
# Data Science Trends and Tools for Measuring Attitudes and Behaviors

Michael Link, Ph.D.
Division Vice President, Data Science, Surveys & Enabling Technologies (DSET)
Abt Associates
BOLD THINKERS DRIVING REAL-WORLD IMPACT

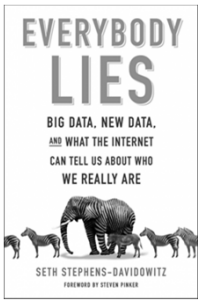
*AAPOR Webinar: October 11, 2017*

## The Truth Is Out There …

"… Google searches are the most important dataset ever collected on the human psyche.

"… mental illness; human sexuality; child abuse; abortion; advertising; religion; health …"

"Too many data scientists today are accumulating massive sets of data and telling us little of importance … You don't always need a ton of data to find important insights. You do need the right data."

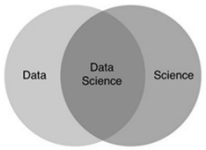





## Focus of Webinar

- What is Data Science?
  - How does it relate to what we typically do?
- Key Areas & Hot Topics:
  - Machine Leaning
  - Text Analytics
  - Data Visualization (Dashboarding)
- Cautions & Challenges
- Potential Resources
- Q&A

## What Is Data Science?




BOLD THINKERS DRIVING REAL-WORLD IMPACT

## What is "Data Science"?

- "Data Science: the scientific study of the creation, validation, and transformation of data to create meaning."
- "Data Scientist: A professional who uses scientific methods to liberate and create meaning from data"

Data Science Association
www.datascienceassn.org

*"Information is produced from data by uses. Data streams have no meaning until they are used. The user finds meaning in data by bringing questions to the data and finding their answers in the data."*

*Robert Groves*
*Provost, Georgetown University*

BOLD THINKERS DRIVING REAL-WORLD IMPACT

## New Sources of Data to Measure Behaviors

- Administrative data
- Transaction records (banking, purchase, etc.)
- Medical Records
- Social Media
- Bluetooth enabled devices
- Mobile devices
- Wearables
- Sensors IOT
- Visual: pictures & video
- Location info: GPS
- Geo-info: Satellites, planes, drones


The New York Times

PERSONAL TECH

The Smartphone's Future: It's All About the Camera

August 30, 2017

BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Design vs Organic Data

**Design Data**

- Traditional data (e.g. surveys)
- From a census or survey
- Collected from specific populations
- For specific purposes
- Often collected by those who will use them
- Respondents asked to answer questions
- Researchers control the data

**Organic Data**

- Arise out of the information ecosystem
- Massive
- Close to "real time" measures
- Not designed for research purpose
- But potentially useful
- Collection unobtrusive to those being measured
- Researchers do not control data

Adopted from Robert Groves (2011). "Census Directors Blog: Designed Data and Organic Data". Accessed at: https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html

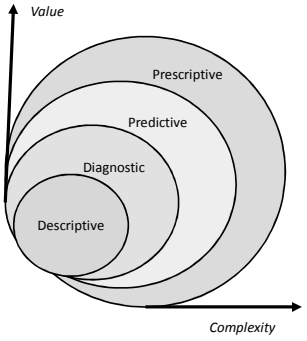


## Types of Organic Data

| | Structured Data - Administrative Records | Other Structured Data | Semi-Structured Data | Unstructured Data |
|---|---|---|---|---|
| Definition | Data with a fixed format easily exportable to a data set for analysis with minimal scrubbing required | Highly organized data easily placed in a data set but require additional scrubbing or transformation before analysis | Data that may have some structure but not complete and cannot be placed in a relational database; requires substantial cleaning | Data which have no standard analytic structure and must have data extracted and transformed before use |
| Examples | • Govt programs<br>• Commercial transactions<br>• Credit card / bank records<br>• Medical records<br>• University / school records | • E-commerce transactions<br>• Mobile phone GPS<br>• Roadside / Weather / pollution sensors | • Computer logs<br>• Text messages<br>• Email<br>• Fitbit / wearable data<br>• Internet of Things | • Social media data<br>• Pictures / videos<br>• Traffic webcams<br>• Drone data<br>• Satellite / radar images |

Adopted from: National Academies of Sciences, Engineering, & Medicine. (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press.

