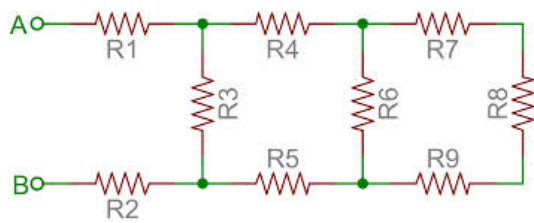# Data Issues

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

How Much Data do You Need? Effort?

- Quantity and quality

- Statistically large vs. computationally large

- Effort?
    - depends on variety of data sources, format, quality e.g. missing/incorrect items,.. Ownership issues,..
    - ETL vs. modern data wrangling

Data Pre-Processing

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

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

Also see:

- 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

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, ..
    - Out-of-date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

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.

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
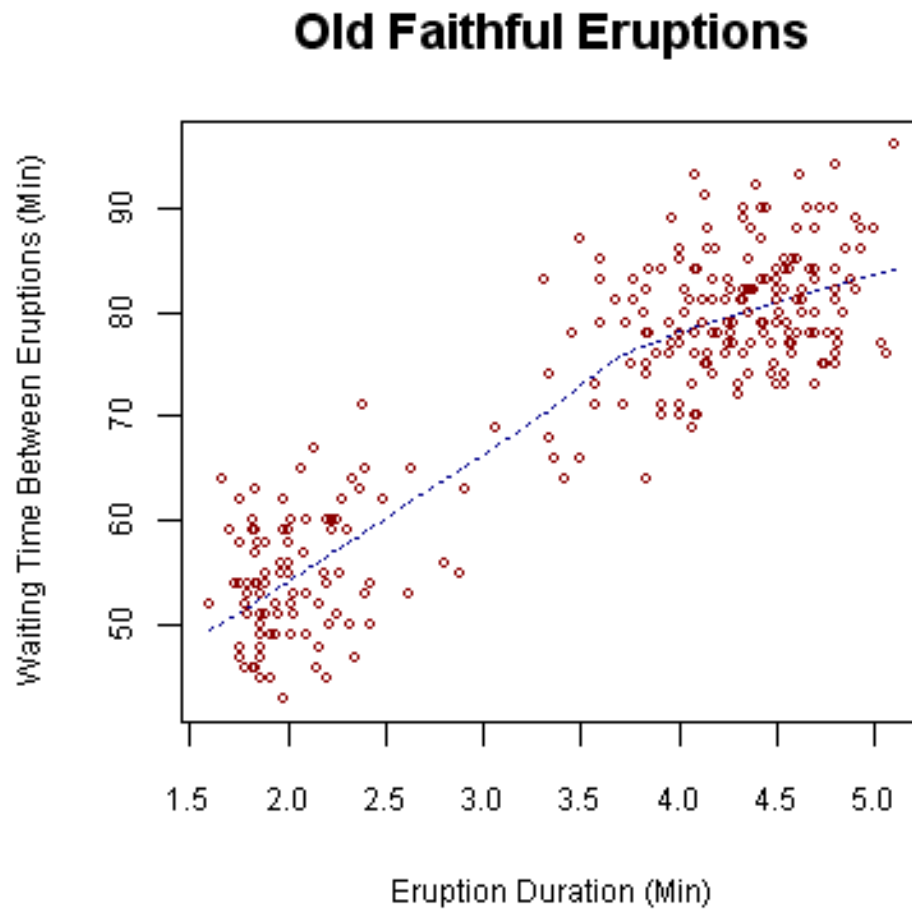   - (scatter plots, correlation,..)

Boxplot

- Shows median, 1st and 3rd quartile (Q), non-outlier extreme points; outliers
- Outliers, <1.5 IQR below 1st Q or >1.5IQR above 3rd Q.

See Wikipedia
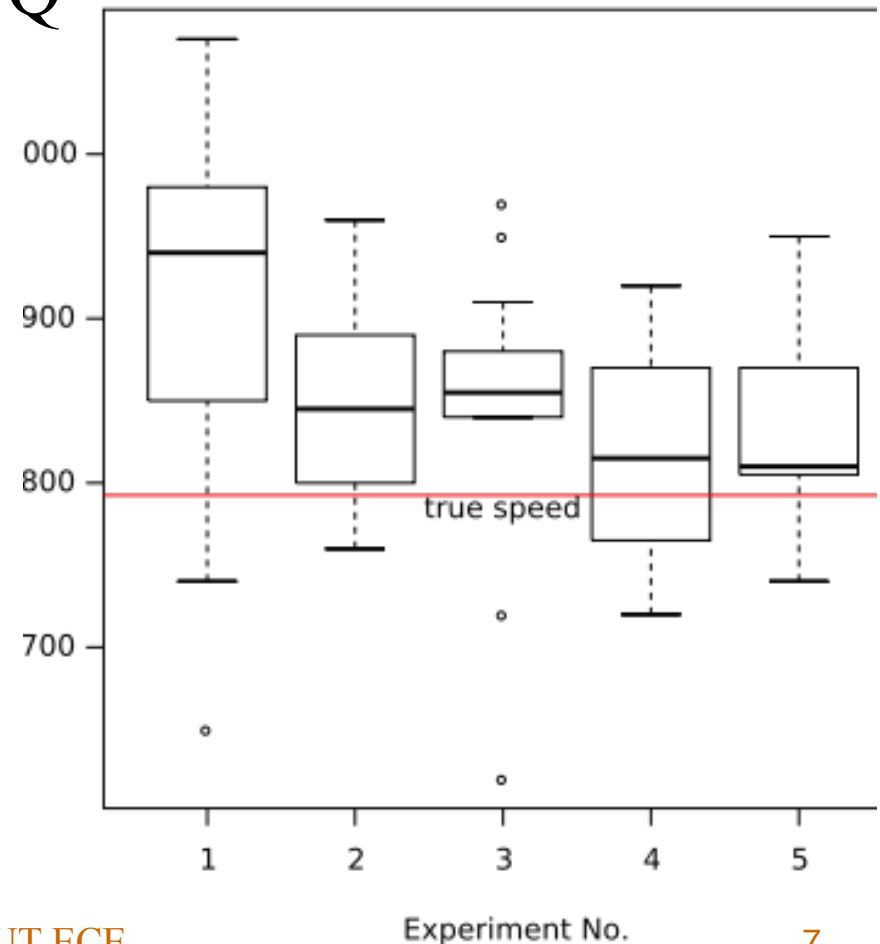
Scatterplot

- Old Faithful Example from Wikipedia


Old Faithful Eruptions

Data Cleaning I

- **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 variable
    - Use K-NN

