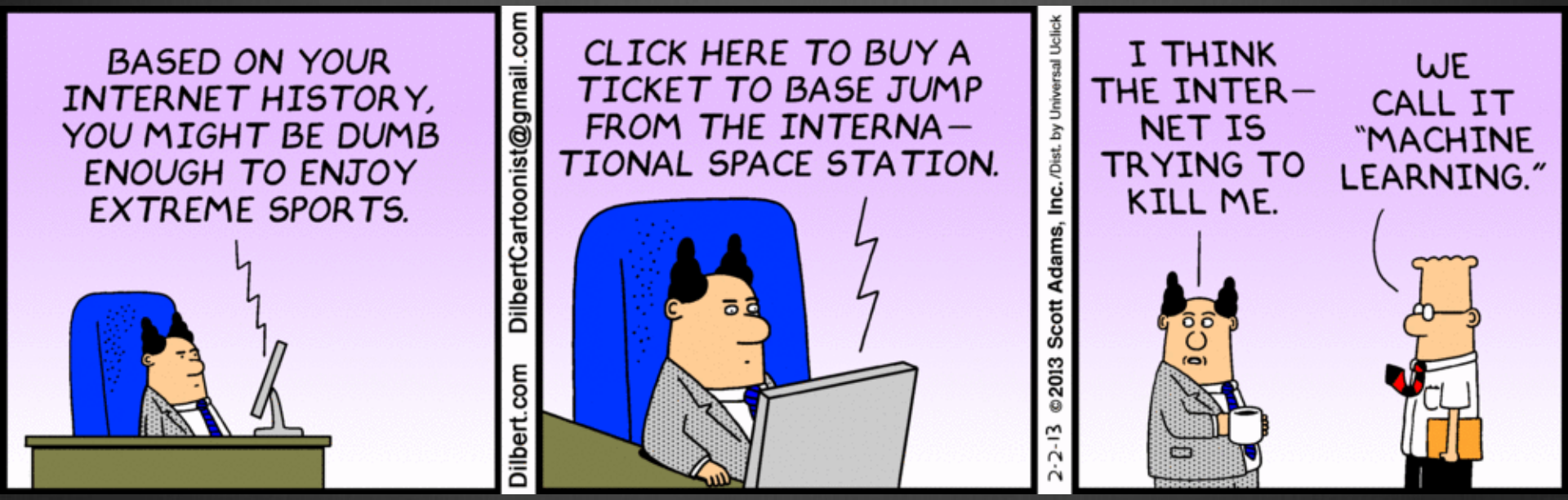
Getting started
Unleash the computer on your data

Amber McKenzie

- Create a Kaggle login: https://www.kaggle.com/
- Download github repo https://github.com/ab6/Codemash2017-UnleashData.git
  - Download from the CodeMash mirror: Codemash2017-UnleashData-master.zip
- We will be working through the Kaggle site so installing the requirements locally is optional.
  - (There are read-only versions of the notebooks if internet troubles crop up)



- GitHub code and readme download:
  - https://github.com/ab6/Codemash2017-UnleashData.git

Who am I?

dialogtech

UNIVERSITY OF SOUTH CAROLINA

C B M

University of South Carolina

"Doc Am"

COMPUTATIONAL DATA ANALYTICS GROUP

AT THE OAK RIDGE NATIONAL LABORATORY

Agenda

- ML applications
- Digit recognizer
  - Data and problem identification
  - Feature identification and data representation
  - Algorithm selection
  - Evaluating models
- Housing prices
  - Data scrubbing and normalization
- Sentiment in movie reviews
  - Basic NLP/text classification
- Titanic
  - Application of concepts

