A BEGINNER'S APPROACH TO NATURAL LANGUAGE PROCESSING

By Sam Fisher

WHO AM I?

- A Beginner.
- A fellow human with curiosity, the internet, and some good books.
- Spent most of my adult life as a professional composer.
- Currently a student in Springboard Data Science Bootcamp.
- Not scary.

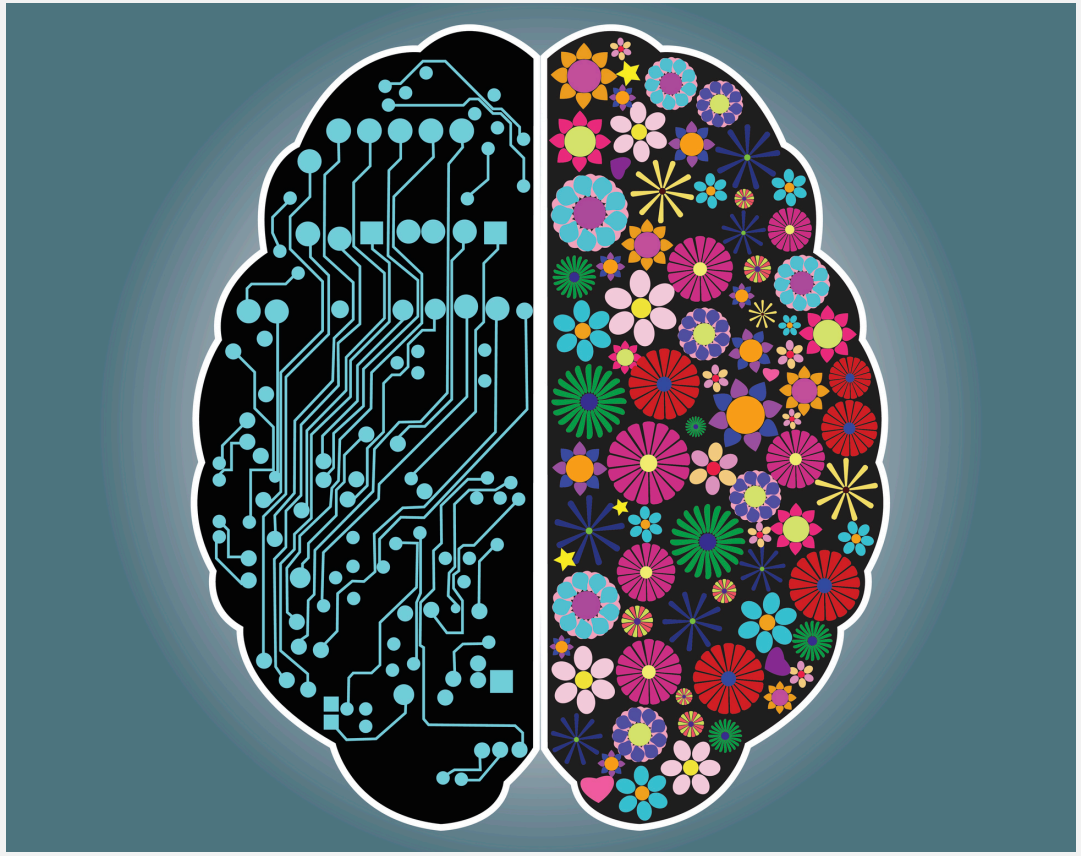
Please feel free to ask questions, challenge things, and engage!

OBJECTIVES FOR TODAY

- Learn why NLP is an interesting and valuable challenge.
- Gain intuition for the fundamental issues in predictive machine learning.
- Implement a classification algorithm to predict document topics in python.

