Big Data & Analytics

Concepts, technologies and the IBM perspective

Alberto Ortiz
Big Data & Analytics Architect

January 2015

What is Big Data?

Definition of Big Data

In short, the term Big Data applies to information that can't be processed or analyzed using traditional processes or tools.

Zikopoulos, Paul, Chris Eaton, Tom Deutsch, Dirk Deroos, and George Lapis. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York: McGraw-Hill, 2012. 3. Print.

Big data is an all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process them using traditional data processing applications.

"Big Data." Wikipedia. Wikimedia Foundation, n.d. Web. 05 Jan. 2015.

... data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

"Bringing Big Data to the Enterprise." IBM. N.p., n.d. Web. 07 Jan. 2015.

- It's not a specific type of software
- It's not a product
- It's not a specific amount of data
- It's not just social media

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 BILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

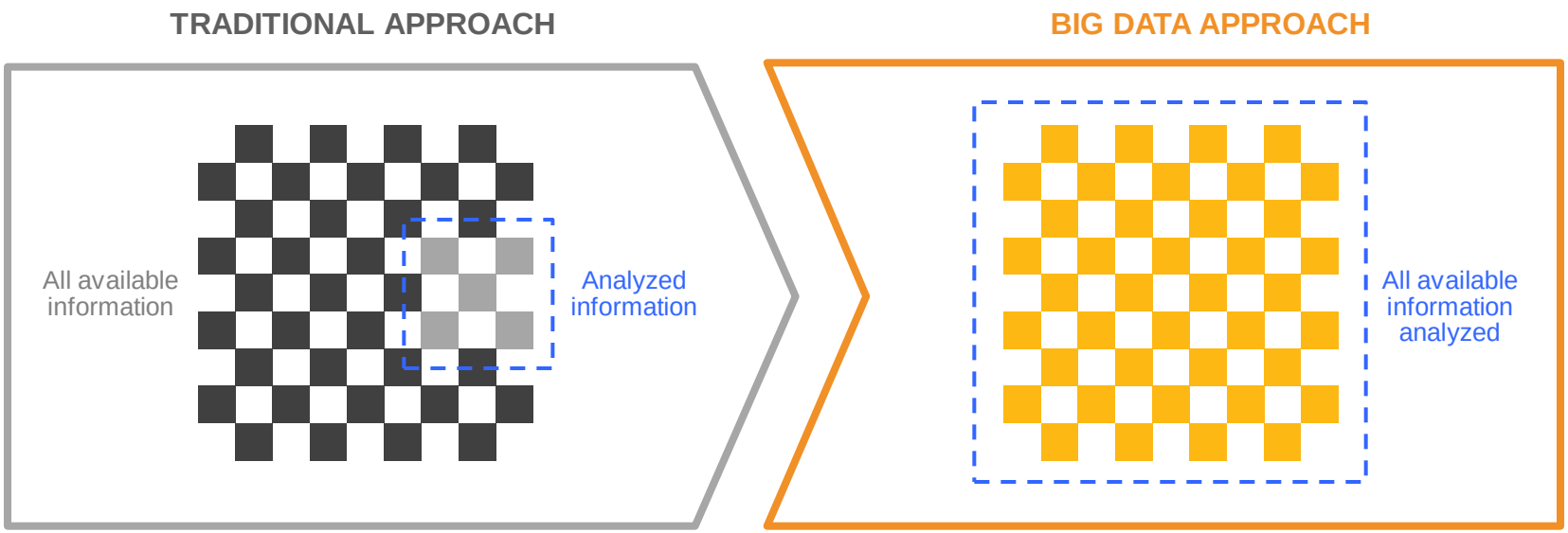
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

What are the characteristics of Big Data?

