Toronto Metropolitan University | TED ROGERS SCHOOL OF MANAGEMENT | Centre for Digital Enterprise Analytics and Leadership

Business Analytics and Data Driven Decision Making

Session#11: Lecture#21: Predictive Analytics: Causal Inferencing

Ravi Vatrapu

Director, Centre for Digital Enterprise Analytics and Leadership (DEAL)

Loretta Rogers Research Chair in Digital Enterprise

Professor, School of Information Technology Management

Ted Rogers School of Management

Toronto Metropolitan University, Canada

Founding Director, Centre for Business Data Analytics (bda.cbs.dk)

Professor (on leave), Copenhagen Business School, Denmark

Visiting Professor, Indian Institute of Management Visakhapatnam, India

Honorary Visiting Professor, GITAM Deemed University, India

Adjunct Faculty, Indian Institute of Management Rohtak, India

Email: vatrapu@torontomu.ca

DEAL Website: https://www.torontomu.ca/tedrogersschool/digital-enterprise-analytics-and-leadership/

Faculty Webpage: https://www.torontomu.ca/information-technology-management/faculty-research/ravi-vatrapu/

IIM VISAKHAPATNAM

GITAM DEEMED TO BE UNIVERSITY

INDIAN INSTITUTE OF MANAGEMENT ROHTAK

CBS CENTRE FOR BUSINESS DATA ANALYTICS

Data Mining: General Methodological Process

Cioffi-Revilla, C. (2013). Introduction to Computational Social Science: Principles and Applications: Springer Science & Business Media.

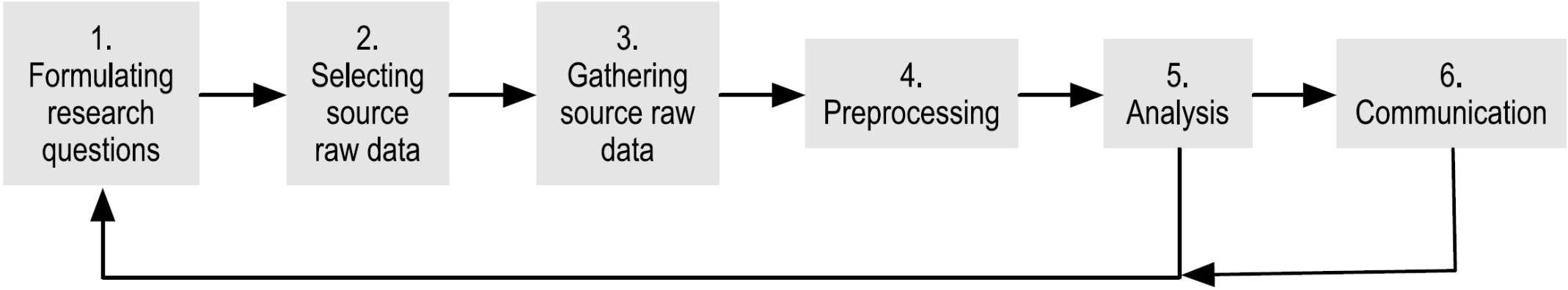
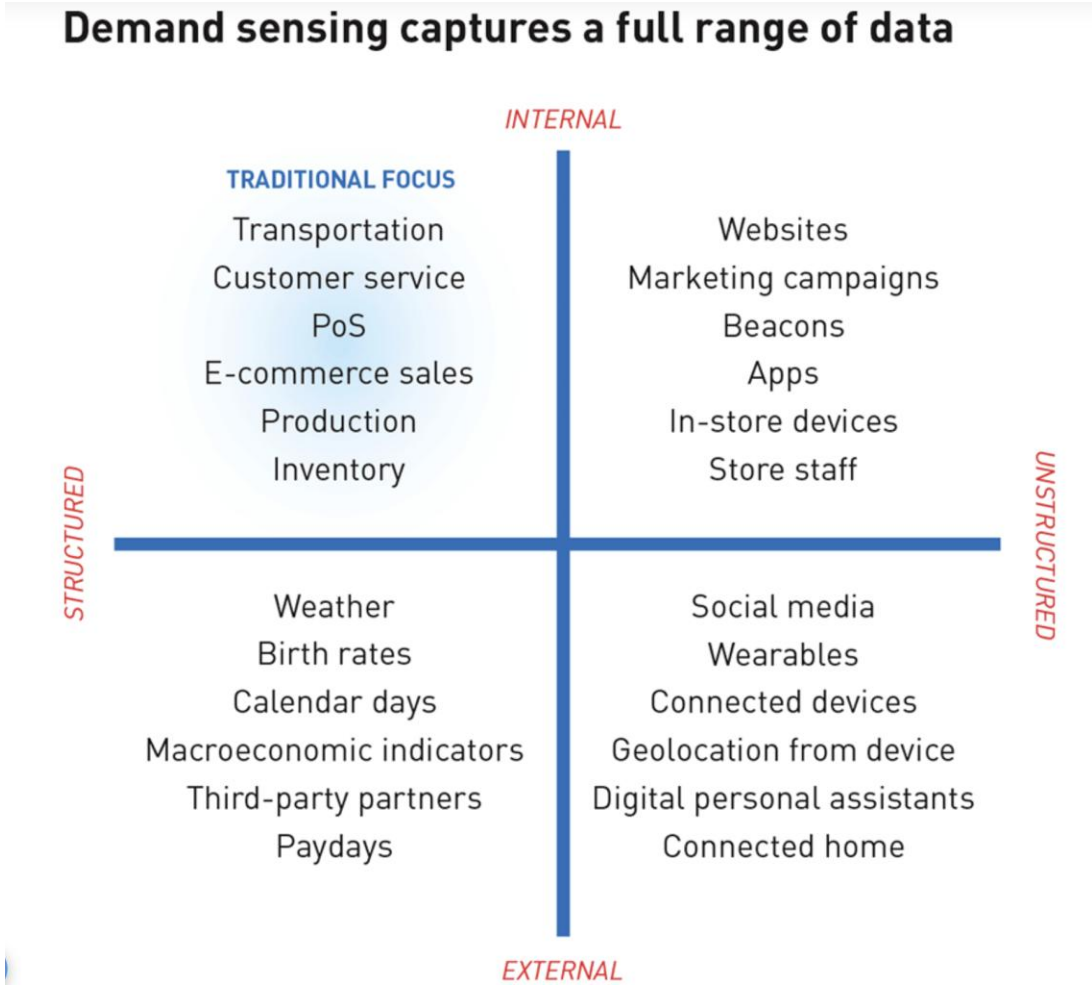


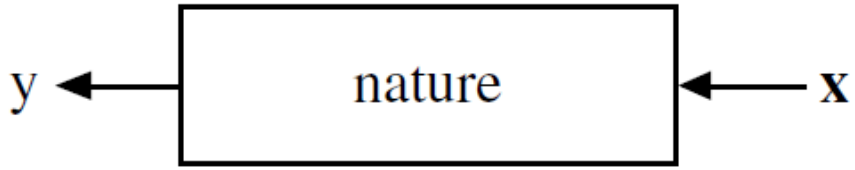
Fig. 3.5 *General Data Mining Methodological Process.* Data mining for automated information extraction involves several stages, the most important being the six highlighted here and discussed below. The core is Analysis for answering research questions, but the other five stages are just as critical for overall quality of the scientific investigation. Each of the six stages involves a variety of procedures, most of them dependent on the research questions being addressed

Data Matrix



https://www.strategy-business.com/article/The-Magic-of-Predicting-Demand-from-Data?gko=94906

Predictive Analytics: Two Cultures

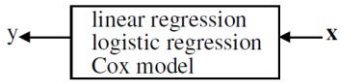


Information. To extract some information about how nature is associating the response variables to the input variables.

