Are You as Smart as a Machine? Learning about Machine Learning

Caleb Bradberry, Ph.D.

RADFORD UNIVERSITY

About Me

- Grew up in West Virginia
- Undergraduate
  - Marshall University: B.S. in Information Technology
  - Marshall University: B.B.A. in Information Systems
- Graduate
  - Marshall University: M.B.A. – Thesis: An Econometric Projection of Coal Sustainability for W.V.
  - UNC Greensboro: Ph.D. – Dissertation: A Design Science Framework for Healthcare Analytics
- Professional
  - Assistant Professor of Information Systems – Radford University
    - ITEC145 – Data Ethics, Privacy, and Security
    - ITEC200 – Healthcare Information Systems
    - ITEC369 – Systems Analysis and Design
    - ITEC375 – Data Science with R and Python
    - ITEC485 – Decision Support Systems
    - ITEC685 – Advanced Data Science

Some Questions to Think About

- How can we proactively predict a heart attack using smart devices?
- How can we predict a patient that is at risk of being readmitted after a heart failure?
- How can we better identify the structure of cancer?
- How can we identify students at risk of failing out of the university?
- Can we use the characteristics of a successful student's schedule to create better schedules?
- How can we predict stock prices? Sports outcomes? Cryptocurrency prices?
- How does the dog filter on Snapchat track my face?
- How does Amazon recommend items?
- WHAT IS WITH THIS FACEBOOK AD?! I ONLY THOUGHT ABOUT BUYING IT
- How is TikTok recommending me this stuff?!

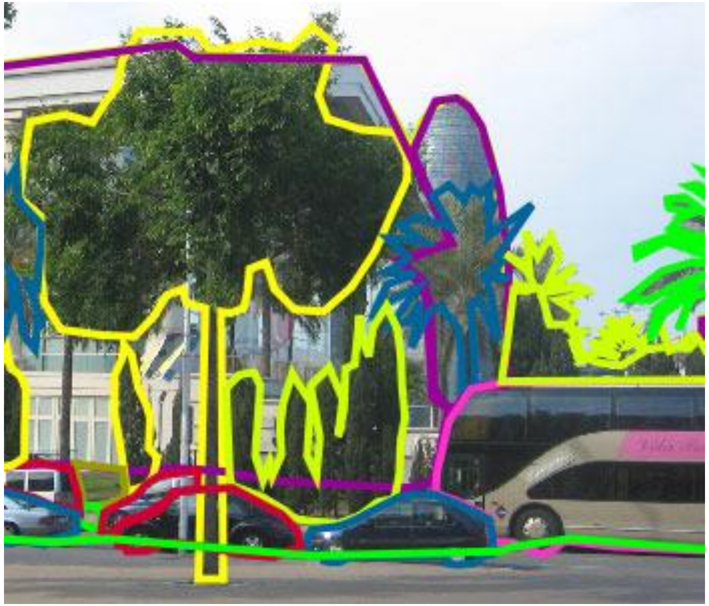
How can a machine learn?

- Ipse se nihil scire id unum sciat
  - """I know that I know nothing" – Plato" – Decartes"
    - -Michael Scott
  - Nothing says a light conversation like philosophy of science
- Big words:
  - Ontological Knowledge
    - "What is the thing"
    - Domains and Boundaries
  - Epistemological knowledge
    - "How is the thing"
    - Belief, truth, and justification
    - A priori, a posteriori, and ex post facto
    - Analytic-Synthetic distinction – Immanuel Kant

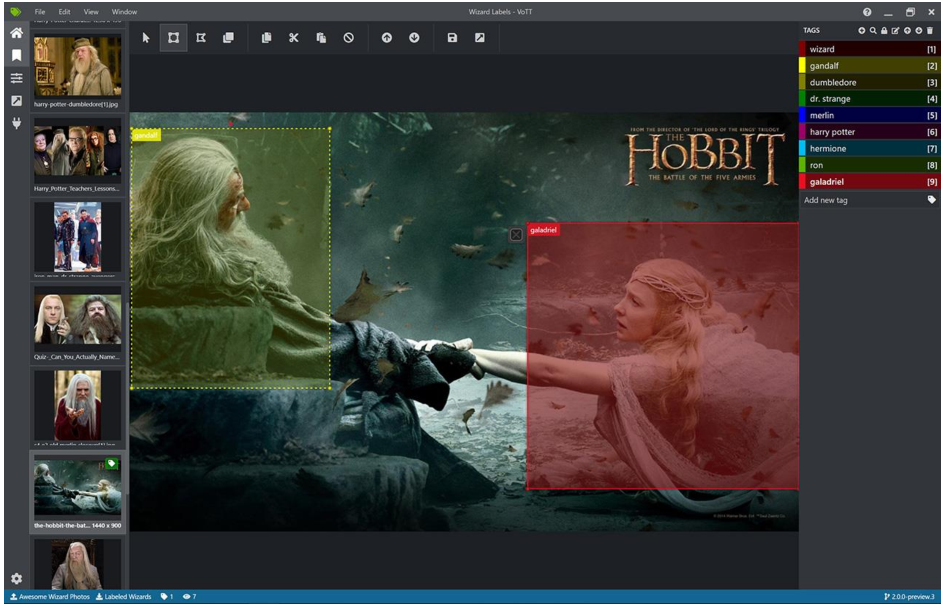
Wait what?

