Chapter 3. Data Processing: Data Reduction

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

Welcome

Notre Dame Lunch

	Calories KCAL	Fat Cal KCAL	Fat Gram	Fat %	Sat Fat Gram	Sat Fat %	Trans Fat Gram	Chol MG	Chol %	Sodium MG	Sodium %	Potassium MG	Potassium %	Carbs Gram	Carbs %	Dietary Fiber Gram	Dietary Fiber %	Sugars Gram	Protein Gram	Vit A %	Vit C %	Calcium %	Iron %
North Daily Menus - 9/5/2017 - Lunch																							
Asian Sauce Bar - Varies with Choices (1g)																							
Pork Tonkatsu - Cutlet (319g)	514	180	20	31%	5	26%	0.01	130	43%	1930	80%	330	9%	53	18%	1	6%	14	28	8%	2%	10%	10%
Agedashi Tofu - 2 Piece Portion (173g)	110	40	4.5	7%	1	4%	0	0	0%	320	13%	260	7%	11	4%	2	8%	1	9	6%	10%	20%	10%
Vegetable Yaki Soba - 6z Portion (167g)	342	81	9	14%	1	6%	0.03	0	0%	1160	48%	180	5%	53	18%	4	15%	5	13	10%	30%	10%	25%
Japanese Hiyashi Chuka Salad - 3z Portion (89g)	230	63	7	11%	1	4%	0.00	70	24%	310	13%	100	3%	38	13%	1	6%	1	9	6%	10%	2%	10%
Brown Rice - 4z Portion (64g)	230	14	1.5	2%	0	0%	0.00	0	0%	0	0%	140	4%	50	17%	3	11%	0	5	0%	0%	0%	6%
Stir Fry * See Specialty Bars - Nutrition in Bars (1g)																							
Steamed Chinese Rice - 3z Portion (85g)	111	0	0	0%	0	0%	0	0	0%	0	0%	30	1%	24	8%	0	1%	0	2	0%	0%	2%	6%
Salad Bar *see Salad Bar - Varies with Choices (1g)																							
Chicken Caesar Salad - 3.27z Portion (187g)	401	216	24	37%	5	27%	0	55	19%	760	32%	180	5%	24	8%	1	5%	1	26	100%	8%	20%	15%
Waffle Bar - See Breakfast Bar (1g)																							
Deli Bar * See Specialty Bars - Nutrition in Bars (1g)																							
Creamy Tomato Soup - 6 Flz Portion (181g)	159	90	10	16%	1.5	7%	0.09	35	11%	520	22%	160	5%	14	5%	1	6%	4	2	10%	8%	2%	4%
Beef Mushroom Burger - Patty (52g)	175	135	15	24%	6	32%	0	35	11%	20	1%	140	4%	< 1	0%	0	1%	0	8	0%	0%	2%	6%
Hamburger Buns - Buns (52g)	146	18	2	3%	0	0%	0	0	0%	250	10%	0	0%	26	9%	0	0%	0	4	0%	0%	40%	2%
Char-Grilled Chicken Breast - 3.25z Portion (123g)	134	22	2.5	4%	1	4%	0.01	85	29%	460	19%	320	9%	2	1%	0	0%	2	25	0%	0%	0%	2%
Breaded Chicken Fillet - Patty+Roll (164g)	404	171	19	30%	2	9%	0	25	8%	860	36%	0	0%	38	13%	< 1	4%	0	20	0%	0%	20%	6%
Malibu Vegan Gardenburger - 3.2z patty (91g)	169	81	9	14%	0	0%	0.00	0	0%	440	18%	280	8%	21	7%	4	16%	0	4	0%	0%	2%	4%
Black Pepper French Fries - 2.5z (71g)	130	54	6	9%	0	0%	0.00	0	0%	400	17%	240	7%	17	6%	2	8%	0	2	0%	8%	0%	4%
Grilled Cheese on White - Sandwich (92g)	369	243	27	42%	7	33%	0	35	11%	720	30%	0	0%	20	7%	0	0%	2	10	8%	0%	35%	4%
Roasted Cod with Chorizo and Black Beans - Fillet (242g)	433	216	24	37%	43	214%	0.03	10	4%	580	24%	780	22%	31	10%	4	17%	2	23	15%	35%	4%	10%
Chicken Pot Pie - Pot Pie (198g)	363	72	8	13%	3	16%	0	105	34%	500	21%	310	9%	57	19%	2	10%	2	13	8%	10%	20%	20%
Roasted Vegetables - 4z Portion (147g)	97	40	4.5	7%	0	0%	0	0	0%	220	9%	420	12%	14	5%	3	11%	4	2	140%	20%	4%	4%
Cooked Quinoa - 4z Portion (48g)	176	27	3	4%	0	2%	0	0	0%	55	2%	270	8%	31	10%	3	14%	0	7	0%	0%	2%	10%
Broccoli Florets - 4z Portion (122g)	34	0	0	0%	0	0%	0	0	0%	15	1%	170	5%	6	2%	4	15%	2	4	25%	80%	4%	4%
Steamed Sliced Carrots - 4z Portion (109g)	40	4	0.5	1%	0	1%	0.00	0	0%	65	3%	210	6%	8	3%	4	14%	4	< 1	370%	4%	4%	4%
Falafel Station - 5z Portion (126g)	255	90	10	16%	0	2%	0	0	1%	790	33%	40	1%	35	12%	3	12%	2	8	6%	8%	15%	15%
Meatballs - Meatball (14g)	45	32	3.5	5%	0	0%	0	5	2%	125	5%	0	0%	1	0%	0	1%	0	2	0%	0%	2%	2%
Garlic Parmesan Breadsticks - Breadstick (49g)	208	63	7	11%	2.5	13%	0.08	10	3%	350	15%	55	2%	29	10%	1	6%	0	7	4%	0%	6%	10%
Multigrain Penne - 4z Portion (114g)	163	22	2.5	4%	0	1%	0.00	0	0%	25	1%	0	0%	31	10%	3	13%	2	8	0%	0%	2%	8%
Small Shells - 2z Portion (57g)	70	4	0.5	1%	0	0%	0.00	0	0%	25	1%	0	0%	13	4%	< 1	2%	0	2	0%	0%	0%	0%
Alfredo Sauce - 2z Portion (55g)	79	45	5	7%	0	0%	0	10	3%	320	13%	80	2%	4	1%	0	0%	0	3	2%	2%	10%	2%
Tomato & Basil Marinara Sauce - 1 Flz Portion (28g)	15	9	1	2%	0	1%	0	0	0%	105	4%	30	1%	2	1%	0	2%	0	0	0%	2%	2%	2%
Sausage Pizza - 16-Cut Slice (78g)	181	72	8	12%	0	0%	0	20	7%	630	26%	105	3%	20	7%	1	5%	< 1	8	4%	2%	15%	2%
Popeye Pizza - 16-Cut Slice (79g)	153	45	5	8%	0	2%	0	15	6%	500	21%	135	4%	20	7%	1	5%	< 1	7	15%	6%	10%	2%
Pepperoni Pizza - Piece (72g)	167	63	7	10%	0	0%	0	20	6%	560	23%	85	2%	19	6%	1	4%	< 1	7	4%	2%	10%	2%
Cheese Pizza - Piece (68g)	145	45	5	7%	0	0%	0	15	4%	490	20%	85	2%	19	6%	1	4%	< 1	6	4%	2%	10%	2%
Tollhouse Bar with Walnuts - Piece (59g)	272	135	15	24%	3	15%	0.00	25	8%	160	7%	100	3%	30	10%	1	4%	20	4	10%	0%	4%	4%
Baker's Choice Cookies - Nutrition Varies (1g)																							
Frozen Yogurt & Ice Cream Bar - Nutrition in Bars (1g)																							
Huevos Rancheros con Chorizo - 7z Portion (140g)	233	144	16	24%	2	10%	0.03	215	72%	350	15%	160	5%	11	4%	1	4%	1	12	20%	10%	4%	8%
Fried Seasoned Potato Cubes - 4z Portion (88g)	114	36	4	6%	0	0%	0.00	0	0%	360	15%	220	6%	18	6%	2	8%	0	2	0%	2%	0%	0%
ND Plain Bagel - Bagels (131g)	303	9	1	2%	0	0%	0	0	0%	560	23%	120	3%	63	21%	2	8%	0	12	0%	0%	2%	20%
Dinner Rolls - Rolls (38g)	100	18	2	3%	1	5%	0.06	5	2%	130	5%	50	1%	17	6%	< 1	3%	2	3	2%	0%	2%	0%
Pastaria Action Station - Varies with Choices (351g)	869	405	45	69%	2	11%	0.06	100	33%	1790	74%	240	7%	81	27%	5	19%	3	39	20%	10%	80%	10%

A Single Chicken Breast (USDA)

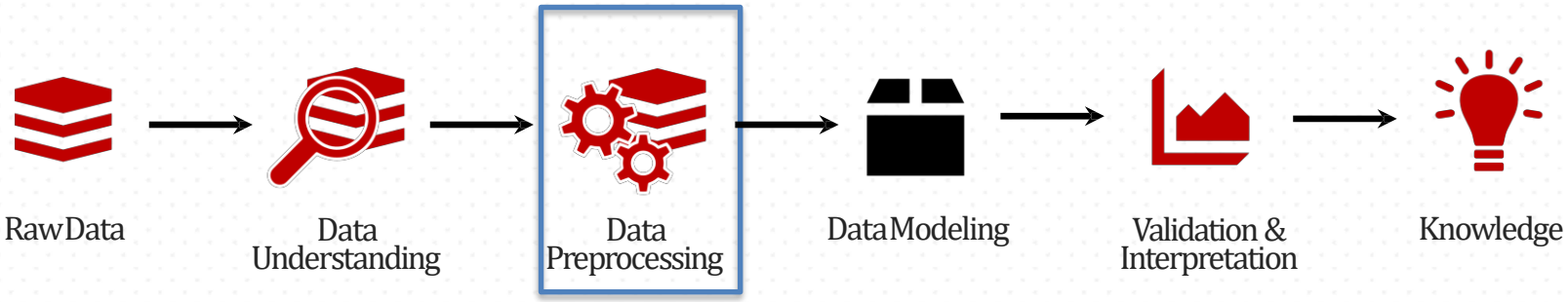
Water	73.24	4	0.107	82.03
Energy	143	--	--	160
Energy	598	--	--	670
Protein	17.44	4	0.189	19.53
Total lipid (fat)	8.1	4	0.243	9.07
Ash	1.17	4	0.173	1.31
Carbohydrate, by difference	0.04	--	--	0.04
Fiber, total dietary	0	--	--	0
Sugars, total	0	--	--	0
Minerals				
Calcium, Ca	6	4	0.494	7
Iron, Fe	0.82	4	0.051	0.92
Magnesium, Mg	21	4	0.295	24
Phosphorus, P	178	4	1.601	199
Potassium, K	522	4	99.392	585
Sodium, Na	60	4	3.659	67
Zinc, Zn	1.47	4	0.059	1.65
Copper, Cu	0.065	4	0.01	0.073
Manganese, Mn	0.016	4	0.002	0.018
Selenium, Se	10.2	4	0.53	11.4
Vitamins				
Vitamin C, total ascorbic acid	0	--	--	0
Thiamin	0.109	4	0.008	0.122
Riboflavin	0.241	4	0.013	0.27
Niacin	5.575	4	0.223	6.244

Pantothenic acid	1.092	4	0.026	1.223
Vitamin B-6	0.512	4	0.017	0.573
Folate, total	1	4	0.408	1
Folic acid	0	4	0	0
Folate, food	1	4	0.408	1
Folate, DFE	1	--	--	1
Choline, total	58.8	--	--	65.9
Betaine	7.7	--	--	8.6
Vitamin B-12	0.56	4	0.065	0.63
Vitamin B-12, added	0	--	--	0
Vitamin A, RAE	0	--	--	0
Retinol	0	2	--	0
Carotene, beta	0	--	--	0
Carotene, alpha	0	--	--	0
Cryptoxanthin, beta	0	--	--	0
Vitamin A, IU	0	--	--	0
Lycopene	0	--	--	0
Lutein + zeaxanthin	0	--	--	0
Vitamin E (alpha-tocopherol)	0.27	4	0.083	0.3
Vitamin E, added	0	--	--	0
Tocopherol, beta	0	4	0	0
Tocopherol, gamma	0.17	4	0.037	0.19
Tocopherol, delta	0.03	4	0.009	0.03
Vitamin K (phylloquinone)	0.8	--	--	0.9
Lipids				

Fatty acids, total saturated	g	2.301	--	--	2.577
4:00	g	0	--	--	0
6:00	g	0	--	--	0
8:00	g	0	4	0	0
10:00	g	0	4	0	0
12:00	g	0	4	0	0
14:00	g	0.041	4	0.001	0.046
15:00	g	0	4	0	0
16:00	g	1.791	4	0.051	2.006
17:00	g	0.007	4	0	0.008
18:00	g	0.456	4	0.01	0.511
20:00	g	0.005	4	0	0.006
22:00	g	0	4	0	0
Fatty acids, total monounsaturated	g	3.611	--	--	4.044
14:01	g	0.016	4	0.001	0.018
15:01	g	0	4	0	0
16:1 undifferentiated	g	0.536	4	0.019	0.6
17:01	g	0	4	0	0
18:1 undifferentiated	g	3.035	4	0.107	3.399
18:1 c	g	2.992	4	0.104	3.351
18:1 t	g	0.042	4	0.002	0.047
20:01	g	0.025	4	0.001	0.028
22:1 undifferentiated	g	0	--	--	0
Fatty acids, total polyunsaturated	g	1.508	--	--	1.689
18:2 undifferentiated	g	1.324	4	0.078	1.483

18:2 n-6 c,c	g	1.303	4	0.077	1.459
18:2 t,t	g	0.022	4	0.001	0.025
18:3 undifferentiated	g	0.071	4	0.003	0.08
18:3 n-3 c,c,c (ALA)	g	0.057	4	0.003	0.064
18:3 n-6 c,c,c	g	0.014	4	0	0.016
18:04	g	0	--	--	0
20:2 n-6 c,c	g	0.011	4	0	0.012
20:3 undifferentiated	g	0	4	0	0
20:4 undifferentiated	g	0.074	4	0.003	0.083
20:5 n-3 (EPA)	g	0.008	--	--	0.009
22:5 n-3 (DPA)	g	0.008	--	--	0.009
22:6 n-3 (DHA)	g	0.023	--	--	0.026
Fatty acids, total trans	g	0.065	--	--	0.073
Fatty acids, total trans-monoenoic	g	0.042	--	--	0.047
Cholesterol	mg	86	4	3.999	96
Amino Acids					
Tryptophan	g	0.147	--	--	0.165
Threonine	g	0.727	--	--	0.814
Isoleucine	g	0.794	--	--	0.889
Leucine	g	1.361	--	--	1.524
Lysine	g	1.509	--	--	1.69
Methionine	g	0.446	--	--	0.5
Cystine	g	0.188	--	--	0.211
Phenylalanine	g	0.683	--	--	0.765
Tyrosine	g	0.604	--	--	0.676

