Data Discovery and Analytics

INFORMS, April, 2012

Michael O’Connell, PhD
Sr Director Analytics
TIBCO Analytics

- 21st Century Trends in Analytics
- Focus Areas
  - Financial Services
  - Life Sciences
  - Telco, Media & Networks
  - Consumer Goods & Retail
- Demonstrations
- Going Forward
Key Components of a 21st Century Platform

**Automation**
- Messaging
- SOA
- BPM
- MDM

**Event Processing**
- Events/CEP
- Rules
- In-memory

**Analytics**
- Visual Analysis
- Data Discovery
- Statistics

**Cloud**
- Private/hybrid Cloud Platform
- Apps
- Deployment Option

**Social**
- Enterprise Social Media
- Collaboration
TIBCO Industry Analytics

Data

Behavior
[Real-time]

Attributes
[Customer]

Context / Facts

Predictive Analytics

Analytic Solutions

Finance / Insurance
- Customer insight
- Risk management & Trading analytics
- Fraud and Compliance

Consumer Goods & Retail
- Category/brand – value/share analysis & forecasting
- Market – product / price analytics, segmentation
- Execution – promotion effectiveness

Energy
- Exploration, drilling analytics; asset recovery prediction
- Trading and financial analytics
- Portfolio optimization

Telco, Media, Networks
- Customer life-cycle management: social churn
- Mobile marketing & advertising
- Fraud detection

Life Sciences
- Research chemistry, biology
- Clinical development
- Pharma sales and marketing
"A little bit of the right information, just a little bit beforehand — whether it is a couple of seconds, minutes or hours — is more valuable than all of the information in the world six months later... this is the two-second advantage."

-Vivek Ranadivé
THE PROPOSITION

THE TWO-SECOND ADVANTAGE

AN EVENT ENABLED ENTERPRISE

BY USING THE TIBCO PLATFORM
"...visualization-based data discovery tools have far-reaching implications for how business information is consumed... end-user organizations should adopt use as a way to improve the success of their BI program."

- Gartner, June 2011
Modeling and Data Mining

- **Supervised Learning**
  - Regression
    - Linear / non-linear models - continuous response (Gaussian underlying errors)
    - Generalized linear models – discrete response (e.g. Poisson, binomial errors)
  - Trees, Nets, SVM, Additive Models, Discriminant Functions, Nearest Neighbors, ...
  - Ensembles – Forests
    - Fit a bunch of models and average (continuous) or vote (categorical)

- **Data Mining – Unsupervised Learning**
  - Hierarchical and K-means clustering
    - Identify segments in the space of X variables
  - Association Rules
    - Identify associations e.g. between purchased products

- **Simulation**
  - Monte Carlo methods

- **Unstructured Data Mining**
  - String matching e.g. bipartite graph algorithm
  - Network analytics
Predictive Analytics in Life Sciences

**Analytics Components**
- Supervised learning: regression, safety signal detection
- Unsupervised learning: clinical trial fraud
- Monte Carlo: trial enrolment
- Unstructured: sentiment
- Data: Clinical, Safety, Operations
- ROI projection

**Real-Time Processes**
- Safety
- Site Monitoring
- Closed Loop Marketing
- Readmissions

**Pharma Research**
- Bioinformatics
- Cheminformatics

**Pharma Development**
- Instream Safety Review
- Trial Operations
- Safety

**Pharma Sales & Marketing**
- Field Force Optimization
- Campaign Effectiveness
- Speaker Lift

**Healthcare**
- Drug Outcomes
- Readmission Prediction
# Adverse Event Treatment Emergence

## Analysis
- Identify AE’s related to Treatment
- High Variable Importance $\Rightarrow$ potential treatment-emergence

## Model
- $y = f(x, \beta) + e$
- $f$ : forest of classification trees (bagging) package: S+forests
- Resample data matrix; fit tree to each sample; average over all trees
- Automatically handles interactions/combinations of AE’s
Adverse Event Treatment Emergence

Variable Importance for AEs
Tree / Forest / Bagging
Real-time Clinical Data Review

- Identify pattern in data
  - Protocol Violations
  - Missing Data
  - Physiological Trends
- Deploy rules/thresholds
- Identify pattern in data
- Escalate?
  - Assess treatment emergence of adverse events, sites
  - Drop Patient
  - Monitor Sites
- Rule Breach
- Event Cloud
  - Data reported from doctors/patients
  - Safety signals
- Build predictive statistical model
- Refine rules
- Systemic?
  - End Study
- Teradata
- Oracle
Life Sciences

