SIOP 2022 ML Friday Seminar

Dr. Scott Withrow, Infor Talent Science

Dr. Rachel T. King, Modern Hire

Dr. Isaac Thompson, Modern Hire

ANNUAL CONFERENCE

2022 | SIOP

SEATTLE ◆ APRIL 28-30

ANNUAL CONFERENCE

2022 | SIOP

SEATTLE ◆ APRIL 28-30

Agenda

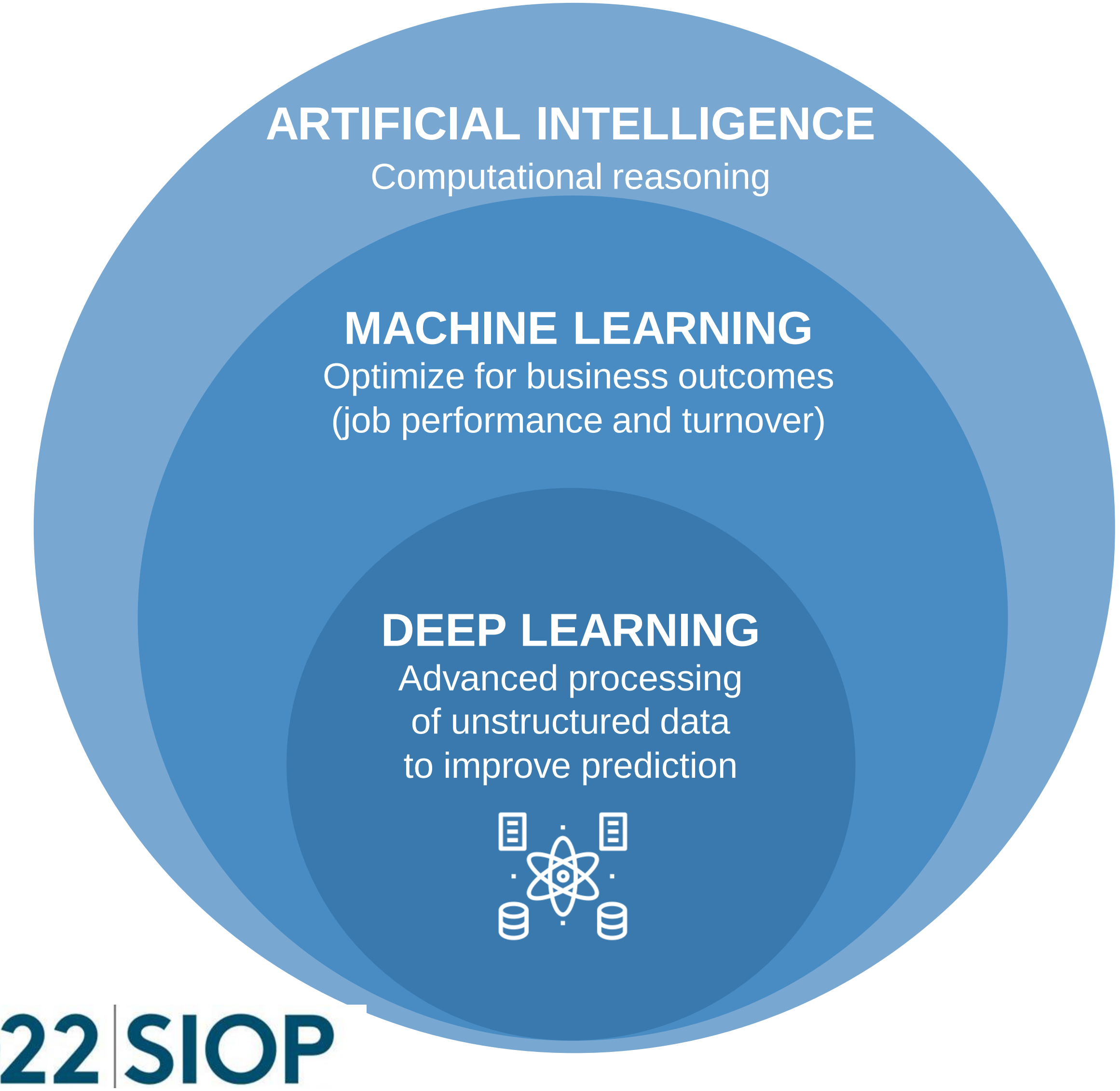
- Overview of Machine Learning
- Data
- Tools
- Break
- Practical Examples
    - Supervised Example
    - Unsupervised Example
- Group Discussion
- Q&A

Machine Learning Overview

What is Machine Learning

- A process where data is transformed into an output.
- The process is a statistical model, series of models, or computer code.
- The output is typically a decision.
- ML is sometimes recursive and can take decisions and consequences and use them as data.
  - Data -> processing -> prediction -> decision -> results/evaluation -> data
  - This is how ML models 'learn'
- ML doesn't have to be recursive and ML models in high stakes settings such as hiring are often frozen to a set of training data and only updated under SME supervision.

Terminology



ARTIFICIAL INTELLIGENCE
Computational reasoning

MACHINE LEARNING
Optimize for business outcomes
(job performance and turnover)

DEEP LEARNING
Advanced processing
of unstructured data
to improve prediction

- Artificial Intelligence (AI)
  - Enabling computers to act, think, and solve problems like a human.
  - May involve external sensors or some way to interact with the world.
  - Taking input and translating it to solutions is called ‘modeling.’
  - A chatbot or social media monitor are common AI tools.
- Machine Learning (ML)
  - A subset of AI
  - The models that AI uses to make decisions are usually machine learning models.
- Deep Learning (DL)
  - A subset of ML
  - Rather than defining the consequence of decisions we give the machine reference material and let the algorithm learn consequence on its own.

