BMI 702: Biomedical Artificial Intelligence

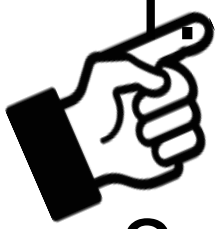
Foundations of Biomedical Informatics II, Spring 2023

Lecture 1: Course overview and introduction to biomedical AI

HARVARD
MEDICAL SCHOOL

Marinka Zitnik
marinka@hms.harvard.edu

Outline for today's class



1. Overview of syllabus
2. What makes biomedical data unique
3. Motivation for machine learning
4. Roadmap for responsible biomedical AI

What will you learn in this course?

- Survey of biomedical AI, covering key data modalities
  - Clinical data
  - Networks, graphs, and multimodal datasets
  - Language and text
  - Images

- Machine learning problems from a practical perspective
  - Problems that drive the adoption of AI in biology and medicine
  - Foundational algorithms and emphasis on the subtleties of working with biomedical data
  - Evaluating and transitioning machine learning systems into biomedical and clinical implementation

- Broader impacts:
  - Bias and fairness
  - Interpretability
  - Ethical and legal considerations

Images generated by DALL-E 2 using text prompts shown on this slide

Course staff

- **Marinka Zitnik** (Instructor)
  - Assistant Professor of Biomedical Informatics
  - Associate Member at the Broad Institute
  - Faculty at Harvard Data Science
  - https://zitniklab.hms.harvard.edu

Course staff: Three TAs

- Yasha Ektefaie
  - 3rd year PhD student in BIG program
  - yasha_ektefaie@g.harvard.edu

- Richard Chen
  - 4th year PhD student in BIG program
  - richardchen@g.harvard.edu

- Yepeng Huang
  - 2nd year MS student in CBQG
  - yepenghuang@hsph.harvard.edu

Dates, times and format

- Website: https://zitniklab.hms.harvard.edu/BMI702
- **Monday, 1:00 PM – 3:00 PM ET**
  - First class: 01/23/23
  - No class: Monday, February 20, President's Day
  - No class or assignments due: Week of March 13
- **Location:** Armenise Modell 100A, 200 Longwood Ave
- **Office hours:**
  - Mon, 3-4 pm, Countway 309
  - Tue, 5-6pm, Countway 423/424 open space
  - Thu, 3-4pm, Countway 423/424 open space
  - Fri, 10-11am, Countway 423/424 open space

Course syllabus

- This course has 14 lectures:
  - Course overview and introduction to biomedical AI
  - Other 12 lectures are divided into six modules
    - The first lecture in each module introduces ML concepts in the area
    - The following lecture introduces advanced topics in the same area
  - Final lecture on ethical & legal considerations of biomedical AI

- Modules:
  - Module 1: Clinical AI
  - Module 2: Trustworthy AI
  - Module 3: Graph Learning
  - Module 4: Language Modeling
  - Module 5: Biomedical imaging
  - Module 6: Therapeutic Science

Assignments

- Problem sets:
  - Pset 1: Bias, explainability, and fairness
  - Pset 2: Graph learning and network embeddings
  - Pset 3: Biomedical imaging
  - Primary form of support are office hours we will host
  - Problems sets must be completed individually
- Pre-class quizzes:
  - Open at 9:00am on Tuesday, due at 11:59pm on Sunday
  - Based on the **Required Reading** section of each lecture
- Class participation and quick checks:
  - Short conceptual questions embedded into each lecture
  - Meant for you to check your understanding of the concepts that were just introduced
  - Your score on them does not matter, you just need to do them

Grading

Grade Components

Component	Percent of grade (%)
Problem Set 1	20
Problem Set 2	20
Problem Set 3	20
Class Participation	14 (1 point for Lecture 1-14)
Pre-Class Quizzes	26 (2 points per quiz; there is no quiz for Lecture 1)

DALL-E 2. Prompt: “Graduate students getting bonus points, van Gogh style”

We want you to succeed!
If you feel overwhelmed, visit our office hours and talk with us. We know graduate school can be stressful and we want to help you succeed

Bonus points for outstanding work on assignments, active participation in discussions on Canvas, …

Course culture and attendance

- Course culture:
  - Students taking this course come from a wide range of backgrounds
  - We hope to foster an inclusive and safe learning environment based on curiosity and research inquiry
  - All members of the course community—the instructor, TAs and students—are expected to treat each other with courtesy and respect
- Attendance:
  - The course will be run in a in-person format
  - Students must attend all classes unless they have explicit permission from the course instructor

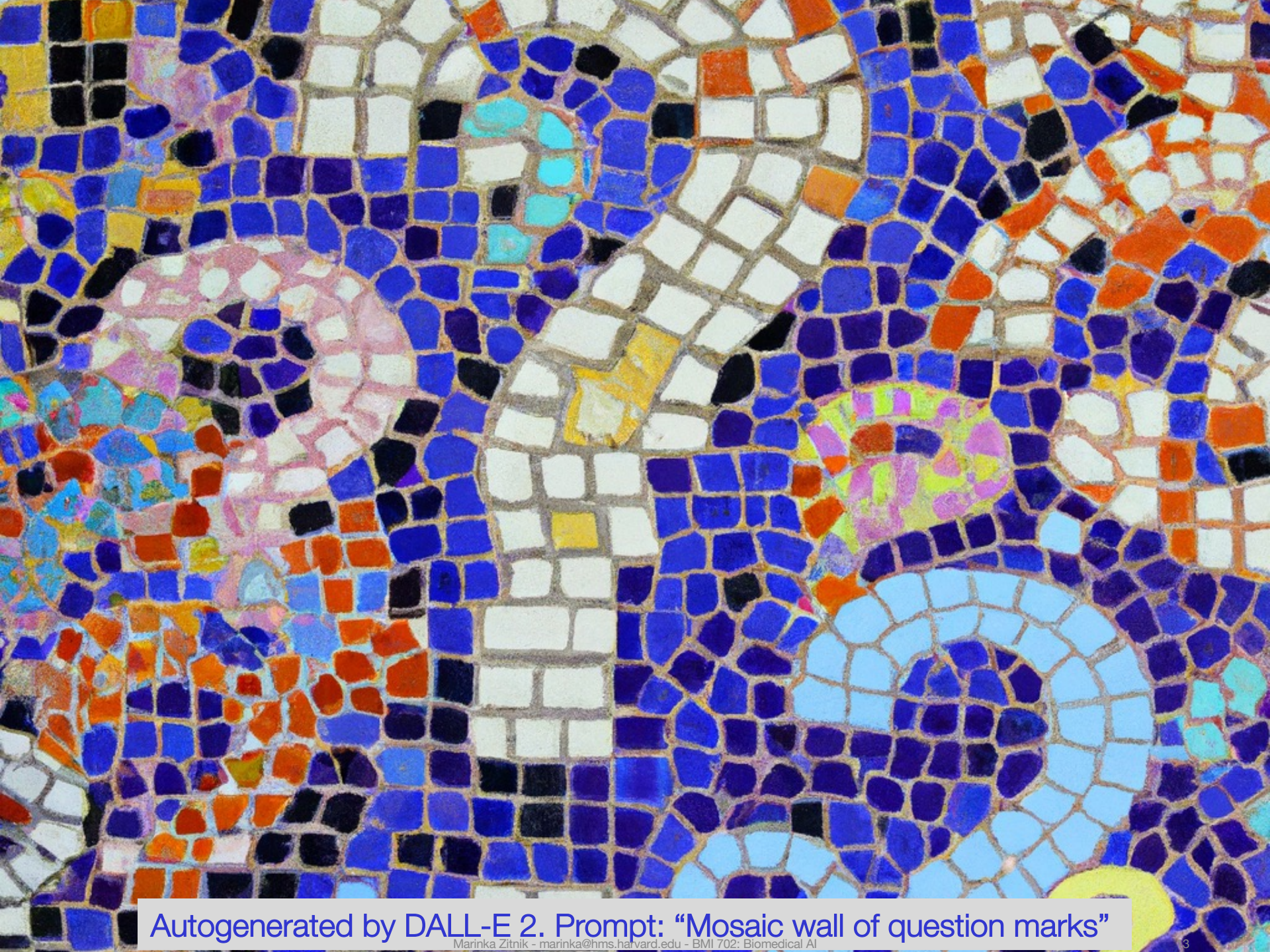
Policies

- Collaboration policy
  - Unless otherwise specified, all work submitted must reflect student's own effort and understanding
  - Clearly distinguish your own ideas and knowledge from information derived from other sources:
    - Students must properly cite all submitted work
    - Unless noted otherwise, students are expected to complete assignments, quizzes, and projects individually, not as teams
    - Discussion about course content and materials is acceptable, but sharing solutions is not acceptable
- Late policy
  - All assignments are due at 11:59 pm on the due date
  - Extensions provided in the case of exceptional circumstances

By the end of this course, …

Goals

- Prepare students for advanced courses in data science, machine learning, and statistics, by providing the necessary foundation and context
- Empower students to apply computational and inferential thinking to address real-world problems
- Understand artificial intelligence methods from a practical perspective
- Understand best practices in implementing, evaluating, and validating ML methods on biomedical data
- Apply ML methods to key data modalities: clinical data, biomedical networks, text, and images
- Understand the pros and cons of different ML methods to select the right method for a given scenario
- Recognize the problem of bias in biomedical data and ML methods in healthcare
- Understand the concept of fairness in biomedical ML
- Become familiar with ethical considerations for biomedical data and algorithms



Autogenerated by DALL-E 2. Prompt: “Mosaic wall of question marks”

Outline for today's class

1. Overview of syllabus

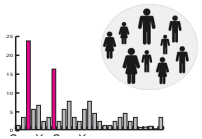
2. What makes biomedical data unique

3. Motivation for machine learning

4. Roadmap for responsible biomedical AI

What makes biomedical data so different?

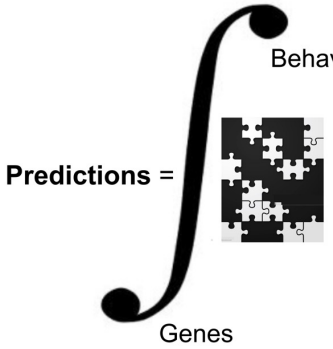
Need to integrate heterogeneous, confounded data that **span from molecules to society**



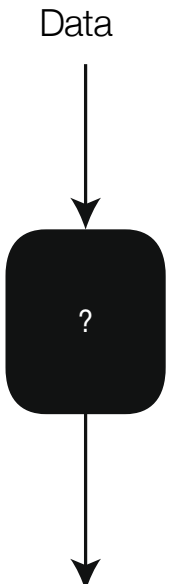
Genome

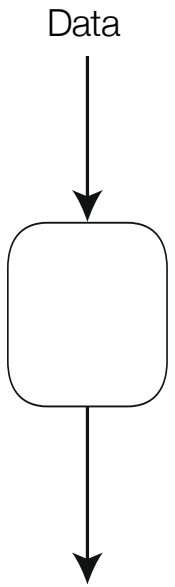
Health records

Behaviors, lifestyle, vital signs



Need to translate predictions into **actionable** hypotheses





Multi-scale: molecules, individuals, populations

Heterogeneous: experimental readouts, curated, self-reported

Confounded: data from different technologies, and measurement platforms

What makes biomedical data so different?

- Life or death decisions
  - 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 biomedical data so different?

- ML models are increasingly deployed in real-world applications and implemented in clinical settings:
  - It is critical to ensure that these models are behaving responsibly and are trustworthy
- Accuracy alone is no longer enough
- Auxiliary criteria are important:
  - Explainable predictions and interpretable models
  - Fair and non-discriminatory predictions
  - Privacy-preserving, causal, and robust predictions
- This broad area is known as trustworthy ML

High-stakes decisions

What makes biomedical data so different?

