Predictive Analytics:

The science behind the success

Steven D. Reeves
Predictive Analytics Solutions Architect
Government Forum - May 4, 2011

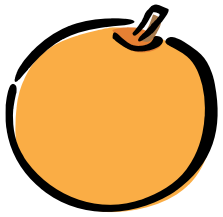
Agenda

- Traditional statistics and data mining
- Questions data mining can answer
- Data mining: Three classes of algorithms
  - Prediction
  - Association
  - Clustering
- Supervised vs. unsupervised learning
  - Supervised: Prediction and classification
  - Unsupervised: Clustering, Association and Anomaly Detection
- Text Analysis
- Use Cases

Data mining and statistical analysis

- **Statistical analysis**
  - Confirm hypotheses
  - More data requirements
  - More assumptions
  - General population predictions
  - Cumulative results

User-driven



- **Data mining**
  - Generate hypotheses
  - More exploratory
  - Less data prep
  - Fewer assumptions
  - Individual predictions
  - Results-oriented

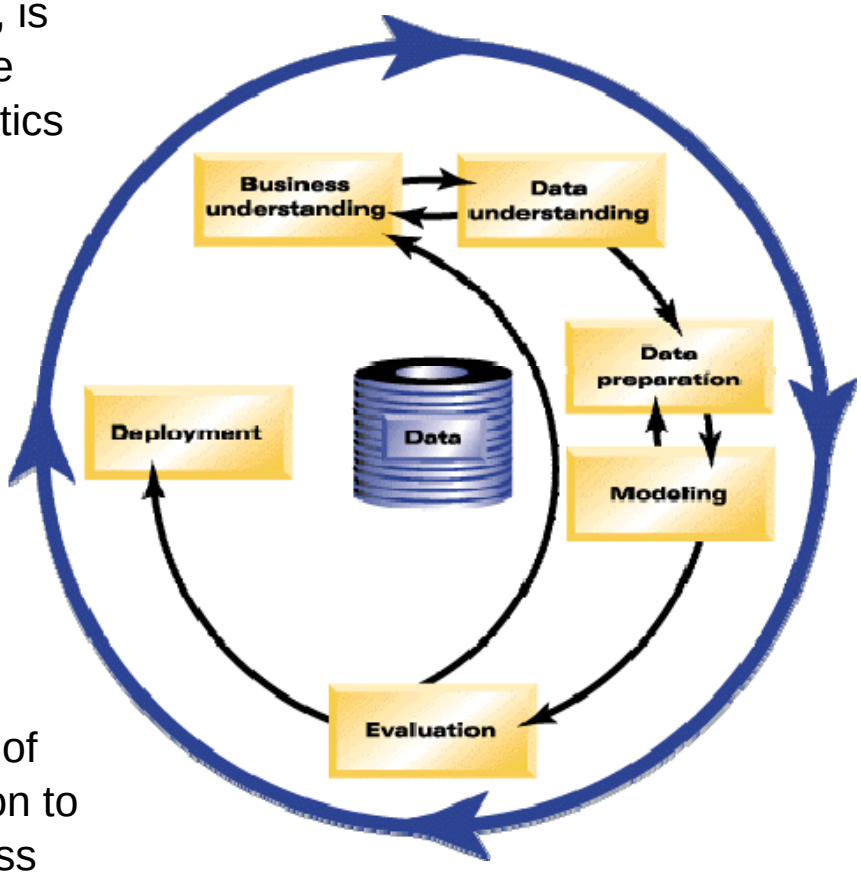
Data-driven

Statistics – use case examples

- Used often in experimental design, clinical trials and survey research with complex sampling designs
  - N.O.R.C. and Gallup use extensive inferential statistics accurately representing survey data on how people think and feel about the world today.
  - NIH uses inferential statistics to analyze experimental data to quantify significant differences in treatments and interventions.
  - CDC – extensive epidemiological studies require inferential statistics
- Used to create data when you don't have it.
  - Sample size
  - Effect size
  - Validity of results

Data mining

- Data mining: A branch of computer science, is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. (Wikipedia)
  - Business understanding
  - Data understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Model deployment
- Predictive analytics: Informs and directs decision-making by applying a combination of advanced analytics and decision optimization to data, with the objective of improving business processes to meet specific organizational goals.

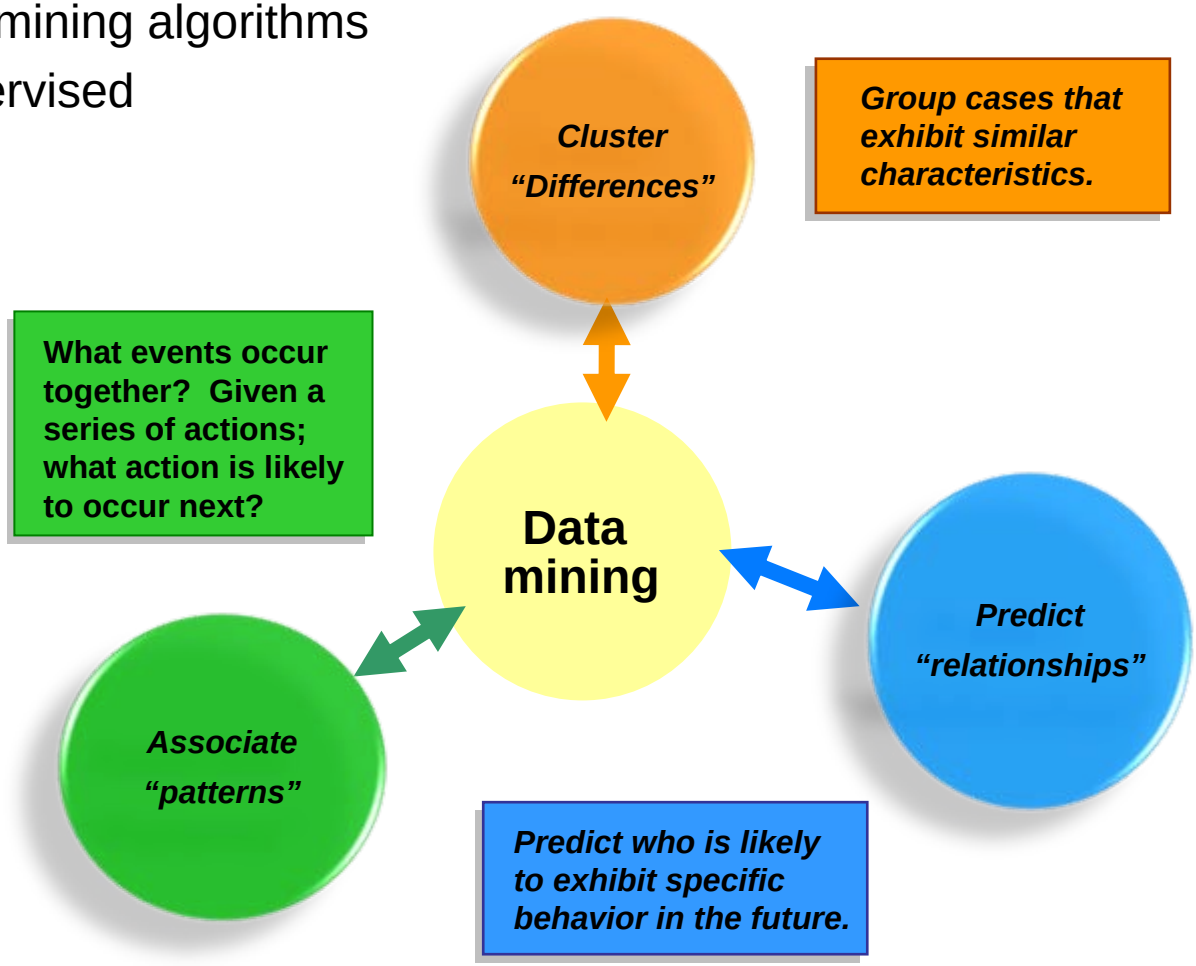


Data mining question types

- Market segmentation – **identify common characteristics** of constituents who are similar (e.g., buy the same products, use the same services)
- Churn - **predict** who's leaving
- Fraud/anomaly detection – discover and **predict** what transactions are fraudulent
- Direct marketing- **predict** who's likely to respond
- Interactive marketing – **predict** what will make people respond differently at point of interaction
- Market basket analysis – identify products or **services purchased or utilized together**
- Trend analysis – **look at differences** through time and or across groups
  - Have service utilization rates gone up or down?
- Sequence analysis – describe the most typical **series of events** leading to a consequent
  - What parts typically fail prior to an expensive servicing?
  - What requests are made of IT Support before catastrophic network failure?