Pharma: CRM

- **Response**
  - Physician behavior (prescription lift)
- **Explanatories**
  - Marketing messages and timing
  - Sales samples leave-behind
  - Call / detailing ratio
  - Speaker (thoughtleader)
- **Analysis**
  - Model behavior as function of marketing campaigns and sales detailing

**Model**

- $y = f(x, \beta) + e$
- $f$: Generalized Additive Model
Predictive Analytics in Telco, Media, Networks

- **Analytic BI**
  - Determine right offer for each target
  - Predict right timing for offer and target
  - Identify channel for dialogue with each target

- **Real-Time Processes**
  - Campaigns and offers
  - Customer care
  - BAM/BPM
  - Dashboards
  - Mobile Advertising

- **Analytics Components**
  - Supervised learning: regression
  - Unsupervised learning: cluster/segmentation
  - Data prep: derived variables, aggregation
  - Data: calls, messages, CRM data (external)

- **Profiling**
  - Customer Segmentation
  - Attribute Models

- **Relationship Growth**
  - Product Launch
  - Product Up/Cross-Sell
  - Top-Up Optimization

- **Churn Prevention**
  - Prepaid Churn
  - Postpaid Churn
  - Product Churn

- **Fraud**
  - Network Fraud
  - Smart Grid Fraud
  - Top-Up Optimization
Fraud in Networks

- **Response**
  - Fraudulent packet (Y/N)

- **Explanatories**
  - Drop-off location, Package Type, service type, Weight, Package Shape, Ship Form, Number Packages, ...
  - Zip: Lat, Long, Income, Pop Density

- **Analysis**
  - Predict package fraud (training set)
  - Score packets: intervene when Pr(fraud) > threshold (e.g. 0.8)

**Model**

- $y = f(x, \beta) + e$
- $f: \text{GLM}$  package: S+glm
Real-time Fraud Detection

Discover pattern in data

Build predictive statistical model

Predict fraud from packet and recipient attributes

Deploy model as web service

Fraud related to packet attributes and zip traits

Event Cloud

Score model on all packages; spin up instances of stats services via Silver Fabric as needed

Refine/Extend

Teradata

Oracle

Measure precision, recall

Make predictions and act on them for all packets

Fraud: Non-staffed drop-off, paper form, heavy, irregular shape
## Telco Churn

### Study historical data to discover risky patterns in customer behavior

### Historical Data: Most recent month

- Join Data
- Adjust variable types
- Calculate call change
- Select relevant service types

### Historical Data: 3 Months Prior

- Chart 1-D
- Hexbin Plot

### Apply model to Current Data. Business value = prediction of risk for each customer

### Current Customers: Most recent month

- Join Data
- Adjust variable types
- Calculate call change

### Current Customers: 3 Months Prior

- Select relevant service types
- Classification Tree

### Variable Estimates

| Variable                | Estimate | Std. Err. | t-Statistic | Pr(>|t|) |
|-------------------------|----------|-----------|-------------|----------|
| (Intercept)             | -4.56    | 0.20      | -23.23      | 1.49E-114|
| call.length             | 0.07     | 0.01      | 13.93       | 0.00     |
| cust. minutes           | 1.15E-3  | 4.41E-5   | 26.06       | 0.00     |
| cust. value             | -0.01    | 0.01      | -0.62       | 0.54     |
| call.change             | -29.57   | 0.87      | -34.10      | 8.39E-234|
| cust.MOB                | -0.66    | 0.03      | -25.65      | 5.65E-138|
| cust.Balance            | 1.05E-4  | 5.27E-4   | 0.20        | 0.84     |
| cust.CycDelq            | 0.29     | 0.09      | 3.22        | 1.27E-3  |
| cust.service(Bundle)     | 0.01     | 0.05      | 0.25        | 0.81     |
| cust.service(Voice)     | -0.01    | 0.05      | -0.25       | 0.81     |
**Telco Churn**

- **Response**
  - Churn (Y/N)

- **Explanatories**
  - Calls: length, change, dropped
  - Customer: Value, balance, delinquent cycles, service bundles
  - Social: friends and family connections

- **Analysis**
  - Predict churn (training set)
  - Score customers: intervene when $Pr(\text{churn}) > \text{threshold}$ (e.g. 0.8) and/or on key predictors

**Model**

- $y = f(x, \beta) + \epsilon$
- $f$: Generalized Linear / Additive Models
Predictive Analytics in Financial Services

Trading
- Equity Trading
- Fixed Income
- Fraud and Compliance

Risk Management
- Enterprise Risk Aggregation
- Credit & Counterparty Risk
- Market Risk

Customer Insight
- Acquisition / Retention
- Customer Growth
- Credit Scoring

Fraud & Compliance
- Anti Money Laundering
- Credit Card Fraud
- Trade Fraud & Compliance