Chapter 3. Data Preprocessing



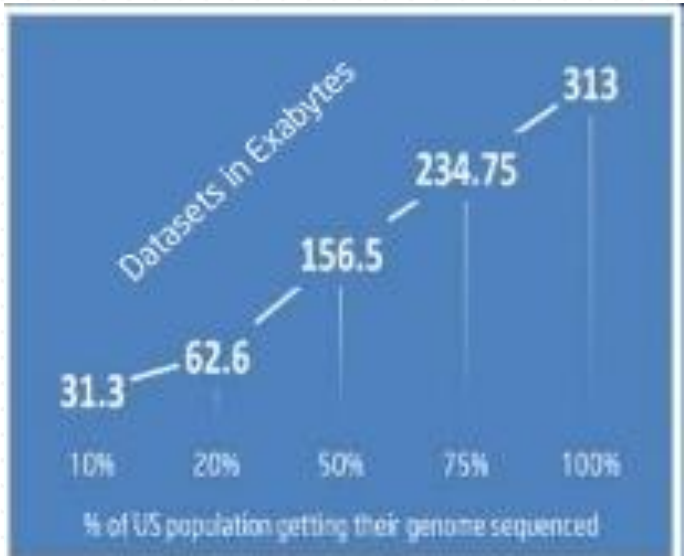
Data Preprocessing

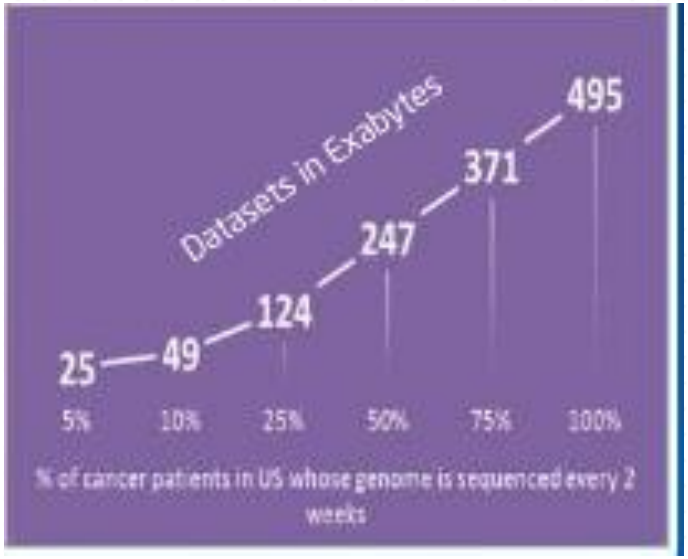
- Data cleaning
- Data integration
- **Data reduction**
  - Reduce data objects
- **Dimensionality reduction**
  - Reduce dimensions and attributes

Data Reduction

Obtain a reduced representation of the data set

Why? Complex analysis may take a very long time to run on the complete data set





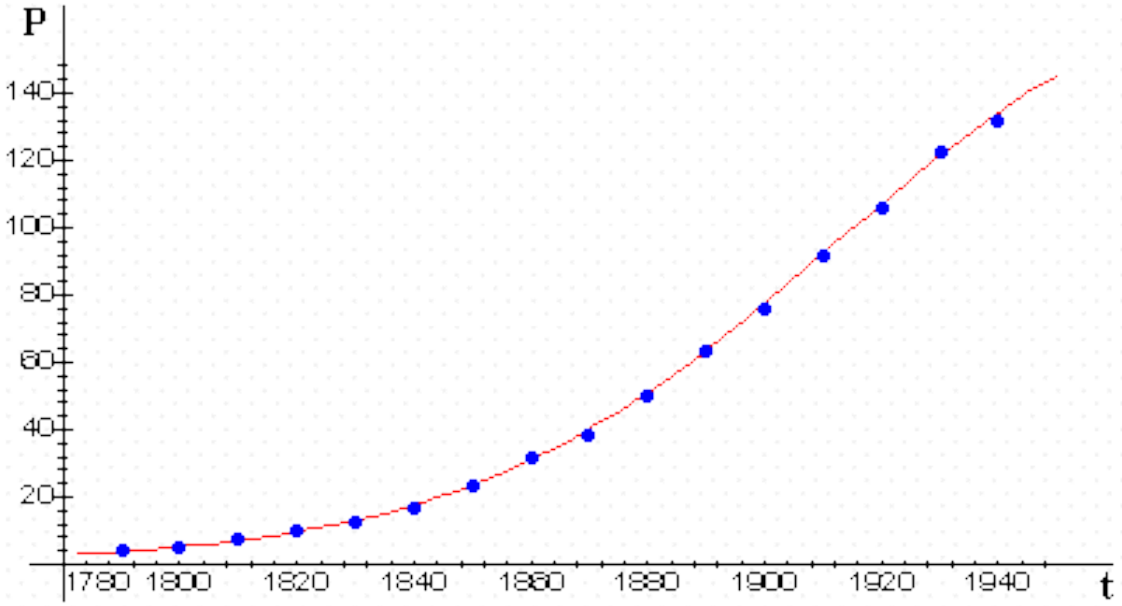
Numerosity reduction
(Data Reduction)

Reduce data volume by choosing alternative, smaller forms of data representation

Data Approximations

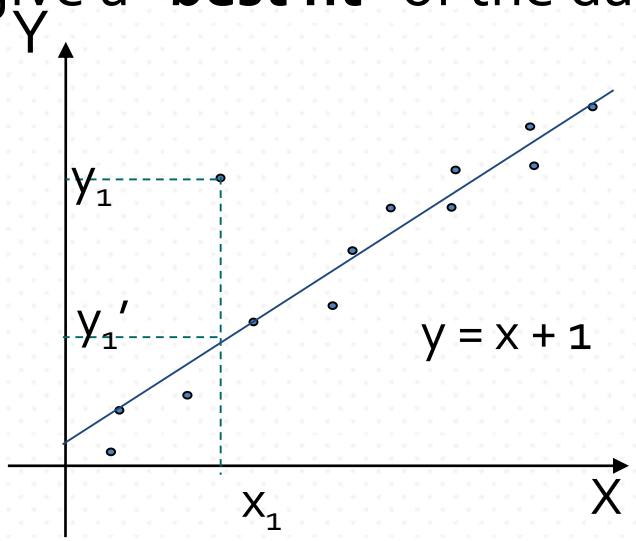
Parametric (model-based) measures

Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)



Regression Analysis

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values
  - of a ***dependent variable*** (also called ***response variable*** or ***measurement***): Y
  - and of one or more *independent variables* (also known as ***explanatory variables*** or ***predictors***): X, or X_1, X_2, …X_n
- Parameters are estimated to give a "**best fit**" of the data
  - Data: (x_1, y_1)
  - Fit of the data: (x_1, y_1')
    - Ex. $y_1' = x_1 + 1$

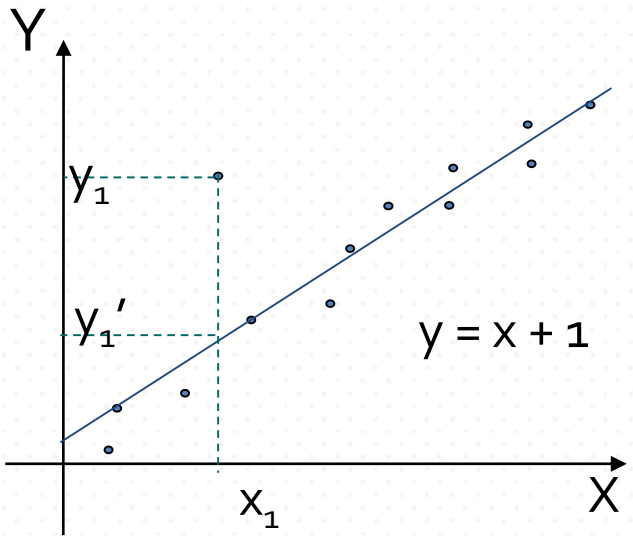


Regression Analysis

- Most commonly the best fit is evaluated by using the ***least square method***, but other criteria have also been used

$$\min g = \sum_{i=1}^{n} (y_i - y'_i)^2, \text{ where } y'_i = f(x_i, \beta)$$

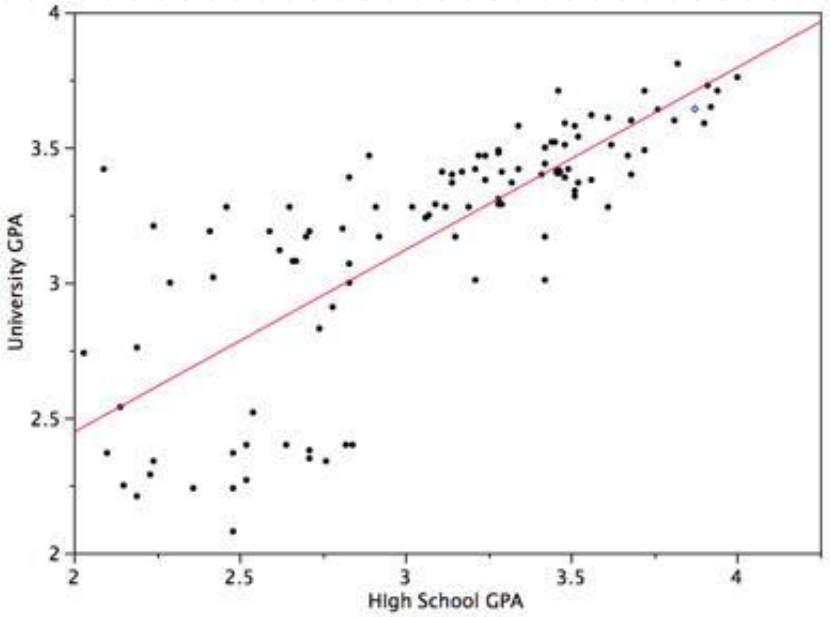
- Used for **prediction** (including forecasting of time-series data), **inference**, **hypothesis testing**, and **modeling of causal relationships**



Set up $y = f(x) = \beta_1 x + \beta_2$
Learn β by minimizing the least square error

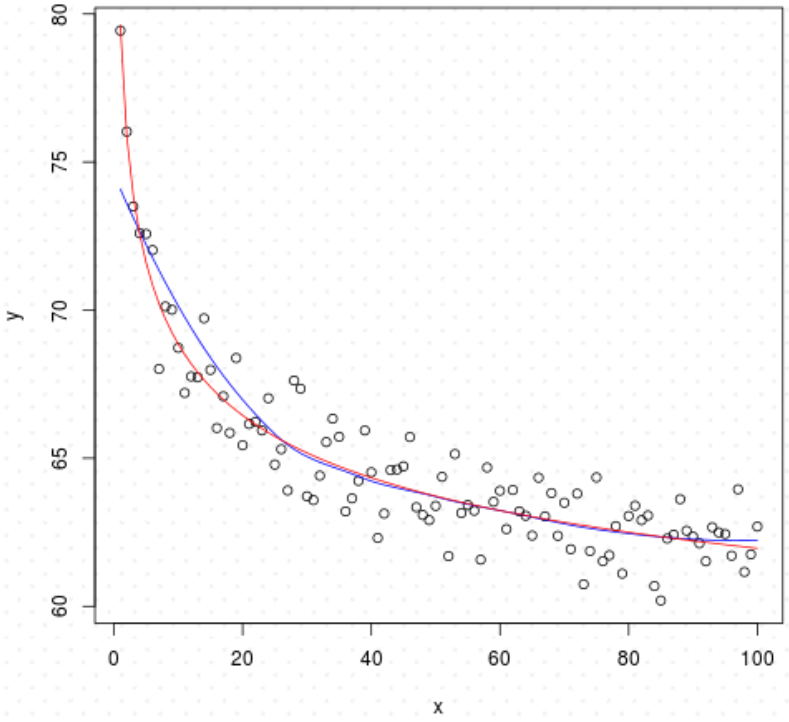
Linear Regression

- Linear regression: $Y = w X + b$
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
  - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand



Nonlinear Regression

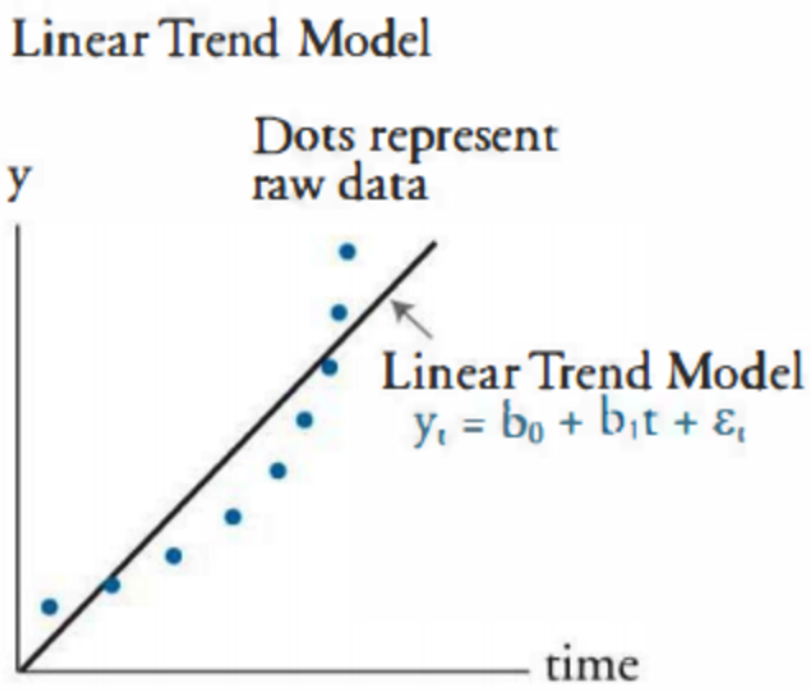
- Nonlinear regression:
  - Data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables

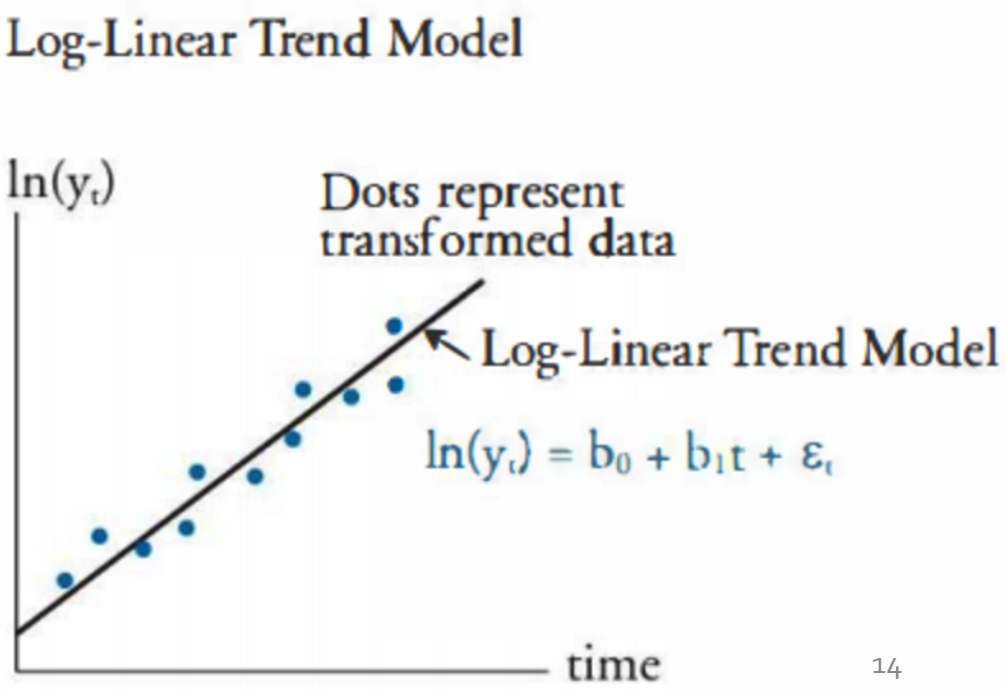


Log-Linear Model

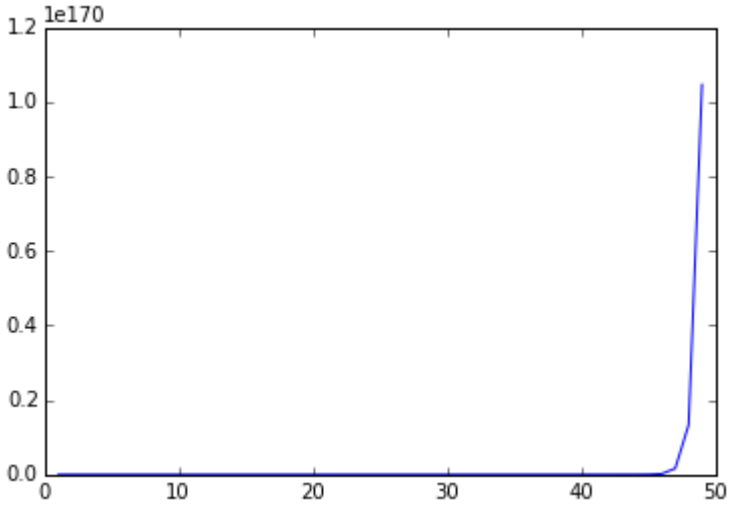
- Log-linear model
  - A math model that takes the form of a **function whose logarithm** is a linear combination of the parameters of the model

