



Lecture 32

Classifiers

Contributions by Vinitra Swamy (vinitra@berkeley.edu) and Fahad Kamran (fhdkmrn@berkeley.edu)
Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)



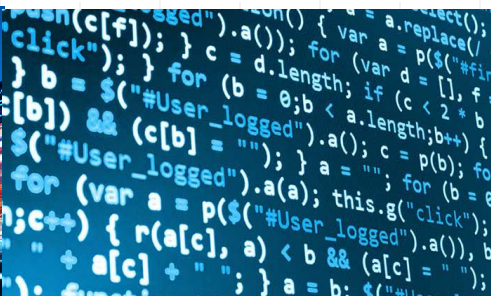
Paths Forward After Data 8

data.berkeley.edu

Data Science Connector Courses in Fall 2018



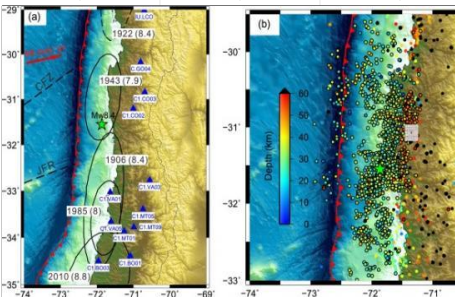
CIVENG 88
Data Science for Smart Cities



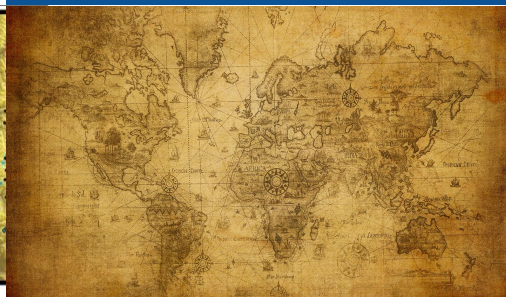
COMPSCI 88
Computational Structures



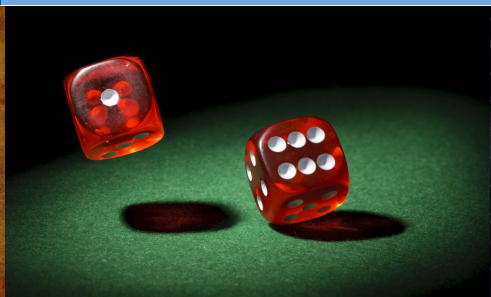
DEMOG 88
Data Science & Immigration



EPS 88
Python for Earth Science



HISTORY 88
How Does History Count?



STAT 88
Probability & Statistics in Data Science



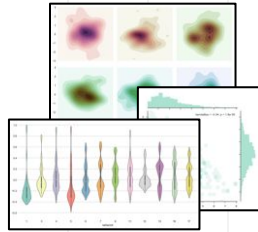
UGBA 96-2 & 3
Data & Decisions



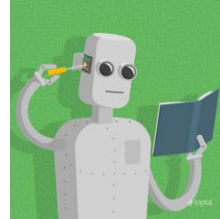
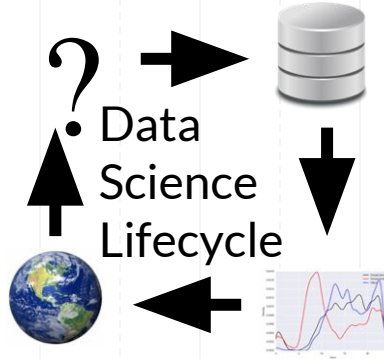
Data 100



