Fogarty International Center

Global mHealth Research Training Institute

June 6- 9, 2016

Center for Global Health Studies



NIH Fogarty International Center

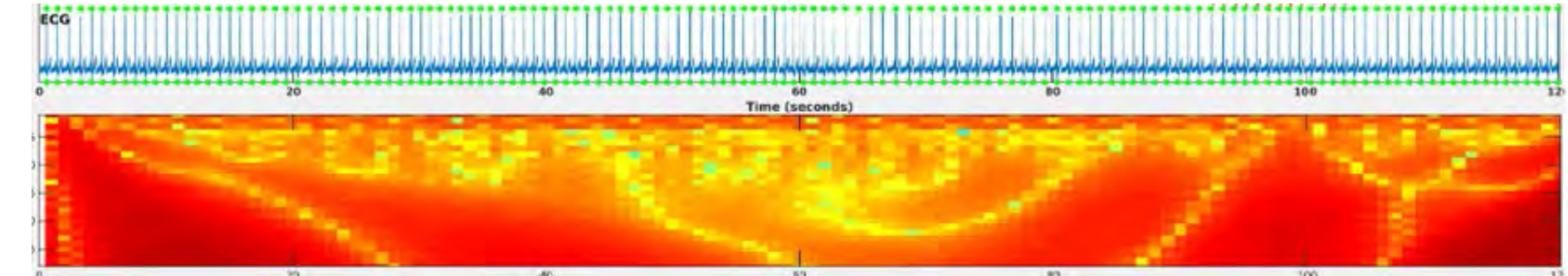
NIH Fogarty International Center
Advancing Science for Global Health

Global mHealth Research Training Institute

How do you do the analysis?

(Signal Processing and Machine Learning)

Gari Clifford



en·gi·neer
/ˌenjəˈnir/

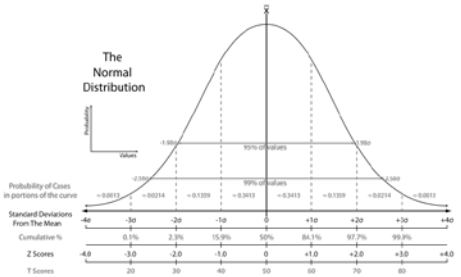
noun

1. a person who designs, builds, or maintains engines, machines, or public works.
   synonyms: originator, deviser, designer, architect, inventor, developer, creator; mastermind
   "the prime engineer of the approach"

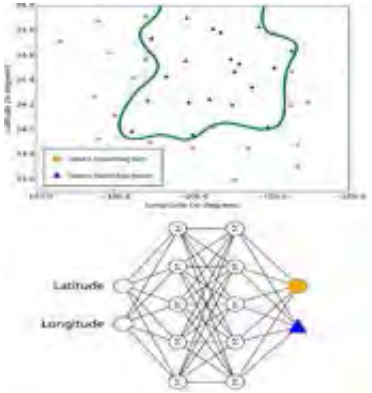
Who should I look for to help me?

- Database or Cloud Computing experts - how do I usefully store my data?
- Electronic Medical Record (EMR) experts
- Experts in ontologies (how you describe my data in a formal way)
- Experts in UI design
- *Experts in Signal Processing*
- *Experts in Data Analytics / Machine Learning (prediction, classification, etc.)*
- *Experts in Crowdsourcing*

Analytics: combining multiple fields



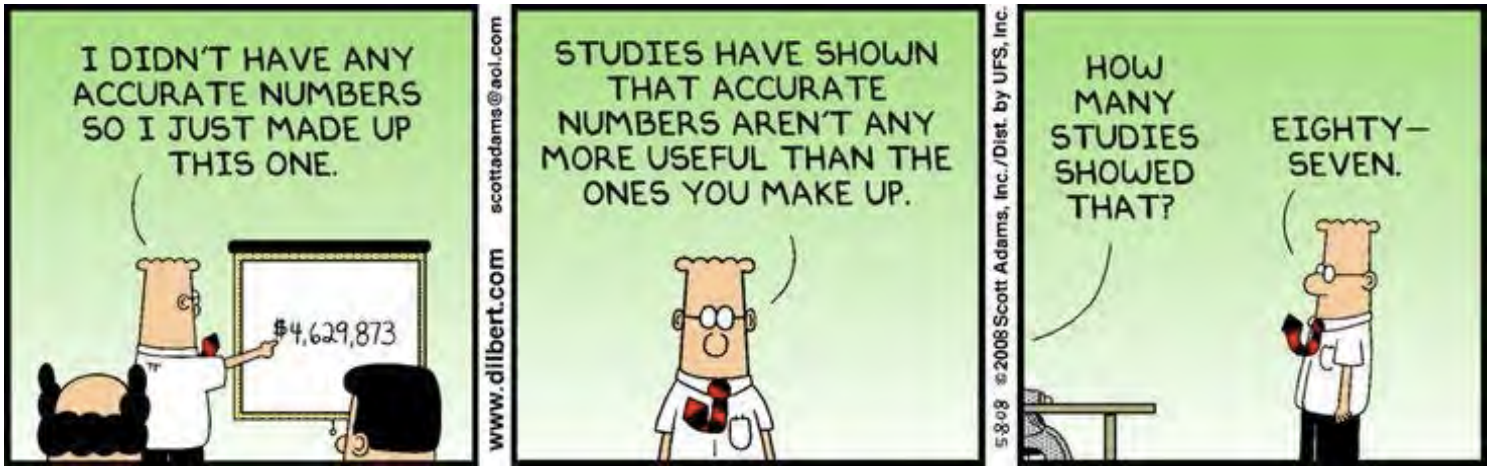
- (Old School) Statistics
  - Usually assume data are independent, linear and have some given distribution (not always)
  - Power Calculations, Inference (what process led to the result?)
- Signal Processing
  - filtering and transforming the series of measurements from your sensors) – not just your standard statistics, because they are temporally correlated - plus **feature extraction**
- Machine Learning
  - Learning from examples
  - Generally used for classification or predictions.
  - Cross validation instead of p-values



Why do I need a (bio) statistician?

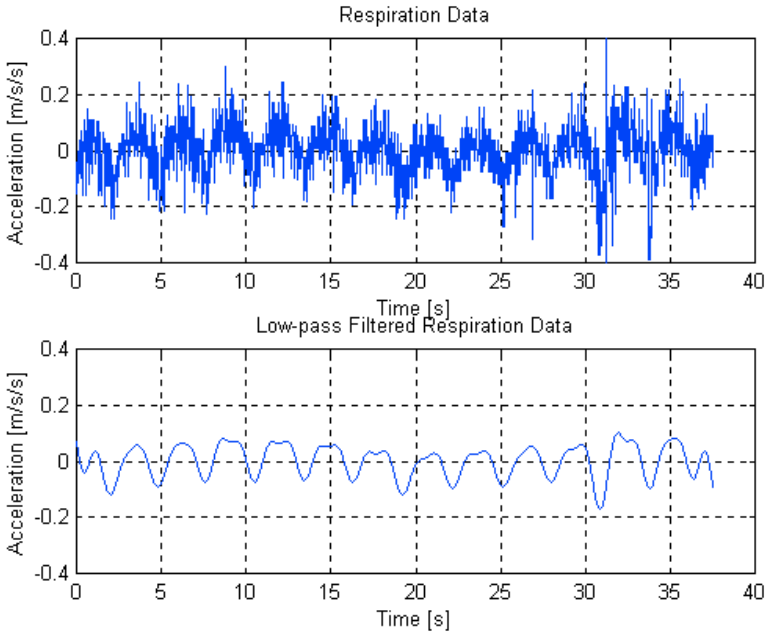
- Power calculations
- Traditional significance tests on pilot data/surveys
- Building (logistic) regression models,

- There's lots in common with data science / machine learning



Signal Processing Experts – What do I need them for?

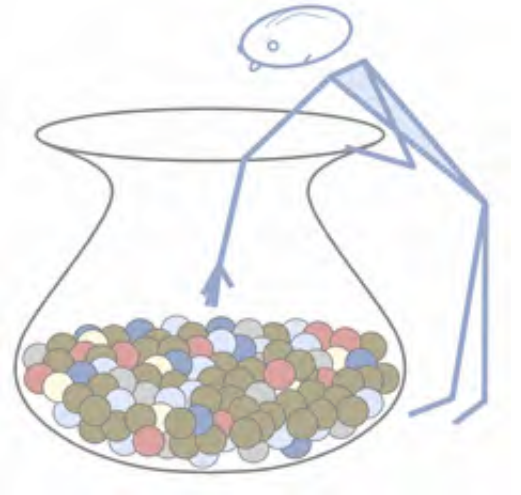
- Identifying which sensors you might need
- Which ones can give you the feature you need …
- Making (non-proprietary) algorithms for:
- filtering (removing noise)
- feature extraction (e.g. respiration rate from accelerometer)
- Avoiding pitfalls (aliasing, clipping, etc)



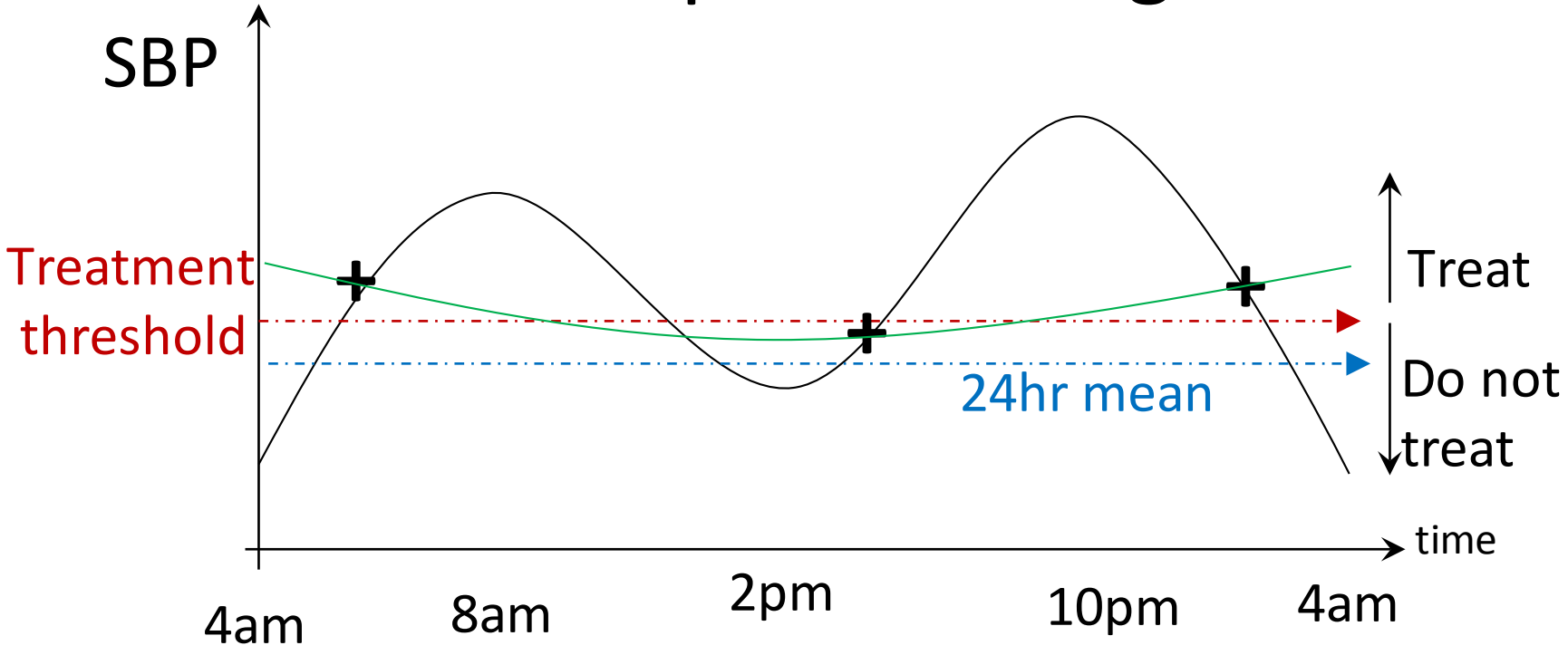
Signal Processing – E.g. Sampling and ‘aliasing’

