ETH zürich

From Big Data to Molecular Insight and Improved Medicine

Gunnar Rätsch

Biomedical Informatics Group

@gxr @gxrlab

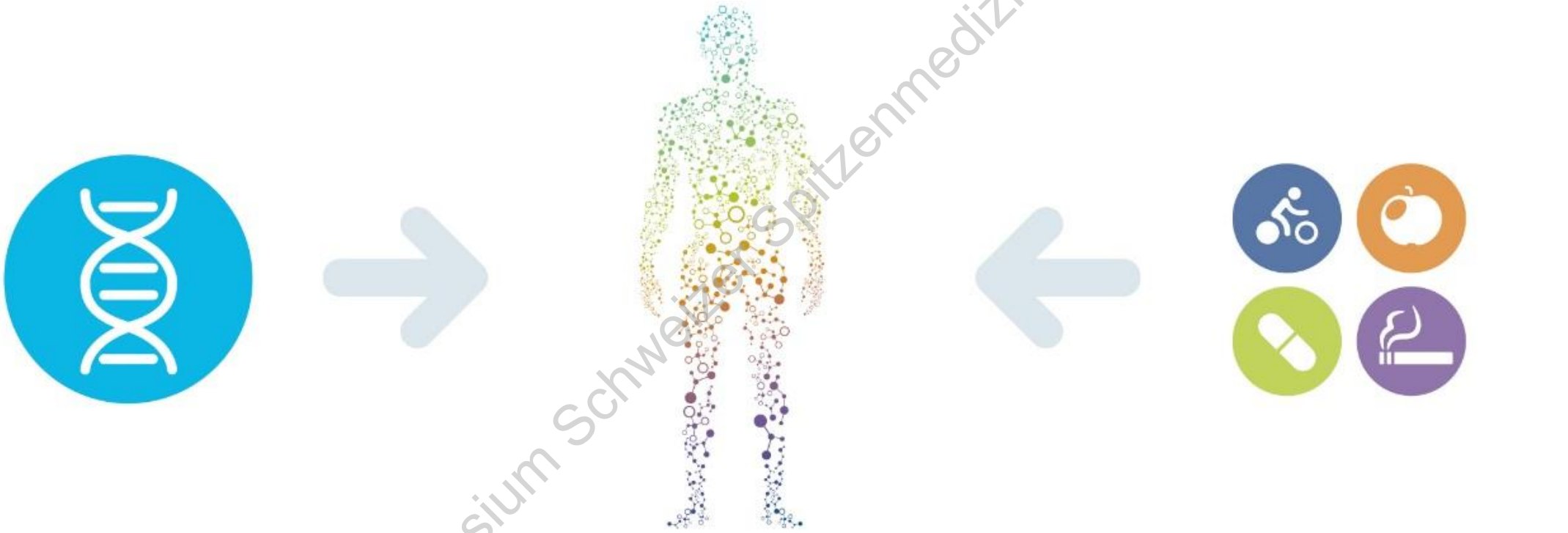
#DataScience #PrecisionMedicine #ClinicalData #Genomics #Cancer #ICU #SPHN

BIOMEDICAL INFORMATICS D INFK D BIOL UniversityHospital Zurich Memorial Sloan-Kettering Cancer Center

Imprecision Medicine

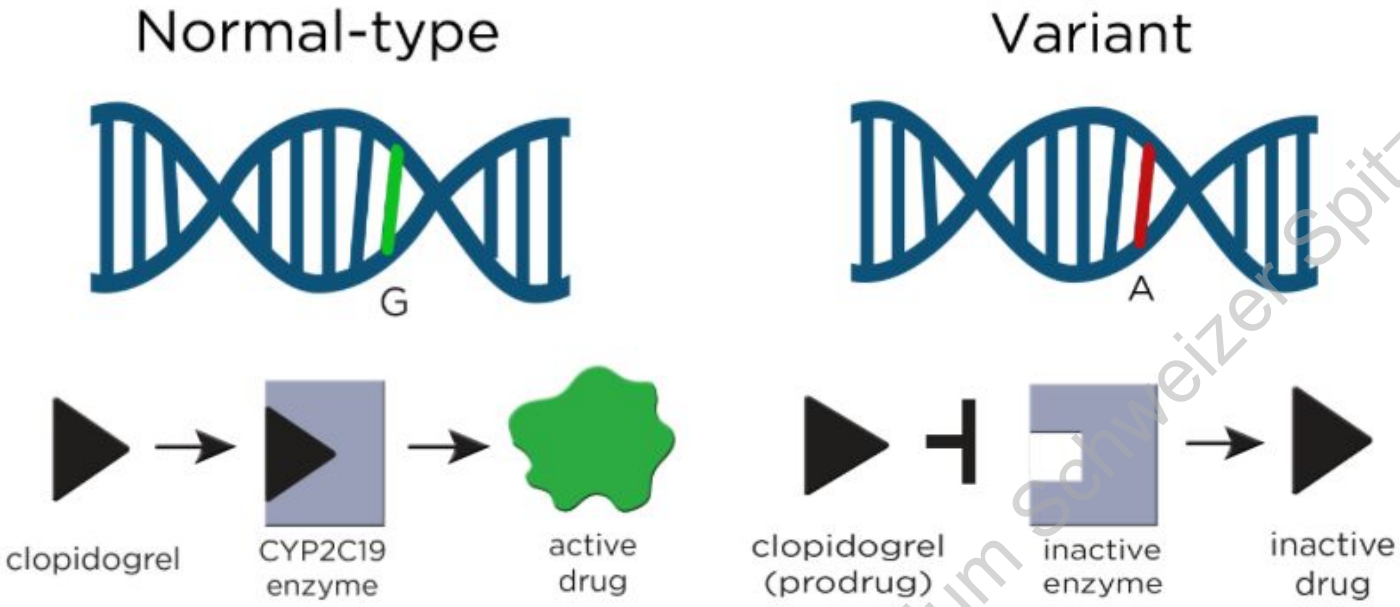
Source: Schork, Nature, 2015

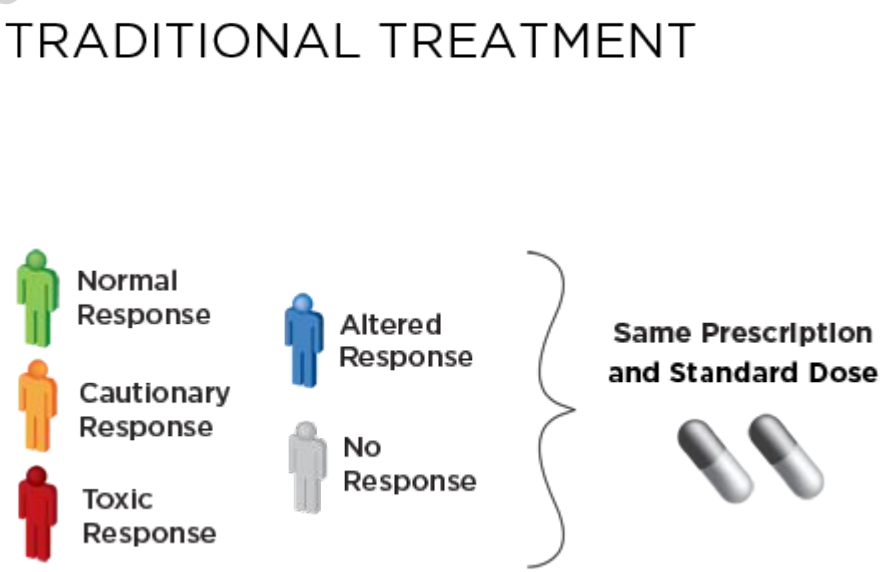
Phenotypes Depend on Molecular Traits, Lifestyle and Environment



Source: Beger et al., Metabolomics, 2016

Example: Pharmacogenomics





Source: Admerahealth.com

Example: Rare Diseases (1)

