A. Data Pre-Processing

Readings for Part A: blogs/videos,
see List on Canvas

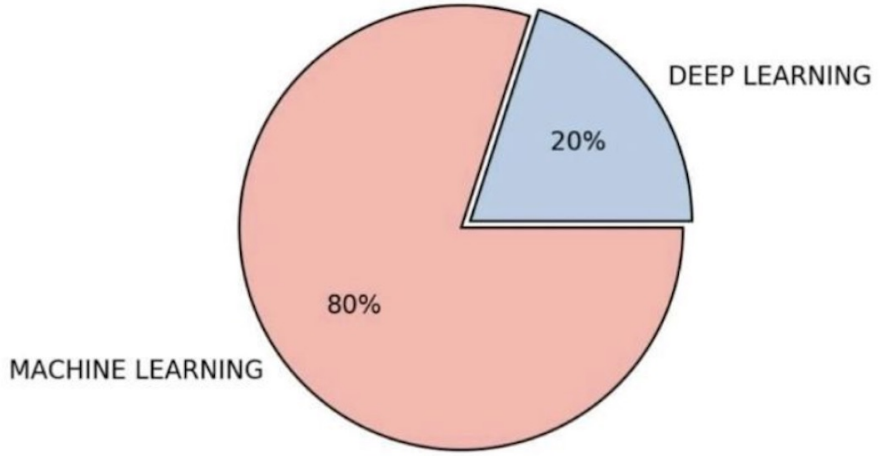
https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/ (Datatypes)
http://datascience.ibm.com/blog/missing-data-conundrum-exploration-and-imputation-techniques/ (Imputation)
https://www.youtube.com/watch?v=V0u6bxQOUJ8 (Outliers)
https://infoactive.co/data-design/ch11.html (Transformations)

B. Dimensionality Reduction

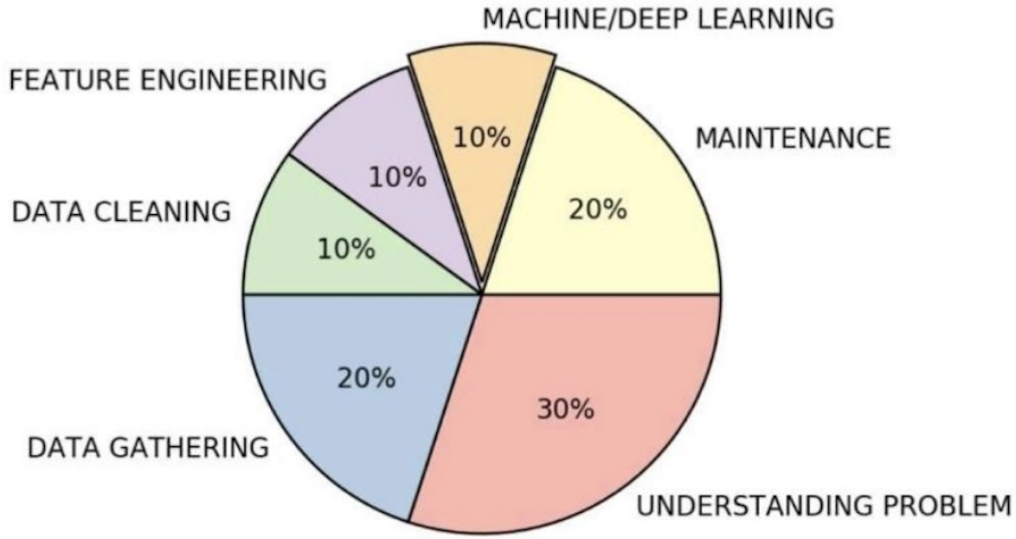
Readings for Part B: EA: 6-6.3, CB 12.1, KM 20.1

Data Scientist's Job

Expectation



Reality



Open sourcing Feathr – LinkedIn's feature store for productive machine learning


David Stein April 12, 2022

Share Tweet Share

We are open sourcing Feathr – the feature store we built to simplify machine learning (ML) feature management and improve developer productivity. At LinkedIn, dozens of applications use Feathr to define features, compute them for training, deploy them in production, and share them across teams. With Feathr, users reported significantly reduced time required to add new features to model training workflows and improved runtime performance compared to previous application-specific feature pipeline solutions.

The problem with scaling feature pipelines

At LinkedIn, we have hundreds of ML models running in applications like Search, Feed, and Ads. Our models are powered by thousands of features about entities in

IBM acquires data observability firm Databand.ai

Databand's data observability platform allows data engineers to tackle challenges associated with bad data at source.


By Anirban Ghoshal

Senior Writer, InfoWorld | JUL 6, 2022 8:43 AM PDT

Types of Data

- A data set is a collection of data objects/records
  - Each object is described by several features/attributes
- Data Types
  - Nominal
    - eye color, hobby
    - Binary: special case
  - Ordinal
    - rankings (e.g., taste of potato chips on a scale from 1-10), grades
  - Interval (numeric)
    - speed, temperature in Celsius

- Categorical: Nominal or Ordinal
- Others: ID, DATE, text/strings, graphs,…

Different data types often need different ways of handling/modeling.

e.g. Proportional odds model for ordinal regression.

Data Issues

- Quantity
- Quality and adequacy
- Acquiring ”labels”
- Big Data Issues

Part A: Why Preprocess Data

- **GIGO!**
  - data may be incomplete, inconsistent, noisy; have outliers, or simply too large
- **Why is data dirty?**
  - Incomplete data may come from
    - Not available or "Not applicable" data value when collected
    - Thoughtless entry (e.g. 0 vs. missing)
  - Noisy data (incorrect values) may come from
    - Faulty data collection instruments
    - Human or computer error at data entry
      - HEB, shoulder surgery, Nursing Homes, ..
    - Out-of-date
  - Inconsistent data may come from
    - Different data sources; formats
    - Inconsistent rules e.g. hotel price on phone vs. internet
  - Duplicate records need to be eliminated

Major Preprocessing Steps(Exploratory Data Analysis)

1) Data cleaning; sanity checks, consistency (already done by ETL tools if data is from warehouse)
2) Exploratory Data Analysis (Often based on a sample)
    1) Fill missing values, remove noise and outliers
    2) transformation/scaling
3) Data reduction
    1) Of records (sampling)
    2) Of attributes (feature selection/extraction)
4) Visualization

- Often takes over 90% of a project's time!
- steps 2-4 often revisited after modeling.
- Repetitive: use pipelines (in scikit; also pandas has pdpipe)

Ref: A Comprehensive Guide to Data Exploration

Pre-Processing in Scikit

- **.3. Preprocessing data**5.3.1. Standardization, or mean removal and variance scaling
  - 5.3.1.1. Scaling features to a range
  - 5.3.1.2. Scaling sparse data
  - 5.3.1.3. Scaling data with outliers
  - 5.3.1.4. Centering kernel matrices
- 5.3.2. Non-linear transformation
  - 5.3.2.1. Mapping to a Uniform distribution
  - 5.3.2.2. Mapping to a Gaussian distribution
- 5.3.3. Normalization
- 5.3.4. Encoding categorical features
- 5.3.5. Discretization
  - 5.3.5.1. K-bins discretization
  - 5.3.5.2. Feature binarization
- 5.3.6. Imputation of missing values
- 5.3.7. Generating polynomial features
- 5.3.8. Custom transformers

- Step 0: see/visualize univariate and bivariate statistics

Dealing with Missing Values (Imputation)

- Missing Completely at Random (MCAR)?
  - Vs. "informative missingness" (e.g doctor's choices)

- ignore record or attribute (often missing values are concentrated in a few instances or attributes)
- Fill in missing values
  - fill with constant, mean or mode
  - conditional mean/ mode
    - Condition on values of a set of related variables
    - Use K-NN
- Use a derived variable, e.g. missing or not

- READ: KM ch. 1.5.5

Data Cleaning I: Dealing with Missing Values (Imputation)

Updated 11m ago

93% REPORTED

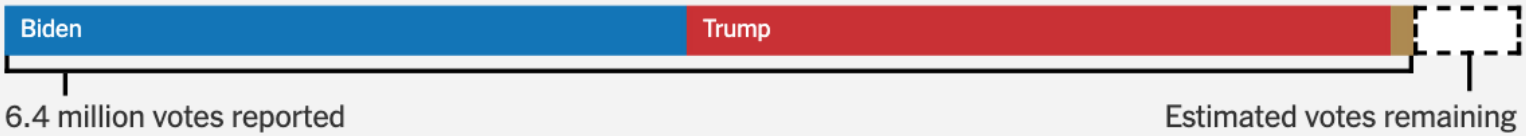
Candidate	Party	Votes	Pct.
 Donald J. Trump*	Republican	3,247,802	**50.1%**
 Joseph R. Biden Jr.	Democrat	3,157,260	**48.7**
Total reported		6,487,875	

+ View all candidates

* Incumbent

93% of the estimated vote total has been reported.

We received 9,526 additional votes from Pennsylvania. Trump is currently up by 90,542 votes. 11m ago



Timing of results: The remaining mail ballots are expected to favor Democrats. Officials have said they expect most votes will be counted by Friday, Nov. 6.