Take the example of periodic blood pressure measurement:

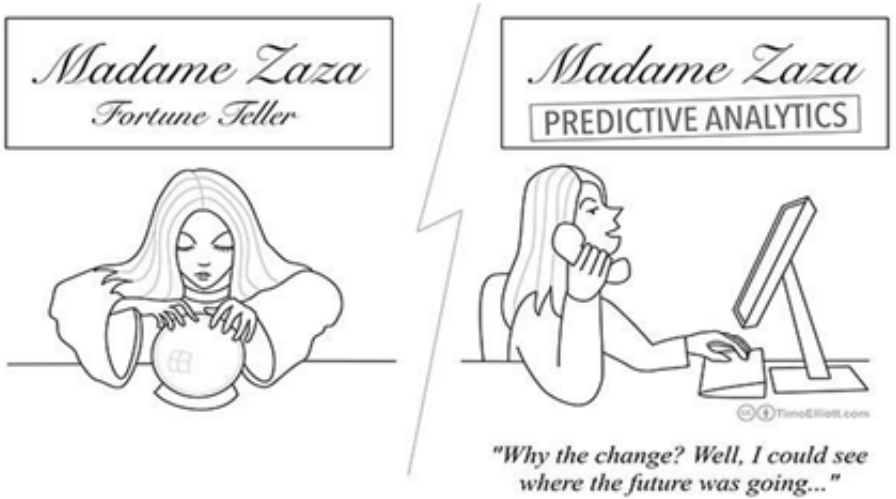
- Is it a relevant device for your population?
    - E.g. there are BP monitors not approved for your population, wrong cuff size, body habitus is wrong [can they sit with elevated arm?]
- How often? Repeated samples to reduce noise …
- When should you measure it? Above Nyquist!!



Example: ‘Aliasing’



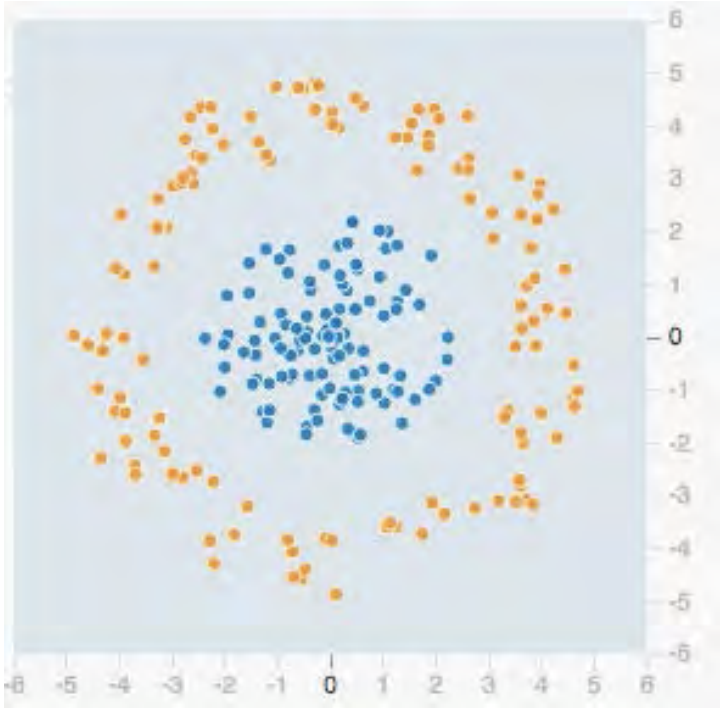
For what do I need a machine learning expert?



Two main reasons:

- Classification (Does the photo of the melanoma look benign or not?)
- Prediction (will the flu outbreak happen this week?)

Machine Learning 101

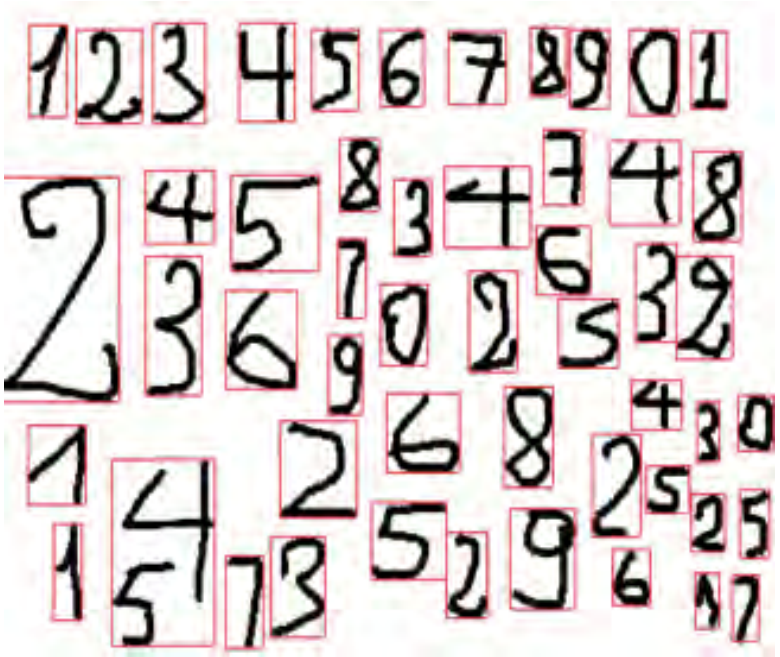


'A type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of **computer programs that can teach themselves to grow and change *when exposed to new data***.'

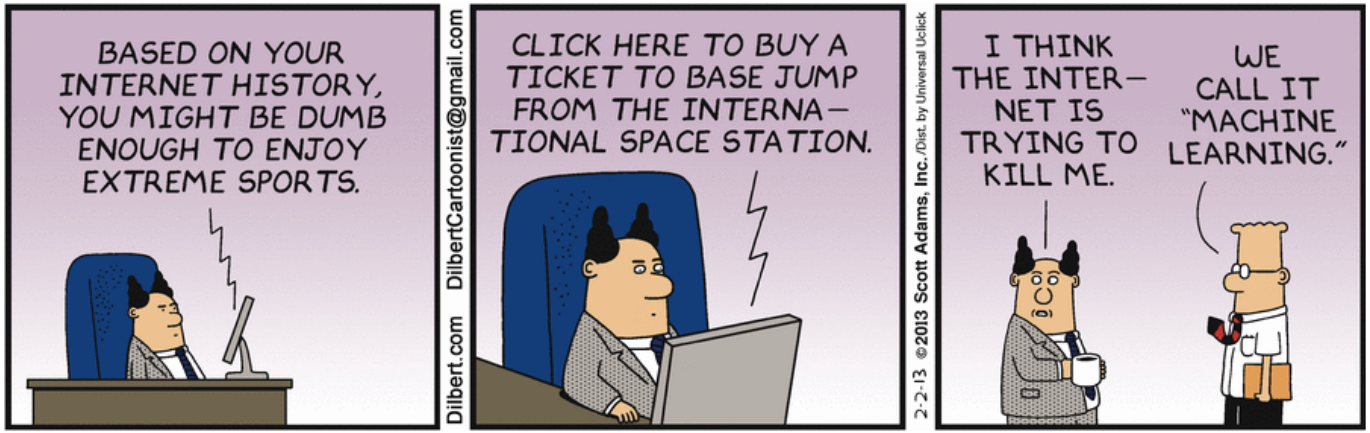
E.g. Tensor Flow (Google) for Deep Learning

Repeated learning of general underlying patterns

- … from lots **(and lots)** of examples:
- Did I say Big Data?



Some real machine learning examples



Recommendation systems (e.g. Amazon 'if you purchased this, why not buy that?')

- Uses: your ratings, viewing history, geography, preferred language, viewing device, time of day, etc,
- Deep Neural Network trained to map these to the 'next' most likely purchase

Natural language processing (NLP)

Many approaches to analysing free text. Promising ML techniques/software:

- Sentiment Analysis
  - https://cloud.google.com/prediction/docs/sentiment_analysis
- Word2Vec:
  - Neural network based
  - http://deeplearning4j.org/word2vec

Mining Twitter Data to Improve Detection of Schizophrenia

Kimberly McManus, BS[*1], Emily K. Mallory, BS[*1], Rachel L. Goldfeder, MS[*1], Winston A. Haynes, BA[*1], Jonathan D. Tatum, BS[*1]
[1]Stanford University, Stanford, CA

Abstract

Individuals who suffer from schizophrenia comprise 1 percent of the United States population and are four times more likely to die of suicide than the general US population. Identification of at-risk individuals with schizophrenia is challenging when they do not seek treatment. Microblogging platforms allow users to share their thoughts and emotions with the world in short snippets of text. In this work, we leveraged the large corpus of Twitter posts and machine-learning methodologies to detect individuals with schizophrenia. Using features from tweets such as emoticon use, posting time of day, and dictionary terms, we trained, built, and validated several machine learning models. Our support vector machine model achieved the best performance with 92% precision and 71% recall on the held-out test set. Additionally, we built a web application that dynamically displays summary statistics between cohorts. This enables outreach to undiagnosed individuals, improved physician diagnoses, and destigmatization of schizophrenia.

