Analytics Role

Jean-François Puget
@JFPuget
Big Data = All Data
_Not just about large volume_

**Volume**
Data at Scale
Terabytes to petabytes of data

**Variety**
Data in Many Forms
RDBMs, objects, free text, multimedia, sensors

**Velocity**
Data in Motion
Analysis of streaming data to enable decisions within fractions of a second.
Video, Social media feeds, Sensor feeds, etc

**Veracity**
Data Uncertainty
Managing the reliability and predictability of inherently imprecise data types.
Measured data, predicted data, etc.
Purpose of Big Data Analytics: Creates New Value from Actionable Insight

- New Mix of Data
- Broader Application
- New Buyers & Decision Types

Data Acquisition | Context | Analyze/Embed | Actionable Insights in the Business Moment

Data
- Traditional data warehouse, transactions, descriptive
- Distributed datamarts, spreadsheets

Big Data
- Unstructured notes, logs
- Social Media pulse, emerging issues
- Survey Research attitudes, opinions
- Sense-making
- Sensors

Data
- Volume
- Velocity
- Variety
- Veracity

Organize
- Model
- Simplify

Analytics

-Decisions
- "consumer oriented agile insight"

Action
- "in the business moment"

Differentiated analytic solutions
- people, process

LEARN

Plan Collaborate

Automation Embed
## Actionable insights examples

<table>
<thead>
<tr>
<th>Data</th>
<th>Insight</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer name, address, past responses</td>
<td>Customer segment</td>
<td>Marketing campaigns</td>
</tr>
<tr>
<td>Customer service requests, customer transactions</td>
<td>Propensity to churn, customer value</td>
<td>Retention action decision</td>
</tr>
<tr>
<td>+ Twitter</td>
<td>Life Event detection</td>
<td>Targeted ads and offers</td>
</tr>
<tr>
<td>Twitter</td>
<td>Demonstration location</td>
<td>Police positioning</td>
</tr>
<tr>
<td>Weather forecast</td>
<td>Energy consumption forecasts</td>
<td>Energy production plan</td>
</tr>
<tr>
<td>Weather forecast</td>
<td>Outage prediction</td>
<td>Crew and resource dispatch</td>
</tr>
<tr>
<td>Weather forecast</td>
<td>Asthma risk score</td>
<td>Patient alerts, ER staffing</td>
</tr>
<tr>
<td>POS data</td>
<td>Category Sales performance</td>
<td>Pricing actions</td>
</tr>
<tr>
<td>POS data + Weather history</td>
<td>Weather adjusted sales performance</td>
<td>Inventory repositioning</td>
</tr>
<tr>
<td>Weather history + Ag yield</td>
<td>Variability and Correlation</td>
<td>Price Crop Insurance</td>
</tr>
<tr>
<td>Equipment sensor data + weather</td>
<td>Failure prediction</td>
<td>Maintenance schedule</td>
</tr>
<tr>
<td>Weather history + Ag yield</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Purpose of Big Data Analytics: Creates New Value from Actionable Insight

New Mix of Data / Broader Application / New Buyers & Decision Types

<table>
<thead>
<tr>
<th>Data Acquisition</th>
<th>Context</th>
<th>Analyze/Embed</th>
<th>Actionable Insights in the Business Moment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traditional data warehouse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organize</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DESCIIONS</td>
<td></td>
<td></td>
<td>consumer oriented agile insight</td>
</tr>
<tr>
<td>&quot;moment of decision&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Input</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Descriptive Analytics**
What has happened?

**Diagnostic Analytics**
Why did it happen?

**Predictive Analytics**
What will happen?

**Prescriptive Analytics**
How can we make it happen?
Irregular Operation Recovery

“... the controlled airspace of many European countries was closed to instrument flight rules traffic, resulting in the largest air-traffic shut-down since World War II. The closures caused millions of passengers to be stranded not only in Europe, but across the world...” Wikipedia

- Automated mass rebooking recommendations
- Customer preferences and priorities
- Alternate routes
- Clear directions for customers
Real time optimization

- Engineering of systems of engagement requires real time decision making
  - We want to provide the best possible action at each interaction
  - Sub second is good enough

Is the answer to make faster solvers?
  Online optimization

Not so sure.
Example: Taxi Dispatch (real customer example, simplified here)

Real time dispatch

<table>
<thead>
<tr>
<th>Time</th>
<th>Car A</th>
<th>Client 1</th>
<th>Car B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client 2</td>
<td>Car A</td>
<td>Client 1</td>
<td>Car B</td>
</tr>
<tr>
<td>Client 2</td>
<td>Car A</td>
<td>Client 1</td>
<td>Car B</td>
</tr>
</tbody>
</table>

Cars are assigned when customers call
For instance, closest car is selected
The Taxi company waits a bit before assigning cars to customers

Without optimization

<table>
<thead>
<tr>
<th>Time</th>
<th>Car A</th>
<th>Client 1</th>
<th>Car B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Car A</th>
<th>Client 1</th>
<th>Car B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

With optimization

<table>
<thead>
<tr>
<th>Time</th>
<th>Car A</th>
<th>Client 1</th>
<th>Car B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Car A</th>
<th>Client 1</th>
<th>Car B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Making decisions one at a time leads to a myopic effect

Gathering data and constraints to understand the ‘big picture’ creates the opportunity for better decisions

Making several decisions at the same time achieves better results
Delay or not?