Prediction. To be able to predict what the responses are going to be to future input variables;

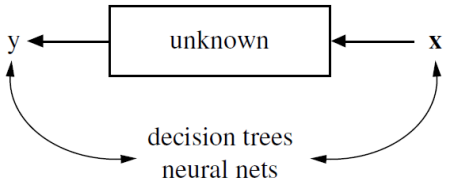
Data Modeling Culture

response variables = f(predictor variables, random noise, parameters)



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

Algorithmic Modeling Culture



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Breiman, L. (2001). Statistical Modelling: The Two Cultures. Statistical Science, Vol. 16(3): 199-231

Predictive Analytics: Three Aspects

- ***Rashomon:*** multiplicity of good models
- ***Occam:*** conflict between simplicity and accuracy
- ***Bellman:*** dimensionality—curse or blessing?

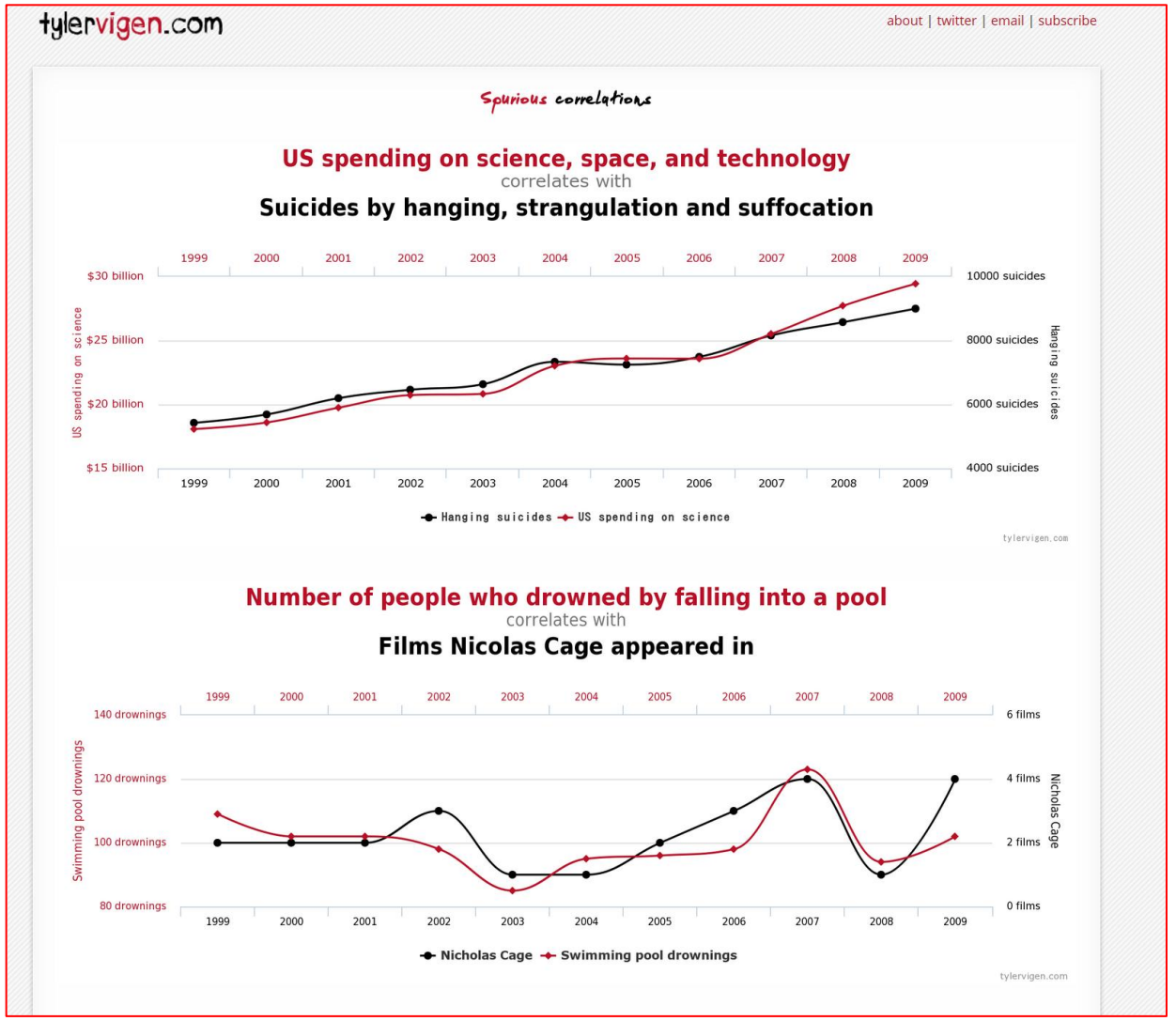
Breiman, L. (2001). Statistical Modelling: The Two Cultures. Statistical Science, Vol. 16(3): 199-231

Correlation vs. Causation

Correlation is NOT necessarily Causation

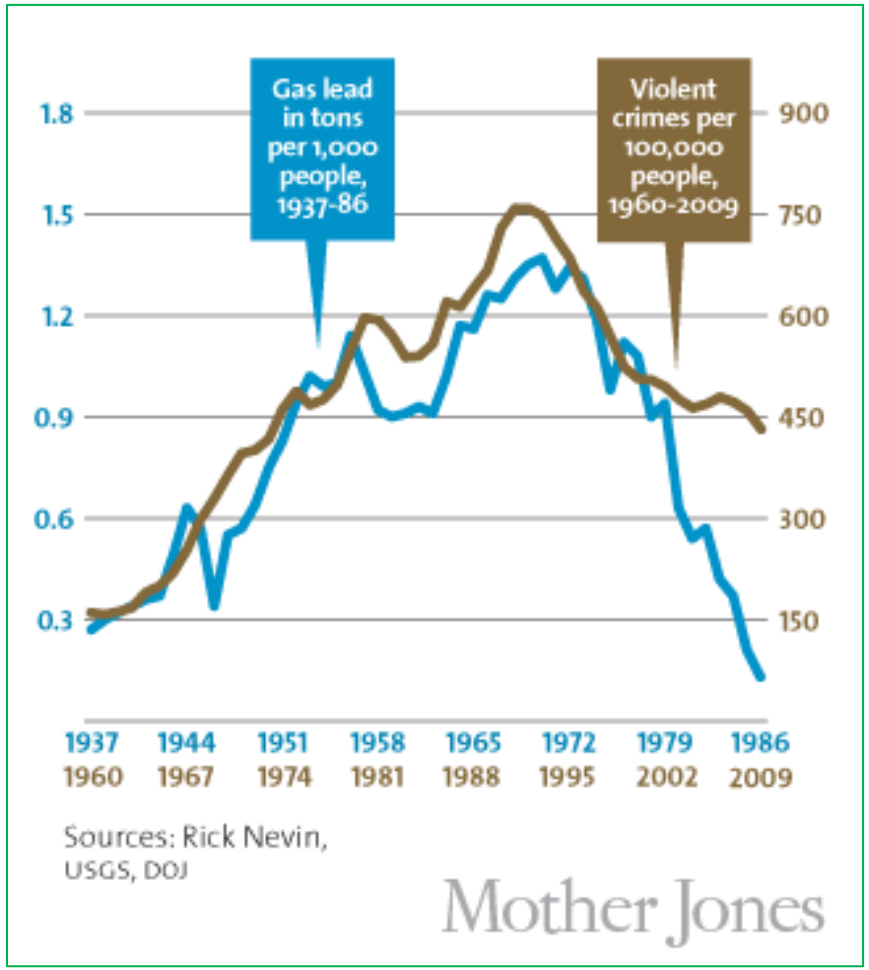
NO Causation without Correlation

Spurious Correlations….



http://tylervigen.com/spurious-correlations

Spooky Correlations….



http://www.motherjones.com/environment/2016/02/lead-exposure-gasoline-crime-increase-children-health

ELSEVIER

Explorations in Economic History

Volume 62, October 2016, Pages 51-86

Lead exposure and violent crime in the early twentieth century ☆

James J. Feigenbaum [a] [1] ✉, Christopher Muller [b] [1] ✉

Show more

https://doi.org/10.1016/j.eeh.2016.03.002 Get rights and content

Abstract

In the second half of the nineteenth century, many American cities built water systems using lead or iron service pipes. Municipal water systems generated significant public health improvements, but these improvements may have been partially offset by the damaging effects of lead exposure through lead water pipes. We study the effect of cities' use of lead pipes on homicide between 1921 and 1936. Lead water pipes exposed entire city populations to much higher doses of lead than have previously been studied in relation to crime. Our estimates suggest that cities' use of lead service pipes considerably increased city-level homicide rates.