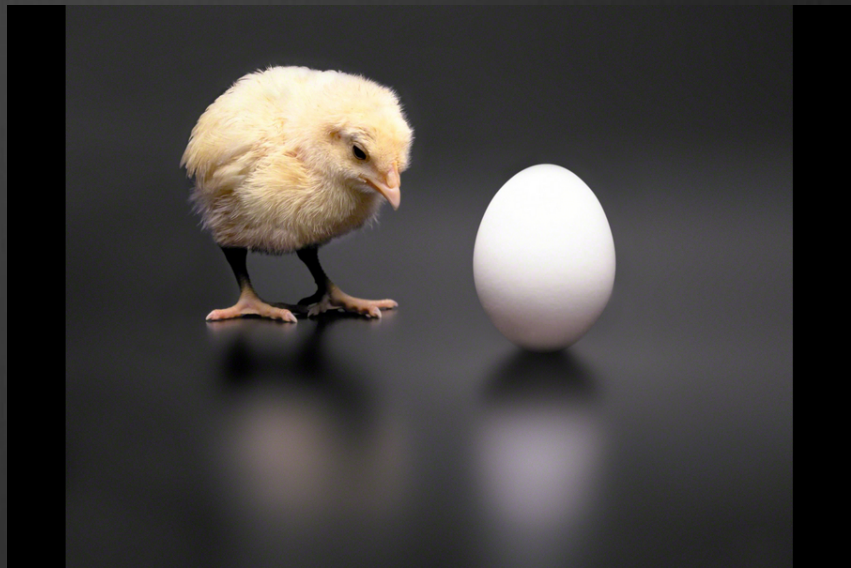
- GitHub code and readme download:
  - https://github.com/ab6/Codemash2017-UnleashData.git

Applications of Machine Learning

- Machine perception
- Computer vision
- object recognition
- Natural language processing
- Syntactic pattern recognition
- Search engines
- Medical diagnosis
- Bioinformatics
- Brain-machine interfaces
- Cheminformatics
- Detecting credit card fraud
- Stock market analysis
- Classifying DNA sequences
- Sequence mining
- Speech and handwriting recognition
- Game playing
- Software engineering
- Adaptive websites
- Robot locomotion
- Computational advertising
- Computational finance
- Structural health monitoring
- opinion mining
- Affective computing
- Information retrieval
- Recommender systems

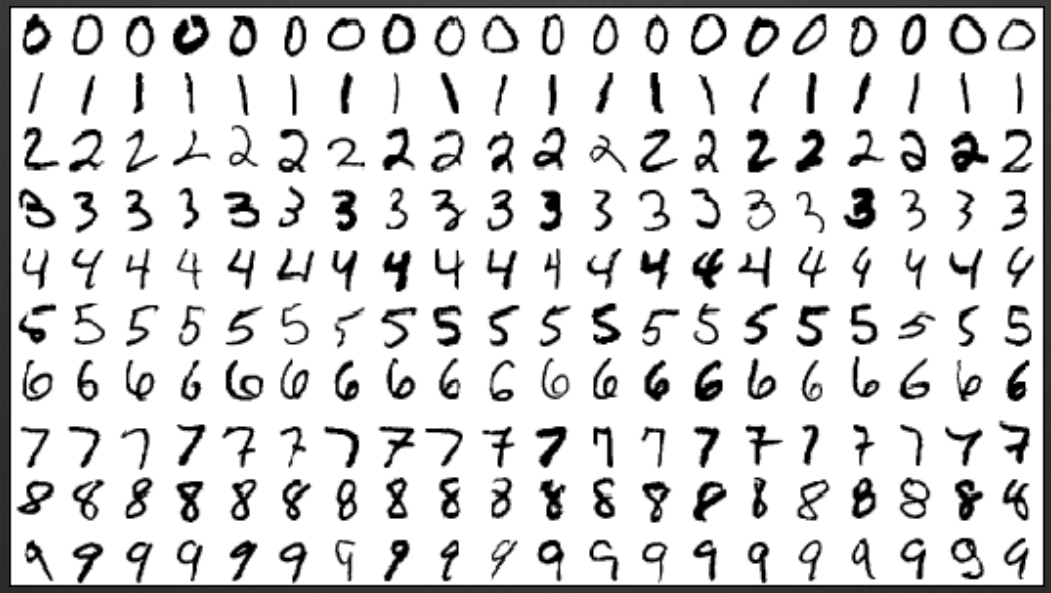
Which comes first: the data or the problem?

- Considerations for each
  - Need the data to address the problem
  - Need a problem that can be addressed by the data
- How do you identify the problem?
- Where do you get data?



Digit recognizer

- https://www.kaggle.com/c/digit-recognizer
- Classify handwritten digits
  - "The goal in this competition is to take an image of a handwritten single digit, and determine what that digit is."

