Machine Learning for Healthcare

6.7930, HST.956

Lecture 1: What makes healthcare unique?

Peter Szolovits

Feb 7, 2023

Many slides from David Sontag

The Problem

- Cost of health care expenditures in the US are over $3 trillion, and rising
- Despite having some of the best clinicians in the world, chronic conditions are
  - Often diagnosed late
  - Often inappropriately managed
- Medical errors are pervasive

What might a solution look like?

Building a Better Delivery System: A New Engineering/Health Care Partnership

National Academy of Engineering (US) and Institute of Medicine (US) Committee on En
and the Health Care System
Proctor P Reid, W Dale Compton, Jerome H Grossman, Gary Fanjiang, editors.

Washington (DC): National Academies Press (US); 2005
The National Academies Collection: Reports funded by National Institutes of Health.

PMID: 20669457 Bookshelf ID: NBK22832 DOI: 10.17226/11378

Free Books & Documents

Excerpt

The report builds on a growing realization within the health care community of the critic
information/ communications technologies, systems engineering tools, and related orga
innovations must play in addressing the interrelated quality and productivity crises faci
care system. The report provides a framework for change and an action plan for a syste
approach to health care delivery based on a partnership between engineers, health car
professionals, and health care managers. The goal of the plan is to transform the U.S. h
sector from an underperforming conglomerate of independent entities (individual pract
small group practices, clinics, hospitals, pharmacies, community health centers, etc.) in
performance "system" in which participating units recognize their interdependence and
implications and repercussions of their actions on the system as a whole. The report de
opportunities and challenges to using systems engineering, information technologies, a
tools to advance a twenty-first century system capable of delivering safe, effective, tim
centered, efficient, equitable health care — a system that embodies the six "equality ai
envisioned in *Crossing the Quality Chasm*.

ChatGPT “passes” USMLE

Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models
Tiffany H. Kung; Morgan Cheatham; ChatGPT; Arielle Medenilla; Czarina Sillos; Lorie De Leon; Camille Elepaño; Maria Madriaga; Rimel Aggabao, Giezel Diaz-Candido; James Maningo; Victor Tseng

We evaluated the performance of a large language model called ChatGPT on the United States Medical Licensing Exam (USMLE), which consists of three exams: Step 1, Step 2CK, and Step 3. ChatGPT performed at or near the passing threshold for all three exams without any specialized training or reinforcement. Additionally, ChatGPT demonstrated a high level of concordance and insight in its explanations. These results suggest that large language models may have the potential to assist with medical education, and potentially, clinical decision-making.

http://medrxiv.org/lookup/doi/10.1101/2022.12.19.22283643

Not (yet) peer-reviewed!

Outline for today's class

1. **Brief history of AI and ML in healthcare**
2. Interlude: Student & faculty introductions
3. Why *now*? What has changed?
4. Examples of how ML will transform healthcare
5. What is *unique* about ML in healthcare?
6. Course logistics & syllabus

1970’s: MYCIN expert system

- 1970’s (Stanford): MYCIN expert system for identifying bacteria causing severe infections
- Proposed a good therapy in ~69% of cases. Better than infectious disease experts

information flow.

to help build a knowledge base, to explain a line of reasoning, and so
The knowledge base is the program’s store of facts and associat
“knows” about a subject area such as medicine. A critical design d
is how such knowledge is to be represented within the program. Th
many choices, in general. For MYCIN, we chose to represent kno
mostly as conditional statements, or rules, of the following form:

IF: There is evidence that A and B are true,

THEN: Conclude there is evidence that C is true.

This form is often abbreviated to one of the following:

If A and B, then C

Dialogue interface

I am ready

** THIS IS A 26 YEAR OLD MALE PATIENT

My understanding is:
The age of the patient is 26
The sex of the patient is male

** FIVE DAYS AGO, HE HAD RESPIRATORY-TRACT SYMPTOMS

What is his name?

** JO

My understanding is:
The name of the patient is Jo
Respiratory-tract is one of the symptoms that the patient had

** A COUPLE OF DAYS BEFORE THE ADMISSION, HE HAD A MALAISE

Please give me the date of admission

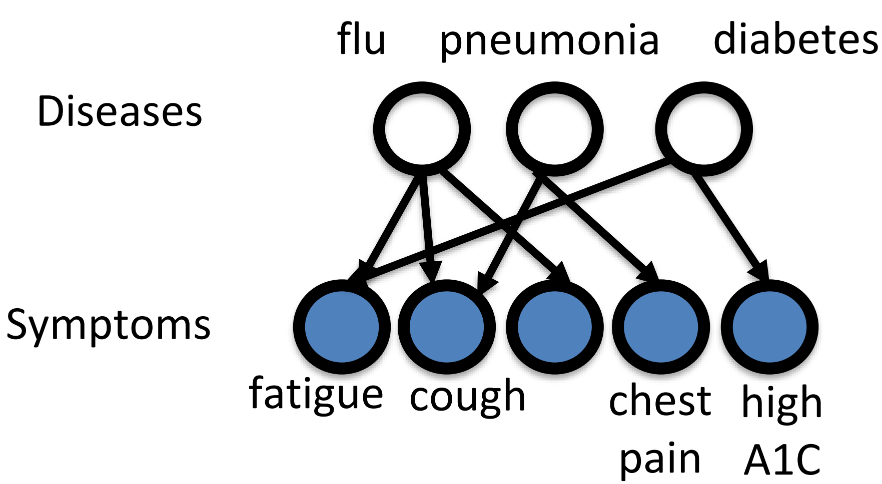
** MARCH 12, 1979

My understanding is:
The patient was admitted at the hospital 3 days ago
Malaise is one of the symptoms that the patient had 5 days ago

FIGURE 33-1 Short sample dialogue. The physician’s inputs appear in capital letters after the double asterisks.

1980's: INTERNIST-1/QMR model

- 1970-80's (Univ. of Pittsburgh): INTERNIST-1/Quick Medical Reference
- Diagnosis for internal medicine



Probabilistic model relating:

570 binary disease variables
4,075 binary symptom variables
45,470 directed edges

Elicited from doctors:
15 person-years of work

Led to advances in ML & AI (Bayesian networks, approximate inference)

Problems:
1. Clinicians entered symptoms *manually*
2. Difficult to maintain, difficult to generalize

[Miller et al., '86, Shwe et al., '91]

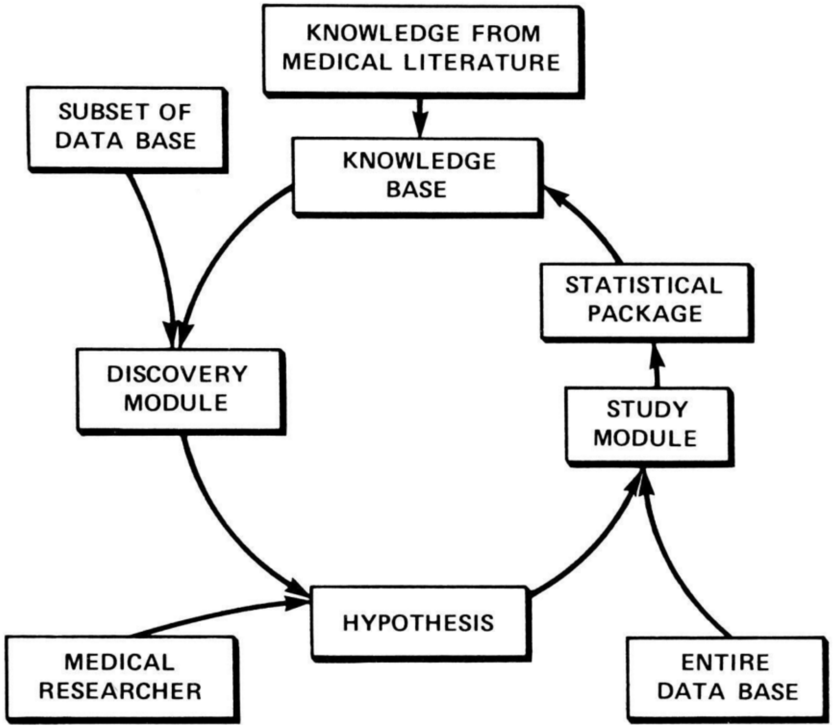
1980's: automating medical discovery

TABLE 1

HYPOTHETICAL TIME-ORIENTED RECORD FOR ONE PATIENT

Visit number	1	2	3
Date	January 17, 79	June 23, 79	July 1, 79
Knee pain	Severe	Mild	Mild
Fatigue	Moderate	—	Moderate
Temperature	38.5	37.5	36.9
Diagnosis	Systemic lupus		
White blood count	3500	4700	4300
Creatinine clearance	45	—	65
Blood urea nitrogen	36	33	—
Prednisone	30	25	20

RX PROJECT: AUTOMATED KNOWLEDGE ACQUISITION



Discovers that prednisone elevates cholesterol

(Annals of Internal Medicine, '86)

[Robert Blum, "Discovery, Confirmation and Incorporation of Causal Relationships from a Large Time-Oriented Clinical Data Base: The RX Project". Dept. of Computer Science, Stanford. 1981]

1990's: neural networks in medicine

- Neural networks with clinical data took off in 1990, with 88 new studies that year
- Small number of features (inputs)
- Data often collected by chart review

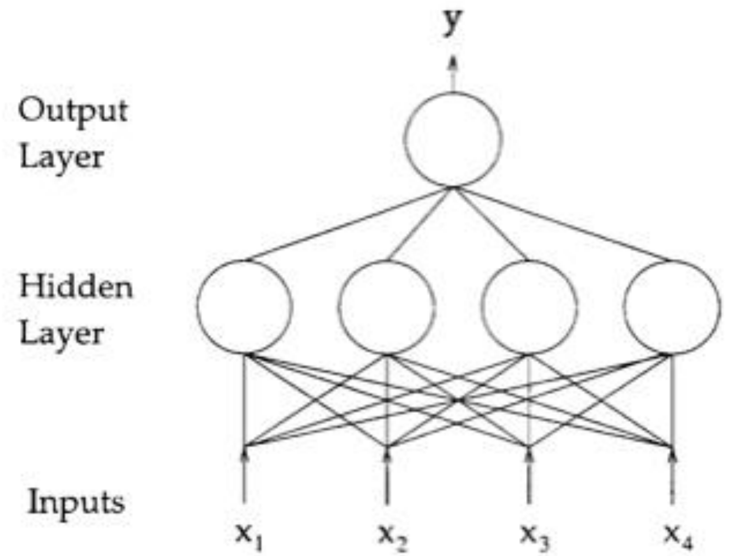


FIGURE 2. A multilayer perceptron. This is a two-layer perceptron with four inputs, four hidden units, and one output unit.

Problems:
1. Did not fit well into clinical workflow
2. Hard to get enough training data
3. Poor generalization to new places

[Penny & Frost, Neural Networks in Clinical Medicine. Med Decis Making, 1996]

Table 1 • 25 Neural Network Studies in Medical Decision Making*

Subject	No. of Examples		P†	Network	D‡	Accuracy§	
	Training	Test				Neural	Other
Breast cancer[4]	57	20	60	9–15–2	0.6	80	75
Vasculitis[2]	404	403	73	8–5–1	8.0	94	—
Myocardial infarction[6]	351	331	89	20–10–10–1	1.1	97	84
Myocardial infarction[8]	356	350	87	20–10–10–1	1.1	97	94
Low back pain[11]	100	100	25	50–48–2	0.2	90	90
Cancer outcome[13]	5,169	3,102	—	54–40–1	1.4	0.779	0.776
Psychiatric length of stay[17]	957	106	73	48–400–4	0.2	74	76
Intensive care outcome[23]	284	138	91	27–18–1	0.5	0.82	0.82
Skin tumor[21]	150	100	80	18	—	80	90
Evoked potentials[35]	100	67	52	14–4–3	3.8	77	77
Head injury[47]	500	500	50	6–3–3	20	66	77
Psychiatric outcome[54]	289	92	60	41–10–1	0.7	79	—
Tumor classification[55]	53	6	38	8–9–3	1.4	99	88
Dementia[57]	75	18	19	80–10–7–7	0.6	61	—
Pulmonary embolism[59]	607	606	69	50–4–1	2.9	0.82	0.83
Heart disease[62]	460	230	54	35–16–8–2	3	83	84
Thyroid function[62]	3,600	1,800	93	21–16–8–3	22	98	93
Breast cancer[62]	350	175	66	9–4–4–2	10	97	96
Diabetes[62]	384	192	65	8–4–4–2	12	77	75
Mycardial infarction[63]	2,856	1,429	56	291–1	9.8	85	—
Hepatitis[65]	39	42	38	4–4–3	3.3	74	79
Psychiatric admission[76]	319	339	85	53–1–1	6.0	91	—
Cardiac length of stay[83]	713	696	73	15–12–1	3.5	0.70	—
Anti-cancer agents[89]	127	141	25	60–7–6	1.5	91	86
Ovarian cancer[91]	75	98	—	6–6–2	2.6	84	81
MEDIAN VALUE	350	175	71	20	2.8		

*For reference citations, see the reference list

†P = prior probability of most prevalent category.

Outline for today's class

1. Brief history of AI and ML in healthcare
2. **Interlude: Student & faculty introductions**
3. Why *now*? What has changed?
4. Examples of how ML will transform healthcare
5. What is *unique* about ML in healthcare?
6. Course logistics & syllabus

Course staff - Professors

- Peter Szolovits (instructor)
  - Professor of Computer Science and Engineering — EECS (course 6)
    - https://people.csail.mit.edu/psz/web/
  - Professor of Health Sciences and Technology — IMES
  - PhD ’74 from Caltech, at MIT since then
  - Leads clinical decision making research group in CSAIL
    - https://mit-medg.github.io

- Manolis Kellis (instructor)
  - Professor of Computer Science and Engineering — EECS (course 6)
    - https://mit.edu/manoli/
  - Genetics, Genomics, Epigenomics, Computational Biology, Machine Learning, Alzheimer’s, Obesity, Schizophrenia, etc
  - Molecular basis of human disease circuitry, single-cell
  - PhD ’03 from MIT
  - Leads MIT computational biology group in CSAIL
    - https://compbio.mit.edu/

Course staff – teaching assistants

- Eric Lehman
  - PhD student in EECS advised by Peter Szolovits
  - Research on clinical natural language processing, predictive models for medicine

- Hussein Mozannar
  - PhD student in Social & Engineering Systems (IDSS), advised by David Sontag
    - https://husseinmozannar.github.io/
  - Research on improving Human-AI interaction by combining machine learning and HCI techniques

Student intros

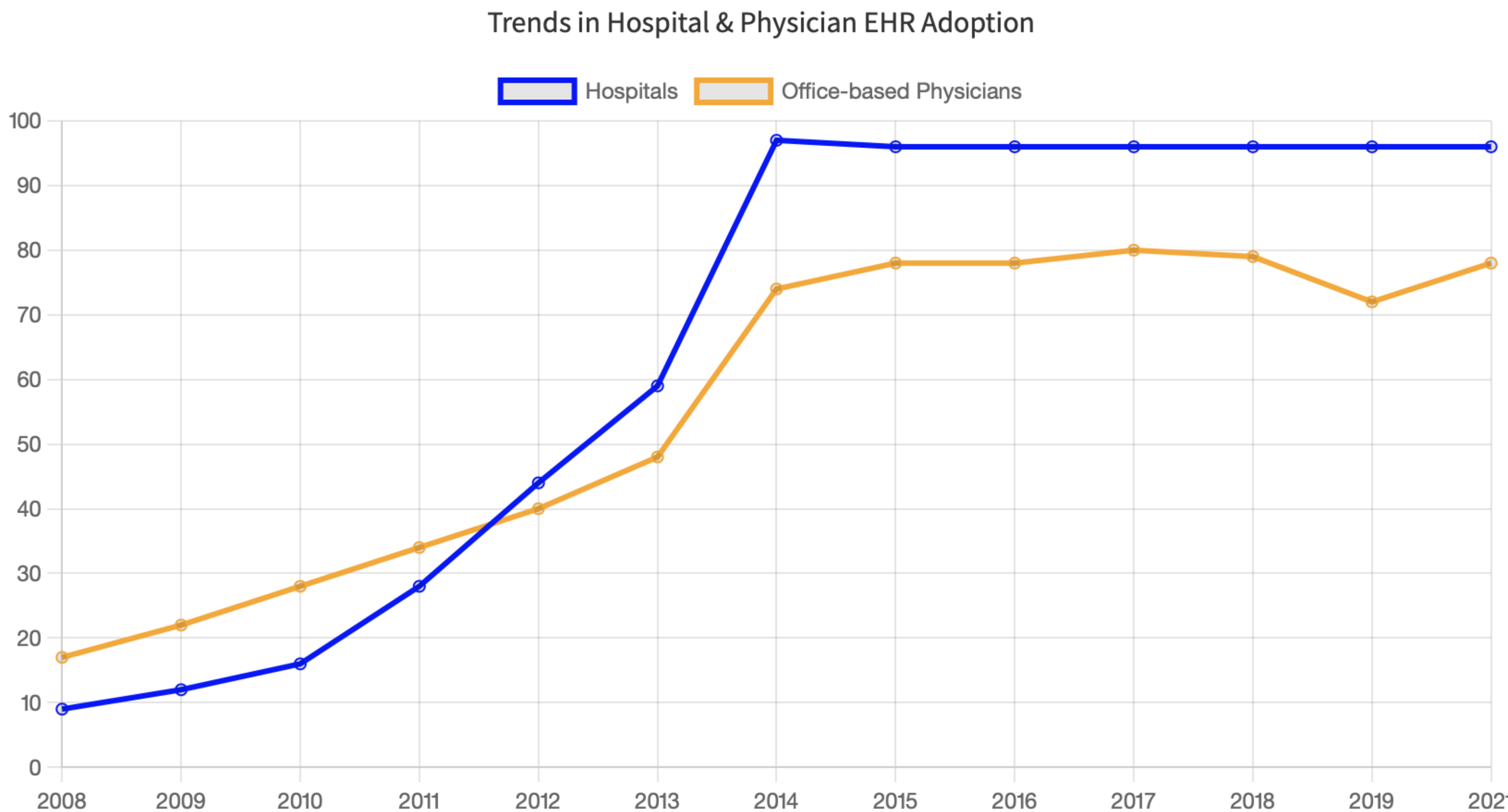
- We have a diverse set of students from MIT, Harvard, and local hospitals – let’s start to get to know each other!

