11-27-2018

Introduction to Data Analytics and Emerging Real-World Use Cases

Art Chaovalitwongse
University of Arkansas, Fayetteville

Citation

Chaovalitwongse, A. (2018). Introduction to Data Analytics and Emerging Real-World Use Cases. *Operations Management Presentations.* Retrieved from https://scholarworks.uark.edu/opmapub/20

Engineering and Operations Management Lunch & Learn Webinar Series

November 27, 2018

UNIVERSITY OF ARKANSAS

Online Degree Options

Expand breadth and depth of engineering knowledge	*Provide leadership and business skills to manage technology teams*	*Improve effectiveness and efficiency of operations*
M.S. in Engineering	**M.S. in Engineering Management**	**M.S. in Operations Management**
MSE Comprehensive Exam	MSEM Comprehensive Exam	MSOM Comprehensive Exam
3 Electives chosen from MSEM, OMGT, and Engineering courses	3 Electives chosen from MSEM, OMGT, and Engineering courses	6 Electives chosen from OMGT and MSEM courses
4 core courses from approved list Computer Applications Mathematics Management Technical Communications	4 core courses EMGT 5033 Intro to Engineering Mgt OMGT 5783 Project Management OMGT 5463 Economic Decision Making INEG/OMGT 5443 Decision Models	4 core courses OMGT 5003 Intro to Operations Mgt OMGT 5123 Finance or 5463 Economic Decision Making OMGT 5623 Strategic Management OMGT 5783 Project Management
Approved 3 course engineering sequence to form cohesive topic area	Approved 3 course engineering sequence	Undergraduate Prereqs (if required) OMGT 4853 Decision Support Tools OMGT 4323 Industrial Cost Analysis OMGT 4333 Statistics OMGT 4313 Law and Ethics
ABET Accredited Bachelors Degree in Engineering		Any Regionally Accredited Bachelors Degree

TODAY'S PRESENTER

Dr. W. Art Chaovalitwongse

21st Century Leadership Chair in Engineering, Professor of Industrial Engineering, and Co-Director of the Institute of Advanced Data Analytics at the UofA

Experienced Professor in Industrial & Systems Engineering & Radiology, Bioengineering, and Operations Research and Financial Engineering

Previously held faculty positions at Rutgers University, Princeton University and University of Washington, Seattle

Industry Experience with Corporate Strategic Research, ExxonMobil Research & Engineering; also holds 3 patents of seizure prediction system, now licensed by Optima Neuroscience, Inc.

Numerous academic honors include National Science Foundation CAREER Award and most recently, the 2018 Technical Innovation in Industrial Engineering Award by the Institute for Industrial & Systems Engineers, among others

Currently serves as Department Editor, Associate Editor and Editorial Board Member of 10 leading international journals. He has edited 4 books and published over 175 research articles

Introduction to Data Analytics and Real-World Use Cases

W. Art Chaovalitwongse

21st Century Research Leadership Endowed Chair

Professor, Department of Industrial Engineering

Co-Director, Institute for Advanced Data Analytics

University of Arkansas, Fayetteville

INTRODUCTION

Data Analytics

90% of the world's current data has been created in the last two years

15 out of 17 industry sectors in the US have more information stored per company than the US Library of Congress

... the best performing firms excel at accessing data, drawing meaningful insights, and transforming this into action.

CEOs agree that deeper customer insight will come from a much better use of data and analytics...

Data is expanding EXPONENTIALLY

How big is enough?

640K ought to be enough for anybody.

Kilobytes (10^3) → Megabytes (10^6) → Gigabytes (10^9) → Terabytes (10^{12})

Petabytes (10^{15}) → Exabytes (10^{18}) → Zettabytes (10^{21})

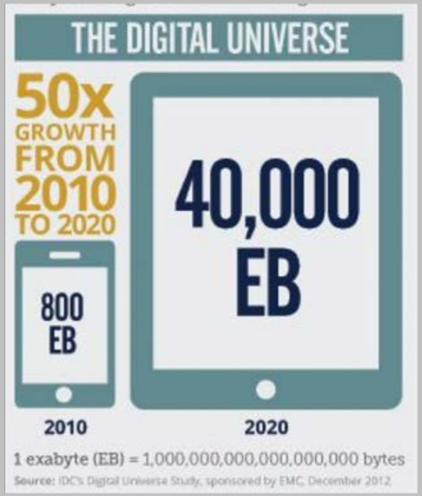
History of Data Storage

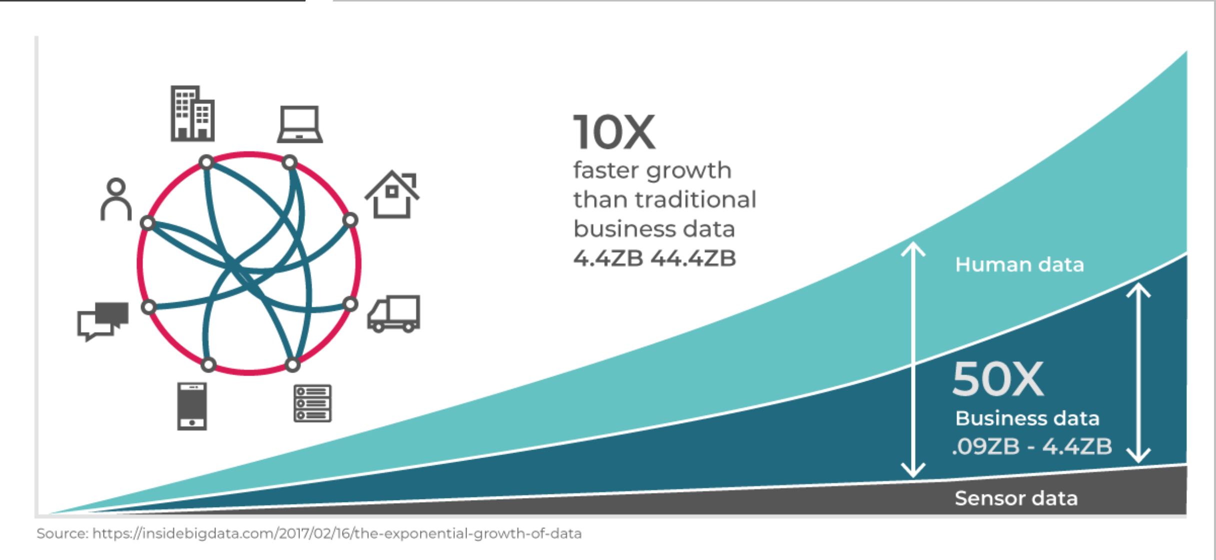
8-inch Floppy	Developed in 1967 by IBM, San Jose, California. a read-only, 8-inch (20 cm) floppy (called the "memory disk“), holding 80 kilobytes (KB).	
5¼-inch drive	Developed in 1975 by Burrough, Glenrothes. A new "double density" format increased it again, to 360 KB of data	
3½-inch disk	widely used in 1984 when Apple Computer selected the Sony 90.0 × 94.0 mm format for their Macintosh computers. A newer "high-density" format storing 1440 KB of data	
Zip Drive	Introduced with a capacity of 100 MB of data. Plans for a lower cost 25 MB version that would work in the same 100MB drive	

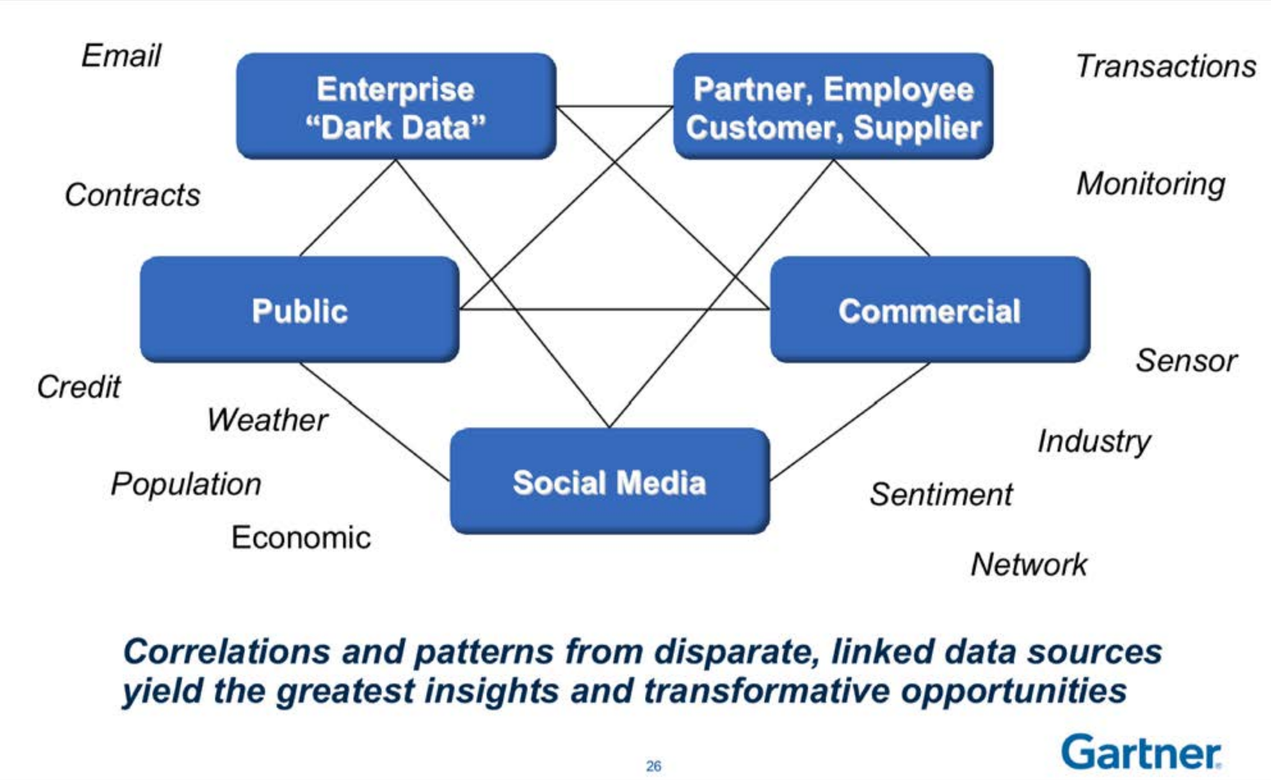
Data is expanding EXPONENTIALLY

There is an explosion of data fueled by cheap and ubiquitous *collection and storage* of everything around us.

- *every single action on websites,*
- *personal and health records,*
- *business transactions,*
- *mobiles,*
- *sensors,*
- *etc.*







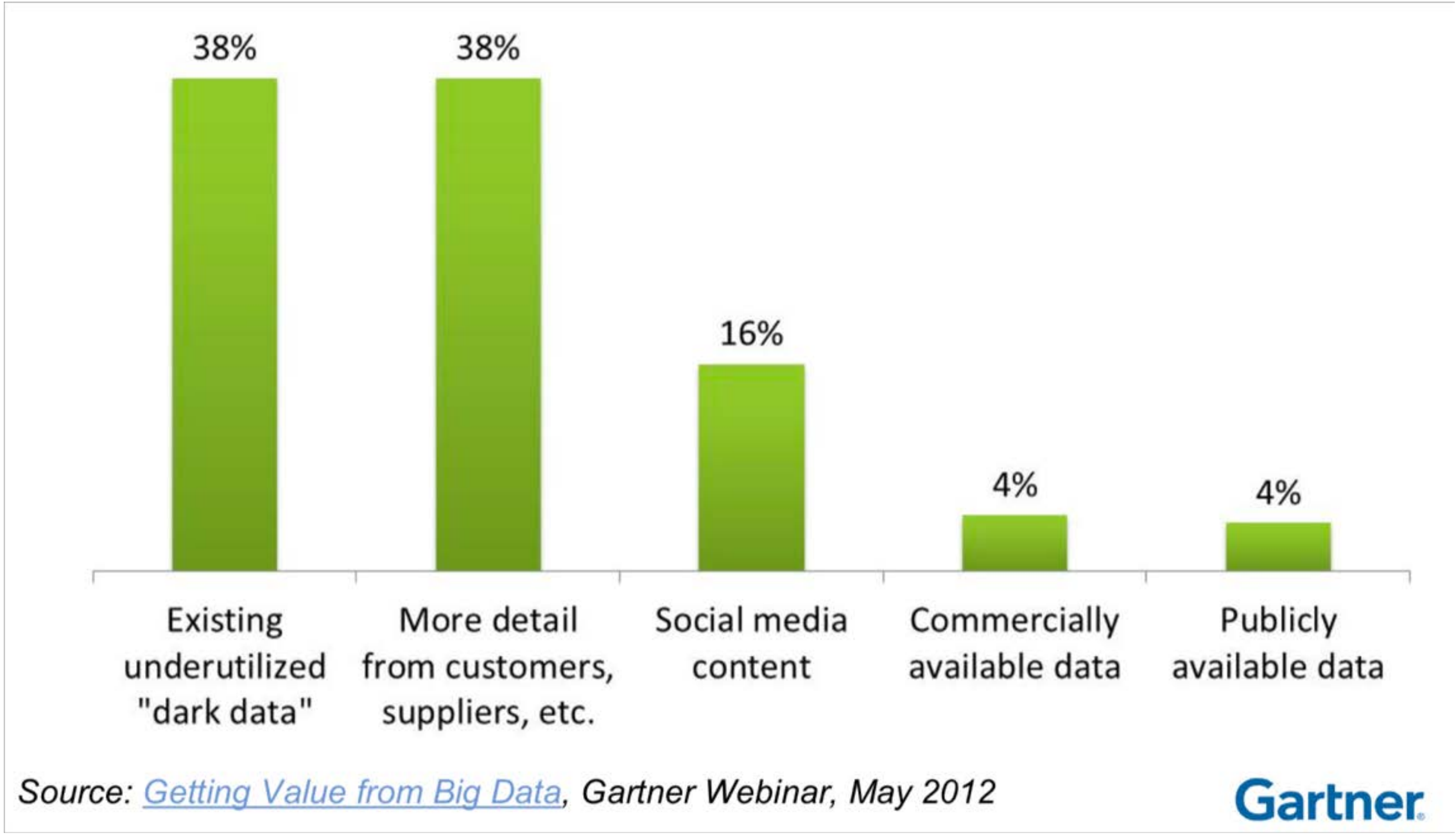
- **Enterprise data**
  - 6 of 10 organizations have more data then they know how to use
- **The Internet (social media)**
- **Communications (VoIP, VDO calls)**
- **Internet of Things (sensors)**

