HUDK 4050:

CORE METHODS

IN EDM

10/30/18 4:09 PM

In the news

How The Education Sector Is Using Data Analytics To Revamp Pedagogy

Analytics India MAGAZINE

Analytics & Machine Learning Expertise Drive Early Detection and Monitoring of Chronic Disease

HEALTHCARE Analytics News

4 human-caused biases we need to fix for machine learning

TNW

QUARTZ

Psychologists have developed a new 10-minute “intelligence” test

Data reveals Amazon has banned more than 5,700 of its top reviewers in the last 2 years as it increasingly cracks down on review abuse

stamford advocate

stackoverflow

Make Stack Overflow More Welcoming

We launched a new Code of Conduct in August that reinforces our commitment to mutual respect and kindness.

Read the Code of Conduct

We improved flagging so it's simpler to report comments that are abusive or unkind. We also introduced the New Contributor Indicator to make it easier to identify and respond to new users.

Events

Event	Date	URL
DSI: Towards Better Reinforcement Learning for High Stakes Domains	5:30pm November 1	https://www.eventbrite.com/e/new-york-data-science-seminar-series-emma-brunskill-stanford-tickets-51551174952
Weekly Coding Session: Viz	1:00pm November 2	Macy 262
TCLA Social Hour	7:00pm November 7	https://goo.gl/forms/uYVHwVUNT0bnDFbI3
Data Law in a Global Digital Economy	November 9	https://www.guariniglobal.org/data-law)
NYAS: Deep Learning to Accelerate Drug Development	November 13	https://www.nyas.org/events/2018/deep-learning-to-accelerate-drug-development-an-nyc-symposium/?utm_source=The+New+York+Academy+of+Sciences&utm_campaign=f0807f47cb-eNews_October_2018-10-18&utm_medium=email&utm_term=0_cba25b11d2-f0807f47cb-184577937&mc_cid=f0807f47cb&mc_eid=cfeeec7fb2
People centric approach to optimize Data Science, Commercial impact and Leadership	10:30am November 14	https://events.columbia.edu/cal/event/eventView.do?b=de&calPath=%2Fpublic%2Fcals%2FMainCal&guid=CAL-00bb9e24-655b8449-0165-5e0ea7e9-00001957events@columbia.edu&recurrenceId=
Cross-device User Clustering at Adobe	5:30 November 29	https://events.columbia.edu/cal/event/eventView.do?b=de&calPath=%2Fpublic%2Fcals%2FMainCal&guid=CAL-00bb9e28-655b8cee-0165-5dd5c72b-00001287events@columbia.edu&recurrenceId=
Machine Learning Innovation Summit	December 12-13	https://www.theinnovationenterprise.com/summits/machine-learning-innovation-summit-new-york-2018

Anonymous Check In

bit.ly/HUDK4050-Checkin

Assessment

- Github contains all assignments: one assignment due per week for the rest of the semester
- Ask question on Stack Overflow (pull requisition to /so-question)
- Final group project will be a video
- Rate project videos

Class 27 - Work Session: Assignment 8, Group Project (12/6/18)

Class 28 - Work Session: Assignment 8, Group Project (12/11/18)

Due: Assignment 7 - Diagnostic Metrics

Class 29 - Rate video presentations (12/13/18)

Class 30 - Rate video presentations (12/18/18)