## Types of Data Analytics



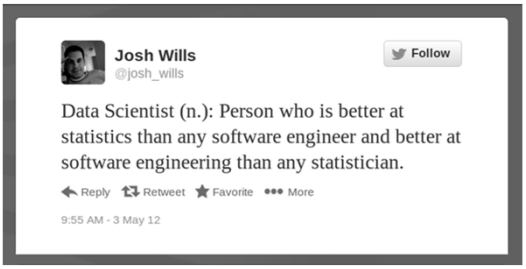

**Descriptive:**
*What is happening?*

**Diagnostic:**
Why is it happening?

**Predictive:**
What is likely to happen

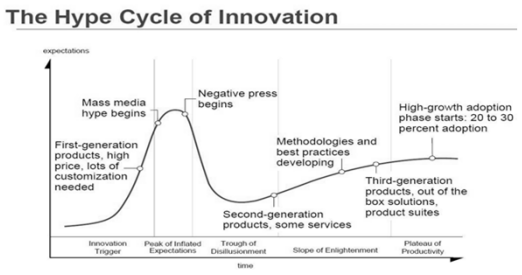
**Prescriptive:**
What do I need to do?

## What Is a Data Scientist?






## Is "Data Science" Just Hype?

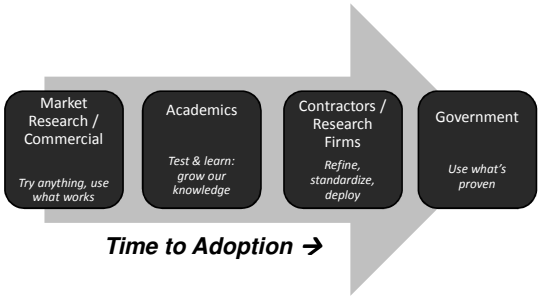



http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

Source: Gartner.com



## (Somewhat) Typical New Tech Adoption Cycle



## Potential Advantages of DS for Measuring Attitudes & Behaviors

- Far wider range of potential data sources
  - Greater granularity (geographic, demographic, etc.)
  - Some sources have near real-time measures (sensors, social media)
  - New measures, new insights (Bluetooth-enabled measures)
  - Potentially more reliable / valid data for some measures (GPS vs Travel Diaries)
- More advanced analytic techniques being developed and applied to social data
  - Techniques for converting unstructured data into analyzable form (videos, images)
  - Methods for managing and reducing large volumes of data into analyzable size and form
  - Algorithms that can be "trained" and "learn" to automate certain practices and process larger volumes

High potential, some of which is currently being realized – but some major caution areas as well – we will explore both sides of that equation …



## Key Areas & Hot Topics…



## Key Terms in Data Science Today

| Often Used Terms / Concepts |
|---|
| **Data Mining:** (Can have a wide range of definitions): A nontrivial and structured process of uncovering new and useful patterns in data that can be generalized. |
| **Web Scraping:** Practice of extracting data from websites |
| **Data Wrangling**: Transforming, mapping, and cleaning data from one "raw" form to another form more appropriate for analytics or other useful purposes, such as inputs for machine learning algorithms |
| **Record Linkage:** Merging together information from two or more sources of data with the goal deriving insights not achievable with separate data sets |
| **Data Visualization**: Any effort to help people understand data insights by placing it in a visual context (as opposed to simple number tables) |
| **Big Data Analytics:** Process of examining large and varied data sets to uncover patterns, unknown correlations, trends and other useful insights |

## Key Terms in Data Science Today (Con't)

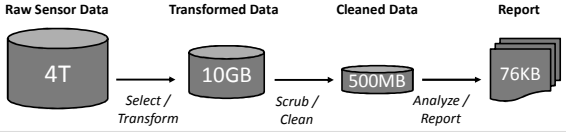
| Often Used Terms / Concepts |
|---|
| **Text Analytics** (aka Text Mining): process of deriving information from text, often patterns and trends |
| **Machine Learning:** Field of computer science that gives computers the ability to improve performance without being continually re-programmed. |
| **Deep Learning**: A class of Machine Learning algorithms that use a cascade of many layers of nonlinear processing units for feature extraction and transformation. |
| **Cognitive Computing:** The simulation of human thought processes in a computerized model – involves self-learning systems that use data mining, pattern recognition, and natural language processing to mimic the way the brain works. |
| **Natural Language Processing:** A way for computers to process, analyze and derive meaning from human language through the use of algorithms placed in semantic and syntactic context. (i.e., words are converted to numeric values or vectors that represent their relative meaning) |
| **Artificial Intelligence**: Development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, & decision-making |



