ILLINOIS

Gies College of Business

"Curse of Dimensionality" Tackling Big Data Challenges: Data Transformation and Dimension Reduction

Sridhar Seshadri

Content

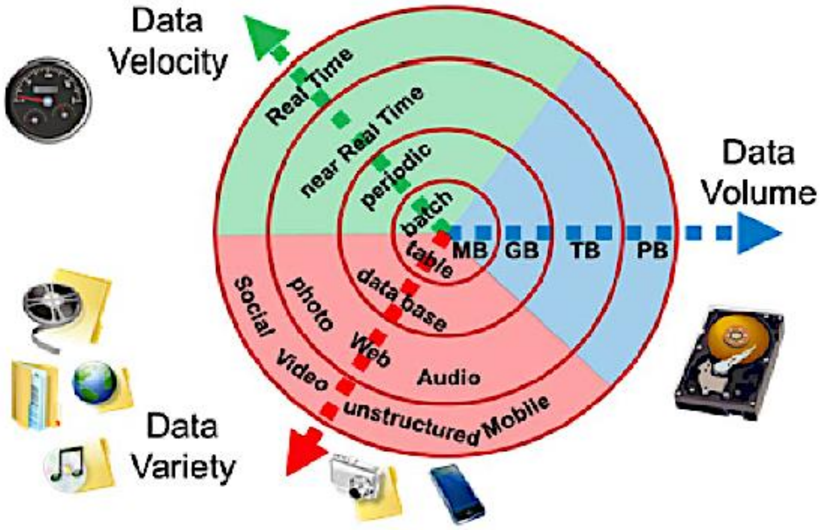
Data types-organization-modality

Increasing complexity, added benefit

Curse of dimensionality

Exploring data with a purpose

Data Comes in Many Forms, Shapes and Sizes!



Velocity: Speed

Volume: Size

Variety: Type

Veracity: Data quality and data value

Data Types

Numeric – a *quantifiable* number

- Type – integer (e.g. age), floating (e.g. price), time, date, …
- Stats – min/max/median/mean/…
- Units – (C/F), (KG/Lb), (Meter/Feet), (Sec/Min/Hrs),
- Distributions – exponential/uniform/…

Data Types

Ordinal – *not quantifiable* but *ordered*

Data Types

Symbolic – *neither quantifiable, nor ordered*

E.g. color = red/green/blue/…

E.g. state/country/region/…

E.g. weather = rainy/cloudy/windy/…

Data Organization

MULTIVARIATE
(rows (**examples**)
of columns (**features**))

	feaure-1	feature-2	feature-3	feature-4	feature-5
example-1					
example-2					
example-3					
example-4					
example-5					
example-6					
example-7					

**Low Dimensional
and Dense**

Data Organization

BASKET
(**sets** of things)
market basket, keyword list

	item-1	item-2	Item-3	Item-4 …	item-10M
example-1	**1**		**1**		
example-2				**1**	**1**
example-3					
example-4	**1**	**1**			
example-5				**1**	
example-6		**1**		**1**	
example-7			**1**		**1**

High Dimensional and Sparse

Data Organization

BAG
(weighted sets of things)
Bag-of-Words,
Bag-of-Visual Words

	item-1	item-2	Item-3	Item-4	item-100M
example-1	**10**		**5**		
example-2				**13**	**11**
example-3					
example-4	**18**	**12**			
example-5				**21**	
example-6		**4**		**51**	
example-7			**1**		**32**

High Dimensional and Sparse

Data Structure

STRUCTURED – fixed columns in a table

- Multivariate data
- Mix of numeric and symbolic features

Data Structure

UNSTRUCTURED -- Arbitrary size data points

- **SEQUENCE** : biological, speech, …
- **SERIES** : stock market, etc.
- **TEXT** : pages, queries, tweets, ads, blogs, news, …
- **IMAGE** : regular, medical, remote sensing,…
- **VIDEO** : regular, movies, security, surveillance,…

Mixture of Feature Types: Is More Better?

Numeric + *Categorical features* → Distance function???

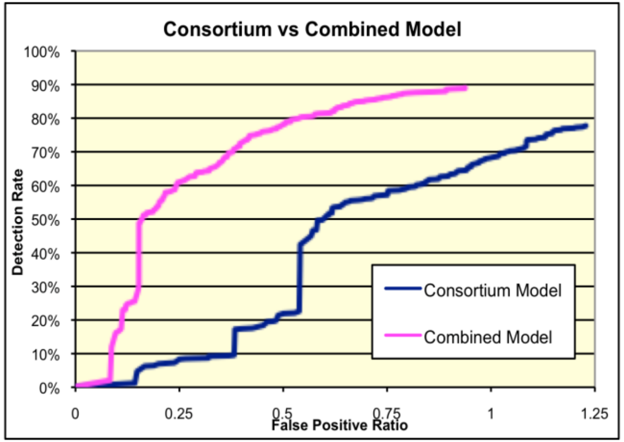
Numeric features → Distance function is well defined.

Categorical features → Distance function is not defined

Numeric + Categorical + *Text features* → Modeling???

Structured data

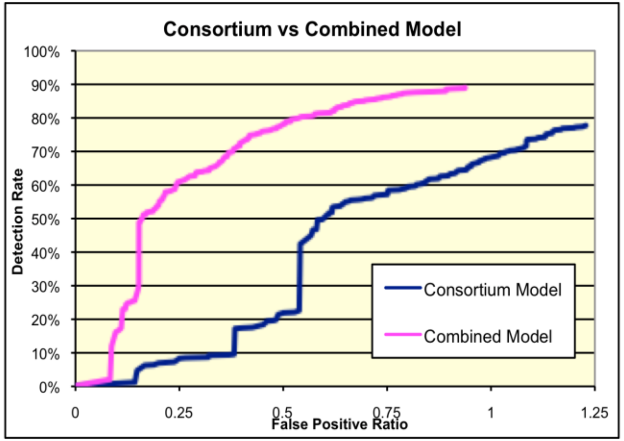
Unstructured data



Mixture of Feature Types: Is More Better?

Collections Case Study

- Application data
- Payment History
- Call Center Notes



Curse of Dimensionality

As variables get added data space becomes very sparse

Establishing causality and prediction accuracy become difficult

Curse of Dimensionality

Examples:

Construct an investment portfolio out of 200 stocks

Determine rankings (university, movie, candidate for election, __) based on available information on different measures

Decide who to serve an ad based on different attributes (wealth, income, cars, preferences, sites visited, marital status, health, age, ….)