EVERYTHING DUE - 12/20/18

Principal Component Analysis

Grouping stuff

By Variables

ID	Var1	Var2	Var3
A			
B			
C			
D			

ID	Var2
A	
B	
C	
D	

Selection

ID	Var2+3
A	
B	
C	
D	

Extraction

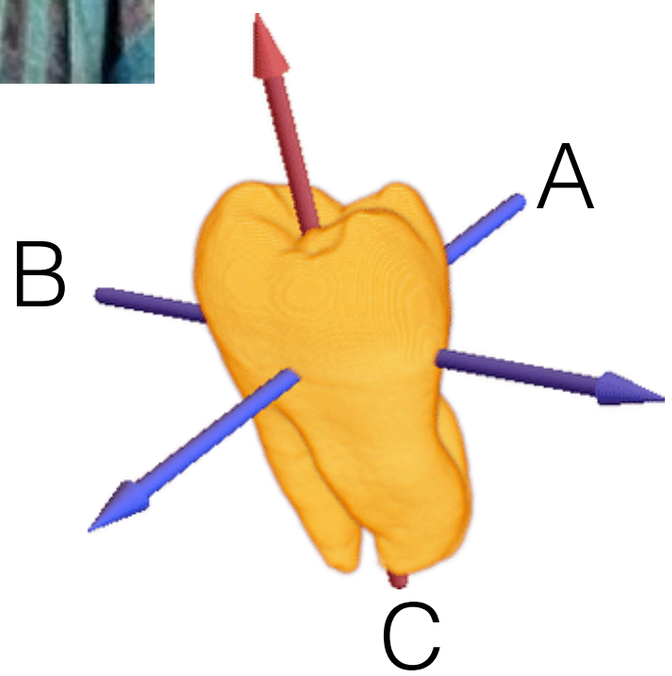
By People

ID	Var1	Var2	Var3
A			
C			

ID	Var1	Var2	Var3
B			
D			

History

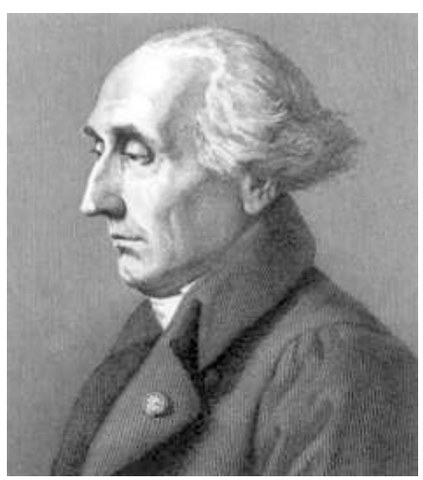
- Part of a set of issues called “Eigen Problems”
- Arose as a subset of phenomena related to differential equations (Your old buddy Euler, c.1750)
- Principal Axes



History

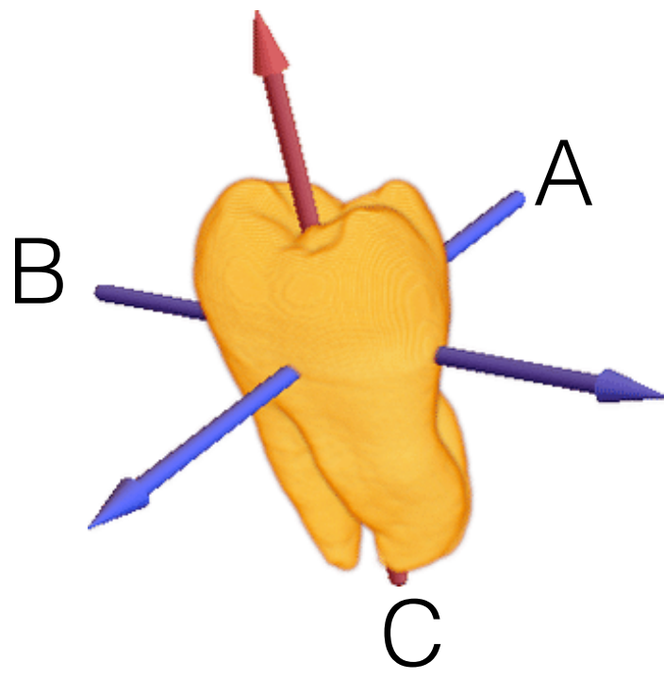
- Describe inertia as a matrix of measurements from the center of an object as it moves

(Joseph-Louis Lagrange)



