

# HUDK 4050: CORE METHODS IN EDM

# In the news

The New York Times

A Big Test of Police Body Cameras Defies Expectations



The inside story on emergencies

Irresponsible data? The risks of registering the Rohingya

**POLITICO**

When Big Data went to war — and lost

University of  
**Kent**

Brain stimulation can improve athletic performance

**AINOW**

AI Now 2017 Report

**NIBLETZ**

How to Use Technology to Prevent School Bullying

**EDUCATION WEEK**

Legends Of Learning Wins AWS  
EdStart Pitch Day New York

The Worldwide Educating for the Future  
Index - A Blueprint for Change

**The Economist**

How to Increase Confidence  
About Digital Learning in  
Schools

**EdTech**  
Focus On K-12™

White House, Tech Companies  
Pledge \$500 Million to Increase  
STEM Opportunities

A combination of grants and donations will support computer science education for underserved students.

**EdSurge**

Pick Your Battles: Edtech Leaders Share Strategies  
for Engaging in Political Discourse

Uber is charging drivers to work

Celebrate Halloween early  
by earning 33% more

Buy a week of accelerated earnings  
for \$115. Opt in below by Saturday,

**NewsCenter.io**

EdTech Startup O.school Brings In \$800,000

# Events

Event	Date	Time	Location	URL
The Provost's Conversation on Online Learning	November 8	4:00pm	Davis Auditorium, Schapiro CEPSR	<a href="https://columbia.us11.list-manage.com/track/click?u=8e1c6110b489f734b9985bec5&amp;id=d0b52b0629&amp;e=1cc15b53eb">https://columbia.us11.list-manage.com/track/click?u=8e1c6110b489f734b9985bec5&amp;id=d0b52b0629&amp;e=1cc15b53eb</a>
Intro to Data Science in R	November 1	11:00am	webinar	<a href="http://info.rstudio.com/g40jaCS0j203XN0YN000VRa">http://info.rstudio.com/g40jaCS0j203XN0YN000VRa</a>
Privacy Localism: A New Research Agenda	November 3	all day	NYU Law School	<a href="http://www.law.nyu.edu/centers/ili/events/privacy-localism">http://www.law.nyu.edu/centers/ili/events/privacy-localism</a>
The People's Disruption: Platform Co-Ops for Global Challenges	November 10/11	all day	The New School	<a href="https://platform.coop/2017">https://platform.coop/2017</a>
TC Microsoft Workshop	November 2	7:00pm	Grace Dodge 179	<a href="http://bit.ly/2yMGAYC">http://bit.ly/2yMGAYC</a>
Bridging Industry and Academia to Tackle Responsible Research and Privacy Practices	November 2/3	all day	Facebook	<a href="https://fpf.org/research-privacy-practices/">https://fpf.org/research-privacy-practices/</a>
KPMG Technology Career Forum	November 14	1:00pm	Online	<a href="https://app.brazenconnect.com/events/X1r8O?utm_medium=Marketing&amp;utm_source=Data+Science+Association#!eventLanding;eventCode=X1r8O">https://app.brazenconnect.com/events/X1r8O?utm_medium=Marketing&amp;utm_source=Data+Science+Association#!eventLanding;eventCode=X1r8O</a>
TCLA Happy Hour	November 1	7:00pm	Tea Magic	

# Opportunities

Woodrow Wilson Foundation/MIT

Senior Assessment Designer

<https://woodrowacademy.org/about/job-opportunities/>

2017 ASSISTments Data Mining Competition

<https://sites.google.com/view/assistmentsdatamining/data-mining-competition-2017>

# Anonymous Check In

**<http://bit.ly/2zjl5hT>**

# Assessment

- Meet with me once
- Github contains all assignments
- Zotero contains notes on readings & assignments
- Ask question on Stack Overflow