Velocity	Volume	Variety	Veracity
Sprint	**Vestas**	**Airbus**	**Italian Bank**
53M subscribers	2.6 Peta Bytes	Technical aircraft information	6th bank in Italy
10B transactions/day	Weeks for placement analysis	Information in CRMs, data bases, ERPs, ECMs, files, etc.	MDM+Social Media+Emails

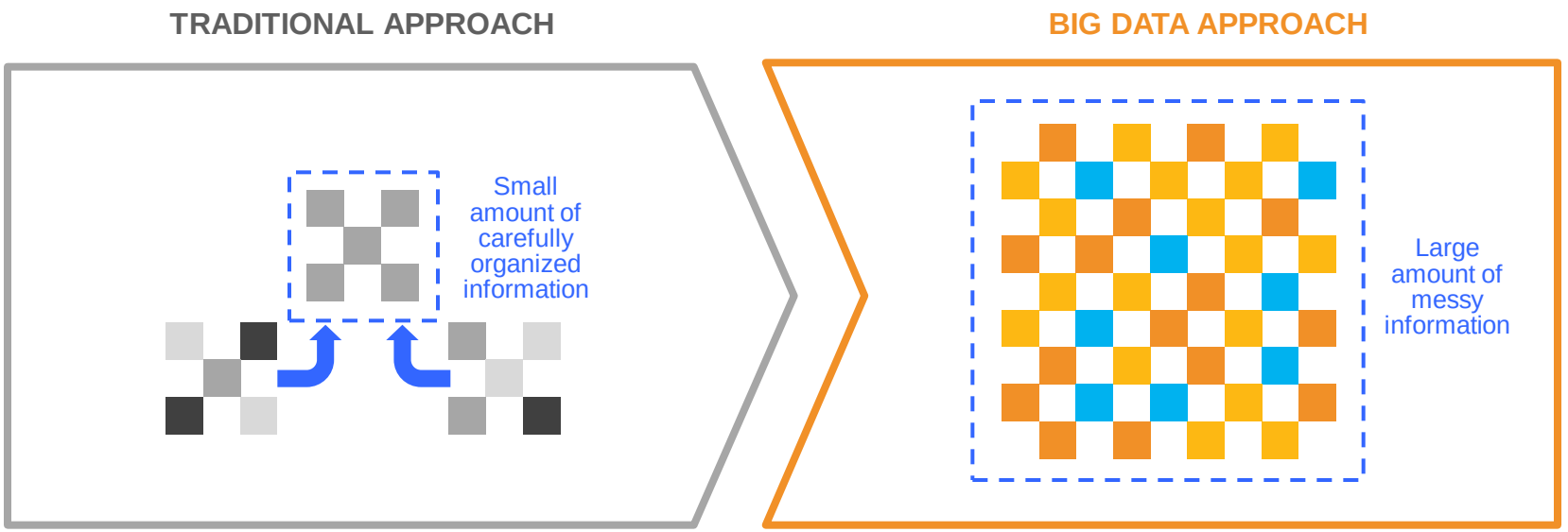
Leverage more of the data being captured



