PH240C: Scope and Introduction

Jingshen Wang

09/01/2021

Class Policy

Homework assignments (65%):

- Biweekly
- The lowest score will be dropped in the final grade, and one late homework (24 hour) is allowed.
- It is encouraged to discuss the problem sets with others, but everyone needs to turn in a unique personal write-up.

Final project write-up and presentation(35%): take home, open books, open notes.

Labs

Supervised Learning (classical approaches)

1. GLM/SVM (09/08)
2. Kernel-based Methods (09/15)
3. Metric Learning (09/22)
4. Tree-based Methods (09/29, 10/06)

Semi-supervised Learning (10/13)

Please fill in the lab time change pool!

3 Labs

1. Neural Networks (10/20)
2. Deep Neural Networks (10/27)

1 Labs

Causal Inference and Clinical Trials

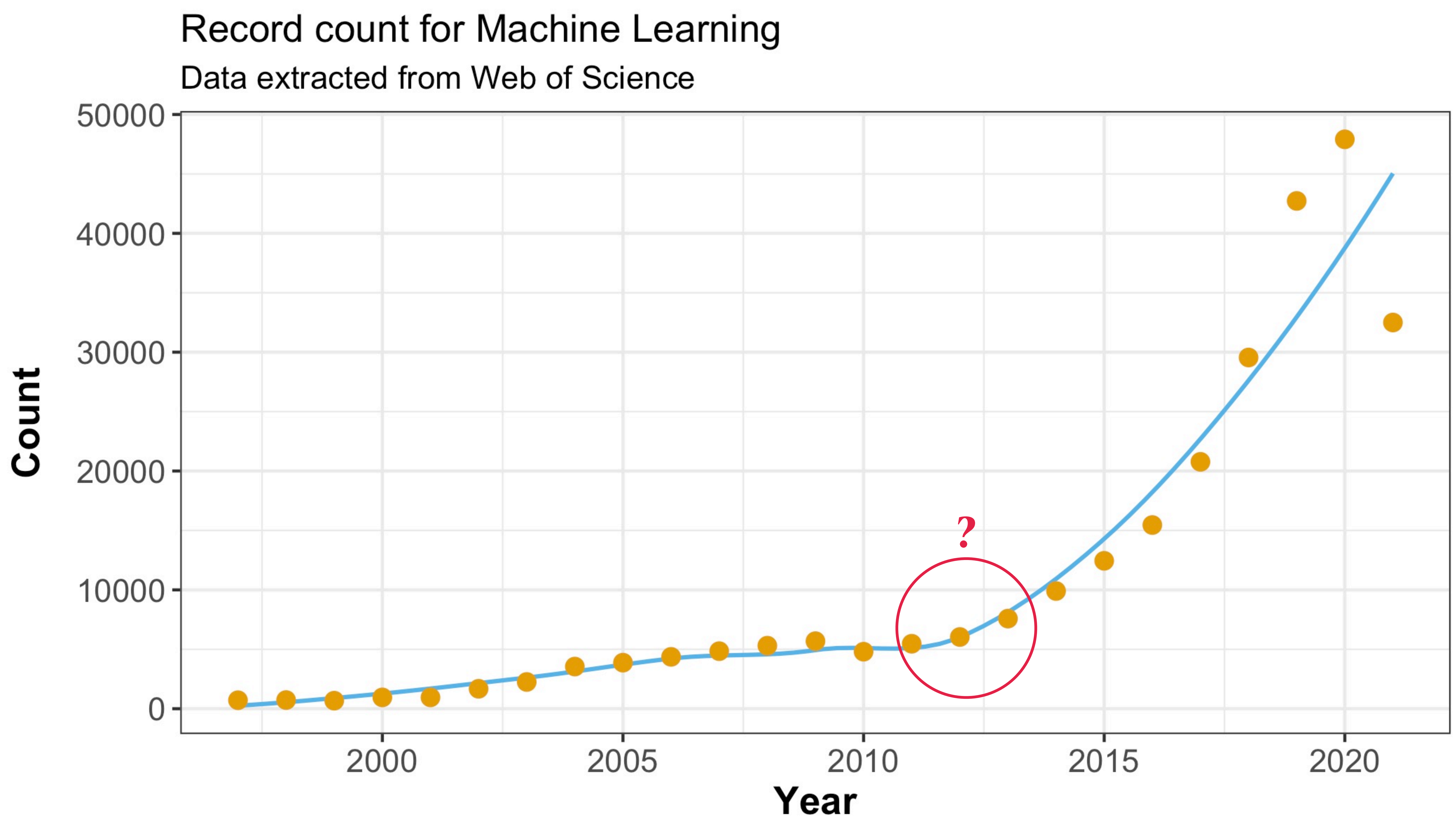
1. Nature's Experiments: Mendelian Randomization (11/10)
2. Bayesian Inference and Design of Experiments (11/17)
3. Adaptive Clinical Trial and Reinforcement Learning (12/01)

2 Labs

What is Machine Learning?

- The concept of ML is not new

- Arthur Samuel (1959): Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed



Change Points of ML — Image Recognition

Imagenet classification with deep convolutional neural networks

[PDF] Imagenet classification with deep convolutional neural networks
A Krizhevsky, I Sutskever… - Advances in neural …, 2012 - proceedings.neurips.cc
We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000 …
Cited by 85795 Related articles All 121 versions

Showing the best result for this search. See all results

The trained deep convolutional neural network provides
much more accurate prediction than the previous methods

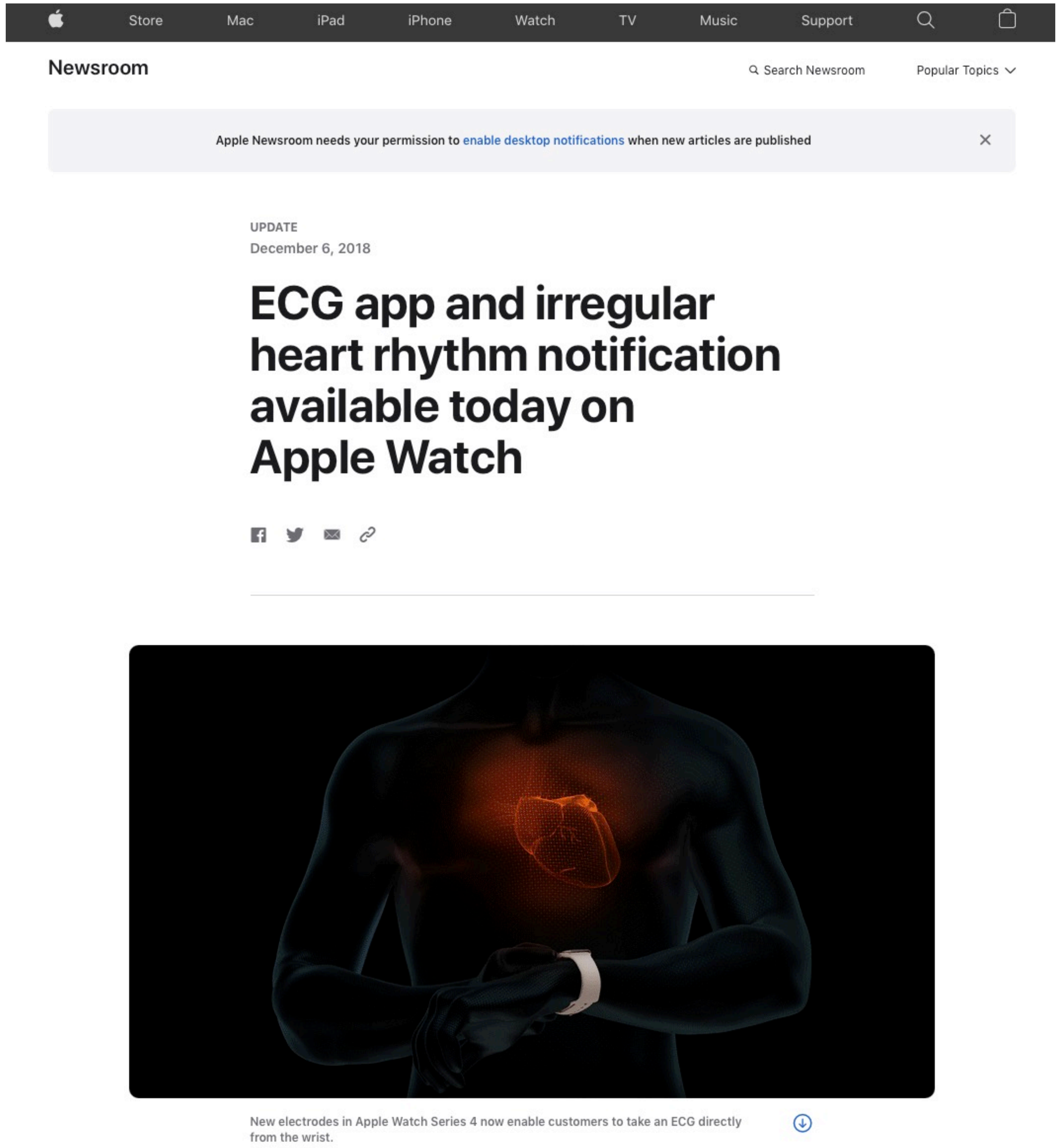
What does it mean for medicine and healthcare?

Studies relied on Deep Learning have showcased its ability to

- Diagnose some types of skin cancer
- Identify specific heart-rhythm abnormality like cardiologists
- Interpret medical scans or pathology slides like highly qualified radiologists
- Diagnose various eye disease as well as ophthalmologist
- ....

