DS - 1001

BRIAN D'ALESSANDRO

VP – DATA SCIENCE, DSTILLERY

ADJUNCT PROFESSOR, NYU

ME

Brian d'Alessandro

Bio

Education:
Undergrad: Rutgers, Math
Grad: NYU Stern, Statistics

Professional Experience
Dstillery (AdTech)
Meetup.com (Social Web)
American Express (Credit/Risk)
TV Guide (Marketing)

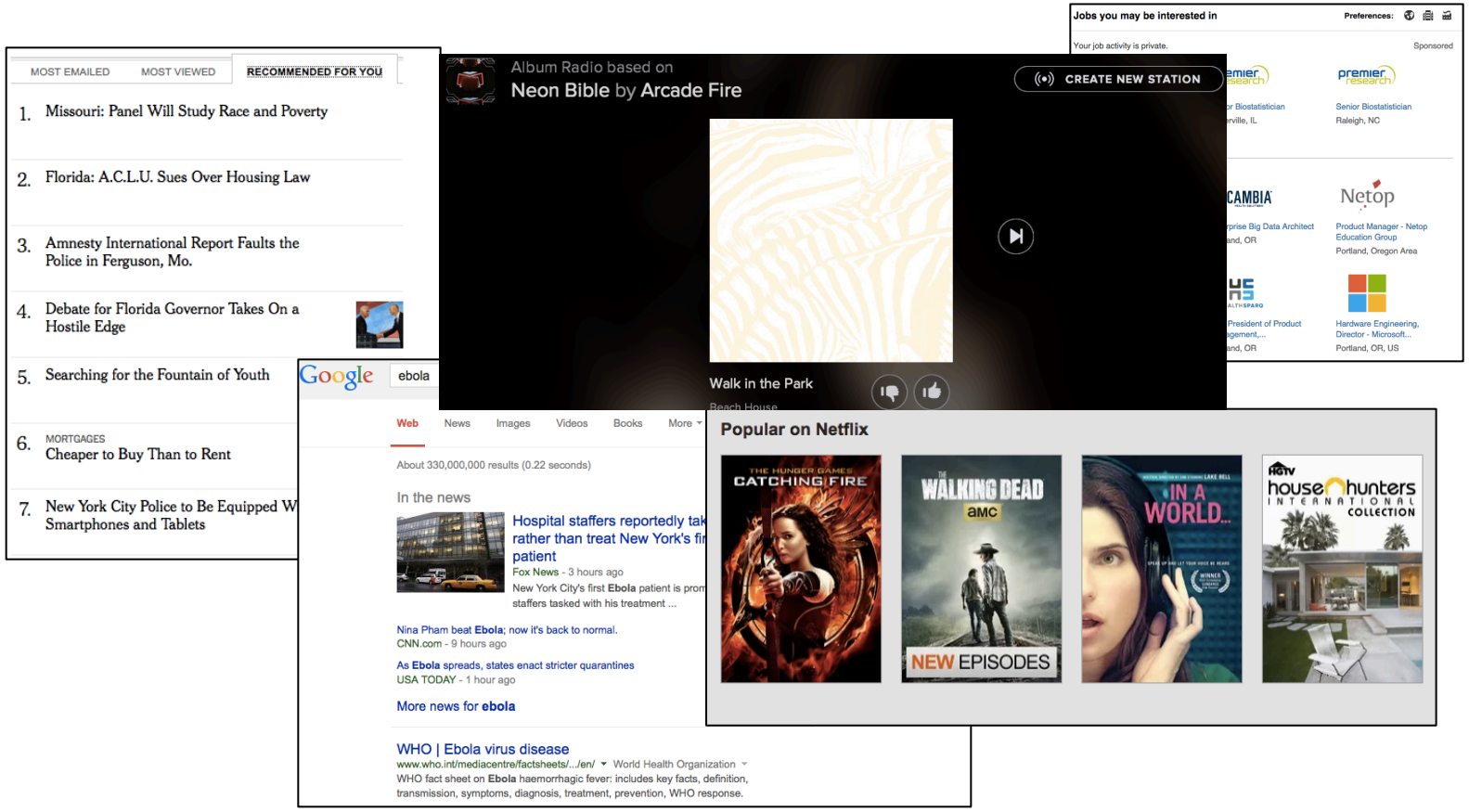
Affiliations/Publications
ACM KDD
Big Data Journal
Machine Learning Journal
SIAM

GOAL1:
DEFINE DATA SCIENCE

DATA SCIENCE IS EVERYWHERE

If you use the internet, you likely suffer from this little problem - *too much information and too little time.*