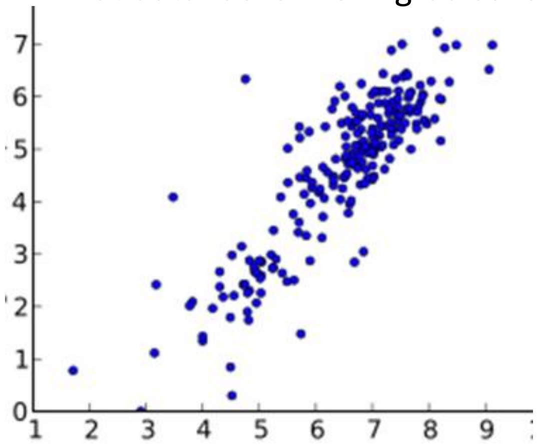
Terminology

Machine Learning Born out of computer science; focused on prediction	Statistical Learning Born out of statistics; focused on theory & interpretation
Algorithm (i.e. Logistic Regression)	Model (i.e. Logistic Regression)
Feature	Predictor/Variable
Label/Target	Criterion/Dependent Variable
Example	Case/Subject
Learning	Estimation

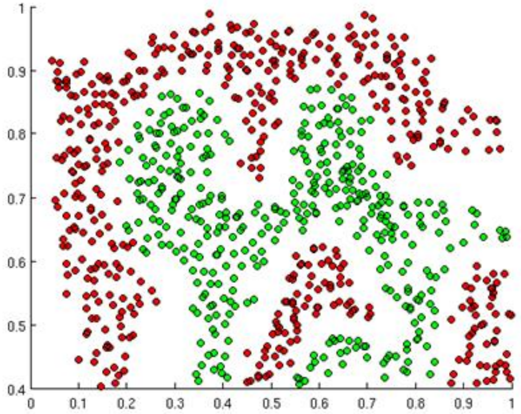
Why Use Machine Learning

- Deals with complex data better.
  - Multiple dependent variables.
  - Uncertain relationships between independent variables.
- Solves some problems traditional modeling simply can’t approach.
  - ○ Non-linear relationships in data
  - ○ Balancing prediction with adverse impact
  - ○ Helping with data issues
  - ○ Avoiding overfitting
- Aggregates massive datasets into human understandable snippets.
- May help avoid over-reliance on p-value based statistics.
- ML is another tool in your toolbox, and may not always be the best solution to every problem

What data looks like in grad school:

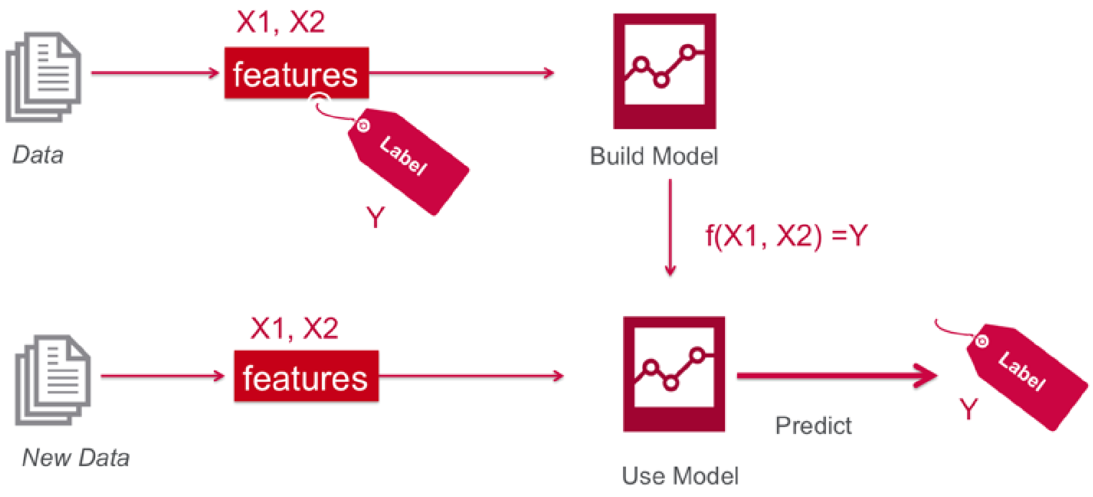


What real data sometimes looks like:



Supervised vs Unsupervised Machine Learning

Supervised



X1, X2

features

Label

Y

Data

Build Model

f(X1, X2) =Y

X1, X2

features

New Data

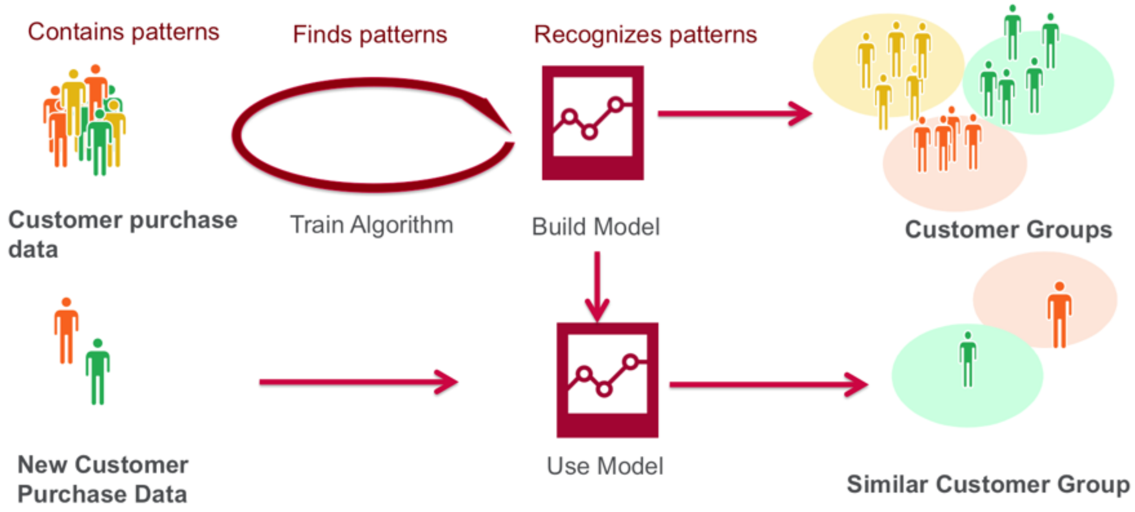
Use Model

Predict

Label

Y

Unsupervised



Contains patterns

Finds patterns

Recognizes patterns

Customer purchase data

Train Algorithm

Build Model

Customer Groups

New Customer Purchase Data

Use Model

Similar Customer Group

Types of Machine Learning Models

Supervised

- The algorithm is trained on labeled data.
- Regression and classification
- Common Models
  - Regression
    - Linear regression
    - Ridge regression
    - Lasso regression
  - SVM (support vector machine)
  - Naïve Bayes
  - Random Forest
  - KNN (K-Nearest Neighbor)

Unsupervised

- Works with unlabeled data.
- Abstracts the relationship between data points without human input.
- Common Models
  - Dimensionality Reduction
  - Clustering
  - Some types of NLP

Data

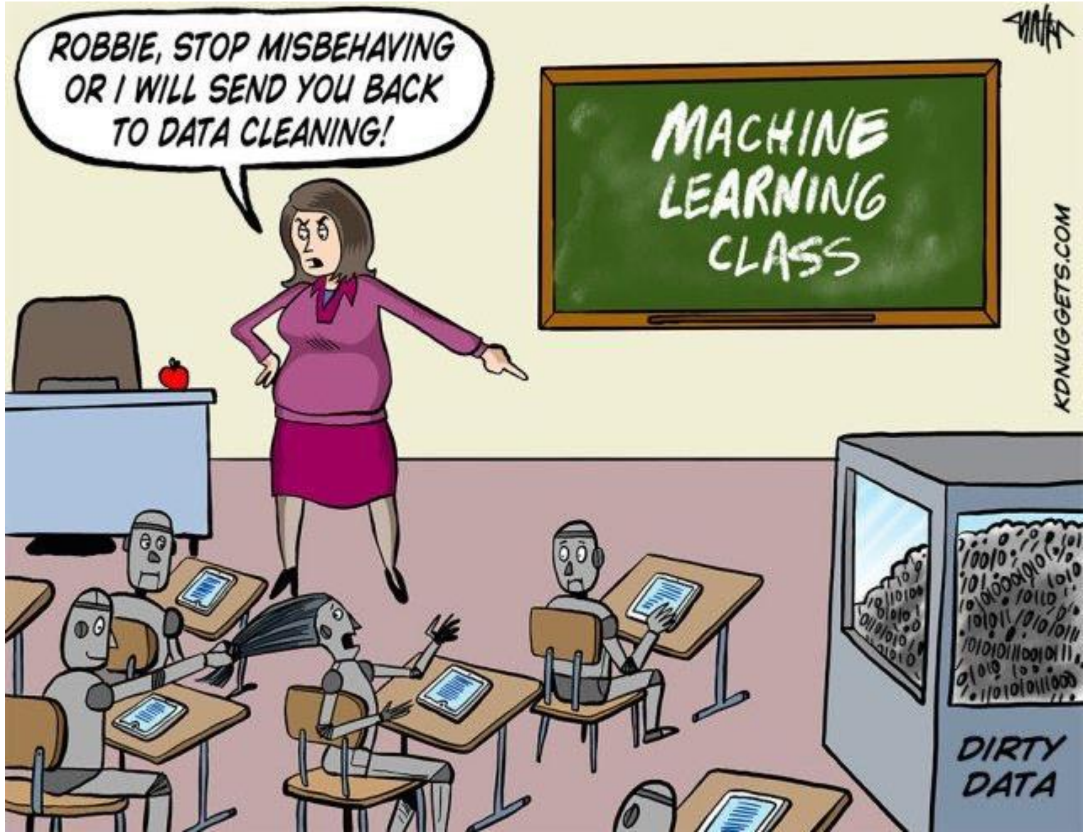
Understanding Datasets

- Sample sizes

- Does the data you’re using allow you to actually answer your question?

- You still need to clean and understand your data

- Is your data representative?

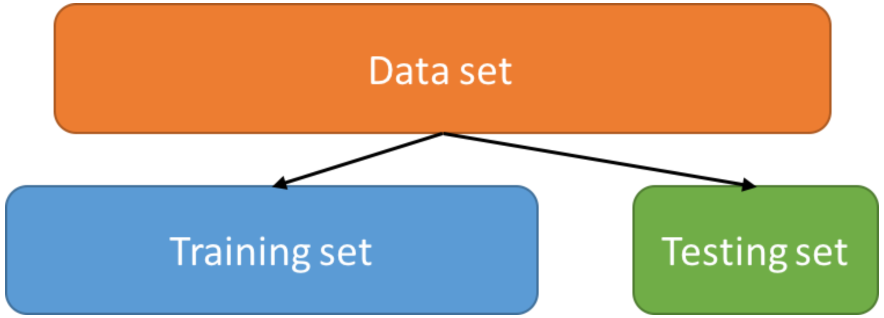

ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

Train/Test Split

Machine Learning tries to mimic the world rather than explain it.

