Prediction vs. Bias: A Debate

Lucy Bernholz

Conference Co-Host; Senior Research Scholar, Stanford PACS; Director, Stanford Digital Civil Society Lab

@p2173

Andrew Means

Conference Co-Host; Head, beyond.uptake

@meansandrew

Kristian Lum

Lead Statistician,
Human Rights Data Analysis Group

@KLdivergence

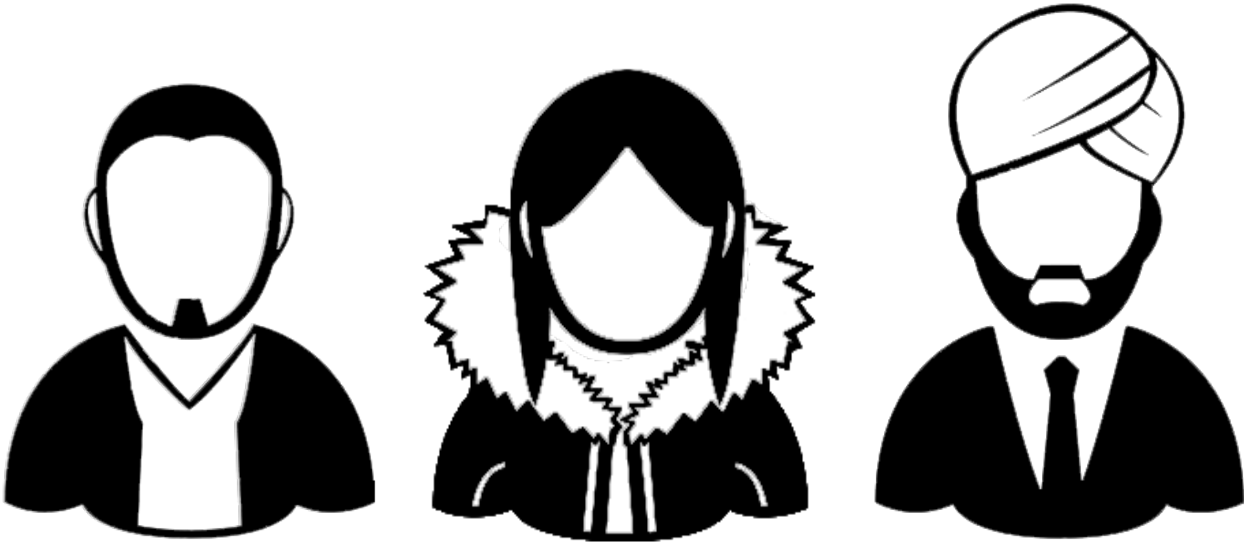
Candace Thille

Founding Director,
Open Learning Initiative,
Carnegie Mellon University and
Stanford University

@StanfordEd

1 START WITH THE LEARNER

DIVERSITY OF THE LEARNER



DIVERSITY OF THE LEARNER

- background knowledge
- relevant skills
- future goals

DIVERSITY OF THE LEARNER

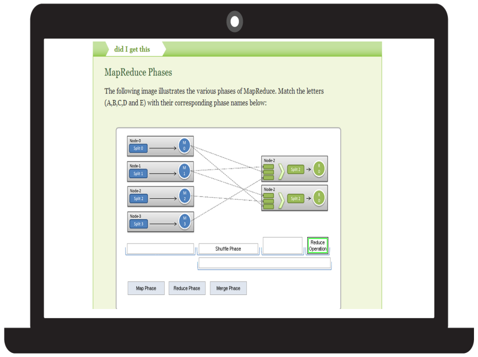
- background knowledge
- relevant skills
- future goals

- attributions—how learners explain the causes of experiences

2

THE INNOVATION

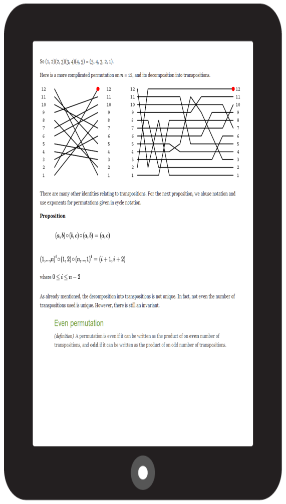
PERSONALIZED AND ADAPTIVE INSTRUCTION



did I get this

MapReduce Phases

The following image illustrates the various phases of MapReduce. Match the letters (A,B,C,D and E) with their corresponding phase names below:



So (1, 2)(2, 3)(3, 4)(4, 5) = (5, 4, 3, 2, 1).

Here is a more complicated permutation on $n = 12$, and its decomposition into transpositions.

There are many other identities relating to transpositions. For the next proposition, we abuse notation and use exponents for permutations given in cycle notation.

Proposition

$$(a,b)\circ(b,c)\circ(a,b) = (a,c)$$

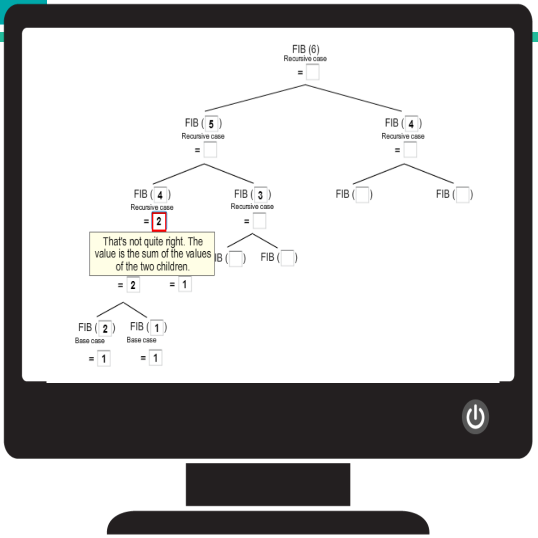
$$(1,\ldots,n)^i\circ(1,2)\circ(n,\ldots,1)^i = (i+1,i+2)$$

where $0 \leq i \leq n-2$

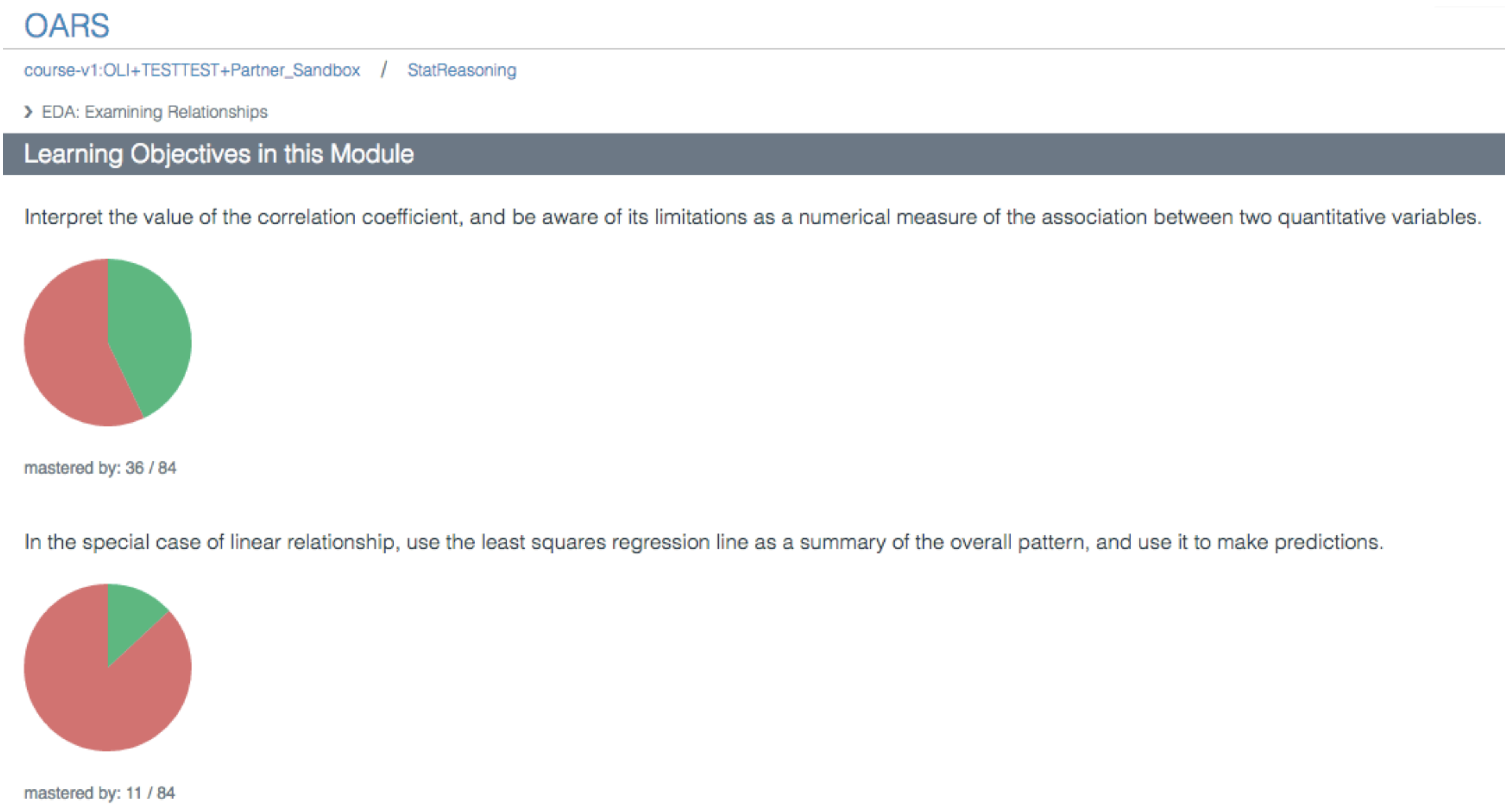
As already mentioned, the decomposition into transpositions is not unique. In fact, not even the number of transpositions used is unique. However, there is still an invariant.