Outline for today's class

1. Brief history of AI and ML in healthcare
2. Interlude: Student & faculty introductions
3. **Why *now*? What has changed?**
4. Examples of how ML will transform healthcare
5. What is *unique* about ML in healthcare?
6. Course logistics & syllabus

The Opportunity:

Adoption of Electronic Health Records (EHR) has increased 9x in US since 2008



As of 2021, nearly 4 in 5 office-based physicians (78%) and nearly all non-federal acute care hospitals (96%) adopted a certified EHR. This marks substantial 10-year progress since 2011 when 28% of hospitals and 34% of physicians had adopted an EHR.

https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records

Large datasets

Massachusetts Institute of Technology

Laboratory for Computational Physiology

De-identified health data from ~60K critical care patients

Demographics, vital signs, laboratory tests, medications, notes, …

MIMIC-IV: ~200K

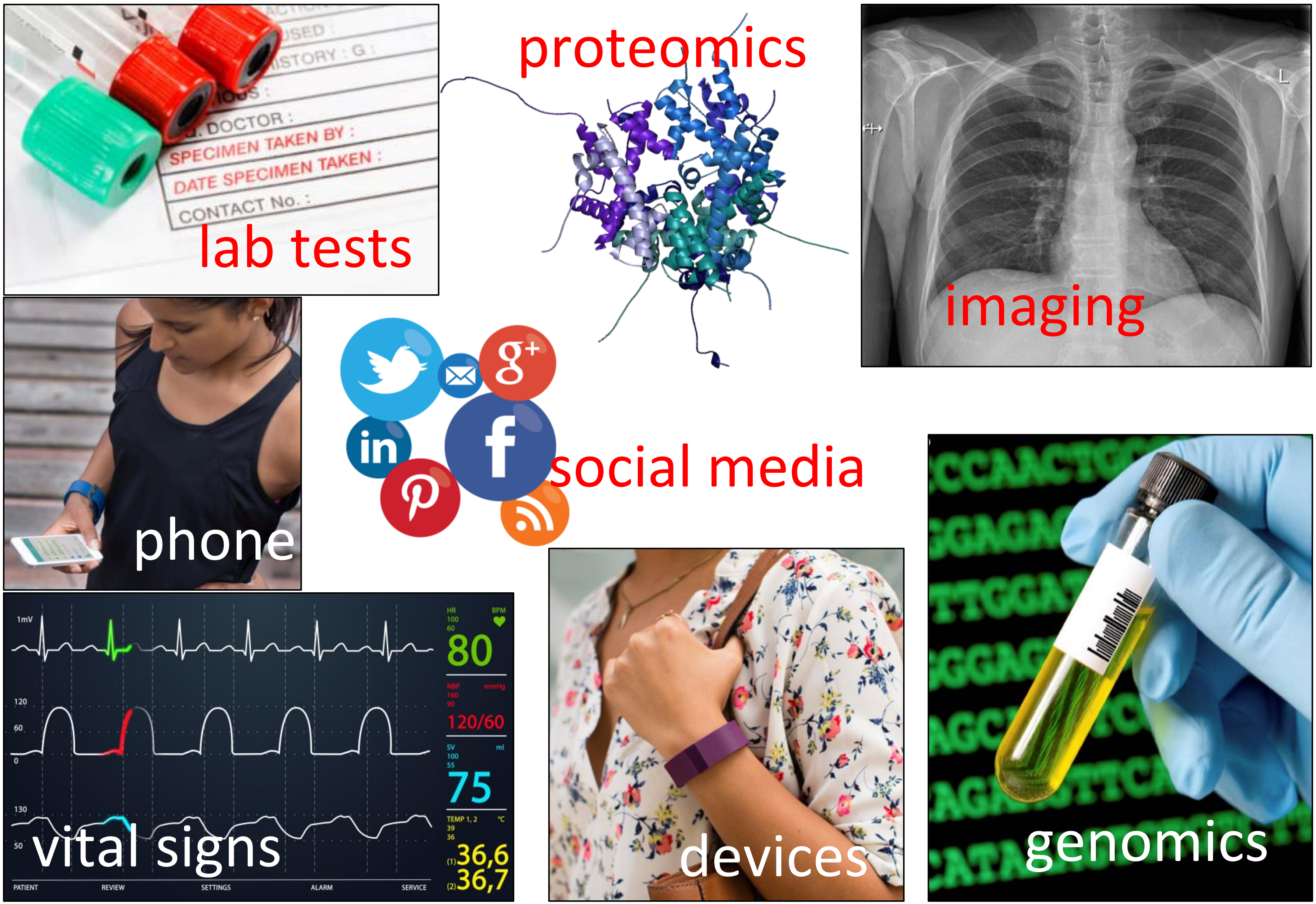
+ ED data, CXR

UK Biobank: ~500K participants

(# of data items in each category)

- Population characteristics:
  - Baseline characteristics: 31
  - Ongoing characteristics: 4
- Assessment centre
  - Recruitment: 17
  - Touchscreen: 396
  - Verbal interview: 37
  - Physical measures: 517
  - Cognitive function: 103
  - Imaging: 2534
  - Biological sampling: 10
  - Procedural metrics: 74
- Biological samples
  - Blood assays: 945
  - Sample inventory: 13
  - Saliva assays: 0
  - Urine assays: 16
- Genomics
  - Polygenic Risk Scores: 91
  - Genetically deduced phenotypes: 1
  - Imputation: 4
  - Genotypes: 35
  - Exome sequences: 32
  - Whole genome sequences: 99
  - Telomeres: 5
- Online follow-up
  - Cognitive function online: 56
  - Diet by 24-hour recall: 473
  - Digestive health: 54
  - Experience of pain: 129
  - Food (and other) preferences 153
  - Mental health: 142
  - Work environment: 100
- Additional exposures
  - Local environment: 37
  - Physical activity measurement: 210
  - Cardiac monitoring: 110
- Health-related outcomes
  - Coronavirus COVID-19: 177
  - Primary care: 3
  - Hospital inpatient: 80
  - Death register: 8
  - Cancer register: 9
  - Algorithmically-defined outcomes: 38
  - First occurrences: 2330

Diversity of digital health data



Standardization

- Diagnosis codes: ICD-9 and ICD-10 (International Classification of Diseases)

…

ICD-9 codes 290–319: mental disorders

ICD-9 codes 320–359: diseases of the nervous system

ICD-9 codes 360–389: diseases of the sense organs

ICD-9 codes 390–459: diseases of the circulatory system

ICD-9 codes 460–519: diseases of the respiratory system

ICD-9 codes 520–579: diseases of the digestive system

ICD-9 codes 580–629: diseases of the genitourinary system

ICD-9 codes 630–679: complications of pregnancy, childbirth,

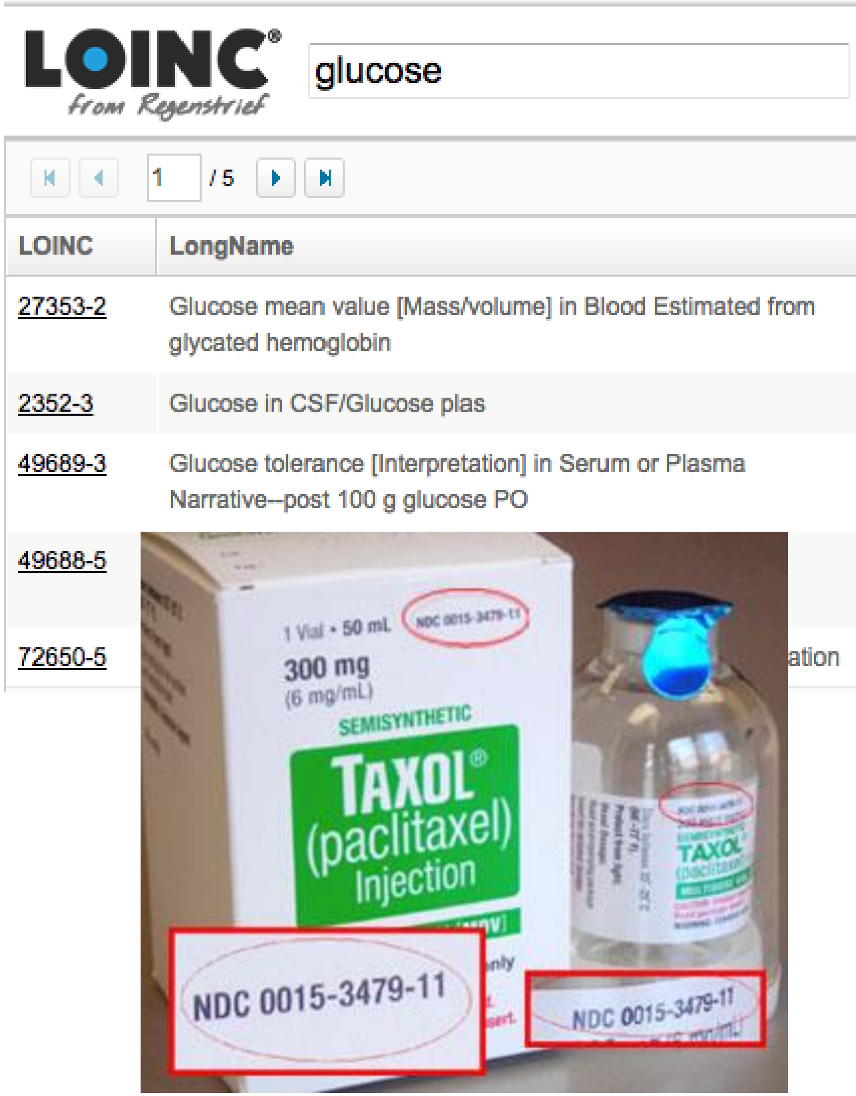
…

[https://en.wikipedia.org/wiki/List_of_ICD-9_codes]

[https://blog.curemd.com/the-most-bizarre-icd-10-codes-infographic/]

Standardization

- Diagnosis codes: ICD-9 and ICD-10 (International Classification of Diseases)
- Laboratory tests: LOINC codes
- Pharmacy: National Drug Codes (NDCs)
- Unified Medical Language System (UMLS): millions of medical concepts



[http://oplinc.com/newsletter/index_May08.htm]

Standardization

FHIR

Level 1 Basic framework on which the specification is built

Foundation
Base Documentation, XML, JSON, REST API + Search, Data Types, Extensions

Level 2 Supporting Implementation, and binding to external specifications

Implementer Support
Downloads, Common Use Cases, Testing

Security & Privacy
Security, Consent Provenance AuditEvent

Conformance
StructureDefinition, CapabilityStatement, ImplementationGuide, Profiling

Terminology
CodeSystem, ValueSet, ConceptMap, Terminology Svc

Linked Data
RDF

Level 3 Linking to real world concepts in the healthcare system

Administration
Patient, Practitioner, Device, Organization, Location, Healthcare Service

Level 4 Record-keeping and Data Exchange for the healthcare process

Clinical
Allergy, Problem, CarePlan, DetectedIssue, RiskAssessment, etc.

Diagnostics
Observation, Report, Specimen, ImagingStudy,Genomics,etc

Medications
Order, Dispense, Administration, Statement, Immunization, etc.

Workflow
Task, Appointment, Schedule, Referral, PlanDefinition, etc.

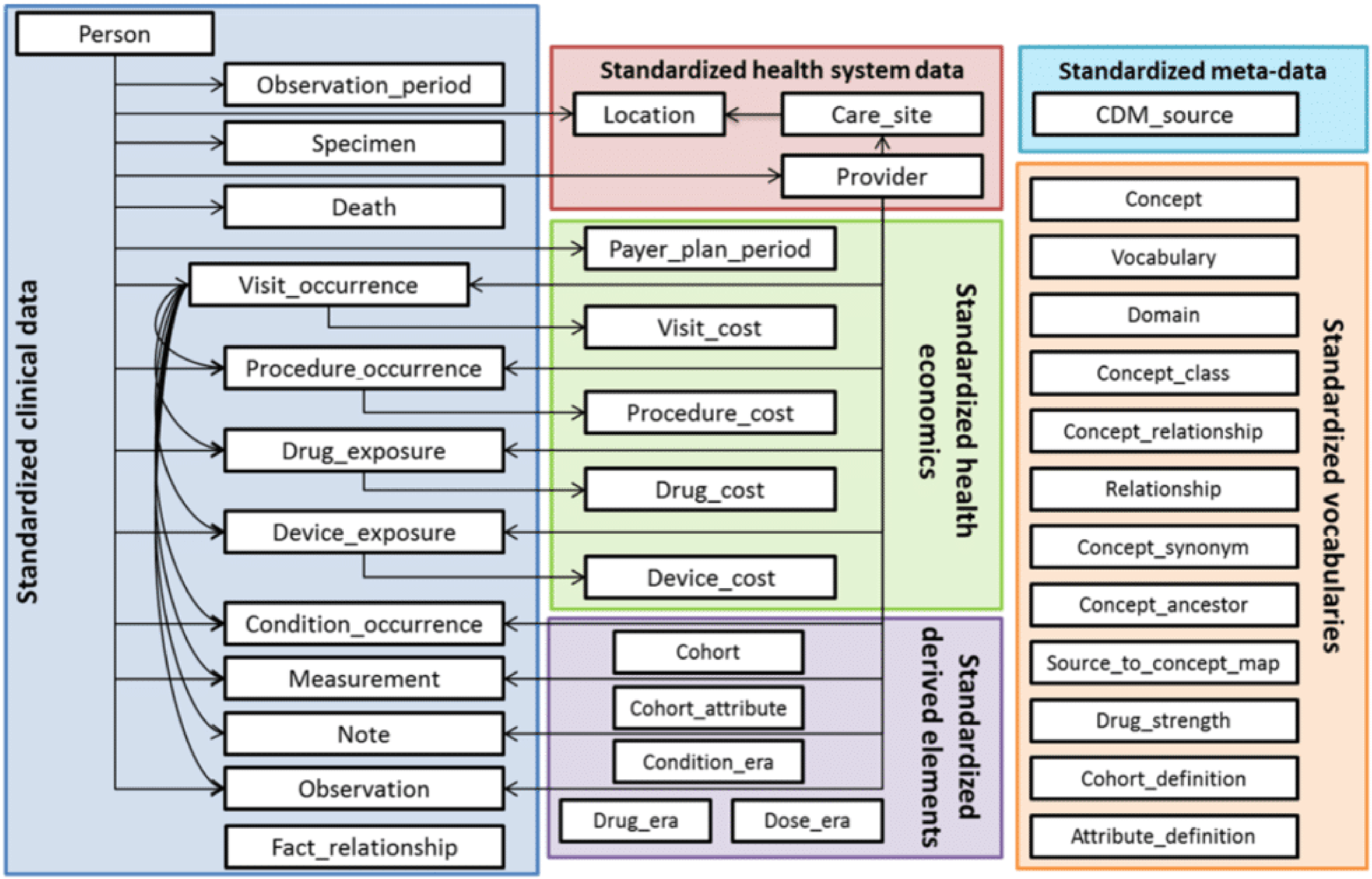
Financial
Claim, Account, Coverage, Claim, EligibilityRequest, ExplanationOfBenefit, etc.

