Data Science at the Davidson School of Chemical Engineering

Curtis P. Martin

13 Nov. 2019

AIChE National Meeting

Data Science In ChE - 20433 - CHE 59700 - 023

Scheduled Meeting Times

Type	Time	Days	Where	Date Range	Schedule Type	Instructors
Class	10:30 am - 11:45 am	TR	Physics Building 111	Aug 19, 2019 - Dec 07, 2019	Lecture	Curtis Patrick Martin (P), Brett Savoie

- One piece in larger school & university initiatives
  - Concentrations for undergraduate & PMP; data science co-op
  - Designed for graduate students & higher-level undergraduates
  - Supplemental non-CS courses popping up around campus

- Focused on effective, responsible application of machine learning
  - Broad survey of methods & applications in chemical engineering. . .
  - . . . deep understanding of strengths, weaknesses & cycle
  - Fundamentals, tools & process

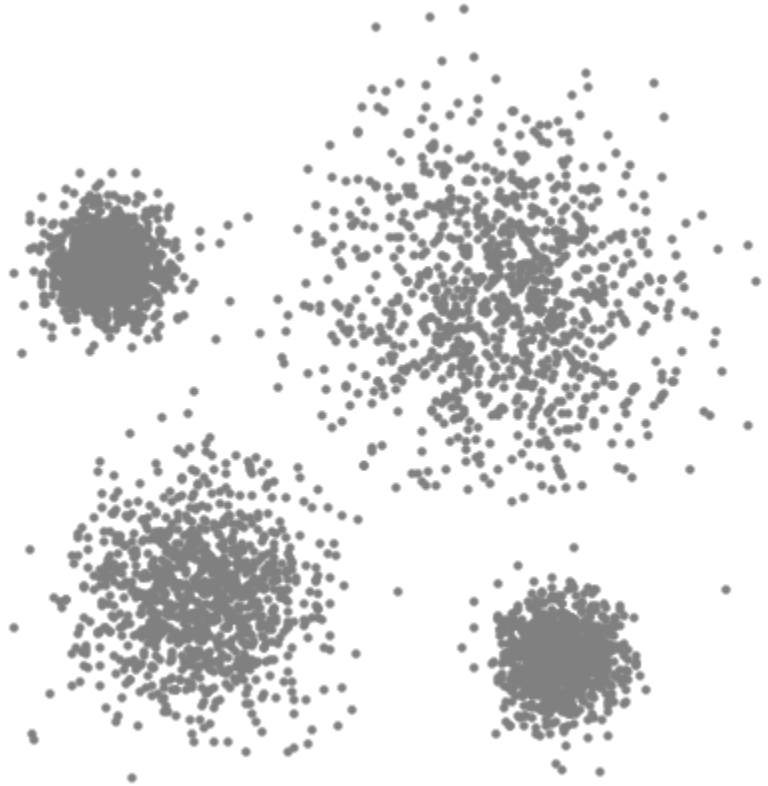
Course Material I - Fundamentals

- Programming proficiency
  - Python (via Jupyter notebook) best option for implementing machine learning quickly
  - Pieces of process discussed concurrently (e.g., formatting, dealing w/ missing data)

- Probability & optimization
  - ML algorithms are often derived from or built on concepts in probability & optimization. . .
  - . . . but generally don't require you to be an expert in either

Course Material II - Tools

- Unsupervised machine learning
  - Dimensionality reduction (principal components analysis)
  - Clustering (k-means, DBSCAN, mixture models, agglomerative)



https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
Kingsford, C. & Salzberg, S. L. "What are decision trees?" Nature Biotech. 26(9), 2008.

