

0 0 1 0 1 1 0 0 0 1 0 0

1 1 1 0 0 0 0 0 0 1 1 0

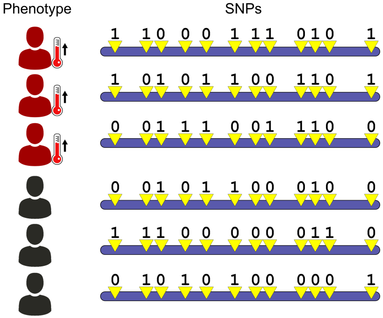
0 1 0 1 0 1 0 0 0 0 0 1

Llinares-López, et al. Bioinformatics (2015)

Significant pattern mining

Definition F. Llinares-López et al. KDD 2015

- The goal of *significant pattern mining* is to identify sets of items that occur statistically significantly more often in one class than in the other.



1 1 0 0 0 1 1 1 0 1 0 1

1 0 1 0 1 1 0 0 1 1 0 1

0 0 1 1 1 0 0 1 1 1 0 0

0 0 1 0 1 1 0 0 0 1 0 0

1 1 1 0 0 0 0 0 0 1 1 0

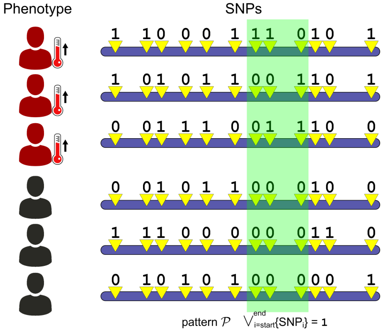
0 1 0 1 0 1 0 0 0 0 0 1

Llinares-López, et al. Bioinformatics (2015)

Significant pattern mining

Definition F. Llinares-López et al. KDD 2015

- The goal of *significant pattern mining* is to identify sets of items that occur statistically significantly more often in one class than in the other.

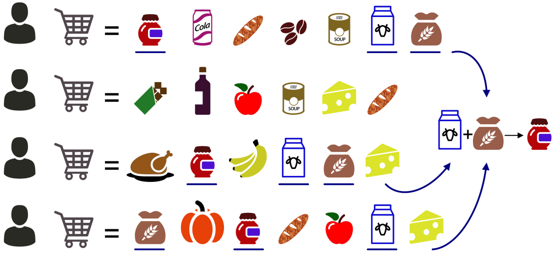

Phenotype
SNPs
pattern $\mathcal{P}$ $\bigvee_{i=\text{start}}^{\text{end}}\{\text{SNP}_i\} = 1$

1 1 0 0 0 1 1 1 0 1 0 1
1 0 1 0 1 1 0 0 1 1 0 1
0 0 1 1 1 0 0 1 1 1 0 0
0 0 1 0 1 1 0 0 0 1 0 0
1 1 1 0 0 0 0 0 0 1 1 0
0 1 0 1 0 1 0 0 0 0 0 1

Llinares-López, et al. Bioinformatics (2015)

What is not significant pattern mining

Frequent itemset mining

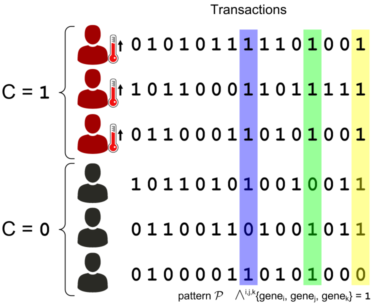


- Goal: Identify sets of products that are jointly bought by most customers
- Construct association rules of frequent itemsets

Significant pattern mining

Key aspects

	Pattern $\mathcal{P}$ is present	Pattern $\mathcal{P}$ is not present	
$C = 1$	a	$n_1 - a$	n_1
$C = 0$	$x - a$	$(n - n_1) - (x - a)$	$n - n_1$
	x	$n - x$	n



Transactions

C = 1

0 1 0 1 0 1 1 1 1 1 0 1 0 0 1

1 0 1 1 0 0 0 1 1 0 1 1 1 1 1

0 1 1 0 0 0 1 1 0 1 0 1 0 0 1

C = 0

1 0 1 1 0 1 0 1 0 0 1 0 0 1 1

0 1 1 0 0 1 1 0 1 0 0 1 0 1 1

0 1 0 0 0 0 1 1 0 1 0 1 0 0 0

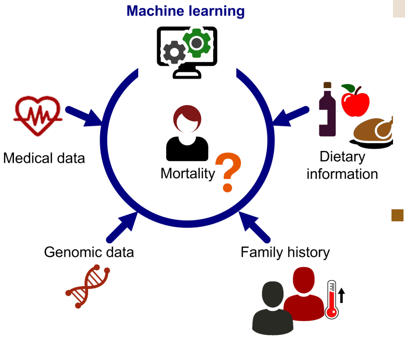
pattern $\mathcal{P}$ $\bigwedge^{i,j,k}\{\text{gene}_i, \text{gene}_j, \text{gene}_k\} = 1$