- Very little labeled data
- Recent breakthroughs in AI depended on lots of labeled data!

Large, diverse data (+ large models) ⟶ Broad generalization

Russakovsky et al. ‘14

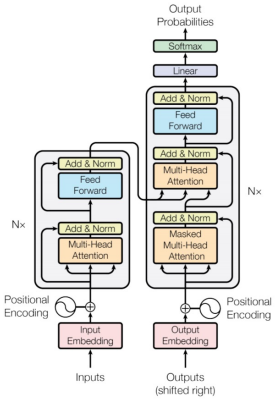


Figure 1: The Transformer - model architecture.

Vaswani et al. ‘18

GPT-2

Radford et al. ‘19

What if you don't have a large dataset?

medical imaging

robotics

personalized education, recommendations

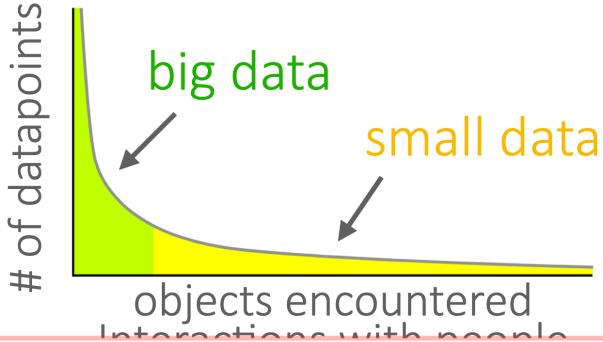
translation for rare languages

What if you want a general-purpose AI system in the real world?

Need to continuously adapt and learn on the job.

Learning each thing from scratch won't cut it.

What if your data has a long tail?



These settings break the supervised learning paradigm.

What makes biomedical data so different?

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

What makes biomedical data so 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 are often in siloes; everyone recognizes need for interoperability, but slow progress
  - Rigorous testing and iteration are needed
- Difficulty of correcting for biases and inequities:
  - Consideration of ethical and legal issues
  - Health data on which algorithms are trained are likely to be influenced by many facets of social inequality

Outline for today's class

1. Overview of syllabus

2. What makes biomedical data unique

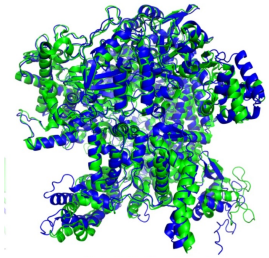
3. Motivation for machine learning

4. Roadmap for responsible biomedical AI

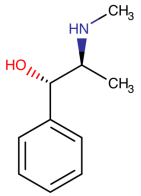
AI opportunities in medicine

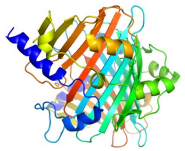
Input AA sequence

Geometric learning

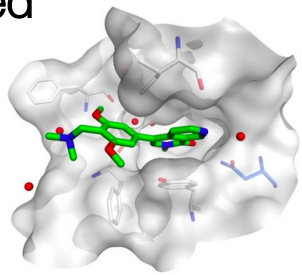


3D coordinates of amino acids in a protein





Self-supervised learning

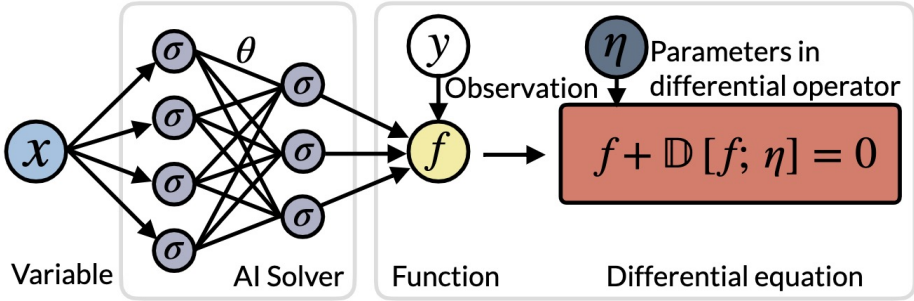


Ligand-target binding affinity

10^8 molecular compounds

Graph learning

Search of new antibiotics



New fundamental results in molecular dynamics

AI opportunities in health

Preliminary diagnosis, early disease detection, self-care

Automated image diagnosis, language modeling

Clinical trial participation, drug discovery, AI-driven medical devices

Comorbidities, chronic disease treatments

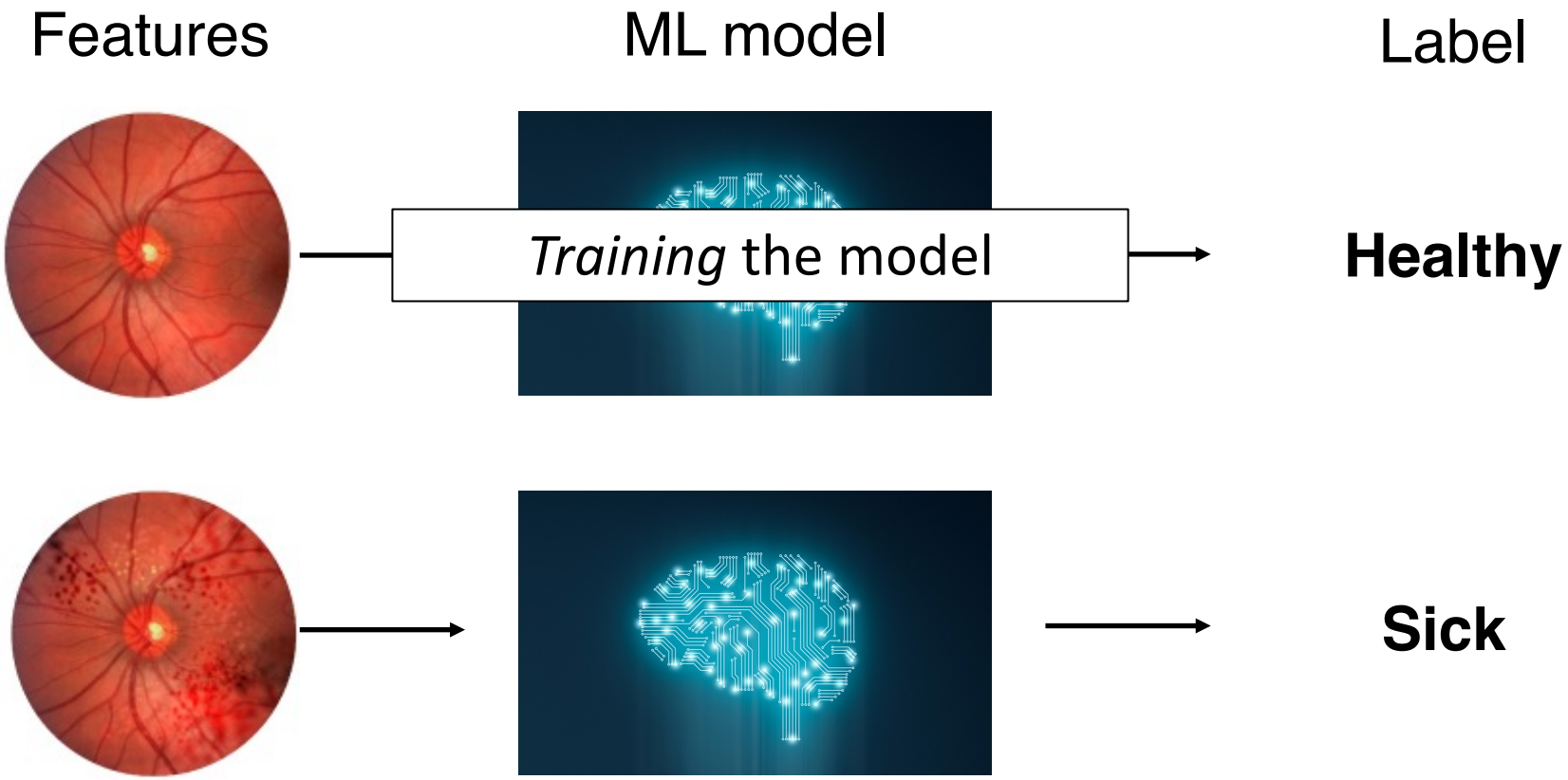
Administrative workflows, costly back-office problems

Inpatient & outpatient policies of care

Real-time patient interventions

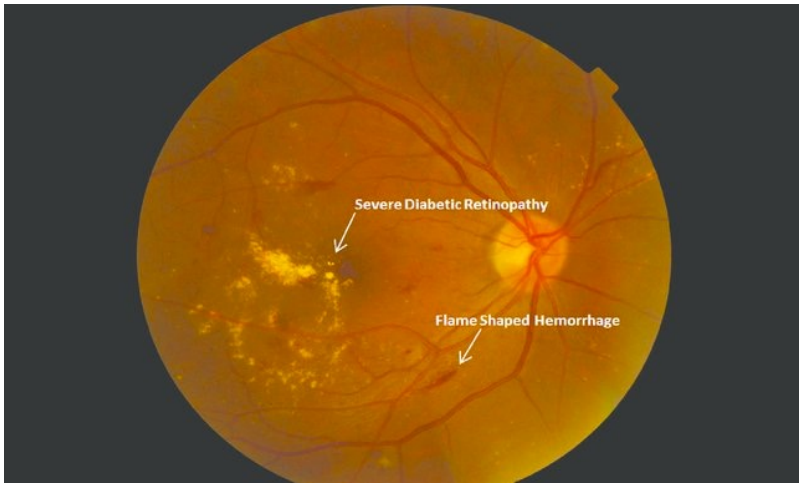
Protection of health data, avoiding medical errors

Machine learning



Diagnosing diabetic retinopathy

- **Diabetic retinopathy** affects blood vessels in the retina that lines the back of the eye
- The most common cause of vision loss among people with diabetes
- Leading cause of vision impairment and blindness among adults



Normal Retina

Diabetic Retina

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *JAMA*, 2016

We will learn about this topic in **Module 1: Clinical AI** and **Module 1: Biomedical Imaging**

Diagnosing diabetic retinopathy

- 128,000 retinal fundus photographs
- Each image was rated by 3-7 ophthalmologists
- "Off the shelf" deep neural network

This Issue | Views 60,378 | Citations 2 | Altmetric 633

Original Investigation | Innovations in Health Care Delivery

December 13, 2016

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram, PhD[1]; et al

» Author Affiliations

JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

Normal
Retina

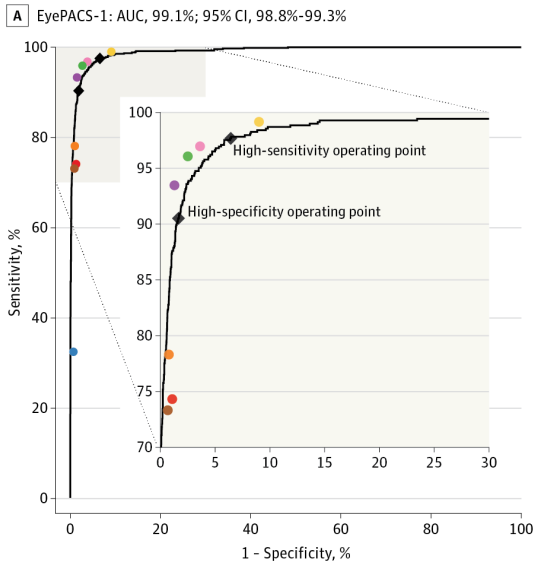
Diabetic
Retina

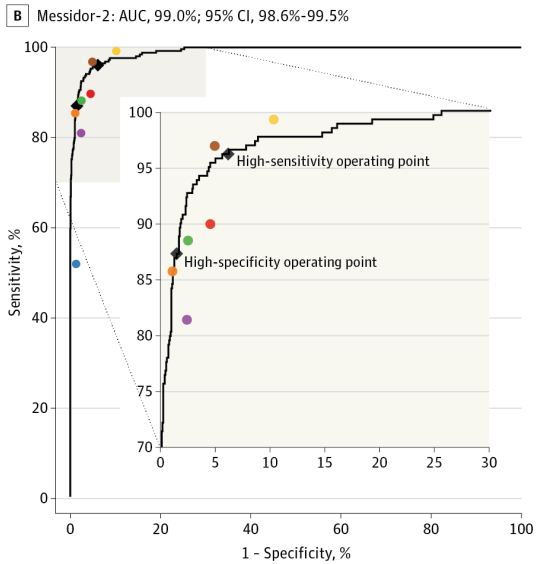
Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *JAMA*, 2016