Level 5 Providing the ability to reason about the healthcare process

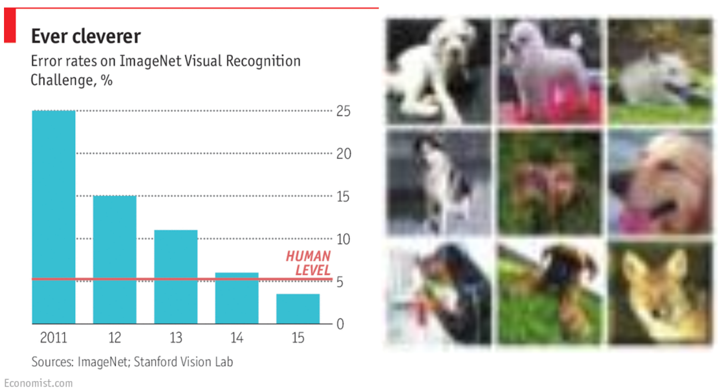
Clinical Reasoning
Library, ServiceDefinition & GuidanceResponse, Measure/MeasureReport, etc

Standardization

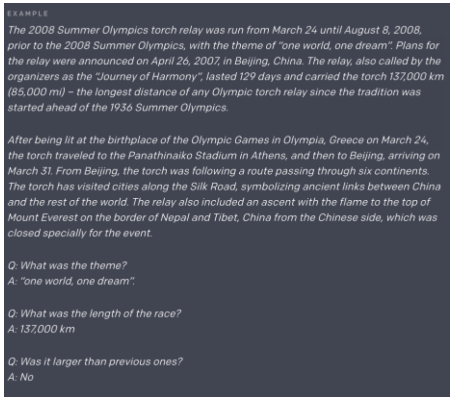
OMOP Common Data Model v5.0



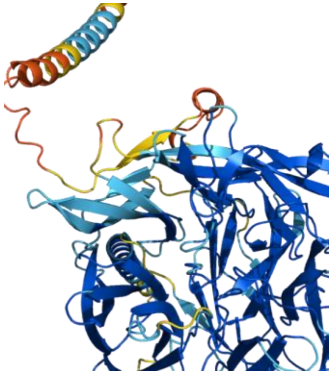
Breakthroughs in machine learning



Object recognition
(deep neural networks)



Text comprehension
(language models)



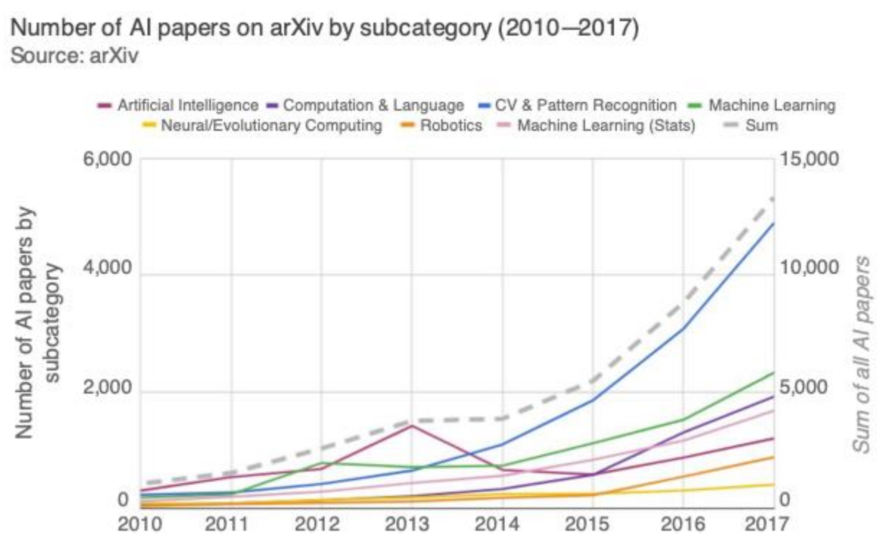
AlphaFold
(attention model learning)

Figure 5: Linear interpolation in latent space between real images.

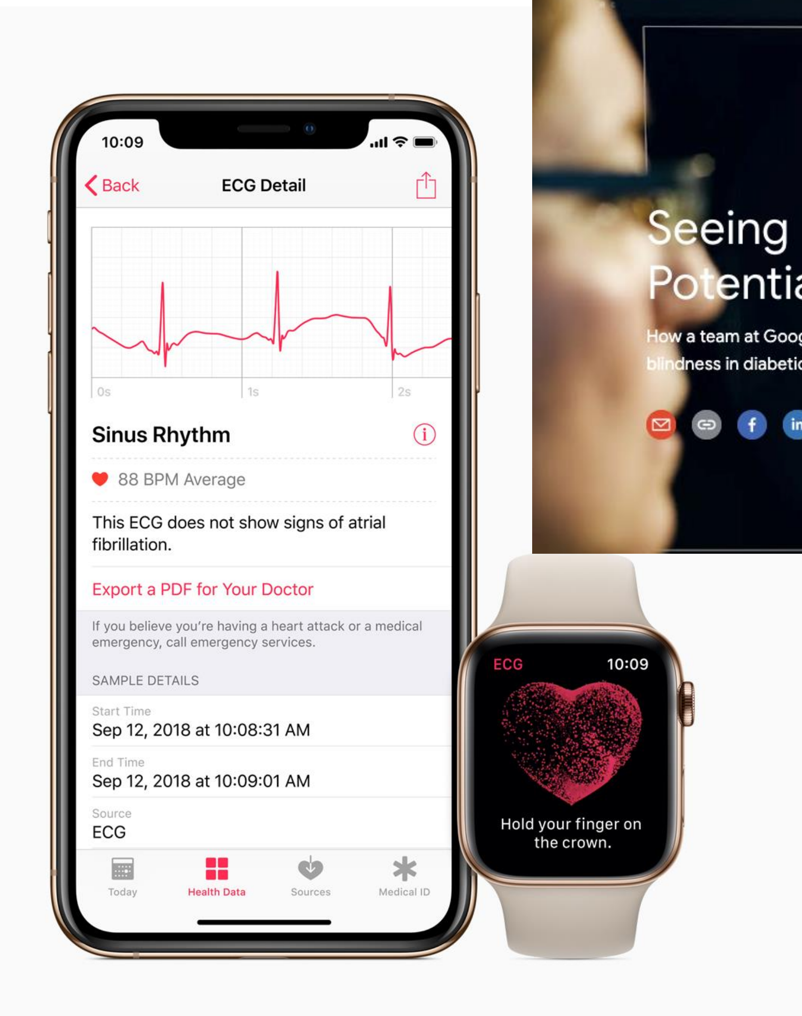
Generating realistic data
(GANs, VAEs)

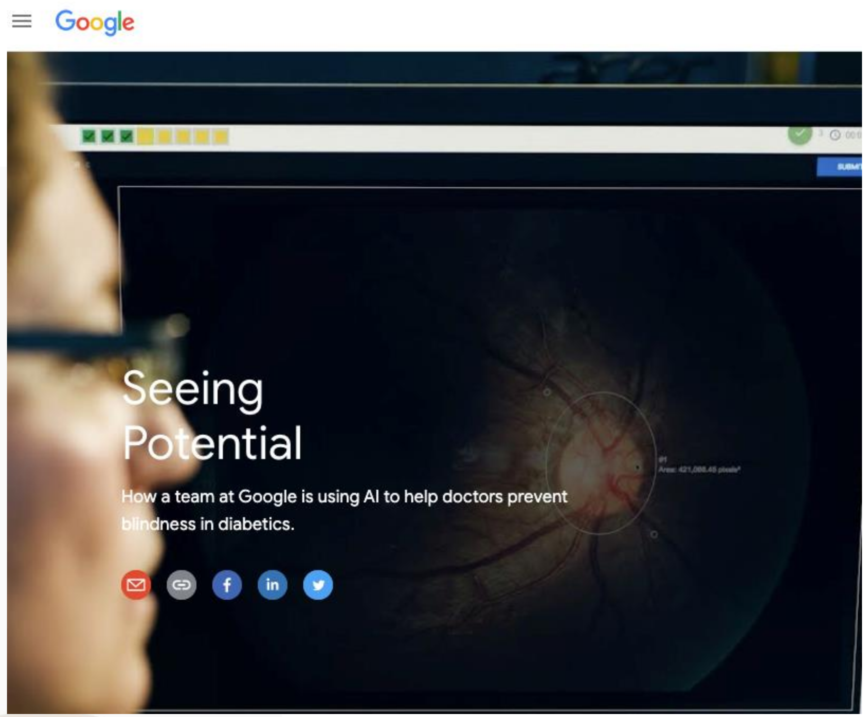
What’s driving these advances?

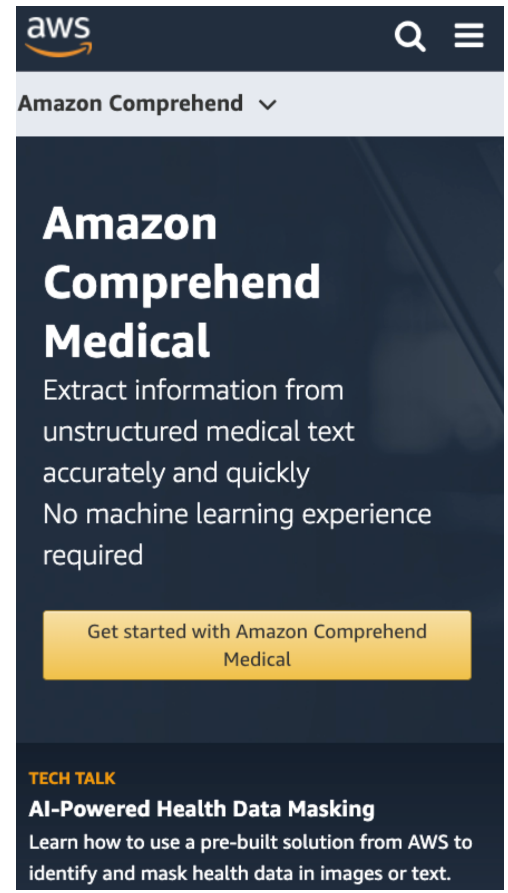
- Democratization of machine learning
  - Large datasets
  - Cheap fast processing (GPUs + TPUs)
  - High-quality open-source software (scikit-learn, PyTorch, TensorFlow)
- More and more researchers



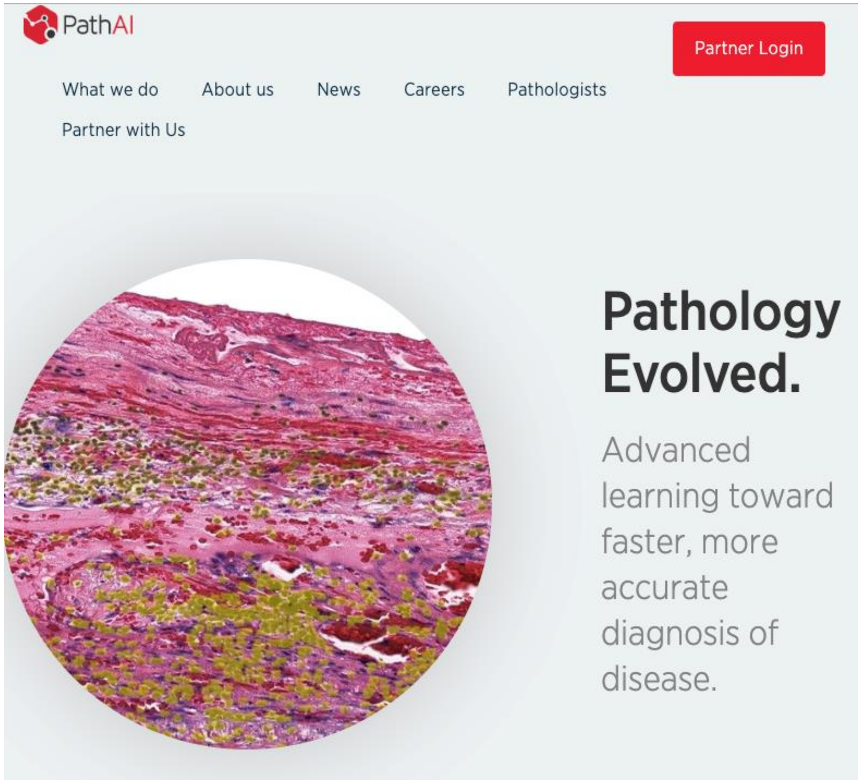
Tech industry interest in health care

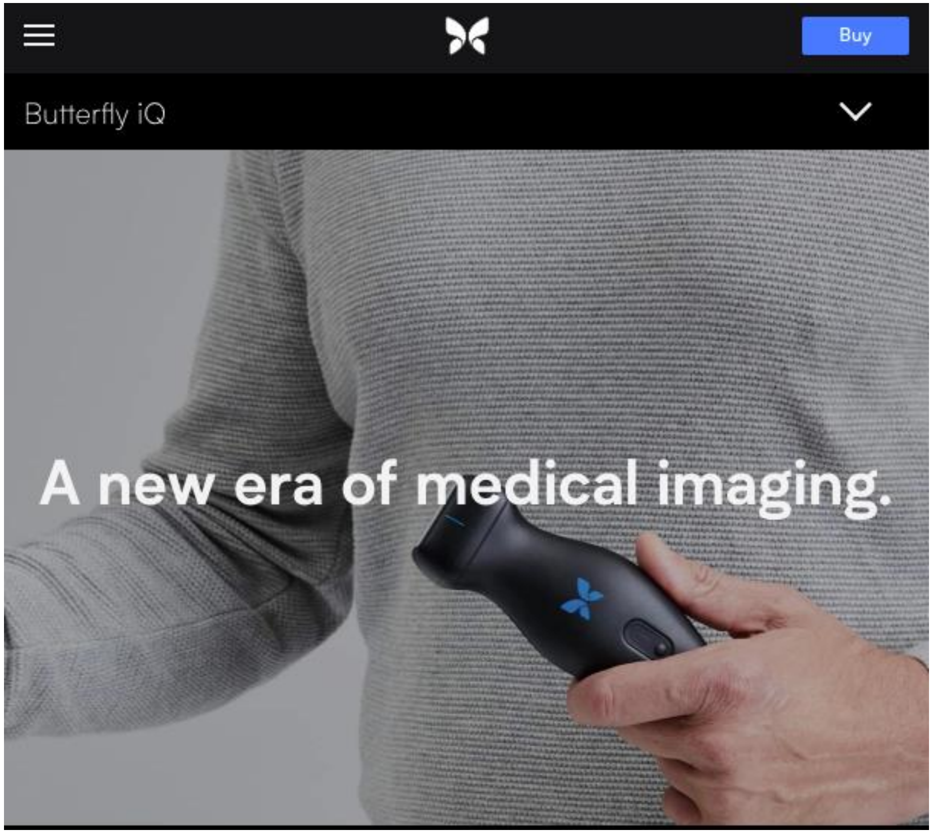






Tech industry interest in health care



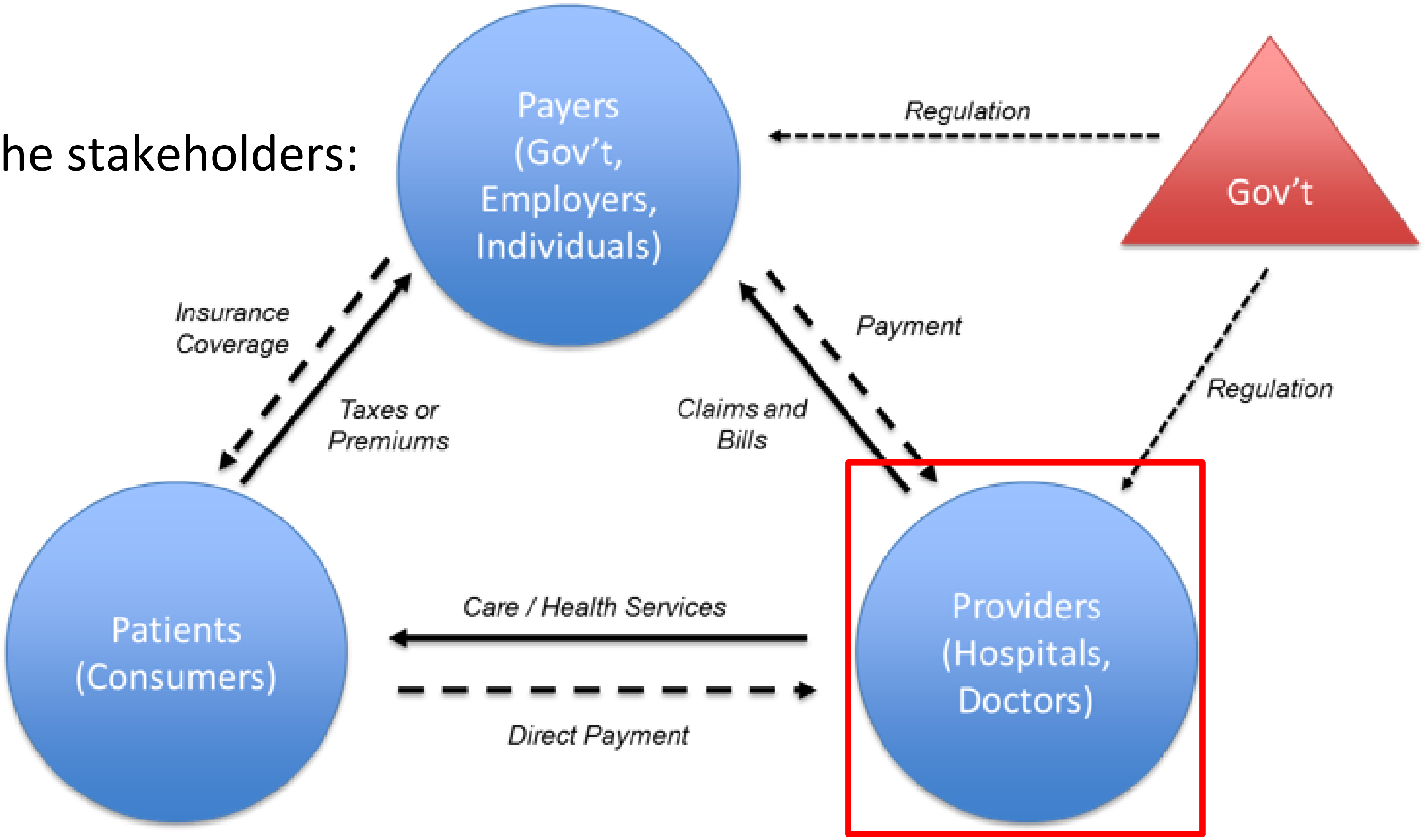


Outline for today's class

1. Brief history of AI and ML in healthcare
2. Interlude: Student & faculty introductions
3. Why *now*? What has changed?
4. **Examples of how ML will transform healthcare**
5. What is *unique* about ML in healthcare?
6. Course logistics & syllabus

ML will transform every aspect of healthcare

The stakeholders:



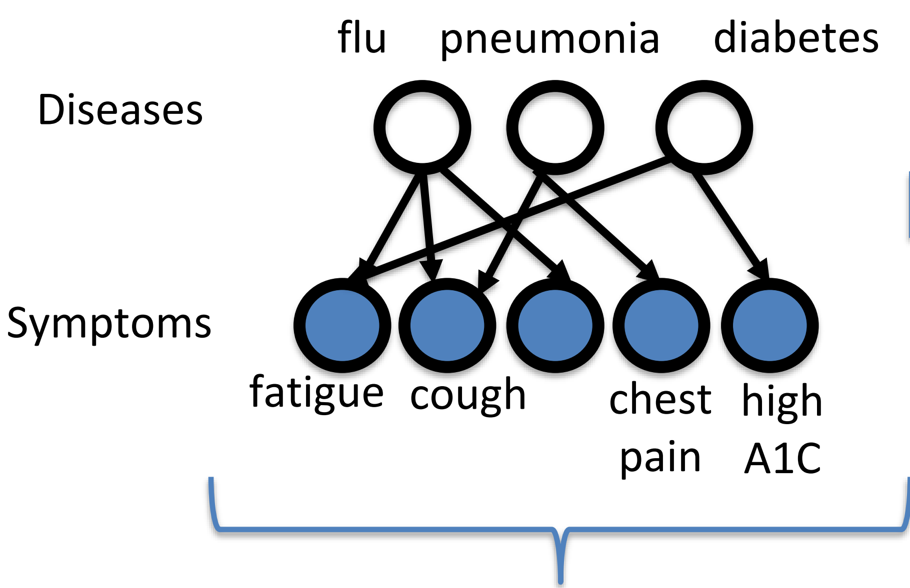
Source for figure:
http://www.mahesh-vc.com/blog/understanding-whos-paying-for-what-in-the-healthcare-industry

Emergency Department:

- Limited resources
- Time sensitive
- Critical decisions

What will the ER of the future be like?

