GLOSSARY

Anika Gupta and Guests

Listen to all episodes at **linktr.ee/thedatapulse**

Attributions: cover art by **Divya Vatsa;** background music: Splash and Memory, released under Creative Commons Zero

Note: some terms may have multiple definitions--the meanings included here are in the context of the podcast's theme; as this document serves as a compilation, definitions are not my own but rather taken and adapted from the web

Speaker-specific terms and resources

Andrew Beck

(PathAI)

Terms: CNN, COMPANION DIAGNOSTIC, FEATURES, GROUND TRUTH, HUMAN GENOME PROJECT, HYPERPARAMETERS, IMMUNOHISTOCHEMISTRY, INDICATION, MEDICAL EVIDENCE GENERATION, METASTATIC CANCER, MODEL ARCHITECTURE, TENSORFLOW, TRAINING / VALIDATION / INDEPENDENT TEST SET, WHOLE SLIDE IMAGES

Resources:

- Company website: pathai.com
- Papers:
    - Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival
    - Deep Learning for Identifying Metastatic Breast Cancer
    - Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer
    - Machine learning-based identification of predictive features of the tumor micro-environment and vasculature in NSCLC patients using the IMpower150 study
    - Association of digital and manual quantification of tumor PD-L1 expression with outcomes in nivolumab-treated patients

Ankit Gupta

(Reverie Labs)

Terms: ASSAY, BLOOD-BRAIN BARRIER, CONSTRAINT, CONTRACT RESEARCH ORGANIZATION (CRO), ENRICHMENT, HIGH THROUGHPUT SCREEN, IMAGENET, LEAD OPTIMIZATION, MULTIPARAMETER OPTIMIZATION, R-SQUARED, RMSE, RESNET, RNNS, SELECTIVITY AND PK PROFILES, SMALL MOLECULE LIBRARY, SOLUBILITY PROBLEM, TENSORFLOW, TRANSCRIPTION FACTOR BINDING

Resources:

- Company website: reverielabs.com
- Papers:
    - Kinase inhibitors review
    - Very broad overview of ML in drug discovery, with a nice overview of several modeling methods to the field
    - Brain penetrant kinase inhibitors
    - Active Learning/Dataset Generation

Anthony Philippakis

(Google Ventures, Broad Institute)

Terms: CARDIOLOGY, CONSUMER TECHNOLOGY, DEEP LEARNING, DEFIBRILLATOR, DEVOPS, DISTRIBUTED SYSTEMS, GENOME BUILD, HUMAN GENOME PROJECT, INFORMATION THEORY, MAP-REDUCE, NEWTONIAN CALCULUS, PATTERN RECOGNITION, REPRESENTATION THEORY, SIGNAL PROCESSING, SPARK, STENT

Ava Soleimany

(Harvard University, MIT)

Terms: ACTIVITY-BASED DIAGNOSTICS, BIOMARKER, ENZYME, FALSE POSITIVE RATE, NANOPARTICLE, PEPTIDE, PROTEASE, RANDOM FOREST CLASSIFIER, RNA-SEQ, SENSE-AND-RESPOND, SENSITIVITY, SPECIFICITY, STATE MACHINE

Resources:

- Personal website
- Papers:
    - Activity-based diagnostics: an emerging paradigm for disease detection and monitoring
    - Urinary detection of lung cancer in mice via noninvasive pulmonary protease profiling
- Introduction to Deep Learning course website

Aviv Regev

(Broad Institute → Genentech)

Terms: ANTIBODY, BIAS, CELL STATE, CELL TYPE, CELLULAR PROGRAM, CIS-REGULATORY ELEMENTS, COMMON VARIATION, DIFFERENTIATION, EFFECT SIZE, ELECTROPHYSIOLOGY, FLUORESCENCE ACTIVATED CELL SORTING (FACS), FUNCTIONAL GENOMICS, GENE EXPRESSION PROFILE, GENETIC INTERACTIONS, GENETIC INTERVENTION, GENETIC VARIATION, GENOME, HIGH DIMENSIONAL DATA, HUMAN CELL ATLAS, HUMAN GENOME PROJECT, HYBRIDIZATION (WITH PROBES), INFERENCE, LATENT REPRESENTATION, MASSIVELY PARALLEL, MEASURE-MODEL-PERTURB, MODEL ORGANISM, MOLECULAR, PROFILE, OLIGONUCLEOTIDE, ORDERS OF MAGNITUDE, ORGANOID, PATTERN RECOGNITION, PERTURB SEQ, PERTURBATION, PHYSIOLOGY, RANDOM SAMPLING, READ DEPTH, REGULATORY SEQUENCE, SEARCH SPACE, SINGLE CELL RNA-SEQ, STATE SPACE

Resources:

- Company website: gene.com
- Human Cell Atlas

Carlos Bustamante

(Stanford University, F-Prime)

Terms: HLA SYSTEM, NEWBORN SCREENING PROGRAMS, ONLINE CONSENT PROCESS, TELEMEDICINE, PATHOGEN, DIRECT-TO-CONSUMER, SEQUENCING

Resources:

- Lab website: bustamantelab.stanford.edu
- F-Prime website

Corey McCann

(Pear Therapeutics)

Terms: API (ACTIVE PHARMACEUTICAL INGREDIENT), BREAKTHROUGH DESIGNATION, CLAIMS DATA, CLINICAL ENDPOINTS, EFFICACY, FDA, LATE STAGE ASSETS, LONGITUDINAL DATA, MARKET AUTHORIZATION, MEDICAL DEVICE, PATIENT ENGAGEMENT, PAYERS, PRESCRIPTION, PRESCRIPTION DIGITAL THERAPEUTICS (PDTS), PSYCHIATRIC CONDITIONS, RESPONDERS/ NON-RESPONDERS, SAFETY, TELEMEDICINE, THERAPEUTIC MODALITY, WEARABLE SENSOR

Resources:

- Company website: peartherapeutics.com

Daphne Koller

(insitro)

Terms: BILINGUAL, BIOMARKER, CELL DIFFERENTIATION, CELL ENGINEERING, DIGITAL BIOLOGY, LESION, MARKET CONDITIONS, MINIMUM VIABLE PRODUCT (MVP), MODEL SYSTEMS, MORTALITY RATE, MULTI-TASK LEARNING, PATIENT POPULATION, PATTERN RECOGNITION, SEMI-SUPERVISED LEARNING, TARGET, TRAINING DATA, TRANSFER LEARNING, UNMET NEED, ZERO-SHOT LEARNING

Resources:

- Company website: insitro.com

Dennis Wall

(Stanford University)

Terms: ACCURACY, ALTERNATING DECISION TREES, AUGMENTED REALITY, BINARY VS CONTINUOUS OUTCOME, CATEGORICAL VARIABLE, CLASSIFIER, CONTINUOUS PHENOTYPE, DIGITAL THERAPEUTICS, FEASIBILITY TEST, FEATURE IMPORTANCE, FEATURES, GENES + ENVIRONMENT = PHENOTYPE, ICD BILLING CODES, INCIDENCE, INTER-RATER RELIABILITY, INTERPRETABILITY, LINEAR MODELS, POLYGENIC, PRECISION HEALTH, RANDOMIZED CONTROL TRIAL, TELEMEDICINE, THERAPEUTIC DOSE

Resources:

- Lab website: wall-lab.stanford.edu
- Company website: cognoa.com

Elaine Nsoesie

(Boston University)

Terms: BUILT ENVIRONMENT, COMPUTATIONAL SOCIAL SCIENTIST, DEEP LEARNING, DIGITAL PRESENCE, EPIDEMIC, FEATURE, FOODBORNE ILLNESSES, INTERVENTION, LINEAR REGRESSION, MISINFORMATION, PANDEMIC, PREVALENCE, SOCIOECONOMIC STATUS, UNITED NATIONS INNOVATION LAB

Resources:

- AI and the built environment
  - JAMA Paper 1
  - JAMA Paper 2
- Representation in digital data for public health research
  - Preventive Medicine paper
  - PLOS Paper
- COVID-19, social distancing and income
- Harvard Kennedy School article on misinformation
- Recent work that wasn't discussed. Disparities in Brazil's homicide trends.