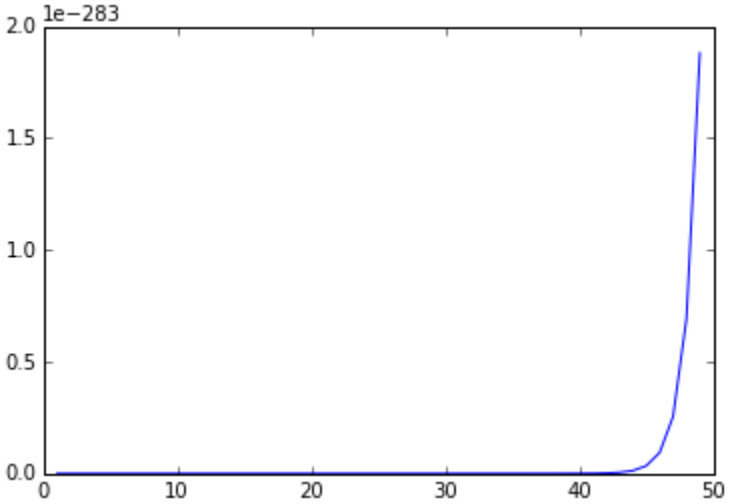
Q: How about Log-Log model?





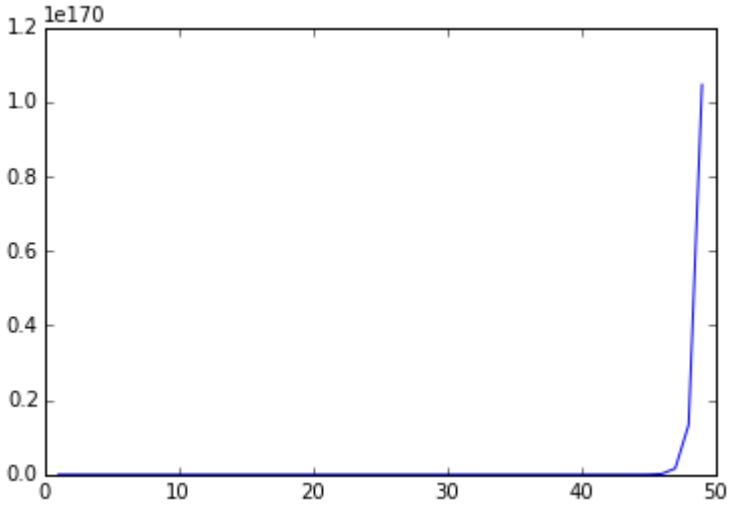
Log-Linear transforms



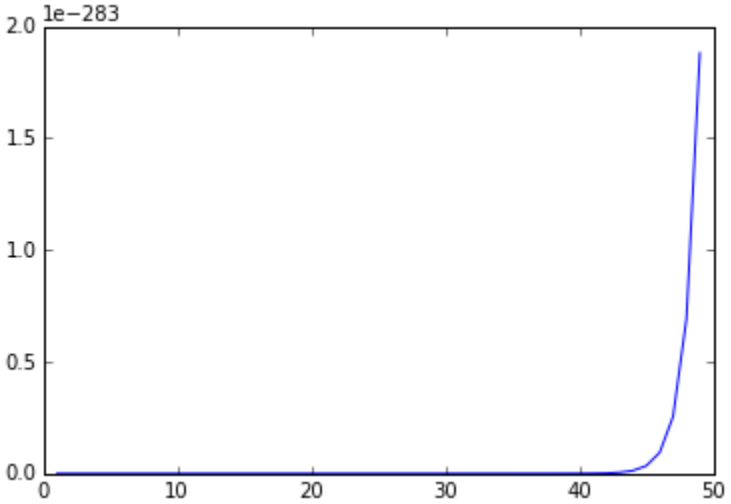


Log-Linear transforms

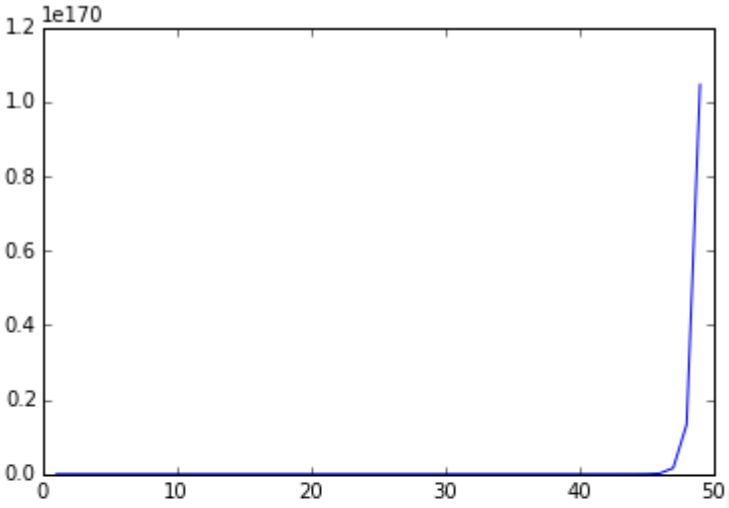
Power-Law

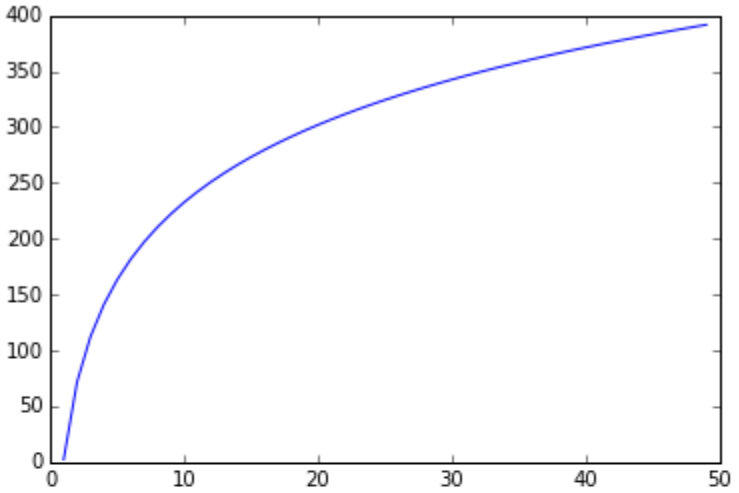


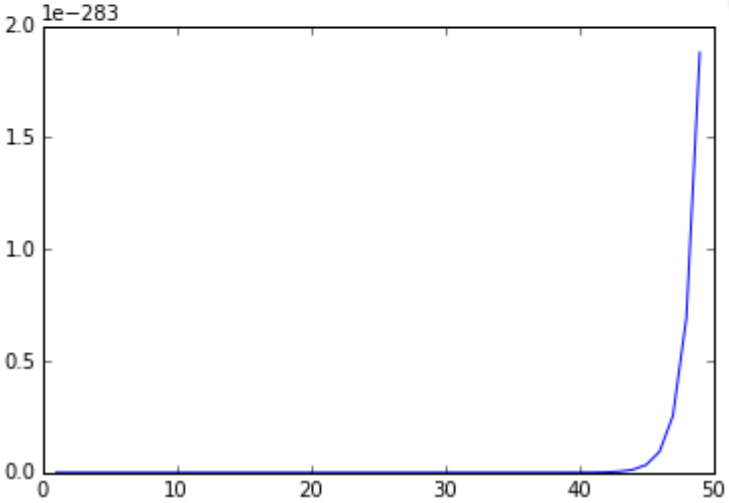
Exponential

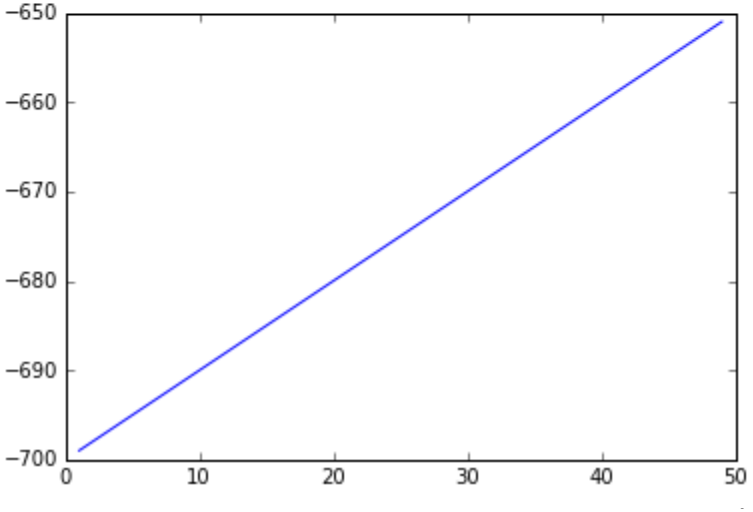


Log Transform



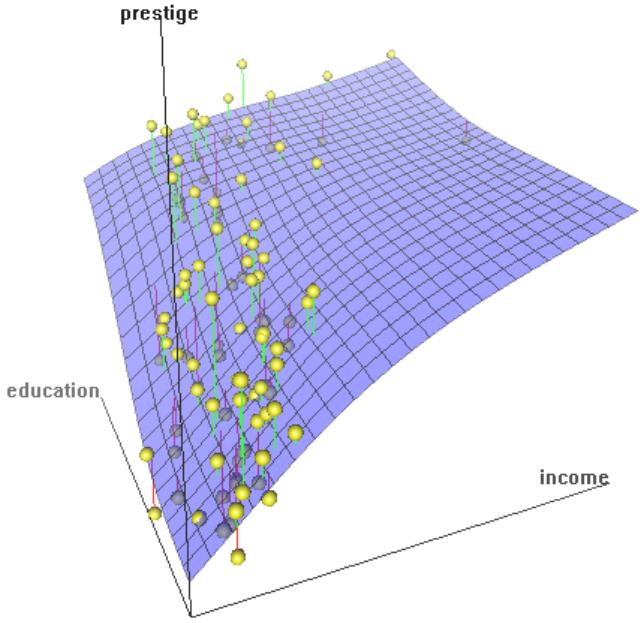




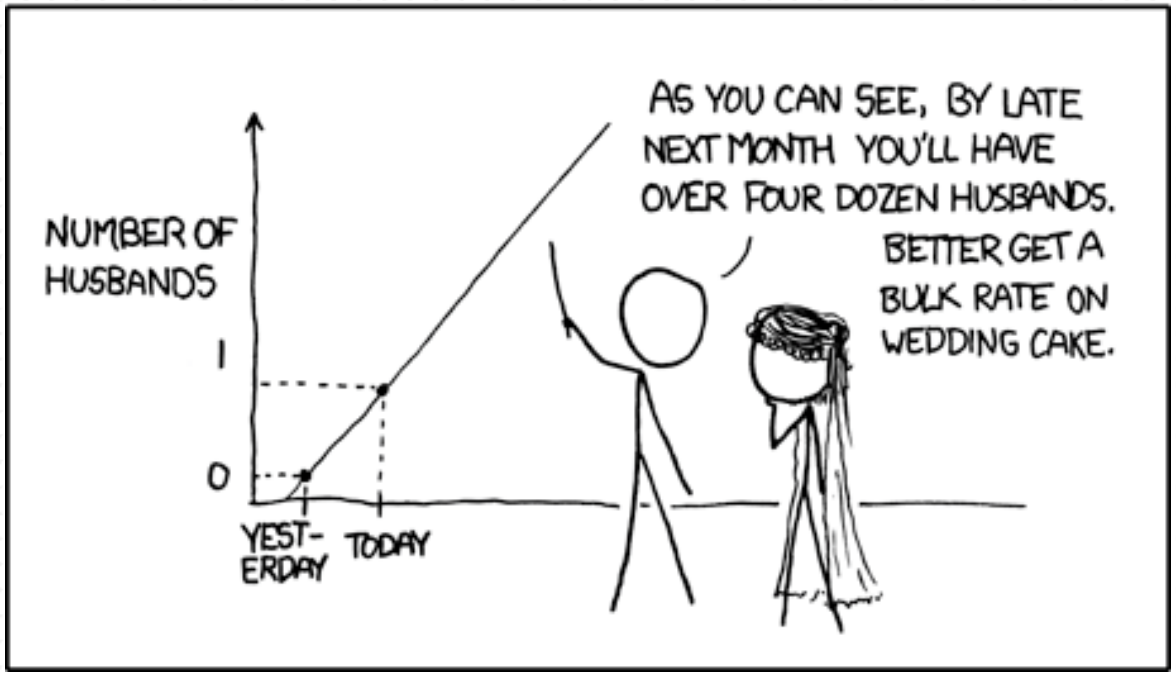


Multiple Regression

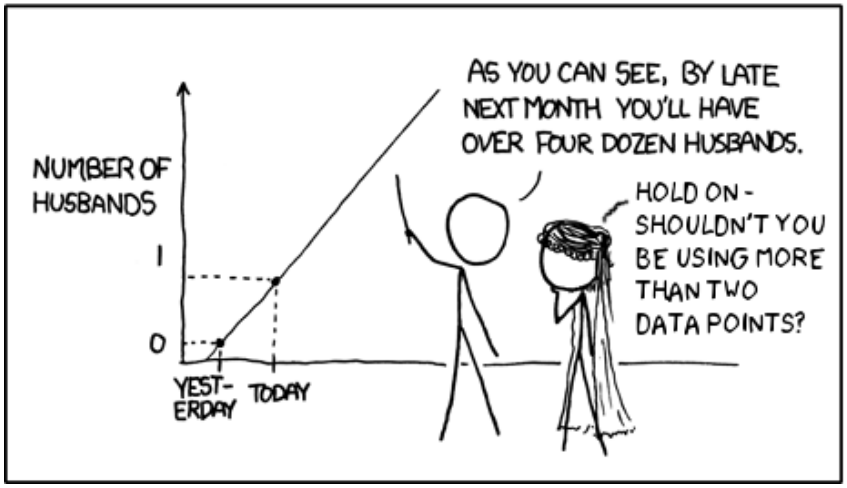
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Allows a response variable Y to be modeled as a linear function of multidimensional feature vector
  - Many nonlinear functions can be **transformed** into the above