Behind-the-scenes reasoning about the patient’s conditions (current and future)



Automatically extracted from electronic health record

- Better triage
- Faster diagnosis
- Early detection of adverse events
- Prevent medical errors

What will the ER of the future be like?

Propagating best practices

The ED Dashboard decision support algorithms have determined that this patient may be eligible for the Atrius Cellulitis pathway. Please choose from the following options:

Enroll in pathway

Decline

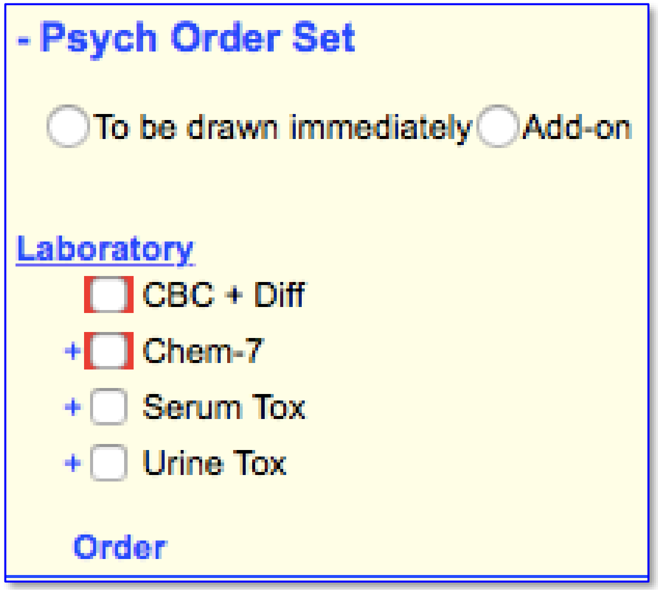
You can include a comment for the reviewers: *Mandatory if Declining*

Below are links to the pathway and/or other supporting documents:

Atrius Cellulitis Pathway

What will the ER of the future be like?

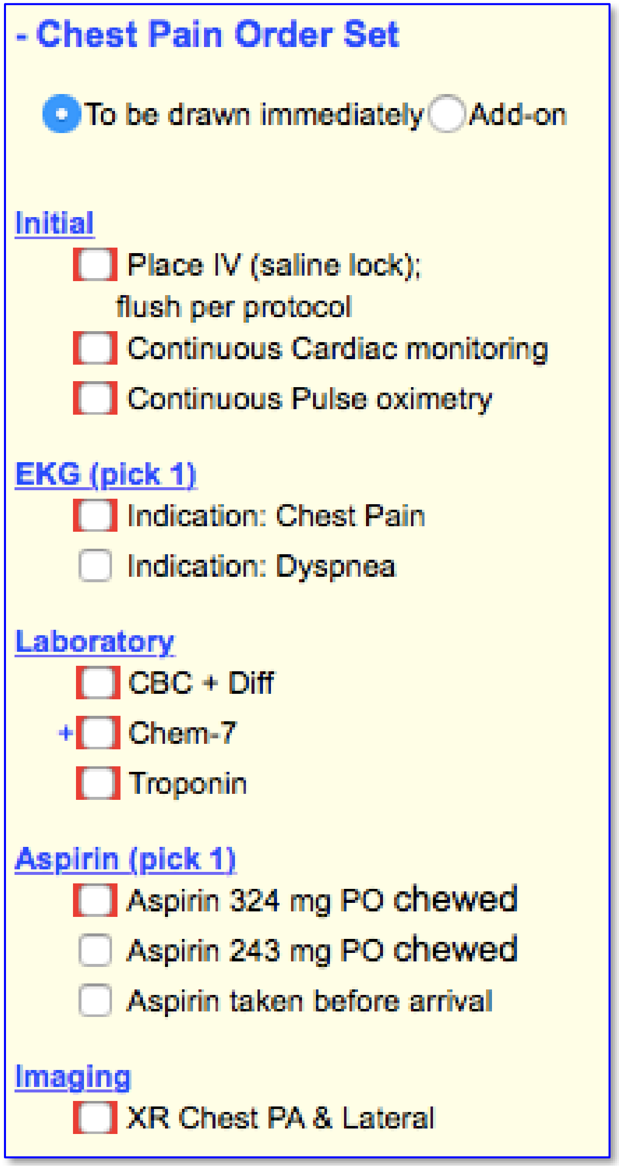
Anticipating the clinicians' needs


- Psych Order Set

To be drawn immediately / Add-on

Laboratory
- CBC + Diff
- + Chem-7
- + Serum Tox
- + Urine Tox

Order


- Chest Pain Order Set

To be drawn immediately / Add-on

Initial
- Place IV (saline lock); flush per protocol
- Continuous Cardiac monitoring
- Continuous Pulse oximetry

EKG (pick 1)
- Indication: Chest Pain
- Indication: Dyspnea

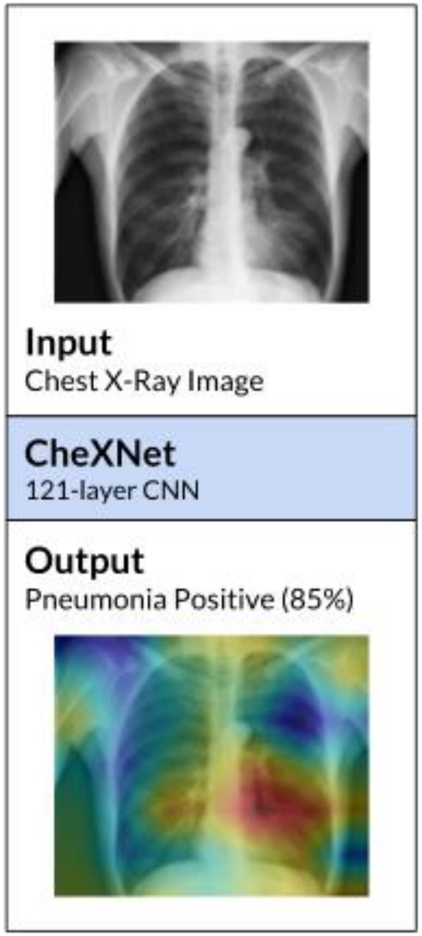
Laboratory
- CBC + Diff
- + Chem-7
- Troponin

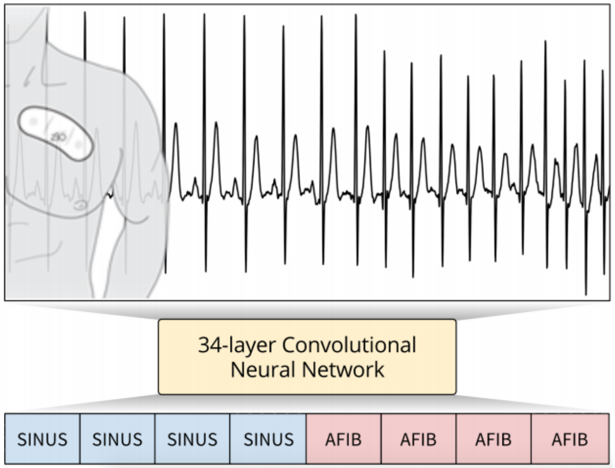
Aspirin (pick 1)
- Aspirin 324 mg PO chewed
- Aspirin 243 mg PO chewed
- Aspirin taken before arrival

Imaging
- XR Chest PA & Lateral

What will the ER of the future be like?

Reducing the need for specialist consults



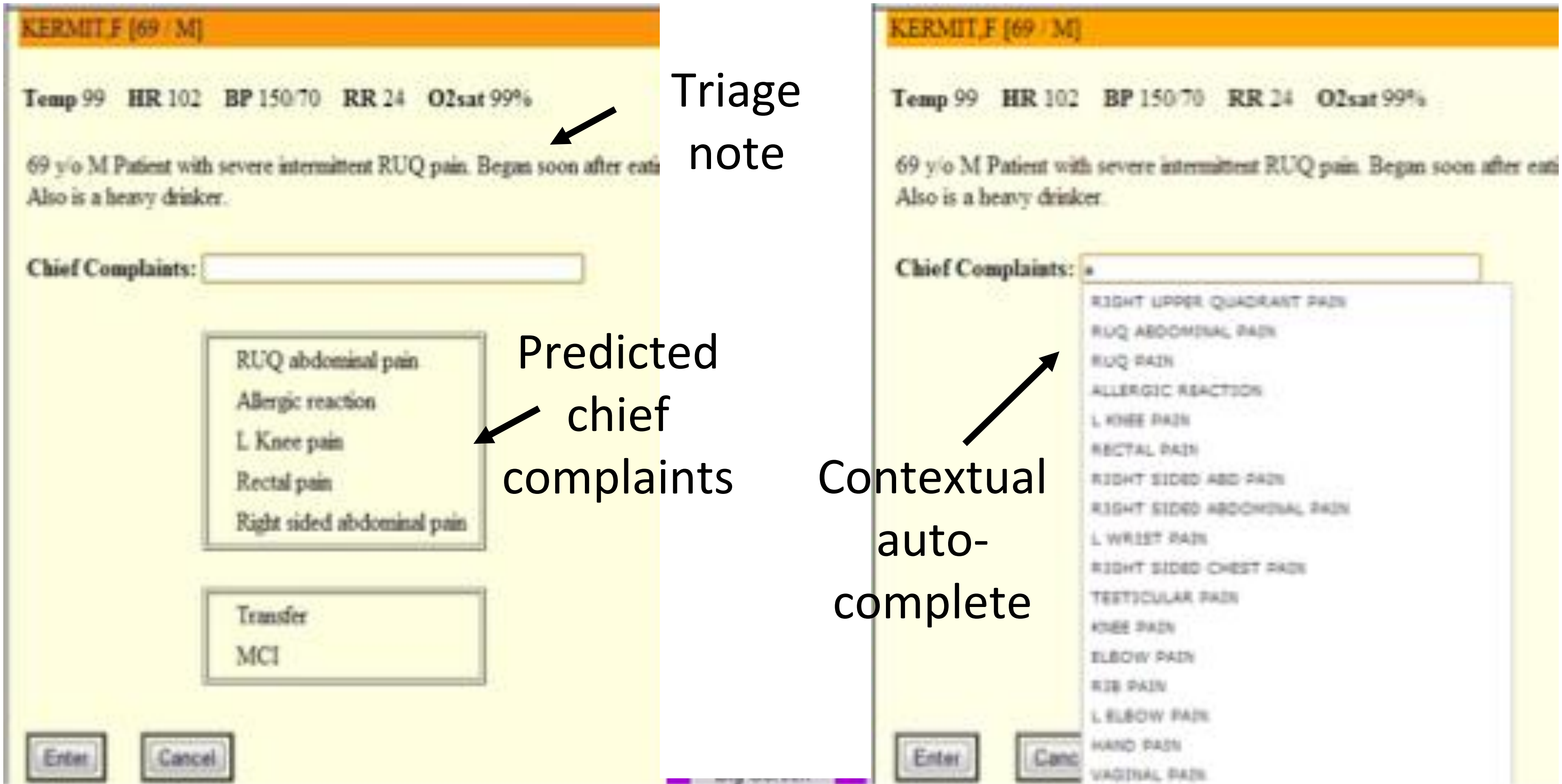


Arrhythmia?

Figure sources: Rajpurkar et al., arXiv:1711.05225 ’17
Rajpurkar et al., arXiv:1707.01836, '17

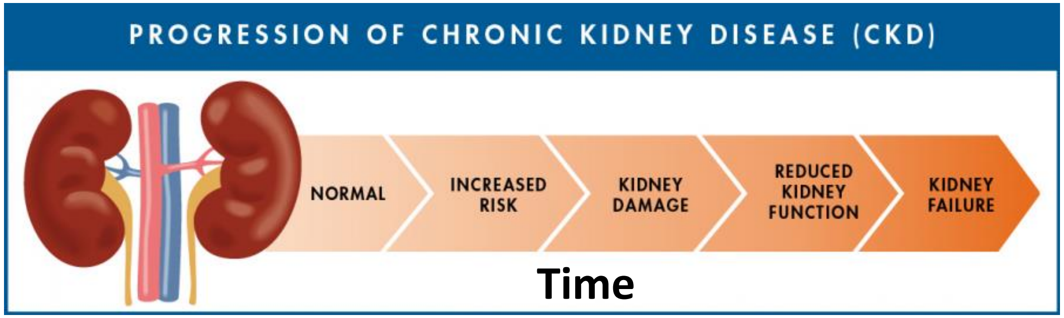
What will the ER of the future be like?

Automated documentation and billing



What is the future of how we treat chronic disease?

- Predicting a patient's future disease progression



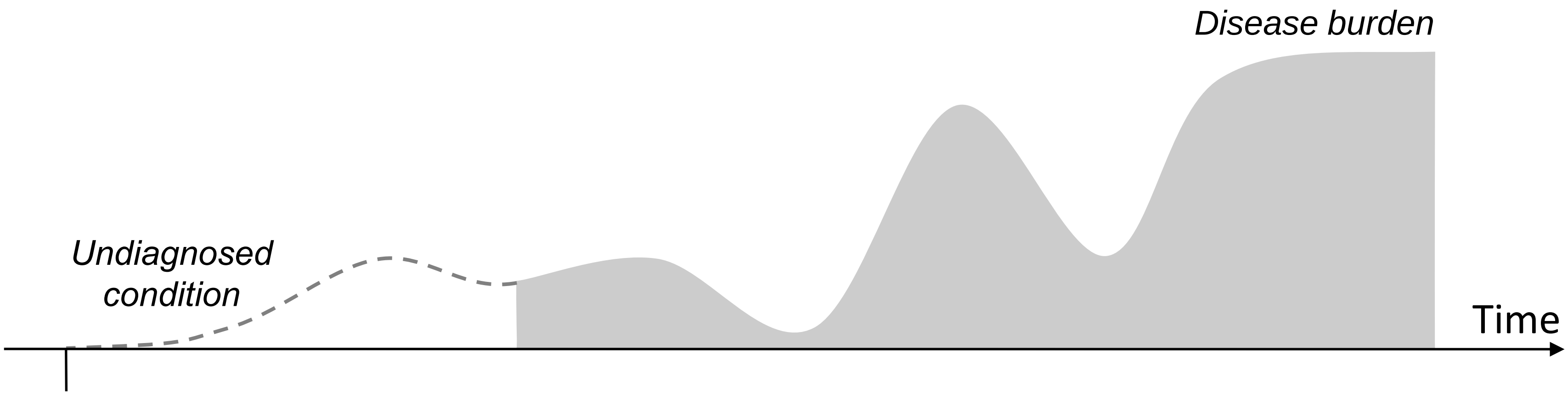


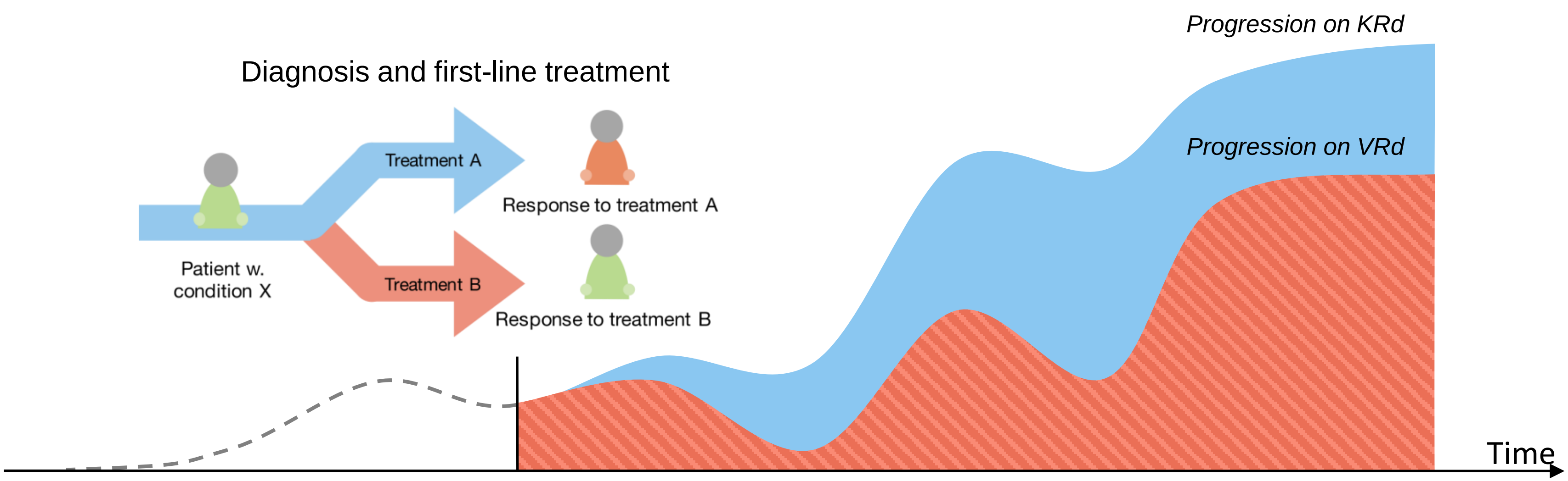
Figure credit: https://www.cdc.gov/kidneydisease/prevention-risk.html

What is the future of how we treat chronic disease?