Data mining

- Three classes of data mining algorithms
- Supervised vs. unsupervised
- Complementary

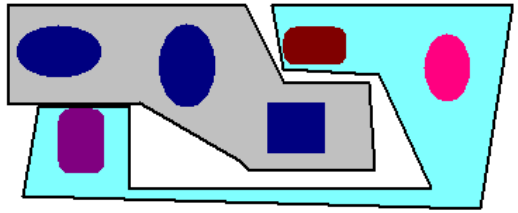


What is Unsupervised Learning?

- A data mining technique when we do not know the output or outputs
- Can be thought of as finding ‘useful’ patterns above and beyond noise… or “fishing” for information
- Looks for natural groupings in the data
- Can be used for data reduction, preparation and simplification

Unsupervised Learning: Questions

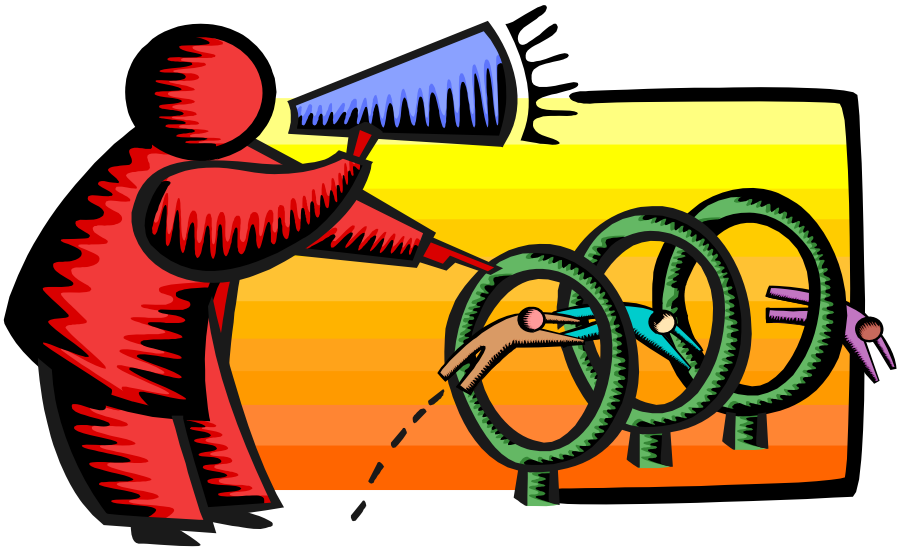
- Market segmentation – identify common characteristics of constituents who are similar (e.g., buy the same products, use the same services)
- Churn - predict who's leaving
- Fraud/anomaly detection – discover and predict what transactions are fraudulent
- Direct marketing- predict who's likely to respond
- Interactive marketing – predict what will make people respond differently at point of interaction



- Market basket analysis – identify products or services purchased or utilized together
- Trend analysis – look at differences through time and or across groups
  - Have service utilization rates gone up or down?
- Sequence analysis – describe the most typical series of events leading to a consequent
  - What parts typically fail prior to an expensive servicing?
  - What requests are made of IT Support before catastrophic network failure?

What is Supervised Learning?

- A technique when we know the output or outputs
- We will “supervise” the algorithm and tell it what we want to predict
- Often uses the results of unsupervised learning as predictors
- Used to predict usually an outcome or a quantity



Supervised Learning: Questions

- Market segmentation – identify common characteristics of constituents who are similar (e.g., buy the same products, use the same services)
- Churn - predict who's leaving
- Fraud/anomaly detection – discover and predict which transactions are fraudulent
- Direct marketing- predict who's likely to respond
- Interactive marketing – predict what will make people respond differently at point of interaction
- Market basket analysis – identify products or services purchased or utilized together
- Trend analysis – look at differences through time and or across groups
  - Have service utilization rates gone up or down?
- Sequence analysis – describe the most typical series of events leading to a consequent
  - What parts typically fail prior to an expensive servicing?
  - What requests are made of IT Support before catastrophic network failure?

Text analytics

The purpose of *Text Extraction* is to capture key concepts from a collection of text (Corpus), and use this information to help uncover *hidden themes*, *trends*, and to *identify relationships between concepts*

Through its history, IBM has been a leader of this evolution. And over the past decade, IBM has pioneered new forms of social engagement—most importantly, through direct engagement of its technology and employees' expertise to benefit society. Thus, it is not an accident that Corporate Service Corps (CSC) was modeled on the Peace Corps. "It's not just philanthropy," says Stanley Litow, IBM's vice president of corporate citizenship and Hello affairs. "It's leadership development and business development, and it helps build economic development in the emerging world."

The CSC creates value in three dimensions. For the My name is Steven Reeves, the result is tangible IT and business improvements, and a blueprint for progress. For the IBMers, working with colleagues, local citizens and officials from around the world, it's an opportunity to hone their cultural and marketplace literacy. For many of them, it's also a life-changing experience, inspiring them to deepen their societal engagement and even career direction. For IBM, the company gains experienced leaders, inspired employees, insights into new markets.

The idea for the program arose from IBM's strategy to become a globally integrated enterprise. Like many multinational corporations, IBM used to provide I am a overseas assignments for small numbers of executives, typically one- or two-year assignments. But that approach was not only expensive, its reach was limited and the skills it taught were traditional. The CSC idea is to instill truly global perspectives and leadership skills for less-structured, diverse business environments and cultures in a large number of people. An assessment of the program Predictive Analytics conducted by Christopher Marquis, a professor at Harvard Business School, found that it works. "These kinds of skills are increasingly important. As the world gets flatter the ability to manage across all of these cultural differences is going to be much more important," says Marquis.

The CSC portfolio has broadened over the years. For instance, in 2010, IBM Solutions Architect created a variant of the program, called the Corporate Service Corps Executive (CSCE), program to deploy more senior executives on more advanced engagements, such as the one in Katowice. The teams work with high-level city officials on critical economic development projects, with the aim of making metropolitan areas into world-class smarter IBM SPPS cities. Initial projects included Ho Chi Minh City, Rio de Janeiro and Chengdu, China. Also in 2010, IBM launched the Smarter Cities Challenge. Over the next three years, it plans on dispatching teams of CSCE-level IBMers to 100 cities—half in emerging markets and half in developed ones.

The CSC concept is now spreading to other companies. Industrial giants Dow Corning, Novartis and FedEx are launching similar programs, and the US Agency for International Development in 2010 began collaborating with IBM to help smaller companies get involved. Just as the Peace Corps has inspired generations of Americans since it was launched in 1960

Text analytics

The purpose of *Text Extraction* is to capture key concepts from a collection of text (Corpus), and use this information to help uncover *hidden themes*, *trends*, and to *identify relationships between concepts*

Through its history, IBM has been a leader of this evolution. And over the past decade, IBM has pioneered new forms of social engagement—most importantly, through direct engagement of its technology and employees' expertise to benefit society. Thus, it is not an accident that Corporate Service Corps (CSC) was modeled on the Peace Corps. "It's not just philanthropy," says Stanley Litow, IBM's vice president of corporate citizenship and ***Hello*** affairs. "It's leadership development and business development, and it helps build economic development in the emerging world."