Potential remedies

Detect small variation in data for some features

Combine features because they correlate highly

Cluster so that features nearly same and eliminate

Domain Knowledge

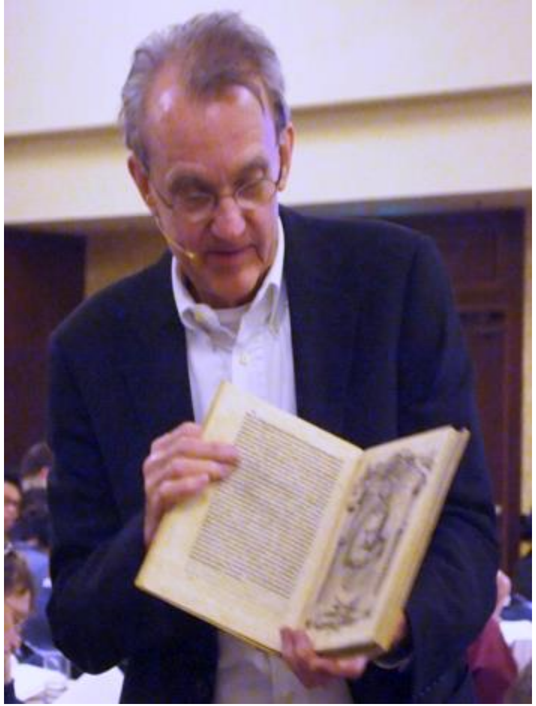
Others: Debate whether feature detection and model development be done together. Domain experts might argue one way Deep Learning proponents the other.

Data Exploration With a Purpose

A Picture is Worth a Million Numbers!

*"Often the most effective way to **describe**, **explore**, and **summarize** a set of **numbers** – even a very large set – is to **look at pictures** of those numbers"*

– Edward R. Tufte



Different Ways of Visualizing Data

- Histograms / Distributions
  - Density Estimation
- Covariance Analysis
- Scatter Plots

Different Ways of Visualizing Data

Principal Components Analysis

(Fisher) Discriminant Analysis*

Multi-Dimensional Scaling

Self-Organizing Maps

Manifold Learning

*Discriminant analysis models the difference between the classes of data. PCA does not take into account difference in classes.

Iris Data – the “Original” Dataset…

How does the data look in various 2-D views?

