DePaul University
College of Computing and Digital Media

Casey Bennett, PhD

Sept.11, 2019

About Me

Casey Bennett, PhD

- Currently a Senior Data Scientist at Cigna
- Received my PhD in Informatics and Computer Science from Indiana University
- Specialize in artificial intelligence, machine learning, robotics, and data science in *healthcare*
- 15 years of industry and academic experience
- Spent about a decade as a data scientist at CVS Health and Centerstone Research Institute (CRI) where he developed national award-winning analytics software and patented AI algorithms
- Served as the Chief Scientific Officer and chief data scientist at a healthcare startup
- Originally from a small town in Kentucky

Mentoring/Teaching

1) **Taught at Indiana University** – for several years. Classes on robotics, health informatics, and machine learning

2) **Led teams** developing software and hardware of sensor systems for monitoring in-home patient health, including algorithms for sensor fusion and machine learning prediction

3) **Ran a learning series** on machine learning techniques for junior analytics staff at CVS Health

4) **Mentoring** junior team members, as well as graduate and undergraduate students. I am even currently mentoring two groups of high school students, one doing virtual avatar research in the national Regeneron Science Talent Search, and second group from the Seattle area building an AI-enabled screening tool for eating disorders

5) Given several **guest lectures** to data science bootcamps and startup events here in Chicago over the past year (e.g. Metis, Depaul, Matter).

6) Currently **serve on the healthcare advisory board** for the Chicago AI Days conference, attempting to build the AI community here in Chicago to better connect universities, startups, and industry, as well as foster more opportunities for young data science students and professionals

Syllabus

Week 1	• Overview of AI, Machine Learning, and Data Science
Week 2	• Methods for Performance Evaluation
Week 3	• Random Forests and Bagging
Week 4	• Ensemble Methods, Boosting, and Voting
Week 5	• Neural Networks and Biologically-Inspired Computing
Week 6	• PCA, Dimension Reduction, and Feature Selection
Week 7	• SVMs, Kernel Methods, and Linear Discriminant Analysis
Week 8	• Bayesian Networks and other Probabilistic Graphical Models
Week 9	• Temporal Modeling and Markov models
Week 10	• Final Project Presentations: part I
Week 11	• Final Project Presentation: part II • Final Report Submission

Four “Soft-Skill” Course Goals

1) Understand the “why” of data science, not just the what
2) Compare model performance *fairly* – is A really better than B
3) Communicate your ML findings effectively
4) Critique other data scientists’ findings

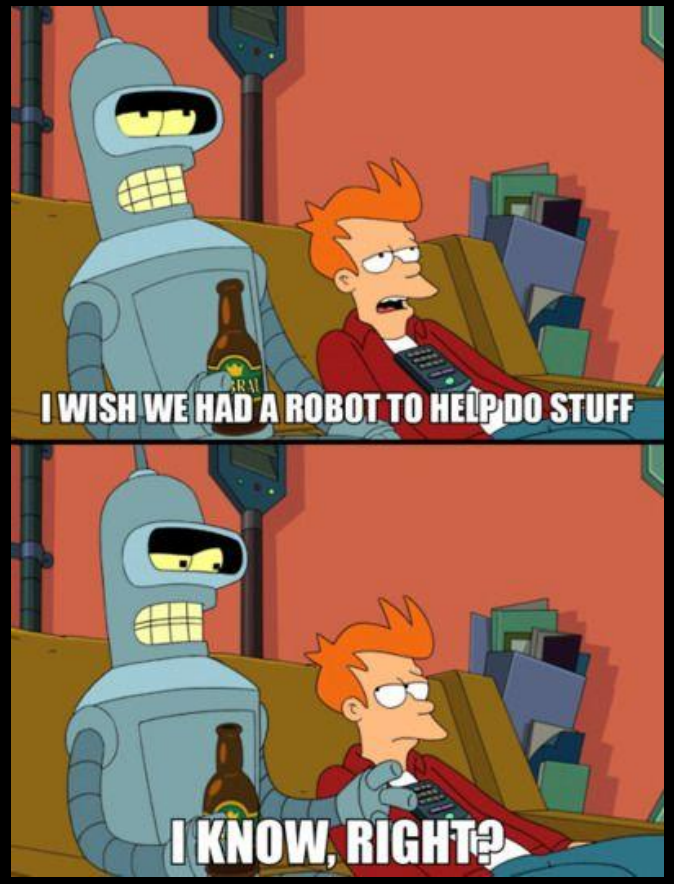
Artificial Intelligence, Machine Learning, and Data Science