https://www.freepik.com/free-icon/oscar-prize-statue-silhouette_767015.htm

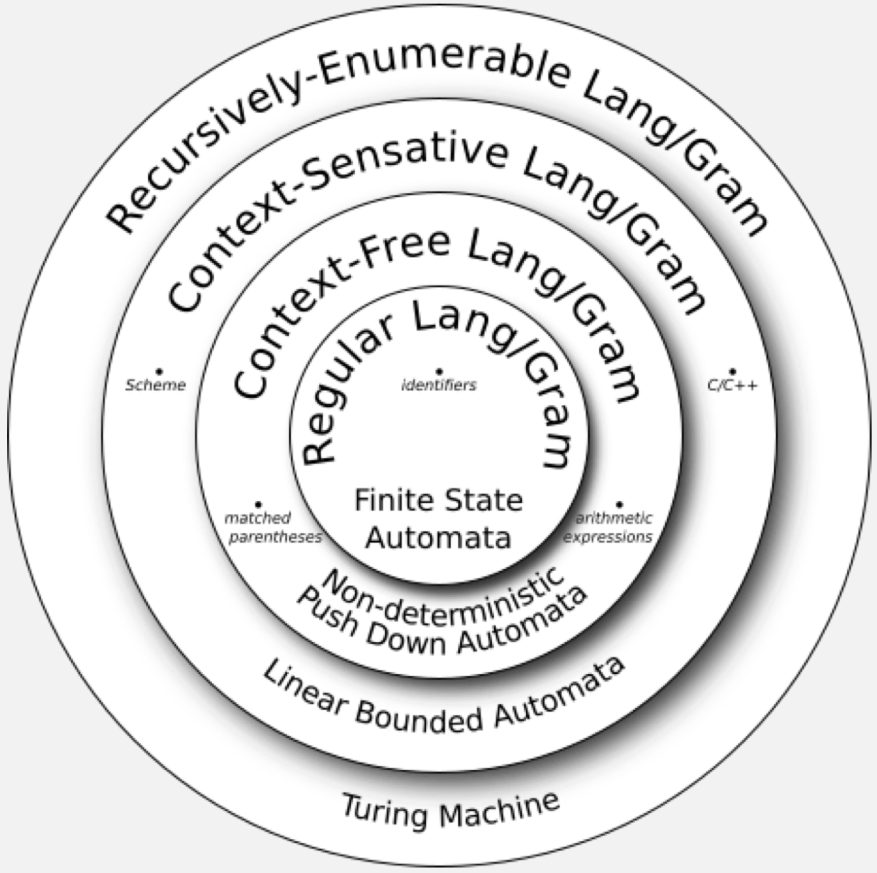
FORMAL VS NATURAL LANGUAGE



https://www.health.harvard.edu/blog/right-brainleft-brain-right-201708251222

FORMAL LANGUAGES

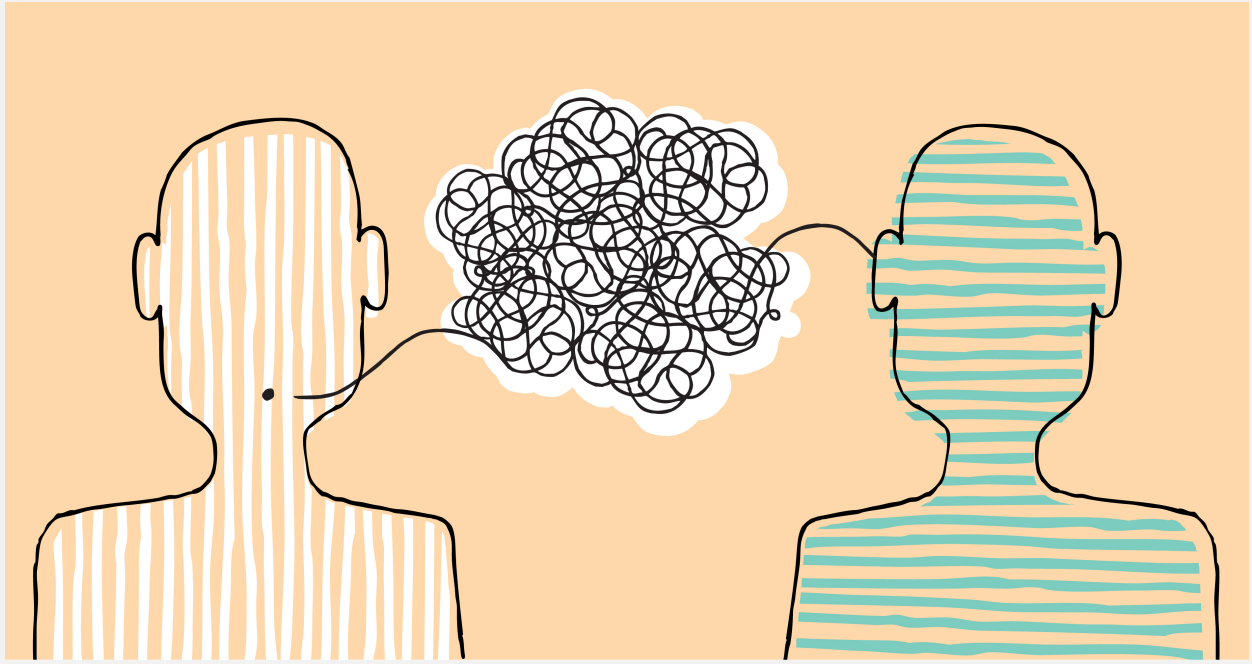
- A set of strings of symbols with a finite set of governing rules.
- From these rules, any valid statement can be produced.
- Ex: Algebra!
- Ex: Python!
- Ex: Any programming language, actually.
- Ex: Most systems that we think of as being fully logical.



http://perugini.cps.udayton.edu/teaching/books/PL/www/lecture_notes/grammars.html

