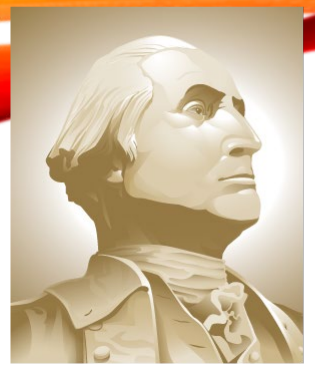


THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

INTRODUCTION TO BIG DATA AND ANALYTICS
CSCI 6444

INTRODUCTION TO DATA ANALYTICS

Prof. Roozbeh Haghnazar

Slides Credit:

Prof. Stephen H. Kaisler & Roozbeh Haghnazar

What is Analytics?

- A word that is widely used, largely misunderstood, and overloaded with many meanings.
- ***Analytics* is the scientific process of transforming data into insight for making better decisions** (INFORMS)
- *Analytics* are a suite of tools for processing data to achieve actionable intelligence to support decision-making processes.
- Analytics requires:
  - Visualization
  - Interactivity
  - Utility
- Often misused in the Business Intelligence World and associated with *data mining*

Why Analytics?

- There is little value in just storing and having data.
- *Value* is created when the data are analyzed, (hopefully) resulting in better decision-making and problem-solving

- **Why Big Data Analytics?**
  - Big Data may/can improve decision-making because there is more data to analyze.
  - Diverse sources of data may conflict, but can also improve the analysis with different perspectives and techniques.
- **But!**
  - Is there glamour in analytics?
  - Of course, there's always the thrill when you discover something you didn't know before.
  - (And, of course, maybe the money!)

THE ANALYTICS

1. Formulate a business or domain question/problem
2. Identify, gather, cleanse and prepare the data available to answer the question
3. Analyze the data
   - Descriptive Analytics:
     - ✓ Familiarize yourself with the data (descriptives, correlations, factor analysis, cluster analysis, etc.)
     - ✓ Generate hypotheses (data mining, patterns, trends, etc.)
     - ✓ What happened?
   - Predictive Analytics
     - ✓ Develop and formulate hypotheses
     - ✓ Identify the most suitable analytic methods
     - ✓ Develop analytic models: multivariate regression, logistic regression, forecasting, non-linear models, classification trees, etc.
     - ✓ Run the models and generate predictions
     - ✓ What is likely to happen
   - Prescriptive Analytics:
     - ✓ Develop decision and optimization models
     - ✓ Use machine learning to program decisions
     - ✓ What do we need to do
4. Write up conclusions and recommendations

Descriptive Analysis (Understanding the past and present data trends)

This type of analysis focuses on summarizing the dataset to generate insights about what has happened.

- **Product Performance Analysis:**
  - Total number of products by category.
  - Average rating per product and category.
  - Top 10 best-selling products and their categories.
- **Customer Review Analysis:**
  - Distribution of review ratings (1-star to 5-star).
  - Number of reviews per product and category.
  - Word cloud of frequent words in positive vs negative reviews (e.g., "bad," "excellent").
- **Time-Based Trends:**
  - Monthly or yearly trend in the number of reviews.
  - Seasonal patterns of product sales (e.g., spikes during holidays).
- **Customer Sentiment Analysis:**
  - Breakdown of positive, neutral, and negative reviews.
  - Distribution of sentiment across product categories.

Predictive Analysis (Forecasting future behavior and trends)

This analysis aims to predict future outcomes based on historical data patterns.

- **Sales Forecasting:**
  - Use time series models (ARIMA, LSTM) to predict future sales trends for a product or category.
- **Rating Prediction:**
  - Build a regression model (e.g., XGBoost) or collaborative filtering (recommendation system) to predict the rating a user will give to a product based on historical data.
- **Review Sentiment Prediction:**
  - Train a classification model (e.g., Random Forest or BERT) on review text to predict the sentiment of new customer reviews.
- **Product Recommendation System:**
  - **Content-based filtering:** Recommend similar products based on product descriptions and categories.
  - **Collaborative filtering:** Recommend products to users based on other customers with similar behavior.
- **Churn Prediction:**
  - Predict whether a customer is likely to stop purchasing based on purchase frequency, review sentiment, and product interaction history.

Prescriptive Analysis (Recommending actions based on predictions)

Prescriptive analysis focuses on recommending actions to achieve desired outcomes.

- **Inventory Optimization:**
  - Based on sales forecasts, recommend inventory levels to reduce stockouts and excess inventory.
- **Product Pricing Recommendations:**
  - Develop price elasticity models to understand how price changes will affect sales and recommend optimal pricing strategies.
- **Marketing Campaign Optimization:**
  - Suggest personalized promotions or discounts to customers based on their past purchase patterns and predicted product preferences.
- **Sentiment-Driven Action Plans:**
  - Use review sentiment analysis to detect early signs of product quality issues and recommend improvements.
- **Customer Retention Strategies:**
  - For customers predicted to churn, recommend personalized incentives like discounts, free shipping, or targeted marketing campaigns.

Data Science

- "A set of fundamental principles that guide the extraction of knowledge from data." (Provost & Fawcett, Data Science for Business)

- The ultimate goal is to make sense of data
  - "It is a *capital mistake* to *theorize* before one has *data*."
  - Sir Arthur Conan Doyle (Who is he?)

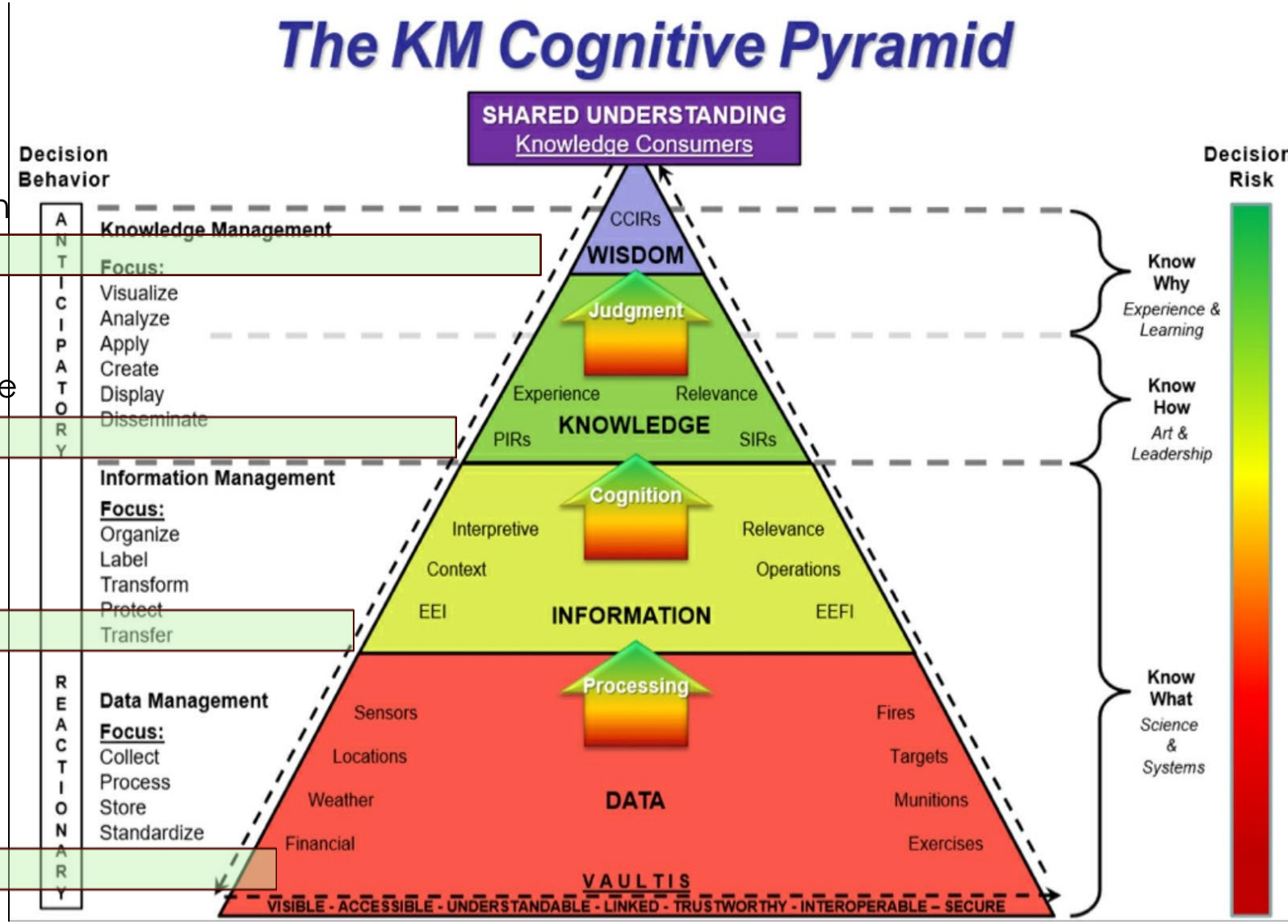
- A *data scientist* has deep knowledge of analytics, all aspects of data, the discipline in which analysis is conducted (marketing, healthcare, social media, etc.), and the underlying core disciplines (e.g., statistics, mathematics, database, etc.)

Wisdom focuses on long-term strategic actions by applying prescriptive insights to optimize business decisions.This layer helps in decision-making based on forecasts and predictions.

Knowledge emerges by understanding patterns in the information to derive actionable insights.This is where Predictive Analysis is applied to forecast trends and predict outcomes.

Information is derived by structuring, summarizing, and processing data into meaningful insights.Here, Descriptive Analysis provides insights into "what has happened" using KPIs and trends.

Raw data collected from Amazon's products, categories, and customer reviews forms the foundation.



DIKW Mapping to Descriptive, Predictive, and Prescriptive Analysis:

- **Data Layer**
  - **What it contains:**
    - Product details: Product IDs, categories, prices, descriptions.
    - Customer reviews: Ratings, review text, timestamps, sentiment tags.
    - Sales data: Units sold, revenue, discounts applied.
    - Customer behavior: Purchase history, browsing behavior.

DIKW Mapping to Descriptive, Predictive, and Prescriptive Analysis:

- **Information Layer**
  - **Examples of information:**
    - **Summary statistics:** Average product rating, total reviews per category.
    - **Visual insights:** Trends in sales over time, seasonal spikes.
    - **Review sentiment breakdown:** Percentage of positive vs. negative reviews across products.
  - **Descriptive Analysis Mapping:**Answer questions like:
    - What are the top 10 highest-rated products?
    - How do reviews vary across product categories?
    - When does product demand increase?

DIKW Mapping to Descriptive, Predictive, and Prescriptive Analysis:

- **Knowledge Layer**
  - **Examples of knowledge:**
    - **Forecasting:** Predicting future sales for each category.
    - **Recommendation models:** Understanding customer behavior to recommend products.
    - **Churn prediction:** Identifying customers likely to stop purchasing.
  - **Predictive Analysis Mapping:**Answer questions like:
    - How many units of a product will be sold next month?
    - Which products are customers likely to buy based on their behavior?
    - What is the predicted sentiment for a new review?

DIKW Mapping to Descriptive, Predictive, and Prescriptive Analysis:

- **Wisdom Layer**
  - **Examples of wisdom:**
    - **Pricing optimization:** Adjusting prices dynamically to maximize revenue.
    - **Inventory management:** Recommending inventory levels based on forecasted demand.
    - **Marketing campaign suggestions:** Offering personalized incentives to prevent churn.
    - **Product improvement:** Identifying areas for product enhancement using review insights.
  - **Prescriptive Analysis Mapping:** Answer questions like:
    - What price should we set to maximize sales and profit?
    - How much stock should be allocated for each product in the upcoming season?
    - What marketing actions will improve customer retention?

DATA SCIENTIST (AND LACK OF)

"The sexy job in the next ten years will be statisticians… The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill."

- Ten years on, it is still one of the jobs that is in high demand.
- But, remember! Over time software will do more and more, so some types of analytics jobs will disappear – but new types of jobs will appear!

Hal Varian, Mckinsey Quarterly, January 2009

http://www.mckinseyquarterly.com/Hal_Varian_on_how the_Web_challenges_managers_2286

Note: This is still true!!

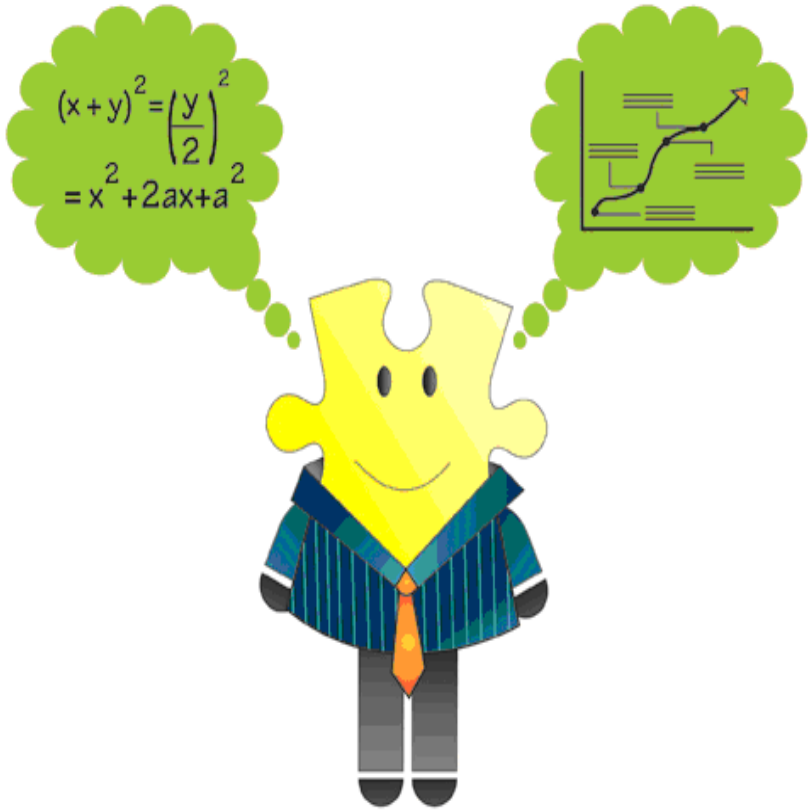
THE ANALYTIC SCIENTIST

"The critical job in the next ***n!*** years will be the ***analytic scientist*** … the individual with the ability to understand a problem domain, to understand and know what data to collect about it, to identify **and develop** analytics to process that data/information, ***to discover its meaning***, and to extract knowledge from it—that's going to be a very critical skill."
- Kaisler, Armour, Espinosa, Money (2014) Amended

Analytic scientists require advanced training in specific domains, data science tools, multiple analytics, and visualization to perform predictive and prescriptive analytics.