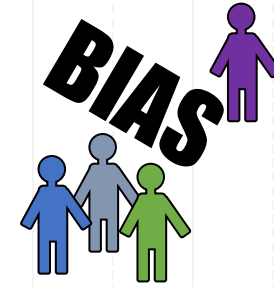
Data Collection
and Preparation



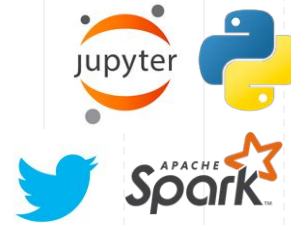
Data
Visualization



Machine
Learning

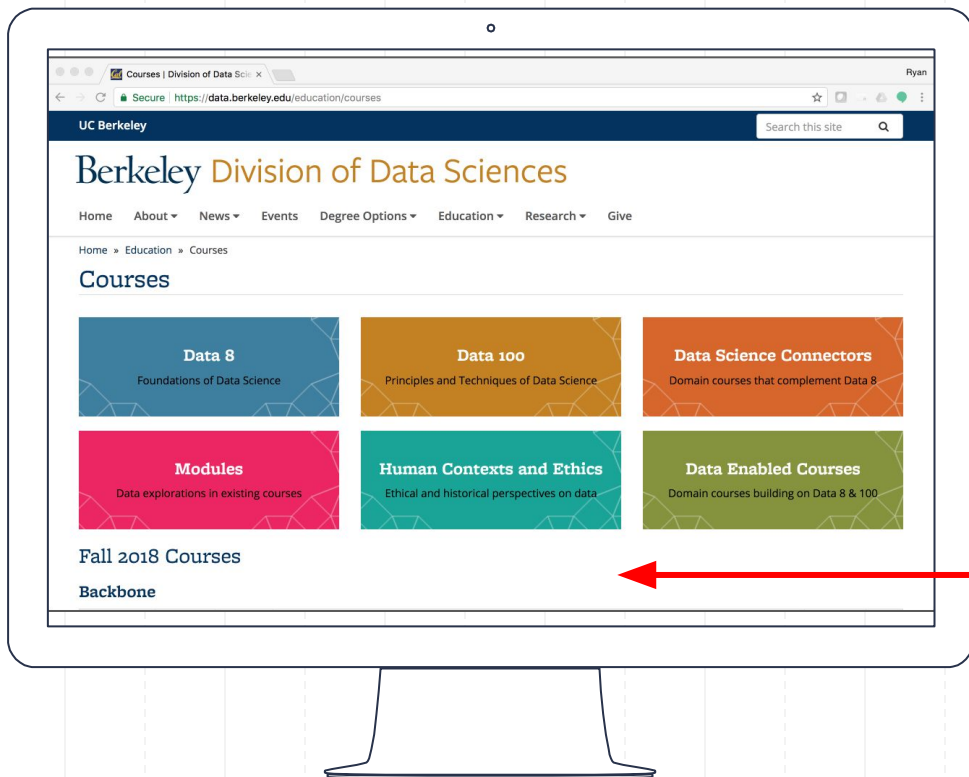


Statistical
Bias & Ethics



Using Real
Tools & Systems



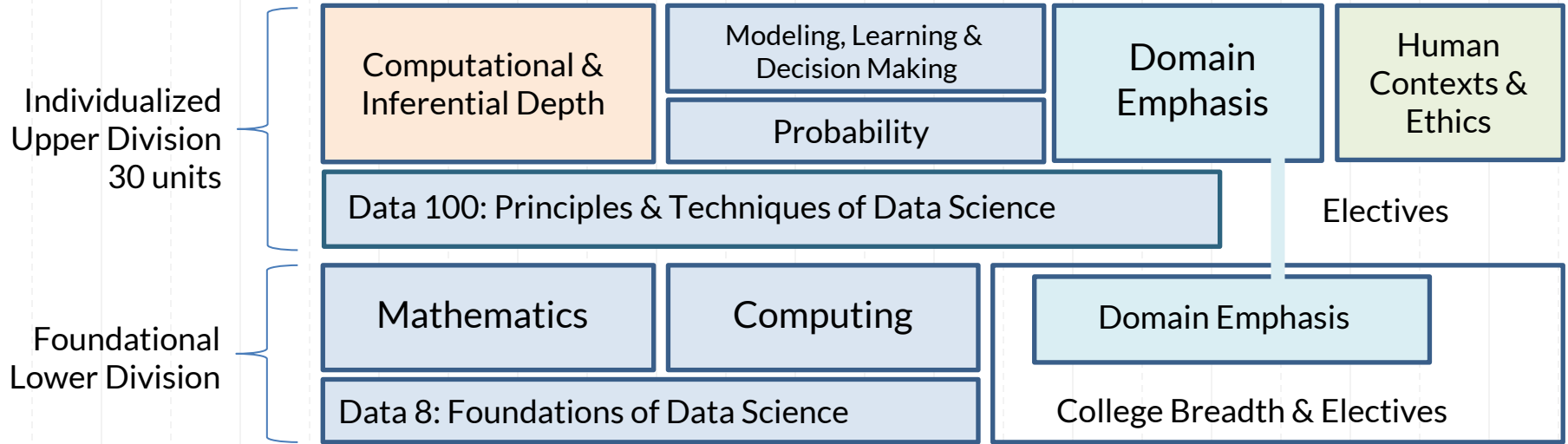


For more information,
check out our website

data.berkeley.edu/education/courses

Scroll down to see the Fall 2018 offerings

Data Science Major (L&S)