Thus, the Train/Test Split method is used to ensure models generalize

- Provide the algorithm with ~75% of the data to learn from. Then use the algorithm to predict unseen remaining ~25%.



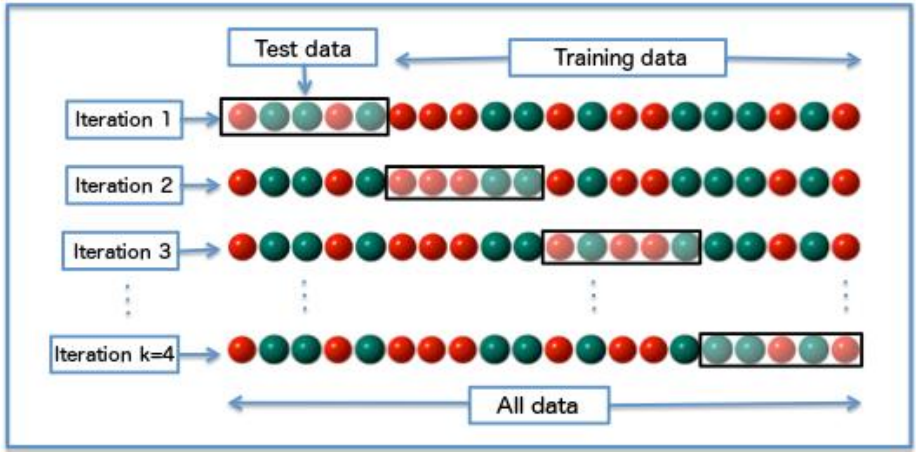
Data set

Training set

Testing set

Another common method used to mimic a train test split is referred to as k-fold cross-validation.

- Provide the algorithm with k-1 folds of the data to train on. Then repeat k times. The final algorithm parameters are the average of k.



Cross Validation

- Resampling methods that help us assess the generalizability of our machine learning models.
- Reduces or flags overfitting and selection bias.
- K-Fold
  - Randomly partition data into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. This is repeated until all the samples have been used as validation once.
- Monte Carlo (Repeated Random Sub-Sampling)
  - Randomly split the dataset into training and validation data. For each split, the model is fit to the training data, and predictive accuracy is assessed using the validation data. The number of splits to perform is up to the user.

Feature Engineering

- The process of transforming and combining raw variables using domain knowledge before running a machine learning model.

- Example: Taking responses to individual items on an assessment and combining them to make scales.

Building your model

- Pick your model
    - What problem are your trying to solve?
    - How much data do you have?
    - How much complexity are you comfortable with?
    - Trial and error
- Hyperparameter tuning
- Ensemble modeling

Common Data Issues and Best Practices

- The more representative your data are the better.
- Have a solid analysis plan before you begin collecting data and use it to guide *how* you collect those data.
- The more automatic your data entry is the fewer problems you will have using those data.
  - For example, you could have a fill in the blank question for age or provide a calendar style entry. The second option will reduce data entry errors.
- Garbage in – Garbage out

Machine Learning Tools

R

Pros

- Designed for statistics.
  - Vectorized by nature
  - IRT
- Tons of great packages.
- Easy to learn.
- Free to use.
- Extensions like RMarkdown and Shiny make dashboards and reports easy.

Cons

- Slow execution.
- Limited parallelization options.
- Fewer options for Machine Learning and Deep Learning than Python.
- Limited use outside of statistics.

Python

Pros

- Tons of great packages.
- Faster than R.
- Many packages dedicated to machine learning.
- You can use Python for other things besides statistics.

Cons

- Reliant on packages for all statistical programming.
- Doesn't have packages like tidyverse
- More difficult to learn than R
- More difficult to reconcile versions of python and packages.
- May be fully integrated with your OS (Linux) so different versions can be very difficult.

IDEs (Integrated Development Environment)

Programming Languages vs. SPSS

Programming Languages

- Open source
- Latest innovations
- Coding (syntax) required
- Multiple ways to solve a problem

SPSS

- Proprietary software
- Point and click interface
- Syntax is program specific
- Limited number of analyses available

Vendor Services

- Drag and drop style machine learning services.
- Usually comes with dashboards and common models pre-built and ready for use.
- Common vendors:
  - Microsoft Azure
  - IBM Watson
  - Amazon Web Services (AWS)

Additional Resources and Tools

- Additional resources and tools are listed on our our GitHub for this seminar
- https://github.com/izk8/2022_siop_fri_seminar

