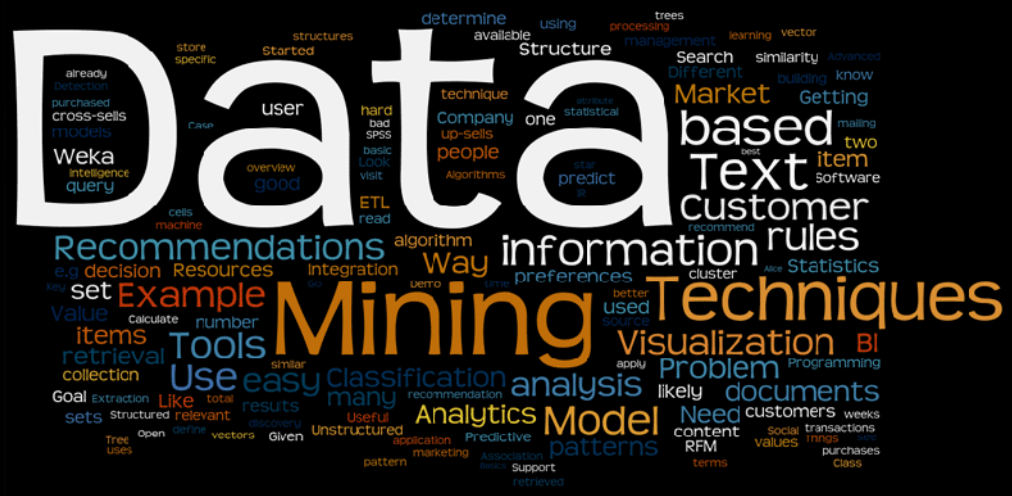
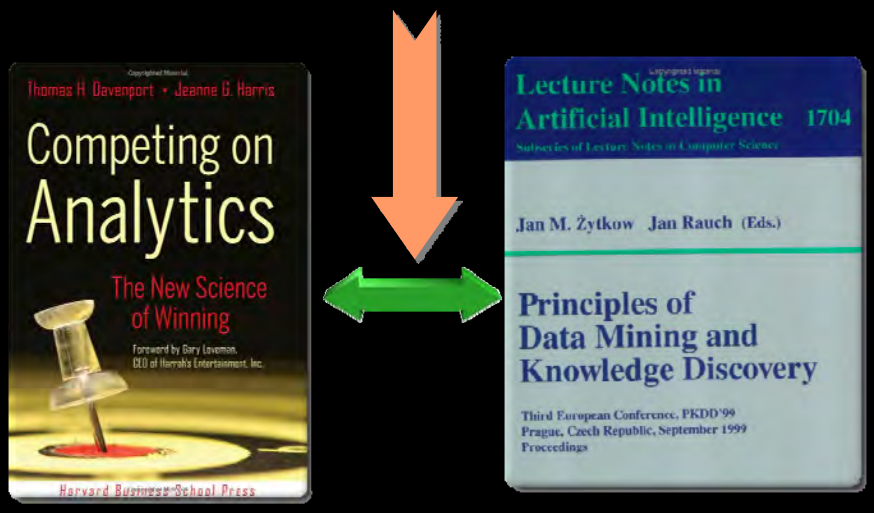
Advanced Analytics:
A Framework for Data Mining, Statistics, Visualization and Other Techniques



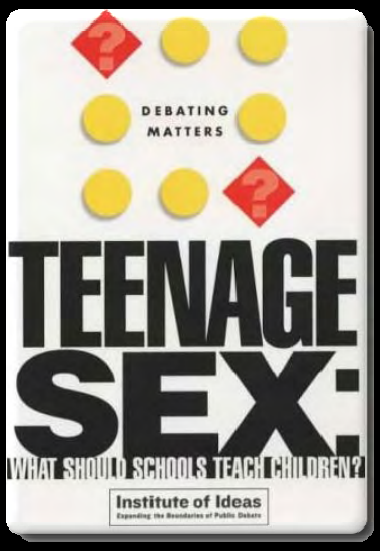
Mark Madsen, Third Nature, Jos van Dongen, Tholis May, 2010

You are Here (sort of)



And so are we...

Advanced analytics is like…



Third Nature

Why the Name Advanced Analytics?

There are more techniques for analyzing data than just data mining.

BI has become synonymous with reporting.

We needed another term so industry analysts could create some new market forecasts.

What most people aren't doing must be advanced since most people aren't doing it. Q.E.D.

Third Nature

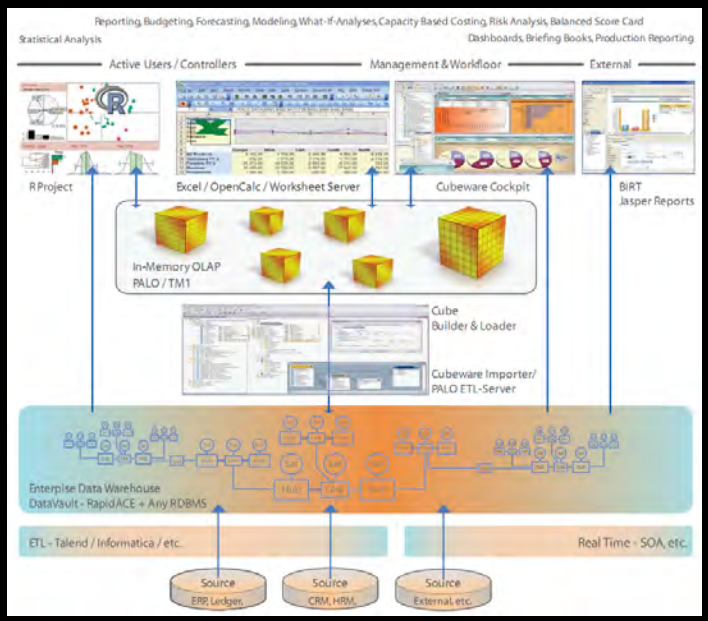
Why is it More Feasible Now?

Data collection, particularly of people's behavior, makes new things possible.

The data volume and complexity require methods other than basic query and reporting.

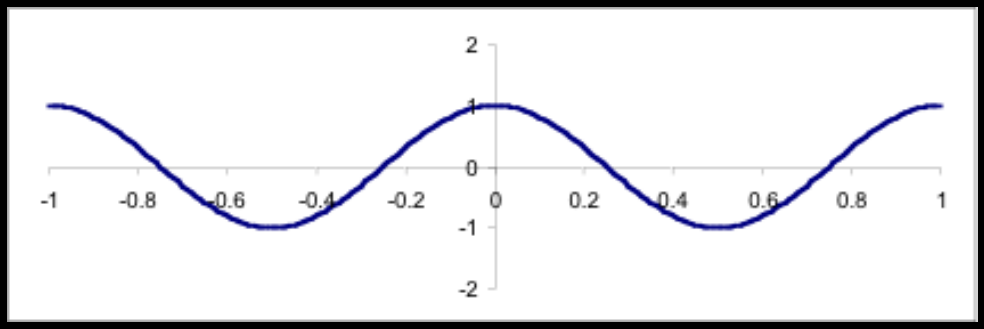
Commodity computing makes complex work possible.

We're Building on Existing Data Infrastructure



Unfortunately, many of the tools still don't talk to or work well with the DW.

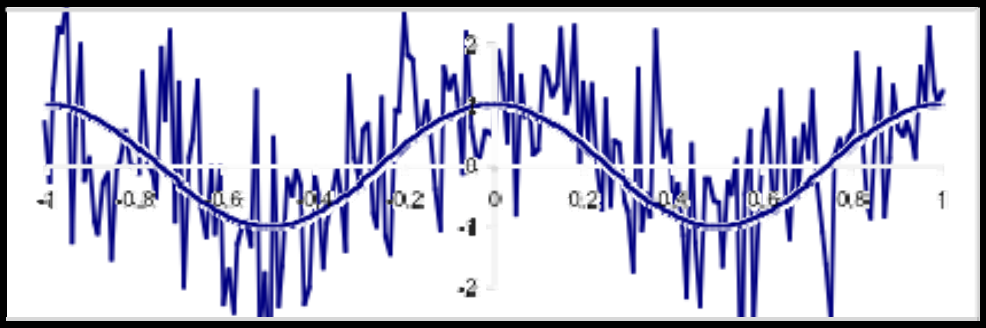
Consequences of Large Data Volumes



Traditional reporting and queries work well when looking at basic exceptions on core transactional data.

Third Nature

Consequences of Large Data Volumes



Sums, counts and sorted results only get you so far.

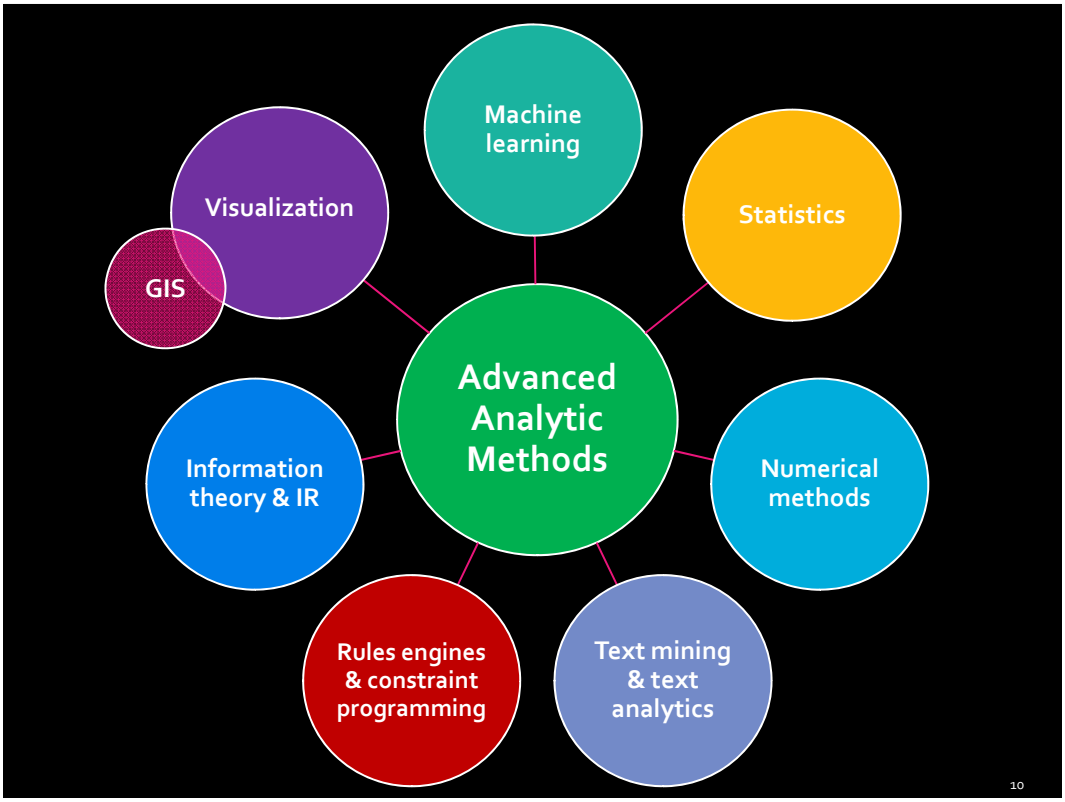
Luckily, with lots of data you don't have to be as clever with statistics as you did in the days of scarce data.