Most companies try to solve this problem for you using data science



THE MAGAZINE

October 2012

Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

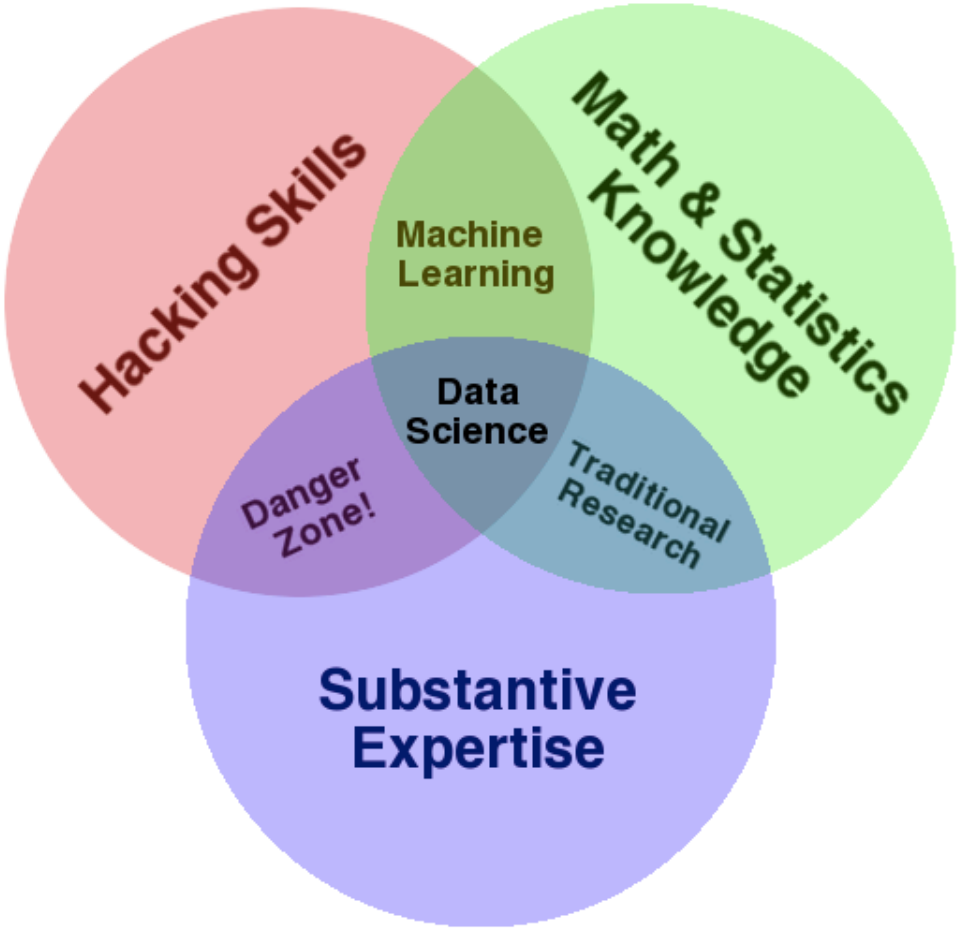
"Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions.

They find the story buried in the data and communicate it. And they don't just deliver reports:

They get at the questions at the heart of problems and devise creative approaches to them."

