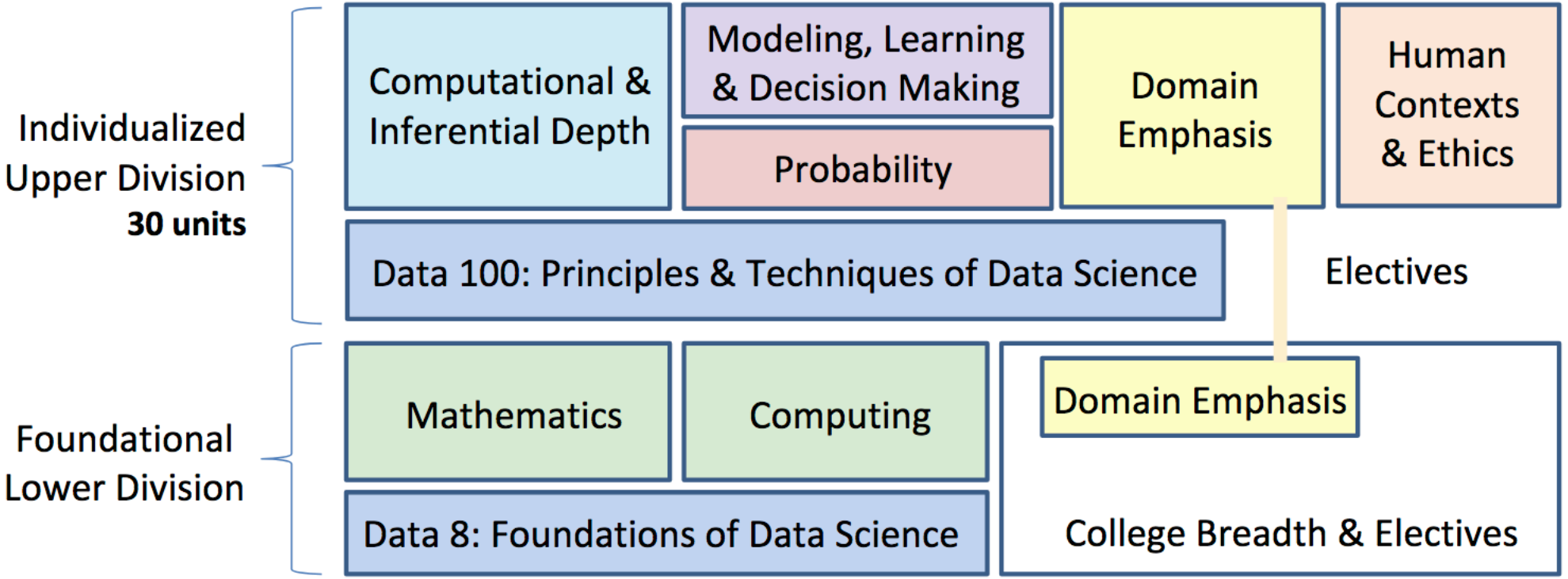
Data Science for All

John DeNero
denero@berkeley.edu

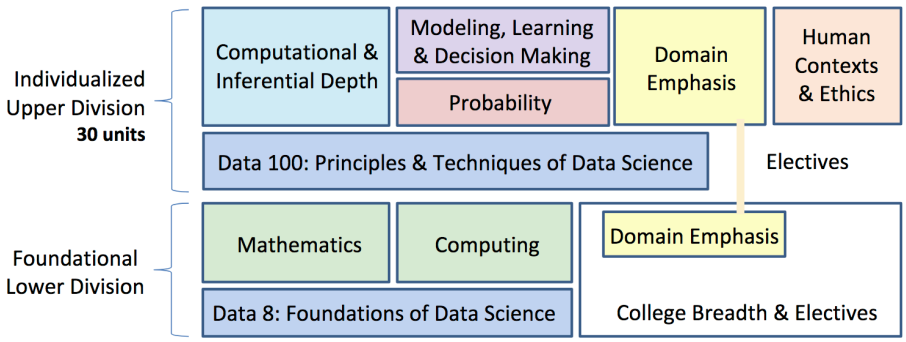
Undergraduate Data Science at UC Berkeley