## Side Note on Data Wrangling: From Big to Small Data

### Road Sensor Data for Official Transportation Statistics

- Leverage data from 60,000 sensors (induction loop, camera, Bluetooth) to develop vehicle lane counts and vehicle size estimates per minute (24/7). System produces more than 230,000,000 records per day.
- Sophisticated systems for extracting & transforming raw sensor data into analyzable information; then extensive cleaning & imputations; finally analysis.
- Converting "Big Data" to "Small Data" then insights.



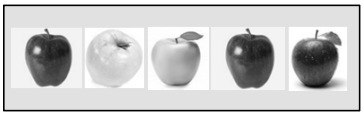



## Hot Topic 1: Machine Learning

- "Machine Learning" is a type of artificial intelligence that allows a computer program or software application to become more accurate in predicting outcomes without being explicitly programmed
- Primary outcomes: predicting classification/groups/values
- Use cases:
  - Categorizing open-ended reviews on consumer websites
  - Potential credit card fraud detection
  - Recommender engines (like Amazon, Netflix)
  - "Fake News" detection on social media sites
  - Exploring large volumes of video data for exo-planet discovery
- In the survey world:
  - Potential interviewer falsification alerts
  - Improving area sampling via satellite imagery (especially in developing world)
  - Exploring new sources of data (social media, medical records, published reports, etc.) for new insights into attitudes and behaviors

## Apples Or Pears?

Apple Training Set



Pear Training Set

***How should we classify these?***

BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Key Machine Learning Outcomes & Techniques

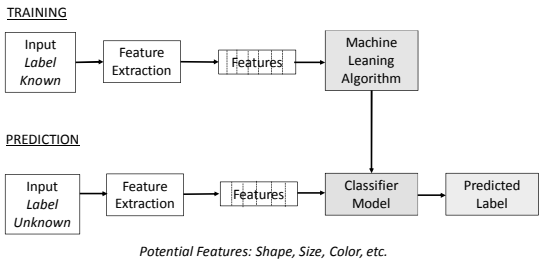
- Classification: predicting an elements class from observations (discrete/categorical)
  - Ex. technique: Naïve Bayes Classifier
- Clustering: Group observations into meaningful groups
  - Ex. techniques: k-means clustering or hierarchical clustering
- Prediction: Predict a value from a set of observations (real number or continuous outcome)
  - Ex techniques: regression

Common Tools Used in Machine Learning

- SAS Enterprise Miner
- R
- Python
- Rapidminer
- Google Prediction API
- Amazon Machine Learning
- Tensorflow
- Apache Singa
- Shogun
- IMB Watson ML

BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Machine Learning Workflow




*Potential Features: Shape, Size, Color, etc.*

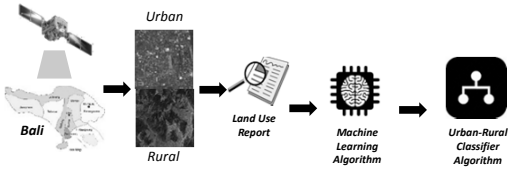
BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Basic Types of Machine Learning

| Machine Learning Methods | Description |
|---|---|
| Supervised Learning | • Correct values for the training data are **known**.<br>• Algorithm trained by human input – reduces level of manual review required for determining relevance and proper coding of outcome.<br>• Predicts an output value (class label or target value)<br>• Models based are predictive |
| Unsupervised Learning | • Correct values for training data are **not known**.<br>• Greater need for manual review to ensure relevance and proper coding of outcome.<br>• Does not predict output value, but rather groups data into patterns based on relationships between variables of an observation<br>• Models are descriptive |
| Reinforcement Learning | • Learning based on feedback from the environment. Learning can occur once or over be continuous (adapting over time).<br>• Algorithm is continually trained by human input then automated once desired accuracy is reached. |

## Machine Learning Example: Use of Satellite Imagery for Stratified Sampling

**Goal: Classify** hard to reach geographies into urban/rural as key strata for accurate sampling & weighting adjustment




Citation: Whye Loon Tung, Jingyuan Zhao and Timothy Banks (2014). "Eye in the Sky: Rural-Urban Classification Using Commercial Satellite Images." Paper presented at the Esomar Asia Pacific Conference, Jakarta Indonesia.

BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Hot Topic2: Text Analytics - Definition

- **Text Analytics** (aka Text Mining): process of deriving information from text, often patterns and trends
  - Applies linguistic and/or statistical techniques to extract features (concepts & patterns) that can be applied to categorize and classify documents, audio, video and images
  - Transforms "unstructured" information (not in easy to analyze form) into data for application of traditional analytic techniques
  - Discerns meaning & relationships in large volumes of information that were previously unprocessed

Uses of Text Analytics

- Google Search
- Social Media content analysis
- Analyzing large blocks of open-ended text (such as customer satisfaction)
- Supporting research planning – expanding literature reviews
- Research tool – identifying condition or disease incidence from digital notes
- Early warning systems – but beware Google flu …

BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Primary Modes of Analysis

- **Word counts:** Simply counting number of mentions of a word/entity of interest
  - Raw counts
  - Percentage of all (or a subset) tweets
- **Sentiment Analysis:** Automated way of assessing and quantifying polarity (positive/negative) orientation of words in text messages
  - Software is developed which can conduct the analysis by calculating the degree to which a text sample contains words belonging to empirically defined psychological and emotional orientations
- **Data visualization:** Graphic representation of networks or linkages between tweets, concepts, topics, people, etc.

Common Tools Used in Text Analytics

- SAS
- IBM SPSS
- Rapidminer
- Lexalytics
- Clarabridge
- NVIVO
- GATE
- WordStat
- NLTK



## Approaches to Sentiment Analysis

Sentiment analysis is the task of identifying positive and negative opinions, emotions and evaluations in text.

"The weather is great today!" — *Positive Sentiment*

"The thermometer reads 82 degrees" — *Neutral or factual statement*

"OMG! It is so freaking humid outside! ☹" — *Negative Sentiment*

Approaches:

1. Lexicon-based approach
   - ✓ categorize by comparing to dictionary of words with known sentiment
2. Machine learning approach
   - ✓ categorize based on syntactic and linguistic features
3. Contextual semantic approach
   - ✓ categorize based on context of a set of words



## Sentiment Is Often a Tricky Thing!

There are often subtle relations, patterns & dependencies among words in tweets:

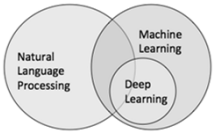
**"Fight Zika! Destroy invading mosquitos!"**

| Fight Zika | Destroy | Invading mosquitos |
|---|---|---|
| ↑ | ↑ | ↑ |
| *Negative Concept* | *Negative Word* | *Negative Concept* |
| ↑ | ↑ | ↑ |

**But the overall sentiment seems a positive one: promoting positive health!**

## Natural Language Processing (NLP)

- Natural Language Processing (or NLP) is an area that is a confluence of Artificial Intelligence and linguistics.
- A way for computers to process, analyze and derive meaning from human language through the use of algorithms placed in semantic and syntactic context. (i.e., words are converted to numeric values or vectors that represent their relative meaning)
- Allows you to sift through large volumes of text to generate insights, such as from sentiment analysis, information extraction, information retrieval, search, etc.
- Machine Learning techniques are often employed as a part of NLP
- Can also help facilitate more sophisticated applications using spoken language

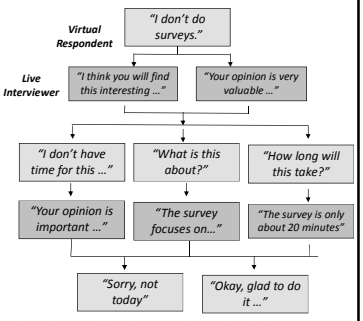



BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Natural Language Processing to Train Interviewers

**Goal:** Provide a realistic virtual environment for interviewers to practice introductions and refusal avoidance

- Natural language processing & a behavior engine drive the interaction
- Sessions can be recorded to review with a supervisor

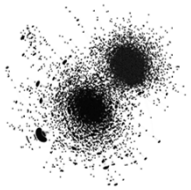



Source: Link, M, R. Hubal, C. Guinn, L. Flicker, and R. Caspar (2003). "Accessibility and Acceptance of a Responsive Virtual Human Technology-based Interviewer Training Application." Paper prepared for presentation at the Fourth International Conference on Survey and Statistical Computing, Warwick, England, UK.

BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Hot Topic 3: Data Visualization

*The communication of information clearly & effectively through graphical means*



Retweets of political posts – 250,000 tweets from 2010 Mid-term Elections (Blue = liberal; red = conservative)

**Typical Visualization Tools:**

- Spotfire
- Tableau
- Shiny
- Cartodb
- Piktochart
- ArcGIS Online
- Spotfire

Source: http://cnets.indiana.edu/groups/nan/truthy/visualizing-the-political-discourse-on-twitter/

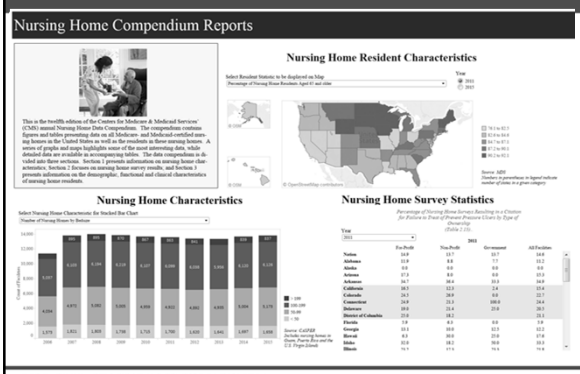
BOLD THINKERS DRIVING REAL-WORLD IMPACT

## Special Case of Data Viz: Dashboarding

- Not just about nice graphics, but the ability to make data more accessible & provide insights to help drive decisions and actions
  - Best designs: communicate a maximum of relevant information as immediately as possible
- "Dashboarding": A visual display of the most important information needed to achieve one or more objectives and which fits entirely on a single computer screen so it can be monitored at a glance
  - Drill downs/toggles to go from high level strategic views to lower level operational views (greater detail with each drill down)



## Use Case: Conversion to Paper Reports to Interactive Dashboard

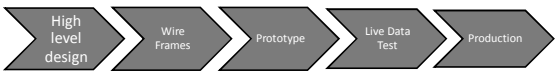





## Key Decisions for Developing Dashboards

- Who is the audience and what is their information need (content, frequency, level of detail)?
- What data and metrics are available that best meet these needs?
- What tools provide the optimal view?

***Typical Dashboarding Workflow***



## Common Dashboard Design Pitfalls

- Exceeding boundaries of single screen
- Inadequate context for the data
- Excessive detail or unnecessary precision
- Choosing inappropriate media for display
- Introducing meaningless variety
- Arranging data poorly
- Ineffective highlighting of what is important
- Screen clutter – useless decorations
- Unappealing visual displays

**Remember:**
Data are useless if they cannot be readily used and understood

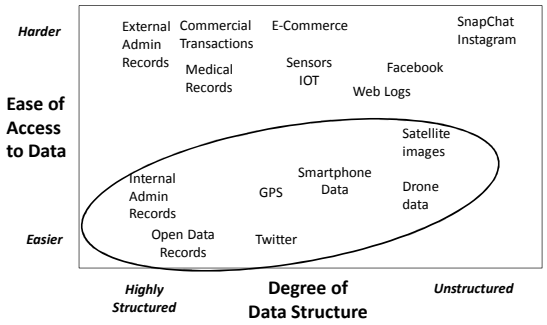


## Some Cautions to Consider in the Data Science World …



## Access Issues

## Data: Degree of Structure x Ease of Access



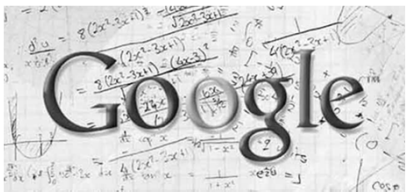

Adopted from: National Academies of Sciences, Engineering, & Medicine. (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press.

