Machine Learning and Artificial Intelligence (1/3)

What is Biomedical & Health Informatics?
William Hersh

Oregon Health & Science University

Machine learning and artificial intelligence

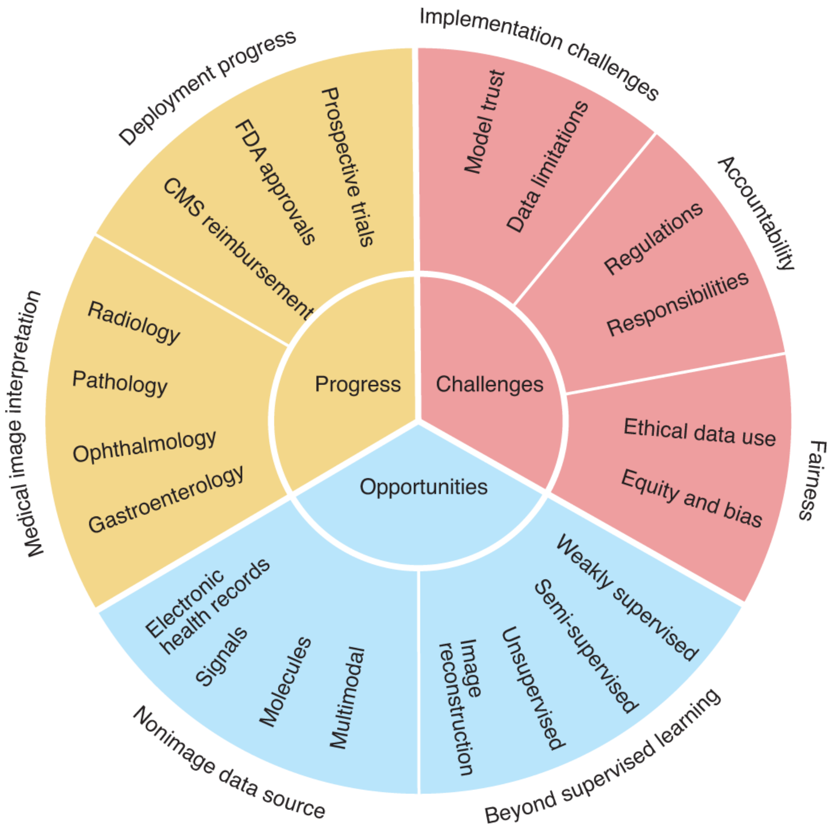
- Overview
- Methods
- Results
- Future directions

Overviews of machine learning (ML)