Third Nature

What You Need for Advanced Analytics

1. Data
   - It needs to be sourced from somewhere. DWs have readily available data and ETL tools to get missing data.
   - BI tools help you identify and isolate the data you need.
2. *Clean* data
   - Your models are as good as the data that goes in. Most data mining techniques are sensitive to noise, inaccuracy, and missing values.
3. *Pre-processed* clean data
   - Most of the time you can't run against your existing data. You need to prepare the data:
     - Group continuous variables, symbolic to numeric transformations, coding categorical or discrete values, suppress correlated data
4. Model building, testing, validation, and operational data storage
   - The process of building, testing, validating is iterative
   - You need data management infrastructure to run your models in production.

Like a DW, ~70% of analytics project effort is spent preparing data.



Before We Start: Some Basics

Probability

The chance of something happening is represented by a value between zero and one, usually the term *p* is used in a formula.

When *p* = 0, the event can't happen

When *p* = 1, the event is a certainty

You can estimate the probability of an event (e) by running a series of N tests and counting the specific outcomes (S) you are interested in:

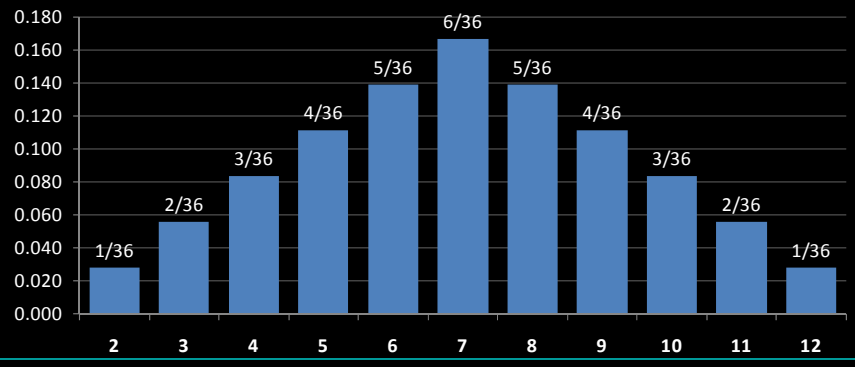
p(e) = S/N

Third Nature

Probability and Distribution of Outcomes

The likelihood of rolling a particular number on a pair of dice follows a distribution defined by a probability density function. It's easy to visualize, and see that rolling 7 is six times more likely than rolling 2 or 12.

Most times you'll have to do more work.



Third Nature

Confidence and Margin of Error

There's some fuzz in our numbers – they aren't exact to the penny like our transactions are.

Margin of error is always +/- a percent

Confidence is the probability that a number will be within the MoE range of the value we're stating.

For example:

"83% of people love our soup"

+/-3% margin of error

So 95% of the time, a survey of this will be in the range of 77% - 86%.

Third Nature

Utility and Accuracy

Usefulness means reducing uncertainty about a decision, but with reasonable cost.

Mailing cost example:

- Probability of a response *r*: 2%
- Expected value of a response *v*: $100
- If mail cost per person > $2 [*p(r) * v*] then create a better offer or find better people to mail to because you'll lose money

More accurate results is great but being useful is more important.

- E.g. a model can be 99% accurate and useless

BI – Statistics – Data Mining?

Business Intelligence

- Answers to predefined questions
- Structure & content of data known beforehand
- Can only show the obvious

Statistics

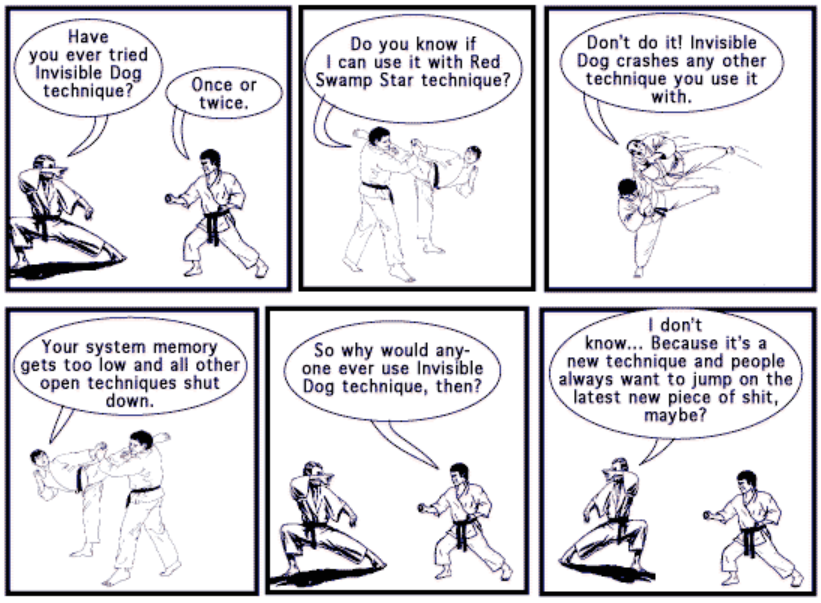
- Scientific application of mathematical principles to the collection, analysis, and presentation of numerical data (ASA)
- Adds the notion of Probability rather than certainty
- One of the foundations for advanced analytics

Data Mining

- Algorithms evolved from pure statistics and other fields
- Non-numerical data added to the equation
- Guided / automated application of techniques

Data Mining Techniques

Every model is wrong. Some are useful.



http://www.mnftiu.cc

What is Data Mining?

"The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"

Prof. Pier Luca Lanzi

"The extraction of hidden predictive information from large databases"

Kurt Thearling

Also:

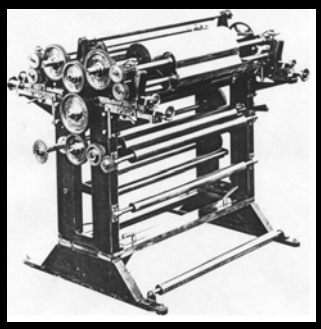
- Knowledge Discovery
- Machine Learning
- "Statistics + Marketing"(frustrated statisticians)
- "Anything the computer does that we don't understand."

Value of Data Mining

Ability to automate detection of patterns in data

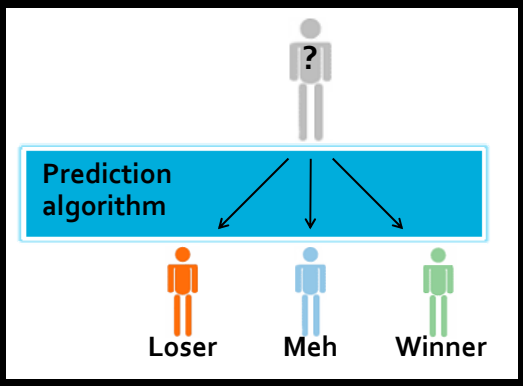
Predictive capabilities

Making *better use of data* that a business already collects in the normal course of business operations



A Useful Way to Look at Techniques

Predictive models



Predictive models can determine an output variable or rank-order based on the input data.

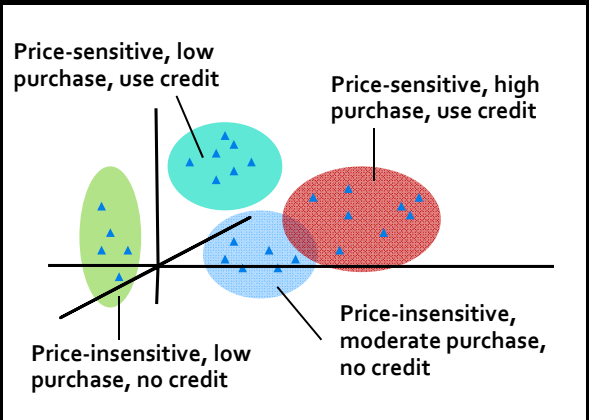
***What*:** Use known variables to predict unknown or future values of other variables.

***For example*:** Determine the odds of an event. Will a customer respond to an offer? Is this a fraudulent transaction?

Techniques:
Linear regression
Neural networks*
Decision, regression trees

A Useful Way to Look at Techniques

Descriptive Models



In theory **you could use these clusters of the existing base to predict the behavior of new customers based on the same attributes.**

***What*:** These find relationships and patterns, or define models that describe the data in question.

For example: Define customer segments based on attributes to better plan features or product development.

Techniques:

- Classification and clustering
- Association rules
- Sequential pattern discovery
- Logistic regression

Another Way to Group Techniques

Most data mining techniques fit into two groups based on how you work with them:

Supervised learning: a person has to define the correct output for some portion of the data. Data is divided into training sets used for model building and test sets for validating the results.

- *What constitutes a good vs. bad credit risk?*

Unsupervised learning: the algorithm functions without an agent specifying the action to take on a given set of input. The algorithm derives the pattern.