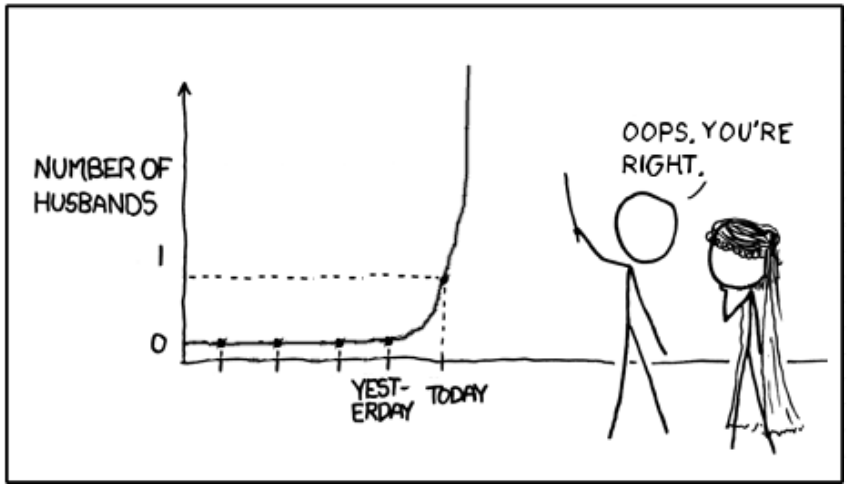


Caution: Extrapolation



Caution: Correct Distribution



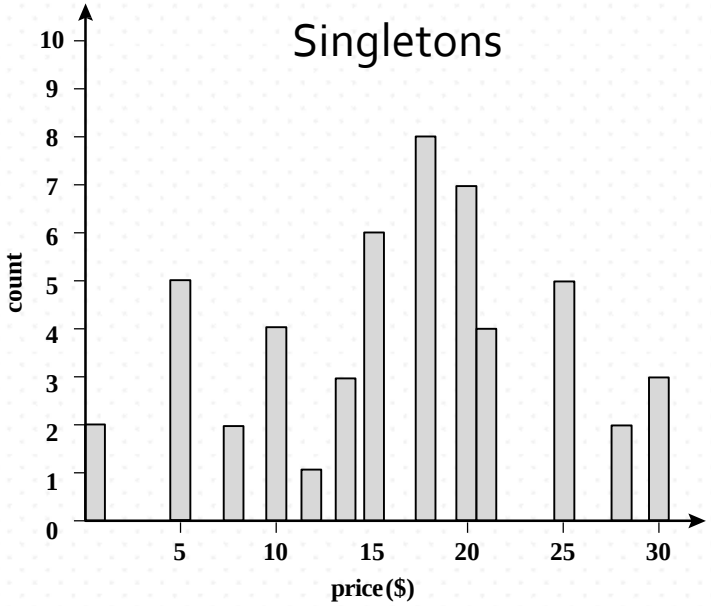


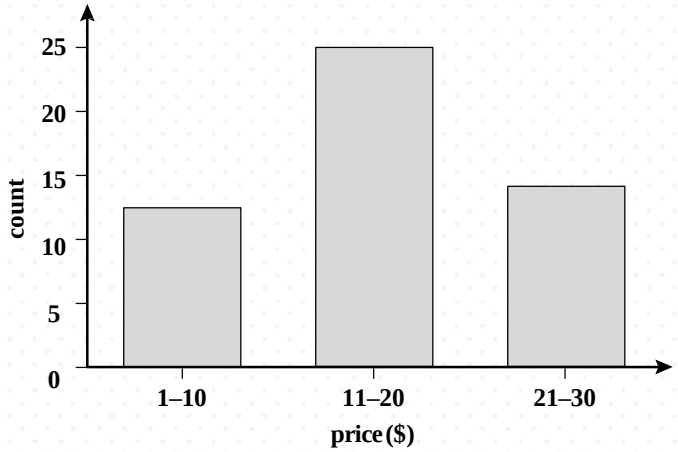
Data Approximations

Non-Parametric measures

Reduces the data itself, by identifying groupings or sampling

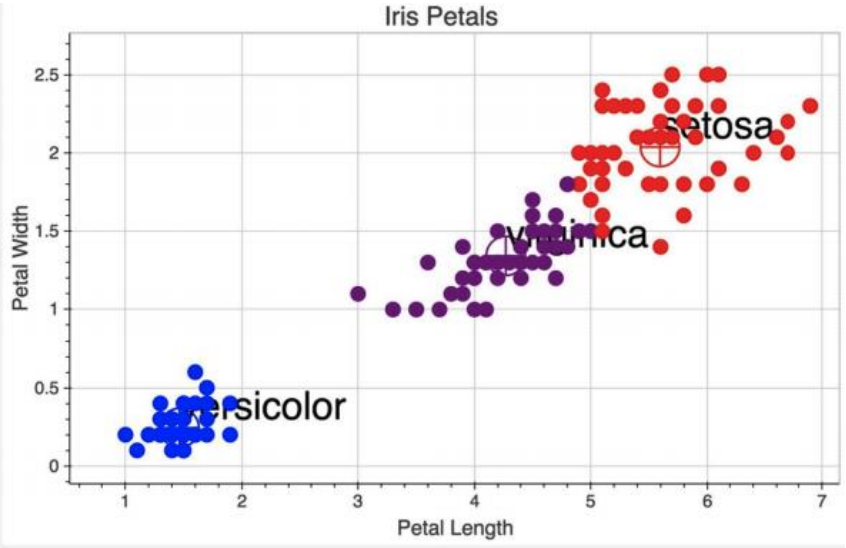
Histogram Analysis



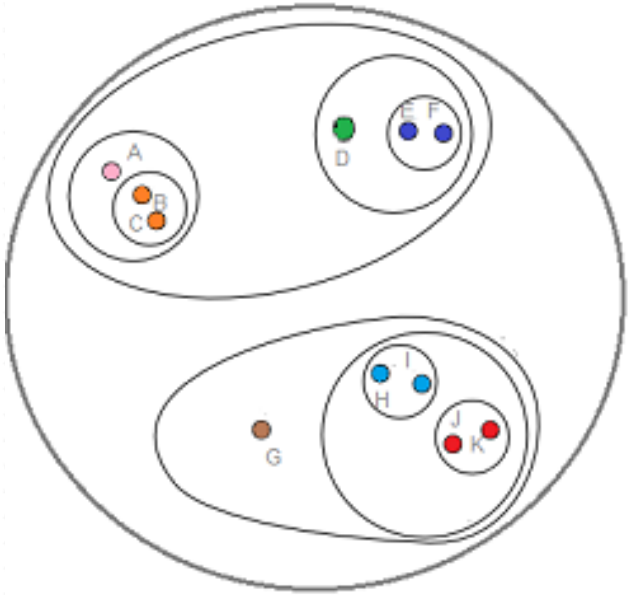


Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- There are many choices of clustering definitions and clustering algorithms
  - Cluster analysis will be studied in depth in Chapter 10



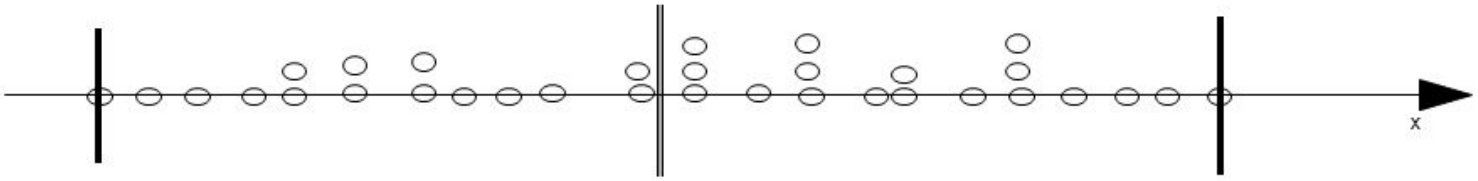
Preview: Getting Relations Between Clusters

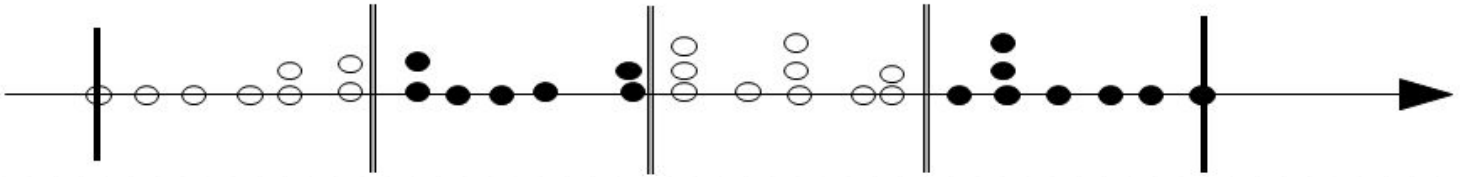


Preview: Discretization

Unsupervised Discretization

- Binning
- Histogram Analysis





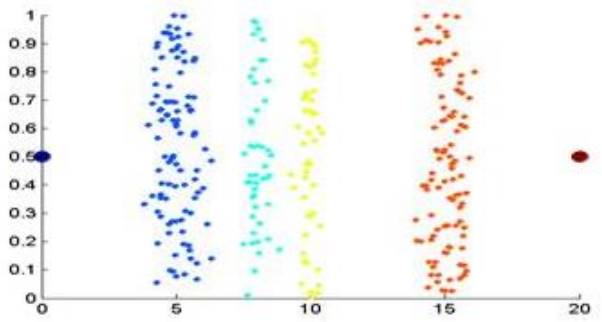
Supervised Discretization

- Cluster
- Decision Tree
- Correlation

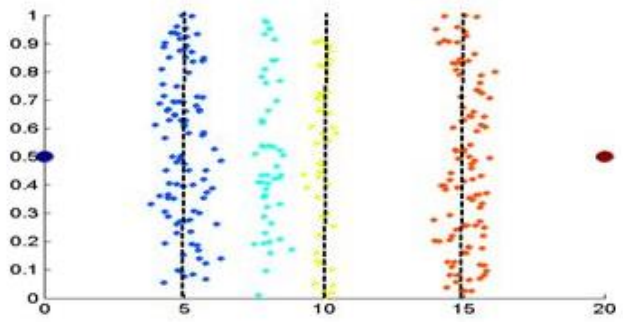
Preview: Discretization

- Top Down:
  - The process starts by first finding one or a few points (called *split points* or *cut points*) to split the entire attribute range, and then repeats this recursively on the resulting intervals, *splitting*.
- Bottom Up
  - Starts by considering all of the continuous values as potential split-points, removes some by merging neighborhood values to form intervals, and then recursively applies this process to the resulting intervals

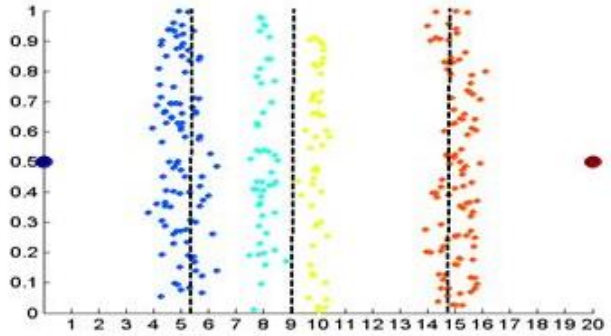
Picking The *Right* Method



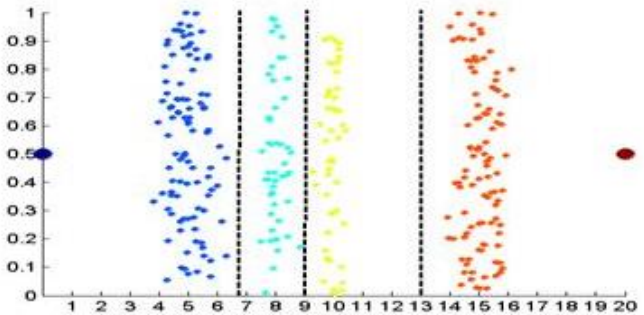
Data



Equal interval width



Equal frequency



K-means

Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew

Simple random sampling:

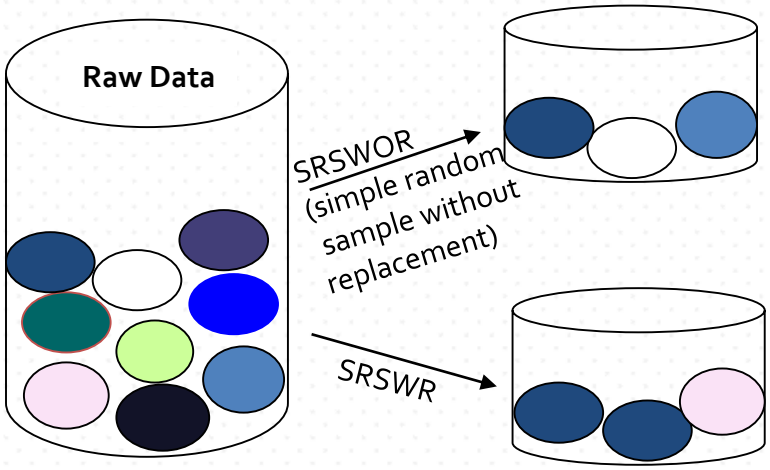
Equal probability of selecting any particular item

Sampling without replacement:

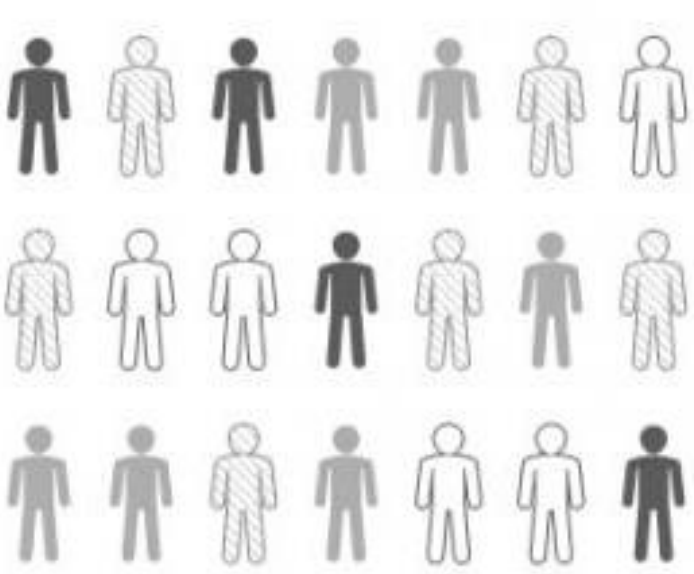
Once an object is selected, it is removed from the population

Sampling with replacement:

A selected object is not removed from the population

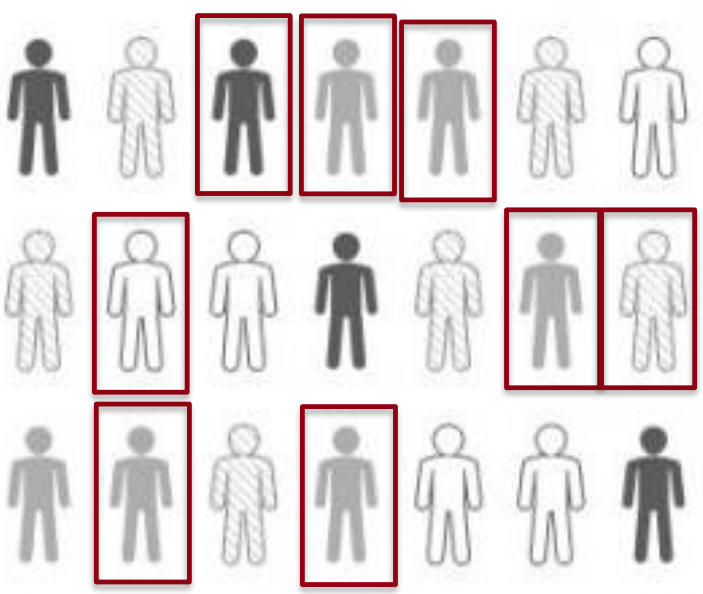


What if the data is imbalanced?

