NYC DATA SCIENCE ACADEMY

Principal Component Analysis

Data Science Bootcamp

Outline

- Part 1: Taking a New Perspective
- Part 2: Dimension Reduction
- Part 3: Vectors of Highest Variance
- Part 4: The PCA Procedure

PART 1

Taking a New Perspective

Taking a New Perspective

"Perspective is everything when you are experiencing the challenges of life."

-Joni Eareckson Tada

Taking a New Perspective

A riddle...

Taking a New Perspective

How many didn’t?

Taking a New Perspective

Let's try that again...

Taking a New Perspective

There are 30 cows in a field,

and 20 ate chickens.

How many didn't?

PART 2

Dimension Reduction

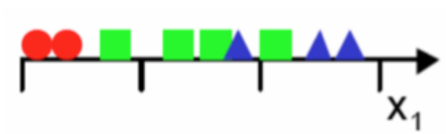
When can High Dimensionality be Adverse?

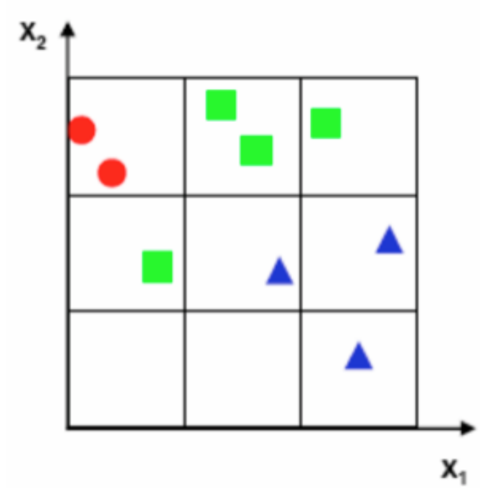
- Sparsity becomes exponentially worse as the dimensionality of our data increases.

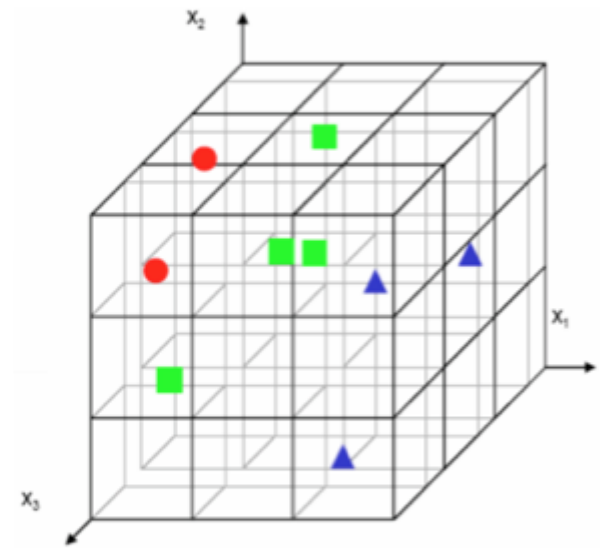
- Given a number of observations, additional dimensions spread the points out further and further from each other.

- There tends to be insufficient repetition in various regions of the high-dimensional space. Less repetition makes inference more difficult:
  - Are the results replicable?
  - What about regions that don’t have any observations at all?

When can High Dimensionality be Adverse?







9 observations 3 sections	9 observations 9 sections	9 observations 27 sections

When can High Dimensionality be Adverse?

- Collecting data is expensive, both monetarily and temporally.
  - You might be working on a budget; the collection of additional variables may not be necessary and could hinder the return on your investment.

- There is too much complexity with higher-order data.
  - Often we not only seek the most accurate solution, but also one that is simple and interpretable.

- We may have redundancy in our measured dimensions.
  - While all the variables in our dataset might be different from one another, the information they contain as a group may overlap.

When can High Dimensionality be Adverse?

- Consider measuring the following on all students at a university:
  - Hours of sleep.
  - Hours of partying.
  - Hours in the library.
  - Number of tests taken.
  - Number of enrolled classes.
  - Number of meals eaten.
  - Amount of physical activity.
  - Amount of printer usage.

- While all these variables measure different things, they might be interrelated; is there a common factor that may help inform measurements on all of these variables? GPA?