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
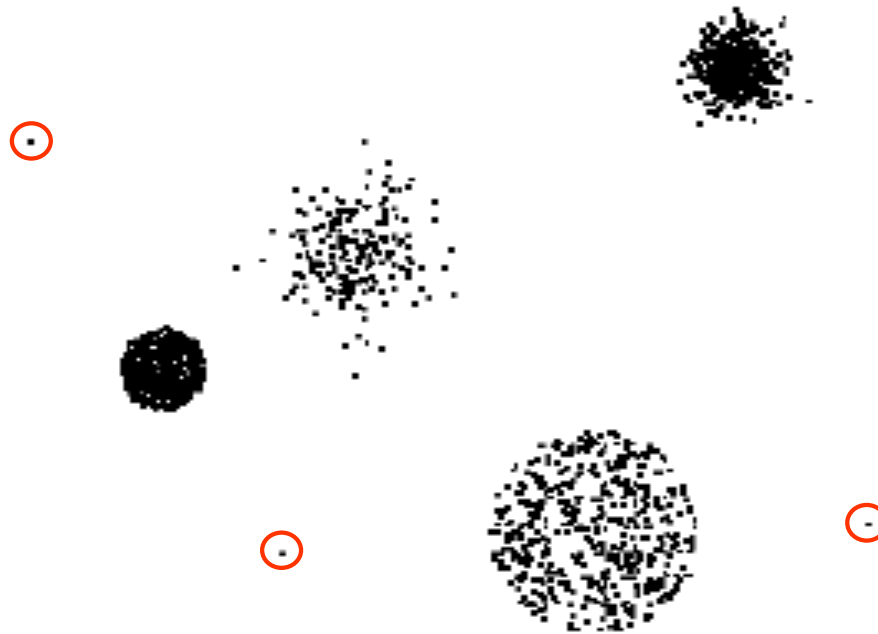
Dealing with Outliers in "X"

- Probability based (old):
  - 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)

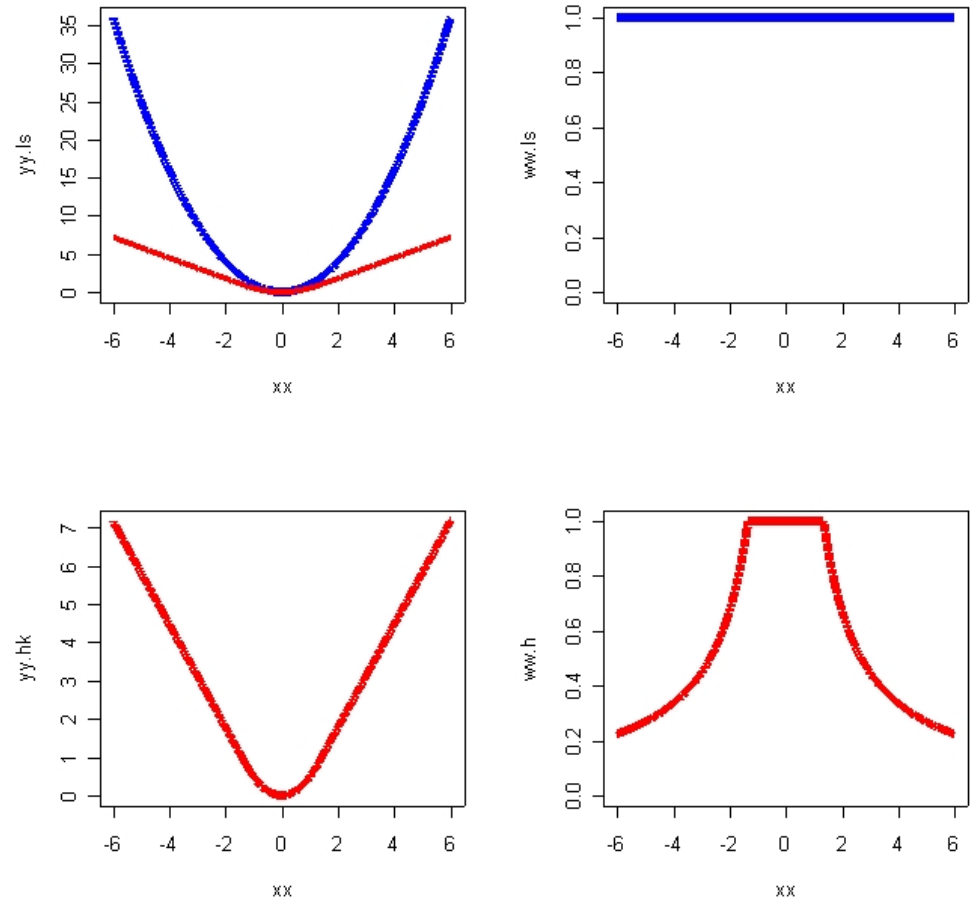Fig: Plotted as a function of residual ($r = y - \hat{y}$):
Blue: Sq. error loss (left) and "equivalent" weights (right)
Red: Huber loss and equivalent weights if Sq. loss was used (right)

Data Transformation

- Scaling

  Normalization by Linear scaling
  - Linear [min, max] $\rightarrow$ [0,1]
  - Centering (e.g. Z-scoring: Normal/Gaussian $\rightarrow$ 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

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
- 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.