Q&A

Practical Examples of Machine Learning

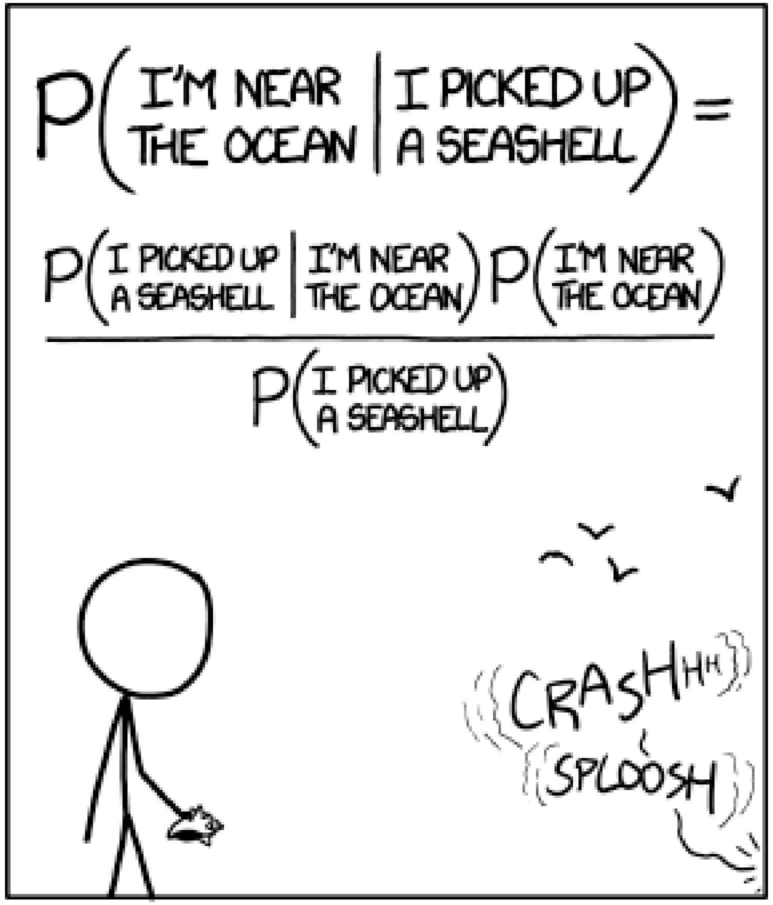
Machine Learning in Selection

- Bayesian regression of personality and cognitive ability on job performance.

- Utilizes Naïve Bayesian Inference

What is Bayesian Statistics

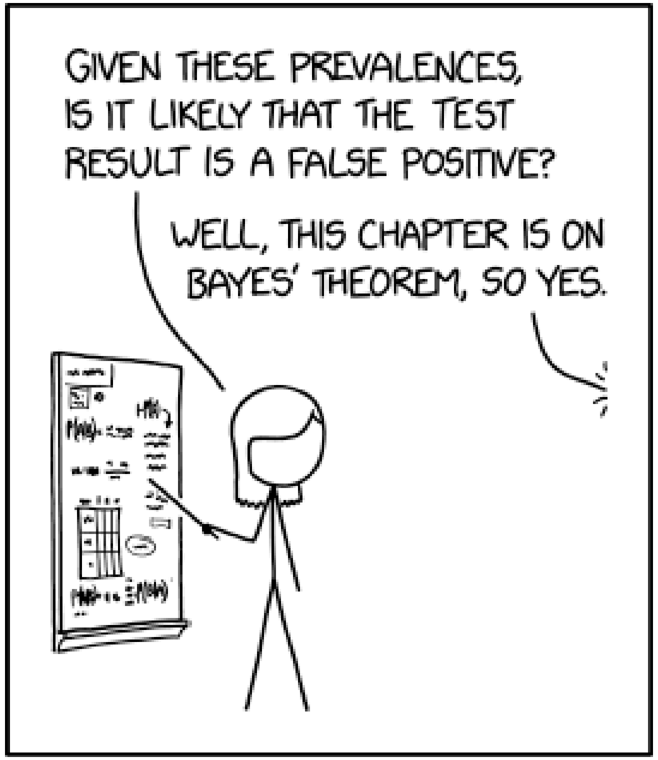
- A family of statistics that assume what has happened before is likely to happen again.
- Based on Bayes' Theorem
- $P[A|B] = \frac{P[A \text{ and } B]}{P[B]} = P[B|A]\frac{P[A]}{P[B]}$
- The conditional probability of A given B is the conditional probability of B given A scaled by the relative probability of A compared to B.


P(I'M NEAR THE OCEAN | I PICKED UP A SEASHELL) =
P(I PICKED UP A SEASHELL | I'M NEAR THE OCEAN) P(I'M NEAR THE OCEAN) / P(I PICKED UP A SEASHELL)
CRASHHH
SPLOOSH

STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

Why does Bayes' Theorem Matter?