http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

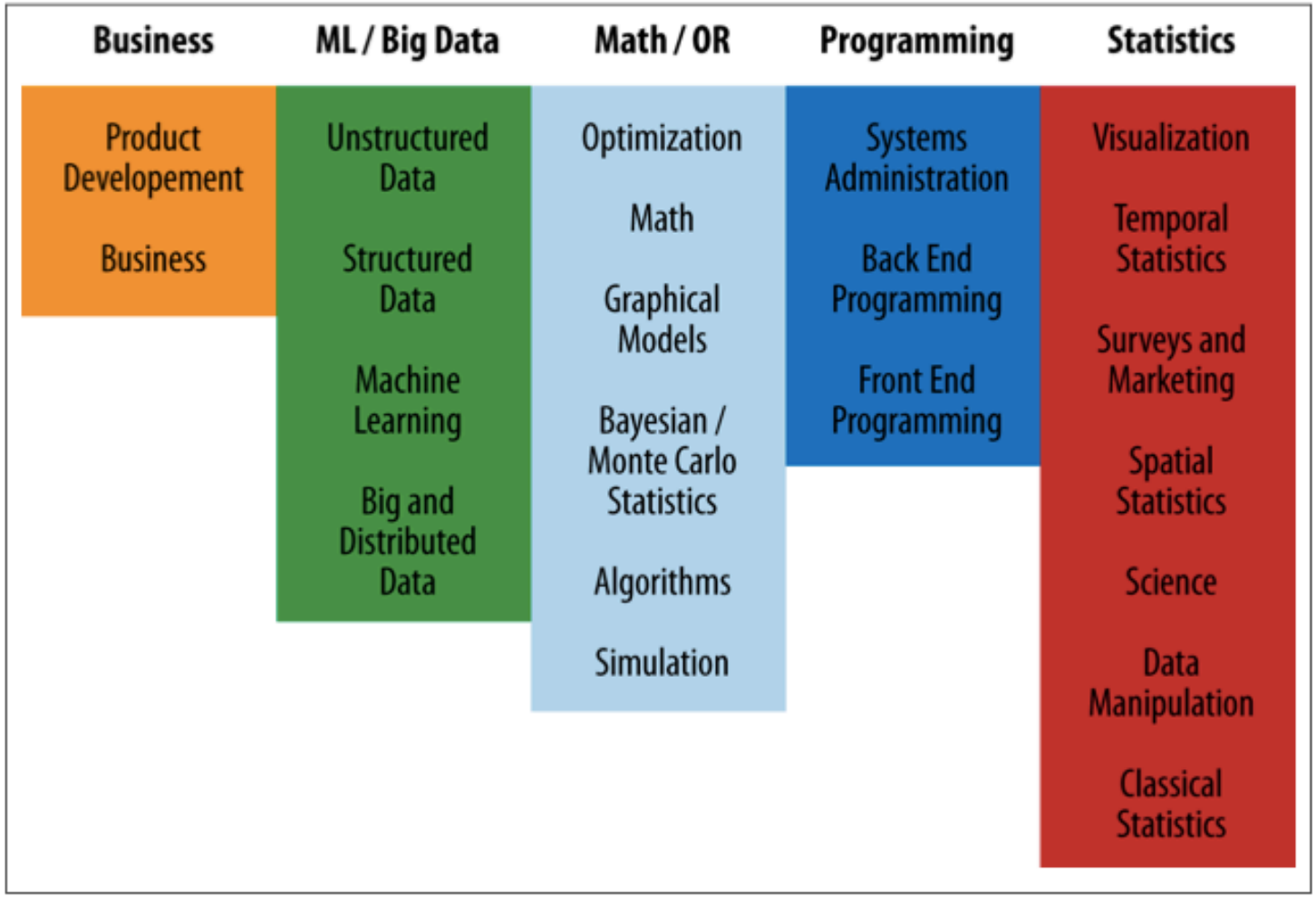
DS IS THE CONFLUENCE OF MANY DISCIPLINES



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

RANGE OF DS SKILLS

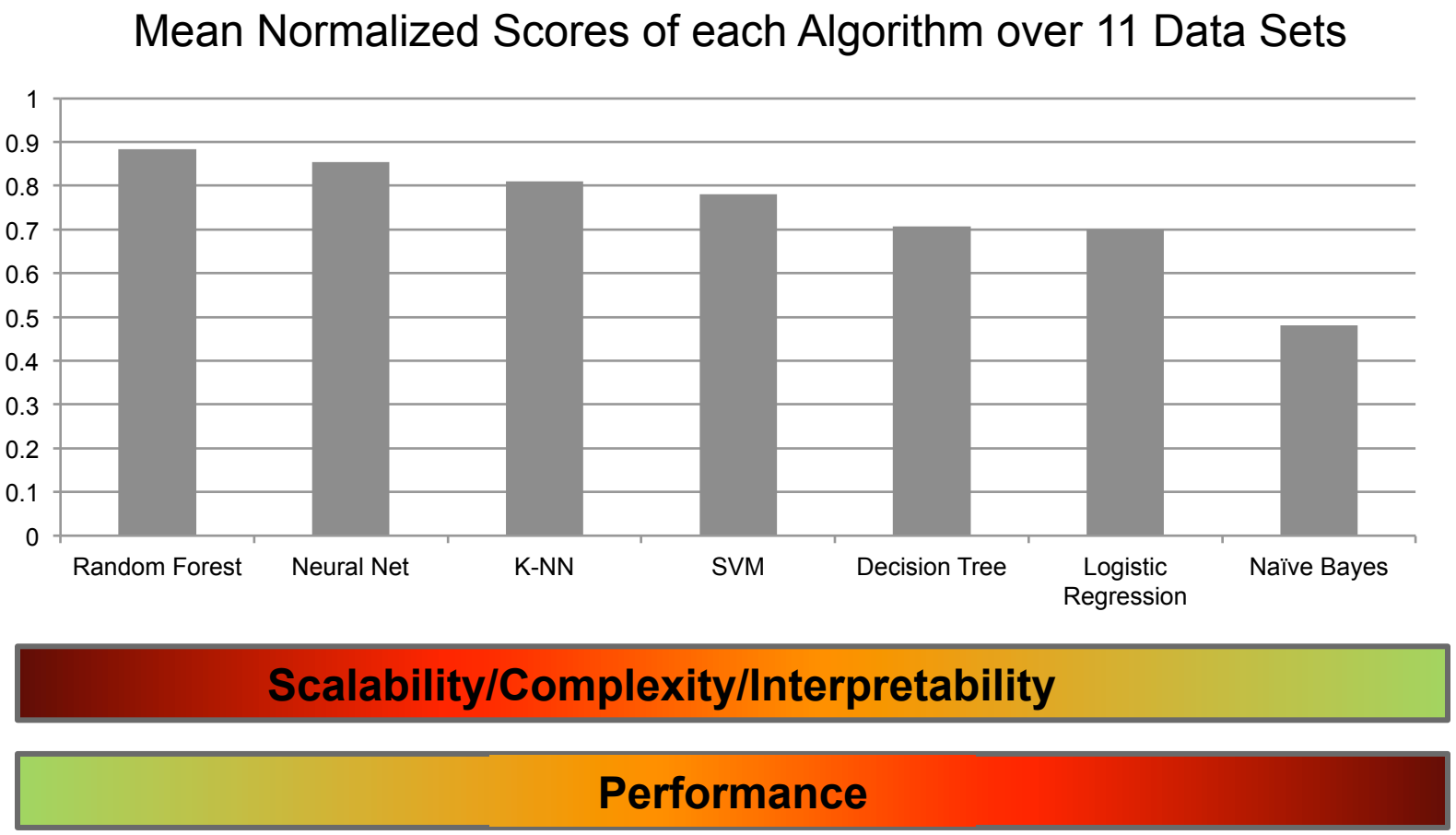
In this class we will develop a foundation for applying all of these skills.