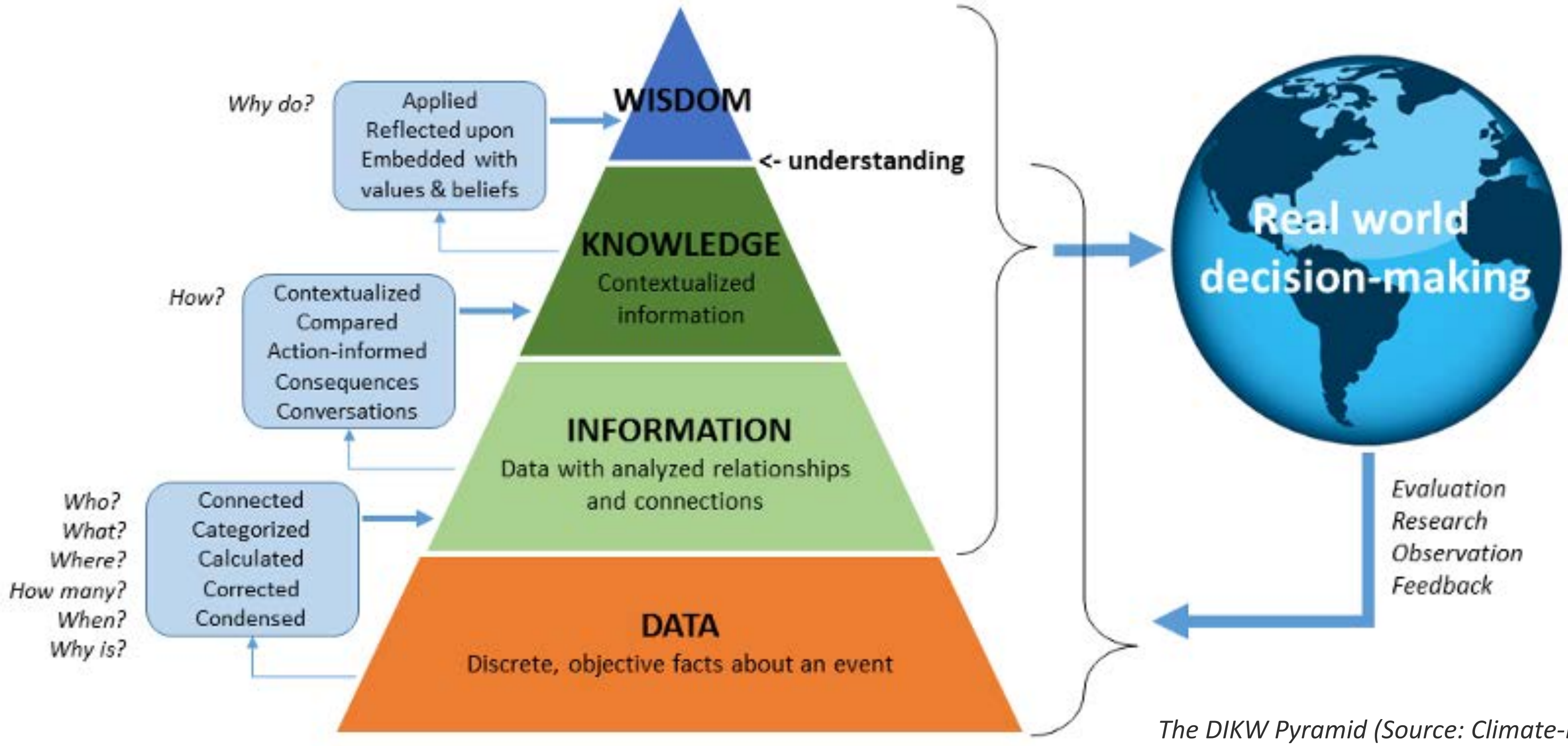
Where do the data come from?

The New Oil

We now need human resources and technologies that can help us make sense of this data, and become more intelligent in our decisions.

Estimated Utility Value



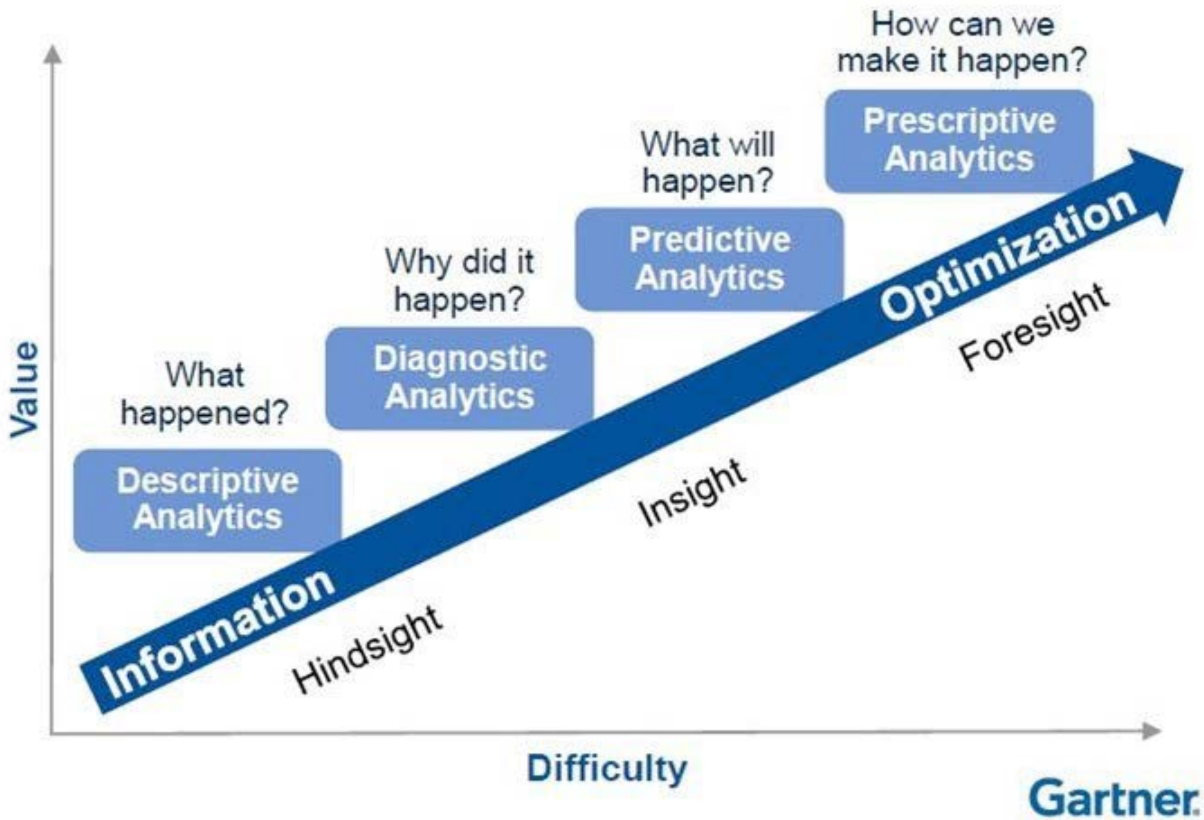


The DIKW Pyramid (Source: Climate-Eval.org)

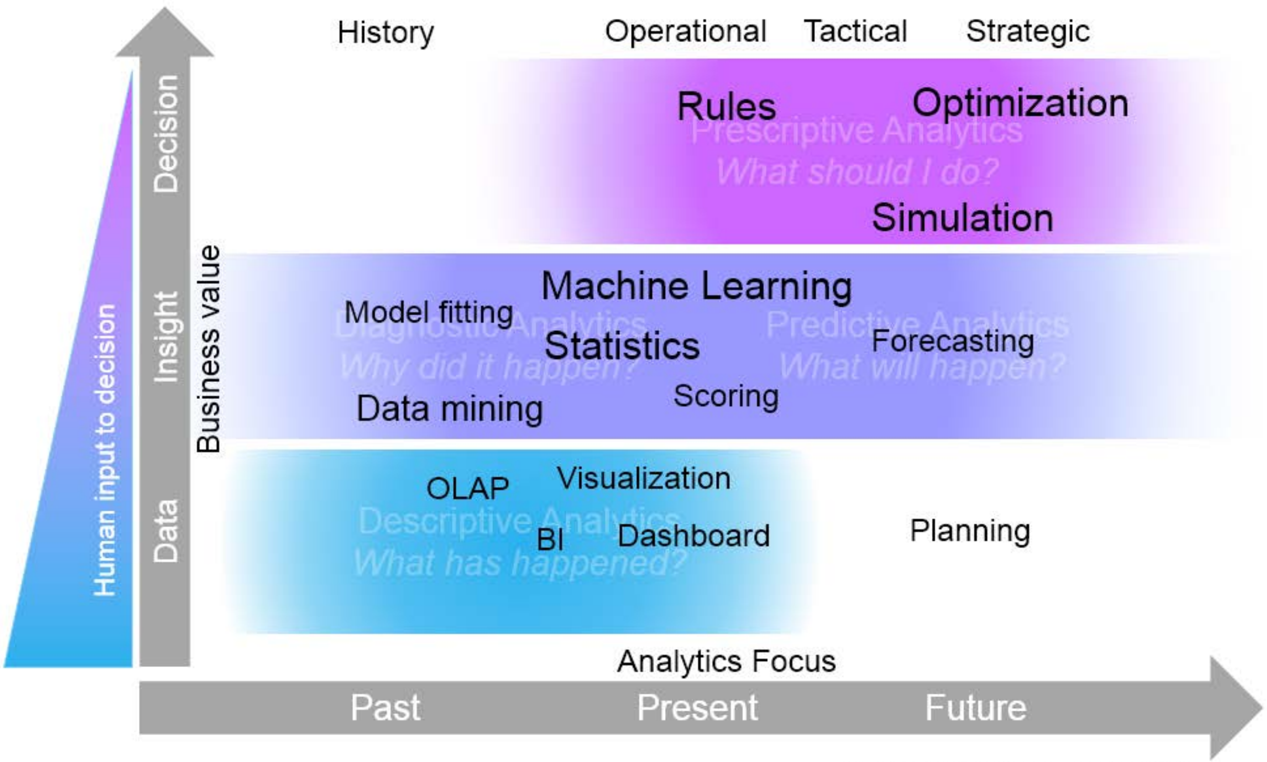
- Data Analytics (or *Data Science*) is one of the fastest growing fields of this decade.
- Data Analytics is the science of examining raw data with the purpose of drawing conclusions about that information.