Even permutation

(definition) A permutation is even if it can be written as the product of on **even** number of transpositions, and **odd** if it can be written as the product of on odd number of transpositions.



That's not quite right. The value is the sum of the values of the two children.



OARS

course-v1:OLI+TESTTEST+Partner_Sandbox / StatReasoning

› EDA: Examining Relationships

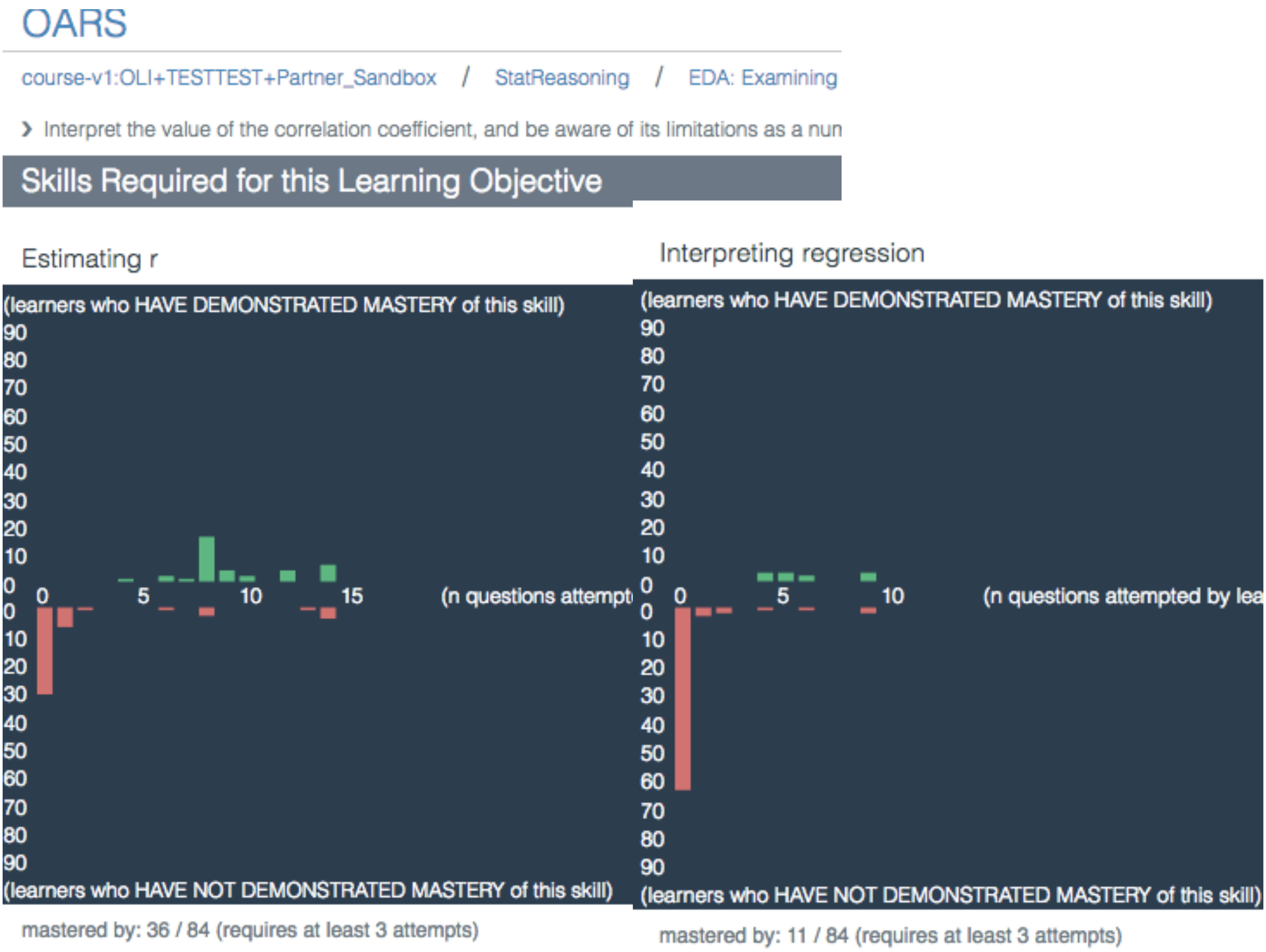
Learning Objectives in this Module

Interpret the value of the correlation coefficient, and be aware of its limitations as a numerical measure of the association between two quantitative variables.

mastered by: 36 / 84

In the special case of linear relationship, use the least squares regression line as a summary of the overall pattern, and use it to make predictions.

mastered by: 11 / 84



OARS

course-v1:OLI+TESTTEST+Partner_Sandbox / StatReasoning / EDA: Examining

› Interpret the value of the correlation coefficient, and be aware of its limitations as a nun

Skills Required for this Learning Objective

Estimating r

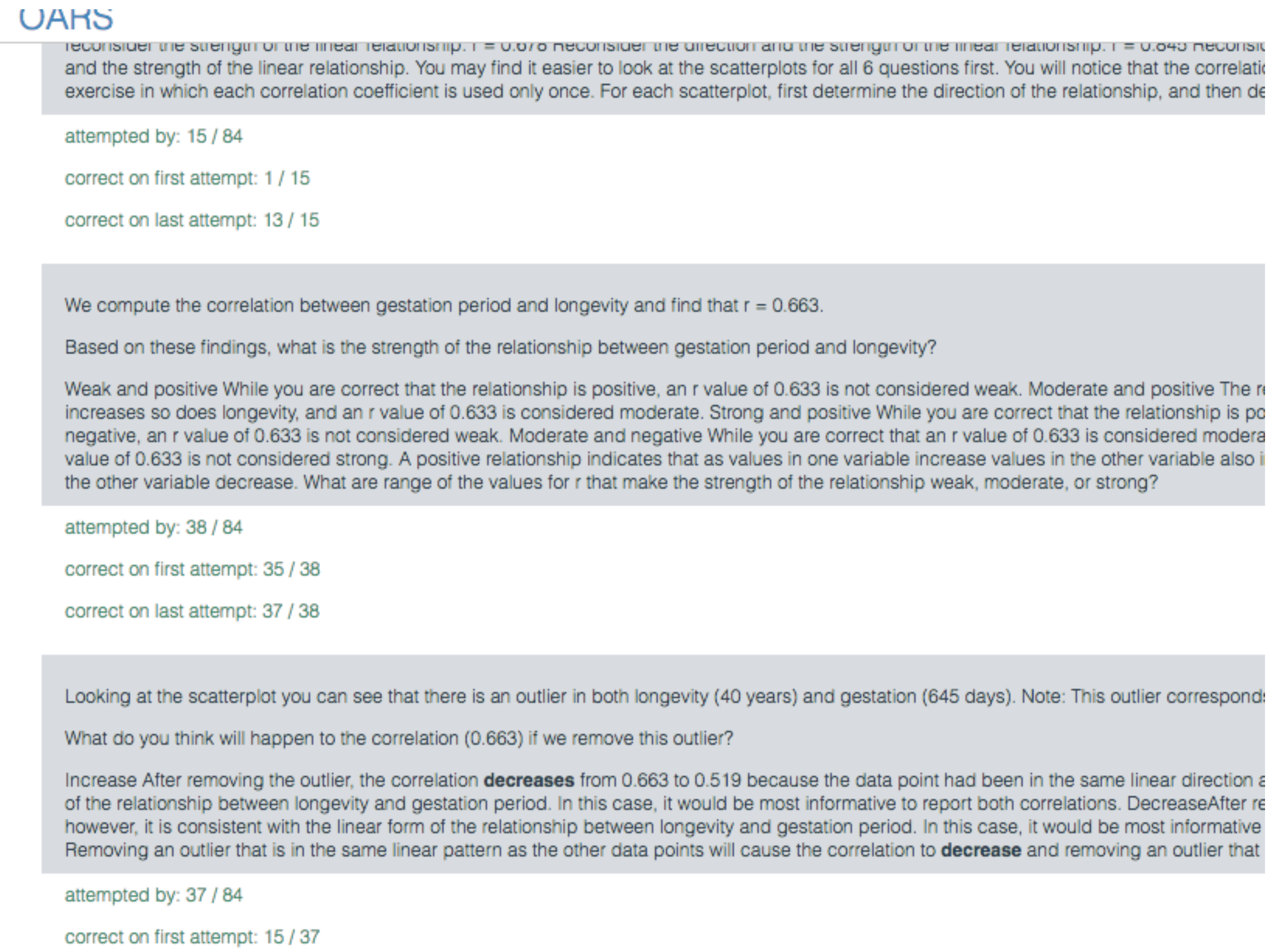
(learners who HAVE DEMONSTRATED MASTERY of this skill)
90
80
70
60
50
40
30
20
10
0
0
5
10
15
(n questions attempt
0
10
20
30
40
50
60
70
80
90
(learners who HAVE NOT DEMONSTRATED MASTERY of this skill)

mastered by: 36 / 84 (requires at least 3 attempts)