Source: http://www.oreilly.com/data/free/analyzing-the-analyzers.csp

GOAL2:
START BUILDING YOUR TOOL CHEST

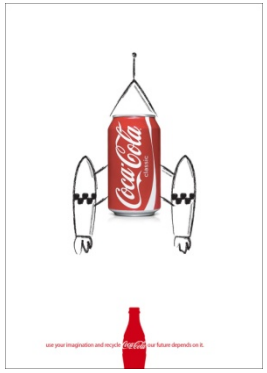
INTRODUCE COMMON ALGORITHMS



Source: An Empirical Comparison of Supervised Learning Algorithms http://www.niculescu-mizil.org/papers/comparison.tr.pdf

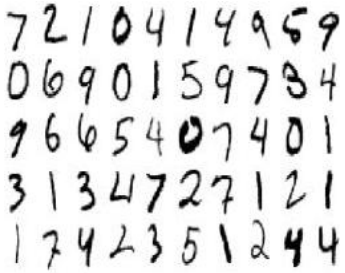
LEARN WHEN TO USE THEM

Will someone click on an ad?:
C=[No,Yes]

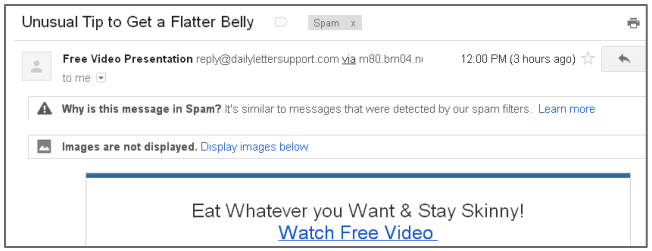


Is this pill good for headaches?:
C=[No,Yes]

What number is this?:
C=[0,1,2,3,4,5,6,7,8,9]