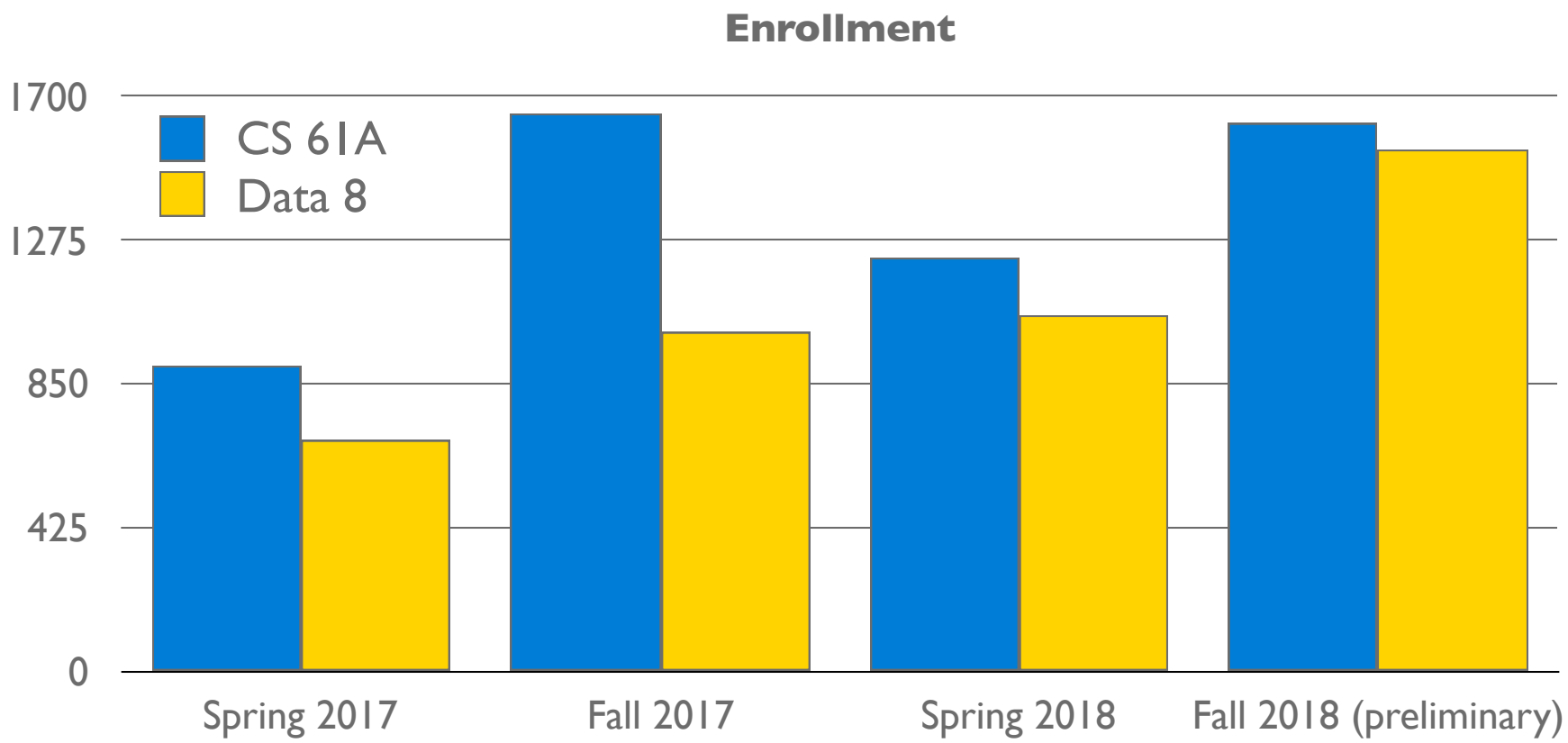
Declarations begin August 15, 2018

- First academic advisor hired
- Faculty committee for policies & appeals
- More than 2,500 messages on the division's electronic forum

Foundations of Data Science



Berkeley's Most Popular Course?



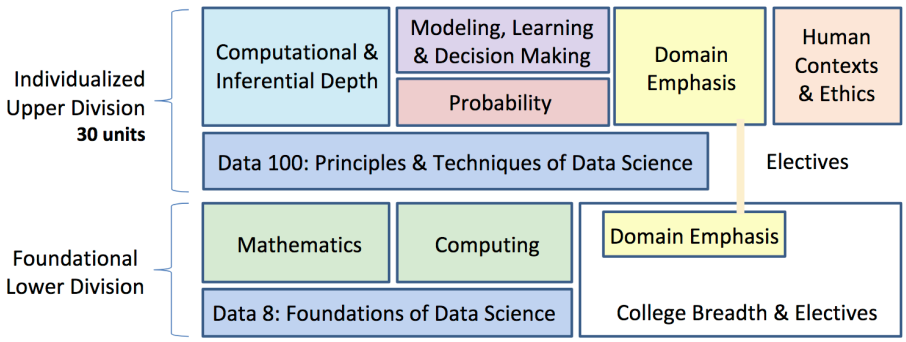
Fall 2018 instructors: David Wagner & Ramesh Sridharan