Note: This is increasingly important for Complex Problems

Role of the Analytic Scientist



- Today's complex problems require more than a single analytic.
- The analytic scientist must also be a software and data architect.
- ***Analysis* is a process of continuous *discovery* – not only of knowledge but of new analytics.**
- They may hold Ph.D.'s, but pragmatic experience in a domain will be equally important.

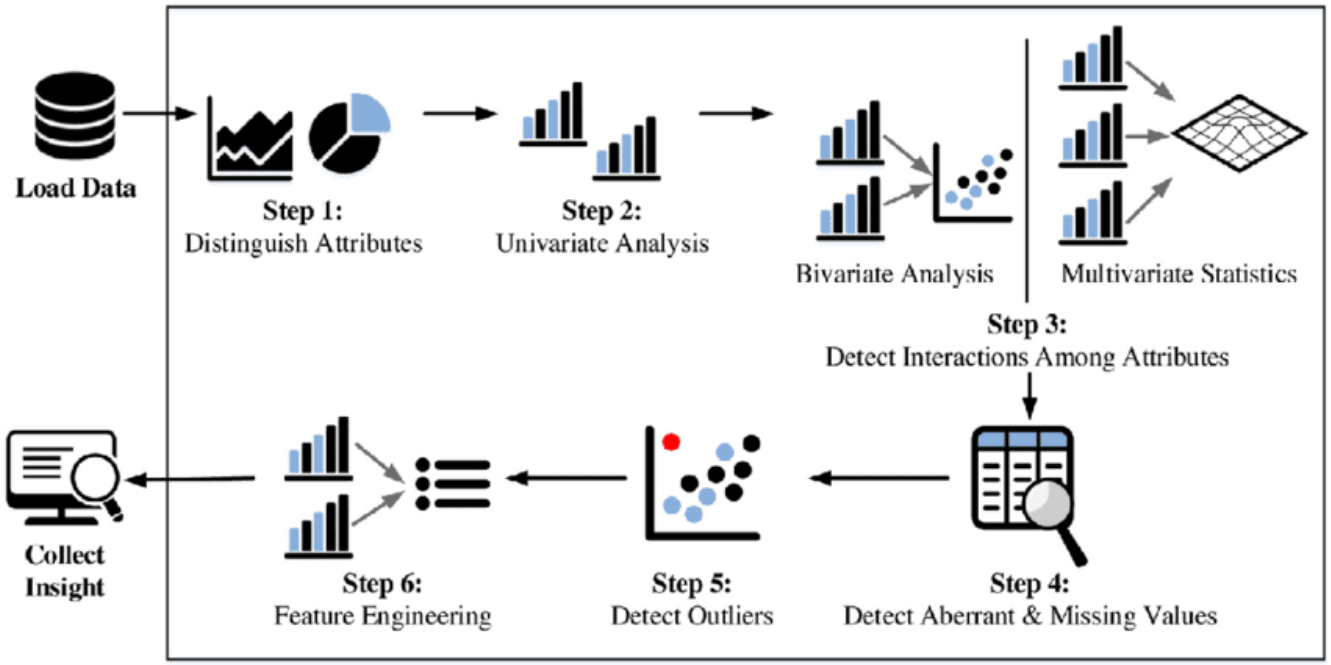
EDA: What and Why?

- An EDA (Exploratory Data Analysis) checklist is a structured approach to the initial phase of data analysis, used to thoroughly understand a dataset before formal modeling. It typically involves steps like checking data structure and types, handling missing values, summarizing variables with statistics and visualizations, and identifying outliers and relationships between variables. Following a checklist ensures data is clean and well-understood, which guides the selection of appropriate subsequent analysis techniques and prevents errors.

- Start from the **business question** to reduce paths and focus the analysis; EDA is the step that turns raw data into usable **information** and quickly checks feasibility.
- **Read/peek early:** confirm schema and types (str()/info()), look at **head/tail/sample**, and ensure expected time coverage/granularity before deeper work.
- Produce a fast **go/no-go** with an assumptions/risks list to carry forward (e.g., time range sufficient? units consistent?).
- Deliverables:
    - (1) One-page EDA summary,
    - (2) short checklist outcome,
    - (3) flagged issues requiring follow-up/extra data.

EDA Checklist (Sanity → Structure → Signals)

- **Assumptions & scope:** write assumptions; define unit of analysis & keys; use "reverse-filter" tests to verify filters (e.g., invert a WHERE clause and expect empty).
- **Packaging/structure & types:** check file bounds and schema; run str()/info(); confirm categorical levels & date bounds.
- **Counts ("n"s):** rows, unique IDs, class balance; duplicates/gaps; basic join cardinalities if multiple tables.
- **Missingness & validity:** missing map / isnull().sum(), out-of-range values, unit consistency.
- **Quick visuals:** univariate (histograms/boxplots) for distributions & outliers; bivariate (scatter, **correlation heatmap**—interpret cautiously); temporal cuts for seasonality/gaps.
- **Cross-checks & skepticism: validate with an external source**; **try the easy solution first**, then **challenge your result** (stress-tests).
- **If modeling next:** note any scaling/normalizing needed (do **after** you understand distributions).



From EDA to Dashboard Brief

- Dashboard Creation Process

The Emerging Factor in Analytics

- Machine Learning:
  - The statistical (quantitative) kind is getting big play!
  - And, has matured in certain areas!
  - DARPA committed over $2 Billion to try and achieve major breakthroughs in ML.
- Two key ideas:
  - How can we build computer systems that automatically improve with experience, and
  - What are the fundamental laws that govern all learning processes?
  - If we discern them, how would we apply them to machine learning?
- Many learning algorithms are not consistent.
- Many performance bounds are not tight.
- The dimensions are high, feature selection is important.
- Most data is unlabeled.
- Executing Multi Layered ML Pipelines are costly

Big Data Issues Affecting Analytics

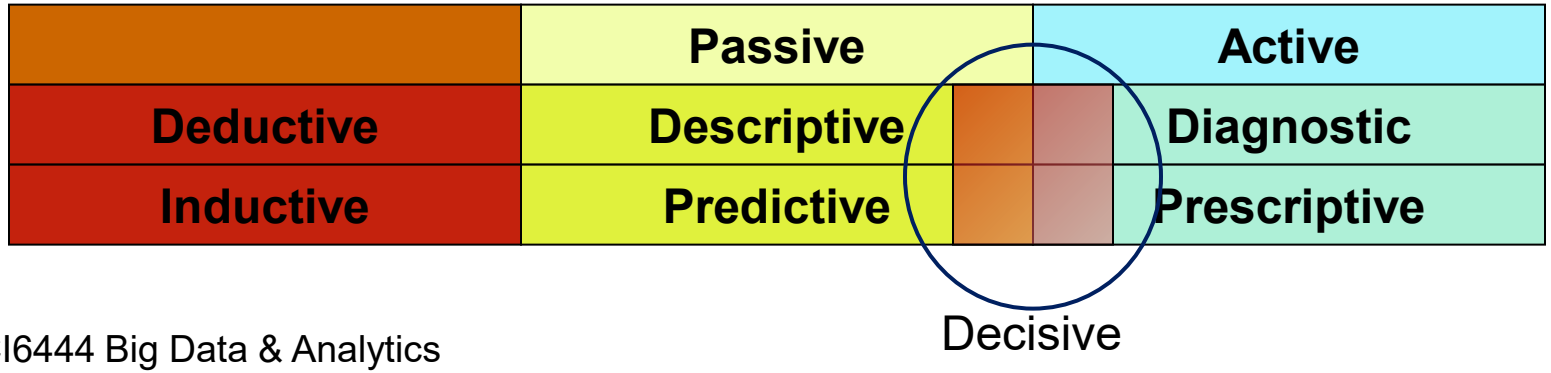
- Volume:
  - How much data is really relevant to the problem solution? Cost of processing?
  - *So, can you really afford to store and process all that data?*
- Velocity:
  - Much data coming in at high speed
  - Need for streaming versus block approach to data analysis
  - *So, how to analyze data in-flight and combine with data at-rest*
- Variety:
  - A small fraction is structured formats, Relational, XML, etc.
  - A fair amount is semi-structured, as web logs, etc.
  - The rest of the data is unstructured text, photographs, etc.
  - *So, no single data model can currently handle the diversity*
- Veracity: cover term for …
  - Accuracy, Precision, Reliability, Integrity, Provenance, Governance
  - *So, what is it that you don't know you don't know about the data?*
- Value:
  - How much value is created for each unit of data (whatever it is)?
  - *So, what is the contribution of subsets of the data to the problem solution?*

Challenges for Big Data Analytics

- **Axiom: Big Data is not just about the size or speed of your data – but, the complexity and quality.**
- Hypothesis: With more data to analyze, Big Data improves decision-making.
- Conjecture: More data beats better algorithms, or does it?
  - Corollary: Big Data with simple algorithms will outperform sampled data with complex algorithms?
- Problem: Many problems require an analytics capability beyond what is offered by traditional business analytics
- Problem: Given a problem (the range), how do we identify the domains of data and the analytic(s) that will enable us to solve the problem?
- Hypothesis: Can good algorithms, models, heuristics make up for data of poor quality? (Our answer: NO!!)

Types of Analytics

- ***Descriptive***: A set of techniques for reviewing and examining the data set(s) to understand the data and analyze business performance. (What happened?)
- ***Diagnostic***: A set of techniques for determine what has happened and why (Why did this happen?)
- ***Predictive***: A set of techniques that analyze current and historical data to determine (What is most likely to (not) happen?)
- ***Prescriptive***: A set of techniques for computationally developing and analyzing alternatives that can become courses of action – either tactical or strategic – that may discover the unexpected (What do we need to do?)
- ***Decisive***: A set of techniques for visualizing information and recommending courses of action to facilitate human decision-making.



Types of Analytics (Example)

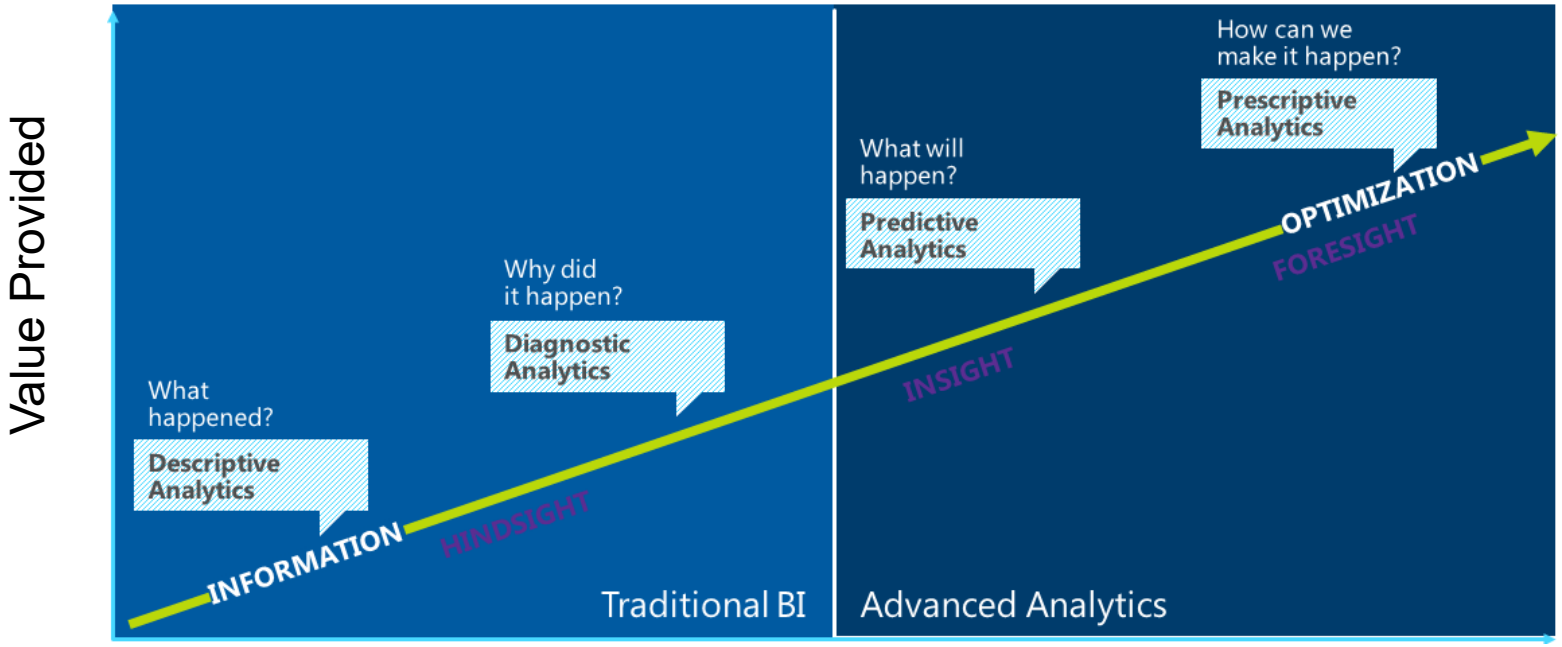
Scenario: weekly sales dropped 12% last month for an e-commerce store

- **Descriptive (What happened?) — passive, deductive/inductive**
  - Tasks: summarize facts.
  - Example: “Weekly sales fell from $1.2M to $1.05M (-12%). Traffic -3%, conversion rate -9%, average order value flat.”
  - Typical artifacts: time-series charts, pivot tables, KPI deltas.
- **Diagnostic (Why did it happen?) — active, mostly deductive**
  - Tasks: look for causes and contributing factors.
  - Example: Segment analysis shows the drop is concentrated in mobile Safari users on iOS 17, primarily on product pages with a new image carousel. Funnel shows add-to-cart -18% only on that segment.
  - Typical tools: cohort cuts, funnel analysis, controlled comparisons, error logs, correlation checks; you’re forming and testing hypotheses.
- **Predictive (What will happen next?) — passive→active, inductive**
  - Tasks: forecast or classify future outcomes.
  - Example: A model predicts that, if unaddressed, mobile conversion will remain ~10% below baseline next 4 weeks, implying a further $320k revenue loss; customers exposed to the carousel are 1.3× more likely to abandon the session.
  - Typical outputs: probability scores, demand forecasts, risk rankings.
- **Prescriptive (What should we do?) — active, decision-oriented**
  - Tasks: choose the best action under constraints.
  - Example: Options evaluated with an optimization/experiment plan: (A) roll back carousel on Safari only; (B) hot-fix image lazy-load; (C) increase mobile promo by 5% to offset. An expected-value analysis says (B) yields the highest net benefit (quick fix, no margin hit).
  - Typical methods: A/B test design, uplift modeling, optimization (LP/MILP), cost–benefit analysis.
- **Decisive (How do we present/operate the decision?) — active, human-in-the-loop**
  - Tasks: communicate options and make the call; monitor impact.
  - Example: A decision dashboard shows: current loss run-rate, predicted loss if no action, and side-by-side EV of A/B/C with uncertainty bands. Product + Growth choose (B) now, (A) as fallback, and schedule an A/B to confirm recovery. Post-decision panel tracks lift vs baseline and automatically alerts if recovery < 80% of expected.

