CMSC/LING/STAT 208: Machine Learning

Abhishek Chakraborty [Much of the content in these slides have been adapted from *An Introduction to Statistical Learning: with Applications in R*, James et al. and *Hands-On Machine Learning with R*, Boehmke and Greenwell]

Unsupervised Learning

- **Supervised learning** problems involve a set of p features $X_1, X_2, \ldots, X_p$ and a response Y measured on n observations.
  - Objective is prediction and explain (if possible) the relation between response and predictors.
- In **unsupervised learning** problems, we observe only the features $X_1, X_2, \ldots, X_p$.
  - Objectives can be to visualize the data, or,
  - discover subgroups among variables or among observations.

We will discuss two methods:

- **Principal Components Analysis (PCA)**: Used for data visualization and pre-processing.
- **Clustering**: Discover unknown subgroups in data.

Unsupervised Learning

- Unsupervised learning problems tend to be more subjective than supervised learning problems.
- Often performed as part of an **exploratory data analysis**.
- Difficult to assess the results obtained from unsupervised learning methods.
- It is easier to obtain **unlabeled data**.

Principal Components Analysis (PCA)

PCA seeks a low-dimensional representation of a dataset that captures as much of the information as possible. PCA serves as a tool for

- Data compression
- Data visualization

The principal components are linear combinations of the p original features subject to certain constraints.

PCA: Example

USArrests dataset

```
library(ISLR2)   # load package
data("USArrests")   # load dataset

head(USArrests)   # first six observations
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

PCA: Example

USArrests dataset

```
cor(USArrests)   # correlation matrix of variables
```

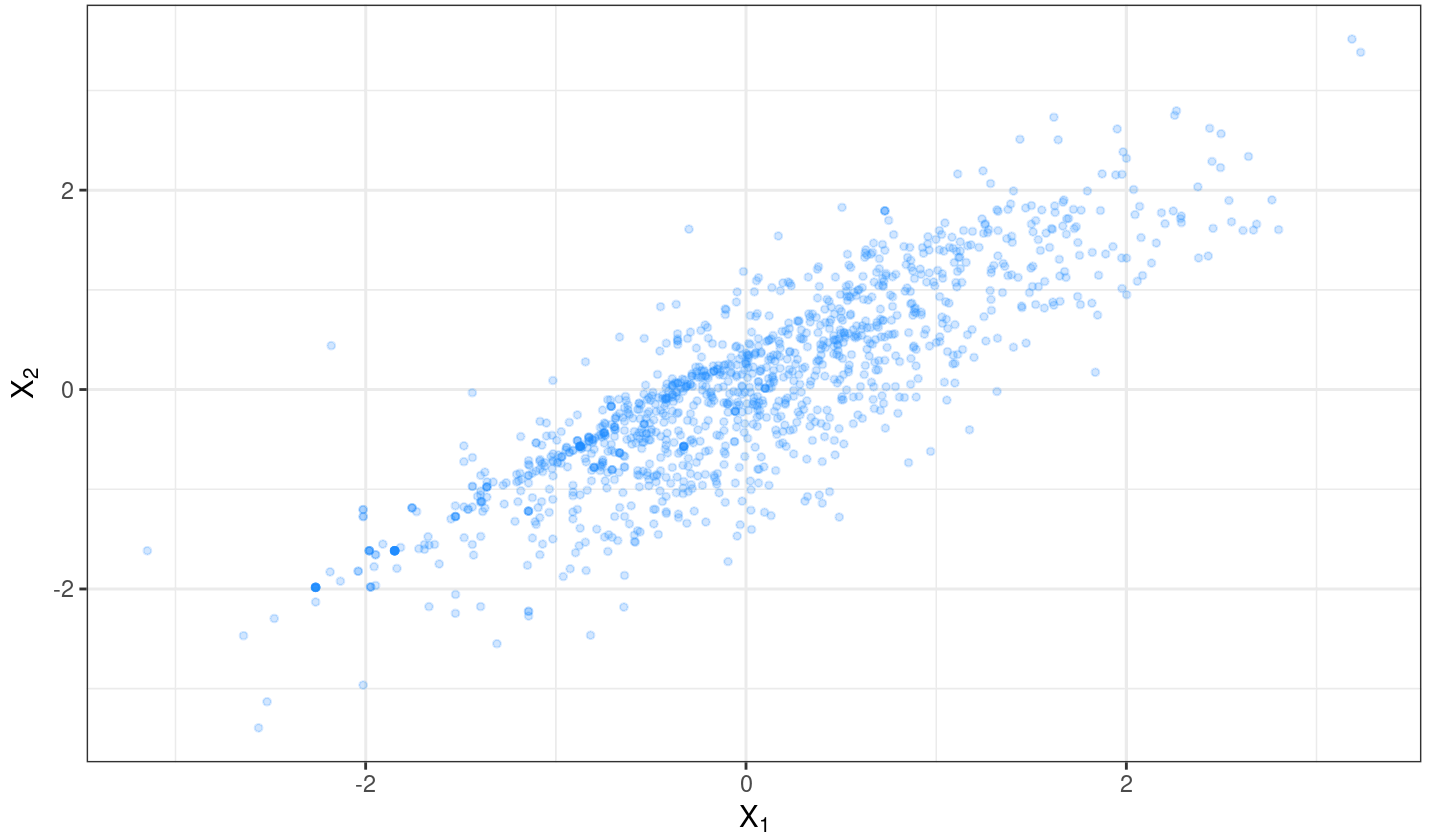
```
##              Murder   Assault   UrbanPop      Rape
## Murder   1.00000000 0.8018733 0.06957262 0.5635788
## Assault  0.80187331 1.0000000 0.25887170 0.6652412
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
## Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