- We’re not done with philosophy of science yet:
- Falsification
  - Karl Popper – “All swans are white and have always been white,” can be proven false
    - “In 1697, black swans were observed on the shore of the Swan River in Australia.”
    - This is the concept of refutation
  - Thomas Kuhn – scientific theories are social constructs
    - Accurate, consistent, broad scope, simple, and fruitful
- How do we ask a machine a question?
  - Modeling.
    - Labels
    - Features
    - Factors
  - Statistics, computer science, and a knowledge domain

Image labeling



http://labelme.csail.mit.edu/Release3.0/



https://github.com/microsoft/VoTT

Types of Machine Learning

- Approaches:
  - Supervised, Unsupervised, and reinforcement
- Processes:
  - Classification
  - Clustering
  - Regression
  - Anomaly Detection
  - Association Rules
  - Structured Prediction
  - Decision Tree

- Models:
  - Neural network
  - Support vector machine
  - Bayesian
  - Genetic
- Problems
  - Bias
  - Model underfit
  - Model overfit
  - Sampling

Supervised, Semi-supervised, Unsupervised Learning

- All types use training data, testing data, and validation
- Some have inputs and a desired output
  - Think of it like regression
    - (or y=mx+b, we are slapping a line through some data)
  - Outcome = Factor1+Factor2
    - Y (outcome) = mx1(factor1) + mx2(factor2) + error
  - StudentRetention ~ GPA + MidtermGrade + Gender + Dorm
- Unsupervised does not have an objective/outcome function
  - How does data group together?
  - Clustering
    - (Correlation matrix)

Fail to Reject the Null

- Machine learning’s underpinnings are statistical, so we work with statistical language
  - Hypothesis testing
  - We reject the null hypothesis when evidence emerges to support the alternative hypothesis
  - When we have no evidence, we failed to reject the null hypothesis
- What did you just call me?
  - Remember falsification and the black swan? We had no evidence of a black swan, until we did. We *failed to reject the null hypothesis*
- Let’s make it more confusing

Confusion and Statistics: A Confusion Matrix

	Observed Positive	Observed Negative
Prediction positive	True Positive	False Positive (Type I error)
Prediction Negative	False Negative (Type II error)	True Negative

Accuracy:

Σ True positive + Σ True negative/Σ Total population

Precision:

Σ True positive/Σ Predicted positive

Recall/Sensitivity

Σ True positive/Σ Condition positive

Specificity

Σ True negative/Σ Condition negative

Associations

- Two types of pitfalls can occur that affect the association between exposure and disease
  - Type 1 error: observing a difference when in truth there is none
  - Type 2 error: failing to observe a difference where there is one.

Interpreting Epidemiologic Results

Four possible outcomes of any epidemiologic study:

	REALITY	
YOUR DECISION	H_0 True (No assoc.)	H_1 True (Yes assoc.)
Do not reject H_0 (not stat. sig.)	**Correct decision**	**Type II (beta error)**
Reject H_0 (stat. sig.)	**Type I (alpha error)**	**Correct decision**

Four possible outcomes of any epidemiologic study:

	REALITY	
YOUR DECISION	H_0 True (No assoc.)	H_1 True (Yes assoc.)
Do not reject H_0 (not stat. sig.)	Correct decision	Failing to find a difference when one exists
Reject H_0 (stat. sig.)	Finding a difference when there is none	Correct decision

Type I and Type II errors

- α is the probability of committing type I error.

- β is the probability of committing type II error.

“Conventional” Guidelines:

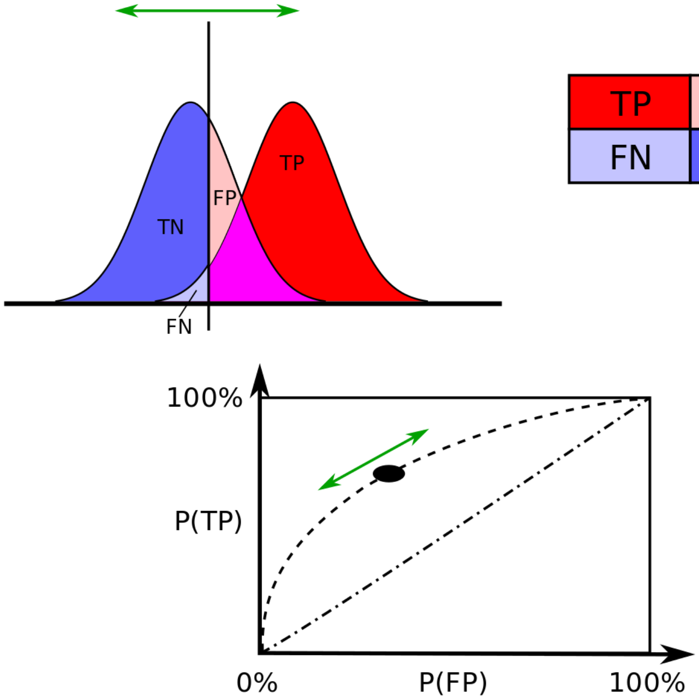
- Set the fixed alpha level (Type I error) to 0.05 This means, if the null hypothesis is true, the probability of incorrectly rejecting it is 5% or less.

	DECISION	**H_0 True**	**H_1 True**
Study Result	**Do not reject H_0** **(not stat. sig.)**		
	Reject H_0 **(stat. sig.)**	**Type I** **(alpha error)**	

Receiver Operating Characteristic (ROC) Curve

- WHY ARE YOU DOING THIS?!
  - World War II radar technology: How did the British allied forces know the difference between a friendly bomber and a German blitzkrieg?
  - The operator of the person using the radar was measured as to how correct or incorrect their detection was