Joydeep Ghosh   UT-ECE

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 $1-\alpha$, 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 p^ for p in above Eqn; if p^ is unknown, use 0.5 for safe answer.

Web Resources

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

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

Introduction to Probability (Undergrad course-notes from MIT).
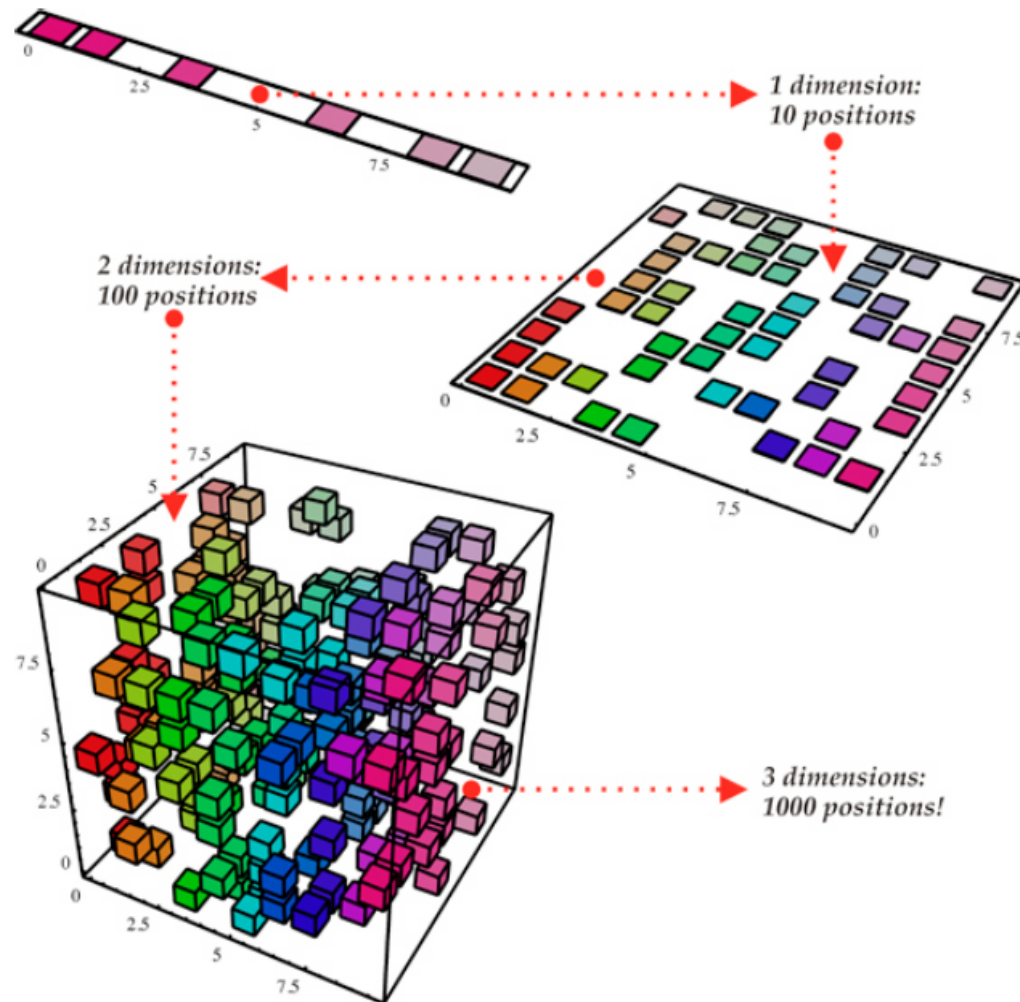http://ocw.mit.edu/OcwWeb/Mathematics/18-05Spring-2005/LectureNotes/index.htm

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?**

- **What happens to inter-point distances for randomly scattered points in high-p?**

Visualizing the Curse

1 dimension: 10 positions

2 dimensions: 100 positions

3 dimensions: 1000 positions!

Joydeep Ghosh   UT-ECE

Reducing # of (Derived) Attributes/Features

- Feature selection (select a subset of original features)

  vs

Feature extraction (use derived features, not original ones)

Why is feature selection often preferred to feature extraction?

Selecting a Subset of Features

NP-complete, so use heuristics

- Filter methods : use intrinsic quality measure
  e.g. correlation with other predictors (cor(data) in R);
  correlation/Chi-sq with target; mutual info with target
- Wrappers (extrinsic evaluation)
  – Greedily evaluate using predictive model (e.g. decision tree)
  – Search strategy for candidate sets to evaluate:
    – Forward inclusion
    – Backward elimination
    – Stepwise (forward, but may remove predictors that no longer meet criterion)
- Embedded (feature selection part of model training, e.g LASSO)