Spring 2019 instructor: Ani Adhikari

Data 8X: 9k learners watched videos in the first week (1k in week 15)

Undergraduate Data Science Pedagogy Workshop with 40 participants

Principles and Techniques of Data Science



Revisiting Exploration, Inference, and Prediction

619 students in Spring 2018; 715 enrolled for Fall 2018

New textbook: www.textbook.ds100.org

Course goals:

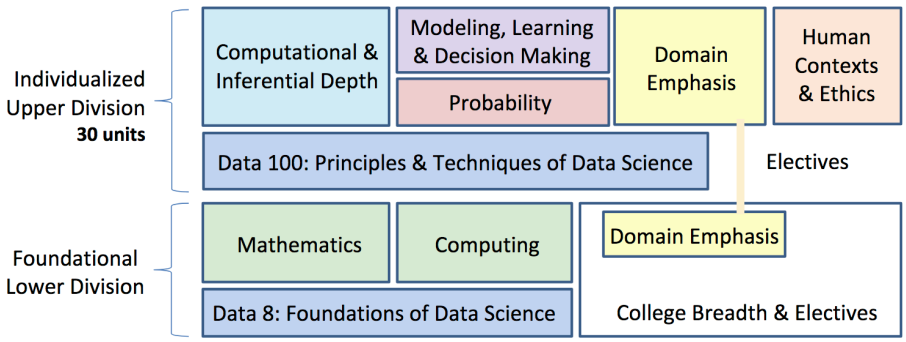
- **Prepare** students for advanced Berkeley courses in data-management, machine learning, and statistics, by providing the necessary foundation and context
- **Enable** students to start careers as data scientists by providing experience working with real-world data, tools, and techniques
- **Empower** students to apply computational and inferential thinking to address real-world problems

Programming: web scraping, regular expressions, databases

Estimation: loss functions, numerical optimization

Learning: multiple regression, logistic regression, multiclass

Theory


Individualized
Upper Division
30 units
Computational & Inferential Depth
Modeling, Learning & Decision Making
Probability
Domain Emphasis
Human Contexts & Ethics
Data 100: Principles & Techniques of Data Science
Electives
Foundational
Lower Division
Mathematics
Computing
Data 8: Foundations of Data Science
Domain Emphasis
College Breadth & Electives

Probability Theory

Probability for Data Science (Stat 140; prob140.org)