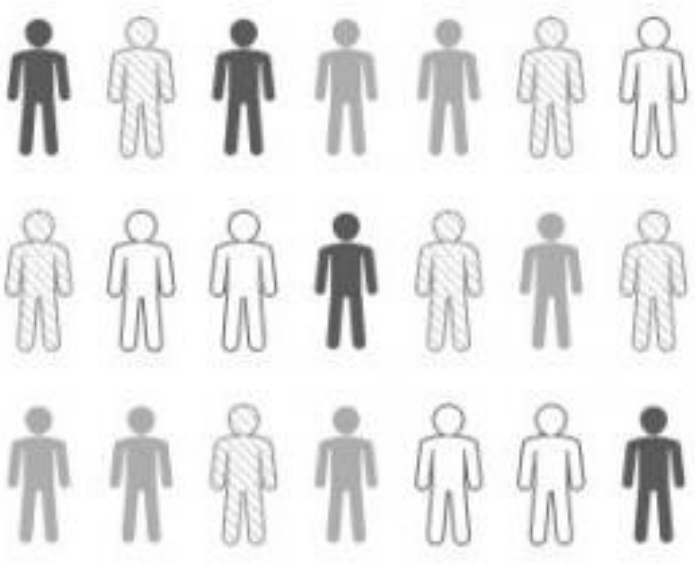


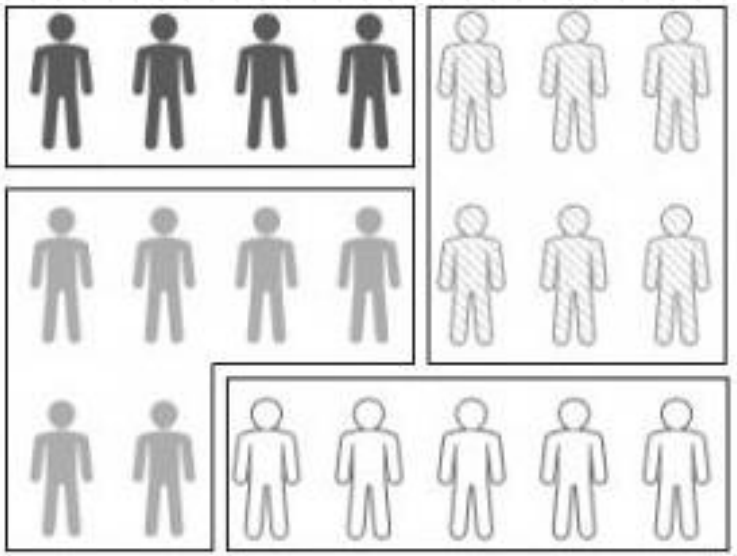
What if the data is imbalanced?

Random Same n=8



Stratified Sampling



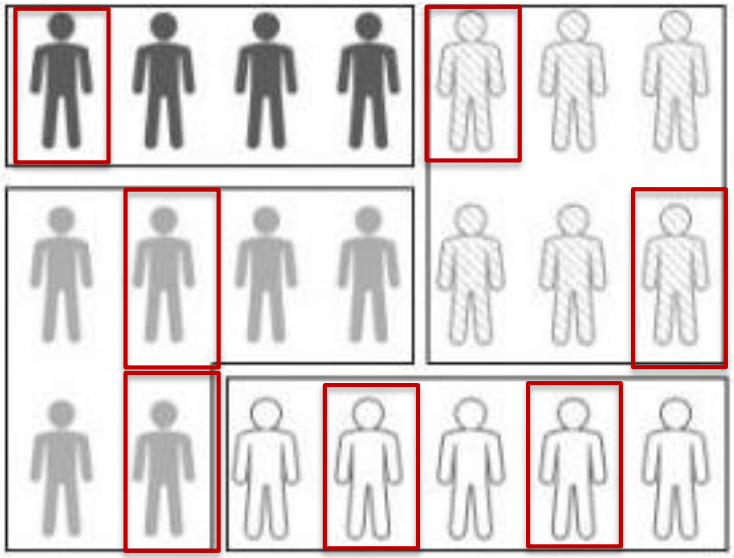


Stratified Population

Partition (or cluster) the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)

Stratified Sampling

Random Same n=8

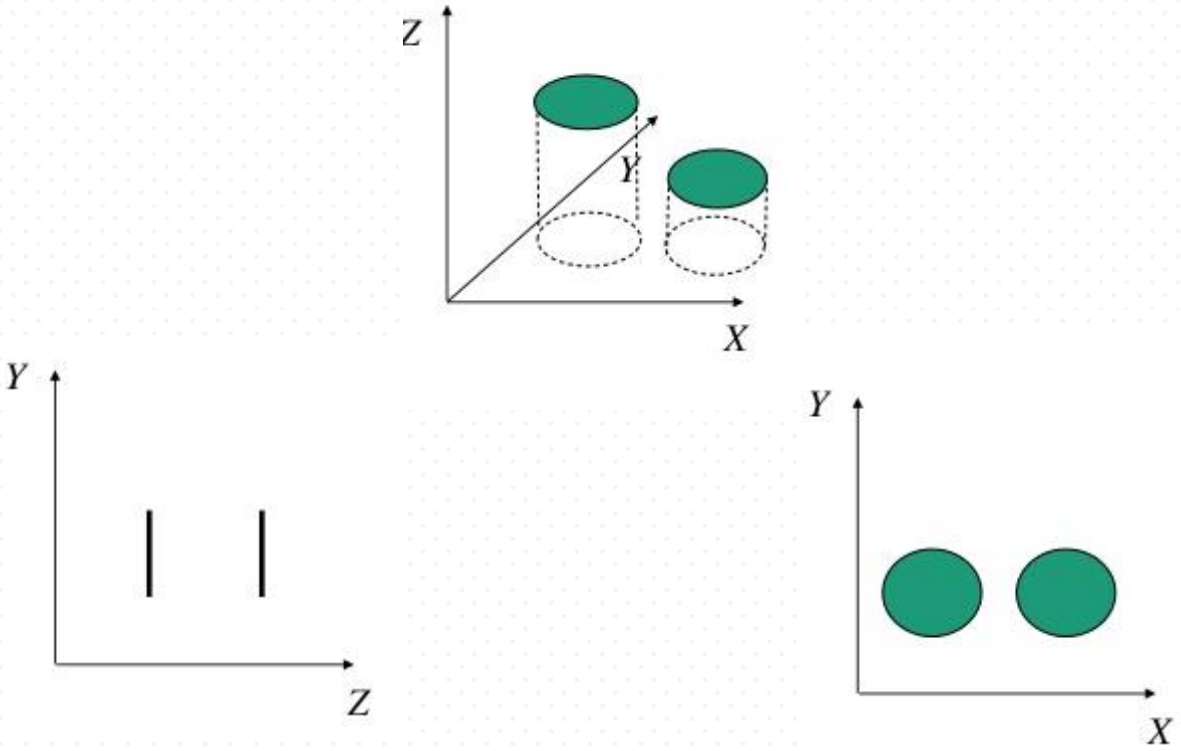


Stratified Population

iPython Examples

Dimensionality reduction

Reduce data attributes



Dimensionality Reduction

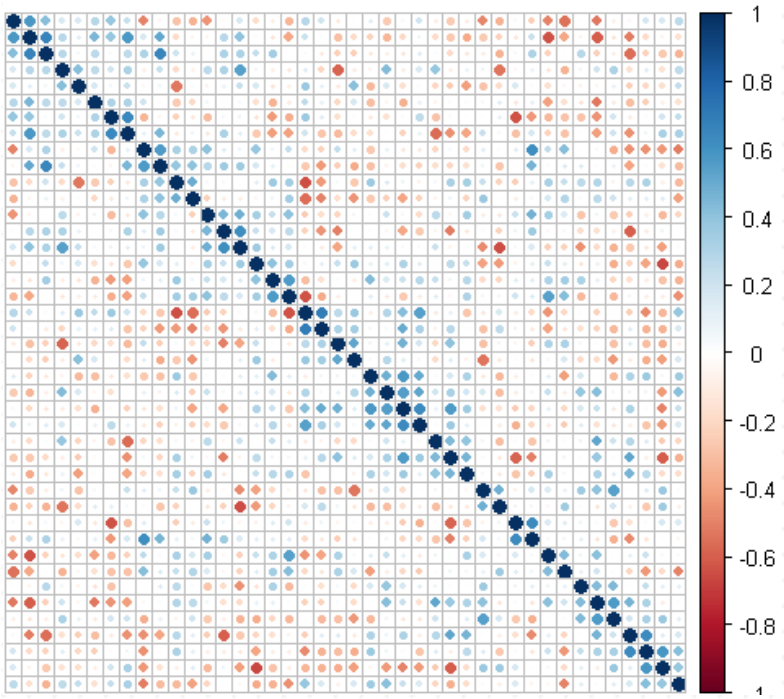
- Thinking Back: Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially

- Advantages of dimensionality reduction
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

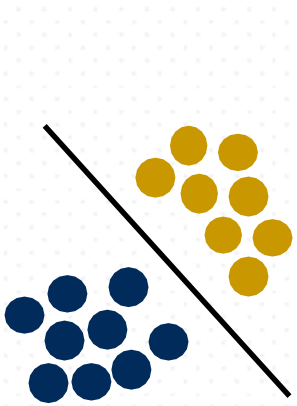
Feature selection (FS):

Find a subset of the original attributes

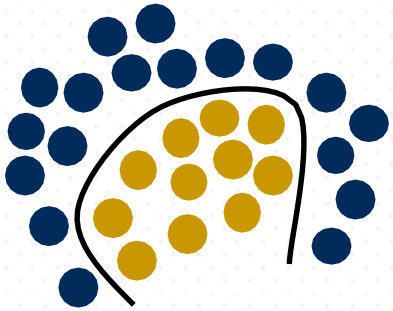
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
    - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
    - Ex. A student's ID is often irrelevant to the task of predicting his/her GPA



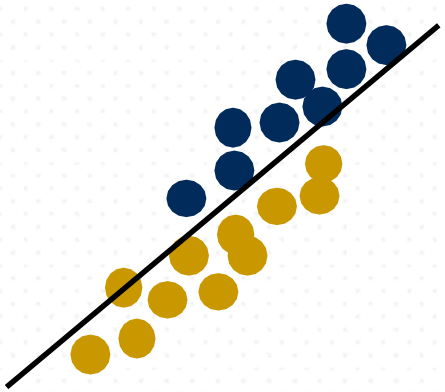
Features Are Complex



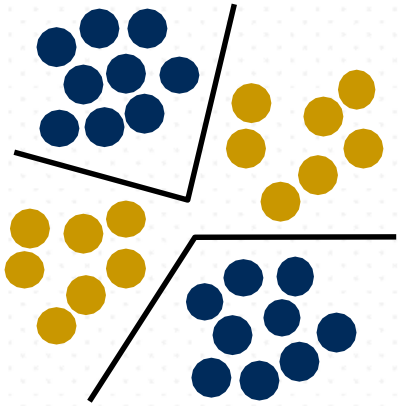
Linear Separability



Non-Linear Separability



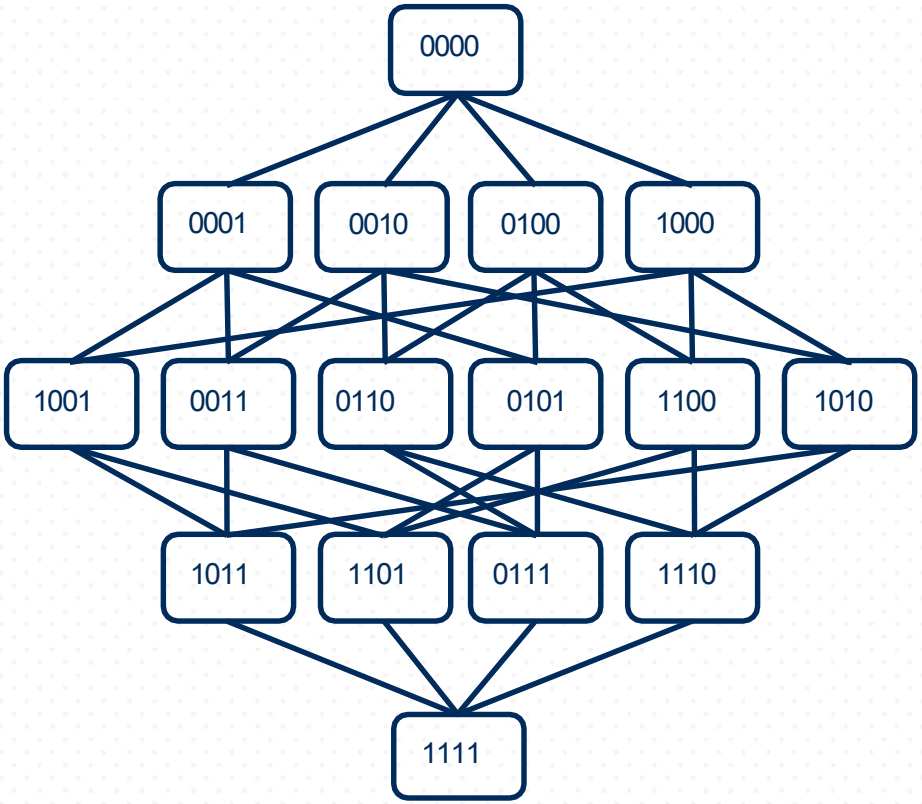
Highly Correlated



Multi-Modal

There are Many Feature Combinations

There are 2^d possible attribute combinations of d attributes



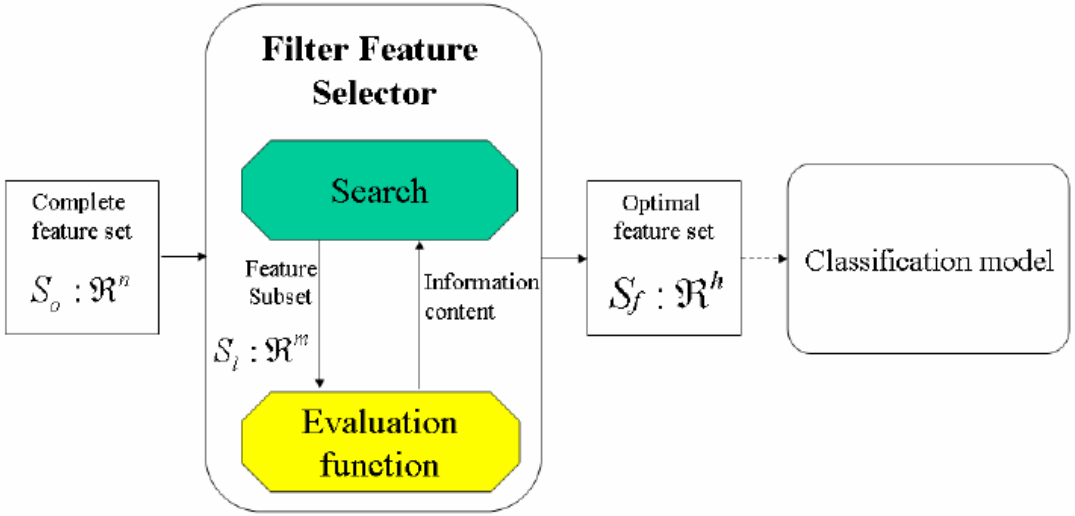
Feature Selection Techniques

- Filters
  - Features selected before algorithm is run using approach independent of task
- Wrappers
  - Features selected with target algorithm used as a "black box", typically without enumerating all subsets.