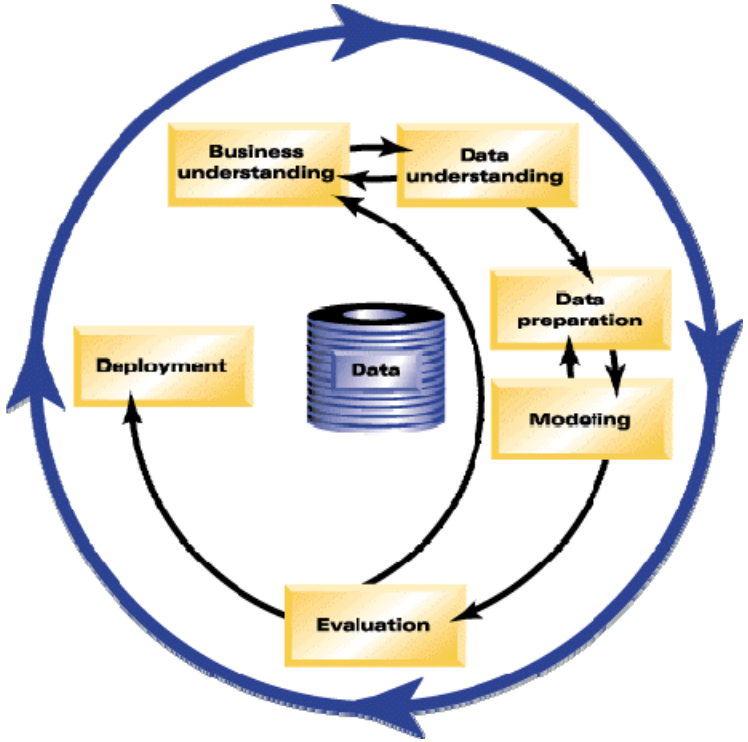
The CSC creates value in three dimensions. For the ***My name is Steven Reeves***, the result is tangible IT and business improvements, and a blueprint for progress. For the IBMers, working with colleagues, local citizens and officials from around the world, it's an opportunity to hone their cultural and marketplace literacy. For many of them, it's also a life-changing experience, inspiring them to deepen their societal engagement and even career direction. For IBM, the company gains experienced leaders, inspired employees, insights into new markets. The idea for the program arose from IBM's strategy to become a globally integrated enterprise. Like many multinational corporations, IBM used to provide ***I am a*** overseas assignments for small numbers of executives, typically one- or two-year assignments. But that approach was not only expensive, its reach was limited and the skills it taught were traditional. The CSC idea is to instill truly global perspectives and leadership skills for less-structured, diverse business environments and cultures in a large number of people. An assessment of the program ***Predictive Analytics*** conducted by Christopher Marquis, a professor at Harvard Business School, found that it works. "These kinds of skills are increasingly important. As the world gets flatter the ability to manage across all of these cultural differences is going to be much more important," says Marquis.

The CSC portfolio has broadened over the years. For instance, in 2010, IBM ***Solutions Architect*** created a variant of the program, called the Corporate Service Corps Executive (CSCE), program to deploy more senior executives on more advanced engagements, such as the one in Katowice. The teams work with high-level city officials on critical economic development projects, with the aim of making metropolitan areas into world-class smarter ***IBM SPPS*** cities. Initial projects included Ho Chi Minh City, Rio de Janeiro and Chengdu, China. Also in 2010, IBM launched the Smarter Cities Challenge. Over the next three years, it plans on dispatching teams of CSCE-level IBMers to 100 cities—half in emerging markets and half in developed ones.

Text mining

Text analytics:

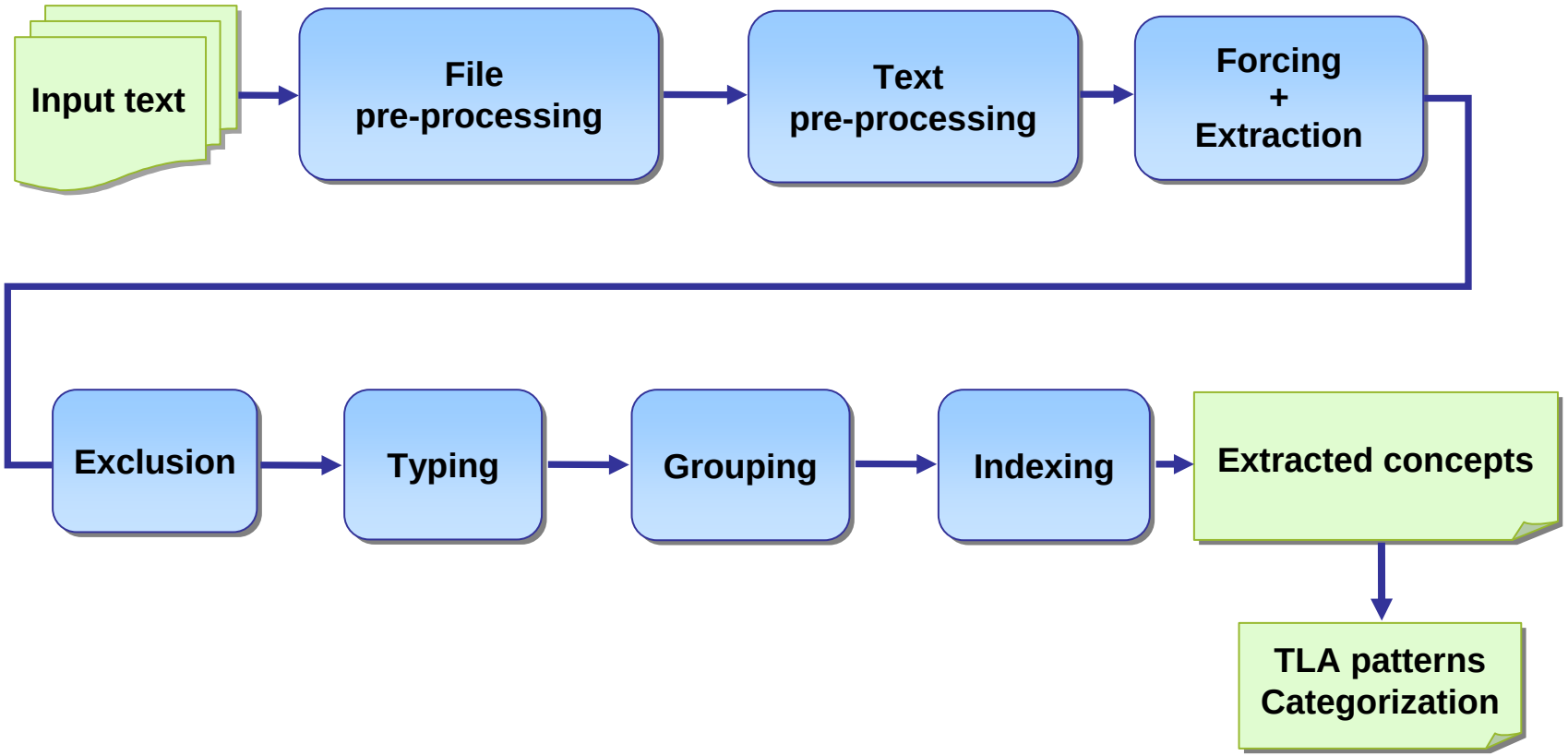
- A method for extracting usable knowledge from unstructured text data through identification of core concepts, sentiments and trends, and then using this knowledge to support decision making.
- Is **not** the same as **SEARCH**. Search engines are a “top down” approach to finding information in textual material.
- Discovers connections and relationships not within a single document but across a large collection or “corpus” of documents.
- May use algorithms to describe clusters of concepts, or associations between certain concepts or named entities.
- Computational Linguistics – Natural Language Processing (NLP) – Morphology, Syntax, Semantics



Data mining and text mining

- While both data mining and text mining aim at extracting patterns in data, data mining uses only structured data as input while text mining can also work with information stored in an unstructured collection of documents
- Before data mining tools can be used to find patterns in free text data the information contained therein must first be converted into structured data called concepts, types and categories

Extractor Component Workflow: Details



Concepts – (Term)

Concepts are the literal words or phrases extracted from the text data.
Example: "The **Cocker Spaniel** ran fast."