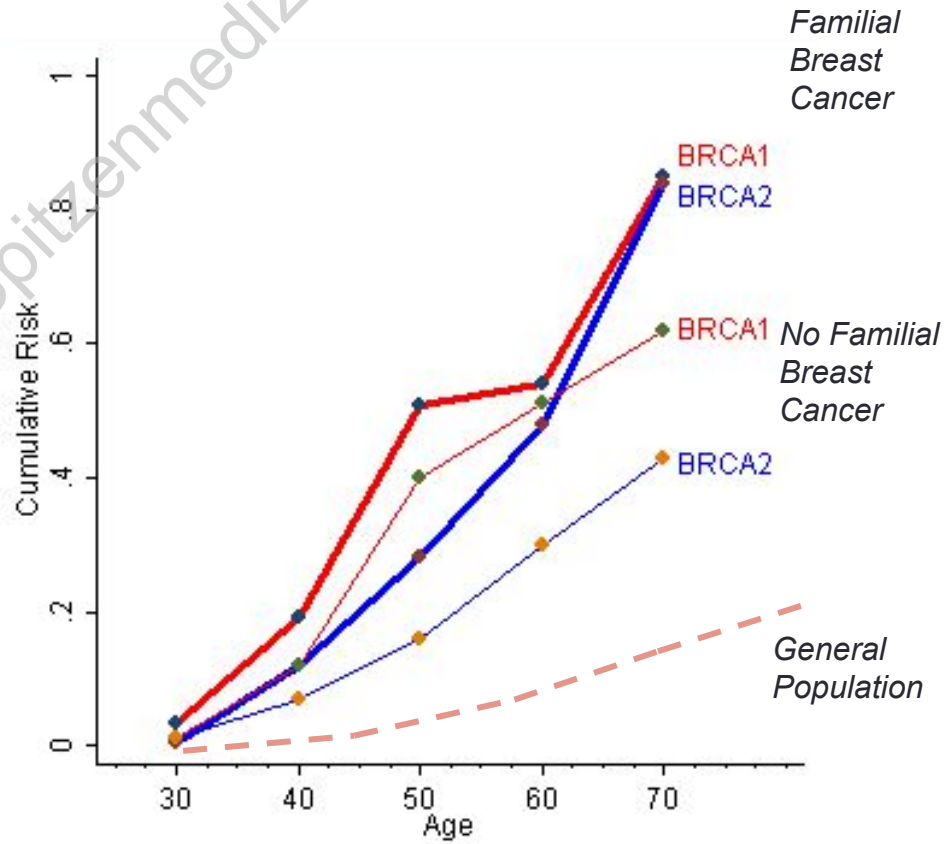
- 20-month-old girl with **rare neurodegenerative disease**
  - Abnormal gait
  - Arm weakness
  - Vision problems
  - Excessive drooling
- Initially misdiagnosed
- Exome sequencing led to correct diagnosis: *Brown-Vialetto-Van Laere Syndrome 2*
  - Cause: Defective cellular vitamin B transport
  - Treatment: Vitamin B supplements

Source: Petrovsky et al., Molecular Case Studies, 2015
News Blog, Columbia University, 2015

Example: Rare Diseases (2)

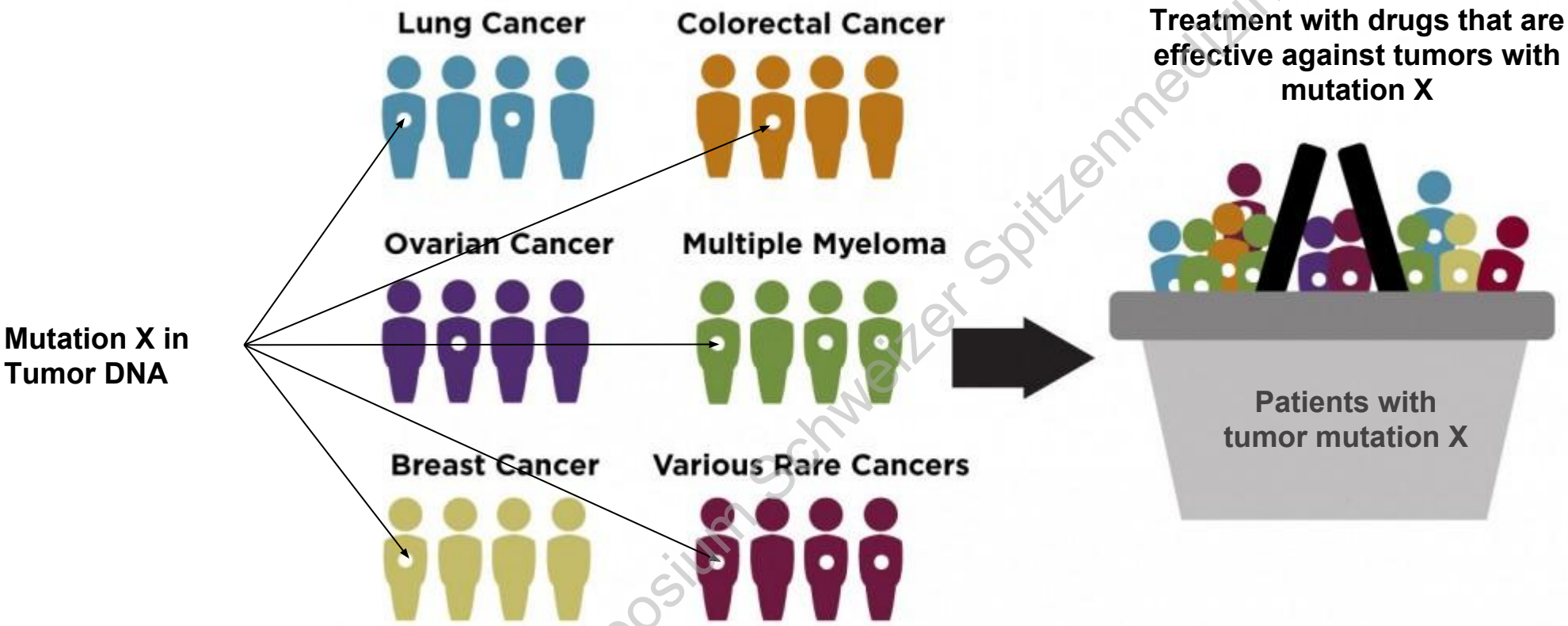
Pathogenic variants in BRCA1/2 have well known medical implications

- Increased lifetime risk of developing breast or ovarian with pathogenic BRCA mutation
- Men with pathogenic BRCA mutations are also at risk for prostate cancer
- Drug treatment: PARP inhibitors show effectiveness for BRCA1/2 patients



Source: Petrovsky et al., Molecular Case Studies, 2015
News Blog, Columbia University, 2015

Example: Genetics Testing for Cancer Treatments & Basket Trials



Source: Stallard et al., MSKCC Blog, 2015

Research with Patient Data

Need & Urgency

Obtain new insights

- Which medication against which mutation in cancer
- Medications for rare diseases

⇨ Often needs data of **thousands of patients** to identify commonalities and statistical relationships

General Consent, Privacy, Ethics

- General and research consents
- Privacy of patients
- Ethics approvals

What Computer Science can contribute

- Secure data infrastructure
- Deidentification of patient data for research
- Encryption methods, computing methods with encrypted data
- **Methodology for data analytics**