$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

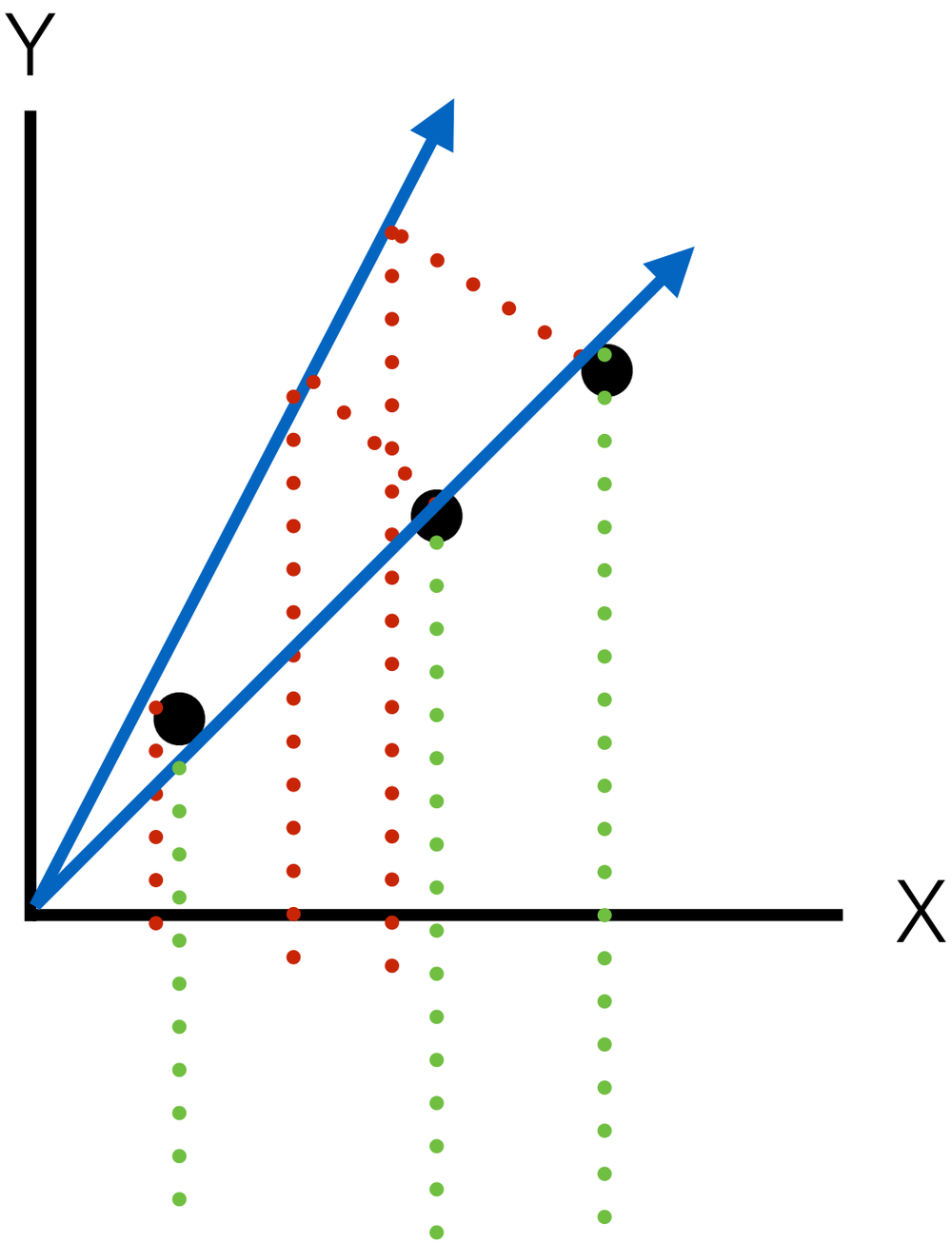
- Principal axes = the lines through which you can describe the object, while maximizing the amount of variation maintained