FEATURES:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

4

50

50

50

Source: R.A. Fisher

Iris Setosa

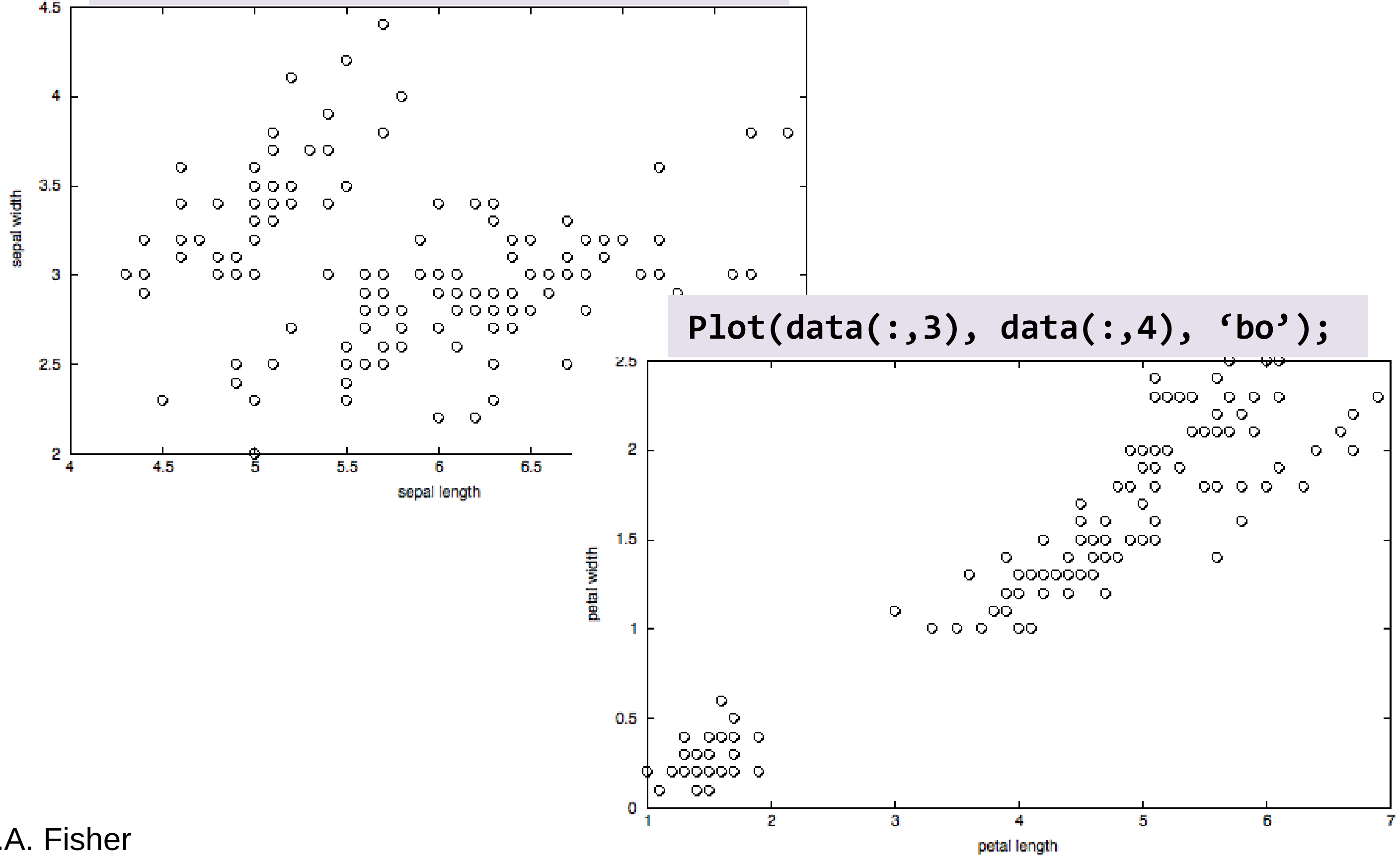
Iris Versicolour

Iris Virginica

Iris – Scatter Plots

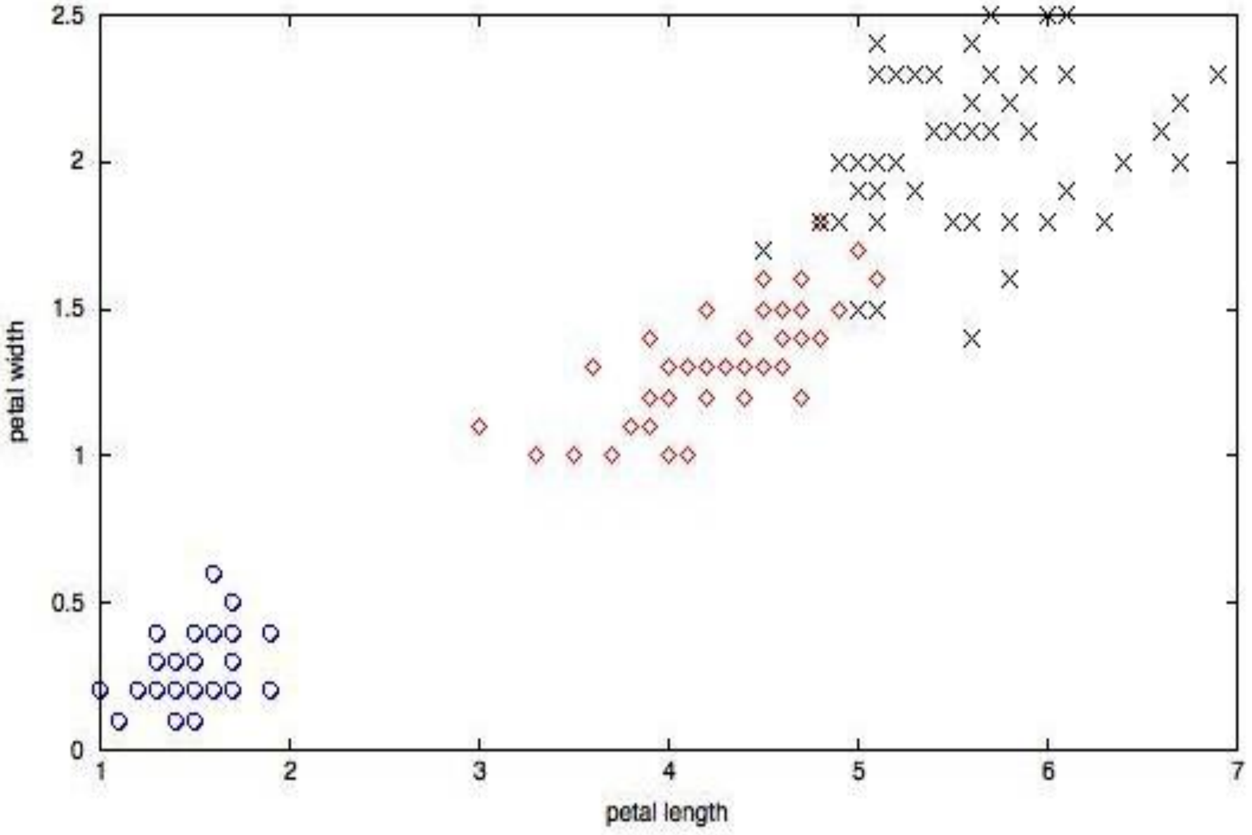
`Plot(data(:,1), data(:,2), ‘bo’);`

`Plot(data(:,3), data(:,4), ‘bo’);`



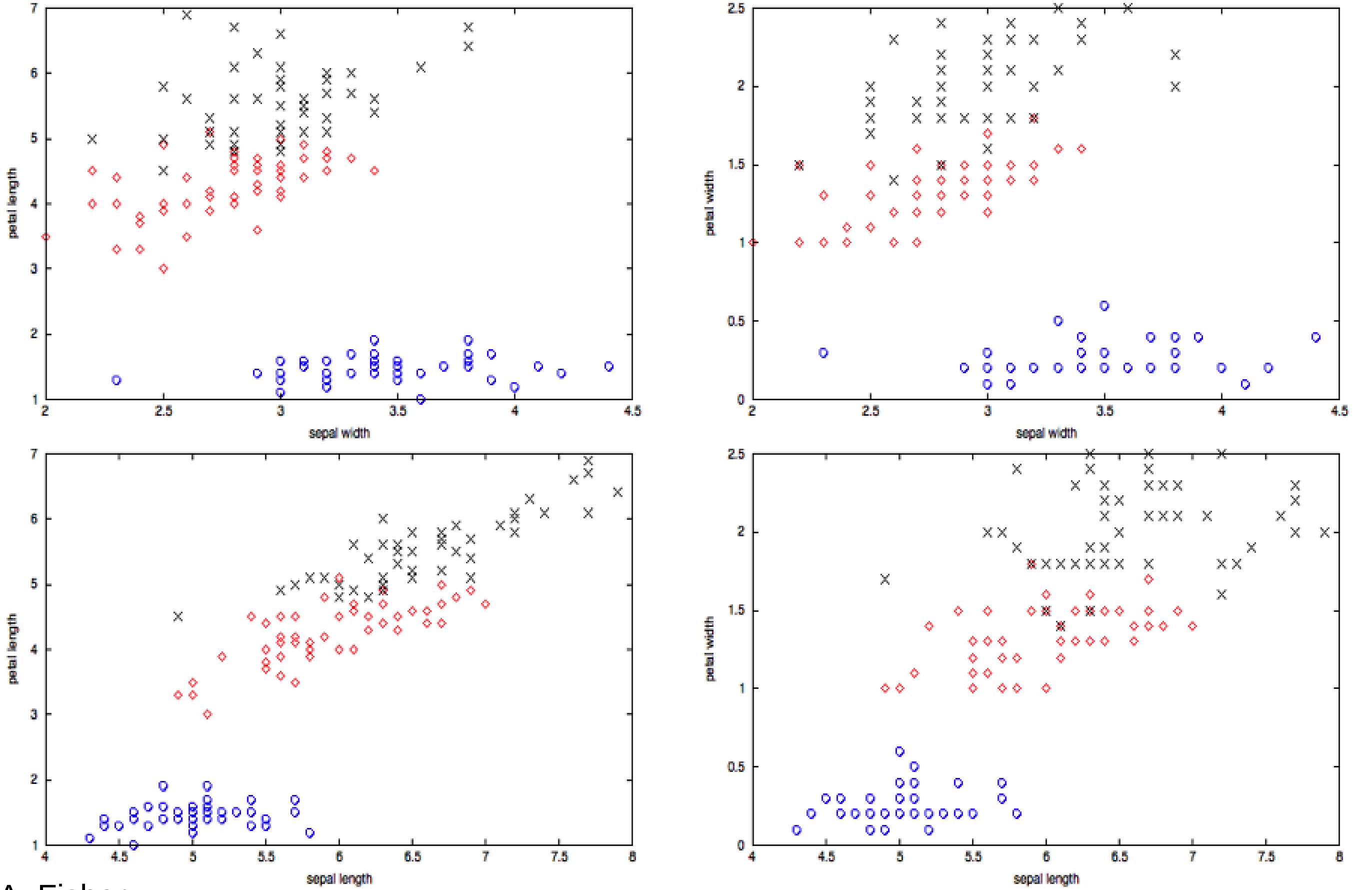
Source: R.A. Fisher

Iris – Scatter Plots + Class Labels



Source: R.A. Fisher

Iris – More Scatter Plots



Source: R.A. Fisher

Scatter Plots

Simple and Powerful Visualization

Limited to 2 or 3 dimensions at a time

Limited to dimensions from among given

What if many dimensions (e.g. 100+)

- We need to see O(10,000) pairs of features

Solution? **Projections**

Different Ways of Visualizing Data

Histograms / Distributions

    Density Estimation (later)

Covariance Analysis

Scatter Plots

Principal Components Analysis

Discriminant Analysis

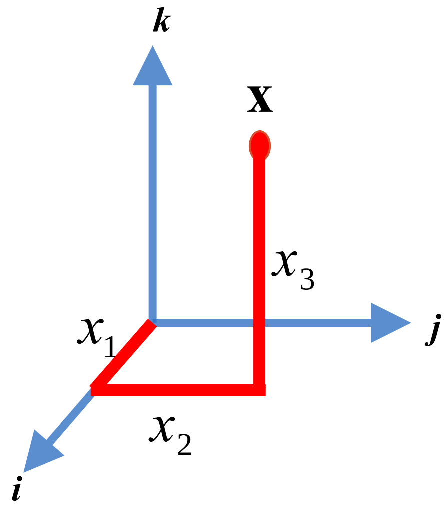
Multi-Dimensional Scaling

Self-Organizing Maps

Manifold Learning

Orthogonal Bases - 101

UNORDERED – all bases equally important!