JEL classification

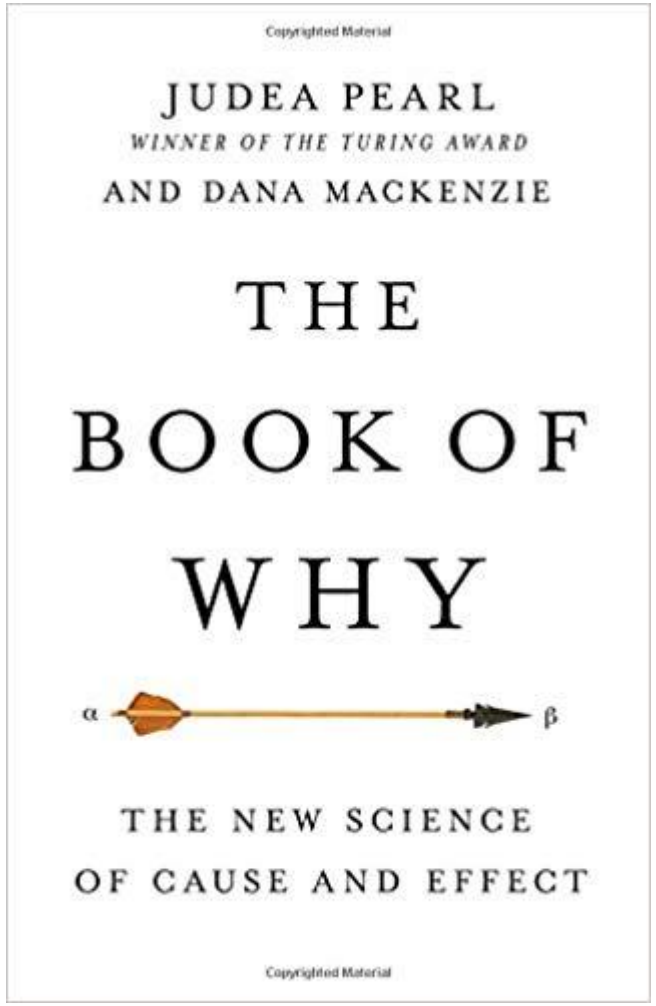
Q53; N32; K42

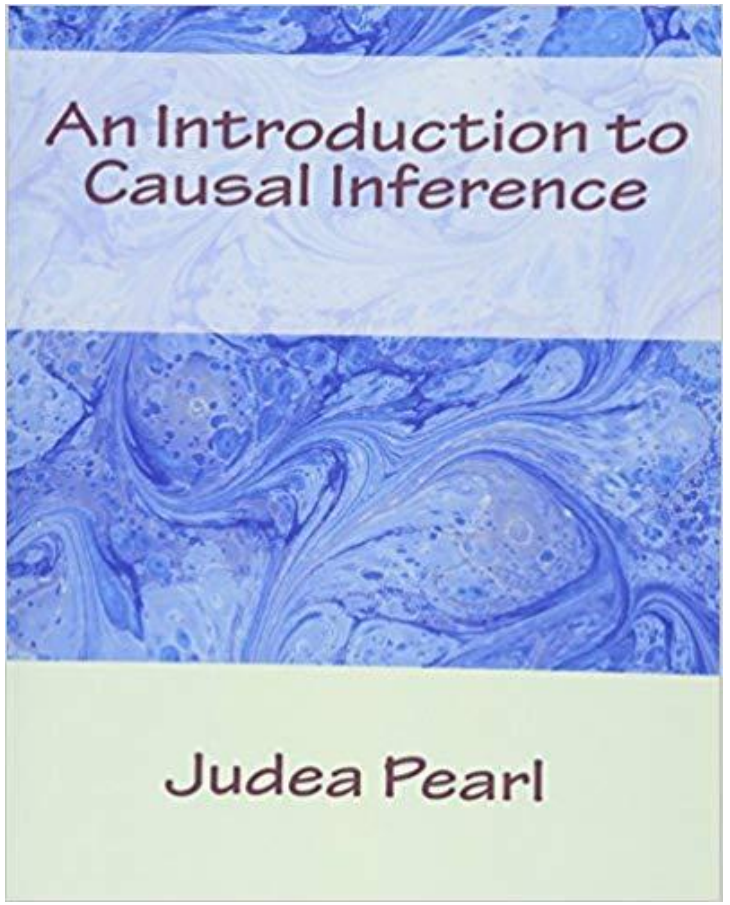
Keywords

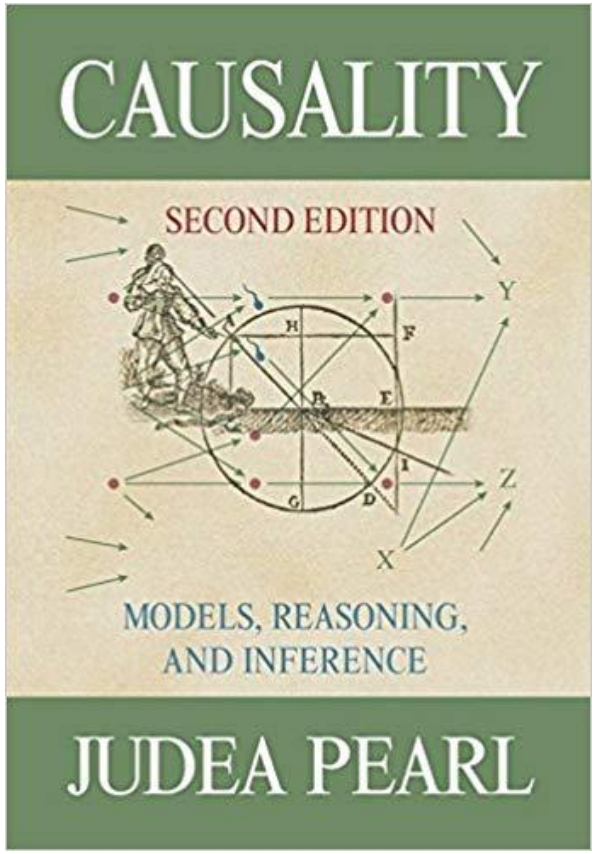
Urban economic history; Pollution; Crime; Lead

http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum_muller_lead_crime.pdf