Diagnosing diabetic retinopathy

Algorithm did better than most individual ophthalmologists in the study

Large data + machine learning ~ human-level performance in diagnostic medical imaging





$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Sensitivity (TPR): probability of positive test result, conditioned on individual truly being positive
Specificity (TNR): probability of a negative test result, conditioned on individual truly being negative

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *JAMA*, 2016

Pivotal trial of an autonomous AI-based diagnostic system

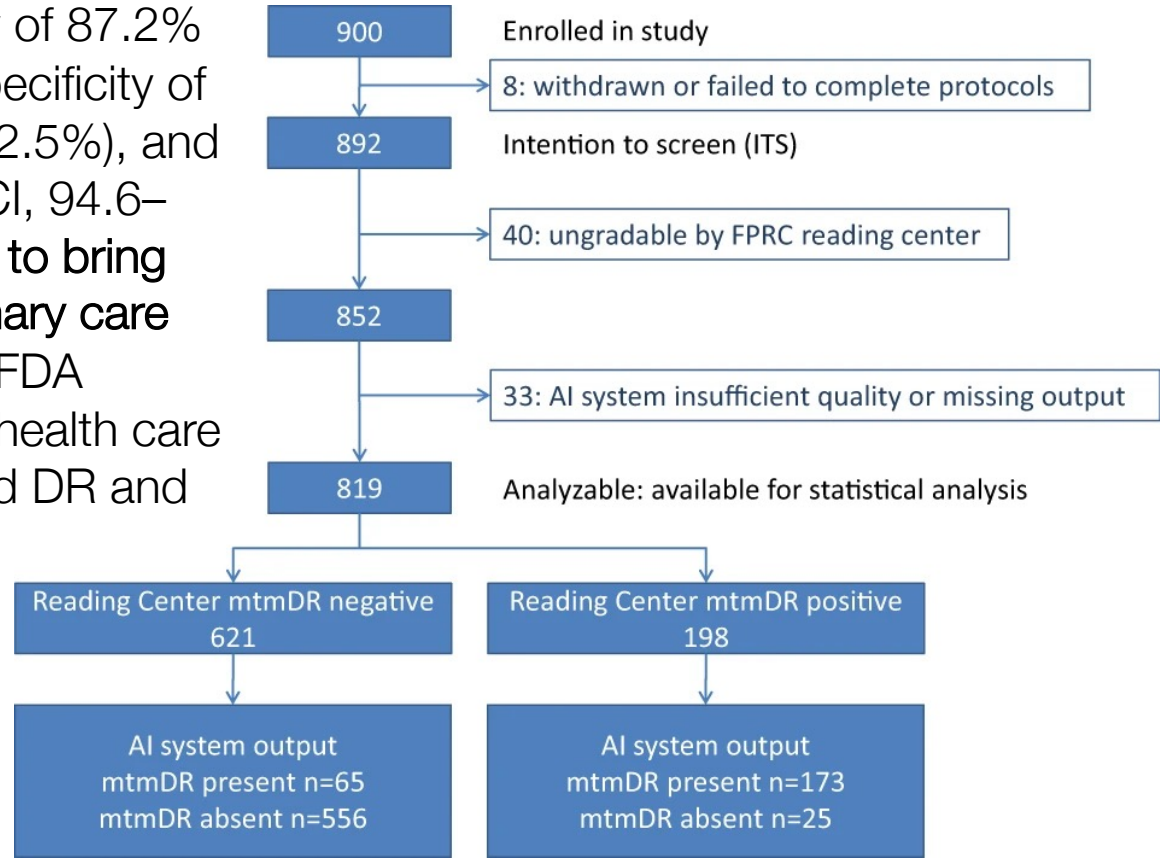
Abstract

Artificial Intelligence (AI) has long promised to increase healthcare affordability, quality and accessibility but FDA, until recently, had never authorized an autonomous AI diagnostic system. This pivotal trial of an AI system to detect diabetic retinopathy (DR) in people with diabetes enrolled 900 subjects, with no history of DR at primary care clinics, by comparing to Wisconsin Fundus Photograph Reading Center (FPRC) widefield stereoscopic photography and macular Optical Coherence Tomography (OCT), by FPRC certified photographers, and FPRC grading of Early Treatment Diabetic Retinopathy Study Severity Scale (ETDRS) and Diabetic Macular Edema (DME). More than mild DR (mtmDR) was defined as ETDRS level 35 or higher, and/or DME, in at least one eye. AI system operators underwent a standardized training protocol before study start. Median age was 59 years (range, 22–84 years); among participants, 47.5% of participants were male; 16.1% were Hispanic, 83.3% not Hispanic; 28.6% African American and 63.4% were not; 198 (23.8%) had mtmDR.

Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 2018

Pivotal trial of an autonomous AI-based diagnostic system

The AI system exceeded all pre-specified superiority endpoints at sensitivity of 87.2% (95% CI, 81.8–91.2%) (>85%), specificity of 90.7% (95% CI, 88.3–92.7%) (>82.5%), and imageability rate of 96.1% (95% CI, 94.6–97.3%), **demonstrating AI's ability to bring specialty-level diagnostics to primary care settings**. Based on these results, FDA authorized the system for use by health care providers to detect more than mild DR and diabetic macular edema.

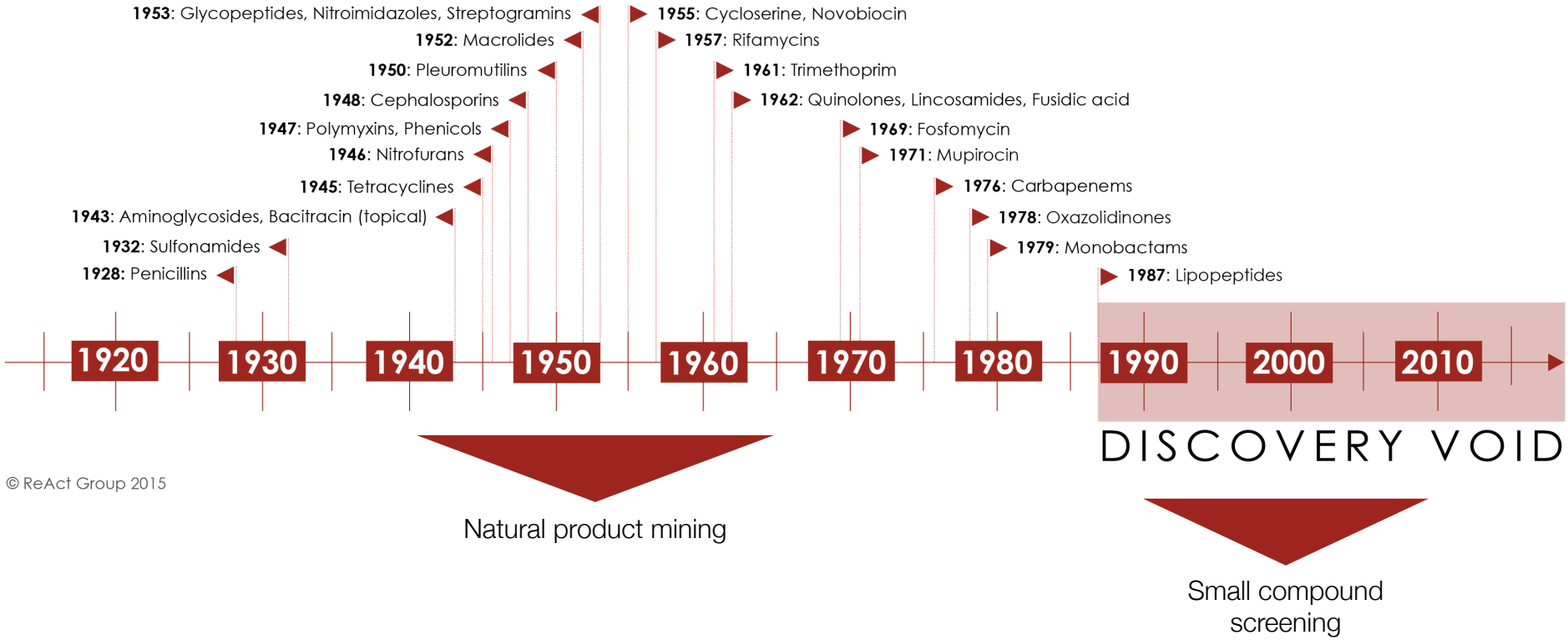


First FDA authorized autonomous AI diagnostic system in any field of medicine

Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 2018

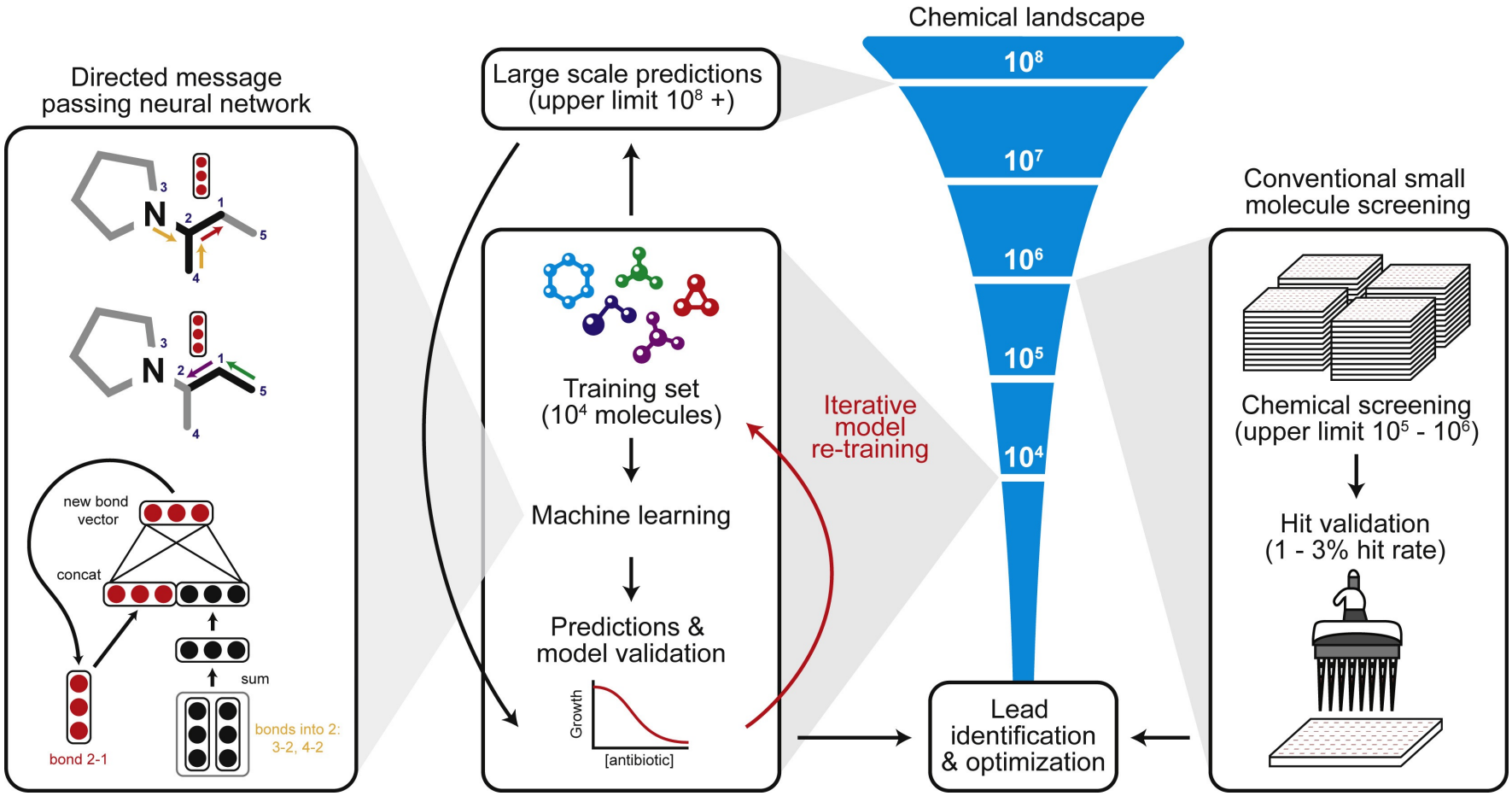
We will learn about this topic in **Module 3: Graph Learning** and **Module 6: Therapeutic Science**