Data Cleaning I: Dealing with Missing Values (Imputation)

County	Margin	2016 margin	▲ Est. votes reported	Total votes	Absentee
Cumberland	**Trump +25**	R+17.8	77%	109,272	—
Mercer	**Trump +38**	R+24.2	78%	47,245	—
Monroe	**Trump +3**	D+0.8	81%	65,327	—
Lehigh	**Biden +3**	D+4.7	84%	157,262	—
Delaware	**Biden +24**	D+22.3	85%	294,896	74,110
Philadelphia	**Biden +61**	D+67	86%	617,696	75,220
Centre	**Biden +4**	D+2.3	86%	76,156	—
Crawford	**Trump +44**	R+37.1	87%	37,007	—

Data Transformation

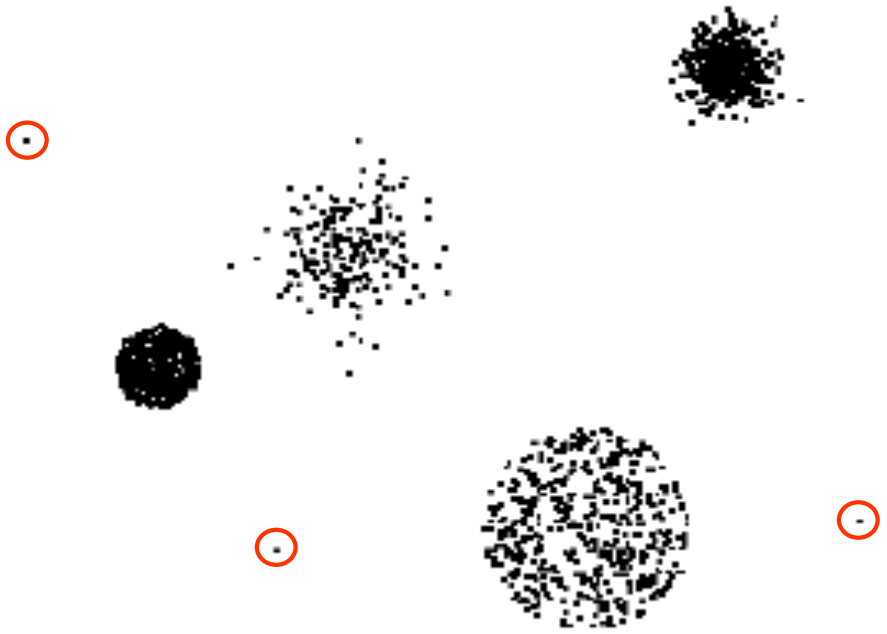
- Scaling

  Normalization by Linear scaling
  - Linear [min, max] → [0,1]
  - Centering (e.g. Z-scoring: Normal/Gaussian → N (0, 1))

- (non-linear) transformation, e.g. to reduce skew or to show a simpler relationship between x and y (for example a power law shows up as a linear relationship in the log space).
  - Log;
  - square;
  - exponential

Cleaning II: Handling Outliers

- Outliers are data objects with characteristics that are considerably different than the vast majority of the other data objects in the data set



Dealing with Outliers in "X"

- Probability based (classical):
  - Estimate pdf of X, using e.g. Parzen windows or mixture of Gaussians
  - Identify low p(x) points

- Discrimination based
  - Rule based, e.g.
    - » less than 1% for categorical variables
    - » Outside 3 sigma for gaussian looking numeric variables
  - Distance based: see if outlier score is > threshold or not
    - » Score could be av. Distance of k-nearest neighbors; distance to the kth neighbor, etc.

Outliers in Y (robust statistics)

Identify outliers and eliminate before applying model

OR

Use models that are little affected by presence of a few outliers

- trimmed means instead of means

- alternatives to "squared error" loss functions
  - e.g. Huber's loss (quad → linear)

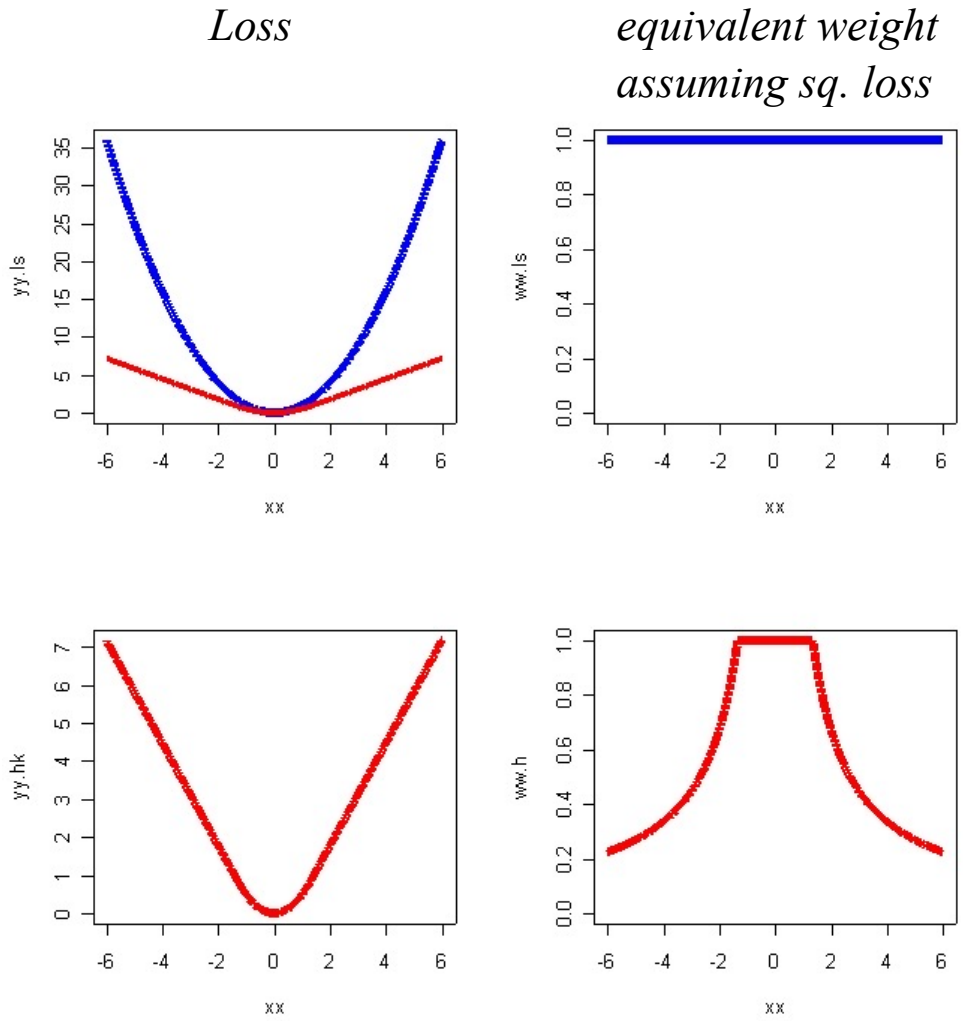


Fig: Plotted as a function of residual ($r = y - \hat{y}$):

Blue: Sq. error loss (left) and "equivalent" weights (right)

Red: Huber loss and equivalent weights if Sq. loss was used (right)

Python Outlier Detection (PyOD)

- PyOD is a comprehensive and scalable **Python toolkit** for **detecting outlying objects** in multivariate data.
- Various dedicated posts/tutorials, including Analytics Vidhya, KDnuggets, Towards Data Science, Computer Vision News, and awesome-machine-learning
- Focus on ensemble methods.

See Table of Implemented Algorithms and Comparison Figure

*** A Survey of Label-noise Representation Learning: Past, Present and Future (2021)) – deep nets oriented.**

Part B: Data Reduction Methods

- Reducing # of rows
- Reducing # of columns
- Reducing "resolution"

Data Reduction Methods

- **Why?**
  - get quicker answers
  - Reducing number of features may (substantially) improve results !!
    - Reduces **"curse-of-dimensionality"**
      - When dimensionality increases, (randomly distributed) data becomes increasingly sparse in the space that it occupies
        - » Problematic for many types of analysis.
    - Collinearity a problem with MLR
      - Tools, e.g. compute all pairwise correlations ("pairs" in R)
      - Heuristics, e.g. eliminate variables till max pairwise correlation < threshold

How to Reduce Data

- Reduce # of records or instances (sampling: see backups)
- Reduce # of attributes or features
- Aggregate (in data cube)
- Reduce resolution of an attribute e.g. discretization of interval variable.

- Note: Data reduction technique will affect quality as well as speed.

Reducing # of (Derived) Attributes/Features (AKA Dimensionality Reduction)

- Feature selection (select a subset of original features)

vs

Feature extraction, aka "feature engineering" (use derived features, not original ones)

Why is feature selection sometimes preferred to feature extraction though it is only a special case?

Read: **EA: 6-6.3**, KM 20.1, B 12.1

Selecting a Subset of Features

NP-complete, so use heuristics

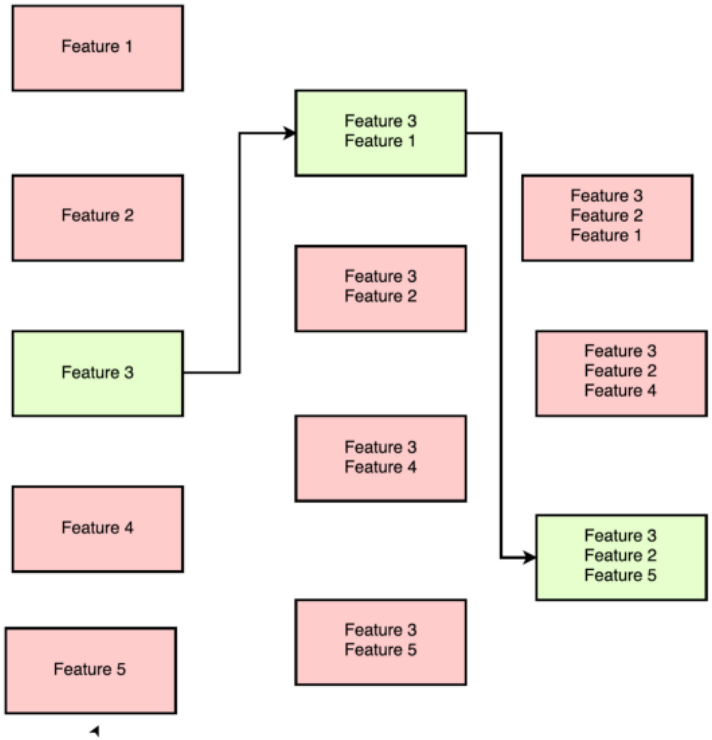
- use intrinsic quality measure (e.g. correlation) with target to score and rank order features
  - No prediction model learnt before features are decided
- (extrinsic evaluation)
  - Greedily evaluate candidate subsets using predictive models
  - Search strategy for candidate sets to evaluate:
    - Forward inclusion
    - Backward elimination (aka recursive feature elimination)
    - Stepwise (forward, but may remove predictors that no longer meet criterion)
- feature selection built into model training (e.g LASSO, decision trees)

Reading: **Feature Selection Methods in Machine Learning**

Extrinsic Search for Features

Search strategy for candidate feature sets to evaluate:

1. Forward inclusion (evaluate each box with a model trained on selected features)



From **Feature Selection Methods in Machine Learning**

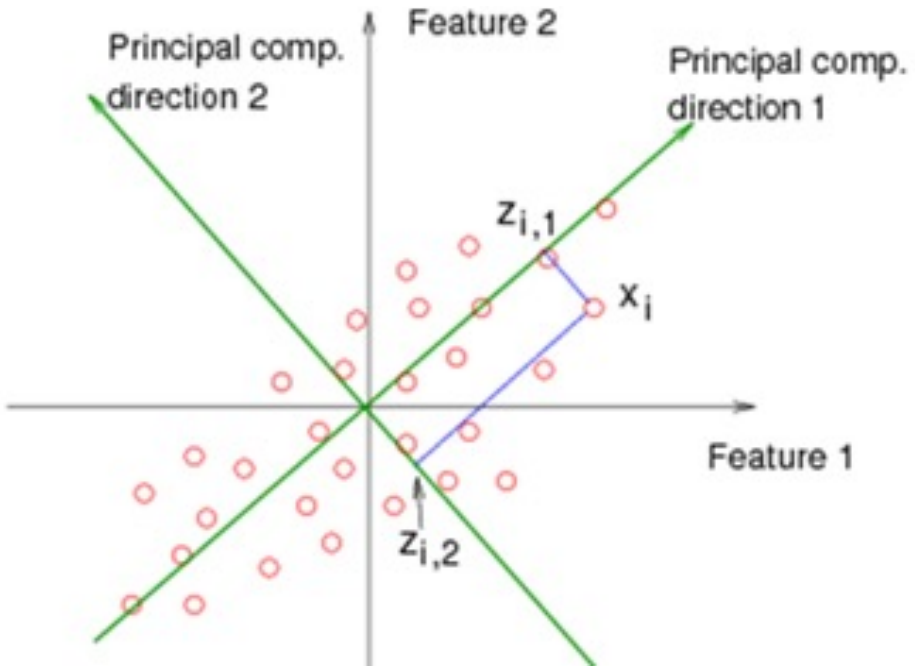
2. Backward elimination (aka recursive feature elimination)
3. Stepwise (forward, but may remove predictors that no longer meet criterion)

Feature Extraction Choices

- Linear ("projections")
  - Unsupervised : PCA
  - Supervised:
    - Fisher's Linear Discriminant (classification)
    - Canonical Correlation (regression)
- Non-Linear
  - Unsupervised : e.g., Principal Curves, T-SNE,…
  - Supervised: Nonlinear discriminant analysis, e.g. using a multi-layered perceptron

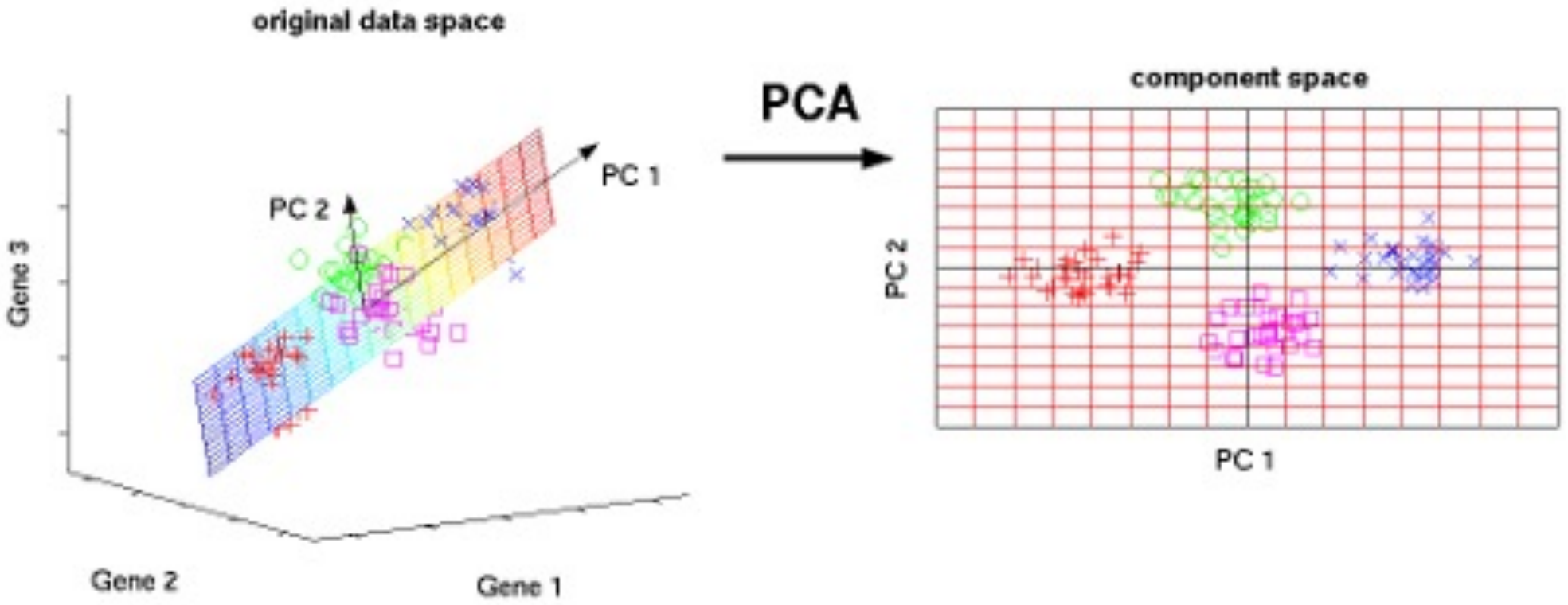
PCA

- Principal Components Analysis:
  - Reduce dimensions while (lossily) retaining info about original data
    - PCA finds the best "subspace" that "retains" as much data variance as possible
  - optimal linear projection/reconstruction in MSE sense
    - MSE of best (linear) reconstruction) = amount of variance lost



Another Example

From http://www.nlpca.org/pca_principal_component_analysis.html

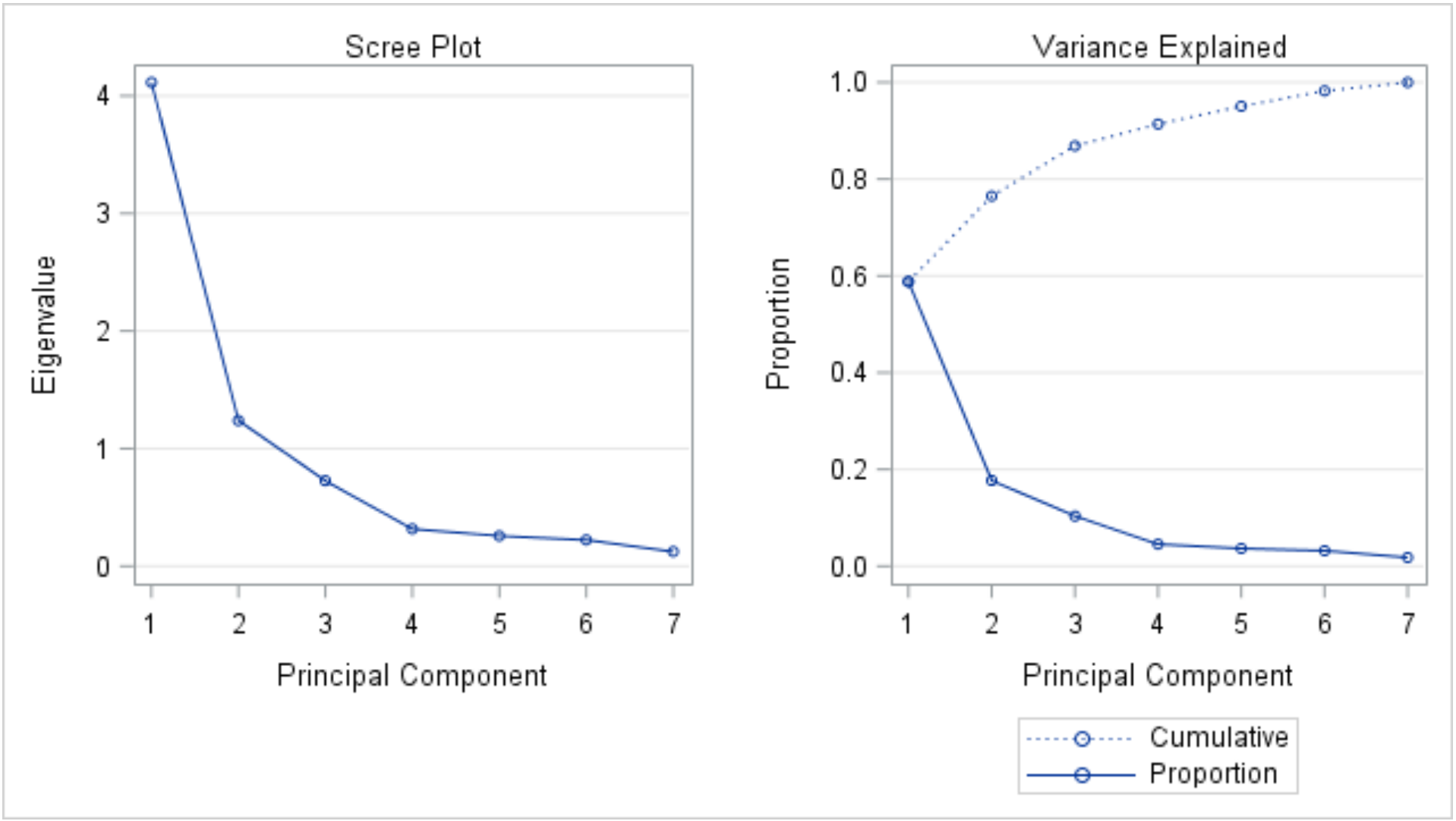


Visualize

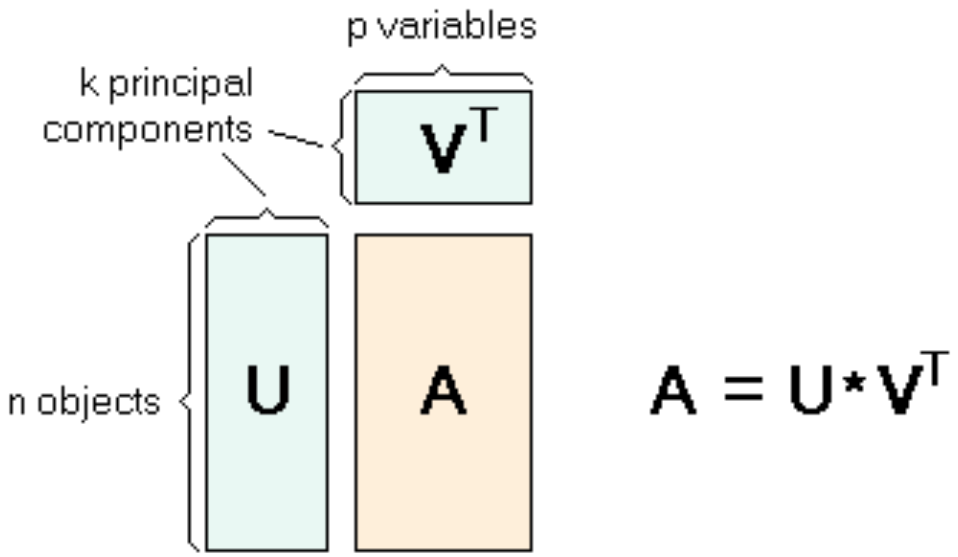
- http://setosa.io/ev/principal-component-analysis/

