Making biomedical data and knowledge work for precision medicine

In this era of digital biomedicine, an unprecedented amount of data is being collected.

Electronic health records (EHRs) capture observations on millions of individuals daily.
By 2025, sequencing will routinely generate 1 zettabyte of data annually.
New sources of data, like mHealth, are providing insight into behaviors in real-world environments.

Current clinical diagnostic imaging produces over 1 exabyte of data each year.

In 2014, more than 1,500 active biological databases covering genomics, proteins, pathways, etc.

We are also producing large volumes of new biomedical knowledge.

The number of publications indexed by PubMed has almost doubled since 2004.

This ability to create a comprehensive view of the individual is driving precision medicine.

UCLA has launched efforts to sequence and deep phenotype more than 100,000 patients.

NIH launched the Precision Medicine Initiative Cohort Program for 1 million individuals.

Verily Life Sciences launched a longitudinal study of 10,000 subjects.

We are also producing large volumes of new biomedical knowledge.

Bringing together all of this data and knowledge is key to enabling discovery and the full promise of precision medicine.

The 2011 Institute of Medicine (IOM) report, Toward Precision Medicine, recognized this need.
Translating insights from data science into usable knowledge and clinical practice is challenging. Integration: Combining what we have and what we know. Usability: Employing our tools to effect change. Prediction: Building a better model. Translating biological data, mHealth, EHR, -omics, clinical studies, pathways, outcomes, heart disease, Endotype discovery: Subgroups and new classifications. Predictive models: Combining data to inform decisions. New knowledge and avenues of discovery. Several computational challenges exist:
- Organizing and managing uncertainty and incomplete data in high dimensional spaces
- Analyzing heterogeneous data
- Learning with limited or no prior knowledge of the domain(s).
Building Knowledge Graphs

New scalable methods for feature extraction from unstructured data
- NLP, topic modeling, semantic analysis
- Signal and image analysis methods

Cutting-edge methods for relation discovery in heterogeneous information networks
- Meta-path and meta-structure algorithms
- Multi-view clustering and network-embedded mining

EHRs, published literature, and social media contain a wealth of untapped insight, often in the form of (unstructured) free-text.

Concept Disambiguation

Text corpus (research papers, etc.) → Concept relational network → Concept disambiguation → Concept type discovery

Improving Stroke Treatment

Construction of an observational, standardized database with UCLA Stroke Center
- Imaging feature analysis
- Bi-convolution neural network (bi-CNN) that attempts to learn the fate of affected tissue
- Three layer architecture (convolution, map stacking, fully-connected) to predict CBV, CBF, MTT, Tmax
- NLP methods to extract findings from clinical notes (e.g., presenting symptoms, past medical history, outcomes)
- Influence diagram for decision support

What relations do we already know?

How do we take advantage of the knowledge contained in published studies and clinical trials?

Which study is relevant for a given patient?
The MedicineMaps project is building a software framework for extracting relationships and context from biomedical literature.

- Provides a systematic means for weighing the strength of evidence for a relation.
- Enable discovery of which studies are relevant to a given patient based on matching characteristics (i.e., evidence-based medicine).

Machine learning methods are commonplace, and we can readily generate classifiers and models...but are they used clinically?

Predictive models rarely perform as well when applied in new clinical environments.

Diseases are complex, evolving entities, and phenotypes change over time and across subgroups.

Real-world observations are noisy and sparse. The context and provenance of the dataset and model are often incompletely described.

- New ways of sharing data that enable comparisons of models on new datasets.
- New shared predictive modeling repository that captures provenance.
- New methods for identifying portions of probabilistic models that can be "shared."
- Multi-modal biomarkers.
Diseases are complex, evolving entities, and phenotypes change over time and across subgroups.

A static model is not likely to accurately represent a disease. The behavior of a disease is often more informative than a single point in time. Models incorporating time are computationally challenging.

- New continuous time models to handle real-world data, integrating new observations.
- Exploring constraint-based approaches to optimize sequential decision-making processes.
- Methods to understand when models need to be retrained over time.

Helping patients understand

The vast majority of Americans now go online to understand their health. But do they understand their medical records and results?

RUMI (Retrieving Understandable Medical Information) provides context around a patient's process of care and his/her EHR.

- Maps the contents of a patient's record to medical problems and the process of care so that correct information is given at the right time.
- For instance, patients recently diagnosed with cancer or those who may get treatment require different information.
- Enables healthcare professionals to understand what information and questions patients have about their disease and care.

Helping patients understand
What if you could predict ahead of time, for a given individual, an asthma attack, and mitigate if not prevent it?

The Los Angeles PRISMS Center is an interdisciplinary effort to understand pediatric asthma

- U54 mHealth informatics center focused on integration platforms to support research and clinical care.
- Comprehensive view of the disease involving sensors (environmental, personal) and the EHR to elucidate individual behaviors and asthma triggers.
- Three interacting projects around secure sensor platforms, data integration and analysis, and field testing with predictive models.

There’s an old joke about pilots and the future plane cockpit...

acknowledgements