Portfolio Management
- Portfolio Optimization
- Regulation Compliance
- Performance & Attribution

Analytics Components
- Supervised learning: regression
- Unsupervised learning: customer segmentation
- Monte Carlo: VaR and Economic Capital
- Unstructured: customer insight
- Data: Trading / Risk / CRM
- ROI projection

Real-Time Processes
- Trading
- Risk Aggregation
- Fraud
- Cross-Sell
- Portfolio
Financial Services Challenges

Need for Business Expansion
- Pressure to reduce cost
- Customer acquisition and retention

Increasing Regulatory Pressure
- Basel III / Economic Capital
- Dodd Frank/Volcker: transparency
- Compliance / Anti-Money Laundering
- Regulatory Reporting

Reports are Information-Poor
- Traders can't obtain suitable views of risk and sensitivities
- Portfolio managers need assets in single view
- Need to explain complex analyses to management
- Rogue excel analyses: need single source of truth!

Gartner reports:
- Volume of information is growing at a minimum rate of 59% annually
- 28% of potential BI users engage with the existing enterprise platform

Sources:
1: Gartner’s Rita Sallam presentation at TUCON 2011, Sept 28, 2011
Customer Growth: Retail Bank

- **Response**
  - Acceptance (Y/N) on credit card offer

- **Explanatories**
  - Banking transactions: ATM, checking; deposit, withdrawal, balance, ...
  - Demographics: sex, age, profession, nationality

- **Analysis**
  - Predict acceptance from transactions and demographics (training set)
  - Score customers: provide offers to “persuadable” customers

**Model**

- \( y = f(x, \beta) + e \)
- \( f : \) Classification Tree
Trading Analytics

Traders need rapid and comprehensive view of risk and sensitivities

Equity Trading
- Analyze price / volume across holdings with drill down into individual securities
- Calculate statistical and technical indicators, benchmark comparisons
- Incorporate indicators and financial data e.g. Bloomberg, Thomson, FactSet
- Collaborate across trading teams

Fixed Income Analysis
- Identify unknowns: bond prices, interest, durations, convexity and yield
- Compare fixed income instruments with economic indicators
- Incorporate statistical methods applied to new and existing instruments

Trade Fraud and Compliance
- Track desk in near real time
- Identify violations, patterns, rule-brushing traders
Risk Management

Risk managers need interactive and aggregate views on risk; and ability to visualize complex risk analyses with management.

Enterprise Risk Aggregation
- Aggregated view of multiple risk positions by geography, product, ...
- Assess loss given default for products and counterparties / days to maturity

Credit and Counterparty Risk
- Handle unexpected changes in markets with rapid rebalancing
- Incorporate predictive analytics e.g. stress tests, economic capital
- Collaborate between traders and managers

Market Risk
- Mark to market, VaR, scenario and sensitivity analysis
- Interest rate and FX-risk
- Collaborate across organization for rapid response to market conditions

Operational Risk
- Track desk in near real time
- Identify violations, patterns, rule-brushing traders
Customer Insight

Business leaders and marketing groups need insight into prospects and customer behavior and to identify “persuadables” for growing the business.

Customer Acquisition/Retention
- Understand net transactions and attractive products for individual investors
- Identify prospects quickly through data-mashups of in-house and industry sources

Customer Growth
- Identify persuadable customers for cross-sell/up-sell programs
- Drive campaign efficiencies
- Discover sensitivities of acceptance to demographics and behavior

Credit Scoring
- Review sequence patterns across credit card, auto and mortgages
Fraud and Compliance

Regulators expect banks to have powerful and robust analytics for compliance and Anti Money Laundering (AML)

Anti-Money Laundering
- Identify / trace unknown and suspect transactions across context eg business lines and geographies
- Set thresholds, define segments and automate alert scoring
- Comply with OCC guidelines

Credit Card Fraud
- Identify fraudulent transactions by vendor, location
- Highlight patterns, context and identify root cause

Trade Fraud and Compliance
- Track desk in near real time
- Identify violations, patterns, rule-brushing traders
Portfolio Management

Managers require multiple assets and funds in single view; need to prove value and comply with new regulations

Portfolio Optimization
- Analyze risk-return sensitivities
- Drill to efficient portfolios of interest, incorporating predictive analytics
- Rebalance on demand

Regulation Compliance
- Comply with new regulations that require thorough understanding of trading across assets and liabilities

Performance and Attribution
- Dimension-free exploration of assets across time
- Real-time analysis and review of fund strategies impact on performance
- Rapid transaction reconciliation with exceptions analysis
- Integrate with CRM data for customer insight
Predictive Analytics in Retail and CPG