- *What is the best way to segment customers for marketing?*

Choosing Data Mining / Statistical Techniques

Clearly define your goal and consult the literature (or an expert) on the best techniques applicable to the problem

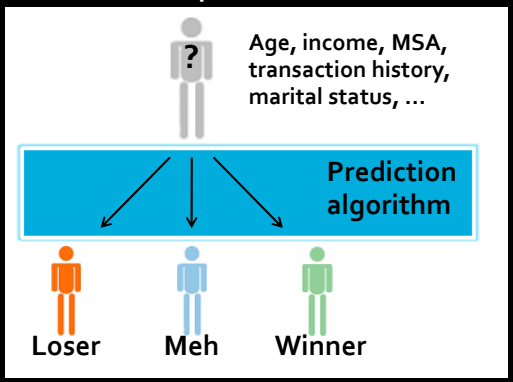
Decide how the output will be delivered or used.

Determine what data you have available - the type, volume and quality will constrain your choice of technique.

Test and compare several techniques for best results.

Classification

Classification takes a set of data (the training set) and assigns a value to the "class" attribute, identifying group membership.



The algorithms define the class as a function of the other attributes so that new, unseen data can be assigned the appropriate class.

The assumption is that you know the categories you want to assign to in advance.

The model is checked for accuracy with a test set.

Classification <> clustering. Sort of.

Classification Techniques

There are many different techniques and categories:

- Density estimation + decision rule
  - Naïve Bayes
- Define a metric space and classify based on proximity
  - Instance-based learning
  - K-nearest neighbor
- Decision regions
  - Decision trees
  - Logistic regression
  - Neural nets

WEKA supports many methods: 0R, 1R, NaiveBayes, DecisionTable, ID3, PRISM, Instance-based learner (IB1, IBk), C4.5 (J48), PART, Support vector machine (SMO)

Working With Machine Learning Algorithms

Classifiers generally rely on training data: you "teach" the system to classify your data.

This means managing your data. Use the following rule of thumb:

Training 50%	• Used to develop the initial models
Validation 25%	• Used to validate the model created • Used to remove noise
Testing 25%	• Used to compare models • Used to predict performance of final model

Third Nature

Looking at the Validity of Results

One way to look at results is to record the actual values in a confusion matrix. “True” values are the ones the model got right, “False” are the ones it got wrong.

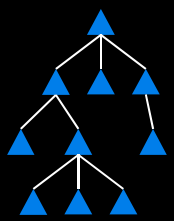
	Positives	Negatives
Positives	True positives	False positives
Negatives	True negatives	False negatives

You do this because models are sometimes good at one ting but not another, and because the cost of false classification varies by type.

Decision Trees

Decisions are rules for the classification of a dataset.

- Sets of decisions based on data attributes are represented in a tree structure.
- The ideal top level decision is based on the attribute with the highest information gain.
- Think of it as a tree representation of “20 questions”
- Useful when there are multiple ways to become a member of a class.



Decision Trees

Trees are easy to understand because the decision rule at each node is visible.

They're relatively easy to build.

Can be used for classification and prediction.

Some of the more common methods:

CART (Classification and Regression Trees: 2 way splits)

ID3/C4.5/C5.0 (C5.0 only commercially available)

J48 (adapted C4.5: Weka Tree algorithm)

CHAID (Chi Square Automatic Interaction Detection)

Association

Goal:

- Given a set of data, find rules that will predict the occurrence of an item based on the occurrences of other items in the data: e.g. {bread, milk} → {jam}

Common applications:

- Core of market basket analysis
- Determine cross-sells
- Catalog design

Challenges:

- Brute force approach (all combinations of all items in all transactions) is computationally prohibitive
- Correlation vs. causality

Association Example

- **Association Rule**
  - An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
  - Example: {Milk, Diaper} → {Beer}

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Source: Tan, Steinbach, Kumar 2004

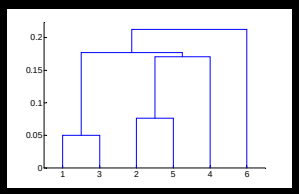
Note: correlation ≠ causality!

Clustering

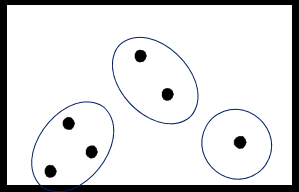
Grouping related sets of data. Like classification, only without knowing the categories in advance.

Primary types of algorithms:

Hierarchical: creates a set of nested clusters, building up larger clusters from smaller, organized as a tree. Arbitrary number of clusters.



Partitional: creates non-overlapping clusters where each element is in only one cluster. Fixed number of clusters.



Third Nature

Clustering

Most simple/popular algorithm: K-Means (1967)

- The algorithm selects k points as the initial cluster centers ("means"). A value for K is chosen by the user.
- Each point in the dataset is assigned to the closest cluster, based upon the Euclidean distance between each point and each cluster center.
- Each cluster center is recomputed as the average of the points in that cluster.
- Steps 2 and 3 repeat until the clusters converge

Determine K value based on experimentation

Techniques Vary: Different Results on Same Data

There is no "best" clustering model. Depends on:

Data types – wavelet is good for continuous data, K-means for numeric, Rock for categorical, etc.

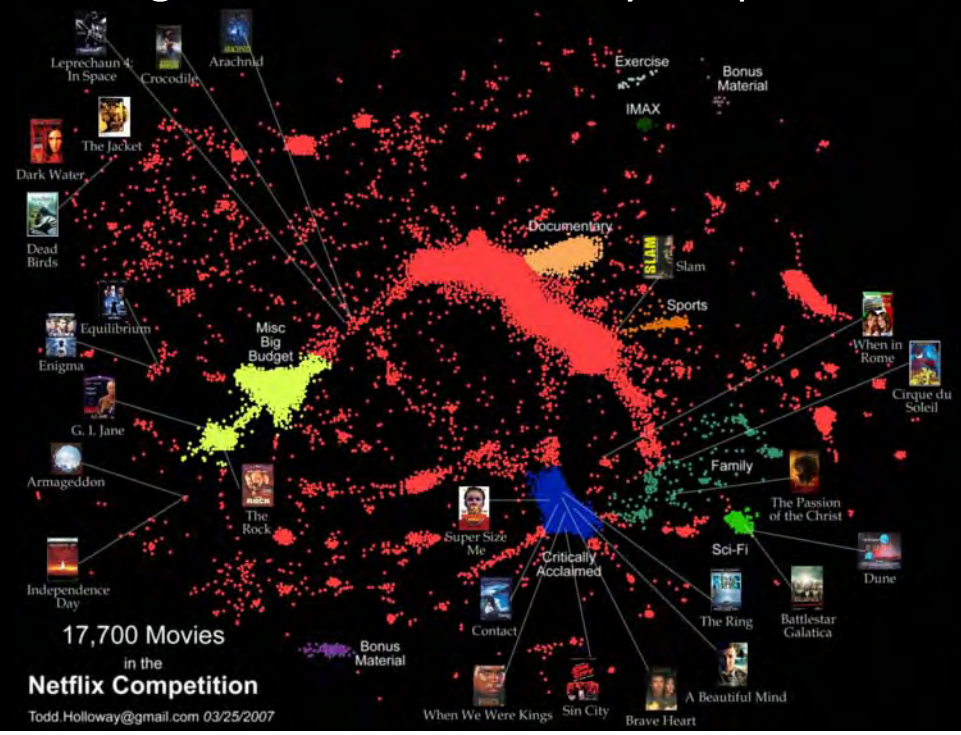
Noise and outliers – K-means and MIN don't work well, DBSCAN handles noise better

Data volume – affects compute cost, many algorithms are $O(n^2)$

Dimensionality – many attributes can cause problems in accuracy or performance

"Shape" – non-globular clusters (as defined by a distance/similarity function) can lead to poor results

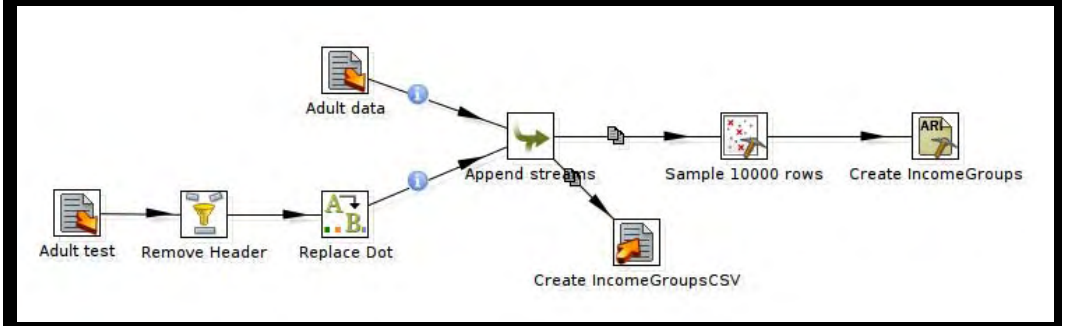
Clustering + Visualization + Query = Explanation



Data Mining & ETL

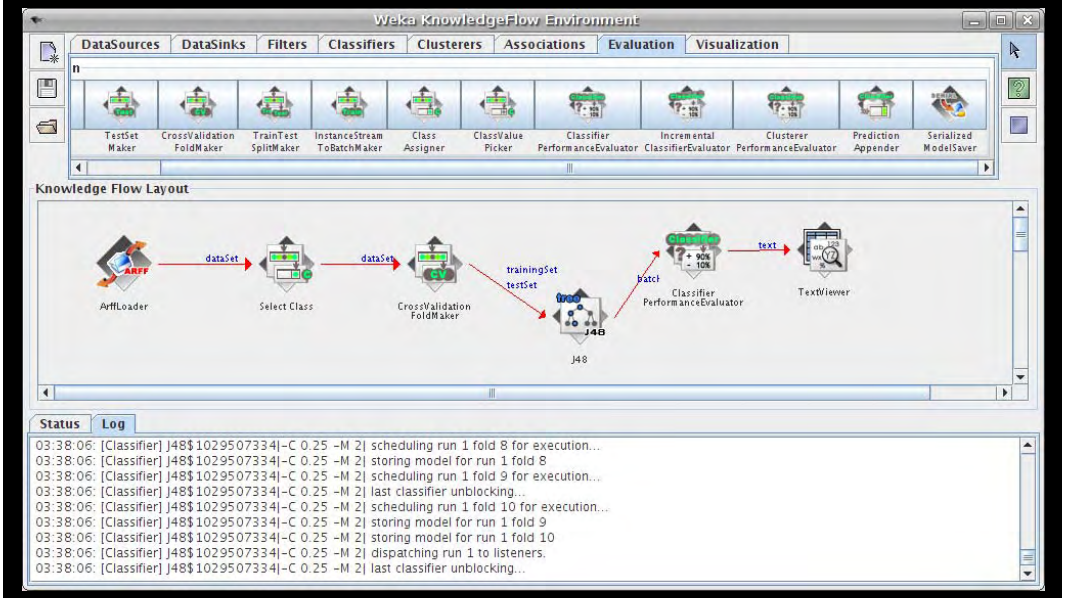
Data mining does not always mean end-user oriented output and standalone tools.