# 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



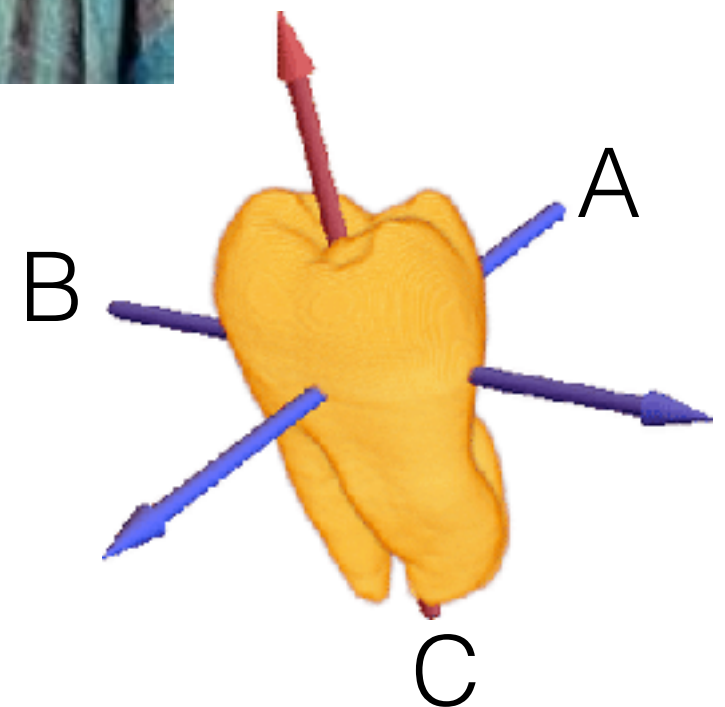
ID	Var1	Var2	Var3
A			
C			

ID	Var1	Var2	Var3
B			
D			



# History

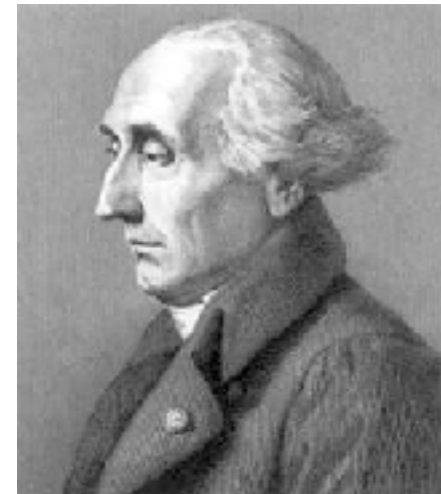
- Part of a set of issues called “Eigen Problems”
- Arose as a subset of phenomena related to differential equations (Your old buddy Euler, c.1750)
- Principal Axes



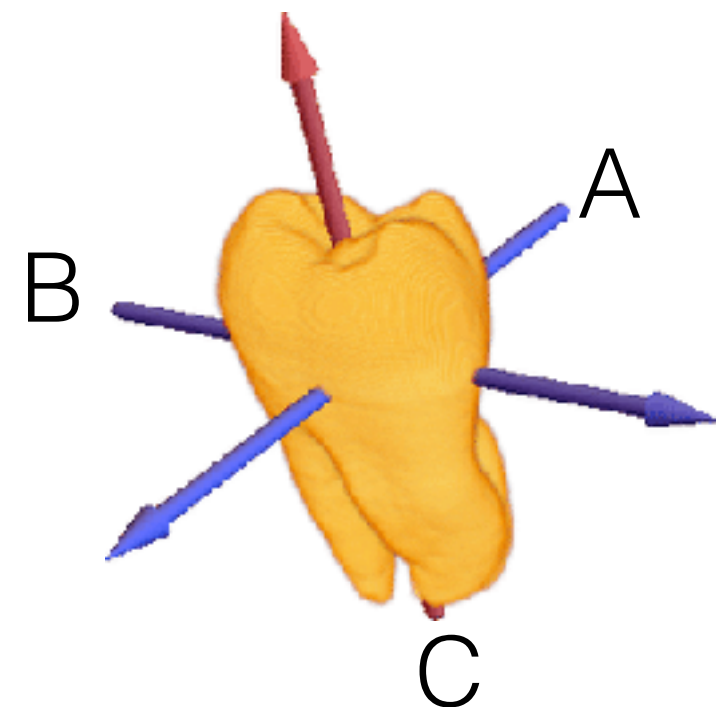
# History

- Describe inertia as a matrix of measurements from the center of an object as it moves
- Principal axes = the lines through which you can describe the object, while maximizing the amount of variation maintained

(Joseph-Louis Lagrange)