Interpreting regression

(learners who HAVE DEMONSTRATED MASTERY of this skill)
90
80
70
60
50
40
30
20
10
0
0
5
10
(n questions attempted by lea
0
10
20
30
40
50
60
70
80
90
(learners who HAVE NOT DEMONSTRATED MASTERY of this skill)

mastered by: 11 / 84 (requires at least 3 attempts)



reconsider the strength of the linear relationship. r = 0.678 Reconsider the direction and the strength of the linear relationship. r = 0.845 Reconsi
and the strength of the linear relationship. You may find it easier to look at the scatterplots for all 6 questions first. You will notice that the correlati
exercise in which each correlation coefficient is used only once. For each scatterplot, first determine the direction of the relationship, and then de

attempted by: 15 / 84

correct on first attempt: 1 / 15

correct on last attempt: 13 / 15

We compute the correlation between gestation period and longevity and find that r = 0.663.

Based on these findings, what is the strength of the relationship between gestation period and longevity?

Weak and positive While you are correct that the relationship is positive, an r value of 0.633 is not considered weak. Moderate and positive The r
increases so does longevity, and an r value of 0.633 is considered moderate. Strong and positive While you are correct that the relationship is po
negative, an r value of 0.633 is not considered weak. Moderate and negative While you are correct that an r value of 0.633 is considered modera
value of 0.633 is not considered strong. A positive relationship indicates that as values in one variable increase values in the other variable also i
the other variable decrease. What are range of the values for r that make the strength of the relationship weak, moderate, or strong?

attempted by: 38 / 84

correct on first attempt: 35 / 38

correct on last attempt: 37 / 38

Looking at the scatterplot you can see that there is an outlier in both longevity (40 years) and gestation (645 days). Note: This outlier correspond

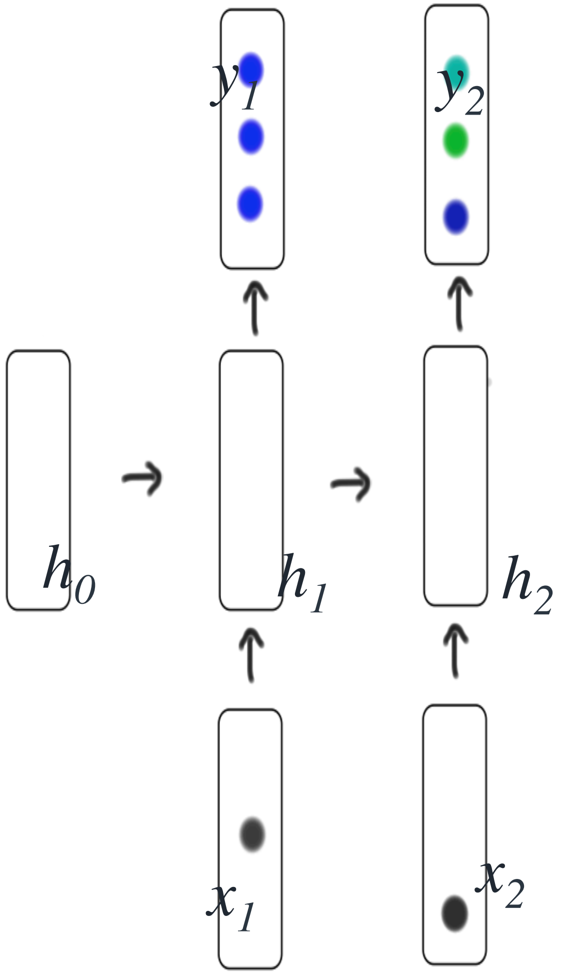
What do you think will happen to the correlation (0.663) if we remove this outlier?

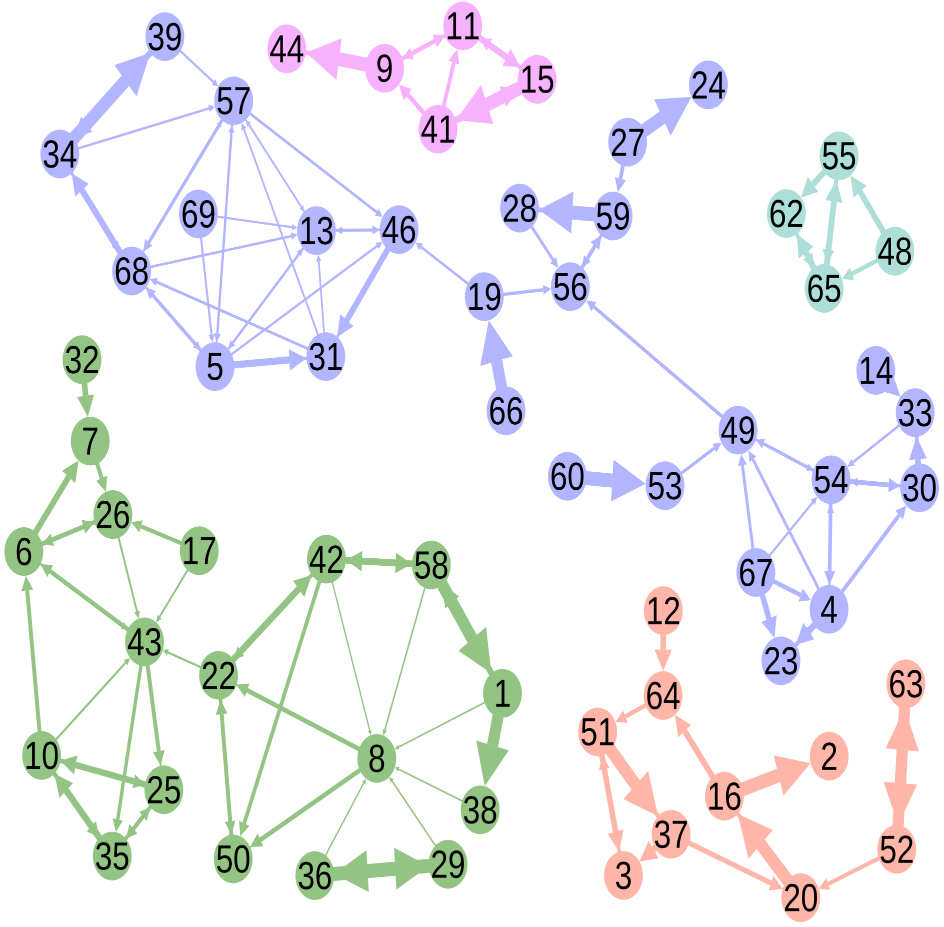
Increase After removing the outlier, the correlation **decreases** from 0.663 to 0.519 because the data point had been in the same linear direction a
of the relationship between longevity and gestation period. In this case, it would be most informative to report both correlations. DecreaseAfter re
however, it is consistent with the linear form of the relationship between longevity and gestation period. In this case, it would be most informative
Removing an outlier that is in the same linear pattern as the other data points will cause the correlation to **decrease** and removing an outlier that