- http://setosa.io/ev/eigenvectors-and-eigenvalues/

PCs sorted by amount of variance explained **(Scree plot)**



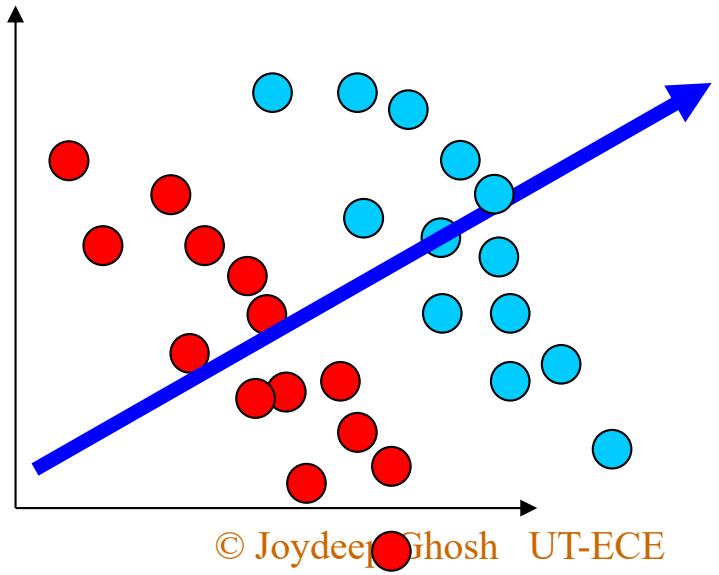
PCA: Loadings & Scores*



V: matrix of loadings: defining the directions of the PCs
U: matrix scores: recording projections of each data point along the different PCs.

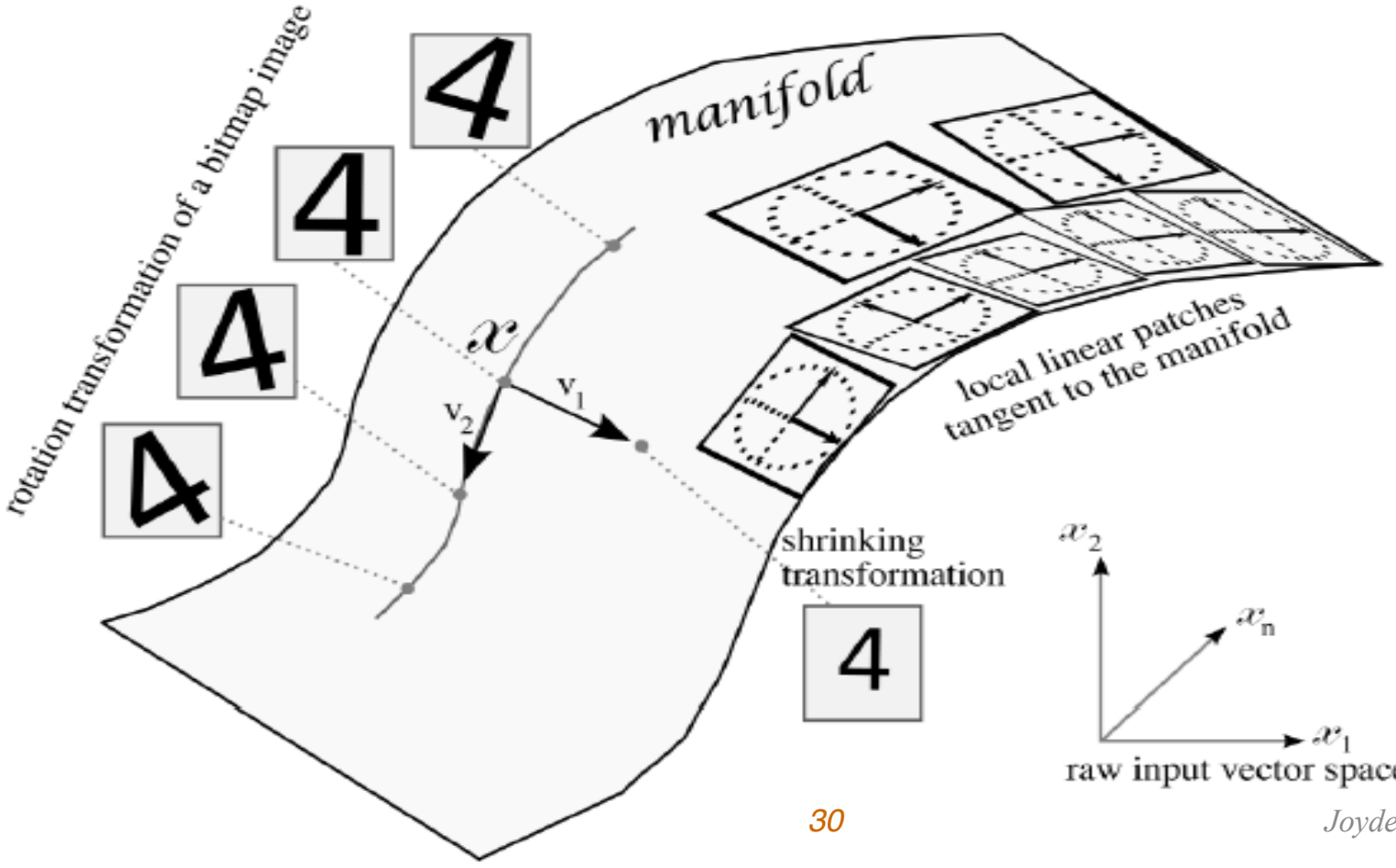
Linear Supervised Method:
Fisher's Linear Discriminant (FLD)

- FLD finds the projection direction that best separates the two classes
- Multiple discriminant analysis (MDA) extends LDA to multiple classes
- For fun: Fisherfaces vs. Eigenfaces https://www.youtube.com/watch?v=x8W_htbct3U (David Mumford at 6:30)



Best projection direction for classification

Handwritten Digits lie near a low-D manifold (why consider non-linear methods?)



Non-Linear Embeddings for Visualization

- Map high-D data into 2 or 3-D so that local properties are "preserved"
  - Kohonen's Self Organizing Feature Map (SOFM). 1988
  - Generative Topological Mapping (GTM). 1997
  - t-distributed stochastic neighbor embedding (t-SNE). 2007.

  - *Visualize: t-SNE and its discontents.
- "Conclusion
- There's a reason that t-SNE has become so popular: it's incredibly flexible, and can often find structure where other dimensionality-reduction algorithms cannot. Unfortunately, that very flexibility makes it tricky to interpret. Out of sight from the user, the algorithm makes all sorts of adjustments that tidy up its visualizations. Don't let the hidden "magic" scare you away from the whole technique, though. The good news is that by studying how t-SNE behaves in simple cases, it's possible to develop an intuition for what's going on."

Tensor Board Visualization

- PCA and t-SNE for MNIST (using Tensorboard's "embedding visualizer")
  - 28x28 images (784 dimensional vectors)

- https://www.youtube.com/watch?v=eBbEDRsCmv4 19:11 onwards
  (Part of Tensorflow dev summit 2017)

Representation Learning*
(popular deep learning approach)

---- is Feature Extraction!

(non-linear embedding to transform to a more useful feature space)

- Auto-encoders
- Language embeddings
- Self-supervised learning

- *What is the relation between auto-encoders and PCA?

Pipelines in Scikit

- Scikit-learn preprocessing package viewed as a list of transforms
- Pipelines assemble a subset of transforms needed for a specific solution
- A Simple Guide to Scikit-learn Pipelines
  - Stated Benefits:
    - They make your workflow much easier to read and understand.
    - They enforce the implementation and ordering of steps in your project.
    - These in turn make your work much more reproducible.

Caution: Validate Data and Results!

- **Bonferroni's Theorem**: if there are too many possible conclusions to draw, some will be true for purely statistical reasons, with no physical validity

- If possible, see that the entries make sense and data was collected properly
  - Ex: milk study at Lanarkshire, Scotland
- Data is often observational and not experimental

- Results validation vs. data dredging, snooping, fishing
  - E.g. S&P index almost perfectly predicted by butter, cheese production and sheep population in US and Bangladesh
  - "parapsychologist" David Rhine found (1950's) found about .1% guessed all 10 card colors correctly, but failed in next round.
    - Concluded that "telling people they have ESP causes them to lose it"!