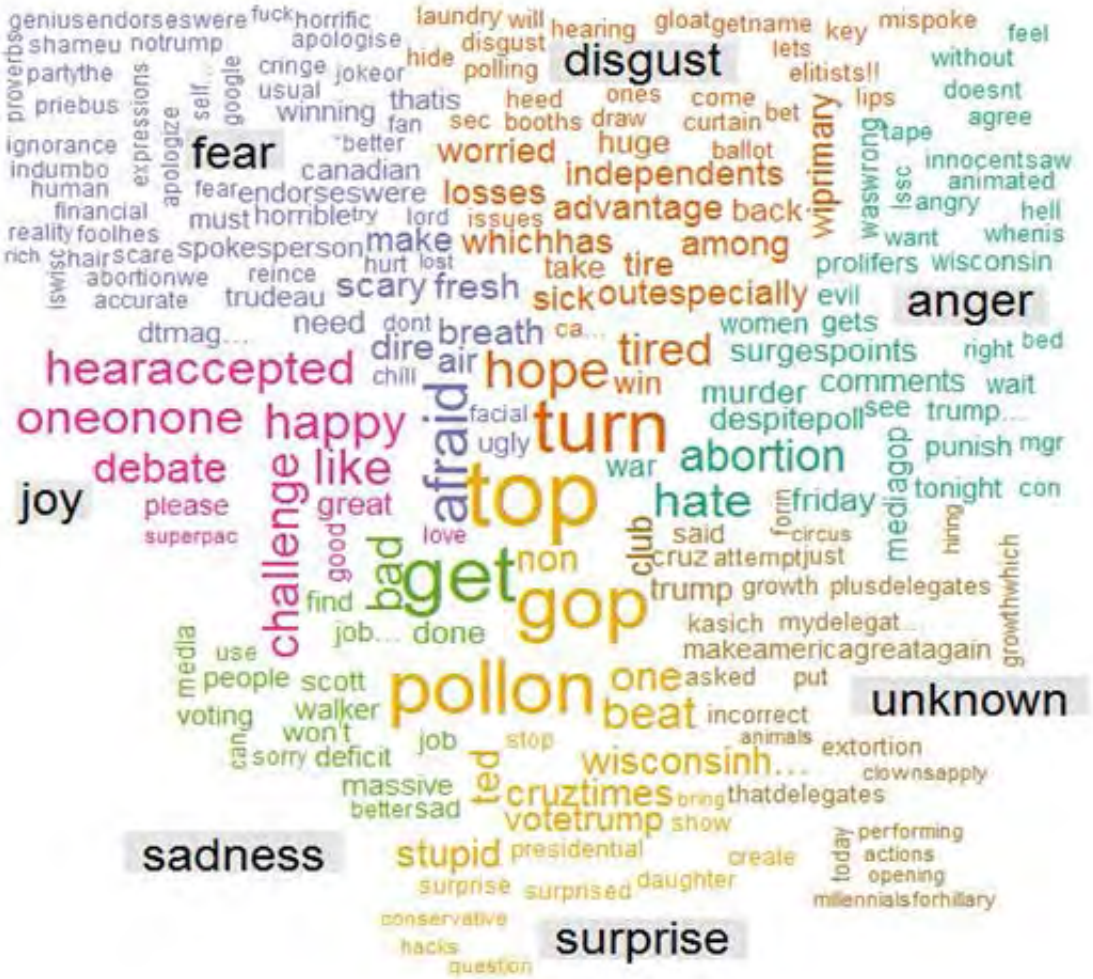
Ref	Media	Cohort Acquisition	Features	Approach	Results
[18,19]	Twitter	Clinical depression surveys	Interactions, emoticons, vocabulary, drugs, linguistic style, behaviors	Support vector machine	0.74 precision 0.63 recall
[20]	Bulletin boards	Prozac post, doctor curation	Vocabulary	2-step support vector machine	0.82 accuracy AUC 0.88
[21]	Sina	Psychologist curation	Pronoun use, emoticons, interactions, behaviors	Weka, BayesNet	0.91 accuracy AUC 0.90

AMIA Jt Summits Transl Sci Proc. 2015; 2015: 122–126.

Natural Language Processing

Trump's Tweets were fed to these NLP engines and clustered using ML to create a 'word cloud' *mapped* to 6 categories …

Donald Trump's cloud shows that his defeat in the Wisconsin primary is classified under **surprise** while the words "Canadian" and 'Trudeau' (referring to Canadian PM Justin Trudeau) are classified under the **fear** category!



Labeling enormous databases

Two key issues:

1. Humans disagree on diagnoses and labels, even if 'event' is well described.

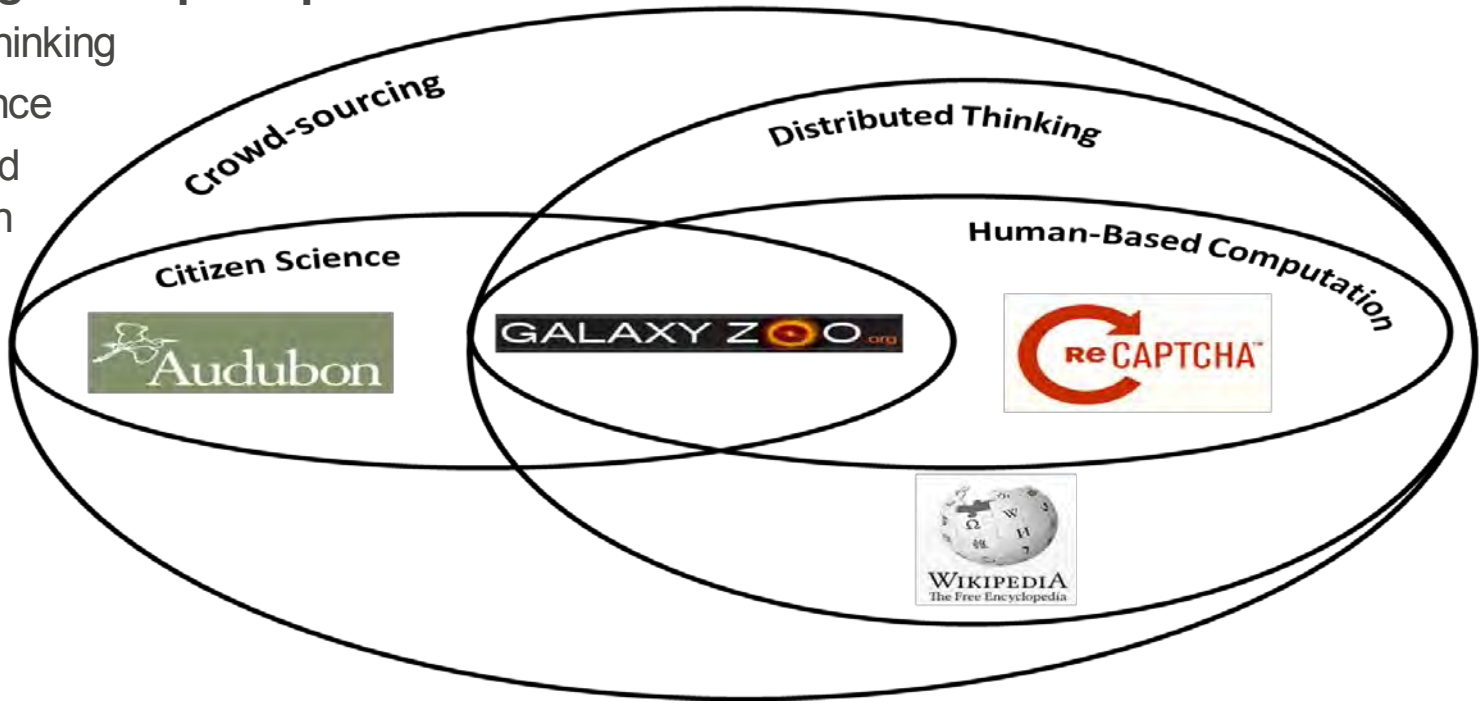
   - Inter- and intra-human bias and variance in diagnoses must be addressed.

1. Labeling of medical data is vast & impossible to do so by hand

   -> Crowd sourcing and intelligent aggregation

Introduction to Crowdsourcing

- **Crowdsourcing concept map**
  - Distributed thinking
  - Citizen Science
  - Human-based computation



-Image adapted from Wong 2011.

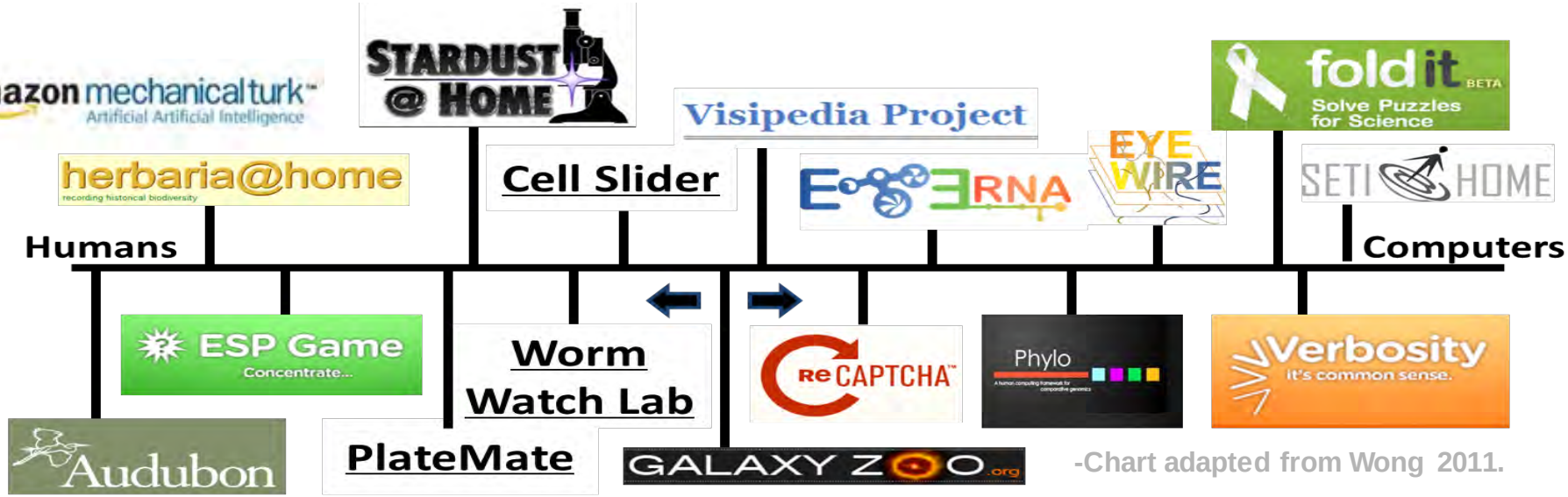
E.g. – ‘Captchas’ or ‘ReCaptchas’

- Run ‘OCR’ software, and when they disagree, send it to humans to process
- When you have a ‘Turing’ challenge you are sometimes transcribing rare books!

- But how many humans (or algorithms) do you need? ….
- Lots! (can be 10-100) … Combine their answers using ML!
- So that’s machine learning of machine learners!

Crowdsourcing Project Continuum

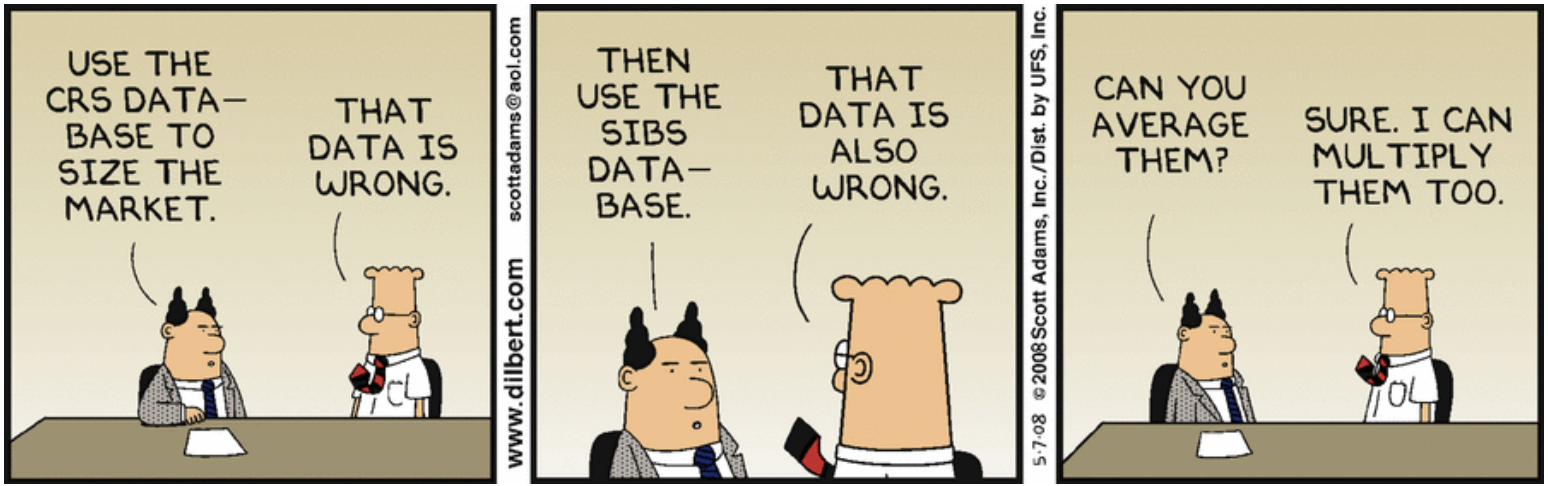
- From Humans to Computers ….



