Making biomedical data and knowledge work for precision medicine

Stanford University, IBIIS Annual Retreat
Alex Bui | Thursday, September 22, 2016

UCLA medical imaging INFORMATICS

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

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

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

---

Verily Life Sciences launched a longitudinal study of 10,00 subjects

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

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

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

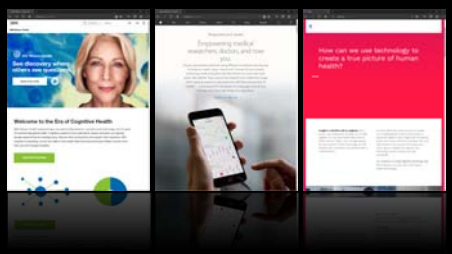
In 2014, more than 1,500 active biological databases covering omics, proteins, pathways, etc.

We are also producing large volumes of new biomedical knowledge.

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

The 2011 Institute of Medicine (IOM) report, *Toward Precision Medicine*, recognized this need

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



Translating insights from data science into usable knowledge and clinical practice is challenging.

Integration
Combining what we have and what we know

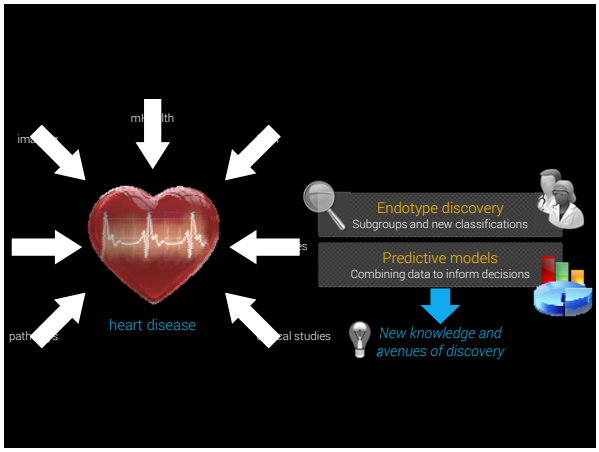
Prediction
Building a better model

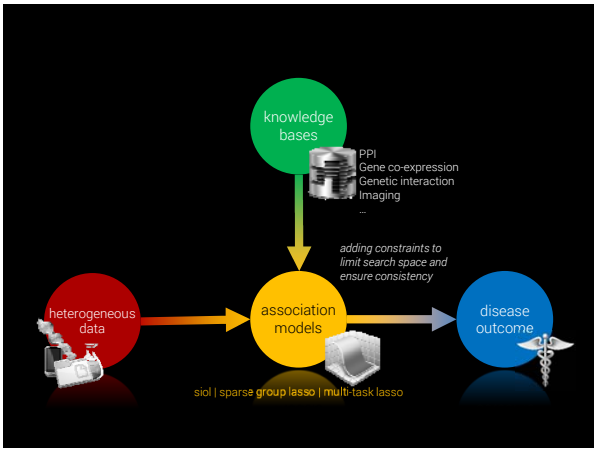
Usability
Employing our tools to effect change

Integration
Combining what we have and what we know

How do we automatically connect across biomedical data and knowledge silos?

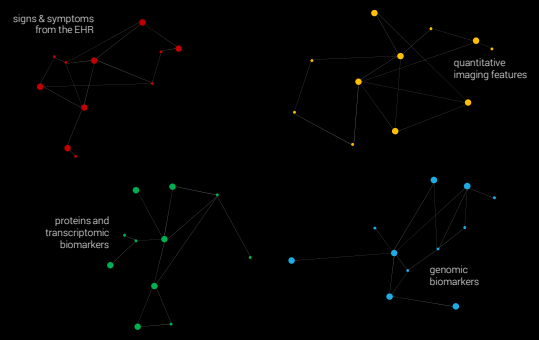
biological data ⟷ medical records ⟷ health outcomes