Francesca Dominici

(Harvard University)

Terms: ASSOCIATION, CAUSAL INFERENCE, CONFOUNDERS, ENSEMBLE LEARNING, FINE PARTICULATE MATTER, HIERARCHICAL MODEL, INVERSE PROBABILITY WEIGHTING, MEDICARE, MORTALITY RATE, MULTI-SITE TIME SERIES STUDY, OBSERVATIONAL STUDY, PROPENSITY SCORE MATCHING, RANDOMIZED TRIAL, REGRESSION, SATELLITE DATA, SPATIAL CORRELATION

Resources:

- About Francesca (HSPH)
- Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly

Gaurav Singal

(previously at Foundation Medicine)

Terms: BIOMARKER, CAUSALITY, CLASS OF MUTATIONS, CLINICAL ENDPOINTS, COMPOSITE BIOMARKERS, CONTINUOUS VARIABLES, CORRELATION, DATA ASSET, DATA GOVERNANCE, DATA LIQUIDITY, GENOME, INFORMATION INSIGHTS COMPANY, MOLECULAR CHARACTERIZATION (OF TUMORS), OBSERVATIONAL/REAL-WORLD DATA, RANDOMIZED CONTROL TRIAL, USE CASE/PROBLEM-BACKED

Resources:

- Company website: foundationmedicine.com

Greg Ryslik

(Celsius Therapeutics)

Terms: CLASSIFIER, EXPRESSION PROFILE, FISH, IHC, LOSS FUNCTION, NOISY DATA, PASSENGER MUTATIONS, SINGLE CELL RNA SEQUENCING, SPARSE DATA, STATISTICAL CONTROL, T-SNE PLOTS

Resources:

- Company website: celsiustx.com
- Single cell RNA sequencing primer
- Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis

Imran Haque

(Recursion Pharmaceuticals)

Terms: ANTIBODY, ASSAY, CELL LINES, CLASSIFICATION, COMPUTER VISION, CONSTRAINT, CYTOTOXICITY, DEEP LEARNING, DENSENET (DENSE CONVOLUTIONAL NETWORK), DROPOUT, DRUG REPURPOSING, EXPRESSION PROFILE, FEATURE IMPORTANCE, FEATURE VECTOR, FEATURES, GENETIC KNOCKOUT, IMPUTATION, INTERPRETABILITY, MODEL ARCHITECTURE, MODEL SYSTEMS, MORPHOLOGICAL PROFILING, MULTI-CHANNEL FLUORESCENCE MICROSCOPY, PERTURBATION, PHENOMICS, PRIMARY CELLS, PSEUDOVIRUS, SMALL INTERFERING RNA (siRNA), STATISTICAL CONTROL, STUDENT-TEACHER MODELS, VIRAL TITER/VIRAL LOAD, WHO (WORLD HEALTH ORGANIZATION)

Resources:

- Company website: recursionpharma.com
- Latest preprint: Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and COVID-19 drug discovery
- Data releases all are available at rxrx.ai; the new ones specifically are rxrx.ai/rxrx2 and rxrx.ai/rxrx19b
- The first morphological imaging dataset on SARS-CoV-2 virus
- Slides on COVID work
- Identification of potential treatments for COVID-19 through artificial intelligence-enabled phenomic analysis of human cells infected with SARS-CoV-2
- Altitude Lab Incubator
- Neural Information Processing Systems conference

Isaac Kohane

(Harvard Medical School)

Terms: API, CENTROID, CLUSTER, COAGULOPATHY, DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS (DSM), EHR/EMR, GROWTH HORMONE DEFICIENCY, HYPOTHYROIDISM, IDIOPATHIC, INFORMATICS FOR INTEGRATING BIOLOGY AND THE BEDSIDE (I2B2), NAIVE BAYES CLASSIFIER, PEDIATRIC ENDOCRINOLOGIST, PENETRANCE, PITUITARY GLAND, PROBABILISTIC MODELING, REGULATORY ELEMENT, SENSITIVITY, SPECIFICITY, TEXT CORPUS, UNDIAGNOSED DISEASES NETWORK (UDN), WORD EMBEDDING

Resources:

- i2b2 efforts to respond to COVID-19 by aggregating EMR/EHR data across hospitals around the world
- Journals: New England Journal of Medicine (NEJM), Association for the Advancement of Artificial Intelligence (AAAI)

Kyle Swanson

(Cambridge University, MIT)

Terms: BI-RADS SCORE, BIOPSY, BREAST DENSITY, CNN (CONVOLUTIONAL NEURAL NET), FEATURE VECTORS, GRAPH NEURAL NETWORKS, MAMMOGRAPHY, MULTI-DRUG RESISTANT BACTERIA, PRE-TRAINING, PRE-TRAINING, RANDOM FOREST, RESNET, RISK FACTORS, SOLUBILITY, SUPPORT VECTOR MACHINES, TOXICITY, VPN

Resources:

- Mammography
    - Paper in *Radiology*: Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation
    - MIT News: Automated system identifies dense tissue, a risk factor for breast cancer, in mammograms
- Molecular Property Prediction
    - Paper in *JCIM* about general property prediction: Analyzing Learned Molecular Representations for Property Prediction
    - Paper in *Cell* about antibiotic discovery: A Deep Learning Approach to Antibiotic Discovery
    - Nature News: Powerful antibiotics discovered using AI
    - Chemprop (ML Model for Property Prediction) Code
    - Chemprop Demo Website
    - AI Cures Website
    - Machine Learning for Pharmaceutical Discovery and Synthesis Consortium Website
- Personal Website

Lily Peng

(Google Health)

Terms: ASYMPTOMATIC, ATTENTION MODELING, AUC, CONFIDENCE INTERVAL, DEEP LEARNING, DIABETIC RETINOPATHY, EXPLAINABILITY, FEATURE IMPORTANCE, GROUND TRUTH, HEMORRHAGES, IMAGENET, MACULA, MICROANEURYSMS, MODEL GENERALIZABILITY, N, OVERFITTING, RETINAL FUNDUS, SALIENCY MAPPING, STEREOSCOPIC PHOTOGRAPHS, TRANSLATIONAL RESEARCH, VERSION CONTROL

Resources:

- Publications
    - Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic

Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
    - Krause, J. *et al.* Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* **125**, 1264–1272 (2018).
    - Raumviboonsuk, P. *et al.* Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med* **2**, 25 (2019).
    - Gulshan, V. *et al.* Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. *JAMA Ophthalmol.* (2019) doi:10.1001/jamaophthalmol.2019.2004.
    - India diabetes report 2010 — 2045. https://www.diabetesatlas.org/data/.
    - Beede, E. *et al.* A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020) doi:10.1145/3313831.3376718.
    - Varadarajan, A. V. *et al.* Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. *Nat. Commun.* **11**, 130 (2020).
- Google story on diagnosing diabetic retinopathy using deep learning
- TEDxGateway: Democratizing Healthcare With AI

Manisha Desai

(Stanford University)

Terms: ANTI-RETROVIRAL DRUGS, ATRIAL FIBRILLATION, CAUSAL INFERENCE, COMBINATION THERAPIES, CONFOUNDERS, CONTROLLED ENVIRONMENT, COUNTERFACTUAL, INDICATION, INVERSE PROBABILITY WEIGHTING, LONGITUDINAL COHORT, MARGINAL STRUCTURAL MODELS, MISSING DATA, MULTIPLE IMPUTATION, NOISY DATA, OBSERVATIONAL DATA, POSITIVE PREDICTIVE VALUE (PPV), PRAGMATIC TRIALS, PROSPECTIVE STUDIES, RANDOMIZED CLINICAL TRIAL, REGRESSION, SENSITIVITY ANALYSES, STUDY DESIGN, T-TEST, WILCOXON TEST

