The Unrealized Power of Data

Roger Barga, Azure Data Platform Group
BIG
DATA
Big Data is Here Today
Maturity Beckons...

http://bit.ly/5a9KL
Competing on Analytics

By 2014, 30 percent of analytic applications will use predictive capabilities.

– Gartner Business Intelligence Summit 2012

If you can predict it, you can own it...

• Forecasting
• Targeting
• Fraud detection, anti-fraud analytics
• Risk
• Customer churn, conversion
• Propensity
• Price Elasticity
Predictive Analytics in Action


How Target Figured Out A Teen Girl Was Pregnant [http://onforb.es/SokL3j](http://onforb.es/SokL3j)
Data Science

65% of enterprises feel they have a **strategic shortage of data scientists**, a role many did not even know existed 12 months ago...

Data Science is far too complex today

- Software Development
- Data Exploration
- Math & Statistics Knowledge
- Machine Learning
- Data Management

this **must get simpler**, it simply won’t scale!
Talk Outline

Predictive Analytics Today
Predictive analytics can be a potent weapon in your business analytics toolbox
With increasing commoditization it will truly be the next differentiator

Advancement of Data Science
Unique role and data science process;
Applying the principles of scientific experimentation to business;

Application of Predictive Analytics Within Microsoft
Demo our internal Azure service for machine learning over data to build predictive models;
Discuss three models built using Data Lab that we are using internally;
Predictive Analytics

“Trying to predict the future is like trying to drive down a country road at night with no lights while looking out the back window.” – Peter Drucker

- We must seek out patterns among seemingly unconnected data and disciplines to gain valuable insights.
- It took 50 years for organizations to fully leverage transactions systems, so it will time to realize significant value from knowledge, information and data.
- What gets measured, gets done.
Predictive Analytics

Successful Predictions

the signal and the noise
why most predictions fail but some don’t
nate silver
## Predictive Analytics

### Successful Predictions

<table>
<thead>
<tr>
<th>Date</th>
<th>Opponent</th>
<th>Score</th>
<th>Outcome</th>
<th>Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-13-2002</td>
<td>vs Toronto Blue Jays</td>
<td>5-4</td>
<td>W</td>
<td>69-51</td>
</tr>
<tr>
<td>08-14-2002</td>
<td>vs Toronto Blue Jays</td>
<td>4-2</td>
<td>W</td>
<td>70-51</td>
</tr>
<tr>
<td>08-16-2002</td>
<td>vs Chicago White Sox</td>
<td>1-0</td>
<td>W</td>
<td>71-51</td>
</tr>
<tr>
<td>08-17-2002</td>
<td>vs Chicago White Sox</td>
<td>9-2</td>
<td>W</td>
<td>72-51</td>
</tr>
<tr>
<td>08-18-2002</td>
<td>vs Chicago White Sox</td>
<td>7-4</td>
<td>W</td>
<td>73-51</td>
</tr>
<tr>
<td>08-19-2002</td>
<td>at Cleveland Indians</td>
<td>8-1</td>
<td>W</td>
<td>74-51</td>
</tr>
<tr>
<td>08-20-2002</td>
<td>at Cleveland Indians</td>
<td>6-3</td>
<td>W</td>
<td>75-51</td>
</tr>
<tr>
<td>08-21-2002</td>
<td>at Cleveland Indians</td>
<td>6-0</td>
<td>W</td>
<td>76-51</td>
</tr>
<tr>
<td>08-22-2002</td>
<td>at Cleveland Indians</td>
<td>9-3</td>
<td>W</td>
<td>77-51</td>
</tr>
<tr>
<td>08-23-2002</td>
<td>at Detroit Tigers</td>
<td>9-1</td>
<td>W</td>
<td>78-51</td>
</tr>
<tr>
<td>08-24-2002</td>
<td>at Detroit Tigers</td>
<td>12-3</td>
<td>W</td>
<td>79-51</td>
</tr>
<tr>
<td>08-25-2002</td>
<td>at Detroit Tigers</td>
<td>10-7</td>
<td>W</td>
<td>80-51</td>
</tr>
<tr>
<td>08-26-2002</td>
<td>at Kansas City Royals</td>
<td>6-3</td>
<td>W</td>
<td>81-51</td>
</tr>
<tr>
<td>08-27-2002</td>
<td>at Kansas City Royals</td>
<td>6-4</td>
<td>W</td>
<td>82-51</td>
</tr>
<tr>
<td>08-28-2002</td>
<td>at Kansas City Royals</td>
<td>7-1</td>
<td>W</td>
<td>83-51</td>
</tr>
<tr>
<td>08-30-2002</td>
<td>vs Minnesota Twins</td>
<td>4-2</td>
<td>W</td>
<td>84-51</td>
</tr>
<tr>
<td>08-31-2002</td>
<td>vs Minnesota Twins</td>
<td>6-3</td>
<td>W</td>
<td>85-51</td>
</tr>
<tr>
<td>09-01-2002</td>
<td>vs Minnesota Twins</td>
<td>7-5</td>
<td>W</td>
<td>86-51</td>
</tr>
<tr>
<td>09-02-2002</td>
<td>vs Kansas City Royals</td>
<td>7-6</td>
<td>W</td>
<td>87-51</td>
</tr>
<tr>
<td>09-04-2002</td>
<td>vs Kansas City Royals</td>
<td>12-11</td>
<td>W</td>
<td>88-51</td>
</tr>
</tbody>
</table>
Predictive Analytics