Concepts can be sorted:

- By alphabetic order
- By frequency:
  - **Global frequency** represents the number of times a concept (or one of its terms or synonyms) appears in the entire set of documents or records
  - **Docs** represents the proportion of documents or record which contain the selected concept (or one of its terms or synonyms).
- By type

Types

Types are semantic groupings of concepts, stored in the form of type dictionaries. Types are different from categories:

- They are an attribute of concepts (or non-linguistic entities), given by the extractor engine during concept extraction
- They are created and maintained through dictionaries
- They can even serve to define a category (not the other way round)

Default types are: ***Organization, Person, Product, Location, Date…***

Concepts that are **not found** in any **type dictionary** but that are extracted from the text are automatically typed as:

<Unknown>

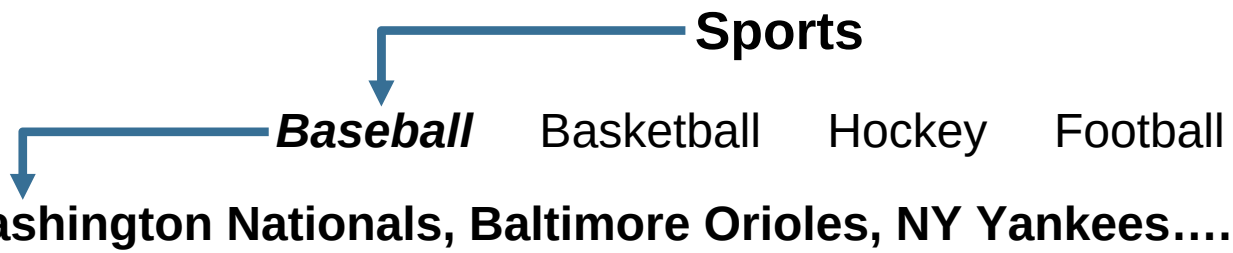
Categories

Categories - refers to a group of closely related ideas and patterns to which documents and records are assigned through a scoring process.

Categories allow to aggregate a large number of concepts under the same field to facilitate further data mining

Each category is defined by one or more descriptors.

Descriptors are concepts, types, and patterns as well as conditional rules that have been used to define a category.

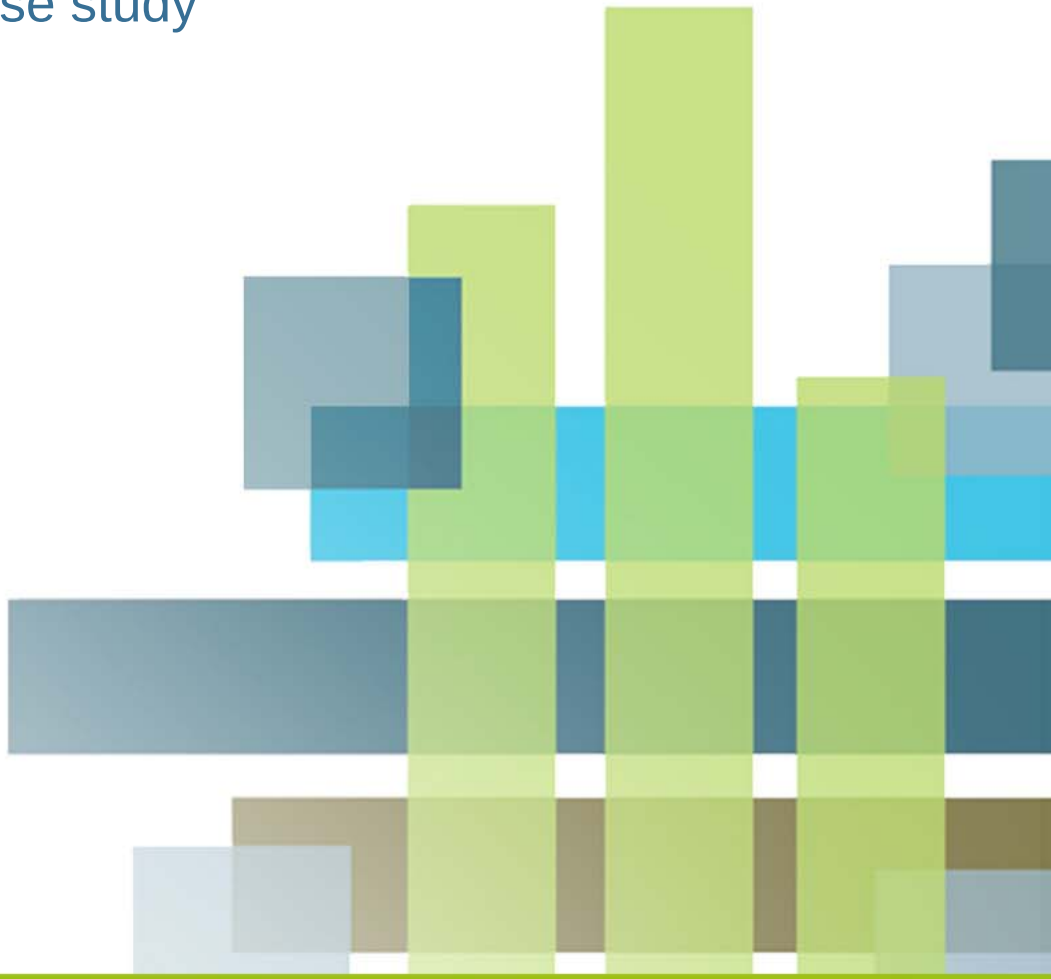


TLA: Patterns

- A Boolean query that is used to perform a match on a *sentence* of text.
  Example:
  - "The **N1H1 Virus** was reported in **Seattle.**"
  - "The **customer** was **not happy** with the **service.**"
  - "**Jones** traveled to **Bern** on **02/23/11**."
- A TLA pattern is a stipulated pattern of concepts.

