# Physics, Health, and Human Trafficking?

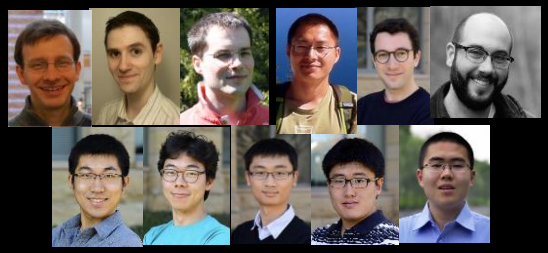
CS341 datasets:

NOT exhaustive

chrismre@cs.stanford.edu

# Extraction

DeepDive



The DeepDive Team

http://deepdive.stanford.edu/


**Dark Data** System: ETL on Steroids

**Quality** that can exceed paid human annotators and volunteers

# Human Trafficking on the (Dark) Web…


**Hypothesis**: Trafficked individuals offer ***lower cost*** and ***riskier*** sexual services.

***In Plain sight***: Web ads for such services

**Challenges**:

1. Need **high-resolution information** to build model.
   - *services for what rate, ethnicity, location, etc.*
2. Scientific papers are **clear**—*dark web is* **obfuscated**.

# In Use by Law Enforcement



*New York DA use MEMEX Data for all trafficking investigations this year.* ***Real Arrests***

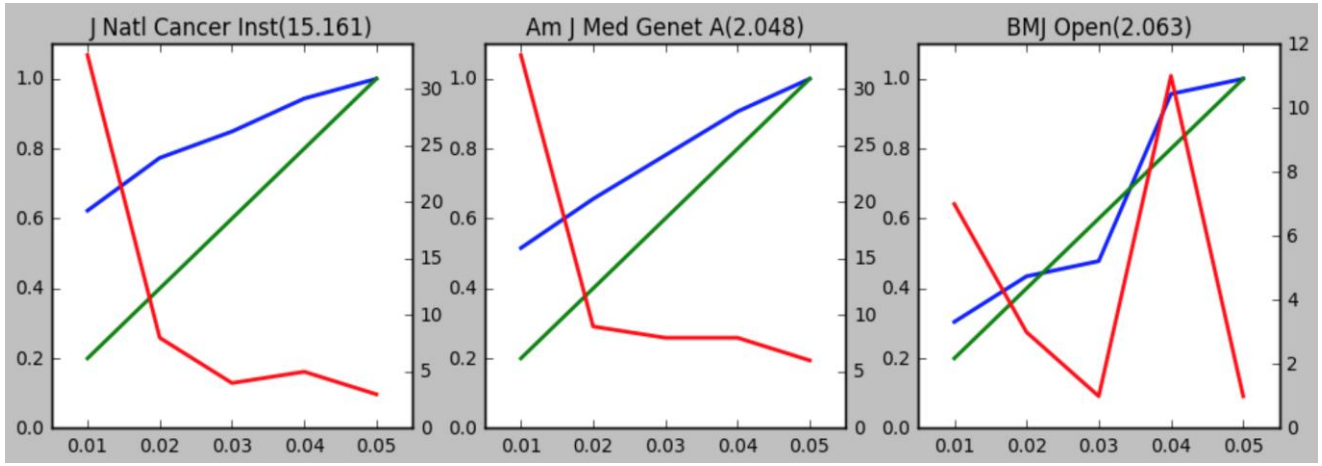
For DARPA MEMEX, we were operational in 6 months

- Processed >35M documents (~26M records)
- Tens of columns (location, phone #, price, etc)
- With compute times of less than a day
- >90% Precision for most relations

Project: Analyze world's biggest (oldest) trade.

# Extracting P-Values from Scientific Literature

- P-value used to show the statistical significance of the study or findings
  - Under no hacking p-values show be uniformly distributed!
- Using DeepDive we analyzed PubMed articles and extracted p-value mentions
  - Processed 500,000 articles and extracted 94,434 p-value mentions
- Automatic generation of p-curves for 2,000 Journals




- **Extraction count**
- **Observed distribution**
- **Uniform distribution**

Do it for real!

# TAIR: A Plant Genome Database

- Model plant: lessons learned in this organism can be translated to plants that are important to humans: food, medicine, biofuel, housing, etc.
- genome completely sequenced in 2000, 30K+ genes, 10K labs all over the world
- 50-70K academic papers describing basic and applied research.  Need many more!
- www.arabidopsis.org

mobilize

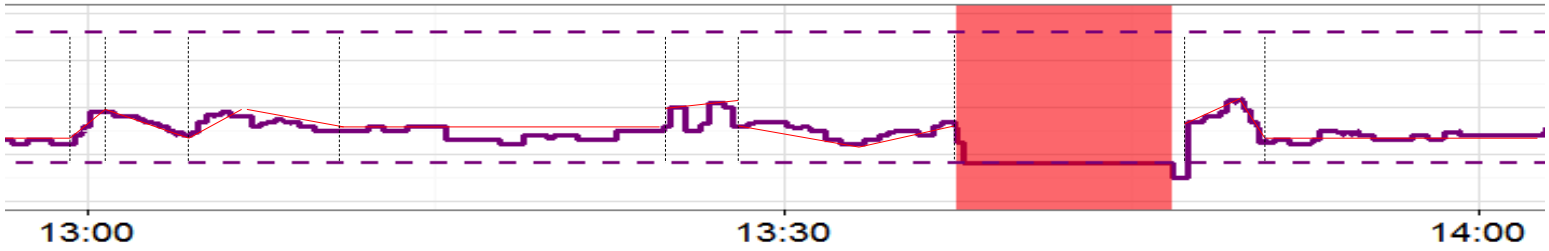
Center for Mobility Data
Integration to Insight