Nevertheless, machine learning methods need to be powered by big data

One example



Store Mac iPad iPhone Watch TV Music Support

Newsroom

Search Newsroom Popular Topics

Apple Newsroom needs your permission to enable desktop notifications when new articles are published

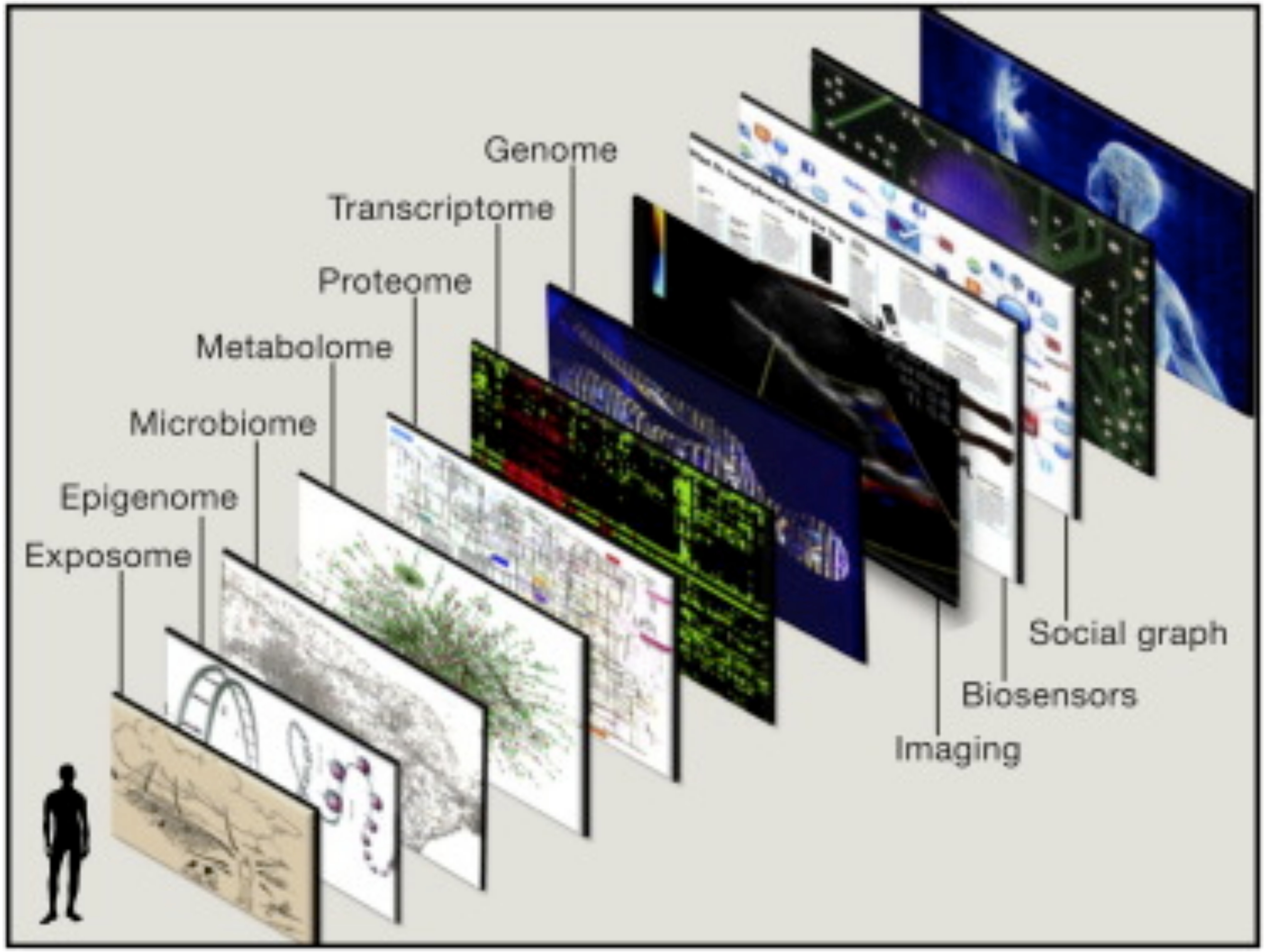
UPDATE

December 6, 2018

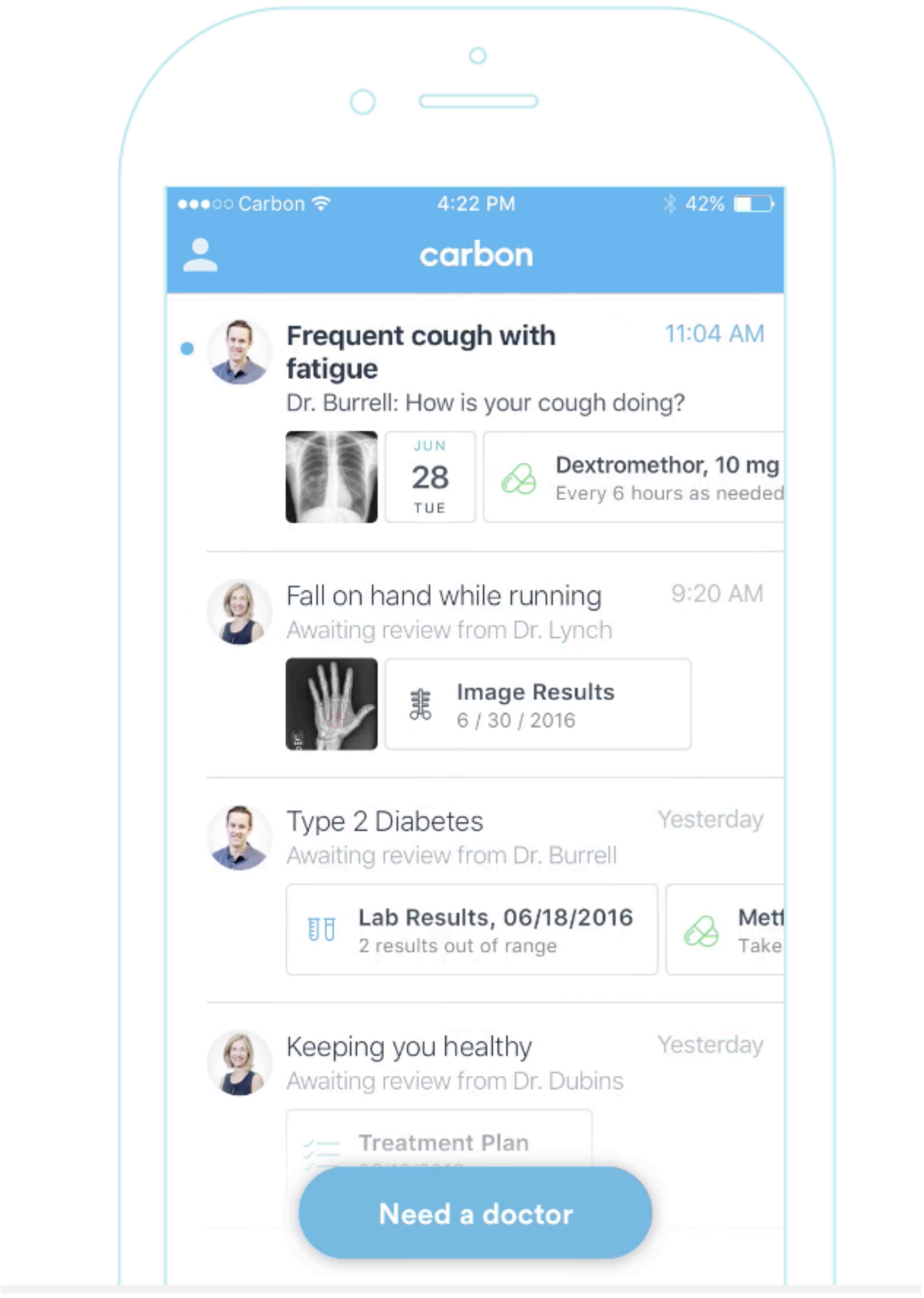
ECG app and irregular heart rhythm notification available today on Apple Watch

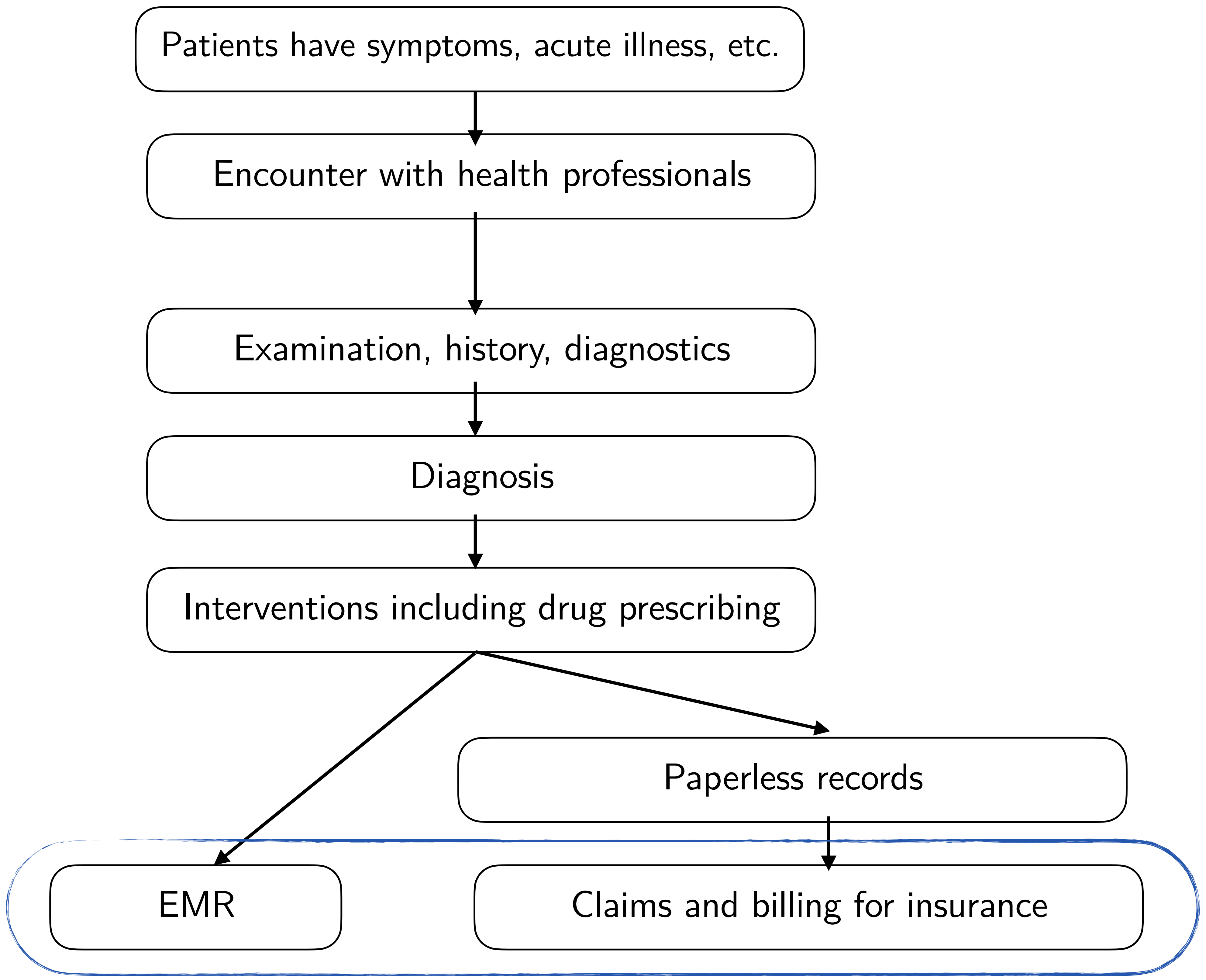
New electrodes in Apple Watch Series 4 now enable customers to take an ECG directly from the wrist.