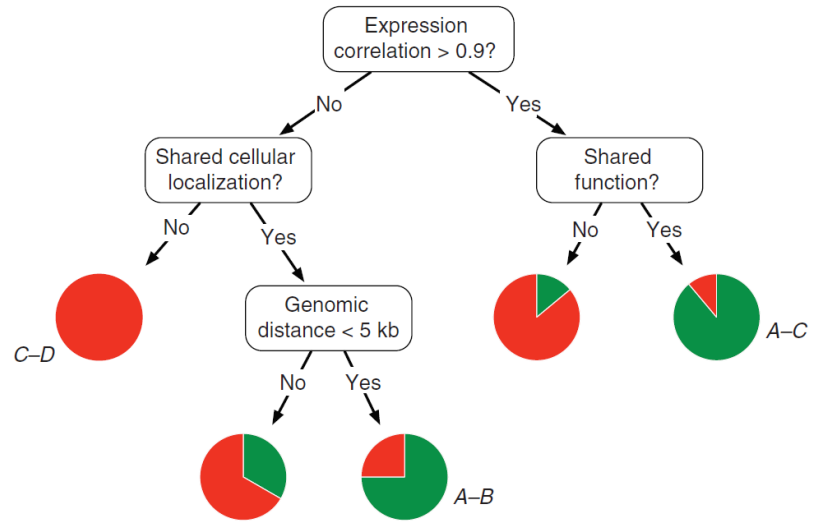
Course Material II - Tools

- Unsupervised machine learning
  - Dimensionality reduction (principal components analysis)
  - Clustering (k-means, DBSCAN, mixture models, agglomerative)
- Supervised machine learning
  - Regression (multiple linear, regularized variants, PLS)
  - Classification (logistic regression, naïve Bayes, nearest neighbors, decision trees & random forests, SVM, neural networks & variants)

a

Gene Pair	Interact?	Expression correlation	Shared localization?	Shared function?	Genomic distance
A-B	Yes	0.77	Yes	No	1 kb
A-C	Yes	0.91	Yes	Yes	10 kb
C-D	No	0.1	No	No	1 Mb

b



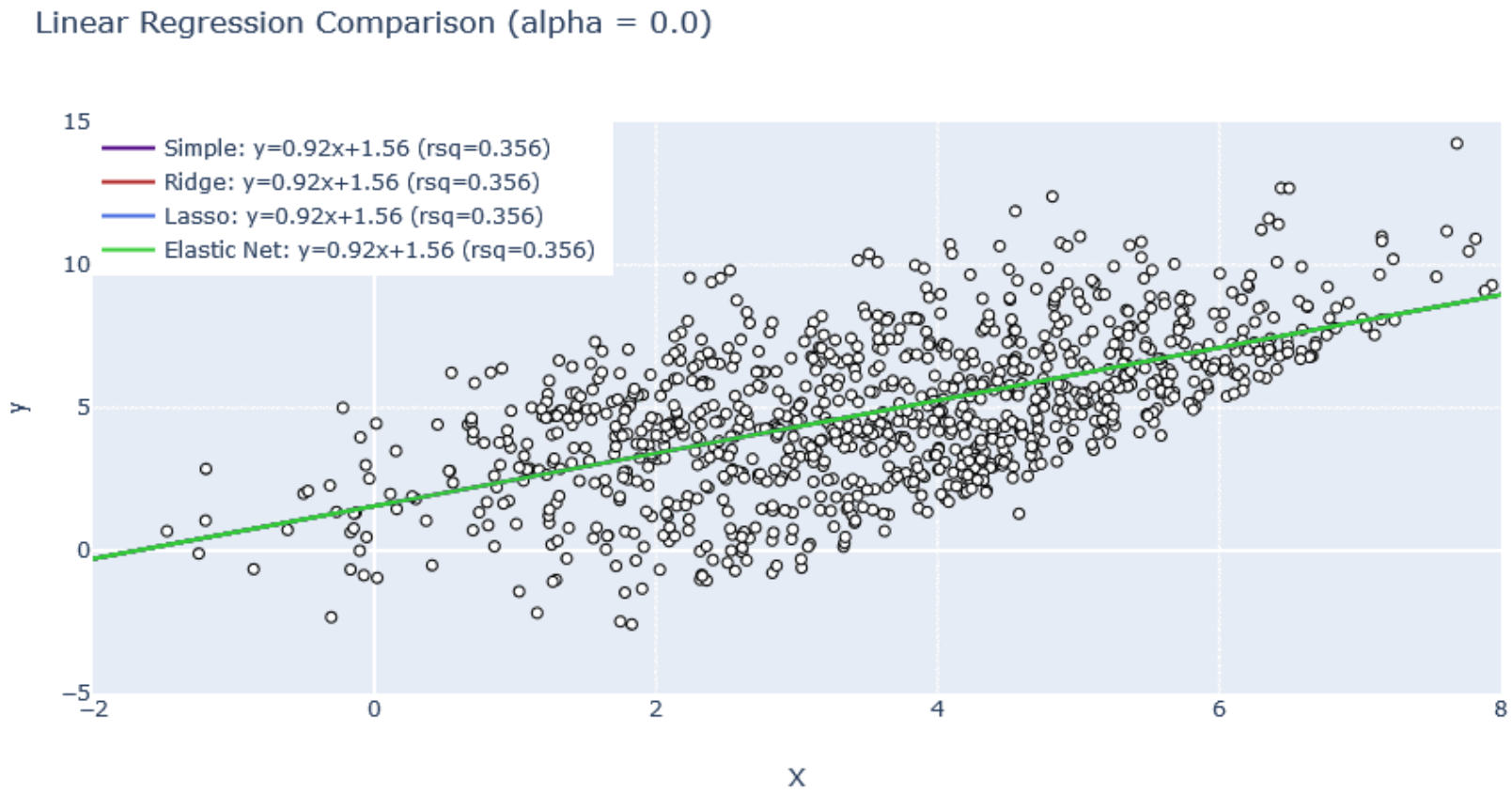
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
Kingsford, C. & Salzberg, S. L. "What are decision trees?" Nature Biotech. 26(9), 2008.