Resources:

- Society of Clinical Trials COVID Research Resource Hub: COVID DSMB registry and COVID Endpoints registry

Marzyeh Ghassemi

(University of Toronto)

Terms: ACCURACY, ACTION SPACE, ASSOCIATION, BIAS, CLASSIFICATION, COMPUTER VISION, CONTEXTUAL LANGUAGE MODELS, DIABETIC RETINOPATHY, DOSE-RESPONSE, FAIRNESS METRICS, GROUND TRUTH, HUMAN COMPUTER INTERACTION, IMPUTATION, MODEL TRAINING, POSITIVE PREDICTIVE VALUE (PPV), PUBMED, SCIBERT MODEL, STATE SPACE, TRANSFORMER MODELS, TURING TEST, WORD EMBEDDINGS

Resources:

- For some examples where models encode biases:
    - [1] Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. InProceedings of the ACM Conference on Health, Inference, and Learning 2020 Apr 2 (pp. 110-120).
    - [2] Chen, Irene Y., Peter Szolovits, and Marzyeh Ghassemi. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?." AMA journal of ethics 21.2 (2019): 167-179.
- For why classifiers tend to discriminate:
    - [3] Chen, Irene, Fredrik D. Johansson, and David Sontag. "Why is my classifier discriminatory?." Advances in Neural Information Processing Systems. 2018.
- For work on auditing models **using race/gender** to ensure fairness post-hoc:
    - [4] Datasheets for datasets. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). arXiv preprint arXiv:1803.09010.
    - [5] Model cards for model reporting. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). ACM.
    - [6] https://research.google.com/bigpicture/attacking-discrimination-in-ml

Peyton Greenside

(BigHat Biosciences)

Terms: ACTIVE LEARNING, ANTIBODY, BIOLOGIC, EXPLORATION-EXPLOITATION TRADEOFF, FEATURES, GENETIC SEQUENCE, GENETIC VARIANT, INTERPRETABILITY, OPTIMIZATION, PARAMETER SPACE, PHARMACEUTICAL

FORMULATION, RESIDUE, SEMI-SUPERVISED LEARNING, TRAINING/VALIDATION/INDEPENDENT TEST SET

Sam Sinai + Jeff Gerold

(Dyno Therapeutics)

Terms: AAV CAPSID PROTEIN, AMINO ACID, CAPSID FITNESS LANDSCAPE, CNN, DELIVERY MECHANISM, EDIT DISTANCE, ENSEMBLE MODELS, EXPLORATION-EXPLOITATION TRADE-OFF, GENE THERAPY, HIGH-THROUGHPUT, IN VIVO, LOGISTIC REGRESSION, MUTANTS (RANDOM + SINGLE SITE-DIRECTED), OFF-TARGET EFFECTS, OPTIMIZATION, PAYLOAD, PRECISION, PROTOTYPE, RNN, SEQUENCE DIVERSITY, TRANSDUCTION, VIABLE, VIRAL VECTORS

Resources:

- Company website: dynotx.com
- A paper in Science demonstrating the power of this approach with George Church, Sam Sinai, and Eric Kelsic
- A review of the approach Dyno is taking from Eric Kelsic and George Church
- Contact info if people can't find what they are looking for on the website: jobs@dynotx.com for applicants, info@dynotx for general inquiries

Vineeta Agarwala

(Andreessen Horowitz)

Terms: API, BIOMARKER, BIOPSY, DEEP CLINICAL ANNOTATION, EHR/EMR, GENETIC VARIANT, IMMUNOTHERAPY, NEXT GENERATION SEQUENCING (NGS), PAYERS, PROVIDER SYSTEMS, SPARSE, TAM (TOTAL ADDRESSABLE MARKET), TARGETED THERAPY, TIME-COURSE DATA, UNSTRUCTURED DATA

Resources:

- Primer on RWD/RWE from FDA
- Aggregating clinical and genomic real-world data (Health Affairs article describing Flatiron/FMI database)
- Technology-enabled chart abstraction and model-assisted cohort selection (Flatiron blog article)
- Can real-world data be used to replicate clinical trials? (JAMA article)

Vyas Ramanan

(Third Rock Ventures/Maze Therapeutics)

Terms: ALLELE, CASE-CONTROL, CONTINUOUS VARIABLES, COPY NUMBER, EXPRESSIVITY, FUNCTIONAL GENOMICS, GENOME-WIDE, GWAS, ISOFORM, LONGITUDINAL DATA, MONOGENIC, MULTI-PARAMETER PHENOTYPE, PENETRANCE, PROXIMITY-BASED INTERACTIONS, SEQUENCE KERNEL ASSOCIATION TEST, SPLICING STRUCTURE, SURVIVAL ANALYSES, TRANS-ACTING GENETIC MODIFIERS

Resources:

- Company websites
    - mazetx.com
    - thirdrockventures.com
- Useful blogs
    - PlengeGen
    - Life Sci VC
- Papers
    - From Maze founders identifying new protective variants in different diseases
        - Paper 1
        - Paper 2
    - A nice recent review on human disease genetics including Maze human genetics founders

Zainab Doctor

(nference)

Terms: ACCURACY, ASSOCIATION, BIOBANK, CLAIMS DATA, CO-OCCURRENCE, COMPOSITE QUERIES, EHR/EMR, FDA, HEURISTIC, ICD CODES, INCLUSION/EXCLUSION CRITERIA, NATURAL LANGUAGE PROCESSING (NLP), PRECISION, PUBMED PUBLICATIONS, RNA SEQUENCING, SEC FILINGS, TEXT CORPUS, WORD EMBEDDINGS-BASED MODEL

Resources:

- Company website: nference.ai

Terminology Definitions

ACCURACY: how close a measured value is to the true value
ACTIVE LEARNING: a special case of machine learning in which a learning algorithm can interactively query a user (or some other information source) to label new data points with the desired outputs; especially useful when unlabeled data is abundant and manual labeling is expensive
ACTION SPACE: the set of actions available to an agent in machine learning
ACTIVITY-BASED DIAGNOSTICS: diagnostics that leverage enzymatic activity to measure or produce biomarkers of disease
ADENO-ASSOCIATED VIRUS (AAV): a small virus that infects humans without causing disease; these viruses have commonly been used as vectors to deliver genetic material into cells for gene therapy, as they persist in cells without integrating into the host genome
ALLELE: one of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome
ALTERNATING DECISION TREES: a machine learning method for classification that consists of an alternation of decision nodes and prediction nodes. An instance is classified by following all paths for which all decision nodes are true
AMINO ACID: an organic compound that comprises the structural units that make up proteins
ANTIRETROVIRAL DRUGS: drugs used to prevent retroviruses such as HIV from replicating
APICAL PROGRAMMING INTERFACE (API): a set of functions and procedures allowing the creation of applications that access the features or data of an operating system or application
ACTIVE PHARMACEUTICAL INGREDIENT (API): any substance or mixture of substances intended to be used in the manufacture of a drug product and that, when used in the production of a drug, becomes an active ingredient in the drug product
ANTIBODY: a blood protein produced in response to and counteracting a specific antigen. Antibodies combine chemically with substances which the body recognizes as foreign
ASSAY: an investigative procedure for qualitatively assessing or measuring the presence, amount, or functional activity of a target entity
ASSOCIATION: any relationship between two measured quantities that renders them statistically dependent
ASYMPTOMATIC: (of a condition or a person) producing or showing no symptoms
ATRIAL FIBRILLATION: an irregular and often rapid heart rate that can increase risk of strokes, heart failure and other heart-related complications
ATTENTION MODELING: input processing technique for neural networks that allows the network to focus on specific aspects of a complex input, one at a time until the entire dataset is categorized
AUC (Area Under the ROC Curve): the expectation that a uniformly drawn random positive is ranked before a uniformly drawn random negative; performance metric for a binary classifier that plots the False Positive Rate vs the True Positive Rate
AUGMENTED REALITY: an interactive experience of a real-world environment where the objects that reside in the real world are enhanced by computer-generated perceptual information
BI-RADS (Breast Imaging Reporting and Database System) SCORE: a comprehensive guide providing standardized breast imaging terminology, report organization, assessment structure and a classification system for mammogram results
BIAS: a phenomenon that occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process
BILINGUAL: fluent in both the languages of biology and computation
BIOBANK: a large collection of biological or medical data and tissue samples, amassed for research purposes
BIOLOGIC: a product that is produced from living organisms or contain components of living organisms
BIOMARKER: a measurable substance in an organism whose presence is indicative of some phenomenon such as disease, infection, or environmental exposure
BIOPSY: an examination of tissue removed from a living body to discover the presence, cause, or extent of a disease
BLOOD-BRAIN BARRIER: a filtering mechanism of the capillaries that carry blood to the brain and spinal cord tissue, blocking the passage of certain substances
BREAKTHROUGH DESIGNATION: a process the FDA uses that is designed to expedite the development and review of drugs that are intended to treat a serious condition and preliminary clinical evidence indicates that the drug may demonstrate substantial improvement over available therapy on a clinically significant endpoint(s)
BREAST DENSITY: reflects the amount of fibrous and glandular tissue in one's breasts compared with the amount of fatty tissue in the breasts, as seen on a mammogram

BUILT ENVIRONMENT: man-made structures, features, and facilities viewed collectively as an environment in which people live and work
CAPSID FITNESS LANDSCAPE: how well suited a viral capsid is for a given task, along the related axes of importance
CAPSID PROTEIN: the protein shell of a virus, enclosing its genetic material
CARDIOLOGY: the branch of medicine that deals with diseases and abnormalities of the heart
CASE-CONTROL: a study that compares patients who have a disease or outcome of interest (cases) with patients who do not have the disease or outcome (controls), and looks back retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the outcome
CATEGORICAL VARIABLE: a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group on the basis of some qualitative property
CAUSAL INFERENCE: the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect
CAUSALITY: the relationship between cause and effect
CELL DIFFERENTIATION: the process in which a cell changes from one cell type to another, usually, to a more specialized type
CELL ENGINEERING: applies the principles and methods of engineering to the problems of cell and molecular biology, often involving modifying cells
CELL LINES: a population of cells from a multicellular organism which, due to mutation, can keep undergoing division and be grown for prolonged periods
CELL STATE: the quantifiable characteristics of a cell, which include gene expression, protein abundances, post-translational modifications, and cell morphology (a combination of cell type and what's happening to the cell right now)
CELL TYPE: a classification used to distinguish between morphologically or phenotypically distinct cell forms within a species; specialization of cells into types is an example of division of labor
CELLULAR PROGRAM: networks of genes acting in a coordinated manner to carry out functions for a cell
CENTROID: the center of mass of a geometric object of uniform density
CIS-REGULATORY ELEMENT: a noncoding DNA sequence in or near a gene required for proper spatiotemporal expression of that gene, often containing binding sites for transcription factors. Often used interchangeably with enhancer
CLAIMS DATA: consists of the billing codes that physicians, pharmacies, hospitals, and other health care providers submit to payers (e.g., insurance companies, Medicare)
CLASS OF MUTATIONS: genetic variants that share certain molecular properties and mechanisms
CLASSIFICATION: a type of supervised machine learning that specifies the class to which data elements belong; best used when the output has finite and discrete values
CLASSIFIER: any machine learning algorithm that sorts unlabeled data into labeled classes, or categories of information
CLINICAL ENDPOINTS: outcome measures referring to occurrence of disease, symptom, sign or laboratory abnormality constituting a target outcome in clinical research trials
CLUSTERS: groups of objects/samples such that members of one group are more similar to each other than members of other groups; clustering analysis is common for exploratory ML analysis
COAGULOPATHY: a condition in which the blood's ability to coagulate (form clots) is impaired
COMMON VARIATION: genetic variants that are present in >5% of the population. Some of those variants lead to susceptibility to complex polygenic diseases. Each variant at each gene influencing a complex disease may have a small additive or multiplicative effect on the disease phenotype
COMPUTATIONAL SOCIAL SCIENCE: the academic sub-disciplines concerned with computational approaches to the social sciences
CONSUMER TECHNOLOGY: any technology that is directly purchased by individuals to meet their needs, rather than through another organization
CONVOLUTIONAL NEURAL NET (CNN): a class of fully connected neural networks that have been most commonly applied to analyze visual image data
CO-OCCURRENCE: two or more variables occurring together or simultaneously
COMBINATION THERAPIES: using multiple therapies to treat a single disease
COMPANION DIAGNOSTIC: diagnostics that are co-developed with drugs to aid in selecting or excluding patient groups for treatment with that particular drug on the basis of their biological characteristics that determine responders and non-responders to the therapy
COMPOSITE BIOMARKERS: the combination of multiple metrics and/or biomarkers that provide insight to disease diagnosis and management
COMPOSITE QUERIES: queries that enable one to combine data from existing queries and then apply filters and aggregate
COMPUTER VISION: a field of artificial intelligence that trains computers to interpret and understand the visual world using digital images and deep learning models

CONFIDENCE INTERVAL: an estimate that proposes a range of plausible values for an unknown parameter (for example, the mean); the interval has an associated confidence level that the true parameter is in the proposed range
CONFOUNDERS: variables that influence both the dependent variable and independent variable, causing a spurious association
CONSTRAINT: a boundary that encourages the learner to emerge with certain behaviors
CONTEXTUAL LANGUAGE MODELS: a probability distribution over a sequence of words that provides context to distinguish between words and phrases that sound similar
CONTINUOUS PHENOTYPE: a phenotype whose distribution varies along a continuum
CONTINUOUS VARIABLES: numeric variables that have an infinite number of values between any two values
CONTRACT RESEARCH ORGANIZATION (CRO): a service organization that provides support to the pharmaceutical and biotechnology industries in the form of outsourced pharmaceutical research services
CONTROLLED ENVIRONMENT: an environment that accounts for extraneous variables that may confound findings
COPY NUMBER: the number of copies of a particular gene
CORRELATION: a mutual relationship or connection between two or more variables
COUNTERFACTUAL: a potential outcome is the outcome that would be realized if the individual received a specific value of the treatment; for each individual, one can generally observe only one, but not both, of the two potential outcomes--the unobserved outcome is the counterfactual
CYTOTOXICITY: the quality of being toxic to cells
DATA GOVERNANCE: a system of decision rights and accountabilities for information-related processes
DATA LIQUIDITY: the ability of data to be entered once and then used downstream by other systems or users
DEEP CLINICAL ANNOTATION: tracking longitudinal information on patients that range from the molecular to the clinical levels
DEEP LEARNING: a subset of machine learning that has networks capable of learning unsupervised from data that is unstructured or unlabeled
DEFIBRILLATOR: a device that restore a normal heartbeat by sending an electric pulse or shock to the heart
DELIVERY MECHANISM: the method by which a drug is administered to a patient
DENSENET (Dense Convolutional Network): an extension of ResNet (see below) that connects each layer to every other layer in a feed-forward fashion; for each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers; original paper here
DEVOPS: a set of practices that combines software development (Dev) and IT operations (Ops). It aims to shorten the systems development life cycle and provide continuous delivery with high software quality
DIABETIC RETINOPATHY: a diabetes complication that affects eyes and is caused by damage to the blood vessels of the light-sensitive tissue at the retina
DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS (DSM): an authoritative volume that defines and classifies mental disorders in order to improve diagnoses, treatment, and research; the product of >10 years of effort by hundreds of international efforts in all aspects of mental health
DIFFERENTIATION: the process by which cells specialize into specific types (i.e. from a pluripotent state)
DIGITAL PRESENCE: the collective existence of a company or individual that can be found online via an online search
DIGITAL THERAPEUTICS: evidence-based therapeutic interventions driven by high quality software programs to prevent, manage, or treat a medical disorder or disease
DIRECT-TO-CONSUMER: selling products directly to customers, bypassing any third-party retailers, wholesalers, or any other middlemen
DISTRIBUTED SYSTEMS: systems whose components are located on different networked computers, which communicate and coordinate their actions by passing messages to one another
DOSE-RESPONSE: the magnitude of the response of an organism as a function of exposure or dose to a stimulus or stressor (typically a chemical) after a certain exposure time
DROPOUT: ignoring units (i.e. neurons) at random during the training phase for a neural network
EDIT DISTANCE: a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other
EFFECT SIZE: in statistics, an effect size is a number measuring the strength of the relationship between two variables in a statistical population, or a sample-based estimate of that quantity
EFFICACY: the maximum response achievable from a pharmaceutical drug in research settings, and the capacity for sufficient therapeutic effect or beneficial change in clinical settings
Electronic Health/Medical Record (EHR/EMR): a digital version of a patient's medical and treatment history
ELECTROPHYSIOLOGY: a test performed to assess the heart's electrical system or activity and is used to diagnose abnormal heartbeats or arrhythmia
ENDOCRINOLOGIST: a physician who specializes in glands, hormones, and metabolism