Geographic Information System of a Human Being



Where do Data Come from?





Research Data Base: Electronic Health Record Data

Electronic Health Record Data

EHR stands for electronic health record. According to Wikipedia:

- EHR is the systematized collection of patient and population electronically-stored health information in a digital format.
- These records can be shared across different health care settings. Records are shared through network-connected, enterprise-wide information systems or other information networks and exchanges.

Reality in EHR data

PatientID	AdmissionID	PrimaryDiagnosisCode	PrimaryDiagnosisDescription
E74E9DF1-D8FD-41BC-8CDE-226CFE318E0B	1	E09.42	Drug or chemical induced diabetes mellitus with neurologi complications with diabetic polyneuropathy
E74E9DF1-D8FD-41BC-8CDE-226CFE318E0B	2	O29.123	Cardiac failure due to anesthesia during pregnancy, third trimester
E74E9DF1-D8FD-41BC-8CDE-226CFE318E0B	3	M84.561	Pathological fracture in neoplastic disease, right tibia
3AB69ECE-65F4-4D04-9E87-54E73C2DB4A8	1	G52.3	Disorders of hypoglossal nerve
3AB69ECE-65F4-4D04-9E87-54E73C2DB4A8	2	C40.31	Malignant neoplasm of short bones of right lower limb
3AB69ECE-65F4-4D04-9E87-54E73C2DB4A8	3	C67.5	Malignant neoplasm of bladder neck
3AB69ECE-65F4-4D04-9E87-54E73C2DB4A8	4	O98.612	Protozoal diseases complicating pregnancy, second trimest
3AB69ECE-65F4-4D04-9E87-54E73C2DB4A8	5	Z13.85	Encounter for screening for nervous system disorders
3AB69ECE-65F4-4D04-9E87-54E73C2DB4A8	6	Z12.2	Encounter for screening for malignant neoplasm of respiratory o
3AB69ECE-65F4-4D04-9E87-54E73C2DB4A8	7	Y71	Cardiovascular devices associated with adverse incidents
1F31FFA4-8EAB-43F9-9C75-B5745E7908CA	1	M90.661	Osteitis deformans in neoplastic diseases, right lower le
1F31FFA4-8EAB-43F9-9C75-B5745E7908CA	2	C79.82	Secondary malignant neoplasm of genital organs
1F31FFA4-8EAB-43F9-9C75-B5745E7908CA	3	D89.81	Graft-versus-host disease
43556DC2-BCFC-45A8-84C3-1D3E4A11B02F	1	M01.X31	Direct infection of right wrist in infectious and parasit diseases classified elsewhere
43556DC2-BCFC-45A8-84C3-1D3E4A11B02F	2	Z22.33	Carrier of bacterial disease due to streptococci
43556DC2-BCFC-45A8-84C3-1D3E4A11B02F	3	M63.82	Disorders of muscle in diseases classified elsewhere, upp
43556DC2-BCFC-45A8-84C3-1D3E4A11B02F	4	O10.12	Pre-existing hypertensive heart disease complicating chil
43556DC2-BCFC-45A8-84C3-1D3E4A11B02F	5	O9A.12	Malignant neoplasm complicating childbirth
43556DC2-BCFC-45A8-84C3-1D3E4A11B02F	6	O9A.11	Malignant neoplasm complicating pregnancy
43556DC2-BCFC-45A8-84C3-1D3E4A11B02F	7	G31.83	Dementia with Lewy bodies
9015B511-F1C9-47F4-AA23-E33E8829151F	1	C90.11	Plasma cell leukemia in remission
9015B511-F1C9-47F4-AA23-E33E8829151F	2	C18.6	Malignant neoplasm of descending colon
9015B511-F1C9-47F4-AA23-E33E8829151F	3	E72.5	Disorders of glycine metabolism

Doctors' Diagnosis (NLP)

PatientID	AdmissionID	LabName	LabValue	LabUnits	LabDateTime
915BC24E-8C44-4D33-A386-CEA965B83F32	1	CBC: HEMATOCRIT	40.7	%	1946-09-07 22:20:26.677
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: ANION GAP	8.4	mmol/L	1946-09-07 11:52:58.600
915BC24E-8C44-4D33-A386-CEA965B83F32	1	CBC: LYMPHOCYTES	4.7	k/cumm	1946-09-07 06:08:57.303
915BC24E-8C44-4D33-A386-CEA965B83F32	1	CBC: HEMOGLOBIN	15.9	gm/dl	1946-09-07 19:16:10.057
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: SODIUM	146.6	mmol/L	1946-09-07 14:03:11.003
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: ALBUMIN	3.3	gm/dL	1946-09-08 04:06:53.967
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: BUN	17.1	mg/dL	1946-09-07 19:41:39.247
915BC24E-8C44-4D33-A386-CEA965B83F32	1	CBC: NEUTROPHILS	8.4	k/cumm	1946-09-07 05:33:34.320
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: CALCIUM	8.7	mg/dL	1946-09-07 16:23:41.213
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: GLUCOSE	110.5	mg/dL	1946-09-07 09:40:39.467
915BC24E-8C44-4D33-A386-CEA965B83F32	1	URINALYSIS: PH	5.5	no unit	1946-09-07 05:37:45.880
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: BILI TOTAL	0.3	mg/dL	1946-09-08 02:06:16.087
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: POTASSIUM	5.3	mmol/L	1946-09-07 10:18:37.693
915BC24E-8C44-4D33-A386-CEA965B83F32	1	URINALYSIS: RED BLOOD CELLS	2.2	rbc/hpf	1946-09-07 09:57:38.487
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: CARBON DIOXIDE	25.5	mmol/L	1946-09-07 08:22:44.910
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: CREATININE	0.6	mg/dL	1946-09-07 22:51:11.803
915BC24E-8C44-4D33-A386-CEA965B83F32	1	URINALYSIS: SPECIFIC GRAVITY	1	no unit	1946-09-07 20:46:39.873
915BC24E-8C44-4D33-A386-CEA965B83F32	1	CBC: MEAN CORPUSCULAR VOLUME	97.5	fl	1946-09-08 00:42:35.567
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: CHLORIDE	109.1	mmol/L	1946-09-07 05:04:43.103
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: ALT/SGPT	39.9	U/L	1946-09-07 15:50:50.170
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: AST/SGOT	14.7	U/L	1946-09-07 07:14:23.897
915BC24E-8C44-4D33-A386-CEA965B83F32	1	METABOLIC: ALK PHOS	44.8	U/L	1946-09-07 10:36:07.923
915BC24E-8C44-4D33-A386-CEA965B83F32	1	CBC: EOSINOPHILS	0.2	k/cumm	1946-09-07 17:56:12.167
915BC24E-8C44-4D33-A386-CEA965B83F32	1	CBC: ABSOLUTE NEUTROPHILS	67.5	%	1946-09-07 17:29:55.997