Task

- **Add the analytical types to your final project**

Analytics Road Map



Q: What is ROI for different types of analytics?

Source: Gartner

Descriptive Analytics

- **Process**:
  - Identify the attributes, then assess/evaluate the attributes
  - Estimate the magnitude to correlate the relative contribution of each attribute to the final solution
  - Accumulate more instances of data from the data sources
  - If possible, perform the steps of evaluation, classification and categorization quickly
  - Yield a measure of adaptability within the Observe, Orient, Decide, Act (OODA) loop
- At some threshold, crossover into diagnostic and predictive analytics

Year 2000				
	Audio Division		Video Division	
Line Items	Budget	Actual	Budget	Actual
Cost of Goods Sold	$6,851,006.49	$7,132,961.38	$4,322,514.74	$4,526,954.71
Marketing Expense	$750,179.20	$756,586.17	$455,048.05	$462,815.40
Research and Development Expense	$538,243.39	$538,014.73	$329,890.95	$336,808.13
Selling Expense	$1,632,921.64	$1,579,790.18	$986,887.49	$927,970.90
Taxes	$314,659.05	$319,390.19	$202,636.67	$200,205.01

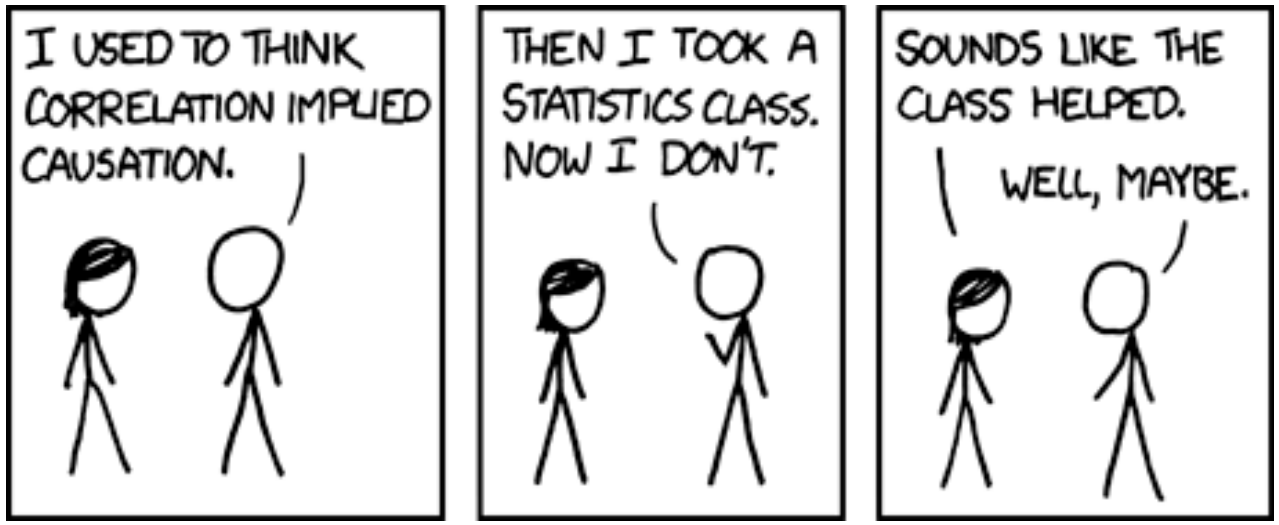
Year 2001				
	Audio Division		Video Division	
Line Items	Budget	Actual	Budget	Actual
Cost of Goods Sold	$2,564,556.31	$2,700,773.16	$1,726,031.16	$1,773,448.08
Marketing Expense	$294,766.22	$290,696.70	$187,757.29	$176,778.55
Research and Development Expense	$200,719.90	$193,236.83	$134,270.95	$125,725.88
Selling Expense	$620,427.30	$611,649.47	$405,092.93	$400,181.91
Taxes	$130,926.70	$122,526.31	$82,450.78	$80,671.87

Descriptive Analytics

- **What has just happened?**
- Examine time series, evaluating past data to predict future demands (level, trend, seasonality):
  - Identify cyclical patterns
  - Isolate the impact of external events (such as, weather)
  - Characterize inherent variability
  - Detect trends.
- Determine "causality" relationships between two or more time series:
  - As an example, forecast the demand for replacement parts for machinery at a manufacturing plant by considering both historical usage rates and known, predicted or seasonal changes in demand.
- BlueFin Technologies → did viewers liked a particular TV show last night, based on Tweets?

Descriptive Analytics

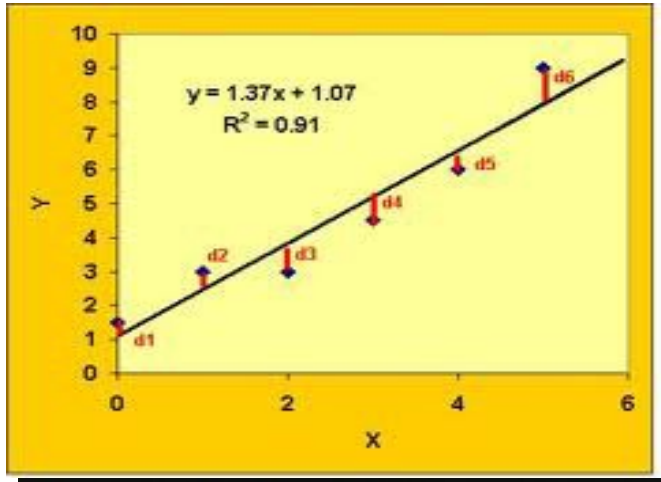
- Identifying **Patterns** in data to determine what happened
- Explaining those patterns by assessing flow across discrete temporal subsets of data
- Be wary of apparent correlations that in fact, are probably just silly.



Source: http://xkcd.com/552/

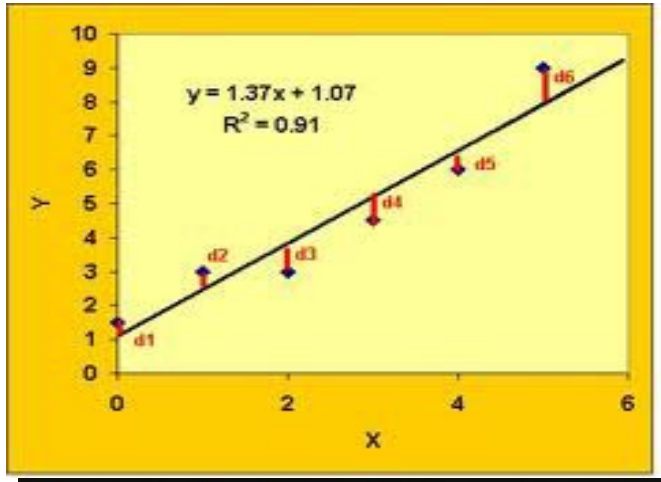
Diagnostic Analytics

- **Why did something happen?**
- Extract patterns from large data quantities via data mining
- Correlate data types for explanation of near-term behavior – past and present
- Estimate linear/non-linear behavior not easily identifiable through other approaches.
- As an example, by classifying past insurance claims, estimate the number of future claims to flag for investigation with a high probability of being fraudulent.
- Diagnostic Analytics functions:
  - Identify anomalies: Based on the results of descriptive analysis, analysts must identify areas that require further study because they raise questions that cannot be answered simply by looking at the data. These could include questions like why sales have increased in a region where there was no change in marketing, or why there was a sudden change in traffic to a website without an obvious cause.
  - Drill into the analytics (discovery): Analysts must identify the data



Diagnostic Analytics

- Diagnostic Analytics functions:
  - **Identify anomalies:** Based on the results of descriptive analysis, analysts must identify areas that require further study because they raise questions that cannot be answered simply by looking at the data. These could include questions like why sales have increased in a region where there was no change in marketing, or why there was a sudden change in traffic to a website without an obvious cause.
  - **Drill into the analytics (discovery):** Analysts must identify the data sources that will help them explain these anomalies. Often, this step requires analysts to look for patterns outside the existing data sets, and it might require pulling in data from external sources to identify correlations and determine if any of them are causal in nature.
  - **Determine causal relationships:** Hidden relationships are uncovered by looking at events that might have resulted in the identified anomalies. Probability theory, regression analysis, filtering, and time-series data analytics can all be useful for uncovering hidden stories in the data.



PREDICTIVE ANALYTICS

- Predictive analytics utilizes a variety of statistical, modeling, data mining, and "machine learning" techniques to study recent and historical data, thereby allowing analysts to make predictions about the future.

- "The purpose of predictive analytics is NOT to tell you what will happen in the future. It cannot do that. In fact, no analytics can do that. Predictive analytics can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature."

  - Michael Wu, Lithium Technologies

- Predictive analytics doesn't predict one possible future, but rather "**multiple futures**".
  - Each prediction may be associated with a likelihood of occurrence.
  - But, the highest likelihood may not be the best rational for what to do next.

Axiom: Predictive ability is only as good as the data available!

Axiom: Bad data can yield bad predictions!

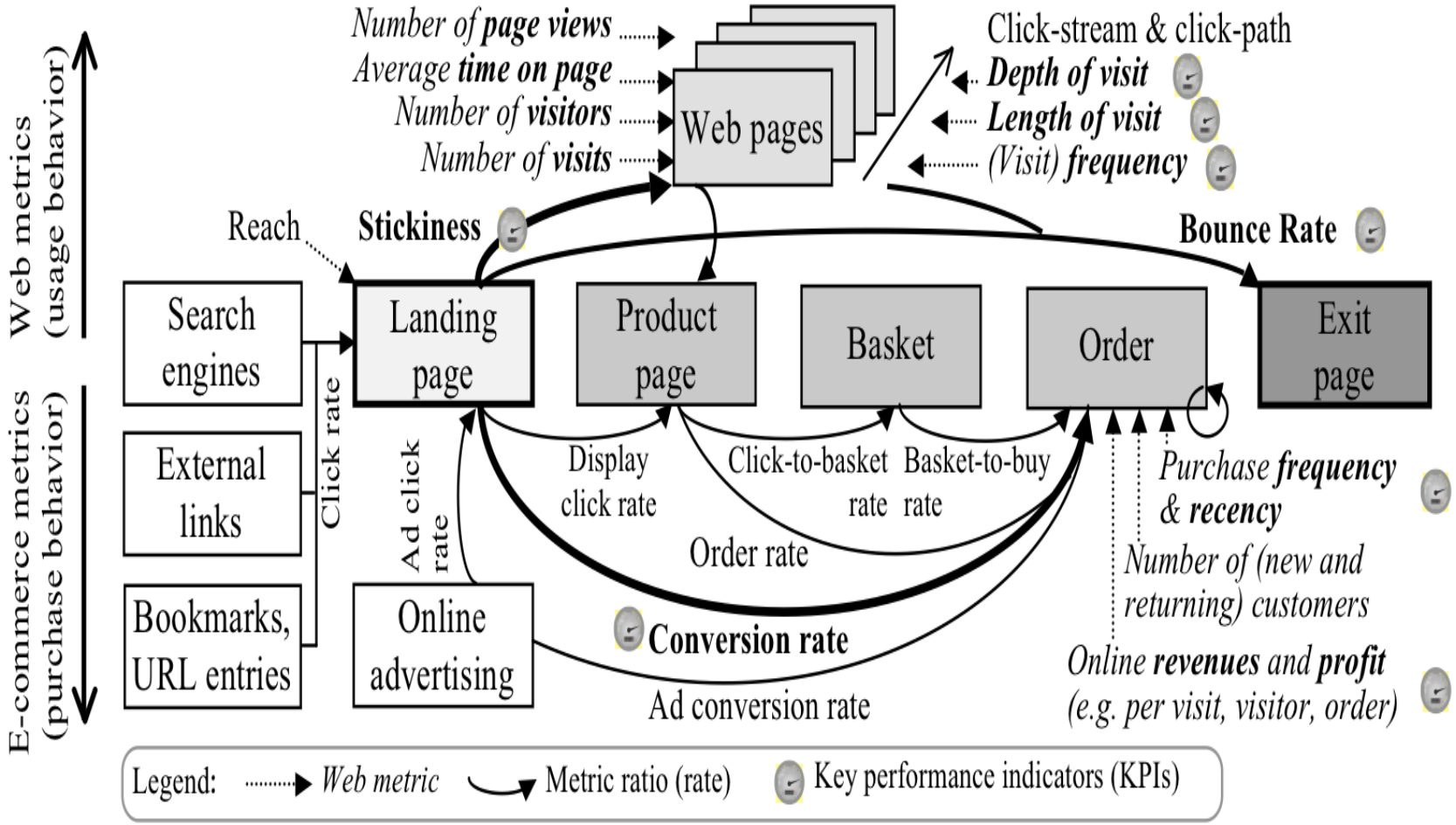
Predictive Analytics

- **What could happen next?**
- Extrapolate through different tools and techniques to long-term view
- Choosing the right data to include in models is important.
- Predictive analytics should not be searching for a diamond in a coal mine. You will get too many spurious findings.
- Rather, it is important have some thoughts as to what variables might be related.
- And once you have findings, domain knowledge is necessary to understand how they can be used.