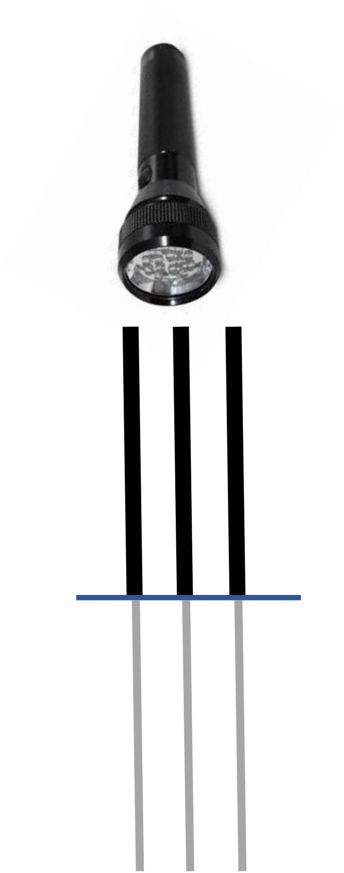


$$\mathbf{x} = x_1\boldsymbol{i} + x_2\boldsymbol{j} + x_3\boldsymbol{k}$$
$$= \left(\boldsymbol{i}^T\mathbf{x}\right)\boldsymbol{i} + \left(\boldsymbol{j}^T\mathbf{x}\right)\boldsymbol{j} + \left(\boldsymbol{k}^T\mathbf{x}\right)\boldsymbol{k}$$

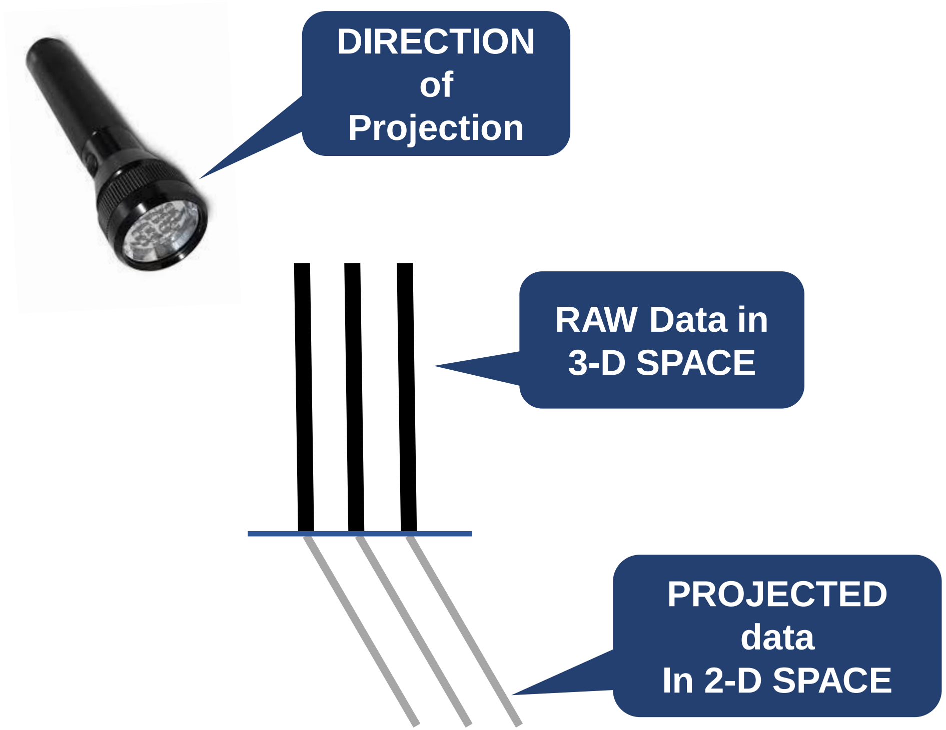
ORDERED – bases have decreasing importance

$$\mathbf{3528 = 3 \times 10^3 + 5 \times 10^2 + 2 \times 10^1 + 8 \times 10^0}$$