-Chart adapted from Wong 2011.

Be careful though …

Data are like people – interrogate them hard enough and they will tell you whatever you want to hear.



I.e. if you do poke around in the data, don't jump to conclusions – use it to inform a hypothesis that you will then test with new data!

So what can I do? … Find the right person

- (Neural) Network to find the right person
    - (ask Wendy, Gari, Donna, ...)
- If they want your data without a discussion and then pump back an answer .. walk away
    - Need to iterate - sit down and explain the data - why it was collected, and where all the problems might be. Then look at results, question if they make sense, and iterate.
    - NIH will want to see a strong partner who is involved and wants to gain domain knowledge
- How do you deal with missing data (if they say ‘don’t worry’ - get worried)
    - Is it missing ‘not-at-random’? E.g., do they turn their phones off when they are sick (as well as on holiday)?
- What other data sources are needed?
    - Get experts involved early on to help define what is collected and why – make them understand the data

Appendices


OUR MACHINE LEARNING TECHNOLOGY ALLOWS US TO TRACK CUSTOMER PREFERENCES AND USE THAT KNOWLEDGE TO MANIPULATE THEM.
THAT SEEMS LIKE THE STEP THAT HAPPENS RIGHT BEFORE THE MACHINES TAKE OVER THE EARTH AND ANNI-HILATE ALL HUMANS.
THERE'S ALWAYS ONE PERSON IN EVERY CROWD WHO SAYS THAT.
NOT FOR MUCH LONGER, APPAR-ENTLY.
Dilbert.com DilbertCartoonist@gmail.com
1-31-13 ©2013 Scott Adams, Inc./Dist. by Universal Uclick

Database Types – How do I choose?

- Data should be stored based on how you intend to analyse it.
- Extremely long temporal signals or images - flat file databases (e.g. PhysioNet.org) [use standard formats: EDF, WFB, DICOM]
- Small (<100 GB!) - Relational (SQL: Oracle, Postgres, MySQL)
- Large (Terabytes) and expandable: NoSQL (Apache Cassandra, HBase, MongoDB, and Couchbase)
- Enormous (petabytes): Oracle Exadata, Amazon Redshift, Netezza, ...
- Enormous with structured+unstructured data: Hadoop / MapReduce

PHRs, EMRs and Schemas

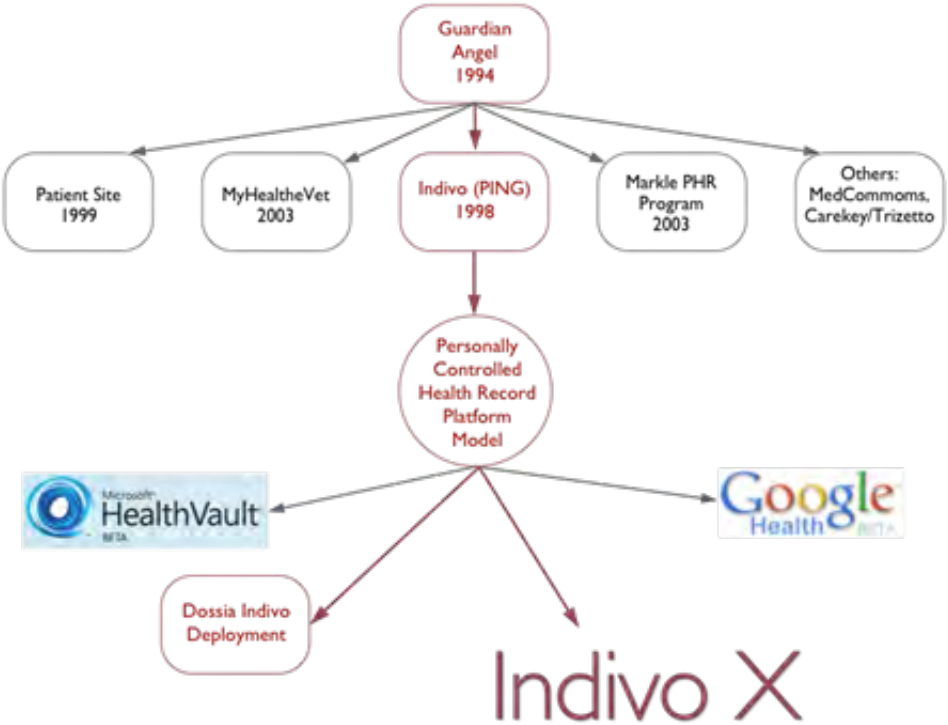
- **PHR**: Personal Health Record
- **EMR/EHR**: Electronic Medical/Health Record
- **Schema**: the skeleton structure that represents the logical view of the entire database. It defines how the data is organized and how the relations among them are associated. It formulates all the constraints that are to be applied on the data.

Open Source PHR: IndivoX

Open Source EMR: OpenMRS

Data interchange schemas: HL7, FHIR …

<u>Why are these important to consider?</u>

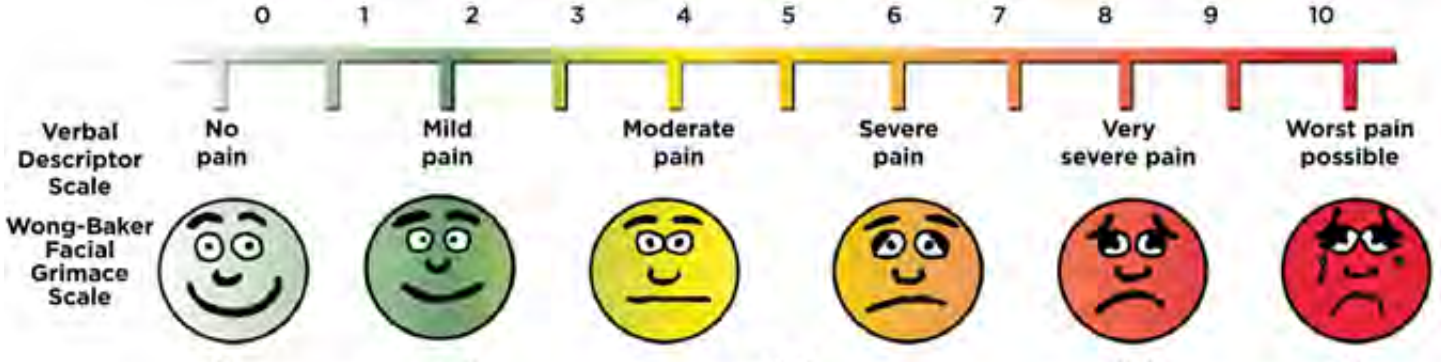


Data Types (in your data)

Free text - *(Avoid if you can?)*: hard to search, full of typos, can include personal identifiers

Ordinal & Categorical: If you can define the right set of labels, then very useful. Beware of casting categorical elements to ordinal ones then attempting to regress… distances between elements may be misleading

Numerical: Most accurate if appropriate. Again, careful not to cast categorical variables to numerical. Represent the pain scale as [0 0 1 0 0 0] not 3/6



Ontologies

What? They are ***explicit formal specifications of the terms in the domain and relations among them … or a hierarchical set of codes to describe the relationship between all the medical concepts in which you are interested.***

Why?

- To share common understanding of the structure of information
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

How? Attach specific alphanumeric codes to each medical concept

Which: SNOMED, LOINC, DSM-IV, RxNORM, ICD-9/10 ….
UMLS: Superset licensed by NLM - includes SNOMED, LOINC etc. Good for research, but mapping is not always unique

General issue: Non-degeneracy ... there are multiple ways to say the same thing, and you often need several codes to describe an issue.
So - **pick a subset of your ontolog(ies) and try to keep the # codes small.**

Continuous Quality Checks on Data Required

- As you are collecting data, you must do continuous quality checks
- Use redundant servers and data channels (e.g. two different mobile phone carriers for SMS, plus a WiFi sync each week)
- Daily email digests listing number of patients, volume of data - look for anomalies (high volumes => hacks, low volumes => dead batteries, patient drop-outs, thefts, etc)
- Consider crowd sourcing approaches to minimising quality audits (of non-protected health data) … e.g. Amazon Mechanical Turk

amazon mechanical turk™
Artificial Artificial Intelligence

Adaptive / Agile Programming

It's really the only way to develop your app – but try to include the communities you are addressing.

Reaching Out to Experts

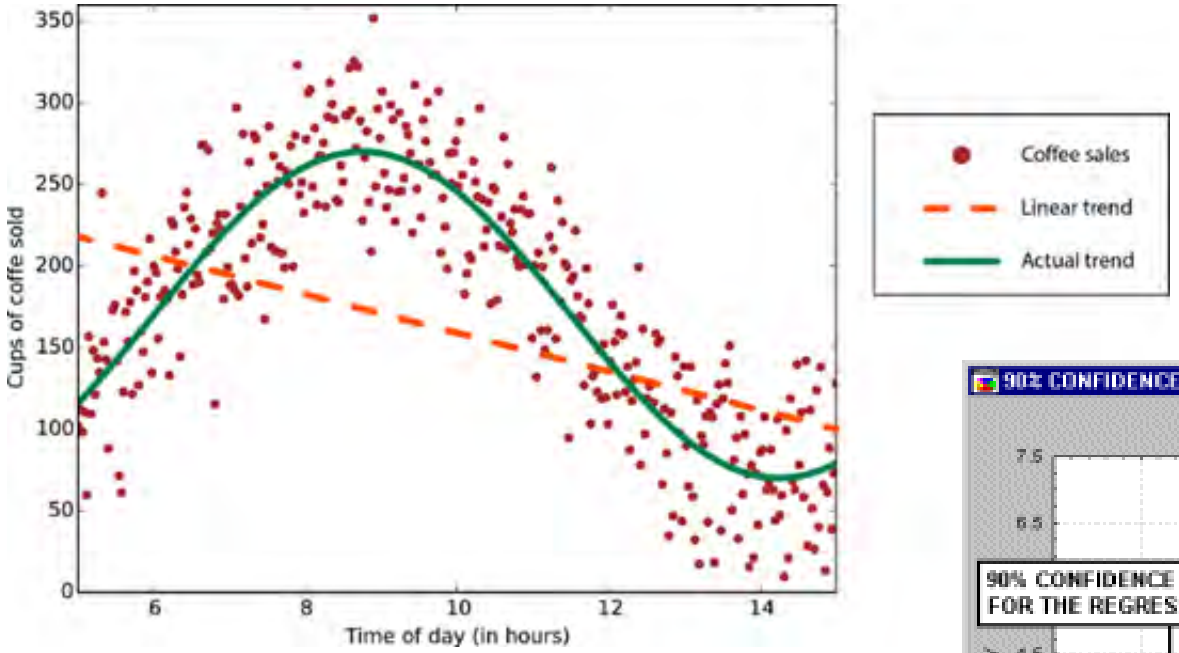
- What are the benefits of working with someone in this field?
  - mHealth has the potential for *scale*. ML really comes into it's own with scale, and can uncover unusual relationships in data that traditional statistical approaches cannot. A Signal processing or ML expert will help you avoid classic errors (such as aliasing, over-training on your data, failing to stratify by patient, etc)
- How might engaging someone in your discipline enhance/strengthen a mHealth application or data?
  - As you move to scale, the utility of your data will increase rapidly, allowing you to generate new hypotheses through your data.
- When should someone reach out to an expert in this area?
  - When they are considering collecting large amounts of data on which you need to make classification or predictions