- Predicting a patient's future disease progression
- Precision medicine

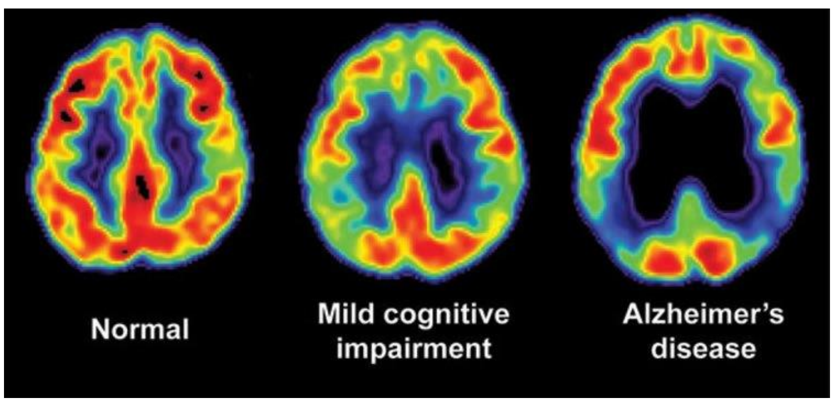
Choosing first line therapy in multiple myeloma

A) KRd: carfilzomib-lenalidomide-dexamethasone, **B) VRd:** bortezomib-lenalidomide-dexamethasone



What is the future of how we treat chronic disease?

- Early diagnosis, e.g. of diabetes, Alzheimer's, cancer



Liquid biopsy

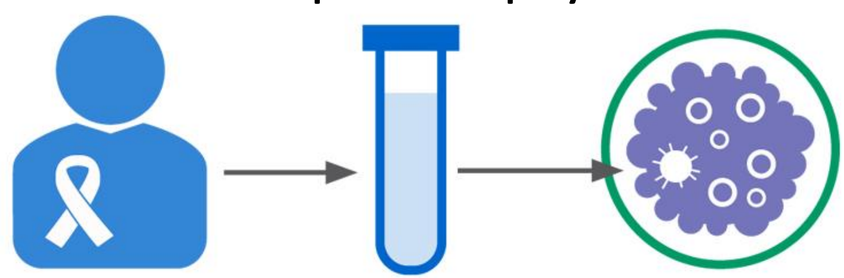


Figure sources: NIH, https://www.roche.com/research_and_development/what_we_are_working_on/oncology/liquid-biopsy.htm

What is the future of how we treat chronic disease?

- Continuous monitoring and coaching, e.g. for the elderly, diabetes, psychiatric disease

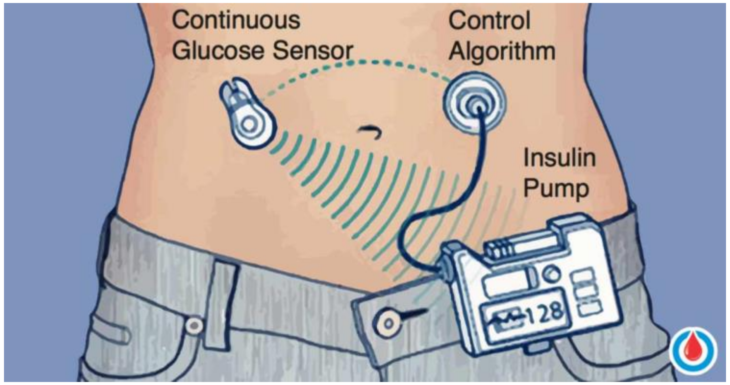
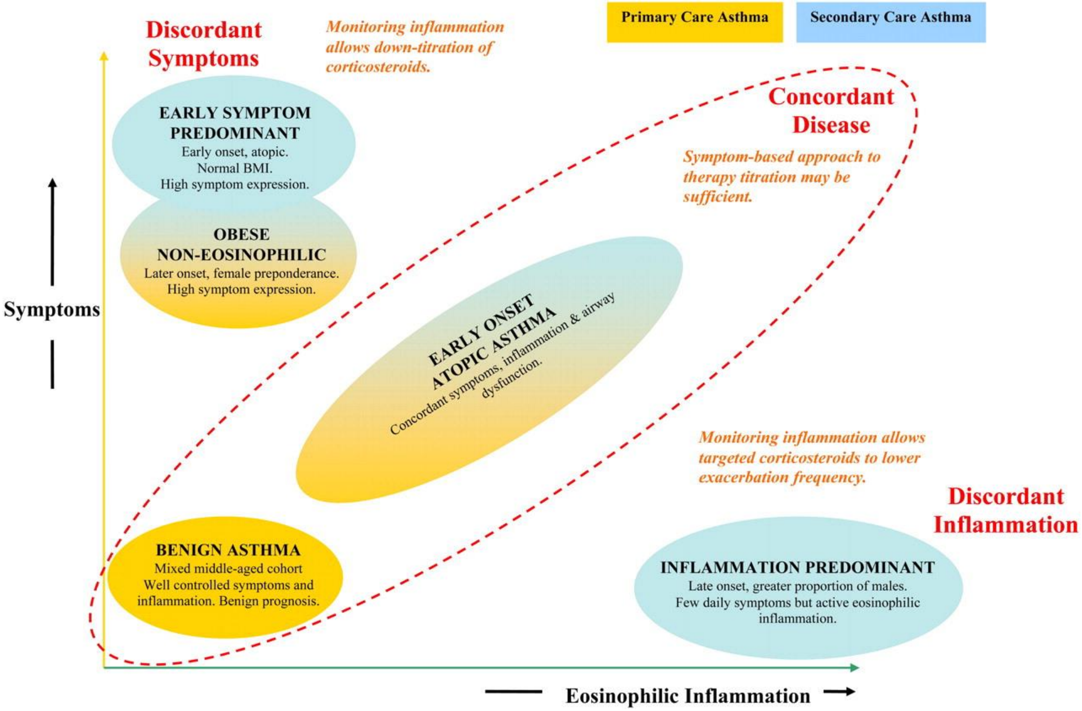


Figure source (left): http://www.emeraldforhome.com/

What is the future of how we treat chronic disease?

- Discovery of new disease subtypes; design of new drugs; better targeted clinical trials



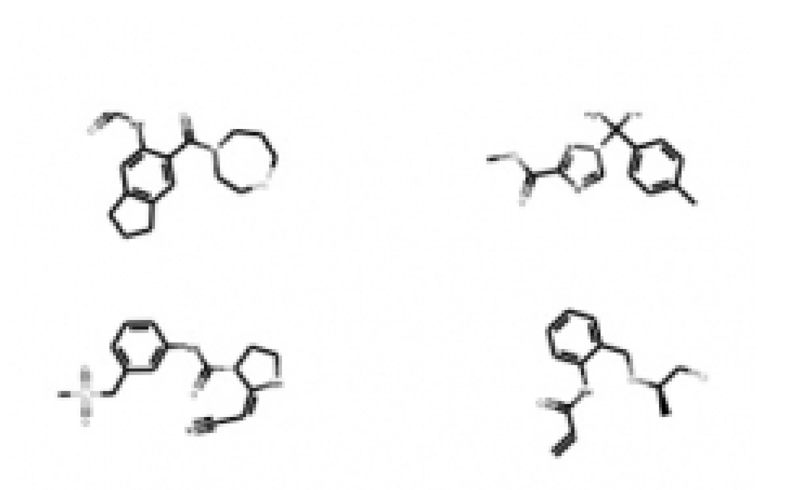
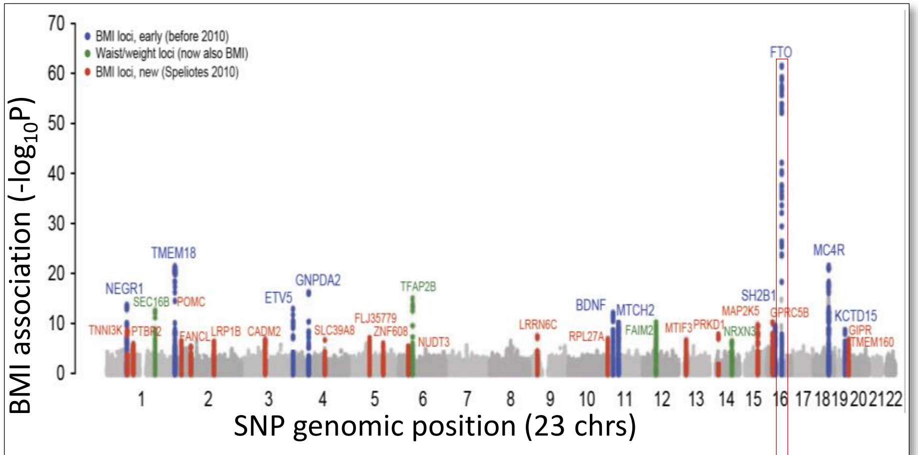


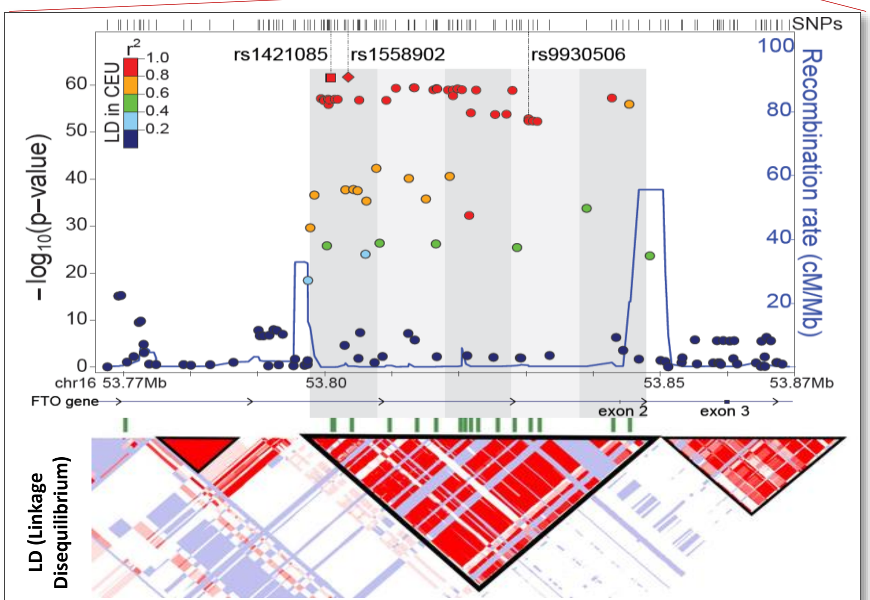
Figure sources: Haldar et al., Am J Respir Crit Care Med, 2008
http://news.mit.edu/2018/automating-molecule-design-speed-drug-development-0706

Genomic medicine: challenge and promises

GWAS Manhattan Plot: simple χ^2 statistical test



Speliotes NG 2010



Dina NG 2007, Frayling Science 2007, Claussnitzer NEJM 2015

The promise of genetics

- Path to causality
- Disease mechanism
- New target genes
- New therapeutics
- Personalized medicine

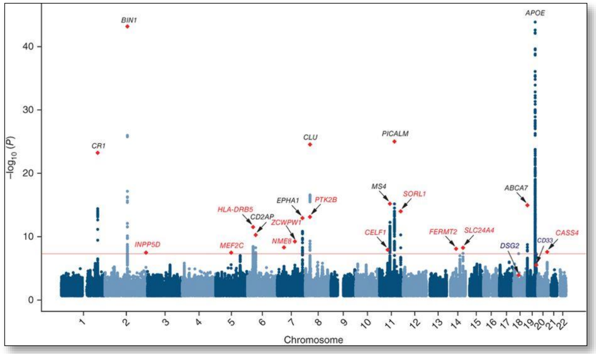
The challenge of mechanism

- **90+% disease hits non-coding**
- Target gene not known
- Causal variant not known
- Cell type of action not known
- Relevant pathways not known
- Mechanism not known

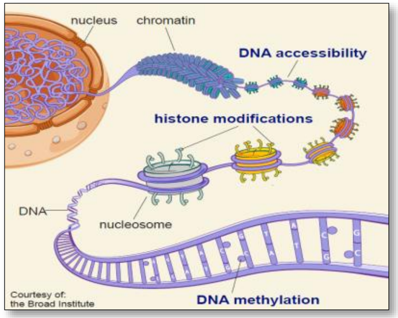
Ward NBT'12

Claussnitzer NEJM'15

Dissect mechanisms of disease-associated regions

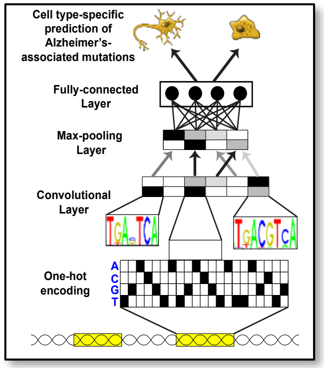


1. Disease genetics reveals common + rare variants/regions

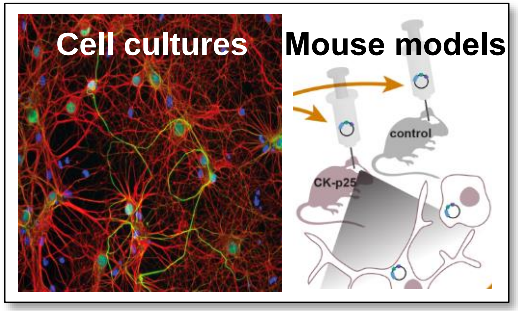


2. Profile RNA + Epigenome in healthy + disease samples

5. Disseminate results



3. Integrate data to predict driver genes, regions, cell types



4. Validate predictions in human cells + mouse models

Roadmap Nature 15

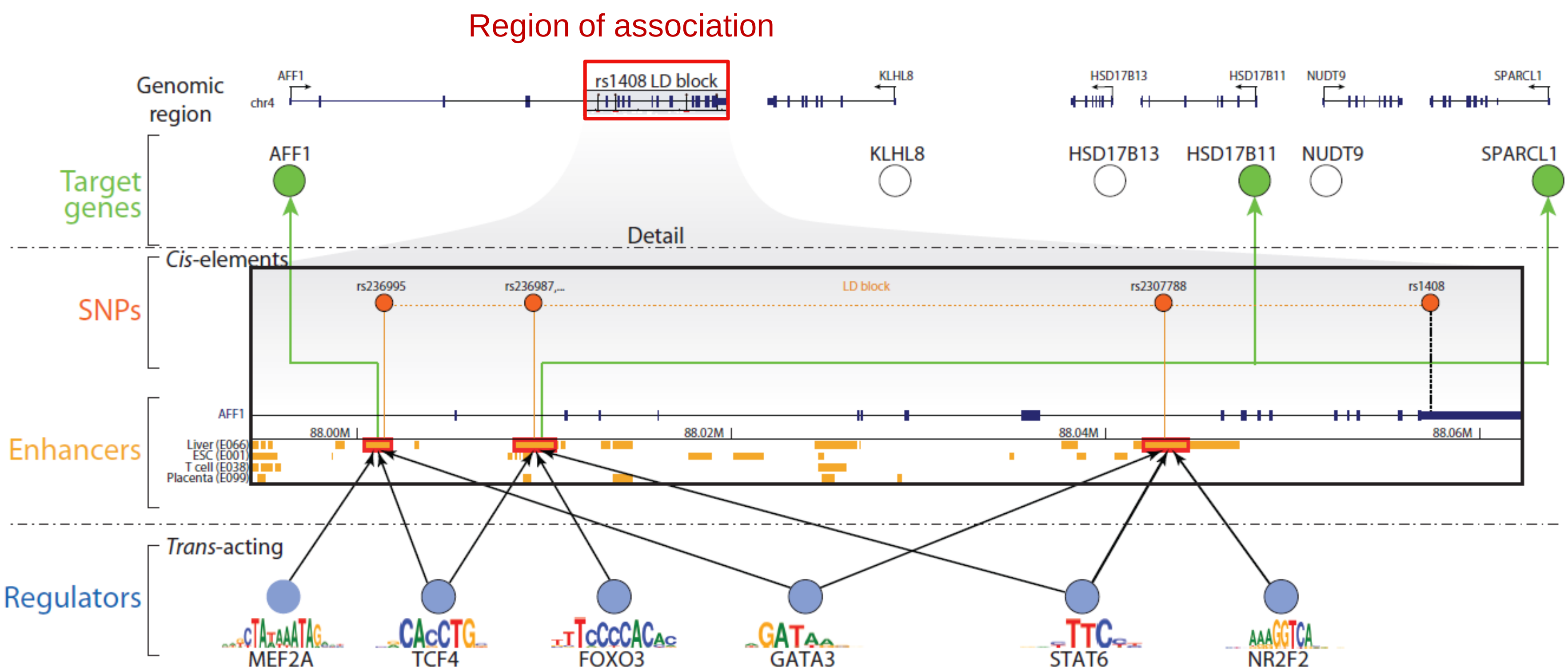
Boix EpiMap Nature 21

Claussnitzer NEJM'15

Blanchard, Nature, 2022

Park NBT 15

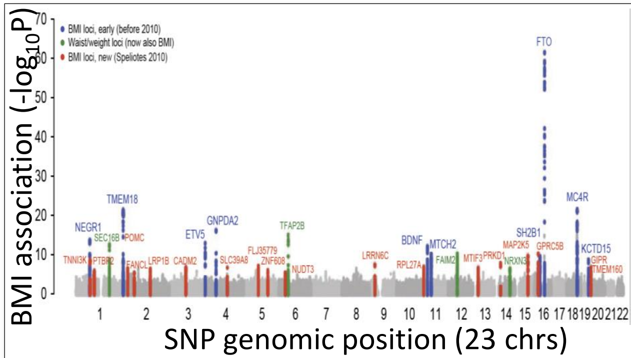
Non-coding circuitry helps interpret disease loci