Data Analytics

Varying Levels of Data Analytics

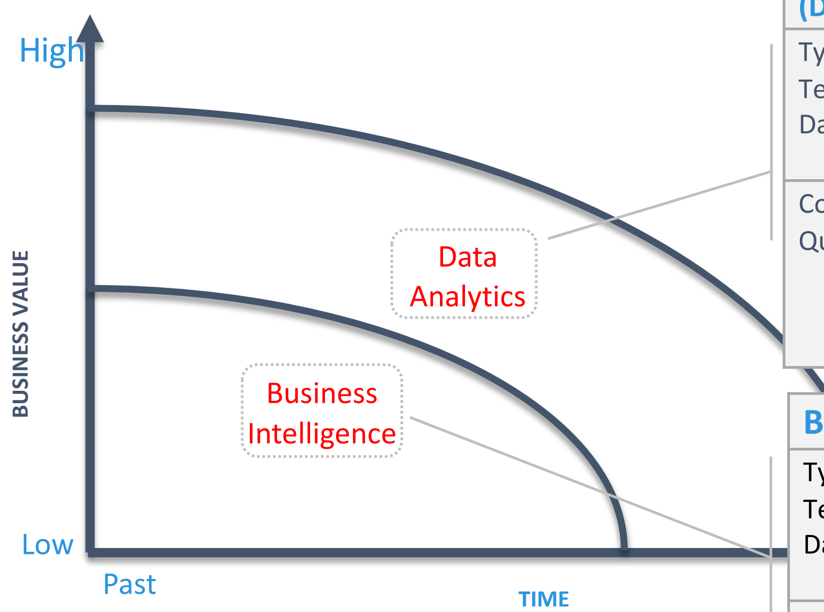


Analytics Lanscape



Source: http://ibm.co/1gJyfl3

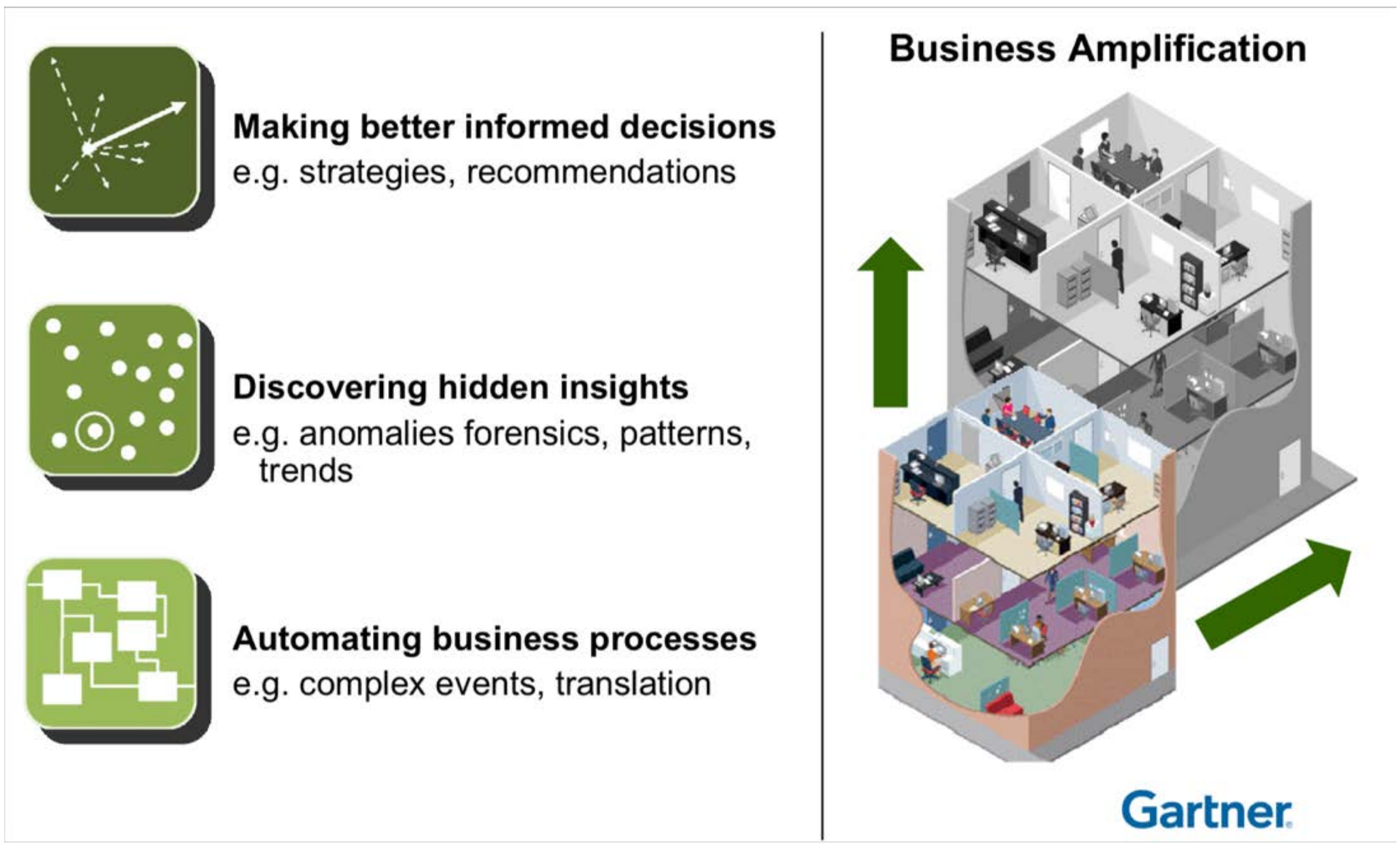
Moving Away from Traditional BI



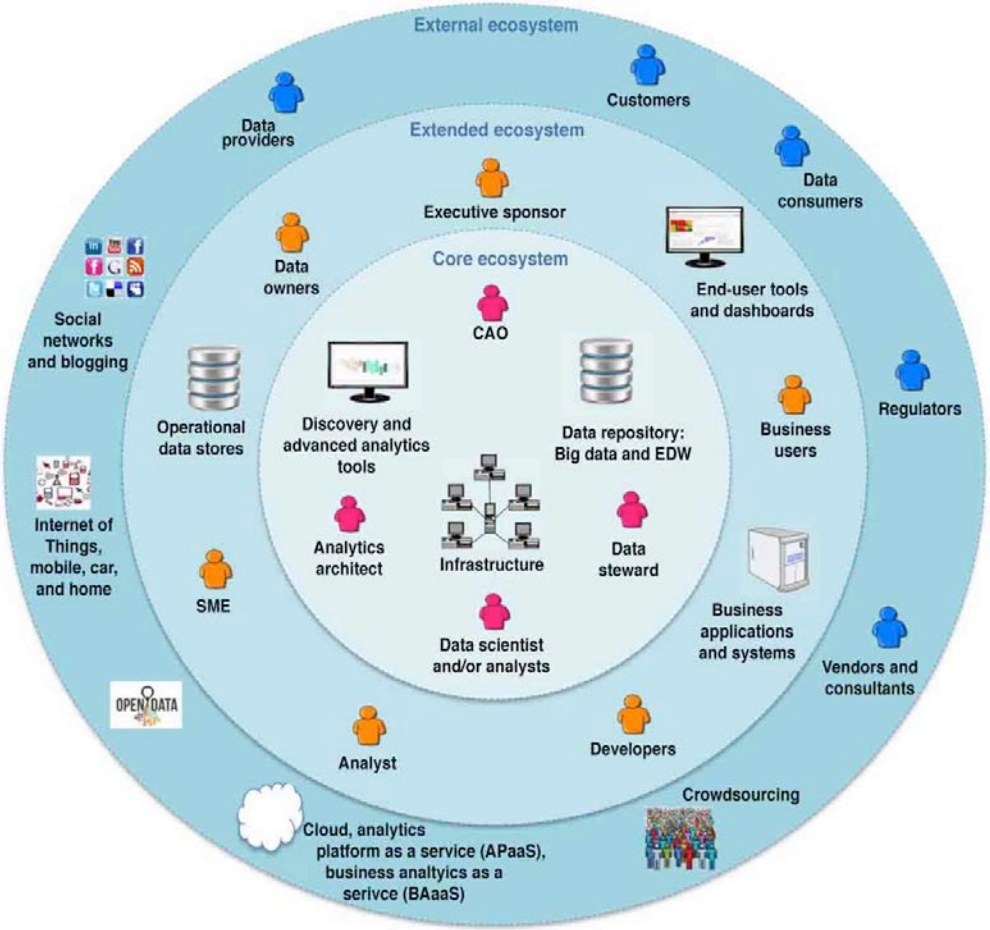
Predictive Analytics & Data Mining (Data Science)	
Typical Techniques & Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large data sets
Common Questions	• What if.....? • What's the optimal scenario for our business ? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques & Data Types	• Standard and ad hoc reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable data sets
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

Opportunities



What about Big Data?



- What if data sets are too large to fit in the memory of your laptop?
- Need to know how to scale analyses and algorithms to large data sets.
- This is where *"Big Data"* technologies come in.

  - Simply Infrastructure

Big Data Ecosystem

The Big Data technology stack is changing rapidly

Source: Informatica blogRead

REAL LIFE PROBLEMS

Data Analytics

Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories.

- ✓**Cost reduction:** identify more efficient ways of doing business.
- ✓**Faster, better decision making:** analyze information immediately – and make decisions based on what they've learned.
- ✓**New products and services:** provide the ability to gauge customer needs and satisfaction so they can create new products to meet customers' needs.

What can Data Analytics do?

Source: Socialnomics.net

Wal-Mart finding out what sells in a hurricane

Netflix finding out what movies a customer might want to watch

An investor finding out anomalies exist in the stock market in order to make a profit to his/her customers

Amazon personalizing and customizing websites

Sprint finding out that a customer might want to drop its service before the customer even knows it *(churn)*

UPS finding the best route for a package in a road network

Harrah's casinos gathering and mining data on gamblers to attract them back

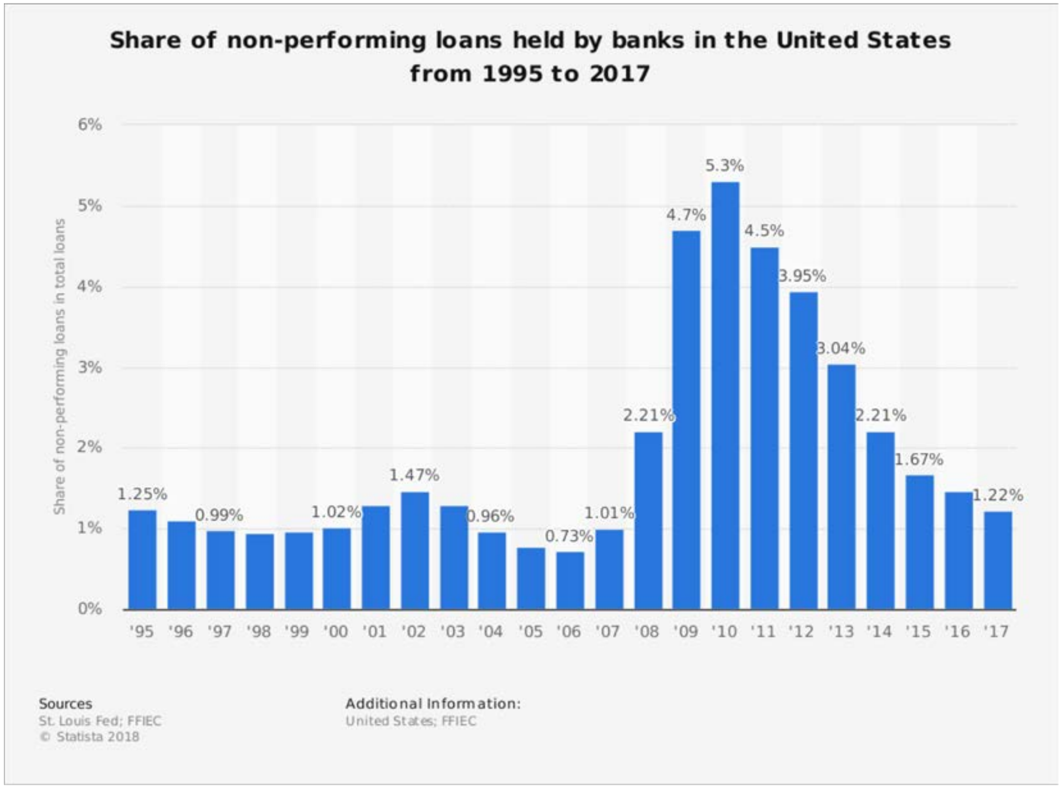
Analytics is everywhere

- ✓ Personalized recommendation/offerings
- ✓ Personal assistants
- ✓ Fraud detection
- ✓ Travels

Data Analytics Applications: Risk Analysis

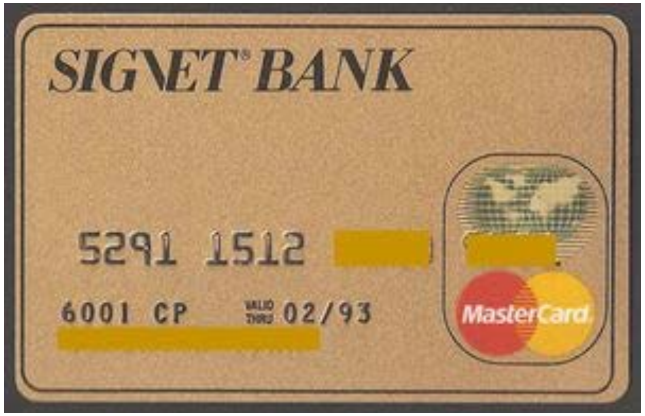
- Business problem: Reduce risk of loans to delinquent customers
- Solution: Use credit scoring models using discriminant analysis to create score functions that separate out risky customers
- Benefit: Decrease in cost of bad debts (non-performing loans)