Antibiotic discovery timeline



A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

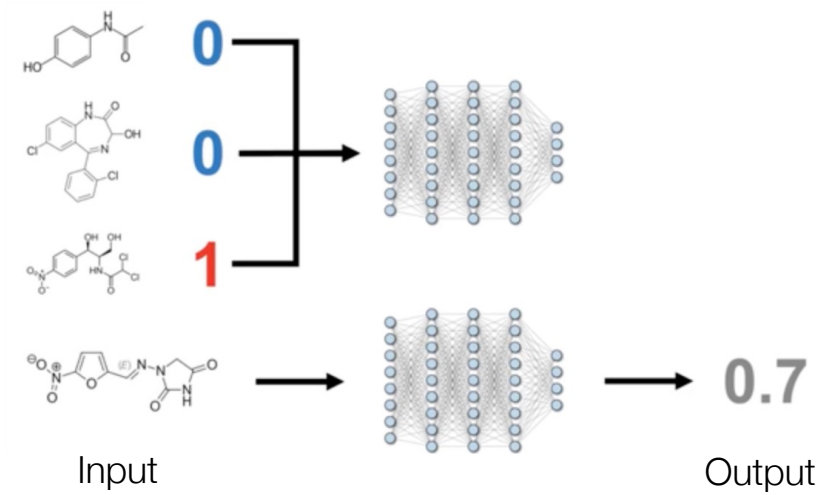
GNN to learn molecular structure



A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

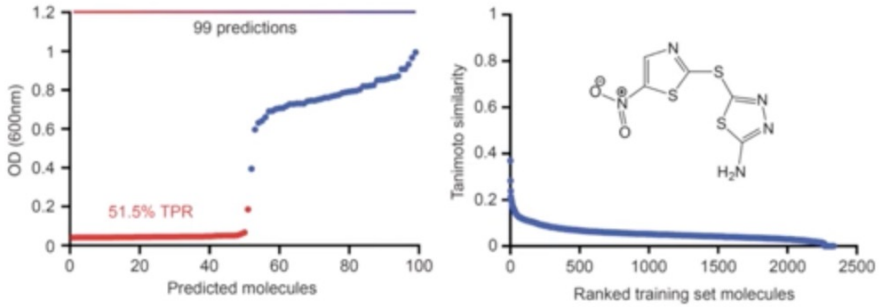
Experimental setup

Training Dataset
(Human Medicines and Natural Products)



Data: 2,335 molecules (human medicines and natural products) screened for growth inhibition

Empirical Validation
(Broad Repurposing Hub)



Data: 6,111 molecules (at various stages of investigation for human diseases) in the Repurposing Hub

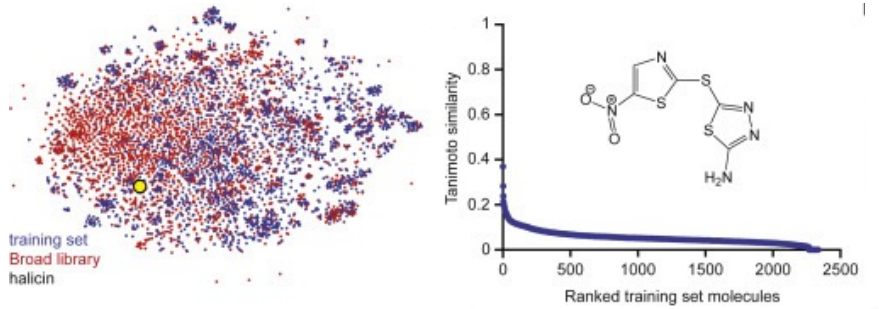
Task: Test top 99 predictions & prioritize based on similarity to known antibiotics or predicted toxicity

A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

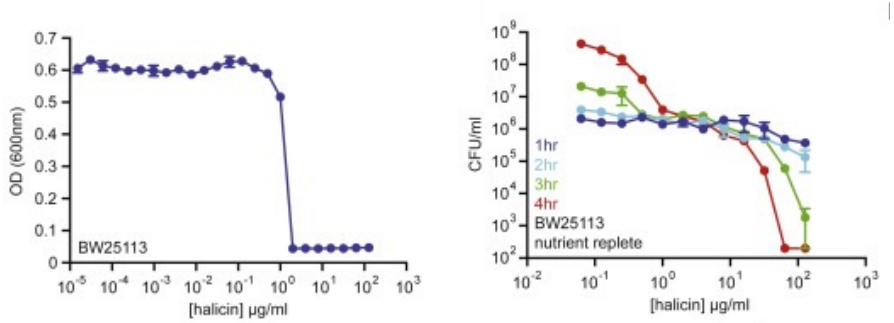
Chemical screening results

Halicin, initially developed as anti-diabetic drug (but discontinued due to poor results in testing), is identified and verified through experiments as a promising antibiotic

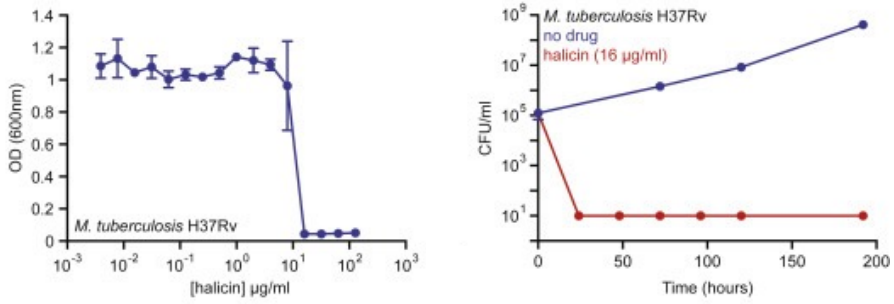
Halicin predicted to be antibacterial



Halicin against *E. coli*



Halicin against *M. tuberculosis*

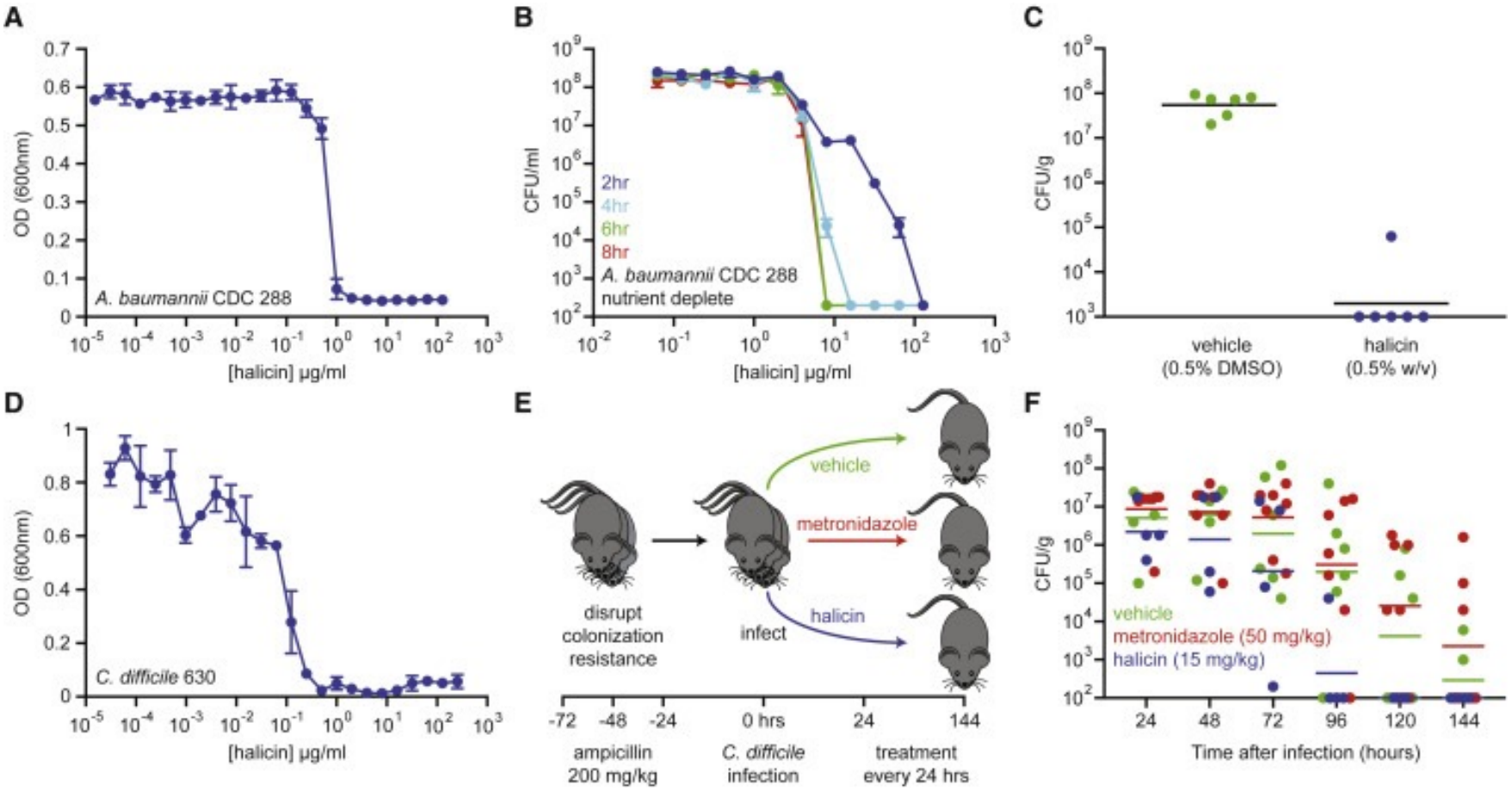


A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

We will learn about this topic in **Module 3: Graph Learning** and **Module 6: Therapeutic Science**

Chemical screening results

Halicin's efficacy in murine models of infection



A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020.