Course Material III - Process

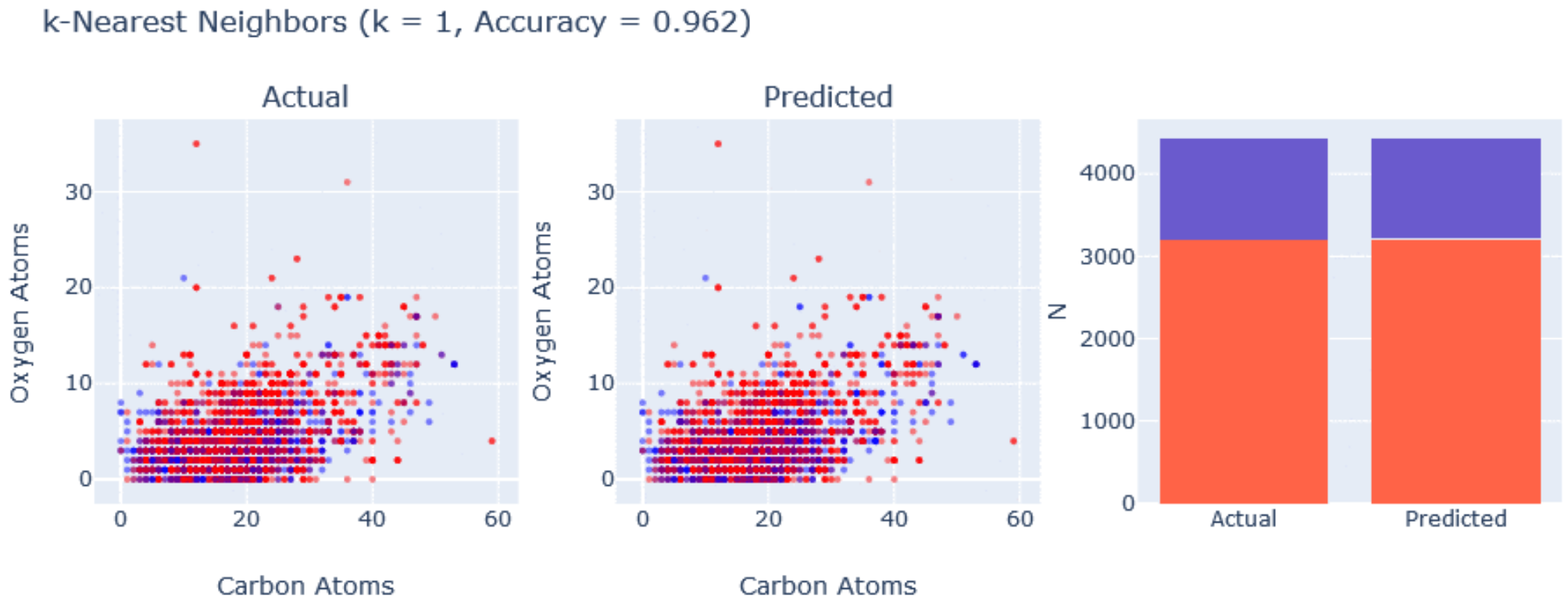
- Model selection
  - Strengths & weaknesses of algorithms; how do they compare
  - Under what circumstances are they likely to be effective
  - Effects of hyperparameters on performance