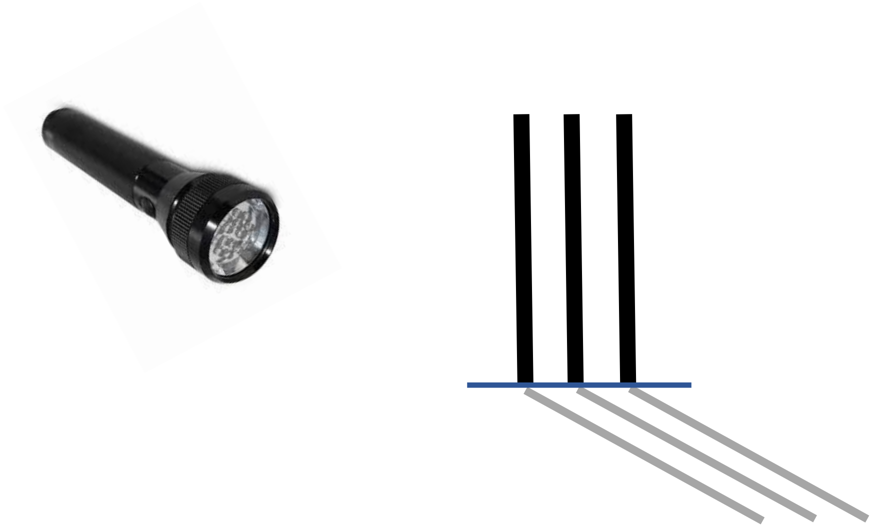
Projection – The Basic Idea



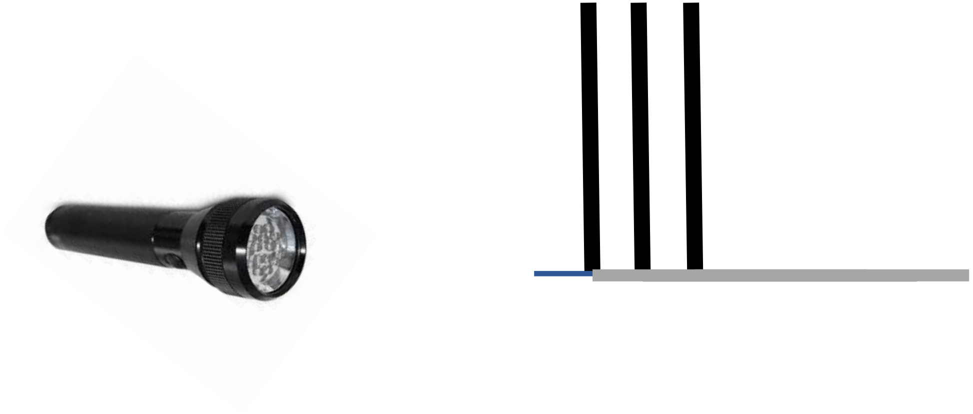
Projection – The Basic Idea



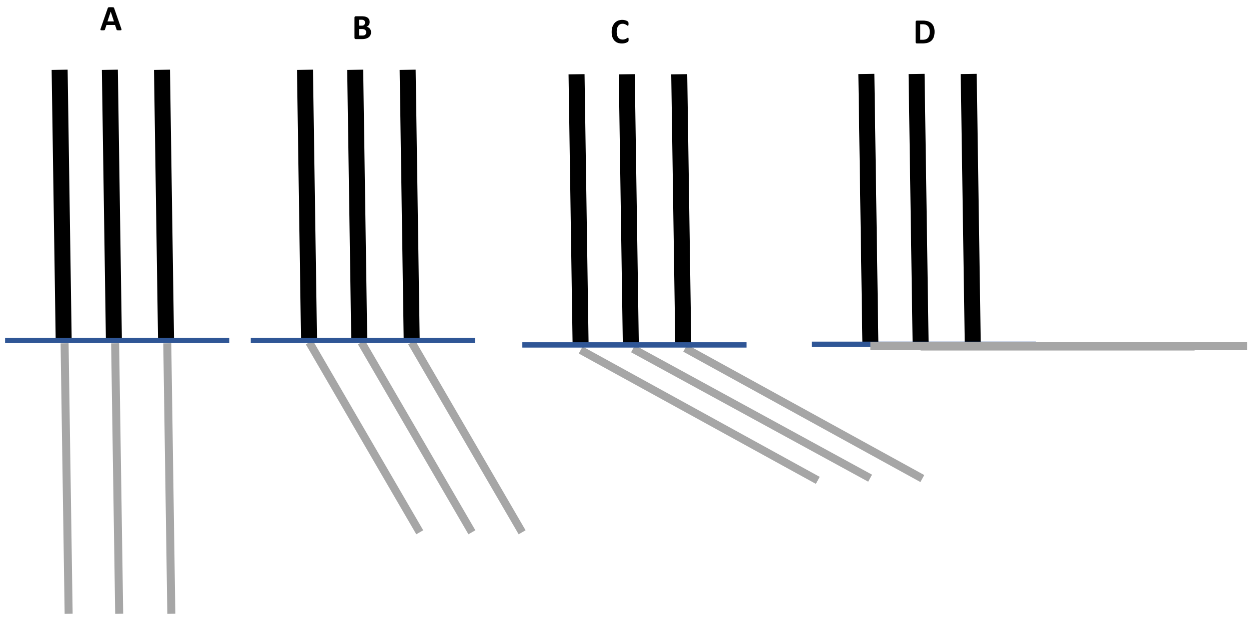
Projection – The Basic Idea



Projection – The Basic Idea



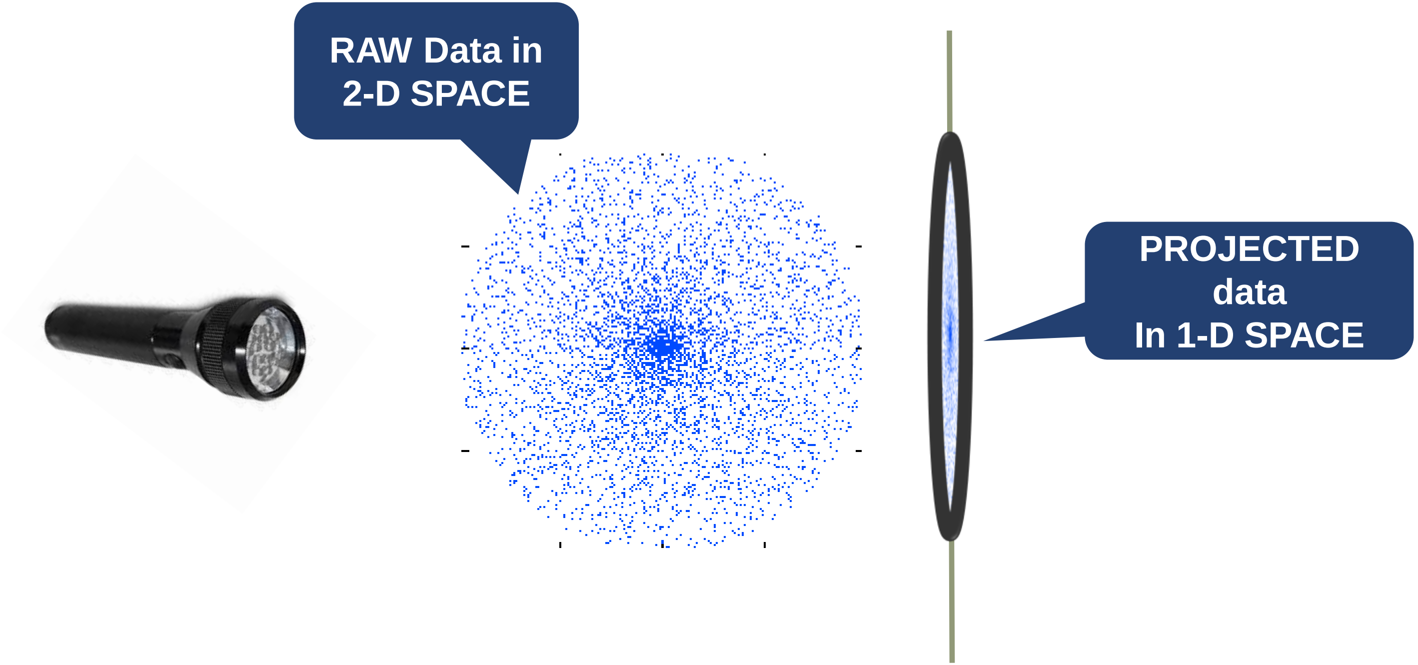
Which is the "Best" Projection?