ROC in Action



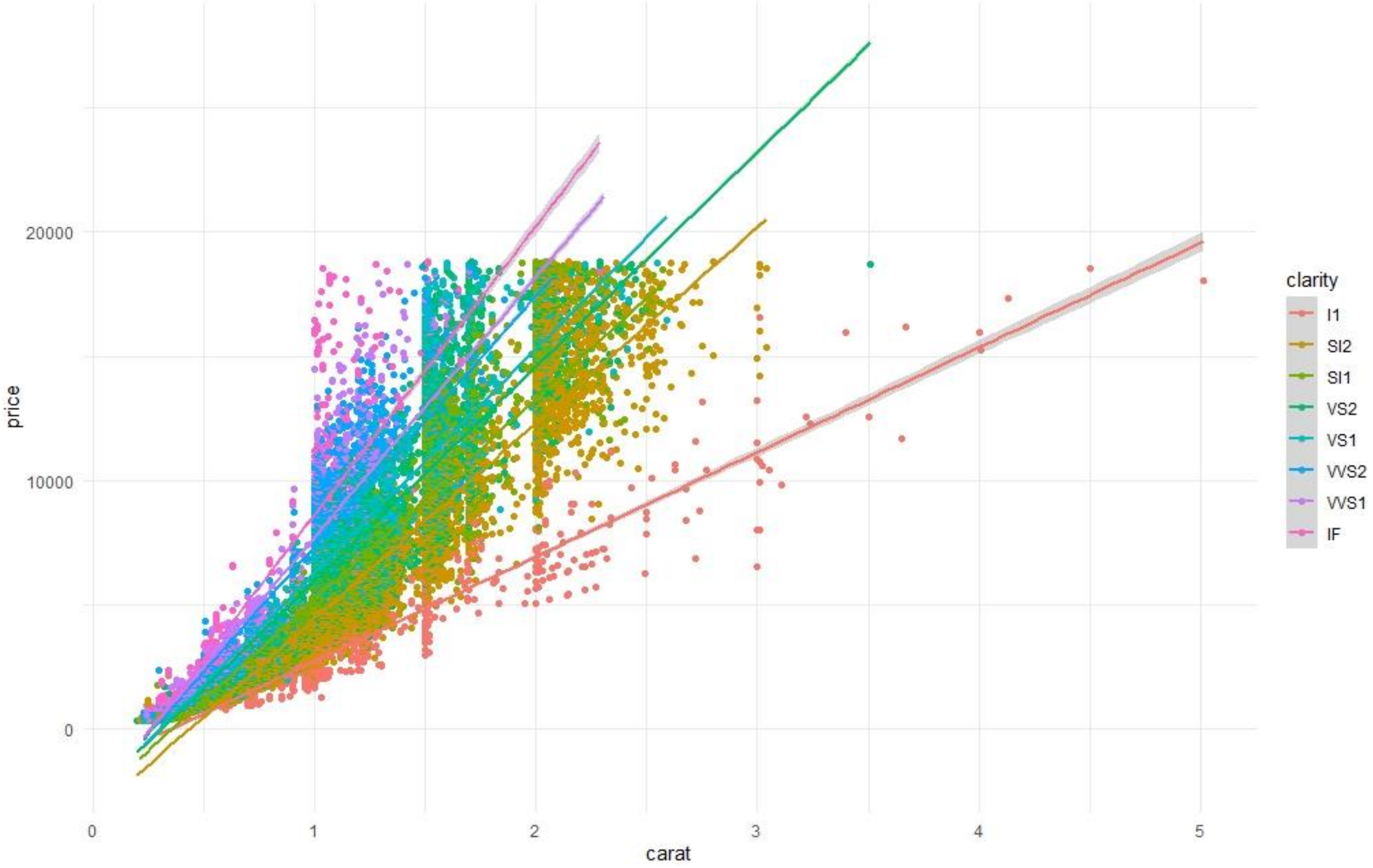
TP	FP
FN	TN

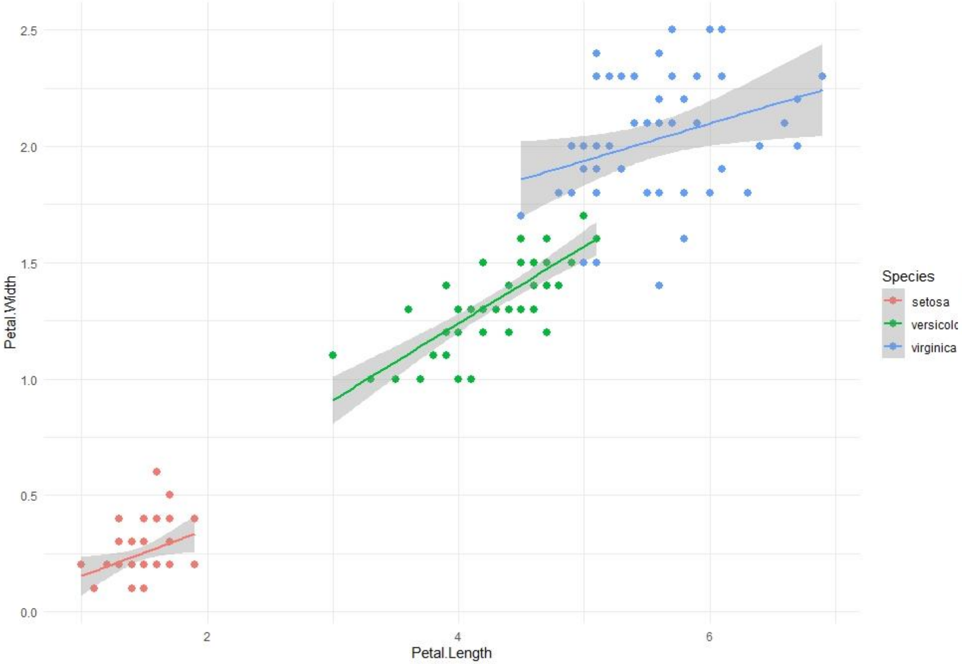
Training and Testing

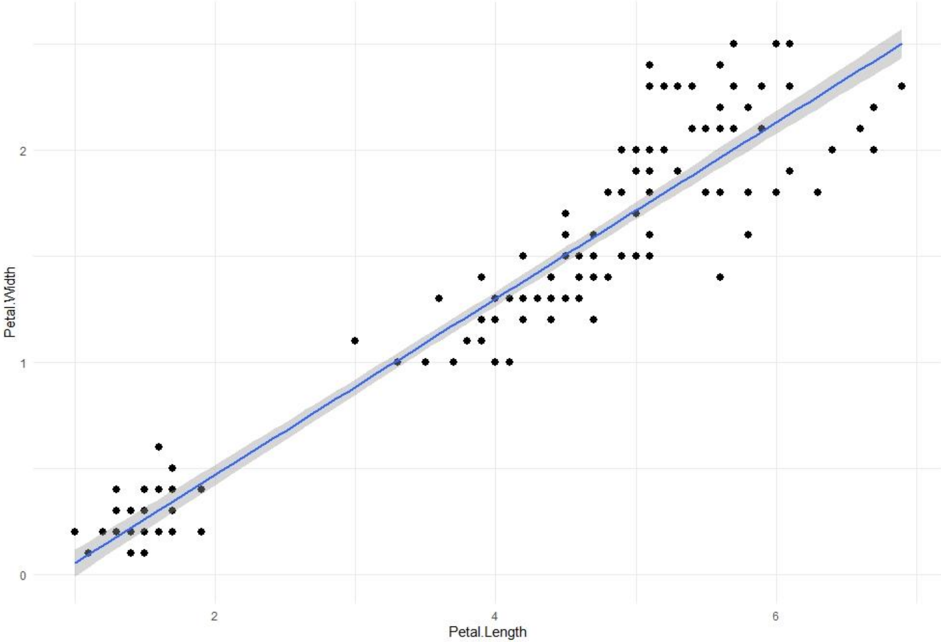
- How can we know if our model works? If our prediction was true?
  - Train the algorithm on a set percentage of data (50-80%)
  - Test the **–trained model-** against the remaining data to see if the data fits the model (imagine drawing a line on a slope, trying to fit that line into the next dot)