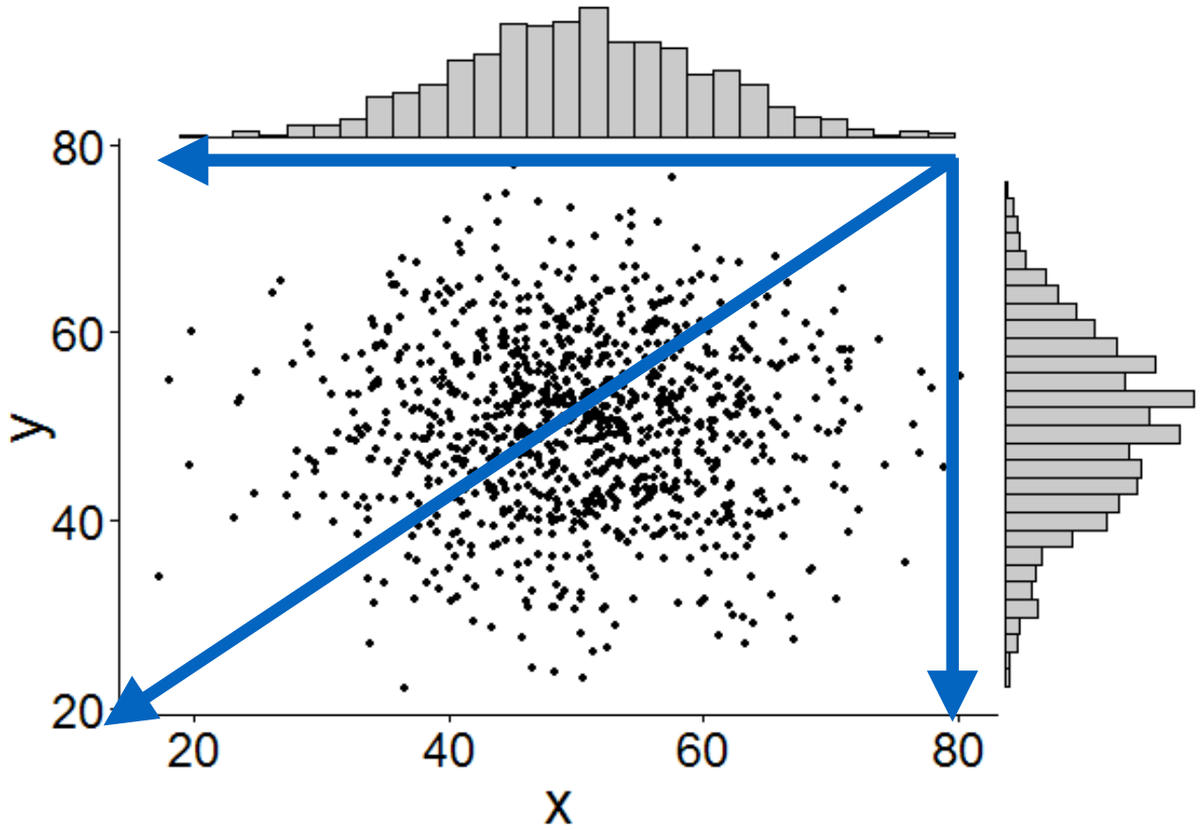


Yada, yada, yada...

