NLP/Information Extraction from Clinical Notes - Welcome!

Atul Butte, MD, PhD
Priscilla Chan and Mark Zuckerberg Distinguished Professor
Director, Institute for Computational Health Sciences, UCSF
Exec. Director, Clinical Informatics, UC Medical System

May 18, 2018
BIG DATA Can....

- Predict the disease before it strikes
- Explain the rare disease that defies experts
- Find drugs for diseases lacking attention
- Make sure we do the right thing for patients
- Be an amazing platform for biomedical innovation
Leading in Biomedical Computation and Health Analytics

Our **Mission**

Bring the power of computational health sciences

… to enable UCSF research in the discovery of a new generation of insights

… in Precision Medicine

… in service of the patient
Combining Healthcare Data From Six UC Medical Centers
Mean Age = 48
A New Generation of Computation for Health Science

Artificial Intelligence (AI) & Machine/Deep Learning → patterns and relationships you could not find using structured data

Tap the (messy) Big Data!
Clinical Notes Contain Invaluable Data

We have 58 million of them from every clinic
Clinical Text Extraction Starting to Benefit Many Disciplines*

Representative Examples

- **Radiology**
  - “NLP in Radiology: A Systematic Review,” Pons et al
  - ”NLP Technologies in Radiology Research and Clinical Applications,” Cai et al
  - “....Mammography reports for decision support,” Bozcurt et al

- **Oncology**
  - “NLP in Oncology,” Yim et al

- **Medications, Drug Safety, Longitudinality**
  - “Advanced Drug Safety Research with Semantic Analysis....,” Hong Yu, UMass
  - “NLP.. Drug Repositioning, Adverse Events....,” Lixia Yao, Mayo Clinic
  - “NLP ... Medication.. Insights...Longitudinal Patient Records,” Murthy D., IBM

* But really, examples coming up in every condition and discipline*

* Quoted sources all at [FDA, UCSF-Stanford CERSI and SFSU Collaborative Workshop](https://www.fda.gov), June 15, 2017
We Plan to Make Extracted Information Available to all Researchers in 9-12 Months

Information Commons (under Development)
Today

- Segment 1: NLP in research practice at UCSF
  - Adams Dudley, on Predicting Mortality in ICU
  - Maryam Panahiazar, on Large Scale Analysis of Clinical Records
  - Milena Gianfrancesco & Suzanne Tamang, Text Mining for Clinical Infections
  - Dean Schillinger, on Patient Health Literacy & Clinician Linguistics

- Break (10:40-10:55)

- Segment 2: Getting Started: NLP tools and UCSF infrastructure
  - Detailed technical review by five data scientists covering several methods in open source
Information Extraction: NLP of Clinical Text
A Simplified Framework

Preprocessing
- Parsing, tokenization, cleaning, POS tagging

Entity Recognition
- Tokens
- Word Vectors
- Features

Relationships
- Sentiments
- Negation
- Categorization
...

Info Extraction for Predictions/Application
- Extracted entities list
- Study-specific predictions (e.g., brain tumor)
- Patient-specific predictions
- Etc. Etc.
Today’s Organizers Represent Central Orgs in Research Data

- Information Commons
  - Advocacy and Navigation (CTSI, SOM Tech, Library)
  - Talent and Training (Library, CTSI, EpiBio)
  - Commons Technical Architecture, Design, Tools (ICHS, CDHI)
  - Shared Computing (ICHS, IT)
  - Data Pipeline (IT, CDHI & Others)

- Research
  - LHS
  - Clinical Practice
  - QI
  - Hospital Operations

- UCSF Clinical Systems
- Public sources
- Data from Apps
- Output data from Research

- UCSF Clinical Data Systems

- Today’s organizers represent central organizations in the research data output.
Symposium Audience

60% have never used NLP

Representatives from UC Berkeley, Davis, Irvine

Your feedback at registration shaped today’s agenda