Questions

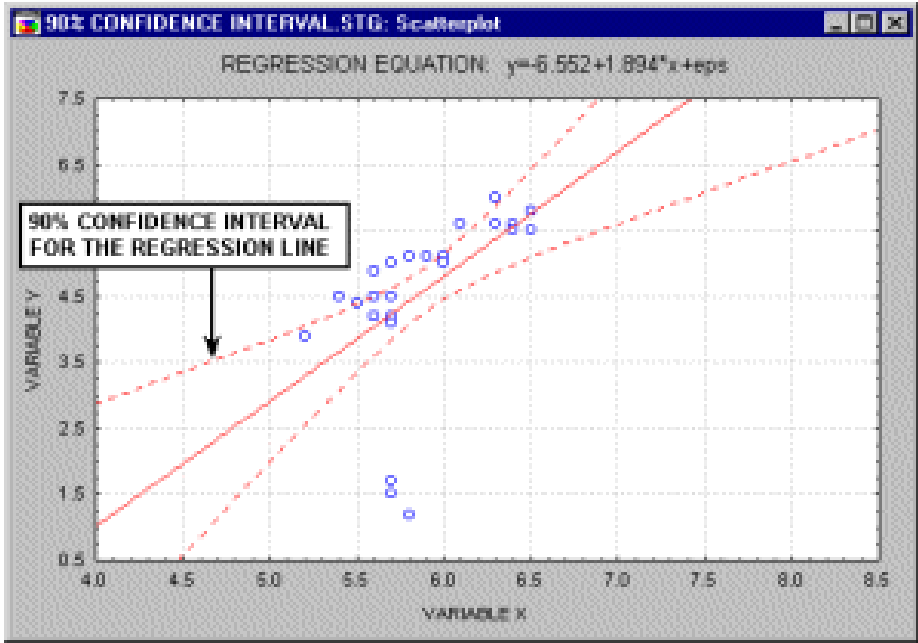
1. Which data will you collect? (Types: ASCII text, voice, numerics, photos, physiology …)
2. Into what systems will you load the data? (EMR, Redcap, flat file systems, Hadoop, Oracle Exadata, Amazon Redshift ...)
3. Do these need to be secure and if so, how? What elements need to be secured?
4. How much data will you store and how expensive will this be? (Tiers on AWS?)
5. If you are using cloud services, how can you ensure it stays in country (if needed)
6. What level of data loss is acceptable? How will you estimate data loss and its impact?
7. How do you deal with human error and inter-observer variability?

More on machine learning

Linear regression

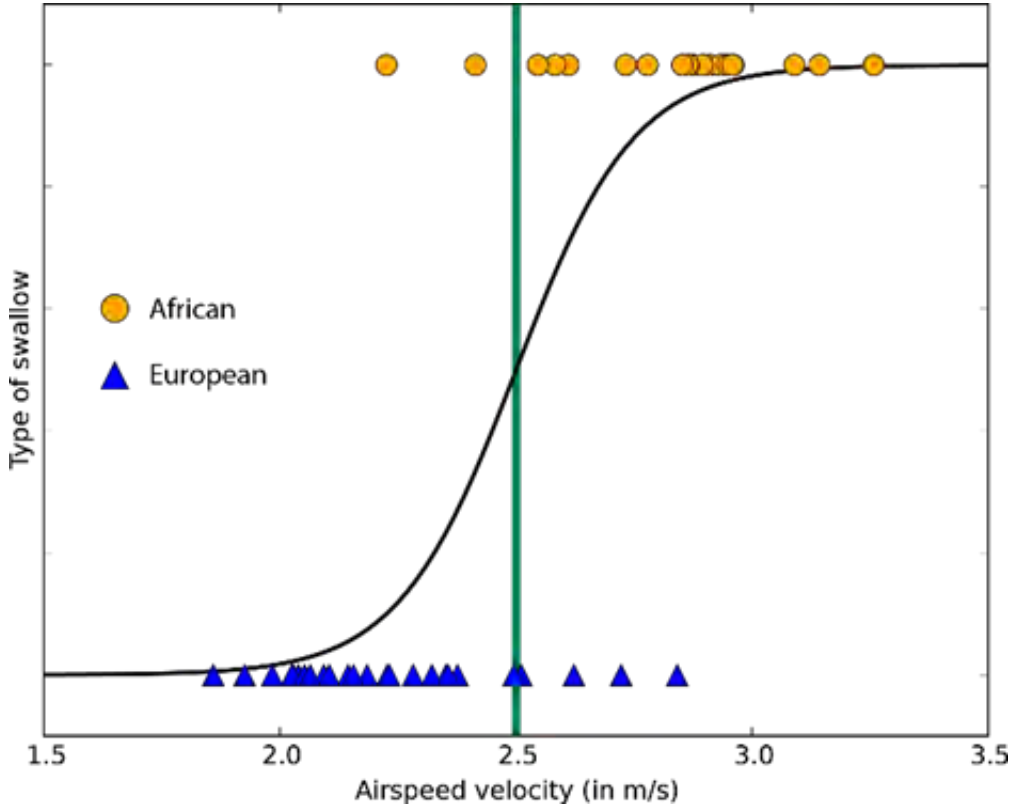


- Doesn't always fit the data well.
- Even with 'robust least squares'!



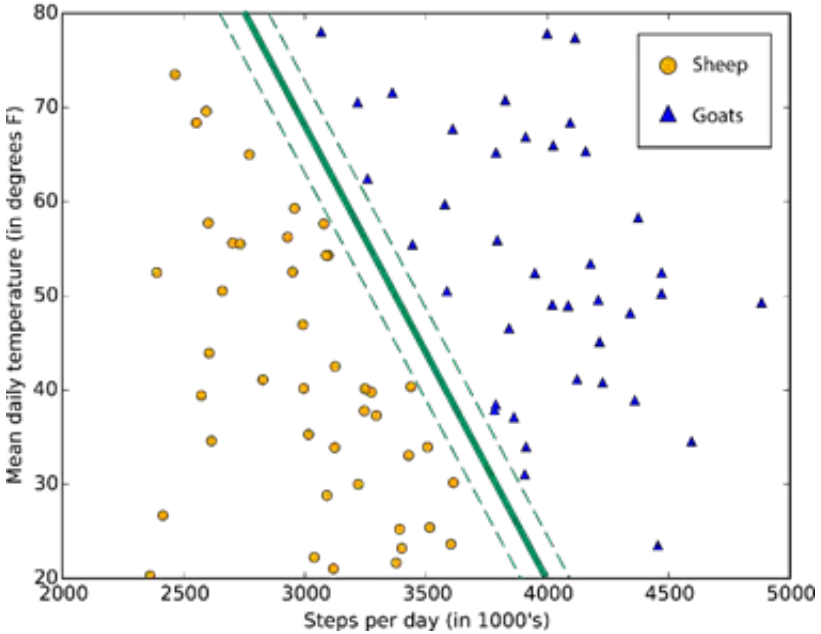
Logistic regression

- Good for binary outputs



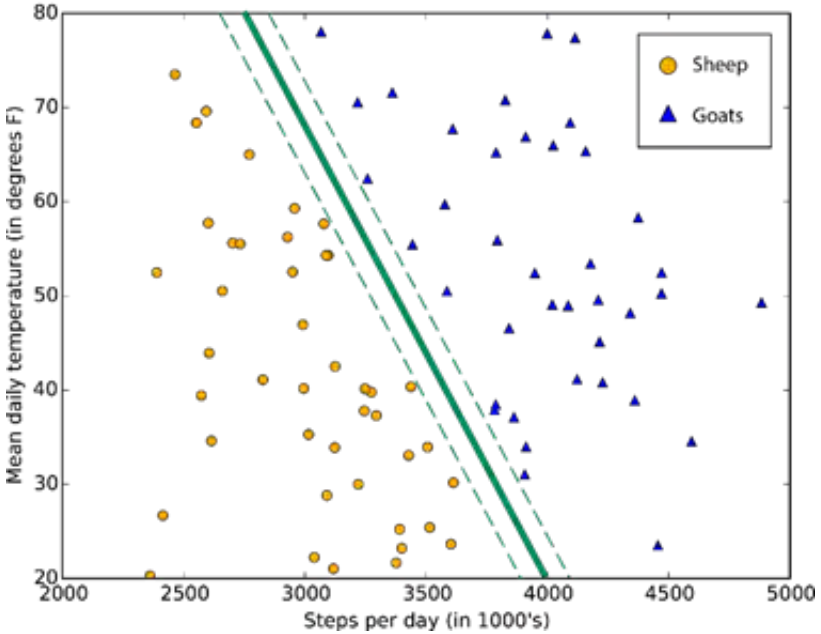
Nonlinear boundaries between classes?

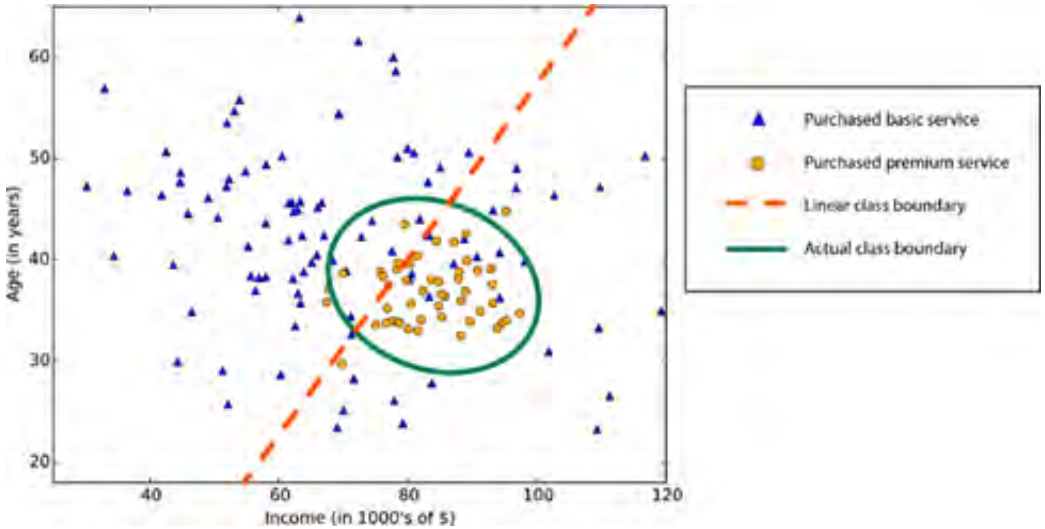
Linear separation / classification



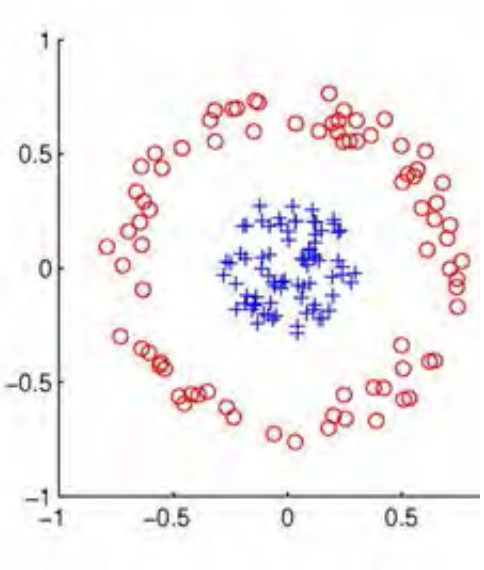
Nonlinear boundaries between classes?