Analyze small subsets of information

Analyze ***all*** information

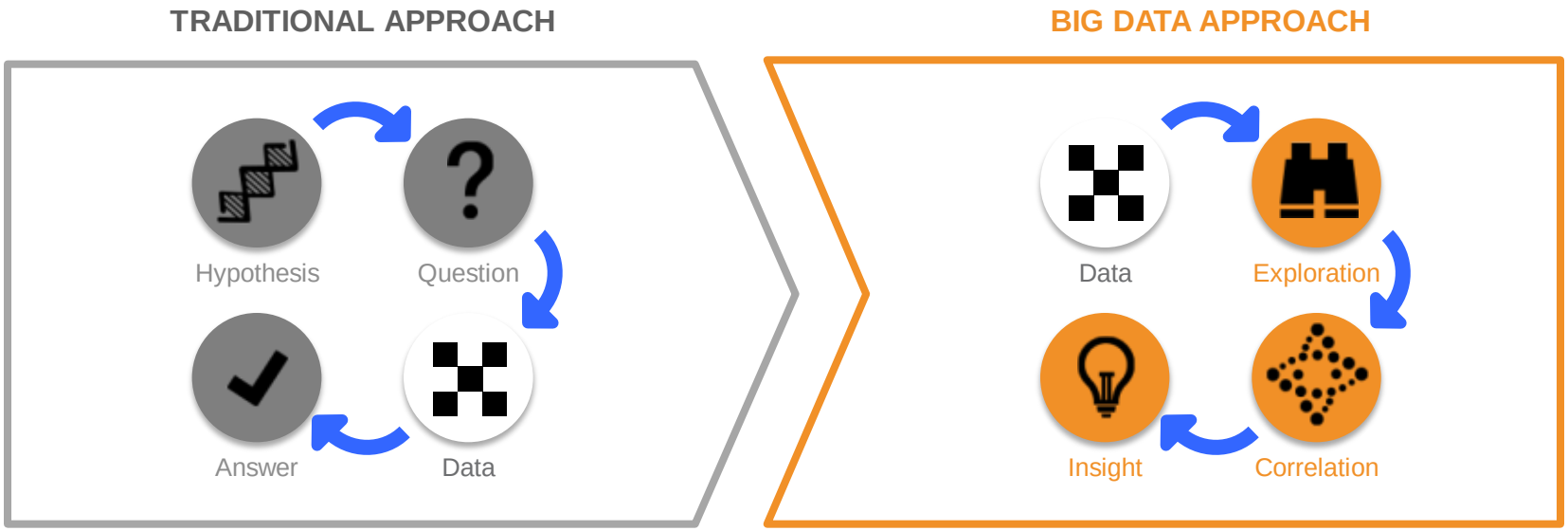
Reduce effort required to leverage data



Carefully cleanse information ***before*** any analysis

Analyze information as is, cleanse as needed

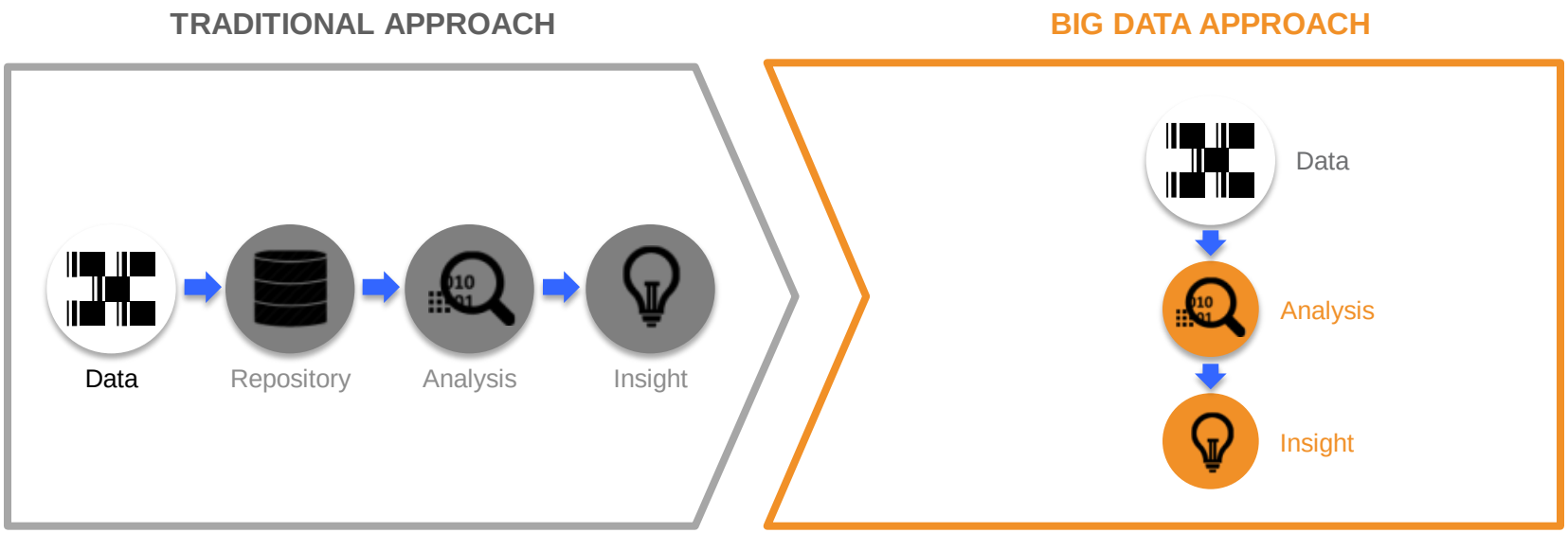
Data leads the way—and sometimes correlations are good enough



