Predictive Analytics

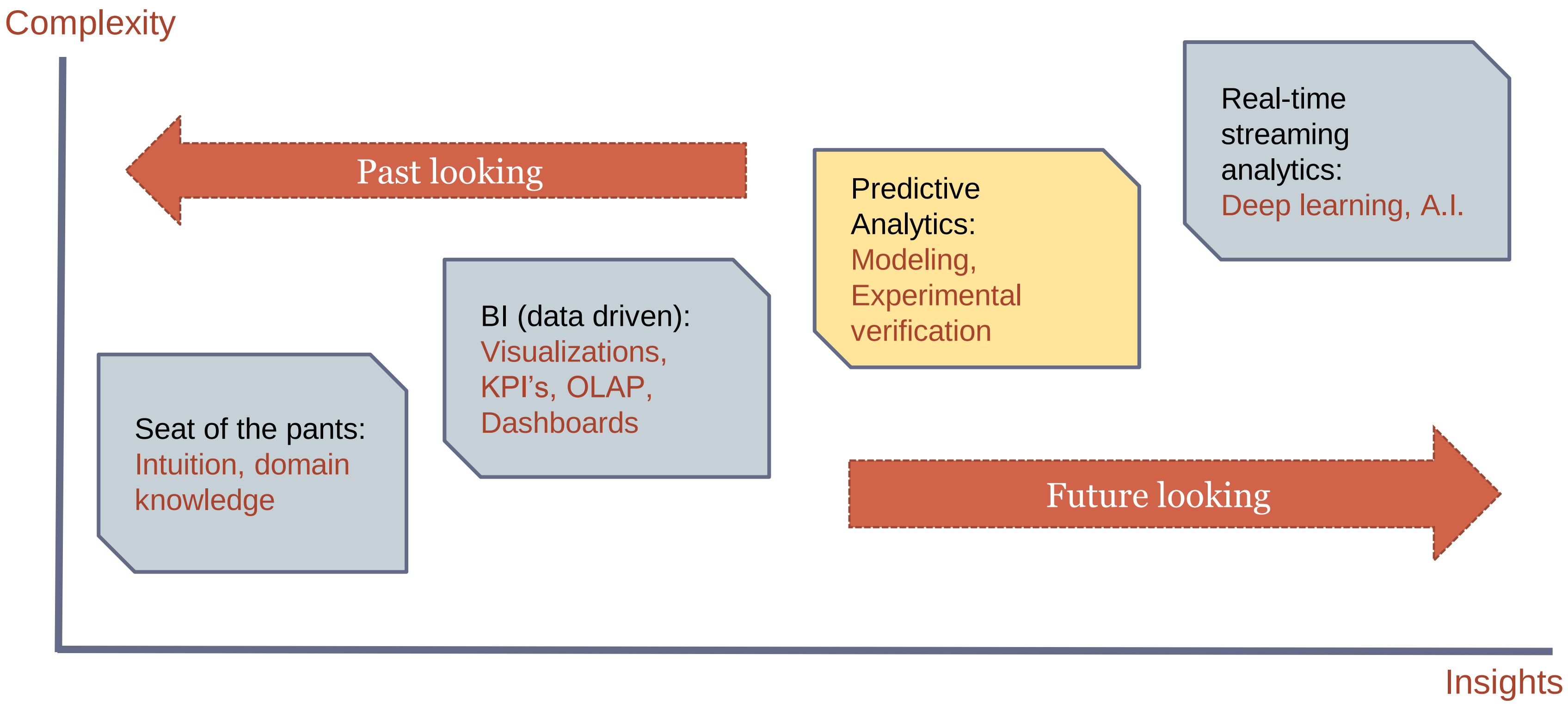
MODELING FOR RESULTS
THE CURRENT STATE OF THE CRAFT

Kimberley Mitchell

Contents

- Evolution of Analytics
- Predictive Analytics: Definition
  - Predictive Analytics: Process
  - Traditional Predictive Analytics
- 21st Century Transition: Big Data Content
  - Who is using big data?
  - Analytic process changes
  - Big Data Tools
  - Big Data Techniques
- Predictive Analytics Use Cases
- Use Cases: Optimizing Outcomes
  - Improve risk calculations
  - Optimize Energy Usage
  - Improve Customer Recommendations
  - Optimize policy interventions
- Business Use Case: Increase Revenue
  - Kraft: Digital Marketing
  - Walgreens: Using data from loyalty program
  - Legendary Pictures: Applied Analytics Division
  - Mashable: Data Driven Digital Publishing
- Business Use Case: Reduce Cost
  - Human Resources: Reducing Attrition
  - Manufacturing Process Improvement
  - Improve Customer Retention
  - Predictive Maintenance
- Predictive Analytics Trends Forecast
- Questions?