Outline for today's class

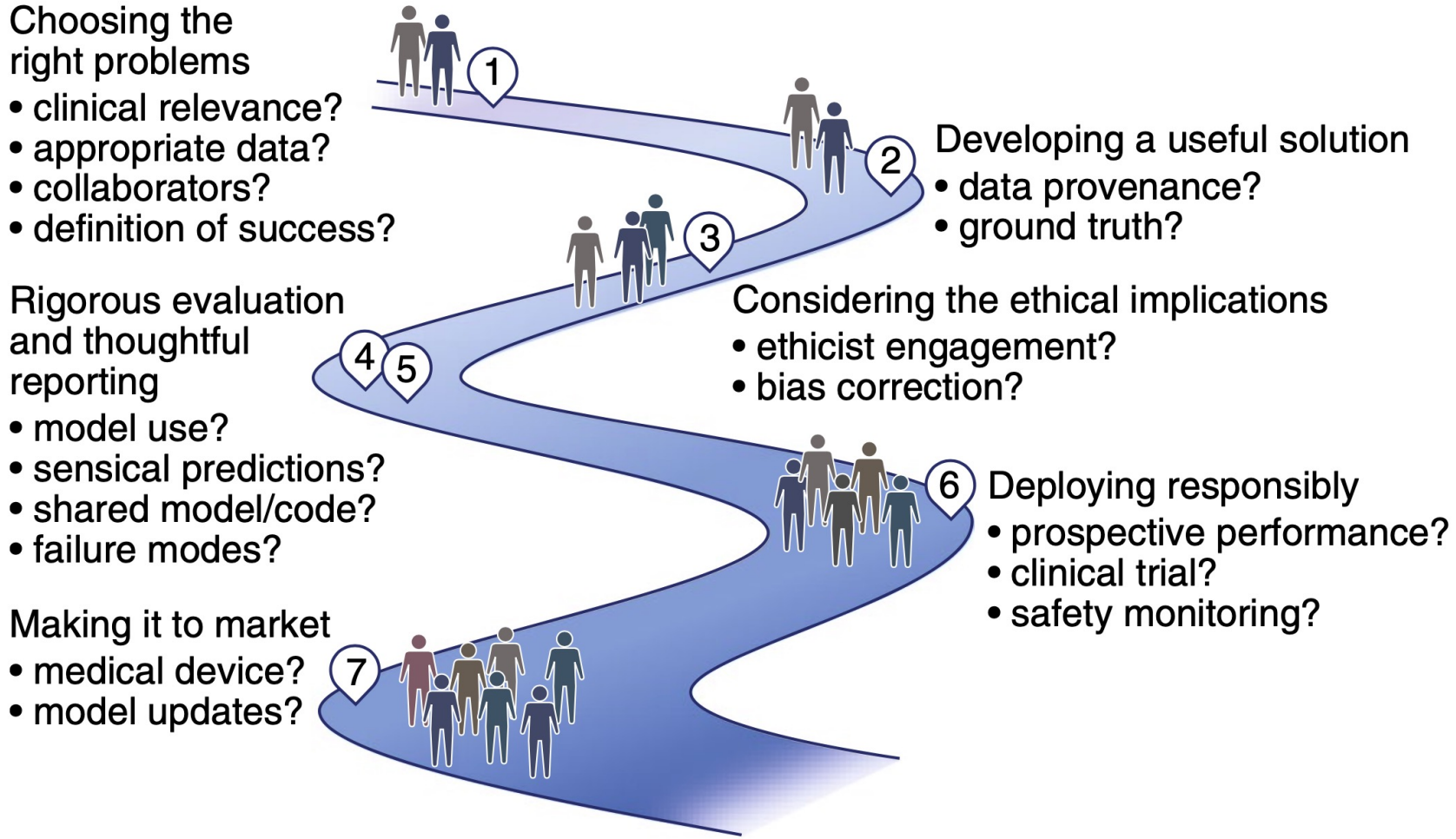
1. Overview of syllabus

2. What makes biomedical data unique

3. Motivation for machine learning

4. Roadmap for responsible biomedical AI

Roadmap to develop, validate and implement ML methods



Do no harm: a roadmap for responsible machine learning for health care, *Nature Medicine*, 2019

Key ML elements (1/2)

- Examples:
  - Also known as ‘samples’ or ‘observations’, basic units being measured
  - Primary data objects being manipulated by an ML model
- Features:
  - Properties of a given example, also known as ‘covariates’
  - For example, the gene expression values associated with a gene or the sequence patterns associated with a genomic window
- Labels/Outcomes:
  - Outcomes are what we want to predict in supervised learning
  - For example, the functional class assigned to a gene or the binary classification of whether a given genomic window contains a promoter
    - In classification, the outcome is a category, known as ‘label’ or ‘class’
    - In regression, the outcome is a real number

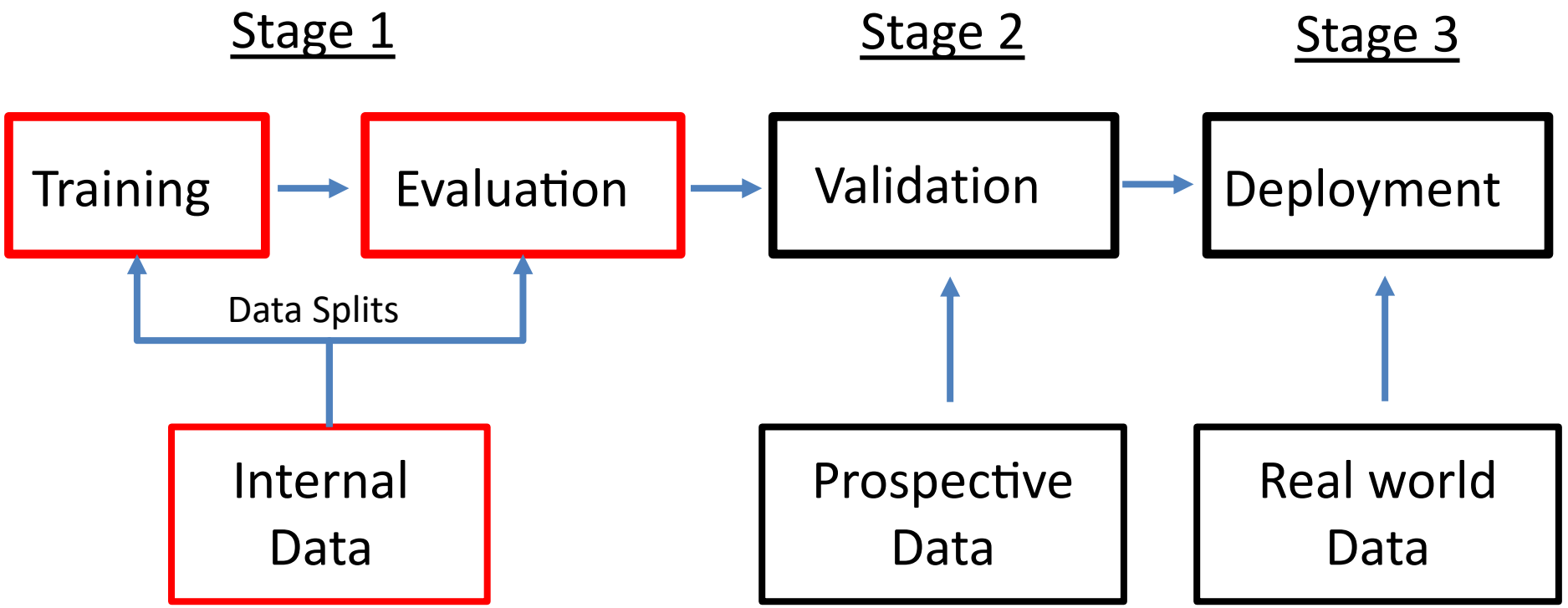
Navigating the pitfalls of applying machine learning in genomics, *Nature Reviews Genet*, 2021

Key ML elements (2/2)

- **Training set:** Examples and associated outcomes used to fit an ML model
  - **Positives:** Examples with the outcome of interest in a binary classifier
  - **Negatives:** Examples with the alternative outcome in a binary classifier
- **Test set:** Examples and associated outcomes that are used to evaluate model performance
  - No examples are shared between training and test sets
- Once we identify the specific ML problem that will be solved, we must train a model and determine how to properly evaluate its performance
- Performance evaluation is often executed using **cross-validation,** whereby examples are iteratively randomized into a training set used to train a model and a held-out test set used to quantify model performance
- **Prediction set:**
  - Examples whose associated outcomes are truly not known, where a fitted model is applied to make predictions
  - Also known as a prospective validation set

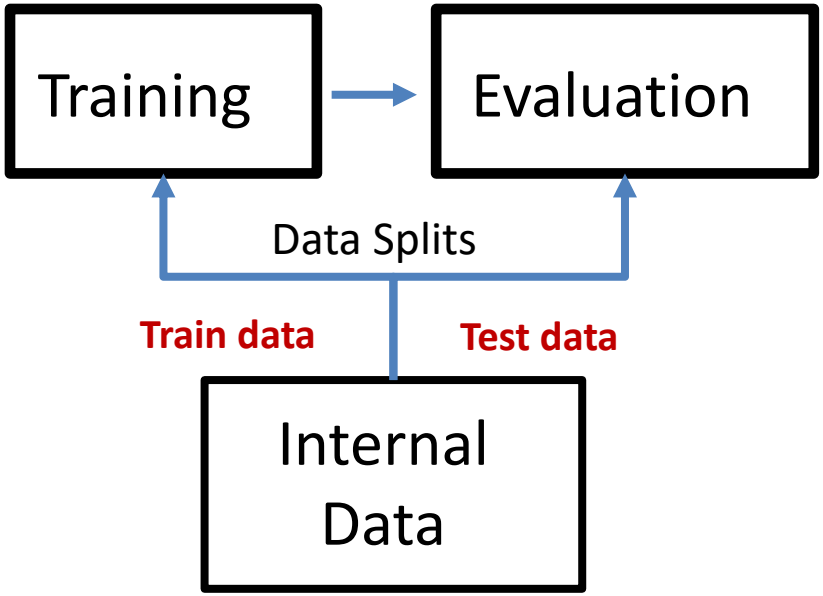
Navigating the pitfalls of applying machine learning in genomics, *Nature Reviews Genet*, 2021

Roadmap for ML development



Stage 1: Algorithm development

- Stage 1 focused on designing and developing initial models
- Use historical or retrospective data not originally collected for developing ML models
- Most of the data used as training data and a small part serve as a held-out test set



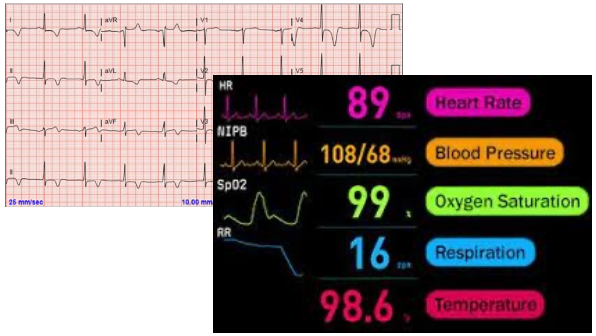
Biomedical data modalities

Images

Text

Reproducibility has been an important and intensely debated topic in science and medicine for the past few decades.[1] As the scientific enterprise has grown in scope and complexity, concerns regarding how well new findings can be reproduced and validated across different scientific teams and study populations have emerged. In some instances,[2] the failure to replicate numerous previous studies has added to the growing concern that science and biomedicine may be in the midst of a "reproducibility crisis." Against this backdrop, high-capacity ma-