PCA: Data Requirements

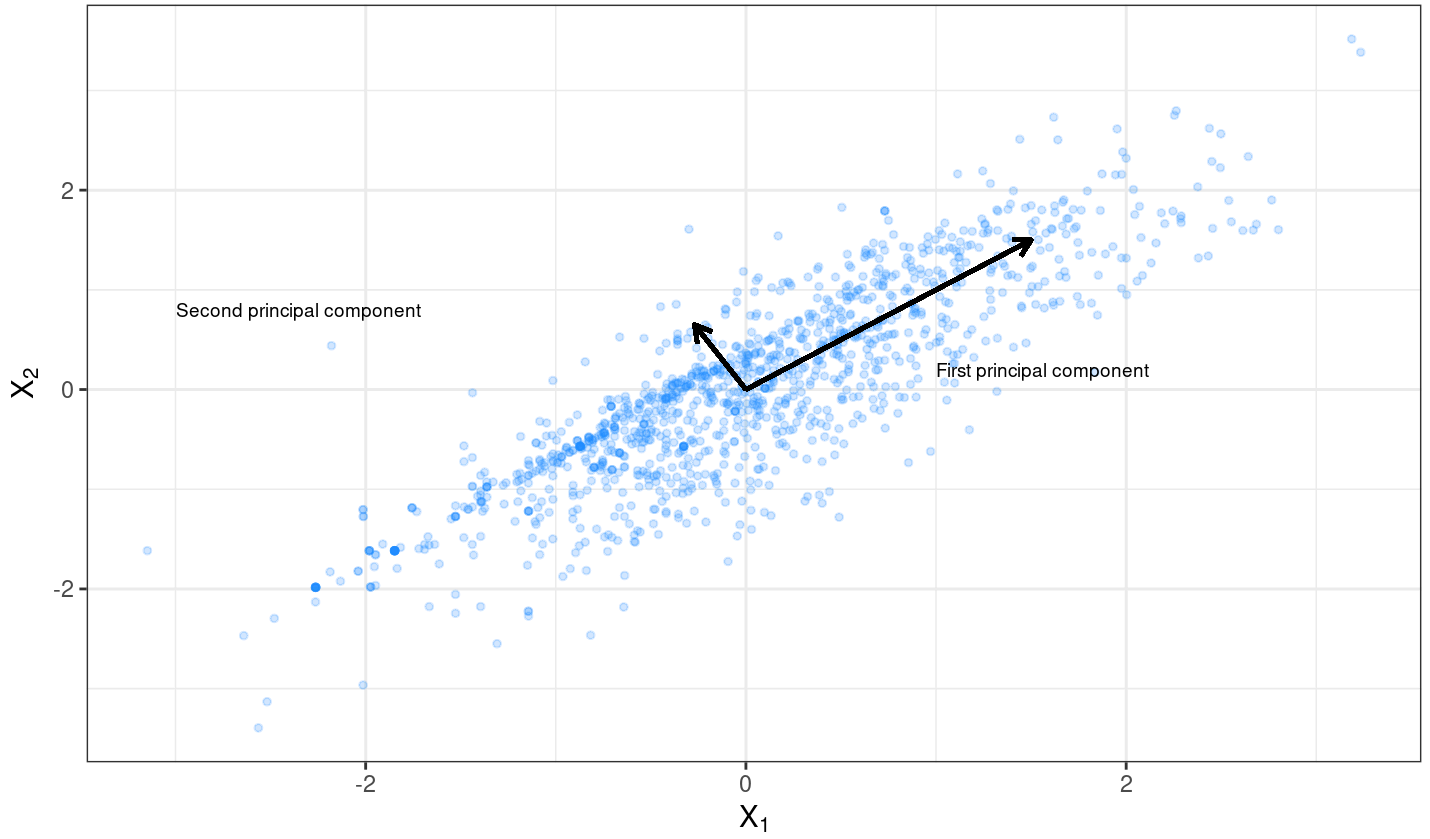
To perform dimension reduction techniques in R, generally, the data should be prepared as follows:

- Data are in tidy format per Wickham et al. (2014);
- Any missing values in the data must be removed or imputed;
- Typically, the data must all be numeric values (e.g., one-hot, label, ordinal encoding categorical features);
- Numeric data should be standardized (e.g., centered and scaled) to make features comparable.