Is this e-mail spam?: *C=[No,Yes]*

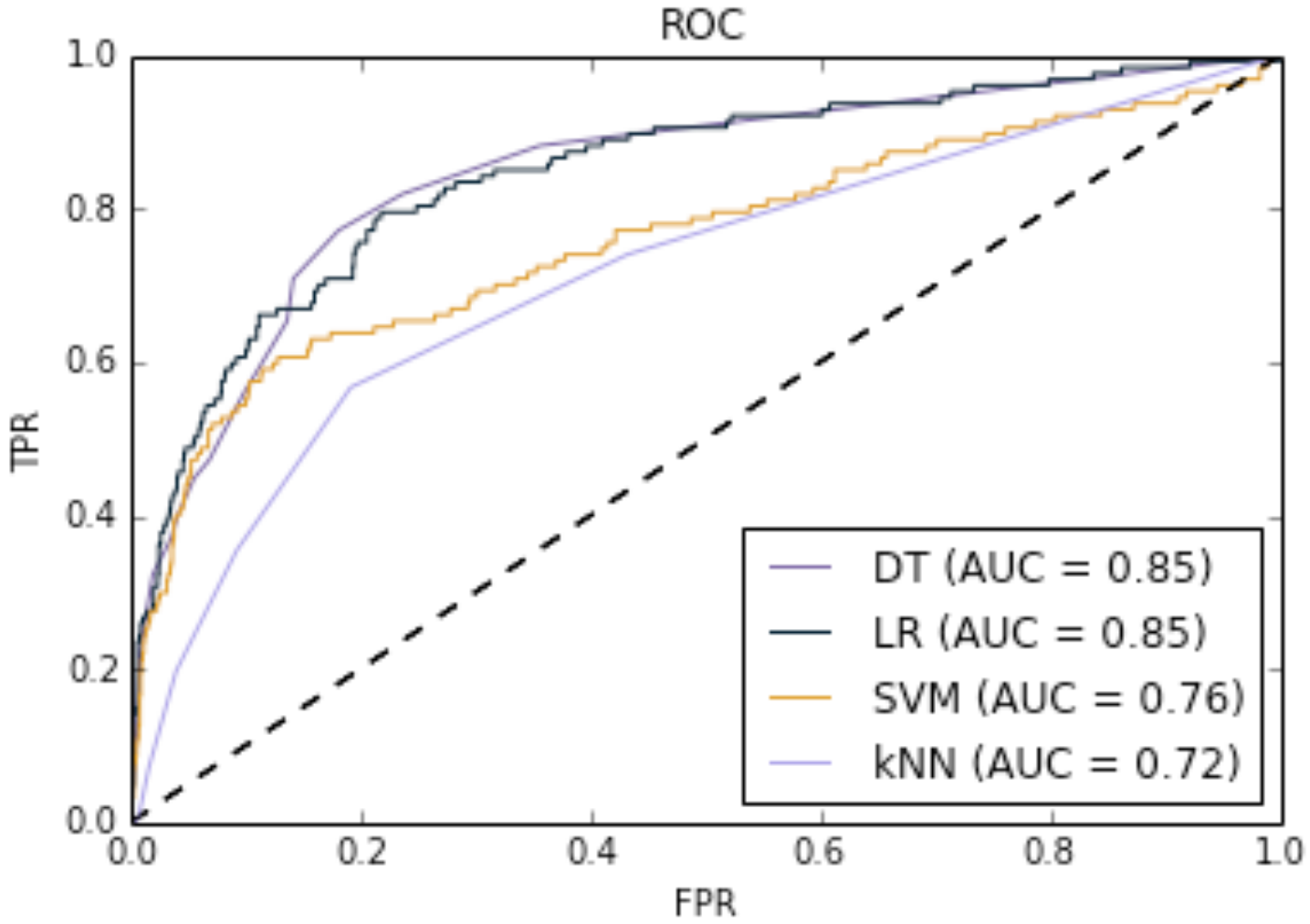


What is this news article about?:
C=[Politics, Sports, Finance …]

LEARN HOW TO USE THEM

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.tree import DecisionTreeClassifier

dt = DecisionTreeClassifier()
dt = dt.fit(X, Y)

lr = linear_model.LogisticRegression()
lr.fit(X, Y)

mm = svm.SVC(kernel='linear')
mm.fit(X, Y)