NATURAL LANGUAGES

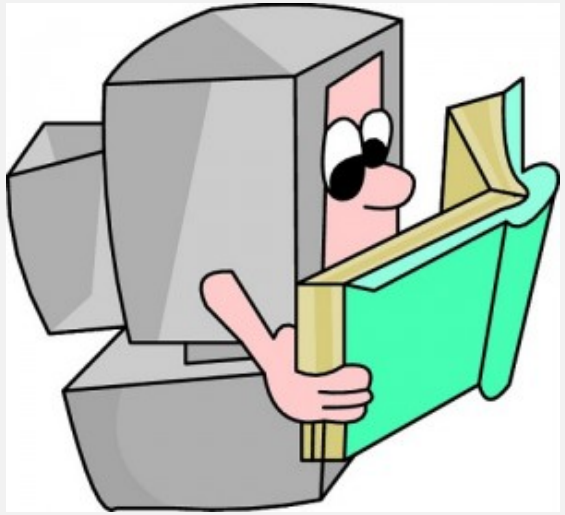
- Language humans use to communicate with each other.
- Evolves naturally, without a premeditated design.
- Riddled with subjectivity, connotation, and ad hoc structure.
- Frustrating when biz people use it on programmers. <- (pandering)



http://ronneb.com/5-cs-to-improve-your-communication-skills/

HOW CAN FORMAL LANGUAGES PARSE NATURAL ONES?

- Parsing natural language means dealing with uncertainty.
- Humans intuitively interpret statements probabilistically when reading or listening.
- We can deploy Machine Learning algorithms to do similar pattern recognition, typically in a highly restrictive context.
- NLP tasks include:
  - **Topic Classification**
  - Entity Extraction
  - Sentiment analysis
  - Error Correction
  - Translation
  - Web Search (Term Relevance)



http://heatstrings.blogspot.com/2015/07/e-ethelbert-miller-aldon-nielsen_14.html

WHY DO WE WANT TO DO THIS, ANYWAY?

- Abstractly, isn't it kind of a cool?
- Concretely, the internet is a mine of valuable text data.
  - Web Pages Relevant to A Search (Ex: Google Search)
  - Customer Opinions on a Product or Brand
  - Voter's Facebook Posts (Ex: Cambridge Analytica (Boo!))
- Also, language is a natural way to interact with devices
  - Chat bots!
  - Alexa!
  - We won't talk about this much today.

https://www.coursera.org/learn/language-processing

TOPIC CLASSIFICATION

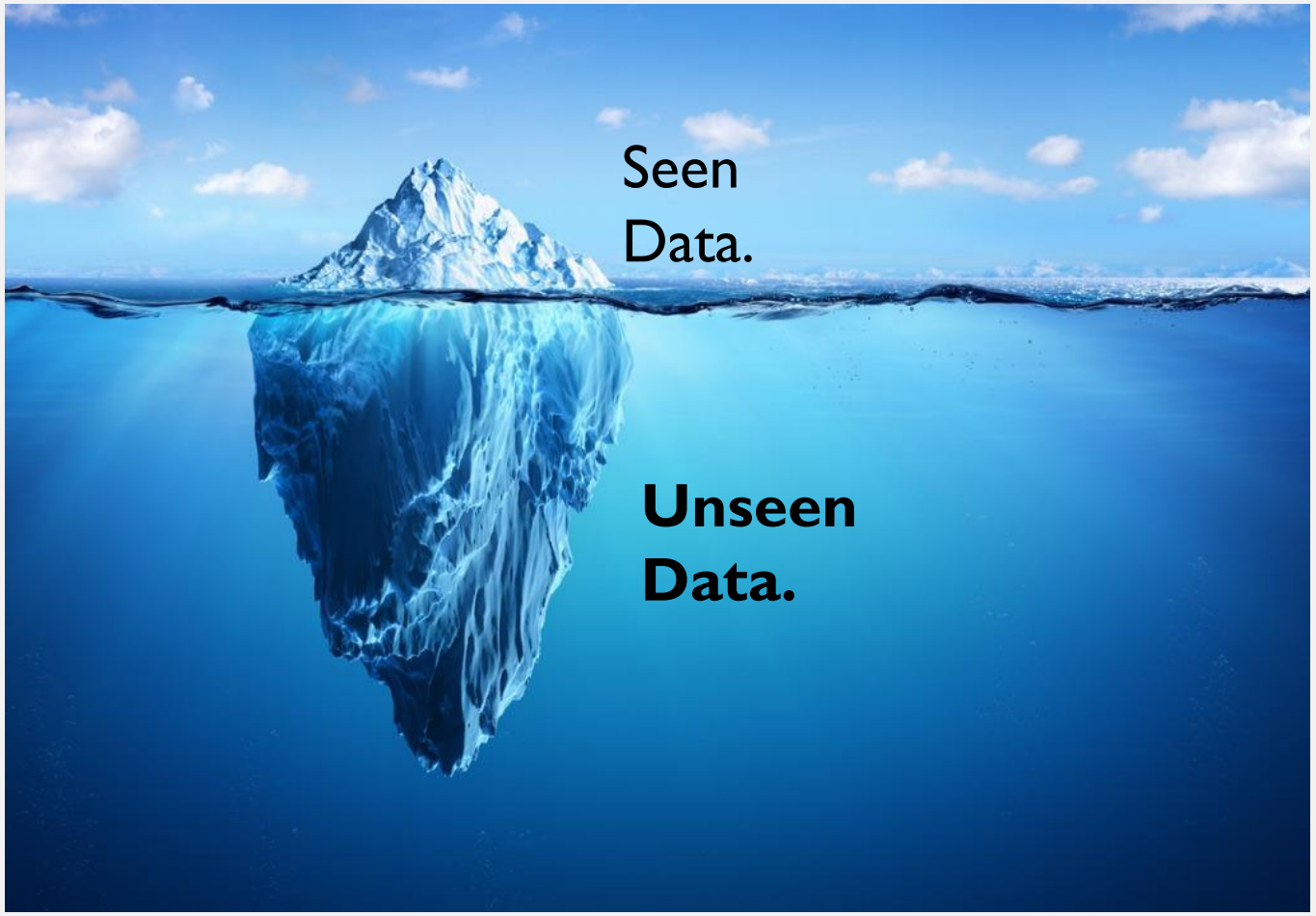
Let's think about what this is for a minute.