Beyond Stats: Machine Learning Types

- Very basic modeling: linear regression
- Generalized linear model (supervised)
  - Basic linear model
    - diamondPrice = cut + carat + clarity + color
    - Interpretation is a straightforward unit increase/decrease
  - Logit model
    - studentRetention = GPA + Midterm + cardSwipes
    - Interpretation is an odds ratio, what is the ‘likelihood of the thing’

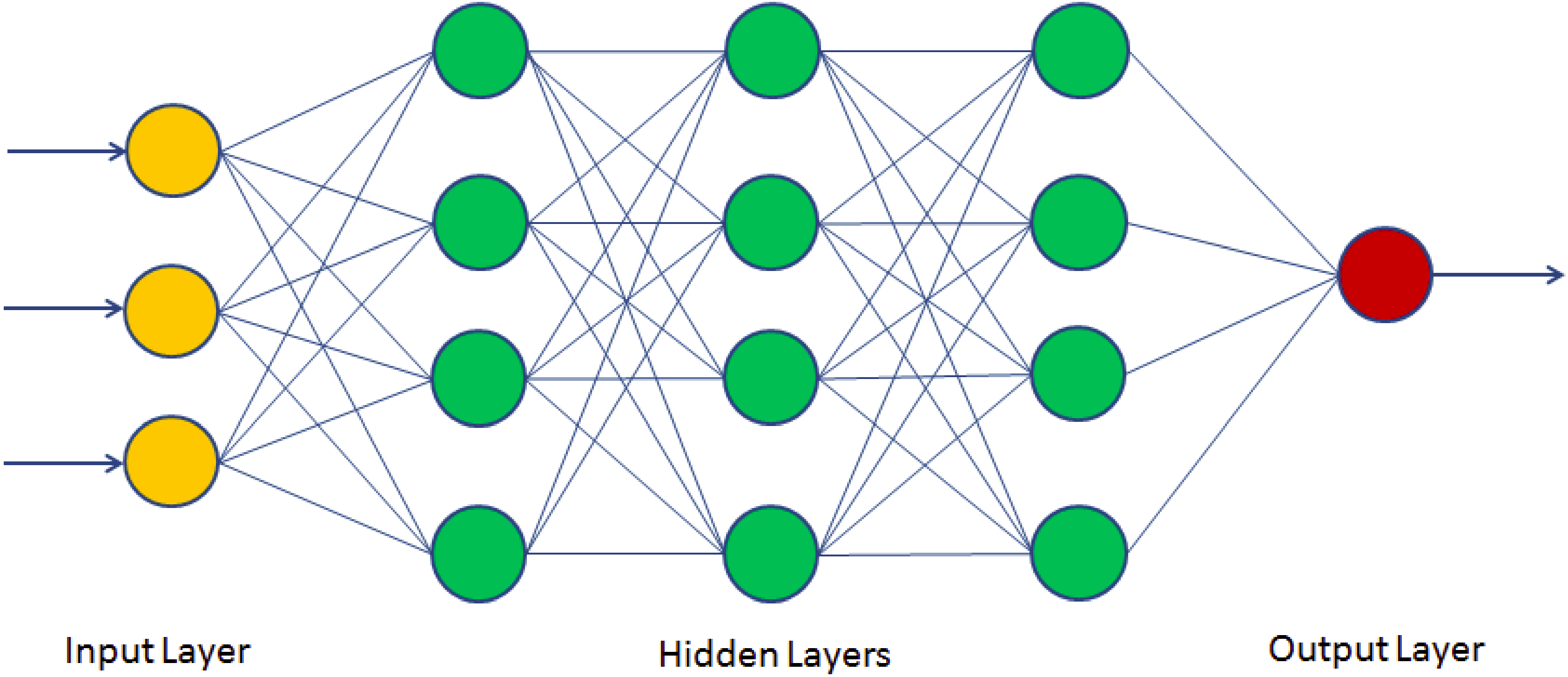

price
20000
10000
0
carat
0
1
2
3
4
5
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