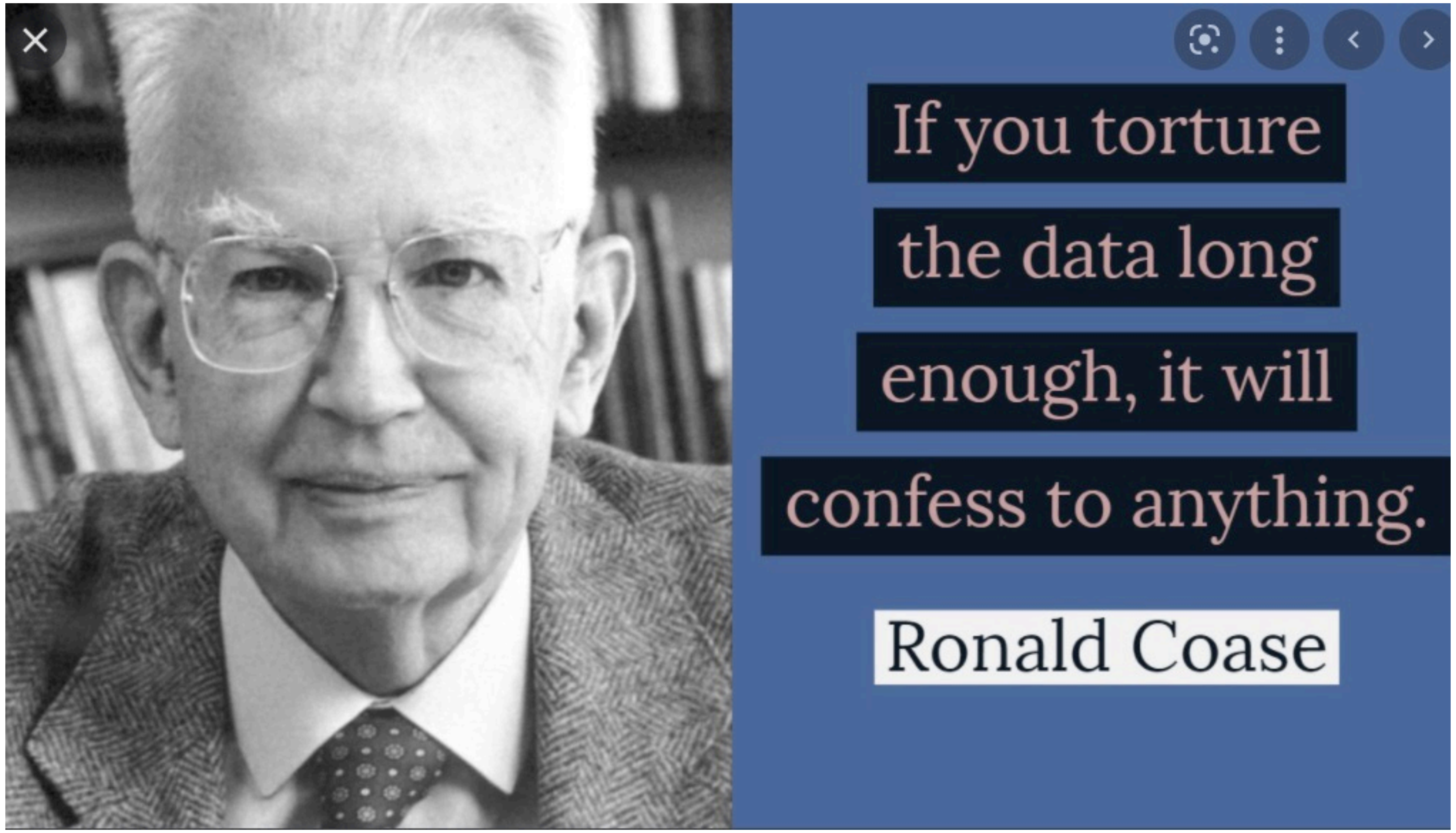
Lab Results

Need to be linked

PatientID	PatientGender	PatientDateOfBirth	PatientRace	PatientMaritalStatus	PatientLanguage	PatientPopulationPercentageBelowPoverty
3A3C2AFB-FFFA-4E69-B4E6-73C1245D5D12	Male	1975-01-04 14:49:59.587	White	Single	Unknown	15.6
801AFB51-036F-40E3-BDFE-FED4844BE275	Male	1964-09-06 13:15:43.043	White	Unknown	English	13.23
366B0CC6-18AC-45DD-9AD4-BE884FE3A299	Female	1953-01-14 06:00:19.330	White	Married	English	12.11
DBB78149-D86C-435E-82C4-341999FD0719	Female	1986-04-28 12:42:02.007	Unknown	Unknown	Unknown	95.8
2C6269F4-71F8-4C07-A905-C08E9A3524C2	Female	1963-08-26 10:57:55.183	White	Unknown	English	18.08
07E4EDF1-3AC8-4A71-AC06-F5DDDD8032BA	Male	1958-11-08 01:27:12.597	White	Single	English	12.38
90A956B2-AAE0-4D39-8A5E-680B36E1EA46	Male	1957-12-27 14:00:00.410	African American	Single	English	19.47
00F8A4BE-3CE3-466A-A90D-0A246D80CE57	Female	1925-03-27 18:58:51.580	White	Divorced	Icelandic	13.47
4731F86A-9225-424F-9944-AEF4001D0C55	Male	1953-10-07 11:56:58.470	African American	Separated	English	19.97
4CF45EFD-ACB3-4199-A974-403F25D219F2	Female	1969-07-02 04:52:46.833	White	Married	Spanish	15.55
75E2A410-6B84-4F88-B23B-D8715B95C8A5	Male	1937-11-03 02:25:53.543	White	Married	Unknown	10.73
30449579-0C45-4047-A755-C13E09F32E23	Female	1938-05-19 13:34:32.603	White	Divorced	Spanish	12.93
8EBF98B7-F043-44DB-874F-DEFBE86176AF	Male	1923-08-13 01:20:41.103	Asian	Married	English	18.81
0F907289-895E-4B76-89E9-8E63C98EE144	Male	1984-06-17 10:56:51.350	White	Married	English	11.3
449A2AC6-170D-415F-9801-8D0FE5F71655	Male	1938-12-21 03:48:25.033	African American	Married	Spanish	17.74
193C35A1-2BF3-4305-A1A6-3006675821F1	Female	1986-01-08 21:43:35.747	Unknown	Married	English	19.96
2C214BAA-3E7E-4D67-A61A-9AF5366F3AAE	Male	1945-07-07 08:10:44.283	Unknown	Divorced	English	14.58
CD30BCDE-ABF6-42C3-B30C-796A86F114ED	Female	1967-12-26 20:41:11.683	African American	Married	English	13.51
82327435-2A63-45EF-AFDE-28A778E5AFAC	Male	1963-05-04 21:58:03.970	White	Separated	English	18.89
16A66D15-D206-4AE9-BA63-A73DAA7F9112	Female	1972-03-21 22:32:32.613	Unknown	Single	Spanish	19.05
2BD263A3-BF89-484F-8C69-AB201F9EDBE7	Male	1936-01-05 16:01:46.253	White	Single	English	14.14
C84E7F47-0BAE-48B1-84FE-92004D2D0A79	Female	1925-12-26 22:12:17.897	African American	Single	English	16.84
463530E1-3469-454E-BA01-8258B9FCFDA2	Female	1964-01-23 06:06:01.210	Unknown	Married	English	8.99
59E199F5-B3B8-4623-877F-49B405C906C6	Male	1989-10-11 07:33:04.960	White	Single	English	19.84

Baseline Biomarkers

What do we want to learn/gain from data?



What do we want to learn/gain from data?

Learning objects in this class:

1. Disease risk (early) prediction/diagnosis — Supervised Learning
2. Utilize massive undiagnosed patient information to improve prediction — Semi-supervised Learning
3. Guide future clinical decisions — causal inference and clinical trial design

We need tools to achieve these goals

Supervised Learning (classical approaches)

1. GLM/SVM
2. Kernel-based Methods
3. Metric Learning
4. Tree-based Methods

Example 1. The prime example: house price prediction

Suppose we have data about

(1) Features X — square footage, number of rooms, features, whether a house has a garden or not

(2) Labels/Outcomes Y — the prices of these houses

By leveraging data coming from thousands of houses, we can train a supervised machine learning model to predict a new house's price based on the examples observed by the model.

Example: Text as data

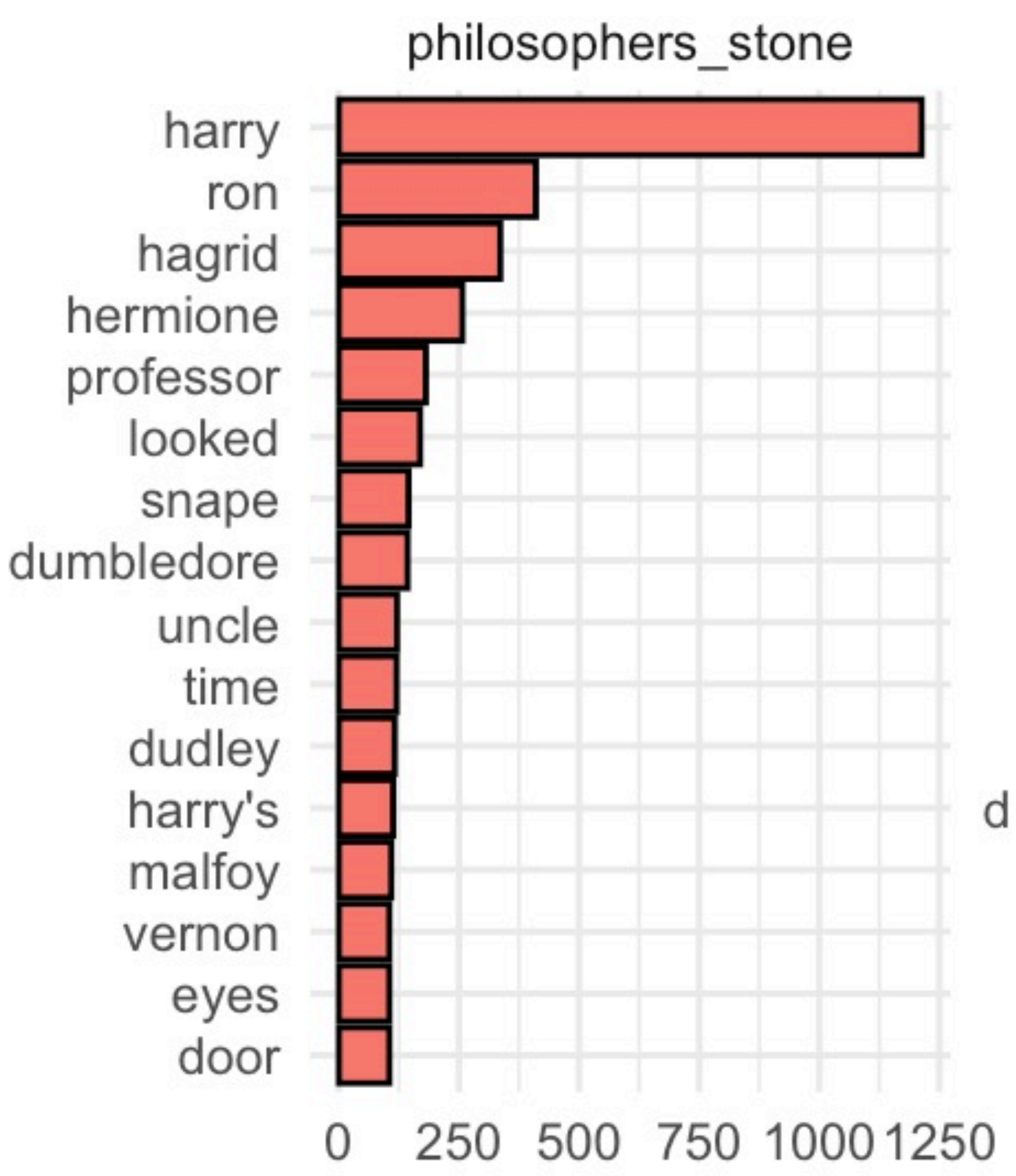
Example 2. Text as data

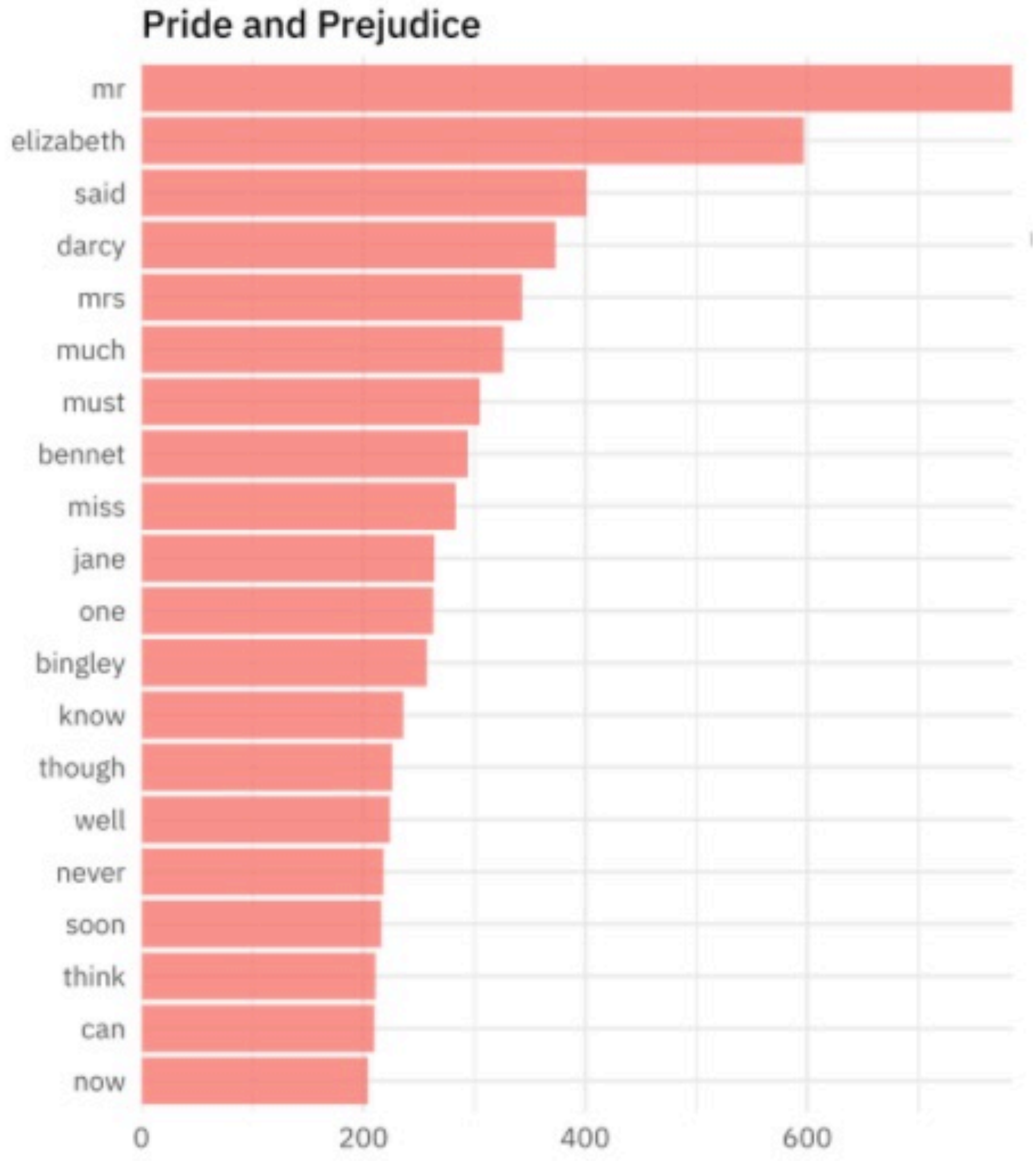
Can you tell who wrote the following sentences? Jane Austen from *Pride and Prejudice* or J. K. Rowling from *Harry Potter*?