- We have a collection of documents with labels.
- The labels are chosen from a finite, predetermined set.
- We train a model to correctly label future documents.

- A prediction problem with labeled data to learn from. (Supervised Learning)
- The prediction is about discrete classes, rather than continuous values. (Classifier)

Any questions so far?

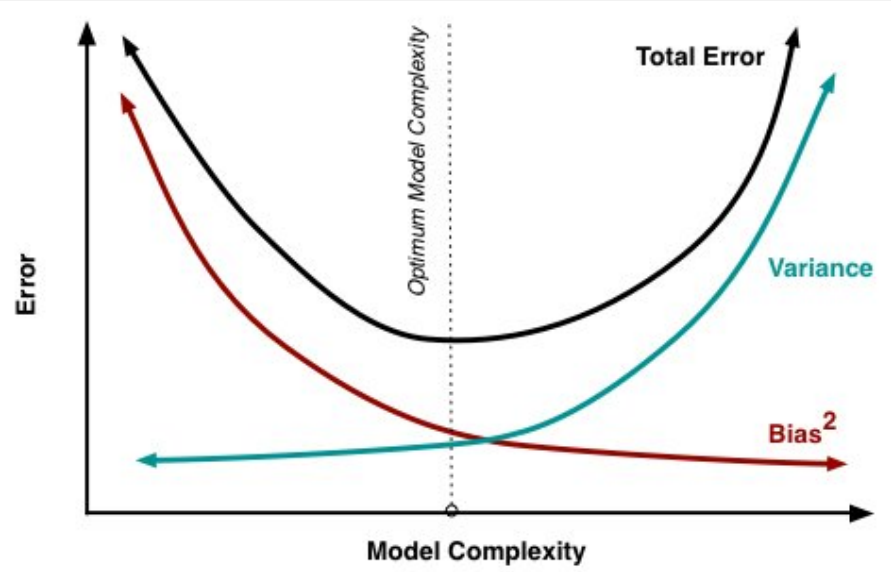
UNCERTAINTY IN PREDICTION.