Today: Information Based Lending



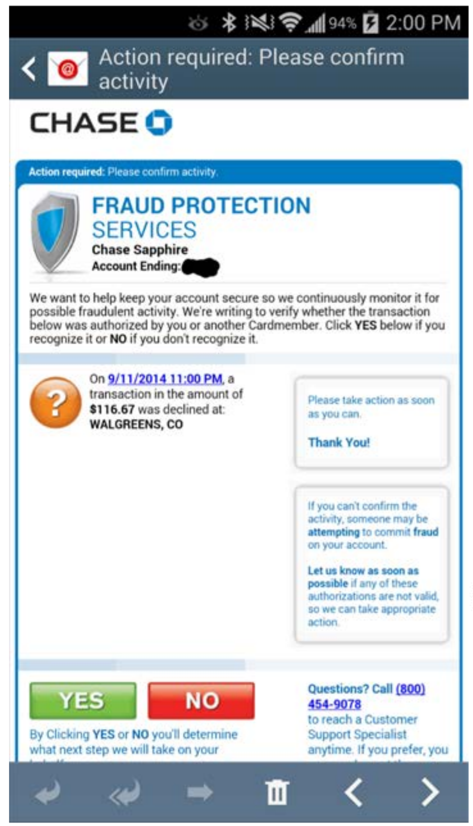
Risk Analysis: Credit Cards

- In the 1980’s credit cards had uniform pricing.
- **Signet**, a small regional bank in Virginia, started thinking about modeling profitability (not just default probability)
  - Offered different terms to different customers (*personalization*)
  - Made better offers to the best customers (*skim the cream*)
- A small proportion of customers actually account for *more than* 100% of a bank’s profit from credit card operations (because the rest are break-even or money-losing)



Data Analytics Applications: Fraud Detection

- Business problem: Fraud increases costs or reduces revenue
- Solution: Use logistic regression, neural nets to identify characteristics of fraudulent cases to prevent in future or prosecute more vigorously
- Benefit: Increased profits by reducing undesirable customers



Fraud Detection: Insurance Analytics

- Opportunity
  - Save and make money by reducing fraudulent auto insurance claims
- Data & Analytics
  - Predictive analytics against years of historical claims and coverage data
  - Text mining adjuster reports for hidden clues, e.g. missing facts, inconsistencies, changed stories
- Results
  - Improved success rate in pursuing fraudulent claims from 50% to 88%; reduced fraudulent claim investigation time by 95%
  - Marketing to individuals with low propensity for fraud

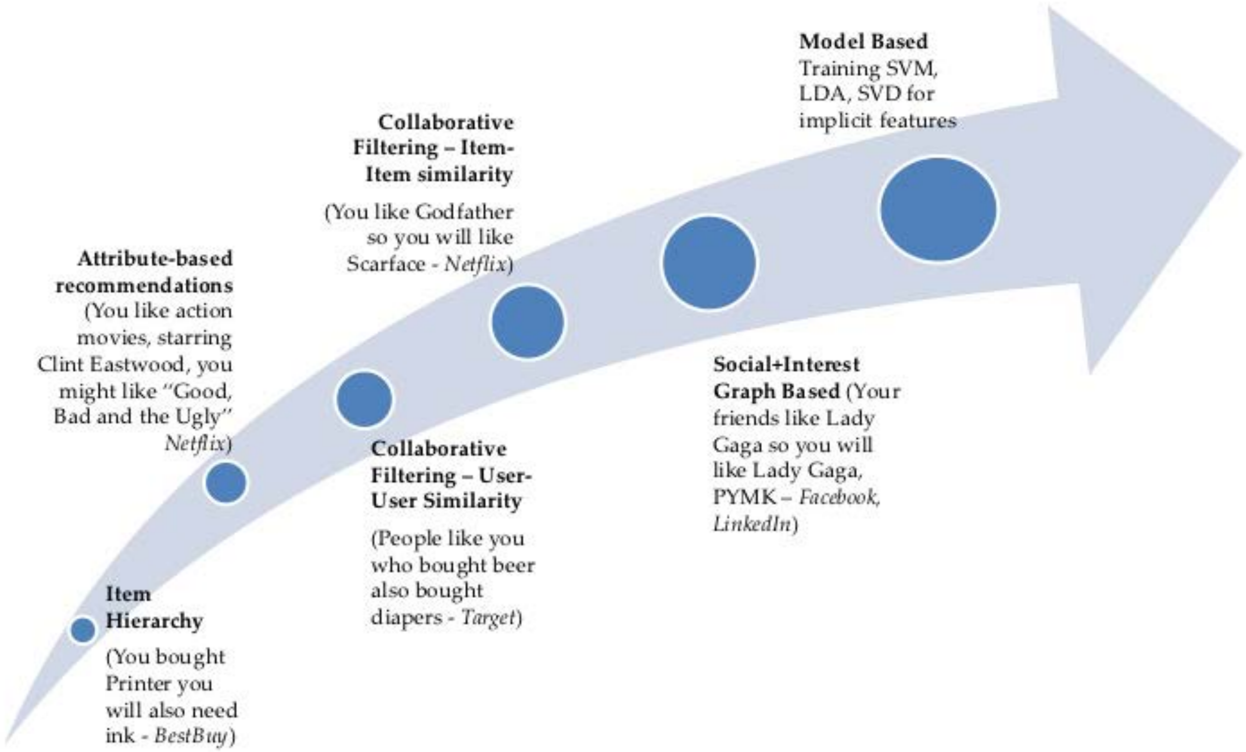
INFINITY.

What "dark data" do you have just laying around that can transform business processes?

Gartner.

Data Analytics Applications: Recommendation System

- Business problem: Users rate items (Amazon Prime, Netflix) on the web. How to use information from other users to infer ratings for a particular user?
- Solution: Use of a technique known as collaborative filtering
- Benefit: Increase revenues by cross selling, up selling



Source: Pinterest

Recommendation System: Netflix

- To help customers find those movies, Netflix developed our world-class movie recommendation system: Cinematch^{SM}.
- It predicts whether someone will enjoy a movie based on how much they liked or disliked other movies.
- Netflix uses those predictions to make personal movie recommendations based on each customer's unique tastes.

- *$1,000,000 Grand Prize!!!*

Recommendation System: Amazon

- Rankings and Recommendations
- Data: *Shopping cart, wish list, previous purchases, items rated and reviewed, geo-location, time-on-site, duration of views, links clicked, telephone inquiries, responses to marketing materials*
- Method and system for anticipatory package shipping

Data Analytics Applications: Marketing

- Business problem: Use list of prospects for direct mailing (or email) campaign.
- Solution: Use data mining to identify the most promising (likely) respondents combining demographic and geographic data with data on past purchase behavior.
- Benefit: Better response rate, savings in campaign cost

Today: Ads on Facebook (especially after your search for a product on Amazon.com)

Prescriptive Analytics: Predicting Demand

Sign In | United States | Contact Us | Search

Industry Solutions | Products | Learn | Support | Customer Stories | Partners | About SAS

Forecasting ATM cash demands

DBS Bank has 80 percent fewer cash-outs, improves process efficiency by 33 percent

In Singapore, customers of banking giant DBS conduct 25 million transactions a month at more than 1,100 ATMs. They rely on DBS – the island nation's largest bank – for convenient, ready access to funds day or night. To make sure that its ATMs – and its customers – don't come up empty-handed (i.e., a cash-out), DBS uses SAS to forecast withdrawal activity and to optimize the reloading process.

As a result, more than 30,000 hours of customer wait time have been eliminated annually as customers are spared the inconvenience of waiting while empty ATMs are reloaded.

Using this innovative solution, a first in the banking world, DBS is now able to convert valuable ATM usage

"We serve over 4 million customers in Singapore, and it is important for us to place customers at the heart of the banking experience across all our touch points," says David Gledhill, Managing Director and Head of Group Technology and Operations at DBS Bank.

"DBS' ATMs have one of the highest utilization rates in the world. Any downtime in a single ATM would mean inconvenience for our customers. Hence, we have to

Challenge

ATM cash-outs dampened the customer experience while increasing the bank's transportation costs.

Solution

- SAS® Forecast Server
- SAS/OR®

Benefits

Accurate forecasts of withdrawal patterns at each machine result in: a 10% reduction in trips to replenish the network, a 33% improvement in the amount of cash sent back to the bank after replenishment, a 80% reduction in cash-outs and more than 350,000 customers a year spared inconvenience.

Prescriptive Analytics: Predicting Demand

- Opportunity
  - Improve heath care and reduce medical costs
- Data & Analytics
  - $5M open contest to predict which patients are most likely to be readmitted to a hospital in the next year, and for how many days
  - Over 10,000 participants and teams
- Result (TBD)
  - Identify advances in diagnoses, treatments, follow-up and release protocols

HERITAGE PROVIDER NETWORK

How can you "gamify" information and analytics to accelerate discoveries?

Gartner

OVERVIEW AND CONCEPTS

Data Analytics

Analytics 101:
Descriptive Analytics

- Provides summary statistics for current and historical data to provide insights into what happened and why
  - Investigating *"associations"* of data
  - *visualization and trend reporting,*
  - *affinity analysis (market basket analysis),*
  - *correlation analysis,*
  - *stylized fact*

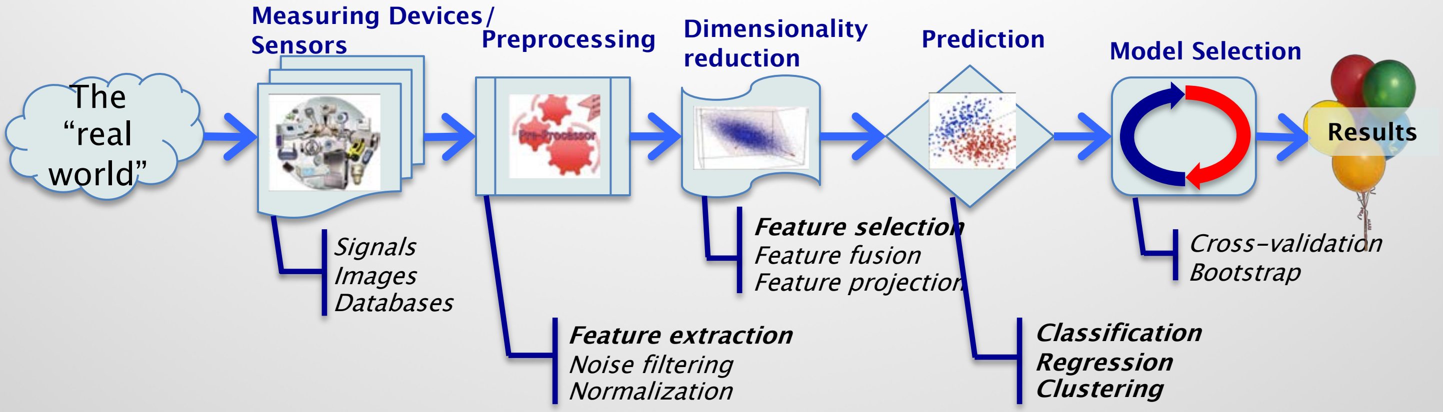
Sample	Feature 1	Feature 2		Feature m
1				
2				
3				
n				

Analytics 101:
Predictive Analytics

- Use machine learning algorithms to build a predictive model from *training (+ validation)* data to make predictions of unseen data *(test data)*.
  - *support vector machines, logistic regression, decision tree, random forest, Bayesian, nearest neighbor, and neural networks*
  - *data = features/predictor variables + response values/target patterns*

Sample	Feature 1	Feature 2		Feature m	Target
1					
2					
3					
n					

Current Landscape: Data Analytics



<u>Analytics/Machine Learning</u>

- Feature Extraction/Engineering
- Feature Selection
- Classification/Regression
- Clustering