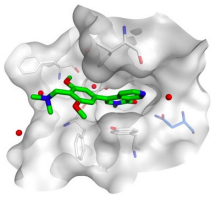
Sensors and time series

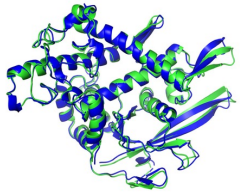


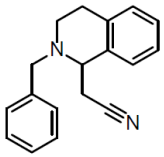
Tabular data

	A	B	C	D	E
1	Outcome	Variable 1	Variable 2	Variable 3	Variable 4
2	0	0.61139659	0.48680968	0.11562259	0.87902761
3	1	0.12829688	0.5486727	0.15348195	0.72252785
4	0	0.71351147	0.75832487	0.90316662	0.88221083
5	1	0.34695312	0.54108338	0.14877813	0.56189187
6	1	0.34725259	0.45194368	0.80122813	0.00010513
7	1	0.92022489	0.88425775	0.74030775	0.17102558
8	0	0.03464291	0.24279388	0.5668997	0.98930859
9	0	0.04666206	0.03866196	0.45542366	0.80195529
10	0	0.97446704	0.69276583	0.83352005	0.59603451
11	0	0.93820124	0.13989022	0.48967383	0.12350255
12	0	0.97684371	0.43805453	0.47441185	0.177821
13	1	0.64077343	0.72296037	0.67427288	0.31867675

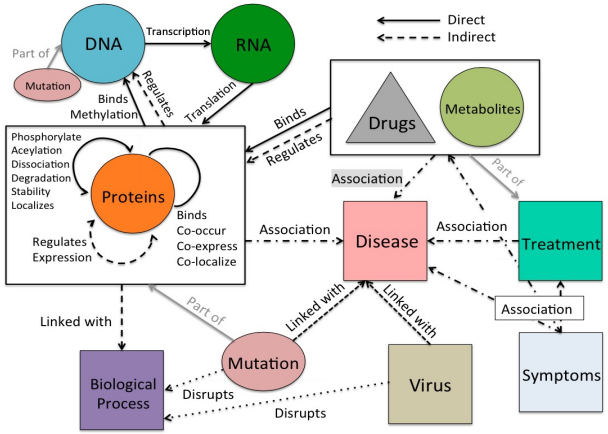
Sequences and molecules







Networks and biomedical knowledge



The importance of labels

- The labels, or ground-truth diagnoses, are often very hard, expensive and time consuming to obtain, but are usually the most important part of building an ML system
- In medicine they are often provided by physicians or other healthcare workers -> time consuming and noisy: Doctors don't always agree!
- The quality of your labels will "upper-bound" the performance of the system – we cannot be more accurate than the labels!

Evaluating performance

- Ideally, you would like the system to optimize for something you care about, e.g., outcomes, cost, etc. but those are only measurable in Stage 2/3
- Instead, we use proxy metrics during Stage 1, e.g.:
  - Classification accuracy
  - Sensitivity/Specificity
  - Precision/Positive predictive value (PPV = #TP / #pos-calls)
  - Area under the ROC curve (AUROC)
  - Area under the precision-recall curve (AUPRC)
- There is not usually an objectively good metric
  - The choice of a metric is application-dependent

Stage 1: Challenges

- ML models can “fail” at this stage for a variety of reasons:
  - Not enough data and insufficient model capability – model is not a meaningful advance over current methods
  - Model is good but is hard to integrate into biomedical and clinical workflows

- Worst case: Model looks good but is subtlety overfit or confounded in a way that is very hard to detect

Stage 1: Example challenge

- We want to develop an ML model to automatically screen for diabetic retinopathy
- We go to a local EHR and download every patient image that had diabetic retinopathy
- We train a model and get an AUROC of 0.99!
- We look at the image with the highest disease risk and see the following image. Is anything wrong?

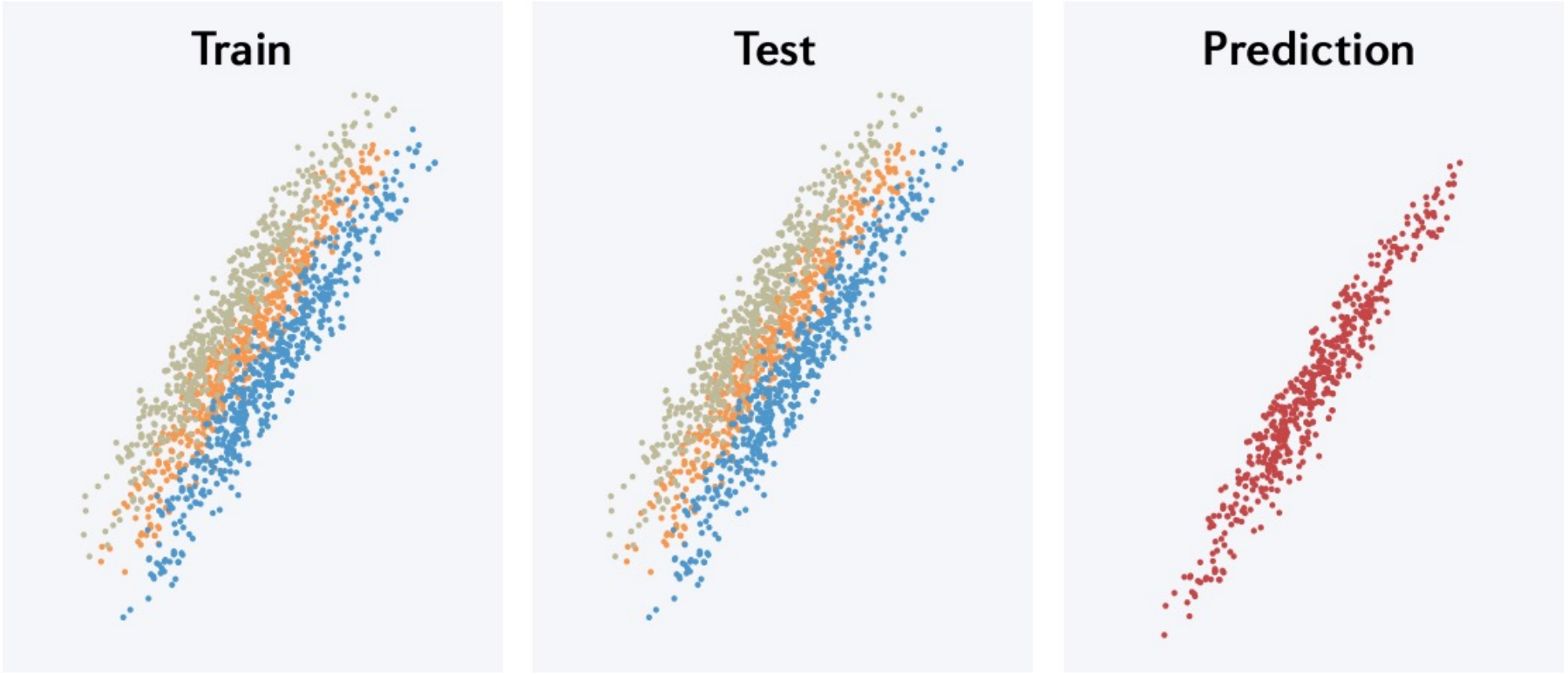
Source: endotext.com

Stage 1: Example challenge

- Training set contained patients who were being treated for the condition and had laser scars
- The algorithm learned to associate the presence of laser scars with diabetic retinopathy
- This will not help in treatment-naïve patients when screening – actual performance is much worse

Source: endotext.com

Quick check: What is the problem with this data split?



Join at **slido.com** with **#033364**

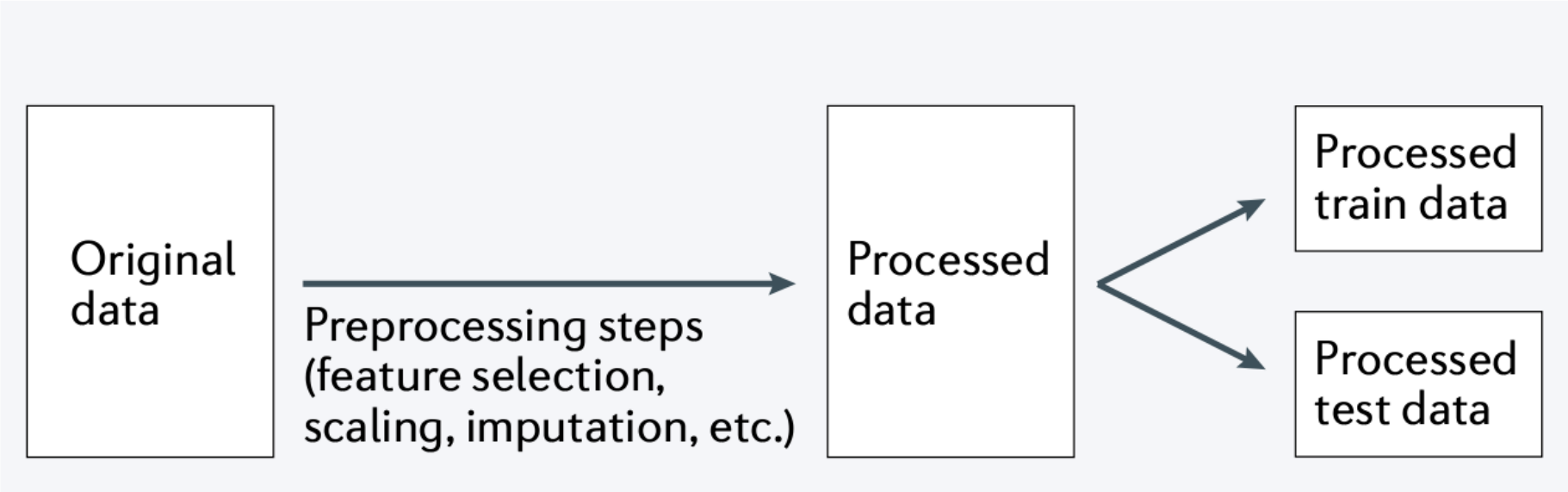
Quick check: What is the problem with this data split?

Distributional differences can arise from various sources, such as batch effects. If training and test sets are a mixture of examples from every batch (left), performance on the test set will be much higher than on a new batch

To fit a model that will generalize to new batches, training and test sets should be composed of different batches (right)

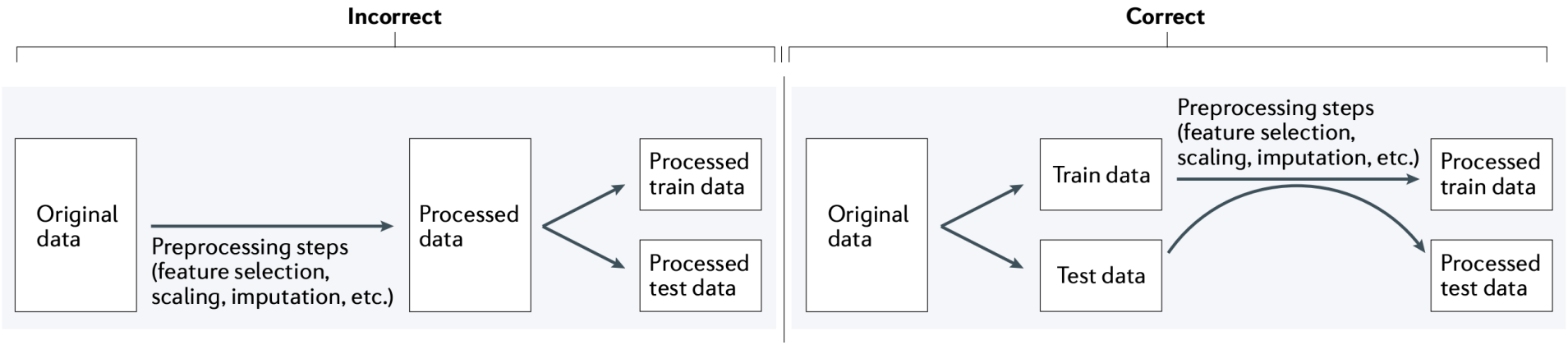
Navigating the pitfalls of applying machine learning in genomics, *Nature Reviews Genet*, 2021

Quick check: What is the problem with this ML workflow?