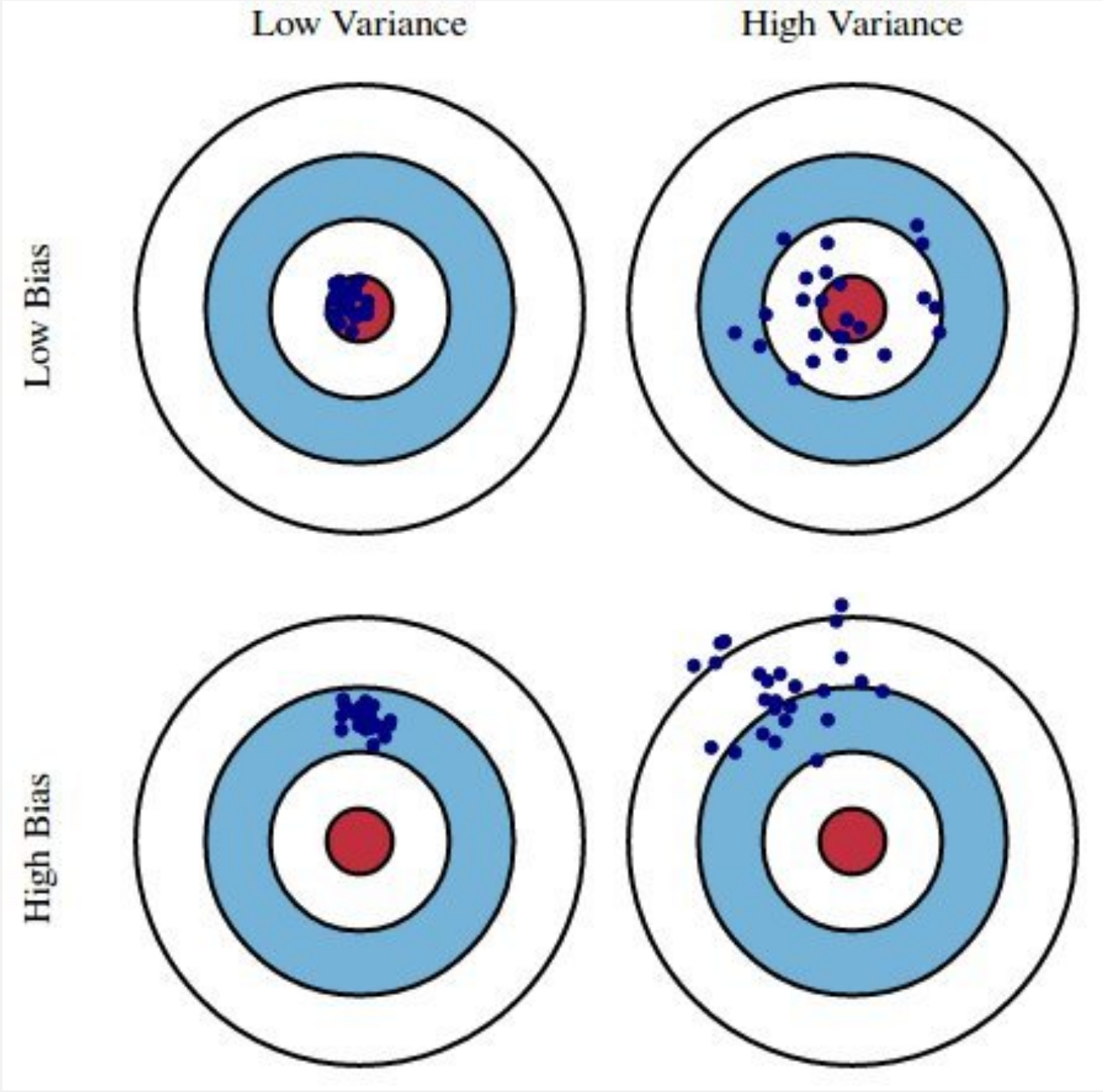
https://www.reward-guide.co.uk/reward-management/the-gender-pay-gap-is-just-the-tip-of-the-iceberg/2370.article

BIAS AND VARIANCE.

- Error in our prediction model comes from two places.
- Bias - Error from incorrectly classifying the seen data. (Underfitting)
- Variance – Error from incorrectly classifying the unseen data. (Overfitting)