Course Material III - Process

- Mode
  - Str
  - Un
  - Effe



Course Material III - Process



Course Material III - Process

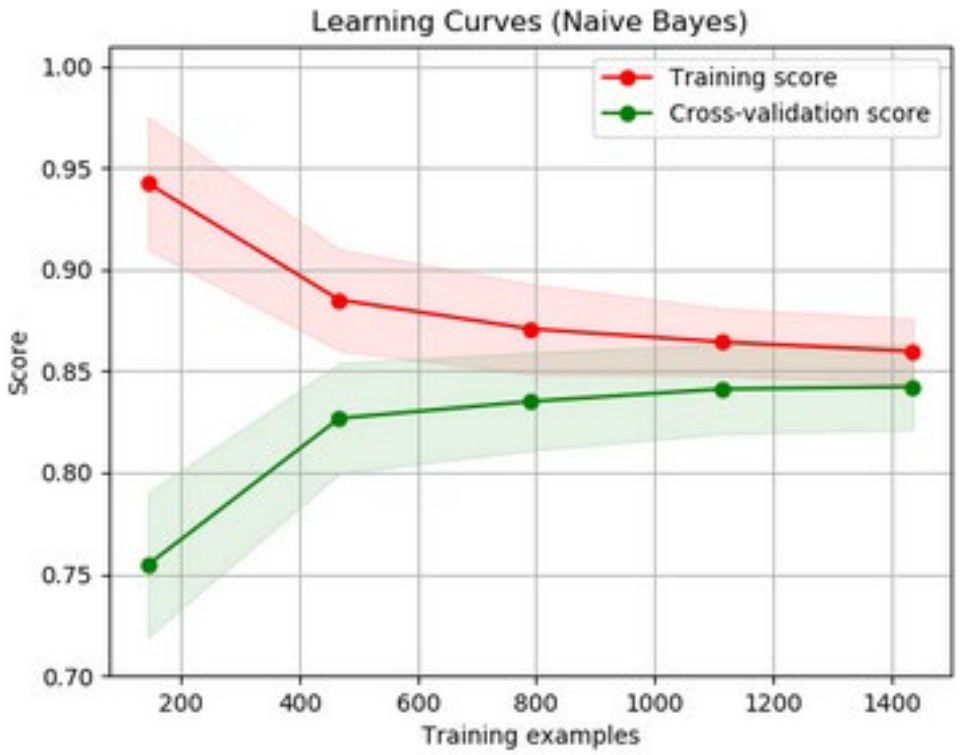
- Model selection
  - Strengths & weaknesses of algorithms; how do they compare
  - Under what circumstances are they likely to be effective
  - Effects of hyperparameters on performance

- Model validation
  - Performance estimation & how metrics affect perception of results
  - Likelihood of generalization
  - Common issues (e.g., overfitting, bias-variance tradeoff, curse of dimensionality)

Figure pulled from nice scikit-learn website on learning curves: https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html#sphx-glr-auto-examples-model-selection-plot-learning-curve-py