Step 1: Prepare Data



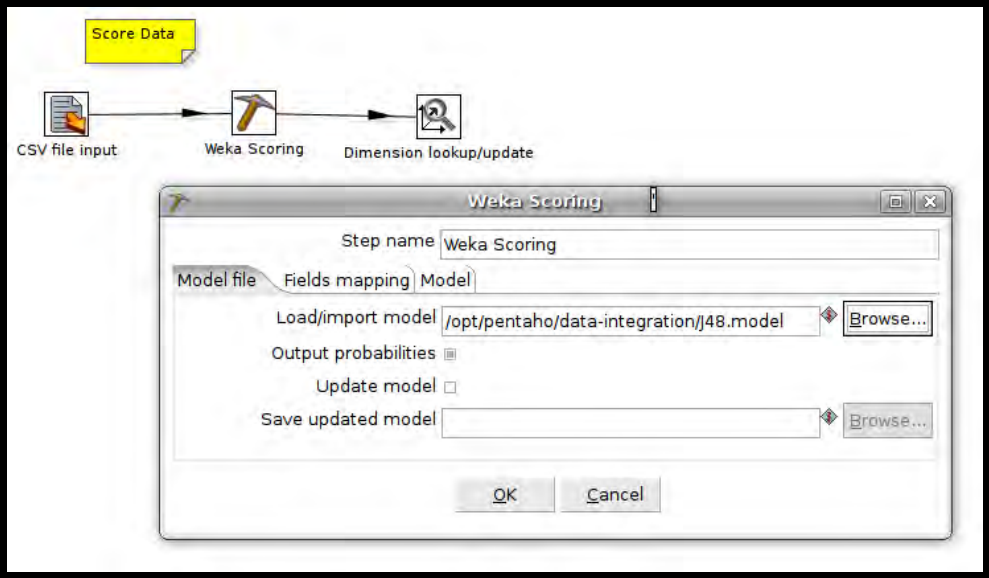
Data Mining & ETL

Step 2: Develop the model



Data Mining & ETL

Step 3: Apply the model



Visualization

Visualization Is …

The conversion of any abstract data into graphical format so the characteristics and relationships of the data can be explored and analyzed.

- Humans have the ability to analyze large amounts of information that is presented visually
- This is good for certain types of pattern and trend analysis and for seeing relationships
- It's often easy to detect outliers and unusual patterns

Useful for exploration, explanation, discovery, but *not* for automated system actions.

Your brain has a GPU, why not use it?

Visualization

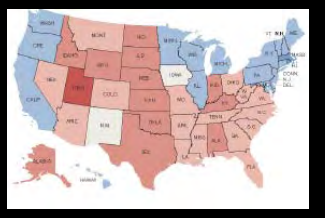
The market today is where data mining was 10 years ago.

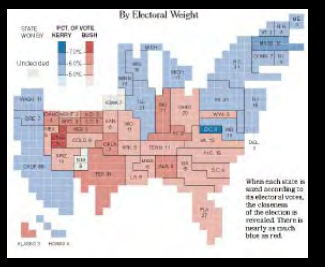
Visualization is more accessible so many amateurs are gravitating to it.

Visualization is deceptive – the topics are deep and broad: cognitive psychology, neuroscience, complex math, design, art. It's easy to do poorly (*see images at right)*

Tool market today:

- Some standalone tools
- Many single purpose tools for specific techniques
- Lots of DIY, libraries
- Limitation of most tools is that they don't talk to databases - file based applications

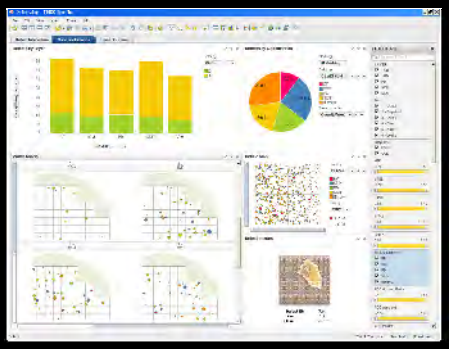


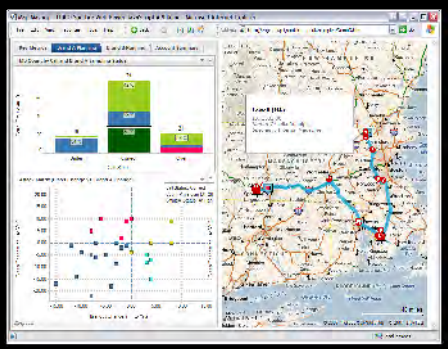


The Basics: Interactive Visualization Example

Interactive (as opposed to static) visualization.

Important elements in visual analysis and exploration are having coordinated views, and "focus + context".

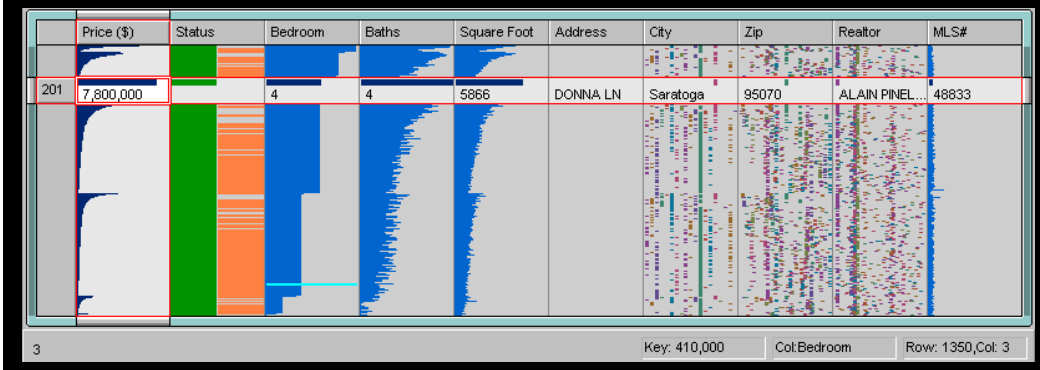




Course Grained Patterns

Viewing every row of data in a table, the old way involved tabular reporting.

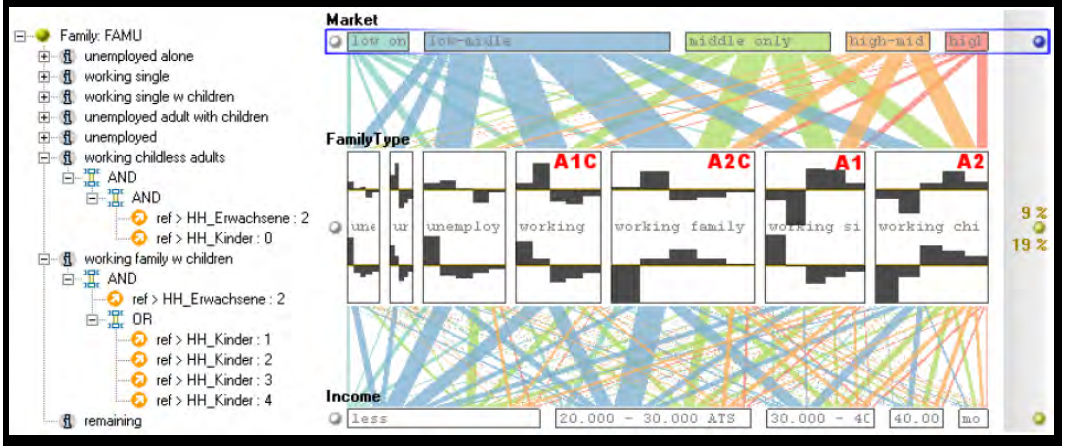
Visualizations offer the ability to view detail and abstraction, and interactively order.



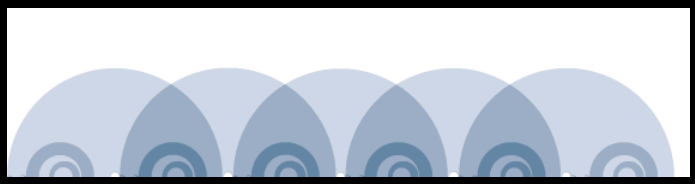
Patterns in Composition

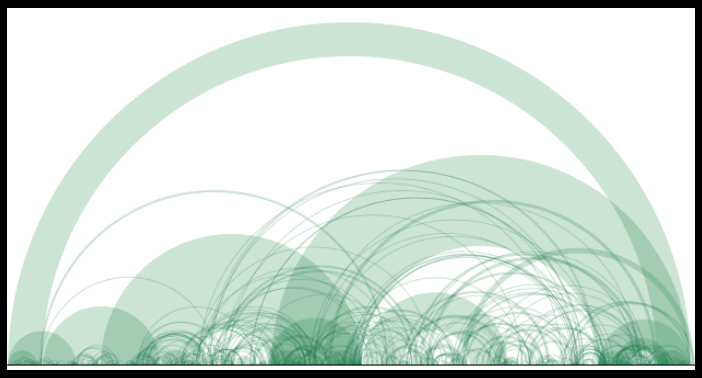
Categorical data (discrete, non-continuous) can be hard to display visually.