Evolution of Analytics



Complexity

Past looking

Seat of the pants: Intuition, domain knowledge

BI (data driven): Visualizations, KPI’s, OLAP, Dashboards

Predictive Analytics: Modeling, Experimental verification

Real-time streaming analytics: Deep learning, A.I.

Future looking

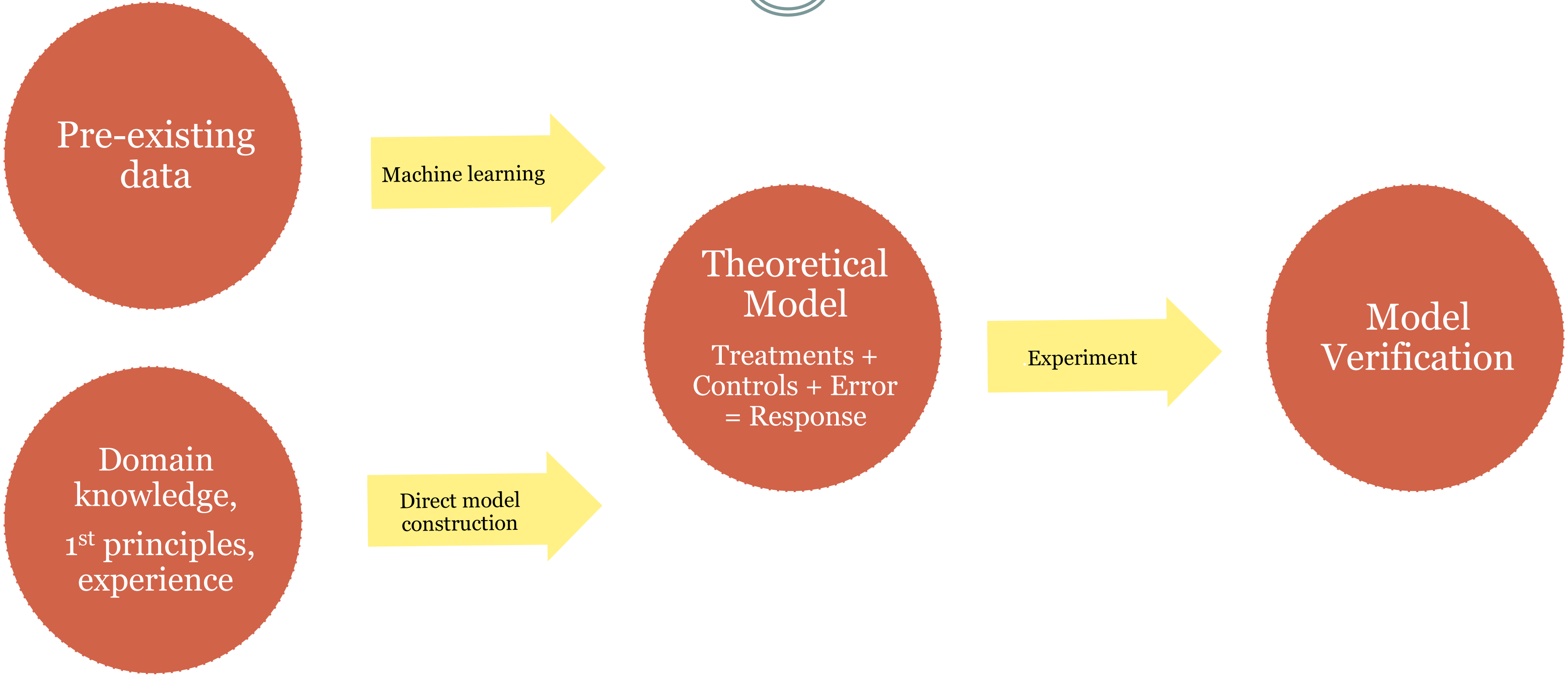
Insights

Predictive Analytics - Definition

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future.