Uses of Big Data: Target

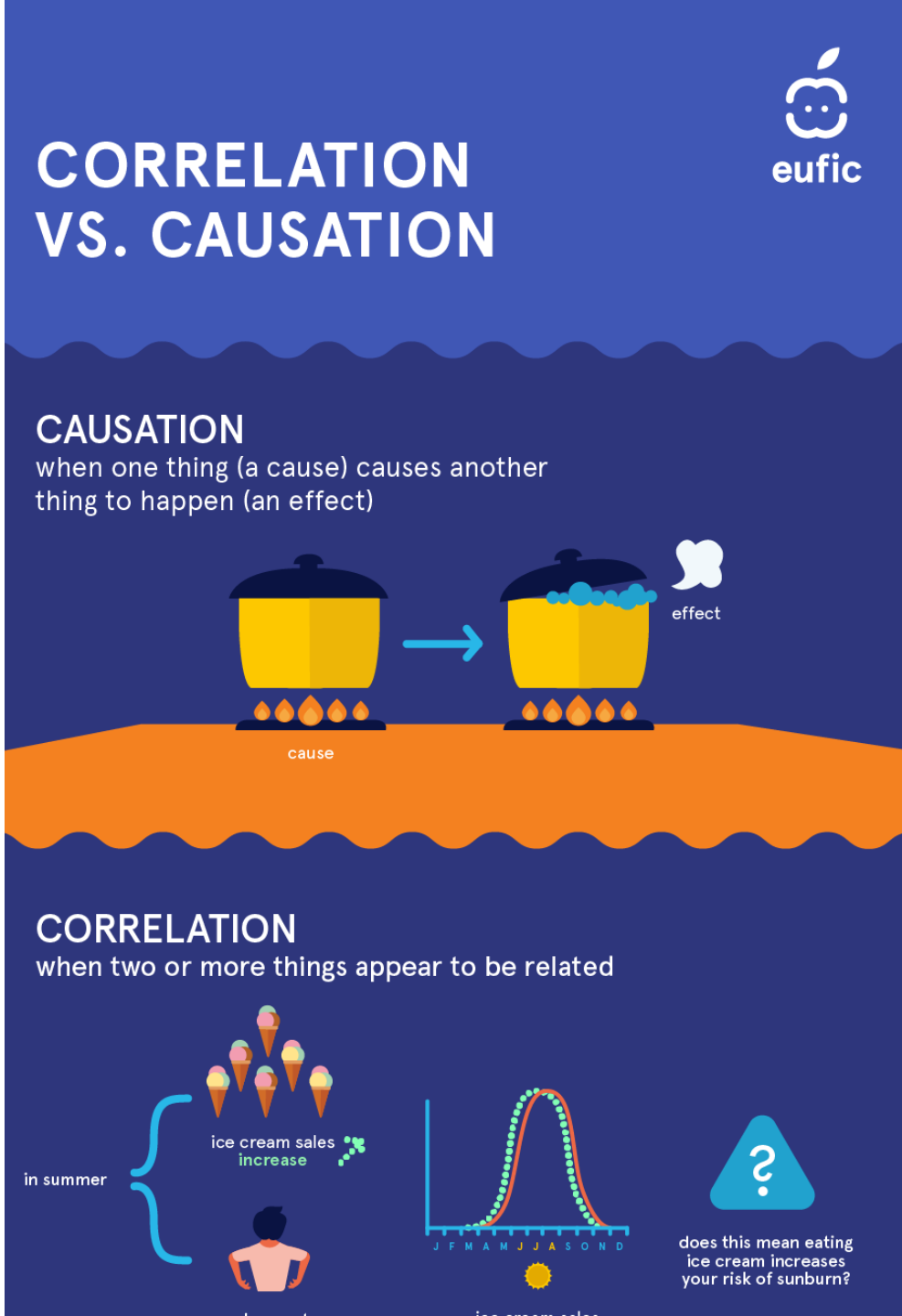
- http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
- Target assigns every customer a guest ID tied to their credit card
  - Stores a history of everything they bought and any demographic info
- Andrew Pole looked at historical data for all women signed up for Target's baby registry
- From Pole's analysis, some key patterns emerged
  - For example, they noticed that women started buying larger quantities of unscented lotion around the second trimester
  - And, in the first 20 weeks, pregnant women loaded up on mineral supplements and large cotton balls, hand sanitizers and wash clothes
- They identified about 25 products that predicted pregnancy and could do so within a narrow window
- So, Target started sending out coupons for baby items
- Which started arriving at this teen's house and caused her father to wonder what was going on.
- He began talking to the store manager, complaining about Target sending his daughter a sale booklet for baby clothes, cribs, and diapers even though she was still in high school.
- Shocked and surprised, the store manager apologized to the angry father.
- Yet, the same store manager received a call from the same father weeks later to find out the father had a talk with his daughter and discovered she was indeed pregnant.

Another Example of Predictive Analytics



Predictive Analytics Issues - I

- There is very little data that is really predictive.
  - At best, using statistics, we can extrapolate from data using numerical methods, most of which is linear extrapolation.
- "Figuring out what is truly causal and what is correlation is very difficult to do".
  - Jan Hutzius, Chief Economist for Goldman Sachs
- **Axiom: Correlation does not imply Causation!**
- Just because two variables have a statistical relationship does not mean that one is responsible for the other.
- Ice cream and forest fires are correlated because they both occur in the summer. But, you don't light a patch of Montana brush on fire every time you buy a pint of Hagen Daz!
- Targeting a particular variable in your analysis:
  - Remember the Hawthorne Effect (also Goldhart's Law)!!



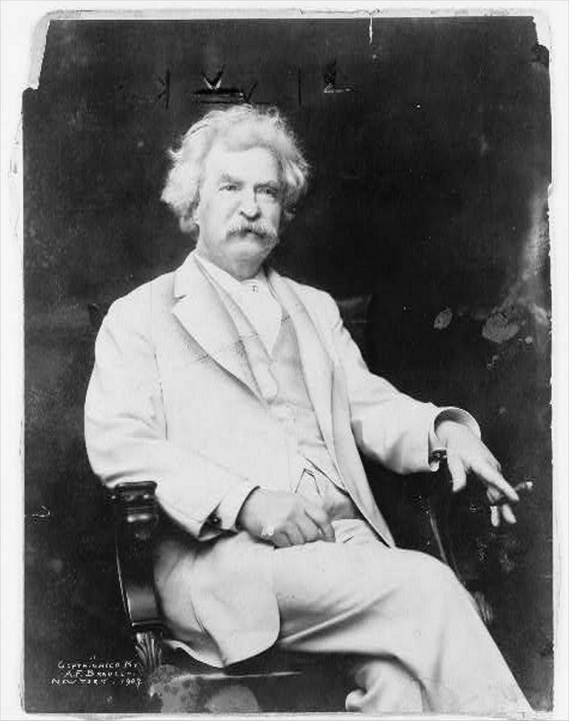
Predictive Analytics Issues - II

Many steps required:

- Data preparation
- Data cleansing
- Identifying important columns
- Recognizing correlations
- Understanding how different algorithms (math) work
- Choosing the right algorithm for the right problem
- Deciding the right properties for the algorithm
- Ensuring the data format is correct
- Understanding the output of the algorithm run
- Re-training the algorithm with new data
- Dealing with imbalanced data
- Deploying/re-deploying the model
- Predicting in real time/batch
- Integrating with your primary application to build data insights into the application and initiate user action (when embedding predictive)

Predictive Analytics: Keep this in Mind!!

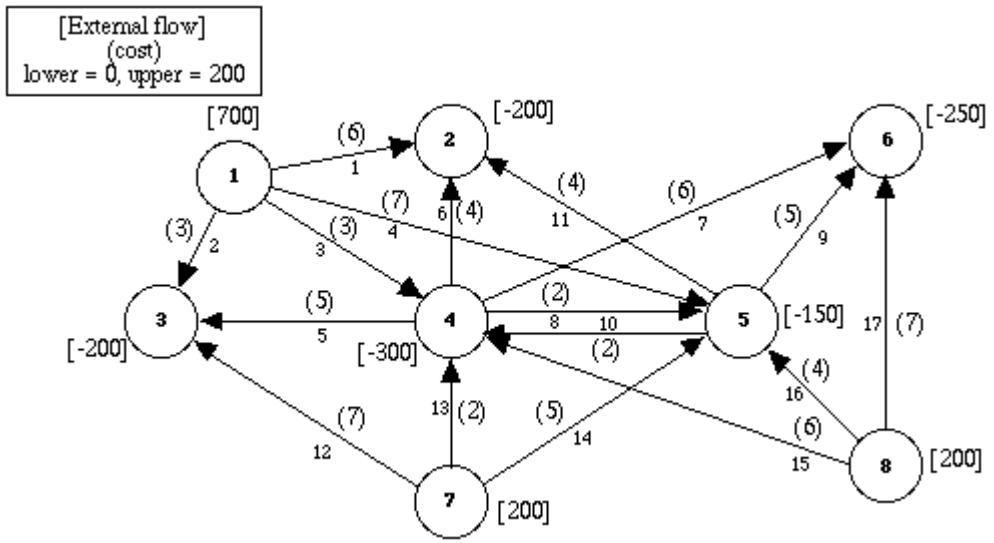
"Lies, damned lies, and statistics."
Mark Twain



- Organizations are skeptical and suspicious of the statistician!
- A phrase describing the persuasive power of numbers, particularly the use of statistics to bolster weak arguments.
- See the book in ReferenceBooks.

PRESCRIPTIVE ANALYTICS

- **What should happen next?**
- Determine what should occur and how to “make it so”
- Determine the mitigating factors that lead to desirable/undesirable outcomes
- “What-if” analysis w/ local or global optimization
  - Ex: Find the best set of prices and advertising to maximize revenue
  - And, the right set of moves to make to achieve that goal



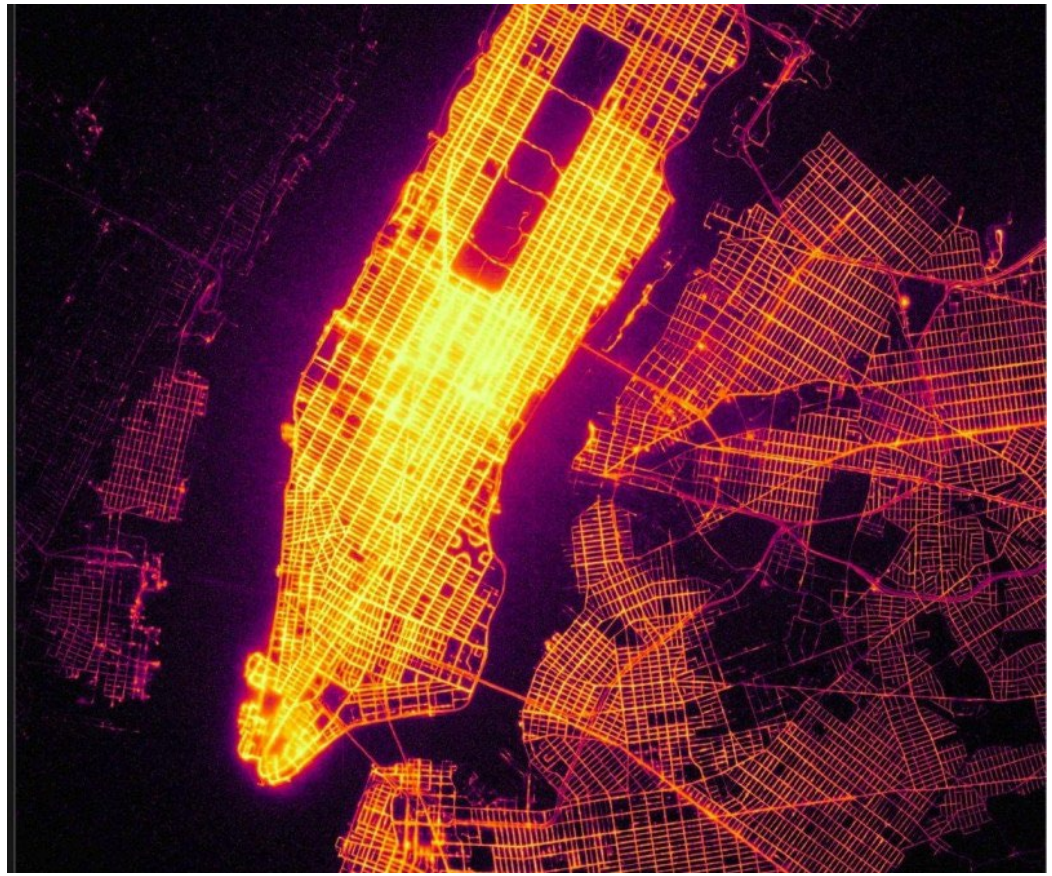
“Make it so”

VISUALIZATION ANALYTICS

- *Visual analytics*: an evolving discipline which is driving new ways of presenting data and information to the user.
- *Visual analytics*: "the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook 2005)
- *Visual analytics*:
  - the formation of visual metaphors in combination with a human information discourse (interaction)
  - enables detection of the expected and discovery of the unexpected within massive, dynamically changing information spaces. (Wong and Thomas 2004)
- *Visual analytics* provides the "last 12 inches" between the masses of information and the human mind that enables us to make decisions.
- It has been said: "A picture is worth a thousand words"

Ref: https://en.wikipedia.org/wiki/Visual_analytics

VISUALIZATION EXAMPLE - I



1.3 Billion Taxi Rides
Data collected, cleaned and plotted the pick-up and drop-off locations of all taxi rides in New York between January 2009 and June 2016.

Decisive Analytics

- **Here is What we will do!**
- Given decision alternatives, choose the one course of action to do from possibly many
- But, it may not be the optimal one.
- Visualize alternatives – whole or partial subset
- Perform exploratory analysis – what-if and why
  - How do I get to there from here?
  - How did I get here from there?

"What it comes down to is this thing is capable of telling us a lot more than we really want to know."

The Role of Analytics

- "Tools and techniques that gear the analyst's mind to apply higher levels of critical thinking can substantially improve analysis… structuring information, challenging assumptions, and exploring alternative interpretations."

  Richards Heuer, Jr., *The Psychology of Intelligence Analysis*

- Beware Frege's Caution:
  - Converse Problems:
    - If you magnify on details, you are losing the overview
    - If you focus on the overview, you don't see the details
  - Problem with Data Mining:
    - Applying statistics to understand the trends may cause a loss of problem understanding (forest vs. trees problem)
  - Problem with Statistical Machine Learning:
    - Changing the parameters, but not the decision processes and heuristics

Read about Gottlob Frege:*https://**en.wikipedia.org**/wiki/Gottlob_**Frege***

Read about Gottlob Frege:** https://**plato.stanford.edu**/entries/**frege

The Analytics Continuum

- Analytics problems span a continuum:
  - Short-term analysis leads to quick fixes and quick results, which may be unsustainable
  - What are the disruptive innovations in the middle-term that provide near-term domain leadership?
  - Long-term leads to strategic changes and innovations that provide sustainable domain dominance but take a long time to realize.