Customer discussions, from the front lines

By applying predictive analytics to data they already have, organizations are uncovering unexpected patterns and associations, and they are beginning to develop predictive models to guide front-line interactions.
Predictive Analytics
Customer discussions, from the front lines

By applying predictive analytics to data they already have, organizations are uncovering unexpected patterns and associations, and they are beginning to develop predictive models to guide front-line interactions.

Business Scenario:

• Downtime due to equipment failure or unnecessary maintenance bears huge cost;
• Goal: Predict failures from collected data and defer maintenance;

Five years of data, time series data from hundreds of sensors on machines throughout the gas field, for dozens of gas fields across different regions of the world...
Predictive Analytics
Customer discussions, from the front lines

By applying predictive analytics to data they already have, organizations are uncovering unexpected patterns and associations, and they are beginning to develop predictive models to guide front-line interactions.

Business Scenario:
• Downtime due to equipment failure or unnecessary maintenance bears huge cost;
• Goal: predict failures from collected data and defer maintenance;

Time series analysis to create features and evaluate various machine learning approaches...
Predictive Analytics
Customer discussions, from the front lines

By applying predictive analytics to data they already have, organizations are uncovering unexpected patterns and associations, and they are beginning to develop predictive models to guide front-line interactions.

Business Scenario:
• Downtime due to equipment failure or unnecessary maintenance bears huge cost;
• Goal: predict failures from collected data and defer maintenance;

One-Class SVM
Predictive Analytics
Customer discussions, from the front lines

By applying predictive analytics to data they already have, organizations are uncovering unexpected patterns and associations, and they are beginning to develop predictive models to guide front-line interactions.

- Oil & Gas customer, building models for predictive maintenance of equipment;
- Telecom customer, building customer churn predictive model to improve retention;

Enterprises are beginning to realize that data is a strategic asset and that they need to start treating it like one...