"The thought of the confined creature was so dreadful to him..."

"...was as if something turned over, and the point of view altered..."

Examples: Text as data (predictors)





You will work on these text data in your labs with your GSI, and make predictions on your own

Example: Disease risk prediction

Example 3. Disease risk prediction

We are given massive biobank data for patients with stroke history that contain information:

1. $Y \in \{0,1\}$: if the patient has developed Alzheimer's disease at the time of recruit

2. $X \in \mathbb{R}^p$: covariates information including individual lifestyles (insomnia, current smoking status, beef lover, etc.), baseline biomarker information (gender, age, education, and family AD history), and genetic information (SNPs)

Goal:

1. Predict which patients that are at high risk of developing AD (why important?)

2. Find any lifestyle factors can reduce the risk of AD

How about the other patients that are not diagnosed? Can we use their information to improve our prediction?

Semi-supervised Learning

Semi-supervised Learning

In medical record database, we often encounter the following scenario:

1. Labeled data $\{Y_i, X_i\}_{i=1}^{n}$ — $Y_i \in \{0,1\}$ represents if the patient is diagnosed with certain disease at the time of recruit, and $X_i \in \mathbb{R}^p$ represents patient all available information

2. Unlabeled data $\{X_j\}_{j=n+1}^{n+N}$ — $X_j \in \mathbb{R}^p$ represents unlabeled patient information

Question: Can we improve the prediction results given massive amount of unlabeled data? If so, when?

Learning objectives in the first half of the semester

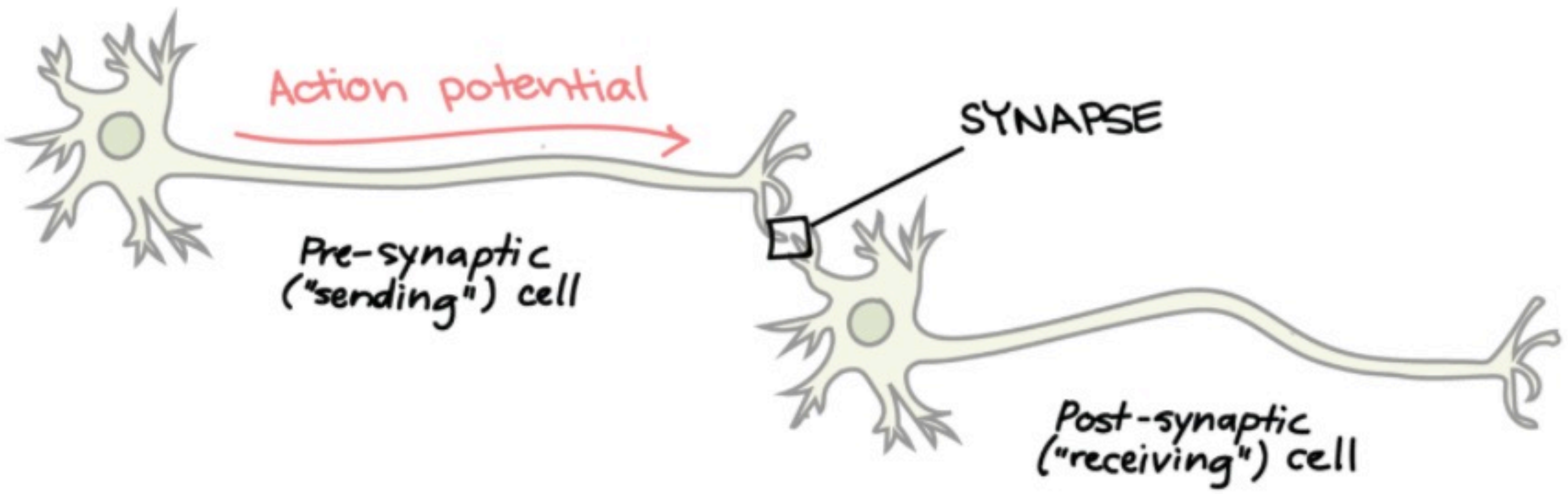
Supervised Learning (classical approaches)

1. GLM/SVM (09/08)
2. Kernel-based Methods (09/15)
3. Metric Learning (09/22)
4. Tree-based Methods (09/29, 10/06)

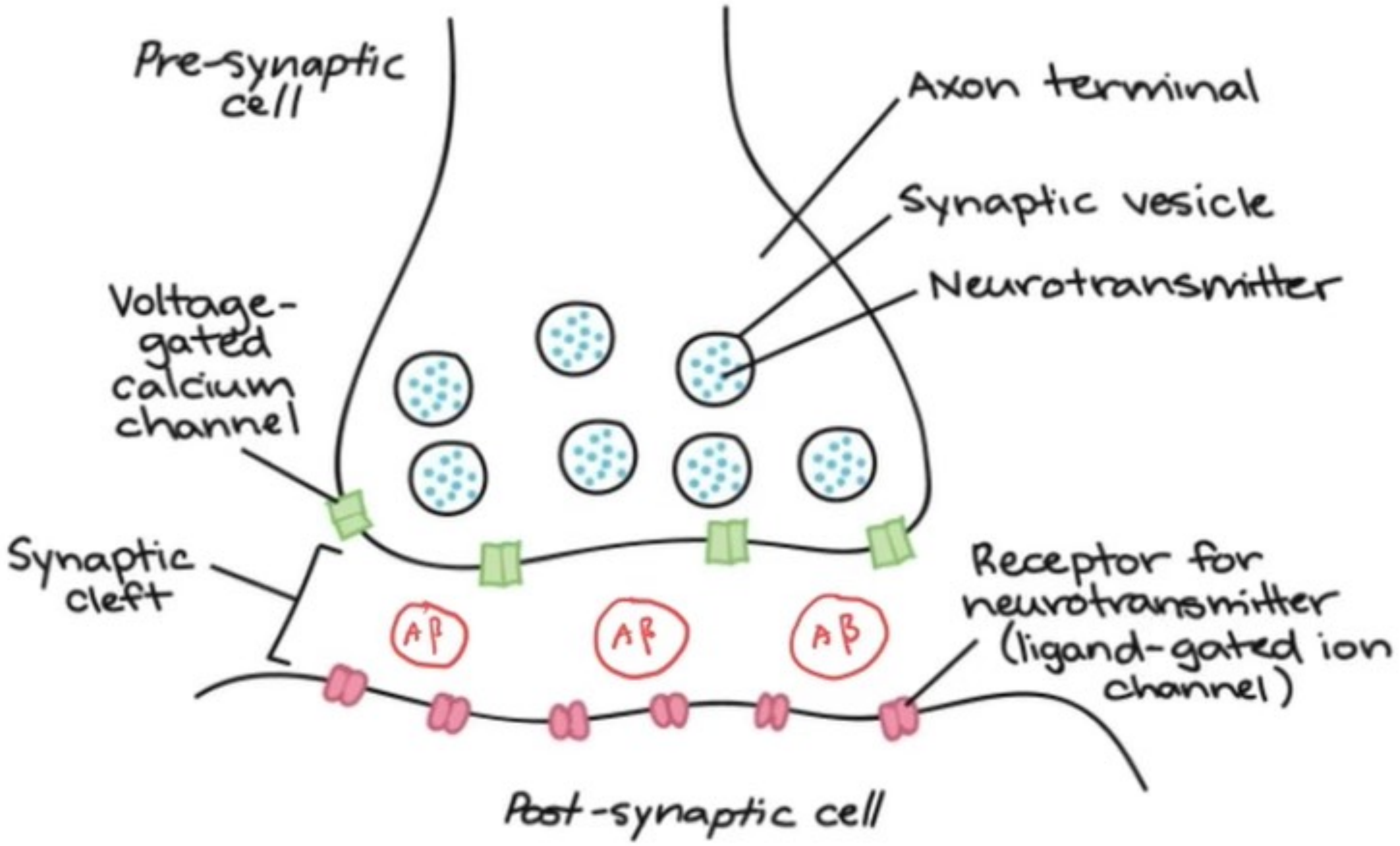
Semi-supervised Learning (10/13)

And then?