Papaxanthos et al. (2016) NIPS

- Compute p-value based on a, x, n_1 and n
  - Use Fisher's Exact Test R.A. Fisher, 1922
- Must guarantee Family-wise Error Rate $< \alpha$
- Correct for multiple hypothesis testing
  - Exclude "untestable" hypotheses R.E. Tarone, 1990

Ongoing projects

Mortality prediction through machine learning (UK Biobank)



Goal

Predict the death risk of a patient in the near/mid future based on current health status data and genotypic data

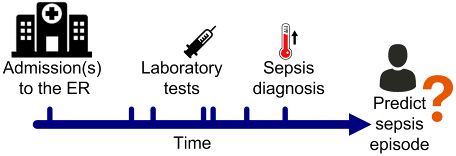
- Integration of different sources of data ($\sim 150,000$ patients)
  - Questionnaires: e.g. average weekly beer intake, happiness level, work/job satisfaction
  - Medical data: hospital admission/diagnoses
  - Genomic data: genotype calls
  - Death register: date of death, primary cause

Ongoing projects

Septic shock prediction (Universitätsspital Basel)

Goal

Predict the occurrence of septic shock based on patients' laboratory tests and previous diagnoses



- No "magic" biomarker exists to predict early stages of the disease
- Mortality can reach up to 70% → early diagnosis has direct impact on life expectancy

- Data collected for patients since 2008.
- Total of $\sim 1,800$ laboratory tests conducted in successive hospital admissions
- Patients' outcomes are known

Part III. Challenges

Ever-increasing amount of data

Need for more complex tools to analyze the data

Insufficient data sharing

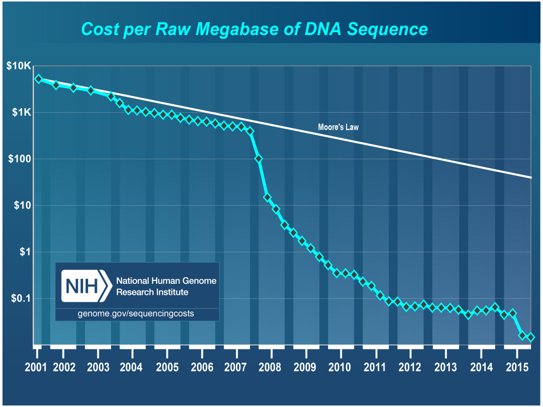
Decaying cost of sequencing/genotyping (in US$)

Cost of sequencing one human genome

- 2003: $500-1,000 million
  - Human Genome Project (estimated)
- 2006: $20-25 million
- 2016: less than $1,000

Cost of genotyping (array)

- ∼$100 per sample
  - HumanOmniExpress-24 BeadChips
  - 713,014 markers



K.A. Wetterstrand. DNA Sequencing Costs. Available at: www.genome.gov/sequencingcostsdata. Accessed January 2017.

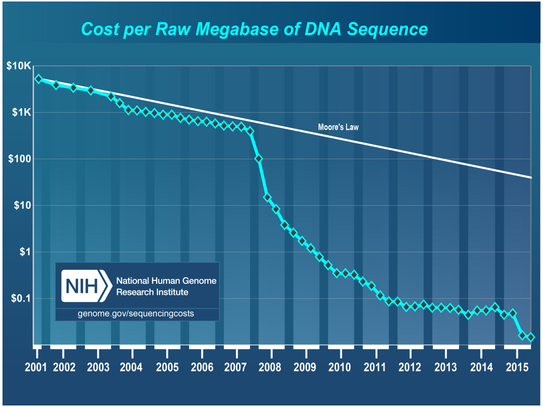
Decaying cost of sequencing/genotyping (in US$)

Cost of sequencing one human genome

- 2003: $500-1,000 million
  - Human Genome Project (estimated)
- 2006: $20-25 million
- 2016: less than $1,000

Cost of genotyping (array)

- ∼$100 per sample
  - HumanOmniExpress-24 BeadChips
  - 713,014 markers

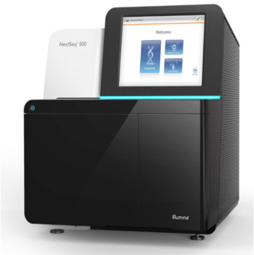


K.A. Wetterstrand. DNA Sequencing Costs. Available at: www.genome.gov/sequencingcostsdata. Accessed January 2017.