attempted by: 37 / 84

correct on first attempt: 15 / 37

KNOWLEDGE MODELING





3

THE SCIENCE

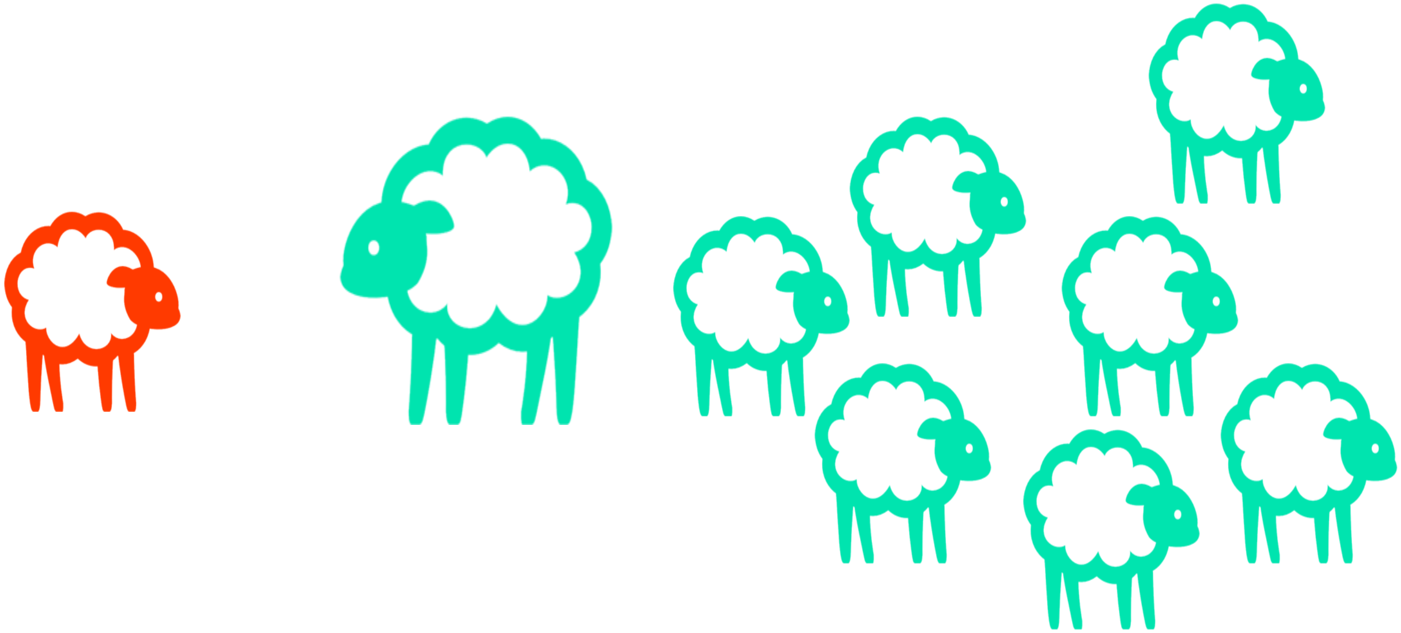
COGNITIVE SCIENCE

- guided learning
- complex cognitive tasks
- targeted hints & feedback

SOCIAL-PSYCHOLOGICAL SCIENCE

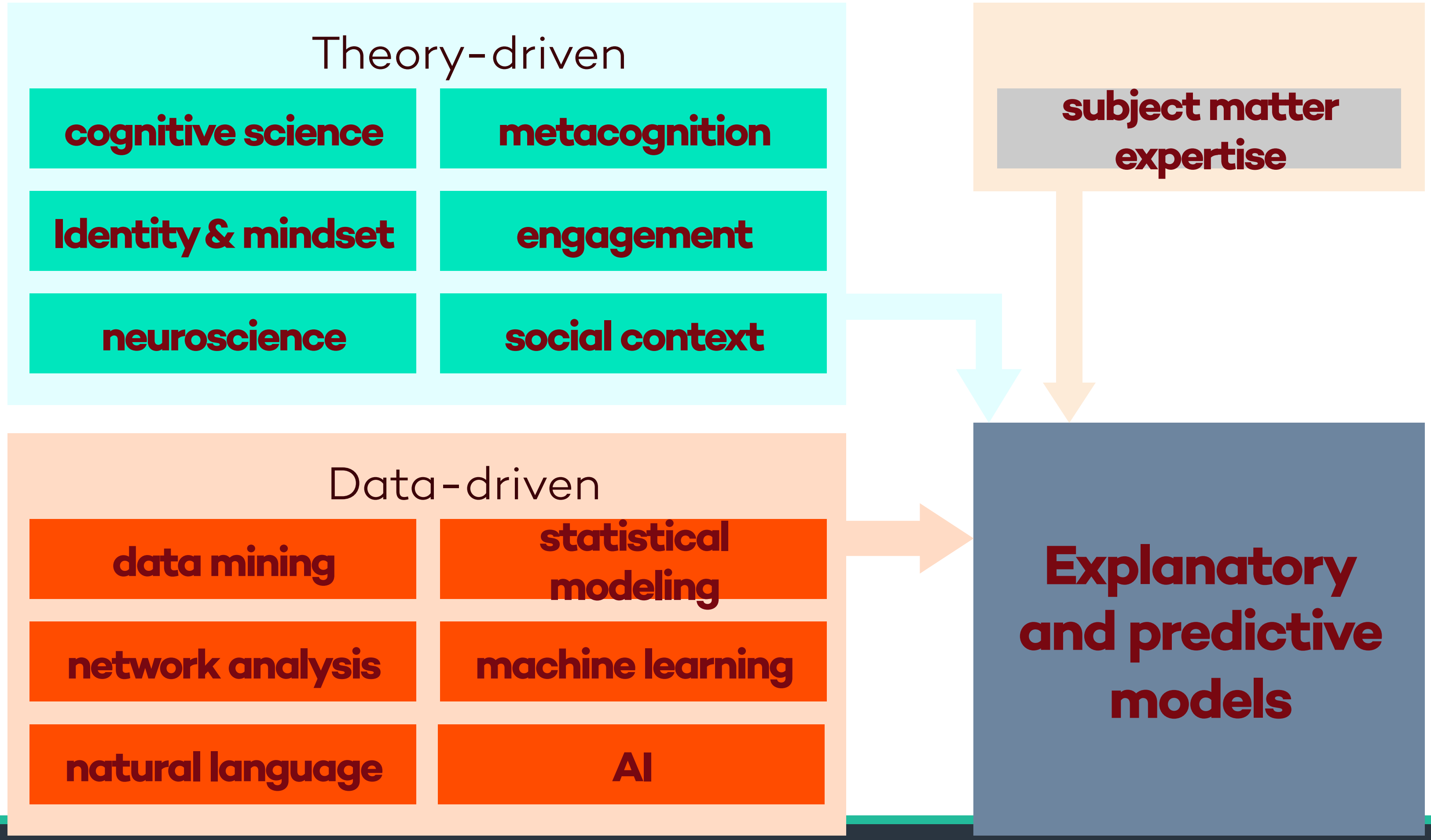
Mindset

SOCIAL-PSYCHOLOGICAL SCIENCE



Social belongingness
Stereotype Threat

DATA SCIENCE



Theory-driven

cognitive science

metacognition

Identity & mindset

engagement

neuroscience

social context

Data-driven

data mining

statistical modeling

network analysis

machine learning

natural language

AI

subject matter expertise

Explanatory and predictive models

4

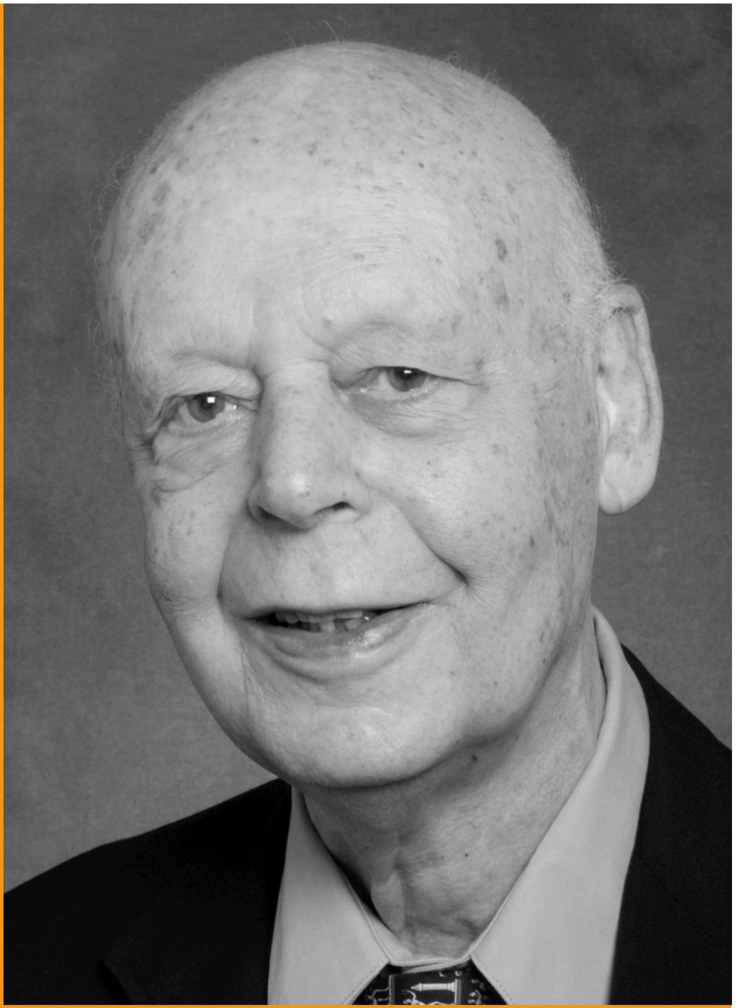
A PART OF

HIGHER EDUCATION

“Improvement in post secondary education will require converting teaching (and courseware, platform & analytic system development) from a solo sport to a community-based research activity.”