- If you don’t have or can’t find the data you need, consider creating it;
Predictive Analytics

Bridging the Gap

Developing predictive analytics must be simpler, today it requires highly specialized skills;

- Tools that their current staff and/or new hires can use for predictive modelling;
- Ability to integrate predictive models in line of business applications;
Developing predictive analytics must be simpler, today it requires highly specialized skills;

- Tools that their current staff and/or new hires can use for predictive modelling;
- Ability to integrate predictive models in line of business applications;

Enterprises require an end-to-end solution for their predictive modelling

- Support for collaboration, lineage tracking;
- Archive for predictive models, model governance;
- Support for search & discovery, reuse;
Predictive Analytics

Bridging the Gap

Developing predictive analytics must be simpler, today it requires highly specialized skills;
- Tools that their current staff and/or new hires can use for predictive modelling;
- Ability to integrate predictive models in line of business applications;

Enterprises require an end-to-end solution for their predictive modelling
- Support for collaboration, lineage tracking;
- Archive for predictive models, model governance;
- Support for search & discovery, reuse;

Predictions need to be operationalized, not locked in the backroom
- Deployable predictions, minimize time to insight;
- Ability to frequently update predictive model given new data, conditions change;
Predictive Analytics
Bridging the Gap

Familiar algorithms;
- Classification and Regression Trees (CART);
- Linear Regression;
- Decision Forests, Boosted Decision Trees, see Kaggle models of choice;
- Logistic Regression or other discrete choice;
- Association Rules;
- K-nearest Neighbors;
- Box Jenkins (ARIMA);
- Support Vector Machines;
- Deep Neural Networks, watch this one carefully...

More data & better features, in the right hands even a simple algorithm can produce a very effective predictive model – enter the data scientist...
Data Scientist
A Relatively New Job Title

DJ Patil and Jeff Hammerbacher

Alternative Job Titles

Signature Skills
- Math and Statistics Knowledge
- Machine Learning
- Programming Skills
- Computer Science
- Data Visualization
- Data Curiosity

Emergence of Data Science Teams
Data Science

Discipline with Deep Academic Roots

Bill Cleveland, Bell Labs
Data Science Action Plan, 2001

The Data Science Journal, 2002 (CODATA)

Journal of Data Science, 2002

National Science Foundation (NSF), 2005
Long-lived Digital Data Collections

JISC, 2008
The Skills, Role & Career Structure of Data Scientists & Curators
“We have found that creative data scientists can solve problems in every field better than experts in those fields can…”

“The expertise you need is strategy in answering these questions.”

Interview with Jeremy Howard, Chief Scientist, Kaggle: http://slate/me/UM0wQ7
Data Science Process

Mind the Gap, the space between the data set and the algorithm

Conventional descriptive analytics goes straight from a data set to applying an algorithm. But in Data Science there’s a huge space in between of very important stuff. It’s easy to run a piece of code that predicts or classifies. That’s not the hard part. The hard part is data wrangling and doing it well.

Handling missing values
Data Science Process

Mind the Gap, the space between the data set and the algorithm

Conventional descriptive analytics goes straight from a data set to applying an algorithm. But in Data Science there’s a huge space in between of very important stuff. It’s easy to run a piece of code that predicts or classifies. That’s not the hard part. The hard part is data wrangling and doing it well.

Handling missing values
Feature selection
Data Science Process
Mind the Gap, the space between the data set and the algorithm

Conventional descriptive analytics goes straight from a data set to applying an algorithm. But in Data Science there’s a huge space in between of very important stuff. It’s easy to run a piece of code that predicts or classifies. That’s not the hard part. The hard part is data wrangling and doing it well.

Handling missing values
Feature selection
Feature creation and imputation
Data Science Process

Mind the Gap, the space between the data set and the algorithm

Conventional descriptive analytics goes straight from a data set to applying an algorithm. But in Data Science there’s a huge space in between of very important stuff. It’s easy to run a piece of code that predicts or classifies. That’s not the hard part. The hard part is data wrangling and doing it well.

- Handling missing values
- Feature selection
- Feature creation and imputation
- Joining with external data to enrich training set
Data Science Process
Requires Agile Experimentation

1. Define the objective and quantify it with a metric – optionally with constraints, if any. This typically requires domain knowledge.

2. Collect and understand the data, deal with vagaries and biases in data acquisition (missing data, outliers due to errors in data collection process, more sophisticated biases due to the data collection procedure, etc.

3. Frame the problem in terms of a ML problem – classification, regression, ranking, clustering, forecasting, outlier detection etc. – combination of domain knowledge and ML knowledge is useful.