Course Material III - Process

- |
- |



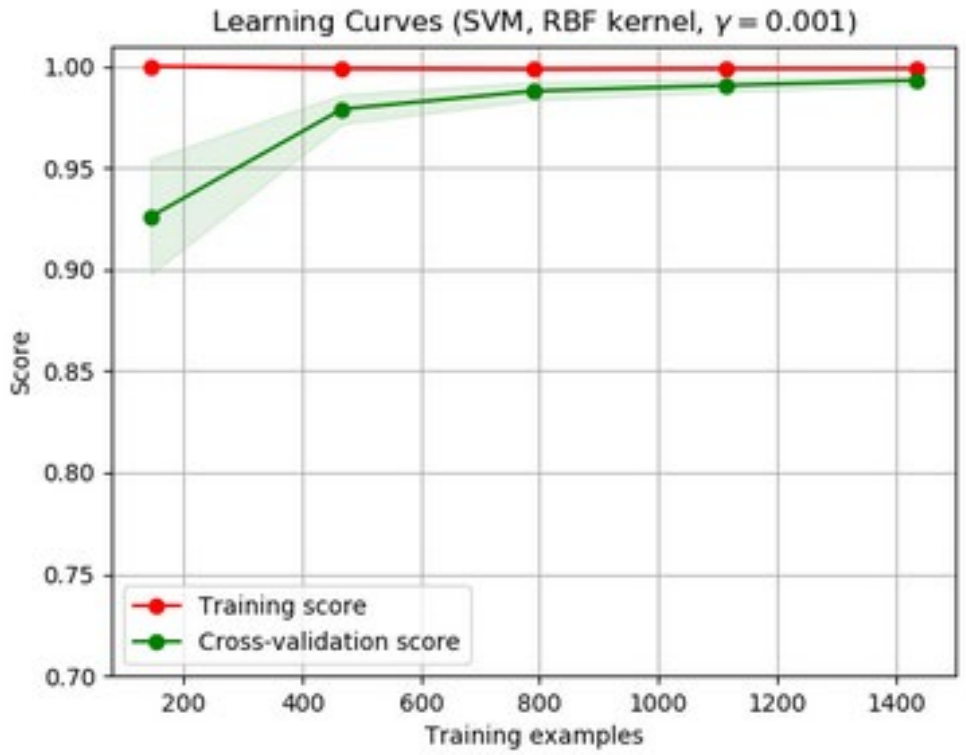
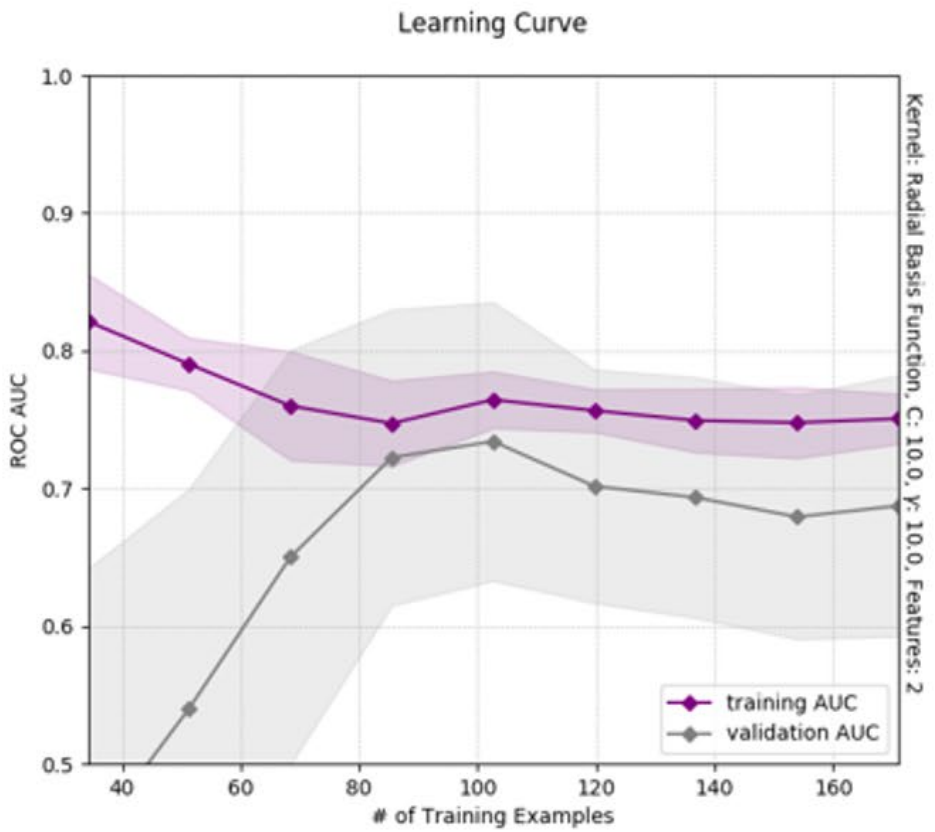