Analytics Components
- Supervised learning: regression cloning
- Unsupervised learning: segmentation, association
- Data prep: derived variables, aggregation
- Data: POS (e.g. Nielsen/Nitro), product, CRM data (e.g. Axicom)
- ROI projection

Real-Time Processes
- Campaigns and offers
- Customer | Retailer Loyalty
- BAM/BPM
- Dashboards
- ERP

Inventory Mgt
Demand Forecasting
Site Planning
Brand/SKU Analysis
Product Mix Analysis
Retail Execution/Monitoring
Pricing Optimization
Campaign mgt
Segmentation
Customer Cloning
Product Up/Cross-Sell
Product Affinity

Product/Store Optimization

Product Analysis

Customer Planning

Relationship Growth/Loyalty

Analytic BI
- Determine right offer for each target
- Predict right timing for offer and target
- Identify channel for dialogue w each target
Retail: Analytics-Driven Offers

Product
- Offer associated product
- Cross-channel, cross-sell
- Product association

Product Association + Offer Acceptance

Customer
- Identify “persuadable” customers
- Personalized offers
- Customer cloning

Product Category Purchases + Customer Demographics

Campaign
- Campaign/Inventory offer: demographics + behavior
- Segment push

Customer Purchases + Demographic Segments
Retail: Recommendations

- **Purchase (Variables)**
  - All items purchased by members in prior year

- **Analysis**
  - Identify associations between products purchased by same customer in given time window
  - Make recommendations for companion / next purchase

**Model**
- supp\((X)\) = \#X in dataset / N
- conf\((X \Rightarrow Y)\) = supp\((X U Y)\) / supp\((Y)\)
- lift\((X \Rightarrow Y)\) = supp\((X U Y)\) / supp\((X)\)*supp\((Y)\)
- Note: supp\((X U Y)\) = supp\((X)\)*supp\((Y)\) iff
  - X, Y are independent itemsets
- Spotfire Miner; very scalable “apriori algorithm”
Retail: Recommendations

Product Associations

Association Network
Retail: Recommendations

Make offers to moderate confidence and lift product associations

Cross-sales: $3M lift in 3 mths
Assuming 1-2% offer redemption
Retail: Rebates

- **Response**
  - Cell phone purchase

- **Explanatories**
  - Customer information (text)

- **Analysis**
  - Customer text processing
  - Default / rebate prediction

- **Models**
  - Text match: bipartite string match
  - (polygrams, alignment, thesaurus)
  - Default: $y = f(x, \beta) + e$
  - $f: tbd$
## Retail: Rebates

**Companies**

17.9 Million US Businesses.

<table>
<thead>
<tr>
<th>ContactName</th>
<th>Title</th>
<th>CompanyName</th>
<th>Industry</th>
<th>ADDRESS</th>
<th>City</th>
<th>State</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ceo</td>
<td>eog</td>
<td>Oil &amp; Gas-Production</td>
<td>333 Clay St # 4200</td>
<td>Houston</td>
<td>TX</td>
<td>77002-4006</td>
</tr>
<tr>
<td>1</td>
<td>0.86666 Mark G Papa</td>
<td>CEO</td>
<td>Eog Resources Inc</td>
<td>Oil &amp; Gas exploration services</td>
<td>P.O. BOX 4362</td>
<td>Houston</td>
<td>TX</td>
</tr>
<tr>
<td>2</td>
<td>0.76187 Mark Papa</td>
<td>Chief Executive Officer</td>
<td>Eog Resources Marketing Inc</td>
<td>16300 Kuykendahl Rd # 330</td>
<td>Houston</td>
<td>TX</td>
<td>77068-2700</td>
</tr>
<tr>
<td>3</td>
<td>0.75947 Joseph Azimi</td>
<td>CEO</td>
<td>Advanced Laser Technologies</td>
<td>1990 Post Oak Blvd # 2100</td>
<td>Houston</td>
<td>TX</td>
<td>77056-3847</td>
</tr>
<tr>
<td>4</td>
<td>0.69323 Michael J Jacobson</td>
<td>CEO</td>
<td>Blue Dolphin Energy Co.</td>
<td>801 Travis St # 2100</td>
<td>Houston</td>
<td>TX</td>
<td>77002-5705</td>
</tr>
<tr>
<td>5</td>
<td>0.69323 John Marlborough</td>
<td>CEO</td>
<td>Alexander Howden Energy Inc</td>
<td>1990 Post Oak Blvd # 2100</td>
<td>Houston</td>
<td>TX</td>
<td>77027-3417</td>
</tr>
<tr>
<td>6</td>
<td>0.69261 K S Adams Jr</td>
<td>CEO</td>
<td>Adams Resources &amp; Energy Inc</td>
<td>4400 Post Oak Pkwy # 2700</td>
<td>Houston</td>
<td>TX</td>
<td>77002-4006</td>
</tr>
</tbody>
</table>
Inside P&G’s digital revolution
Spotfire and Procter & Gamble