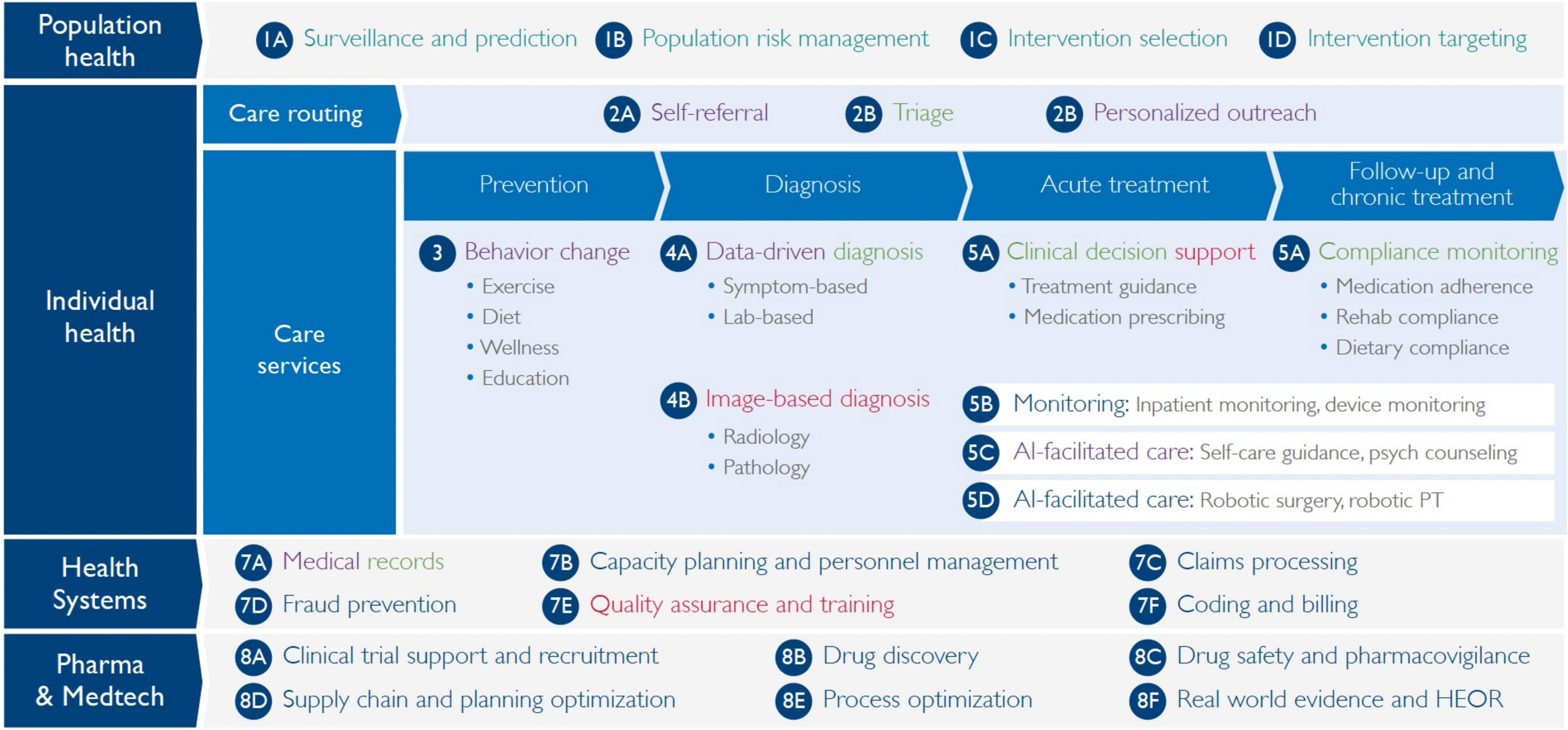
- Blogs
  - Chugh, 2018; Shin, 2020
- Monographs
  - Alpaydin, 2020
- Books
  - Scarlat, 2019; Topol, 2019
- Math important but not necessary for understanding big picture
  - Statistical learning (James, 2017)
  - Math for ML (Deisenroth, 2020)
  - Probability in machine learning (Chan, 2021; Murphy, 2022; Murphy, 2023)
  - Causal inference (Hernán, 2023)
- Course – https://www.cs197.seas.harvard.edu/

Overviews of artificial intelligence (AI)

- Overviews
  - National Academy of Medicine (Matheny, 2019)
  - Progress, challenges, and opportunities (Rajpurkar, 2022)
  - Textbook (Cohen, 2022)
- Many biomedical and health application areas
  - Global Health (USAID, 2019)
  - Automating production of systematic reviews (Marshall, 2019)
  - Medical imaging (Esteva, 2021)
  - Uses in biology (Greener, 2021)
  - Reducing ocular health disparities (Campbell, 2021)
  - Improving patient safety (Bates, 2021)
  - Use in clinical decision support (Adlung, 2021; Chen, 2022)
  - Clinical and translational research (Bernstam, 2021)
  - Healthcare (Davenport, 2022; Busnatu, 2022)
- HHS use cases inventory
  - https://www.hhs.gov/about/agencies/asa/ocio/ai/use-cases



Applications of AI (USAID, 2019)



Population health

1A Surveillance and prediction 1B Population risk management 1C Intervention selection 1D Intervention targeting