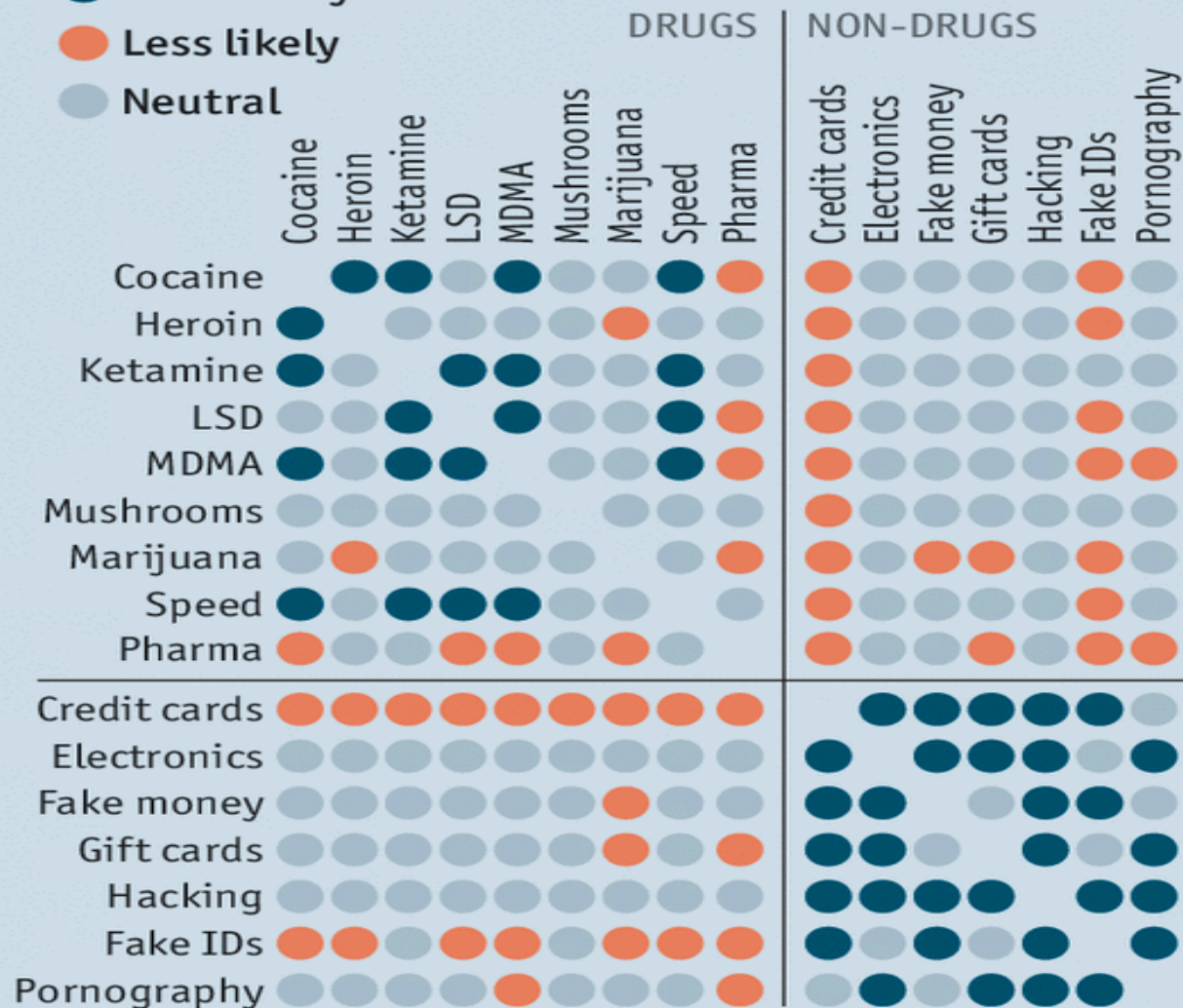- Advanced Methods : http://featureselection.asu.edu/index.php

Perfect pairings

Likelihood that vendors selling one product on dark-web markets will sell another
December 2013-July 2015

Sources: Gwern Branwen's dark-web archive; *The Economist*

Feature Extraction Choices

- Linear
  - Unsupervised : PCA
    - 5 functions to do Principal Components Analysis in R
    http://gastonsanchez.com/blog/how-to/2012/06/17/PCA-in-R.html
  - Supervised:
    - Fisher's Linear Discriminant (classification)
    - Canonical Correlation (regression)

- Non-Linear
  - Unsupervised : Principal Curves, Sammon's Map, Kohonen's SOM
  - Supervised: Nonlinear discriminant analysis, e.g. using a multi-layered perceptron.

PCA

- Principal Components Analysis:
  - Reduce dimensions while retaining info about original data
    - PCA finds the best "subspace" that captures as much data variance as possible
  - optimal linear projection/reconstruction in MSE sense
  - Based on eigen-decomposition of data covariance matrix; can also be obtained sequentially
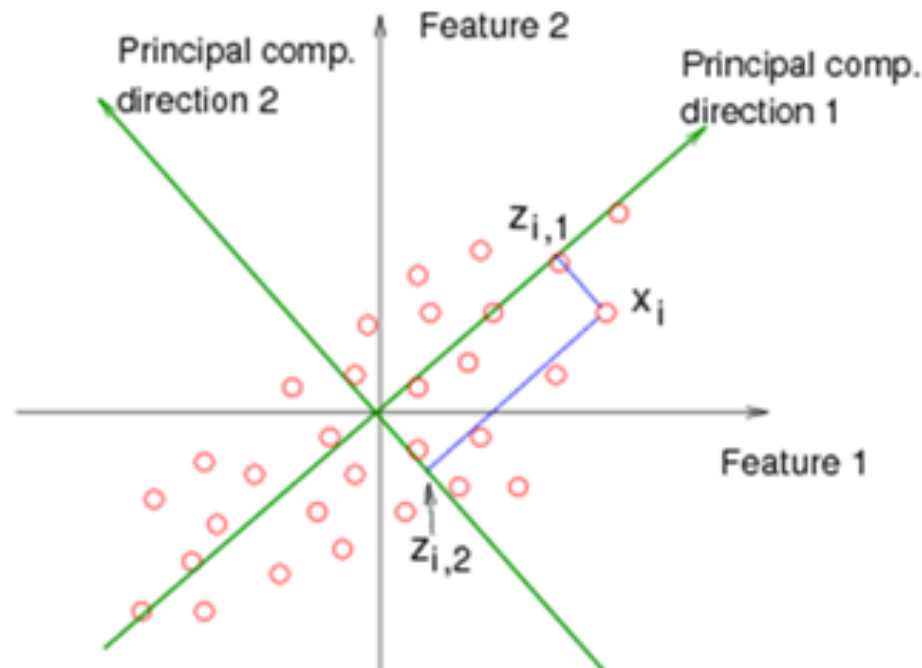
Another Example