- Herbert Simon 1991 (modifications 2015)

"Without a complete revolution…in our approach to teaching… we cannot go beyond (current levels) of productivity" William Baumol, 1967



Our Message:

Such a revolution is ~~possible~~ happening

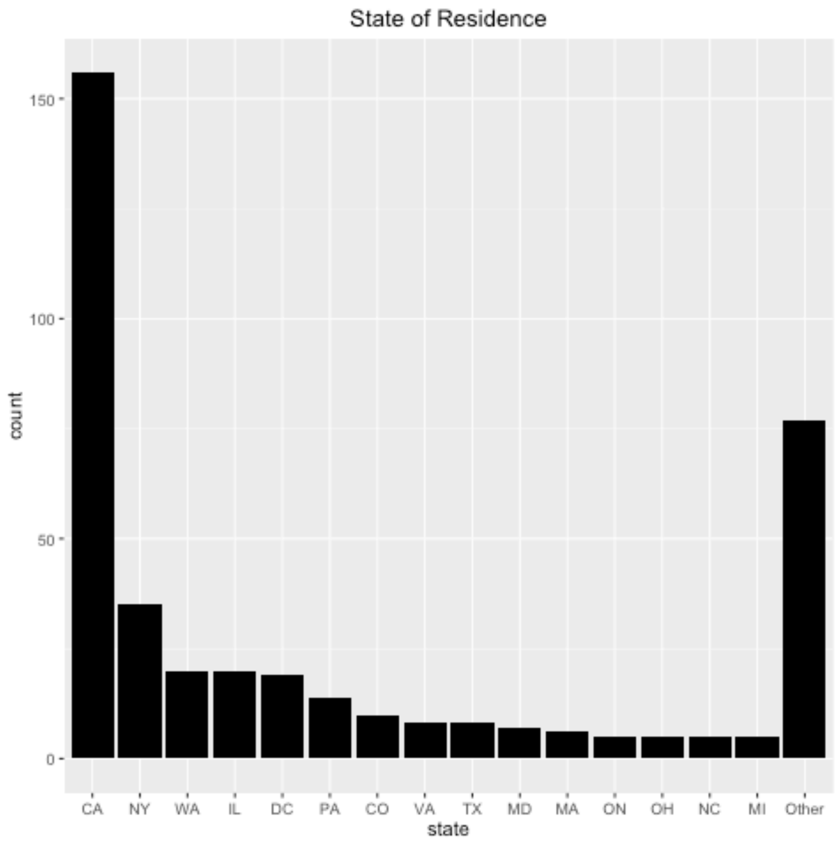
Our Question:

Who will lead it?

5

START WITH AND END WITH THE LEARNER

Registrant Data

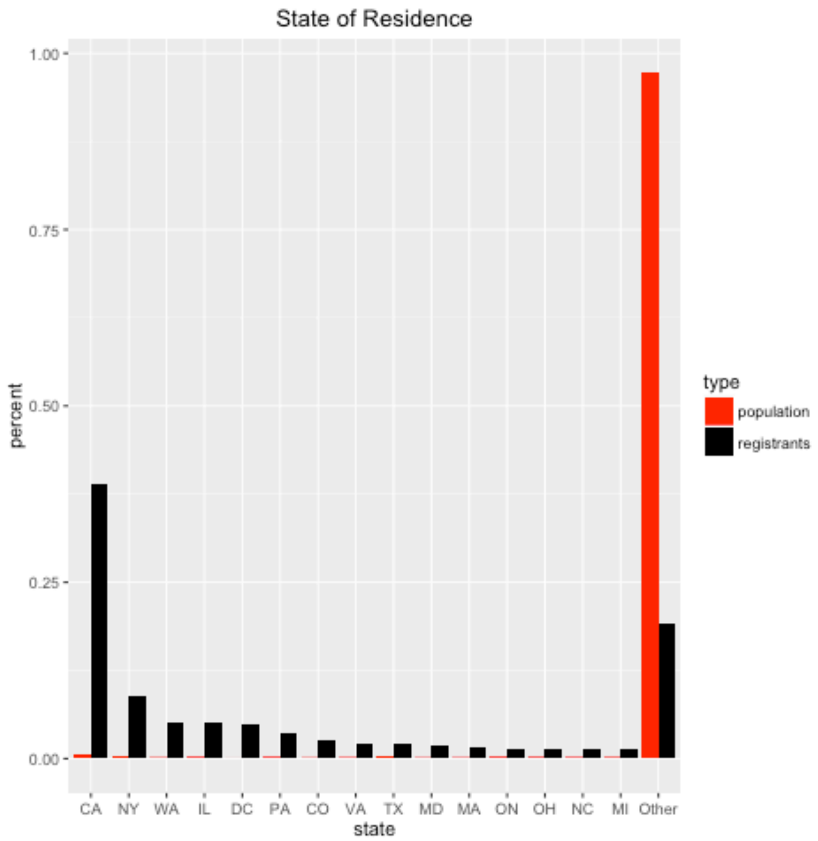


Is this data biased?

A data set is *non-representative* or *biased** with respect to a population of interest if not all elements of the population had the same probability of appearing in the dataset.

*This is not a value judgment about the dataset or those who compiled it. A “statistically biased” dataset does not require that anyone acted with bias in the colloquial sense of the word.

Registrant Data

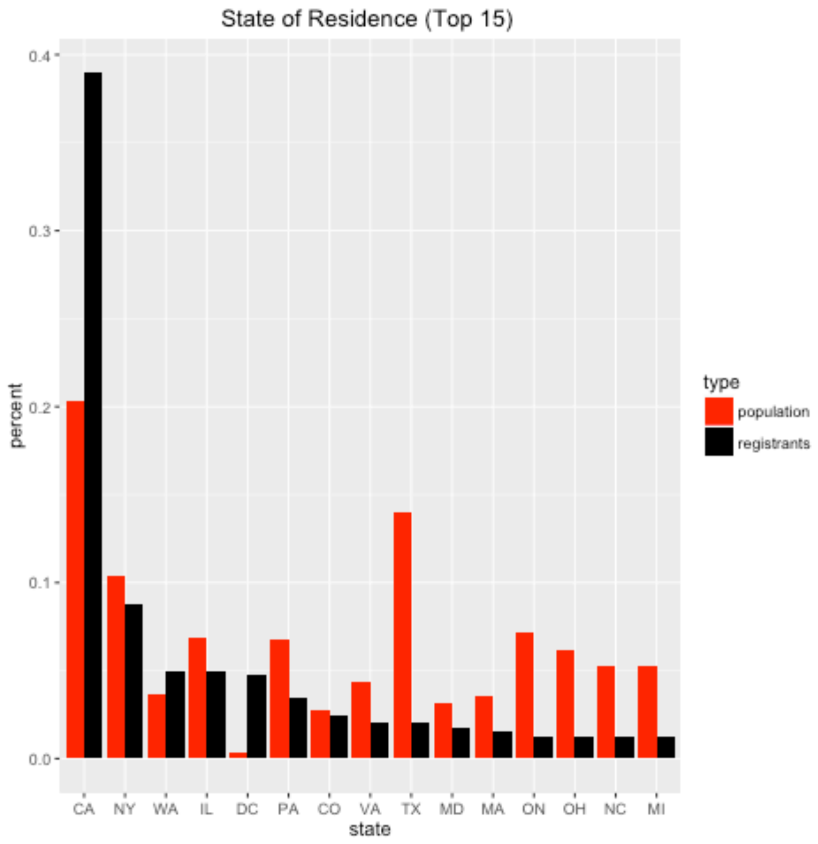


Is this data biased?

A data set is *non-representative* or *biased** with respect to a population of interest if not all elements of the population had the same probability of appearing in the dataset.

*This is not a value judgment about the dataset or those who compiled it. A “statistically biased” dataset does not require that anyone acted with bias in the colloquial sense of the word.

Registrant Data



Is this data biased?

A data set is *non-representative* or *biased** with respect to a population of interest if not all elements of the population had the same probability of appearing in the dataset.

*This is not a value judgment about the dataset or those who compiled it. A “statistically biased” dataset does not require that anyone acted with bias in the colloquial sense of the word.