The Causal Revolution



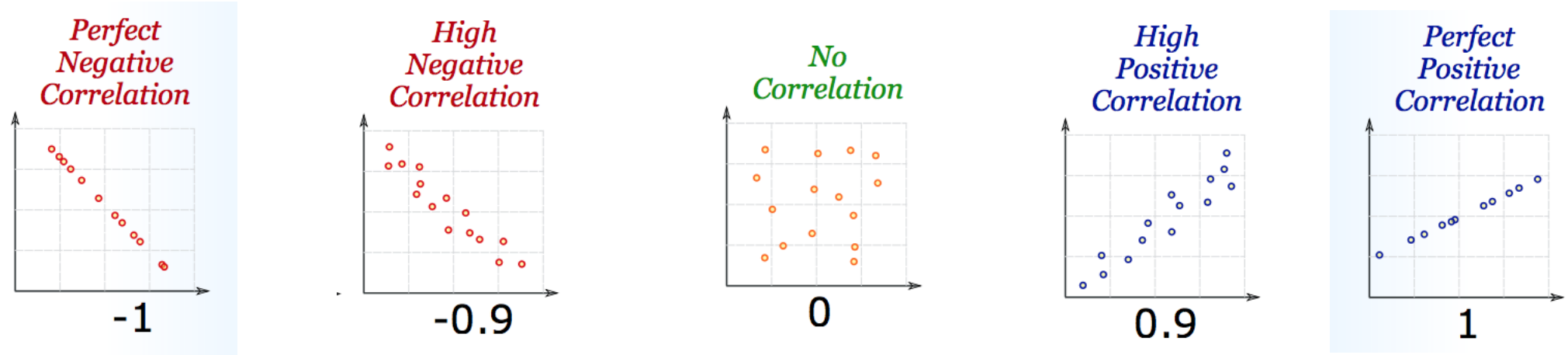


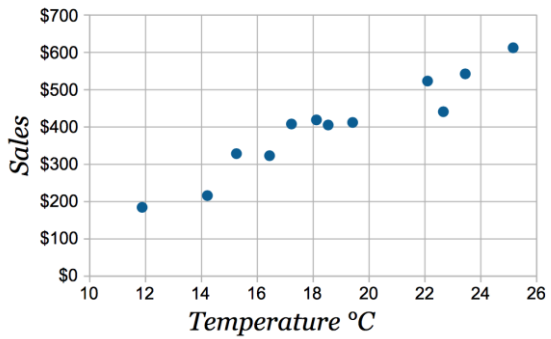


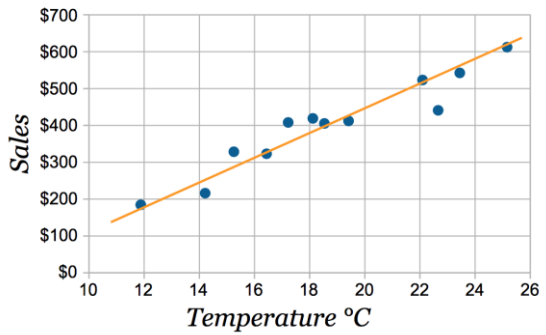
http://bayes.cs.ucla.edu/jp_home.html

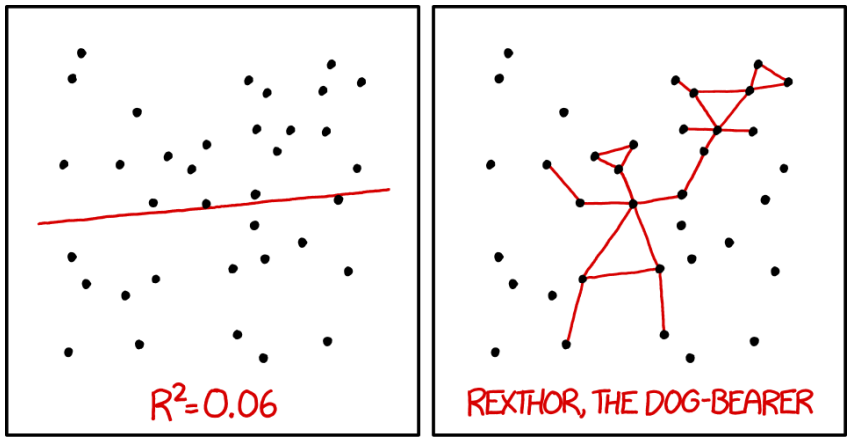
Predictive Analytics: Data-Modelling Approach

Predictive Analytics: Data-Modelling Approach







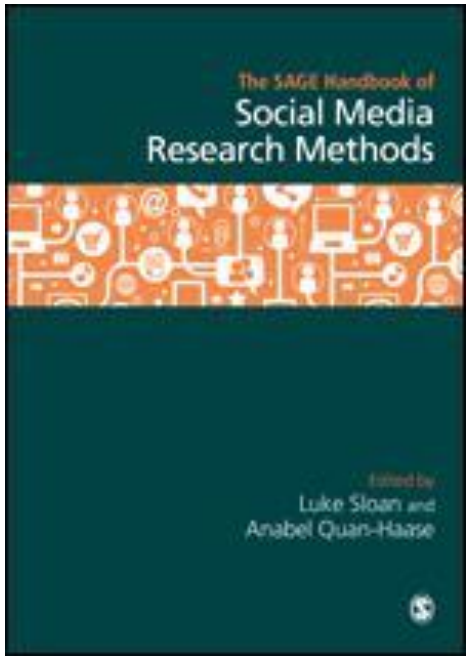


https://xkcd.com/1725/

Systematic Review of 40 Predictive Models using Big Social Data -1/3

table 20.1 categorization of research Publications on Predictive analytics with social Media data

Reference	*Social Data*	*Dependent Variables*	*Independent Variables*	*Statistical Methods*
Asur & Huberman (2010)	Twitter	Movie revenue	Twitter activity, sentiment and theatre distribution	Time-Series Multiple Regression Model
Lassen et al. (2014)	Twitter	iPhone sales	Twitter activity and sentiment	Time-Series Multiple Regression Model
Bollen & Mao (2011)	Twitter	Dow Jones Industrial Average	Calm, Alert, Sure, Vital, Kind and Happy	Time-Series Multiple Regression Model
Voortman (2015)	Google Trends	Car sales	Google trend data car names	Time Series Linear Regression Model
Vosen & Schmidt (2011)	Google Trends	Consumer spending	Real personal income y, interest rates on 3-month Treasury Bills I and stock prices s (measured on S&P 500), Google Trend, and consumer spending t-1	ARIMA/Time Series Multiple Regression Model
Choi & Varian (2012)	Google Trends	Sales of cars, homes and travel	Historical sales and Google trend variable	Simple Seasonal AR Models and Fixed-Effects Models
Chung & Mustafaraj (2011)	Twitter	Political election outcome	Twitter collective sentiment	Linear Regression
Conover, Gonçalves, Ratkiewicz, Flammini, & Menczer (2011)	Twitter	Political alignment	Twitter hashtags	SVM trained on hashtag metadata
Bothos, Apostolou, & Mentzas (2010)	IMDB, Flixster, Yahoo Movies, HSX, Twitter, RottenTomatoes.com	Movie Academy Award winners	Measures from IMDB, Flixster, Yahoomovies, HSX, Twitter, RottenTomatoes.com	Multivariate Distribution Models
Culotta (2010)	Twitter	Detecting influenza outbreaks	Twitter keywords	Time-Series Multiple Regression Model
Dijkman, Ipeirotis, Aertsen, & van Helden (2015)	Twitter	Many types of sales	Twitter activity and sentiment	Time-Series Multiple Regression Model
Eysenbach (2011)	Twitter	Total number of citations	Twimpact variable (number of tweetations within n days after publication)	Multi-Variate/Linear Regression
Gruhl, Guha, Kumar, Novak, & Tomkins (2005)	Blogs	Sales	Product/brand mentions	Time-Series using Cross-Correlation
Jansen, Zhang, Sobel, & Chowdury (2009)	Twitter	Brand variables	Twitter sentiment variables	Time-Series Linear Regression Models
Li & Cardie (2013)	Twitter	Early stage influenza detection	Twitter texts about flu	Unsupervised Bayesian Model based on Markov Network
Radosavljevic, Grbovic, Djuric, & Bhamidipati (2014)	Tumblr	Sport results and number of goals	Team and player mentions	Poisson Regression Model using Maximum Likelihood Principle