Individual health

Care routing

2A Self-referral 2B Triage 2B Personalized outreach

Care services

Prevention

3 Behavior change
- Exercise
- Diet
- Wellness
- Education

Diagnosis

4A Data-driven diagnosis
- Symptom-based
- Lab-based

4B Image-based diagnosis
- Radiology
- Pathology

Acute treatment

5A Clinical decision support
- Treatment guidance
- Medication prescribing

Follow-up and chronic treatment

5A Compliance monitoring
- Medication adherence
- Rehab compliance
- Dietary compliance

5B Monitoring: Inpatient monitoring, device monitoring

5C AI-facilitated care: Self-care guidance, psych counseling

5D AI-facilitated care: Robotic surgery, robotic PT

Health Systems

7A Medical records 7B Capacity planning and personnel management 7C Claims processing

7D Fraud prevention 7E Quality assurance and training 7F Coding and billing

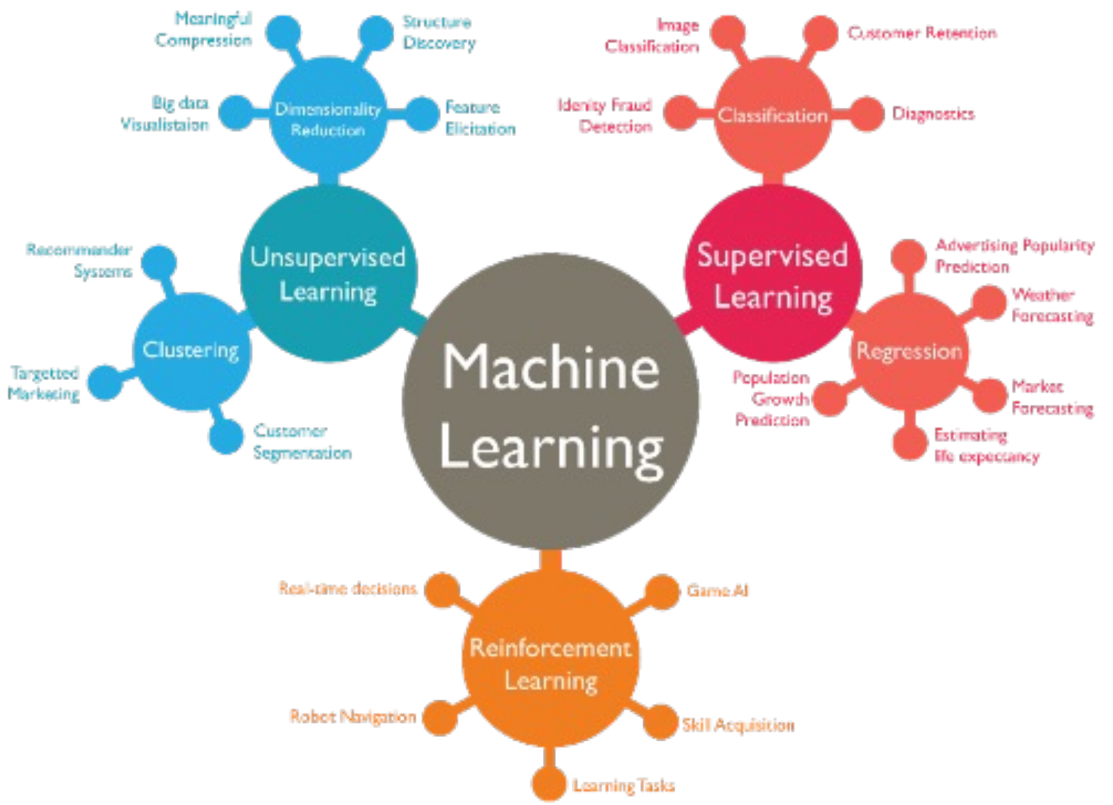
Pharma & Medtech

8A Clinical trial support and recruitment 8B Drug discovery 8C Drug safety and pharmacovigilance

8D Supply chain and planning optimization 8E Process optimization 8F Real world evidence and HEOR