-- Wikipedia

Predictive Analytics Process



Pre-existing data

Machine learning

Domain knowledge,

1st principles, experience

Direct model construction

Theoretical Model

Treatments + Controls + Error = Response

Experiment

Model Verification

Traditional Predictive Analytics

- Applications
  - Research and development in agriculture and industry
    - Designed experiments for product and process improvement
  - Verifying the effectiveness of healthcare treatments
    - Clinical trials (randomized double blind are best)
  - Predicting the weather
  - Polls: opinions, election politics, Nielson ratings.
  - Standardized tests: e.g. SAT's to predict college success
  - Actuarial Science: life expectancy, etc.

- Limitations
  - Each data point must be planned and collected intentionally
  - Expensive to design, collect the data

21st Century Transformation: Big Data Content

1. Internet behavior and logs
   - Every click and keystroke stored, server logs
   - Automatic byproduct of the Internet

2. Internet content
   - Text, emails, statuses, images, audio, video, etc.
   - Voluntary product of Internet use, automatically harvested

3. Internet of things
   - Sensors of all types: temperature, pressure, kinematics, images, audio, etc.
   - Must be consciously set up and designed

4. Access to thousands of sources of traditionally collected data
   - Open data movement (gov't, academia)

Who is using big data?

“Big data is like teenag everye lks about it, nobody really knows how to do it, ev thinks e yone else is doing it, so everyone claims they do …”—D Ariely (2012)

Analytic process changes

- Huge quantity and variety of data available with little effort/cost
- Substantial extra effort needed to extract and process data that have been collected without a specific end use in mind (ETL)
  - Unstructured data images, videos, audio, etc.
  - Combination of a variety of data sources
  - Some creativity required to compose relevant metrics
- Data availability leads to increased data mining, exploratory data analysis and data visualization requirements
  - Hundreds or even thousands of variables are "thrown in" to analyses
  - More "statistically significant" relationships found
    - Some new insights that are real
    - Some mistakes due to over-fitting, random clusters (some algorithms exist to help filter out false "hits")
  - Domain knowledge / insights still crucial to distinguish real effects from random chance or allied effects

Traditional Tools and Techniques

- Traditional Tools
  - Mostly commercial – SAS, SPSS, STATA, Minitab, Excel
  - Open Source – R, Python

- Traditional Techniques
  - Various flavors of regression analyses
    - Linear, multinomial, logistic, general linear models, regression discontinuity, instrumental variables, panel models, time series, etc.
    - Designed experiments, ANOVA
      - Controls built in for known covariates
      - Randomization to accommodate unknown covariates
      - Sometimes designed to extract maximum possible info from fewest observations

Big Data Tools

- Big Data Predictive Analytics Tools
  - Open Source
    - Apache Spark / MLlib
    - Python scripts interfaced to Hadoop or other distributed data source
    - Sometimes R or other traditional tools if data can be extracted and reduced to a size to fit on a single processor.
    - Mondrian (OLAP – more BI than predictive analytics)
  - Commercial / Proprietary
    - RapidMiner (Radoop for Hadoop analytics)
    - Pentaho – compare effectiveness of various models: operates against many data sources (being acquired by Hitachi)
    - HP Vertica Distributed R – ML algorithms pre-loaded
    - BeyondCore - 1st through 4th order interactions automatically calculated
    - SAS Enterprise Miner
    - SPSS adding big data analytics modules
    - IBM Predictive Maintenance and Quality
    - Homegrown solutions internal to companies (hand-coded Python or Java algorithms)

Big Data Techniques

- Big Data Predictive Analytics Techniques
  - Numerical responses / Categorical responses
  - Traditional Analytical Techniques – Linear, logistic regressions, GLM regressions, simulations
  - Distributed Analytics – i.e. MapReduce
  - Machine Learning Algorithms - Bayes network, decision trees, rule engines
  - A/B Testing – Designed randomized experiments to confirm model's predictive power
  - More use of natural experiments (due to increased sharing of data sets)

- Reference Books
  - Kuhn, M. and Johnson, K., **Applied Predictive Modeling**, Springer, 2013.
    - Many worked examples in R with data available
    - Guide to understanding various models and how to apply them.
    - Johns Hopkins/Coursera MOOC on Practical Machine Learning uses this book

  - Leskovec, Rajaraman, and Ullman, **Mining of Massive Datasets**, Cambridge University Press, 2nd edition, 2014.
    - Available commercially and also as a free pdf download from http://www.mmds.org/index2.html with associated materials.
    - Teaches the mathematical models and logical constructs (algorithms) underlying data mining and machine learning algorithms, including many exercises .
    - Stanford/Coursera MOOC on Mining Massive Datasets uses this book