Lassen, N. B., la Cour, L., & Vatrapu, R. (2017). Predictive Analytics with Social Media Data. *The SAGE Handbook of Social Media Research Methods, 328-341.*
https://www.dropbox.com/s/49ur6lkbmkfyenf/2017-BookChapter-SageHandbook-PredictiveAnalyticsSocialMedia.pdf?dl=0

Systematic Review of 40 Predictive Models using Big Social Data -2/3

- 2008: Influenza, GFT, **Google** Flu Trend, failed 2011-2012, shut down 2013.
- 2009: Brand Variables, Chowdury et al, **(Twitter)**
- 2010: Influenza, Culotta, **(Twitter)**
- 2010: Movie revenues, Asur & Huberman, HP Labs**. (Twitter)**
- 2010: Election outcomes, **(Twitter, Facebook, Google)**
- 2011: Dow Jones Index, Bollen & Mao, Indiana University **(Twitter)**
- 2011: Consumer spending, Vosen & Schmidt, **(Google)**
- 2012: Sales of cars, homes and travel, Choi & Varian, **(Google)**
- 2012: Personalities, Hughes, Rowe, Batey, & Lee, Uni of Bath, **(Twitter & Facebook)**
- 2013: US Stock price Index, Karabulut, Geothe University **(Facebook)**
- 2013: Depression, De Choudhury et al (2013), **(Twitter)**
- 2013: News spreading SoMe, Weeks & Holbert, **(Twitter, Facebook , YouTube)**
- 2014: UK, US, and Canadian stock markets, Mao, Counts, & Bollen **(Twitter)**
- 2014**:** Sport results and number of goals, Radosavljevic et al, **(Tumblr)**
- 2014: Sales of iPhones, Lassen et al. (2014), **(Twitter)**
- 2015: Heart attacks, Eichstaedt et al., **(Twitter)**

Systematic Review of 40 Predictive Models using Big Social Data -3/3

- **App. 20% of examined models are within sales & marketing.**
- **App. 15% of examined models are within Election behavior.**
- **App. 15% of examined models are within Diseases.**
- **App. 20% of examined models are within Stock Price Indexes, Financial Markets**
- **App. 5% of examined models are within News, Info exchange**
- **App. 15% of examined models are within Human behavior, psychology.**
- **App. 10% of examined models are within Sport results, Academy awards etc.**

50% of all models are regression models.

Predictive Analytics with Big Social Data: General Regression Model

$$y_t = \beta_a * A_t + \beta_p * P_t + \beta_d * D_t + \beta_o * O_t + \varepsilon_t$$

Where:

y_t = Outcome variable of interest

A_t = Accumulated time-lagged social media activity associated with outcome variable at time *t*

$A_t = \sum A_{st}$

A_{st} = Social media activity in terms of actions by actors on artifacts associated with outcome variable at time *t*

P_t = Individual or social psychological attribute(s) at time *t*

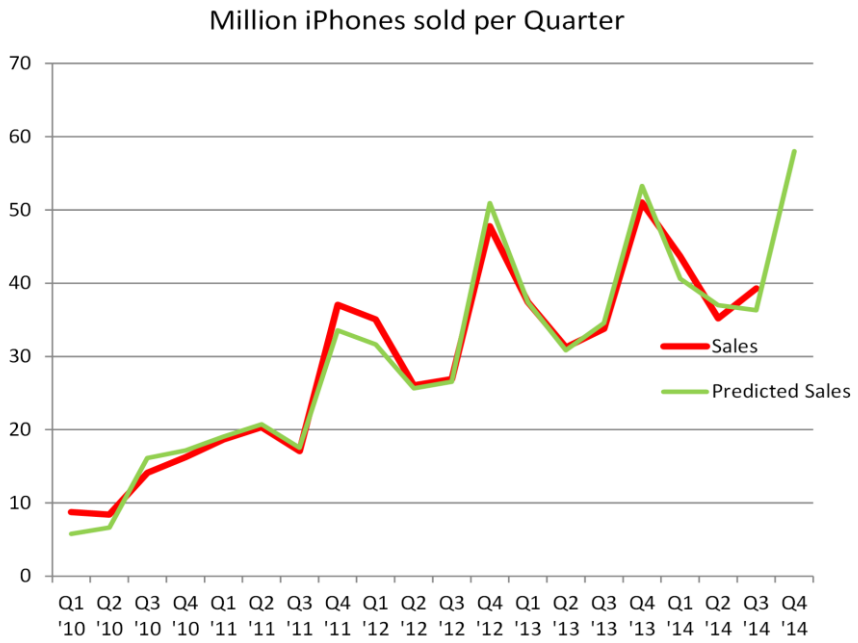
D_t = Social media dissemination factors

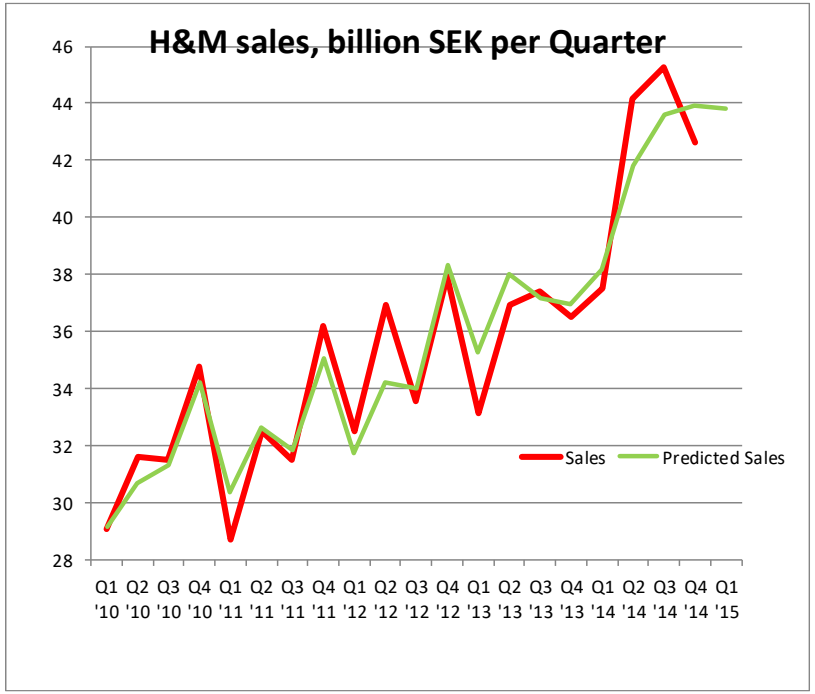
O_t = Other explanatory factors

Business Value: Sales and Revenue Predictive Models

Company	Data Source	Time Period	Size of Dataset
Apple	Twitter	2007 → October 12, 2014	500 million+ tweets containing “iPhone”
H & M	Facebook	January 01, 2009 → October 12, 2014	~15 million Facebook events

Stages	AIDA	Hierarchy of effects
Cognition	Attention	Awareness Knowledge
Affect	Interest Desire	Liking Preference Conviction
Behavior	Action	Purchase