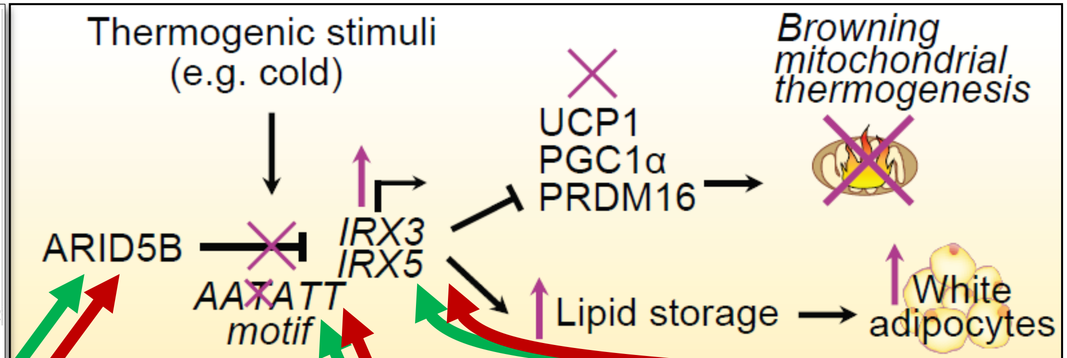


- Expand each GWAS locus using SNP linkage disequilibrium (LD)
  - Recognize **relevant cell types**: tissue-specific enhancer enrichment
  - Recognize **driver TFs**: enriched motifs in multiple GWAS loci
  - Recognize **target genes**: linked to causal enhancers

Quon bioRxiv 467852

FTO & Obesity: Uncover & manipulate circuitry ➔ reverse disease phenotypes

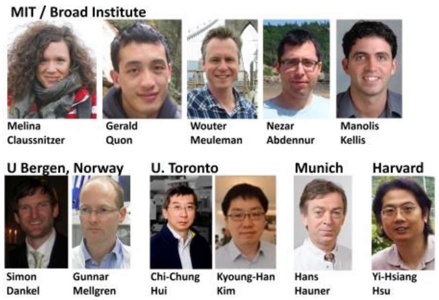




Lean

Obese

Speliotes NG 2010



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 — SEPTEMBER 3, 2015 — VOL. 373 NO. 10

FTO Obesity Variant Circuitry and Adipocyte Browning in Humans

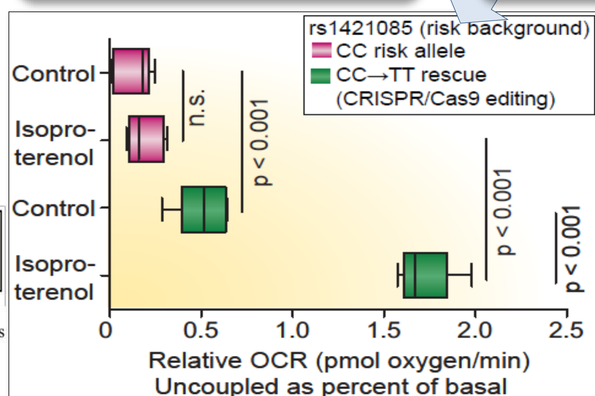
Melina Claussnitzer, Ph.D., Simon N. Dankel, Ph.D., Kyoung-Han Kim, Ph.D., Gerald Quon, Ph.D., Wouter Meuleman, Ph.D., Christine Haugen, M.Sc., Viktoria Glunk, M.Sc., Isabel S. Sousa, M.Sc., Jacqueline L. Beaudry, Ph.D., Vijitha Puviindran, B.Sc., Nezar A. Abdennur, M.Sc., Jannel Liu, B.Sc., Per-Arne Svensson, Ph.D., Yi-Hsiang Hsu, Ph.D., Daniel J. Drucker, M.D., Gunnar Mellgren, M.D., Ph.D., Chi-Chung Hui, Ph.D., Hans Hauner, M.D., and Manolis Kellis, Ph.D.

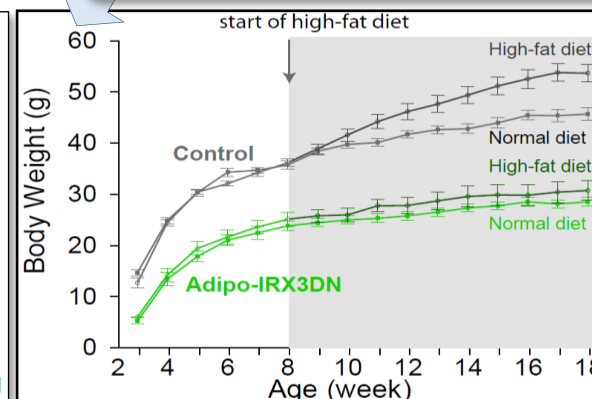
Claussnitzer, NEJM 2015

Incr. ARID5B ➔ Lean
Decr ARID5B ➔ Obese

C-to-T ➔ Lean
T-to-C ➔ Obese

Decrease IRX3, IRX5 ➔ Lean
Increase IRX3, IRX5 ➔ Obese

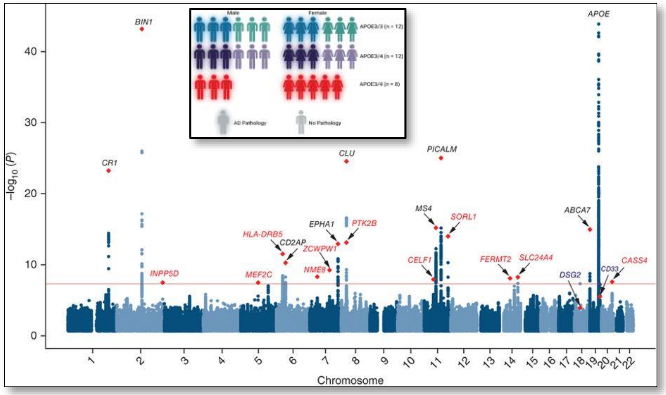




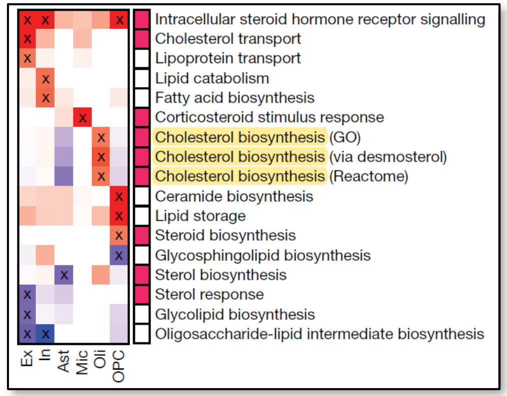
CRISPR-edit human fat cells
➔ able to burn calories again

IRX3 KD ➔ Burn calories in their sleep
➔ 54% weight loss. Can't gain weight

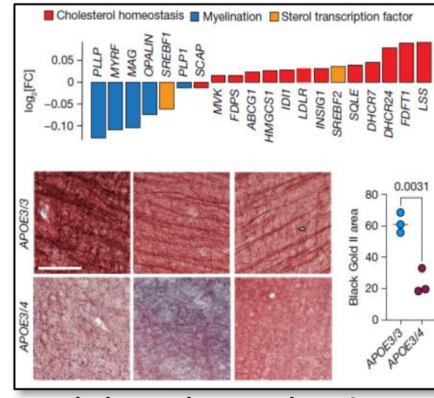
ApoE4 & Alzheimer's: Cholesterol transport ➔ Oligo ER accumulation ➔ Myelin ➔ Cognition

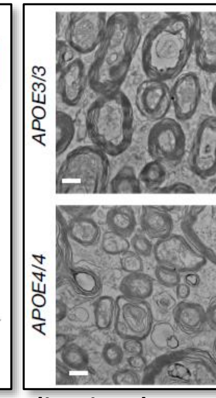


scRNA of ApoE33, ApoE34, ApoE44 individuals

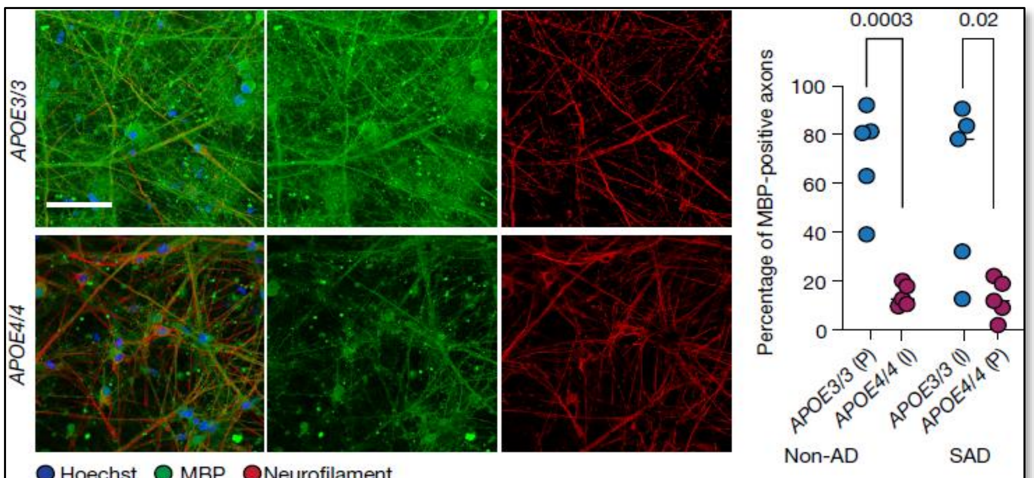


Cholesterol transport & biosynthesis in oligos



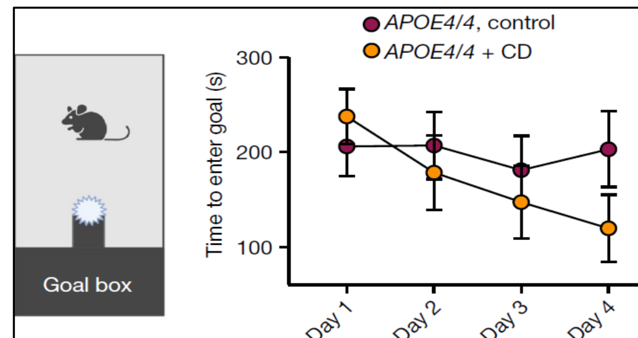


Cholesterol accumulates in ER, Myelination decrease

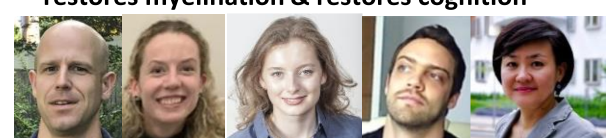


Causality: Lack of myelination recapitulated in ApoE4 iPSC-derived oligodendrocytes

Blanchard, Nature, 2022

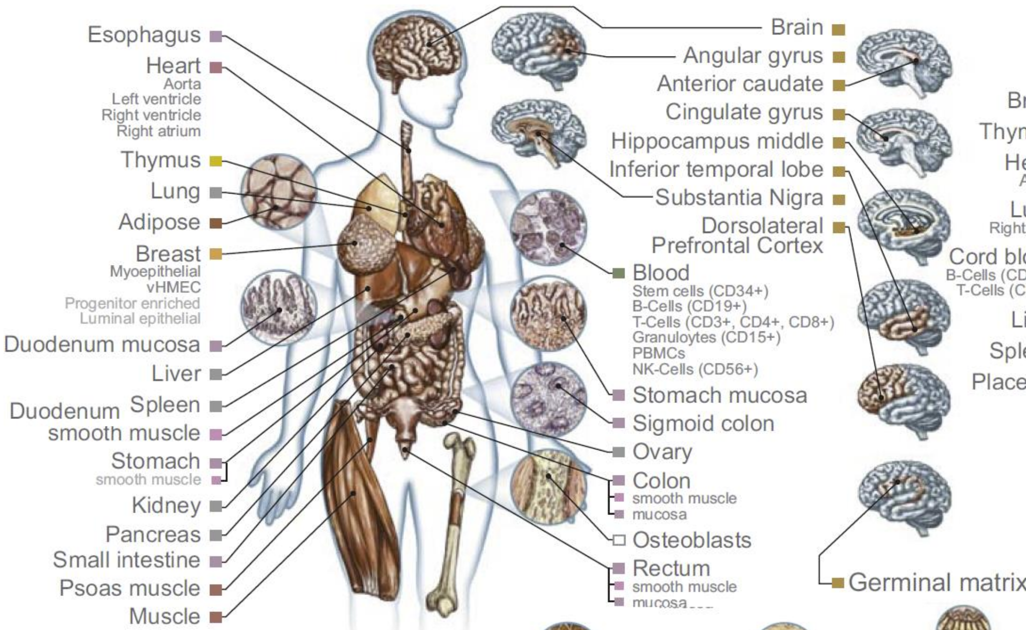


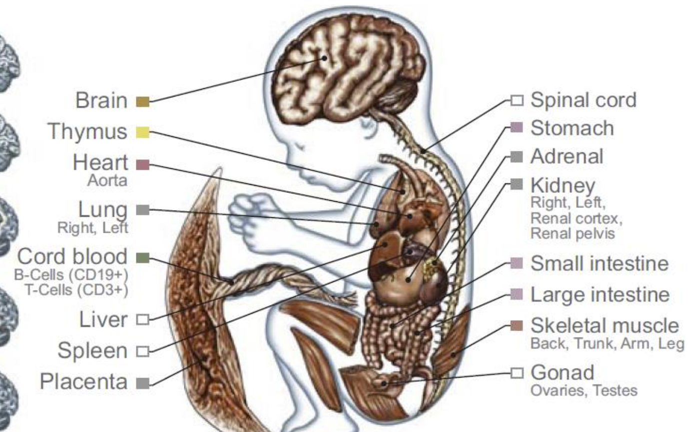
Restoring cholesterol transport (Cyclodextrine) restores myelination & restores cognition



With: Joel Blanchard, Leyla Akay, Jose Davila-Velderrain, Djuna von Maydel, Li-Huei Tsai

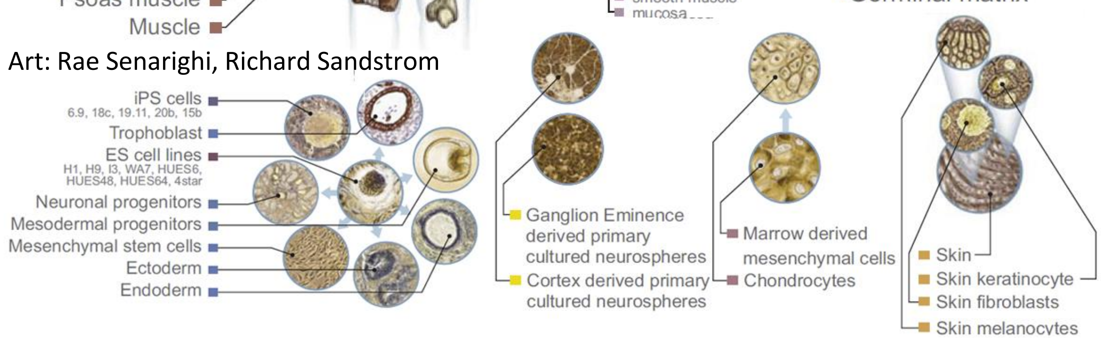
Epigenomics Roadmap across 100+ tissues/cell types





Roadmap Nature, 2015

Art: Rae Senarighi, Richard Sandstrom



Diverse epigenomic assays:

1. Histone modifications
- **H3K4me3, H3K4me1**
- **H3K36me3**
- **H3K27me3, H3K9me3**
- **H3K27/9ac, +20 more**

2. Open chromatin:
- **DNase**

3. DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

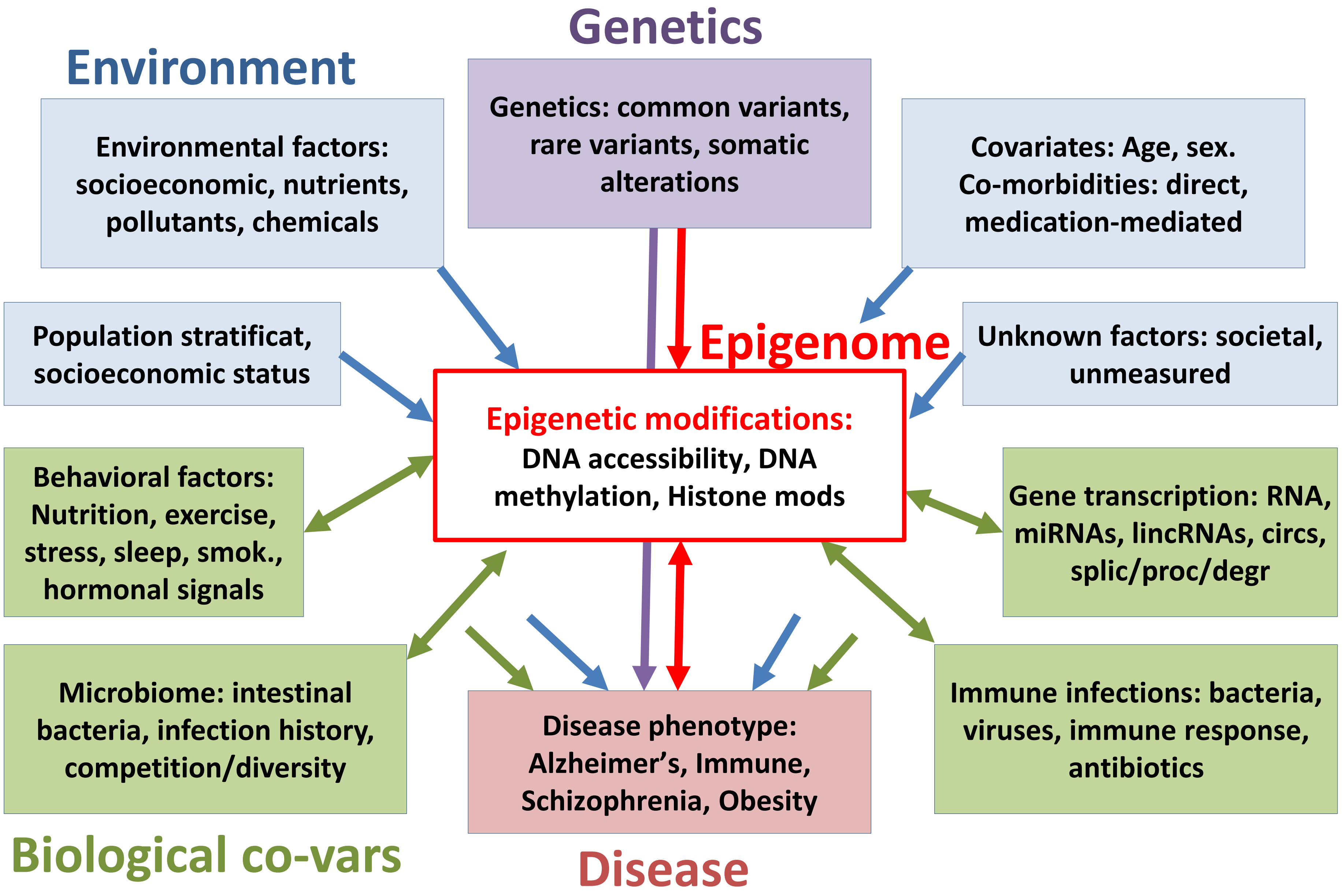
4. Gene expression
- **RNA-seq, Exon Arrays**

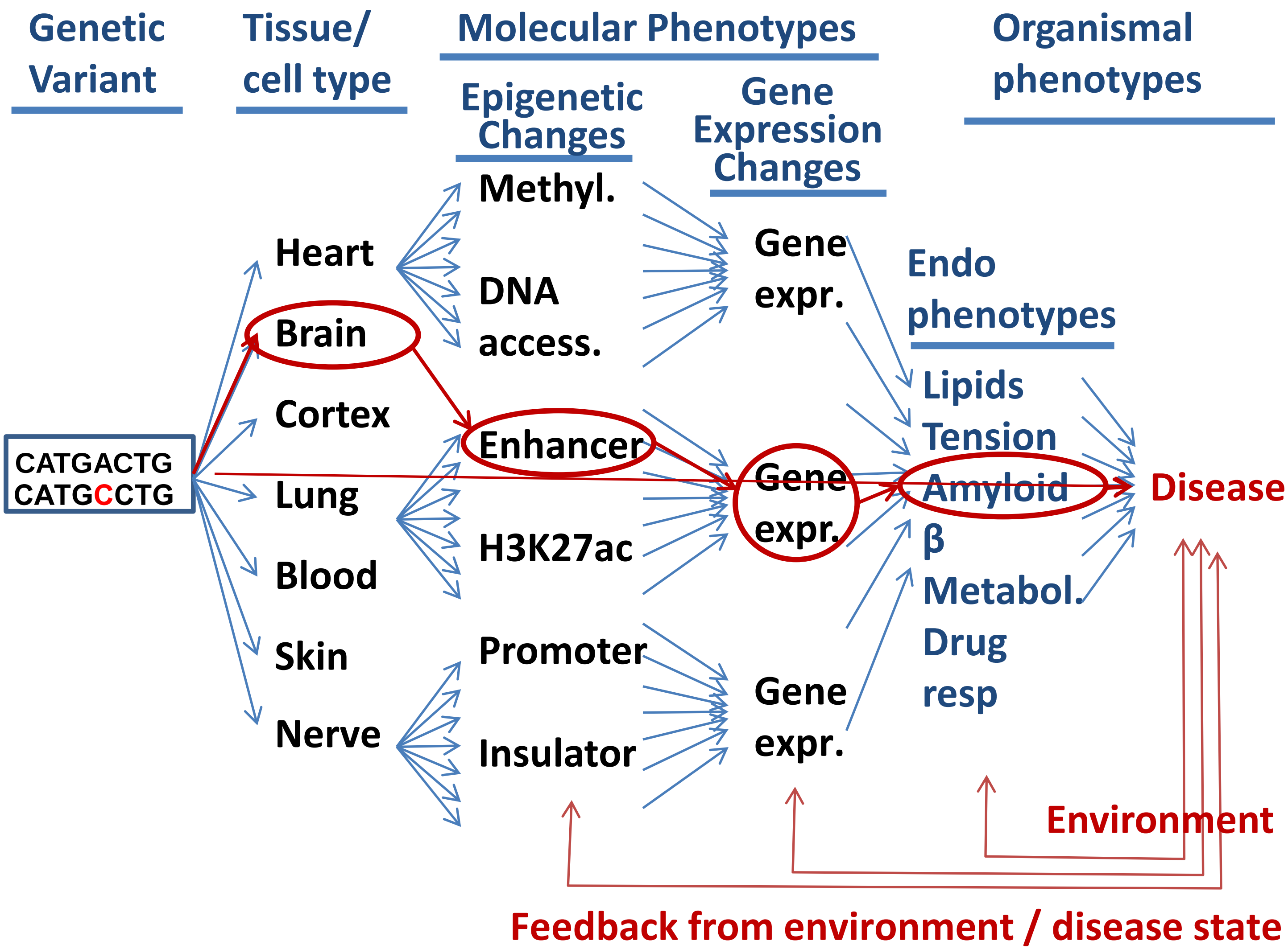
Diverse tissues and cells:

1. **Adult tissues and cells** (brain, muscle, heart, digestive, skin, adipose, lung, blood…)
2. **Fetal tissues** (brain, skeletal muscle, heart, digestive, lung, cord blood…)
3. **ES cells, iPS, differentiated cells** (meso/endo/ectoderm, neural, mesench, trophobl)

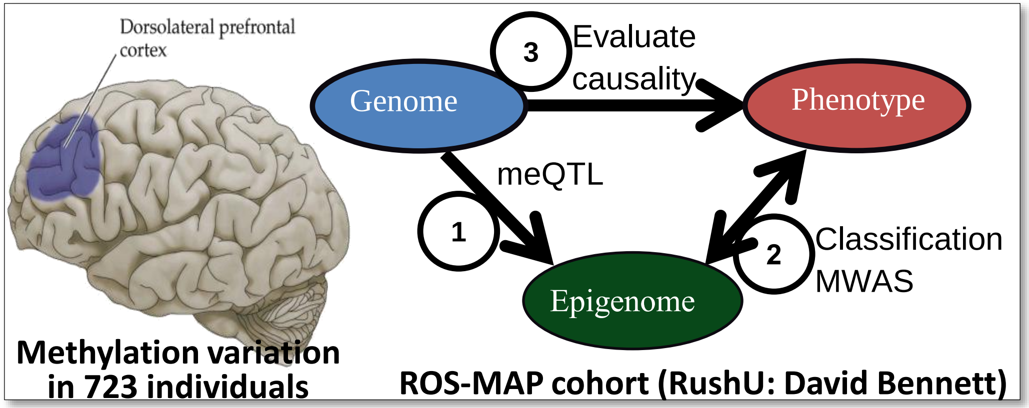
Roadmap Nature 15

Epigenome integrates genetic + env + dis + bio signals

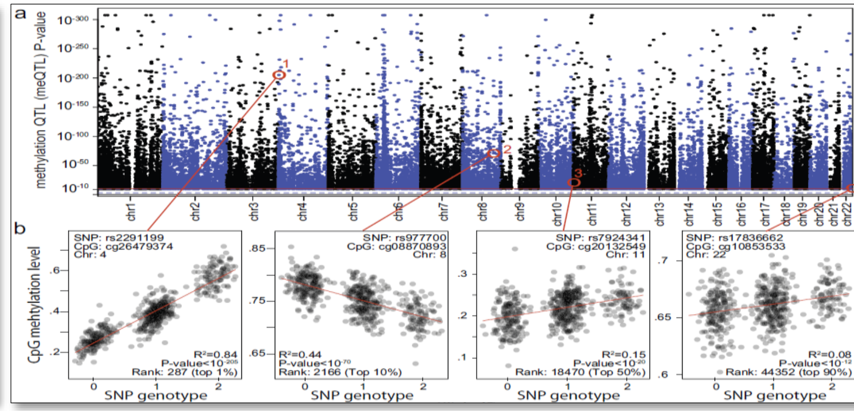



Genetic Variant
Tissue/ cell type
Molecular Phenotypes
Organismal phenotypes
Epigenetic Changes
Gene Expression Changes
CATGACTG
CATGCCTG
Heart
Brain
Cortex
Lung
Blood
Skin
Nerve
Methyl.
DNA access.
Enhancer
H3K27ac
Promoter
Insulator
Gene expr.
Gene expr.
Gene expr.
Endo phenotypes
Lipids
Tension
Amyloid β
Metabol.
Drug resp
Disease
Environment
Feedback from environment / disease state

Mediation analysis across 750 Alzheimer patients/controls: iMWAS



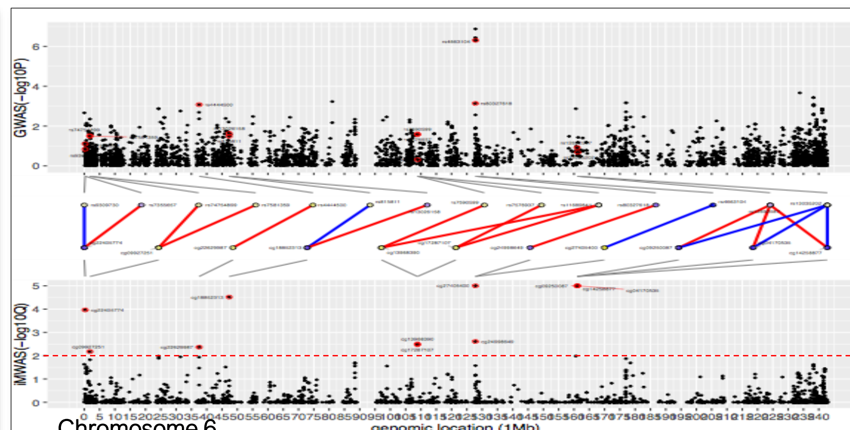
Relate: Genotype ⇔ Methylation ⇔ Phenotype



50,000 significant meQTLs (after Bonferroni)

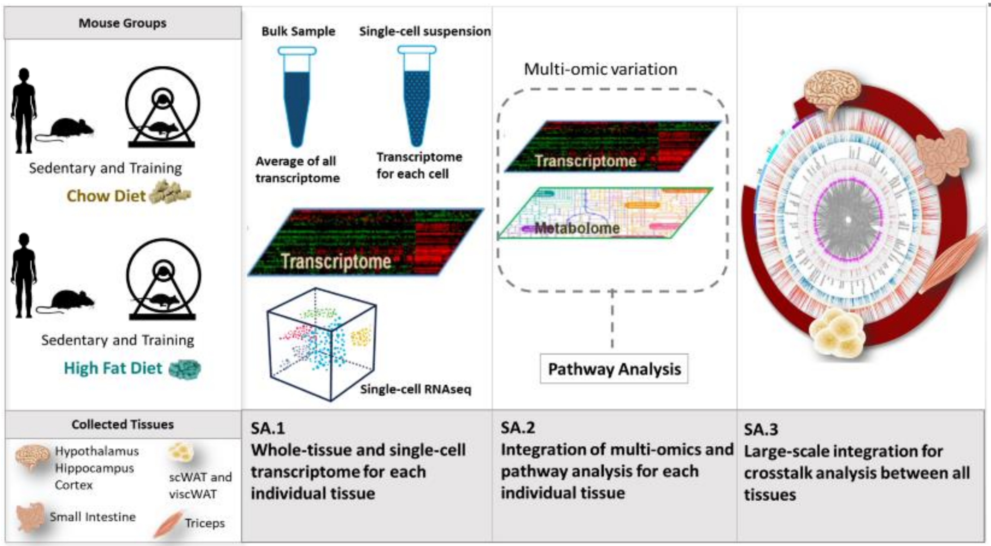
GWAS:	G ⟶ D		N=74k	**Learn G→D directly (complex phenotype)**
meQTL:	G → M		N=800	**Learn G→M (simpler phenotype)**
MWAS:	M ⟷ D		N=800	**M⇔D (no causality)**
iMWAS:	G → iM ⟶ D		N=74k	**Apply G→M to get iM iM→D (causality)**

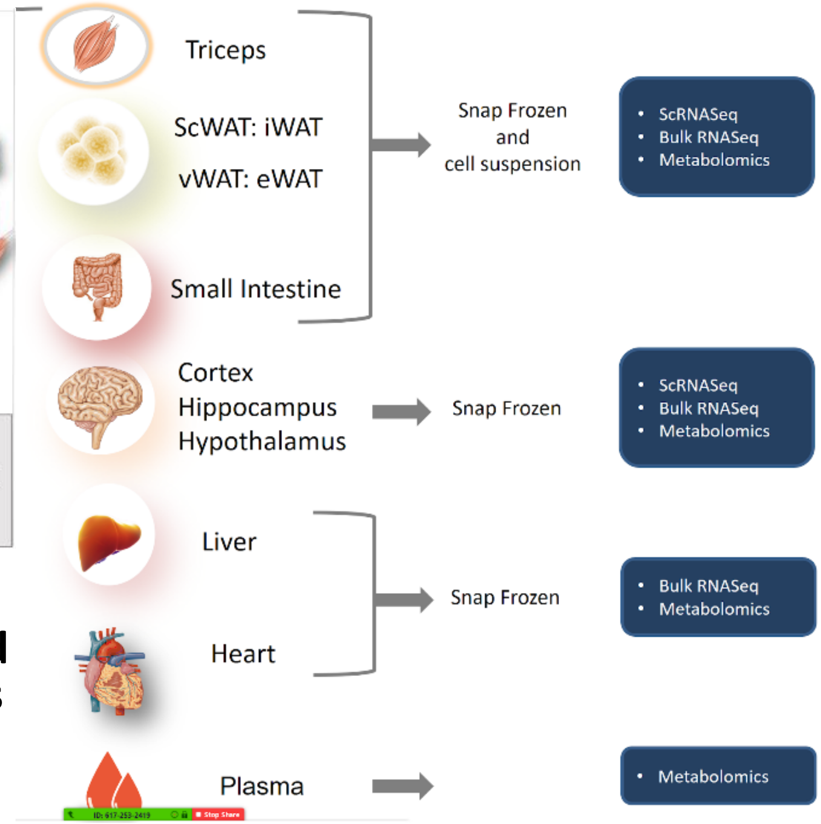
Imputed MWAS: incr. power, causality



iMWAS: new loci, multi-SNP effects

Multi-tissue multi-omics of exercise/diet in hum/mou





- Omics: Transcriptomic/epigenomic/metabolic
- Tis:Muscle, fat, digest, brain, liver, heart, blood
- Ph: Exercise-sedentary vs. diet-overeating axes
- Species: Human/mouse parallel studies
- Cell type: scRNA/scATAC in each, imm. enrich


Lydia Lynch


Laurie Goodyear

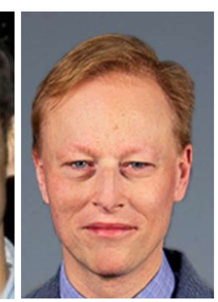

Maria Vamvini


Leandro Agudelo


Jackie Yang


Na Sun

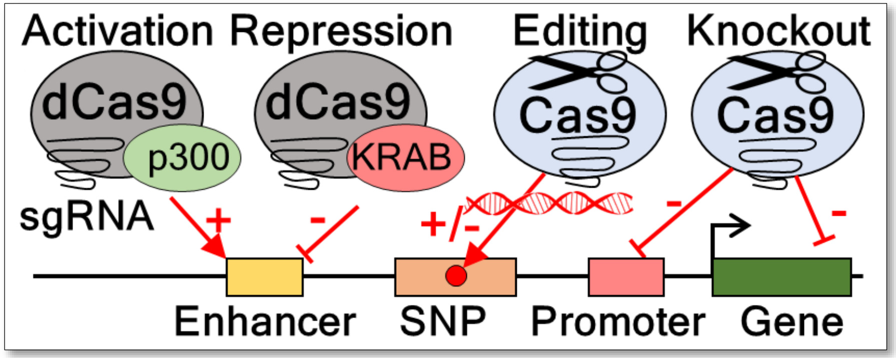

Pasquale Nigro


Jan-Willem Middelbeek

Y
Cell

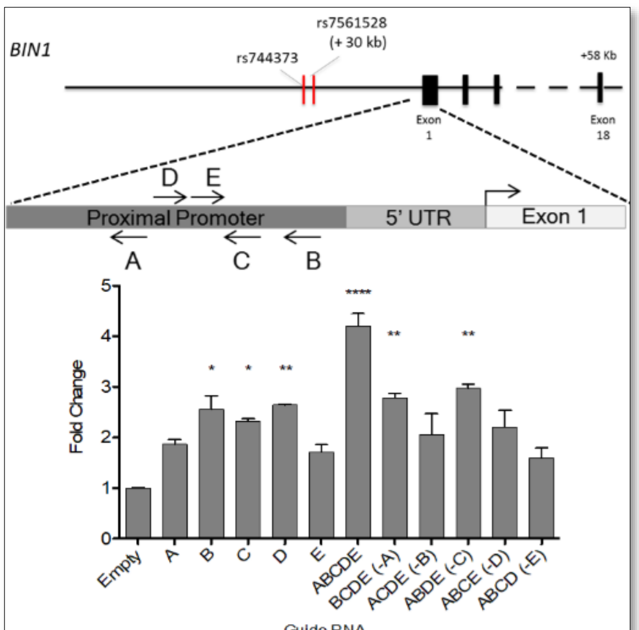
Yang, Vamvini, Nigro *et al*, In revisions