Parallel sets allow multiple category variables to be displayed in relation to each other.

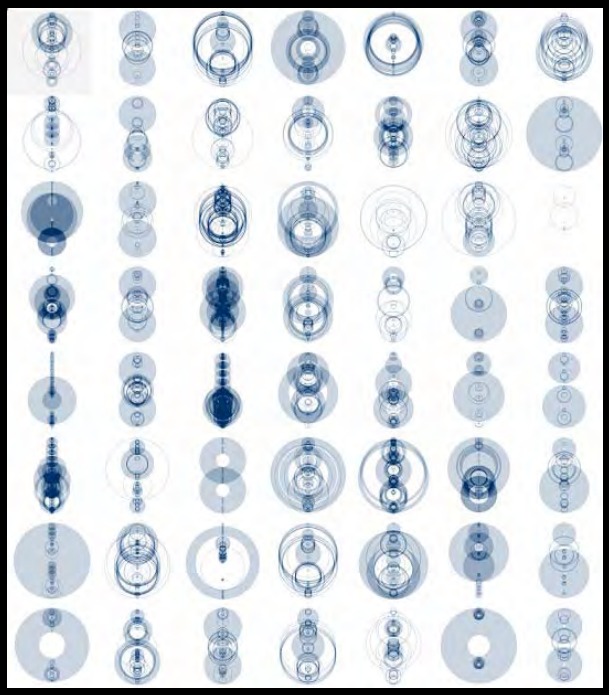


Patterns of Structure





Patterns of Structure



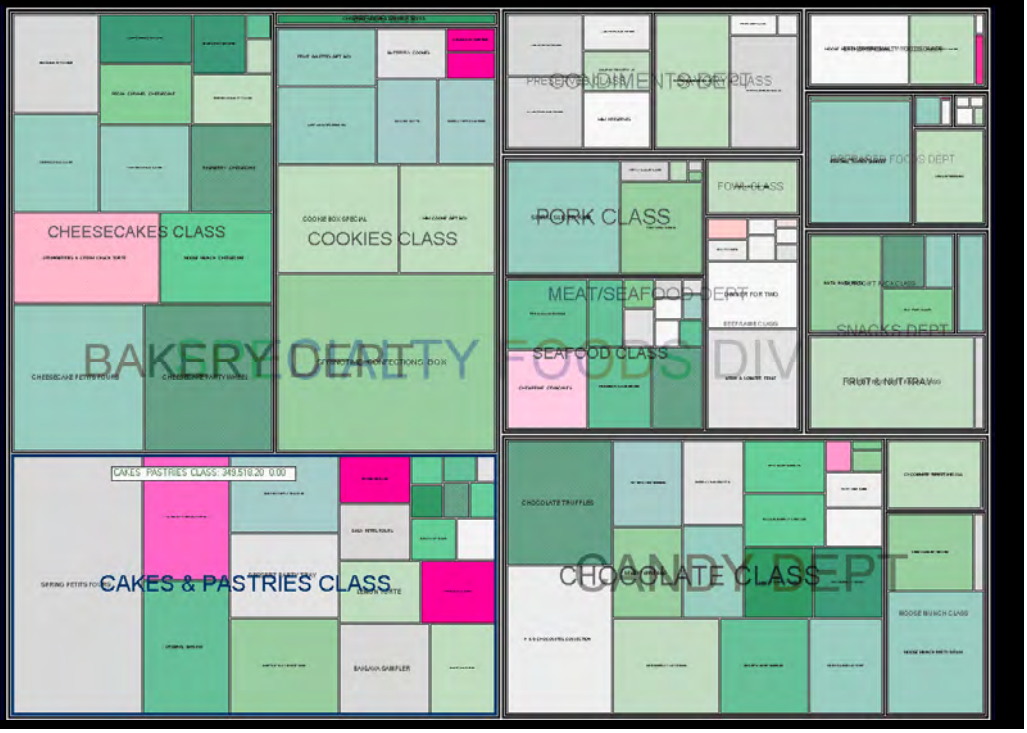
Patterns of Containment: Treemaps

Treemaps visualize hierarchical structures, more useful for showing containment than connection.

- Good: shows outliers pretty well, and more than one variable.
- Bad: hard to see the structure compared to other projections

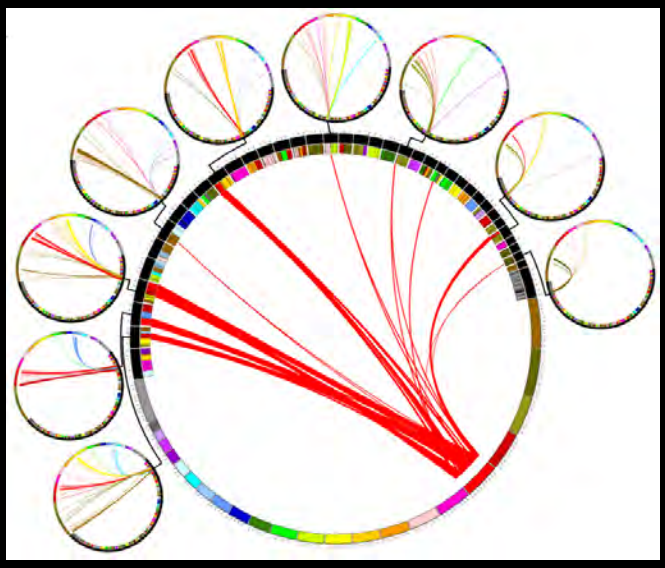
Treemap Example: Overview

Treemap Example: Drilling Down



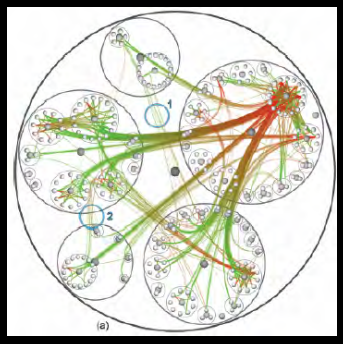
Treemap Example: Drilling Down

Patterns of Connection: Arc Diagram Example



Three example uses:

- Bioinformatics
- Law enforcement
- Retail



Some Notes About Visualization

It's not "chart junk" – although it is easy to make bad or inappropriate visualizations.

- There is lots of science involved (math, psychology, perception) in doing it right, along with a dose of intuition.
- Most people aren't aware of the science.

The field is still fairly new, with most information and tools being generated by academic research.

The proper technique can make some tasks easy.

The proper technique can communicate the story in data in a compelling way.

Visualization is very useful when combined with other techniques and technologies.

GIS and Spatial Data

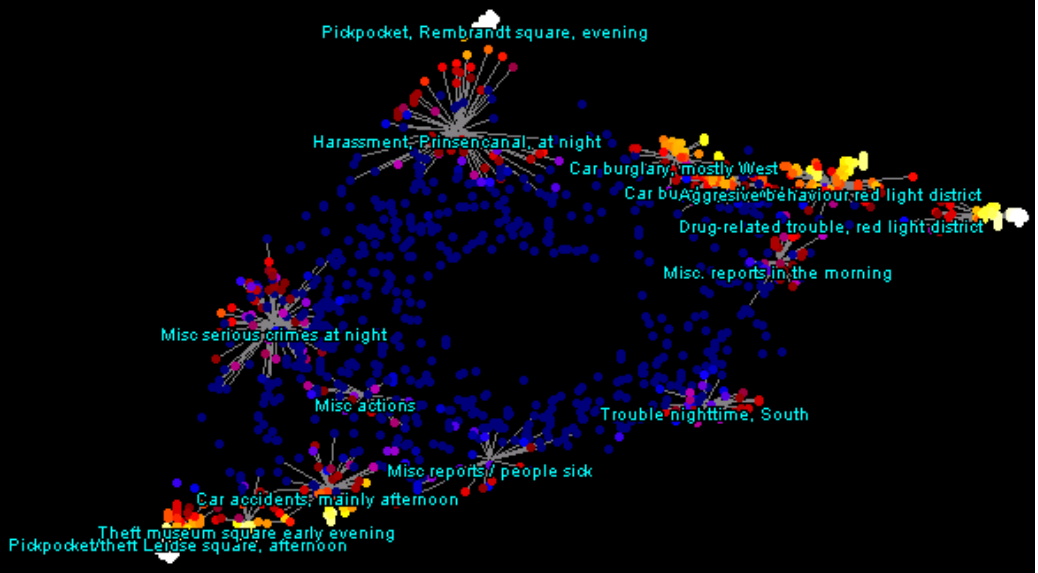
GIS displays are more complicated than slapping data points on Google maps.

- What about lots of data points (like 50K)?
- Aggregates and zoom levels?
- Spatial feature queries?

You'll need a GIS-ready database, GIS server, map APIs, base layer (map tiles), and geocoding.

Image source: kagii.com/

Combining Techniques: Clustering and GIS



Combining Techniques: Clustering and GIS

Deployment clusters (location, day of the week, time)

Unstructured is Not Really Unstructured

"Data is all surface and no insides. It's all handles and no suitcase. It's a folder whose content is just another label."

David Weinberger

Unstructured data isn't really unstructured: language has structure. Text can contain traditional structured data elements. The problem is that the content is *unmodeled.*

It's also a problem of what metadata is when applied to text

Basics: Just Sourcing Text Can Be Hard

Heterogeneous or homogenous sources?

Documents or databases?

- clobs, blobs ,or varchars

Text or bitmapped?

Exhaustive or focused extraction?

Then there's getting the text itself out of the structure it's stored in so the tools can process it.