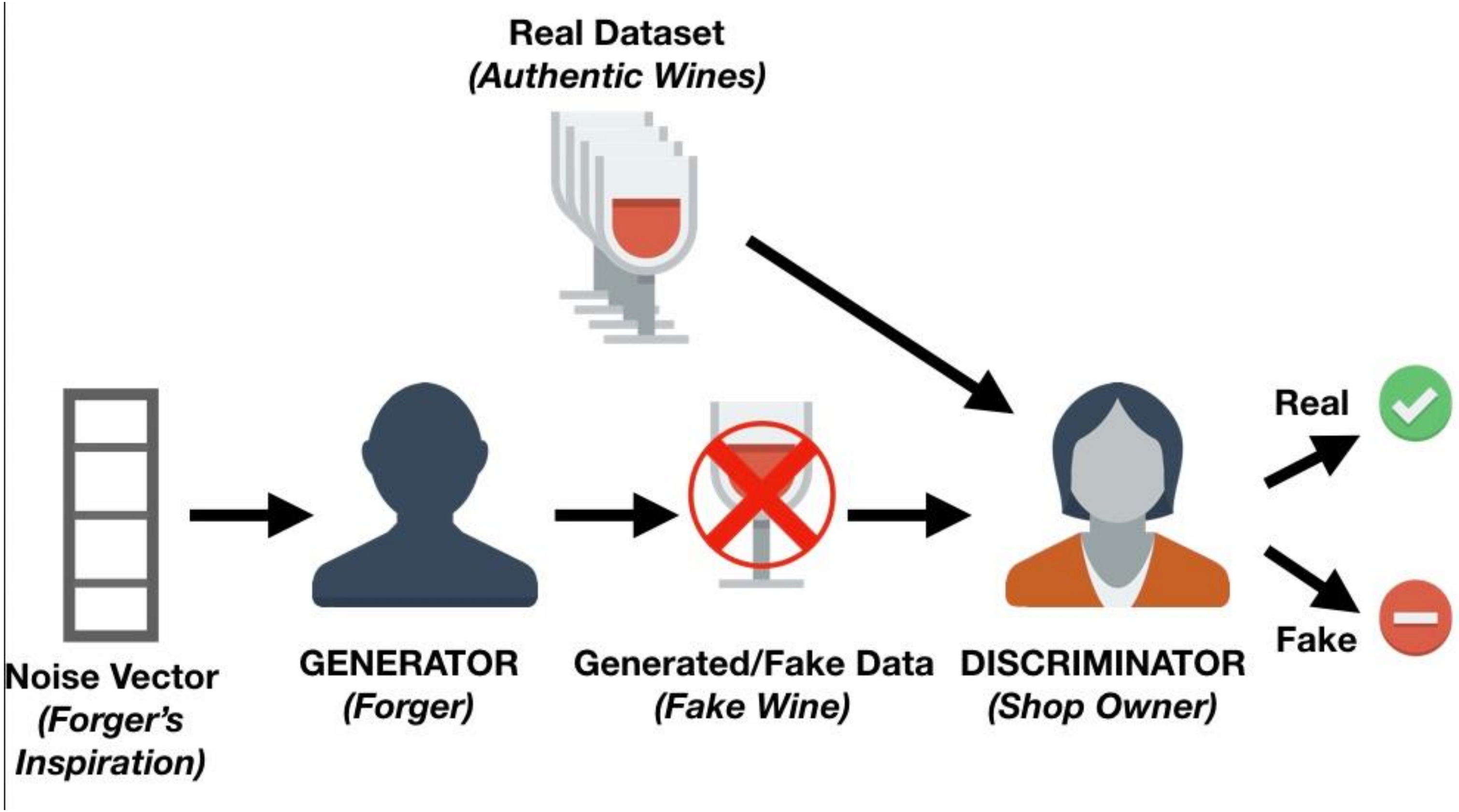

Petal.Width
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
Petal.Length
Species
setosa
versicolc
virginica