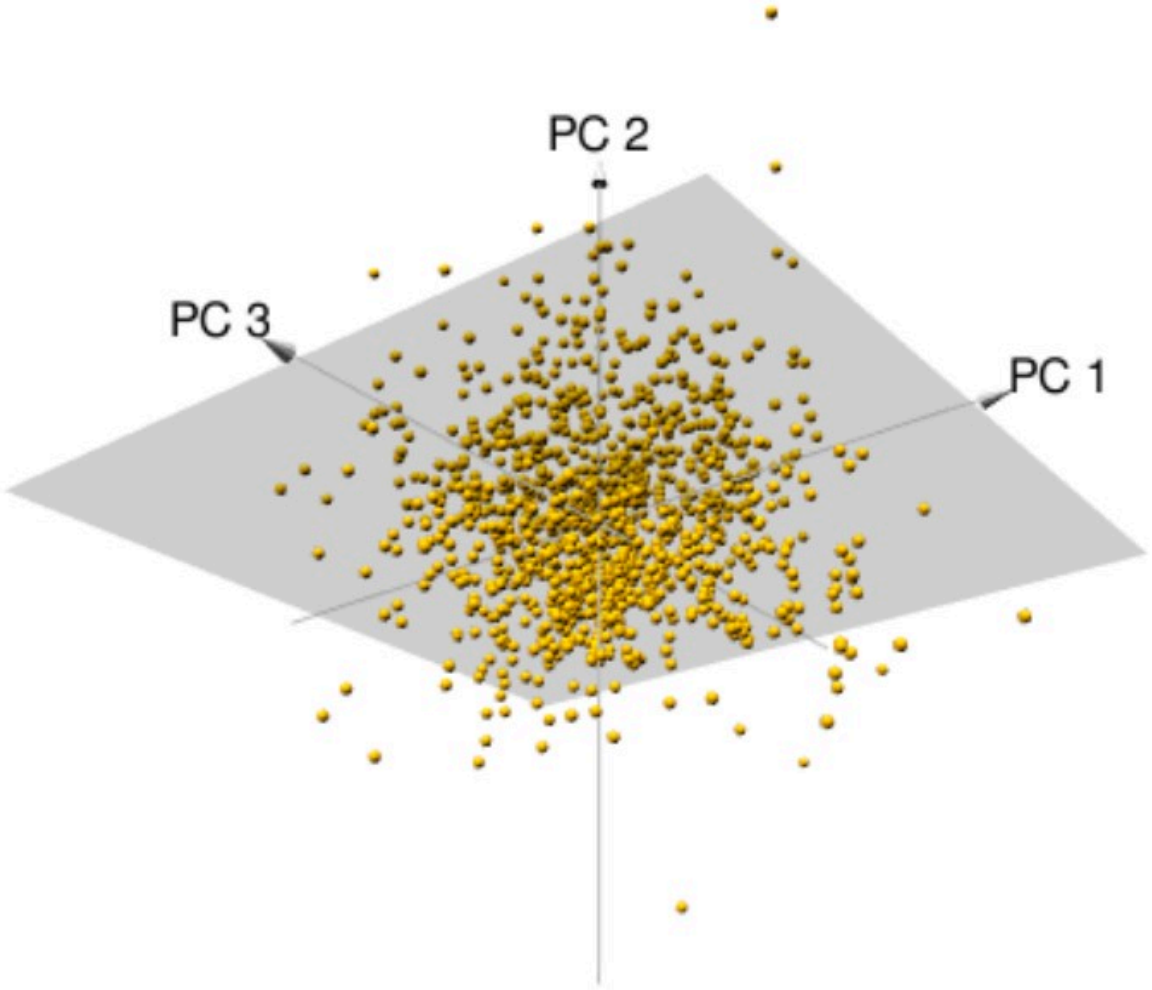
PCA: Toy Example



PCA: Toy Example



PCA: Toy Example



PCs with 3 features. [Adapted from HMLR, Boehmke & Greenwell]

PCA: First PC

The **first principal component** Z_1 of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features that has the largest variance.

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p$$

By **normalized**, we mean $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

For the i^{th} observation,

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \ldots + \phi_{p1} x_{ip}$$

The elements $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ are **loadings of the first PC**. The loadings make up the **first PC loading vector** $\phi_1 = \begin{pmatrix} \phi_{11} & \phi_{21} & \ldots & \phi_{p1} \end{pmatrix}^T$.

$z_{11}, z_{21}, \ldots, z_{n1}$ are the **first PC scores**.

PCA: First PC

Suppose an $n \times p$ feature matrix $\mathbf{X}$.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

The first PC is obtained by solving

$$\underset{\phi_{11},\dots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

PCA: Second PC

The **second principal component** Z_2 is the linear combination of $X_1, \ldots, X_p$ that has maximal variance among all linear combinations that are **uncorrelated** with Z_1.

The second PC scores are $z_{12}, z_{22}, \ldots, z_{n2}$ where

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \ldots + \phi_{p2} x_{ip}$$

ϕ_2 is the **second PC loading vector** with **loadings** $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$.

Z_2 **uncorrelated** with Z_1 is equivalent to ϕ_2 being orthogonal (perpendicular) with ϕ_1.

PCA: How Many PCs to Use?