Filters

Use feature selection technique or heuristic to rank-order relevant features, selecting only those above a threshold.

- Examples:
  - information gain
  - correlation measures
  - RELIEF-F
  - odds ratio



Filters

Advantages:

- Typically faster execution
- Non-iterative
- Generalizable

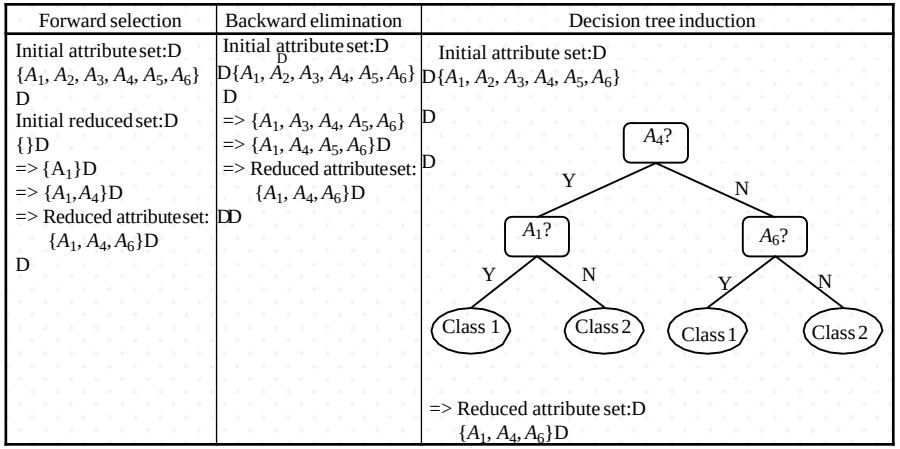
Disadvantages:

- Require arbitrary cut-off
- Don't capture inter-feature interactions

Wrappers

Use the performance of algorithm using one or more features to identify ideal combination of features.

- Best step-wise feature selection:
  - The best single-attribute is picked first
  - Then next best attribute condition to the first, ...
- Step-wise attribute elimination:
  - Repeatedly eliminate the worst attribute
- Best combined attribute selection and elimination



Wrappers

Advantages:

- Tuned for the underlying inductive bias.
- Can capture inter-feature interactions

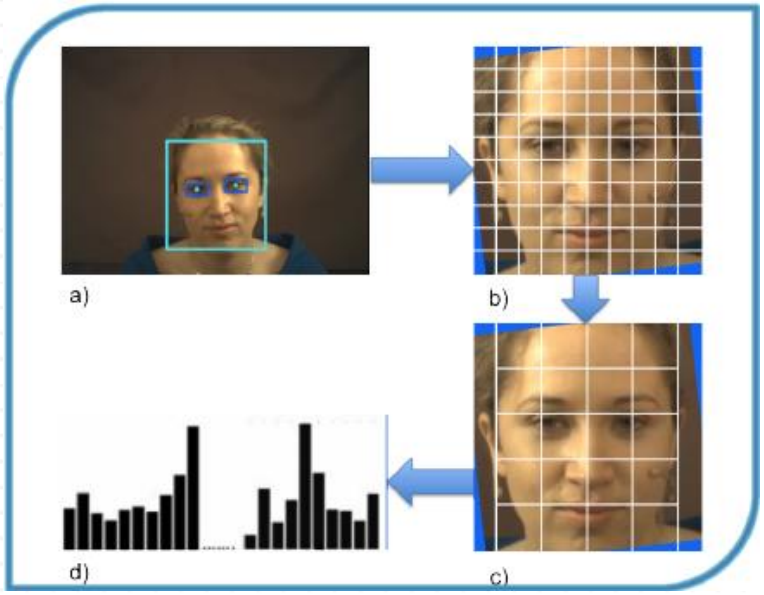
Disadvantages:

- Slow
- Lack of generality

iPython Examples

Feature extraction (FE)

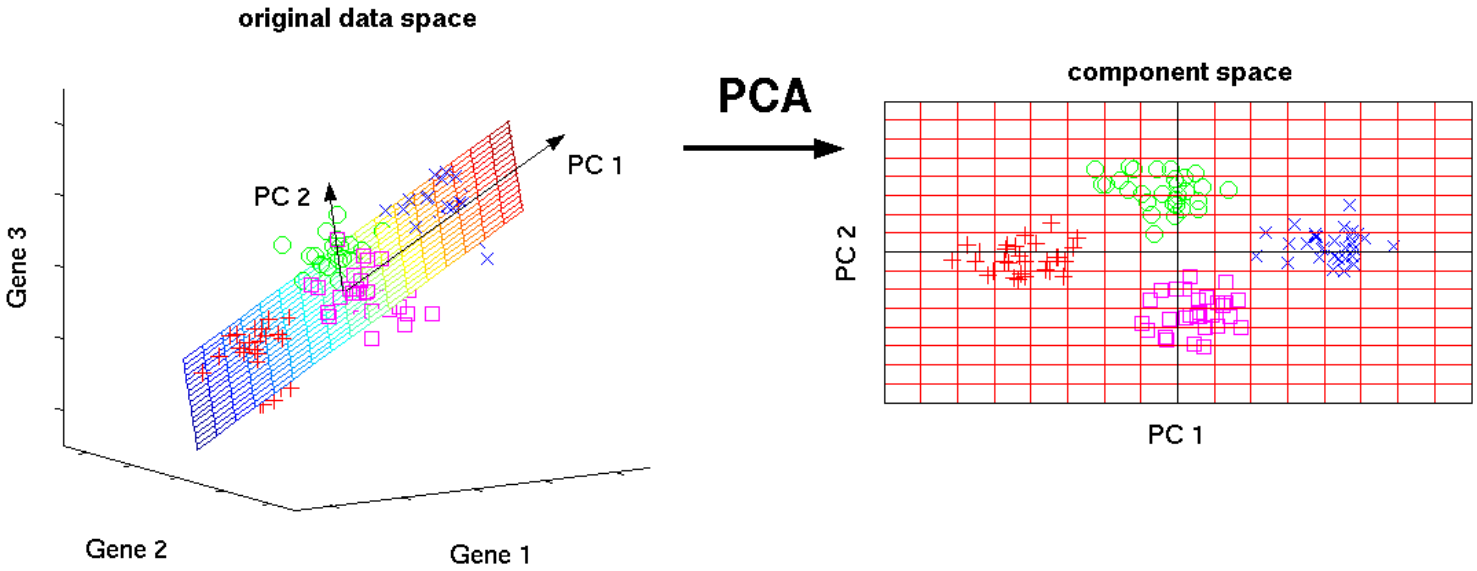
Transform the data in the high-dimensional space to a space of fewer dimensions



Create a new subset features that are representative of the original features.

Principal Component Analysis (PCA)

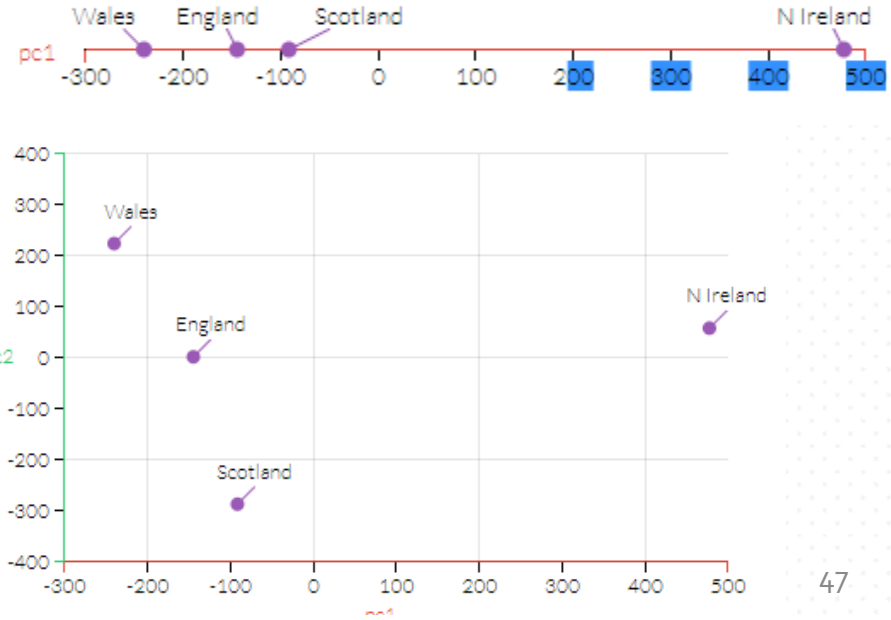
- PCA: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called ***principal components***
- The original data are projected onto a **much smaller space**, resulting in dimensionality reduction (e.g., n=3 to k=2)



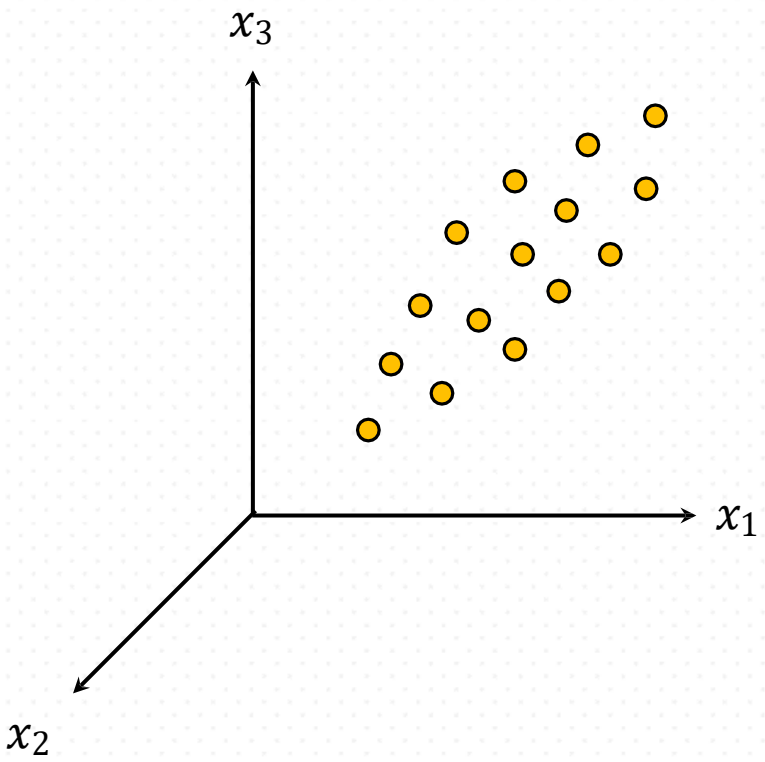
PCA Intuition

- Given a data set with N attributes
  - Can we create a set of new features that combine the existing attributes in such a way that we can highlight the differences between instances in less than N combinations.

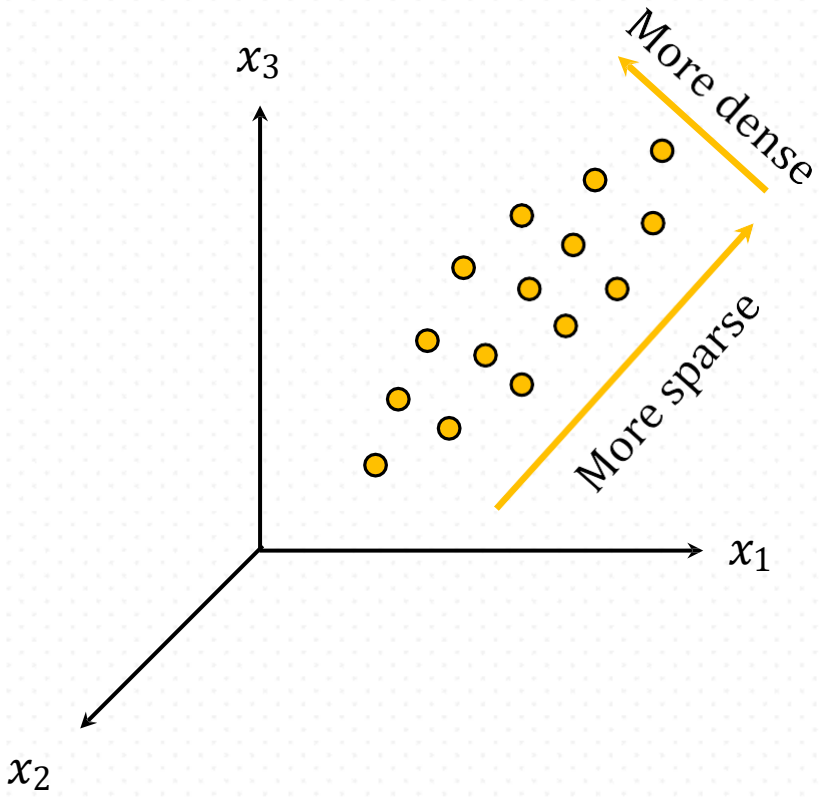
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