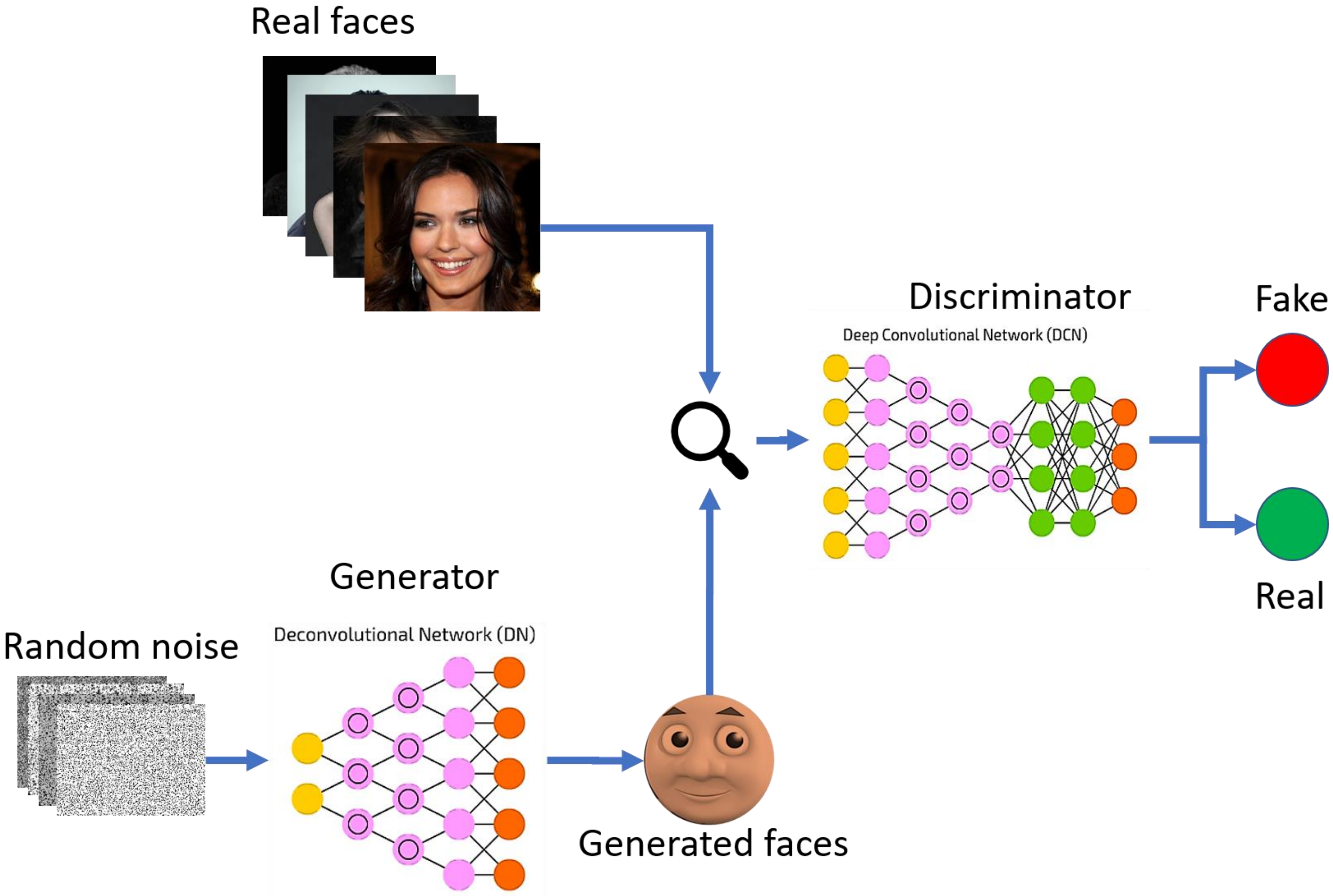

Petal.Width
2
1
0
2
4
6
Petal.Length

ML Types: Neural Networks

- Artificial neural network
  - Linear modeling
  - Classification and prediction
- Convolutional neural network
  - Visual modeling (Style transfer)
  - Image generation (Deep dream, Deep Fake)
  - Generative adversarial neural network [GANN] (DeepFake, ThisPersonDoesNotExist)
- Recurrent neural network
  - Time-varying, backwards propagation
  - Gradient descent
  - Self organizing map
  - Music generation
- Adversarial Neural Network
  - Typically a GANN, but other configurations exist


Input Layer
Hidden Layers
Output Layer


Real Dataset
(Authentic Wines)
Real
Fake
Noise Vector
(Forger's
Inspiration)
GENERATOR
(Forger)
Generated/Fake Data
(Fake Wine)
DISCRIMINATOR
(Shop Owner)


Real faces
Discriminator
Deep Convolutional Network (DCN)
Fake
Real
Generator
Random noise
Deconvolutional Network (DN)
Generated faces

Reinforcement Learning (Recurrent Neural Network)

- Actions to maximize an outcome
  - Game theory
  - Information theory
  - Simulation
  - Markov Decisions
- Rain, sprinklers, and is the grass wet?
  - Bayesian network
  - Email filters
- Examples:
  - LSTM algorithm (stock prediction)
  - TensorFlow Magenta https://magenta.tensorflow.org/