Predictive Analytics: Current Use Cases

DIRECTLY RELATED TO THE SPONSORING ORGANIZATION'S GOALS

Use Cases

- What determines actual usage of analytics?
  - As always: organizational goals
- Business / organizational goals drive initiatives
  1. Improve outcomes (not obviously dollar related)
     - For non-profits: improve client outcomes
       - Fulfill the non-market based goal
     - Improve products, customer satisfaction
       - Provide the best product/experience and the customer will come
  2. Increase profits by increasing revenue
  3. Increase profits by reducing costs

Use Cases: Optimize Products or Outcomes

- More precise risk calculations in Auto Insurance (IoT)
  - Progressive - Snapshot measures risk of driving habits to offer lower rates
    - http://www.dailyfinance.com/2012/08/27/the-hidden-costs-of-cheap-car-insurance/
- Optimize comfort and energy usage in commercial buildings
  - BuildingIQ - Predictive Energy Optimization™ uses advanced algorithms to automatically fine tune and control HVAC systems resulting in savings while maintaining or improving comfort.
    - http://www.buildingiq.com/
- Improve Customer Recommendations (many examples)
  - Nordstrom - Scoring customer to product segment interaction. Output: Affinity scores of each customer to the product segment, For each segment, count of customers who have above average affinity scores compared to the general population in that segment.
    - https://ieondemand.com/divisions/analytics/events/275/presentations/multi-channel-analytics-at-nordstrom-data-labs
- Policy interventions – development economics (non-profit)
  - Data-Pop Alliance: Using cell phone data to predict socioeconomic levels in developing areas
    - https://ieondemand.com/divisions/analytics/events/275/presentations/big-data-for-development-technocratic-democratic-considerations