- **Business Sphere**
  - Business decisions made with Spotfire
- **GBUs run their business on Spotfire**
- **Sales and Marketing**
  - Optimize marketing campaigns and spend – initiative tracking
  - Track retailers | product categories by traits, location...
  - Forecast share, volume, profit
- **Supply chain**
  - Keep shelves stocked; shipments
- **R&D**
  - Product development and testing
  - Product claims
Retailer Trends | Campaigns Mashup

POS - Category
POS - Category
POS - Brand
Panel - Category

Market - Promotions

Retailers / Region / Category => Drill on Share Drivers, Price, Media
Key Unknowns and Business Insights

- Which stores of which retailers are best at selling each category/brand?
- Is Kroger beating Target in selling laundry products in rural areas?
- Which retailers are doing well (or poorly) with what products in what stores/regions and why?
- What promotion/pricing campaigns drive share/value?
TIBCO Spotfire & Big Data

Hybrid In-Memory
- In-memory, high-performance

Data on Demand
- Replace data in memory on demand
- Drill-down out-of-memory

Statistical Services
- Server-side calculations
- R/S+, Matlab, SAS

Application Data Services
- Server-side computations in database
- Hadoop, Teradata, Netezza, Vertica, ...
- Data virtualization: updating cache
- Federation
- Aggregation

Sources:
NY Times: “The Age of Big Data” (Feb 11)
NY Times: “How Companies Learn Your Secrets” (Feb 16)
AREAS of FOCUS

Dimension-Free Data Exploration
- Visual
- Interactive
- No Constraints

Data Mashup
- Combine Data Sources
- No Scripting
- IT-Free

Predictive & Event Driven
- Data at Rest & Data In Motion
- Open Source & 3rd Party
- “Two Second Advantage”

Contextual Collaboration
- Bookmarks
- Guided Apps
- Portals & Social Platforms

Enterprise-Class
- Unmatched Performance
- Massive Scalability
- 24x7 Expertise
Product (Spotfire) Stack

Diagram showing the setup of TIBCO Spotfire Stack, including components like TIBCO Spotfire Developer, TIBCO Spotfire Professional, TIBCO Spotfire Enterprise Player, TIBCO Spotfire Web Player client, SMTP Server, Scheduler, TIBCO Spotfire Automation Services, TIBCO Spotfire Server cluster, Spotfire repository/Library DB, RDB/Data warehouse, Active Directory.
Integrating Statistics into Spotfire

Statisticians

Develop analytic

R/S+  Spotfire Miner  Matlab/SAS
Integrating Statistics into Spotfire

Statisticians
Develop analytic

Make analytic available to application authors

Spotfire Statistics Services
Spotfire Server

Deploy and register analytic
Integrating Statistics into Spotfire

Statistician

- Develop analytic

Spotfire Statistics Services
- Deploy and register analytic

Spotfire Server
- Configure Application to use analytic

Analysts and Application Authors

No coding required
Integrating Statistics into Spotfire

Managers, Consumers, Executives

One-click deployment of web applications

Centrally-managed application with governance managed by statistics / analytics team

Statistician

Develop analytic

Spotfire Statistics Services

Spotfire Server

Deploy and register analytic

Configure Application to use analytic

Analysts, Application Authors

TIBCO
SPOTFIRE VALUE DRIVERS

**Universal Adaptability**
Leverage a single analytics and data discovery platform to empower anyone, anywhere to make insightful decisions.

**Visibility Into the Unknown**
Discover unexpected insights hidden in Big Data and Real-Time Events to immediately identify strategic business opportunities or threats.

**Self-Service Discovery**
Freely explore data to any level of detail, radically accelerating decision making, while dramatically reducing dependence on IT.

**Fastest to Actionable Insight**
Instantly turn insight into action by enabling anyone to rapidly discover hidden insights and quickly collaborate in context.
Contact Information

Michael O’Connell, PhD
Sr. Director Analytics
TIBCO
moconnel@tibco.com
http://about.me/moconnell
If we get the **RIGHT INFORMATION**
to the **RIGHT PLACE**, at the **RIGHT TIME**
and put it in the **RIGHT CONTEXT**

then we can make the **WORLD A BETTER PLACE**.