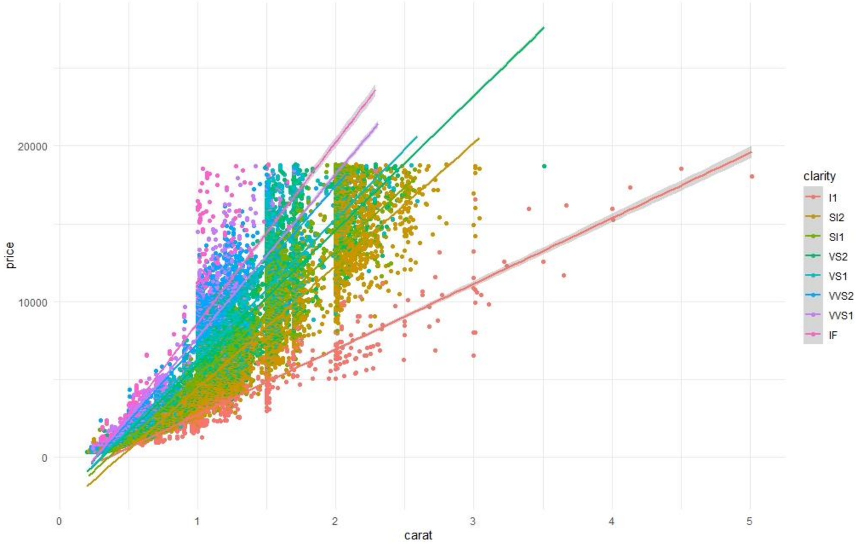
NLP: Natural Language Processing

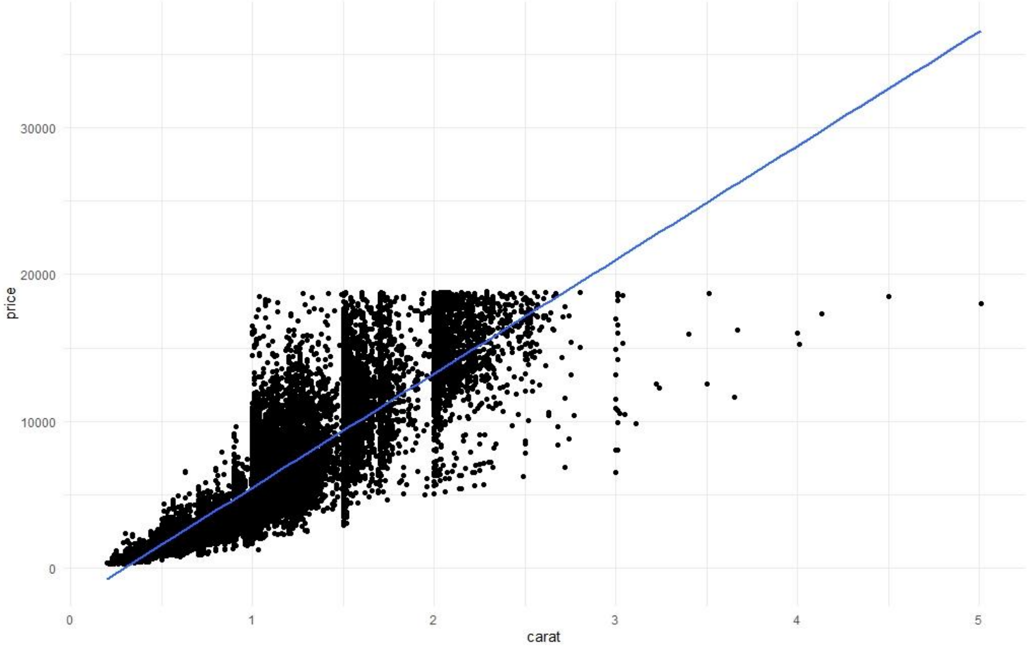
- Two types: rule based, statistical
  - Evolved into statistical using probabilistic decisions
- Okay Google, Hey Siri, Alexa reorder…
- Grammar Induction
  - Rules of the language
- Morphological segmentation
  - Root: the word open. Opens, opened, opening
- Parts-of-speech labeling
  - Subject noun verb predicate
- Stemming
- Latent Dirichlet allocation
- This is why you get ads for stuff you were talking about or thinking about.

Limitations: Bias

- Facial recognition trained on a subset of demographics
  - We use data available, and if that data is 93% Caucasian, the network will classify Caucasian better than other skin tones
- Microsoft's Twitter Chatbot
  - People were able to send it messages of any kind: it evolved to pick up on racist and sexist language
- Identifying student retention
  - I trained a neural network to predict retention, but the goal was to predict attrition

Limitations: Model Fit





ML and the Real World

- Hospital Readmissions
- Heart Failure Prediction
- Student Retention
- Student Scheduling
- Auto-coloring filters
- Face recognition
- Purchase likelihood

ML: Ethics

- Ownership
  - Who owns data?
- Transparency
  - How is the data collected, aggregated used?
- Consent
  - If your personal data is being used, did you consent?
- Privacy
  - Who can see what about you?
- Currency
  - Who is profiting from your data?
- Openness
  - Can you access the data about you?

Some starting resources

- AWS Compute Time - $100
  - https://aws.amazon.com/education/awseducate/
- Microsoft Azure - $100 credit
  - https://azure.microsoft.com/en-us/free/students/
- Google GPU - $300 free
  - https://cloud.google.com/free/
- R and R Studio
  - https://www.r-project.org/
  - https://www.rstudio.com/
- Anaconda Python Installer (Spyder IDE recommended)
  - https://www.anaconda.com/distribution/
- Keras Deep Learning Library
  - https://keras.io/
- Reddit Learn Machine Learning
  - https://www.reddit.com/r/learnmachinelearning/