4. Transform raw data into a “modeling dataset”, with features, targets etc., which can be used for modeling. Feature construction often improved with domain knowledge. Target must be identical or proxy of metric in step 1.
Data Science Process
Requires Agile Experimentation

1. Train, test and evaluate, taking care to control bias/variance and ensure the metrics are reported with the right confidence intervals (cross-validation helps here), be vigilant against target leaks (which typically leads to unbelievably good test metrics) – this is the machine learning heavy step.
Data Science Process
Requires Agile Experimentation

1. Define Objective
2. Access and Understand the data
3. Pre-processing
4. Feature and/or Target construction
5. Train/Test split
6. Iterate steps (2) – (5) until the test metrics are satisfactory

- Feature selection
- Model training
- Model scoring
- Evaluation
Data Science Process
Support Algorithmic Creativity and Exploration

• Neural Networks
• Logistic Regression
• Support Vector Machine
• Decision Trees
• Ensemble Methods
• adaBoost
• Bayesian Networks

• Genetic Algorithms
• Random Forest
• Monte Carlo methods
• Principal Component Analysis
• Kalman Filter
• Evolutionary Fuzzy Modelling
• Deep Neural Networks
Data Science Process

Ability to Determine if a Model Will Predict With Accuracy...

Overfitting has many faces, always measure performance on out-of-sample data.

- At least one test, train on some of your data and test on the rest;
- Do several, splitting the data differently each time;

“Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”

George E.P. Box
Demonstration

Microsoft Data Lab

The place where you experiment on data.
Data Lab

So simple that anyone can use it, so powerful our data scientists want to use it...

**Machine learning** over data to build predictive model;

**Visual composition** using modules, no coding required;

**Modules** to author the end-to-end machine learning workflow from reading data, to training and validating the predictive model;

**Support for experimentation**, persist, rollback, retry to built the optimal model in the least amount of time;

**Support for collaboration**, from anywhere in the world jointly author a model from a web browser, with lineage tracking, isolation,…

A machine learning service on Window Azure to build predictive models from data…
Data Lab brings engineering best practices to data science...

Archive for predictive models, ensuring models are not lost, deleted, or corrupted.

Search, discovery and reuse existing models to build on the work of others;

Deploy predictive model into operation, from Data Lab to minimize time to insight;

Frequently update the predictive model, to adapt to changing business conditions.

Every new algorithm added as a module, every new predictive model deployed will flow through Data Lab to build up the knowledgebase and make Data Lab more valuable.
Data Lab
Analyze SQL Azure database traces, predict rack capacity for planning purposes

Map DW Data to Hive tables.


Run CapacityForPrediction (output used by R-module) and CalcPLEMachineZones.
Data Lab

Analyze SQL Azure database characterization model: bully, resource profiling, etc...
Data Lab

Protection Services (Windows Defender), reduce false positives in malware detection
Closing Comments
Emerging literacies and requirements....

Data is a core strategic asset, it encodes your business’ collective experience
• It is imperative to learn from your data, learn about your customers, products,...
• As your org learns from its collective experience it will discover strategic insights ...
...and can put this knowledge into action.

Predictive Analytics capitalizes on all available data
• Adds smarter decisions to systems, drives better outcomes, delivers greater value...
• It requires specialized expertise, talent, and tools to execute well.

Rise of the Data Scientist
• Applying the principles of scientific experimentation to business processes;
• Adept at interpretation & use of numeric data, appreciation for quantitative methods

Better tooling for data science productivity and end-to-end management...
Win a Microsoft Surface Pro!

Complete an online SESSION EVALUATION to be entered into the draw.

Draw closes April 12, 11:59pm CT
Winners will be announced on the PASS BA Conference website and on Twitter.

Go to passbaconference.com/evals or follow the QR code link displayed on session signage throughout the conference venue.

Your feedback is important and valuable. All feedback will be used to improve and select sessions for future events.
Thank you!