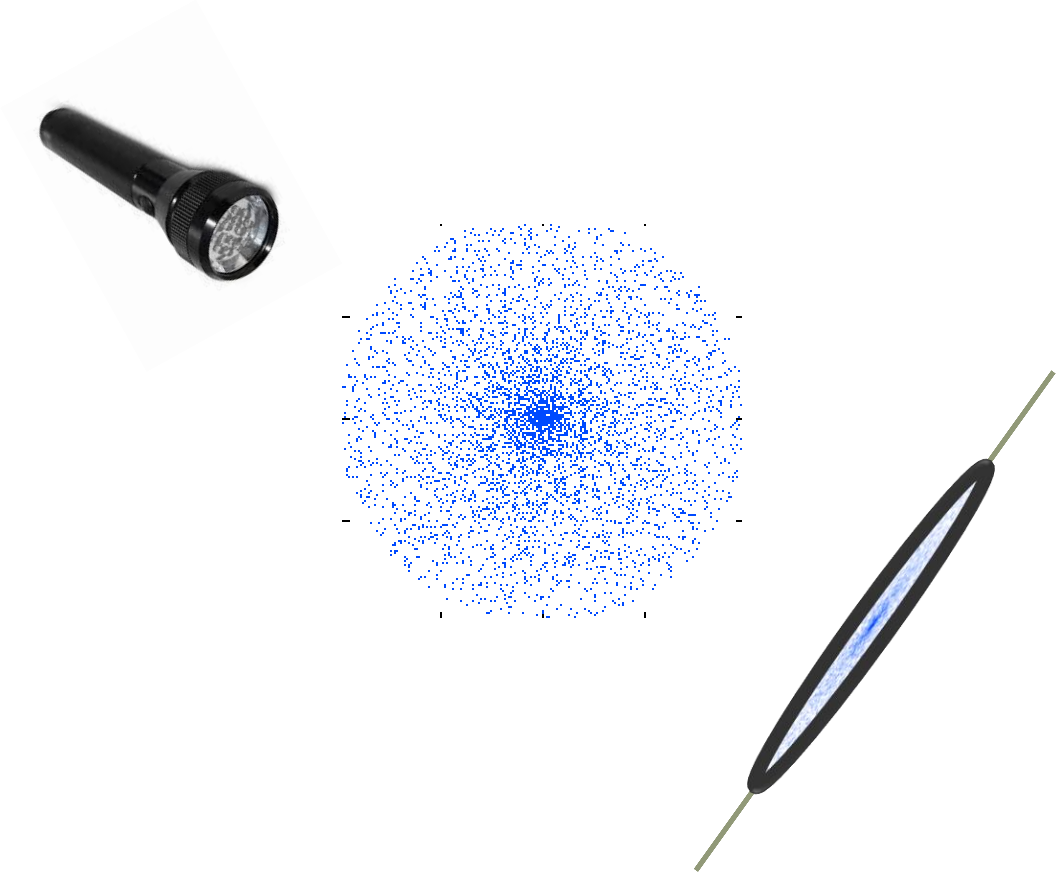
How do we "**measure**" the "**goodness**" of a projection?

The one that "**preserves**" the maximum "**information**"

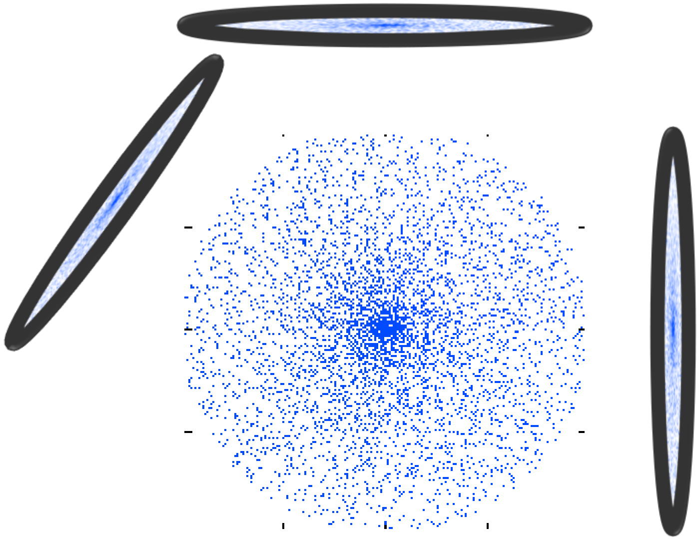
Projection: Spherical Data Cloud



Which “Measure” Do We Use Here?



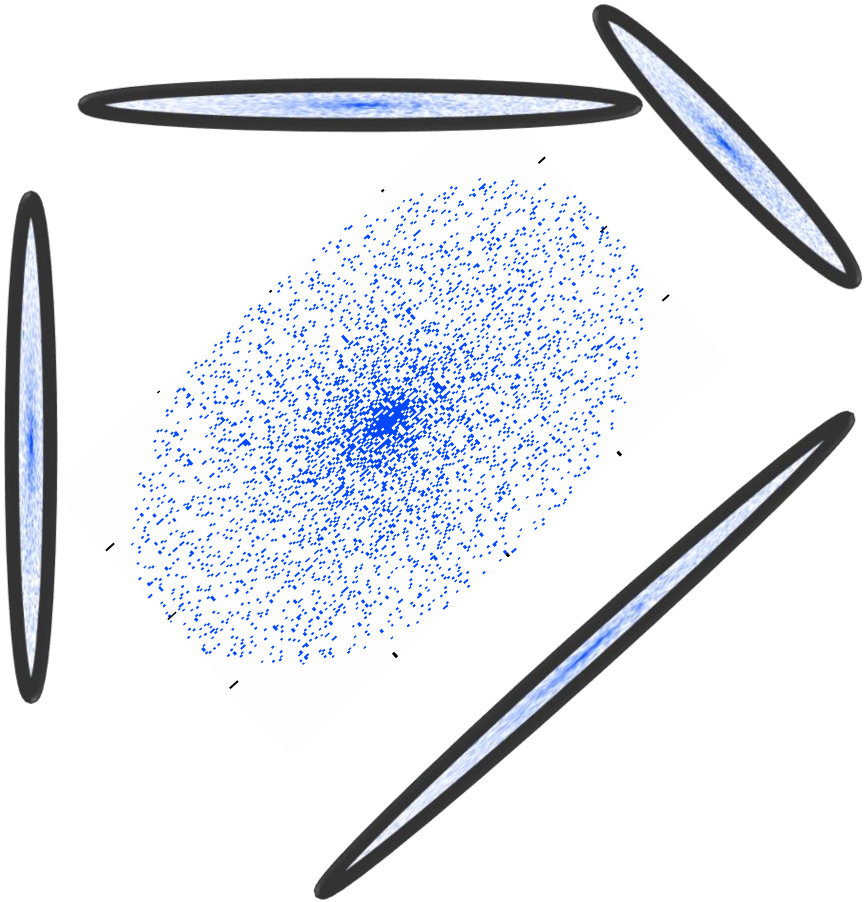
Which is the “Best” projection?



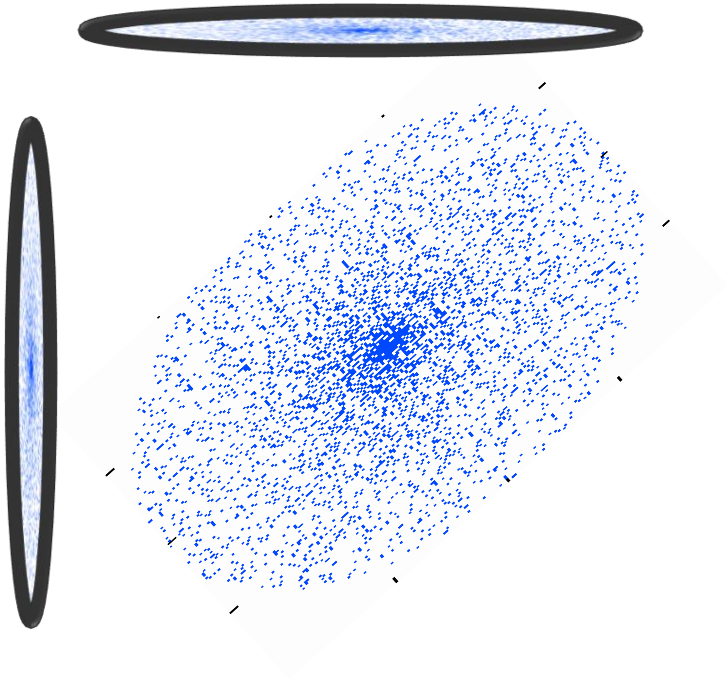
How do we “**measure**” the “**goodness**” of a projection?