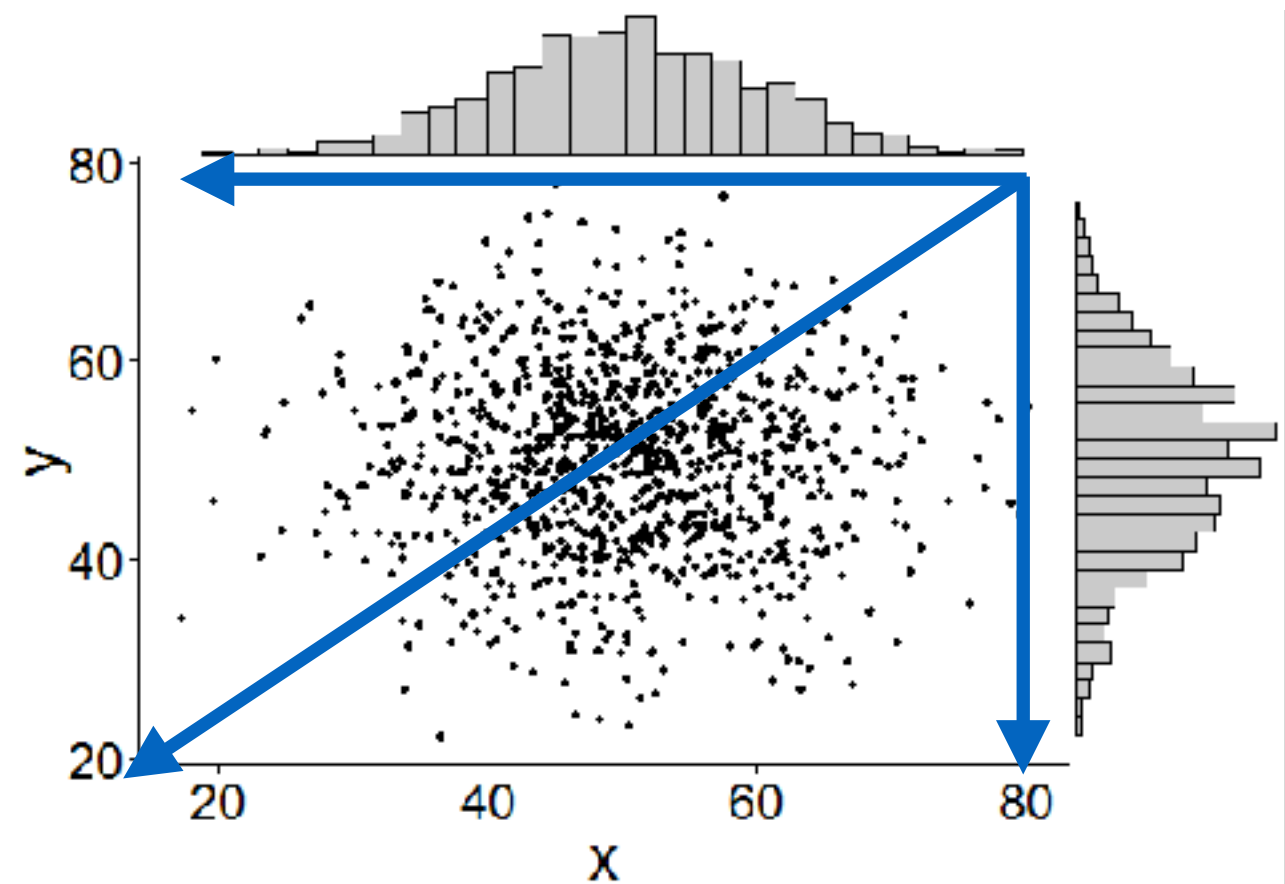
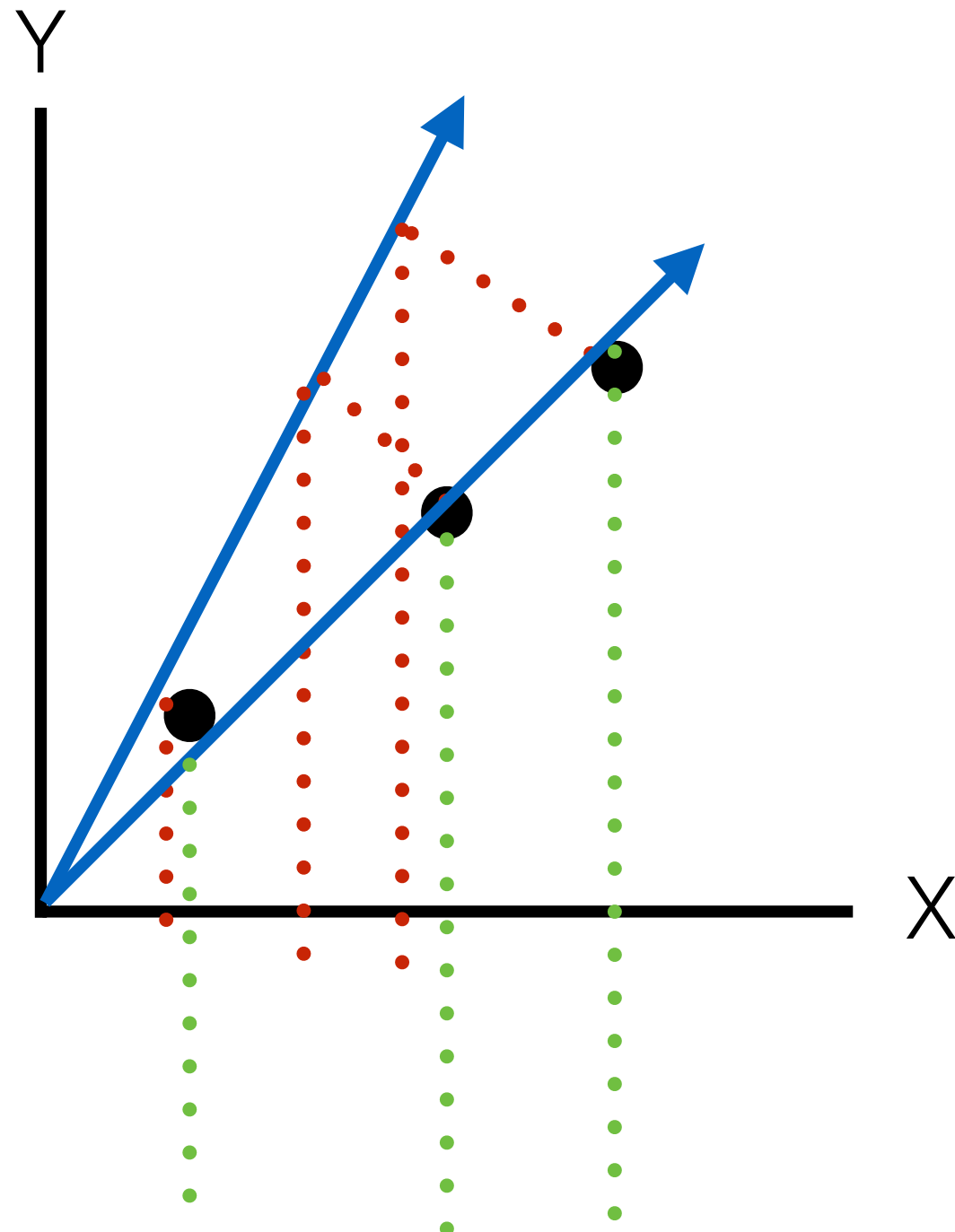
$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$