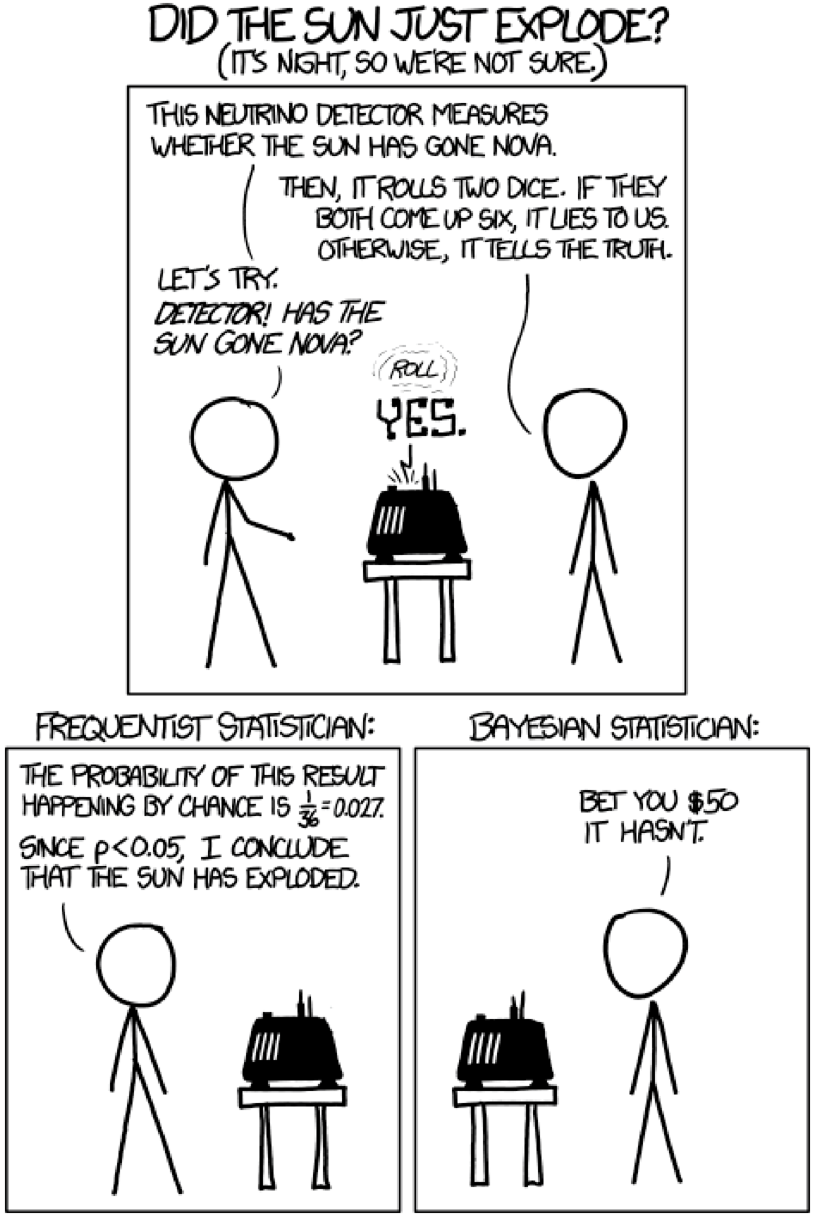
- Past performance is the best predictor of future performance.
- Life is full of patterns. Let's consider predicting simple behavior.
  - I am off to the store to buy ice cream.
  - My favorite ice cream is Tin Lizzy.
  - What ice cream do I buy?


GIVEN THESE PREVALENCES, IS IT LIKELY THAT THE TEST RESULT IS A FALSE POSITIVE?
WELL, THIS CHAPTER IS ON BAYES' THEOREM, SO YES.

SOMETIMES, IF YOU UNDERSTAND BAYES' THEOREM WELL ENOUGH, YOU DON'T NEED IT.

What Isn’t Bayes?

- Frequentist inference is the most common form of statistics we use in psychology.
  - You are likely used to seeing things like p-values, confidence intervals, and error bands.
  - The underlying assumption of frequentist inference is the use of multiple repetitions of the same experiment to derive the probability of the results not being chance.
  - New data is always subjected to the probability of being incorrect with no ability to ever show it was incorrect.
    - I.e., You can never prove the null hypothesis correct.
- With Bayesian inference we take what we know and update our assumptions based on new data to calibrate the probability of being true.


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027.
SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Naïve Bayesian Inference

- Under the hood the same theorem is at play.
- However, this time we want to add in a whole bunch of predictors (features if we want to use machine learning terminology) and the outcome is a *classification.*
  - Classifications are groups, bins, silos, etc. that describe something.

$$P(y|x_1,\dots,x_n) = \frac{P(x_1,\dots,x_n|y)\cdot P(y)}{P(x_1,\dots,x_n)} = \frac{P(x_1|y)\cdot P(x_2|y)\cdot\ldots\cdot P(x_n|y)\cdot P(y)}{P(x_1,\dots,x_n)}$$

- For example, we could add any number of features (as above) in order to see how likely someone would be to fall into a category.
  - The model assumes that the predictors are independent which is a rather **naïve** assumption.

Key Takeaway

- Bayes' is a process where we update known information with new information.
- We can use machine learning models to make decisions well when we have lots of predictors.
- Comparing different machine learning models can be very easy with the correct packages.

Q&A

Break Time

10 Minutes

Unstructured & Unsupervised Data Analytics

Overview

Key concepts

- messy data, today's example data set

Job title example (sbert)

- Embedding a large set of text and querying it

Job description example (berTopic)

- Embedding job descriptions
- Clustering, exploring, pruning clusters, seeding w/ labels, seeding w/ raw text

Key takeaways

Messy Data

Unstructured

- Open-ended, e.g. interview, resume, qualitative data

Unsupervised

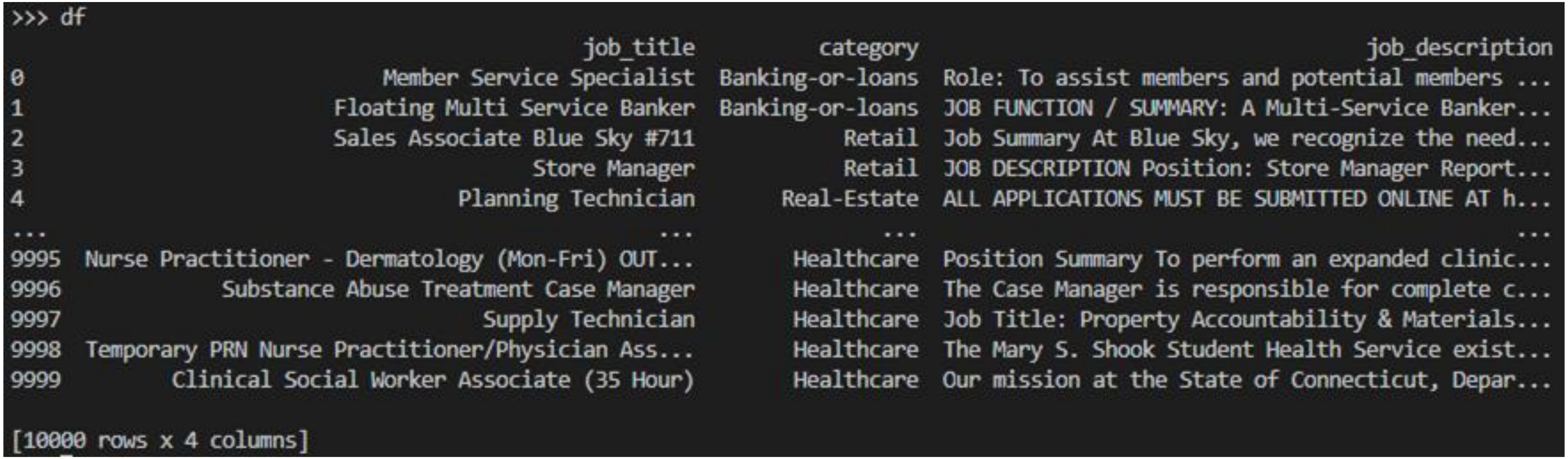
- No labels, no local criteria to be maximized for, e.g. google search