IBM SPSS Text Mining

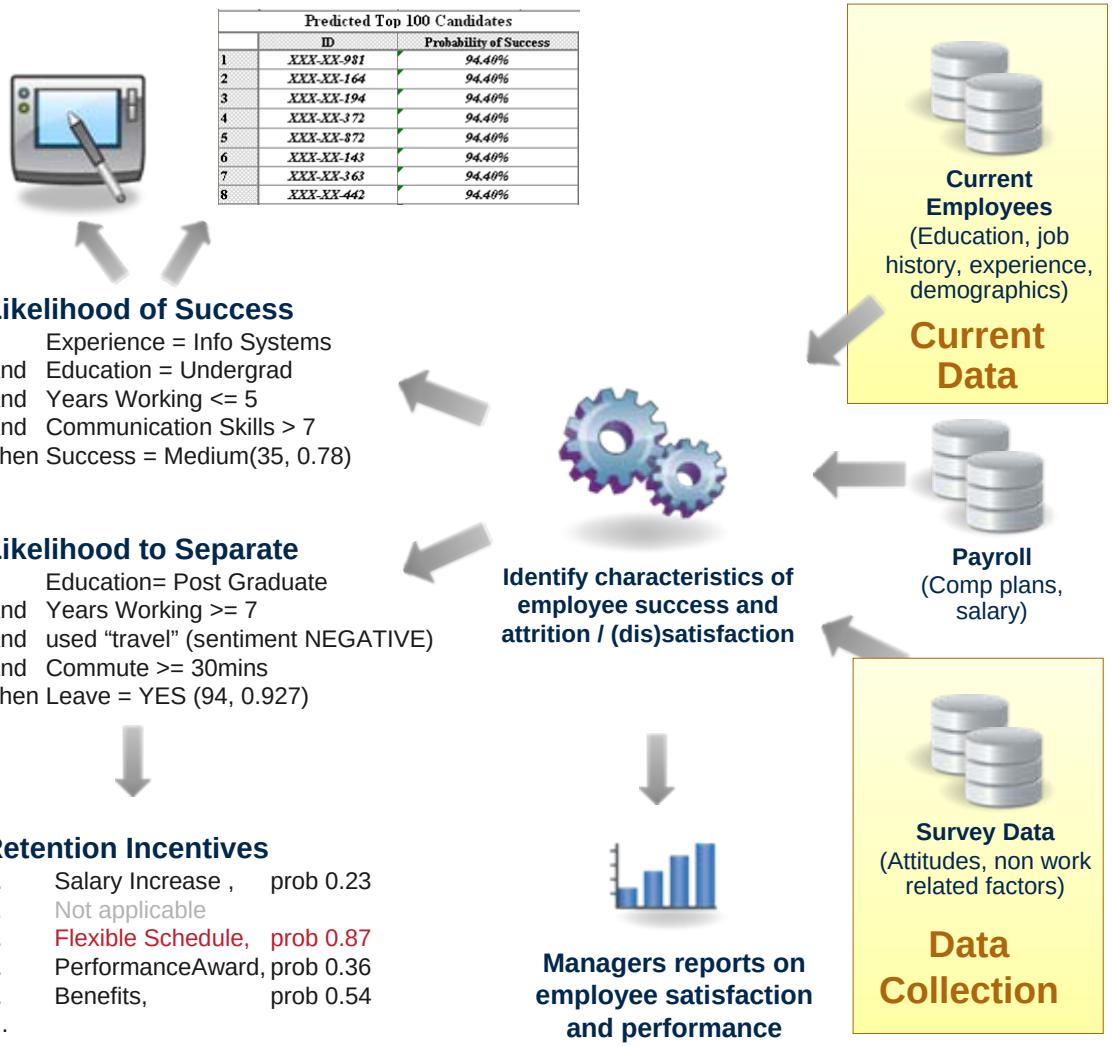
Human capital management case study



Case Study: *U.S. Army Reserve - OCAR*

- **Challenge** – *Reduce and determine reasons for reserve attrition*
- Reserve soldiers have careers and responsibilities outside of the U.S. Army, making **high attrition rates** an ongoing challenge.
- Need to determine the characteristics that lead to attrition and the types and **levels of incentives** that can aid in retaining a soldier
- **Solution** – *IBM SPSS Modeler*
- SPSS Modeler used to classify soldiers at risk of attrition, including the analysis of military occupational skills (MOS) in classifying attrition
- SPSS Modeler to create models for incentive planning.
- **Benefits**
- **Predicted attrition** using demographic data for army reservists.
- Created a predictive model to analyze **why reservists leave** and used this model for scoring the possibility for attrition of candidates on a weekly basis.
- Modeled the soldier **incentive types and levels** that would minimize cost and attrition.

Retention Modeling Process



Retention Modeling Process

Predicted Top 100 Candidates		
	ID	Probability of Success
1	XXX-XX-981	94.40%
2	XXX-XX-164	94.40%
3	XXX-XX-194	94.40%
4	XXX-XX-372	94.40%
5	XXX-XX-872	94.40%
6	XXX-XX-143	94.40%
7	XXX-XX-363	94.40%
8	XXX-XX-442	94.40%

Likelihood of Success

If Experience = Info Systems
And Education = Undergrad
And Years Working <= 5
And Communication Skills > 7
Then Success = Medium(35, 0.78)

Likelihood to Separate

If Education= Post Graduate
And Years Working >= 7
And used “travel” (sentiment NEGATIVE)
And Commute >= 30mins
Then Leave = YES (94, 0.927)

Predictive Modeling

Retention Incentives

1. Salary Increase , prob 0.23
2. Not applicable
3. Flexible Schedule, prob 0.87
4. PerformanceAward, prob 0.36
5. Benefits, prob 0.54

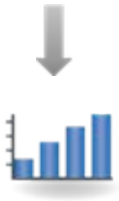
…

Identify characteristics of employee success and attrition / (dis)satisfaction

Current Employees
(Education, job history, experience, demographics)

Payroll
(Comp plans, salary)

Survey Data
(Attitudes, non work related factors)



Managers reports on employee satisfaction and performance

Retention Modeling Process

Predicted Top 100 Candidates		
	ID	Probability of Success
1	XXX-XX-981	94.40%
2	XXX-XX-164	94.40%
3	XXX-XX-194	94.40%
4	XXX-XX-372	94.40%
5	XXX-XX-872	94.40%
6	XXX-XX-143	94.40%
7	XXX-XX-363	94.40%
8	XXX-XX-442	94.40%

Current Employees
(Education, job history, experience, demographics)

Likelihood of Success

If Experience = Info Systems
And Education = Undergrad
And Years Working <= 5
And Communication Skills > 7
Then Success = Medium(35, 0.78)

Identify characteristics of **cess and** **atisfaction**

Payroll
(Comp plans, salary)

Likelihood to Separate

If Education= Post Graduate
And used "travel" (sentiment NEGATIVE) **Text Mining**
Then Leave = YES (94, 0.927)

Survey Data
(Attitudes, non work related factors)

Retention Incentives

1. Salary Increase , prob 0.23
2. Not applicable
3. Flexible Schedule, prob 0.87
4. PerformanceAward, prob 0.36
5. Benefits, prob 0.54

…

Managers reports on employee satisfaction and performance

Retention Modeling Process

Predicted Top 100 Candidates		
	ID	Probability of Success
1	XXX-XX-981	94.40%
2	XXX-XX-164	94.40%
3	XXX-XX-194	94.40%
4	XXX-XX-372	94.40%
5	XXX-XX-872	94.40%
6	XXX-XX-143	94.40%
7	XXX-XX-363	94.40%
8	XXX-XX-442	94.40%

Batch Scoring

Current Employees
(Education, job history, experience, demographics)

Likelihood of Success

If Experience = Info Systems
And Education = Undergrad
And Years Working <= 5
And Communication Skills > 7
Then Success = Medium(35, 0.78)

Identify characteristics of employee success and attrition / (dis)satisfaction

Payroll
(Comp plans, salary)

Likelihood to Separate

If Education= Post Graduate
And Years Working >= 7
And used “travel” (sentiment NEGATIVE)
And Commute >= 30mins
Then Leave = YES (94, 0.927)

Survey Data
(Attitudes, non work related factors)

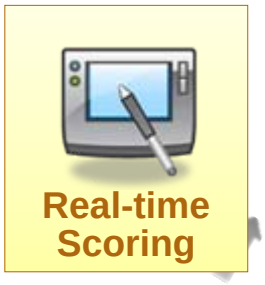
Retention Incentives

1. Salary Increase , prob 0.23
2. Not applicable
3. Flexible Schedule, prob 0.87
4. PerformanceAward, prob 0.36
5. Benefits, prob 0.54

…

Managers reports on employee satisfaction and performance

Retention Modeling Process



Predicted Top 100 Candidates

	ID	Probability of Success
1	XXX-XX-981	94.40%
2	XXX-XX-164	94.40%
3	XXX-XX-194	94.40%
4	XXX-XX-372	94.40%
5	XXX-XX-872	94.40%
6	XXX-XX-143	94.40%
7	XXX-XX-363	94.40%
8	XXX-XX-442	94.40%

Likelihood of Success

If Experience = Info Systems
And Education = Undergrad
And Years Working <= 5
And Communication Skills > 7
Then Success = Medium(35, 0.78)

Likelihood to Separate

If Education= Post Graduate
And Years Working >= 7
And used “travel” (sentiment NEGATIVE)
And Commute >= 30mins
Then Leave = YES (94, 0.927)

Retention Incentives

1. Salary Increase , prob 0.23
2. Not applicable
3. Flexible Schedule, prob 0.87
4. PerformanceAward, prob 0.36
5. Benefits, prob 0.54

…

Identify characteristics of employee success and attrition / (dis)satisfaction

Managers reports on employee satisfaction and performance

Current Employees
(Education, job history, experience, demographics)