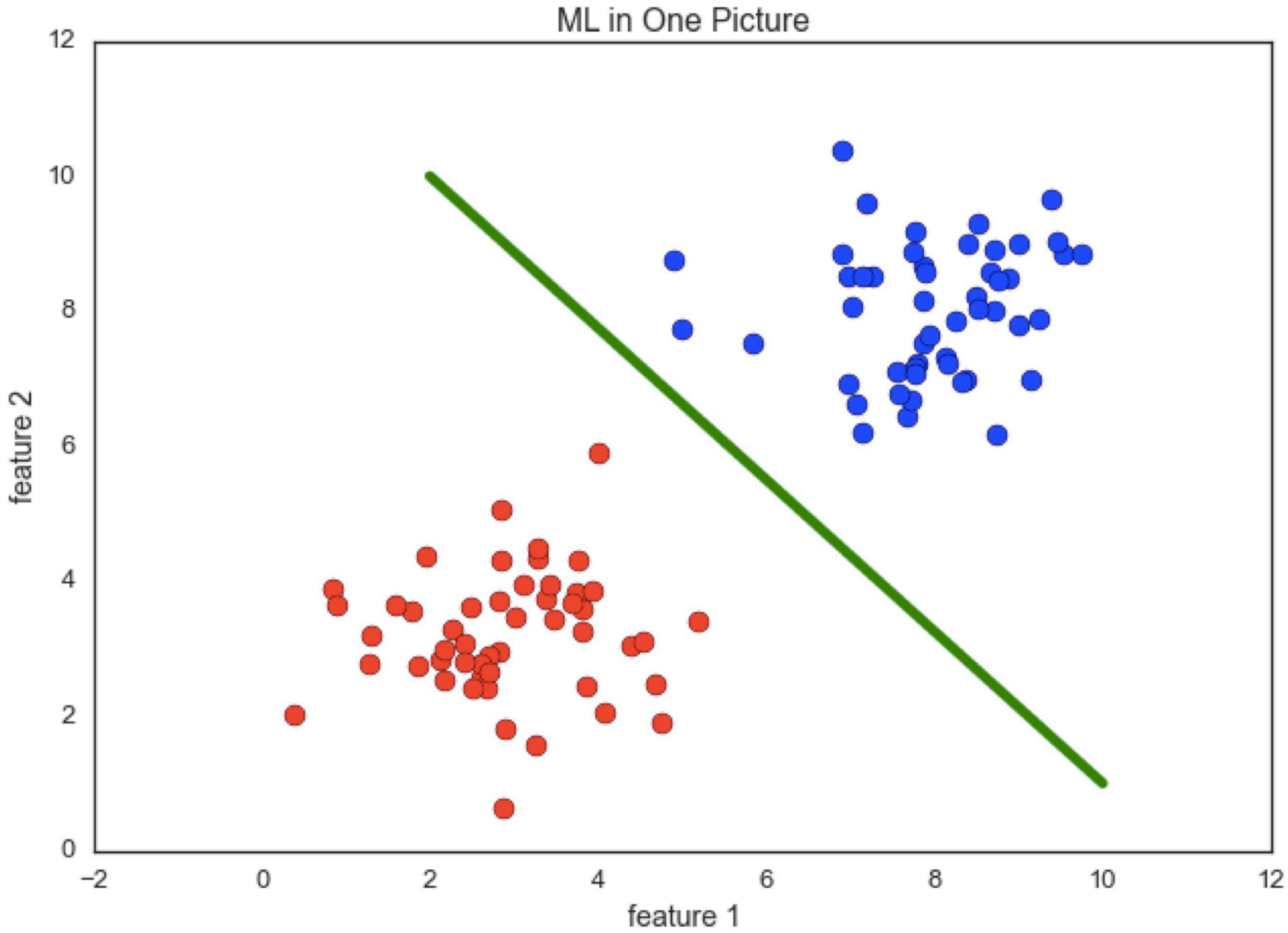


https://www.kdnuggets.com/2016/08/bias-variance-tradeoff-overview.html



https://www.kdnuggets.com/2016/08/bias-variance-tradeoff-overview.html

THE BIG PICTURE OF CLASSIFICATION