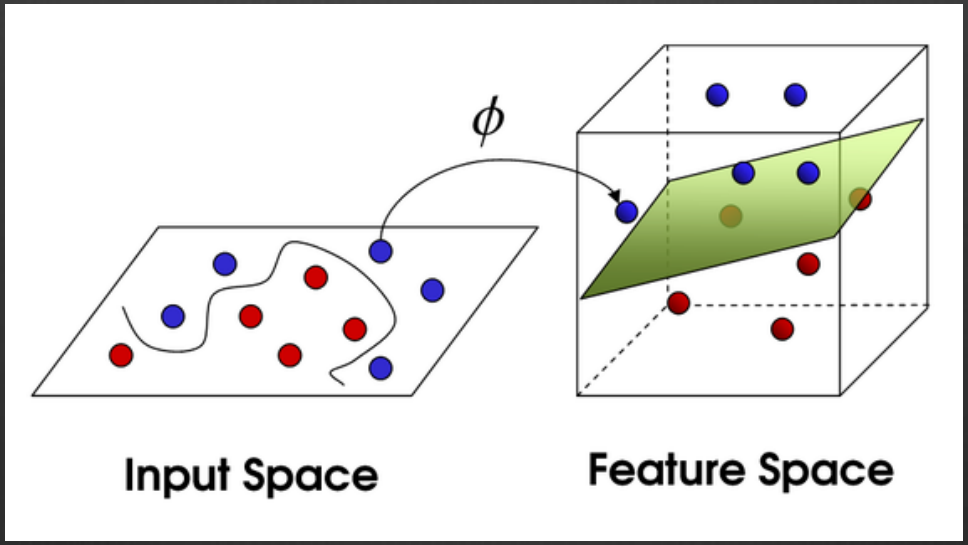


Feature identification and data representation

- Data exploration
- Data representations
  - Numerical
  - Text-based
- Splitting train and test sets
- Validation set
  - Terminology interchangeable when you don't have validation set
  - Some models use validation set in model building process

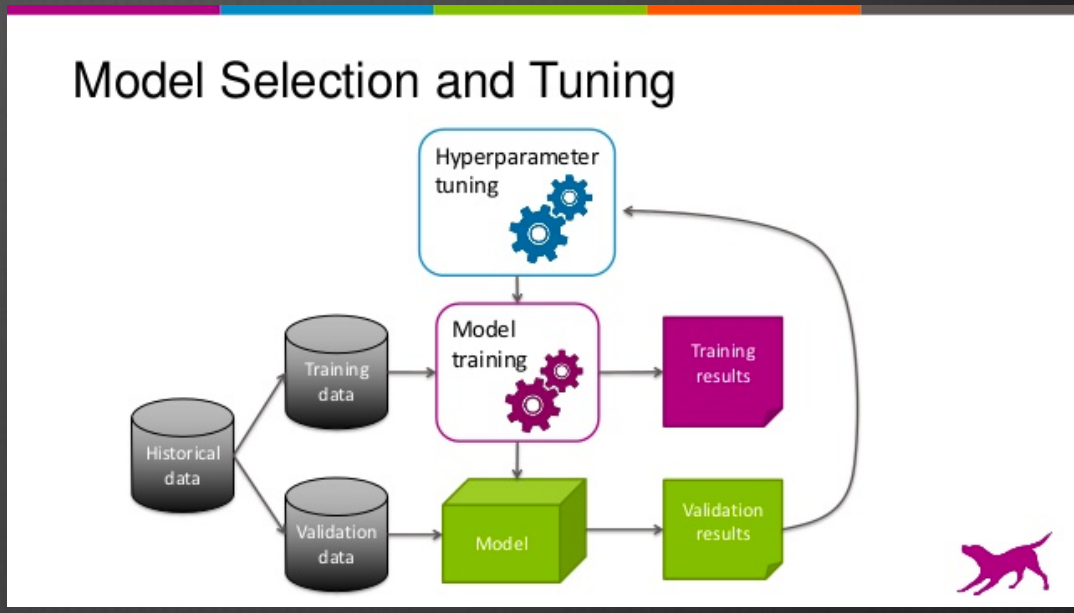
Algorithm selection

- Classification
  - KNearestNeighbor
  - Naïve Bayes
  - Support Vector Machines
- Prediction
  - Regression: linear, ridge, etc.
  - Gradient boosting
  - Random Forest



Evaluating models

- Scoring metrics
  - Gold standard
  - Classification
    - Precision and recall
    - F-measure
  - Linear prediction
    - Mean squared error
- Cross validation