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

Joydeep Ghosh   UT-ECE
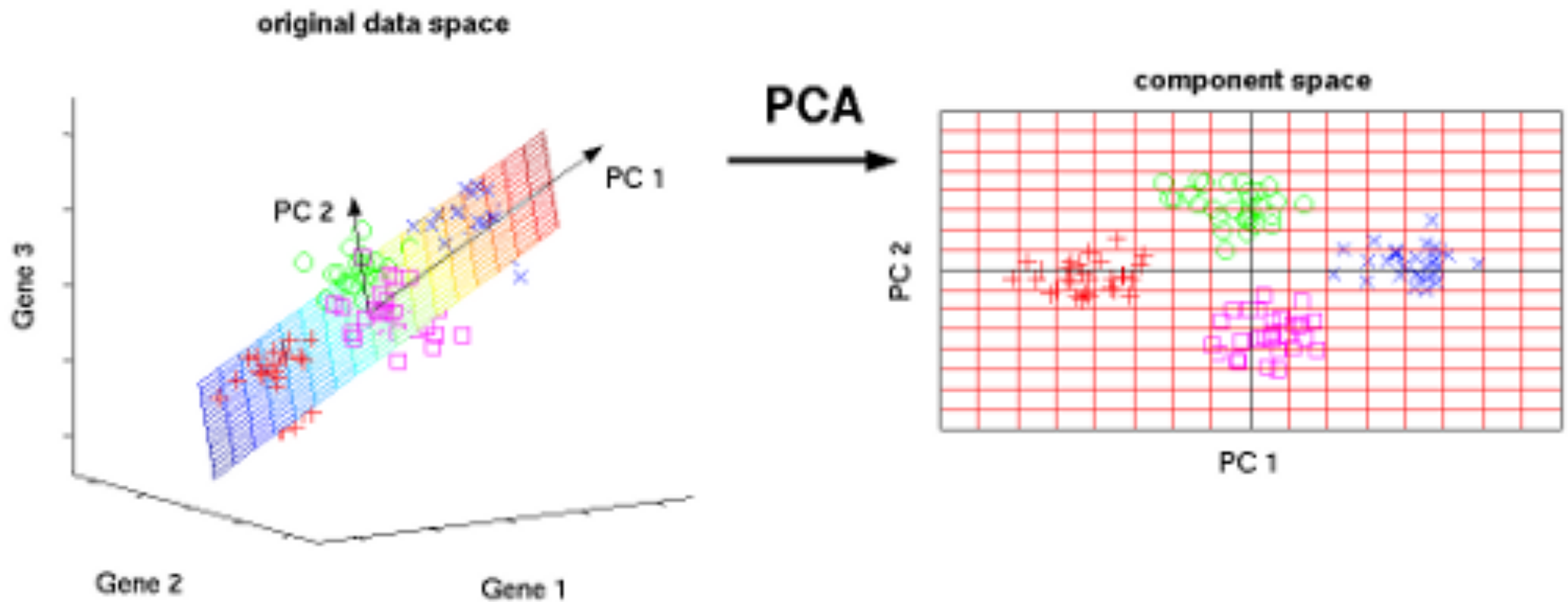
PCs sorted by amount of variance explained

Example scree plot
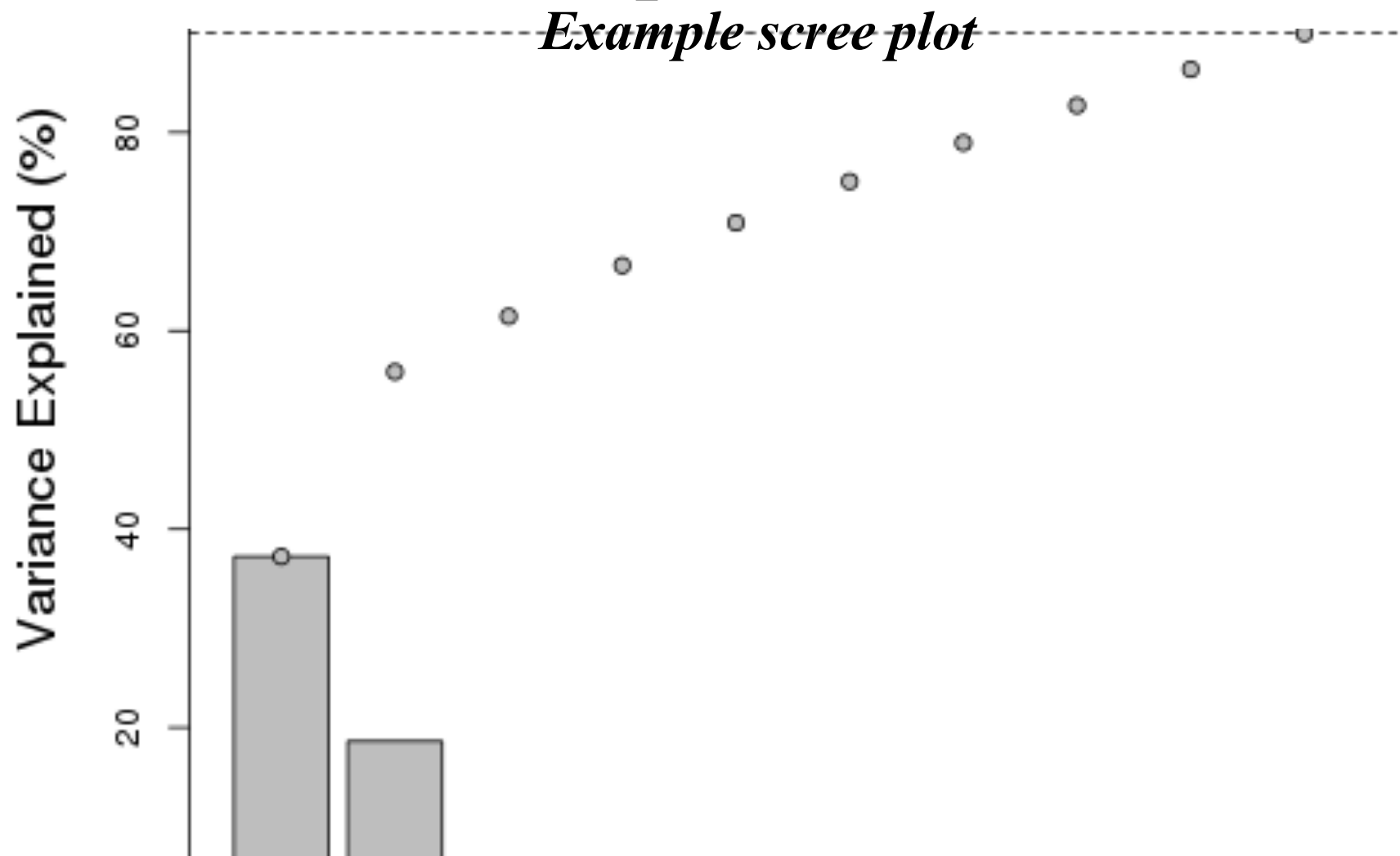
Tensor Board Visualization