Finding a Needle
In a Haystack
Top-Down Analysis
Deductive

A little bit of this; a little bit of that
Middle-Out Analysis
Abductive?

Spinning Straw into Gold
Bottom-Up Analysis
Inductive

Finding A Needle in a Haystack

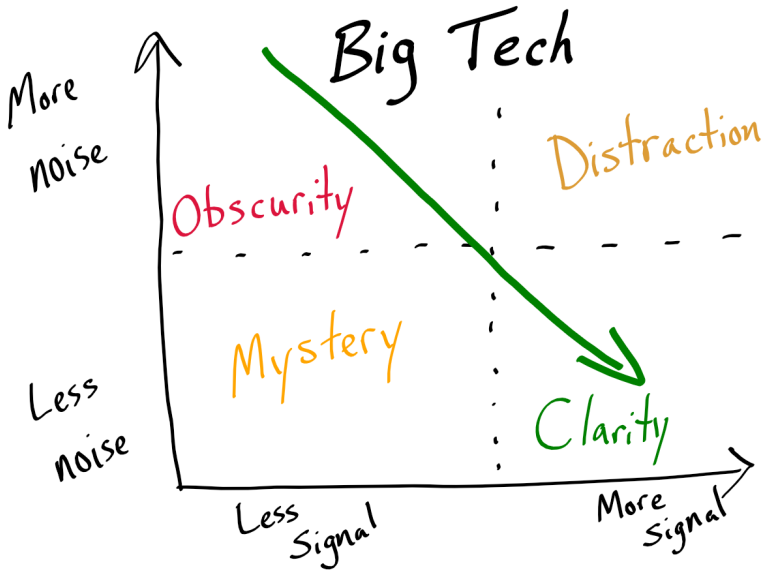
- With (all) the data available, find the/a key pattern that indicates a situational change
  - A single event
  - Perhaps, a sequence of events
  - Not the signal in the noise problem!! Where the level of noise overwhelms the signal.
- Have we seen this pattern before?
  - Determine its characteristics, not just that it exists
- Predict what event occurs next because this event occurred in the pattern
- How to identify relevant fragments of data easily from a multitude of data sources?
- Difficult to determine what the right answer is in advance

Problem: The needle hasn't grown as fast as the haystack!!

Signal and Noise

- **Signal =** the meaningful pattern that helps you predict or decide; **noise =** randomness, measurement error, spurious correlations. In the "big data" era, having more data often adds more noise unless you model uncertainty well. Silver's thesis is: think probabilistically, update beliefs as new evidence arrives (Bayesian mindset), and stay humble about model limits.

- Real-world datasets are messy (multiple sources, shifting definitions, biased instruments). If your measuring tool or assumptions are biased, collecting more data won't fix it, noise just scales up. Hence the emphasis on context, assumptions, and uncertainty rather than searching for an idealized, one-shot model.



- **Prior (baseline):** Before the campaign, a fundamentals model says Candidate A has a 55% chance.
- **New evidence:** Polls roll in; some are noisy/outliers. A Bayesian forecaster down-weights low-quality polls, accounts for house effects, and updates the probability *gradually* instead of "flipping" on a single poll.
- **Outcome:** As better polls arrive nearer to election day, uncertainty shrinks; the probability tightens (e.g., 72% → 79% → 83%). This is how Nate Silver produced accurate state-by-state calls **not** by declaring certainties, but by **updating** with new data and keeping error bars visible.

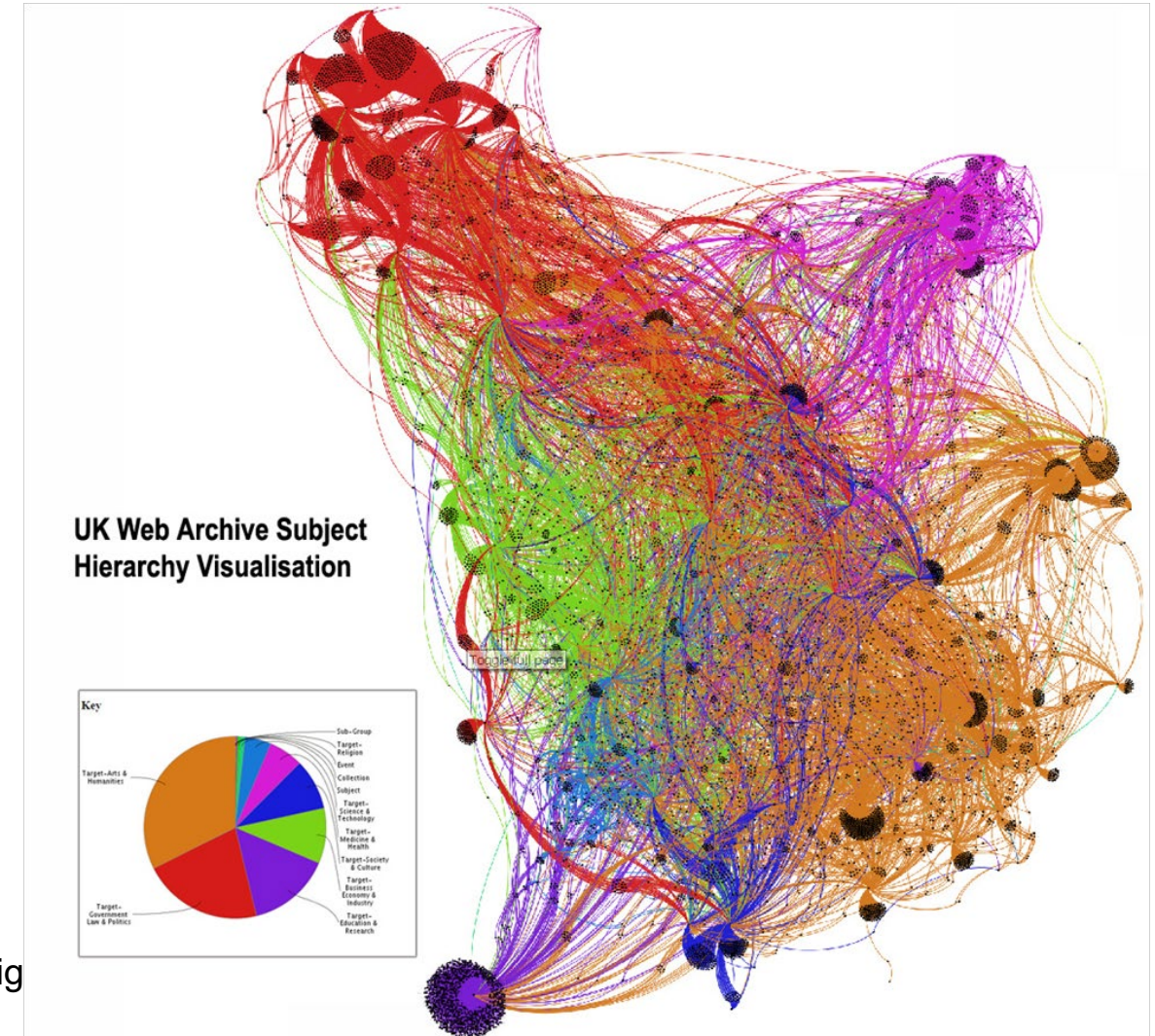
Finding A Needle in a Haystack

- What if the "needle" happens to be a complex data structure?
  - Brute force search and computation are unlikely to succeed due to inefficiency
  - Complexity increases with streaming data as opposed to a static data set

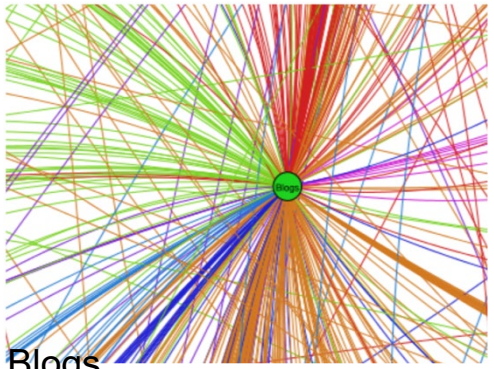
Axiom: Absence of evidence (so far) is not evidence of absence! (Borne 2013)

- What preprocessing do we need to do before searching?
  - Quality vs. Quantity: What data are required to satisfy the given value proposition?
  - At what precision and accuracy and reliability and veracity
- What if the needle must be derived rather than found?
  - How do we track the provenance of the derived data/information?
  - Is the process repeatable as we change algorithms and data structures?
  - What are the proxy variables that will be used in the derivation
- *Challenge*: Consider finding the few packets in the trillions flowing through a network that carry a virus or malware.

NEEDLE(S) IN A HAYSTACK



Ref: Crawford, L. Access and Analytics to the UK Archive, British Library, 2010



Blogs

Uhmm! What are we supposed to glean from this picture?

Relevance to Big Data

- We are drowning in the deluge of data that are being collected world-wide, while starving for knowledge at the same time*
- Anomalous events occur relatively infrequently
- However, when they do occur, their consequences can be quite dramatic and quite often in a negative sense

"There's a pony in here somewhere"

Ref: J. Naisbitt, Megatrends: Ten New Directions Transforming Our Lives. New York: Warner Books, 1982

Spinning Straw Into Gold

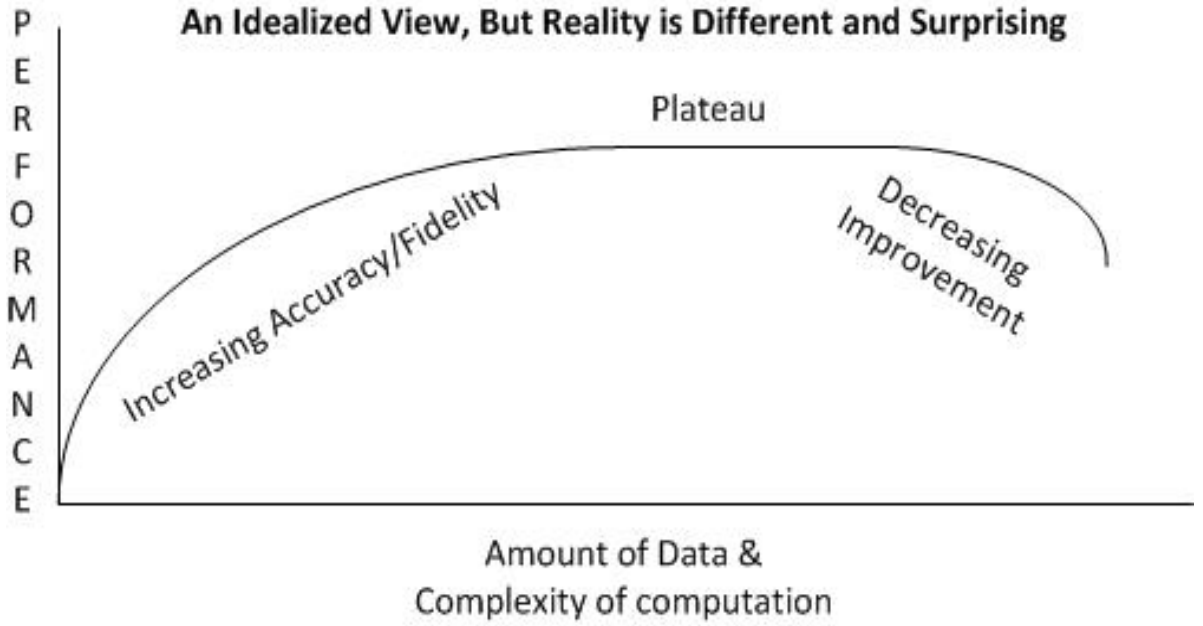
- With (all) the data available, describe a situation in a generalized form such that predictions for future events and prescriptions for courses of actions can be made.
- Objective: Identify one or more patterns that characterize the behavior of the system.

- **Axiom**: ***All data has value to someone, but not all data has value to everyone.***

  - Patterns may be unknown or ambiguously defined.
  - Patterns may be morphing over time.
  - The problem is sensemaking: the dual process of trying to fit data to a model and of fitting a model around the data to explain it.
  - Neither data nor model comes first!
  - Must co-evolve concurrently!

Finding the Knees

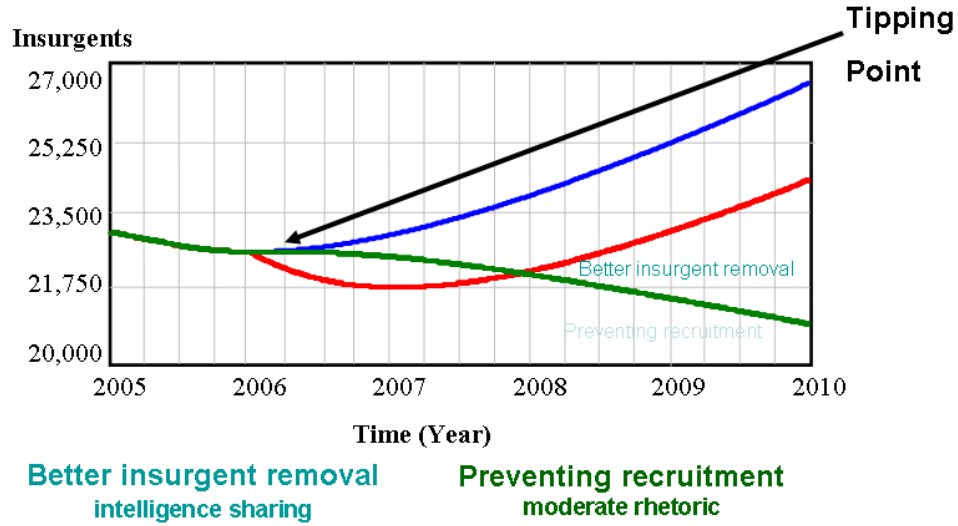
- The *knee* of an algorithm or analytic is the scale value at which the performance begins to degrade as larger data volumes are processed.
  - Every analytic and algorithm has one (or more?)
  - Where positive slope increases begin to flatten out
  - Where positive or flat slopes transition to negative slopes
- Factors affecting the knee:
  - data structure, volume, and variety
  - algorithm complexity, and
  - infrastructure implementation, including architecture.
- What is/are the corollaries for non-algorithmic analytics?