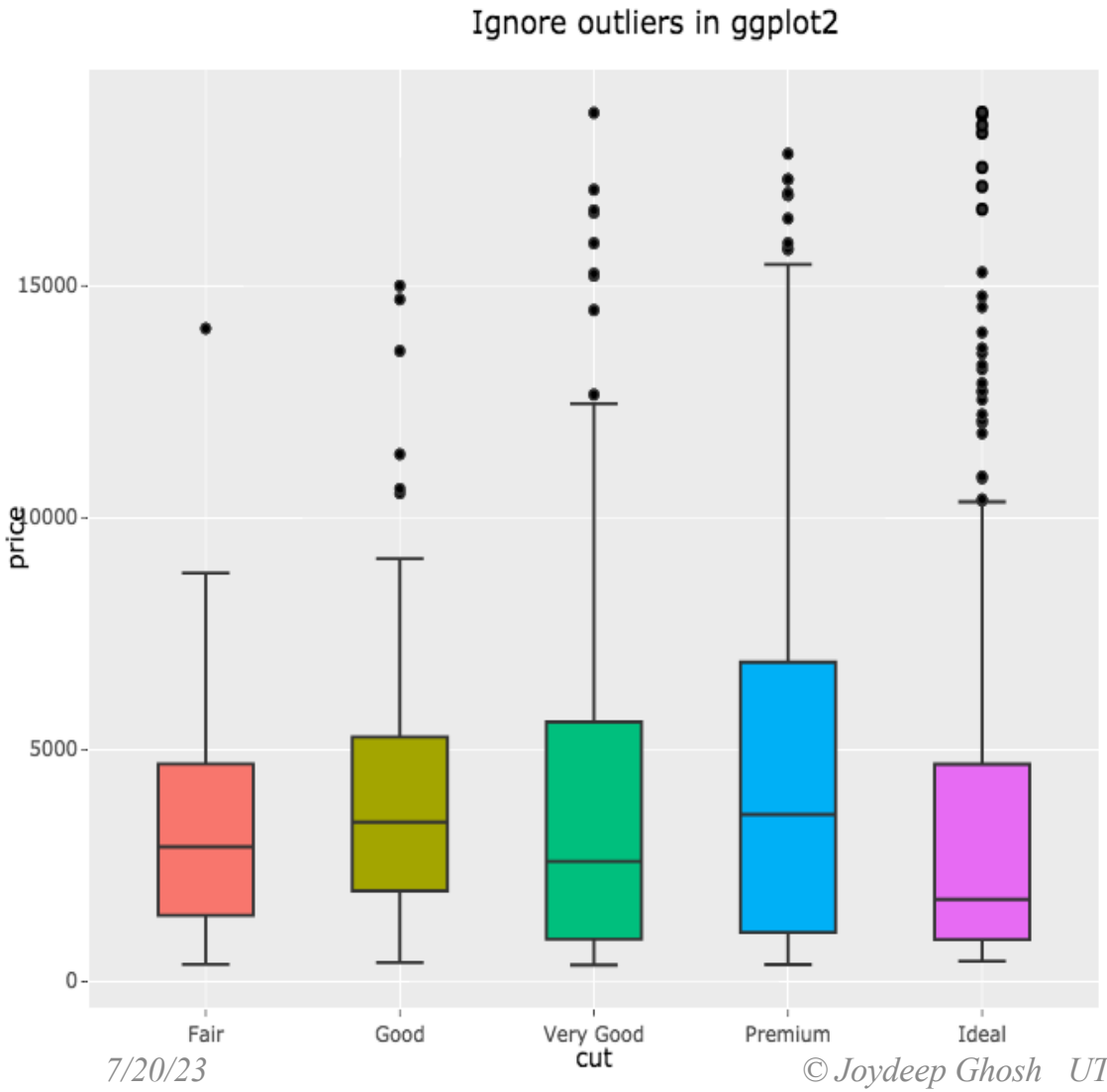
    - www.tylervigen.com

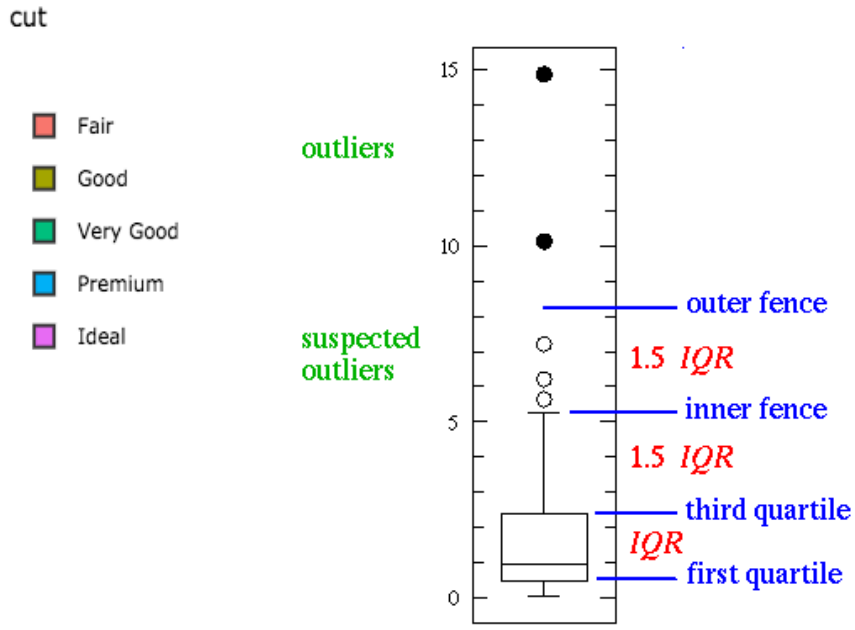
Extras

Before You Clean the Data..

- .. Do a quick summarization/visualization
  - Single "input" variable summaries
    - Variable type, mean, range, %missing, skewness, histograms, boxplots, ….
  - Bivariate (X_i vs. Y or X_i vs. X_k) visuals
    - (scatter plots, correlation,..)