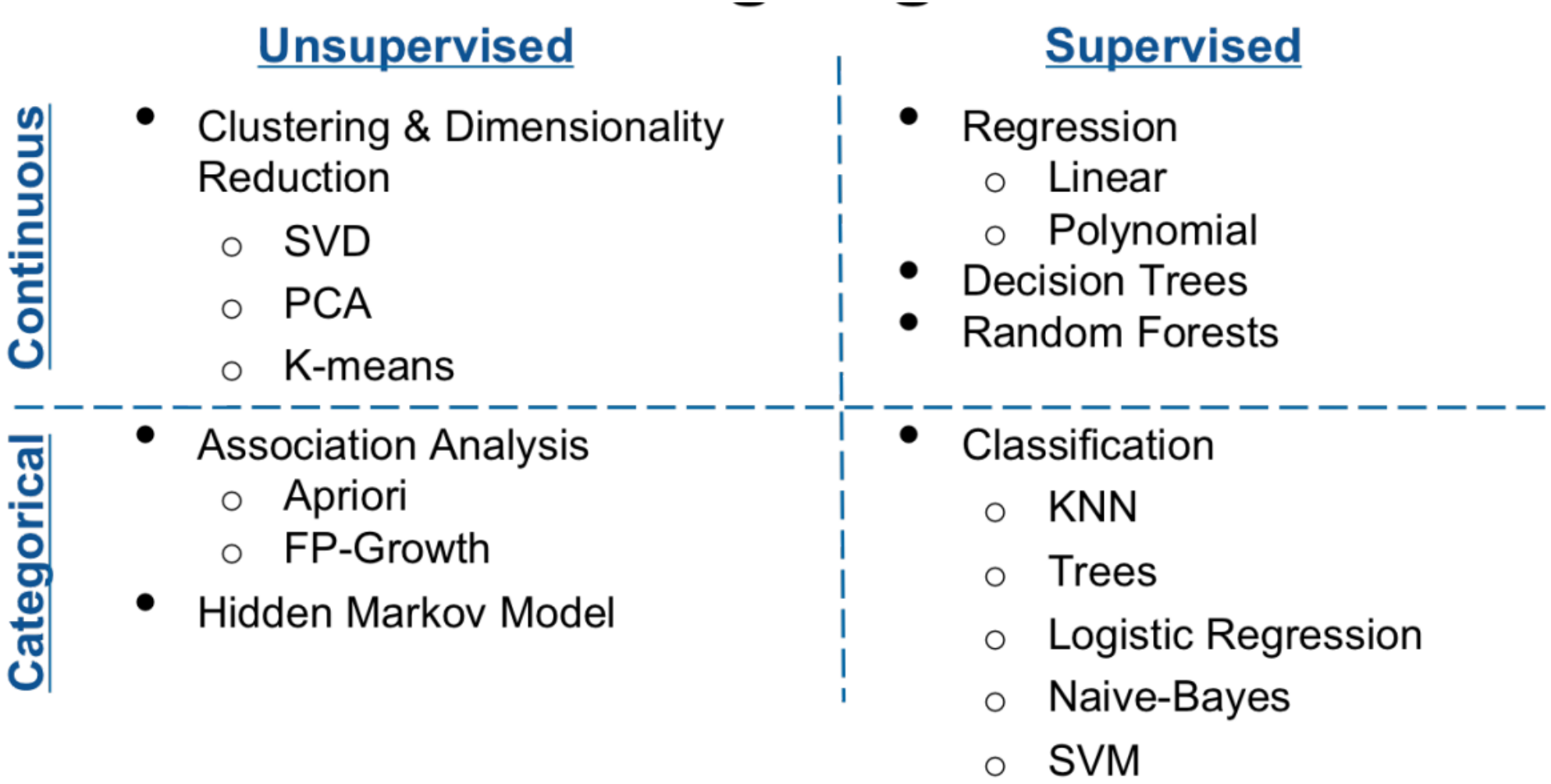




(IEEE EDOC 2014) (ICCSS 2015)

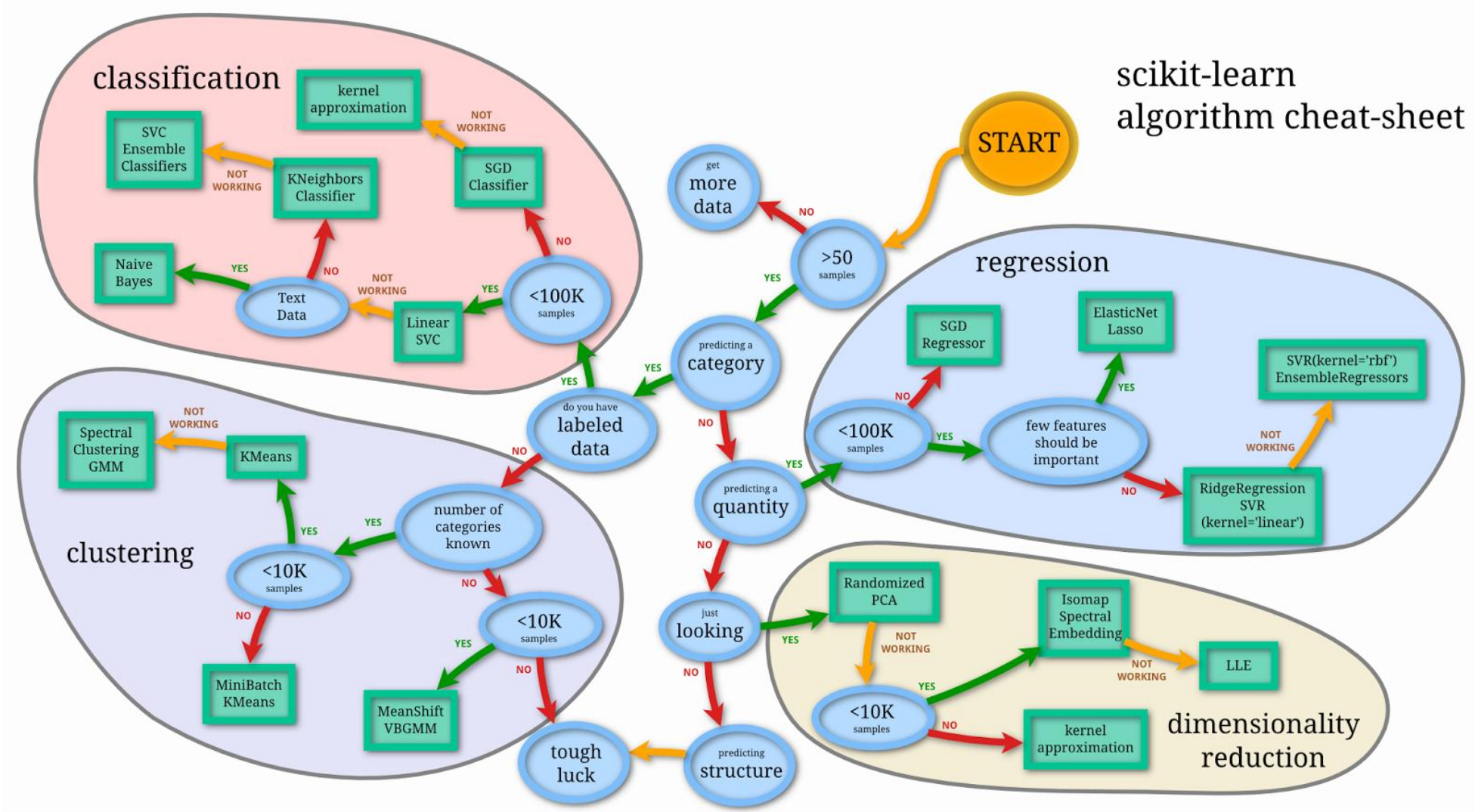
Predictive Analytics: Algorithmic Modelling Approach