Linear separation / classification …. Or not?





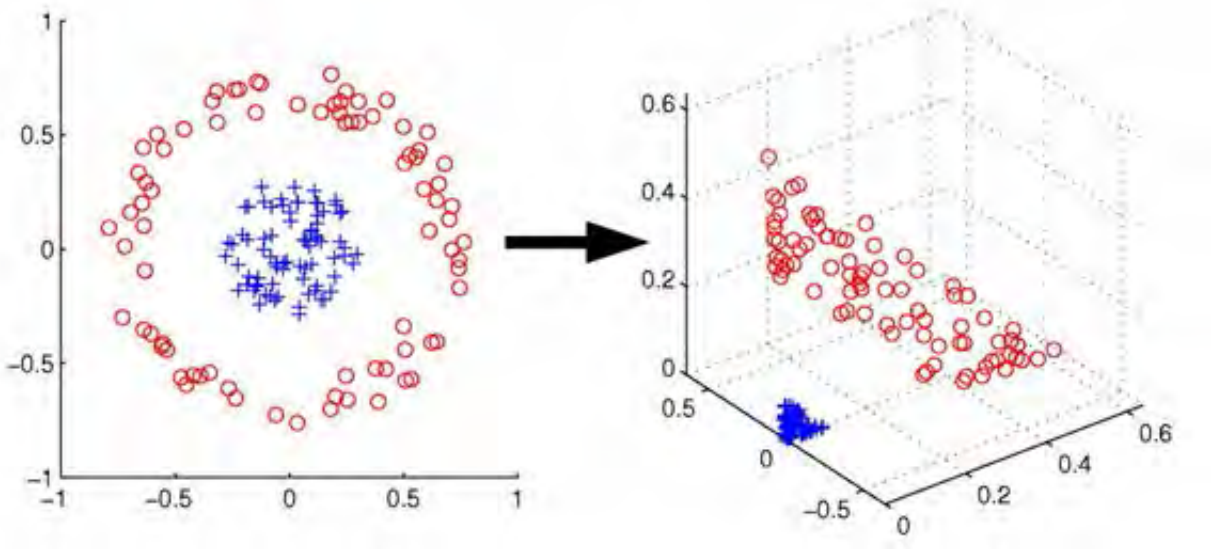
The Support Vector Machine



What if the data are not linearly separable?

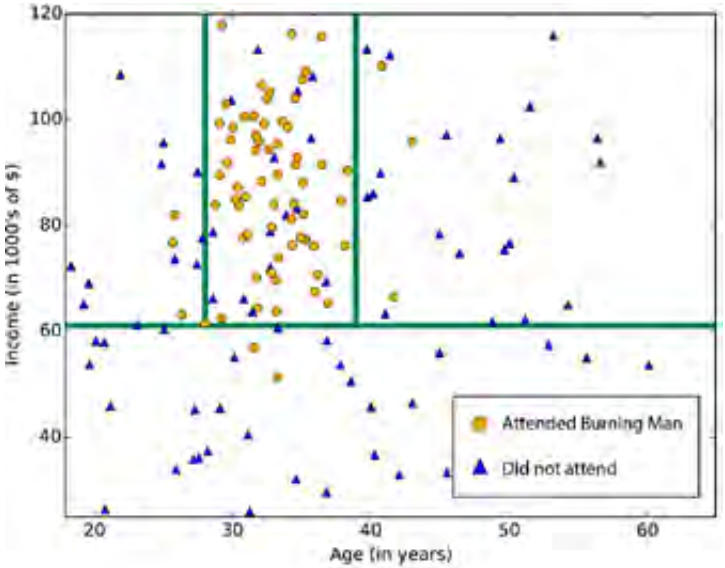
The Support Vector Machine

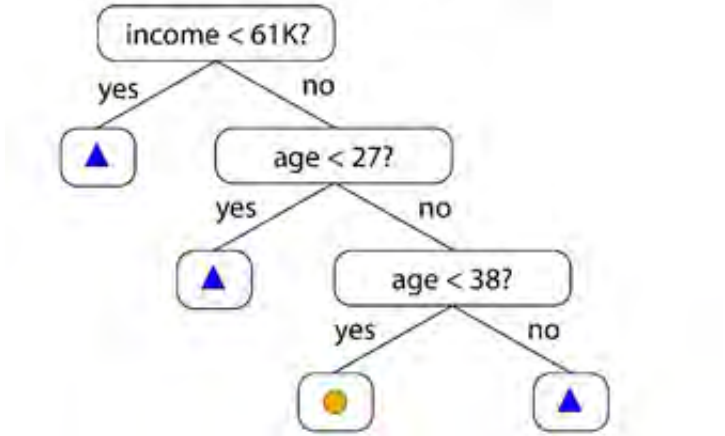
The ‘Kernel Trick’



Decision trees, random forests, ...

Good with mixed data types



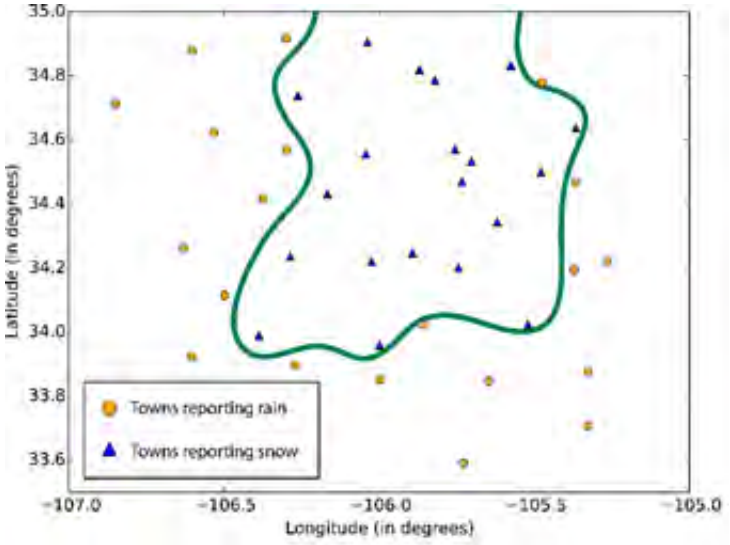


Neural networks?

The boundaries learned by neural networks can be complex and irregular

Deep learning?

- Can *auto-encode* data and create a hybrid supervised-unsupervised approach

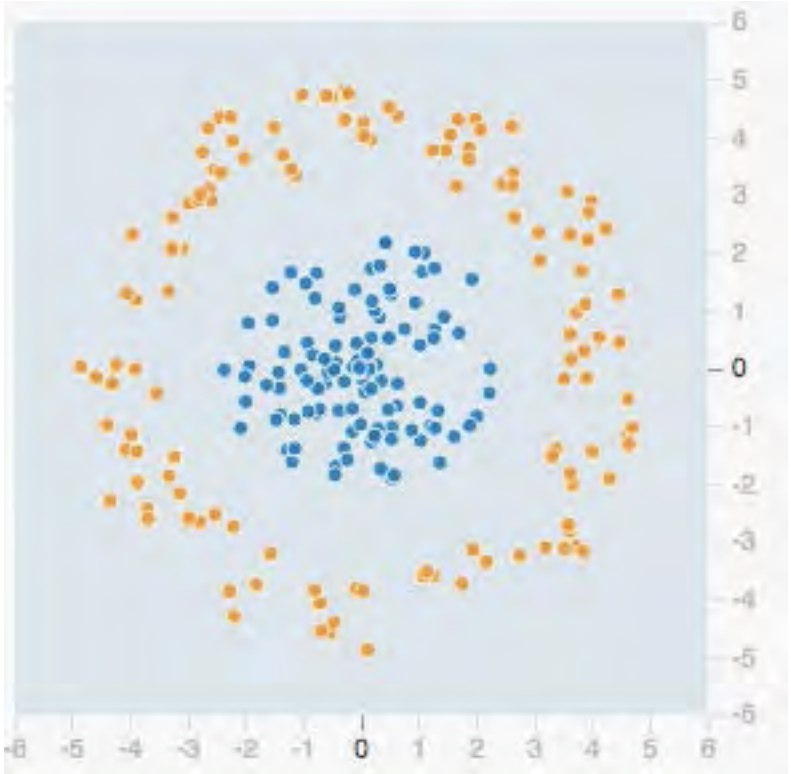


Deep Neural Networks

- E.g. Tensor Flow (Google)

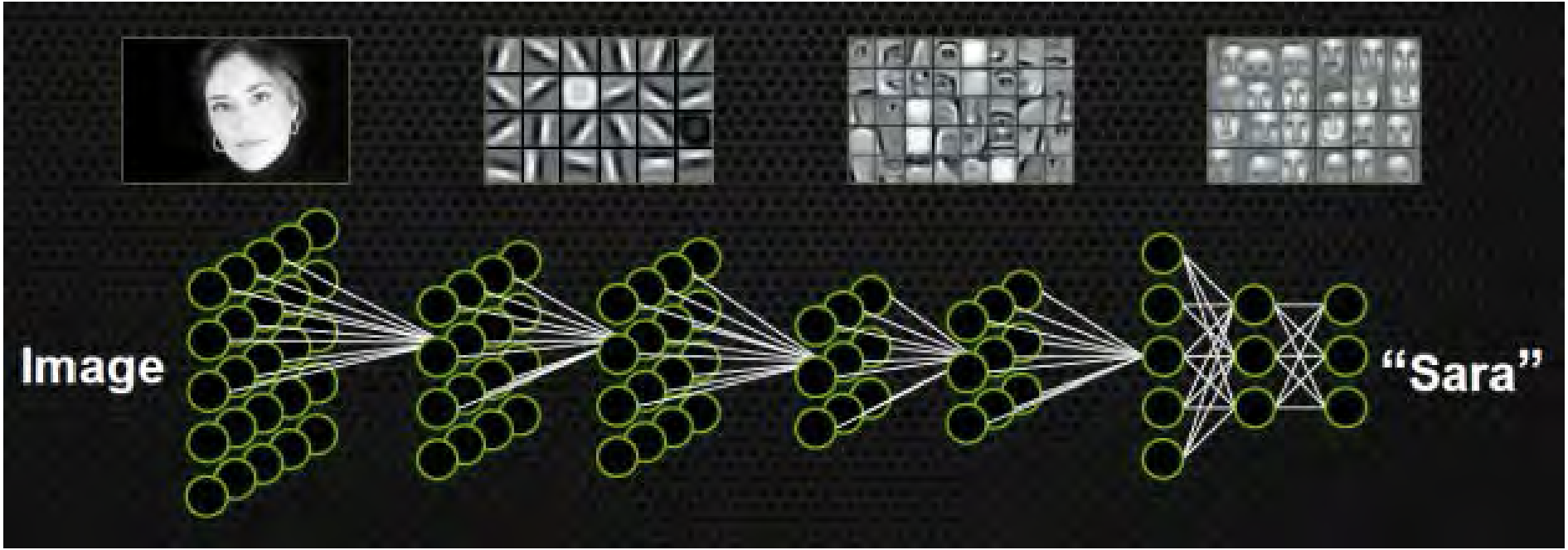
or

- Deep Scalable Sparse Tensor Network Engine, (DSSTNE or 'Destiny'), from Amazon



Successive layers encode ‘features’