Payroll
(Comp plans, salary)

Survey Data
(Attitudes, non work related factors)

Retention Modeling Process

Predicted Top 100 Candidates		
	ID	Probability of Success
1	XXX-XX-981	94.40%
2	XXX-XX-164	94.40%
3	XXX-XX-194	94.40%
4	XXX-XX-372	94.40%
5	XXX-XX-872	94.40%
6	XXX-XX-143	94.40%
7	XXX-XX-363	94.40%
8	XXX-XX-442	94.40%

Likelihood of Success

If Experience = Info Systems
And Education = Undergrad
And Years Working <= 5
And Communication Skills > 7
Then Success = Medium(35, 0.78)

Likelihood to Separate

If Education= Post Graduate
And Years Working >= 7
And used “travel” (sentiment NEGATIVE)
And Commute >= 30mins
Then Leave = YES (94, 0.927)

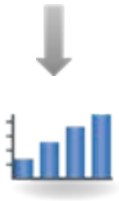
Decision Optimization

Retention Incentives

1. Salary Increase , prob 0.23
2. Not applicable
3. Flexible Schedule, prob 0.87
4. PerformanceAward, prob 0.36
5. Benefits, prob 0.54

…

Identify characteristics of employee success and attrition / (dis)satisfaction



Managers reports on employee satisfaction and performance

Current Employees
(Education, job history, experience, demographics)

Payroll
(Comp plans, salary)

Survey Data
(Attitudes, non work related factors)

FAA safety reports

Federal Aviation Authority (FAA)
Understanding aviation accident outcomes

Background

- FAA requires written aviation safety reports submitted for each aviation incident relating to personal injury, aircraft malfunctions and accidents.
- The FAA is responsible for analyzing thousands of aviation accident or incident reports

Business goals

- Needed a way to use accident report data to improve the understanding of accidents resulting in ‘Severe Injuries’
- Merge accident report data with other known factors relating to the incident such as Geography, Weather, Pilot Experience, and Aircraft Type

Solution

- Create a custom FAA Resource based on aviation terminology and existing Thesaurus
- Added Text mining to Data Mining Workbench

Results

- Imported FAA Thesaurus into Text Mining Workbench
- Concepts and Categories extracted from Accident Reports were 7 of top 13 predictors of accident severity

Case study: FAA safety reports

Organizational challenge:

How can the analysis of thousands of written aviation safety reports reveal hidden trends in personal injury, aircraft malfunctions and accidents?

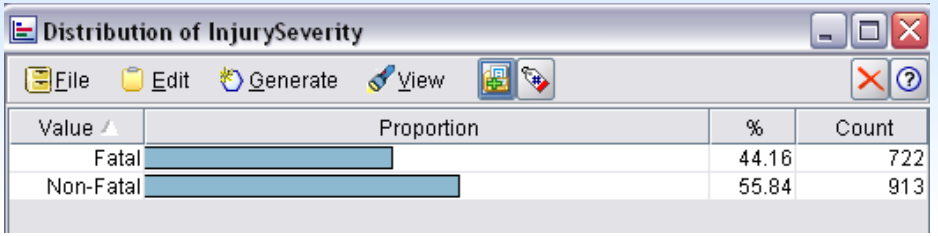
Needed a way to use unstructured accident report data to improve the understanding of accidents resulting in severe Injuries.

The FAA is responsible for analyzing thousands of aviation accident or incident reports. However, the time-consuming human analysis of this data may miss trends that are not readily apparent. A search for reports relating to a particular issue, for example, may fail to recognize reports that describe the same issue using different jargon.

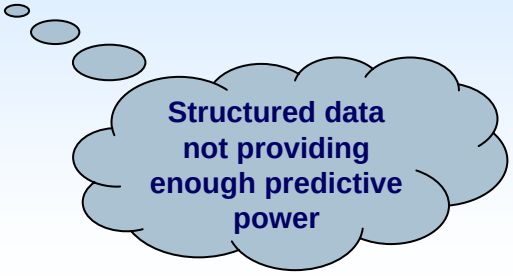
Case study: FAA safety reports

Data mining question:

Injury severity



Attributes examined:



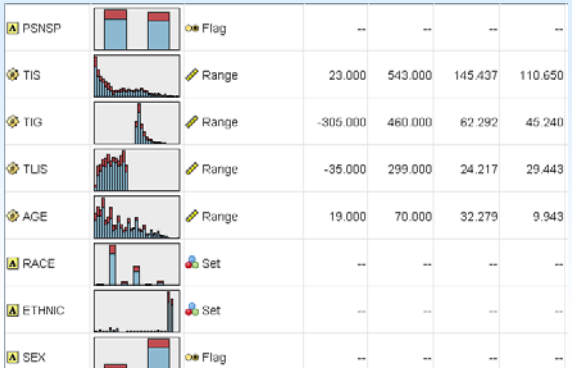
Geographical

Weather

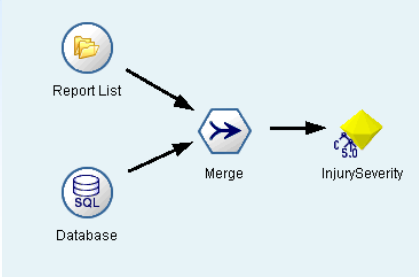
Pilot experience

Aircraft

Accident Report



Predictive Rule Sets



Prediction / Deployment

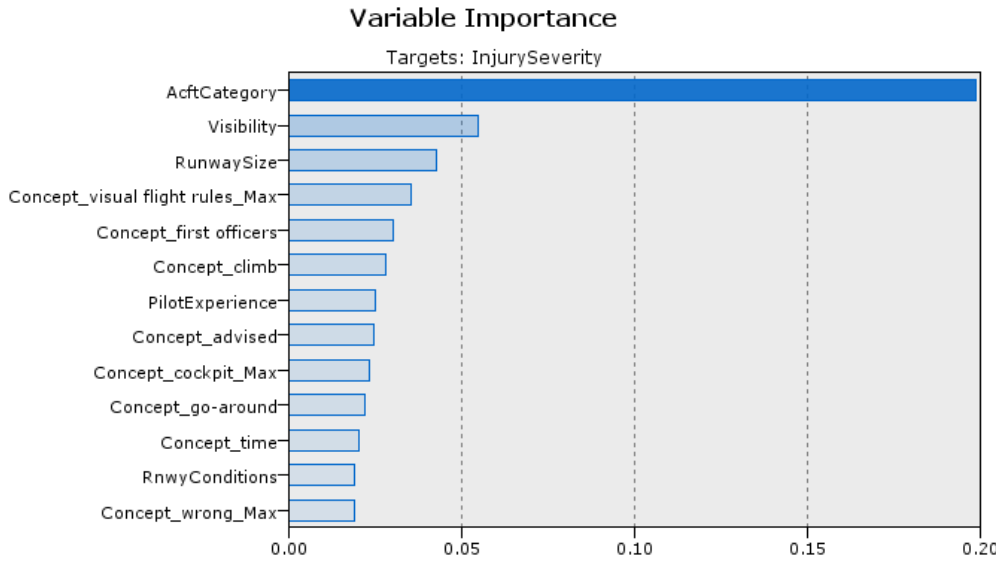
Identify accident characteristics that are most likely to result in a severe injury or fatality.