- We would like to use the smallest number of PCs required to get a good understanding of the data.
- CV cannot be implemented to answer this question.
- Two common approaches in helping to make this decision (depends on the objective and analytic workflow):
    - Proportion of variance explained (PVE)
    - Screeplot. Look for an **elbow**.

Clustering

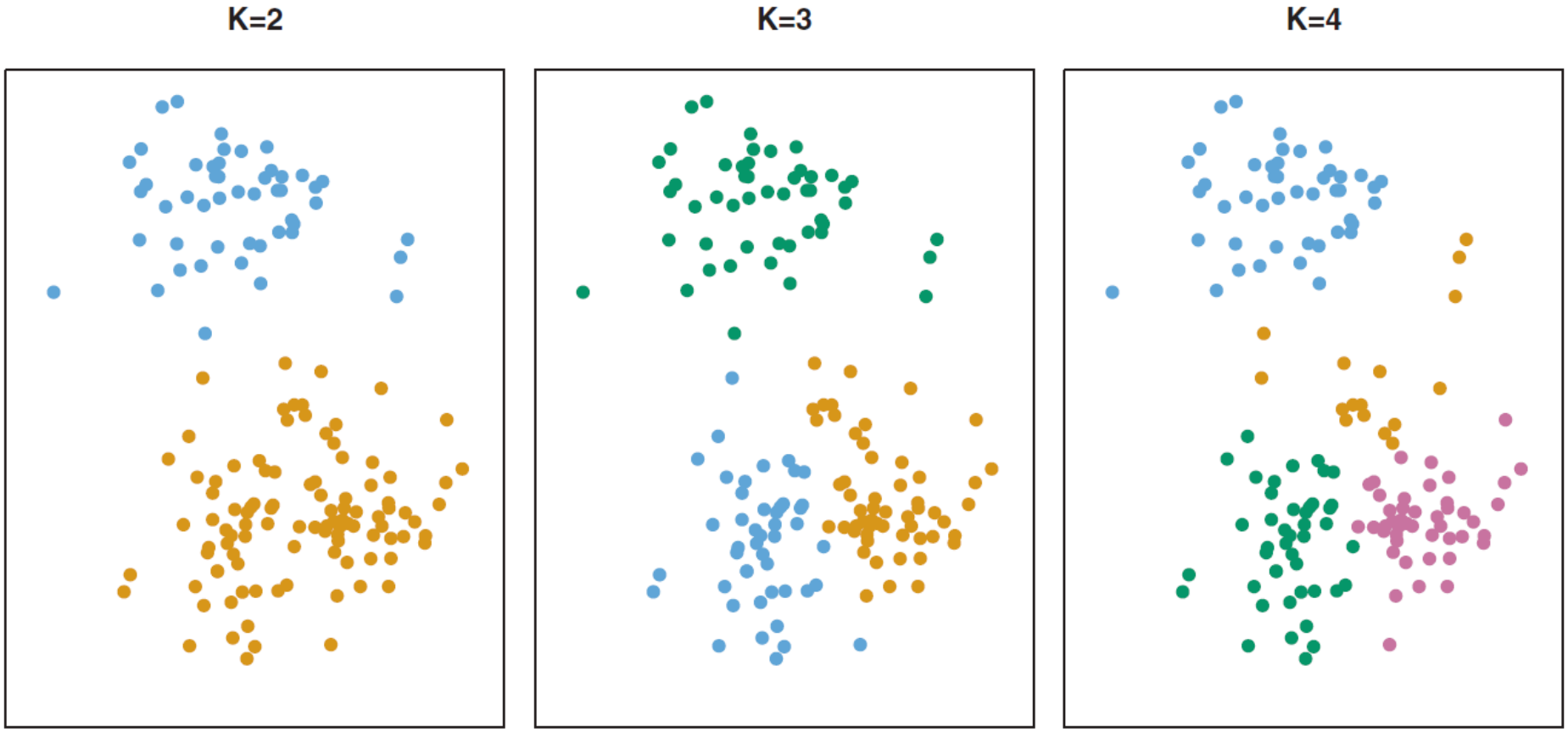
- Broad class of techniques for finding **subgroups** or **clusters** in a dataset.
- Partition the data into distinct groups so that observations within each group are similar to each other.
- Definition of similarity depends on the context and the dataset being studied.
- We will talk about:
  - K-means clustering
  - Hierarchical clustering

Clustering: Applications

- Cancer research: n observations correspond to tissue samples for patients with different types of cancer, p features correspond to gene expression measurements.
- Market segmentation: n observations on p variables. Identify subgroups of people who might be more receptive to a particular form of advertising.
- Social network analysis, astronomical data analysis, organizing computing clusters etc.