Need for Large-scale Clinical Genomics

Optimistic estimates predict that by 2022, internationally,

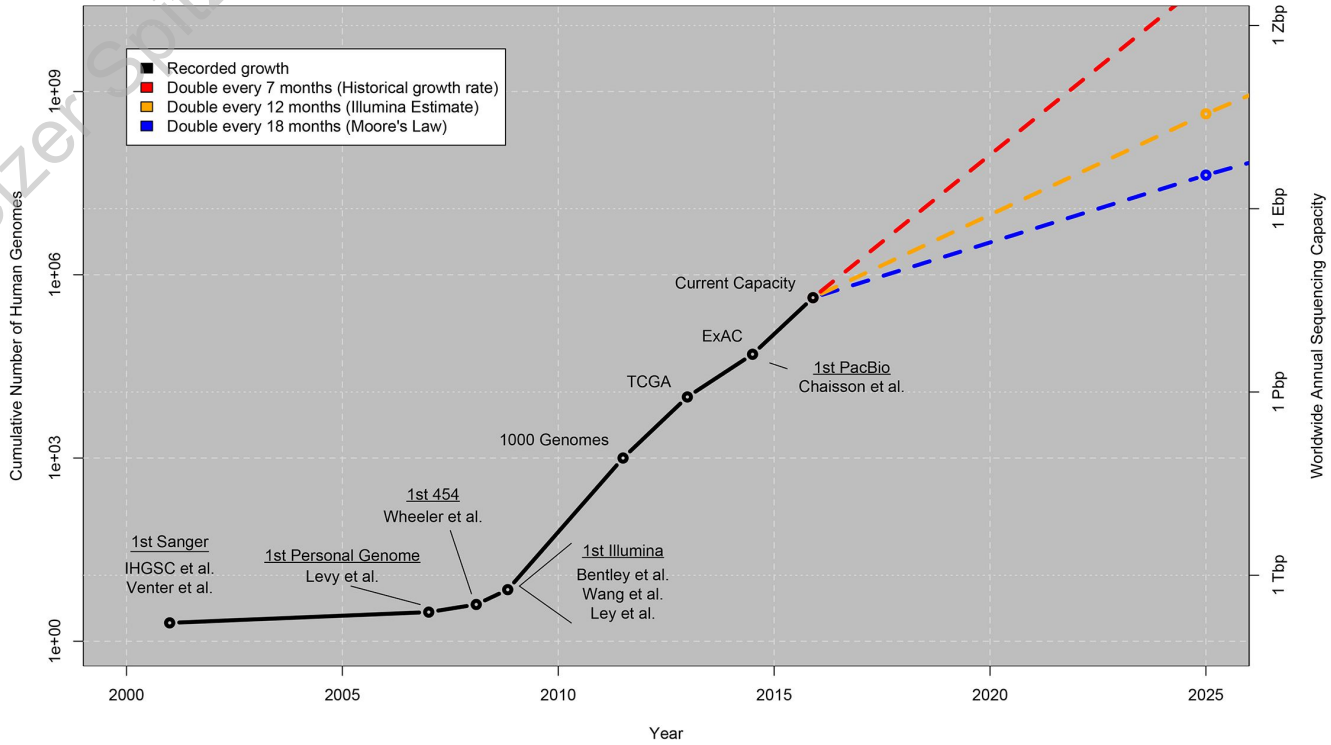
- ≈20 mid- to large-scale healthcare systems will sequence all individuals with rare diseases or cancer
- **≈15% of developed world population will be sequenced**

15% of about 1 billion people: **150M people**

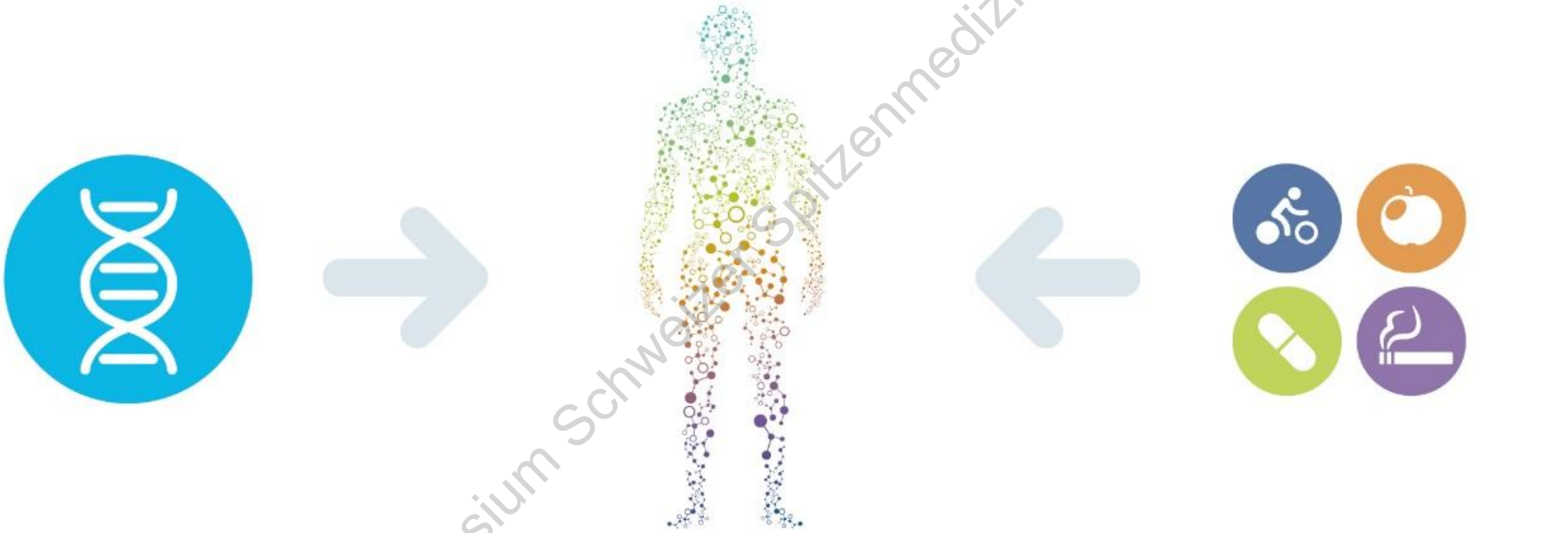
≈28 Exabyte of raw sequence data (≈Google-scale)