Semi-supervised

- Some labels

Today’s Example Data

- **Job data**
- **10k rows**
- **3 Columns: job titles, job descriptions, categories**



```
>>> df
                                        job_title          category                                  job_description
0                     Member Service Specialist  Banking-or-loans  Role: To assist members and potential members ...
1                 Floating Multi Service Banker  Banking-or-loans  JOB FUNCTION / SUMMARY: A Multi-Service Banker...
2                 Sales Associate Blue Sky #711            Retail  Job Summary At Blue Sky, we recognize the need...
3                                 Store Manager            Retail  JOB DESCRIPTION Position: Store Manager Report...
4                          Planning Technician       Real-Estate  ALL APPLICATIONS MUST BE SUBMITTED ONLINE AT h...
...                                         ...               ...                                              ...
9995  Nurse Practitioner - Dermatology (Mon-Fri) OUT...        Healthcare  Position Summary To perform an expanded clinic...
9996        Substance Abuse Treatment Case Manager        Healthcare  The Case Manager is responsible for complete c...
9997                          Supply Technician        Healthcare  Job Title: Property Accountability & Materials...
9998  Temporary PRN Nurse Practitioner/Physician Ass...        Healthcare  The Mary S. Shook Student Health Service exist...
9999       Clinical Social Worker Associate (35 Hour)        Healthcare  Our mission at the State of Connecticut, Depar...

[10000 rows x 4 columns]
```

Data Structure Cont.

Python “dictionary”

- ['df', 'category_names', 'all_cat_indicies', 'labels_to_add', 'indices_to_add', 'y']

Category names

- ['Banking-or-loans', 'Retail', 'Real-Estate', 'Law-Enforcement-or-security', 'Administrative', 'Arts-or-entertainment-or-publishing', 'Customer-Service', 'Computer-or-internet', 'Human-Resources', 'Healthcare', 'Sales', 'Insurance', 'Manufacturing-or-mechanical', 'Construction-or-facilities', 'Hospitality-or-travel', 'Restaurant-or-food-Service', 'Transportation-or-logistics', 'Education-or-training', 'Telecommunications', 'Legal', 'Accounting-or-finance', 'Non-profit/volunteering', 'Engineering-or-architecture', 'Pharmaceutical/bio-tech', 'Upper-Management-or-consulting', 'Marketing-or-advertising-or-pr', 'Government-or-military']

Indices

- number representation of category i.e. [0,1,2,...]

Embeddings

Tokens

- Spliced down words/parts of words fit for mathematical substitution
- e.g. embeddings. = ‘em’, ‘##bed’, ‘##ing’, ‘##s’, ‘.’, ‘[SEP]’, ‘=’

Embeddings

- Captures token, position, & segment embeddings (pre-trained or fine tuned)
- Captures **word meaning** and **word context** (‘I code in Python’ vs ‘I saw a Python at the zoo’)
- BERT, RoBERTa, SBERT, etc.