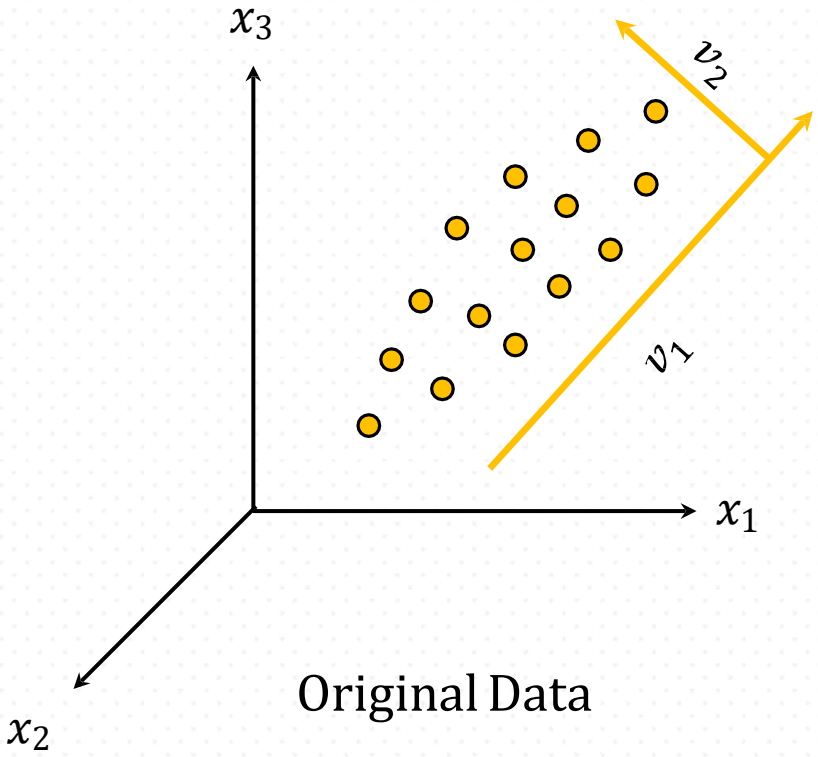
PCA: Visual Intuition



PCA: Visual Intuition



PCA: Visual Intuition



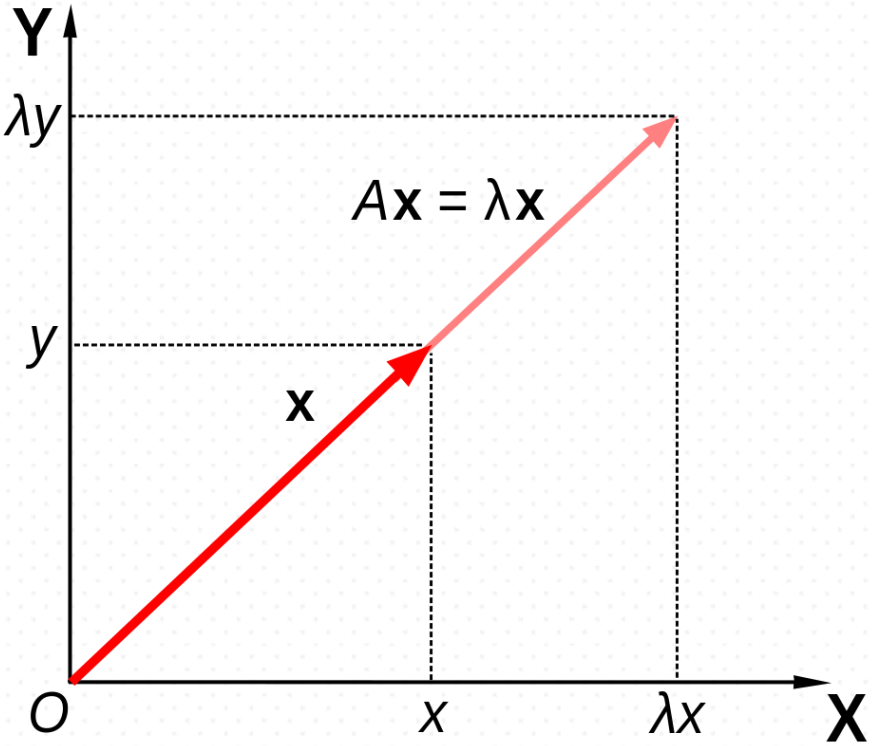
Original Data

- **V1**: Chosen to capture highest variance in the data
- **V2**: The next projection should maximize the remaining variance, while being orthogonal (perpendicular) to the previous projection(s).
  - If the projection is not orthogonal, it will be capturing the same variability already captured by previous projection(s).

How Do We Get There?

Eigenvectors

- For a square matrix **A** (n*n), find the eigenvector **x** (n*1).
  - **A** represents the linear transformation (from n to n)
- Matrix **A** acts by stretching the vector **x**, not changing its direction, so **x** is an eigenvector of **A**.



PCA and Eigenvectors (cont.)

- Method: Find the **eigenvectors of (square) matrix**, and these eigenvectors define the new space

$$\begin{aligned} \mathbf{Ax} = \lambda\mathbf{x} &\Leftrightarrow \mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow \mathbf{Ax} - \lambda\mathbf{Ix} = \mathbf{0} \\ &\Leftrightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \end{aligned}$$

The equation $\mathbf{Ax} = \lambda\mathbf{x}$ has nonzero solutions for the vector x if and only if the matrix $\mathbf{A} - \lambda\mathbf{I}$ has zero determinant.

Can We Use Raw Data?

Input Features: X
(attributes)

	x_1	x_2	..	x_m
1				
2				
.				
.				
n				

Data Objects (Instances)

We can't use this quite yet, why?

Looking Back- Covariance!

*Covariance Matrix – (**Square**)*

Combinations of all attributes (n-dimensional)

$$C = \begin{pmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) & \mathrm{cov}(y,z) \\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z) \end{pmatrix}$$

Eigenvalues

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

Eigenvalues

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

$$\begin{aligned} \mathbf{Ax} = \lambda\mathbf{x} &\Leftrightarrow \mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow \mathbf{Ax} - \lambda\mathbf{Ix} = \mathbf{0} \\ &\Leftrightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \end{aligned}$$

$$\det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \lambda I\right) = 0$$

Eigenvalues

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

$$\begin{aligned} \mathbf{Ax} = \lambda\mathbf{x} &\Leftrightarrow \mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow \mathbf{Ax} - \lambda\mathbf{Ix} = \mathbf{0} \\ &\Leftrightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \end{aligned}$$

$$\det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \lambda I\right) = 0$$

For a 2x2 Matrix, the Identity matrix can be replaced with

$$\det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$$

Eigenvalues

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

$$\det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$$

$$\det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = 0$$

$$\begin{vmatrix} (2-\lambda) & 2-0 \\ 5-0 & -1-\lambda \end{vmatrix}$$

$$\lambda^2 - \lambda - 12 = 0$$

Eigenvalues

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

$$\det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = 0$$

$$\begin{vmatrix} (2-\lambda) & 2-0 \\ 5-0 & -1-\lambda \end{vmatrix}$$

$$\lambda^2 - \lambda - 12 = 0$$

The eigenvalues of $\mathbf{A}$ are the solutions of the quadratic equation $\lambda^2 - \lambda - 12 = 0$, namely $\lambda_1 = -3$ and $\lambda_2 = 4$.

Ex. Eigenvalues 3x3

Example: Find the eigenvalues and associated eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} 7 & 0 & -3 \\ -9 & -2 & 3 \\ 18 & 0 & -8 \end{bmatrix}.$$

First we compute $\det(\mathbf{A} - \lambda\mathbf{I})$ via a cofactor expansion along the second column:

$$\begin{aligned} \begin{vmatrix} 7-\lambda & 0 & -3 \\ -9 & -2-\lambda & 3 \\ 18 & 0 & -8-\lambda \end{vmatrix} &= (-2-\lambda)(-1)^4 \begin{vmatrix} 7-\lambda & -3 \\ 18 & -8-\lambda \end{vmatrix} \\ &= -(2+\lambda)[(7-\lambda)(-8-\lambda)+54] \\ &= -(\lambda+2)(\lambda^2+\lambda-2) \\ &= -(\lambda+2)^2(\lambda-1). \end{aligned}$$

Thus $\mathbf{A}$ has two distinct eigenvalues, $\lambda_1 = -2$ and $\lambda_3 = 1$. (Note that we might say $\lambda_2 = -2$, since, as a root, -2 has multiplicity two. This is why we labelled the eigenvalue 1 as λ_3.)

Eigenvalues → Eigenvectors

First, we work with $\lambda = -3$. The equation $\mathbf{Ax} = \lambda\mathbf{x}$ becomes $\mathbf{Ax} = -3\mathbf{x}$. Writing

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and using the matrix $\mathbf{A}$ from above, we have

$$\mathbf{Ax} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix},$$

while

$$-3\mathbf{x} = \begin{bmatrix} -3x_1 \\ -3x_2 \end{bmatrix}.$$

Setting these equal, we get

$$\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} -3x_1 \\ -3x_2 \end{bmatrix} \Rightarrow \quad 2x_1 + 2x_2 = -3x_1 \quad \text{and} \quad 5x_1 - x_2 = -3x_2$$

$$\Rightarrow \quad 5x_1 = -2x_2$$

$$\Rightarrow \quad x_1 = -\frac{2}{5}x_2.$$

$$\mathbf{u_1} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}$$

Eigenvalues → Eigenvectors

Similarly, we can find eigenvectors associated with the eigenvalue $\lambda = 4$ by solving $\mathbf{Ax} = 4\mathbf{x}$:

$$\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} 4x_1 \\ 4x_2 \end{bmatrix} \Rightarrow 2x_1 + 2x_2 = 4x_1 \quad \text{and} \quad 5x_1 - x_2 = 4x_2$$
$$\Rightarrow x_1 = x_2.$$

Hence the set of eigenvectors associated with $\lambda = 4$ is spanned by

$$\mathbf{u_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Eigenvalues, Eigenvectors → PCA

Eigenvalues

$$e_{values} = \begin{bmatrix} -3 \\ 4 \end{bmatrix}$$

Eigenvectors

$$e_{vector} = \begin{bmatrix} 2 & -5 \\ 1 & 1 \end{bmatrix}$$

PCA components

Eigenvalues

$$e_{values} = \begin{bmatrix} -3 \\ 4 \end{bmatrix}$$

Eigenvectors

$$e_{vector} = \begin{bmatrix} 2 & -5 \\ 1 & 1 \end{bmatrix}$$

Sort Abs(e_{values})

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

PCA components

Eigenvalues

$$e_{values} = \begin{bmatrix} -3 \\ 4 \end{bmatrix}$$

Eigenvectors

$$e_{vector} = \begin{bmatrix} 2 & -5 \\ 1 & 1 \end{bmatrix}$$

Sort Abs(evalues)

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

Sort evector based on evalues

$$\begin{bmatrix} 1 & 1 \\ 2 & -5 \end{bmatrix}$$

Transform Data

Sorted Eigenvalues

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

Sorted Eigenvectors

$$\begin{bmatrix} 1 & 1 \\ 2 & -5 \end{bmatrix}$$

(Fake) Original Data

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

1st Principle Component

$$\begin{bmatrix} 4 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 \end{bmatrix}$$

Transform Data

Generalized

$$\begin{bmatrix} \text{Instance}_1 \\ \text{Instance}_2 \\ \text{Instance}_k \end{bmatrix} \bullet \begin{bmatrix} PC_1 & PC_2 & PC_n \end{bmatrix}$$

(Fake) Original Data

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$\bullet$

1st Principle Component

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

(Fake) Original Data

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

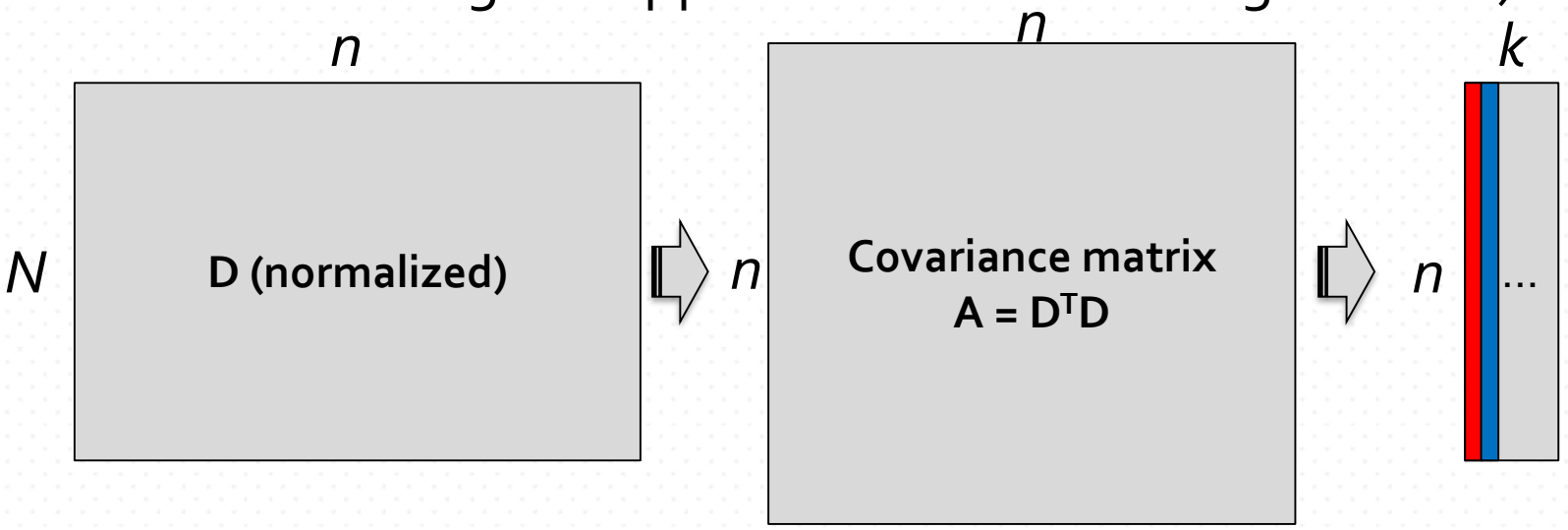
$\bullet$

1st and 2nd Principle Component

$$\begin{bmatrix} 1 & 2 \\ 1 & -5 \end{bmatrix}$$

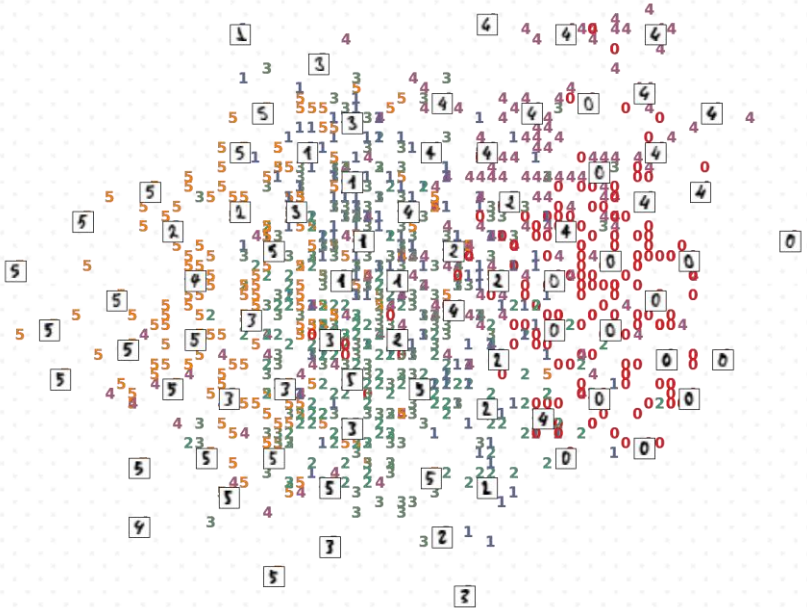
PCA and Eigenvectors

- For ***Square Matrix***: Data matrix to Covariance matrix
- The principal components are sorted in order of **decreasing "significance" or strength**
- **From n to k:** Since the components are sorted, the size of the data can be reduced by eliminating the weak components (i.e., using the strongest principal components, to reconstruct a good approximation of the original data)

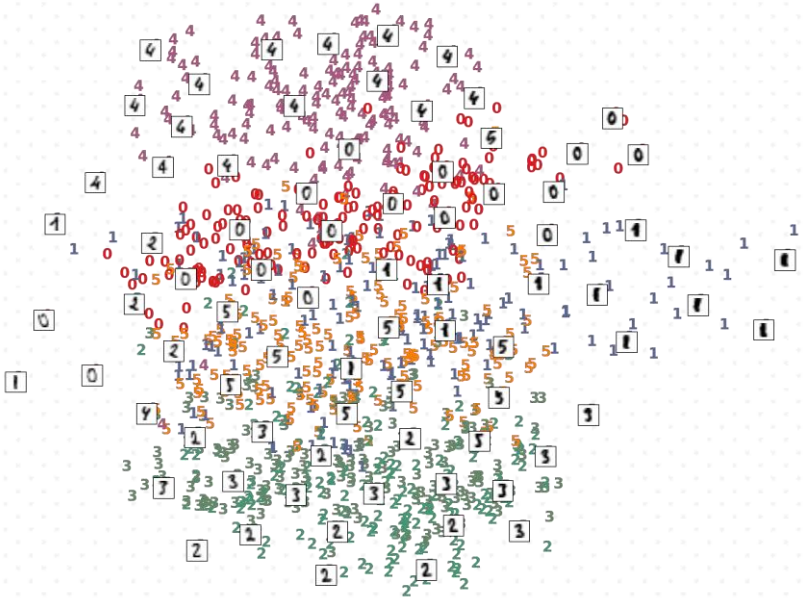


OptDigits Dataset: A Comparison

Random Projection (2D)