Join at **slido.com** with **#033364**

Quick check: What is the problem with this ML workflow?

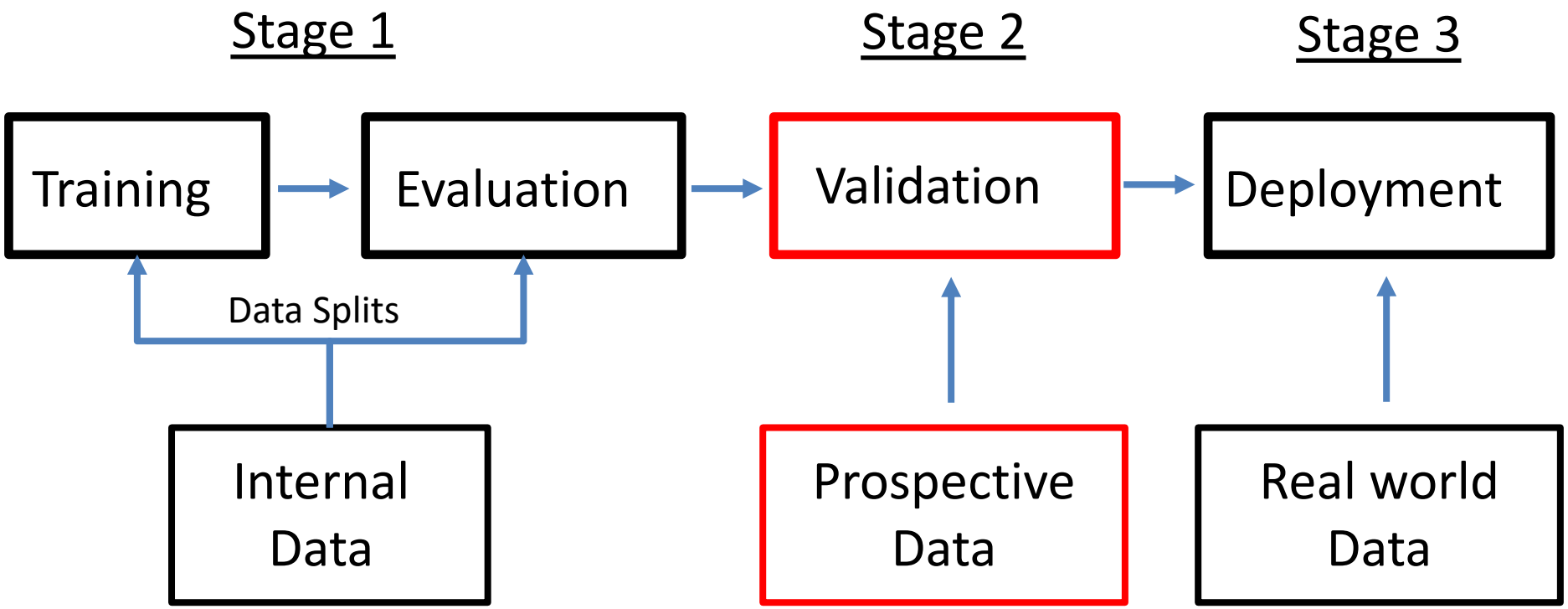


Information leakage can happen when information is leaked from the test set into the training as a result of the training and test sets being preprocessed together (left)

Instead, the raw data should be split into training and test sets with preprocessing performed separately (right)

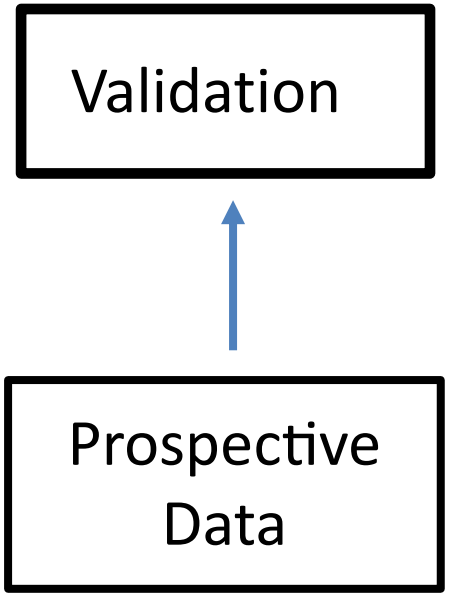
Navigating the pitfalls of applying machine learning in genomics, *Nature Reviews Genet*, 2021

Roadmap for ML development

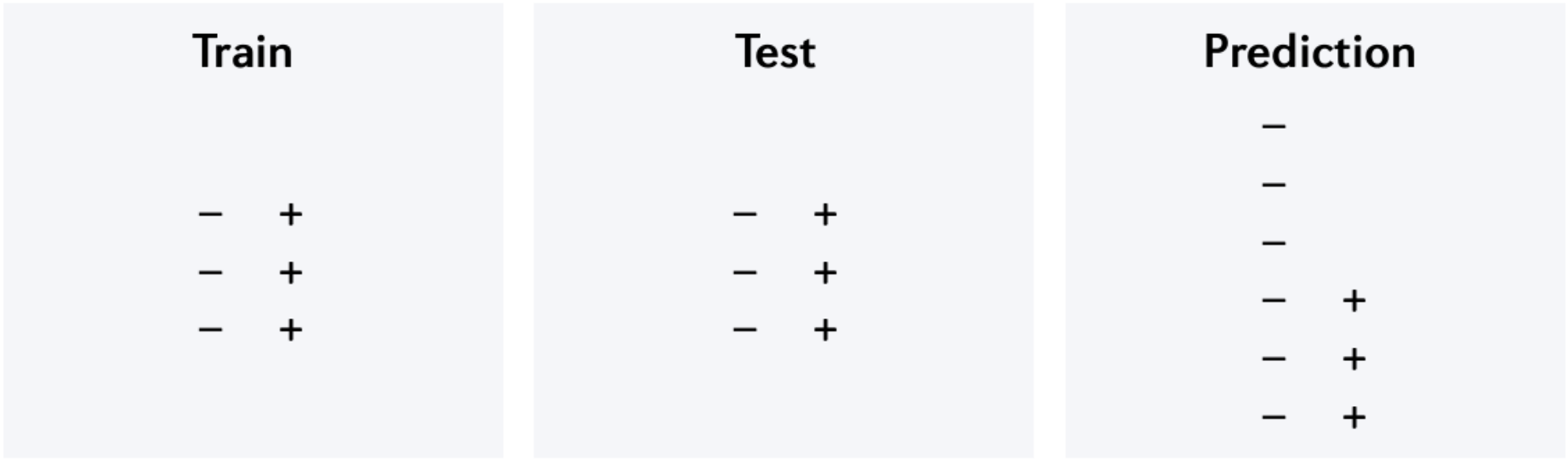


Stage 2: Prospective validation

- Stage 2 is focused on prospectively validating the model from Stage 1 on “live” data coming in under a (somewhat) controlled setting
- Often follows a trial format with pre-registered endpoints and set for a fixed amount of time
- Goal is to show the model performs well and generalizes to real-world data

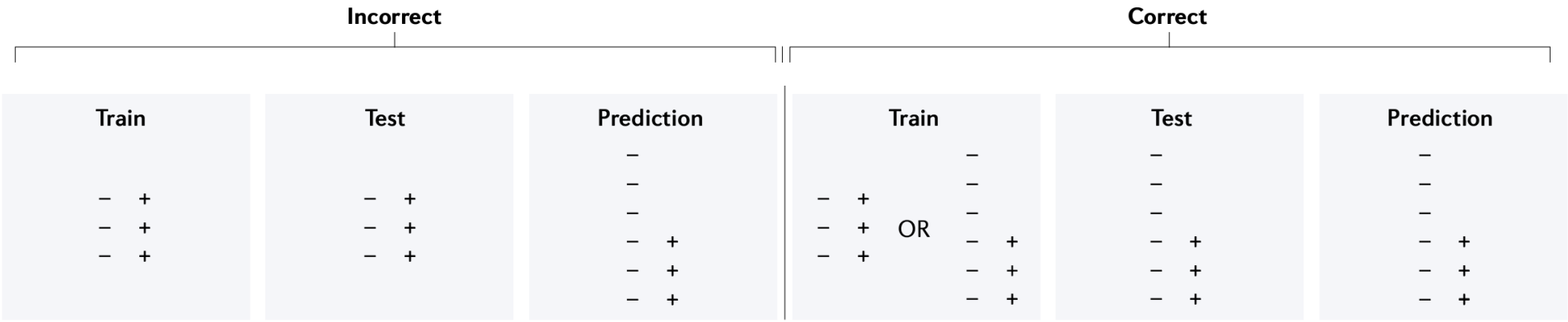


Quick check: What is the problem with this experimental setup?



Join at **slido.com** with **#033364**

Quick check: What is the problem with this experimental setup?



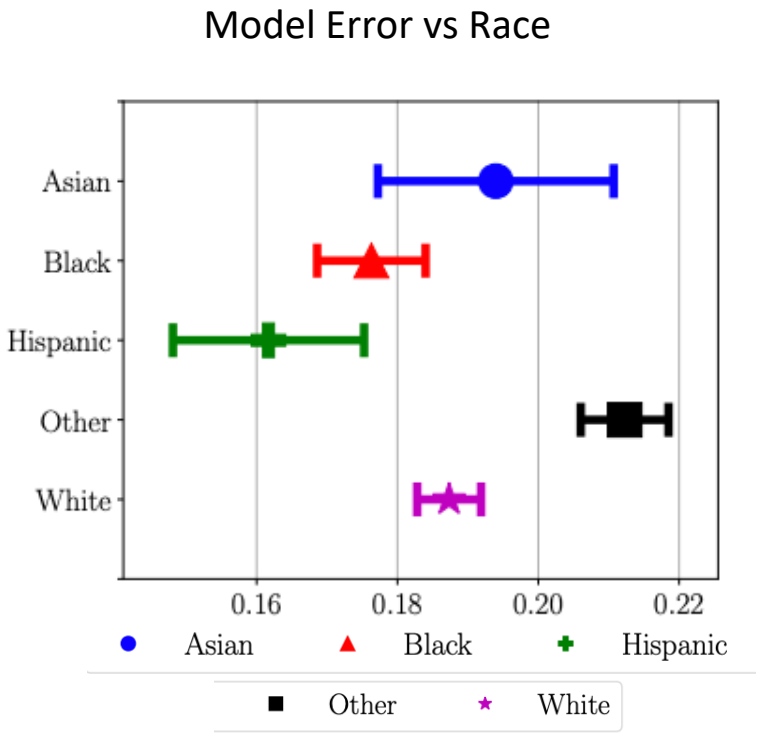
Unbalanced data can make model training and evaluation difficult. If the training and test sets are balanced but the prediction set is unbalanced, test set performance will not reflect prediction set performance (left)

Regardless of whether a balanced or unbalanced training set is used, the imbalance in the test set should be reflective of the imbalance in the prediction set (right). Ideally, one should also use a performance measure that can handle imbalance

Navigating the pitfalls of applying machine learning in genomics, *Nature Reviews Genet*, 2021