Modular and programmable CRISPR-Cas9/dCas9 system

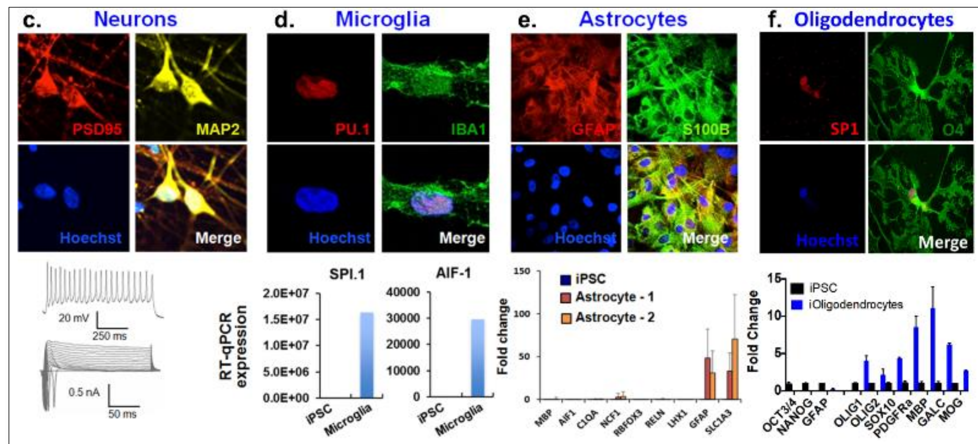


Modularity:
- Pick perturbation type (3 lines)
- Pick cell type (differentiation)
- Pick target (sgRNA + repair template)
- Induce (Dox/Tet control)
- Environmental modulations (+Aβ)
- Cross-cell-type effects (2D/3D co-culture)

- ***Activation: CRISPR-dCas9+p300***
- ***Repression: CRISPR-dCas9+KRAB***
- ***Editing: CRISPR-Cas9 + repair template***
- ***Knockout: CRISPR-Cas9 cutting***



AD: Bin1 enhancer activation
with multiple sgRNAs



Apply in iPSCs, differentiate into NPCs, neurons, astrocytes, oligodendrocytes, microglia

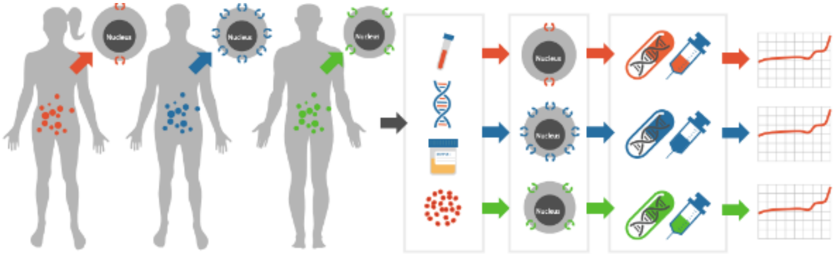
Collaboration with Li-Huei Tsai Lab

Disease still reigns

- ➢ **My own family:** Obesity, cancer, stroke, diabetes
- ➢ **My own predispositions:** obesity, blindness, cancer.
- ➢ **Genetics:** Each of us in this room carries mutations
- ➢ **Environment:** pollution, nutrition, sedentary lifestyle
- ➢ **Systemic disorders**: obesity, diabetes, cancer, heart
- ➢ **Pathogens**: infections, immune dysregulation, cancer
- ➢ **Lifespan**: Alzheimer's, new diseases

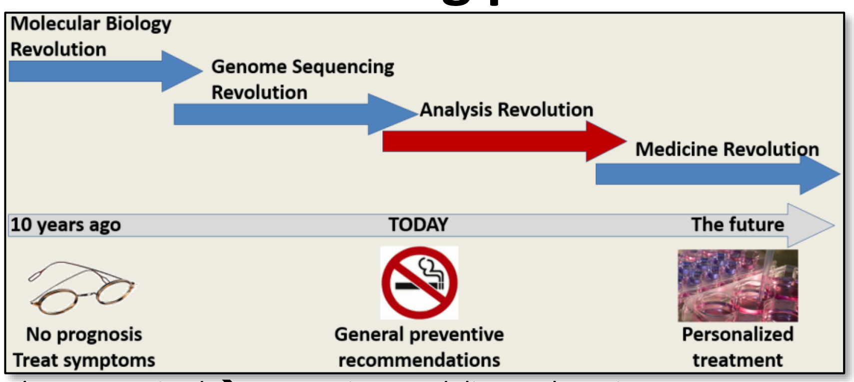
Kellis TEDx 2010

Personalizing Medicine



- ➢ **Polygenicity:** Thousands of variants
- ➢ **Convergence:** Small number of common pathways
- ➢ **Hallmarks** of disease: causal pathways
- ➢ **Manipulation**: reverse disease circuitry
- ➢ **Individualized** treatment: combine pathways
- ➢ **Each Patient**: different combination
- ➢ **Burden**: Accumulation of pathway perturbations
- ➢ **Omics**: Genetic, epigenomic, transcript, proteomic

Transforming pharma



- Always surprised ➔ Prognosis: Mendelian, Polygenic Scores
- Misdiagnosis ➔ Better biomarkers, Multi-modal diagnosis
- Treat manifestations ➔ Address root causes, causal hallmarks
- Monolithic: AD,T2D,Cancer ➔ Heterogeneity: symptoms+causes
- Monolithic: AD ➔ Understand components: Ab, tau, infl, lipids
- Silos: tissues, departments ➔ Interplay, commonalities, sharing
- Treatment too late ➔ Preventive personalized interventions

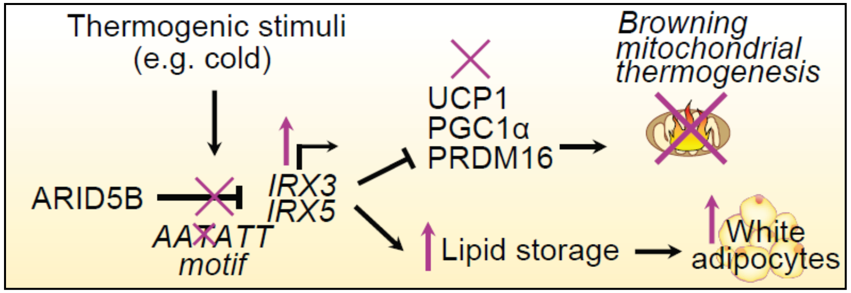
Kellis TEDx 2021

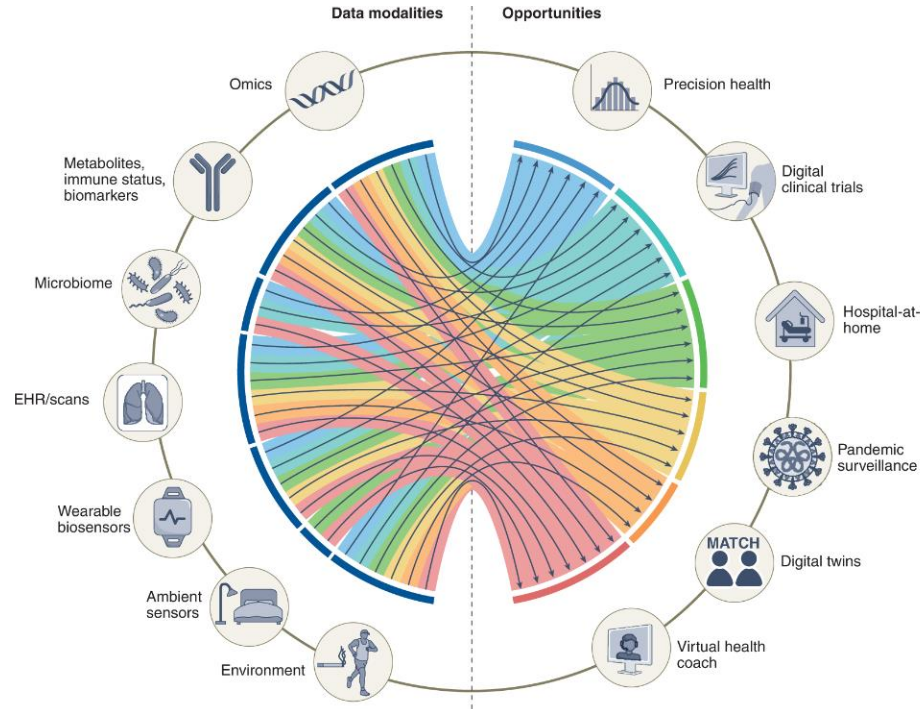
Call to action: Coalition

- ➢ **CS**: ML, DeepNN, DNA code, circuitry, big data
- ➢ **Bio**: High-throughput profiling + manipulation
- ➢ **Chemistry**: Libraries, synthesis, modularity
- ➢ **Biotech**: New technol. for rewiring, delivery
- ➢ **Finance**: long-term 10-year 20-year 'biobonds'
- ➢ **Pharma**: partnership, pre-competitive sharing
- ➢ **Patients**: empowrmnt, personalization, sharing
- ➢ **Hospitals**: combine cohorts, increase power

Deep Learning, Circuitry Inference, Decoding of Human Health

- Human body as a dynamic reconfigurable system
  - Systems circuitry view of human health
- Disease prognosis and early intervention and measurement before symptoms appear
- Use of biomarkers indicative and prognostic of disease onset
- Biomarker modeling of intervention to gauge success of treatments intervention
- Dynamic sensor modeling and selection of data to measure
- Electronic health record data mining with awareness of bias in data gathering
  - Doctors will prescribe specific tests that they expected to have abnormal results
  - Thus, the values distribution of measured variables is dramatically different from that of unmeasured variables
- Treatment interventions are guided by biomarkers that assess disease statues leading to coupling of positive and negative outcomes with opposite predictive values
- Measuring differences between observed outcomes of gene expression levels and predicted outcomes based on genetic variables to infer the impact of environmental effects
- Deep learning and translation between gene expression
- Systems-level convergence of mutations
- Smart sensors and system monitoring for human health
- The quantified life





Outline for today's class

1. Brief history of AI and ML in healthcare
2. Interlude: Student & faculty introductions
3. Why *now*? What has changed?
4. Examples of how ML will transform healthcare
5. **What is *unique* about ML in healthcare?**
6. Course logistics & syllabus

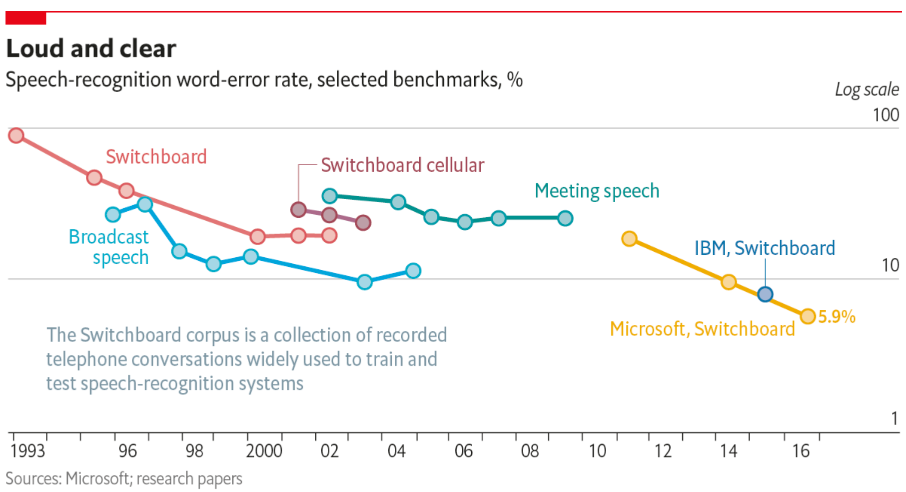
What makes healthcare different?

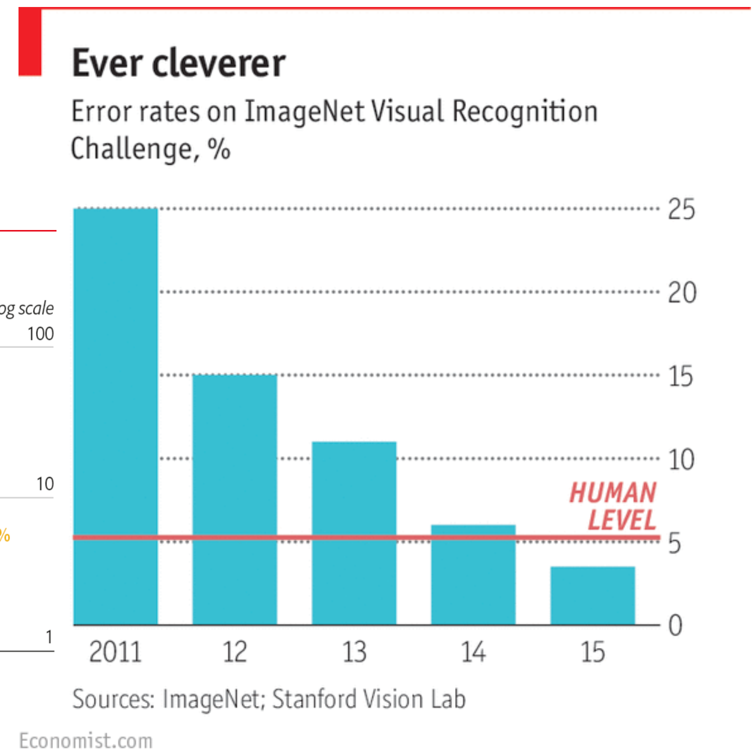
- Life or death decisions
  - Non-fungibility of patients
  - Need **robust** algorithms
  - Checks and balances built into ML deployment
  - (Also arises in other applications of AI such as autonomous driving)
  - Need **fair** and **accountable** algorithms
- Many questions are about unsupervised learning
  - Discovering disease subtypes, or answering question such as “characterize the types of people that are highly likely to be readmitted to the hospital”?
- Many of the questions we want to answer are *causal*
  - Naïve use of supervised machine learning is insufficient

What makes healthcare different?

- Very little labeled data

Recent breakthroughs in AI depended on *lots* of labeled data!





What makes healthcare different?

- Very little labeled data
  - Motivates semi-supervised learning algorithms
- Sometimes small numbers of samples (e.g., a rare disease)
  - Learn as much as possible from other data (e.g. healthy patients)
  - Model the problem carefully
- Lots of missing data, varying time intervals, censored labels

What makes healthcare different?

- Difficulty of de-identifying data
  - Need for data sharing agreements and sensitivity
- Difficulty of deploying ML
  - Commercial electronic health record software is difficult to modify
  - Data is often in silos; everyone recognizes need for interoperability, but slow progress
  - Careful testing and iteration is needed

Goals for the semester

- Intuition for working with healthcare data
- How to set up as machine learning problems
- Understand which learning algorithms are likely to be useful and when
- Appreciate subtleties in safely & robustly applying ML in healthcare
- Set the research agenda for the next decade

Outline for today's class

1. Brief history of AI and ML in healthcare
2. Interlude: Student & faculty introductions
3. Why *now*? What has changed?
4. Examples of how ML will transform healthcare
5. What is *unique* about ML in healthcare?
6. **Course logistics & syllabus**

Prerequisites

- Previous undergraduate-level ML (e.g. $6.390_{[6.036]}$):
  - Machine learning methodology (e.g. generalization, cross-validation)
  - Supervised machine learning techniques (e.g. linear and logistic regression, neural networks)
  - Loss functions, regularization, and optimization (e.g. stochastic gradient descent)
  - Statistical modeling (e.g. Gaussian mixture models)

  Want a quick review? See Videos 1, 2, 3, 4 (~1 hr.)
- Python

Logistics

- Course website: https://mlhcmit.github.io/

- All announcements made via Canvas
- Use Piazza for Q&A with staff and each other

- Recitation (required): Fridays 3-4pm in 4-270 (starts this week)
- Office hours TBD

Grading

- 40% course project
- 35% homework (~4 problem sets; both theory & practice)
- 20% final exam (date to be scheduled by registrar)
- 5% participation – note: class attendance is required*

* Exceptions will be made for quarantine/isolation.

This week’s assignments/readings

- PS0 (due **Tue, 2/14 11:59pm**): human subjects training & MIMIC data use agreement
- Reading response (due **Fri 2/4, 1pm)**

AI in Health and Medicine – required reading
Pranav Rajpurkar, Emma Chen, Oishi Banerjee & Eric J. Topol
Nature Medicine, 2022

Machine Learning in Medicine – optional reading
Alvin Rajkomar, Jeffrey Dean, Isaac Kohane
New England Journal of Medicine, 2019

Course project

- Teams of ~4-8 students each
- Each project will have one or more clinicians involved as mentors and/or students

- Project descriptions during class of **Feb 16**
- Project poster presentations **May 16**