1 Exabyte ≈ 1'000 Petabyte ≈ 1'000'000 Terabyte

Source: Ewan Birney (European Bioinformatics Instit



Phenotypes Depend on Molecular Traits, Lifestyle and Environment

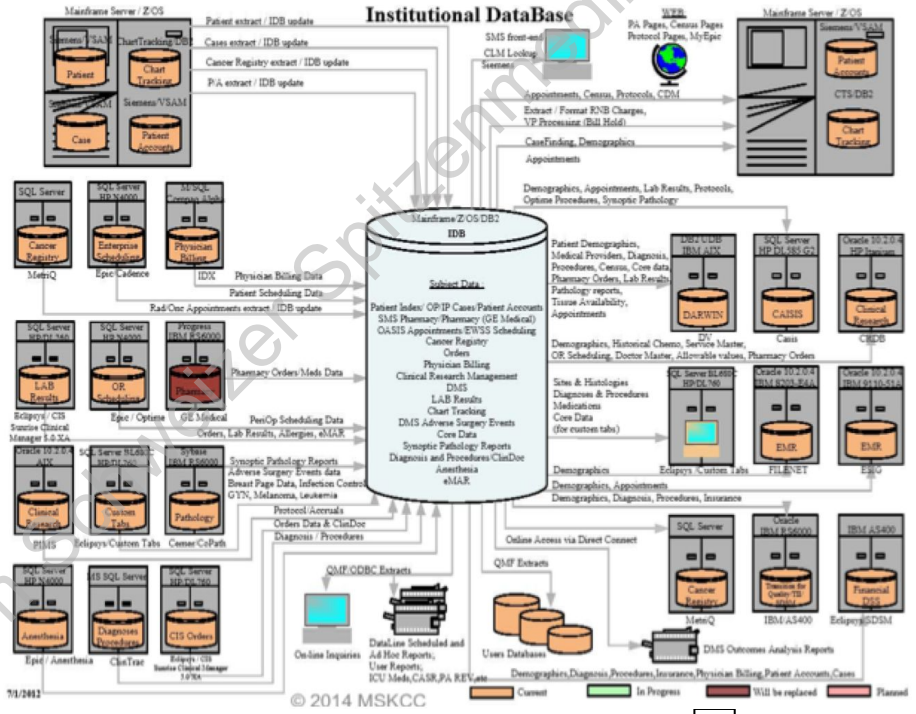


Source: Beger et al., Metabolomics, 2016

Patient Data Yesterday, Today, Tomorrow

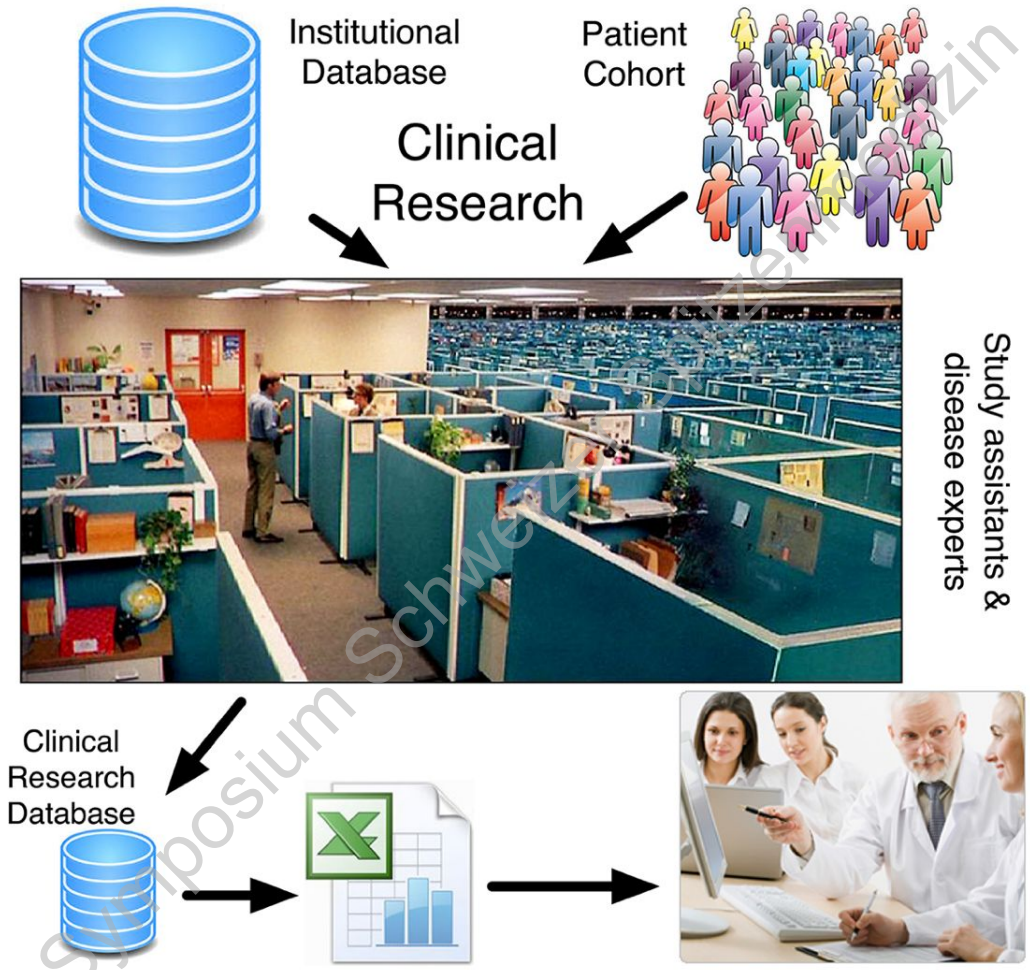
Data yesterday

Data today



"Data" tomorrow

Clinical Studies Today are Needed yet Inefficient

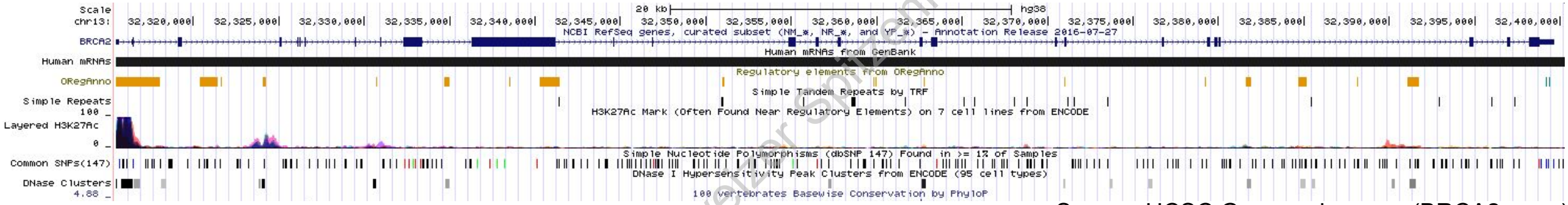


Source: Center: Tron, 1982, rest: Google images

Data Science Challenges at Medical Centers

- Efficient search and information retrieval
- Exploration of complex data by visualization

Source: Google.ch



Source: UCSC Genome browser (BRCA2 gene)

- High performance data access and computing
  - Efficient data structures & scalable computing
  - Advanced computational models
- Data Science training for medical personnel

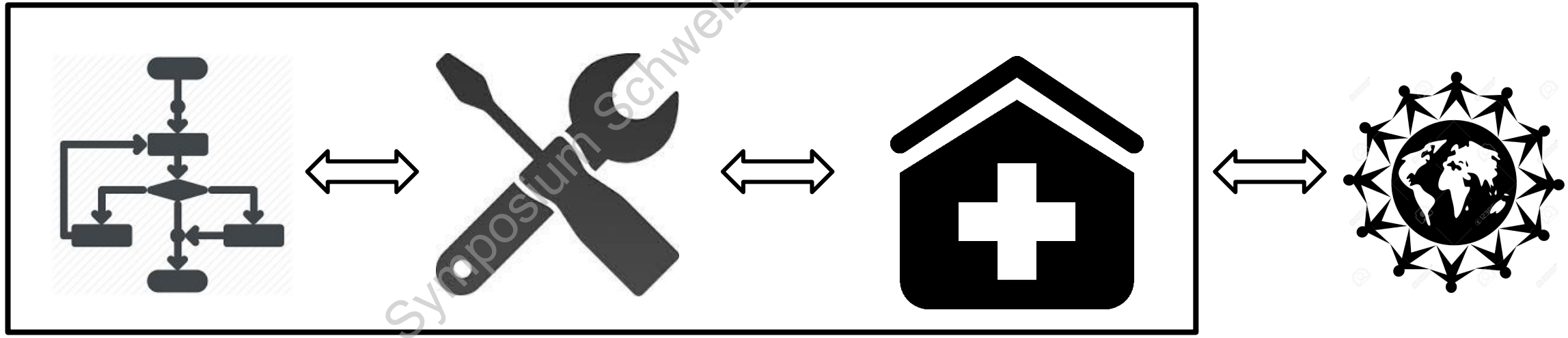
Data Science Research Challenges

Challenge 1: Develop novel data science approaches for medical data

Challenge 2: Provide tools for the community

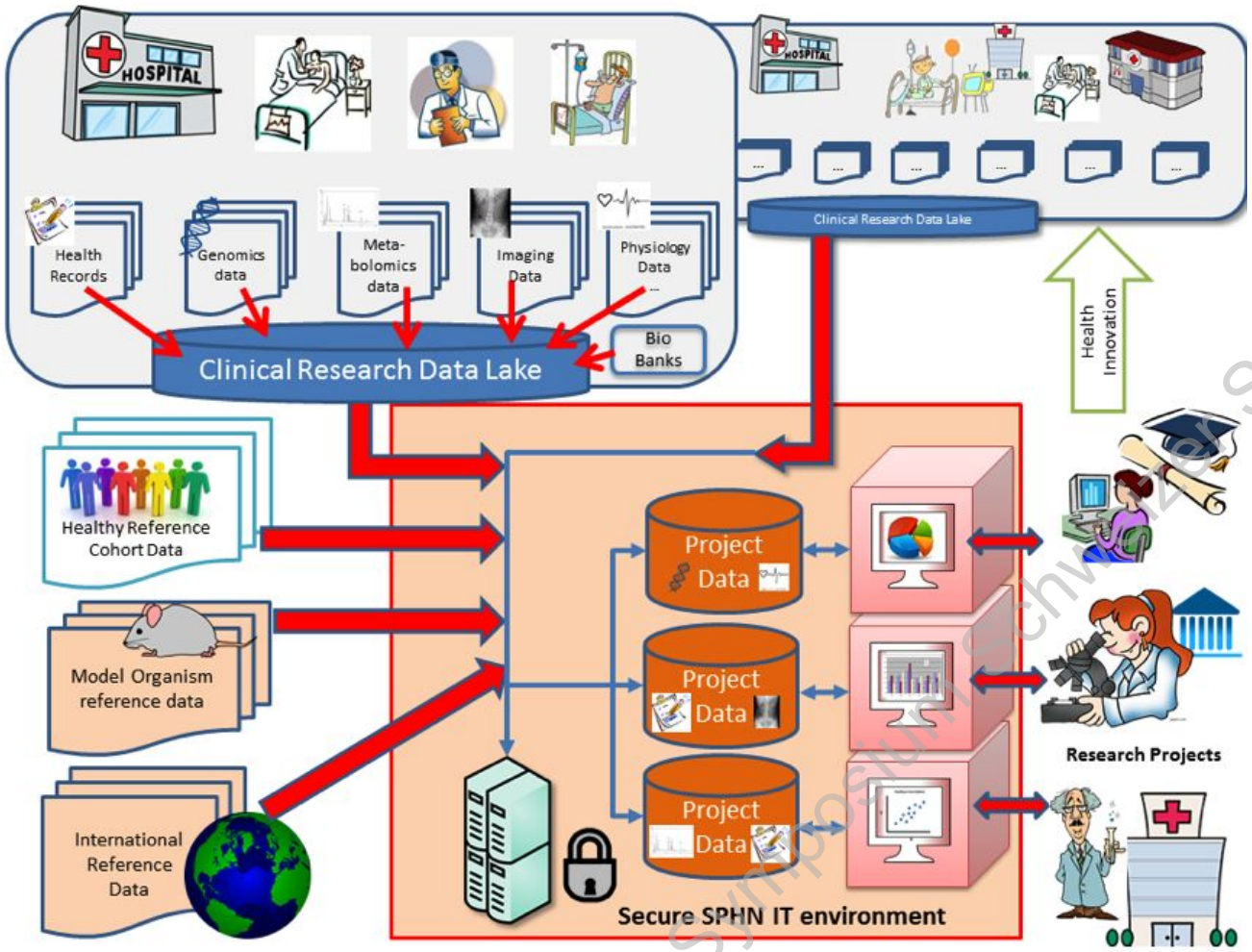
Challenge 3: Solve important biomedical problems through collaborations