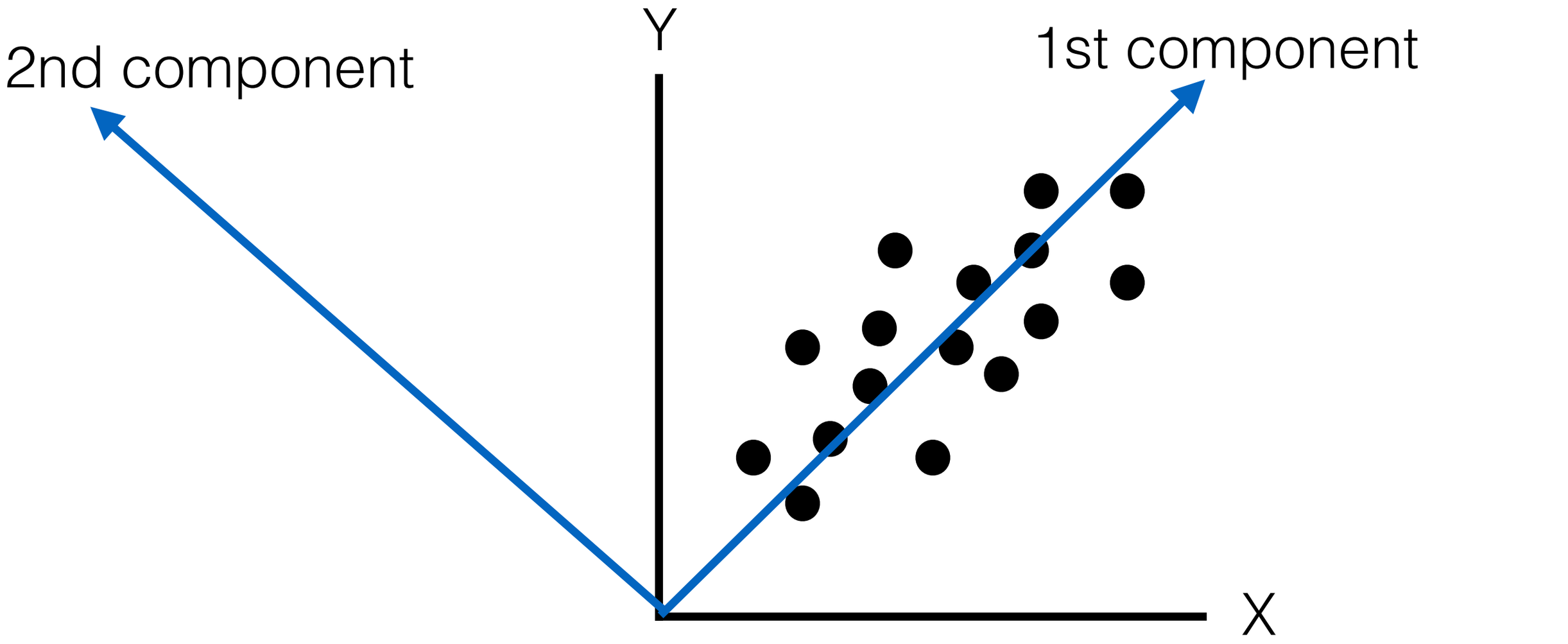
Google

PCA is about Finding the Direction of Maximal Variance

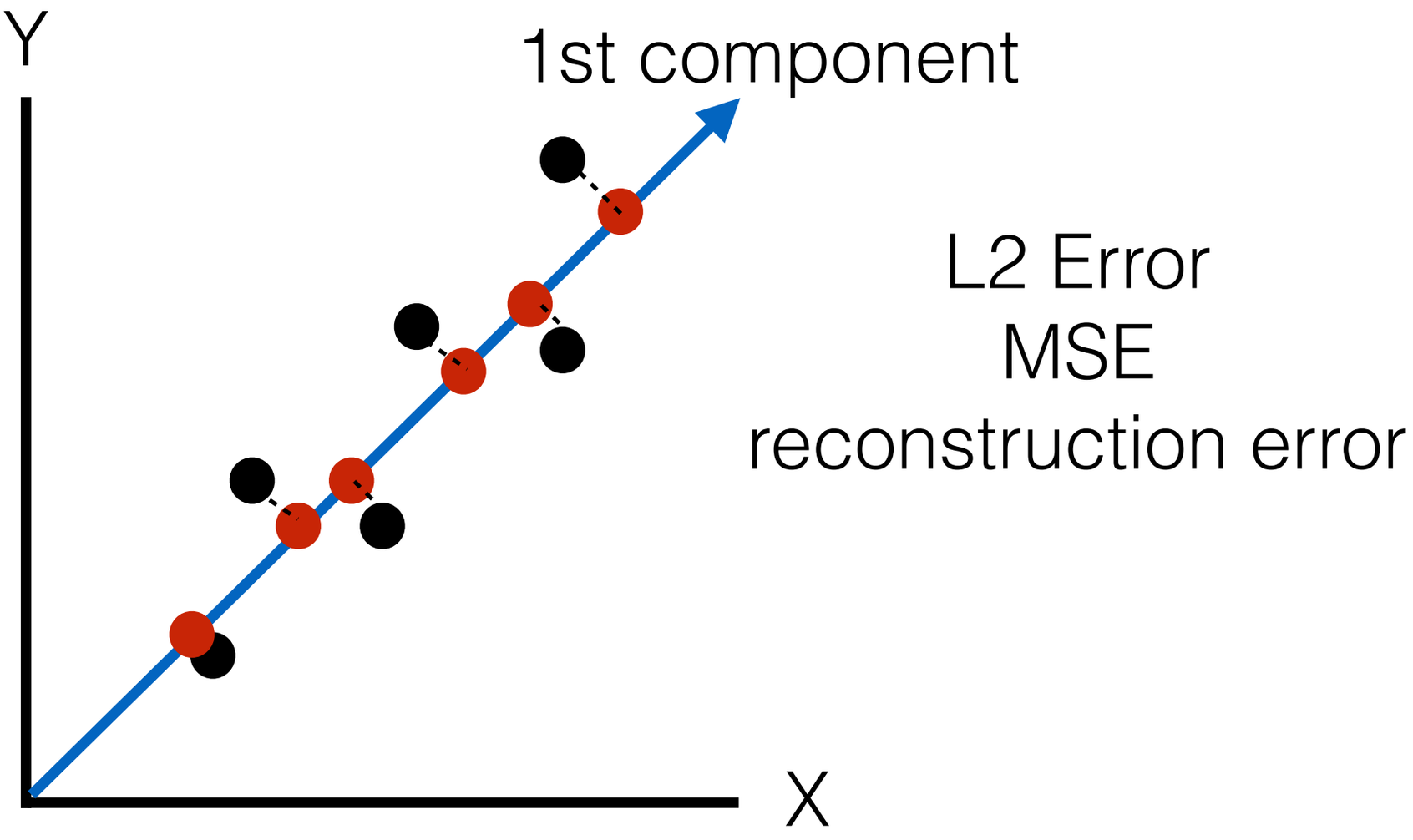




Global Constraint: Orthogonal Components



- “Best” reconstruction of the data (because not really doing anything)
- But also true for linear reconstruction of the data