Predictive Analytics: Algorthmic Modelling Approach



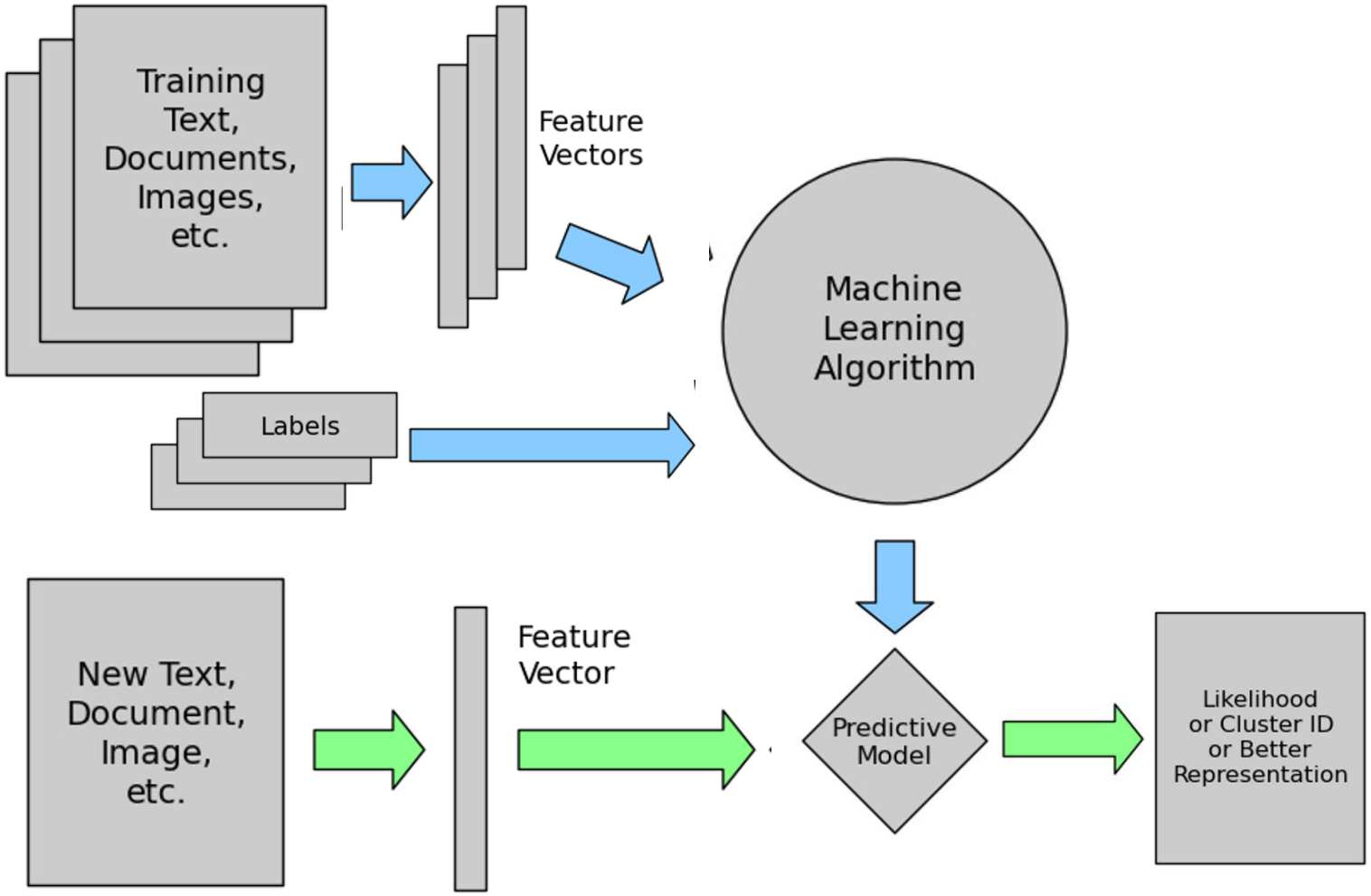
http://blog.hackbrightacademy.com/2014/05/pycon-2014-melanie-warrick-machine-learning-talk/

Machine Learning: Algorithms: Cheat Sheet



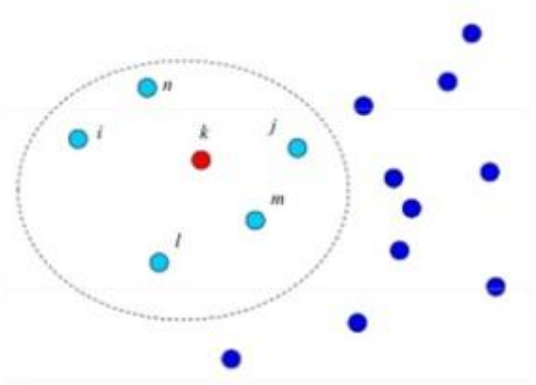
http://n-chandra.blogspot.dk/2013_01_01_archive.html

Supervised Machine Learning

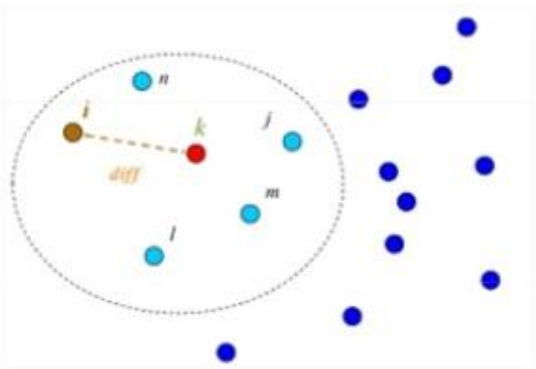


http://www.astroml.org/sklearn_tutorial/general_concepts.html

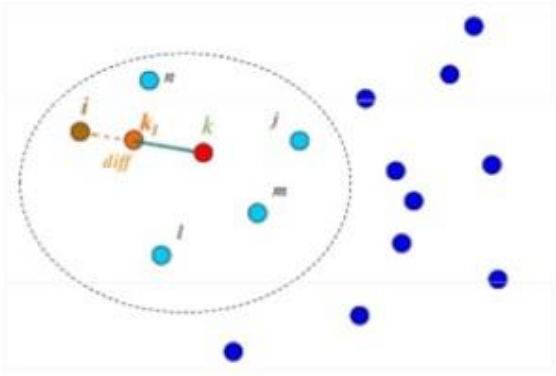
Problem: Unbalanced Classes in Big Datasets



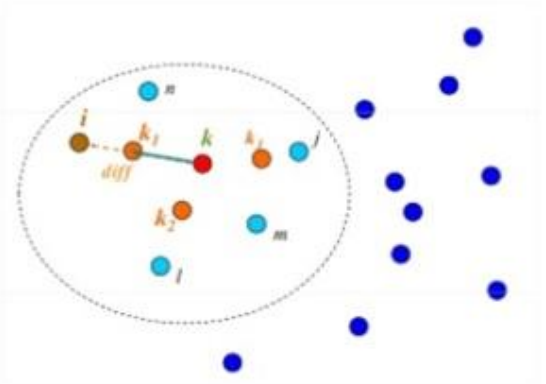
1. For each minority example k compute nearest minority class examples (i, j, l, n, m)



2. Randomly choose an example out of 5 closest points



3. Synthetically generate event k_1, such that k_1 lies between k and i



4. Dataset after applying SMOTE 3 times

https://www.slideshare.net/dalpozz/racing-for-unbalanced-methods-selection

SMOTE: Synthetic Minority Over-sampling Technique: https://jair.org/index.php/jair/article/view/10302