Start with hypothesis and test against selected data

Explore ***all*** data and identify correlations

Leverage data as it is captured



Analyze data ***after*** it's been processed and landed in a warehouse or mart

Analyze data ***in motion*** as it's generated, in real-time

What is analytics?

an·a·lyt·ics

/ˌanəˈlidiks/

noun

the systematic computational analysis of data or statistics.
"content analytics is relevant in many industries"

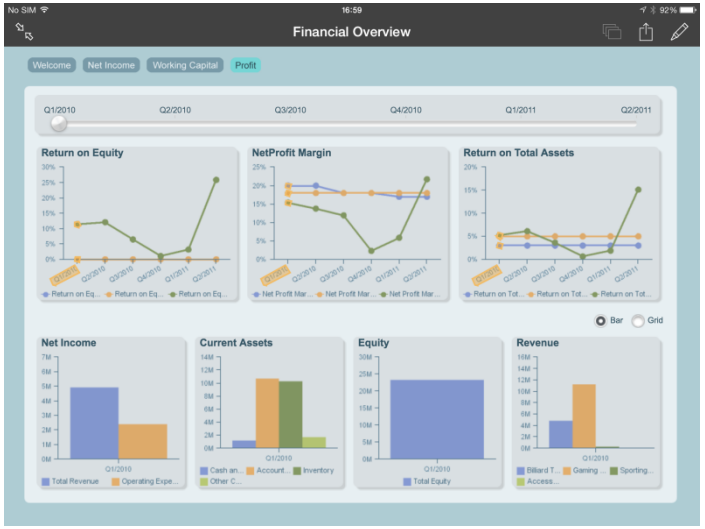
- information resulting from the systematic analysis of data or statistics.
  "these analytics can help you decide if it's time to deliver content in different ways"

https://www.google.com/search?q=analytics+definition

Source: http://d3js.org/

Why did it happen?

Descriptive analytics



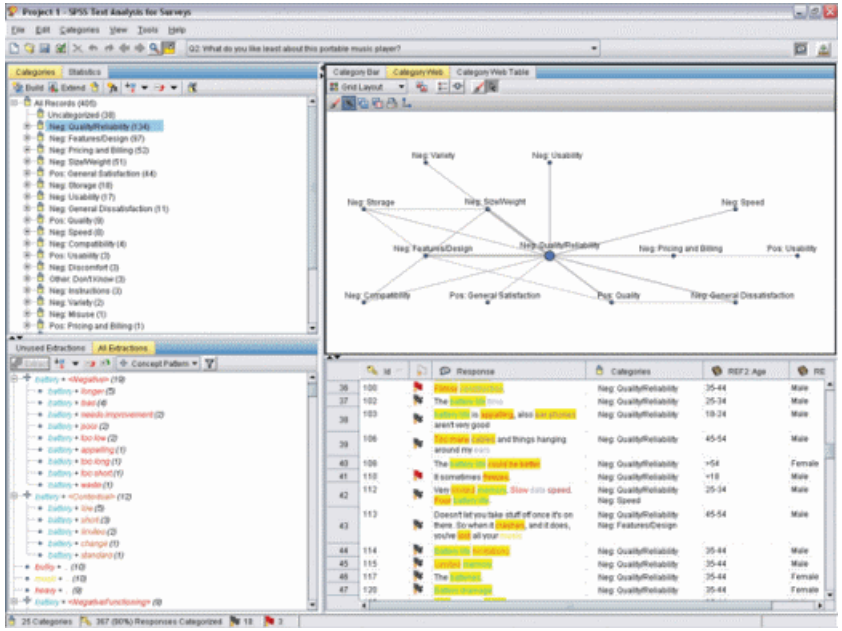
What is happening?