Machine Learning, Statistical Modeling, Predictive Analytics, etc.

logistic regression

k-means clustering

random forest

support vector machines

kernel density estimation

reinforcement learning

neural nets

nearest neighbors

generalized linear model

boosting

deep learning

ensemble models

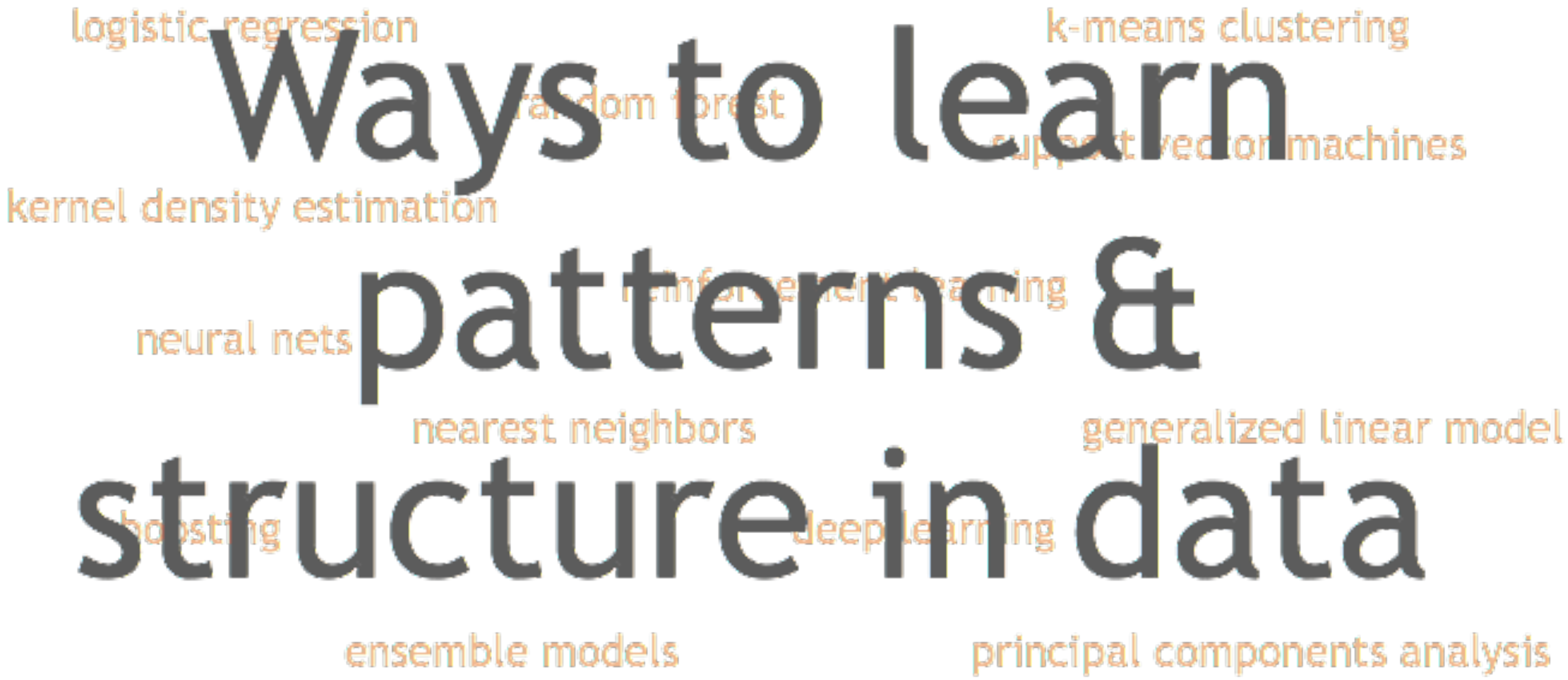
principal components analysis

Police Data

Are police records and unbiased sample of all crimes?

- **Variation in reporting rates**
  - » NCVS indicates that reporting rates vary substantially by demographic characteristics, i.e. some crimes are more likely than others to be reported to police depending on who was victimized.
  - » In this case, the bias derives not from the police themselves but from the community the police serve.
- **Variation in police attention**
  - » Crimes that are committed in areas that are highly patrolled by police are more likely to be discovered by police than those committed in less patrolled areas.
  - » Police are not tasked with collecting a random sample, so bias in the data may come from legitimate police strategy.
- **Variation in rates of *enforcement* for similar criminal behavior**
  - » While white and black populations use marijuana at similar rates, blacks are arrested for marijuana possession at a rate several times that of whites.*

Machine Learning, Statistical Modeling, Predictive Analytics, etc.



Ways to learn patterns & structure in data

**Which model you pick defines *how* you learn the patterns and the types of patterns we can learn.

What is predictive policing?

Predictive policing uses police records to learn patterns in the ~~occurrence of crime.~~

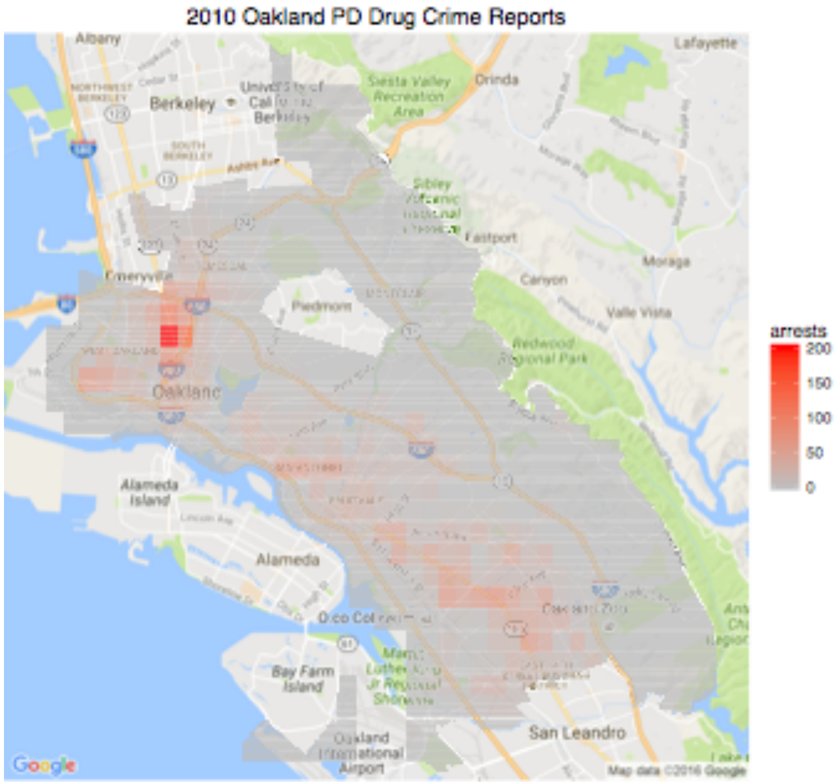
police records or the recording of crime

Using these patterns, the computer then predicts the most likely ~~locations of future crimes~~.

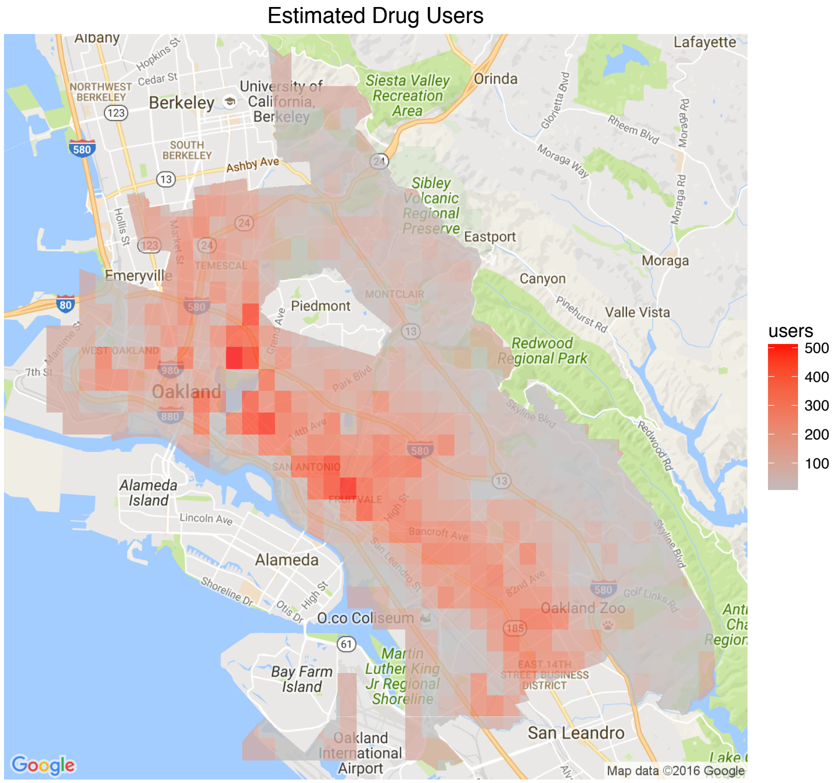
where crime will be detected in the future.

Additional police are dispatched to the locations with the highest predicted rate of crime.