FINDING THE TIPPING POINT - I

- A *tipping point* is one in which change in a system becomes potentially irreversible and maybe even unstoppable.
  - Maybe associated with negative or positive effects
  - In social systems, a buildup to a critical mass at which point a seminal change occurs.
  - Ex: MySpace was a formidable opponent of Facebook, but once the Facebook membership reached its "tipping point" people started abandoning MySpace and signing up for Facebook.

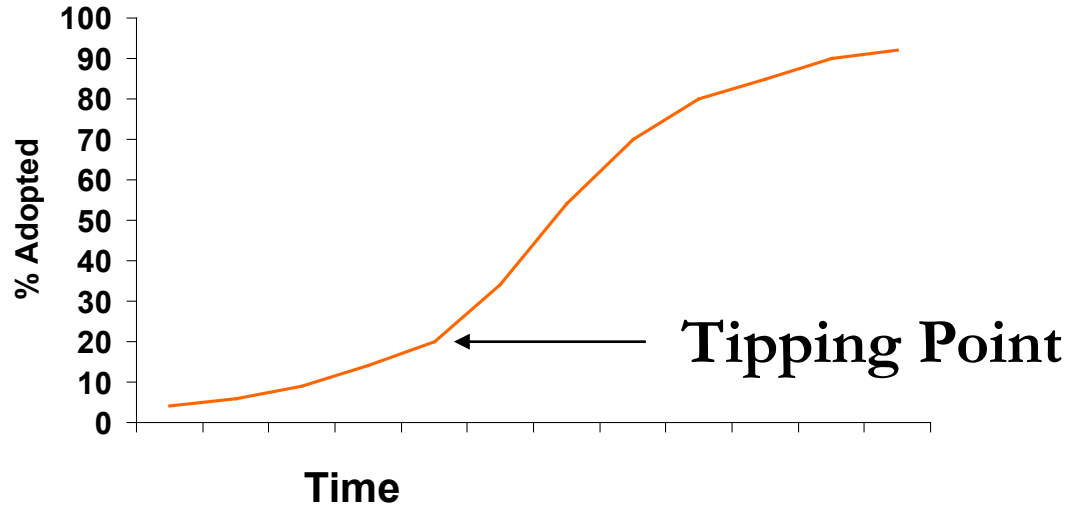
- Small events can create ripple effects – may be linear or non-linear, chaotic or perturbative
- Concept of emerging trends in the commercial marketplace
- The explosion of a viral infection into an epidemic



Ref: Choucri, N., et al. (2006) Understanding and Modeling State stability: Exploiting System Dynamics. MIT Sloan Research Papers, No. 4574-06, Jan. 2006.

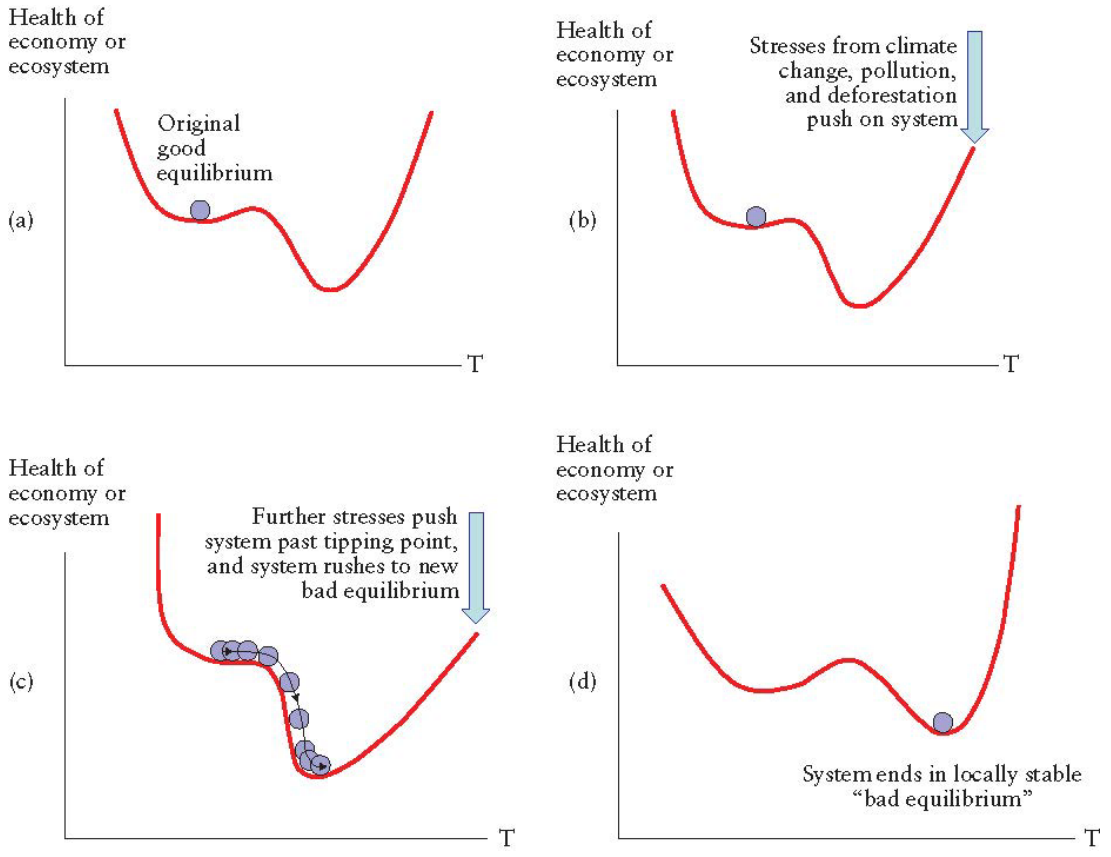
Finding the Tipping Point - II

- Tipping Points occur in every field of endeavor:
  - Epidemiology, re: the spear of epidemics such the Great Flu of 1918, or Ebola virus in the late 20th Century
  - The recent meltdown in the Housing industry and collapse of the bond markets of 2008-2009.
  - The Diffusion Curve: how info/memes propagate through a social field



Finding the Tipping Point - III

- Tipping points: moving from good to a bad equilibrium.
- Stresses can change a system slowly until a tipping point is reached, after which there are rapid and potentially catastrophic changes. Note that there are two equilibria - a good equilibrium in (a), and a bad equilibrium in (d).



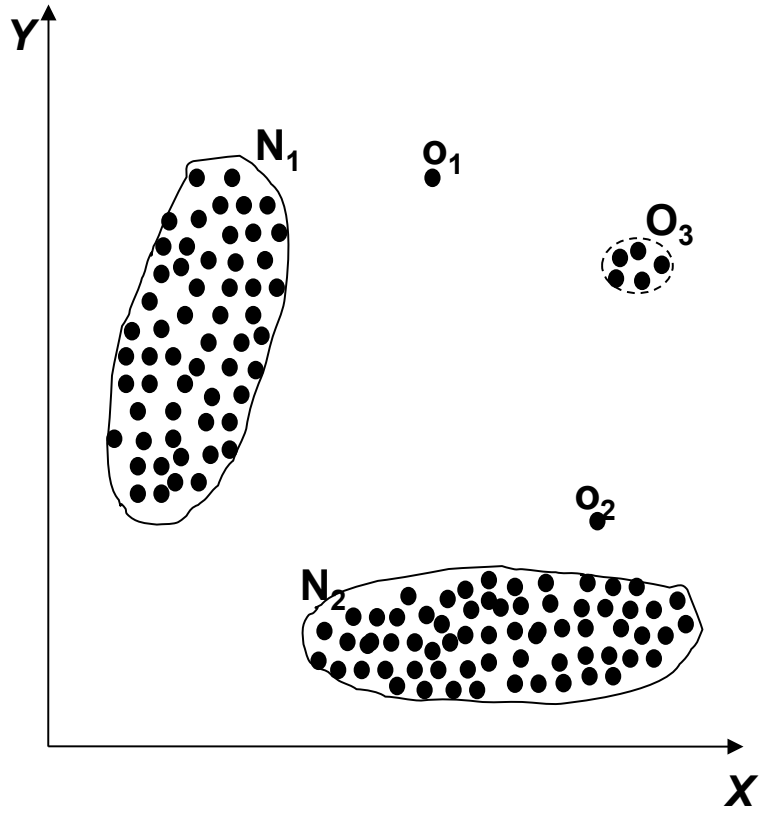
ANOMALIES - I

- An *anomaly* is a pattern in the data that does not conform to the expected behaviour
  - Behavior that occurs when it should not or is out of the ordinary
  - Behavior that does not occur when it should

    (Sherlock Holmes, "The Hound of the Baskervilles")
- Also referred to as outliers, exceptions, peculiarities, surprise, etc.
- Anomalies translate to significant (often critical) real life entities
  - Cyber intrusions
    - A web server involved in *ftp* traffic
  - Credit card fraud
    - An abnormally high purchase made on a credit card

Anomalies - II

- N_1 and N_2 are regions of normal behaviour
- Examples of Point Anomalies:
  - Points O_1 and O_2 are anomalies
  - Points in region O_3 are anomalies
- Anomalies are difficult to find using statistical techniques.
- Rule-based systems often work very well
  - But, require explicit domain knowledge
- How to Find?
  - K-Means Nearest Neighbour
  - Different clustering techniques
  - Activity vs time where sudden bursts occur



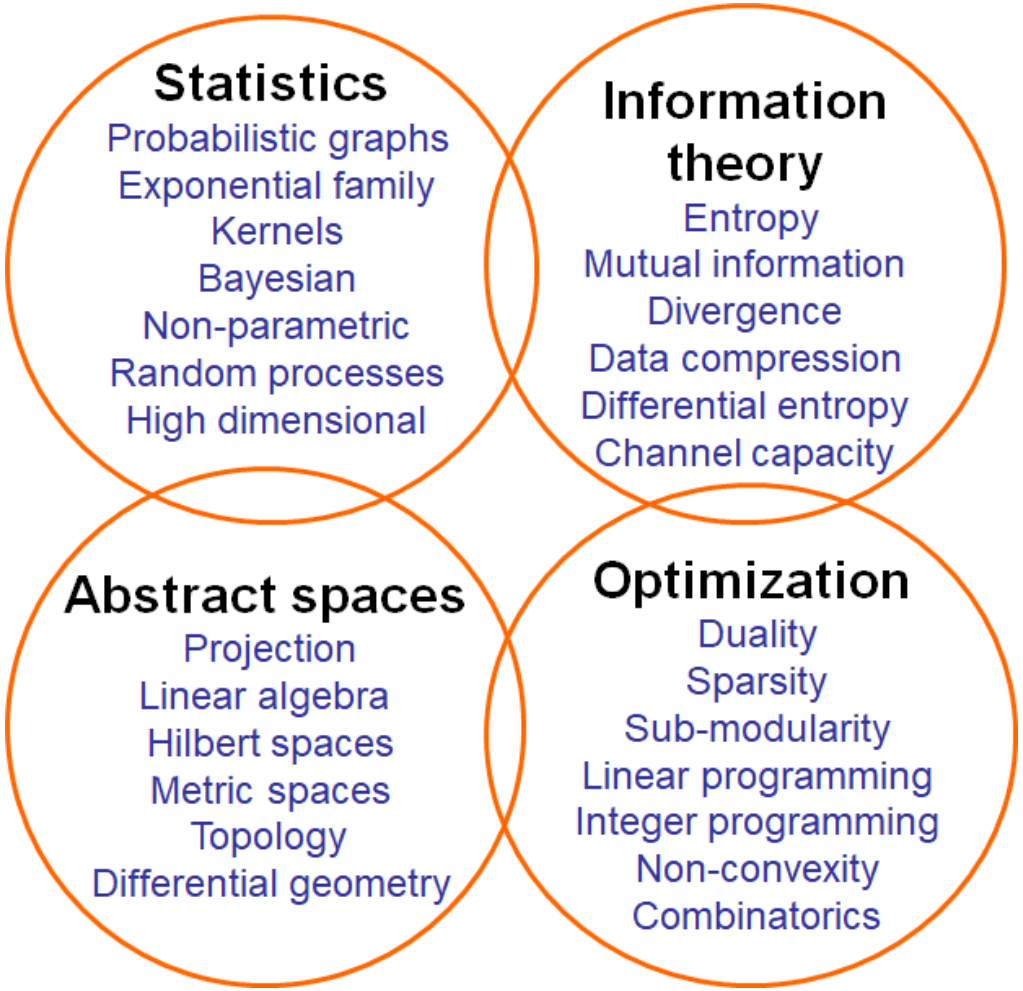
Key Skills for Analytics Scientists

- A high sense of intellectual curiousity
  - The need to know why and how
- Mathematically and Logically oriented
  - Must be comfortable with mathematics and associated tools
- Big Picture Vision:
  - Focus on the business decisions while dealing with the data
  - Q: What does it all mean?
- Detail oriented:
  - Able to pay attention to the details no matter how mind-numbing they become
- Able to differentiate between tools and methods
  - SAS, SPSS, Excel, etc. are "dumb" tools
  - Analytics is not SAS; it is using SAS to obtain results using analytical methods
- Interpretation skills:
  - Numbers by themselves mean nothing
  - Expert and domain knowledge are required to interpret the results

Training for Analytics Scientists (My View)

- 4 year college degree in Economics, Mathematics, Physics, Statistics, Ops Research, Decision Sciences, etc.
- **At least a Masters Degree in one of the above**
  - **Or, the new MS in Data Analytics**
- High proficiency in databases (e.g., NoSQL, MySQL) and query languages (e.g., SQL, Pig, etc.)
- Knowledge of programming languages: Java, Javascript, Spark, Python, R, Scala, C/C++ (?)
- Experience with domain analytics is a big plus (TBD)
- Data modeling
- Outstanding Excel and PowerPoint skills
- Concise written and verbal communication
- Good oral presentation skills
- Ability to think creatively, analytically, and abstractly
- Highly motivated and results-oriented

Mathematics for Analytic Science



Making Decisions Based on models

- "A model is a representation of a system that can be used to answer questions about the system"
- "All decisions are based on models …and all models are wrong"
  - John Stearman, Eminent Statistician
- "All models are wrong, but some are useful"
  - George Box, , Eminent Statistician

- **Axiom: It is better to be approximately right than to be absolutely wrong to infinite precision!**

REFERENCES

- Acharjya, D. P., S. Dehuri, and S. Sanyal, Eds. 2015. *Computational Intelligence for Big Data Analysis*. Springer Cham, Heidelberg, Germany
- Ahlemeyer-Stubbe, A. and S. Coleman. 2014. *A Practical Guide to Data Mining for Business and Industry*, John Wiley & Sons, New York
- Borne, K. 2013. Statistical Truiisms in the Age of Big Data, retrieved December 2013 from http://www.statisticsviews.com/details/feature/4911381/Statistical-Truisms-in-the-Age-of-Big-Data.html
- Heuer, R.J., Jr. 1999. *Psychology of Intelligence Analysis*, Center for the Study of Intelligence, Central Intelligence Agency, Washington, D.C.
- Myatt, G. J. and W. P. Johnson. 2014. *Making Sense of Data* I, 2nd Ed., John Wiley & Sons, New York
- Provost, F. and T. Faucett. ?? *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*
- Stanton, J. 2013. *Introduction to Data Science,* v3., Syracuse University, Syracuse, NY
- *Upton, G. J. G. 2017. Categorical Data Analysis By Example,* John Wiley & Sons, New York
- Nate Silver, fivethirtyeight.com ← Take a look!!