Figure pulled from nice scikit-learn website on learning curves: https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html#sphx-glr-auto-examples-model-selection-plot-learning-curve-py

Course Material III - Process

- |
- |



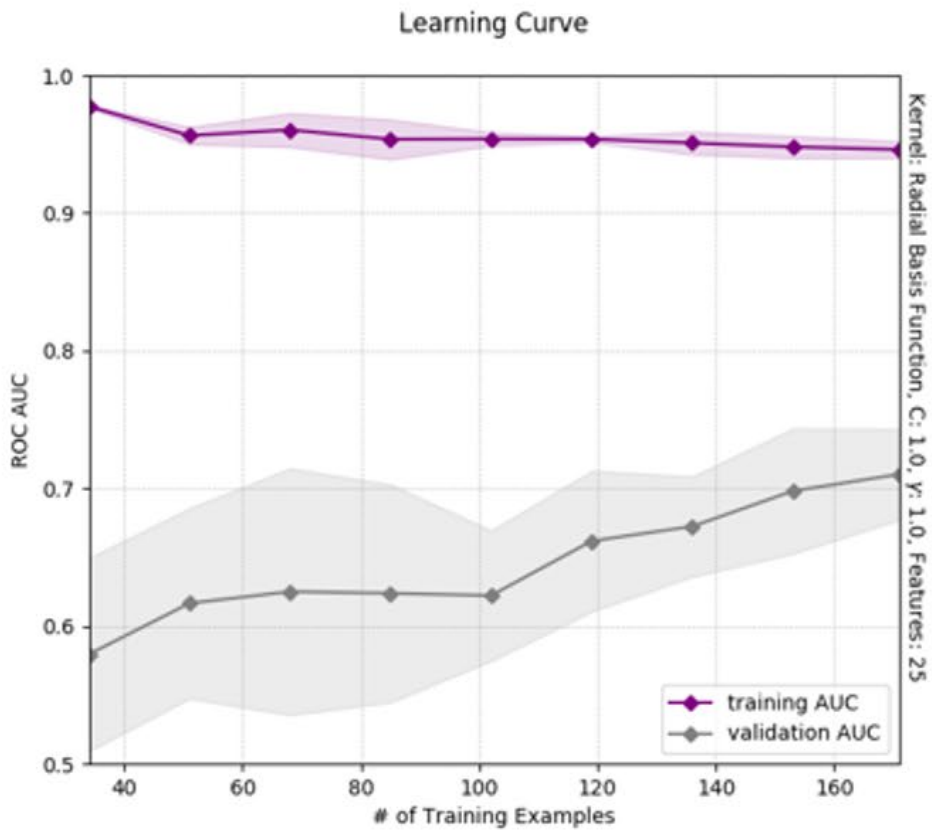
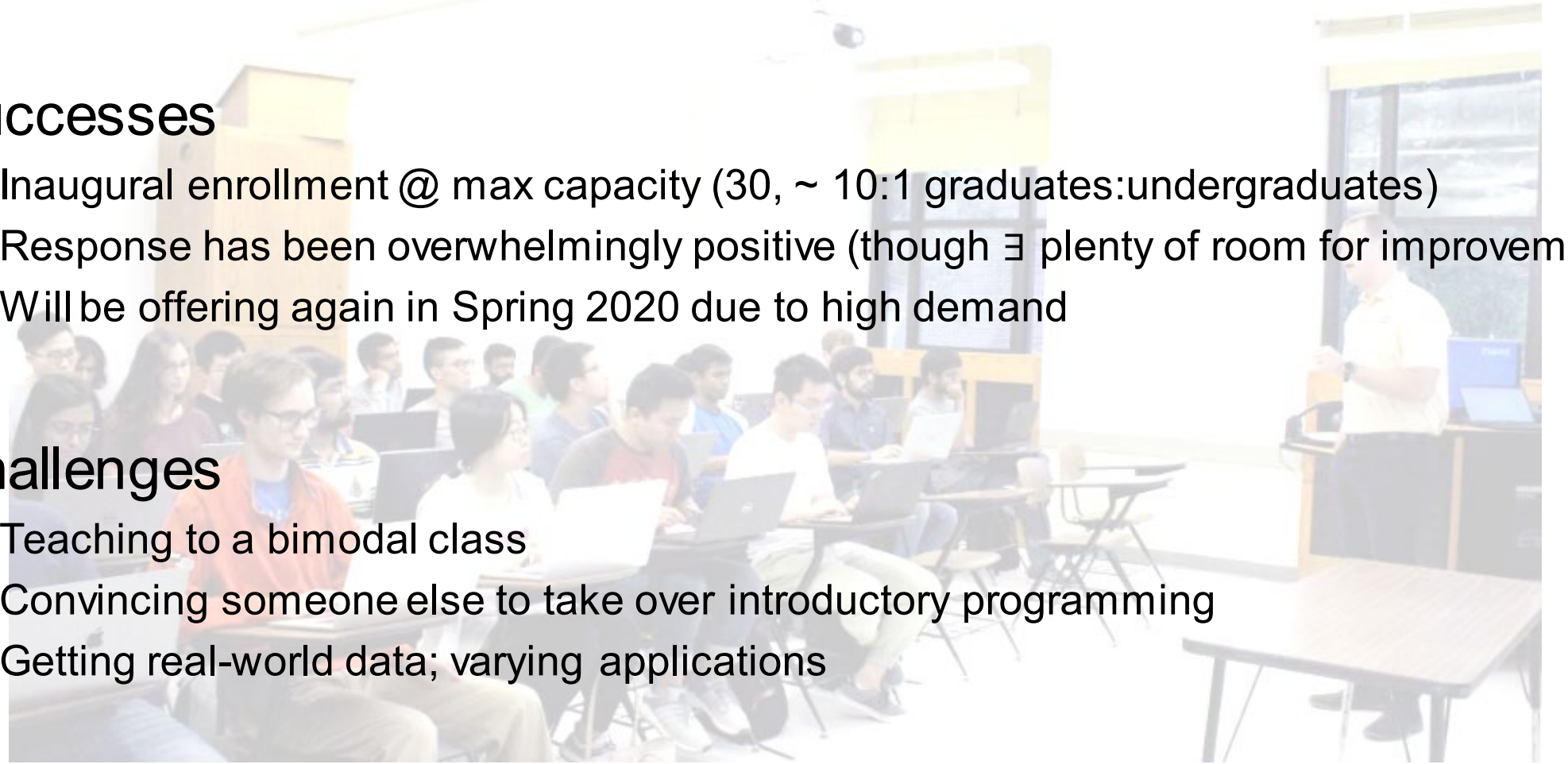


Figure pulled from nice scikit-learn website on learning curves: https://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html#sphx-glr-auto-examples-model-selection-plot-learning-curve-py

The Experience Thus Far



- Successes
  - Inaugural enrollment @ max capacity (30, ~ 10:1 graduates:undergraduates)
  - Response has been overwhelmingly positive (though ∃ plenty of room for improvement)
  - Will be offering again in Spring 2020 due to high demand
- Challenges
  - Teaching to a bimodal class
  - Convincing someone else to take over introductory programming
  - Getting real-world data; varying applications

Thanks to Many

DOW®

Lilly

P&G

RA Rockwell Automation

OSIsoft.

ExxonMobil