- PCA and t-SNE for MNIST

- https://www.youtube.com/watch?v=eBbEDRsCmv4 19:11 onwards

- Tensorflow dev summit 2017
- https://www.youtube.com/playlist?list=PLOU2XLYxmsIKGc_NBoIhTn2Qhraji53cv

Visualize

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

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

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

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

The Google Flu-Trends Fiasco

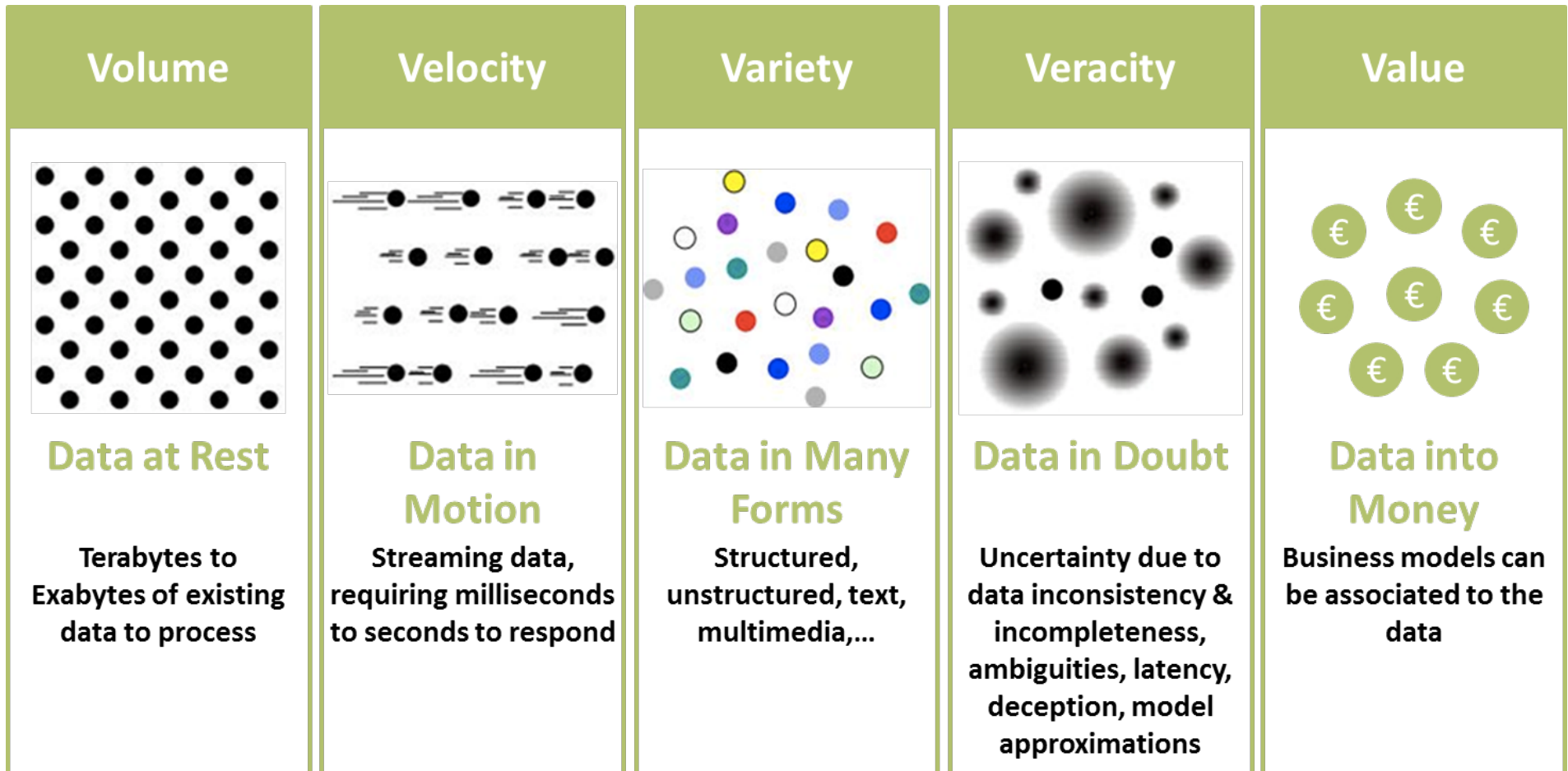The Parable of Google Flu: Traps in Big Data Analysis
http://science.sciencemag.org/content/343/6176/1203
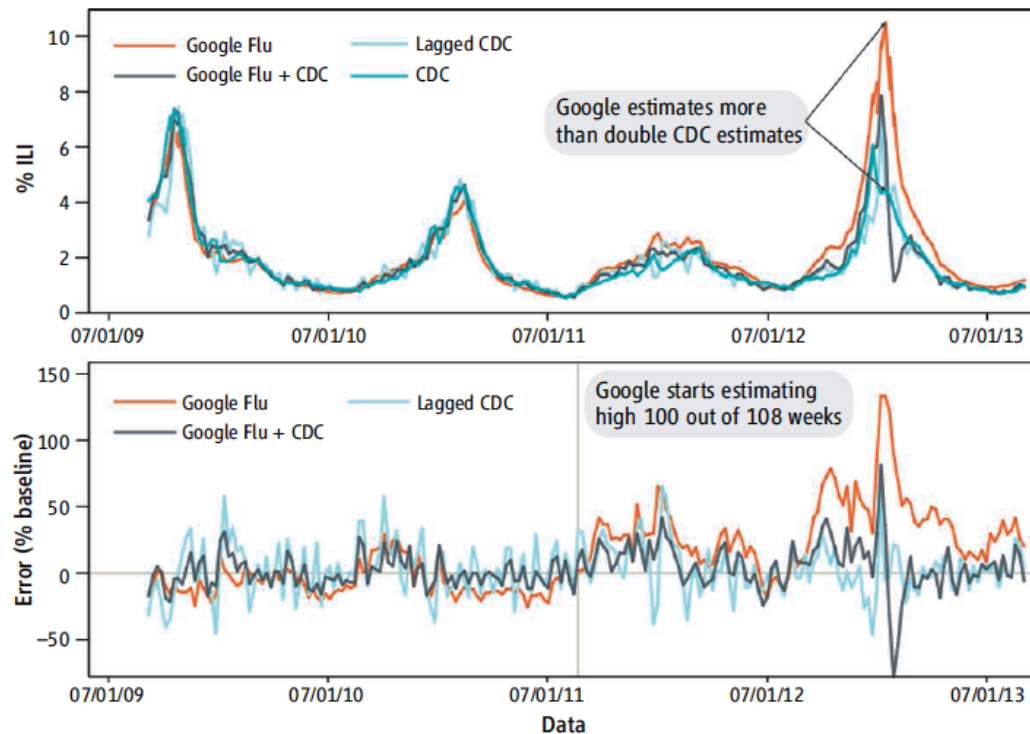
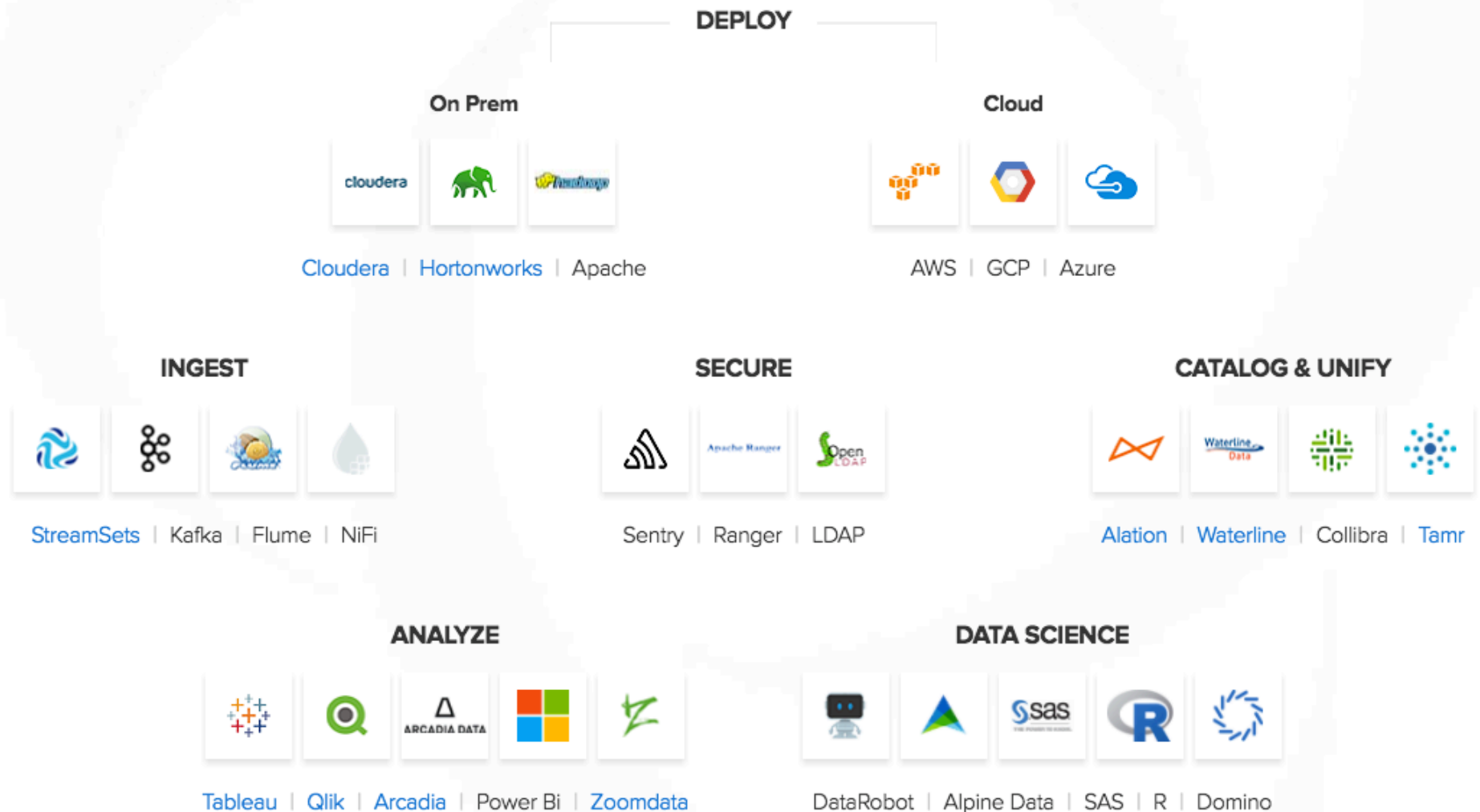

GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged 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.