ENRICHMENT: overrepresentation of an element or set of elements in a list
ENSEMBLE LEARNING: the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem
ENZYME: a substance produced by a living organism which acts as a catalyst to bring about a specific biochemical reaction
EPIDEMIC: a widespread occurrence of an infectious disease in a community at a particular time
EXPLAINABILITY: how interpretable machine learning models are to humans; how well humans can explain why a model reaches the conclusions it does
EXPLORATION-EXPLOITATION TRADE-OFF: a dilemma for a decision-making system with incomplete knowledge that must decide whether to repeat decisions that have worked well so far (exploit) or to make novel decisions, hoping to gain even greater rewards (explore)
EXPRESSION PROFILE: the landscape of activity (expression) of thousands of genes, creating a global picture of cellular function
EXPRESSIVITY: the degree to which a phenotype is expressed by individuals having a particular genotype
FAIRNESS METRICS: a set of metrics used to assess whether a machine learning model is fair (wiki on a sample list can be found here)
FALSE POSITIVE RATE: the probability of falsely rejecting the null hypothesis for a particular test
Federal Drug Administration (FDA): an agency within the U.S. Department of Health and Human Services that protects the public health by assuring the safety, effectiveness, and security of human and veterinary drugs, vaccines and other biological products for human use, and medical devices.
FEASIBILITY TEST: an assessment of the practicality of a proposed project or system
FEATURE: an individual measurable property or characteristic of a phenomenon being observed
FEATURE IMPORTANCE: the weight a particular feature is given when a machine learning model makes a prediction
FEATURE VECTOR: an n-dimensional vector of features that represent the input data
FINE PARTICULATE MATTER: tiny particles or droplets in the air that are two and one half microns or less in width; often contribute to environmental pollution
FLUORESCENCE ACTIVATED CELL SORTING (FACS): a specialized type of flow cytometry that provides a method for sorting a heterogeneous mixture of biological cells into two or more containers, one cell at a time, based upon the specific light scattering and fluorescent characteristics of each cell
FLUORESCENCE *IN SITU* HYBRIDIZATION (FISH): a molecular cytogenetic technique that uses fluorescent probes that bind to only those parts of a nucleic acid sequence with a high degree of sequence complementarity
FOODBORNE ILLNESSES: illness caused by consuming contaminated foods or beverages
FUNCTIONAL GENOMICS: a field of molecular biology that attempts to describe gene (and protein) functions and interactions; make use of the vast data generated by genomic and transcriptomic projects
GENE EXPRESSION PROFILE: in the field of molecular biology, gene expression profiling is the measurement of the activity (the expression) of thousands of genes at once, to create a global picture of cellular function
GENE THERAPY: an experimental technique that uses genes to treat or prevent disease; allows treating certain disorders by inserting a gene into a patient's cells instead of using drugs or surgery
GENES + ENVIRONMENT = PHENOTYPE: the interplay between two major drivers of each individual's phenotype
GENETIC INTERVENTION: interventions that include transgenic engineering, gene editing, and other forms of genome modification aimed at altering the information in the genetic code
GENETIC INTERACTIONS: the phenomenon where the effects of one gene are modified by one or several other genes
GENETIC KNOCKOUT: a genetic technique in which one of an organism's genes is made inoperative
GENETIC SEQUENCE: the order of DNA bases in a given region of the genome
GENETIC VARIANTS: the differences that make each person's genome unique; can be single nucleotide or longer stretches of genetic differences
GENETIC VARIATION: the variation in the DNA sequence in each of our genomes. Mutation is the ultimate source of genetic variation, but mechanisms such as sexual reproduction and genetic drift contribute to it as well
GENOME BUILD: a release of assemblies made from contiguous sequences (contigs) assembled in what is thought to be chromosomal order. This allows researchers to have a common scaffold on which they can annotate genomic features in chromosomal coordinates
GENOME-WIDE: across all ~20k genes in the human genome
GRAPH NEURAL NETWORKS: connectionist models that capture the dependence of graphs via message passing between the nodes of graphs; can represent information from its neighborhood with arbitrary depth
GROUND TRUTH: the ideal expected result; objective information provided by direct observation (i.e. empirical evidence) as opposed to information provided by inference
GROWTH HORMONE DEFICIENCY: a rare disorder characterized as the secretion of growth hormone from the anterior pituitary gland