Methods of ML – types of learning

- Supervised – learn to predict a known output
  - Learns from training data
  - Evaluated on test data
    - To avoid "overfitting"
- Unsupervised – find naturally occurring patterns or groupings within data
- Semi-supervised – mixture of two, with combination of labeled and unlabeled inputs
  - Algorithms find structure and patterns on their own with help from labeled inputs
- Reinforcement learning learns from ongoing data and results, e.g., from ongoing use in a clinical setting (Gottesman, 2019; Ahilan, 2023)
- Transfer learning – applying learning trained for one task to another (Yang, 2020)
  - Large foundational models for generative AI (Bommasani, 2022)



(Chugh, 2018)

Tasks of supervised learning

- Classification – predict class from one or more features of data, e.g., diagnosis or patient outcome
  - k-Nearest Neighbors (kNN) – aim to find category having “closest” number of attributes
  - Naïve Bayes – derive conditional probabilities that classify into categories
  - Support vector machines (SVMs) – for binary classification, draw “line” that separates one category from other
  - Decision trees – develop set of rules that classify into categories
- Regression – predict numerical value from data, e.g., risk of disease or poor outcome or benefit from treatment
  - Linear – fit a line to data
  - Multivariate (polynomial) – fit many variables to model
  - Logistic regression – binary output

Tasks of other types of learning

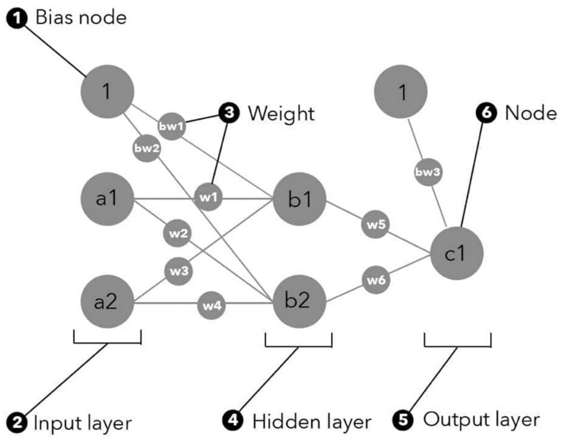
- Unsupervised learning
  - Clustering – group items together
  - Density estimation – find statistical values
  - Dimensionality reduction – reduce many to few features
- Growing use of transfer learning
  - Large language models developed for one task applied to others (Mwiti, 2022)

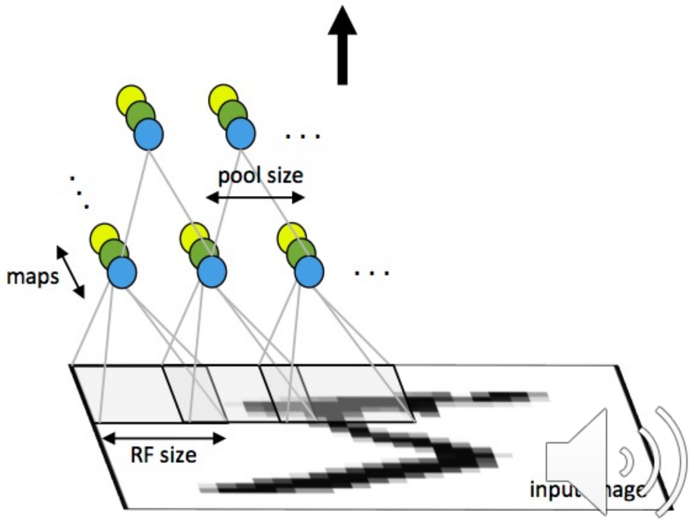
Artificial neural networks (ANNs)

- Have come to fore as main approach for ML with large amounts of data and increased modern computing power (Choi, 2020)
  - Particular success has been achieved with deep learning, with much internal complexity to networks
  - ANNs had been around for many decades (McCulloch, 1943), but deep learning successes often attributed to work of Hinton (2006)
- Mathematics complex, but can understand what they do in context of ML tasks

Anatomy and physiology of neural networks (Taylor, 2017)