Challenge 4: Create an environment which allows us to work on the above

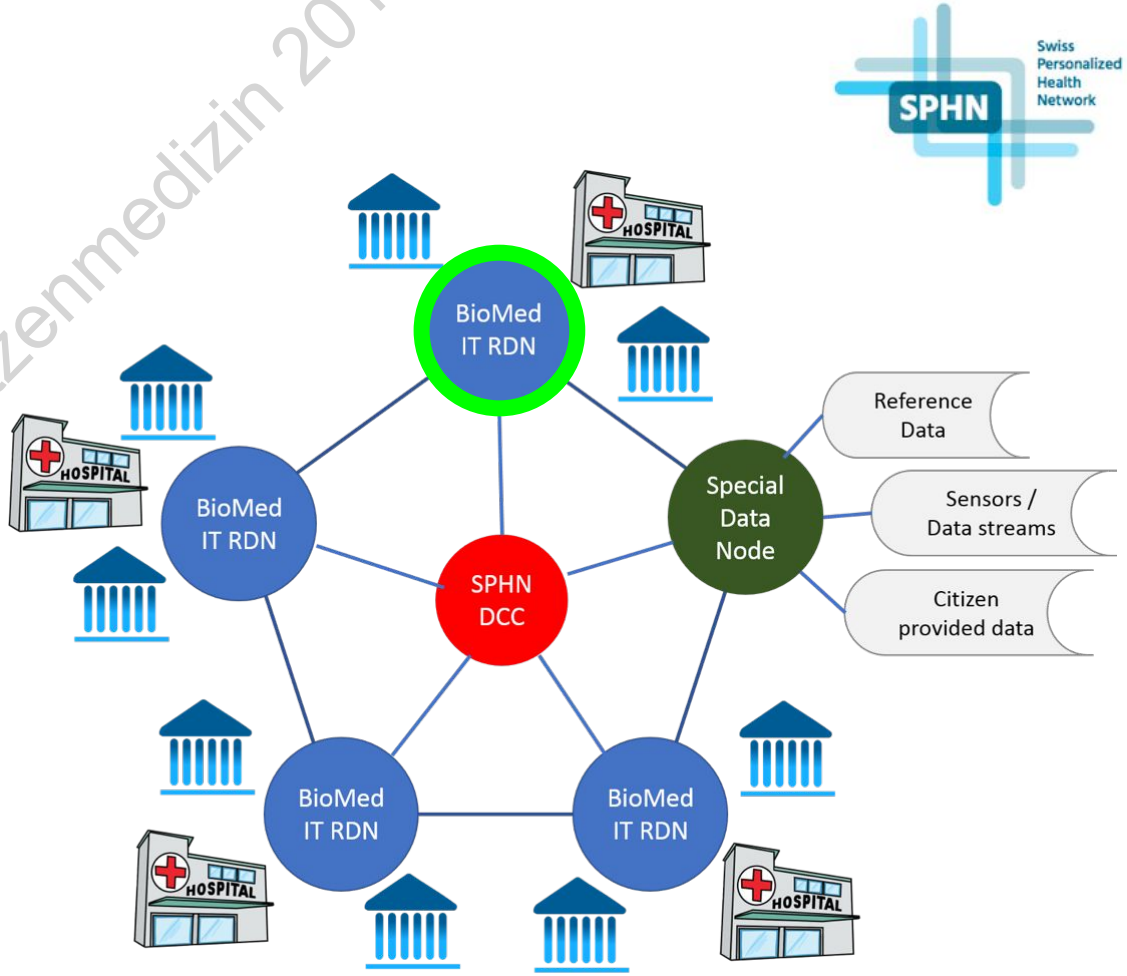


Source: Center: Google icons search

Effort 1: Swiss Personalized Health Network & Data Coordination Center



Source: Courtesy of Torsten Schwede



Source: Report of SPHN Data Expert Group, March 2017

ETH zürich

Effort 2: Research Collaboration Between ETH & University Hospital Bern

PD Dr. med. Tobias Merz

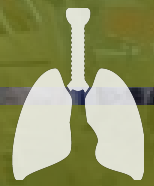
ETH zürich

Early Warning System for Intensive Care Patients

organ function parameters circulation

treatment parameters circulation

organ function parameters pulmonary

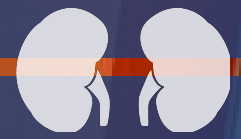


treatment parameters pulmonary

organ function parameters neuro

treatment parameters neuro

organ function parameters renal

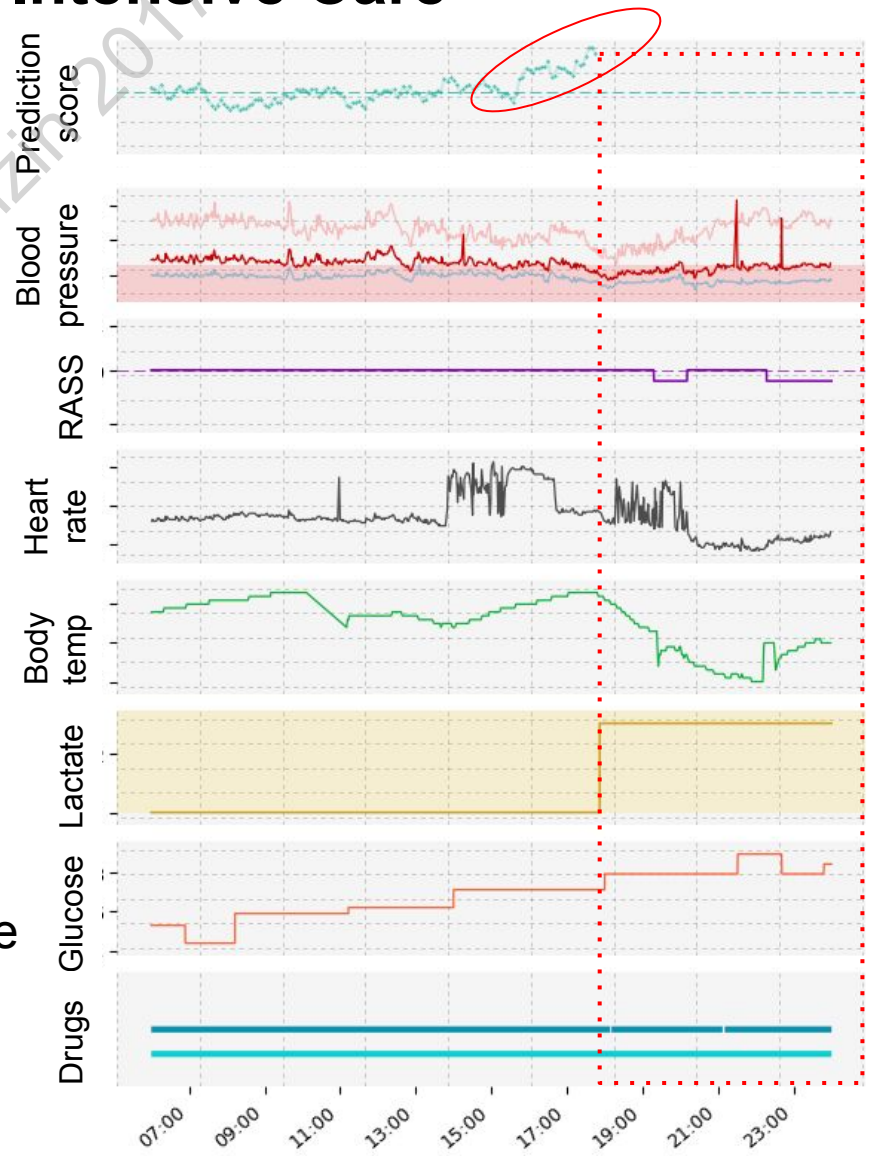


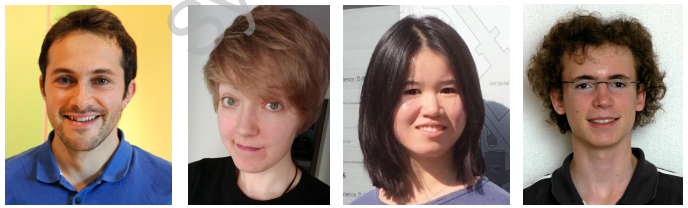
treatment parameters renal

treatment parameters infection

Goal: Early Detection of Organ Systems Failure in Intensive Care

- Research collaboration started fall 2016, obtained data in January
  - 2 months discussions, 2 months legal agreement, <2 months analysis
- Data from ≈54'000 patients
  - 189 vital signs and lab test values
  - 267 medication event values
- **500GB of raw data** (3.5 billion measurements)
- Various medical endpoints, e.g.:
  - circulatory shock (see right),
  - renal failure,
  - respiratory system failure
- Can detect 80% of circulatory shocks 4 hours in advance*
- Top features: glucose level, blood pressure, lactate levels, RAS score





BIOMEDICAL INFORMATICS

*20% FPR

Effort 3: Joint Analysis of Cancer Clinical Notes and Somatic Mutations

CHIEF COMPLAINT: Ms. NAME is a AGE-year-old woman who presents with newly diagnosed stage IV metastatic non-small cell lung carcinoma here for further treatment options.

HISTORY OF PRESENT ILLNESS: Here today for evaluation. She developed dyspnea and was found to have a right sided pleural effusion on chest x-ray. Thoracentesis cytology was indicative of malignant cells consistent with adenocarcinoma.

