SI 618
Exploratory Data Analysis

Machine Learning

Review and Q&A

Instructor: Dr. Chris Teplovs (cteplovs@umich.edu)
Lead Developer, Digital Innovation Greenhouse, Office of Academic Innovation
Adjunct Lecturer of Information, School of Information
GSI: SungJin Nam (sjnam@umich.edu)

Reminder

- Final project report due Friday, Dec 16, 1:00pm (late days can apply)
- Homework 5 (Factor analysis) is an optional bonus assignment, due Friday, Dec. 16, 1:00 pm
  - Up to +10% on course grade

We're hiring!

- A Data Science Student Fellow: http://ai.umich.edu/students/student-opportunities/

SI 618 Data Exploration: Class Schedule

Date	Topic	Assignments Due
Week 1	Course introduction Basics of Programming with R	
Week 2	Basic analysis and visualization using ggplot2: qplot() Manipulating data frames using plyr	Homework 1
Week 3	Smoothing and Trend-finding. Building ggplot Layer by Layer	Homework 2
Week 4	Cluster analysis	Homework 3
Week 5	(Thanksgiving: no class!)	
Week 6	Factor Analysis Methods (PCA, EFA)	Homework 4
Week 7	Machine Learning, Review, Evaluations	

Class Schedule for Today

- Machine Learning:
  - Introduction
  - Graphical Overview
- MonkeyLearn
- Review
- Q&A
- Evaluations

Machine Learning

- From Wikipedia:
  - a subfield of computer science that "gives computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959).
- Classification, regression, and clustering are common goals
- Pattern recognition, text and speech processing

Machine Learning is “HOT”

- Lots of interest in ML:
  - Diverse fields: autonomous vehicles, learning analytics, business intelligence
  - Anything that has to do with predictive modelling

Machine Learning is a HUGE field

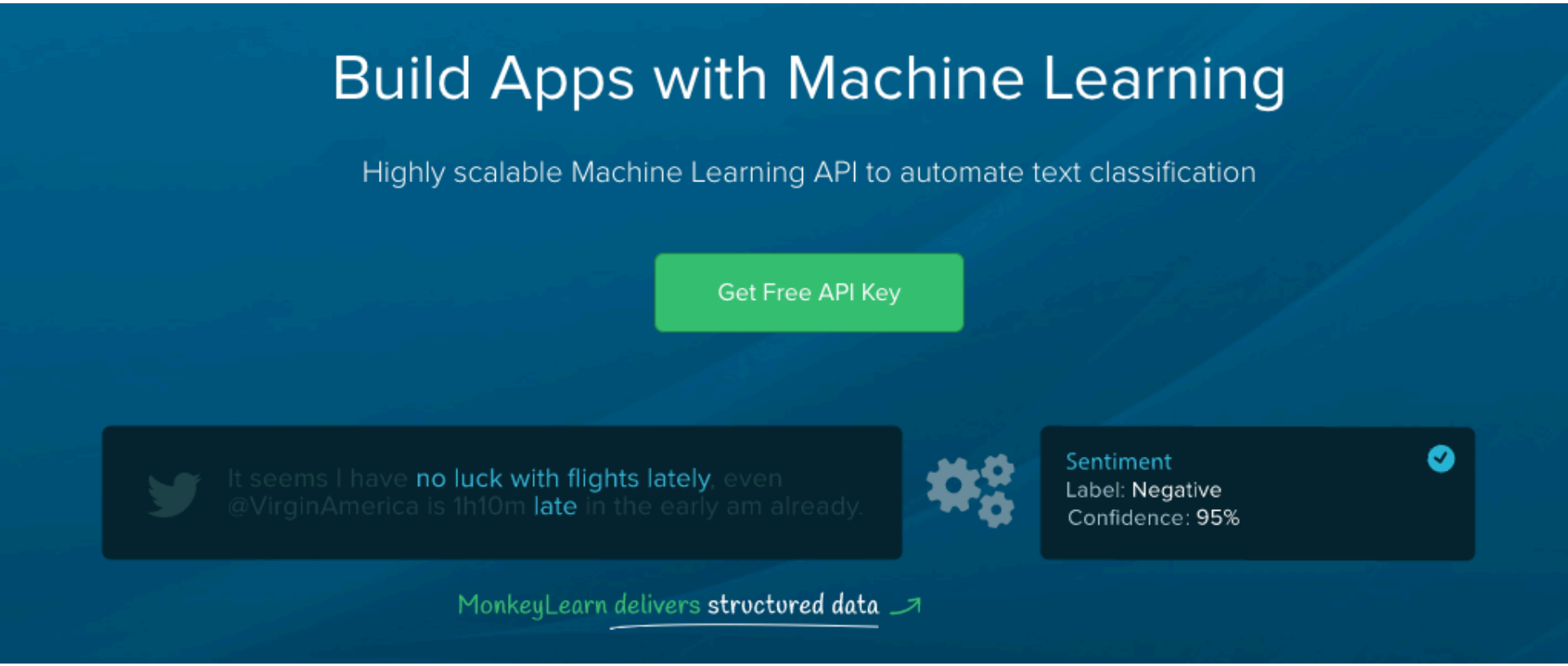
- Entire courses, programs, fields of study
- We are just going to get a general idea of ML, and then move onto some very practical examples using a third-party API provider
- If you're interested in ML consider taking one of the MOOCs (check out Coursera or edX)

A Visual Introduction to Machine Learning

- www.r2d3.us/

Leveraging a Machine Learning Service

- MonkeyLearn: monkeylearn.com



Some ideas:

- Generic Topic Classifier
- Business Classifier
- Hotel Review Sentiment Analysis
- Restaurant Review Sentiment Analysis

On to a practical example (#1)

- Can you deduce sentiment without actually running a sentiment analysis?
  - https://app.monkeylearn.com/main/classifiers/cl_qkjxv9Ly/tab/classify-sandbox/

Practical Example #2

- How can you characterize a twitter user?
  - By their followers' bios?
  - https://blog.monkeylearn.com/know-followers-machine-learning/

Where to go from here:

- Play around with MonkeyLearn
  - See "Understanding Users Through Twitter Data and Machine Learning" blog post
- Look at Machine Learning in R and python:
  - Variety of packages in R:
    - https://www.datacamp.com/community/tutorials/machine-learning-in-r
  - Scikit-learn for Python
- Weka: http://www.cs.waikato.ac.nz/ml/weka/
- Google Brain Team's TensorFlow: https://www.tensorflow.org/

SI 618 Data Exploration: Class Schedule

Date	Topic	Assignments Due
Week 1	Course introduction Basics of Programming with R	
Week 2	Basic analysis and visualization using ggplot2: qplot() Manipulating data frames using plyr	Homework 1
Week 3	Smoothing and Trend-finding. Building ggplot Layer by Layer	Homework 2
Week 4	Cluster analysis	Homework 3
Week 5	(Thanksgiving: no class!)	
Week 6	Factor Analysis Methods (PCA, EFA)	Homework 4
Week 7	Machine Learning, Review, Evaluations	

We've covered a lot of computational ground!

Skills:

Compute and visualize a dataset's key summary statistics

Explore relationships between variables

Find trends over time

Discover clusters and outliers

Use factors to analyze underlying variables in data

Produce polished presentations for publication/display

How to apply R coding and packages to solve the above problems

.. And much more!

Tools:

The R language

RStudio integrated development environment

RMarkdown authoring tool

ggplot2 visualization package

Several other useful R packages: SQL access, plyr