Huang and Chaovalitwongse (2015): *Tutorials in Operations Research.*

DESCRIPTIVE ANALYTICS

Statistical Inference

Affinity Analysis

Descriptive Analytics

Univariate Analysis

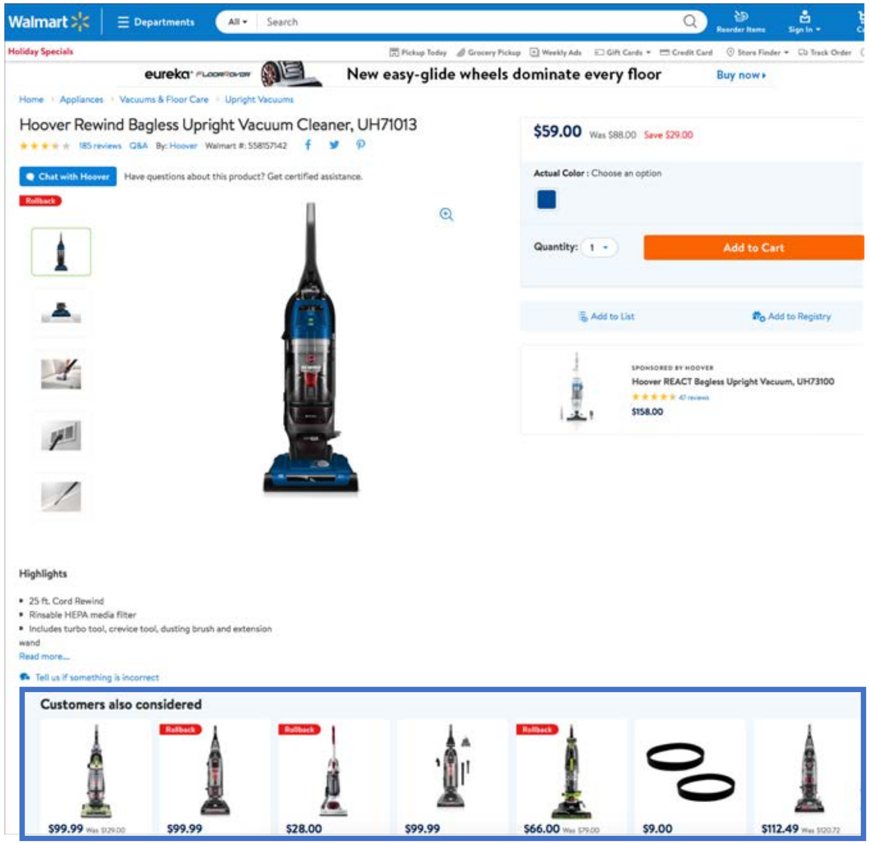
- Describing a single set (column) of data.
- Summary statistics of data features.
- Histogram
  - Number of data points
  - Min; Max
- Central Tendencies
  - Mean; Median; Mode; Quantile
- Dispersion
  - Range; Variance

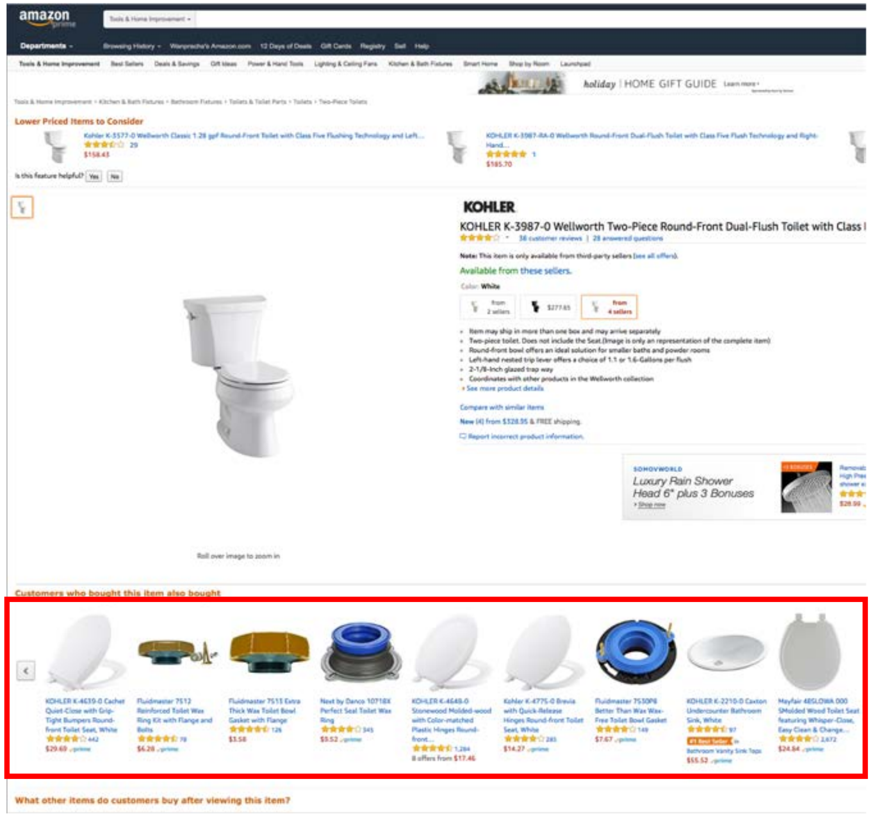
Multivariate Analysis

- Investigates how two or more variables are connected or related.
- Visualization is used to observe the relation between two variables.
- Correlation and covariance are often used to quantify the relation between two variables.
- Covariance measures how two variables vary in tandem from their means correlation
- Multivariate analysis can be extend to analyze time series data.

Affinity Analysis (*aka* association rules, market basket analysis):

Used in many recommender systems





Benefits of Market Basket Analysis

Affinity Positioning

- coffee, coffee makers in close proximity

selection of promotions, merchandising strategy

- sensitive to price: Italian entrees, pizza, pies, Oriental entrees, orange juice

uncover consumer spending patterns

- correlations: orange juice & waffles

joint promotional opportunities

Cross-Selling

- cold medicines, kleenex, orange juice
- Monday Night Football kiosks on Monday p.m.

Customer 1: *beer, pretzels, potato chips, aspirin*
Customer 2: *diapers, baby lotion, grapefruit juice, baby food, milk*
Customer 3: *soda, potato chips, milk*
Customer 4: *soup, beer, milk, ice cream*
Customer 5: *soda, coffee, milk, bread*
Customer 6: *beer, potato chips*

Co-occurrence Table

- Let's focus on *beer, potato chips, milk, diapers, and soda*

	Beer	Pot Chips	Milk	Diapers	Soda
Beer	3	2	1	0	0
Pot Chips	2	3	1	0	1
Milk	1	2	4	1	2
Diapers	0	0	1	1	0
Soda	0	1	2	0	2

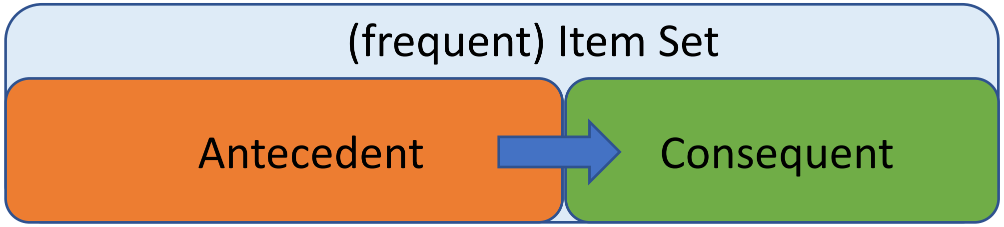
Market Basket: Example

"IF" part = **antecedent;** "THEN" part = **consequent**

"Item set" = the items (e.g., products) comprising the antecedent or consequent

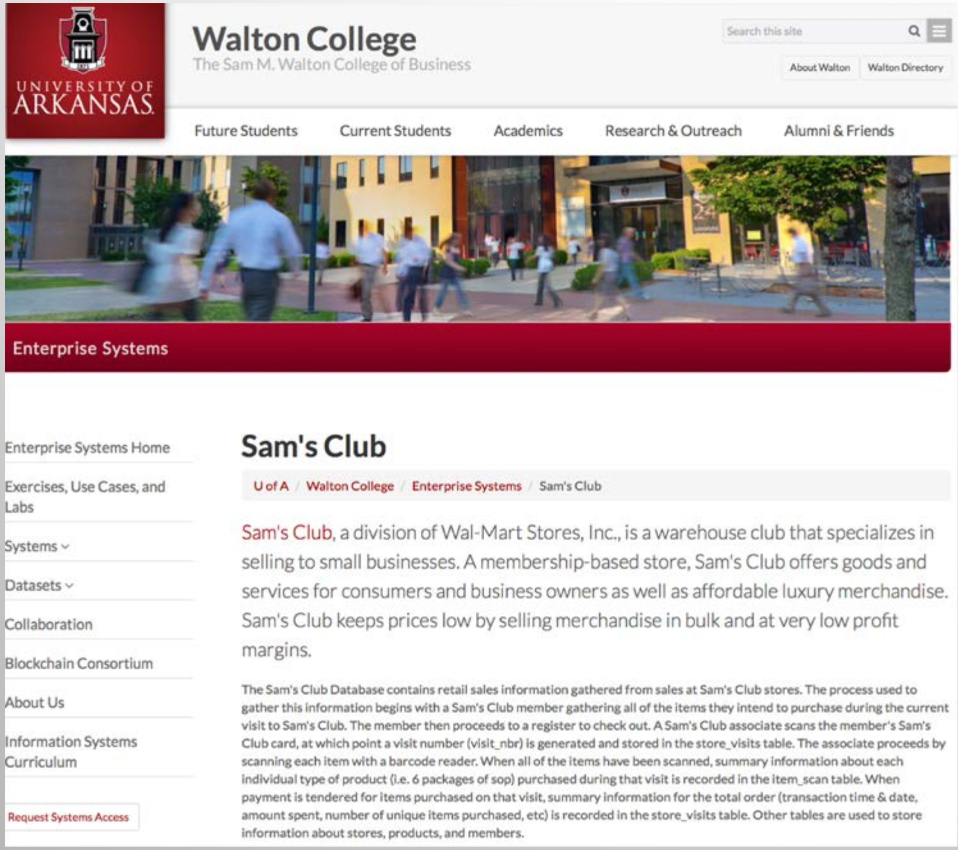
Use Apriori algorithm to find frequent item sets (*support*)

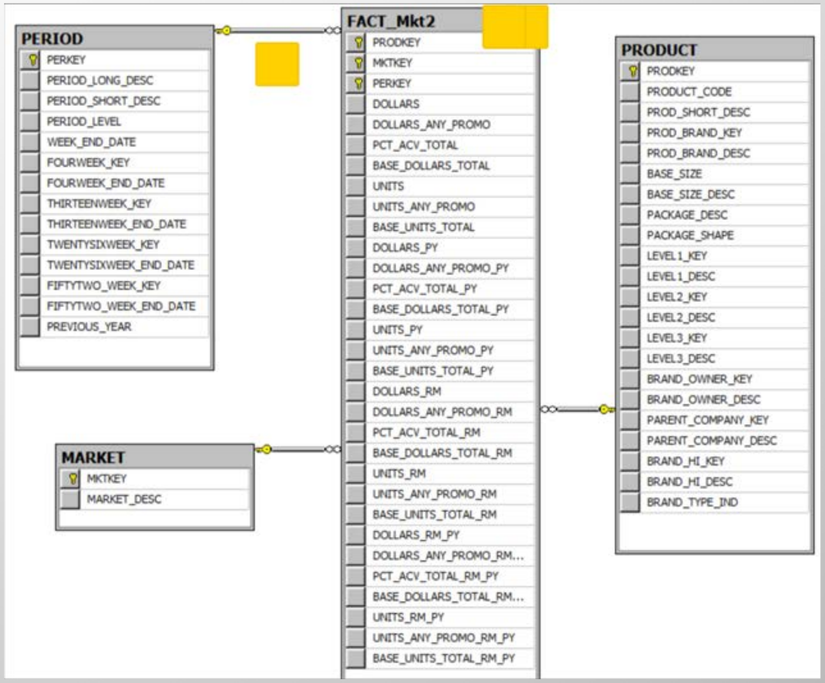
Mine rules with high *confidence* or *lift*