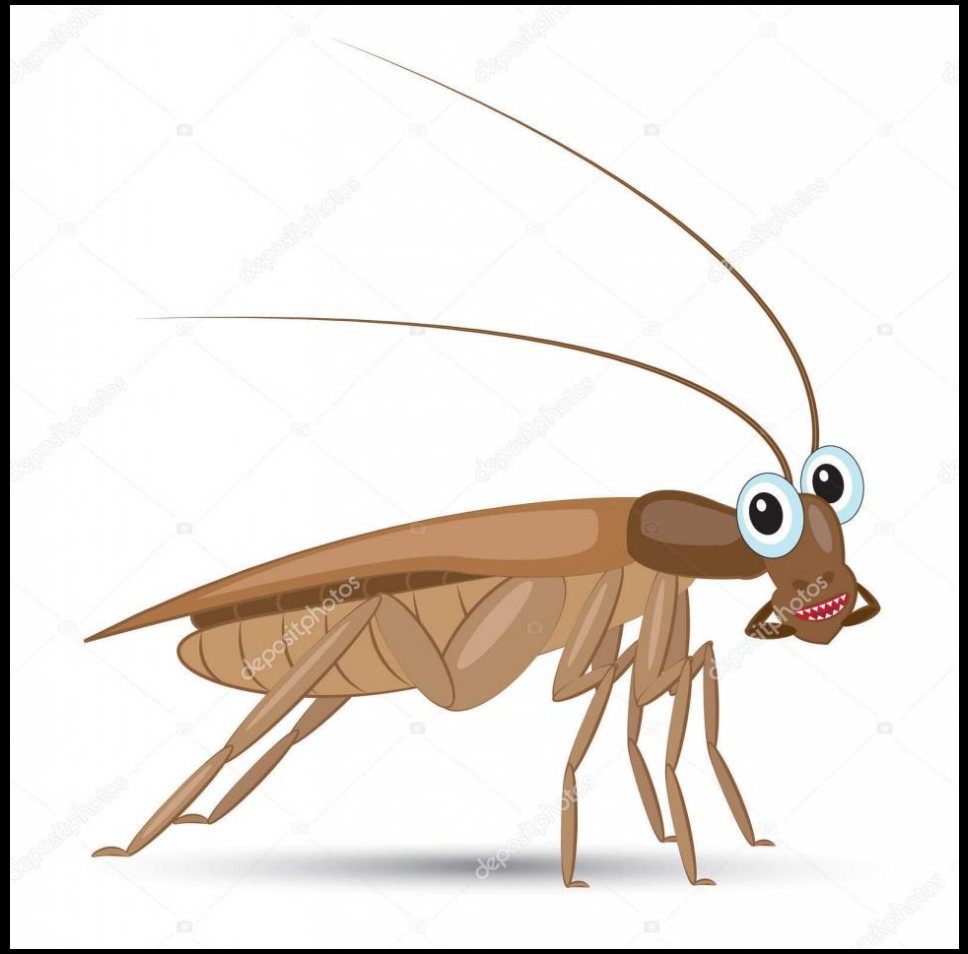
https://pollev.com/caseybennett801

or text “caseybennett801” to 37607

“A year spent in artificial intelligence is enough to make one believe in God.”

– Alan Perlis





Cockroach

Instructor

The ability to respond to novel stimuli in the environment

If I wanted to anticipate and respond to my environment before something happened, how could I do that?

In a physical world, those stimuli might be physical things … but in a *digital* world, those stimuli might be data

Data >> Model >> Prediction

Whether I’m driving a car, or building robots, or doing data science, am I following the same process?

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

“It doesn’t matter what the external thing is, the value we place on it subjugates us ... where our heart is set, there our impediment lies ”

– Epictetus

Thinking like a data scientist

ML & Coding

1) ML Outline

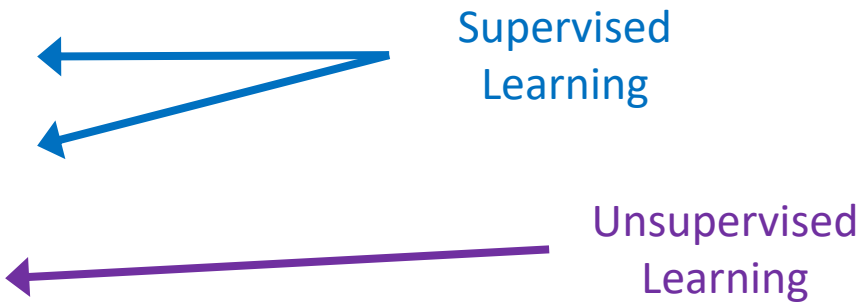
2) Python/Scikit Tutorial

ML Stages

1) **Load Data**
   - *Read File, Parse header and row data*
2) **Preprocess**
   - *Normalize, Discretize, Impute, etc.*
3) **Feature Selection**
   - *Select subset of relevant features*
4) **Train Model**
   - *Fit some model(s) to the dataset*
5) **Evaluate Performance**
   - Did it work?

Different Types of ML Models

1) Classification
2) Regression
3) Clustering
4) Time Series/Temporal
5) Search
6) Optimization



Different Types of Scores

1) Accuracy
2) AUC (ROC Analysis)
3) RMSE
4) Explained Variance
5) AIC/BIC
6) Silhouette Scores
7) etc. etc. etc.

Different Types of Evaluation

1) **Cross Validation**
   - Split the dataset into *k* folds, run *k* times
2) **Test/Train Split**
   - Simple split, e.g. 65% of data used to train, 35% to test

- **Also some alternative approaches (e.g. leave-one-out validation, train/test/validation, bootstrap OOB error) but we'll focus on the above two**

ML Stages

1) Load Data
2) Preprocess
3) Feature Selection
4) Train Model
5) Evaluate Performance

Types of ML Models

1) Classification
2) Regression
3) Clustering
4) Time Series/Temporal
5) Search
6) Optimization

Types of Scores

1) Accuracy
2) AUC (ROC Analysis)
3) RMSE
4) Explained Variance
5) AIC/BIC
6) Silhouette Scores
7) etc. etc. etc.

Types of Evaluation

1) Cross-Validation
2) Test/Train split

Python/Scikit Tutorial

- Scikit method (simple)
- Scikit method (complex)
- Scikit via Jupyter Notebook
- Spark method

  - Idle, Spyder, Jupyter Notebook
  - Using Pip and Conda for Libraries
  - Installing Anaconda

Python/Scikit Tutorial

- Environment setup
- Control flags and global parameters
- How to load data (csv files)
- Data must be float
- Missing values need to be imputed or filtered
- Discuss 5 ML Stages
- Scoring
- Printing output
- Plotting ROC scores, confusion matrices, graphs

For next week

1) Make sure you have Python and Scikit installed (or Anaconda)
2) Familiarize yourself with the syllabus and schedule
3) Read Papers (posted in Content on D2L)
4) Post on online Discussion Forum (Week 2)