Bitcoin Blockchain: Deanonymizing Entity Types using Supervised Machine Learning (JMIS 2019)

https://www.dropbox.com/s/7374w6tfe13ikkd/Regulating%20Cryptocurrencies%20A%20Supervised%20Machine%20Learning%20Approach%20to%20De%20Anonymizing%20the%20Bitcoin%20Blockchain.pdf?dl=0

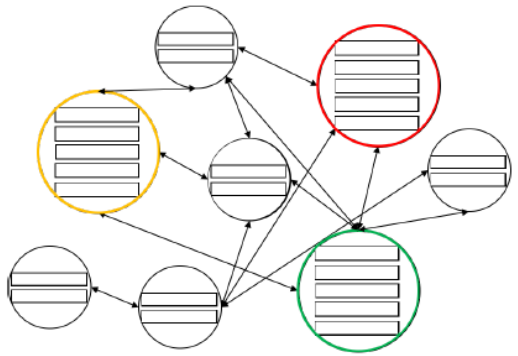


Figure 2. Visualisation Network of Different Categories

Figure 2. Data Preparation and Analysis Process

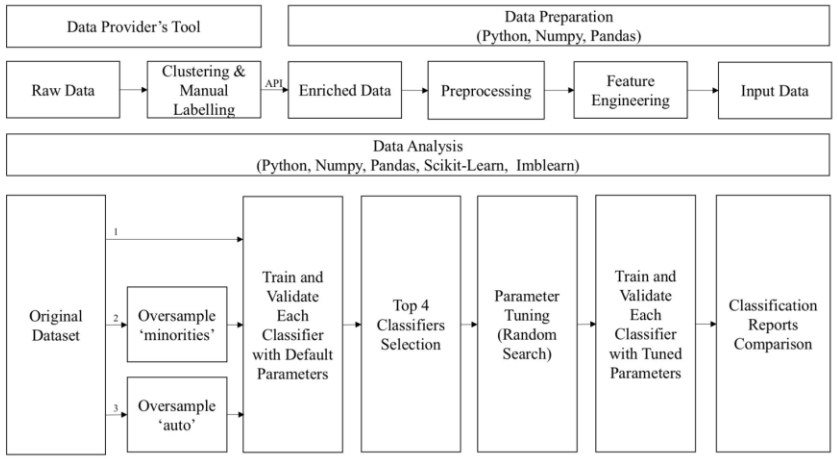
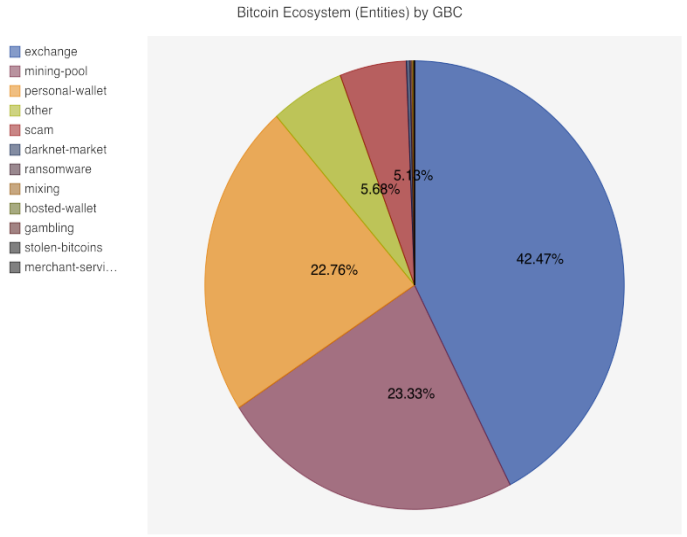


Table 5. Prediction Results

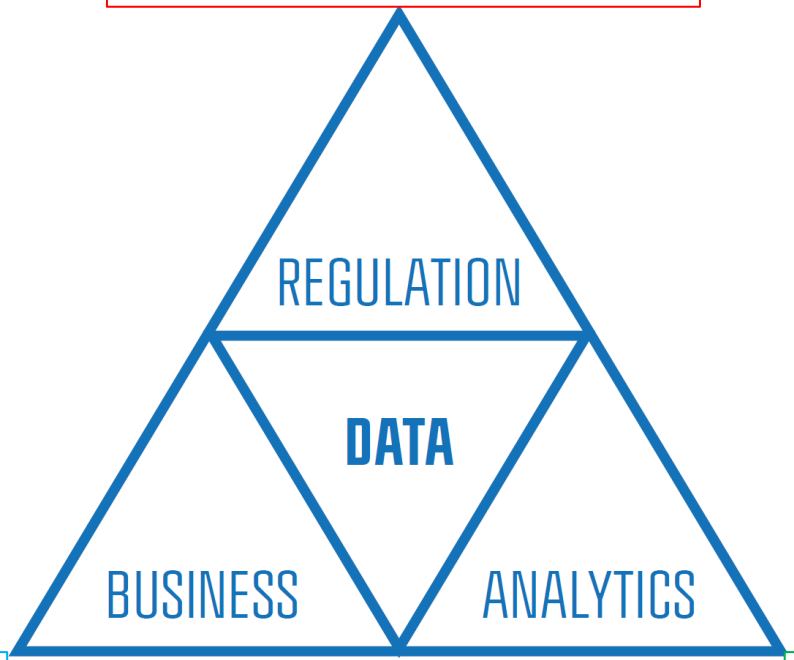
Address	RFC	ETC	BGC	GBC
add1	personal-wallet	personal-wallet	personal-wallet	exchange
add2	ransomware	personal-wallet	gambling	ransomware
add3	ransomware	personal-wallet	ransomware	exchange
add4	personal-wallet	personal-wallet	exchange	ransomware
add5	exchange	personal-wallet	gambling	ransomware
add6	ransomware	ransomware	ransomware	ransomware
add7	exchange	personal-wallet	ransomware	personal-wallet
add8	gambling	gambling	other	gambling
add9	exchange	other	exchange	exchange
add10	gambling	personal-wallet	gambling	ransomware
add11	exchange	other	exchange	exchange
add12	personal-wallet	personal-wallet	personal-wallet	personal-wallet
add13	ransomware	ransomware	ransomware	darknet-market
add14	personal-wallet	exchange	ransomware	ransomware
add15	ransomware	personal-wallet	ransomware	ransomware
add16	ransomware	personal-wallet	ransomware	ransomware
add17	exchange	personal-wallet	gambling	ransomware
add18	exchange	personal-wallet	exchange	exchange
add19	exchange	personal-wallet	gambling	ransomware
add20	exchange	personal-wallet	gambling	ransomware
add21	ransomware	personal-wallet	gambling	ransomware
add22	exchange	personal-wallet	gambling	ransomware



CBS CENTRE FOR BUSINESS DATA ANALYTICS
TRANSFORMING BIG DATA SETS INTO BUSINESS ASSETS

MSc in Business Data Analytics

Datafication: Regulation, Governance, Security, Privacy and Ethics (7.5 ECTS)



Predictive Analytics (7.5 ECTS)

Text Analytics (7.5 ECTS)

Data Mining, Machine Learning and Deep Learning (7.5 ECTS)

Visual Analytics (7.5 ECTS)

Foundations of Business Data Analytics: Architecture, Statistics and Programming (7.5 ECTS)

Data Economics (7.5 ECTS)

Innovation and Strategy in the Data Economy (7.5 ECTS)

Electives, Exchange or Internship (30 ECTS)

Master Thesis (30 ECTS)

CBS CENTRE FOR BUSINESS DATA ANALYTICS
TRANSFORMING BIG DATA SETS INTO BUSINESS ASSETS

Causal Modeling Workshop

Reflections