The one that “**preserves**” the maximum “**information**”

Which is the Best/Worst Projection?

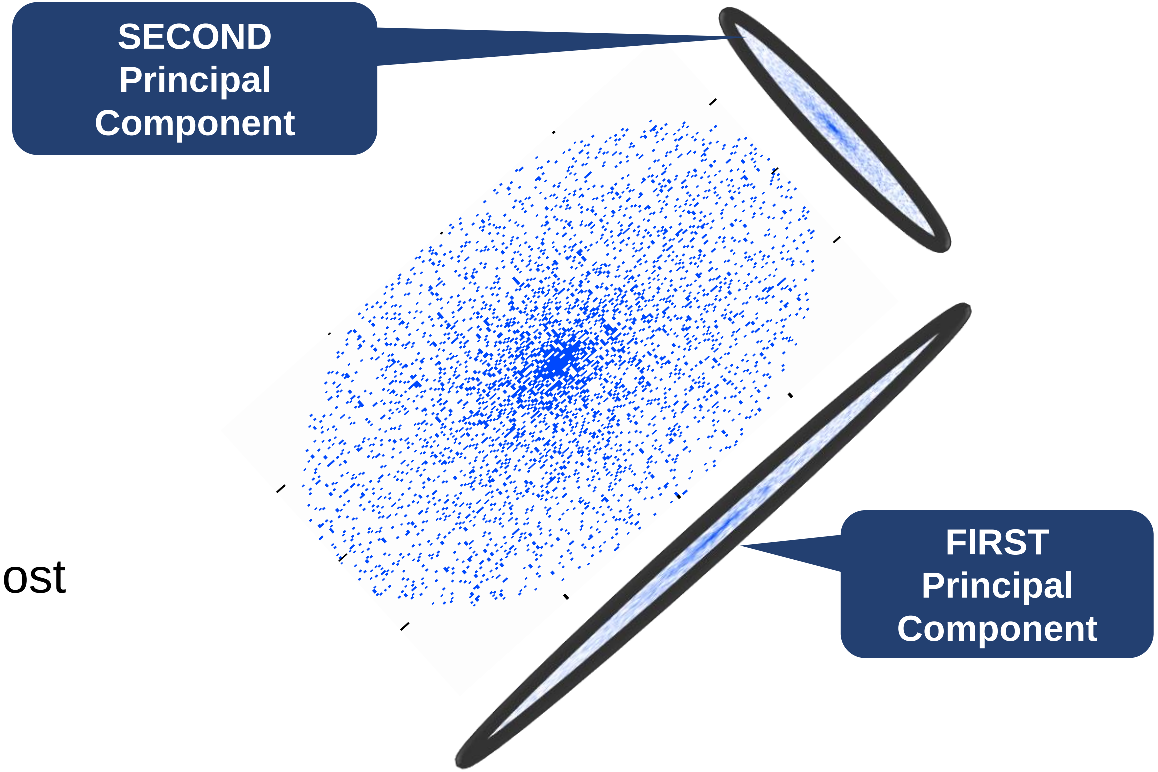


Complete and Orthogonal



Complete Set of
Orthogonal Projections
Capture All Information

Complete Projection



Capture the most
Then the next
Then the next
Until complete

Rattle example

PCA on University Data

Download instructions given in University_Data_Downloader-1.2_updated.pdf

Variable	Description
Univ	University name
SAT	Average SAT scores of new freshmen
Top10	% new freshmen in top 10% of highschool class
Accept	% of applicants accepted
SFRatio	Student-to-faculty ratio
Expenses	Estimated annual expenses
GradRate	Graduation rate (%)

[Download from http://users.stat.umn.edu/~kb/classes/8401/files/data/JWData5.txt](http://users.stat.umn.edu/~kb/classes/8401/files/data/JWData5.txt)

Data Viewer

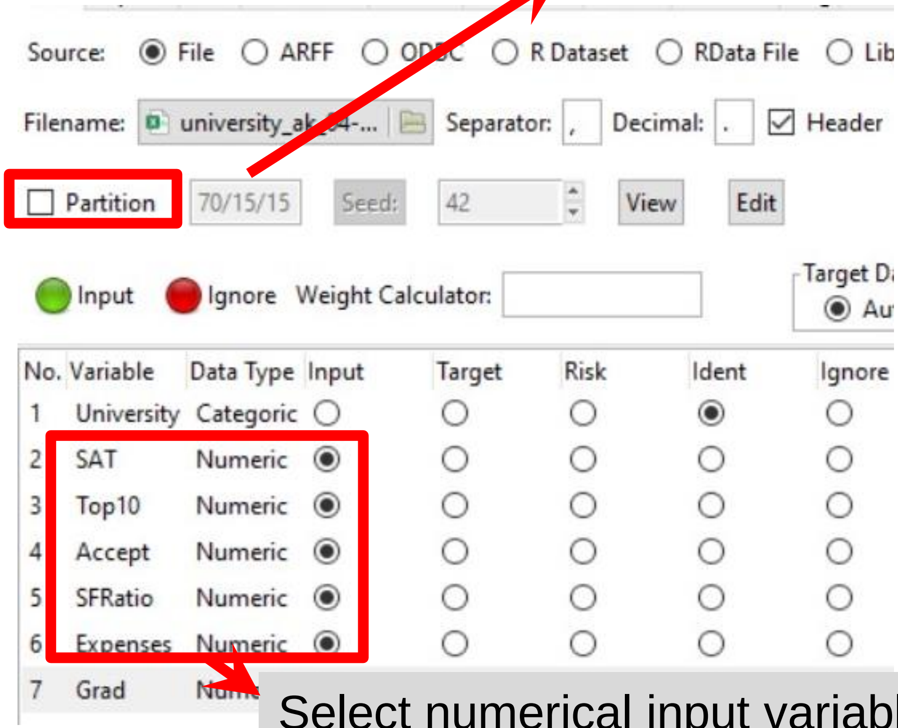
File

	University	SAT	Top10	Accept	SFRatio	Expenses	Grad
1	Harvard	14.00	91	14	11	39.525	97
2	Princeton	13.75	91	14	8	30.220	95
3	Yale	13.75	95	19	11	43.514	96
4	Stanford	13.60	90	20	12	36.450	93
5	MIT	13.80	94	30	10	34.870	91
6	Duke	13.15	90	30	12	31.585	95
7	Cal_Tech	14.15	100	25	6	63.575	81