100 students in Spring 2018; 195 enrolled for Fall 2018

New textbook: https://textbook.prob140.org/

Prerequisites: Data 8, one year of calculus

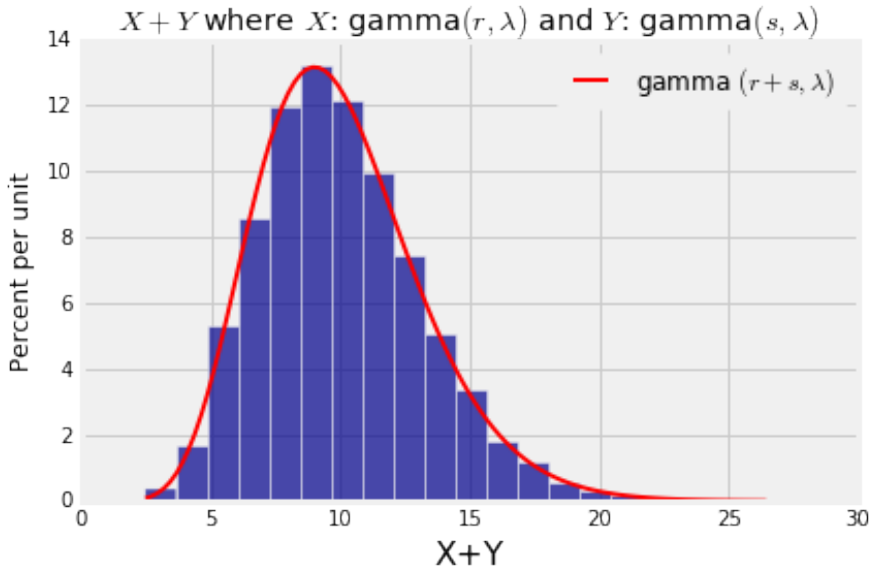
Co-requisite: Linear Algebra

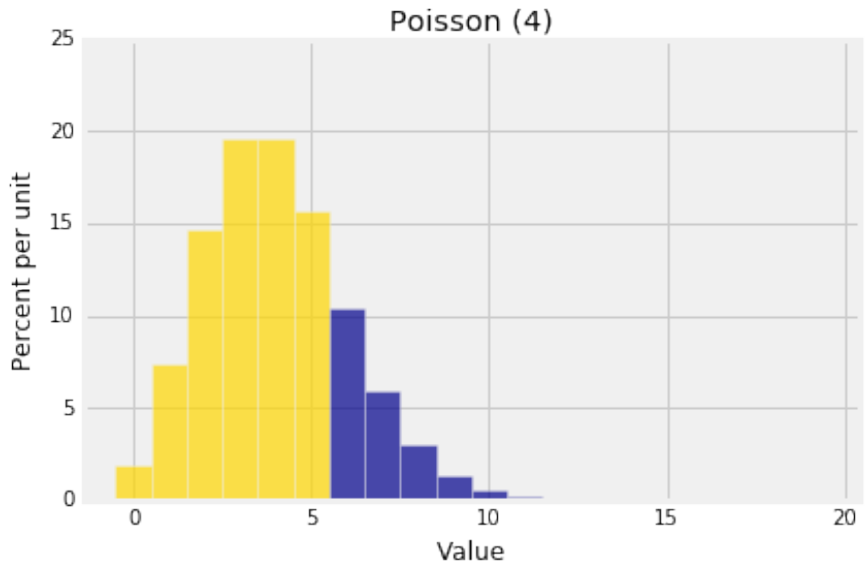
Course Goals:

- Develop the theoretical underpinnings of data science
- Cover selected topics in inference so that students can go directly to machine learning classes
- Develop an appreciation for probability theory

Syllabus: Probability for Data Science

- Basics; all major discrete and continuous distribution families; bounds; transforms; Central Limit Theorem
- Conditioning; random permutations and symmetry
- Markov Chains; Metropolis algorithm for MCMC
- MLE and MAP estimates; conjugate priors; beta-binomial distribution
- Prediction: least squares, multivariate normal distribution, multiple regression





Approach: Probability for Data Science

- Data 8 background provides extensive experience with sampling variability and the motivation to delve deeper.
- Computation and visualization help reinforce the math and make it concrete.
- Patterns observed by computation provide motivation for mathematical justification, along with appreciation of the math.
- Students solve more substantive problems than they could have with just pen and paper.

(Demo)

Linear Algebra for Data Science

35 students in Spring 2018 (Stat 89A).

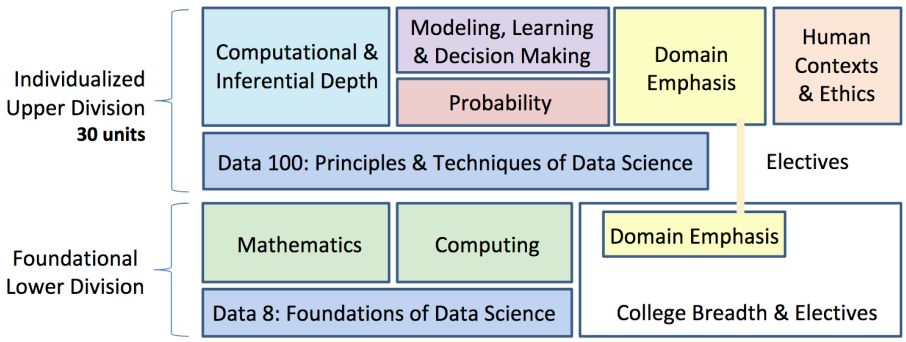
Increased emphasis on connections to probability theory.

Reduced emphasis on linear equations & subspaces.

Topics:

- quadratic forms,
- eigenvalue decompositions of symmetric matrices,
- least squares regression,
- principal components analysis,
- spectral clustering and ranking,
- solving linear equations.

Modeling, Learning, and Decision Making



Machine Learning Depth

Data 8: Nearest-neighbor classification & simple linear regression

Data 100: Multiple regression, logistic regression, regularization

What's left to learn?

- Techniques: max margin, trees, neural networks, kernels, LDA/QDA
- Non-linear models
- Dimensionality, model capacity, feature selection
- Unsupervised learning

CS 189: Introduction to Machine Learning

Stat 154: Modern Statistical Prediction and Machine Learning

IEOR 142: Introduction to Machine Learning and Data Analytics

CS 182: Designing, Visualizing and Understanding Deep Neural Networks

Data 102: TBD

Coming soon: Human Context & Ethics