Discussion Questions

- Slide 11 described the characteristics of Big Data.

1. Does Big Data have to have all of these characteristics?

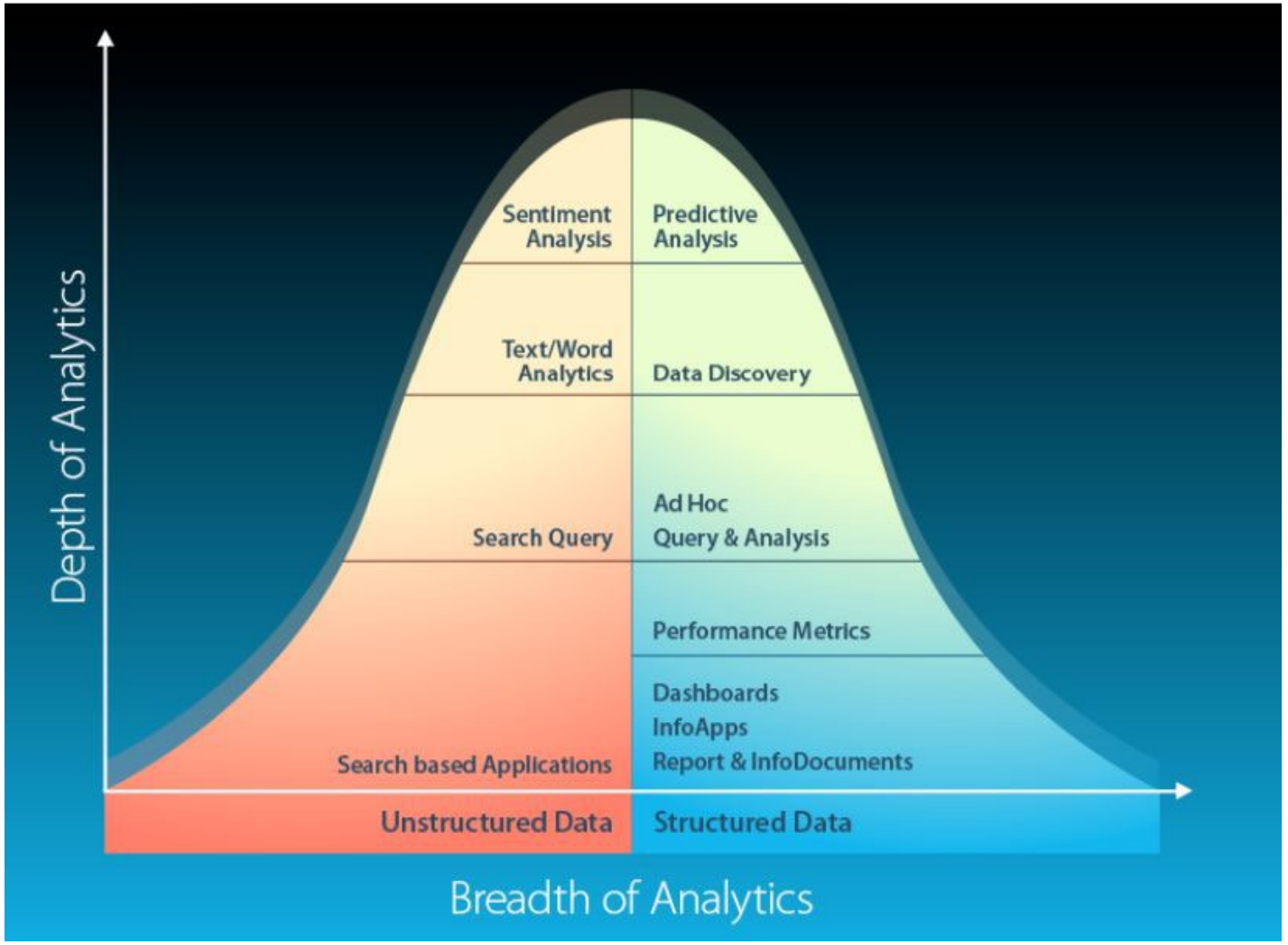
- If not, how many must it have to be Big Data?
- And, which ones are most important?

- Slide 12:

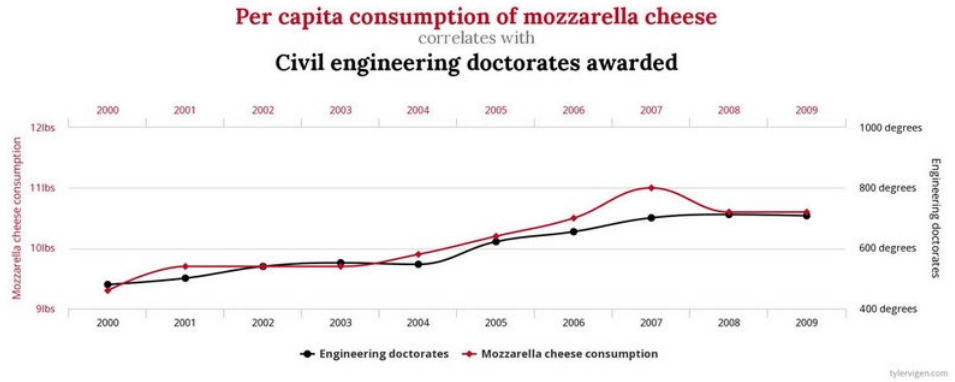
2. Argue pro or con whether more data beats better algorithms.

- How does "veracity" affect the results of algorithms?
- Is more of the same (data) useful or not? (Think Value!)

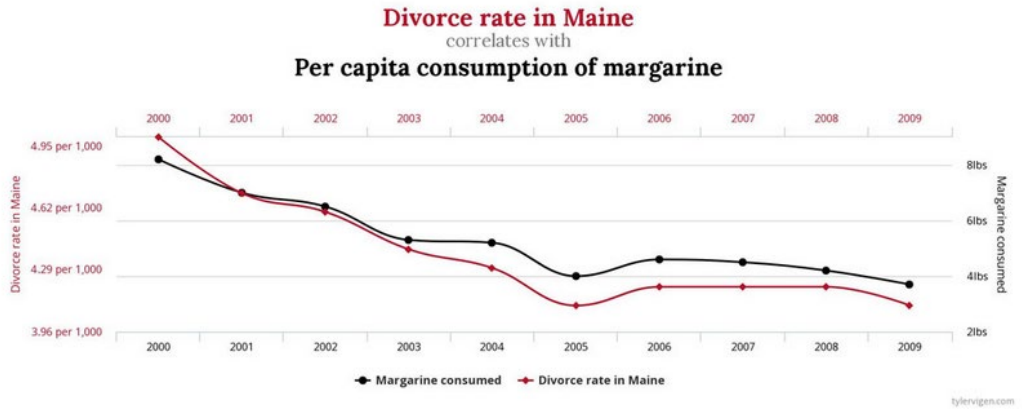
Additional Material


Depth of Analytics
Sentiment Analysis
Predictive Analysis
Text/Word Analytics
Data Discovery
Search Query
Ad Hoc Query & Analysis
Performance Metrics
Dashboards
InfoApps
Report & InfoDocuments
Search based Applications
Unstructured Data
Structured Data
Breadth of Analytics

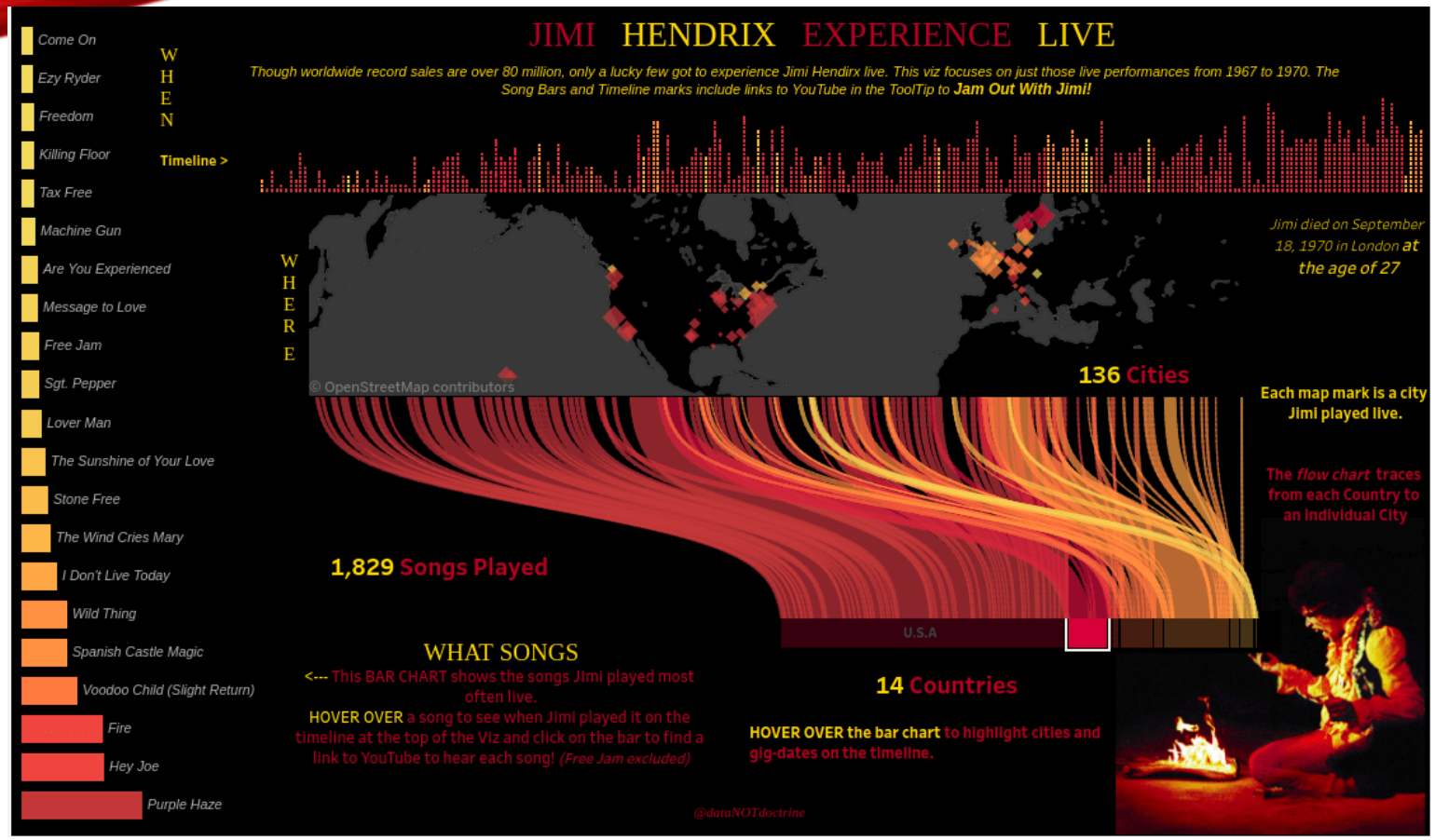
Silly Causation vs. Correlation



Source: tylervigen.com (link to license)



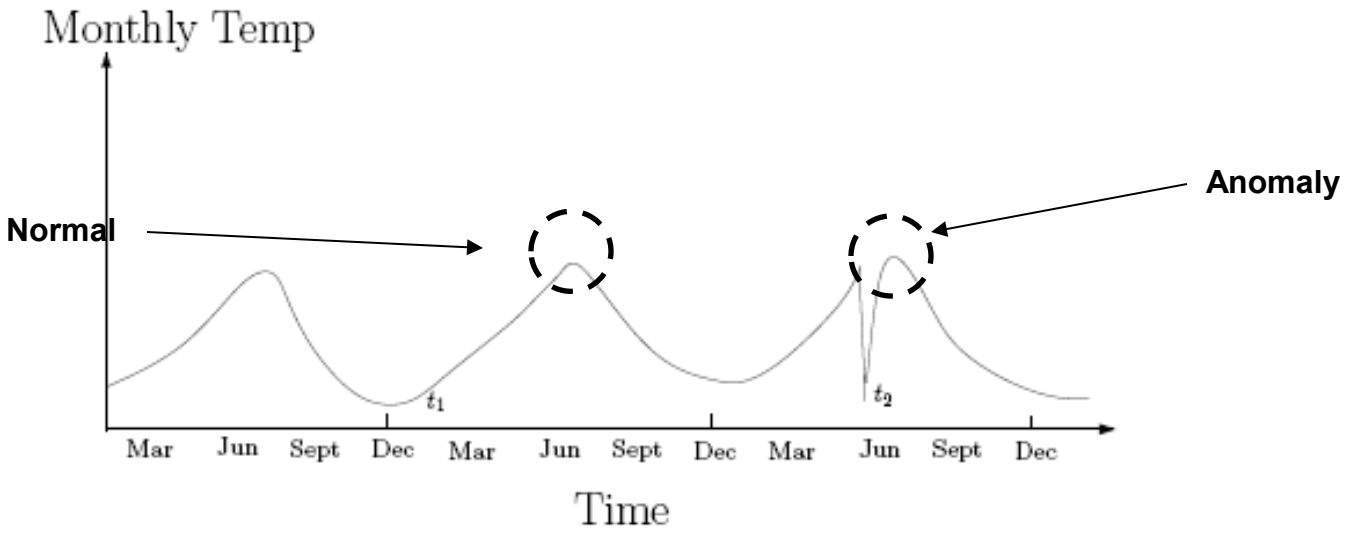
Source: tylervigen.com (link to license)



This Tableau visualization contains all of Jimi Hendrix's live performances between 1967 and 1970. It includes which songs were played and their frequency, where the concerts were held, among other insightful data.