knn = KNeighborsClassifier(n_neighbors=10)
knn.fit(X, Y)
```

AND HOW TO EVALUATE THEM EMPIRICALLY



THIS IS NOT A THEORY COURSE

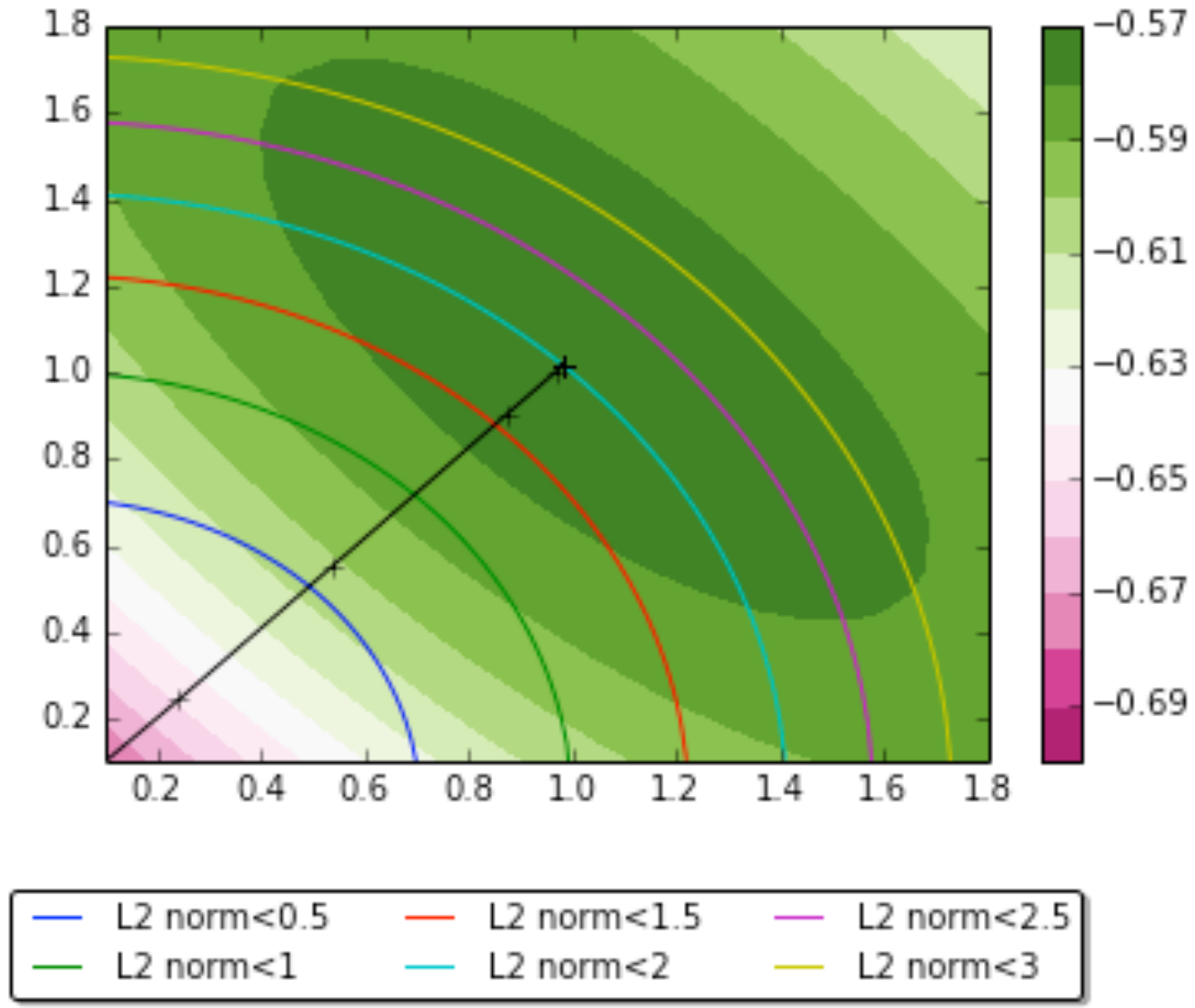
$$E[f] \leq \hat{E}_S[f] + 2\mathcal{R}_m(\mathcal{F}) + O\sqrt{\frac{ln(\frac{1}{\delta})}{m}}$$

$$E[f] \leq \hat{E}_S[f] + 2\hat{\mathcal{R}}_S(\mathcal{F}) + O\sqrt{\frac{ln(\frac{1}{\delta})}{m}}$$

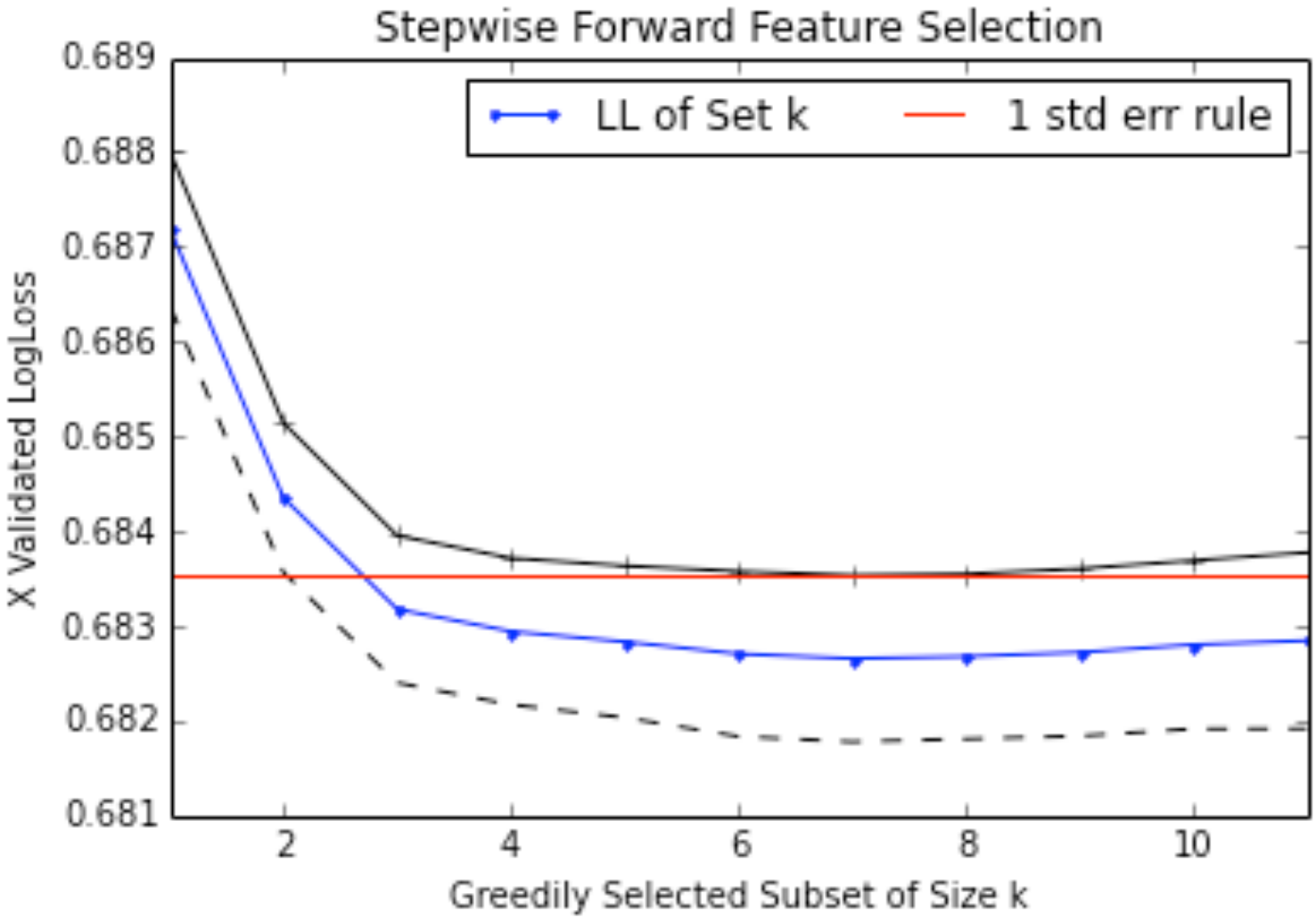
(BUT WE WILL DEVELOP AN INTUITION BEHIND KEY THEORETICAL CONCEPTS)

SUCH AS BIAS-VARIANCE TRADEOFFS

REGULARIZATION



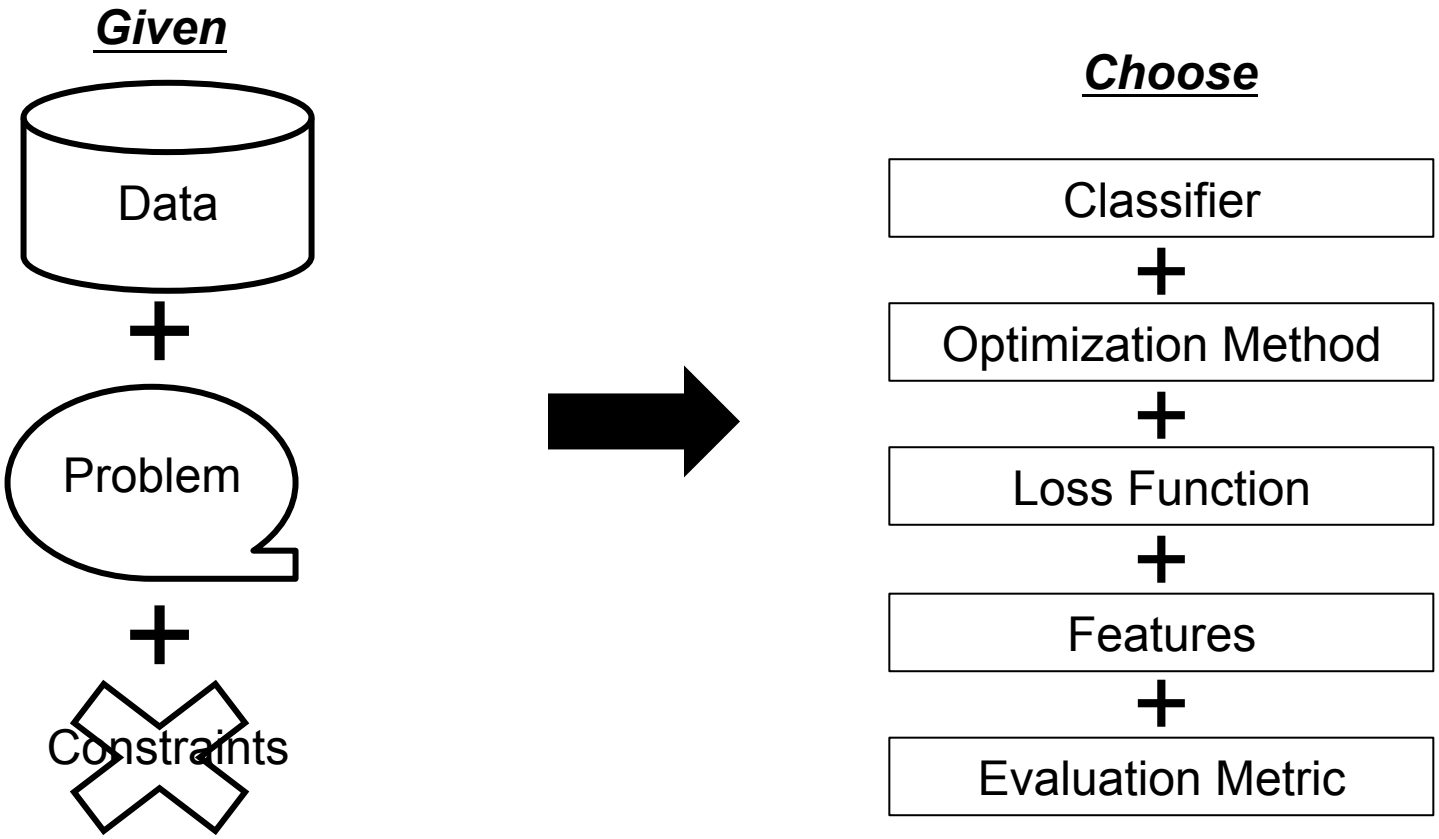
AND MODEL SELECTION



GOAL 3:
BETTER DATA DRIVEN DECISIONS

A COMMON THEME

Few problems have out of the box solutions



The Data Scientist has to navigate these choices

BECOMING A SCIENTIST

<u>The scientific method:</u> evaluating the merit of a hypothesis with rigorous empirical testing.

I.e.,

Given raw data, constraints and a problem statement, you have an infinite set of models to choose from, with which you will use to maximize performance on some evaluation metric, that you will have to specify.