Drug Crimes in Oakland, CA

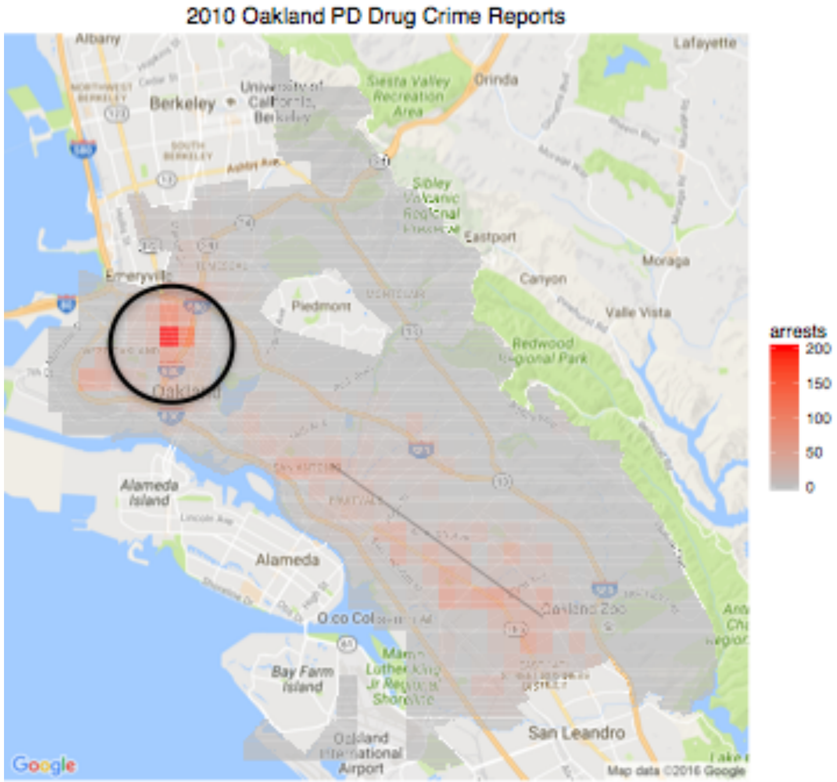

2010 Oakland PD Drug Crime Reports

Data collected and cleaned by OpenOakland.org


Estimated Drug Users

United States. Office of Applied Studies. Substance Abuse and Mental Health Archives. National Survey on Drug Use and Health.2010 RTI U.S. Synthetic Population Ver. 1.0 RTI International. May, 2014. URL: https://www.epimodels.org/midas/pubsyntdata1.do

Drug Crimes in Oakland, CA



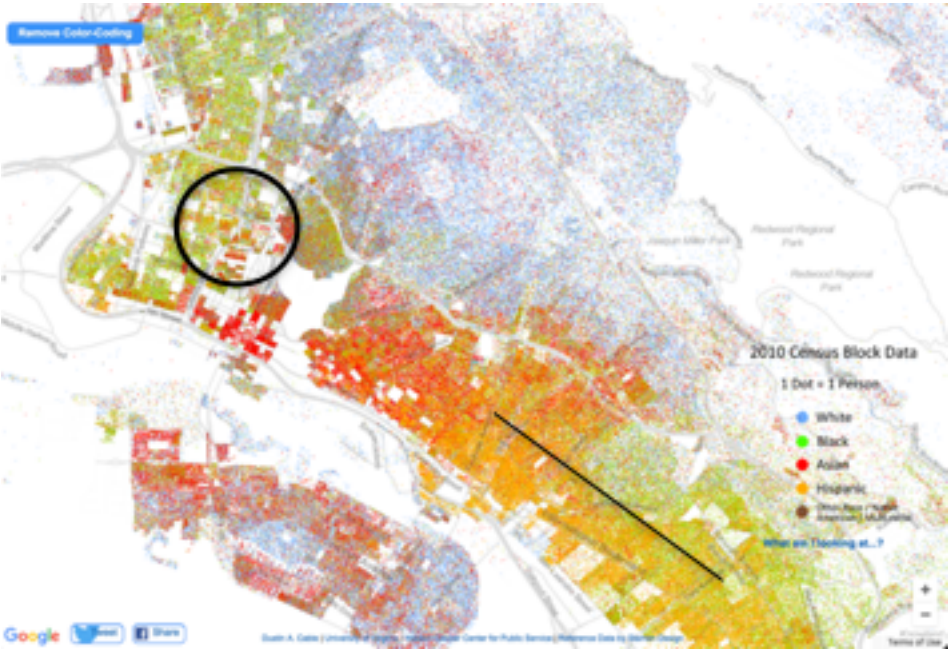
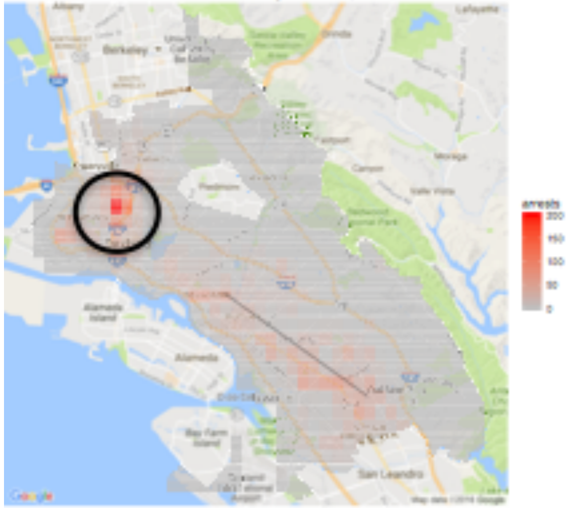
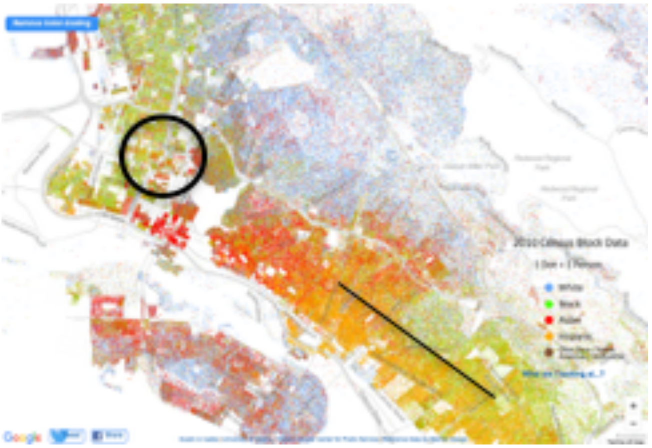


Image Copyright, 2013, Weldon Cooper Center for Public Service, Rector and Visitors of the University of Virginia (Dustin A. Cable, creator)