Don’t Just Throw Away Data!

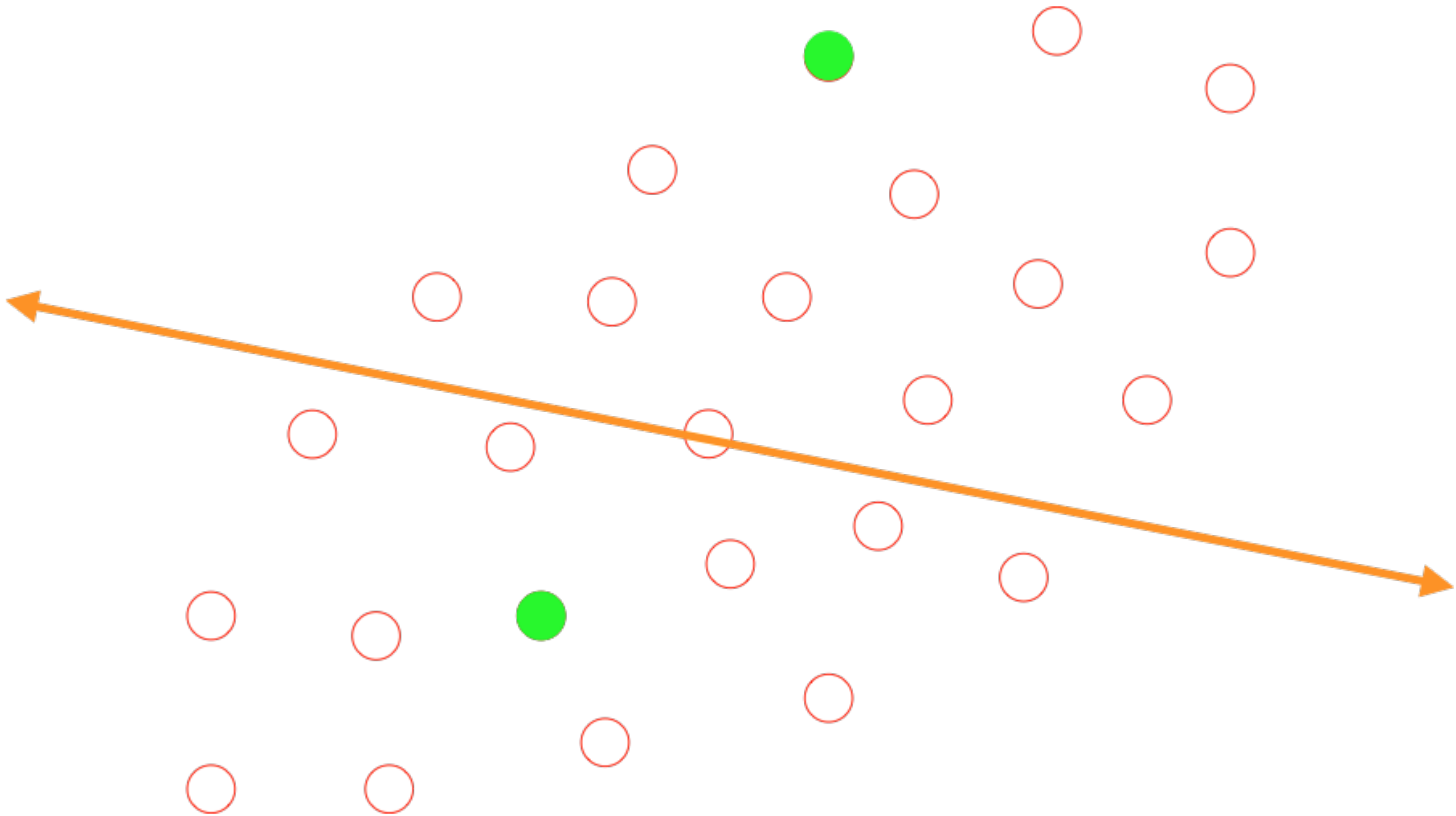
- Remember, data is a commodity. It is important to not frivolously disregard variables or observations without careful consideration.
  - Instead, be careful and selective in the dimensions we choose to analyze.