Genome-Wide Association Study (GWAS): an approach used in genetics research to associate specific genetic variations with particular diseases
HEMORRHAGE: an escape of blood from a ruptured blood vessel, especially when profuse
HEURISTIC: any approach to problem solving or self-discovery that employs a practical method that is not guaranteed to be optimal, perfect, or rational, but is nevertheless sufficient for reaching an immediate, short-term conclusion
HIERARCHICAL MODEL: a model in which lower levels are sorted under a hierarchy of successively higher-level units; data are grouped into clusters at one or more levels
HIGH DIMENSIONAL DATA: data where the number of dimensions are staggeringly high — so high that calculations can become extremely difficult. With high dimensional data, the number of features can exceed the number of observations
HIGH THROUGHPUT SCREEN: the use of automated equipment to rapidly test thousands to millions of samples for biological activity at the model organism, cellular, pathway, or molecular level
HLA (HUMAN LEUKOCYTE ANTIGEN) SYSTEM: an important part of the immune system that encodes cell surface molecules specialized to present antigenic peptides to the T-cell receptor (TCR) on T cells
HUMAN CELL ATLAS: an international consortium whose mission is to create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease
HUMAN COMPUTER INTERACTION: studies the design and use of computer technology, focused on the interfaces between people (users) and computers
HUMAN GENOME PROJECT: the international research effort to determine the DNA sequence of the entire human genome, completed in 2003
HYBRIDIZATION WITH PROBES: a hybridization probe is a fragment of DNA or RNA of variable length (usually 100–10000 bases long) which can be radioactively or fluorescently labeled. It can then be used in DNA or RNA samples to detect the presence of nucleotide substances (the RNA target) that are complementary to the sequence in the probe under correct salt and temperature conditions
HYPERPARAMETERS: a parameter whose value is used to control the learning process; by contrast, the values of other parameters (typically node weights) are derived via training
HYPOTHYROIDISM: a condition in which the thyroid gland is underactive
ICD (International Classification of Diseases) BILLING CODES: the common system of codes used in billing work; each diagnosis code provides a general description of the disease or injury that led to the patient/physician encounter
IDIOPATHIC: relating to or denoting any disease or condition which arises spontaneously or for which the cause is unknown
IHC (Immunohistochemistry): an application of immunostaining, which involves selectively identifying antigens (proteins) in cells of a tissue using antibodies
IMAGENET: an image database organized according to the WordNet hierarchy, in which each node of the hierarchy is depicted by hundreds and thousands of images
IMMUNOTHERAPY: a type of cancer treatment that helps your immune system fight cancer
IMPUTATION: the process of replacing missing data with substituted values
IN VIVO: performed or taking place in a living organism
INCIDENCE: the occurrence of new cases of disease or injury in a population over a specified period of time
INCLUSION/EXCLUSION CRITERIA: characteristics that the prospective subjects must have if they are to be included in the study/characteristics that disqualify prospective subjects from inclusion in the study
INDICATION: the use of a drug for a particular disease
INFERENCE: a conclusion reached on the basis of evidence and reasoning from experimental data
INFORMATICS FOR INTEGRATING BIOLOGY AND THE BEDSIDE (I2B2): a member-driven non-profit foundation that aims to enable effective collaboration for precision medicine, through the sharing, integration, standardization and analysis of heterogeneous data from healthcare and research; through engagement and mobilization of a life sciences focused open-source, open-data community
INFORMATION INSIGHTS COMPANY: a company that uses its data and data science capabilities to enable insights for its customers
INFORMATION THEORY: a field that studies the quantification, storage, and communication of information
INTER-RATER RELIABILITY: the degree of agreement among raters; a score of how much homogeneity or consensus exists in the ratings given by various judges
INTERPRETABILITY: how understandable a machine learning algorithm is to humans
INTERVENTION: a manipulation of the subject or subject's environment for the purpose of modifying one or more health-related biomedical or behavioral processes and/or endpoints
INVERSE PROBABILITY WEIGHTING: a technique for calculating statistics standardized to a pseudo-population different from that in which the data was collected

ISOFORM: any of two or more functionally similar proteins that have a similar but not identical amino acid sequence and are either encoded by different genes or by RNA transcripts from the same gene which have had different exons removed
LATE STAGE ASSETS: pharmaceutical products that are in a late stage of clinical development
LATENT REPRESENTATION: latent dimensions/latent variables are variables which we do not directly observe, but which we assume to exist (in at least some instrumental sense) in order to explain patterns of variation in observed or manifest variables
LEAD OPTIMIZATION: the process by which a drug candidate is designed after an initial lead compound is identified
LESION: a region in an organ or tissue which has suffered damage through injury or disease, such as a wound, ulcer, abscess, or tumor
LINEAR MODELS: models that are specified as a linear combination of features
LOGISTIC REGRESSION: a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome, typically binary
LONGITUDINAL COHORT: a study on a group of people who share a defining characteristic and upon whom cross-sectional observational are performed at intervals over time
LONGITUDINAL DATA: data that tracks samples over time
LOSS FUNCTION: a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event; the goal is the minimize this cost
MACULA: the central area of the retina
MAMMOGRAPHY: a technique using X-rays to diagnose and locate tumors of the breasts
MAPREDUCE: a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster
MARGIN OF SAFETY: The difference between the usual effective dose and the dose that causes severe or life-threatening side effects
MARGINAL STRUCTURAL MODELS: a class of statistical models used for causal inference in epidemiology; they handle the issue of time-dependent confounding in evaluation of the efficacy of interventions by inverse probability weighting for receipt of treatment
MARKET AUTHORIZATION: all approvals from the relevant Regulatory Authority necessary to market and sell a Licensed Product in a country
MARKET CONDITIONS: the factors that influence a given market
MASSIVELY PARALLEL: the term for using a large number of processes (computational or experimental) to simultaneously perform a set of coordinated computations in parallel
MEASURE-MODEL-PERTURB: an iterative loop paradigm of computational biology that involves measuring signal in an experiment, modeling the phenomenon computationally, and perturbing the biological system based on computational predictions
MEDICAL DEVICE: an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory
MEDICAL EVIDENCE GENERATION: obtaining medical proof of concept for the clinical efficacy of a given product
MEDICARE: USA's health insurance program for people age 65 or older
METASTATIC CANCER: cancer that has spread to a different part of the body from where it started
MICROANEURYSMS: localised capillary dilatations which are usually saccular; the earliest clinically visible changes of diabetic retinopathy
MINIMUM VIABLE PRODUCT (MVP): a version of a product with just enough features to satisfy early customers and provide feedback for future product development
MISINFORMATION: false or inaccurate information, especially that which is deliberately intended to deceive
MISSING DATA: when no data value is stored for certain variables in an observation
MODEL ARCHITECTURE: the design and structure of a machine learning algorithm
MODEL GENERALIZABILITY: how well the model explains different data sets that were all generated from the same underlying process
MODEL SYSTEMS/ORGANISMS: cell- and/or animal-based systems that aim to reflect the underlying biology that is being targeted
MODEL TRAINING: determining good values for all the weights and the bias from labeled examples
MOLECULAR PROFILE: the genetic makeup that characterizes a cell or tumor
MONOGENIC: traits that are determined by a single gene
MORPHOLOGICAL PROFILING: quantitative data are extracted from microscopy images of cells to identify biologically relevant similarities and differences among samples based on these profiles
MORTALITY RATE: a measure of the number of deaths in a particular population, scaled to the size of that population, per unit of time

MULTI-CHANNEL FLUORESCENCE MICROSCOPY: optical microscopy that uses fluorescence to study the properties of organic or inorganic substances; channels enable tracking different colors
MULTI-DRUG RESISTANT BACTERIA: bacteria that are resistant to numerous antibiotics
MULTI-PARAMETER PHENOTYPING: broad and quantitative molecular and physiological measurements of cellular responses to compound treatment are used to provide information on compound activity and target mechanisms
MULTI-SITE TIME SERIES STUDY: a clinical study that tracks individuals at multiple sites (around the country or world) over time
MULTI-TASK LEARNING: a subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks
MULTIPARAMETER OPTIMIZATION: approaches to simultaneously optimizing many factors in a design
MULTIPLE IMPUTATION: an iterative form of stochastic imputation, where the distribution of the observed data is used to estimate multiple values that reflect the uncertainty around the true value
MUTANTS (RANDOM + SINGLE SITE-DIRECTED): an alteration in a genetic sequence that results from either spontaneous changes or perturbations
N: the number of samples/data points in an experiment
NAIVE BAYES CLASSIFIERS: a family of probabilistic classifiers and Bayesian network models with strong independence assumptions between the features;
NANOPARTICLE: a particle of matter that is between 1 and 100 nanometres (nm) in diameter
NATURAL LANGUAGE PROCESSING (NLP): a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data
NEWBORN SCREENING PROGRAMS: public health programs that require testing all newborns to identify conditions that can affect a child's long-term health
NEXT GENERATION SEQUENCING (NGS): a massively-parallel DNA sequencing technology that enables rapid identification of the sequence of an entire human genome
NOISY DATA: data with a large amount of additional meaningless information in it called noise (in biological experiments, can have technical noise, biological noise, etc)
OBSERVATIONAL DATA: information gathered without the subject of the research (for example an individual customer, patient, employee, etc.) having to be explicitly involved in recording what they are doing; in contrast to randomized controlled trials in clinical settings
OFF-TARGET EFFECTS: effects that can occur as a result of a drug or perturbation binding to non-targets or leading to unintended downstream consequences
OLIGONUCLEOTIDE: a short DNA or RNA molecule (oligomer) that has a wide range of applications in genetic testing, research, and forensics
ONLINE CONSENT PROCESS: agreements individuals must sign prior to enrolling in a clinical trial
OPTIMIZATION: the process of improving a program's performance characteristics such as code size (compactness) and execution speed; also used in the context of improving an algorithm's predictions to closely capture reality
ORDER OF MAGNITUDE: a class in a system of classification determined by size, each class being a number of times (usually ten) greater or smaller than the one before
ORGANOID: a miniaturized and simplified version of an organ produced in vitro in three dimensions that shows realistic micro-anatomy
OVERFITTING: a model that models the training data too well; happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data
PANDEMIC: (of a disease) prevalent over a whole country or the world
PARAMETER SPACE: the space of possible parameter values that define a particular mathematical model, often a subset of finite-dimensional Euclidean space
PASSENGER MUTATIONS: mutations that do not directly drive cancer initiation and progression, as opposed to driver mutations, such as mutations in oncogenes
PATHOGEN: a bacterium, virus, or other microorganism that can cause disease
PATIENT ENGAGEMENT: encouraging patients in their own care to help improve health outcomes, drive better patient care, and achieve lower costs
PATIENT POPULATION: the intended population a new biomedical innovation intends to help
PATTERN RECOGNITION: automated recognition of patterns and regularities in data
PAYERS: a person or organization that gives someone money that is due for work done, goods received, or a debt incurred (typically, insurance companies)
PAYLOAD: the DNA or RNA being transferred (in gene therapy)