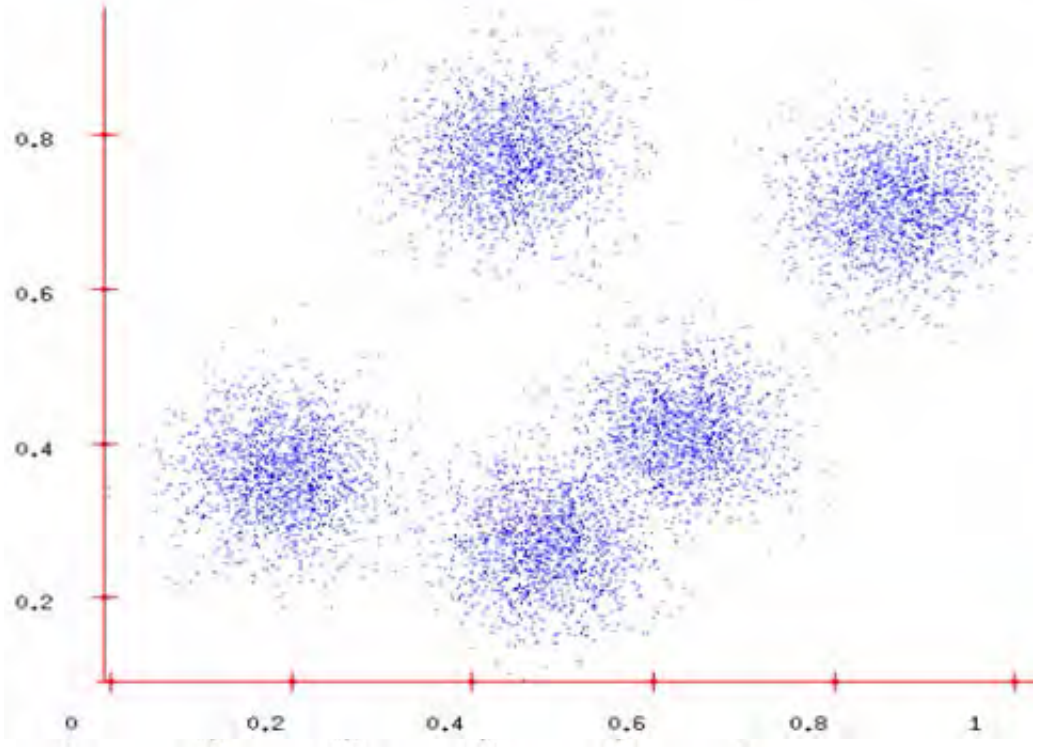
First layer could be linear regression, second encodes squares of factors, etc, … and last layer *could be* logistic regression!



Unsupervised classification

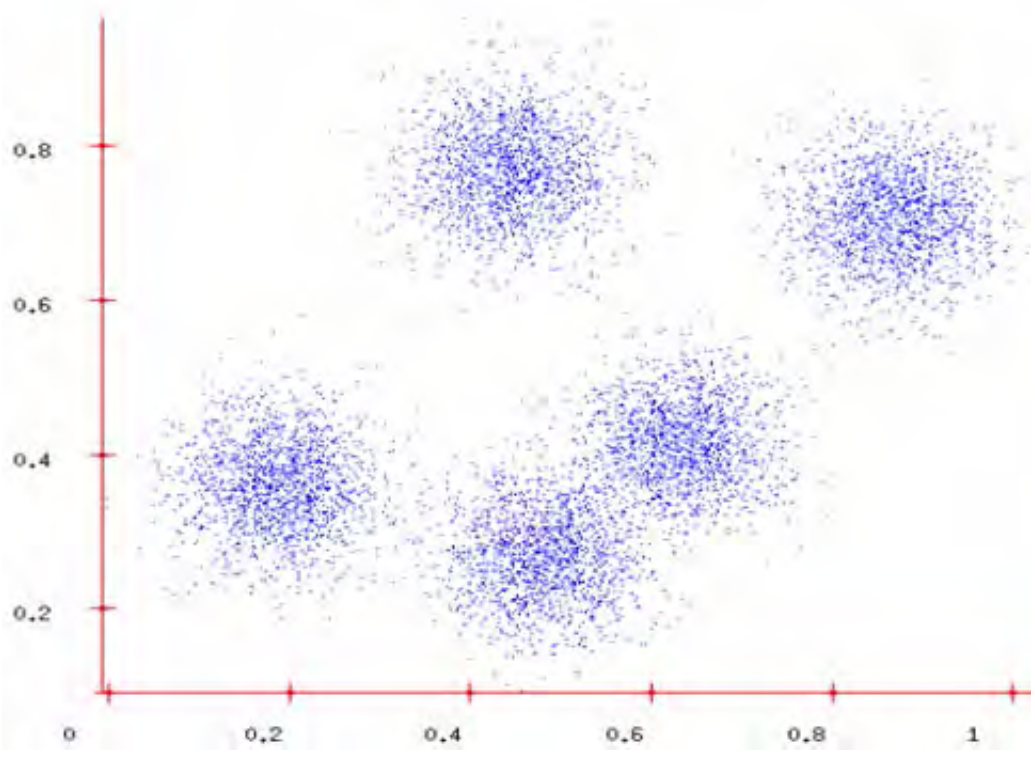
- Suppose you have some 2D data with this relationship?
- How do you ‘discover’ the underlying unknown classes?
- You need an unsupervised learning technique ….

e.g. **K-means**



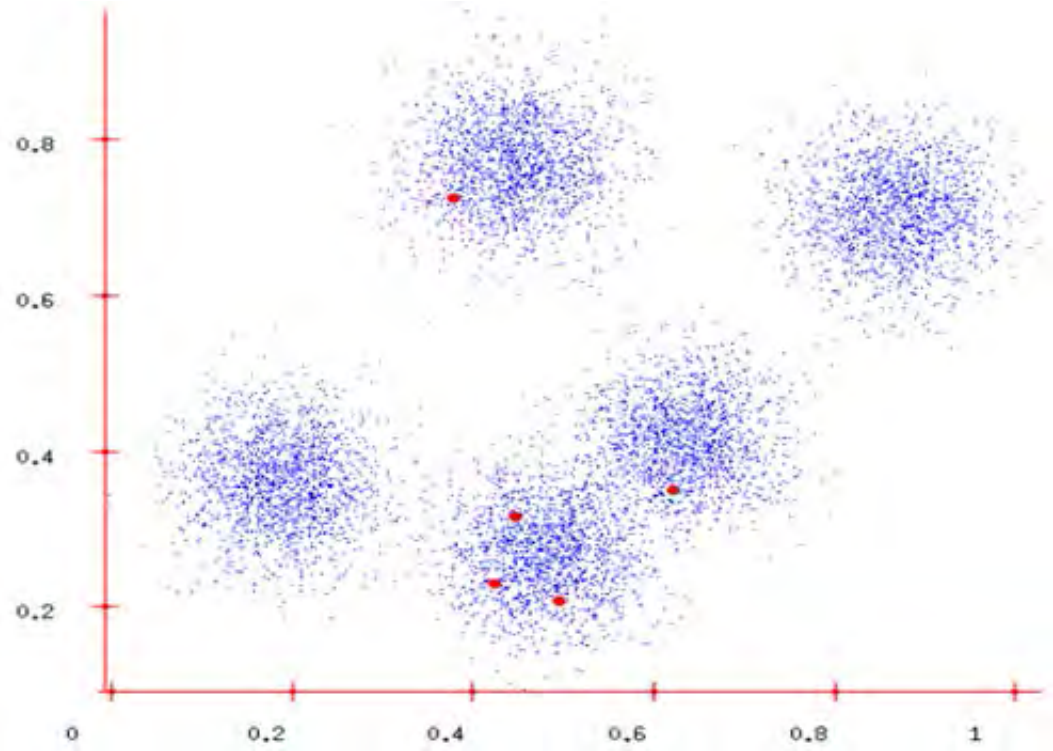
K-means clustering

- Guess a number (k) of clusters (e.g. k=5)



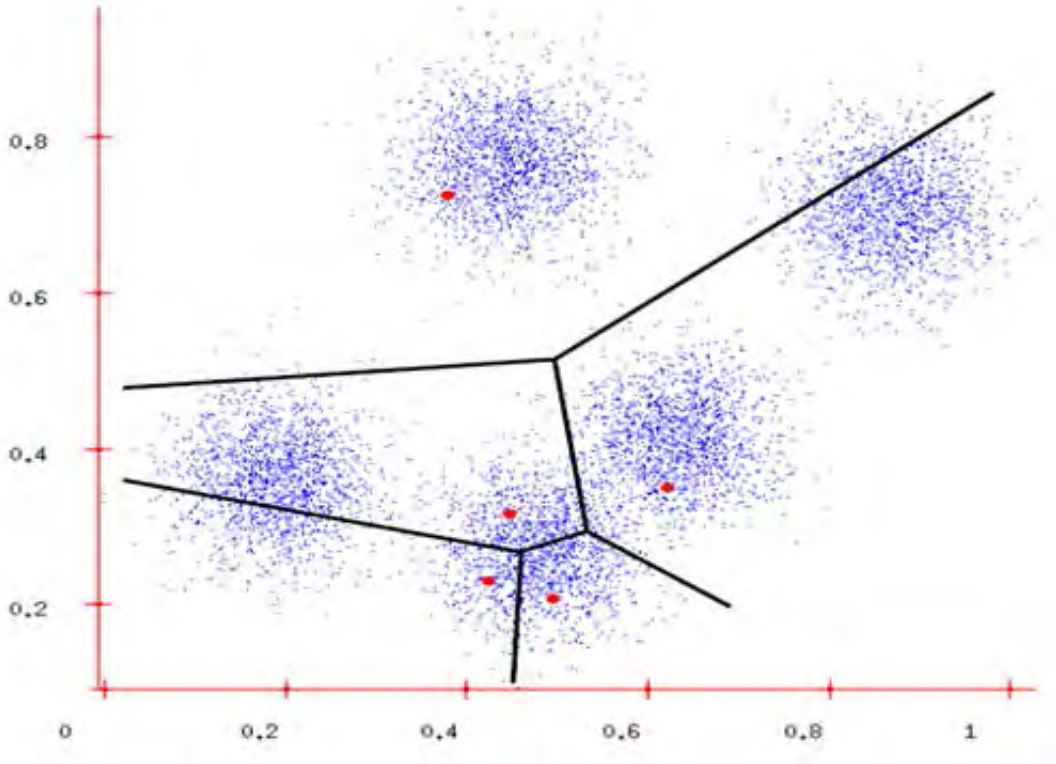
K-means clustering

- Guess a number (k) of clusters (e.g. k=5)
- Randomly guess the k cluster centre locations