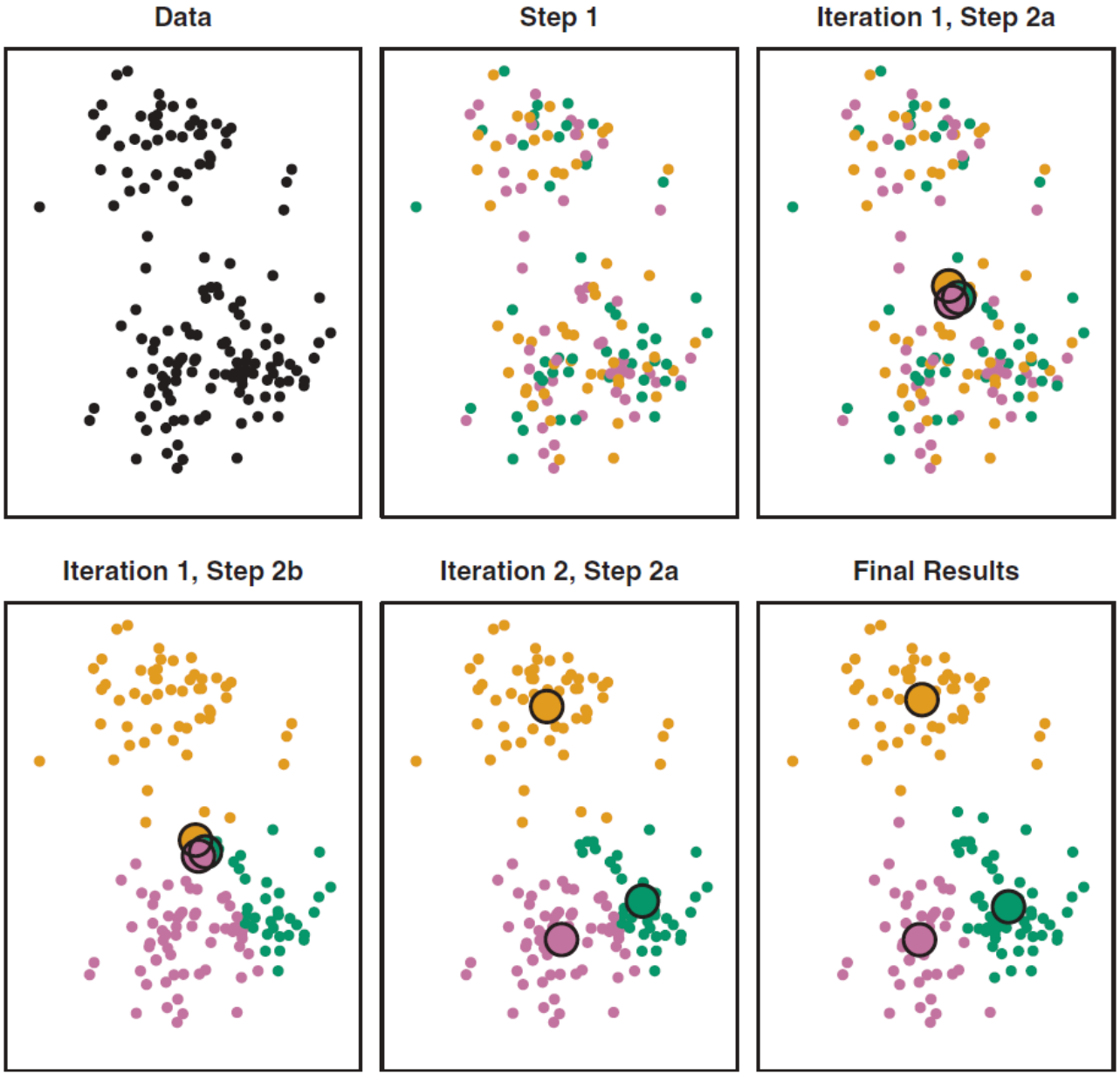
K-Means Clustering

Partition the dataset into a pre-specified number of K distinct, non-overlapping clusters.



The coloring (ordering) of the clusters is arbitrary.

K-Means Clustering



K-Means Clustering

The idea behind K-means clustering is that a **good clustering** is one for which the **within-cluster variation** is as small as possible. The resulting clusters are such that

- each observation belongs to at least one cluster, and
- clusters are non-overlapping, no observation belongs to more than one cluster.

K-Means Clustering Formulation

The **within-cluster variation** for cluster C_k is a measure $W(C_k)$ of the amount by which the observations within a cluster differ from each other. Thus the objective is to

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

where

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Combining, we have,

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

K-Means Clustering Algorithm

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

K-Means Clustering Algorithm

- The algorithm is guaranteed to decrease the value of the objective at each step since

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

where $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$: mean of j^{th} feature in cluster C_k.