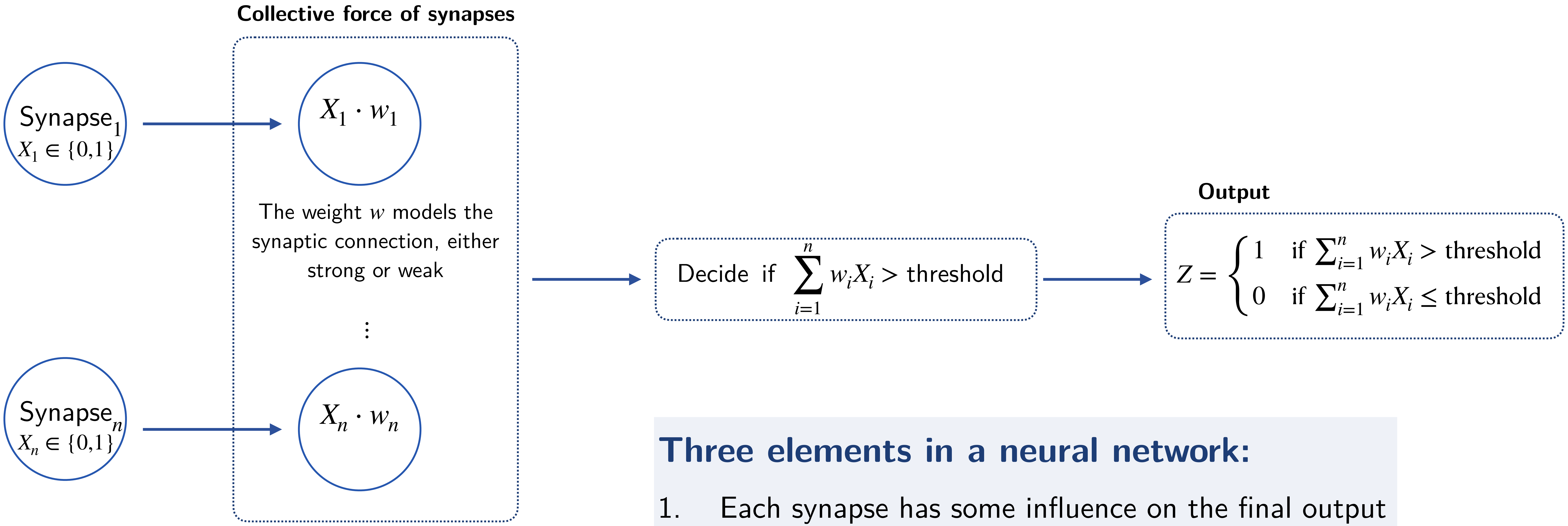
Alzheimer's disease: a little more detail..



Alzheimer's disease is a neurodegenerative disease often characterized by dementia, accumulation of beta-amyloid $(A\beta)$ plaques and tau proteins on neurons, and brain inflammation and atrophy.



Neural Networks (1)

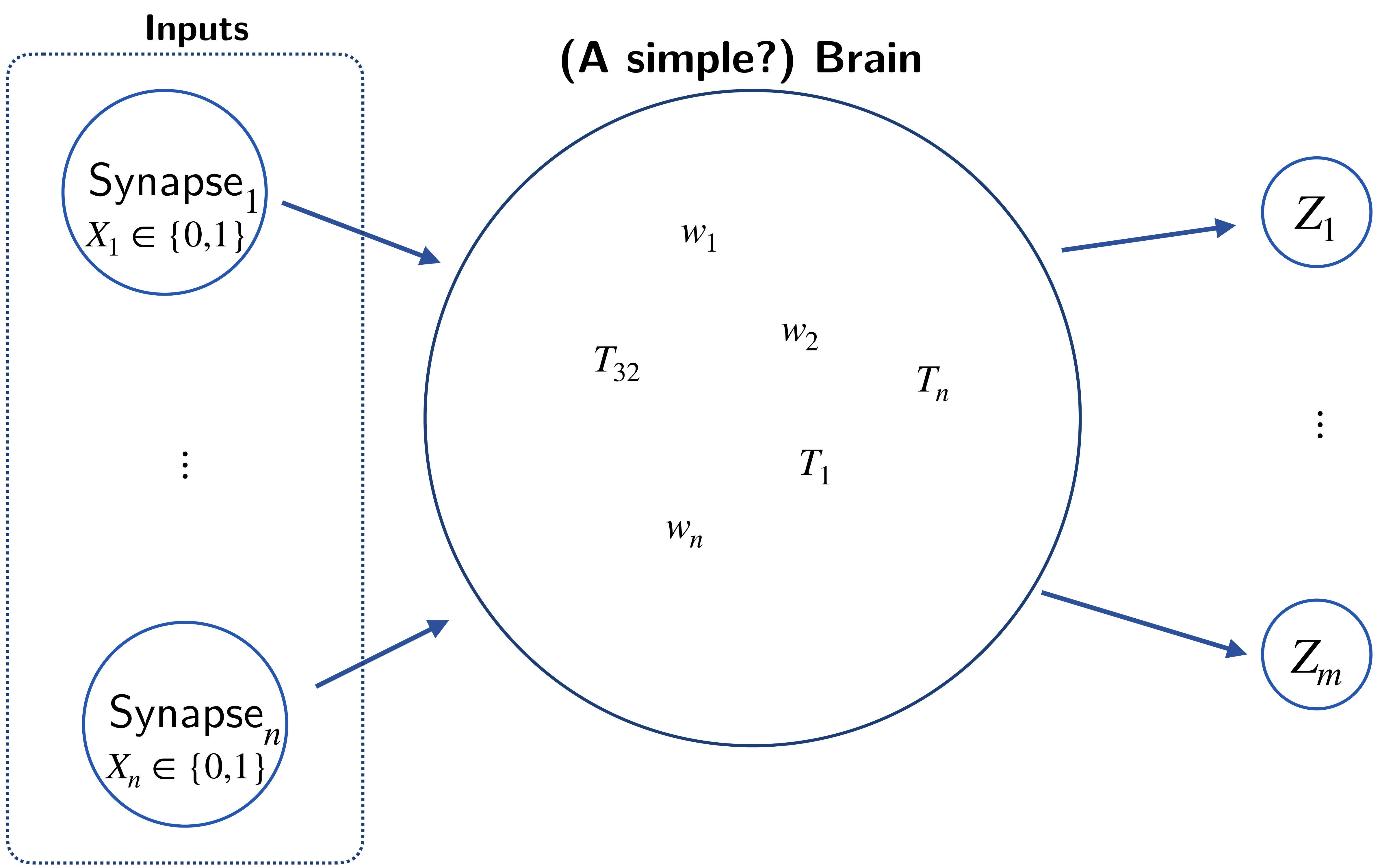


Measure the influence of the synapses—decide whether or not the whole axon is stimulated

Three elements in a neural network:

1. Each synapse has some influence on the final output
2. Cumulative influence simulate the axon
3. Synaptic weight

Neural Networks (2)



Neural Network (NN)

A NN learns the functional form of

$$Z_j = f_j(X_1, \ldots, X_n, w_1, \ldots, w_n, T_1, \ldots, T_n), \quad j = 1, \ldots, m,$$

where we need to adjust the weights w_i and the thresholds T_i so that what we get out is the outcome Z_j.

Example: Disease risk prediction with NN

Example 3. Disease risk prediction (revisit)

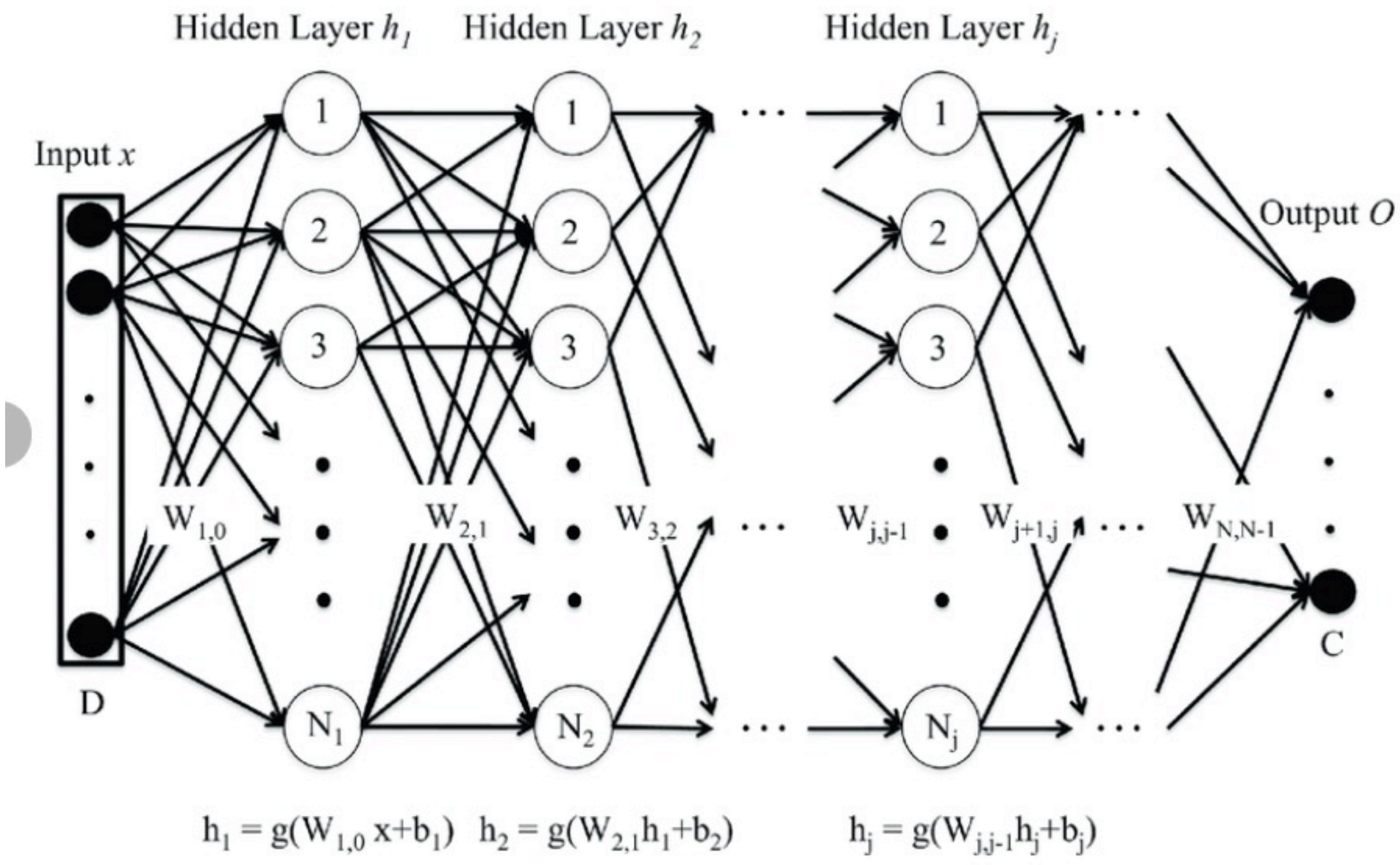
We are given massive biobank data for patients with stroke history that contain information:

1. $Y \in \{0,1\}$: if the patient has developed Alzheimer's disease at the time of recruit