- Delaying response can be good
  - Too quick an answer: we get suboptimal resource allocation
  - Too long an answer: we do not deliver a usable service

When delay is not possible?

- Pre position vehicles so that assigning the closest one yields good resource allocation on average

- Two steps
  - Predictive analytics: analyze history to predict demand
  - Compute optimal vehicle positions (set covering problem)

- We did this for ambulances at Ottawa
- We did this for police patrols
Emergency vehicle pre positioning

Training (offline)

- Emergency events history
- Incident Trends

Online

- Weather, time of week
- Available vehicles

ML, Stats

Incident Trends

Emergencies forecast

Optimization

Vehicle positions
Optimization at scale

- Optimization problems get larger and larger

- Two main drivers

- Traditional optimization problems, e.g. supply chain
  - Companies are integrating
  - Companies want to optimize the overall supply chain, directly using point of sale data to drive inventory optimization along the chain, and the manufacturing planning.

- Machine Learning
  - Most ML algorithms are optimization problem: find the model that best fits training data
  - We begin to see ML folks using mathematical programming techniques
    - Boyd
    - Bertsimas
Optimization at scale

- Optimization algorithms **scale up**
  - Larger machines
  - Shared memory parallel algorithms (multi threading)

- Limits of scaling up
  - Sending data to a central compute machine introduces latency
  - Memory can become a bottleneck
  - Cost of machine goes up quickly

- Big data algorithms **scale out**
  - Leverage large number of commodity hardware
  - Move computation to where data is in a reliable way
    - Duplicate storage
    - Node failure resilience

- Hadoop/MapReduce, Spark
  - Store data 3 times
  - Maps computation to data
  - Reduce (aggregates) results in a meaningful way
  - Needs a scheduler
Optimization of big data jobs

Engineering Resource Management Middleware for Optimizing the Performance of Clouds Processing MapReduce Jobs with Deadlines
Lim et al, 2015
Scaling out optimization algorithms?

- Distributing the search space: each worker gets a piece of it
  - Distributed MIP
    - CPLEX (2013) and Gurobi (2014) distributed mip solvers
  - Regin et al, Embarrassingly Parallel Search, 2013
  - Matteo Fischetti, Michele Monaci, and Domenico Salvagnin, Self Split, 2013

- The above assumes the problem is duplicated. Can we split the problem data as well?
  - Convex Optimization problems can be partitioned via ADMM (Boyd)
  - Decomposition methods, eg Benders
    - (Nilsen 97)
Price Optimization

Training (offline)
- Past sales
- Past prices
- ... (ML, Stats)
  - Price elasticity

Online
- Current prices
  - Predicted sales
  - Optimization
  - New prices
Optimization using predicted data

Training (offline)

- History (big data)
  - ML, Stats
  - Predictive model

Online

- Predictive model
  - Predicted data (Small)
  - Current state (Small)
  - Optimization
    - Decisions

Veracity

This is uncertain data
Stable decisions, stable profits

- **Examples**
  - Supply chain planning for a motorcycle vendor
    - 2% increase in profits vs. deterministic optimization
  - Inventory optimization for IBM Microelectronics Division
    - Greater than 7x increase in feasibility vs. deterministic optimization

- **Case studies**
  - Energy cost minimization for Cork County Council
    - Estimated 30% value-add in cost reduction vs. deterministic optimization
  - Leakage reduction for Dublin City Council
    - Estimated 10 times increased stability vs. deterministic optimization

- **Other benefits**
  - Automated toolkit reduces dependence on PhD-level experts & statistical data
  - Visualize trade-off between multiple KPIs across multiple scenarios and plans
On Feb. 14, 2011, the Watson computer made history!
Watson analyzes the human genome to battle brain cancer

a cancer mutation shown on a cell protein pathway from genome sequencing.

The New York Genome Center and IBM are partnering in a first-of-a-kind program to accelerate the race to personalized, life-saving treatment for cancer patients.
Optimization and Big Data: Lots of opportunities!

**Volume**
Optimization at Scale
Distributed computing
Decomposition methods

**Variety**
Data in Many Forms
Watson

**Velocity**
Data in Motion
Online optimization using predicted data

**Veracity**
Data Uncertainty
Robust optimization
Stochastic Programming
Beyond Smarter Planet: Smarter Comet

- Philae operations were scheduled using IBM Constraint Programming technology
  - Key bottleneck is data transmission
  - 25 minutes
  - Limited storage on Philae
- ESA/CNES had to quickly adjust plan because the robot landed in a tilted position
- CP was used to check the feasibility of adjusted plans