AI at Scale for Health

Gina Tourassi
Group Leader – Biomedical Sciences, Engineering and Computing
Director – Health Data Sciences Institute

August 20, 2019
Health: A grand challenge for our nation’s economic and national security

- US healthcare spending already comprises far more of the national GDP than any other sector including defense, education, transportation, and energy together.
- By 2030, US will experience the highest increase in GDP loss (8.1%) due to aging workforce and high burden of chronic diseases.*
- Health care technologies can be an international stabilizing force

Health: A grand data challenge

By 2025, genomic data alone:

- Human genome sequencing data doubling every 7mos
- Projected 1 exabase/year within 5 years
- Projected 100 million to 2 billion human genomes sequenced by 2025
- Data storage needs of 2 to 40 exabytes (@30xcoverage)

Volume of health data:

- Biomedical and social research information collected by scientists in academia, government, insurance agencies and industry doubles every 12-14 months.
  - * PLOS Biology. 2015; 1002195
### Example: Neuroscience

Increase of Data Volume and Complexity relative to Computational Power.

<table>
<thead>
<tr>
<th>Neuroimaging (annually)</th>
<th>Genomics (BP/Yr)</th>
<th>Moore's Law (transistor counts)</th>
<th>Bandwidth (Edholm's Law)</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Complexity</td>
<td>Size</td>
<td>Complexity</td>
<td></td>
</tr>
<tr>
<td>200 GB</td>
<td>1</td>
<td>10 MB</td>
<td>1</td>
<td>1985-1989</td>
</tr>
<tr>
<td>1 TB</td>
<td>2</td>
<td>100 MB</td>
<td>2</td>
<td>1990-1994</td>
</tr>
<tr>
<td>50 TB</td>
<td>5</td>
<td>10 GB</td>
<td>3</td>
<td>1995-1999</td>
</tr>
<tr>
<td>250 TB</td>
<td>6</td>
<td>1TB</td>
<td>4</td>
<td>2000-2004</td>
</tr>
<tr>
<td>1 PB</td>
<td>7</td>
<td>30TB</td>
<td>5</td>
<td>2005-2009</td>
</tr>
<tr>
<td>5 PB</td>
<td>8</td>
<td>1 PB</td>
<td>7</td>
<td>2010-2014</td>
</tr>
<tr>
<td>10+ PB</td>
<td>9</td>
<td>20+ PB</td>
<td>8</td>
<td>2015-2019 (estimated)</td>
</tr>
</tbody>
</table>

Complexity = measure of data heterogeneity (e.g., new imaging data acquisition modalities or sequence coverage depth; complexity of 5 indicates a 5-fold increase of the data diversity over 1985)
The role of AI in medicine is broad (but still fragmented)

Many promising examples already and many more will continue to emerge with growing computing resources, computational and algorithmic advances...
THE PRESENT: Many AI applications in disease risk prediction, diagnosis, treatment response, outcomes, e.g.,

MIT/MGH's image-based deep learning model can predict breast cancer up to five years in advance. (May 2019 - Retrospective study with ~90K mammograms),

Deep learning can detect basal cell carcinoma and differentiate malignant from benign lesions in histopathological images of the skin with diagnostic accuracy of >90% comparable with experts.
But are we targeting the right problem?
But are we targeting the right problem?

US Average Lifespan is 79 years

US Average Healthspan is 65 years

* 80% of Americans >65yrs have at least one chronic condition and 50% have two

PRESENT: Disease Care

GOAL: Wellness Care

Annual total costs of age-related diseases expected exceed $1.5 trillion in the US by 2030 (considering heart disease, Alzheimer's, and cancer alone)!

Genes and biology are responsible for 10% of our health and well being
Can we model health trajectories?

“High Definition Medicine” Cell, Volume 170, Issue 5, 24 August 2017
THE FUTURE: Combining models and machines to execute in silico longitudinal randomized clinical trials by generating lifelong health trajectories of synthetic patients.
Scientific Drivers in AI for Health

- **Learning with limited data**
  - Interesting health events can be rare, expert annotation is expensive

- **Learning with noisy labeled data (e.g., due to human error)**
  - Inter- and intra-annotator variability is a well-known challenge in medicine

- **Learning to foresee the unseen**
  - How can we predict the future, if it hasn’t happened in the past

- **Learning across tasks and domains**
  - e.g., learn one imaging task/domain and extrapolate to a new one

- **Integrating biomedical domain knowledge**
  - Data-driven AI models benefit from domain understanding

- **AI explainability/interpretability/UQ**
  - Its importance will vary across problems and will be debated for some time

- **AI needs causality**
Infrastructure and Policy Drivers in AI for Health

- **Infrastructure**
  - Handling heterogeneous data (text + images + genomics + physiology + environment + ….)
  - Supporting federated AI learning with sensitive data
  - Handling time-critical and risk-critical AI inference
  - Energy-efficient AI solutions for deployment to point-of-care

- **Policy**
  - Cybersecurity concerns for both data and algorithms
  - Benchmarking of data quality and algorithms based on domain-relevant performance metrics
### Final Thoughts on AI for Health

<table>
<thead>
<tr>
<th>Hope</th>
<th>Hype</th>
<th>Hard Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>The convergence of big data and AI will enable the accumulation and automation of functional knowledge in biomedicine.</td>
<td>AI solutions are superior to collective intelligence of the experts</td>
<td>We need a sustainable infrastructure to support the continuum of AI-driven scientific discovery and clinical application.</td>
</tr>
<tr>
<td></td>
<td>Practical translation of AI tools is straightforward</td>
<td>Human-in-the-Loop challenges will impact real-world value.</td>
</tr>
<tr>
<td></td>
<td>Correlation is enough when big data is involved</td>
<td>AI interpretability and (real-time) UQ are important future directions but true impact will depend on the application.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vulnerability issues for AI models and AI users (cognitive hacking) are critical and should be considered from the beginning.</td>
</tr>
</tbody>
</table>
Amara’s Law
We tend to overestimate the effect of a technology in the short run and underestimate its effect in the long run.
Thank you