Extraction of Semantic Content

Assuming the "easy" part of getting the text out is done…

You have to extract the elements you are interested in (assuming you know these in advance, not a good assumption)

Product

Concept

Complaint	Response	Comment
PEACH & BERRY PIE arrived bubbling and swelling; customer ate it anyway and got food poisoning.	Customer asked for replacement pie.	OK, that's one more flag for the Customer dimension (Clueless and Hungry, Y/N).
VICTORIAN BIRDBATH: Is it common for bowl to melt in hot weather?	Refunded, don't know what temp is too hot	I guess Bush was wrong about global warming

Concept

Product

Event

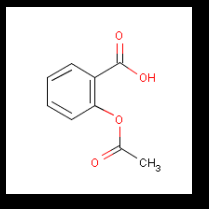
Entity Recognition Challenges

Forms and reference:

- Search for "phone number" will not return "541-555-1212", nor will it or pattern match return "011-541-555-1212"

Common Names

- Aspirin, Acetylsalicyclic acid, Excedrin



Structural formulas

- Commonly used to communicate between chemists

Systematic nomenclatures:

- Mass formula: C9H8O4
- SMILES: OC(=O)C1=C(C=CC=C1)OC(=O)C
- InChI: 1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
- IUPAC: pyrido[1",2":1',2']imidazo[4',5':5,6]pyrazino[2,3-b]phenazine

There Are Different Techniques

Four basic categories:

1. Manual
2. Statistical ("bag of words")
3. Machine learning
4. Natural language processing

One of the news reports got it wrong:

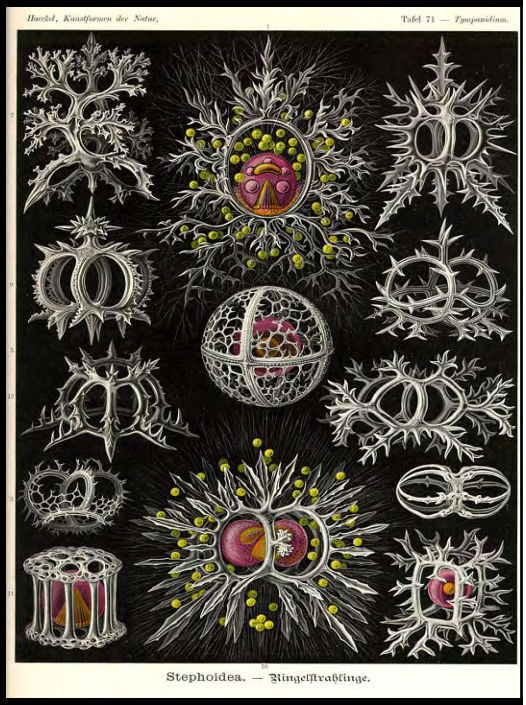
i. "Company A was acquired by Company B"
ii. "Company B acquired Company A"
iii. "Company A's acquisition of Company B is complete"

Statistical and ML techniques likely will miss this.

There are a Gazillion Scenarios for Use

Clinical trials, warranty analysis, hospital admissions, insurance fraud, call center effectiveness, email support, legal discovery, compliance, governance, competitive intelligence, brand monitoring, brand management, voice of the customer, market intelligence, criminal investigation, ECM enrichment, mining model improvement, data quality, patent research, marketing campaign management, customer assurance, customer retention, product management, up-selling, cross-selling, buzz tracking, contract analysis, treatment efficacy, bid tracking, claims management, mortgage assessment, sentiment analysis

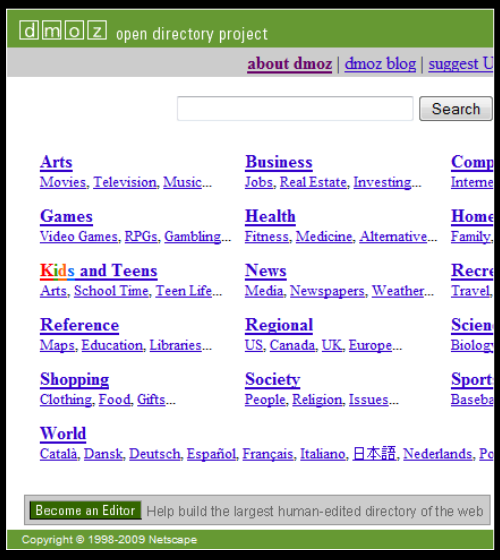
They are as innumerable as for BI so I won't enumerate them all.



Information theory & information retrieval

Go back and read Vannevar Bush, circa 1945…

Which Do You Use, Directory or Search?



Search indexes are based on information retrieval theory.
The relevance ranking of a SERP is largely due to a social algorithm.

The Secret Ingredient in Search

Information Retrieval

An umbrella term for finding specific or relevant information, usually text, in a larger collection.

IR queries are often seeking information about something, not the something itself.

Data vs. information retrieval:

Characteristics	Data Retrieval	Information Retrieval
Matching model	Exact match	Partial or best match
Items retrieved	Matching only	Relevant
General model	Deterministic	Probabilistic
Classification of items	Monothetic	Polythetic
Error tolerance	Sensitive	Insensitive
Inference model	Deduction	Induction
Query specification	Complete	Incomplete
Query language	Artificial	Natural

The Basics: You need Four Things

You need:

- a machine representation for the information in your collection (documents, words, table columns, values)
- a way to store and retrieve the structures
- mechanisms to search and retrieve the structures
- a way to measure the effectiveness of retrieval

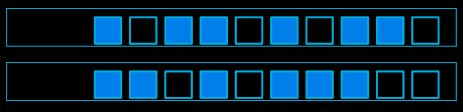
The first two are easy: usually vectors or inverted indexes, stored in databases or file structures

The rest is math and statistics. Sorry.

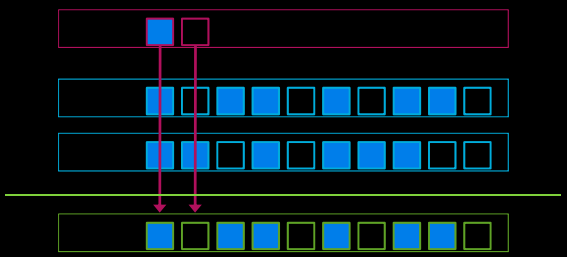
Basic Concepts: Encoding Data

Think of a document as a vector containing 1's and 0's that represent the terms or values in the document:

- A collection of documents is a collection of term vectors



- Make the query a vector too and it's easy to retrieve exact matches to query terms, it's Boolean presence/absence



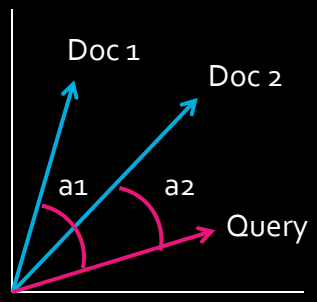
Basic Concepts: Scoring for Retrieval

IR can't use exact matching most of the time because we are seeking relevance.

- For example, terms may have overloaded meaning, e.g. star = celebrity and star = stellar object and star = shape
- Combine "rock" with "star" and the meaning is likely narrowed.

This means measuring query *similarity*

- We can use the vectors: measure the distance (cosine of the angle) between the documents and the query

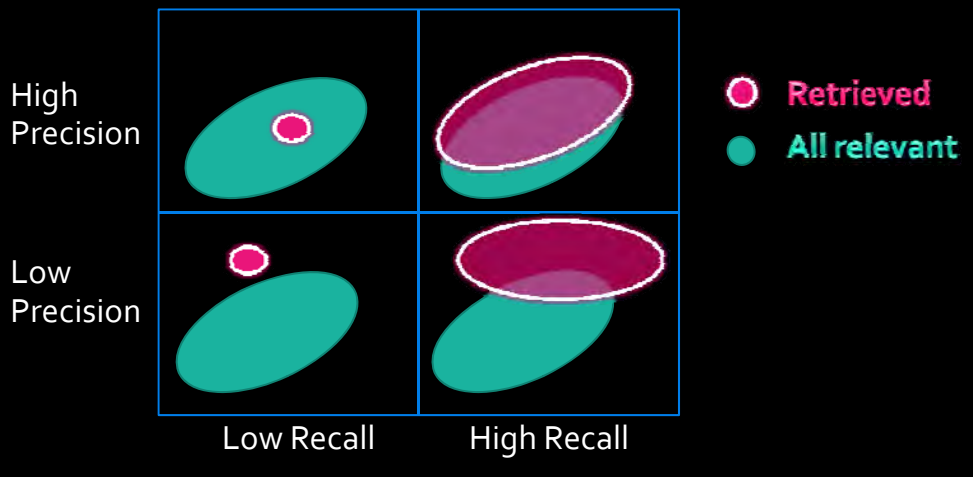


$$sim(a,b) = \sum_{\text{word } i} \frac{v(a,j)}{\sqrt{\sum_{j'} v^2(a,j')}} \cdot \frac{v(b,j)}{\sqrt{\sum_{j'} v^2(b,j')}} = A' \cdot B'$$

Measuring Effectiveness: Two Key Concepts

Precision- The ratio of the number of relevant documents retrieved to the total number of documents. Answering "is it relevant to my query?"

Recall - The ratio of the number of relevant documents retrieved to the total number of relevant documents (both retrieved and not retrieved). Answering "what fraction of the total was returned for my query?"



Where IR Shows Up

This may appear to apply only to searching in or for documents, but we'll see it also applies to things like cross-sells, up-sells and sales predictions.