PENETRANCE: the extent to which a particular gene or set of genes is expressed in the phenotypes of individuals carrying it, measured by the proportion of carriers showing the characteristic phenotype
PEPTIDE: a compound consisting of two or more amino acids linked in a chain
PERTURB SEQ: a reverse genetics approach that allows for the investigation of phenotypes at the level of the transcriptome, to elucidate gene functions in many cells, in a massively parallel fashion
PERTURBATION: an alteration of the function of a biological system by external or internal means such as environmental stimuli, drug inhibition, and gene knockdown
PHARMACEUTICAL FORMULATION: the process in which different chemical substances are combined to produce a final medicinal product
PHENOMICS: measurement of phenomes, where a phenome is the set of phenotypes (physical and biochemical traits) that can be produced by a given organism over the course of development and in response to genetic mutation and environmental influences
PHENOTYPE: the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment
PHYSIOLOGY: the way in which a living organism or bodily part functions
PITUITARY GLAND: a pea-sized gland that plays a major role in regulating vital body functions by controlling the activity of other hormone-secreting glands
POLYGENIC: a trait that is defined by many genes
POSITIVE PREDICTIVE VALUE (PPV): the probability that subjects with a positive screening test truly have the disease
PRAGMATIC TRIALS: trials that are easy to collect constant data from and easy to intervene in
PRE-TRAINING: training a machine learning model *before* it attempts a certain task
PRECISION: how close the measured values are to each other
PRECISION HEALTH: reimagines medicine to focus on predicting, preventing, and curing disease precisely
PRESCRIPTION DIGITAL THERAPEUTICS (PDTS): software designed for and intended to treat human disease
PREVALENCE: the proportion of persons in a population who have a particular disease or attribute at a specified point in time or over a specified period of time
PRIMARY CELLS: cells taken directly from living tissue (e.g. biopsy material) and established for growth in vitro
PROBABILISTIC CLASSIFIER: a classifier that is able to predict a probability distribution over a set of classes, rather than only outputting the most likely class that the input observation belongs to
PROPENSITY SCORE MATCHING: a statistical matching technique that attempts to estimate the effect of a treatment, policy, or other intervention by accounting for the covariates that predict receiving the treatment
PROSPECTIVE STUDIES: watches for outcomes, such as the development of a disease, during the study period and relates this to other factors such as suspected risk or protection factor(s); the study usually involves taking a cohort of subjects and watching them over a long period
PROTEASE: an enzyme which breaks down proteins and peptides
PROTOTYPE: a first, typical or preliminary model of something, especially a machine, from which other forms are developed or copied
PROVIDER SYSTEMS: also known as healthcare systems; organization or policies in place that are designed to plan and provide medical care for people
PROXIMITY-BASED INTERACTIONS: systems that rely on protein-protein interactions to label and map proteins of interest
PSEUDOVIRUS: a retrovirus that can integrate the envelope glycoprotein of another virus to form a virus with an exogenous viral envelope
PSYCHIATRIC CONDITIONS: mental illness diagnosed by a mental health professional that greatly disturbs one's thinking, moods, and/or behavior and seriously increases risk of disability, pain, death, or loss of freedom
PUBMED: free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics
RANDOM SAMPLING: a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population
R-SQUARED: relative measure of fit; represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model
RANDOM FOREST CLASSIFIER: ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees
RANDOMIZED CLINICAL TRIAL: a study in which the participants are divided by chance into separate groups that compare different treatments or other interventions; using chance to divide people into groups means that the groups will be similar and that the effects of the treatments they receive can be compared more fairly

READ DEPTH: in most sequencing protocols, the genome is fragmented into short sections of a few hundred base pairs. Sequencing depth (also known as read depth) describes the number of times that a given nucleotide in the genome has been read in an experiment
REGRESSION: a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables
REGULATORY ELEMENT: region of non-coding DNA which regulate the transcription of neighboring or far (in genomic distance) genes
REGULATORY SEQUENCE: a segment of a nucleic acid molecule which is capable of increasing or decreasing the expression of specific genes within an organism
REPRESENTATION THEORY: a branch of mathematics that studies abstract algebraic structures by representing their elements as linear transformations of vector spaces, and studies modules over these abstract algebraic structures
REPURPOSING: finding alternative use cases/indications for an existing drug
RESIDUE: a single unit that makes up a polymer, such as an amino acid in a polypeptide or protein
RESNET: a residual neural network is an artificial neural network of a kind that builds on constructs known from pyramidal cells in the cerebral cortex; residual neural networks do this by utilizing skip connections, or shortcuts to jump over some layers
RESPONDERS/NON-RESPONDERS/SUPER RESPONDERS: a patient who reacts positively to stimulus/a person who does not respond to stimulus**/**a patient with incurable disease who has a complete response or remission for more than one year after treatment, or someone who has stable disease for at least three years
RETINAL FUNDUS: the interior lining of the eyeball, including the retina (the light-sensitive screen), optic disc (the head of the nerve to the eye), and the macula (the small spot in the retina where vision is keenest)
RISK FACTORS: a variable associated with an increased risk of disease or infection
RMSE: square root of the variance of the residuals; indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values; lower values of RMSE indicate better fit
RNA SEQUENCING: technology-based sequencing technique which uses next-generation sequencing to reveal the presence and quantity of RNA in a biological sample at a given moment, analyzing the continuously changing cellular transcriptome
RECURRENT NEURAL NETWORK (RNN): a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence; this allows it to exhibit temporal dynamic behavior
SALIENCY MAPPING: an image that shows each pixel's unique quality; the goal of a saliency map is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze
SATELLITE DATA: images of Earth or other planets collected by imaging satellites operated by governments and businesses around the world
SCIBERT MODEL: an NLP model pre-trained on scientific text (paper here)
SEARCH SPACE: the feasible region (the set of all possible points that satisfy the problem's constraints) defining the set of all possible solutions
SEC FILING: a financial statement or other formal document submitted to the U.S. Securities and Exchange Commission; public companies are required to make these
SELECTIVITY AND PK PROFILES: how selective a drug is to its target, and its pharmacokinetic profile (time course of absorption, distribution, metabolism, and excretion)
SEMI-SUPERVISED LEARNING: an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training
SENSE-AND-RESPOND: identify a potential threat, then analyze the threat and take the appropriate actions; in the context of diagnostics, can engineer systems to sense a pathogen and release a signal that can be measured
SENSITIVITY: the proportion of people with the disease who will have a positive result (AKA the True Positive Rate)
SENSITIVITY ANALYSES: determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions
SEQUENCE DIVERSITY: the average number of nucleotide differences per site between two DNA sequences in all possible pairs in the sample population
SEQUENCE KERNEL ASSOCIATION TEST (SKAT): a supervised test for the joint effects of multiple variants in a region on a phenotype
SIGNAL PROCESSING: an electrical engineering subfield that focuses on analysing, modifying, and synthesizing signals such as sound, images, and scientific measurements
SINGLE CELL RNA-SEQ: examines the sequence information from individual cells with optimized next-generation sequencing technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment. See RNA sequencing for bulk version
SMALL INTERFERING RNA (siRNA): a class of double-stranded RNA non-coding RNA molecules, 20-25 base pairs in length, similar to miRNA, and operating within the RNA interference pathway