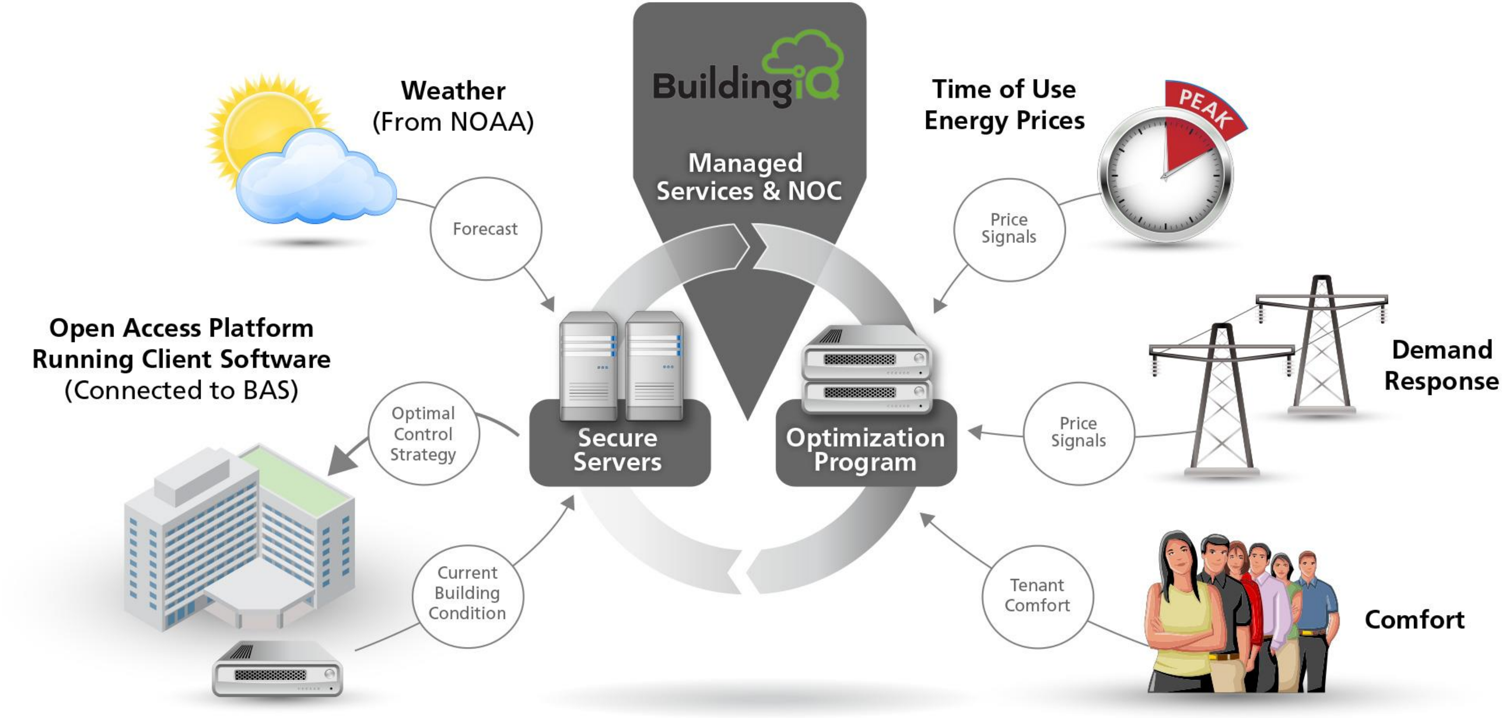
Improve risk calculations

- Progressive Auto Insurance: Snapshot measures risk of driving habits to allow the offer of lower rates
  - Automated device tracks actual driver patterns for 30 days (IoT)
  - Discounts offered based on time of day, pattern of hard braking, amount of driving
  - Response: Much clearer picture of actual driving risks than self-reporting
  - http://www.dailyfinance.com/2012/08/27/the-hidden-costs-of-cheap-car-insurance

Optimize Energy Usage

- Optimize comfort and energy usage in commercial buildings (IoT)
  - BuildingIQ - Predictive Energy Optimization™ uses advanced algorithms to automatically fine tune and control HVAC systems resulting in savings while maintaining or improving comfort.
  - Interfaces with existing building energy management system and uses predictive analytics to reduce energy costs while maintaining comfort.
    - http://www.buildingiq.com/

Optimize Energy Usage - BuildingIQ



Improve Customer Recommendations

- Many examples with “long tail” product offerings
  - Nordstrom - Scoring customer to product segment interaction.
    - Output: Affinity scores of each customer to the product segment
    - For each segment, count of customers who have above average affinity scores compared to the general population in that segment.
    - Probable tools: Clustering, recommendation engines
    - https://ieondemand.com/divisions/analytics/events/275/presentations/multi-channel-analytics-at-nordstrom-data-labs

Optimize policy interventions

- Development economics (non-profit)
  - Data-Pop Alliance: ”Our mission statement and raison d'être is to promote a humanistic, people-centered ‘Big Data revolution’: one that fosters human development and social progress through the ethical use of personal data and empowerment of the poor and most vulnerable and avoids the pitfalls of a new digital divide, de-humanization and de-democratization.”
  - Using cell phone data to predict socioeconomic levels in developing areas
  - Traditionally difficult to measure level of economic well-being; involving expensive surveys and in-person travel
  - Since many developing societies are bypassing landlines and going directly to cell phone communications, quantities of cell phone units and traffic can be used to estimate changes in societal well being at a vastly reduced cost.
    - https://ieondemand.com/divisions/analytics/events/275/presentations/big-data-for-development-technocratic-democratic-considerations

Business Use Case: Increase Revenue

- Digital Marketing
  - Kraft: How Kraft uses Digital Marketing
    - http://www.clickz.com/clickz/news/2392789/how-kraft-harnessed-data-to-transform-marketing
  - Walgreens: Using data from loyalty program to deliver value and drive customer loyalty
    - https://ieondemand.com/divisions/analytics/events/275/presentations/delivering-value-driving-best-customer-loyalty-through-pricing-promotions
  - Legendary Entertainment: Inform Creative decisions (primarily concept evaluation and cast evaluation (social media) transform marketing
    - https://ieondemand.com/divisions/analytics/events/275/presentations/changing-hollywood-paradigms-with-analytics
  - Mashable: Data Driven Digital Publishing - Mashable's proprietary Velocity platform predicts what's going viral next
    - https://ieondemand.com/divisions/analytics/events/275/presentations/data-driving-digital-publishing

Kraft: Digital Marketing

- How Kraft uses Digital Marketing
- Leveraging 18 years worth of customer connections, first from print, then from kraftrecipes.com
- Using the combined content from kraftrecipes.com with their data management systems to develop insights about their customer base.
- Slogan: “data is people”
- Now recording more than 34,000 attributes from 100 million online visitors each year, forming 800 segments of customers to buy ads against
- http://www.clickz.com/clickz/news/2392789/how-kraft-harnessed-data-to-transform-marketing