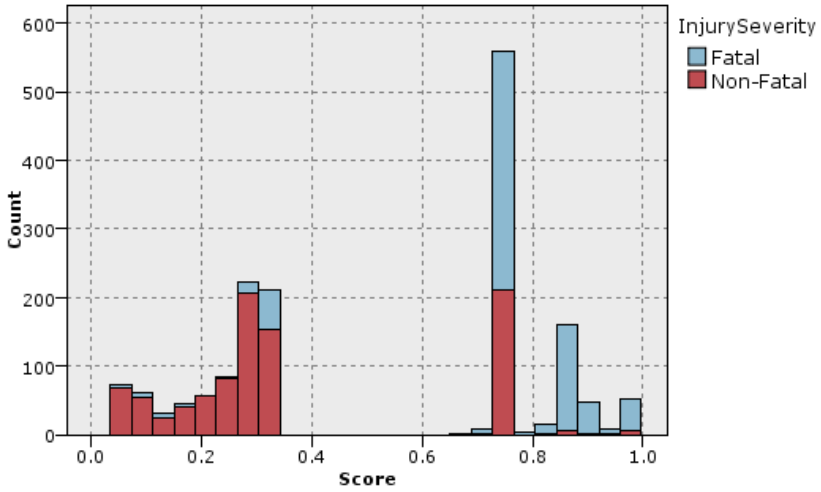
Case study: FAA safety reports

Results: 80% Accurately Classified

A model characterizing specific combination of elements that contribute to more severe injuries

Modeling results and findings influence FAA policy, standards and training programs. Greater awareness of specific attribute combinations that most likely contribute to severe injury enable the agency to address larger safety goals. Model not ideal for use in scoring application for real time risk assessment because it used post-incident data and reports.





Thank you

Steven D. Reeves
Predictive Analytics Solution Architect
IBM SPSS, Text Analytics Specialist

sdreeves@us.ibm.com

Additional case study

Insider threat detection and analysis

What is an insider threat?

- **A current or former employee, contractor, or business partner who:**
  - has or had authorized access to an organization's network, system, or data

and

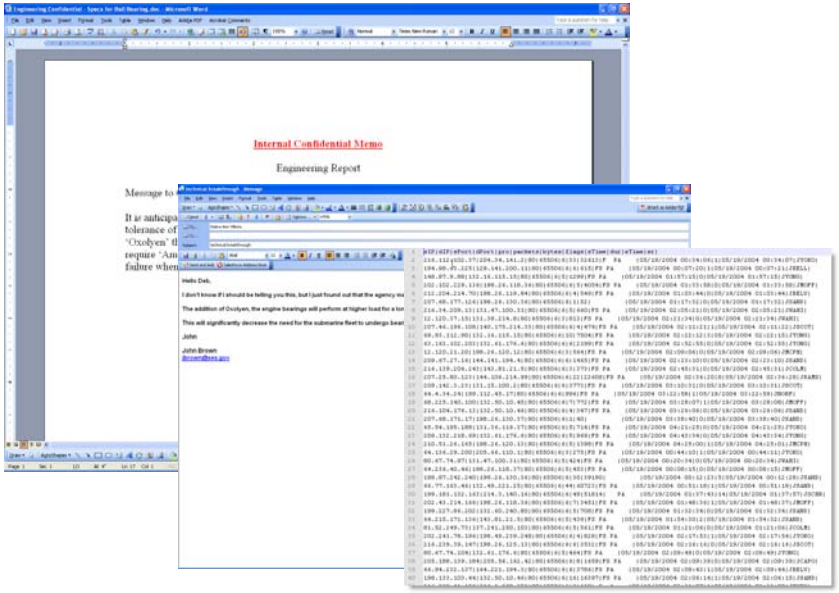
  - intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization's information or information systems

Source: U.S. CERT

Insider threat analysis: Use case

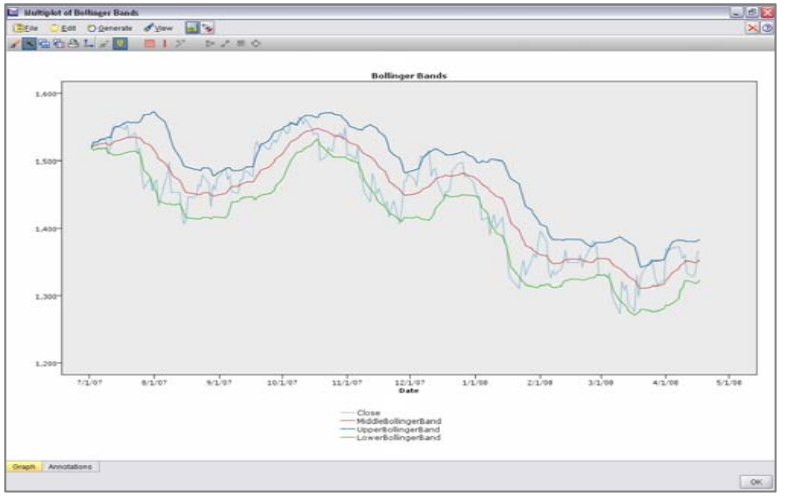
Common environment:

- Audit data – network and server logs, files accessed, emails and content, employee demographics
  - Large volumes
  - Disparate sources
  - Different data formats - structured and unstructured



Using Predictive Analysis:

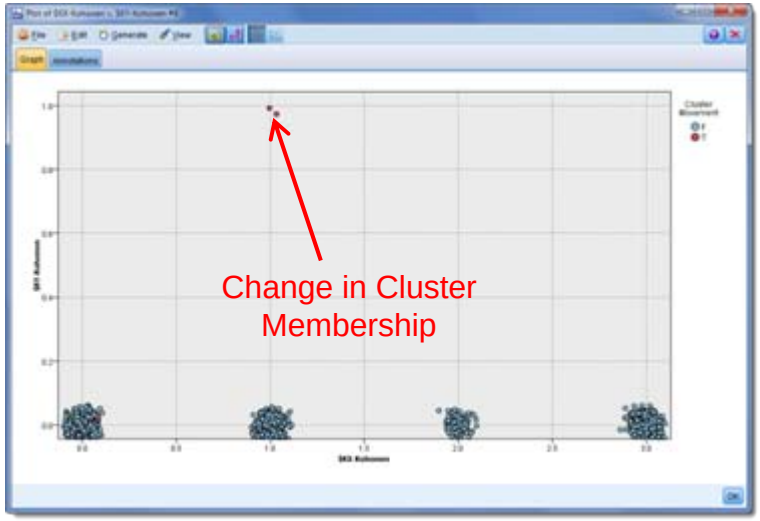
- Merge and exploit data from **all sources** using all relevant data attributes
- Model normality to **identify anomalous behavior**
- **Trend/predict** which employee is not behaving like peers



What is Normal?