SMALL MOLECULE LIBRARY: used for screening for drug discovery
SOCIOECONOMIC STATUS: the social standing or class of an individual or group. It is often measured as a combination of education, income and occupation
SOLUBILITY: a property referring to the ability for a given substance, the solute, to dissolve in a solvent
SPARK: a formally defined computer programming language based on the Ada programming language, intended for the development of high integrity software
SPARSE DATA: has a relatively high percentage of the variable's cells do not contain actual data
SPATIAL CORRELATION: there is a correlation between the received average signal gain and the angle of arrival of a signal
SPECIFICITY: the extent to which a diagnostic test is specific for a particular condition, trait, etc
SPLICING: a post-transcriptional modification in which a single gene can code for multiple proteins; an important source of protein diversity
STATE MACHINE: a behavior model that consists of a finite number of states and is therefore also called finite-state machine (FSM); based on the current state and a given input the machine performs state transitions and produces outputs
STATE SPACE: euclidean space in which the variables on the axes are the state variables; the state of the system can be represented as a vector within that space
STATISTICAL CONTROL: a method of quality control which employs statistical methods to monitor and control a process. Negative controls are particular samples included in the experiment that are treated the same as all the other samples but are not expected to change due to any variable in the experiment--the proper selection and use of controls ensures that experimental results are valid and not due to unforeseen confounders
STENT: a tubular support placed temporarily inside a blood vessel, canal, or duct to aid healing or relieve an obstruction
STEREOSCOPIC PHOTOGRAPHS: the production of the illusion of depth in a photograph, movie, or other two-dimensional image by the presentation of a slightly different image to each eye, which adds the first of these cues (stereopsis); the two images are then combined in the brain to give the perception of depth
STUDENT-TEACHER MODEL: a training method involving model distillation; once a Deep CNN is trained, one can 'compress' it into a much smaller network, which does a reasonable job of approximating the original function
STUDY DESIGN: a process wherein the trial methodology and statistical analysis are organized to ensure that the null hypothesis is either accepted or rejected and the conclusions arrived at reflect the truth
SUPPORT VECTOR MACHINES: supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis
SURVIVAL ANALYSES: a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems
T-TEST: one type of inferential statistics; it is used to determine whether there is a significant difference between the means of two groups
TAM (TOTAL ADDRESSABLE MARKET): the revenue opportunity available for a product or service; TAM helps to prioritize business opportunities by serving as a quick metric of the underlying potential of a given opportunity
TARGET: anything within a living organism to which some other entity is directed and/or binds, resulting in a change in its behavior or function
TARGETED THERAPY: a cancer treatment that uses drugs to target specific genes and proteins that are involved in the growth and survival of cancer cells
TELEMEDICINE: refers to the practice of caring for patients remotely when the provider and patient are not physically present with each other
TENSORFLOW: an end-to-end open source machine learning platform where much of the latest deep learning neural network architectures are developed and stem from
TEXT CORPUS: a language resource consisting of a large and structured set of texts
THERAPEUTIC DOSE:
THERAPEUTIC MODALITY: tools physical therapists might use to help generate healing and assist with muscle reeducation
TIME-COURSE: refers to the evolution of a measurement over time
TOXICITY: the quality of being toxic or poisonous
TRAINING/VALIDATION/INDEPENDENT TEST SET: implemented to build up a model; validate the model built; test the built model on new (previously not seen) data
TRANS-ACTING GENETIC MODIFIERS: a gene that regulates another gene on a different chromosome
TRANSCRIPTION FACTOR BINDING: studying where and how transcription factors, which are master genetic regulators, bind to their targets along the genome
TRANSDUCTION: the process by which foreign DNA is introduced into a cell by a virus or viral vector
TRANSFER LEARNING: a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem

TRANSFORMER MODELS: a deep machine learning model introduced in 2017, used primarily in the field of natural language processing; designed to handle sequential data for tasks such as translation and text summarization
TRANSLATIONAL RESEARCH: a term often used interchangeably with translational medicine or translational science or bench to bedside – is an effort to build on basic scientific research to create new therapies, medical procedures, or diagnostics
T-SNE PLOTS: t-distributed Stochastic Neighbor Embedding is a dimensionality reduction technique that allows compressing a high-dimensional dataset into two dimensions for visualization
TURING TEST: a method of inquiry in artificial intelligence for determining whether or not a computer is capable of thinking like a human being
UNDIAGNOSED DISEASES NETWORK (UDN): a research study backed by the National Institutes of Health Common Fund that seeks to provide answers for patients and families affected by these mysterious conditions (undiagnosed.hms.harvard.edu)
UNITED NATIONS TECHNOLOGY INNOVATION LABS: a global organization designed to move humanity forward, faster by focusing on the use of innovative technology to solve some of humanity's most pressing needs. Each UNTIL is based on different humanitarian themes that are central to the needs of individual Lab's specific geolocation which, in turn, are aligned with the UN Mandates in Peace and Security, Human Rights and Sustainable Development.
UNMET NEED: the lack of availability and adequacy of existing treatments
UNSTRUCTURED DATA: information that either does not have a predefined data model or is not organized in a pre-defined manner
USE CASE-/PROBLEM-BACKED: a problem-solving approach that starts with the end goal in mind; a specific situation in which a product or service could potentially be used
VERSION CONTROL: a system that records changes to a file or set of files over time so that you can recall specific versions later
VIABLE: capable of working successfully; feasible
VIRAL TITER/VIRAL LOAD: a numerical expression of the quantity of virus in a given volume of fluid (e.g. blood or sputum)
VIRAL VECTORS: tools commonly used by molecular biologists to deliver genetic material into cells
Virtual Private Network (VPN): extends a private network across a public network and enables users to send and receive data across shared or public networks as if their computing devices were directly connected to the private network
WEARABLE SENSOR: integrated into wearable objects or directly with the body in order to help monitor health and/or provide clinically relevant data for care
WHOLE SLIDE IMAGES: refers to scanning of conventional glass slides in order to produce digital slides, is the most recent imaging modality being employed by pathology departments worldwide
WILCOXON TEST: a nonparametric statistical test that compares two paired groups, and comes in two versions the Rank Sum test or the Signed Rank test; the goal of the test is to determine if two or more sets of pairs are different from one another in a statistically significant manner
WORD EMBEDDINGS: a learned representation for text where words that have the same meaning have a similar representation
WORLD HEALTH ORGANIZATION (WHO): WHO's primary role is to direct international health within the United Nations' system and to lead partners in global health responses
ZERO-SHOT LEARNING: a problem setup in machine learning, where at test time, a learner observes samples from classes that were not observed during training, and needs to predict the category they belong to