Walgreens: Using data from loyalty program

- Primarily "brick and mortar" stores; so how to get individual customer data for analytics?
- Walgreens launched "Balance Rewards" in September 2012, to combine insights from a loyalty program with consumer behavior.
  - Loyalty programs are an illustration of how non-website data can be used for predictive analytics
- Driving direct mail advertising, brick and mortar display positioning, inventory, etc.
  - Assortment: Carry products that meet the needs of targeted customer segments
  - Mass promotions: Drive loyalty with offers that attract and drive sales with targeted segments
  - Direct Marketing: Target "best" customers with "best" products for them.
  - Pricing: Understand customer pricing sensitivity.
  - Store Layout and Space Allocation: Align layout with the way customers shop.
- https://ieondemand.com/divisions/analytics/events/275/presentations/delivering-value-driving-best-customer-loyalty-through-pricing-promotions

Legendary Pictures: Applied Analytics Division

- Legendary Pictures
  - Movies such as Interstellar, Godzilla, Dracula Untold, Man of Steel, Inception, Unbroken
  - Global scale and reach
  - Innovation factory, not a warehouse

- Use of predictive analytics now a part of key strategies to inform creative decisions
  - Concept and cast evaluation
    - Green lighting concepts
    - Cast strengths and weaknesses
    - Build trailers on analytics

- Proprietary, commercial-grade software constantly analyzes unique, extensive data across people, social media, and content.
  - 500 million emails, 250 million households
  - Entertainment data
    - Metadata on films since 1989
    - Theatre data for 10 years
    - Social media data: sentiment, zeitgeist, topics

- Transform marketing
  - Old fashioned: 4 quadrants: male/female/over 25/under 25; 4 groups of 80 million each
  - Targeting: 80 million groups of 4
  - Aiming at “swing groups” not fans nor never interested “Mom”

- https://ieondemand.com/divisions/analytics/events/275/presentations/changing-hollywood-paradigms-with-analytics

Mashable: Data Driven Digital Publishing

- For article placement, mashable.com needs to predict which items will go viral in real time!
- Mashable's proprietary Velocity platform predicts what's going viral
- Mashable Velocity: proprietary software: 130,000,000 urls, with 1,000,000 urls crawled per day
- Saw an increase of 55% in web traffic over the first 2 years using the Velocity technology.
- https://ieondemand.com/divisions/analytics/events/275/presentations/data-driving-digital-publishing

Predict:
Mashable's proprietary Velocity platform predicts what's going viral next

Business Use Case: Reducing Cost

- Human Resources: Reducing Attrition
  - Talent Analytics: Case Study: “Raw Talent Traits” Correlated to Attrition - Attrition Dropped by 30% Yielding Multimillion Dollar Cost Savings
    - http://www.talentanalytics.com/resources/case-studies/
- Manufacturing Process Improvement
  - Intel - Using IoT and predictive analytics in chip manufacturing saved $3 million in core processor testing
    - http://newsroom.intel.com/community/apac_en/blog/2014/10/09/big-data-and-iot-in-manufacturing-in-everyday-life
- Improve Customer Retention
  - Time-Warner Cable: Using analytics to optimize price changes
    - https://ieondemand.com/divisions/analytics/events/275/presentations/using-analytics-to-optimize-subscription-price-changes
- Predictive Maintenance
  - How Predictive Analytics Improves Wind Turbine Maintenance
    - http://algoritmica.nl/how-predictive-analytics-improves-wind-turbine-maintenance/

Human Resources: Reducing Attrition

- Talent Analytics: Case Study: “Raw Talent Traits” Correlated to Attrition - Attrition Dropped by 30% Yielding Multimillion Dollar Cost Savings
  - CSR roles required a 12-week training plus passing a series 7 test before starting work
  - Voluntary attrition was greater than 60%; reduced to < 40% as a result of analytics intervention
  - Success linked to combining data sets; traditional HR metrics with proprietary survey to identify raw talent traits
  - http://www.talentanalytics.com/resources/case-studies/