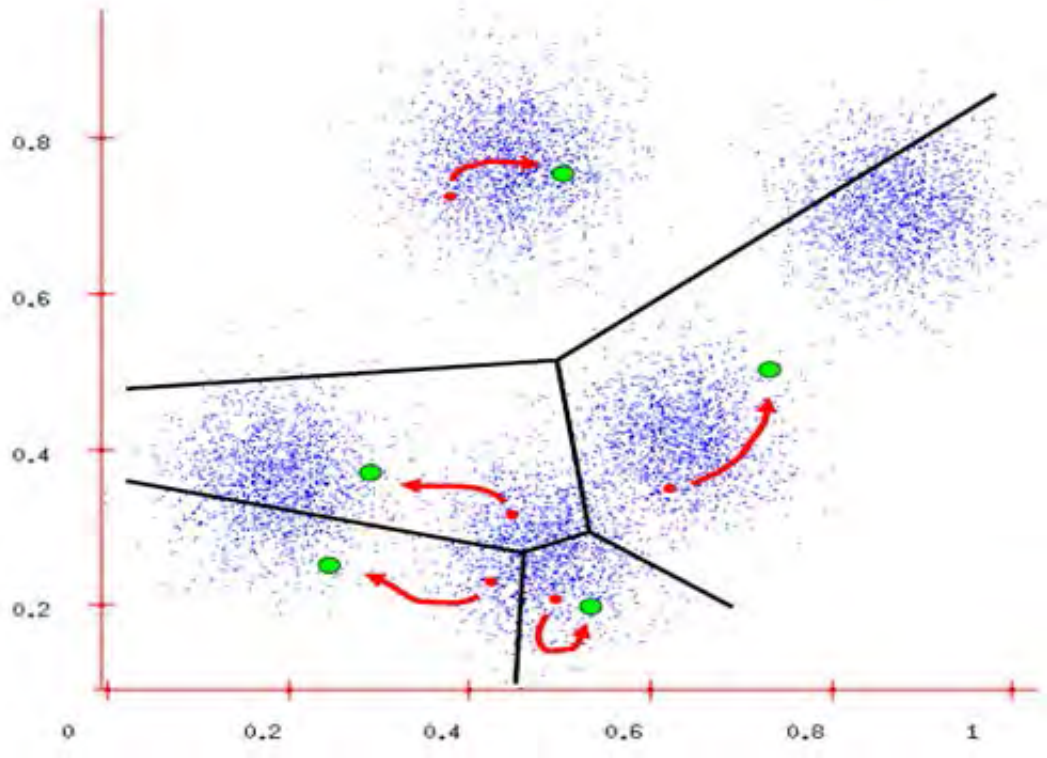
K-means clustering

- Guess a number (k) of clusters (e.g. k=5)
- Randomly guess the k cluste centre locations
- Associate each datapoint with closet centre



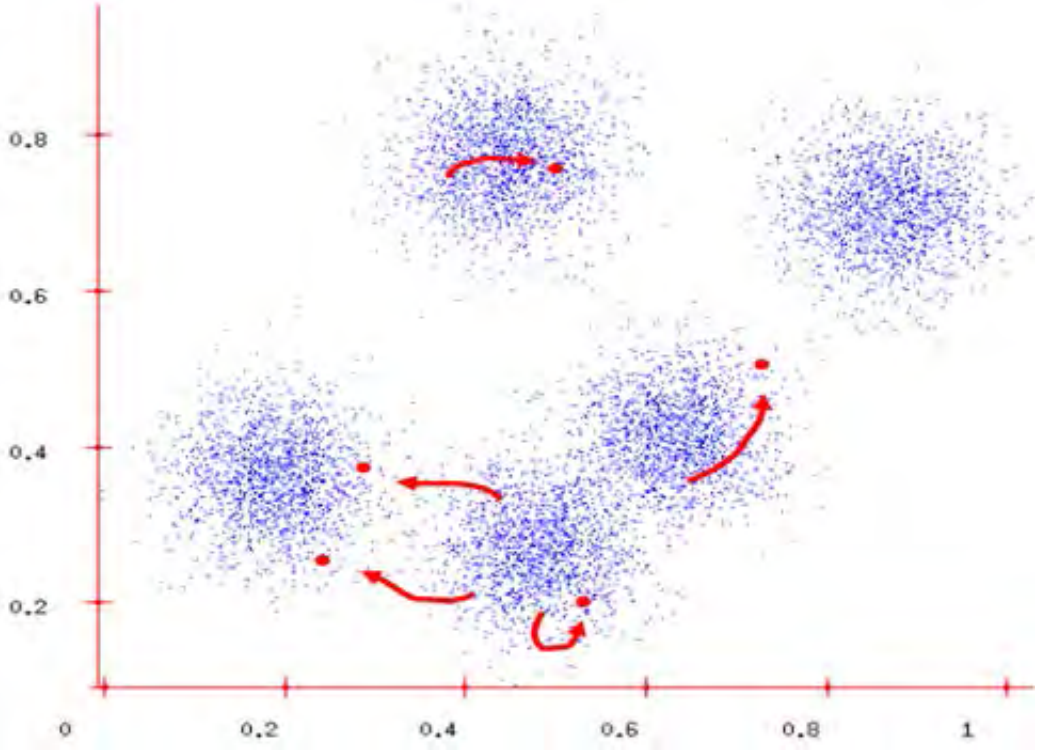
K-means clustering

- Guess a number (k) of clusters (e.g. k=5)
- Randomly guess the k cluster centre locations
- Associate each datapoint with closet centre
- Find the mean point in each region



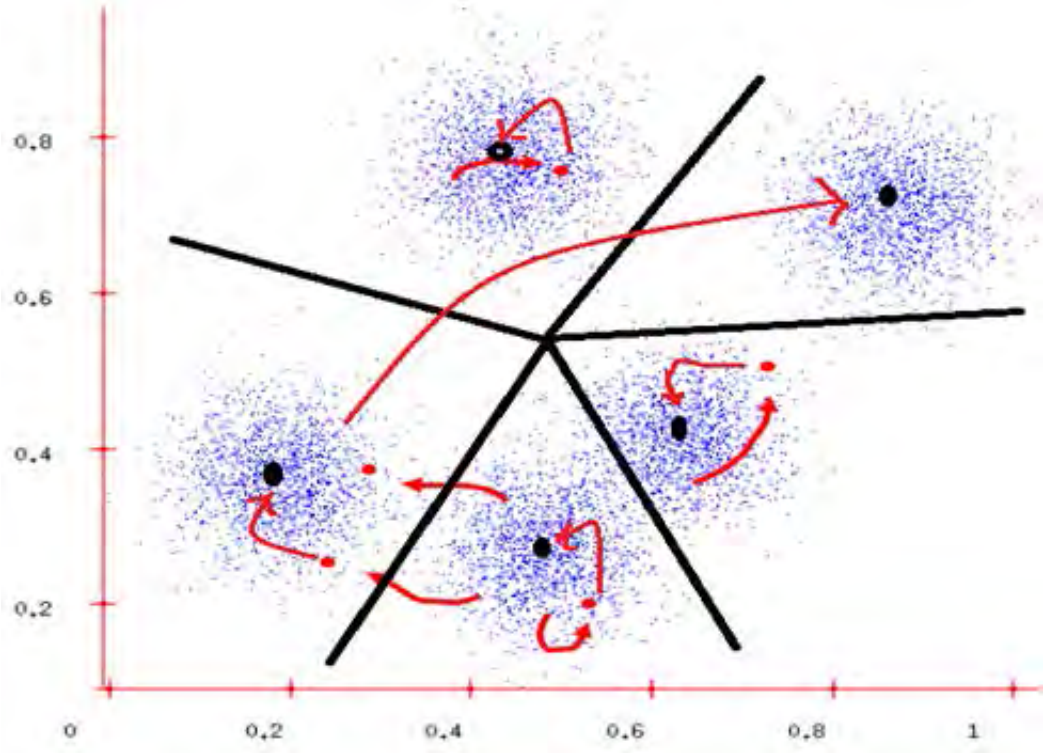
K-means clustering

- Guess a number (k) of clusters (e.g. k=5)
- Randomly guess the k cluster centre locations
- Associate each datapoint with closet centre
- Find the mean point in each region
- … and move the cluster center there



K-means clustering

- Guess a number (k) of clusters (e.g. k=5)
- Randomly guess the k cluster centre locations
- Associate each datapoint with closet centre
- Find the mean point in each region
- … and move the cluster center there
- Repeat until change is small



Choosing the right ML technique?

- There’s no formal solution
    - Are my labels accurate enough?
    - Is the labelling task impossible? - Unsupervised approaches?

- So what do we do with multiple answers?
    - Voting algorithms together
    - Learn which algorithms perform well in what contexts?

Signal Processing
Machine Learning

E.g.: Speech analytics:

- Pitch
- Tone
- Complexity of speech
- Delay time, (Zlochower & Cohn, 1996)

Speech-Based Assessment of PTSD in a Military Population using Diverse Feature Classes

Dimitra Vergyri[1], Bruce Knoth[1], Elizabeth Shriberg[1], Vikramjit Mitra[1], Mitchell McLaren[1], Luciana Ferrer[1,2], Pablo Garcia[1], Charles Marmar[3]

[1]SRI International, Menlo Park, CA
[2]CONICET and University of Buenos Aires, Argentina
[3]NYU Langone Medical Center, Department of Psychiatry, New York, NY
{dimitra.vergyri,bruce.knoth,elizabeth.shriberg,vikramjit.mitra,mitchell.mclarren}@sri.com,
lferrer@dc.uba.ar, pablo.garcia@sri.com, Charles.Marmar@nyumc.org

Abstract

There is a critical need for detection and monitoring of Post-Traumatic Stress Disorder (PTSD) in both military and civilian populations. Current diagnosis is based on clinical interviews, but clinicians cannot keep up with the growing need. We examined the feasibility of using speech for assessment in a military population. We analyzed recordings of the Clinician-Administered PTSD Scale (CAPS) interview from military personnel diagnosed as PTSD positive versus negative. Three feature types were explored: frame-level spectral features, longer-range prosodic features, and lexical features. Results using gaussian backend, decision tree and neural network classifiers (for spectral and prosodic features) and boosting (for lexical features) showed an accuracy of 77% correct in split-half cross validation experiments, a figure significantly above chance (which was 61.5% for our dataset). Spectral and prosodic features outperformed lexical features, and feature combination yielded further gains. An important finding was that sparser prosodic features offered more robustness than acoustic features to channel based variation in the interview recordings. Implications and future work are discussed.

self-reporting, which may be degraded by distortions in memory and self-perception [5], or financial and social incentives [6]. Moreover, the interview requires a visit to the clinician's office, which some patients may not be willing or able to make. Thus, there is a need for a more objective, cost- and time-efficient means of PTSD assessment.

In this paper, we use automatic analysis of the patient's speech to determine PTSD status. Speech is natural, noninvasive, cheap, and can be obtained via phone for analysis at a distance. This work builds on prior research that focused on using speech signal to detect mental health [7] and emotion. For emotion detection, previous research used prosodic features such as speaking rate, pitch, energy or intensity, and pause duration [8-13] as well as other acoustic features such as voice quality [10,11], spectral features [8] and Mel frequency cepstral coefficients (MFCCs) [12].. Changes in prosodic features (pitch, energy, speaking rate), spectral features (formants, their corresponding bandwidths, power spectral density, spectral tilt), and MFCCs were also found useful in depression detection [14-19] and also recent PTSD-detection work [20, 21].

2. Data