Student Opportunities

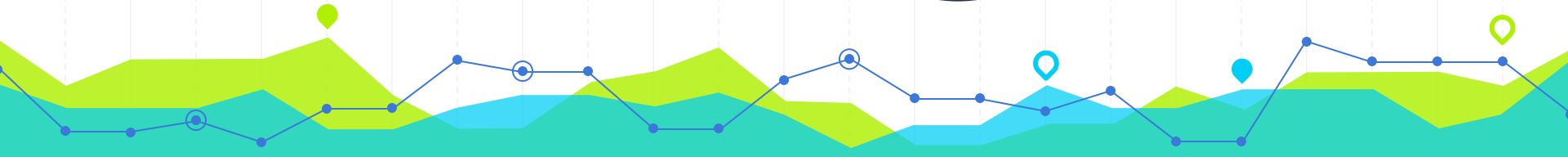
Available Opportunities

Curriculum

Peer
Advising

Discovery
Research

Internal
Operations



Curriculum

Hone your skills as an educator and data scientist by developing educational materials for courses

Recruiting for 2 positions:



Connector
Assistants

Help instructors of
Data Science
Connector courses
to deliver and
teach material

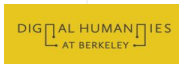


Curriculum
Development

Create curriculum
materials for
Connectors,
Data-Enabled Courses,
or short explorations
into DS (modules)

Discovery Research

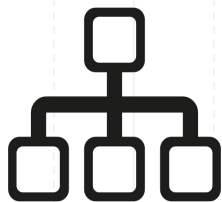
Connecting students with hands-on data science research- non-profits, start-ups, insitutions, etc



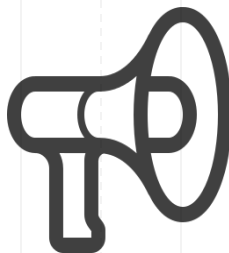
- Opportunities for All
- High Impact Research
- Career Advancement

Internal Operations

Help organize and manage the initiatives that draw all our student programs together



Mapping & Analytics



Diversity, Outreach & Communications



Technology
Engagement & Support

Peer Advising

First point of contact for students who are interested in the major and DS in general

Provide advice, support, and mentorship on topics like:

- Student involvement and extracurriculars
- Undergraduate research opportunities
- Tutoring/studying services
- Schedule planning and course recommendations
- Knowledge and experience of major courses



How to Apply

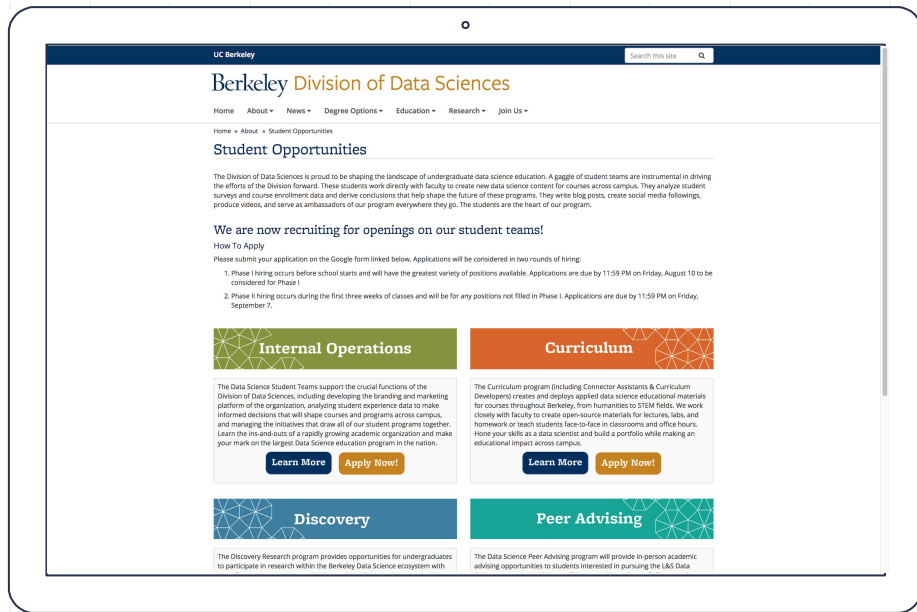
Phase 1: Applications due by **Friday, August 10th**

Greatest variety of positions available

Phase 2: Applications due by **Friday, September 7th**

For positions not filled in Phase 1





For more information,
check out our website at
bit.ly/ds-teams-apply

Data Scholars

For students who are

- Low income
- First generation college
- Historically underrepresented

Aim: Support diversity in the Data Science student community all the way through.