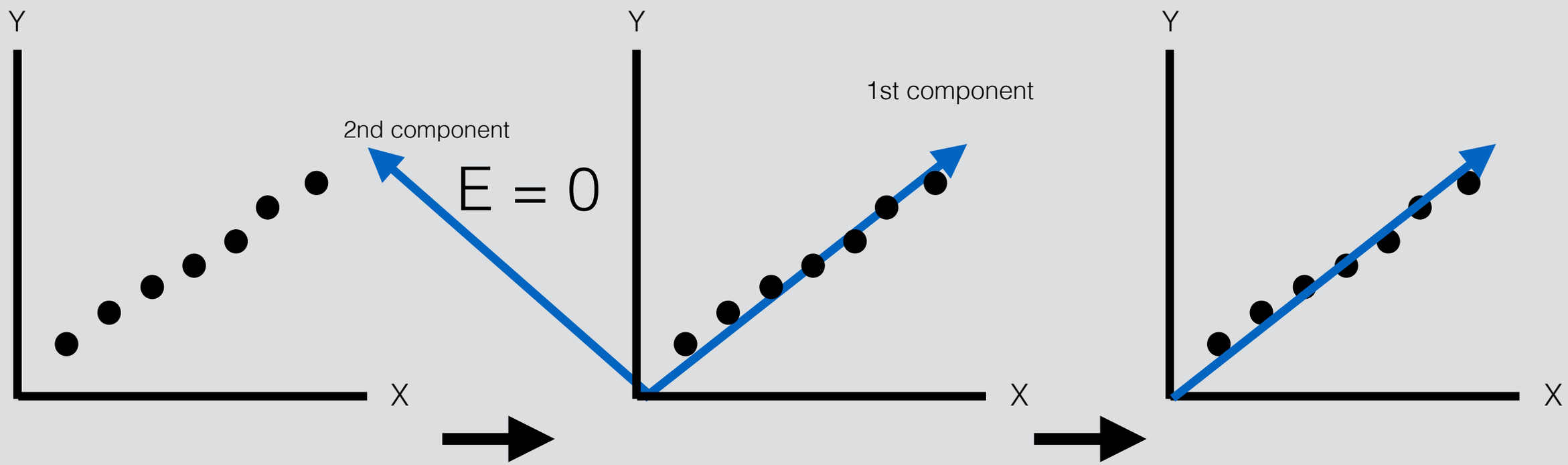
Component is a description of X & Y

Eigenvalues

- Every component has an associated eigenvalue
- Eigen- = “characteristic”
- Created when linear transformations are applied to a matrix
- Take away: the size of the eigenvalue is relative to how well the component maximizes variance