Class project pitches

MP1

# Learning representations from time series data

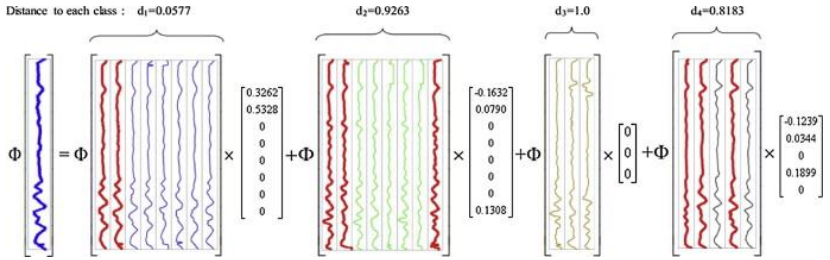
- Extracting features from time series
  - is a painstaking process
  - relies on domain knowledge
  - often requires multiple iterations
- Project aim: automate the extraction process
- Data: accelerometer step counts,
  motion tracking wave forms, vital sign data



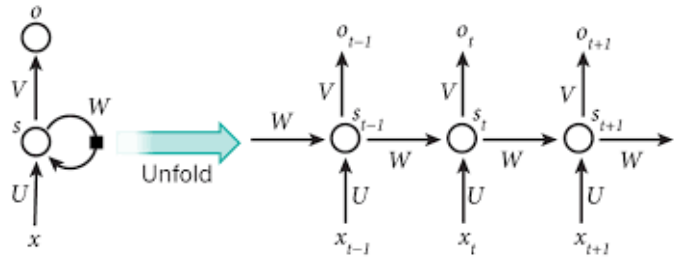
MP1

# Learning representations from time series data

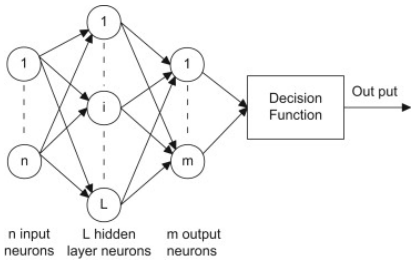
mobilize
Center for Mobility Data Integration to Insight




Kernel representation




Recurrent neural networks




Extreme learning machines

Time series representations

Contacts: Ina Fiterau (mfiterau@cs.stanford.edu), Jason Fries (jfries@stanford.edu)

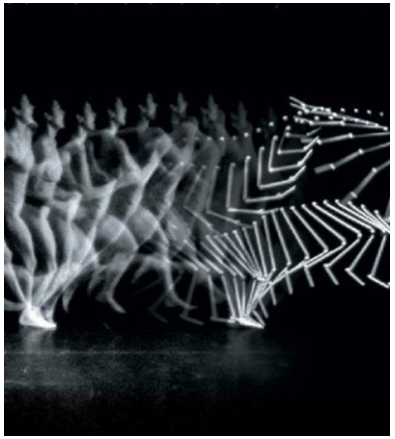
References:

[1] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

[2] Zhihua Chen, Wangmeng Zuo, Qinghua Hu, Liang Lin, Kernel sparse representation for time series classification, Information Sciences, Volume 292, 20 January 2015

MP5

# Detecting Running Injury Patterns

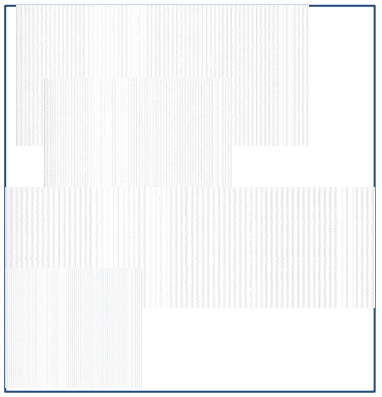


Motion tracking
(time series)



Extracted features
(structured)
obtained by applying
domain expertise

Running
Injury
Clinic



Medical chart (structured)
NOT UNIFORM
(collection depends on study)

Stereo recordings (video)

MP5

# Detecting Running Injury Patterns

- 1645 sessions (clinic visits), 1424 patients, 18 studies
- Studies handle patients with different conditions
- Aim: predict injury type/severity based on running pattern
- Topics: missing data imputation, domain adaptation
- Collaborators: Jennifer Hicks, Kevin Thomas (MD PhD student)
- Data agreement signature required

Session Data

- subject height, gender, study
- running frequency, distance
- skeletal structure measurements
- muscle strength, flexibility
- injury description

Kinematics session (as many as 10 per subject)

- at least one per session
- different test conditions
- 3 types of markersets
- features extracted from motion tracking (foot, ankle, knee, hip, pelvis)

Contact: Ina Fiterau (mfiterau@cs.stanford.edu)

SLAC.
Imaging and Physics

Email chrismre@cs.stanford.edu to get access to physics projects.