- Anatomy
  - Layers
  - Nodes and weights – connected like neurons
- Physiology
  - Feedforward – processing from input to output
    - Convolutional neural networks (CNNs) particularly effective for image analysis
  - Feedback – processing loops backwards
    - Sometimes called recurrent neural networks (RNNs), most useful for sequential data, such as text





Tools for ML and AI

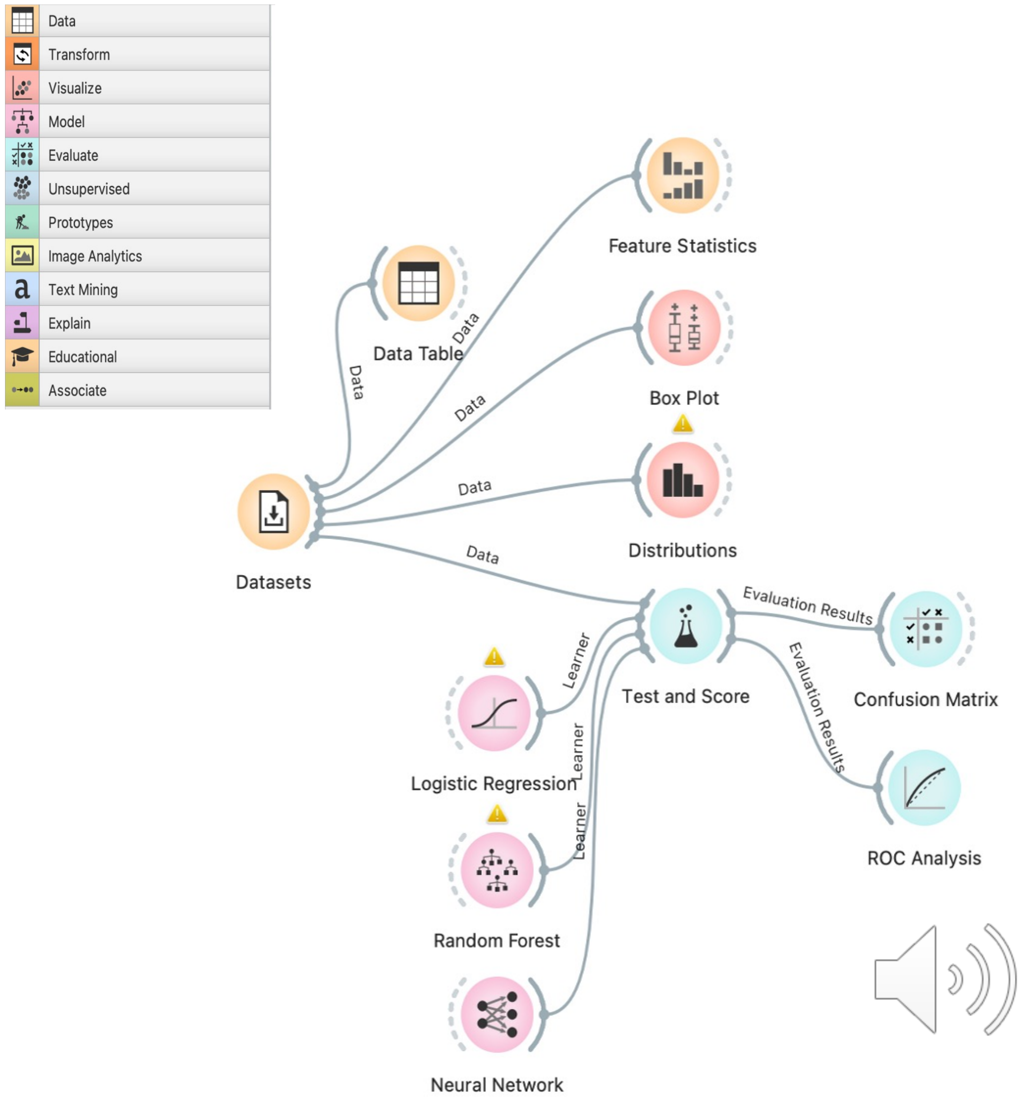
- Overview with biomedical focus (Hoyt, 2019)
- Many programming languages but 2 most widely used (both open-source)
  - R – focused on statistical computing and graphics, especially with ”tidy” data (Wickham, 2017)
  - Python – easy to use and read language has gained popularity for data science and ML (Downey, 2016)
- Jupyter notebooks – locally run Web pages that contain live code, equations, figures, interactive apps, and Markdown text (Galea, 2018)
  - Initially developed for Python but now can use other languages, including R

Tools (cont.)