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

Joydeep Ghosh   UT-ECE

Trifacta Integrates into the Modern Data Ecosystem

DEPLOY

On Prem

cloudera | Hortonworks | Apache

Cloud

AWS | GCP | Azure

INGEST

StreamSets | Kafka | Flume | NiFi

SECURE

Sentry | Ranger | LDAP

CATALOG & UNIFY

Alation | Waterline | Collibra | Tamr

ANALYZE

Tableau | Qlik | Arcadia | Power Bi | Zoomdata

DATA SCIENCE

DataRobot | Alpine Data | SAS | R | Domino

Joydeep Ghosh   UT-ECE

Extras

Joydeep Ghosh   UT-ECE

Singular Value Decomposition (SVD)

- Practical way of obtaining Principal compo

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

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

SVD

- Singular Value Decomposition (SVD)
  $A = U \ x \ \Lambda \ x \ V^T$
  - 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
  - is diagonal matrix of singular values (sorted)
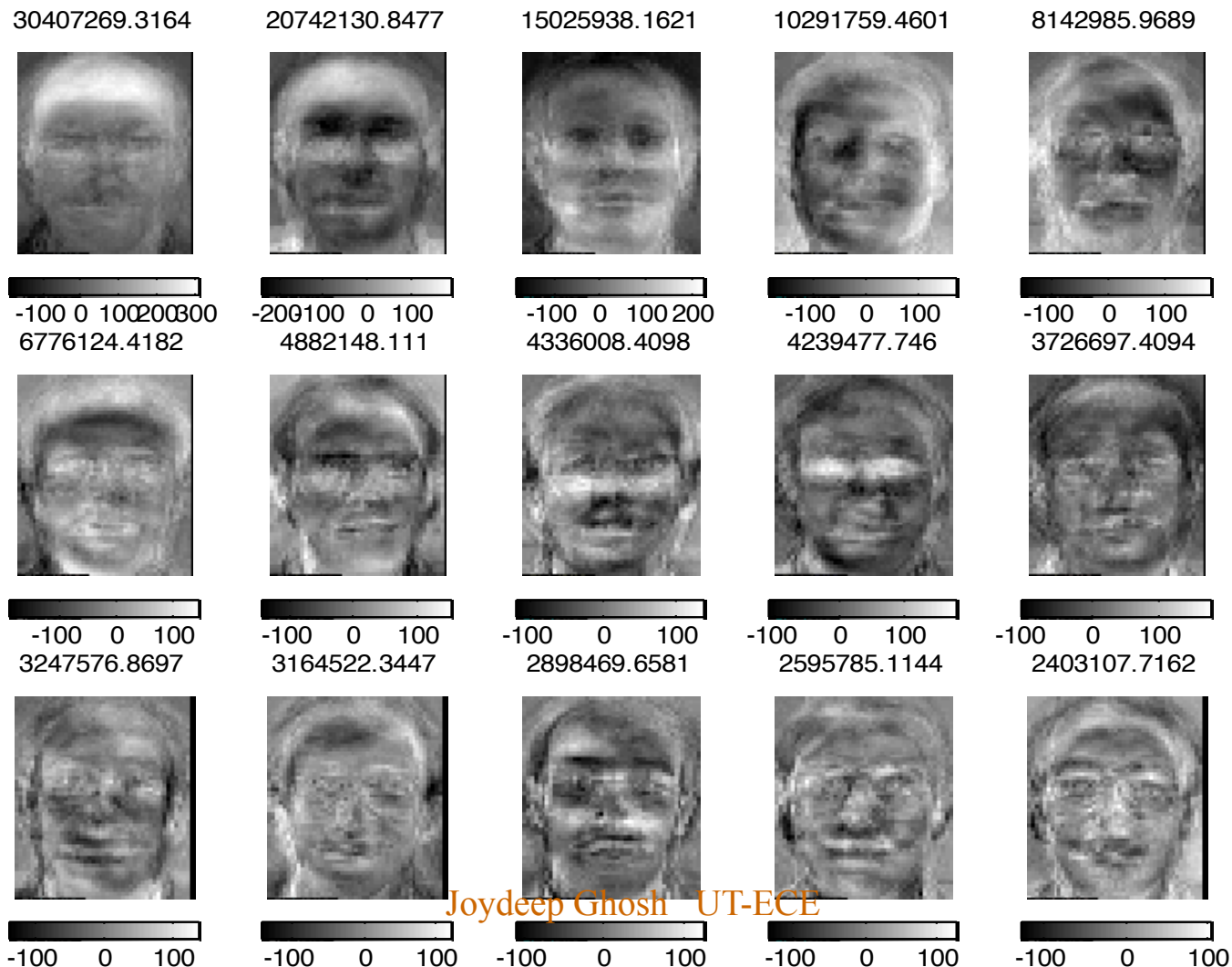   - (sq. root of eigen-values of AA^T Or A^TA)

Image Database

EigenFaces

Reconstruction

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

Effect of using more bases: each consecutive image uses 8 more bases

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

Joydeep Ghosh   UT-ECE