2. $X \in \mathbb{R}^p$: covariates information including individual lifestyles (insomnia, current smoking status, beef lover, etc.), baseline biomarker information (gender, age, education, and family AD history), and genetic information (SNPs)

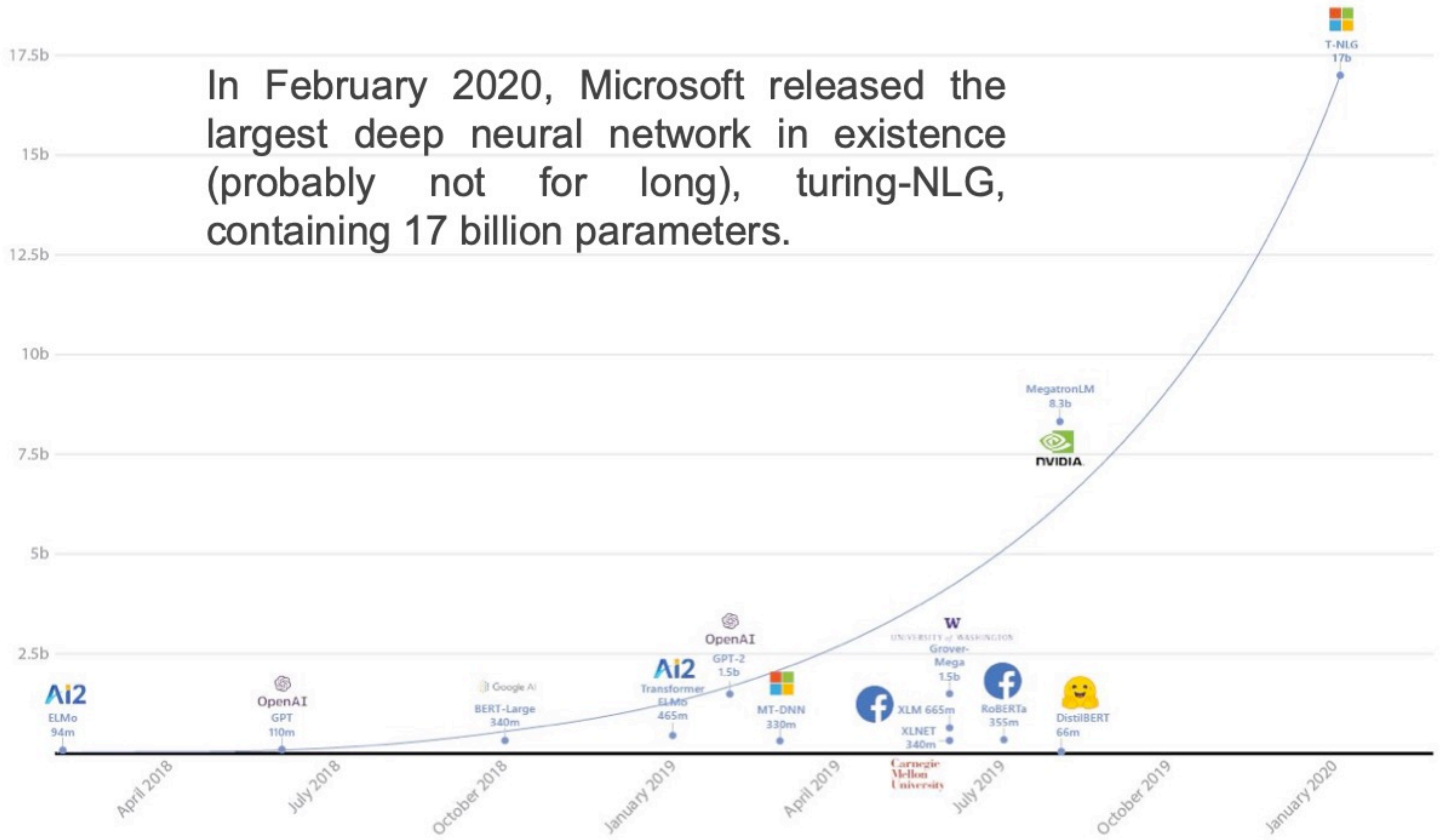
To have binary input, we can transform the covariates into dummy variables. Can we still obtain the effect of certain lifestyle on lowering the disease risk?

Deep Neural Networks



Deep Neural Networks

In February 2020, Microsoft released the largest deep neural network in existence (probably not for long), turing-NLG, containing 17 billion parameters.



Challenges in Research

IBM Watson for oncology

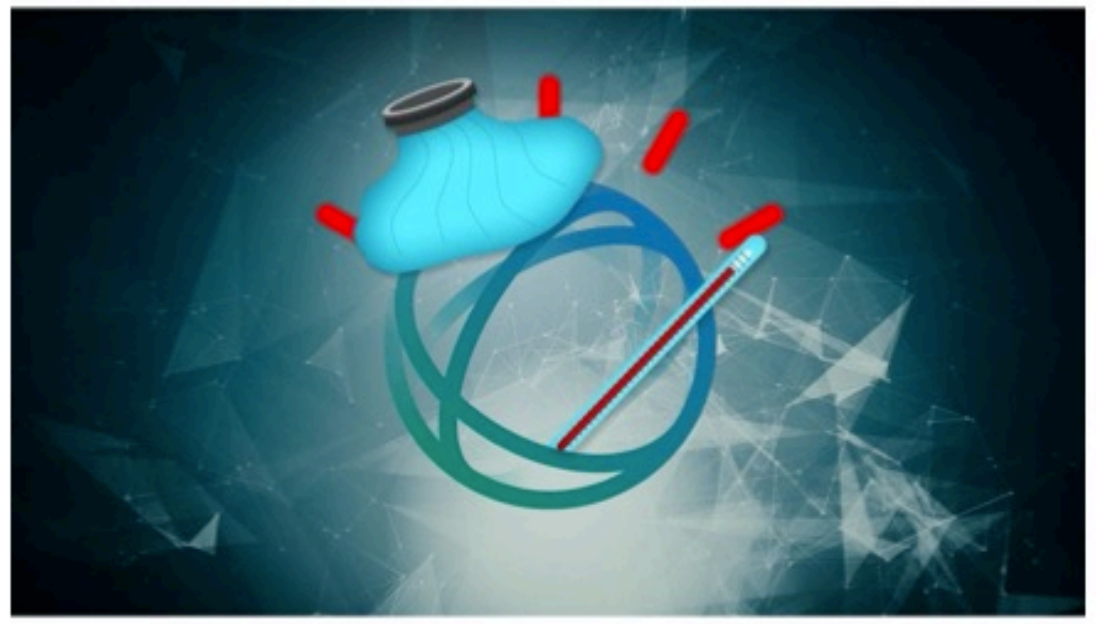
EDITORS' PICK | Feb 19, 2017,03:48pm EST

MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine

Matthew Herper Former Staff
Healthcare
I cover science and medicine, and believe this is biology's century.

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By Casey Ross @caseymross and Ike Swetlitz • July 25, 2018



Alex Hogan/STAT

Internal IBM documents show that its Watson supercomputer often spit out erroneous cancer treatment advice and that company medical specialists and customers identified "multiple examples of unsafe and incorrect treatment recommendations" as IBM was promoting the product to hospitals and physicians around the world.

1. MD Anderson taps IBM Watson to power "Moon Shots" mission aimed at ending cancer, starting with Leukemia
2. Big data insights to help accelerate translation of cancer-fighting knowledge to cutting edge medical practices
3. Link to the news: https://www.ibm.com/products/clinical-decision-support-oncology
4. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

How to make machine learning methods more trustworthy?

Learning objectives

1. Neural Networks (10/20)

   Convolutional Networks, Recurrent Networks, and their applications in medical research

2. Deep Neural Networks (10/27)

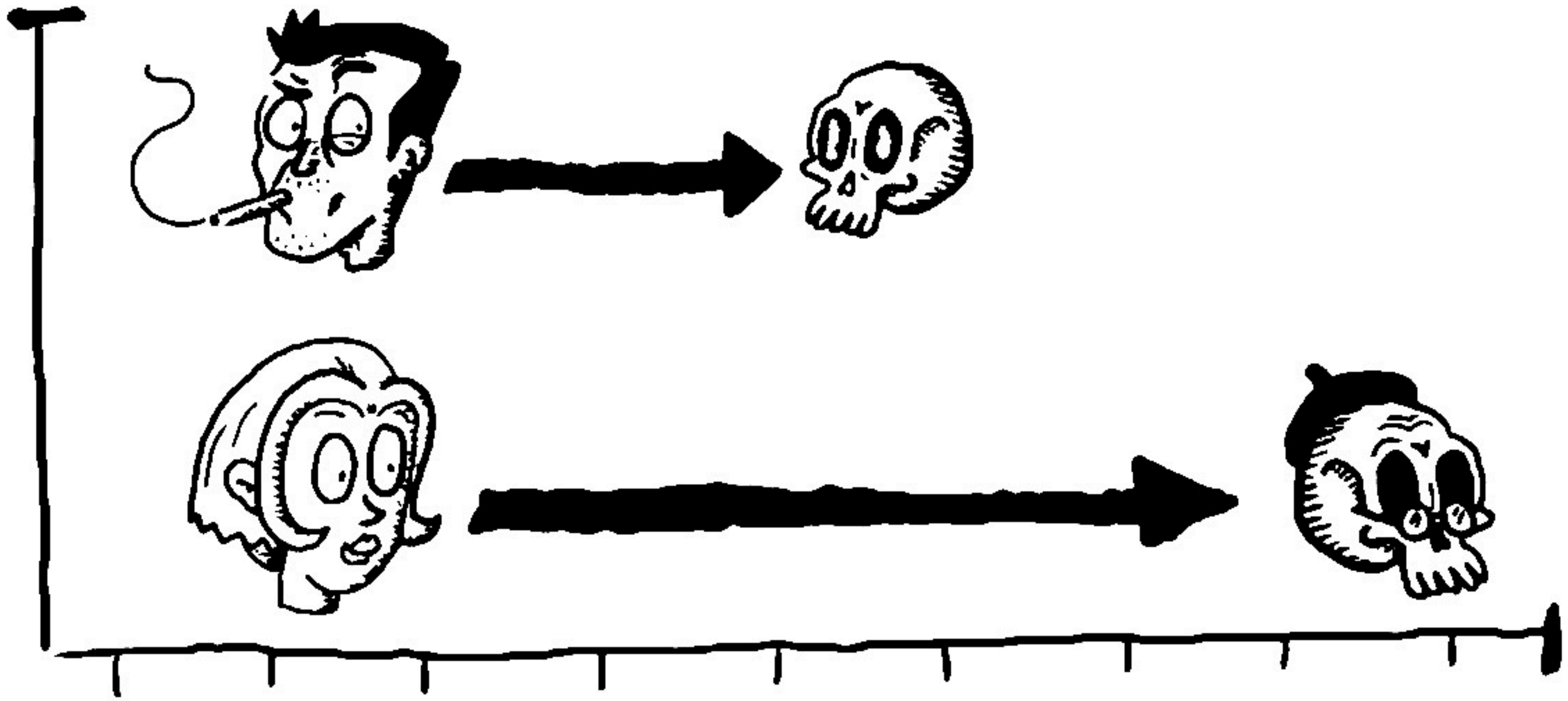
Observational data

1. Disease risk prediction — supervised learning (including NN and DNN), semi-supervised learning
2. Learning effective treatments for different patients

?