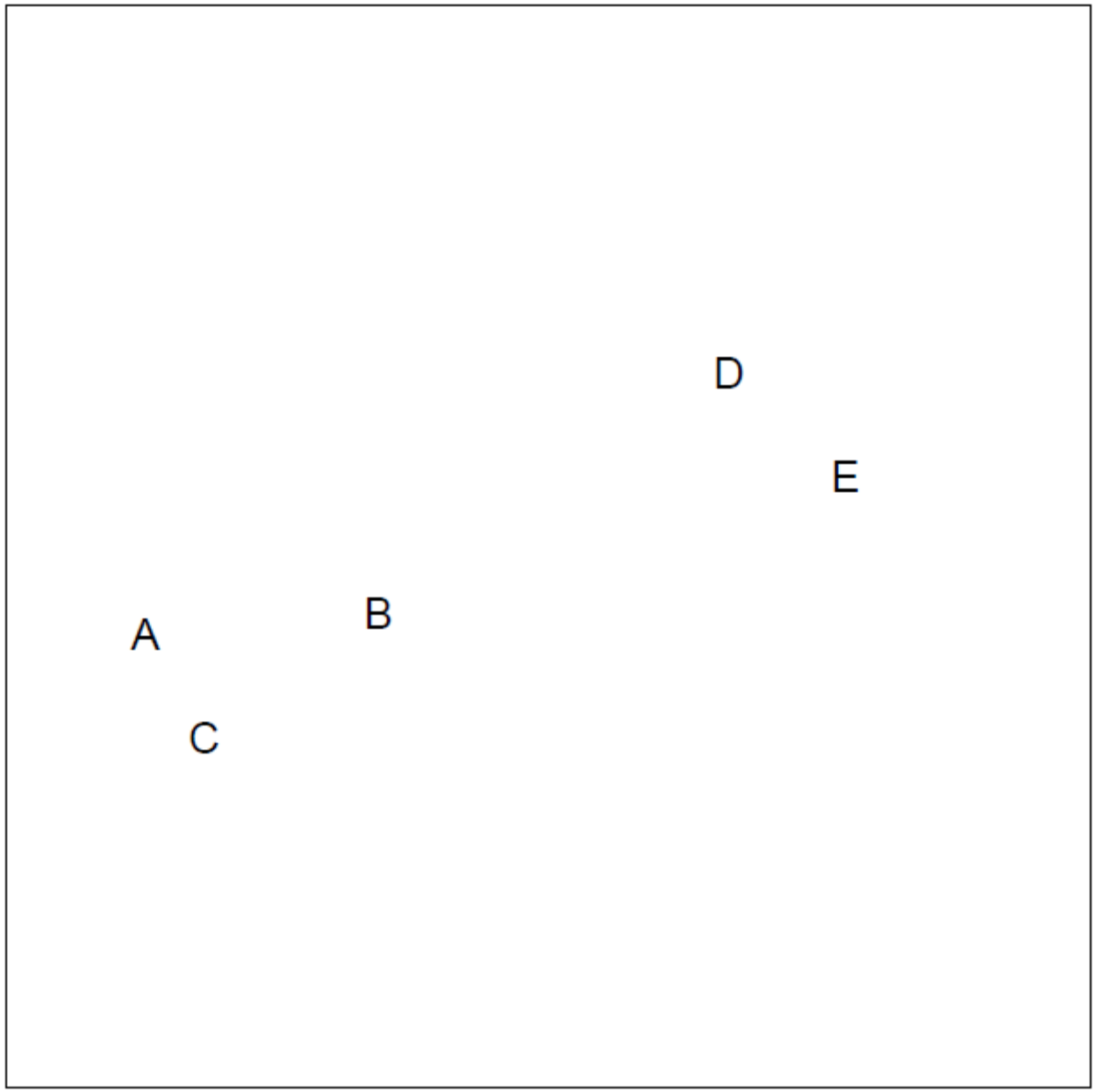
- The algorithm is not guaranteed to find the global optimum.
- Results depend on the initial (random) cluster assignments. It is recommended to run the algorithm multiple times from different random initial configurations.