Association Rule Mining

In-House Data at UA



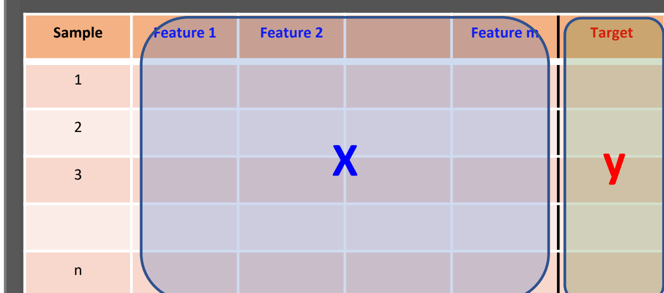


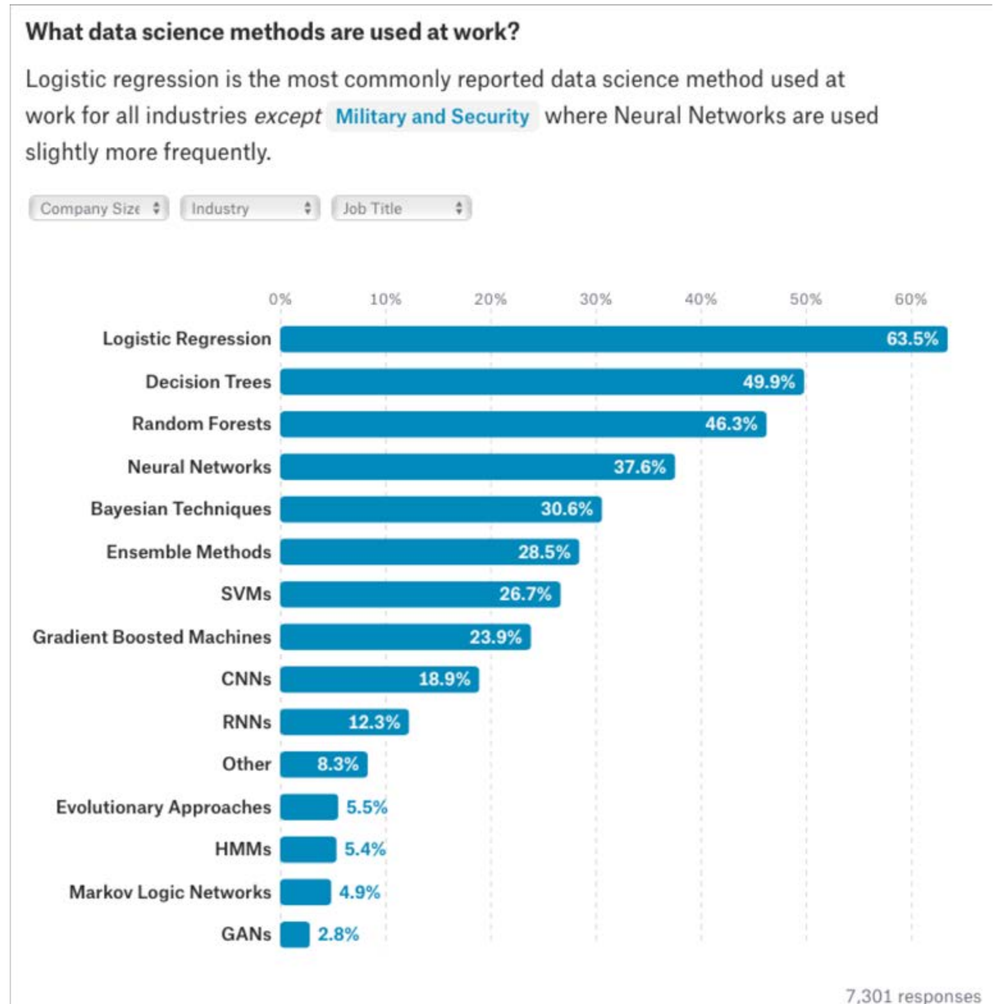
PREDICTIVE ANALYTICS

Machine Learning Algorithms:

Classification and Regression

Popularity of Machine Learning Methods





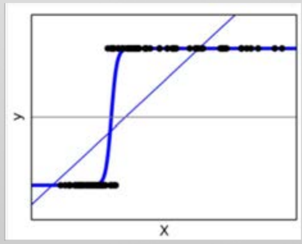
Data from Kaggle.com

Logistics Regression

- Odds – like probability.
- Odds are usually written as “5 to 1 odds” which is equivalent to 1 out of five or .20 probability, etc.
- *Here we consider Posterior Probability = P(y|X)*
- Odds ratio – the ratio of the odds over 1, e.g., the probability of winning over the probability of losing.

- Logit – this is the natural log of an odds ratio. The logit scale is linear (probability is not) and functions much like a z-score scale.

$$\ln\left(\frac{P(Y \mid X)}{1-P(Y \mid X)}\right) = \beta_o + \beta_1 X$$



- Moving things around, the logistic regression model is given by

$$P(Y \mid X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

Naïve Bayes Classifier

- The naive Bayes classifier is designed for use when predictors are independent of one another within each class.

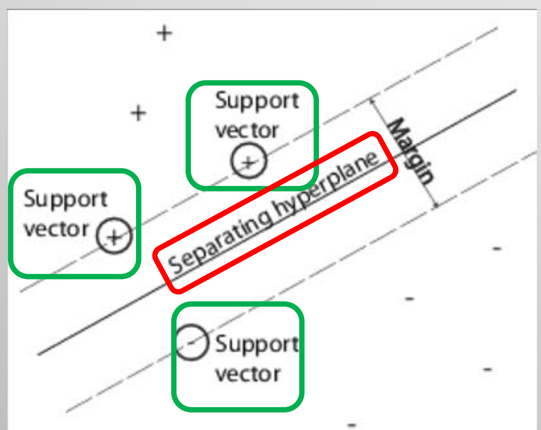
$$P\left(\omega_j \middle| X\right) = \frac{P\left(X \middle| \omega_j\right) \cdot P\left(\omega_j\right)}{P\left(X\right)}$$

Posterior = (Likelihood x Prior) / Evidence

- The key idea is to estimate the distributions.
  - Normal (Gaussian) Distribution
  - Kernel Distribution
  - Multinomial Distribution
  - Multivariate Multinomial Distribution

Support Vector Machines:
Concepts and Models

- An SVM classifies data by finding **the best hyperplane** that separates all data points of one class from those of the other class.
- The **support vectors** are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab.
- The best hyperplane for an SVM is the one with **the largest margin** between the two classes.
  - Margin means the maximal width of the slab parallel to the hyperplane.



$$\min_{w,b,s}\left(\boxed{\frac{1}{2}\langle w,w\rangle}+C\sum_i s_i\right)$$

such that

$$y_i\left(\boxed{\langle w,x_i\rangle+b}\right)\ge 1-s_i$$

$$s_i\ge 0.$$

Penalty to move misclassified points to where they belong

$$L_P=\frac{1}{2}\langle w,w\rangle+C\sum_i s_i-\sum_i\alpha_i\left(y_i\left(\langle w,x_i\rangle+b\right)-\left(1-s_i\right)\right)-\sum_i\mu_i s_i,$$

Support Vector Machines: *Solution Approaches*

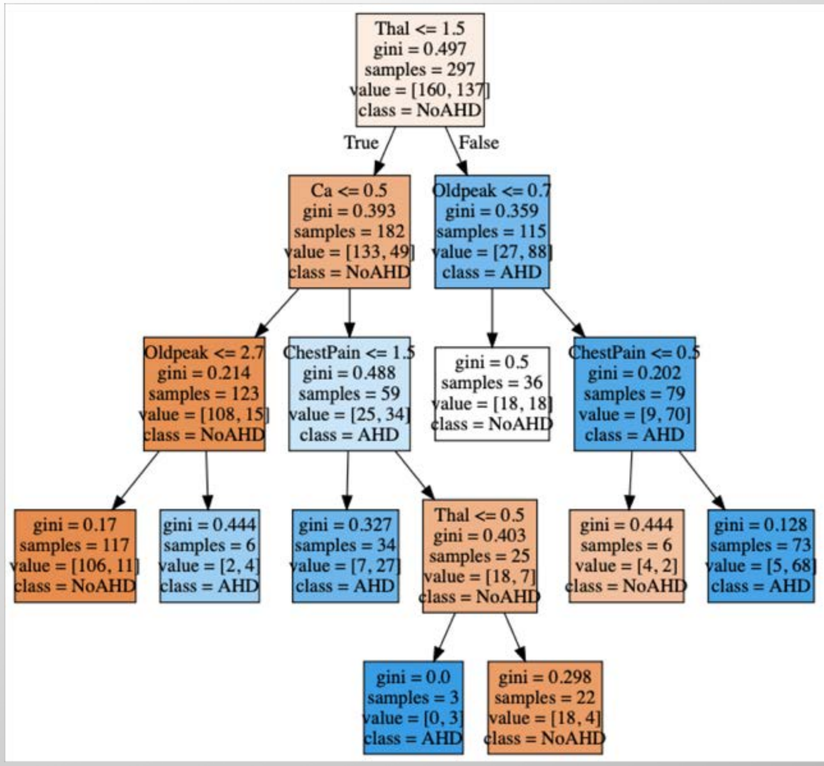
- Sequential Minimal Optimization (SMO) minimizes the one-norm problem by a series of two-point minimizations.

- Iterative Single Data Algorithm (ISDA) solves the one-norm problem using a series on one-point minimizations but does not respect the linear constraint, and does not explicitly include the bias term in the model.

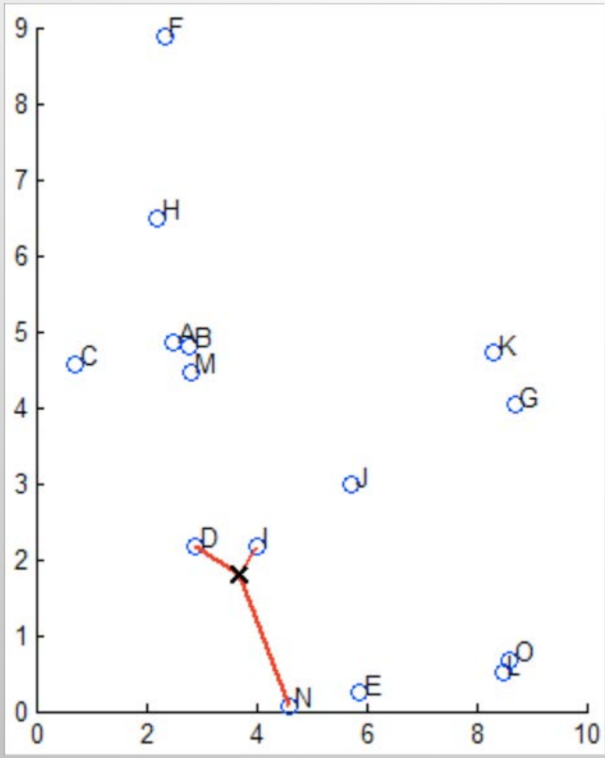
- You can solve the one-norm problem using any quadratic programming solver (e.g., *quadprog* in Matlab's Optimization Toolbox)

Classification/Regression Tree

- Start with all input data, and examine all possible binary splits on every feature.
- Select a split with best optimization criterion.
  - Gini's Diversity Index: $1 - \sum_i p^2(i)$
  - Entropy (information): $-\sum_i p(i) \log p(i)$
  - subject to the MinLeaf constraint – min # of observations in the child node
- Impose the split.
- Repeat recursively for the two child nodes.



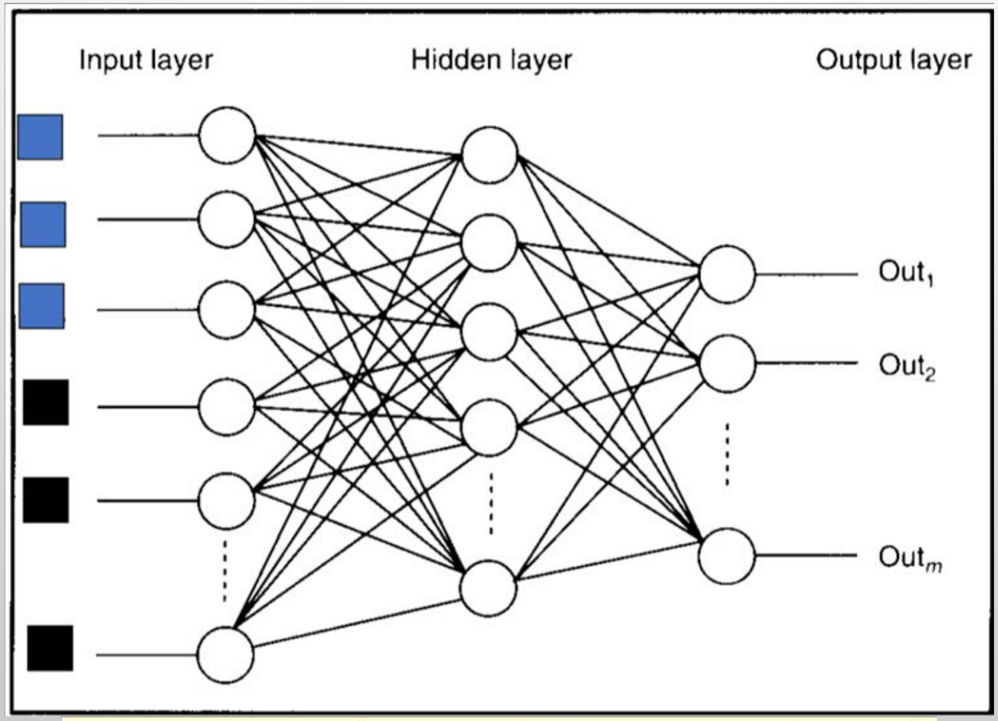
Nearest Neighbor



Key factors:

- # of neighbors
- Distance matrix
  - Euclidean
  - Mahalanobis
  - Cosine
  - Correlation
  - Spearman
  - Hamming
  - Jaccard

Neural Network/Deep Learning



$$g_k(x) \equiv z_k = f\left(\sum_{j=1}^{n_H} w_{kj} f\left(\sum_{i=1}^{d} w_{ji} x_i + w_{j0}\right) + w_{k0}\right)$$

A good-rule-of-thumb

Algorithm	Predictive Accuracy	Fitting Speed	Prediction Speed	Memory Usage	Easy to Interpret	Handles Categorical Predictors
Trees	Medium	Fast	Fast	Low	Yes	Yes
SVM	High	Medium	*	*	*	No
Naive Bayes	Medium	**	**	**	Yes	Yes
Nearest Neighbor	***	Fast***	Medium	High	No	Yes***
Discriminant Analysis	****	Fast	Fast	Low	Yes	No