1. Will you trust the findings from observational data?
2. If not, what can we do?
3. How can we guide future clinical decisions?

Effect of Smoking on Life Expectancy?



What hinders our understanding?

Smoking

Confounding factors

Smokers tend to drink more heavily and have a less healthy diet

Randomized experiment is not feasible

Smoking cannot be randomly assigned to individual — neither practical nor ethical

Reverse Causation

When people become ill, they tend to give up smoking. This wrongly suggests smoking worsen health.

Shortened Life Expectancy

Last learning objectives

Causal Inference and Clinical Trials

1. Nature's Experiments: Mendelian Randomization (11/10)
2. Bayesian Inference and Design of Experiments (11/17)
3. Adaptive Clinical Trial and Reinforcement Learning (12/01)

Observational data

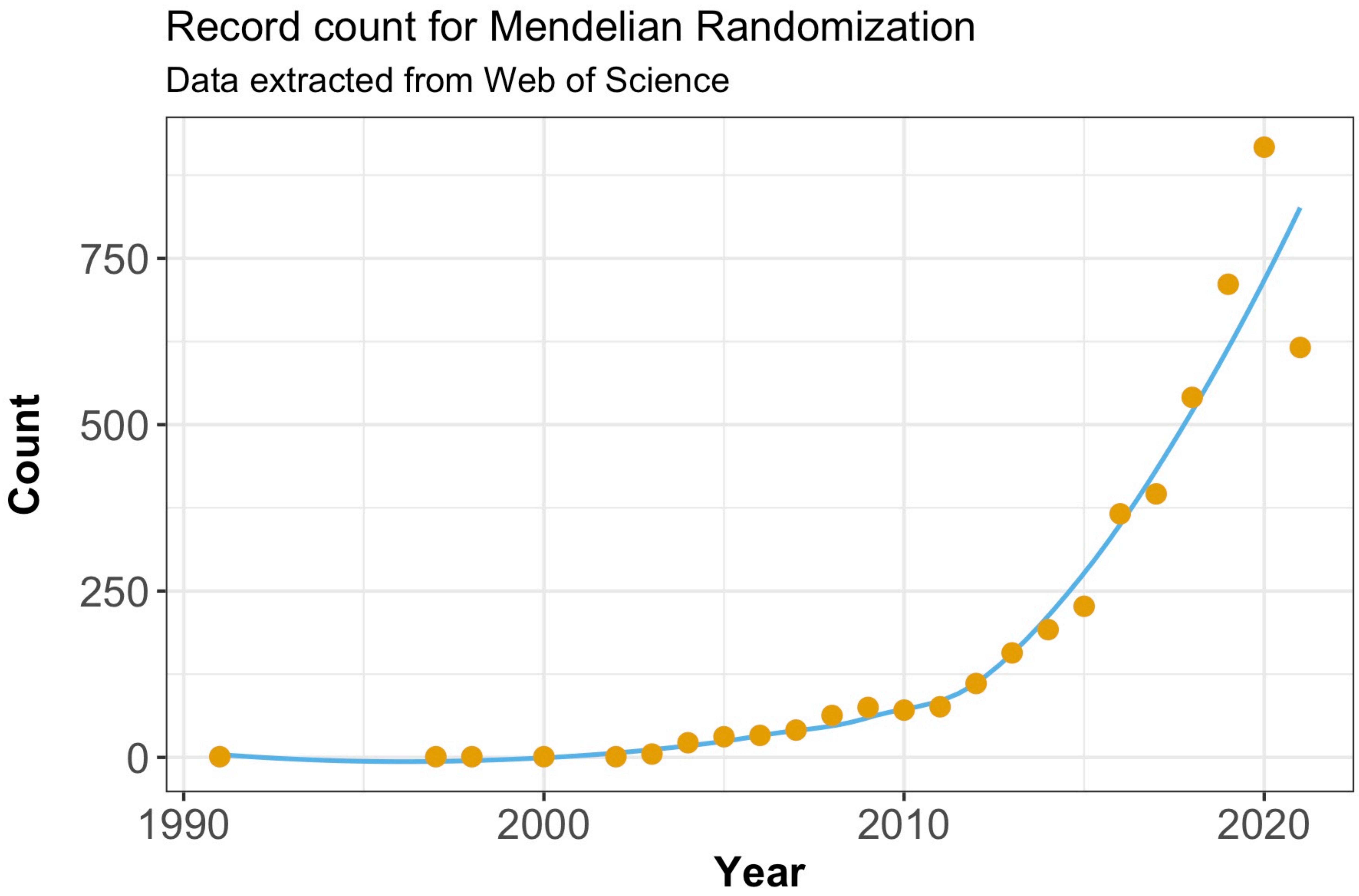
1. Disease risk prediction — supervised learning (including NN and DNN), semi-supervised learning
2. Learning effective treatments for different patients

Causal Inference

1. Randomized control trials (RCT) can help us to verify our finds from observation study
2. When RCT is not available, we may rely on Nature's experiment (Mendelian Randomization)

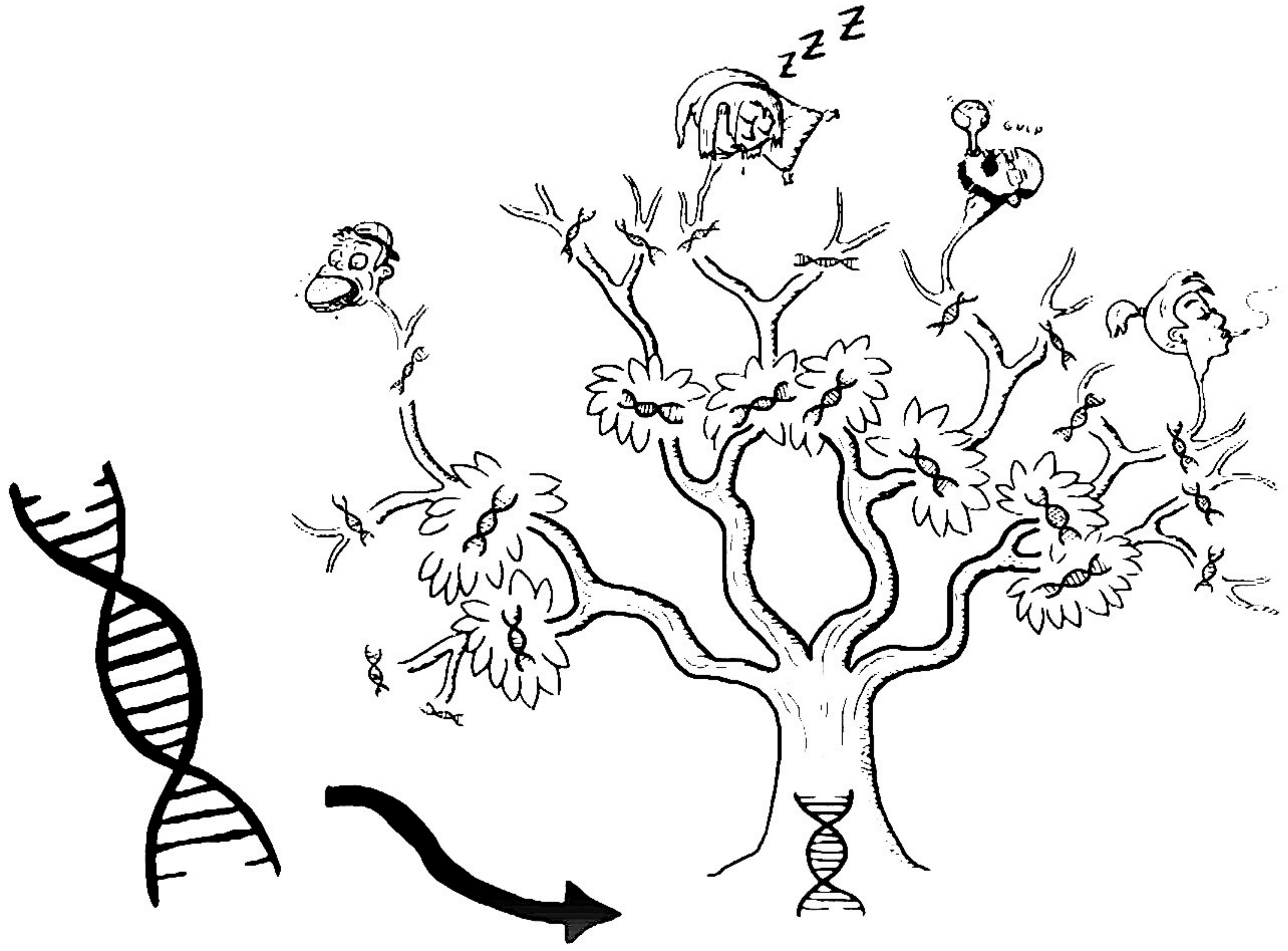
Reinforcement

Mendelian Randomization: Popularity



Mendelian Randomization

Randomly inherited genes are **not** associated with any confounding factors



CHRNA5

Smokers who carry one version of CHRNA5 tend to smoke less heavily than those who carry a different version

Smoking

Shortened Life Expectancy

Different variant of CHRNA5 in non-smokers has no effect on life expectancy

Labs

Supervised Learning (classical approaches)

1. GLM/SVM (09/08)
2. Kernel-based Methods (09/15)
3. Metric Learning (09/22)
4. Tree-based Methods (09/29, 10/06)

Semi-supervised Learning (10/13)

3 Labs

1. Neural Networks (10/20)
2. Deep Neural Networks (10/27)

1 Labs

Causal Inference and Clinical Trials

1. Nature's Experiments: Mendelian Randomization (11/10)
2. Bayesian Inference and Design of Experiments (11/17)
3. Adaptive Clinical Trial and Reinforcement Learning (12/01)

2 Labs

Please fill in the lab time change pool!