Stage 2: Example challenge

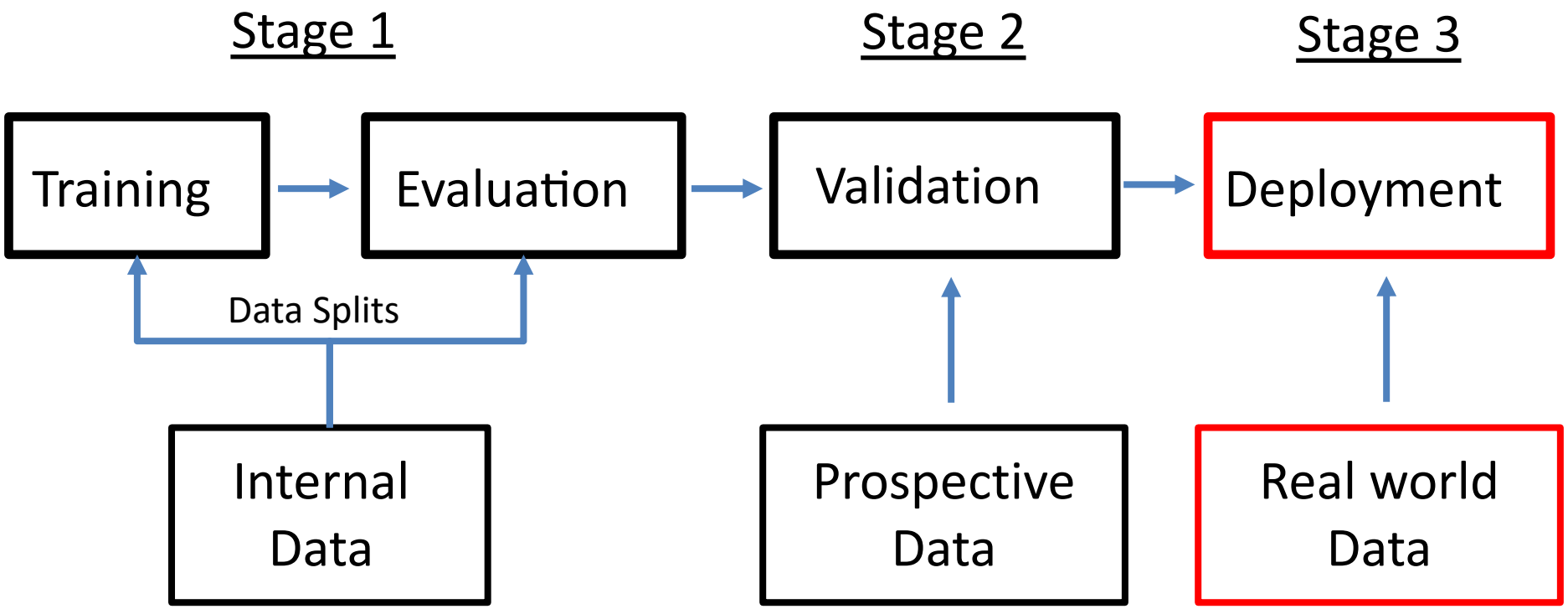
- Prospective evaluation might include many new clinical site in order to assess the generalizability of ML model
- These sites include populations rarely seen in retrospective, single-site data
- Results stratified by demographics produce a graph shown on the right



Source: Chen et al, NeurIPS '18

Deep transfer learning for reducing health care disparities arising from biomedical data inequality, *Nature Commun*, 2020

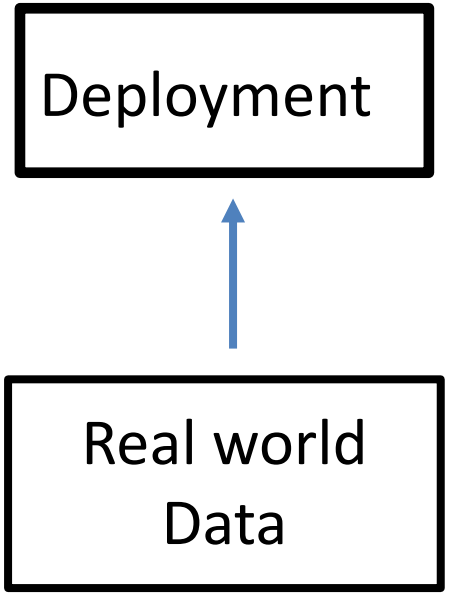
Roadmap for ML development



Stage 3: Deployment

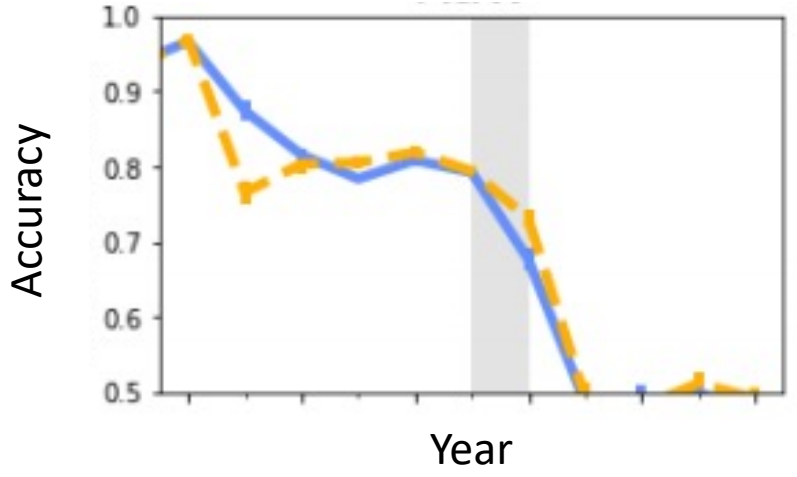
- Stage 3 has likely been the ultimate goal

- ML model is implemented in a biomedical or clinical settings and used to guide experiments in the laboratory or provide decision support



Stage 3: Example challenge

- ML model has been extensively validated and shown to be very accurate
- We implement the model and suddenly notice a huge drop in performance during an audit
- What could be going on?

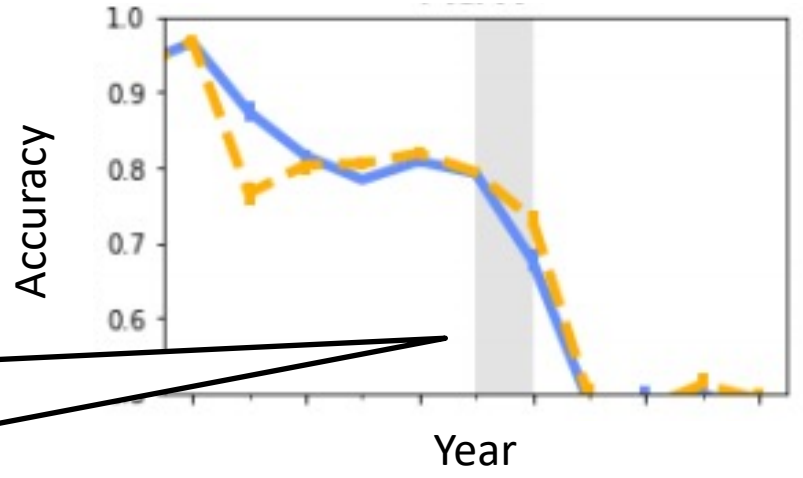


Source: Nestor et al, 2018

Stage 3: Example challenge

- ML model has been extensively validated and shown to be very accurate
- We implement the model and suddenly notice a huge drop in performance during an audit
- What could be going on?

Answer: Your hospital updated its EHR to a new version.
Your AI system was completely tied to the old way the data were recorded and now no longer works.



Source: Nestor et al, 2018

Stage 3: Example challenge

Two examples of problematic AI systems:

1. **Unfairness:** Criminal Risk Assessment Tools
   - Defendants are assigned scores that predict the risk of re-committing crimes.
   - These scores inform decisions about bail, sentencing, and parole. Current systems have been accused of being biased against black people
2. **Biases:** Face Recognition Systems
   - Considered for surveillance and self-driving cars.
   - Current systems have been reported to perform poorly, especially on minorities

Criminal Risk Assessment Tools
The COMPAS debate

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPu

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

COMPAS debate

- Correctional Offender Management Profiling for Alternative Sanctions
- Used in prisons across country: AZ, CO, DL, KY, LA, OK, VA, WA, WI
- “Evaluation of a defendant’s rehabilitation needs”
- Recidivism = likelihood of criminal to reoffend

COMPAS (continued)

"Software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result […] algorithms can reinforce human prejudices." (Miller (2015))

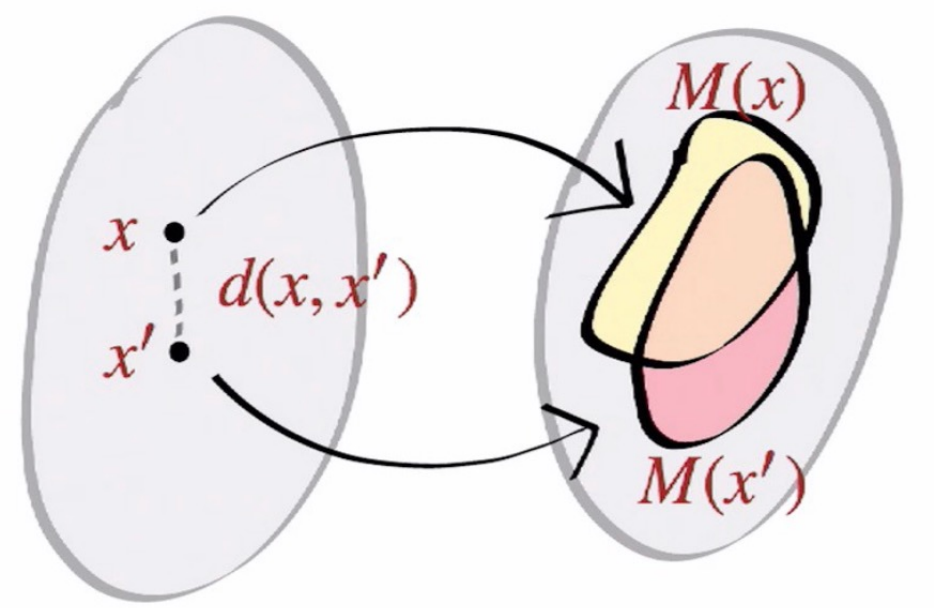
How to mitigate this issue?
1) Fairness through unawareness

- **Idea:** Don't record protected attributes, and don't use them in your algorithm
  - Predict Y from features X and group A using $P(\hat{Y} = Y|X)$ instead of $P(\hat{Y} = Y|X, A)$
- **Pros:** Guaranteed to not be making a judgement on protected attribute A
- **Cons:** Other proxies may still be included in a "race-blind" setting, e.g. zip code or conditions

We will learn about this topic in **Module 2: Trustworthy AI**

How to mitigate this issue? 2) Individual fairness

- **Idea:** Similar individuals should be treated similarly
- **Pros:** Can model heterogeneity within each group
- **Cons:** Notion of "similar" is hard to define mathematically, especially in high dimensions



We will learn about this topic in **Module 2: Trustworthy AI**

How to mitigate this issue? 3) Group fairness

- **Idea:** Require prediction rate be the same across protected groups
  - E.g. "20% of the resources should go to the group that has 20% of population"
- Predict Y from features X and group A such that $P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$
- **Pros:** Literally treats each race equally
- **Cons:**
  - Too strong: Groups might have different base rates. Then, even a perfect classifier wouldn't qualify as "fair"
  - Too weak: Doesn't control error rate. Could be perfectly biased (correct for A=0 and wrong for A=1) and still satisfy

We will learn about this topic in **Module 2: Trustworthy AI**

Outline for today's class

1. Overview of syllabus ✓
2. What makes biomedical data unique ✓
3. Motivation for machine learning ✓
4. Roadmap for responsible biomedical AI ✓