She underwent a CT scan of the chest that demonstrated a left lower lung nodule measuring 1.2cm. A CT scan of the abdomen and pelvis was negative in detail.

PAST MEDICAL HISTORY: Hypertension, kidney stones. Breast lump removed DATE, hysterectomy DATE, cesarean section DATE. Right leg surgery after an accident. Hyperlipidemia.

SOCIAL HISTORY: No history of alcohol or tobacco use. Patient lives alone in Manhattan. She has two adult children who live nearby. She works at a law firm.

FAMILY HISTORY: No family history in first-degree relatives. History of esophageal cancer in aunt, melanoma in uncle. Father died of heart attack at AGE.

× 2,000,000

× 200,000

Source: MSKCC Clinical Notes (2004-2014)

Reduction to Sentences

CHIEF COMPLAINT: Ms. NAME is a AGE-year-old woman who presents with newly diagnosed stage IV metastatic non-small cell lung carcinoma here for further treatment options.

HISTORY OF PRESENT ILLNESS: Here today for evaluation. She developed dyspnea and was found to have a right sided pleural effusion on chest x-ray. Thoracentesis cytology was indicative of malignant cells consistent with adenocarcinoma.

She underwent a CT scan of the chest that demonstrated a left lower lung nodule measuring 1.2cm. A CT scan of the abdomen and pelvis was negative in detail.

PAST MEDICAL HISTORY: Hypertension, kidney stones. Breast lump removed DATE, hysterectomy DATE, cesarean section DATE. Right leg surgery after an accident. Hyperlipidemia.

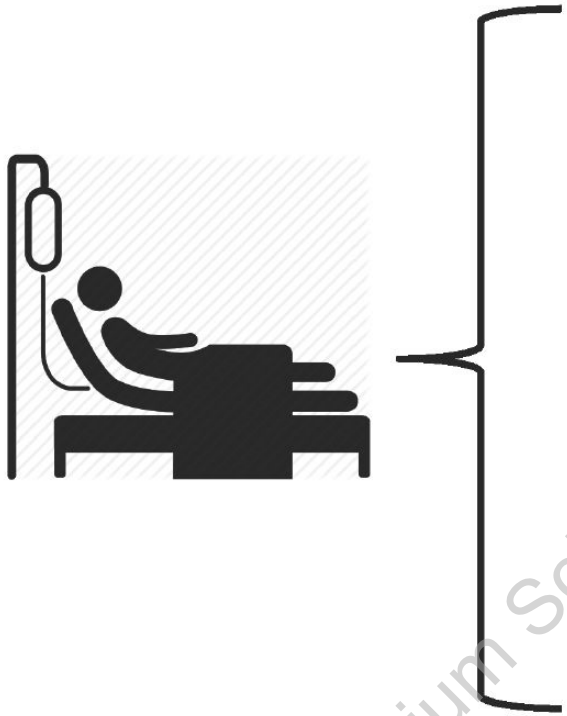
SOCIAL HISTORY: No history of alcohol or tobacco use. Patient lives alone in Manhattan. She has two adult children who live nearby. She works at a law firm.

FAMILY HISTORY: No family history in first-degree relatives. History of esophageal cancer in aunt, melanoma in uncle. Father died of heart attack at AGE.

100,000,000 sentences

Source: MSKCC Clinical Notes (2004-2014)

Reduction to Sentences



given peripheral nature pulmonary nodules feel bronchoscopy sufficiently sensitive affect management

feel bronchoscopy sufficiently sensitive affect management regardless cause

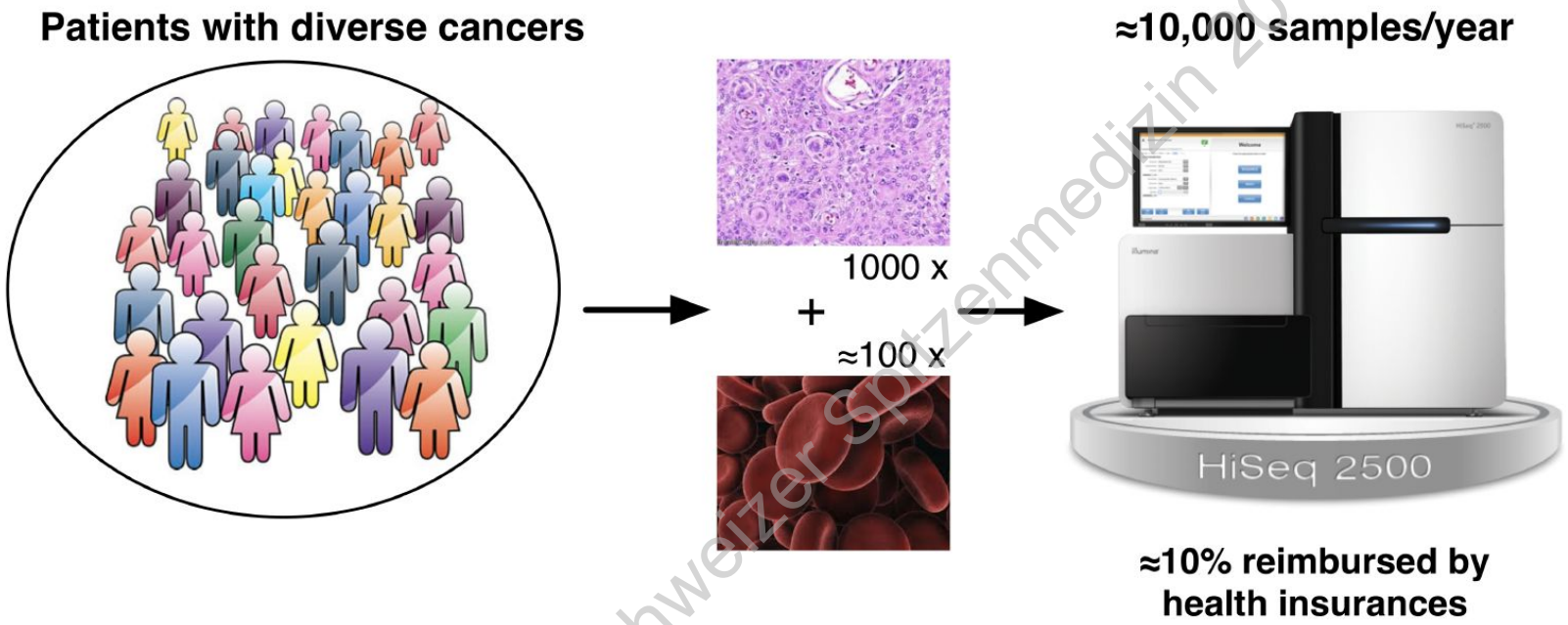
opinion changes distal render bronchoscopy sufficiently sensitive alter management