BoxPlot

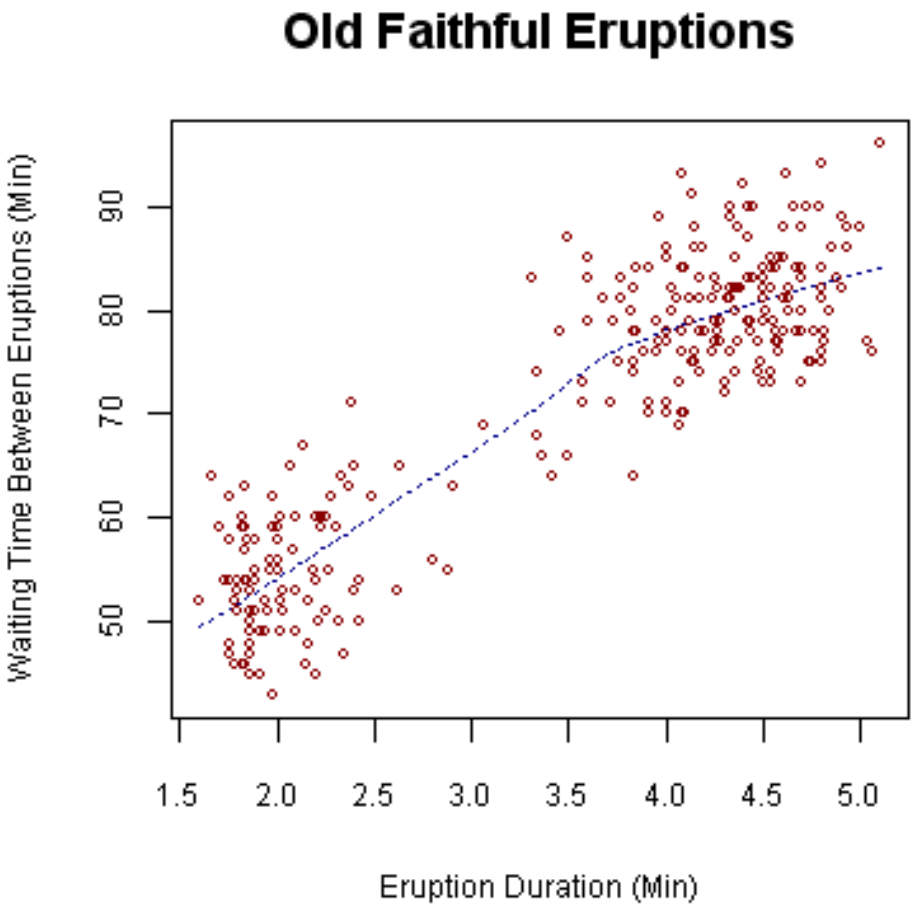




Using Plotly (interactive)
https://plot.ly/ggplot2/box-plots/

Scatterplot

- Old Faithful Example from Wikipedia



Importance of Good Data

3:18 PM Fri Jun 5

Headlines

CNN Live TV

Two coronavirus studies retracted after questions emerge about data

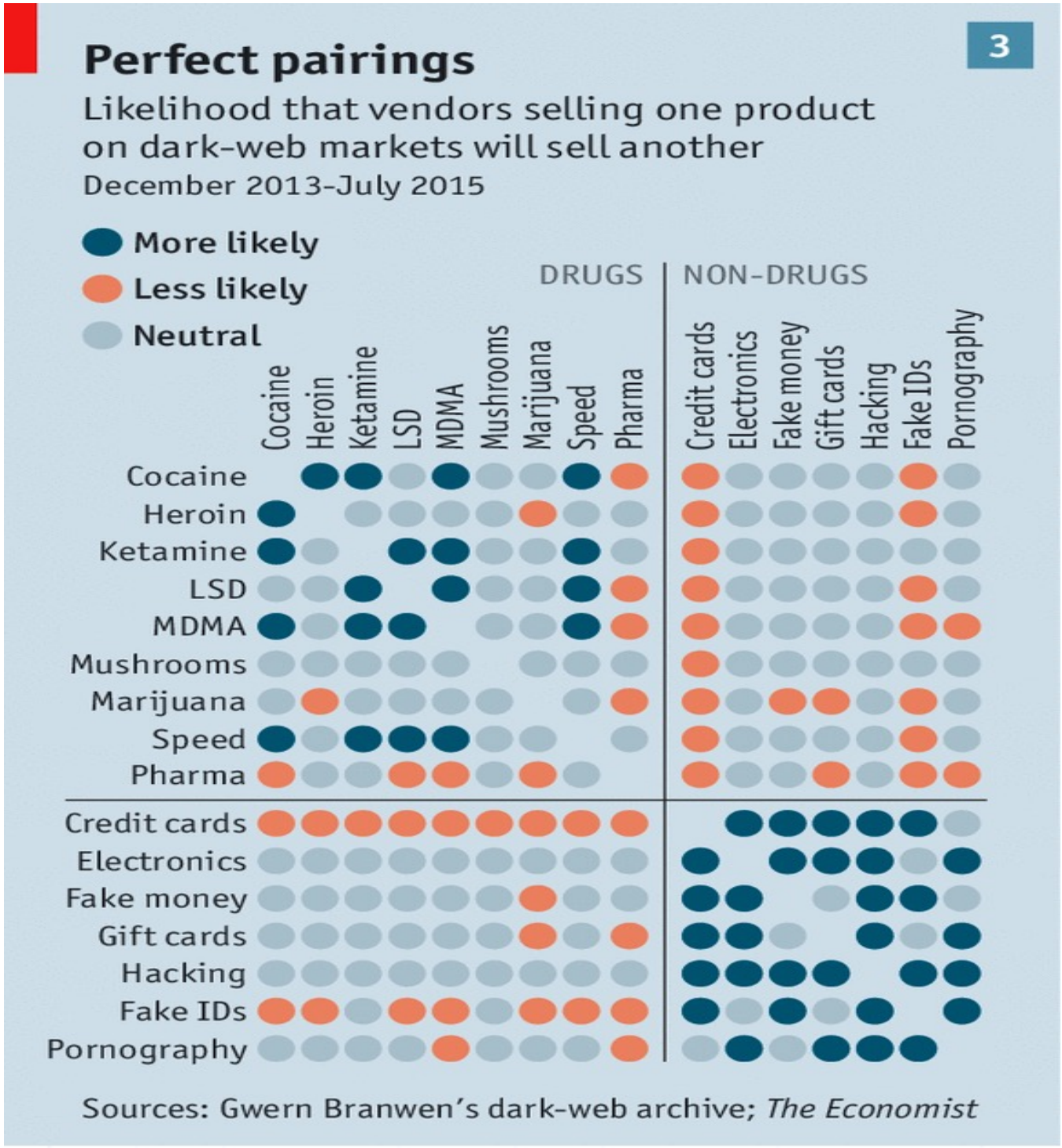
By Jamie Gumbrecht and Maggie Fox, CNN

Updated 7:16 PM EDT, Thu June 04, 2020

(CNN) — Two influential medical journals retracted separate coronavirus studies Thursday over concerns about the data used in both studies -- data that came from the same international registry.

The authors of the studies, one published in The Lancet and another in The New England Journal of Medicine, requested the studies be retracted because independent auditors weren't able to access all the information needed to verify the data. Both studies used data from data analytics company Surgisphere Corporation.

The retracted Lancet study, published May 22, found Covid-19 patients treated with hydroxychloroquine and chloroquine were more likely to die or suffer dangerous side effects.


Perfect pairings
3
Likelihood that vendors selling one product on dark-web markets will sell another
December 2013-July 2015
More likely
Less likely
Neutral
DRUGS
NON-DRUGS
Cocaine
Heroin
Ketamine
LSD
MDMA
Mushrooms
Marijuana
Speed
Pharma
Credit cards
Electronics
Fake money
Gift cards
Hacking
Fake IDs
Pornography
Sources: Gwern Branwen's dark-web archive; The Economist

Sampling

A recent Texas Public Employees Association (TPEA) survey found that 11.7 percent of state employee households received public assistance in the past year. More than 16,000 state employees responded to our survey, and because our sample size was so large, our results can be considered representative of all general state government — approximately 149,000 employees — with a 99 percent confidence level and a 1 percent margin of error.

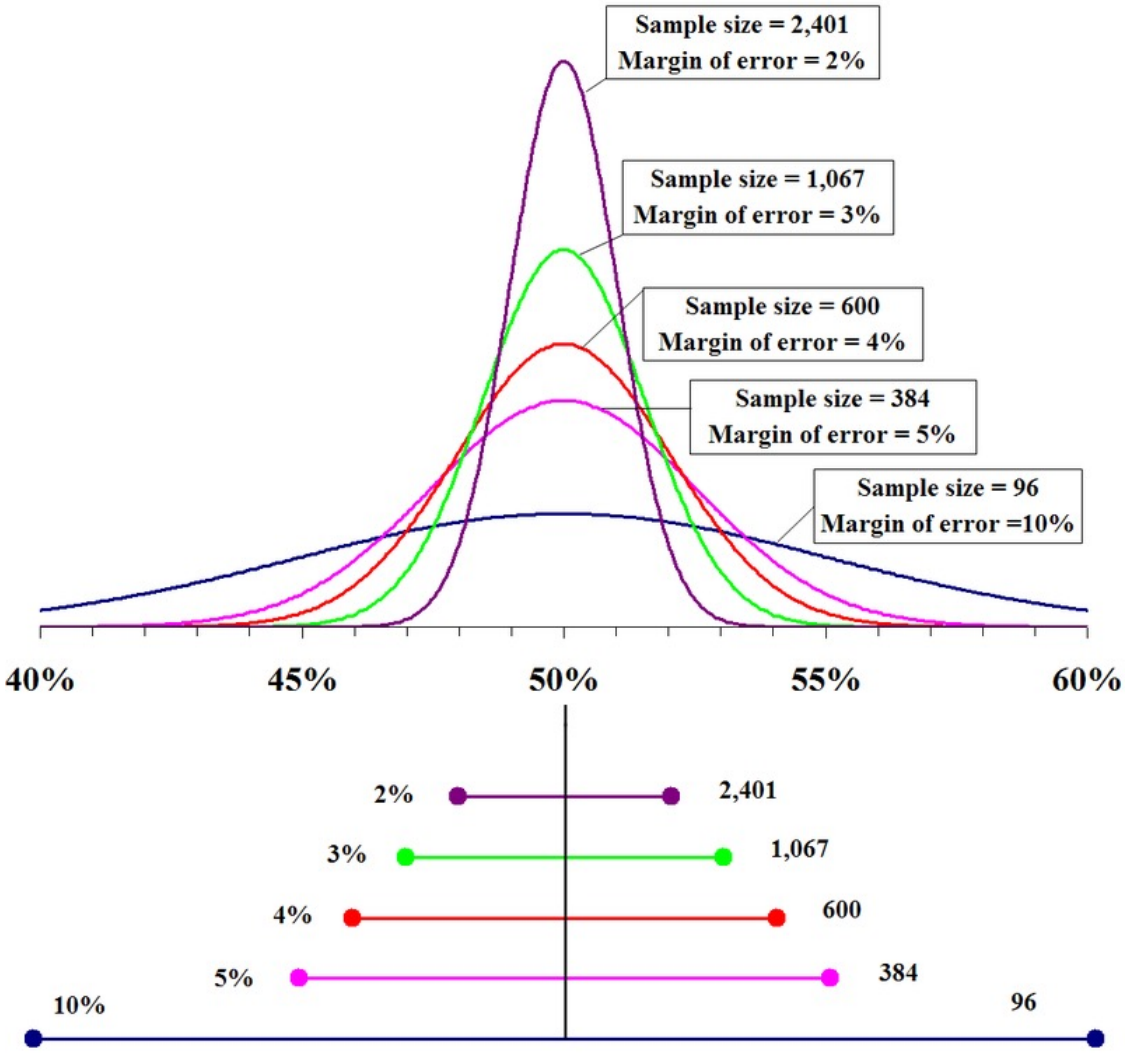
From AAS, April 24, 2015

Basics of Simple Random Sampling

- Estimating the proportion of a binary choice
- Estimate is based on a sample size n, much smaller than size of underlying population N
- Answer can only be "Probably Approximately Correct"
  - Quantify via ε, the margin of error, and 1-α, the confidence level
    - # of samples required depends on pre-specified "epsilon" and "alpha"

Estimating Sample Size

- Want: **within ε of mean** with high **probability (1- α)**
  - Normal: 90% of probability within +/- 1.65 σ of mean
    - 95% of probability within +/- 1.96 σ of mean
    - 99% of probability within +/- 2.58 σ of mean
    - Margin of error is ε; critical value (for standardized curve) is denoted by $z_{\alpha/2}$
      - » If **α** =0.05, then $z_{\alpha/2}$ is 1.96
- Minimum Sample size needed, $n = p(1-p)\,(z_{\alpha/2} / \varepsilon)^2$
  - **independent of N!!**
  - Use $\hat{p}$ for p in above Eqn; if $\hat{p}$ is unknown, use 0.5 for safe answer.



The top portion of this graphic depicts probability densities that show the relative likelihood that the "true" percentage is in a particular area given a reported percentage of 50%. The bottom portion shows the 95% confidence intervals (horizontal line segments), the corresponding margins of error (on the left), and sample sizes (on the right). In other words, for each sample size, one is 95% confident that the "true" percentage is in the region indicated by the corresponding segment. The larger the sample is, the smaller the margin of error is.

From https://en.wikipedia.org/wiki/Margin_of_error

Web Resources

- Many good web resources to understanding sampling, confidence intervals, etc.

Understanding confidence intervals:
http://www.lordsutch.com/pol251/schacht-08-web.pdf

Several lectures at OCW/MIT, e.g. see

- Lecture 8: Sampling and Standard Error

- NYTimes website for the polling demo:
https://www.nytimes.com/interactive/2018/upshot/elections-poll-wa08-1.html

Data Pre-Processing

Cleaning, integration, exploration, reduction/transformation, visualization….