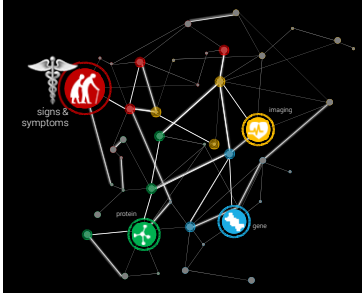
---

building knowledge graphs



---

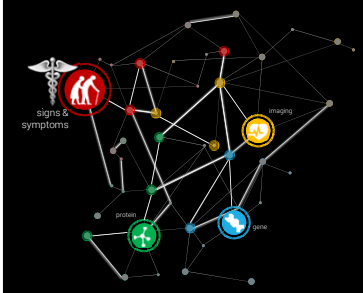
building knowledge graphs



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

Biomedical NLP continues to be an active area of research, especially around context, coreference, and disambiguation

Active learning is being applied to enhance classifier training in the NLP space, selecting high-value examples for annotation

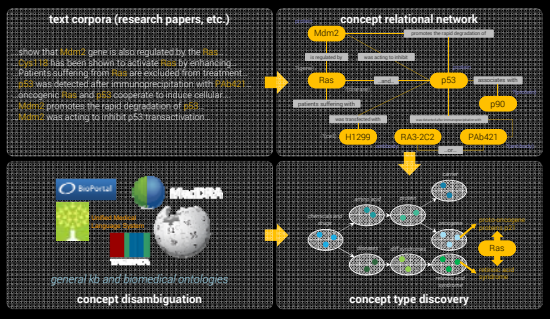
Topic models provide an *unsupervised* statistical method for discovery of common concepts within large corpora

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

RadACC: Applications include assessment of radiologist performance over time, given different types of cases

---

concept disambiguation



---

improving stroke treatment

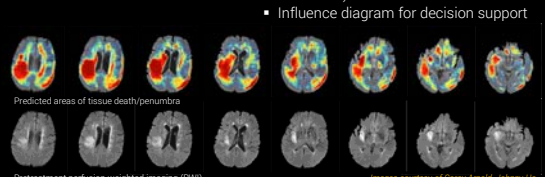
While early intervention with thrombolytic agents and clot retrieval improves outcomes in many stroke patients, it is unclear which individuals will benefit from which treatment

---

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, T_{max}
- NLP methods to extract findings from clinical notes (e.g., presenting symptoms, past medical history, outcomes)
- Influence diagram for decision support



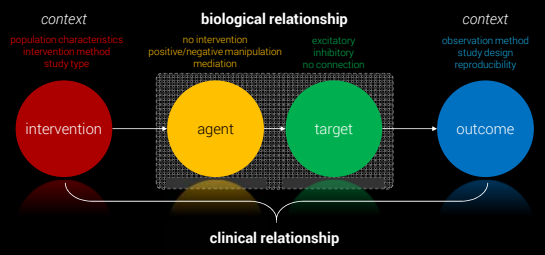
Images courtesy of Corey Arnold, Johnny Ho
A Temporal Deep Learning Approach for MR Perfusion Parameter Estimation in Stroke, ICPR 2016

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?

mapping biomedical literature



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).

Prediction

Building a better model

How do we use integrated data to inform improved predictive models for healthcare?

Prediction

Building a better model

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.

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

Real-world observations are noisy and sparse.

The context and provenance of the dataset and model are often incompletely described.

What is the reproducibility of prognostic models for brain cancer (glioblastoma multiforme) patients seen at UCLA?

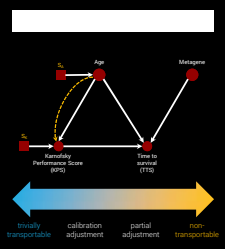
- Unclear/different semantics around predictive model variables.
- Differences in patient populations (and its impact) is not well-defined.
- Statistical methods to parameterize models are often unreproducible.

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

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.



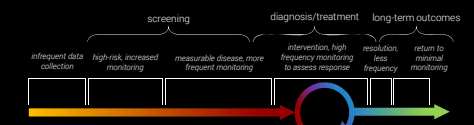
Work with William Hsu, Kyle Singleton

Motivating the additional use of external validity: Examining transportability in a model of glioblastoma multiforme. AMIA 2014

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.

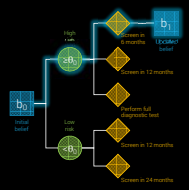


Diseases are complex, evolving entities, and phenotypes change over time and across subgroups.