## Beware "Big Data Hubris"

*Source: Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). "The Parable of Google Flu: Traps in Big Data Analysis," Science, 343: 1203-1205.*

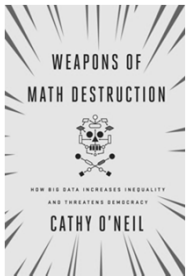


## Dilemma of the Perpetual, Dynamic Algorithm




*Source: Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). "The Parable of Google Flu: Traps in Big Data Analysis," Science, 343: 1203-1205.*

## Coverage, Representation, Potential Bias

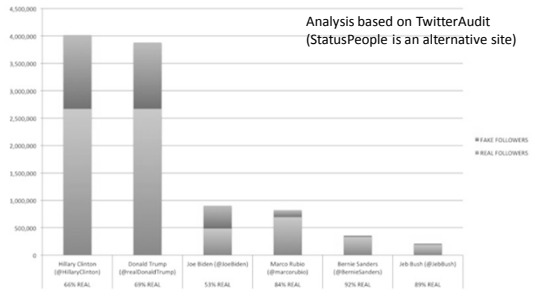




- Cathy O'Neil (2016). "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy." Crown, New York, NY.
- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



## Bots & "Fake" Accounts




Analysis based on TwitterAudit
(StatusPeople is an alternative site)

Source: Wayne Rash (8/27/15) "Hillary Clinton Has More Than a Million Fake Twitter Followers: What Does It Mean?" Yahoo Tech.



## What's Behind the Curtin …? Transparency

*Source: Wizard of Oz (1939), Film by Metro-Goldwyn-Mayer, Inc.*

## Privacy Concerns




## Cannot Answer All Questions




## We Need a New Evaluation Framework

TOTAL DATA ERROR

## University Training Programs for Data Science and Advanced Analytics

- Carnegie Mellon University
- Georgetown University
- Georgia Institute of Technology
- Harvard University
- New York University
- Northwestern University
- Temple University
- University of Maryland
- University of Michigan

*Disclaimer: These are universities I've worked with over the past 5 years, there are many others with Data Science and/or Advanced Analytics programs*



## Online Resources

- http://www.datasciencecentral.com
- https://www.coursera.org/
- https://ocw.mit.edu/index.htm
- https://www.edx.org/
- https://www.udemy.com/
- https://Alison.com
- https://www.kaggle.com
- https://www.datacamp.com

*Disclaimer: Here's a list to get you started, not an endorsement of any specific site*



**AAPOR**

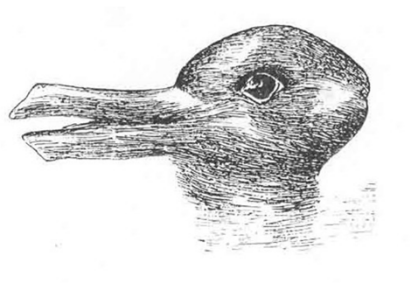
American Association for Public Opinion Research

*Aapor.org*

- Task Force reports on wide range of topics
- AAPOR-Sponsored Outlets for Publishing Research:



## In the End: Are We Seeing the Same Thing? Or Something New?



*Source: Joseph Jastrow (1899), "The Mind's Eye", Popular Science Monthly 54: 299-312*



# Q&A

Michael Link
Contact Info:

Michael_Link@Abtassoc.com
Twitter: @MLink01

BOLD THINKERS DRIVING REAL-WORLD IMPACT



## Appendix: Data Mining Tasks and Examples

| Tasks | Description | Algorithms | Use Cases |
|---|---|---|---|
| Classification | Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known data set. | Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbors | Assigning voters into known buckets by political parties, e.g. soccer moms<br><br>Bucketing new customers into one of the known customer groups |
| Regression | Predict the numeric target label of a data point. The prediction will be based on learning from a known data set. | Linear regression, logistic regression | Predicting unemployment rate for next year<br><br>Estimating insurance premiums |
| Anomaly Detection | Predict if a data point is an outlier compared to other data points in the data set. | Distance based, density based, local outlier factor (LOF) | Fraud transaction detection in credit cards<br><br>Network intrusion detection |
| Time Series | Predict the value of the target variable for a future time frame based on historical values. | Exponential smoothing, autoregressive integrated moving average (ARIMA), regression | Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated |
| Clustering | Identify natural clusters within the data set based on inherit properties within the data set. | K-means, density-based clustering (e.g., density-based spatial clustering of applications with noise [DBSCAN]) | Finding customer segments in a company based on transaction, web, and customer call data |
| Association Analysis | Identify relationships within an item set based on transaction data. | Frequent Pattern Growth (FP-Growth) algorithm, Apriori algorithm | Find cross-selling opportunities for a retailer based on transaction purchase history |

Source: Vijay Kotu and Bala Deshpande, *Predictive Analytics and Data Mining* (Boston: Morgan Kaufmann, Elsevier, 2015), 11.