Machine Learning and Artificial Intelligence (1/3)
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/](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)

<table>
<thead>
<tr>
<th>Population Health</th>
<th>Individual Health</th>
<th>Care Services</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surveillance and prediction</strong></td>
<td><strong>Self-referral</strong></td>
<td><strong>Behavior change</strong></td>
</tr>
<tr>
<td><strong>Population risk management</strong></td>
<td><strong>Triage</strong></td>
<td><strong>Data-driven diagnosis</strong></td>
</tr>
<tr>
<td><strong>Intervention selection</strong></td>
<td><strong>Personalized outreach</strong></td>
<td><strong>Image-based diagnosis</strong></td>
</tr>
<tr>
<td><strong>Intervention targeting</strong></td>
<td></td>
<td><strong>Clinical decision support</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Compliance monitoring</strong></td>
</tr>
</tbody>
</table>

### Health Systems

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Medical records</strong></td>
<td><strong>Capacity planning and personnel management</strong></td>
<td><strong>Claims processing</strong></td>
</tr>
<tr>
<td><strong>Fraud prevention</strong></td>
<td><strong>Quality assurance and training</strong></td>
<td><strong>Coding and billing</strong></td>
</tr>
</tbody>
</table>

### Pharma & Medtech

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical trial support and recruitment</strong></td>
<td><strong>Drug discovery</strong></td>
<td><strong>Drug safety and pharmacovigilance</strong></td>
</tr>
<tr>
<td><strong>Supply chain and planning optimization</strong></td>
<td><strong>Process optimization</strong></td>
<td><strong>Real world evidence and HEOR</strong></td>
</tr>
</tbody>
</table>
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)
  – 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/
• 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)