unfortunately feel bronchoscopy setting sensitive enough alter management regardless nodules cause

largest nodule opinion peripheral render bronchoscopy sufficiently sensitive alter management.

Similar sentences have similar words -> cluster sentences
Patient can be summarized as set of sentence clusters

Routine Molecular Diagnostics of Tumors at MSKCC

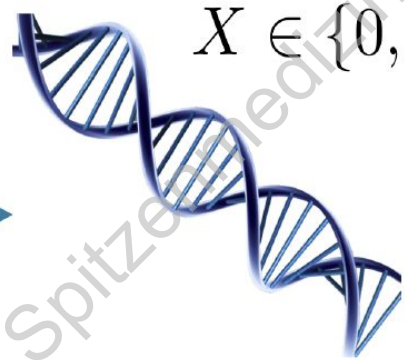


MSK IMPACT Panel (342 genes)

ABL1 AKT1 AKT2 AKT3 ALK ALOX12B APC AR ARAF ARID1A ARID1B ARID2 ARID5B ASXL1 ASXL2 ATM ATR ATRX AURKA AURKB AXIN1 AXIN2 AXL B2M BAP1 BARD1 BBC3 BCL2 BCL2L1 BCL2L11 BCL6 BCOR BLM BMPR1A BRAF BRCA1 BRCA2 BRD4 BRIP1 BTK CARD11 CASP8 CBFB CBL CCND1 CCND2 CCND3 CCNE1 CD274 CD276 CD79B CDC73 CDH1 CDK12 CDK4 CDK6 CDK8 CDKN1A CDKN1B CDKN2A CDKN2B CDKN2C CHEK1 CHEK2 CIC CREBBP CRKL CRLF2 CSF1R CTCF CTLA4 CTNNB1 CUL3 DAXX DCUN1D1 DDR2 DICER1 DIS3 DNMT1 DNMT3A DNMT3B DOT1L E2F3 EED EGFL7 EGFR EIF1AX EP300 EPCAM EPHA3 EPHA5 EPHB1 ERBB2 ERBB3 ERBB4 ERCC2 ERCC3 ERCC4 ERCC5 ERG ESR1 ETV1 ETV6 EZH2 FAM123B FAM175A FAM46C FANCA FANCC FAT1 FBXW7 FGF19 FGF3 FGF4 FGFR1 FGFR2 FGFR3 FGFR4 FH FLCN FLT1 FLT3 FLT4 FOXA1 FOXL2 FOXP1 FUBP1 GATA1 GATA2 GATA3 GNA11 GNAQ GNAS GREM1 GRIN2A GSK3B H3F3C HGF HIST1H1C HIST1H2BD HIST1H3B HNF1A HRAS ICOSLG IDH1 IDH2 IFNGR1 IGF1 IGF1R IGF2 IKBKE IKZF1 IL10 IL7R INPP4A INPP4B INSR IRF4 IRS1 IRS2 JAK1 JAK2 JAK3 JUN KDM5A KDM5C KDM6A KDR KEAP1 KIT KLF4 KRAS LATS1 LATS2 LMO1 MAP2K1 MAP2K2 MAP2K4 MAP3K1 MAP3K13 MAPK1 MAX MCL1 MDC1 MDM2 MDM4 MED12 MEF2B MEN1 MET MITF MLH1 MLL MLL2 MLL3 MPL MRE11A MSH2 MSH6 MTOR MUTYH MYC MYCL1 MYCN MYD88 MYOD1 NBN NCOR1 NF1 NF2 NFE2L2 NKX2-1 NKX3-1 NOTCH1 NOTCH2 NOTCH3 NOTCH4 NPM1 NRAS NSD1 NTRK1 NTRK2 NTRK3 PAK1 PAK7 PALB2 PARK2 PARP1 PAX5 PBRM1 PDCD1 PDGFRA PDGFRB PDPK1 PHOX2B PIK3C2G PIK3C3 PIK3CA PIK3CB PIK3CD PIK3CG PIK3R1 PIK3R2 PIK3R3 PIM1 PLK2 PMAIP1 PMS1 PMS2 PNRC1 POLE PPP2R1A PRDM1 PRKAR1A PTCH1 PTEN PTPN11 PTPRD PTPRS PTPRT RAC1 RAD50 RAD51 RAD51B RAD51C RAD51D RAD52 RAD54L RAF1 RARA RASA1 RB1 RBM10 RECQL4 REL RET RFWD2 RHOA RICTOR RIT1 RNF43 ROS1 RPS6KA4 RPS6KB2 RPTOR RUNX1 RYBP SDHA SDHAF2 SDHB SDHC SDHD SETD2 SF3B1 SH2D1A SHQ1 SMAD2 SMAD3 SMAD4 SMARCA4 SMARCB1 SMARCD1 SMO SOCS1 SOX17 SOX2 SOX9 SPEN SPOP SRC STAG2 STK11 STK40 SUFU SUZ12 SYK TBX3 TERT TET1 TET2 TGFBR1 TGFBR2 TMEM127 TMPRSS2 TNFAIP3 TNFRSF14 TOP1 TP53 TP63 TRAF7 TSC1 TSC2 TSHR U2AF1 VHL VTCN1 WT1 XIAP XPO1 YAP1 YES1

Association Study: Clinical Notes vs. Somatic Mutations

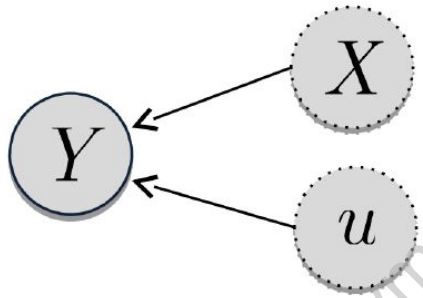
$Y \in \mathbb{N}^{P \times Q}$



$X \in \{0,1\}^{P \times G}$

$C \in \{0,1\}^{P \times L}$

P patients
Q medical obs
G genes
L covariates



Linear Mixed Model

Association Study: Clinical Notes vs. Somatic Mutations

Study with ~2000 patients with diverse cancer types (62 sub-types)

Gene	MAF	q-value	β	Sentence prototype
APC	112	0.0037	0.33	He underwent a colonoscopy which revealed a pedunculated polyp in the ascending colon.
ALK	40	0.0063	0.57	The patient showed a mild decrease in her blood counts.
HNF1A	13	0.0028	0.70	The patient was tearful presented with depressed affect and mood.
TRAF7	11	0.0008	0.59	He has a history of adenoid cystic carcinoma of the salivary gland.
NOTCH2	57	0.0009	0.24	History of multiple nonmelanoma skin cancers and melanoma.
SUFU	14	0.003	0.31	The patient has a solitary fibrous tumor.
ERBB4	93	0.06	0.08	This is a man/lady with metastatic colon carcinoma.