Every design choice you make can be formulated as a hypothesis, upon which you will use rigorous testing and experimentation to either validate or refute.

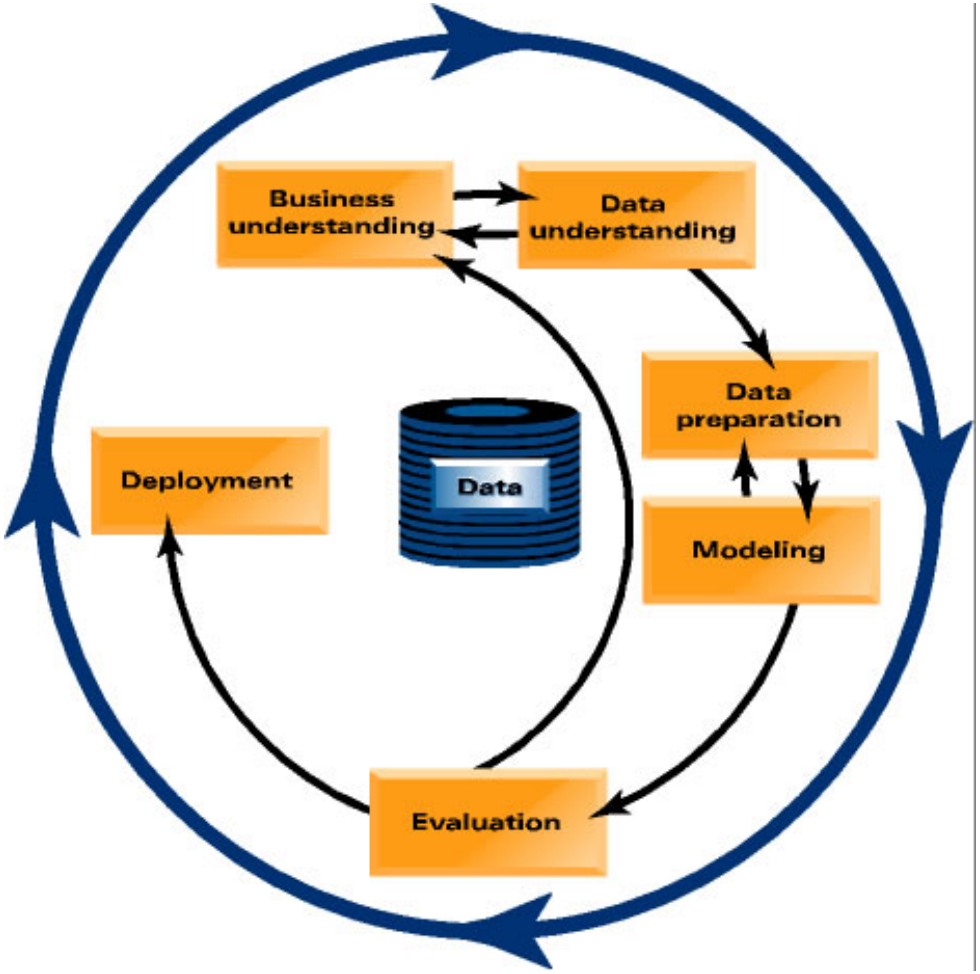
BUT ITS STILL AN ART

Outside of modeling competitions, seldom is a well-posed problem and clean dataset presented to you.

Putting the art into your practice means…

- Translating problems into the language of data science
- Formulating reasonable hypotheses
- Developing an intuition for good vs. bad data, good vs. bad models.
- Abstracting problems to identify similarities
- Managing the DS process from end to end

MAKING DECISIONS WITH THE BIG PICTURE IN MIND



GOAL 4:
PREPARATION

FOR THE NEXT TWO YEARS

Year 1 – Fall

Course Title	Credits
DS-GA-1001 Intro to Data Science	3
DS-GA-1002 Statistical and Mathematical Methods for Data Science	3
Data Science Elective 1	3
TOTAL CREDITS	**9**

Year 1 – Spring

Course Title	Credits
DS-GA-1003 Machine Learning and Computational Statistics	3
DS-GA-1004 Big Data	3
Data Science Elective 2	3
TOTAL CREDITS	**9**

Year 2 – Fall

Course Title	Credits
DS-GA-1005 Inference and Representation	3
DS-GA-1006 Capstone Project in Data Science	3
Data Science Elective 3	3
TOTAL CREDITS	**9**

Year 2 – Spring

Course Title	Credits
Data Science Elective 4	3
Data Science Elective 5	3
Data Science Elective 6	3
TOTAL CREDITS	**9**

YOUR FIRST INTERNSHIP

A Typical Data Scientist Job Description

Requirements

- Ph.D. in a relevant technical field, or 4+ years experience in a relevant role
- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- Ability to communicate complex quantitative analysis and analytic approaches in a clear, precise, and actionable manner
- Fluency with scripting languages such as Python, Ruby, or PHP
- Familiarity with relational databases and SQL-like query languages
- Expert knowledge of a scientific computing language such as R, Python, or Julia
- Experience working with data-distributed query tools a plus (Hadoop, Hive, Presto, etc.)