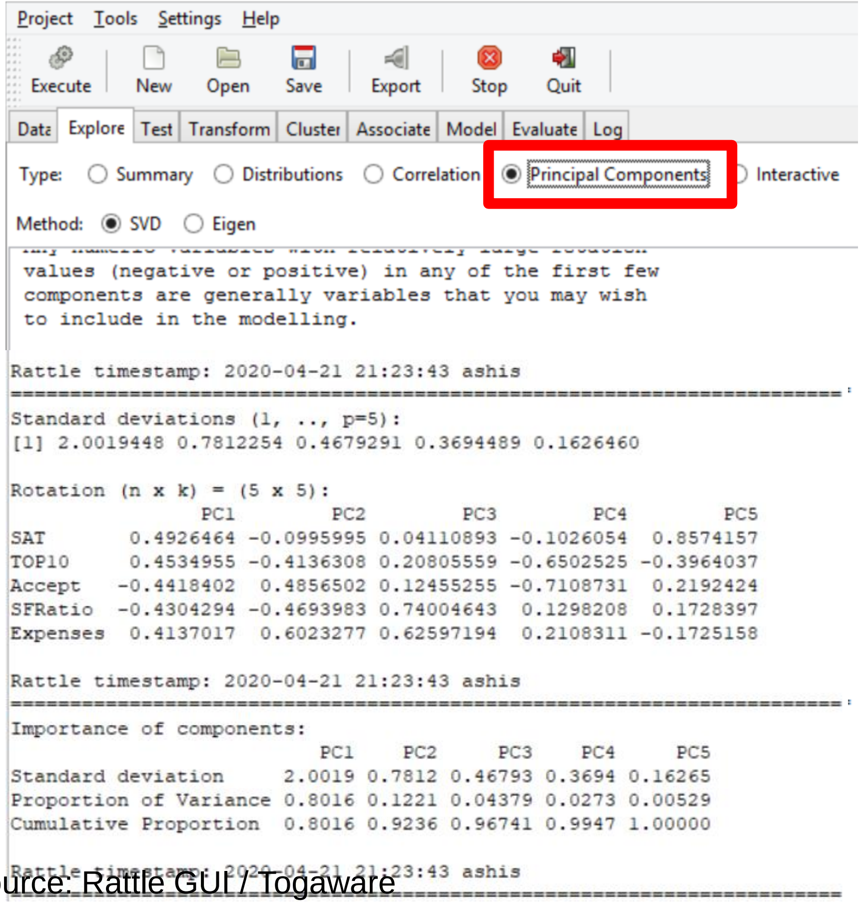
Source - US News & World Report, Sept 18, 1995, p. 126 (in Johnson & Wichern)

PCA in Rattle

Keep whole data



Select numerical input variable
for dimension reduction



Source: Rattle GUI / Togaware

PCA in R – Value of PCA in a Regression (Predicting Graduation Rate) *pca_2020.R file*

university <- read.csv("Universities.csv") # read data in R

pca_univ <- prcomp(university[,c(3:7)], center = TRUE, scale. = TRUE) # PCA model

Feature Scaling and Centering done to prevent one feature from dominating another. Feature Transformation/weights done to prevent skewed data affecting outcomes.

PCs <- pca_univ$x # extracting the components

university_pca <- cbind(university,PCs) # saving the PCs in the dataset

summary(pca_univ) # get PCA summary

```
> summary(pca_univ)
Importance of components:
                          PC1    PC2     PC3    PC4     PC5
Standard deviation     2.0019 0.7812 0.46793 0.3694 0.16265
Proportion of Variance 0.8016 0.1221 0.04379 0.0273 0.00529
Cumulative Proportion  0.8016 0.9236 0.96741 0.9947 1.00000
```

Source: Output in RStudio

PCA in R – Value of PCA in a Regression (Predicting Graduation Rate)

 # Model with all five variables (without PCA)

```
model <- lm(GradRate~ SAT+Top10+Accept+SFRatio+Expenses, data = university_pca)

summary(model)

# Multiple R-squared:  0.7104
```

 # Adjusted R-squared: 0.6342

PCA in R – Value of PCA in a Regression (Predicting Graduation Rate)

 # Model with PCA - 1st component

model_pca1 <-lm(GradRate~ PC1, data = university_pca)

summary(model_pca1)

Multiple R-squared: 0.5414

 # Adjusted R-squared: 0.5215

PCA in R – Value of PCA in a Regression (Predicting Graduation Rate)

 # Model with PCA - 2 components

model_pca2 <-lm(GradRate~ PC1+PC2, data = university_pca)

summary(model_pca2)

Multiple R-squared: 0.6705

 # Adjusted R-squared: 0.6405

PCA in R – Value of PCA in a Regression (Predicting Graduation Rate)

We find that the Adj Rsq is even higher when we use two Principal Components instead of using the five variables.

PCA Exercise

For the PCA model analyzed using Rattle, run a linear regression (using R commands) to predict the graduation rate based on the first three principal components.

Report the Adj Rsq. It should be approximately 0.66.

What About Iris?

Run PCA_IRIS_commented.R script

Iris data set is in R library. If you get an error keep running and it will use iris dataset in the library.

You will see the Principal Components tell a story similar to what we saw when we examined correlations!

When to Use PCA?

When the **DATA** is **MULI-VARIATE** and **NUMERIC**

When Number of **FEATURES** is **LARGE**

When Data is **Unimodal**

When **CLASS** labels are **NOT** present / ignored

When to Use PCA?

To **VISUALIZE** the data – top 2 or top 3 PC’s.

To **REDUCE** #Dimensions/Features for next stages

To **REMOVE Noise** in features and **Outliers** in data

Some Other Methods

Fisher Discriminant Analysis

Multi-dimensional Scaling (MDS)

- t-SNE

Self Organizing Maps

Summary

Curse of dimensionality: creates issues when predicting and extending results

Data visualization comes to our help! Helps identify clusters and features for further analysis

PCA : For numeric data provides handy guidance to identifying important features

References

Dlanglois. (2005, July 9). Iris versicolor3. Wikimedia Commons. Retrieved from https://bit.ly2KQOo30

Ender005. (2016, July 3). *Big data*. Wikimedia Commons. Retrieved from https://bit.ly/2vC7akq

Fisher, R. A. Iris dataset.

Fulkerson, A. (2011, February 7). *Edware Tufte giving a class*. Wikimedia Commons. Retrieved from https://bit.ly/2VkIbk2

KMJ. (2009, April 7). Flashlight. Wikimedia Commons. Retrieved from https://bit.ly/2KS7s0V

Radomil. (2015, May 15). *Kosaciec szczecinkowaty Iris setosa*. Wikimedia Commons. Retrieved from https://bit.ly/2vhLYQk

Mayfield, F. (2007, May 28). Iris virginica. Wikimedia Commons. Retrieved from https://bit.ly/2XCKcWd

Rattle GUI / Togaware (https://rattle.togaware.com/)