K-Means Clustering Algorithm

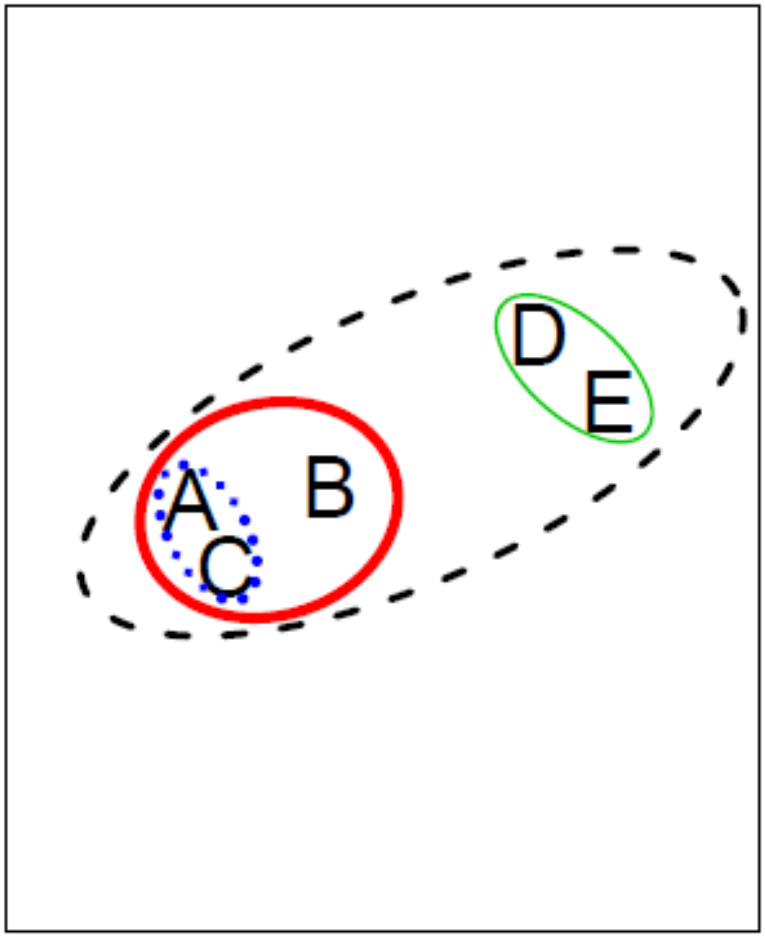
Hierarchical Clustering

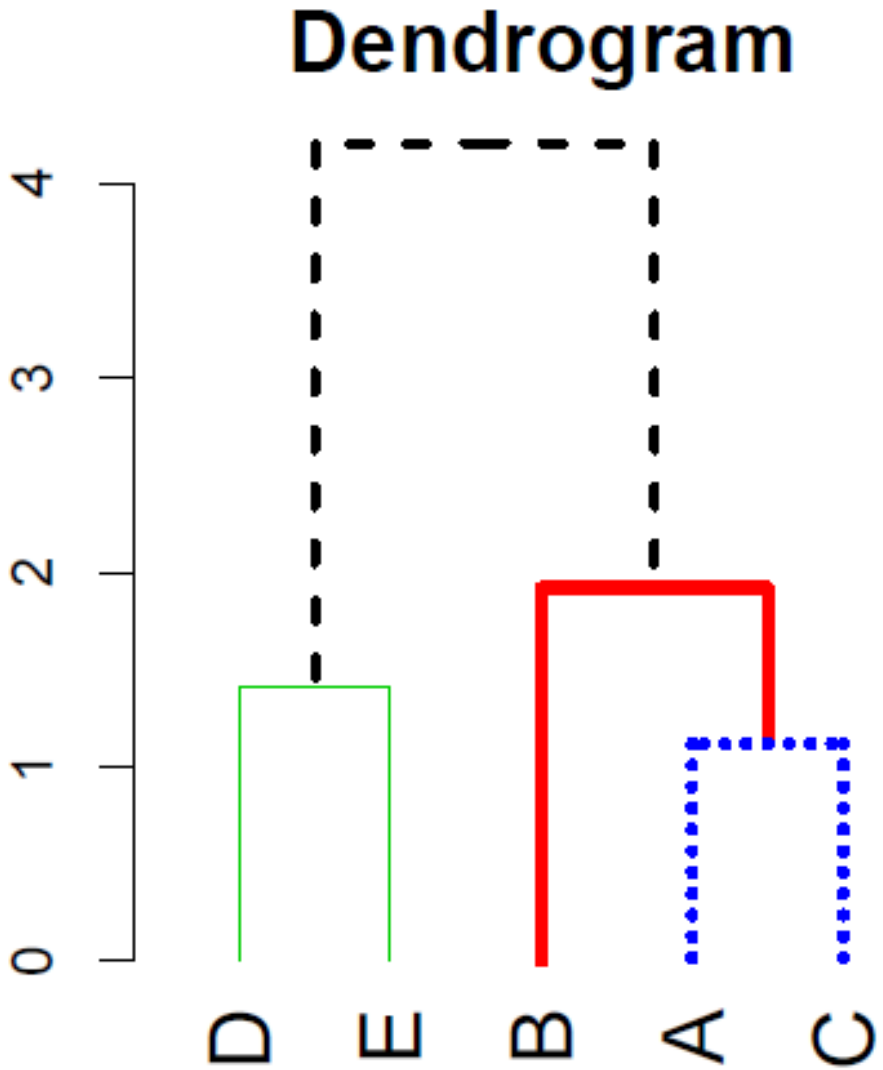
- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage.
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.
- Hierarchical clustering results in a tree-based representation of the observations, called a **dendrogram**.
- We discuss **bottom-up** or **agglomerative** clustering. This is the most common type of hierarchical clustering. The dendrogram is built starting from the leaves (bottom) and combining clusters up to the trunk (top).

Hierarchical Clustering



Hierarchical Clustering





Hierarchical Clustering: Types of Linkage

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Hierarchical Clustering Algorithm

Algorithm 10.2 *Hierarchical Clustering*

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

Practical Issues in Clustering

- Scaling of the variables matters.
- Choices in hierarchical clustering.
  - Dissimilarity measure
  - Type of linkage
  - Where to cut the dendrogram?
- Choices in K-means clustering.
  - What should be the value of K?
- Clustering methods are not very robust to perturbations to the data.
- Which features should be used for clustering?
- For more details, see *Elements of Statistical Learning*, Chapter 14.

To Sum It All Up

Other Things To Watch Out For

Spring works at a juice-packing company where she uses a machine-learning model to predict the demand for different juice flavors. After training and evaluating the model, Spring was confident that it was ready for production. The model was deployed, and the company started using it to plan its production and distribution.

During the first few months, everything was working as expected. But then, the company noticed that the model consistently overestimated the demand for certain flavors. What could be the cause of the problem with the model?