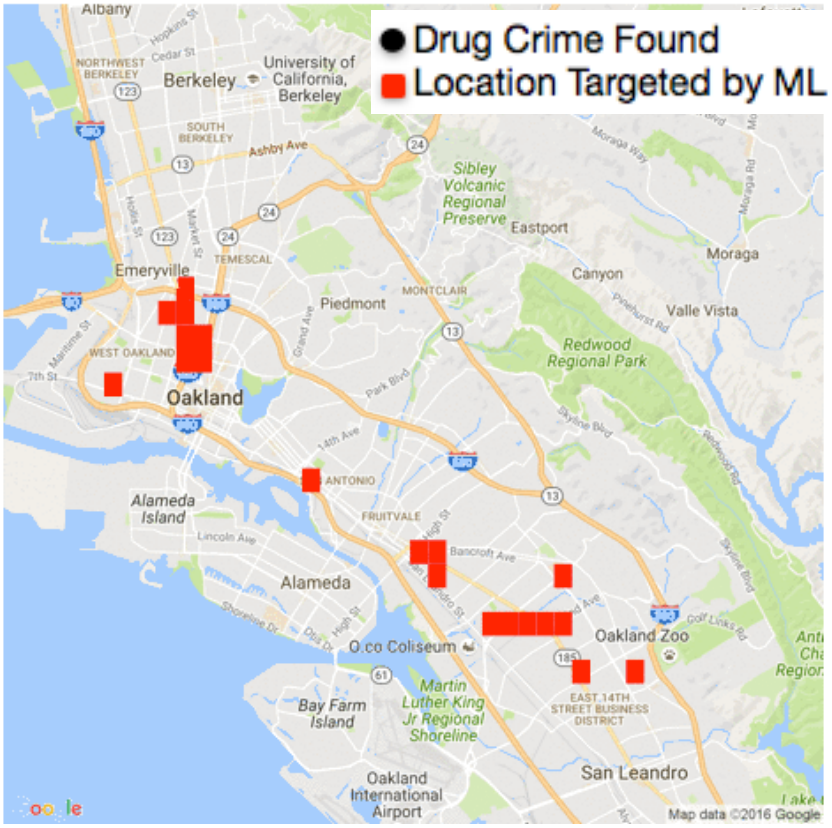
Drug Crimes in Oakland, CA

2010 Oakland PD Drug Crime Reports

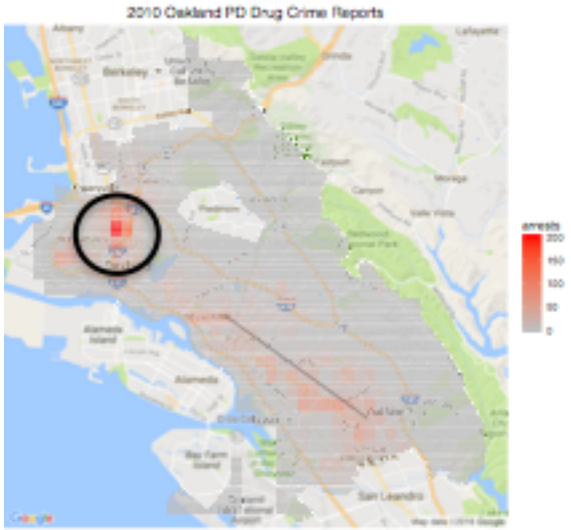


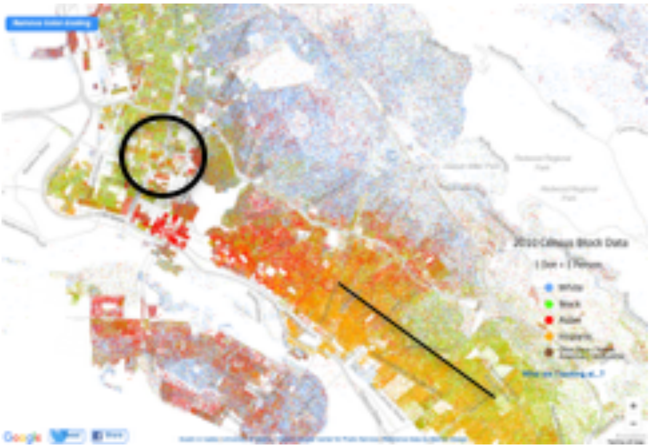


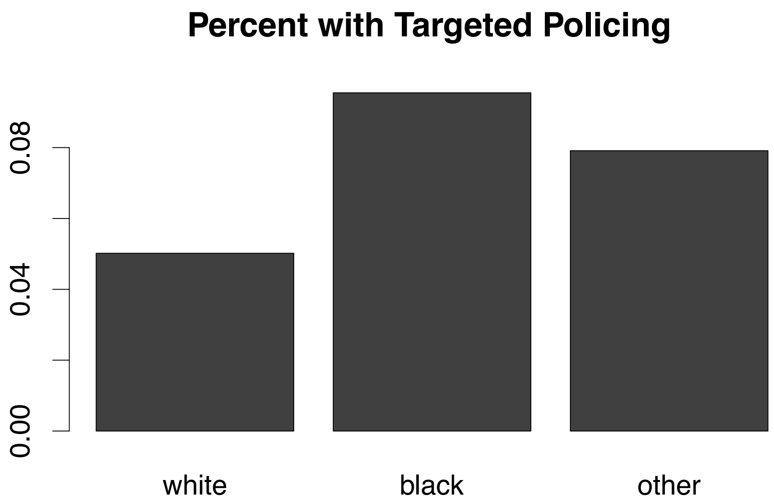
2011-Jan-01

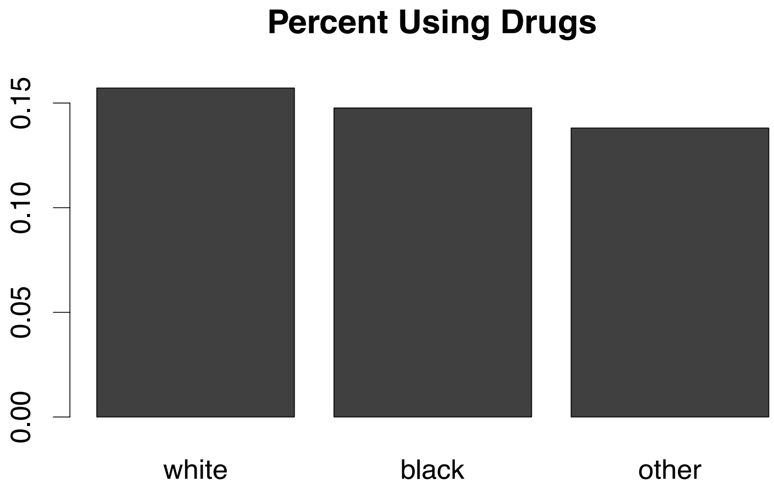


Drug Crimes in Oakland, CA



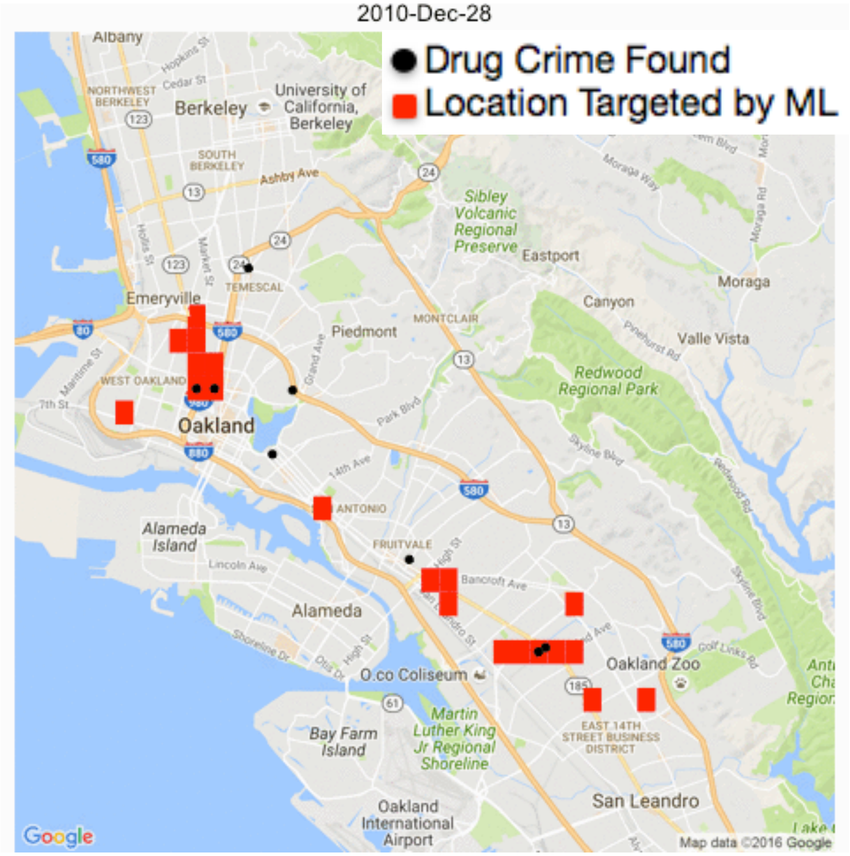


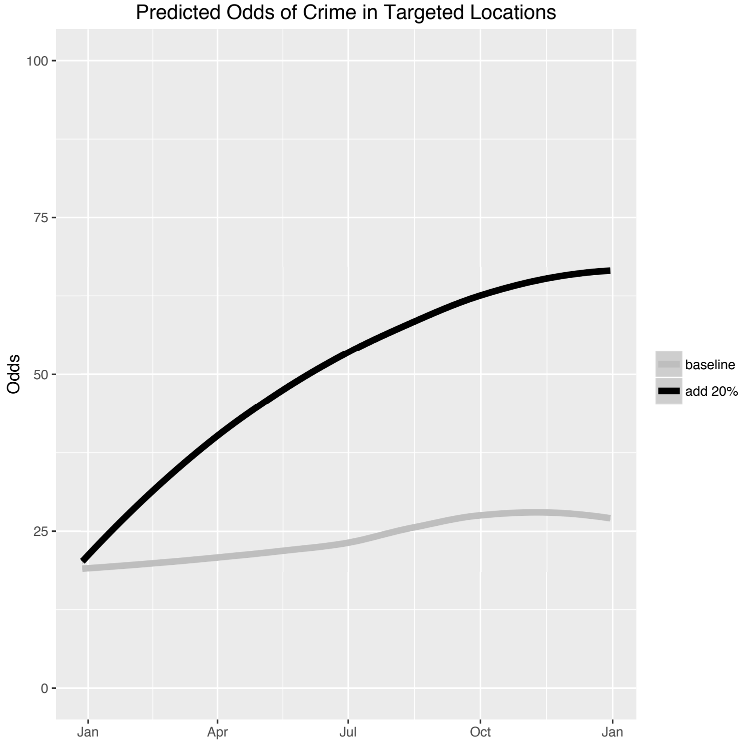




What if...

When police are sent to a location, they find a little more crime than they would have?





Audience Q&A

Lucy Bernholz

Conference Co-Host; Senior Research Scholar, Stanford PACS; Director, Stanford Digital Civil Society Lab

@p2173

Andrew Means

Conference Co-Host; Head, beyond.uptake

@meansandrew

Kristian Lum

Lead Statistician,
Human Rights Data Analysis Group

@KLdivergence

Candace Thille

Founding Director,
Open Learning Initiative,
Carnegie Mellon University and
Stanford University

@StanfordEd