IR techniques are making their way into BI due to:

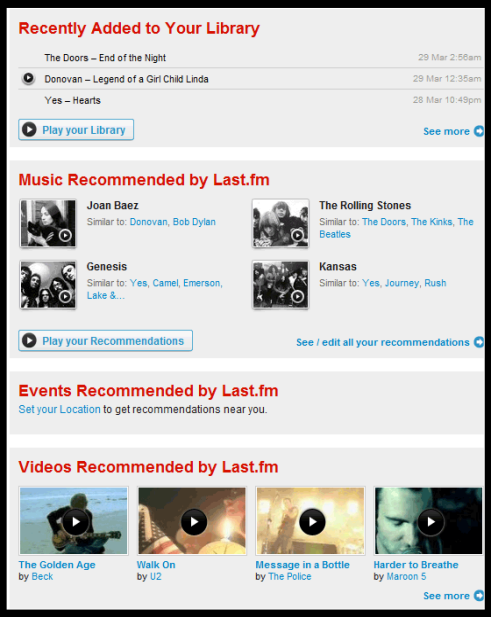
- Text as a data source
- Text as augmented context
- Advanced analytic techniques for predictions
- Recommendations

Recommendations

Recommendations Can Tell You…

What to read

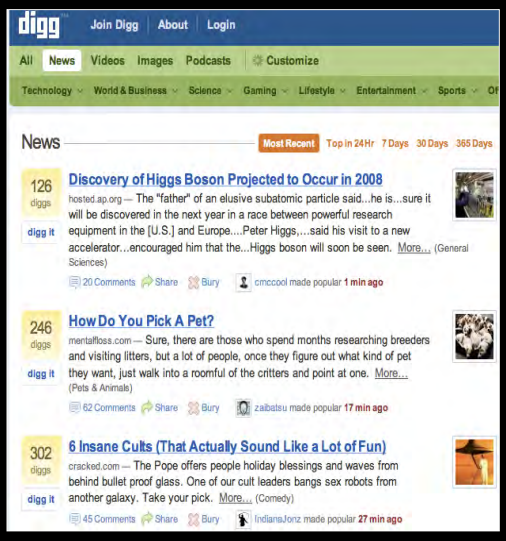
What to listen to



Recommendations Can Tell You…

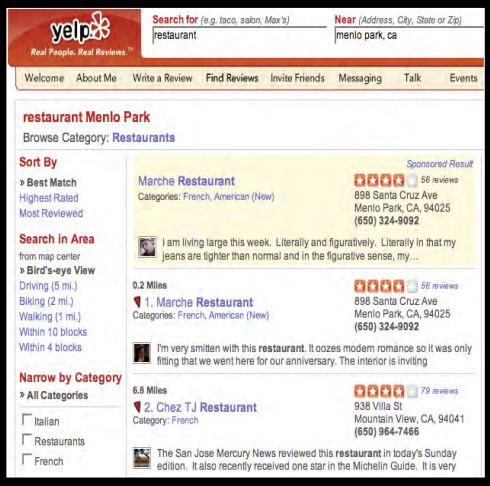
What to watch

What's newsworthy



Recommendations Can Tell You…

Where to eat



What to eat

Recommendations Aren't Perfect

Recommendations can backfire:

- All paths lead to the Beatles
- The Harry Potter problem
- Relevance

On page descriptions you should show the rationale for "why this product?" or risk frustrating buyers

- "people who bought"
- "highly rated"
- "books by this author"

Always log user activity: not measuring personalization leads to a depersonalization strategy.

Recommendations as a BI Problem

Three primary approaches:

- **Personalized recommendation** - recommend based on the individual's preferences
- **Social recommendation** - recommend based on the preferences of similar users
- **Attribute recommendation** - recommend based on item attributes and either item similarities or a utility function

They have big data requirements

Profile data, preferences, attention data, transaction history, behavior data for *everyone*

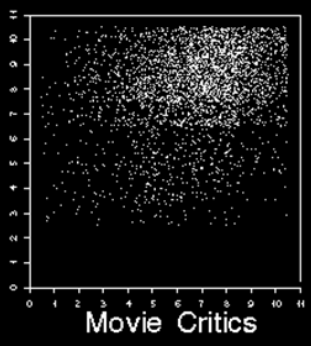
What Are Recommendations?

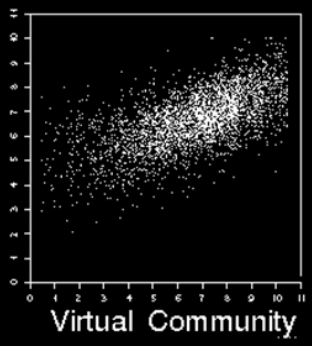
They're predictions about your preferences or behavior.

We can apply different techniques to predict these.

The most common (online) is collaborative filtering; personal tastes are usually correlated:

- If Alice and Bob both like beer and Alice likes handkäse then Bob is more likely to like handkäse.
- This is even more likely if Bob knows Alice.

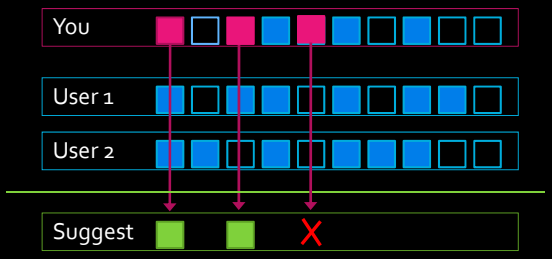




Collaborative Filtering

In general (and that's a lot of generalization) your preferences can be used to assume similarity to others, hence recommend something to you based on what they prefer that we don't know you prefer.

Ratings, or purchases, or preferences

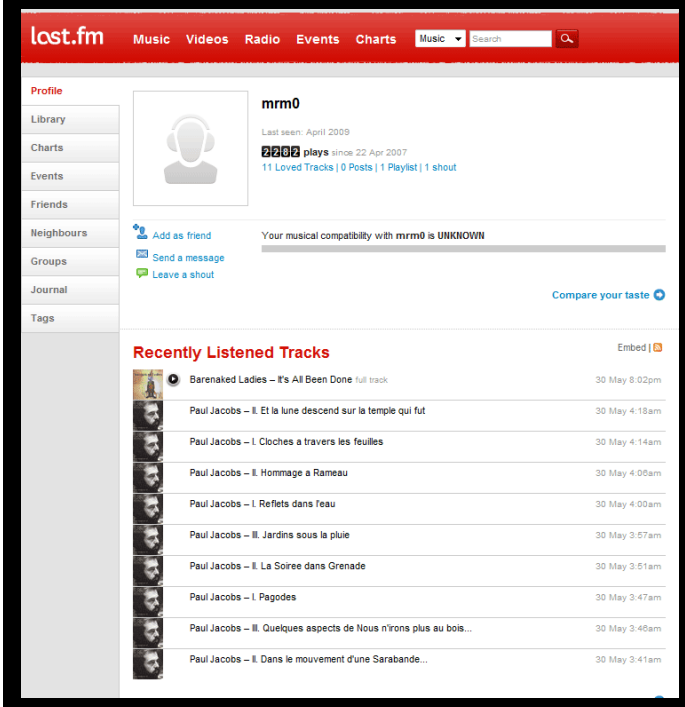


Example 1: Music Recommendations

Based on music similarity and what you specify about the songs.

Not social, tends to deliver more of what you already know and like.

Example 2: Music Recommendations



Based on user preference similarity and implicit preferences (based on listening habits).

Social, tends to deliver more unknown finds.

Precision & recall fail to measure this.

Up-selling and Cross-selling

Up-sells as Recommendations

An up-sell is a prediction that the recommended item will be purchased *in place of* the item already selected.

Upsells: Rules and Rules Engines

Most up-sells implemented as hard-coded or data-driven rules.

- Rules models are data driven – you apply rules on known data to produce sets of results.
- Defining rules is almost the same thing as programming.

Problems:

- Many rules, or rules with overlapping conditions makes it hard for a person to predict the impact of rule changes.
- Hard to validate and identify conflicts and dependencies manually so it's best to use a rules engine with a conflict resolver which will invariably be based on Rete.

There are some good open source rules engines available.

Upsells: Constraint Programming

Mostly used in operations research for optimization problems, it's a method of declarative programming:

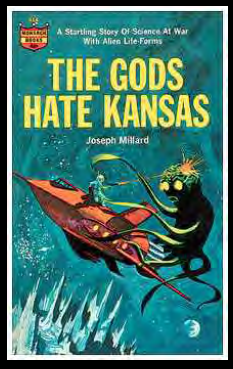
- Unlike rules, it's goal driven: it works backward from a goal to find unknown data and produces a set that meets the constraints, e.g. here's the goal, go expand the set of products that apply
- Define the problem, define the boundaries, determine the goal, and it finds the solution. A Zen way of programming.
- Generally produce optimal results for the desired goal.

Problems:

- CP tends to be challenging for many people because there's formal technique involved in defining the models.
- Most tools have poor UIs and rely on languages.
- Performance can sometimes be slow and unpredictable.

Cross-sells as Recommendations

A cross-sell is a prediction that the recommended item will be purchased *with* the item already selected.



+

=

Cross Sells Are a Classic Data Mining Example

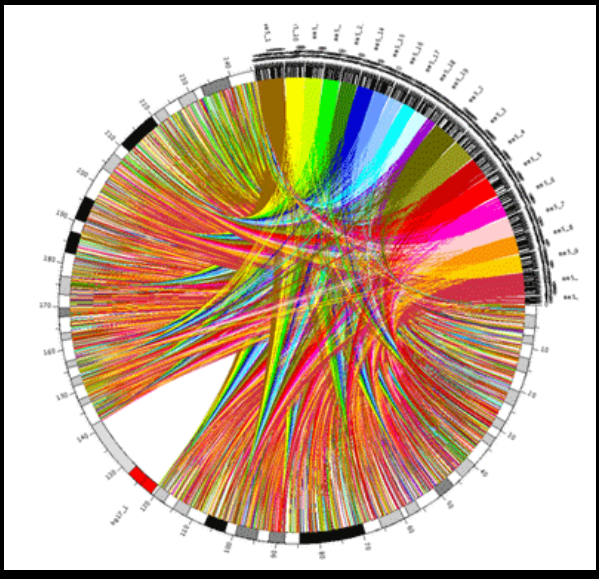
Two ways to use data mining for up-sells and cross-sells.

- *Item-oriented*: find the items most often chosen when up-sells are offered, or which items contribute to secondary purchases.
- *User-oriented*: look at the people who are most likely to respond to an up- or cross-sell. Different set of techniques to accomplish the same task. You can also do a combination of both.

BI can rank items purchased together, but it's slow and not easy to do exhaustively. Can't determine probabilities or likely buyers.

Association rules work better for cross-sells. Plenty of examples to read from so I'll talk about an infrequently used visualization technique instead.

