

# **Introduction to Data Science**

**COURSE FUNDAMENTALS**

**BRIAN D'ALESSANDRO**

*Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.*

# INSTRUCTOR

Brian d'Alessandro



## Bio

*Education:*

Undergrad: Rutgers, Math  
Grad: NYU Stern, Statistics

*Professional Experience*

Capital One (Credit/Risk)  
ZocDoc (Healthcare/Tech)  
Facebook (Social/Tech)  
Dstillery (Advertising/Tech)  
Meetup.com (Social/Tech)  
American Express (Credit/Risk)

*Affiliations/Publications*

ACM KDD  
Big Data Journal  
Machine Learning Journal  
SIAM

# LEARNING OBJECTIVES OF THIS SERIES

Design effective ML solutions, by...

## Preparing data for modeling

1. Constructing meaningful and unbiased labels and
2. Sampling without bias
3. Feature Engineering, Regularization and the avoidance of leakage
4. Data splitting for out-of-sample valuation

## Building generalizable models

1. Identifying the appropriate set of algorithms for a given problem statement observing problem constraints
2. Finding an optimal model fit by tuning hyper-parameters on a variety of classification algorithms

## Optimizing Decision Systems

1. Setting up ML research experiments
2. Identifying the appropriate evaluation metric based on a given problem goal
3. Designing and implementing a decision function that reflects the cost of classification errors that might be made
4. Bayesian updating and experimentation

# WHAT IS NOT IN SCOPE

- **This is not a math or ML theory course** (though some calculus, linear algebra and probability are used)
- **This is not a programming or Python course** (refer to supplemental learning materials for this)
- **This is not an ML system deployment course**