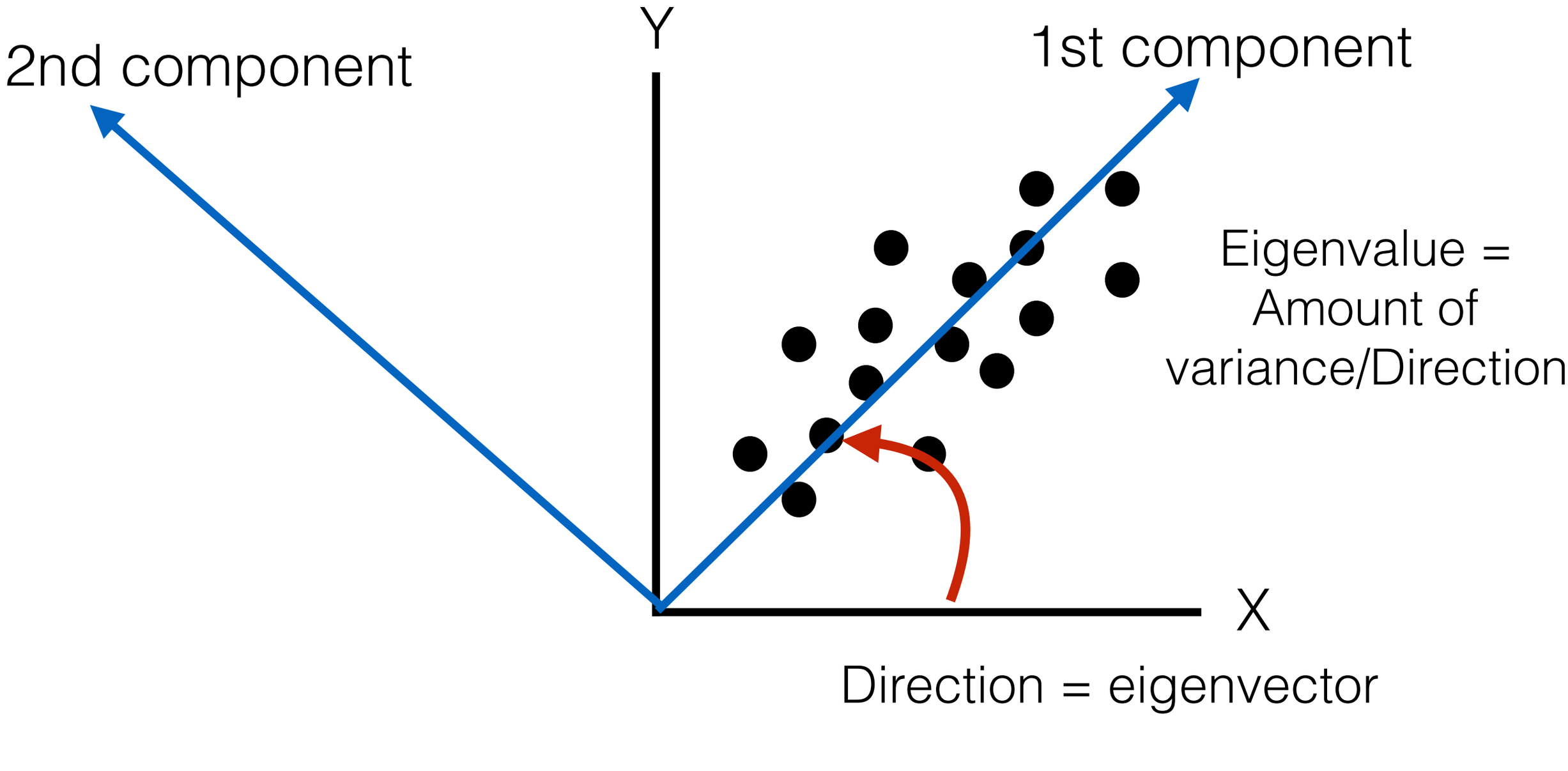
Feature Selection

- If a component has an eigenvalue of zero = non-informative (will not effect reconstruction error)
- Therefore, we can delete it = reduce features



Questions?

Orthogonal Components



Eigenvectors

pca$rotation

	PC1	PC2
V1	0.34	-1.6
V2	0.13	-0.07
V3	0.01	0.6
V4	0.02	1.5

Eigenvectors

Creating Composites

Because the eigenvectors represent the shift of each dimension, accounting for max variance, we can use these numbers to weight the construction of a composite.

Composite1(PC1) =
(V1 x E1) + (V2 x E2) + (V3 x E3) + (V4 x E4)

HOWEVER: You must make substantive sense of the component!

Gotchas

- Data needs to be scaled
- Often centered so that the direction goes through zero
- Outliers have an outsized impact on your results
- Continuous variables (or binary but be careful)
- Linear relationships between variables (sometimes impractical)
- Better with larger samples (no real way to test though)
- Components will be uncorrelated!