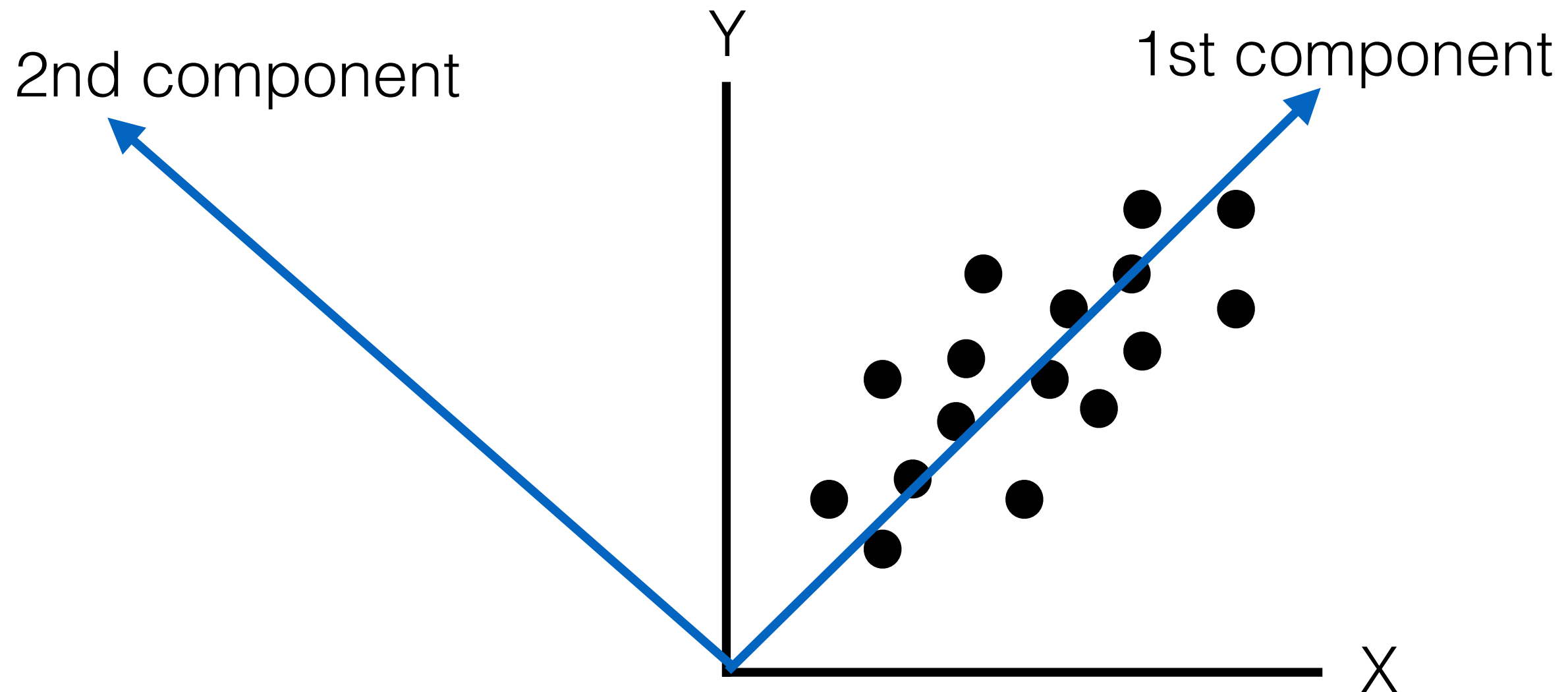
Yada, yada, yada...