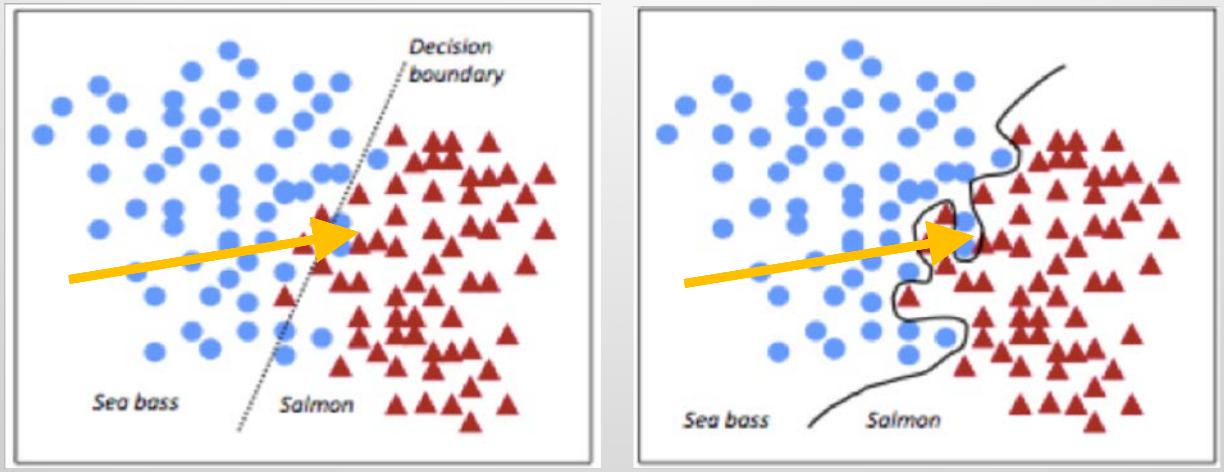
- * — **SVM** prediction speed and memory usage are good if there are few support vectors, but can be poor if there are many support vectors.
- ** — **Naive Bayes** speed and memory usage are good for simple distributions, but poor for kernel distributions and large data sets.
- *** — **Nearest Neighbor** usually has good predictions in low dimensions, but poor predictions in high dimensions. Nearest Neighbor can have either continuous or categorical predictors, but not both.
- **** — **Discriminant Analysis** is accurate when the modeling assumptions are satisfied (multivariate normal by class). Otherwise, the predictive accuracy varies.

PREDICTIVE ANALYTICS

Feature Selection and Regularization

Why is feature selection so important/challenging?

Interpretable/generalizable prediction model



Occam's razor **(law of parsimony)**

- simplicity is a goal in itself
- simplicity leads to greater accuracy
- simplicity leaders better generalization

Practical Decision Models

- When the number of features far exceeds the number of samples

The # of predictor vars (*p*) >> The # of observations (*n*)
The # of unknown vars >> The # of linear equations

Ill-posed problem = Overfitting

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- This is quite common as we *never collect enough data samples*
- *Feature Selection* is thus used to construct **generalizable decision models**

Feature Selection vs. Feature Transformation

- Feature transformation methods create new features (*predictor variables*) that are hoped to have a descriptive power that is more easily ordered than the original features
  - Principal component analysis
  - Independent component analysis
  - Factor analysis

- Feature selection reduces the dimensionality of data by selecting only a subset of measured features to create a model.
  - Preferable when original features are important and the modeling goal

Feature Selection:
Combinatorial optimization problem

X = features; y = response values; w = weights (decision variables)

$(\sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))))$

Logistic loss

$(\sum_{i=1}^{n} \max(1 - y_i(\mathbf{w}^T\mathbf{x}_i), 0)^2)$

Hinge loss

# of features with non-zero weight

$\|\mathbf{w}\|_0 = \sum_{j=1}^{m} I_{\{w_j \neq 0\}}$

- Filter approach
  - Screening/removing irrelevant features using a pre-determined criterion (e.g., *FDR – False Discovery Rate*)
- Wrapper approach – (greedy approach)
  - Heuristic method to iteratively search for the (local) best combination of features that optimizes a pre-determined criterion (e.g., stepwise, sequential – *knapsack heuristic*)
- Embedded approach
  - Integrate feature selection with prediction model (e.g., LASSO)

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \gamma\|\mathbf{w}\|_0, \text{ where } \gamma \in \mathbb{R} \text{ is a regularization parameter}$$

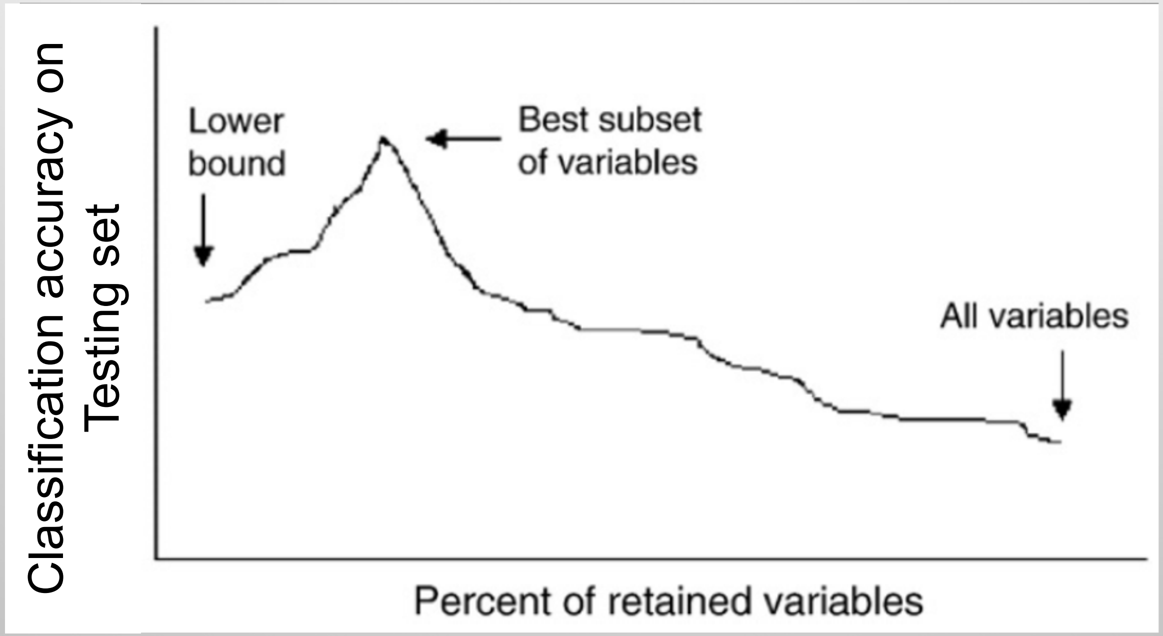
Feature Selection:
Sequential Approach

- Stepwise (sequential) procedure
  - Sequential forward selection (SFS): features are sequentially added to an empty candidate set until the addition of further features does not decrease the **criterion**.
  - Sequential backward selection (SBS): features are sequentially removed from a full candidate set until the removal of further features increase the **criterion**.

- Used **criterion** is often based on statistical significance in
  - Correlation with trained targets (e.g., *partial least square, regression weights*)
  - Separation between two classes (e.g., *t-test, mutual information, Fisher's, Chi's square*)

Sequential Feature Selection:

Hypothetical testing accuracy profile with sequential feature selection



Sequential Feature Selection:
Multicriteria selection

European Journal of Operational Research 218 (2012) 97–105

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

ELSEVIER

Production, Manufacturing and Logistics

Multicriteria variable selection for classification of production batches

Michel J. Anzanello[a,*], Susan L. Albin[b,1], Wanpracha A. Chaovalitwongse[b,2]

[a] Department of Industrial Engineering, Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99, 5 andar, Porto Alegre, Brazil
[b] Department of Industrial and Systems Engineering, Rutgers University, 96 Frelinghuysen Road, CoRE Building, Room 201, Piscataway, NJ, USA

ARTICLE INFO

Article history:
Received 20 January 2011
Accepted 19 October 2011
Available online 26 October 2011

Keywords:
Multivariate statistics
Variable selection
Multiple criteria
Data mining
Batch manufacturing

ABSTRACT

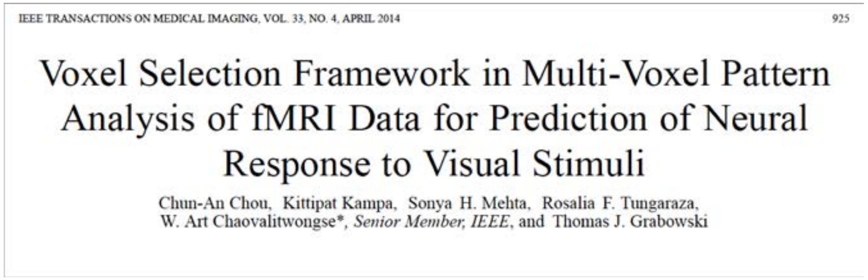
In many industrial processes hundreds of noisy and correlated process variables are collected for monitoring and control purposes. The goal is often to correctly classify production batches into classes, such as good or failed, based on the process variables. We propose a method for selecting the best process variables for classification of process batches using multiple criteria including classification performance measures (i.e., sensitivity and specificity) and the measurement cost. The method applies Partial Least Squares (PLS) regression on the training set to derive an importance index for each variable. Then an iterative classification/elimination procedure using *k*-Nearest Neighbor is carried out. Finally, Pareto analysis is used to select the best set of variables and avoid excessive retention of variables. The method proposed here consistently selects process variables important for classification, regardless of the batches included in the training data. Further, we demonstrate the advantages of the proposed method using six industrial datasets.

Respiratory trace feature analysis for the prediction of respiratory-gated PET quantification

Shouyi Wang[1,2], Stephen R Bowen[3,4], W Art Chaovalitwongse[1,2], George A Sandison[4], Thomas J Grabowski[2,3] and Paul E Kinahan[3]

[1] Department of Industrial and Systems Engineering, 3900 Stevens Way, Seattle, WA 98195, USA
[2] Integrated Brain Imaging Center, 1959 NE Pacific St, Seattle, WA 98195, USA
[3] Department of Radiology, 1959 NE Pacific St, Seattle, WA 98195, USA
[4] Department of Radiation Oncology, 1959 NE Pacific St, Seattle, WA 98195, USA

Sequential Feature Selection:
Prediction of Visual Stimuli

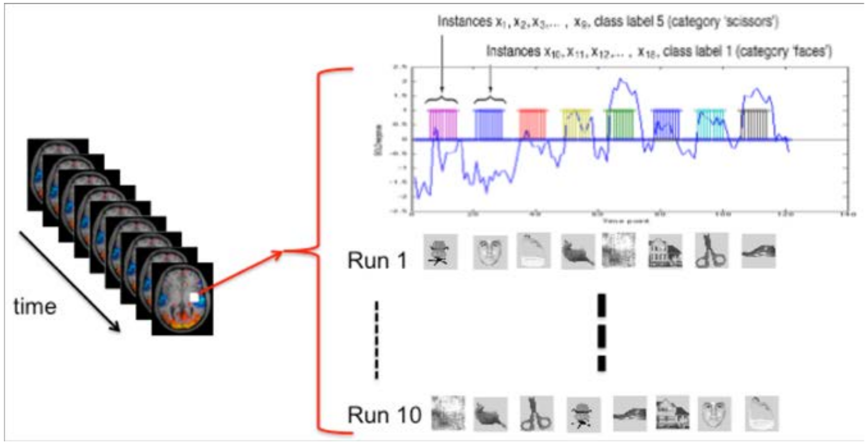


IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 33, NO. 4, APRIL 2014 925

Voxel Selection Framework in Multi-Voxel Pattern Analysis of fMRI Data for Prediction of Neural Response to Visual Stimuli

Chun-An Chou, Kittipat Kampa, Sonya H. Mehta, Rosalia F. Tungaraza, W. Art Chaovalitwongse*, *Senior Member, IEEE*, and Thomas J. Grabowski

- There are 10 runs (blocks), each producing 121 fMRI data points.
- Each block displayed image exemplars from all 8 conceptual categories: *1) face, 2) house, 3) cat, 4) bottle, 5) scissor, 6) shoe, 7) chair, and 8) 'scrambled picture'*.