- Reducing the dimensionality of our data can often inform interpretability and statistical inference in the long run, but we can’t avoid losing some information.
  - Try to preserve as much structure from the original data as possible.

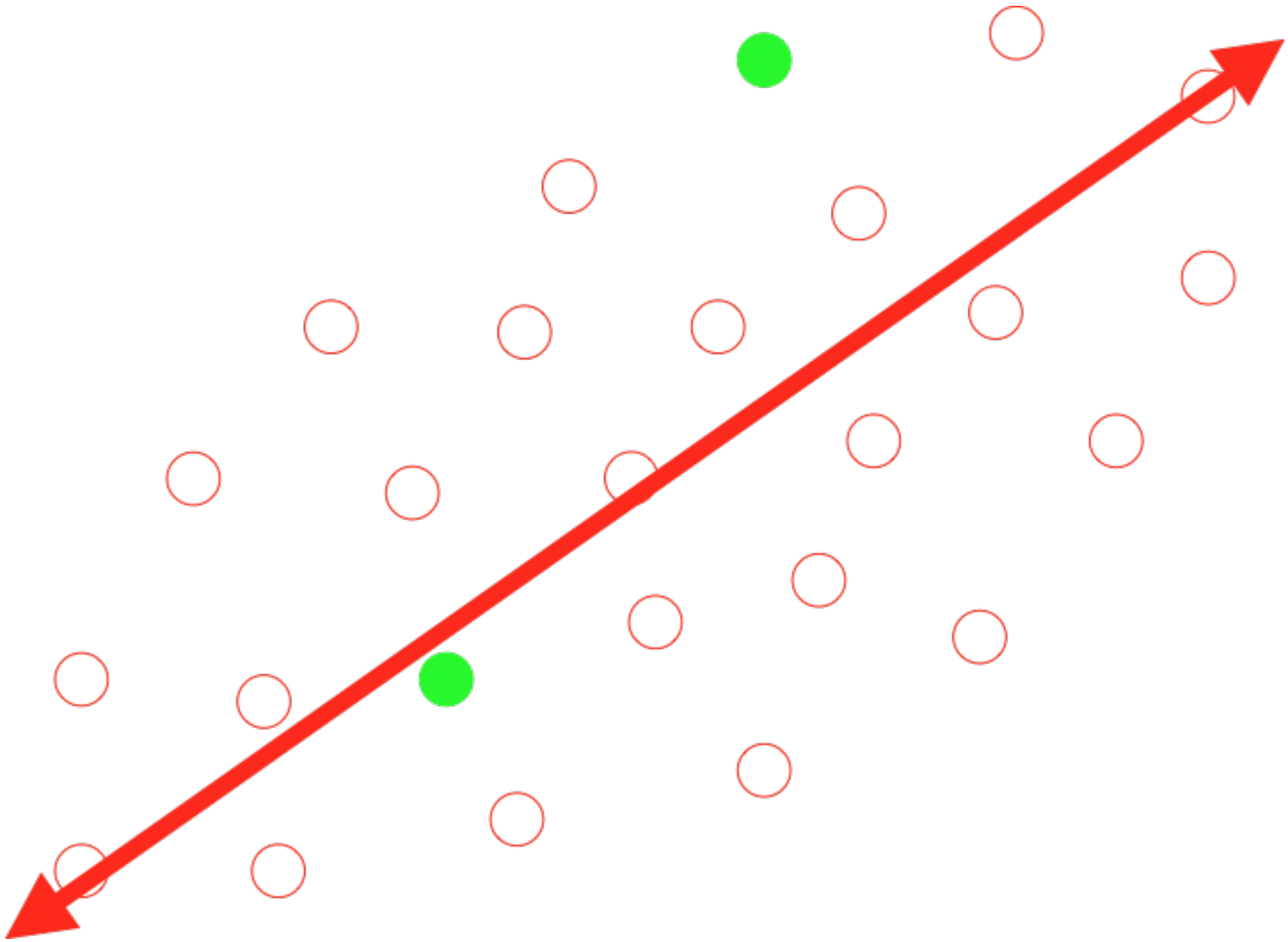