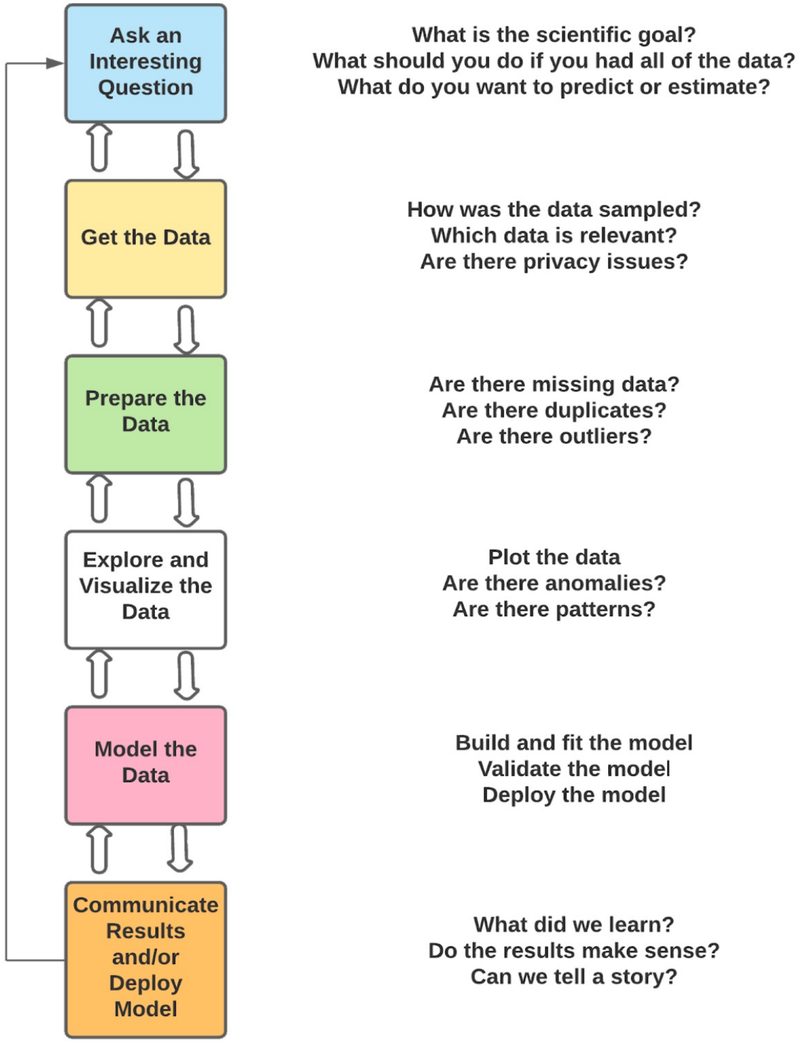
- Code libraries – several open source
  - TensorFlow – Google
    - https://www.tensorflow.org/
  - Scikit-learn – for Python
    - https://scikit-learn.org/
  - Tidyverse – libraries for analyzing (dplyr) and visualizing (ggplot) “tidy” data in R
    - https://www.tidyverse.org/
- ML data sets
  - Many (Hoyt, 2019; Altexsoft, 2022)
  - UCI ML Repository – https://archive.ics.uci.edu/ml/index.php
  - Physionet.org, including Medical Information Mart for Intensive Care (MIMIC) – https://physionet.org/ (Johnson, 2023)

No-code programming – Orange data mining

- “No-code” model development – visual programming packages
  - Orange – https://orangedatamining.com/
  - RapidMiner – https://rapidminer.com/
- Orange is open-source with large community of support (Smith, 2022; Hoyt, 2022; Hoyt, 2022)



Steps in data analysis or “wrangling” (Hoyt, 2019; Anaconda, 2022)



Ask an Interesting Question

What is the scientific goal?
What should you do if you had all of the data?
What do you want to predict or estimate?