Google

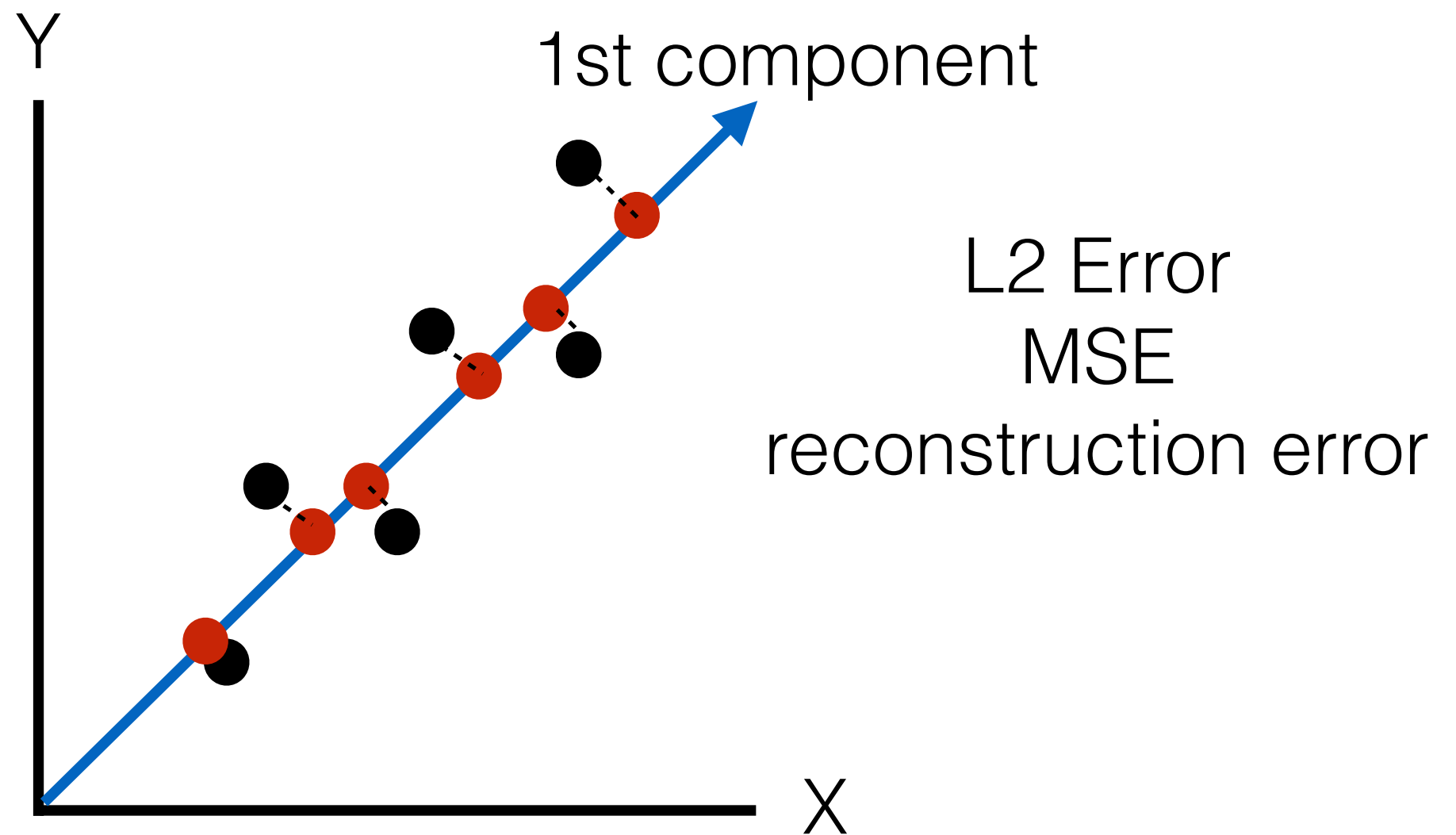
# PCA is about Finding the Direction of Maximal Variance