Foundations

Concurrent to Data 8

Pathways

Further technical
skill & career development

Discovery

Support for research
Experiences



Thank You!



data.berkeley.edu



[ds-curriculum](#)
[ds-teams](#)
[ds-discovery](#)
[ds-peer-advising](#)

[@berkeley.edu](#)



Lecture 32

Classifiers

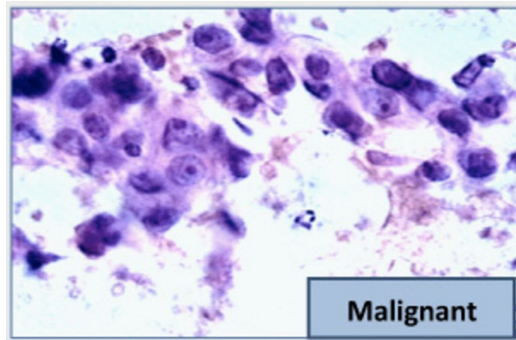
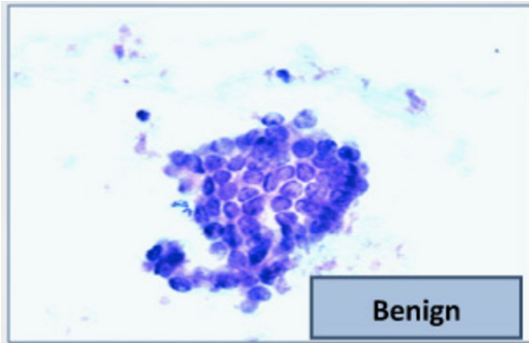
Contributions by Vinitra Swamy (vinitra@berkeley.edu) and Fahad Kamran (fhdkmrn@berkeley.edu)
Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)

Announcements

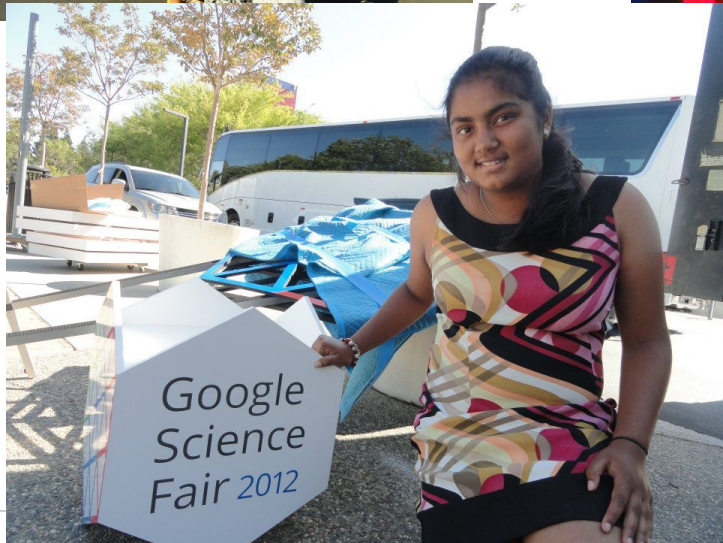
Classifiers

The Google Science Fair

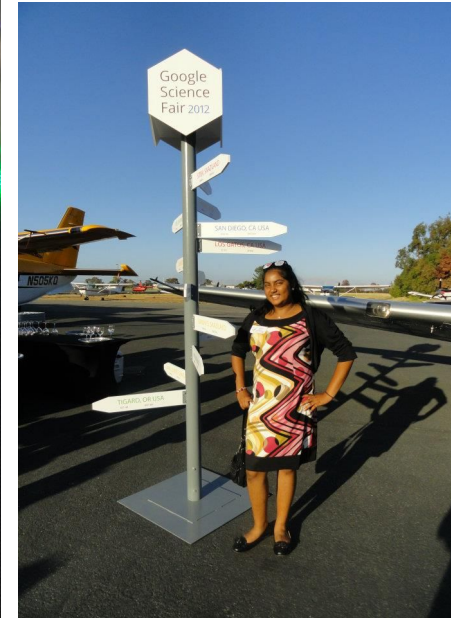
- Brittany Wenger, a 17-year-old high school student in 2012
- Won by building a breast cancer classifier with 99% accuracy