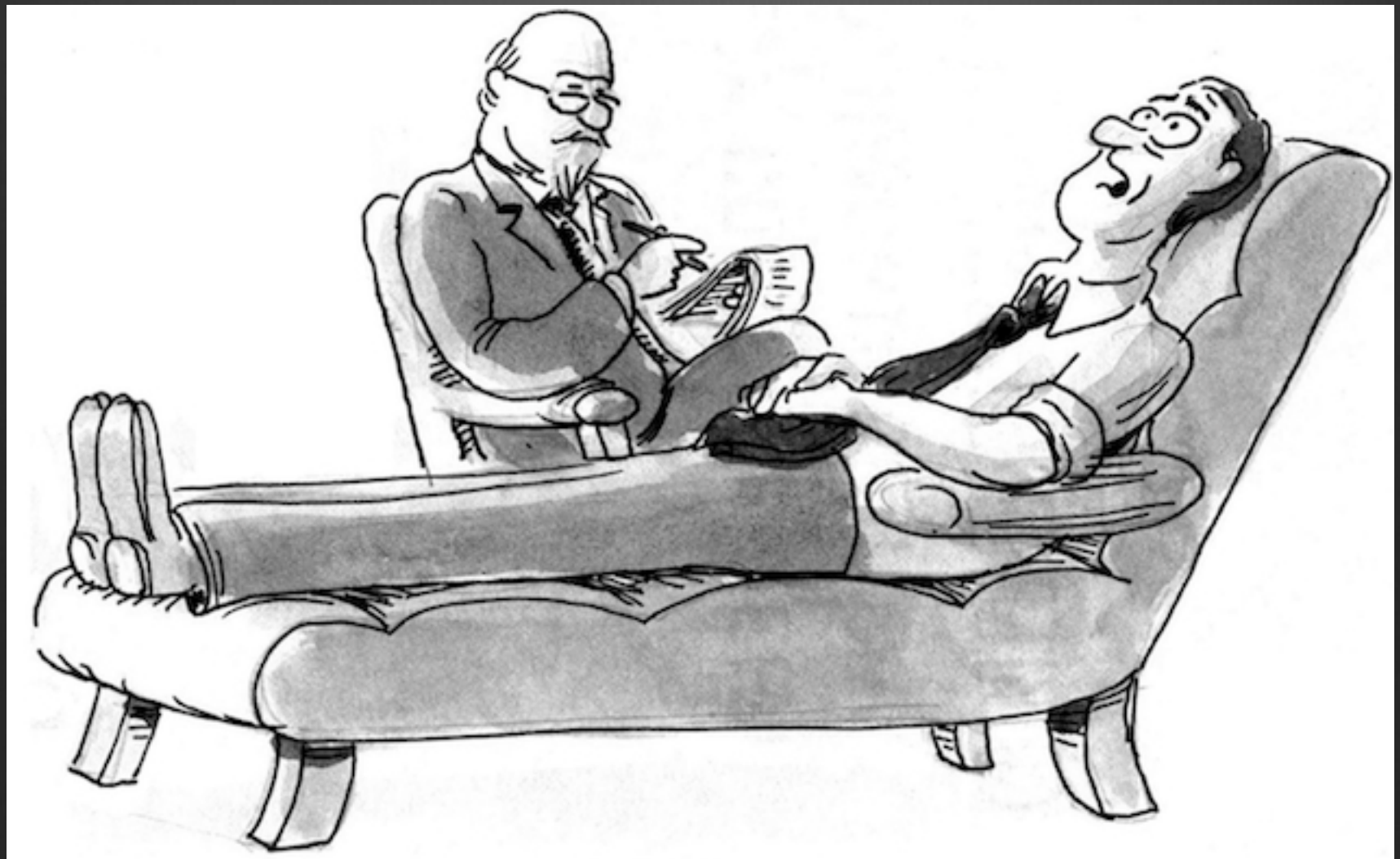


Housing prices prediction

- https://www.kaggle.com/c/house-prices-advanced-regression-techniques
- “With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.”