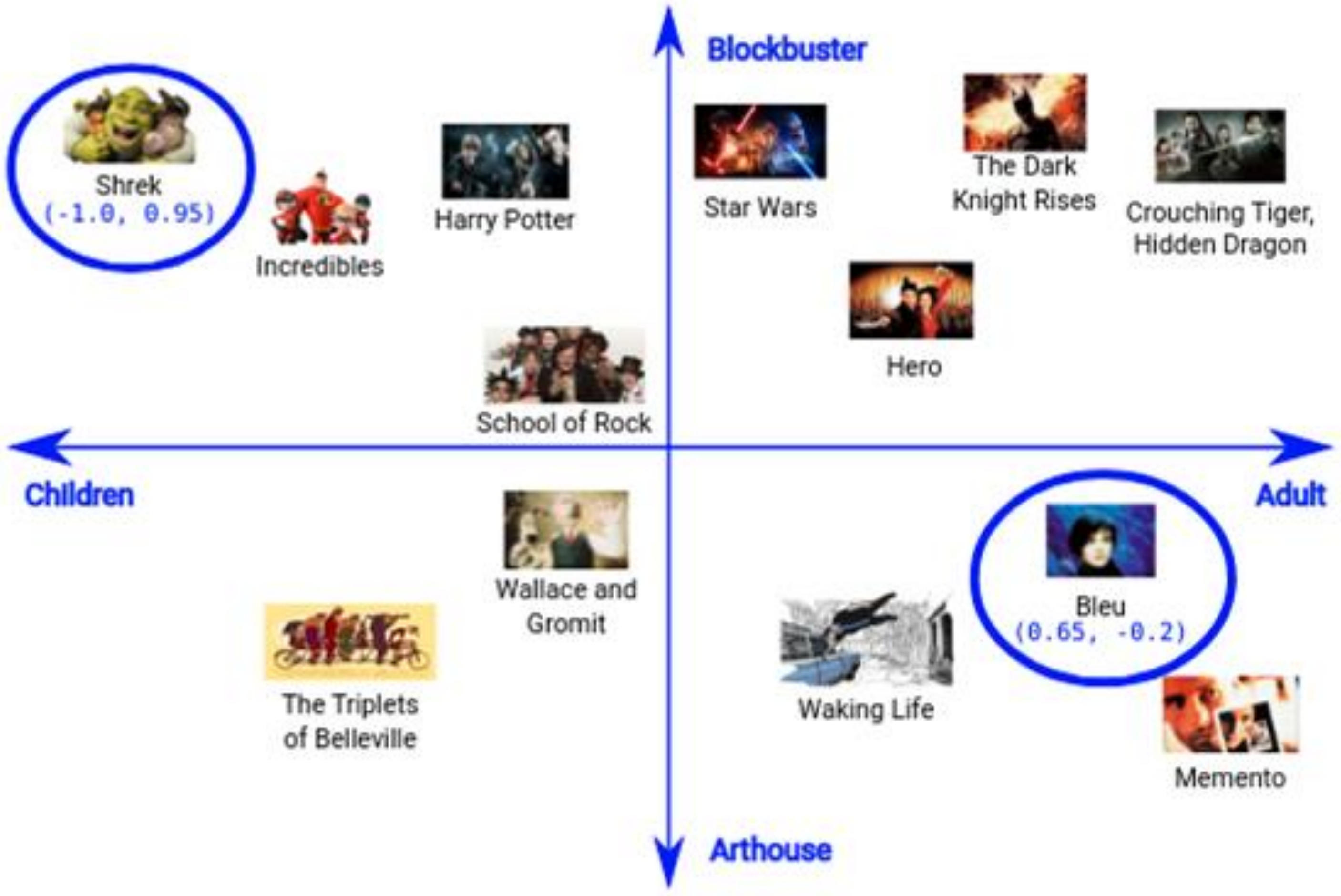
Results

- Vector with numbers representing meaning

Embeddings

Vector space

- dimensions on similar axis
- movies in terms of similarity
- words, to phrases to 400 dimensions



Tutorial 1: Embedding Job Titles

Key concepts

- Demonstrate how to embed text, what it looks like, basic interactions

Model used

- Sentence bert (sbert)
- Good speed/performance, good for phrases/sentences

Results

- Function that takes any text and returns the similar jobs titles from our example data set

Tutorial 2: Clustering Job Descriptions

Key concepts

- Demonstrate how to extract meaning from unstructured data

Model used

- BerTopic
- Prepackaged dimension reduction visualization interfaces

Results

- Clustering topics, reducing clusters, partially labeled data, seeded topics

Key Takeaways

- Embeddings are a way to “understand text”
- We turn that understanding into a number (vector)
- Can then “google search” our own data with new text
- Can cluster our text data based on meaning
- Can assign partial labels or seed phrases*
- Output can be used in many ML/analytics outputs

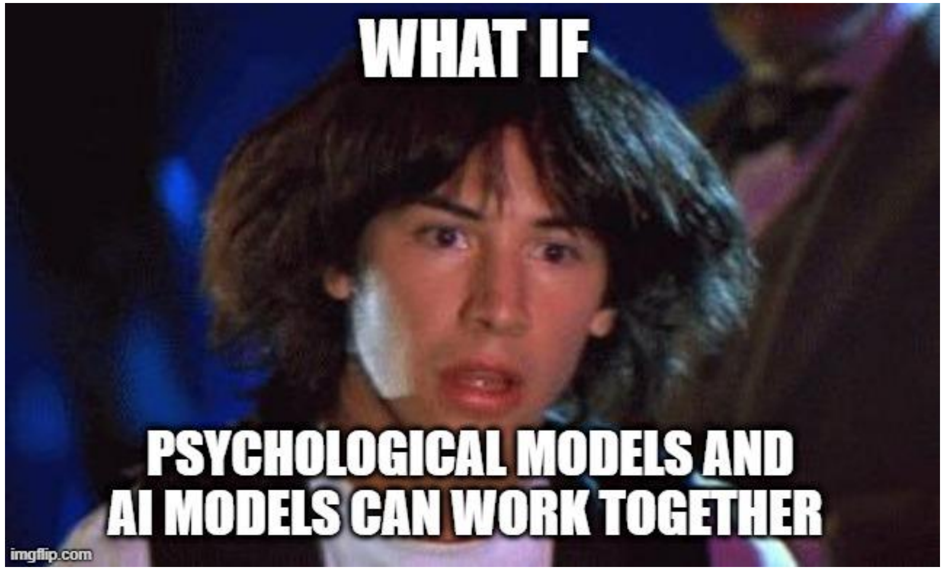
What excites me the most: AI + IO

Theoretical driven AI

- Merge psychological constructs/models with unstructured large scale data
- Cheap (don’t have get human labels) scalable, and fast

Example:

- Personality inventory of words, used as seeds
- Unstructured responses input
- Cluster assignment as outputs
- New features/labels can be used in other predictive models



Application Examples

- Engagement surveys
- Resumes
- Job descriptions
- Interviews
- Open ended assessment
- Emails
- Tweets

Q&A

General Q&A and Discussion

Small Group Discussion Questions

- What kinds of organizational data do you feel would be a good fit for a machine learning approach?
- What potential applications do you see in your work for machine learning?
- What issues have you run into applying or talking about machine learning in your professional life?