Fate works in mysterious ways

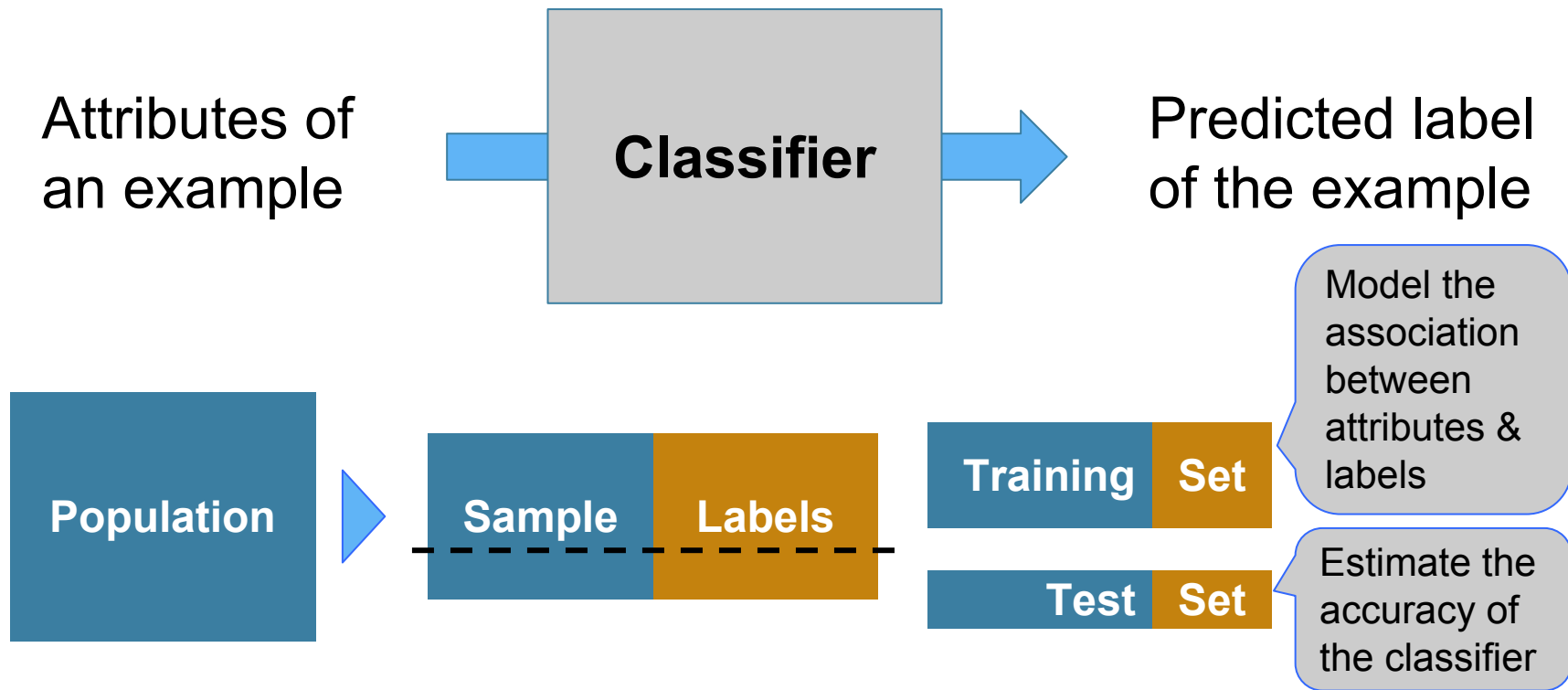


And the finalists?