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.



For each individual, can we determine the set of decisions over time that maximizes quality of life while minimizing observations?

Usability

Employing our tools to effect change

How do we use our models and tools to improve healthcare?

helping patients understand

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

helping patients understand

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

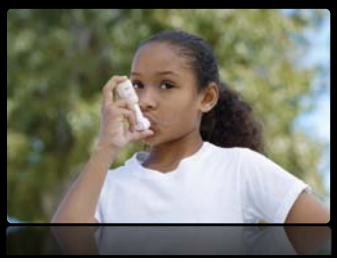
- Maps the contents of a patient's record to medical problems and the process of care so that correct information is given at the right time
  - For instance, patients recently diagnosed with cancer vs. those who are post-treatment require different information
- Enables informaticians and physicians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

UCLA USC

Los Angeles PRISMS Center

Pediatric Research with Integrated Sensor Monitoring Systems

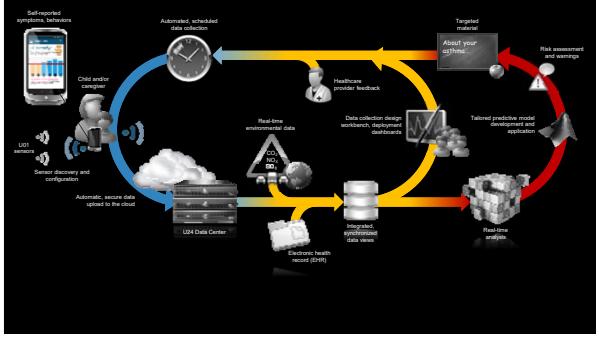
What if you could predict ahead of time, for a given individual, an asthma attack, and mitigate if not prevent it?



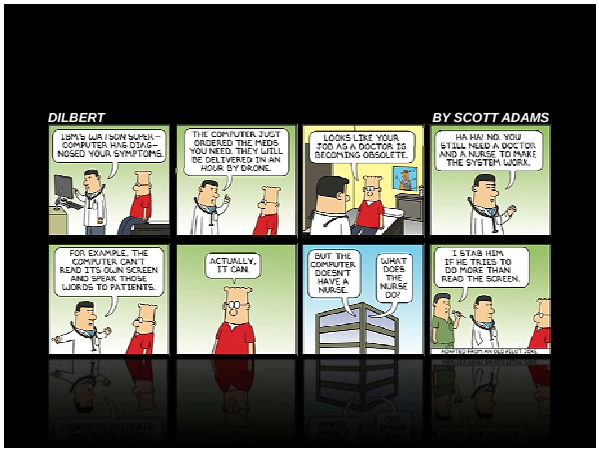
The Los Angeles PRISMS Center is an interdisciplinary effort to understand pediatric asthma

- 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.

BREATHE: Biomedical REAL-Time Health Evaluation



There's an old joke about pilots and the future plane cockpit…



acknowledgements

Faculty

Denise Aberle
Corey Arnold
Suzie El-Saden
William Hsu
Frank Meng
Craig Moriioka
James Sayre
Ricky Taira

Staff

Lew Andrada
Shawn Chen
Patrick Langdon
Denise Luna
Cleo Maharta
Isabel Rippy
Weixia Yu
Bing Zhu

Students & Postdocs

King (Johnny) Chung Ho
Edgar Rios-Piedra
Shiwen Shen
Jiayun Li
Tianran Zhang
William Speier
Karthik Sarma
Simon Han
Panayiotis Petousis
Nicholas Matiasz
Nova Smedley

Funding Sources

NIH/NINDS R01 NS076534
NIH/NLM R01 LM011333
NIH/NIBIB T32 EB016640
NIH/NCI R01 CA157553
NIH/NIBIB R01 EB00362
NIH/NIBIB U54 EB022002
NIH/NIGMS GM114833-02S1
UCOP 285247
NSF CCF-1436827

Training programs

T32 Imaging Informatics
T32 Biomedical Big Data (BD2K)
UCLA MII RISE (UC-HBCU)

Collaborative programs

Center for Domain-Specific Computing (CDSC)
Clinical and Translational Science Institute (CTSI)
BD2K Centers Coordination Center