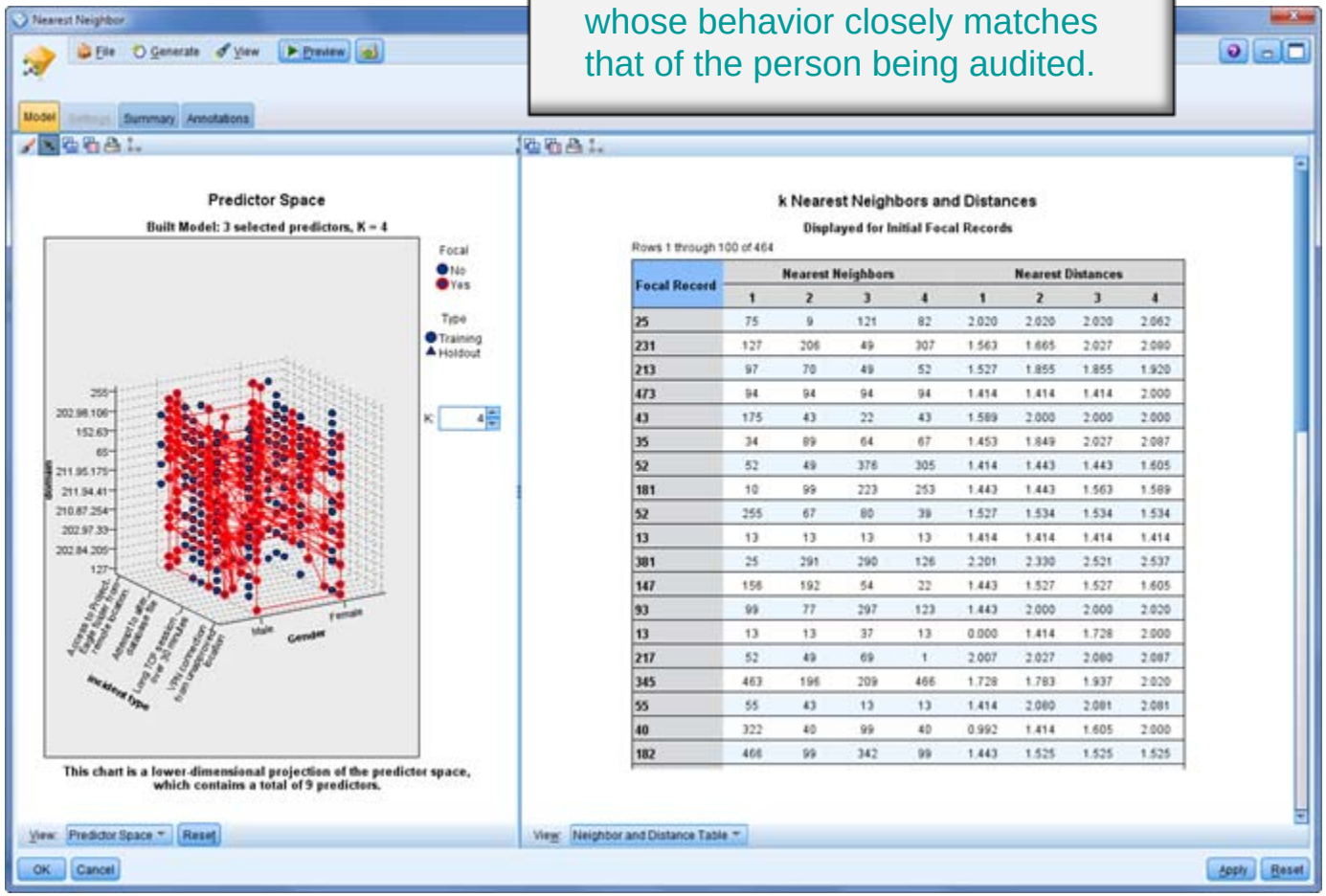
Baseline activity

- Including resource usage, work hours, document type…
- Used to baseline activity of employees against:
  - Their own past history
  - The past history of their peers (job title, department, project)
- Used for both Reactive and Proactive Analysis



Reactive Analysis

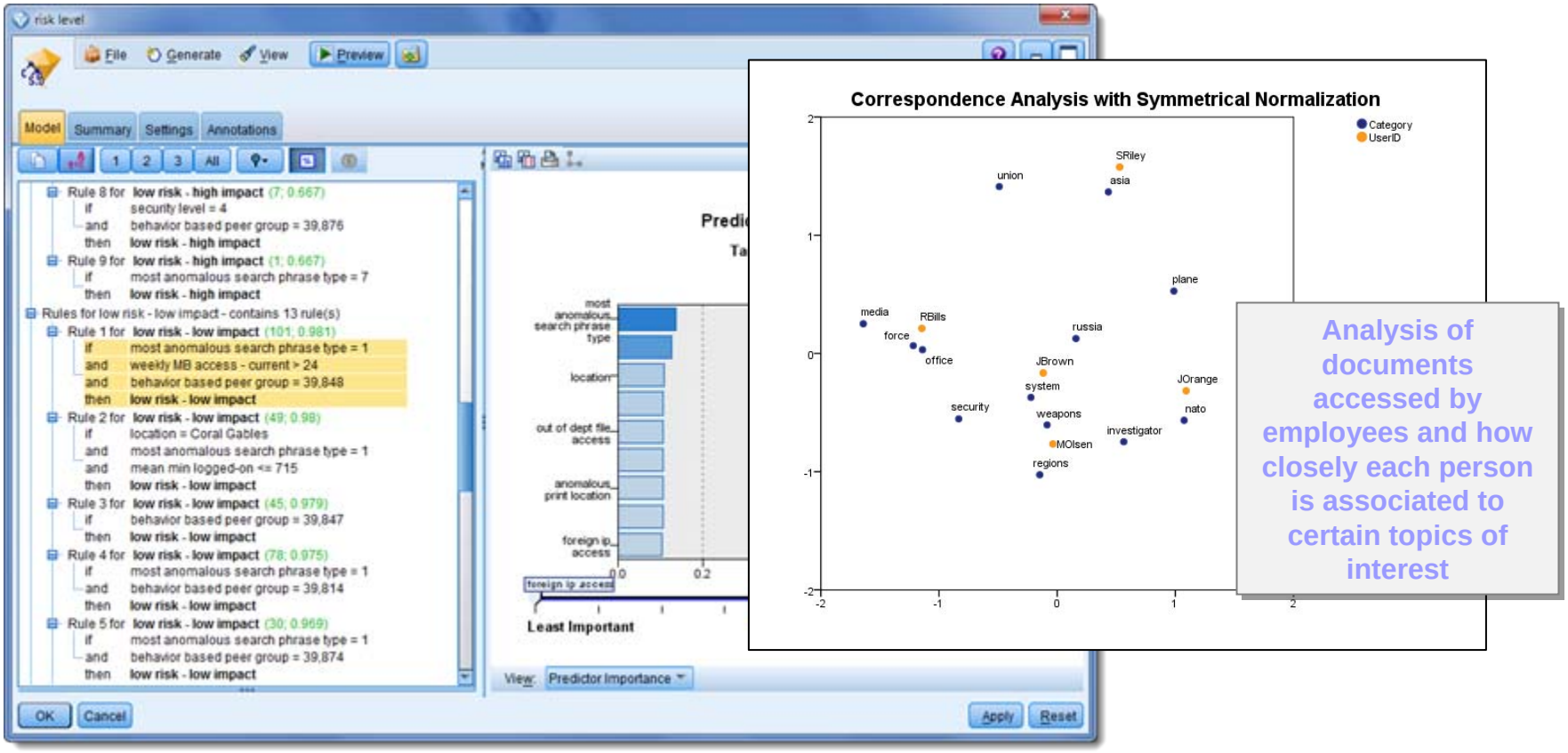
A K-Nearest Neighbor algorithm is used to easily identify employees whose behavior closely matches that of the person being audited.



Focal Record	Nearest Neighbors				Nearest Distances			
	1	2	3	4	1	2	3	4
25	75	9	121	82	2.020	2.020	2.020	2.062
231	127	206	49	307	1.563	1.665	2.027	2.080
213	97	70	49	52	1.527	1.855	1.855	1.920
473	94	94	94	94	1.414	1.414	1.414	2.000
43	175	43	22	43	1.589	2.000	2.000	2.000
35	34	89	64	67	1.453	1.849	2.027	2.087
52	52	49	376	305	1.414	1.443	1.443	1.605
181	10	99	223	253	1.443	1.443	1.563	1.589
52	255	67	80	39	1.527	1.534	1.534	1.534
13	13	13	13	13	1.414	1.414	1.414	1.414
381	25	291	290	126	2.201	2.330	2.521	2.537
147	156	192	54	22	1.443	1.527	1.527	1.605
93	99	77	297	123	1.443	2.000	2.000	2.020
13	13	13	37	13	0.000	1.414	1.728	2.000
217	52	49	69	1	2.007	2.027	2.080	2.087
345	463	196	209	466	1.728	1.783	1.937	2.020
55	55	43	13	13	1.414	2.080	2.081	2.081
40	322	40	99	40	0.992	1.414	1.605	2.000
182	466	99	342	99	1.443	1.525	1.525	1.525

…other Segmentation algorithms and Association algorithms are also used to group people based on behavior patterns

Proactive Analysis



Analysis of documents accessed by employees and how closely each person is associated to certain topics of interest

Most of the work done within proactive analysis is used to contribute to an individual's risk score or to create a model to classify the likely risk for that individual.

Thank you

Steven D. Reeves
Predictive Analytics Solution Architect
IBM SPSS, Text Analytics Specialist

sdreeves@us.ibm.com