Regularization:
Objective Function

- Common objective function of prediction model (regression/classification):

  Minimize *Classification/Regression Error*

- ***Regularization*** – process of introducing a penalty term in the objective function to avoid overfitting in an ill-posed problem

  Minimize *Classification/Regression Error +* Penalty

- Prediction Error
  - Regression error: L-1, L-2 norms
  - Classification error: L-0 norm (logistic regression, hinge loss)
- Penalty on the feature weights
  - Continuous: L-1, L-2 norms
  - Discrete: L-0 norms (control the # of features/vars)

Least Absolute Shrinkage and Selection Operator (Lasso)

- Lasso (Tibshirani, 1996) is a very popular technique for variable selection for high-dimensional data.
  - a shrinkage and selection method for linear regression that minimizes the sum of squared errors, with a L1-norm penalty

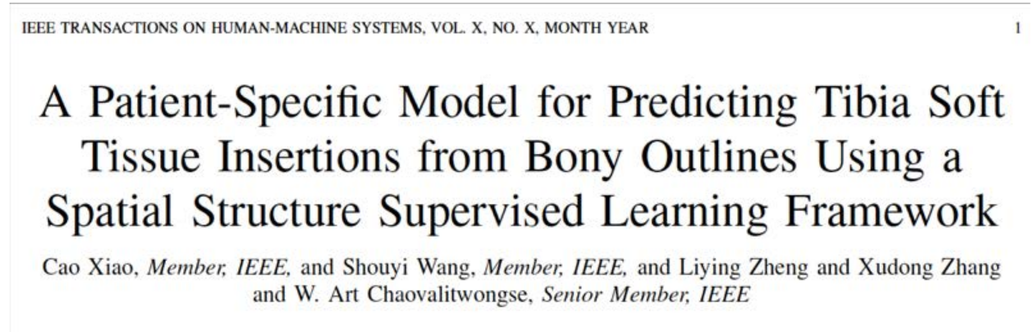
$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2} ||y - X\beta||_2^2 + \lambda ||\beta||_1$$

Lasso vs. Ridge regression vs. Elastic net

$$\lambda \sum_{j=1}^{p} |\beta_j| \qquad \lambda \sum_{j=1}^{p} \beta_j^2 \qquad \lambda \sum_{j=1}^{p} (\alpha |\beta_j| + (1-\alpha)\beta_j^2)$$

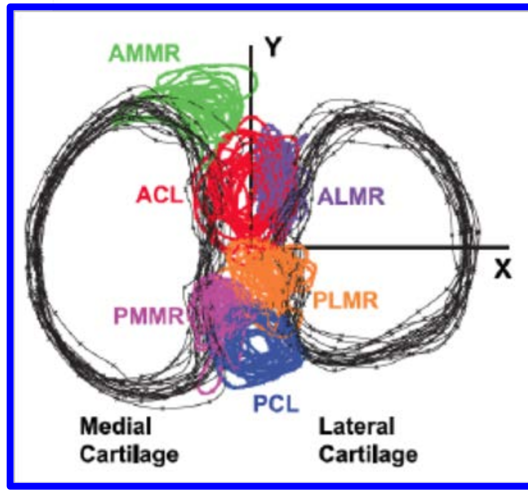
- If the loss function is replaced by hinge loss, it is L-1 norm SVM
- If the loss function is replaced by logistic function, it's called logistic regression

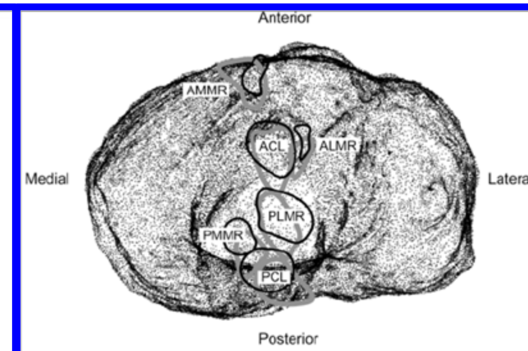
Regularization: *Prediction of Soft Tissue Locations*

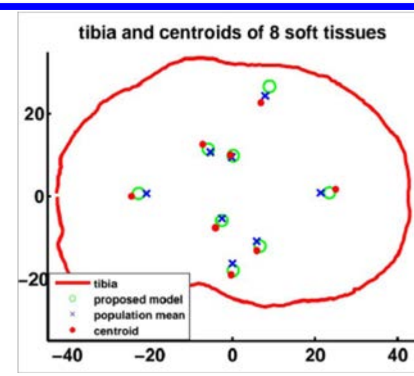

IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. X, NO. X, MONTH YEAR 1

A Patient-Specific Model for Predicting Tibia Soft Tissue Insertions from Bony Outlines Using a Spatial Structure Supervised Learning Framework

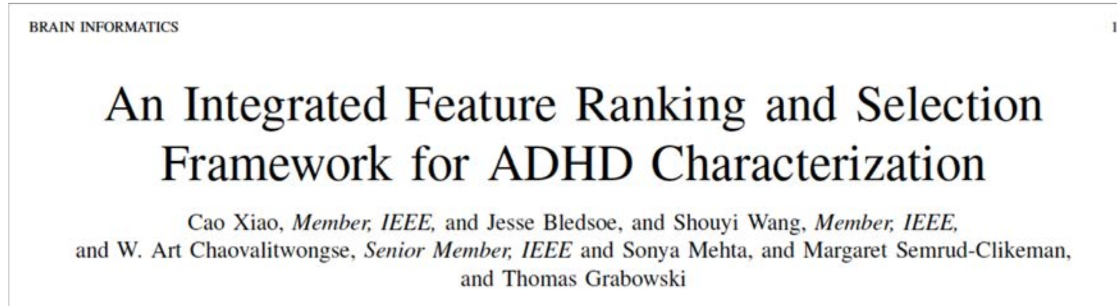
Cao Xiao, *Member, IEEE,* and Shouyi Wang, *Member, IEEE,* and Liying Zheng and Xudong Zhang and W. Art Chaovalitwongse, *Senior Member, IEEE*





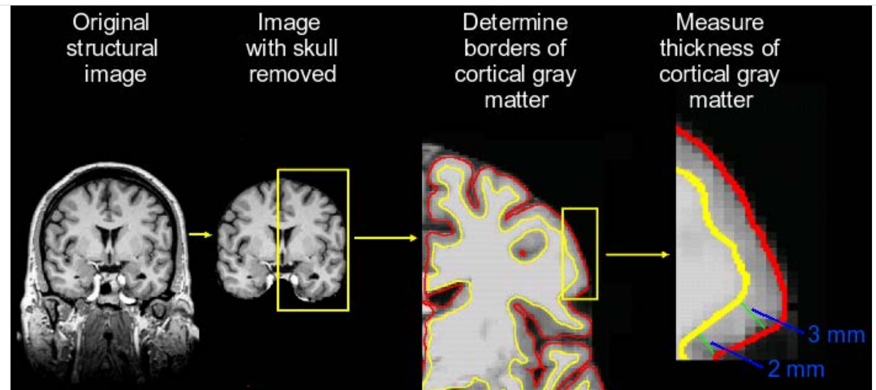


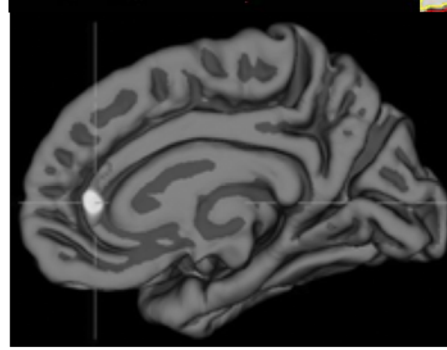
Regularization: *Prediction of ADHD Diagnosis*


BRAIN INFORMATICS 1

An Integrated Feature Ranking and Selection Framework for ADHD Characterization

Cao Xiao, *Member, IEEE,* and Jesse Bledsoe, and Shouyi Wang, *Member, IEEE,* and W. Art Chaovalitwongse, *Senior Member, IEEE* and Sonya Mehta, and Margaret Semrud-Clikeman, and Thomas Grabowski





Right rostral anterior cingulate cortex (Right ACC): significant differences between ADHD and control groups.

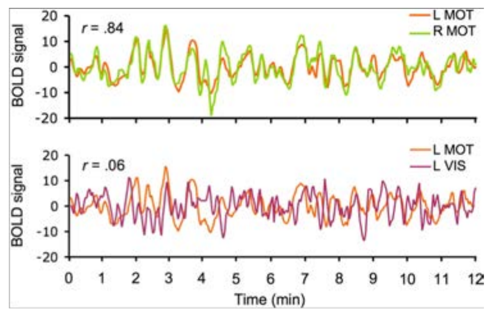
Regularization: *Prediction of Brain Diagnosis*

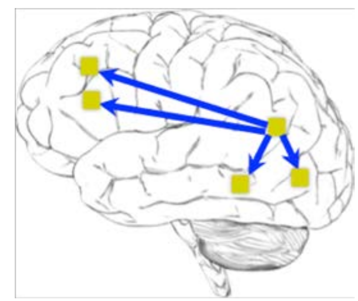
Convex Optimization for Group Feature Selection of Networked Data

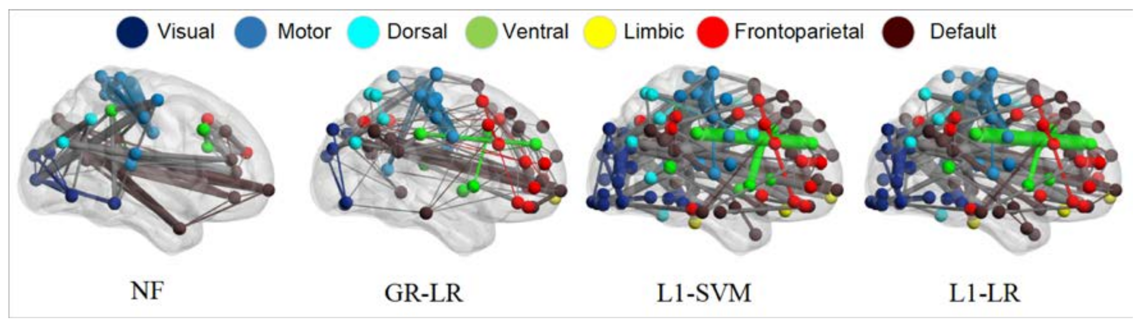
Daehan Won
Systems Science & Industrial Engineering Department, Binghamton University, the State University of New York, NY, dhwon@binghamton.edu

Hasan Manzour
Department of Industrial & Systems Engineering, University of Washington, Seattle, WA, hmanzour@uw.edu

W. Art Chaovalitwongse
Institute for Advanced Data Analytics, Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, artchao@uark.edu







Credits

- “Decision Model for Patient-Specific Motion Management in Radiation Therapy Planning”
- “Network Optimization of Functional Connectivity in Neuroimaging for Differential Diagnoses of Brain Diseases”
- “Computational Framework of Robust Intelligent System for Mental State Identification and Human Performance Prediction with Biofeedback”

National Science Foundation
WHERE DISCOVERIES BEGIN

- “IBIC: Integrated Brain Imaging Center for the University of Washington”

NIH National Institutes of Health
Turning Discovery Into Health

- “Continuous Assessment of Cognitive Load in Information Seeking”

INSTITUTE of Museum and Library SERVICES

Thank you


obrigado
Dank U
Merci
mahalo
Köszi
спасибо
Grazie
Thank you
mauruuru
Takk
Gracias
Dziękuję
Děkuju
danke
Kiitos

M.S. in Operations Management At a Glance:

- Online and Live Course Options
- 30 Credit Hours (10 Graduate Courses)
  - With up to 4 pre-requisite classes
- Five 8-week Sessions Per Year
- No GRE/GMAT required with 3.0 Bachelor's GPA
- Total Program Cost is $12,000 to $15,000 (depending on pre-reqs needed)
- Can be completed in one year, but you have up to six years to complete

Project Management Graduate Certificate At a Glance:

- Online and Live Course Options
- 12 Credit Hours (4 Graduate Courses)
- 8-week sessions
- Five Enrollment Periods: Aug, Oct, Jan, Mar, May
- Entire Program Cost: Approximately $5,000
- 2.5 GPA with a Bachelor's Degree required for admission
- Certificate courses can also count toward MSOM degree

Full Webinar Schedule:

Date:	Webinar Title:	Presenter:
Wednesday, August 29th	*Presenting to Sr. Decision Makers: Clear, Concise, & Complete*	Kirk Michealson
Tuesday, September 25th	*Leading Through Change **Live Presentation at Walmart Home Office***	Travis McNeal
Thursday, October 25th	*Stop "Droning" on about Unmanned Aircraft Systems and Do Something About It*	Dr. Ham
Tuesday, November 27th	*Introduction to Data Analytics and Emerging Real-World Use Cases*	Dr. Chaovalitwongse
Tuesday, December 18th	*Group Facilitation*	Terry Bresnick
Wednesday, January 23rd	*Machine Learning*	Dr. Rainwater
Thursday, February 21st	*Project Selection: The $1 Trillion Decisions*	Leonard Nethercutt
Wednesday, March 27th	*BlockChain*	Dr. Ed Pohl
Thursday, April 25th	*An Engineered Approach to Site Selection: Determining Where Facilities Should Be Located*	Kerry Melton

Thanks for attending!

- For information about our flexible degree program options, email Mindy Hunthrop, hunthrop@uark.edu.

- The video from today's webinar will be available on our website within about a week, *registered* participants will receive an email with the video link.

- We hope to see you online next month!