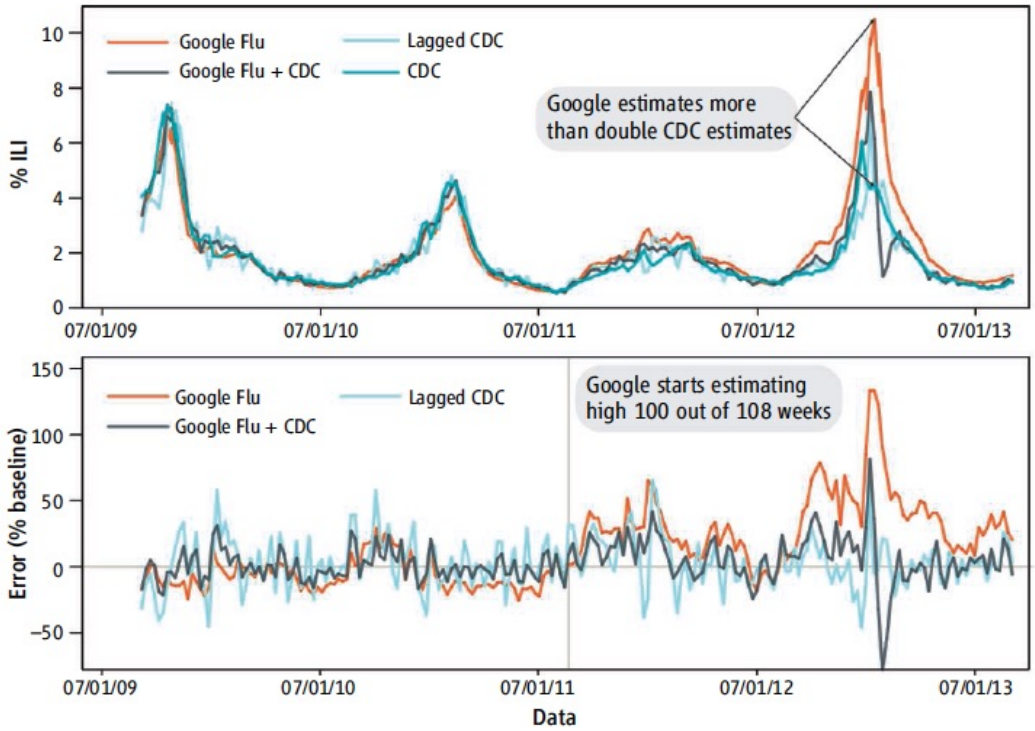
- **Optional Readings:**
  - **KJ Ch 3, 19**

Other sources:

- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- Garcia, Luengo, Herrera, "Data Preprocessing in Data Mining", Springer 2015.
- Xu Chu, Ihab Ilyas, Sanjay Krishnan, Jiannan Wang, "Data Cleaning: Overview and Emerging Challenges" SIGMOD Tutorial, Jun. 2016.
  slides at https://sites.google.com/site/datacleaningtutorialsigmod16/home/slides
- **Explore** segmentationOriginal (KJ, 3.1) and German Credit Card datasets

The Google Flu-Trends Fiasco

The Parable of Google Flu: Traps in Big Data Analysis
http://science.sciencemag.org/content/343/6176/1203

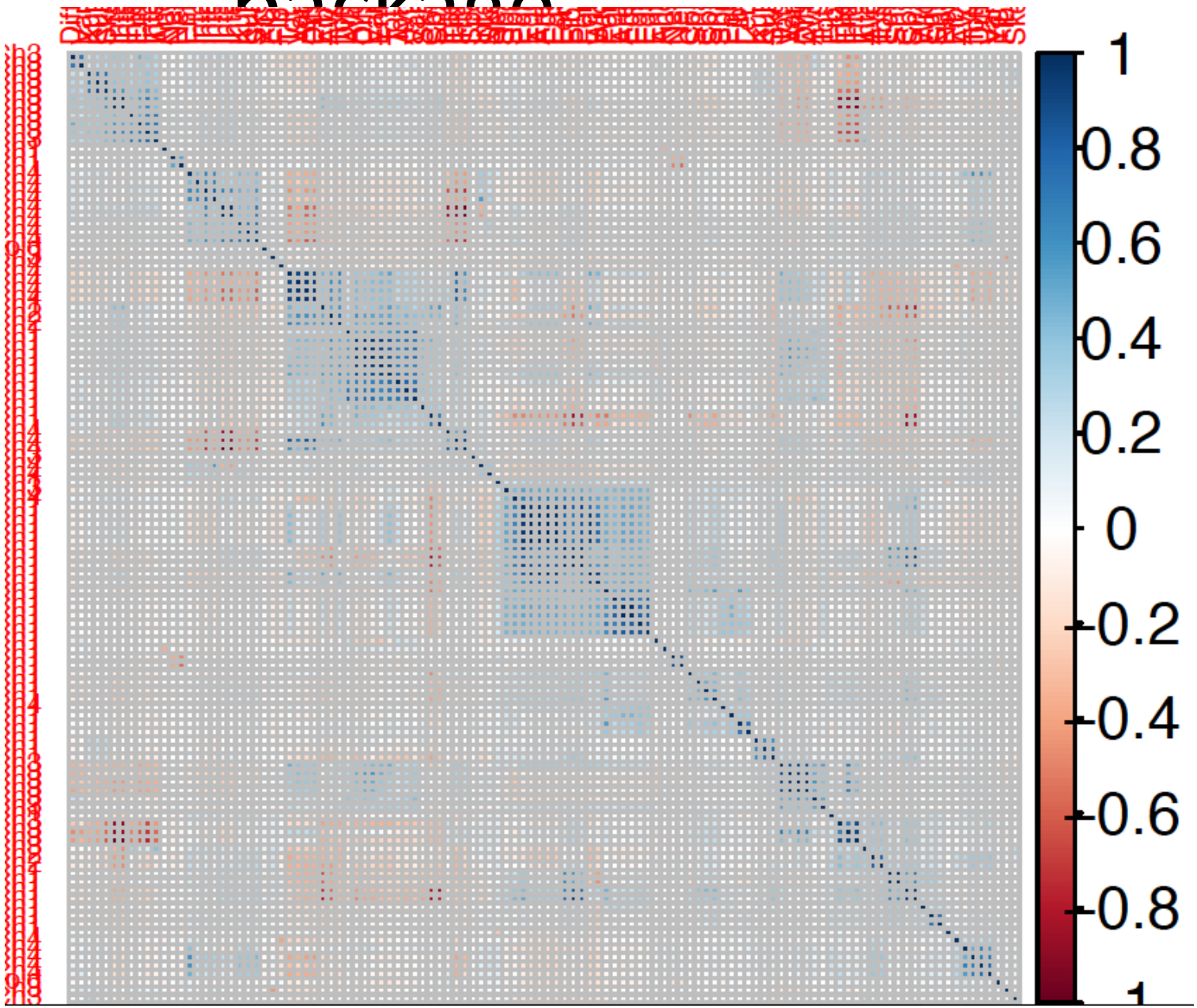


GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. **(Top)** Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. **(Bottom)** Error [as a percentage {[Non-CDC estmate) − (CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

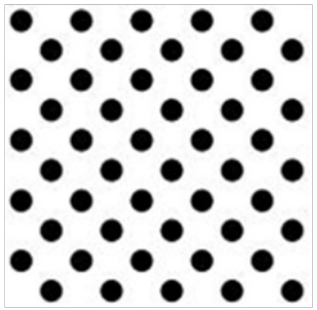
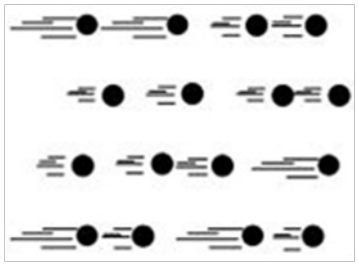
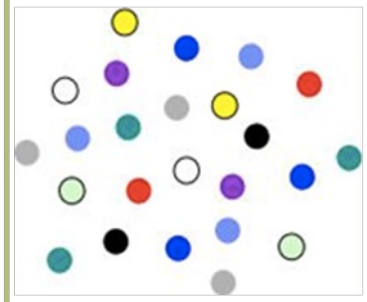
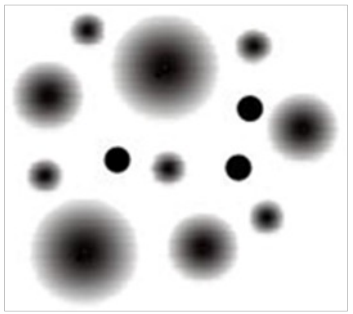
Feature Selection Using Corrplot package