Association Study: Clinical Notes vs. Somatic Mutations

Study with ~2000 patients with diverse cancer types (62 sub-types)

"The APC gene in colorectal cancer" (Eur J Cancer. 2002 May;38(7):867-71)

" all patients presented with systemic symptoms and signs, including fever, **anemia**, and thrombocytosis. ... **ALK** expression was associated with localized disease" (Mod Pathol 2003;16(8):828-832)

Genomic Analysis of Non-NF2 **Meningiomas** Reveals Mutations in **TRAF7**, KLF4, AKT1, and SMO (Science 1 March 2013: Vol. 339 no. 6123 pp. 1077-1080)

Gene	MAF	q-value	β	Sentence prototype
APC	112	0.0037	0.33	He underwent a colonoscopy which revealed a pedunculated polyp in the ascending colon.
ALK	40	0.0063	0.57	The patient showed a mild decrease in her blood counts.
HNF1A	13	0.0028	0.70	The patient was tearful presented with depressed affect and mood.
TRAF7	11	0.0008	0.59	He has a history of adenoid cystic carcinoma of the salivary gland.
NOTCH2	57	0.0009	0.24	History of multiple nonmelanoma skin cancers and melanoma.
SUFU	14	0.003	0.31	The patient has a solitary fibrous tumor.
ERBB4	93	0.06	0.08	This is a man/lady with metastatic colon carcinoma.

Association found in 4 diff. cancers: Bladder Cancer | Head and Neck Carcinoma | Melanoma | Skin Cancer (Non-Melanoma)

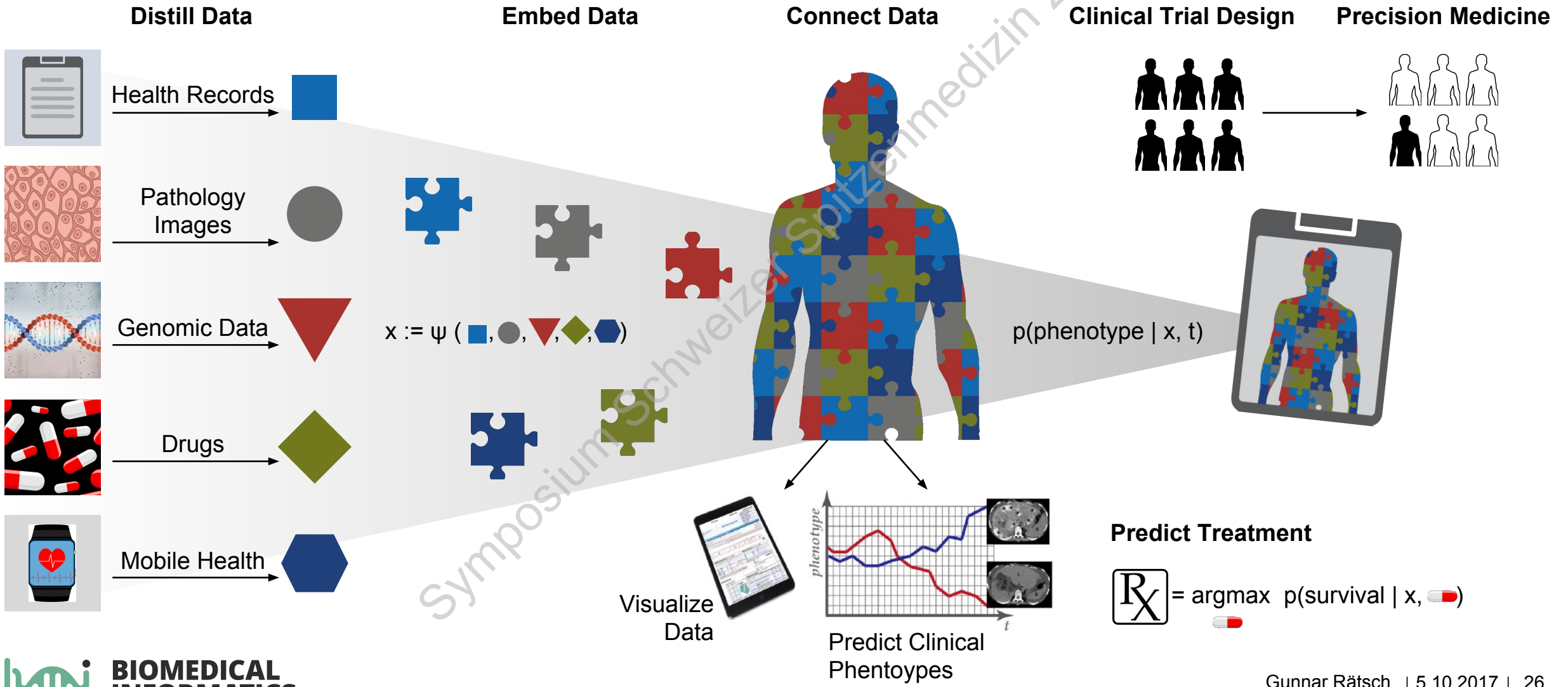
ERBB4 is over-expressed in human **colon cancer** and enhances cellular transformation. (Carcinogenesis (2015) doi: 10.1093/carcin/bgv049)

"39 **Notch2** promotes **bladder cancer** progression: Pre-clinical rationale for a novel targeted therapy" (European Urology Supplement 2014; 1569-9056)

"Evidence for differential expression of **Notch** receptors and their ligands in melanocytic nevi and cutaneous malignant **melanoma**" (Modern Pathology (2006) 19, 246–254)




Comprehensive Patient Models for "Computational Medicine"



Summary

- Group works at interface of data science and biomedicine
- Research on clinical and molecular data is needed
  - Understand molecular effects on disease and treatments
- Many scientific, technical, ethical and societal challenge around research with patient data

- Genomic, text & lab data require sophisticated analysis techniques to distill relevant information
- Data Science can provide integrative models for accurate predictions and to gain new insights
- Needs relatively big (patient) datasets usable for research
  - Need for large research cohort with clinical and molecular data

- Collaboration with physicians needed to translate into new science and better healthcare
- Data Science technologies need to get closer to the patient data
- SPHN Data Coordination Center will make collaborations easier

Thanks to my team!

Basic group statistics

Role	4 Postdocs
	10 Graduate students
	1 Scientific coordinator
	1 Administrative assistant
	1 Student Assistant
Gender	5 Female
	12 Male
Origin	3 USA
	2 Germany
	2 Switzerland
	1 Canada/Cyprus
	1 China
	1 Germany/Korea
	1 Hungary/Serbia
	1 Iran
	1 Ireland
	1 Italy
	1 Poland
	1 Russia
	1 Spain

ETH zürich

Acknowledgements

Biomedical Informatics
Cristóbal Esteban
Andre Kahles
Kjong Lehmann
Stephanie Hyland
Natalie Davidson
Gideon Dresdner
Stefan Stark
Xinrui Liu
Matthias Hüser
Vipin Sreedharan
David Kuo
Francesco Locatello

Alumni
Julia Vogt
Yi Zhong
Linda Sundermann
Melanie Fernandez
Theofanis Karaletsos
Katherine Redfield-Chan

MSKCC Cancer Biology
Guido Wendel
Kamini Singh

MSKCC Molecular Oncology Center
Niki Schultz
David Solit
David Hyman

MSKCC IT Services
Chris Crosbie
Stuart Gardos
Juan Perin

ETH IT Services
Bernd Rinn
Olivier Byrde
Stefan Walter

Global Alliance for Genomics and Health
David Haussler/UCSC
Benedict Paten/UCSC
Melissa Cline/UCSC
Stephen Chanock/NCI
John Burn/University of Newcastle

International Cancer Genome Consortium
Angela Brooks/UCSC
Alvis Brazma/EBI
Oliver Stegle/EBI

NEXUS@ETH
Nora Toussaint
Daniel Stekhoven

ETH BSSE
Dean Bodenham
Karsten Borgwardt

University Hospital Bern
Martin Faltys
Tobias Merz

Funding: ETH Zürich, Sloan Kettering Institute, Memorial Hospital, National Institute of Health, National Cancer Institute, Swiss National Science Foundation, Max Planck Society, German Research Foundation, European Union, Geoffrey Beene Foundation, Lucille Castori Center