Manufacturing Process Improvement

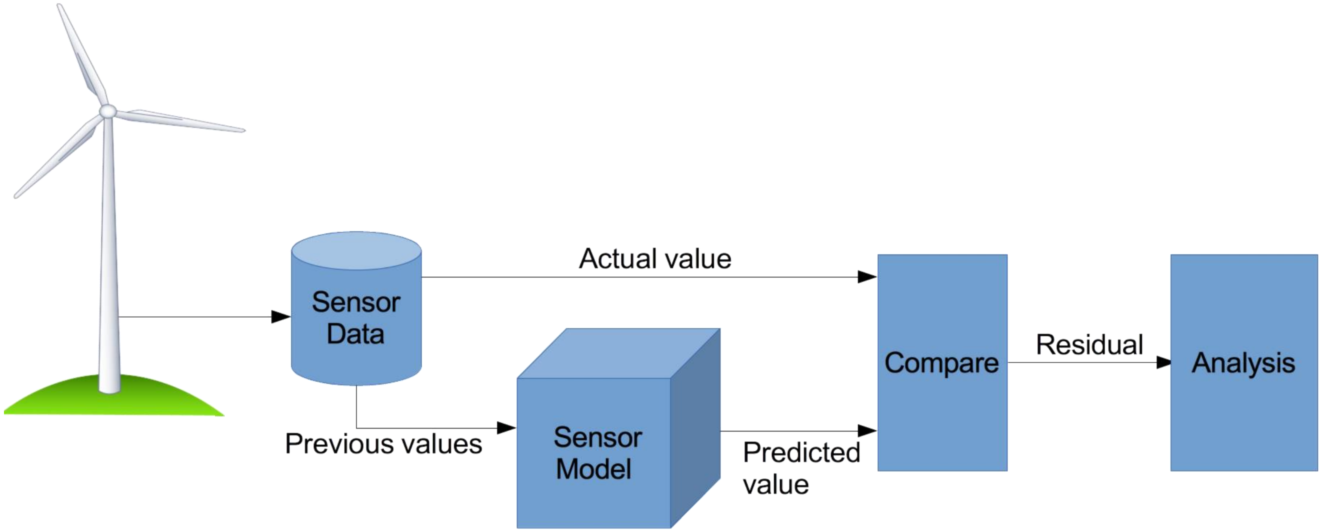
- Intel - Using IoT and predictive analytics in chip manufacturing
  - Process and product measurements taken upstream during the manufacturing process are able to predict whether the chip will pass final test leading to a reduction in final testing requirements.
  - "Instead of running every single chip through 19,000 tests, we can focus tests on specific chips to cut down test time.“
  - This predictive analytics process, implemented on a single line of Intel Core processors in 2012, allowed Intel to save $3 million in manufacturing costs. In 2013-14, Intel expects to extend the process to more chip lines and save an additional $30 million, the company said.
    - http://www.informationweek.com/software/information-management/intel-cuts-manufacturing-costs-with-big-data/d/d-id/1109111?

Improve Customer Retention

- Time-Warner Cable: Using analytics to optimize price changes
  - Using analytics, predict time-oriented price sensitivity
  - Use promotion expiration as "natural experiments" due to regulatory prohibitions against actual experimentation with customers pricing
  - Measure churn response of "test" group against pre-selected similar "control" group
    - https://ieondemand.com/divisions/analytics/events/275/presentations/using-analytics-to-optimize-subscription-price-changes

Predictive Maintenance

- How Predictive Analytics Improves Wind Turbine Maintenance
  - http://algoritmica.nl/how-predictive-analytics-improves-wind-turbine-maintenance/



Predictive Analytics Trends

- Cost of analytics will be driven down by:
  - Standardization of IoT
  - Spread of expertise
  - Commercialization and widespread use of successful techniques
  - Development of “best practices” and user-friendly software to support them in specific key applications
- Web and web-content data
  - Profusion of analytical tools will consolidate to a couple of winners
  - Most used and useful tools will be identified and commercialized
  - New applications will continue to target identifiable individuals in the “persuadable middle”
- Internet of Things
  - Development of interface standards will lead to more commercially available tools as well as open source solutions
- Healthcare analytics
  - Will increasingly use Big Data as privacy and regulatory issues resolved
- Leading edge
  - Move from manual to dynamic model generation as understanding of relevant factors improves (especially in IoT, since things tend to have more consistent behaviors than people)
  - Analysis of streaming data becomes more mainstream

Questions?

Prediction is very difficult, especially if it’s about the future.

-- Niels Bohr