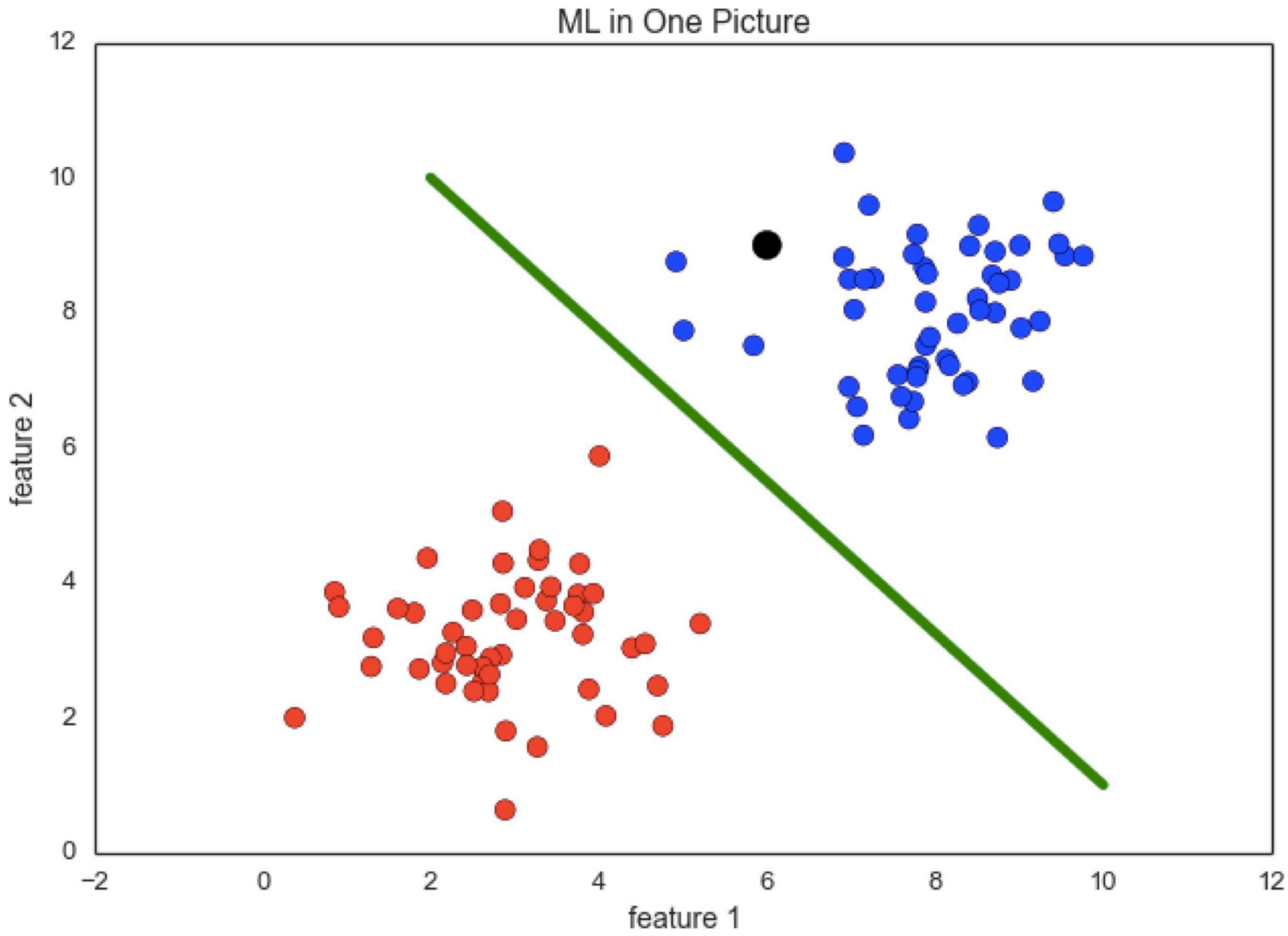
x: data points

y: labels

features

decision boundary

http://cs109.github.io/2015/pages/videos.html



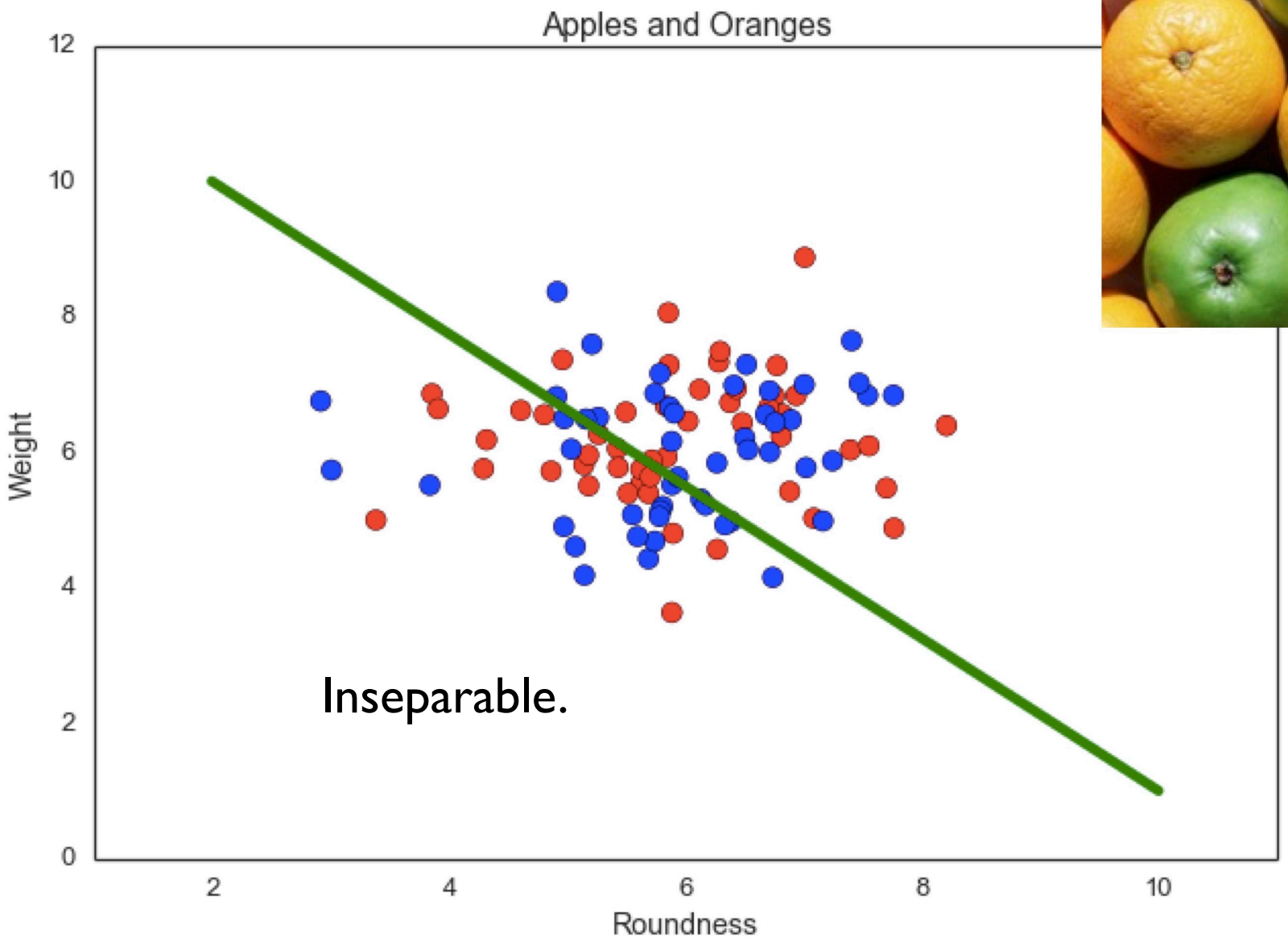
x: data points

y: labels

features

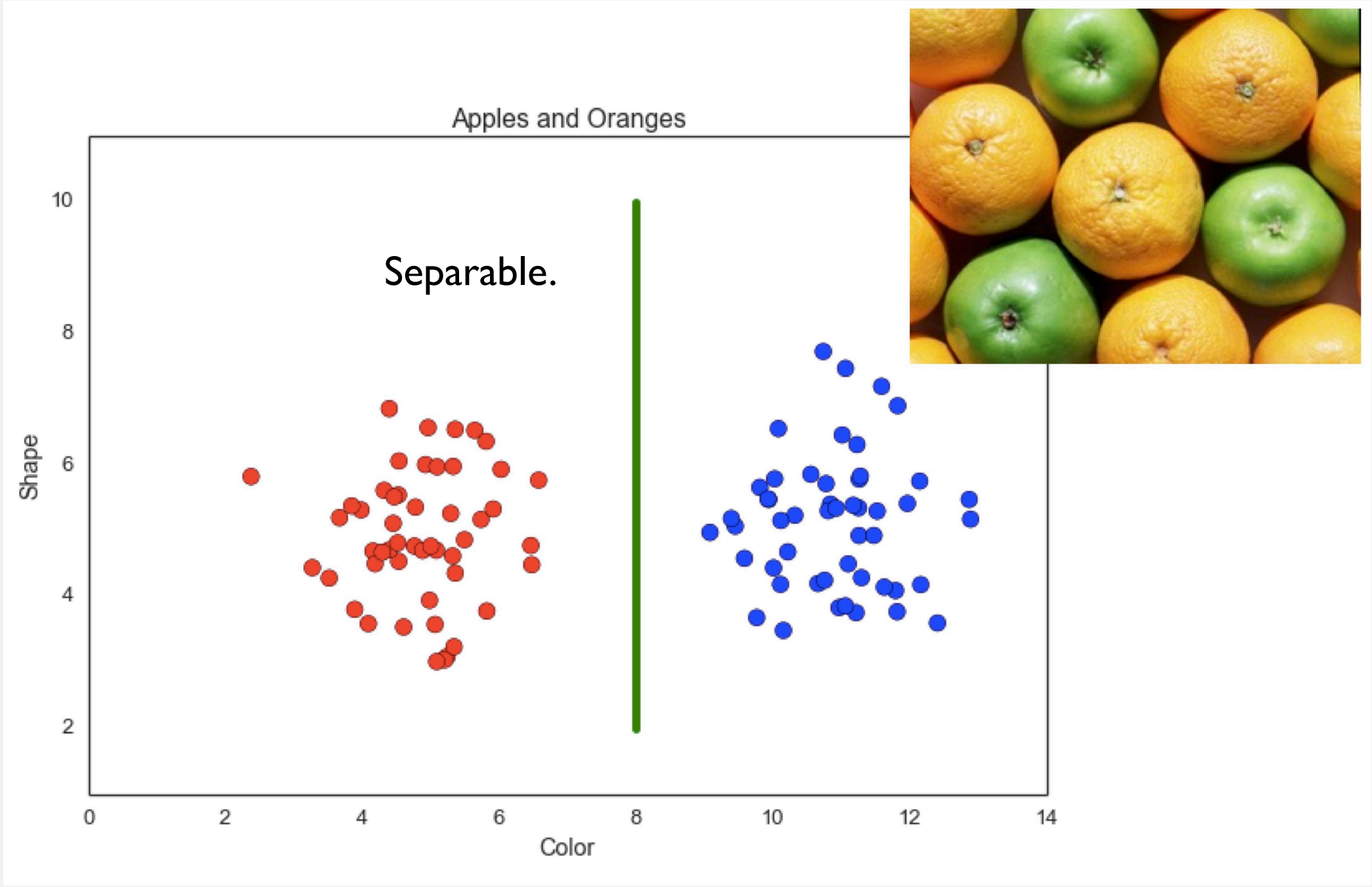
decision boundary

http://cs109.github.io/2015/pages/videos.html



Inseparable.

http://cs109.github.io/2015/pages/videos.html



http://cs109.github.io/2015/pages/videos.html

Ok, let's go do some machine learning in python.