Data scrubbing and normalization

- Need to do preprocessing on all data: important if you are given a separate test set
- Numerical data
  - Null values
    - Imputer: fill in with mean, median or most frequent values
    - Sometimes dropping rows or using marker values is more effective
  - Scaling: Standardize features by removing the mean and scaling to unit variance
- Categorical data: expand out columns into binary features
- PCA: Reduce size of matrix without losing information

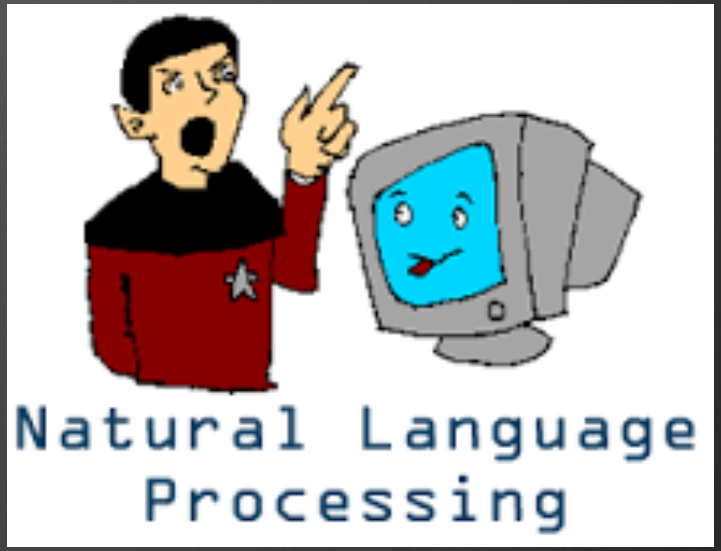


"My profession has probably been transformed again just since we started this session."

Sentiment in movie reviews

- https://www.kaggle.com/c/word2vec-nlp-tutorial
- "The labeled data set consists of 50,000 IMDB movie reviews, specially selected for sentiment analysis. The sentiment of reviews is binary, meaning the IMDB rating < 5 results in a sentiment score of 0, and rating >=7 have a sentiment score of 1."

NLP



- Data representation
  - Bag of words
  - Frequency count
  - TF-IDF
  - Word2vec
- Feature extraction/text manipulation
  - Stop words
  - Stemming

Titanic practice problem

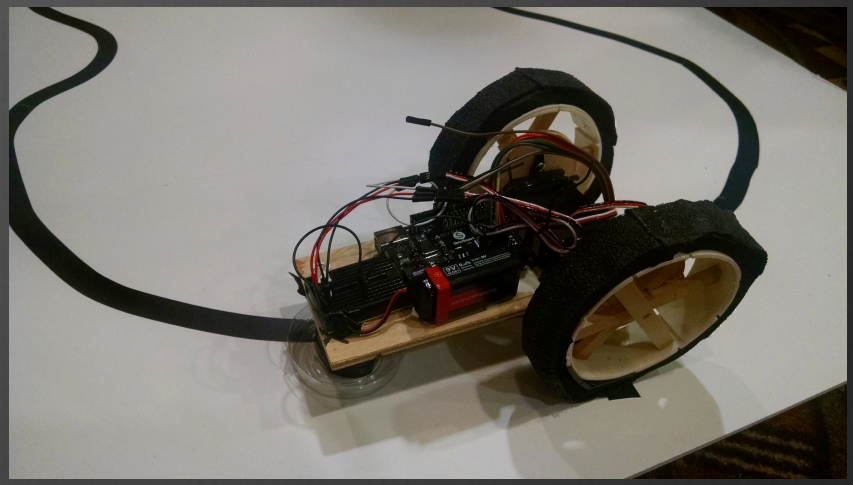
- https://www.kaggle.com/c/titanic
- ” In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.”

Contact information

Amber McKenzie, Ph.D.
Data Scientist

DialogTech
www.dialogtech.com

mckenzie.amber@gmail.com
amber.mckenzie@dialogtech.com
https://nlprunner.wordpress.com



Accuracy

- Gold standards
  - Sampling of data that has been categorized by a human expert
- Inverse is classification error
- Not good when a large portion of documents are not relevant: can just not return any and get a good score

$$\text{accuracy} = (tp + tn)/(tp + fp + fn + tn)$$

Precision and recall

- Recall: how many results that we were supposed to find did we find
  - Bad on its own because you can get 100% by returning all of them, i.e. bad to measure classes that have most of the data in them
- Precision: how many of the results that we found were right
  - Bad on its own because you can get good results by only returning one correct result

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

F-measures

- F measure is a means of representing both precision and recall.
  - **β** < 1 emphasizes precision
  - **β** > 1 emphasizes recall

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$