Visualizing Cross-sells

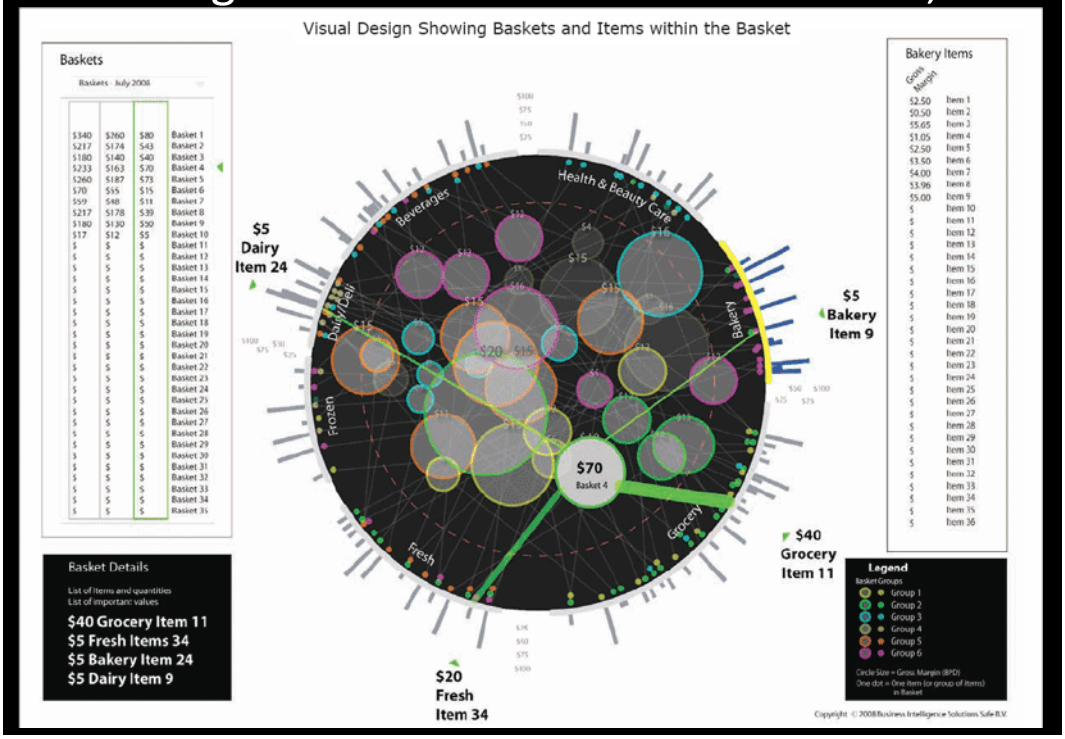


It's easy to see related purchases visually over the entire set of data, even tens of thousands of purchases.

Interactive visualization allows you to drill into particular items or filter the data to those meeting key criteria.

Plus this is more fun.

Visualizing Cross-sells: lots of data interactive, fun



Getting Started: Choosing Techniques

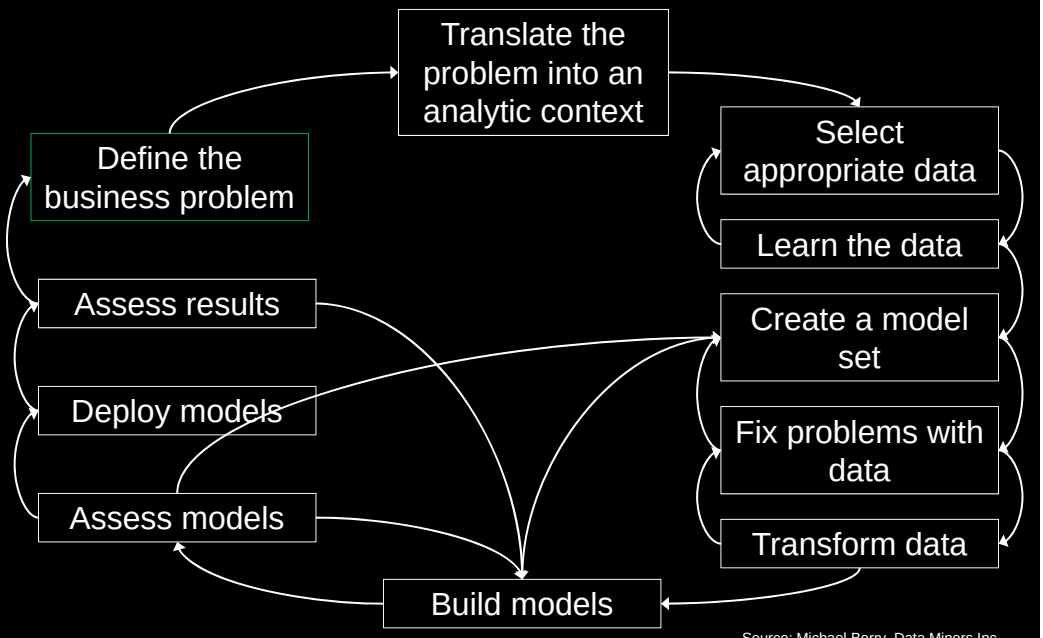
What do you know in advance?

- If you have data that defines a good vs. bad outcome you can use supervised methods.

What type of problem are you solving?

- Predictive techniques are suitable for building a model to determine a specific value.
- Descriptive techniques are suitable for building a model that defines categories or is multi-valued.
- Visualization is useful for exploration, analysis and communicating insights to others.
- Processing text or making recommendations have their own fields and techniques to be used in combination with these.

The Analytic Process is Highly Iterative



Source: Michael Berry, Data Miners Inc.

Third Nature

Where to go from here?

"How to Measure Anything" is a useful text

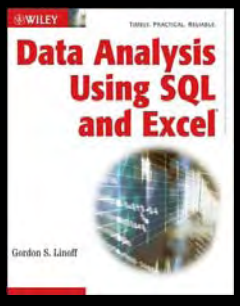
Move on to 'Data Analysis Using SQL and Excel'

Statistics for business:

- http://home.ubalt.edu/ntsbarsh/Business-stat/opre504.htm

Data Mining:

- www.rapid-i.com (RapidMiner)
- http://www.thearling.com
- http://www.autonlab.org/tutorials/



For free data mining e-books, search www.scribd.com

Free Tools So You Can Get Started Inexpensively

WEKA http://www.cs.waikato.ac.nz/ml/weka/
IND decision tree software http://opensource.arc.nasa.gov/software/ind/
Clustering http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/
Parallel Sets http://eagereyes.org/parallel-sets#download
RapidMinerhttp://rapid-i.com/content/blogcategory/38/69/
Knimehttp://www.knime.org/
Orange http://www.ailab.si/Orange/
R statistics software http://www.r-project.org/
ARC statistics software http://www.stat.umn.edu/arc/software.html
Octave numerical and matrix computation
http://www.gnu.org/software/octave/
Processing http://www.processing.org/
Circoshttp://mkweb.bcgsc.ca/circos/
Treemaphttp://www.cs.umd.edu/hcil/treemap/
Microsoft Data Mining Add-Ins for Excel *(free if you have Office and SS)*
Oracle, IBM, SqlServer built-in data mining functions

More Resources to Get You Started

Books:

- Data Mining Techniques: For Marketing, Sales and Customer Support, Michael J. Barry and Gordon Linoff
- Data Preparation for Data Mining, Dorian Pyle
- Data Mining Algorithms, Elbe Frank, Ian Witten, Jim Gray
- An Introduction to Information Retrieval, Christopher D. Manning, PrabhakarRaghavan, HinrichSchütze
- Information Retrieval, C. J. van Rijsbergen
- The Visual Display of Quantitative Information, Edward R. Tufte

Journals, Newsletters, Web Sites:

- SIG KDD Explorations, Newsletter of the ACM SIG on Knowledge Discovery and Data Mining
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- KDNuggets data mining resources: www.kdnuggets.com
- Flowing Data, visualization resources: http://flowingdata.com/
- Infoaesthetics, visual design resources: http://infosthetics.com/
- Visual Complexity, visualization resources: www.visualcomplexity.com/vc/index.cfm
- Recommendation systems resources: http://www.deitel.com/ResourceCenters/Web20/RecommenderSystems/tabid/1229/Default.aspx
- The Impoverished Social Scientist's Guide to Free Statistical Software and Resources: http://maltman.hmdc.harvard.edu/socsci.shtml

KTHXBAI!

Creative Commons and Other Image Attributions

Thanks to the people who supplied the creative commons licensed images used in this presentation:

Word cloud - generated by http://www.wordle.net/ from the text of this presentation

My new fighting technique is unstoppable - © 2006 http://www.mnftiu.cc/2006/11/08/081/

Magician_poster2.jpg - http://www.flickr.com/photos/trialsanderrors/2831258177/

Crime clustering images - Sentient Systems, Netherlands

Tibetan_sand_painting.jpg - http://www.flickr.com/photos/wonderlane/3242519210/

Maps of the Kerry-Bush election - New York Times information graphics department

"What music looks like" - Martin Wattenberg

Diagrams drawn by Circos http://mkweb.bcgsc.ca/circos/

Map zoom levels - http://www.kagii.com/?p=112

Book of hours manuscript2.jpg - http://flickr.com/photos/jeffrey/89461374/

Klein_bottle_blue.jpg - http://flickr.com/photos/candiedwomanire/60224567/

Open_air_market_bologna.jpg - http://flickr.com/photos/pattchi/181259150/

Four cupcake frogs.jpg - http://www.flickr.com/photos/abielskas/114946978/

Bound documents - http://flickr.com/photos/peterkaminski/1688635/

Old man union square selling.jpg - http://flickr.com/photos/ktb/4683842/

Spiny frog and Fiji alligator – Takeshi Yamada

Diagram drawn by Circos http://mkweb.bcgsc.ca/circos/

Retail visualization courtesy of Bis2 http://www.bis2.net

snake_eyes_red.jpg - http://www.flickr.com/photos/t3rmin4t0r/3823260539/

About the Presenter

Mark Madsen is president of Third Nature, a technology research and consulting firm focused on business intelligence, analytics and data management. Mark is an award-winning author, architect and CTO whose work has been featured in numerous industry publications. Over the past ten years Mark received awards for his work from the American Productivity & Quality Center, TDWI, and the Smithsonian Institute. He is an international speaker, a contributing editor at Intelligent Enterprise, and manages the open source channel at the Business Intelligence Network. For more information or to contact Mark, visit http://ThirdNature.net.