- The model is underfitting and needs more complexity.
- The model is overfitting and needs more regularization.
- The model is suffering from data drift.
- The model is suffering from sampling bias.

Other Things To Watch Out For

Spring's been working on a model to classify photos of food. Her company is building an application that will let users snap a picture of a plate at a restaurant and show them a potential recipe so they can cook it at home. After a year of work, Spring's model was working great. The company launched the model worldwide and started monitoring user feedback.

Unfortunately, users from an Asian country complained because the model wasn't working for them. What is the most likely reason for the problem?

- Spring's model didn't have enough complexity to learn all the data, so it's normal to have problems with certain regions.
- Spring needed to train the model for more time to fully capture the dataset's information.
- Spring's model is suffering from data drift.
- Spring's model is suffering from sampling bias.

ML Ethics and Fairness

- As recently as 2015, a major corporation reportedly utilized a model to evaluate applicants' resumes for technical posts. They scrapped this model upon discovering that, by building this model using resume data from its current technical employees (mostly men), it reinforced a preference for male job applicants. (see Dastin 2018)
- Facial recognition models, increasingly used in police surveillance, are often built using image data of researchers that do not represent the whole of society. Thus, when applied in practice, misidentification is more common among people that are underrepresented in the research process. Given the severe consequences of misidentification, including false arrest, citizens are pushing back on the use of this technology in their communities (see Harmon 2019)
- In 2020, the New York Civil Liberties Union filed a lawsuit against the U.S. Immigration and Customs Enforcement's (ICE) use of a "risk classification assessment" model that evaluates whether a subject should be detained or released (see Hadavas 2020). This model is notably unfair, recommending detention in nearly all cases.

ML Ethics and Fairness

Facial recognition

Interview

'A white mask worked better': why algorithms are not colour blind

Ian Tucker

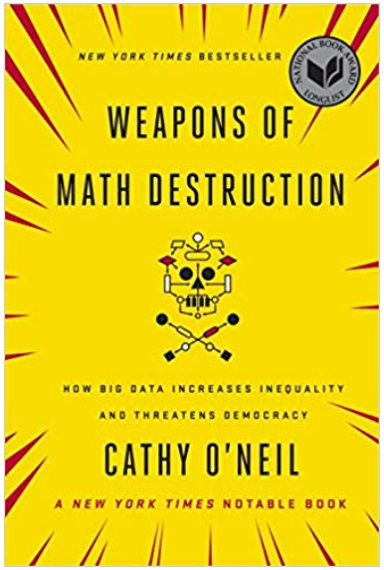
When Joy Buolamwini found that a robot recognised her face better when she wore a white mask, she knew a problem needed fixing

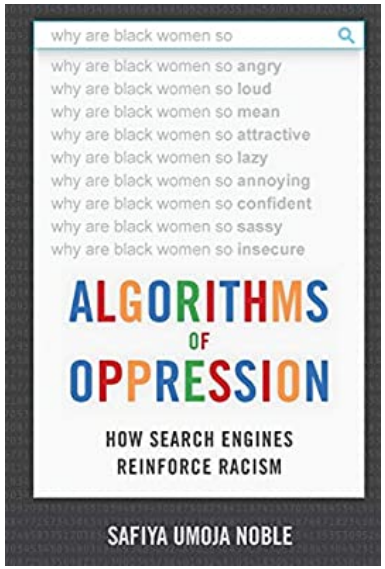
Sun 28 May 2017 13.27 BST

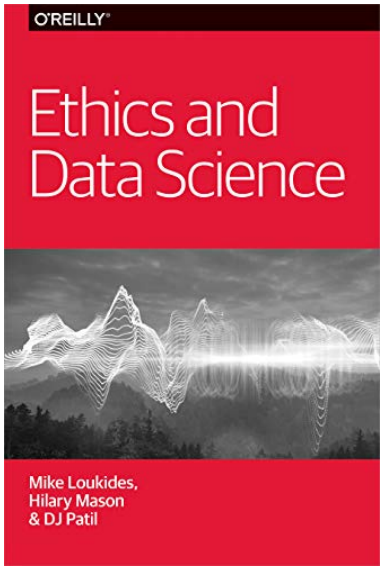
Joy Buolamwini is a graduate researcher at the MIT Media Lab and founder of the Algorithmic Justice League - an organisation that aims to challenge the biases in decision-making software. She grew up in Mississippi, gained a Rhodes scholarship, and she is also a Fulbright fellow, an Astronaut scholar and a Google Anita Borg scholar. Earlier this year she won a $50,000 scholarship funded by the makers of the film *Hidden Figures* for her work fighting coded discrimination.

Ian Tucker. 'A white mask worked better': why algorithms are not colour blind.

Further Readings







Next Steps

- Statistics and Data Science Minor
- Future readings
  - *Elements of Statistical Learning*, HTF
  - *Feature Engineering & Selection*, KJ
  - *Deep Learning*, GBC
  - *Supervised Machine Learning for Text Analysis in R*, HS
  - Optimization Algorithms
  - Linear Algebra
  - ML with Python