# Global Constraint: Orthogonal Components



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



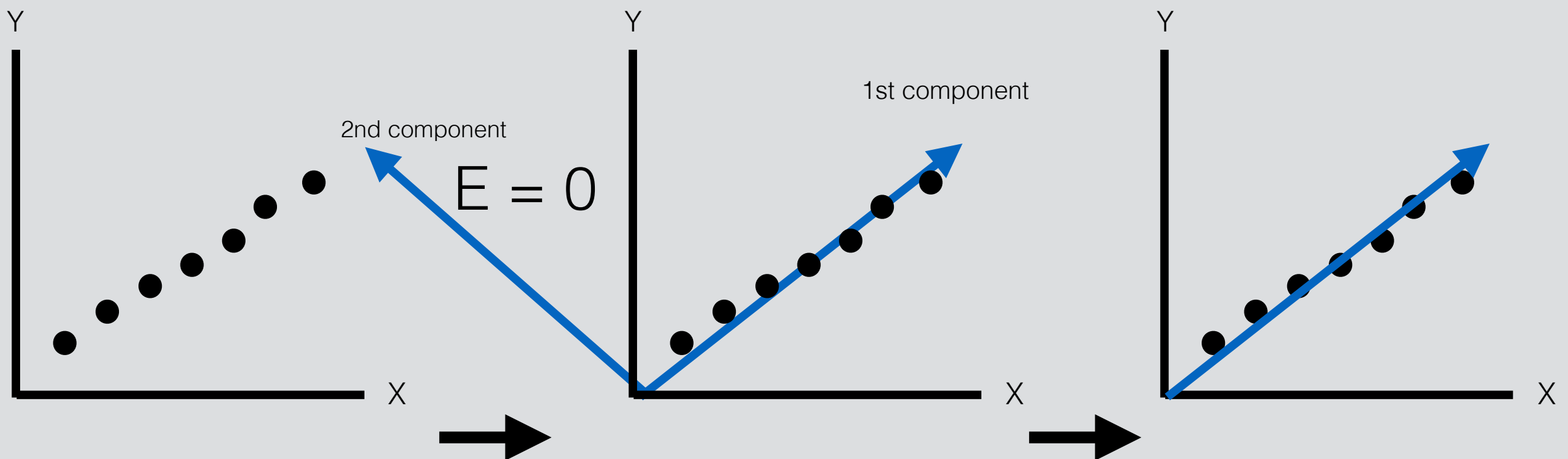
Component is a description of X & Y

# 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

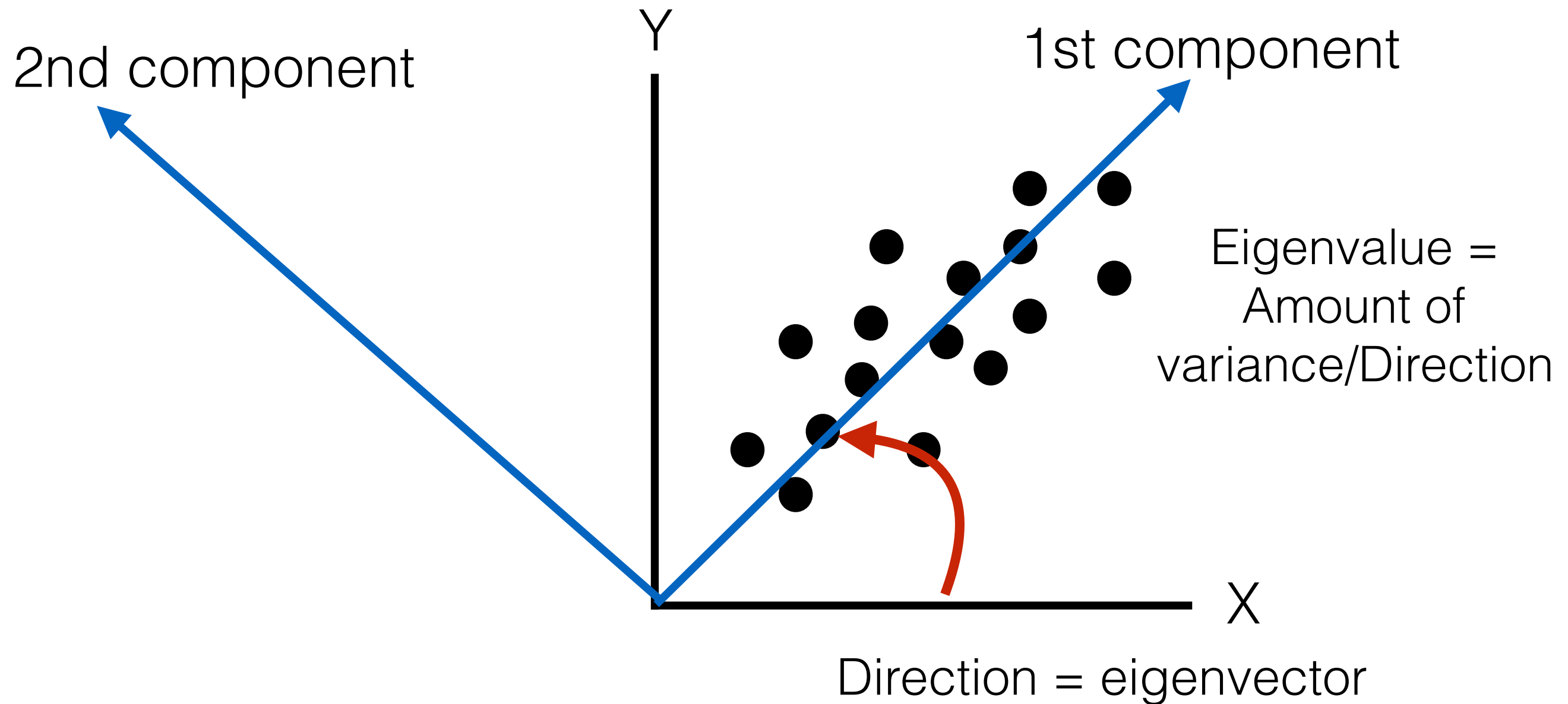
- If a component has an eigenvalue of zero = non-informative (will not effect reconstruction error)
- Therefore, we can delete it = reduce features





Questions?

# Orthogonal Components



# Eigenvectors

pca\$rotation

Eigenvectors

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

# 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.

$$\text{Composite1(PC1)} = (V1 \times E1) + (V2 \times E2) + (V3 \times E3) + (V4 \times E4)$$

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!