Get the Data

How was the data sampled?
Which data is relevant?
Are there privacy issues?

Prepare the Data

Are there missing data?
Are there duplicates?
Are there outliers?

Explore and Visualize the Data

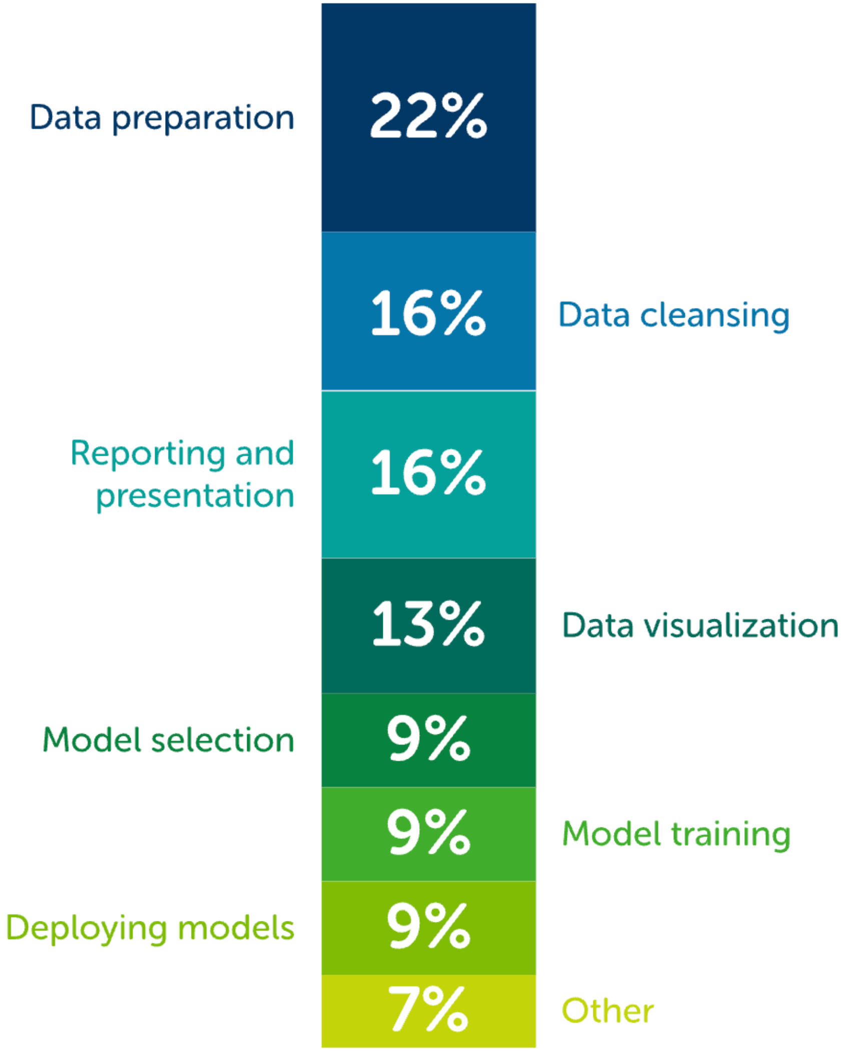
Plot the data
Are there anomalies?
Are there patterns?

Model the Data

Build and fit the model
Validate the model
Deploy the model

Communicate Results and/or Deploy Model

What did we learn?
Do the results make sense?
Can we tell a story?


Data preparation
Data cleansing
Reporting and presentation
Data visualization
Model selection
Model training
Deploying models
Other