Plethora of data

High throughput technologies (to name a few...)

- Measure expression levels of mRNA, rRNA, tRNA, and other non-coding RNA
- Identify genomic locations of DNA-binding proteins
- Measure abundance of metabolites
- Determine the 3-dimensional structure of chromosome in the nucleus



Illumina NextSeq 500, datasheet

Machine learning challenge

- Design better tools for data analysis
- Integration of patients' clinical information and *omics* data
- Design models for early disease diagnoses and predictors of response to treatment

Insufficient data sharing

Editorial in the New England Journal of Medicine

- "Research parasites"
  - Not involved in the study design
  - Potentially steal from research productivity envisioned by data gatherers, or
  - Disprove conclusions of the study

Longo, D.L., Drazen, J.M. Data Sharing. N Engl J Med 2016; 374(3): 276-7

Cartoon by Pécub. SIB.

Summary

Part I - Motivation

- Personalized medicine initiatives in the US and Switzerland

Part II - Application

- Significant pattern mining.
- Application I: Intervals of consecutive point mutations with pattern → search space in $O(d^2)$
- Application II: Subsets of point mutations → search space in $O(2^d)$
- Use **testability** to discard hypotheses and prune search space

Part III - Challenges

- Large amounts of *omics* data produced at unprecedented scale. Yet, insufficient data sharing

Thank You

Machine Learning and Computational Biology Lab

- Karsten Borgwardt
- Dean Bodenham
- Lukas Folkman
- Elisabetta Ghisu
- Udo Gieraths
- Thomas Gumbsch
- Anja Gumpinger
- Xiao He
- Katharina Heinrich
- Felipe Llinares López
- Laetitia Papaxanthos
- Matteo Togninalli
- Caroline Weis

@MLCBResearch

@AGKBorgwardt

References

- Title slide: Photo by Peter Gartmann, http://susanneminder.ch/susanne-minder-bildarchiv-basel-bilder/
- Figures without acknowledgement were created by Damian Roqueiro using icons downloaded & altered from: flaticon.com (designed by Freepik)
- Slide 3: Tom Mitchell. 2006. The Discipline of Machine Learning. `http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf`
- Slide 4: Picture extracted from `https://obamawhitehouse.archives.gov/administration/president-obama`

  Text extracted from `http://obamawhitehouse.archives.gov/precision-medicine`

  Definition of precision medicine: extracted from Science Magazine. doi:10.1126/science.aaa6436

  Details about NIH Research Project "All of Us" `https://www.nih.gov/allofus-research-program`
- Slide 5: Swiss Academy of Medical Sciences, Bulletin 01.2016. `http://www.samw.ch/en/Projects/Personalized-Health.html`
- Slide 7: Definition of significant pattern mining: F. Llinares-López, M. Sugiyama, L. Papaxanthos and K. Borgwardt. (2015). Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). ACM, New York, NY, USA, 725–734

  Reference to interval search: F. Llinares-López, D.G. Grimm, D.A. Bodenham, U. Gieraths, M. Sugiyama, B. Rowan, and K. Borgwardt. (2015). Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. Bioinformatics, 31(12), i240–i249

References

- Slide 8: Association rules: R. Agrawal, T. Imielienski and A. Swami. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (SIGMOD '93), Peter Buneman and Sushil Jajodia (Eds.). ACM, New York, NY, USA, 207–216.
- Slide 9: R.A. Fisher. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85 (1): 87–94

  R.E. Tarone.(1990). A modified Bonferroni method for discrete data. Biometrics 46, 515

  L. Papaxanthos, F. Llinares-López, D. Bodenham and K. Borgwardt. (2016). Finding significant combinations of features in the presence of categorical covariates. Advances in Neural Information Processing Systems 29. 2271–2279
- Slide 13: K.A. Wetterstrand. DNA Sequencing Costs. Available at: https://www.genome.gov/sequencingcostsdata/. Accessed January 2017.
- Slide 14: Image of Illumina NextSeq System 500 extracted from datasheet: https://www.illumina.com/systems/nextseq-sequencer.html
- Slide 15: D.L. Longo and J.M. Drazen. (2016). Data Sharing. N Engl J Med, 374:276–277

  Cartoon created by Pécub. Property of the Swiss Institute of Bioinformatics.