ANOMALIES-III

- Contextual Anomalies:
  - An individual data instance is anomalous within a context
  - Requires a notion of context
  - Also referred to as conditional anomalies*

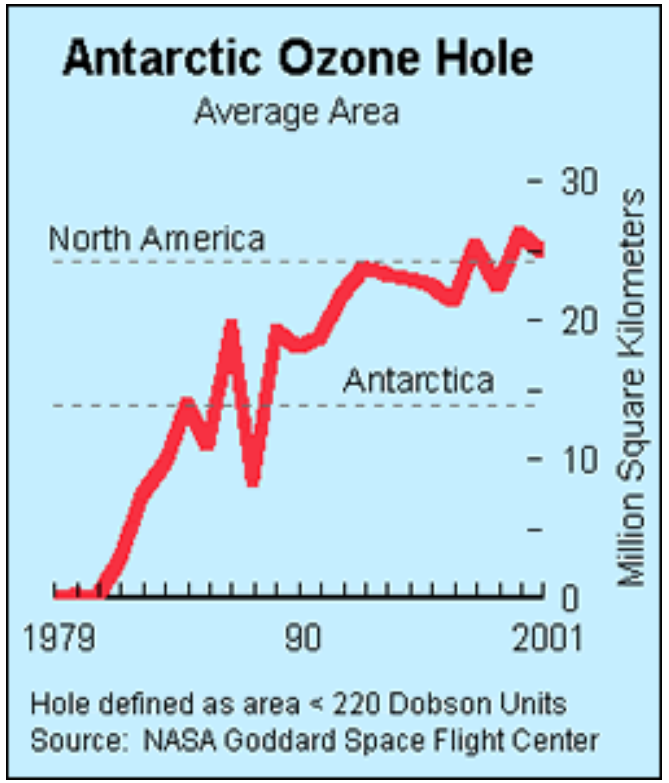


*Xiao Song, Mingxi Wu, Christopher Jermaine, Sanjay Ranka, Conditional Anomaly Detection, IEEE Transactions on Data and Knowledge Engineering, 2006.

Importance of Anomaly Detection

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels
- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?
- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Anomalies: Challenges

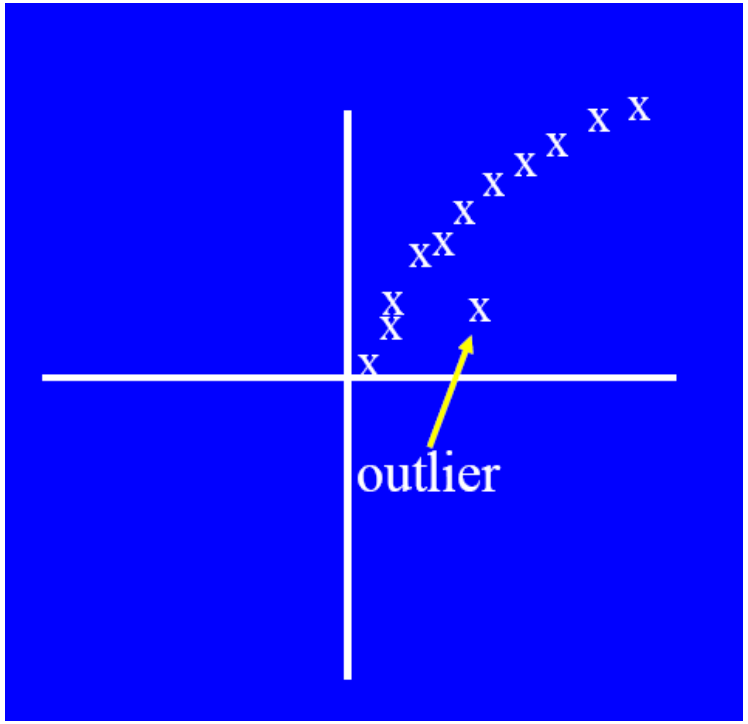
- Defining a representative normal region is challenging
- The boundary between normal and outlying behaviour is often not precise
- The exact notion of an outlier is different for different application domains
- Availability of labelled data for training/validation
- Malicious adversaries
- Data might contain noise
- Normal behaviour keeps evolving
- Type of anomaly: point, contextual, structural
- How many outliers are there in the data? Can we predict roughly?

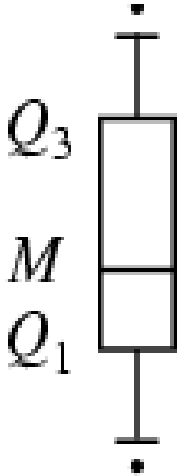
ANOMALIES-IV

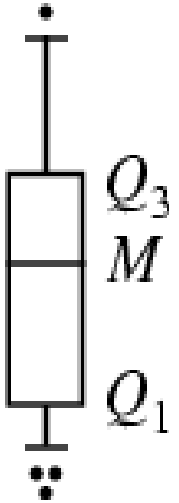
- Our Working Assumption:
  - There are considerably more "normal(N)" observations than "abnormal (A)" observations (outliers/anomalies) in the data, e.g., N >> A
- General Approach:
  - Build a profile of the "normal" behavior
    - Profile can be patterns or summary statistics for the overall population
    - Must be domain-based; otherwise, how can you explain why the anomaly occurs other than to say "it is just different"
  - Use the "normal" profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile
    - Need development versus validation data sets

ANOMALIES-V

- Graphical Approaches to Detection:
  - Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)
- Limitations
  - Time consuming; Subjective

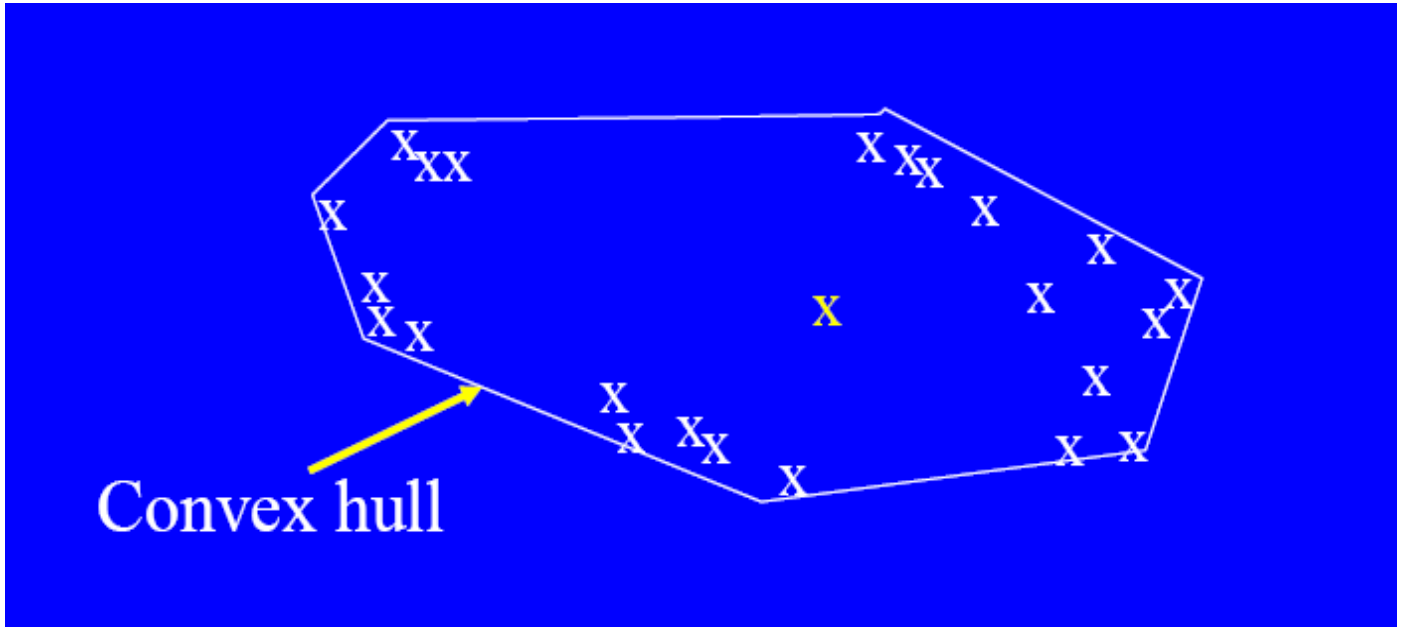






ANOMALIES-VI

- Convex Hull Method:
  - Extreme points, e.g., distant from most other points, are assumed to be outliers
  - Challenge: When data is non-numerical, how do you measure difference?



SENSEMAKING - I

- In the end, what analytics is really about is *sensemaking*:
  - What does that event really mean to me/him/her/my friends/etc.
  - What is plausible?

- Sensemaking fits data into a mental model and sets a model around the data.
  - Neither data nor comes first, but evolve concurrently
  - Data evoke models and models select and connect data

- Sensemaking requires situational awareness that helps us to adapt and respond to known and unexpected or unknown situations
  - Interpreting – something is there that is waiting to be discovered or approximated
  - Comparison to previous experience - retrospectively
  - Requires a higher level of intellectual engagement, **not a passive translation**

SENSEMAKING - II

- **"It is change, continuing change, inevitable change, that is the dominant factor in society today. No sensible decision can be made any longer without taking into account not only the world as it is, but the world as it will be."**
  - **Isaac Aasimov asserted….**

- ***Axiom: The absence of nonsense is not sufficient to make something absolutely true!***

SENSEMAKING CHALLENGES

- Lillian Wu (IBM) has noted that everything is becoming:
  - *Instrumented*: We now have the ability to measure, sense and see the exact condition of practically everything.
  - *Interconnected*: People, systems and objects can communicate and interact with each other in entirely new ways
  - *Intelligent*: People, systems and objects can respond to changes quickly and accurately, and get better results by predicting and optimizing for future events.
- Dealing w/ Ambiguity:
  - Resolving ambiguity in human languages (much data is unstructured text)
    - Ex: The word "strike" has over 30 meanings in English
  - Entity resolution is a multi-level process (Talburt 2009-2011)
    - Ex: There are more than 45,000 people named "John Smith" in the U.S.
    - Computational complexity increases with knowledge level.
- Tradeoff is end-to-end processing time versus number of entities to be resolved
  - Scaling may be/is problematic.

Ref: Technology, Data. Analytics, PSM workshop -- October 14, 2011

Ten Principles of Analytics

1. Visual + Interactive:
   - Be able to quickly view and manipulate data
   - Invoke tools as understanding evolves

2. Zero Code:
   - Tools should be designed for business or operations users
   - Require no programming (or even knowledge of it)
   - Tools should provide a lot of capabilities

3. Easy to Use (Actually!)
   - The interface must be understandable
   - GUIs are sometimes much harder to use than command lines if you don't understand what is going to happen hen you click on something
   - Need to understand what it means when you type a value into a field
   - Develop tools for the least common denominator:
     (The Homer Simpson Factor)

Ten Principles of Analytics

4. Fast:
   - A relative term, but you will know it when you see it
   - Generally, results must be returned within the attention span of the user.
   - Cycle time versus response time

5. Disposable/Persistent:
   - Some analytics are performed only once; some may be repeated
   - Be able to conceive of and execute analysis quickly, and dispose of the results when no longer needed.

6. Collaborative:
   - Analysis is often a collaborative process, requiring results to be shared among a group of users
   - What may be important is to just the results, but the steps to get there

Ten Principles of Analytics

7. Conceptually Sound:
   - Interactive analysis of data must be based on a sound set of relationships between the tools used to process the data
   - Hide the data models and storage requirements from the user

8. Depth:
   - User must be able to see, if necessary, all of the attributes of the data
   - User must be able to filter the data to obtain workable, usable subsets

9. Good Software Citizen
   - Analytics need to fit within the enterprise architecture
   - Observe security, integrity and so forth – be well-behaved

10. Expressive:
    - Moving from idea to visualization to action has to be simple

A Starter List - I

- **Understanding and Targeting Customers**
  - One of the biggest and most publicized areas of big data use today.
  - Companies are keen to expand their traditional data sets with social media data, browser logs and text analytics and sensor data to get a more complete picture of their customers.
  - The objective – create predictive models of customer behavior.

- **Understanding and Optimizing Business Processes**
  - Optimizing stock on hand based on purchase patterns, supply chains and delivery routes
  - Geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic data, etc.
  - Optimize talent acquisition – Moneyball style

A Starter List - II

- **Improving Healthcare and Public Health**
  - Analysis of DNA and genome data will enable the detection of bases for diseases and predisposition to wards certain illnesses, etc.
  - Think about the impact on clinical trials with 100,000s of people wearing different types of monitors
- **Improving Sports Performance**
  - Most elite sports have now embraced big data analytics.
  - Video analytics are used to track the performance of every player in a football or baseball game
- **Financial Trading**:
  - In High-Frequency Trading (HFT), big data algorithms are used to make trading decisions that humans are unable to make.

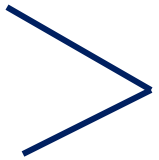
A Starter List - III

- **Science and Research**:
  - Too numerous to do justice to all of the applications.
  - CERN uses 65,000 processors to analyze the 30 PBytes of data it has collected from the Large Hadron Collider
- **Optimizing Machine and Device Performance**:
  - The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings
- **Improving Security and Law Enforcement**
  - Preventing terrorist and cyber attacks
  - Credit card companies use big data to predict fraud and crime

Tools for Data/Analytic Science (September 2018)

- Most frequently referenced tools in Glassdoor Report for job postings:

1. Python (72%)
2. R (64%)
3. SQL (51%)
4. Hadoop (39%)
5. Java (33%)
6. SAS (30%)
7. Spark (27%)
8. MATLAB (20%)
9. Hive (17%)
10. Tableau (14%)



Need these three as "bread and butter skills"

→ Fast emerging as the "go to" interactive environment

Note that Python is interpreted while R is compiled. For very large
Data sets, there may be scalability and performance differences that must be taken into account.

Top 10 Frameworks

- No framework is ubiquitous!

1. Spark/Scala (31%)
2. Hive (17%)
3. HBase (17%)
4. MapReduce (15%)
5. Presto (13%)
6. Kafka (13%)
7. Impala (11%)
8. Storm (11%)
9. Flink (9%)
10. Pig (6%)