- See KJ

Fig 3.10

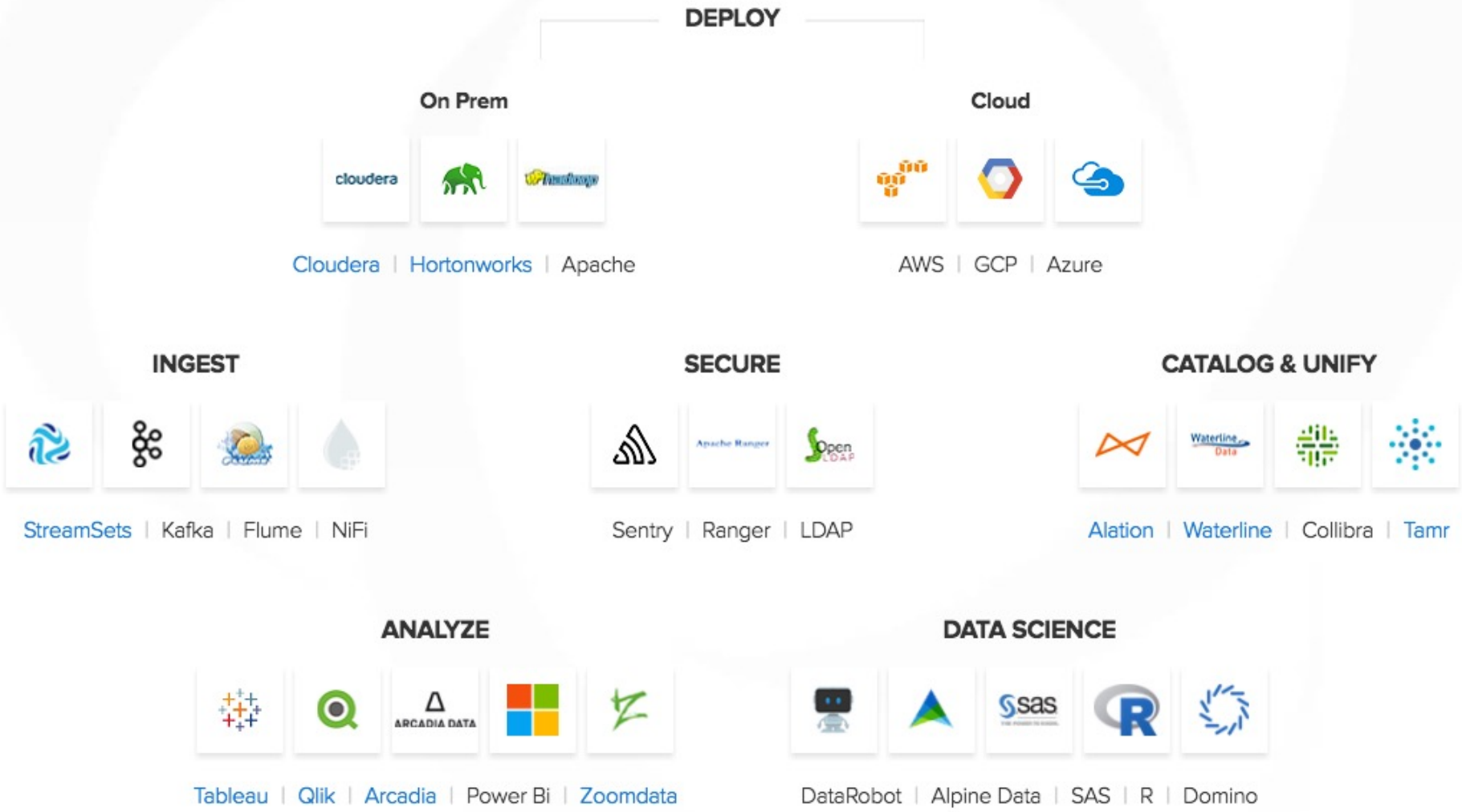


5 Vs of Big Data

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,…	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

Trifacta Integrates into the Modern Data Ecosystem



Curse of Dimensionality

See HTF pp 22-26.

- Exponential growth of # of cells with # of dimensions, p
  - implications
- Where is probability mass concentrated in "hyper"cubes/spheres, as p gets large?
- Practical: Target Encoding of Categorical Variables
  - And watching out for "leakage"

Visualizing the Curse

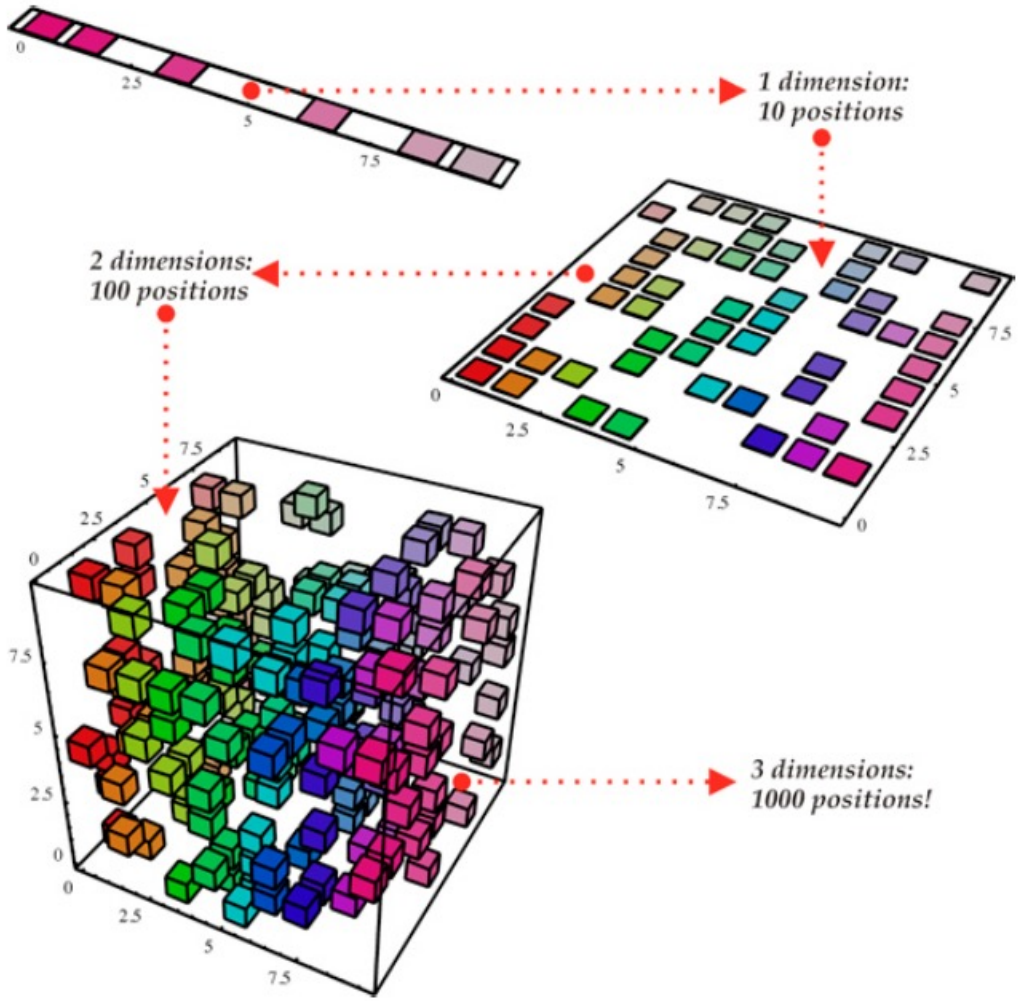


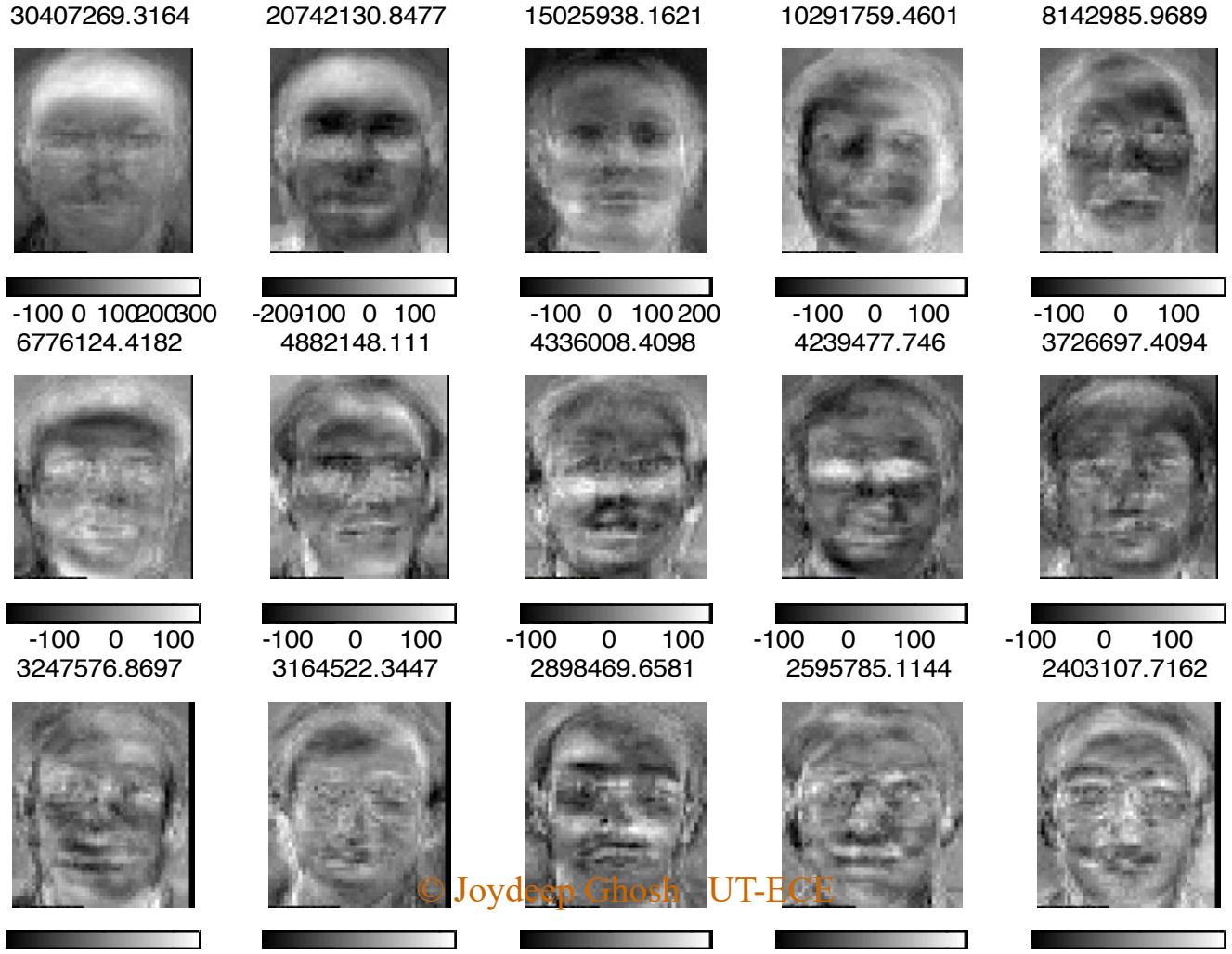
Image Database

EigenFaces

https://github.com/ml4a/ml4a-guides/blob/master/notebooks/eigenfaces.ipynb

http://www.cs.princeton.edu/~cdecoro/eigenfaces/

(also search for Eigenfaces on Youtube, e.g.
https://www.youtube.com/watch?v=_lY74pXWlS8&t=38s



Reconstruction

an example of reconstruction [from Turk and Pentland, 91] using different numbers of bases.

Singular Value Decomposition (SVD)

- Practical way of obtaining Principal components

day / customer	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

$$\mathbf{A} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

SVD

- Singular Value Decomposition (SVD)
  $A = U \times \Lambda \times V^T$
  - for A = customer -day matrix, interpret
  - U as customer-to-pattern similarity matrix
    - Columns of U are (orthonormal) eigen-"days"
      - Eigenvectors of AA^T
  - V as day-to-pattern similarity matrix
    - Rows of V are (orthonormal) eigen-"customers"
      - Eigenvectors of A^TA
  - Λ is diagonal matrix of singular values (sorted)
    - (sq. root of eigen-values of AA^T Or A^TA)