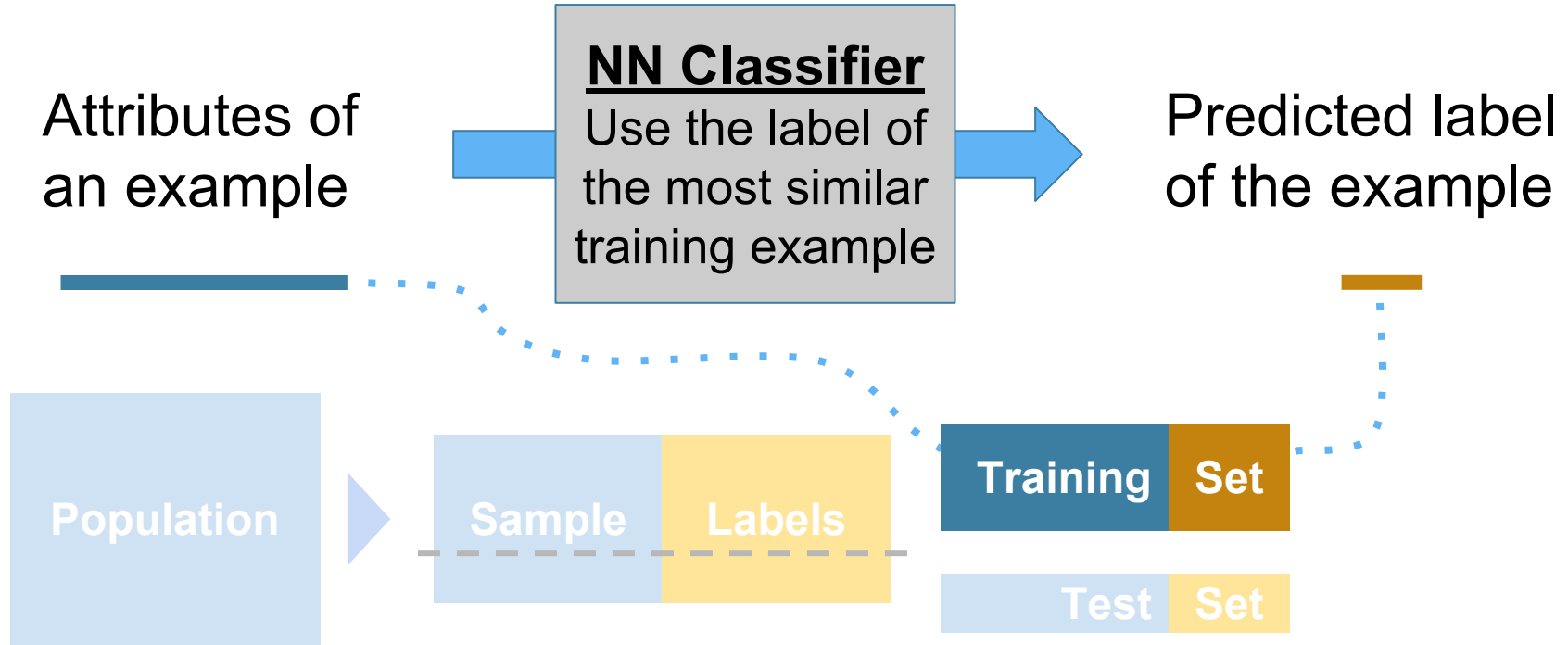


(Demo)

Training a Classifier



Nearest Neighbor Classifier



Distance

Rows of Tables

Each row contains all the data for one individual

- `t.row(i)` evaluates to *i*th row of table *t*
- `t.row(i).item(j)` is the value of column *j* in row *i*
- If all values are numbers, then `np.array(t.row(i))` evaluates to an array of all the numbers in the row.
- To consider each row individually, use

```
for row in t.rows:
    ... row.item(j) ...
```

Distance Between Two Points

- Two attributes x and y :

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}.$$

- Three attributes x , y , and z :

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$$

- and so on ...

(Demo)

Nearest Neighbors

Finding the k Nearest Neighbors

To find the k nearest neighbors of an example:

- Find the distance between the example and each example in the training set
 - Augment the training data table with a column containing all the distances
 - Sort the augmented table in increasing order of the distances
 - Take the top k rows of the sorted table (Demo)
-

The Classifier

To classify a point:

- Find its k nearest neighbors
- Take a majority vote of the k nearest neighbors to see which of the two classes appears more often
- Assign the point the class that wins the majority vote

(Demo)

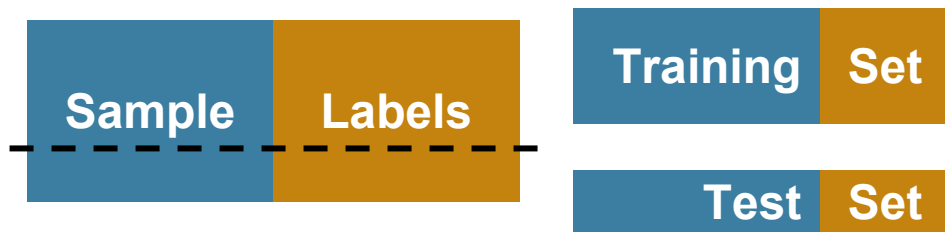
Evaluation

Accuracy of a Classifier

The accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly

Need to compare classifier predictions to true labels

If the labeled data set is sampled at random from a population, then we can infer accuracy on that population



(Demo)