PCA Projection (2D)

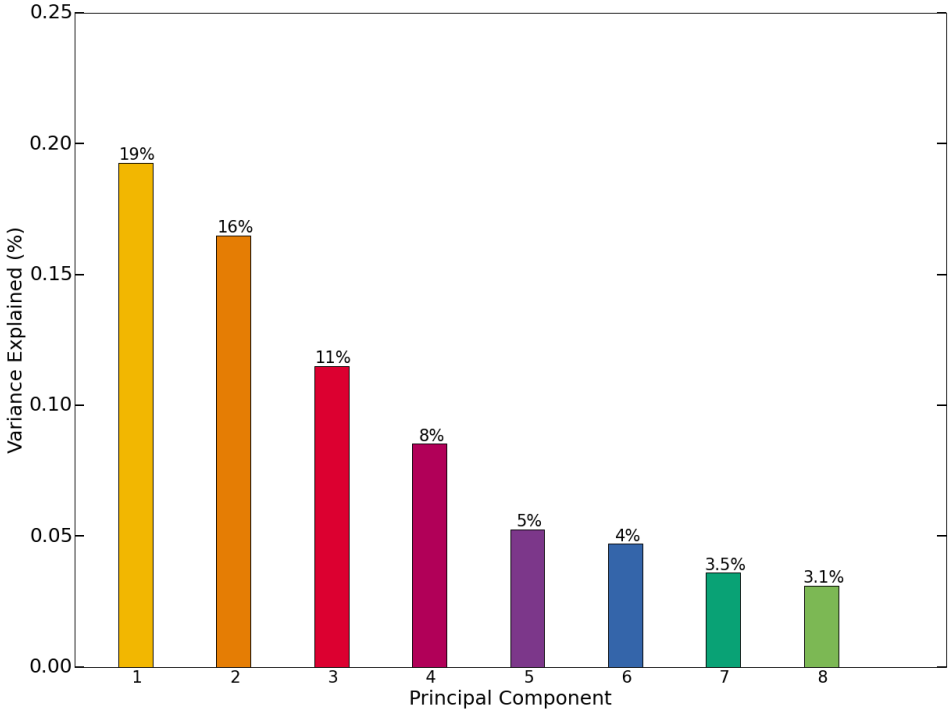


How Many Principal Components?

- If all of the principal components are kept, then there is no data reduction.
- Rather, the objective is to minimize the number of retained principal components.

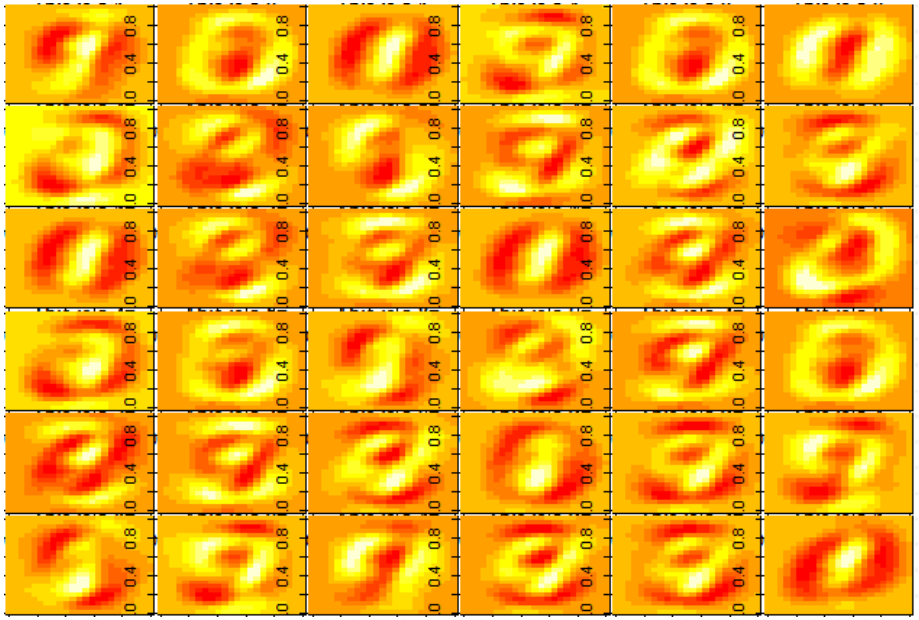
Scree Plots

- A Scree Plot is a simple line segment plot that shows the fraction of total variance in the data as explained or represented by each principal component.
- The point where there is a significant drop in the explained variance is sometimes called the "knee" or "elbow" point of the plot.

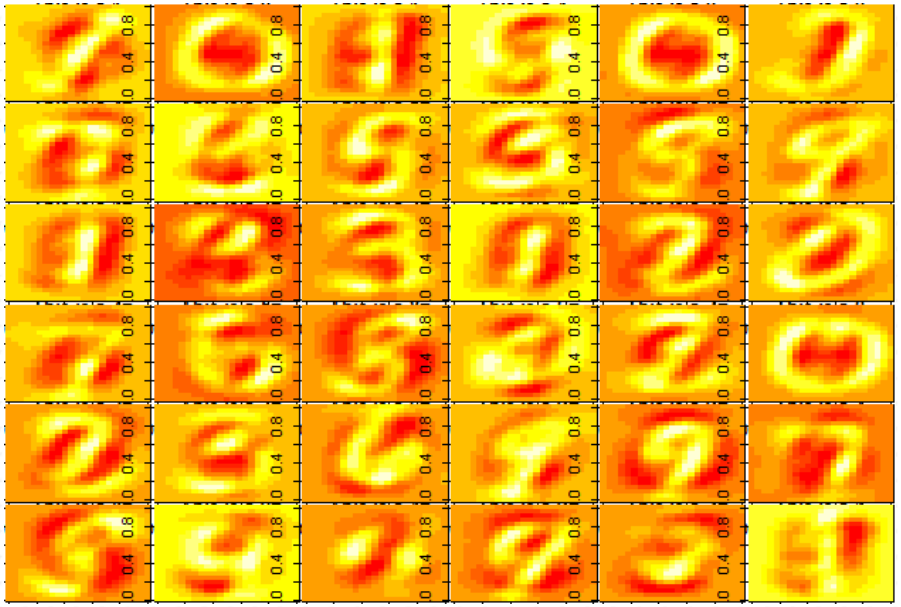


PCA

3 - PCA



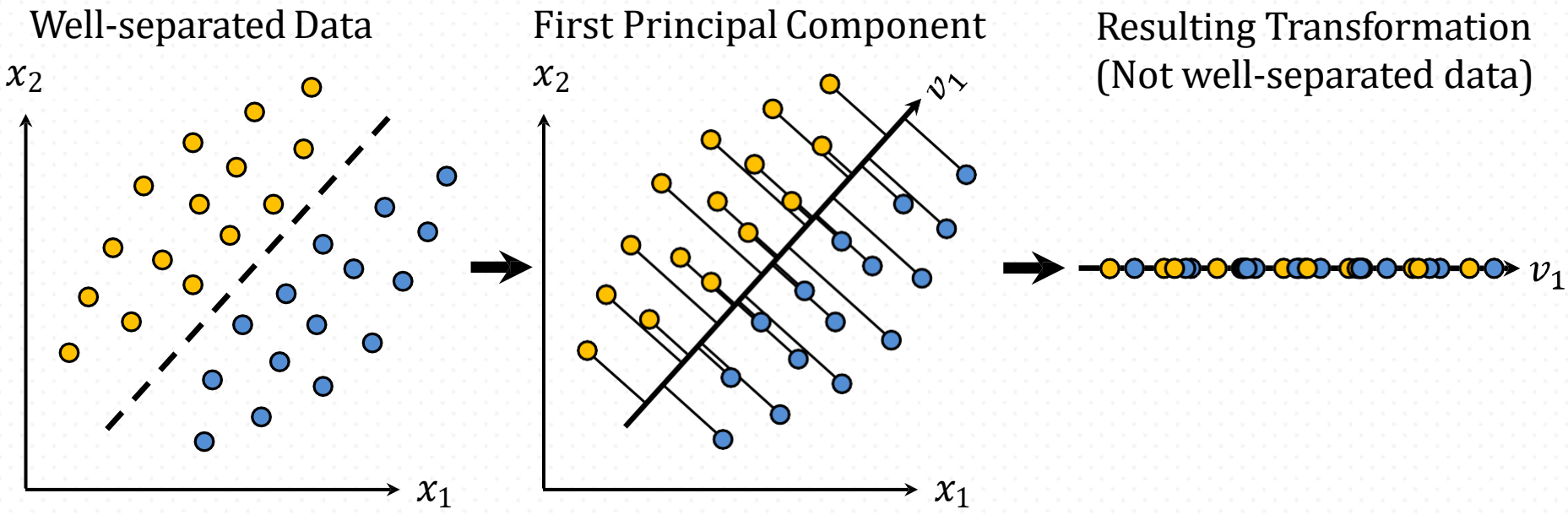
10 - PCA



iPython Examples

Limitations of PCA

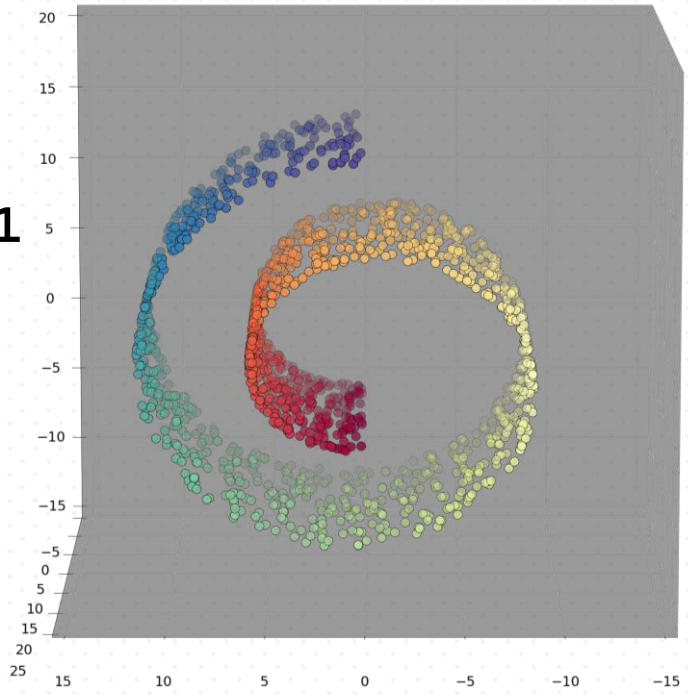
- PCA does not necessarily help segmenting or separating data.
- PCA generates principal components, which are linear combinations of the original features.
  - Non-linear structure may not be captured.



Limitations of PCA: Separating Data

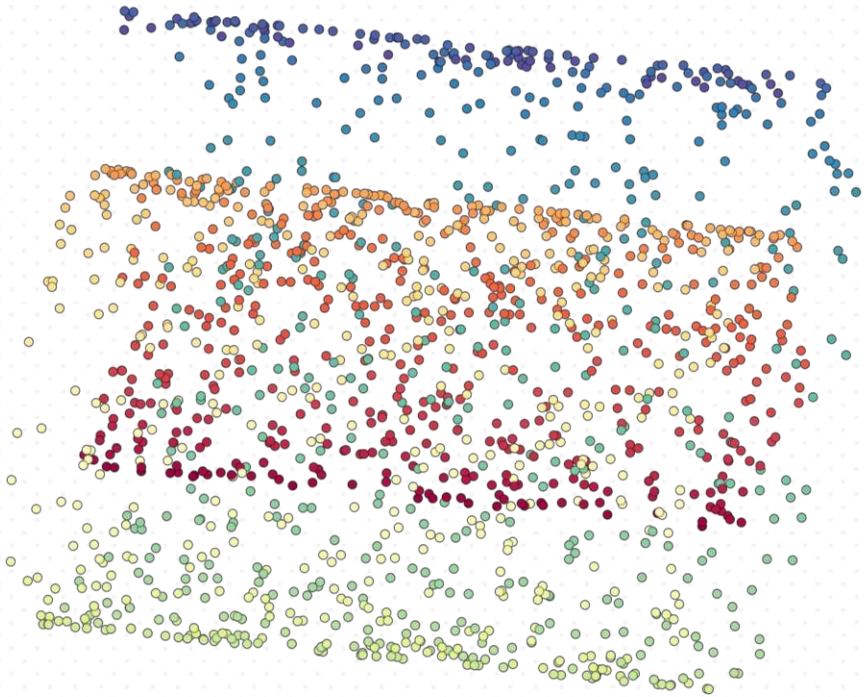
- 1,500 instances
- 3 features or dimensions
  - X is based on cosine function
  - Y is random real value 0 … 21
  - Z is based on sine function

Swiss Roll Data (3D)



Limitations of PCA: Separating Data

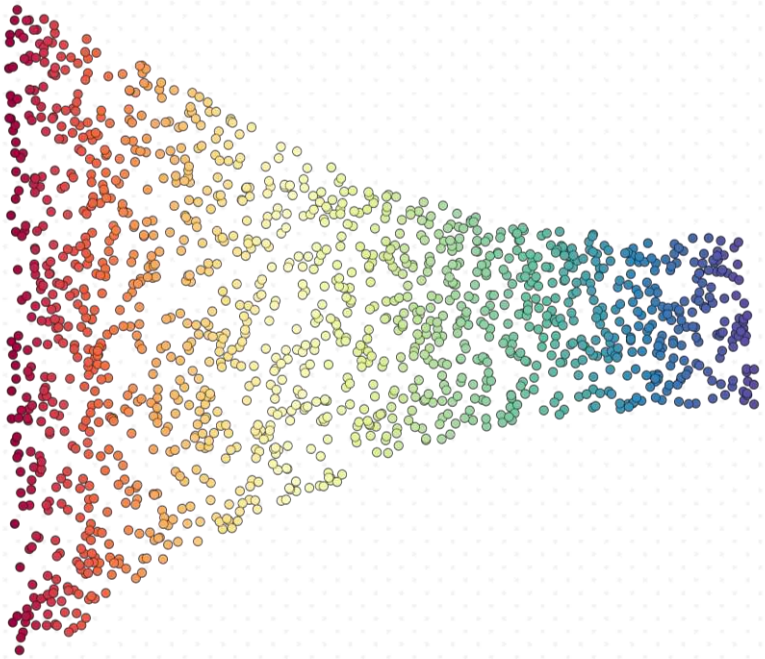
Swiss Roll Data Projection (2D)



- Notice that as a result of PCA, the projected data is not well-separated.
- The principal components maximize the variance, but not necessarily the separation of the data.

Other Methods: Local Linear Embedding (LLE)

Swiss Roll Data Projection (2D)



Reduce dimensionality by analyzing overlapping local neighborhoods to determine local structure:

1. Find the nearest neighbors of each data point.
2. Express each point x_i as a linear combination of the other points, i.e., $x_i = \sum_j w_{ij} x_j$, where $w_{ij} = 0$ if x_j is not a near neighbor of x_i.
3. Find the coordinates of each point in lower-dimensional space of specified dimension p by using weights found in step 2.

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Correlation analysis: Chi-Square test, Covariance
- **Data reduction and data transformation**
  - Normalization: Z-score normalization
- **Dimensionality reduction**
  - PCA, Heuristic Search in Attribute Selection

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4*
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995