PART 3

Vectors of Highest Variance

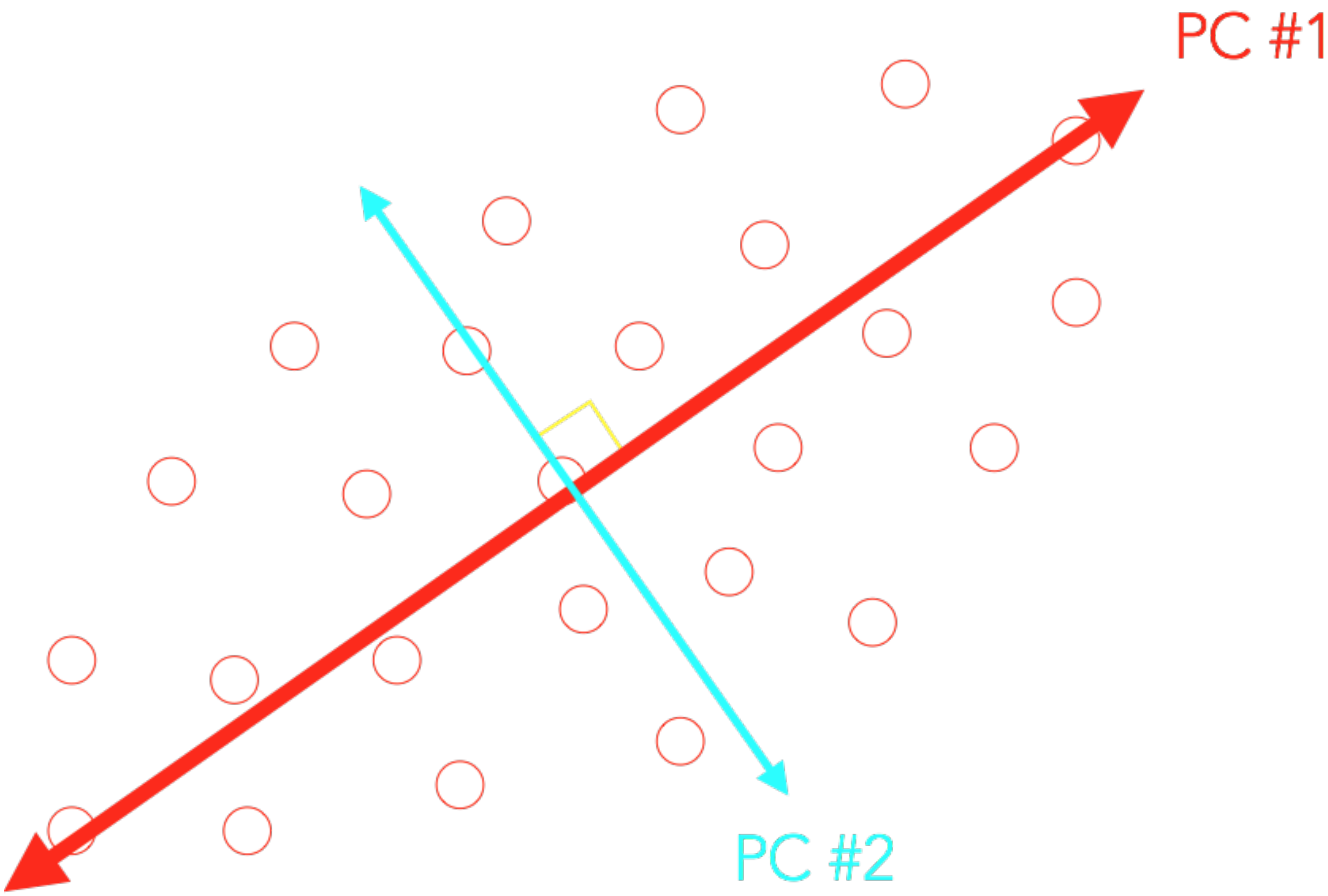
Vectors of Highest Variance



Vectors of Highest Variance



Vectors of Highest Variance



PART 4

The PCA Procedure

PCA Mathematically

- An overview of the PCA procedure mathematically:

- Center the data at 0 by subtracting off the mean from each variable:
  - Pragmatically, this allows the future mathematical processes to be easier.
  - Conceptually, PCA is modeling the variances of the data -- the mean doesn't matter as much. We can always add the mean back in later if we desire to do a bit of back-construction.

$$x'_{i,j} = x_{i,j} - \mu_j$$

PCA Mathematically

- Compute the covariance matrix Σ:
  - Observe a unique property of convergence.

$$\Sigma v$$

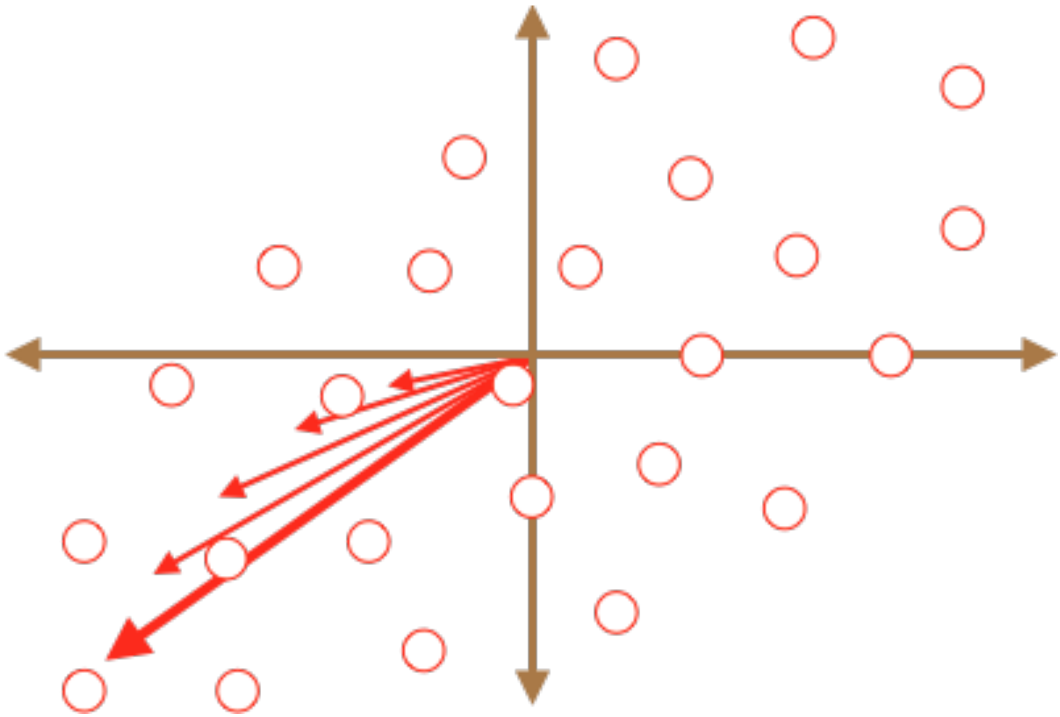
$$\Sigma(\Sigma v)$$

$$\Sigma(\Sigma(\Sigma v))$$

$$\Sigma \ldots \Sigma(\Sigma(\Sigma v)) \approx e$$

PCA Mathematically

- Compute the covariance matrix Σ:
  - Observe a unique property of convergence.



PCA Mathematically

- Find the eigenvectors e of Σ:
  - Solve the equation:

$$det(\Sigma - \lambda I) = 0$$

  - Compute the eigenvectors by finding the solutions to:

$$\Sigma e = \lambda e$$

  - The principal components are the eigenvectors e.
  - The eigenvectors are ordered by the magnitude of the corresponding eigenvalues λ.

PCA Mathematically

- Determine how many principal components to use:
  - Strike a balance between the total amount of variance that is captured by the principal components and the number of principal components selected.
  - Use the first k principal components.

- Project the original data onto the chosen k principal components.

The Result of PCA

- What do we get?
  - Transformed data that straddles only k carefully selected dimensions that preserve as much original structure as possible.

- Same data, new perspective.

Other Properties of PCA

- The following results are useful properties that can be proved using calculus and linear algebra (omitted for brevity).

- The eigenvectors of Σ yield orthogonal directions of greatest variability (principal components).

- The eigenvalues λ correspond to the magnitude of variance along the principal components.