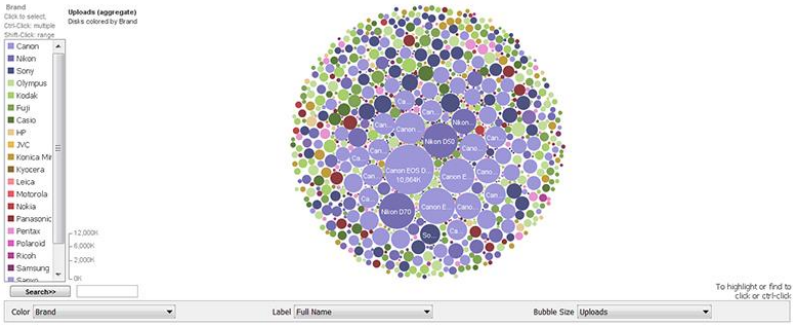
Text analytics

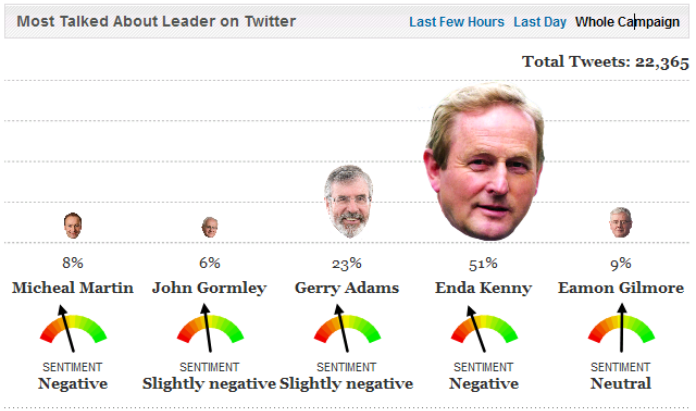
Sentiment analysis

Entity extraction

Natural language processing

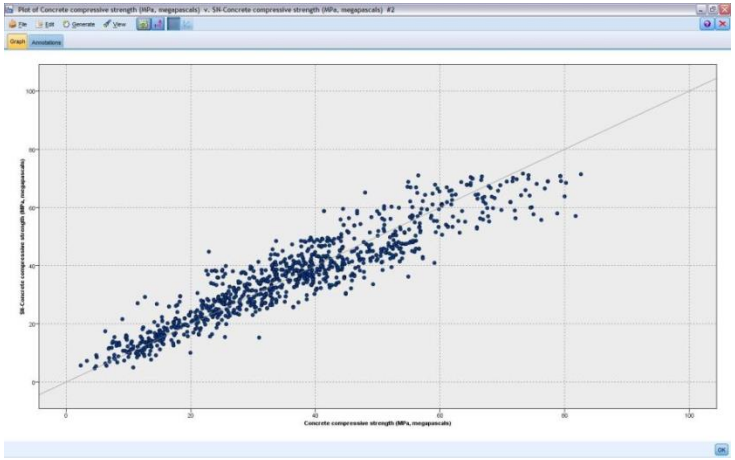


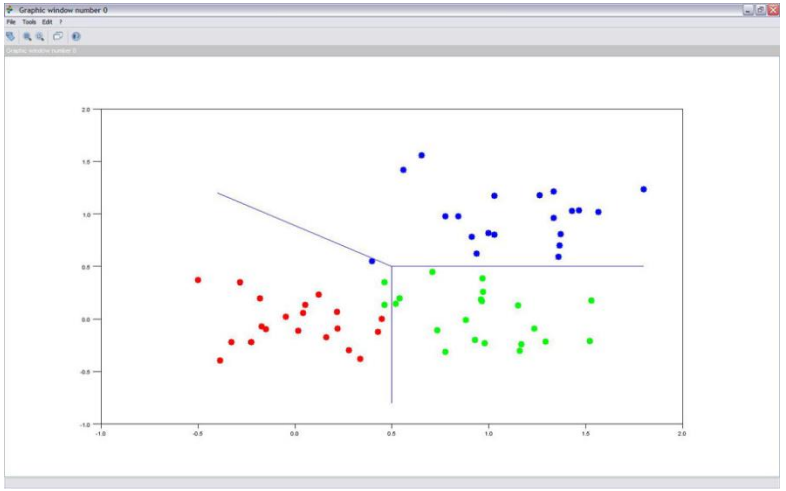


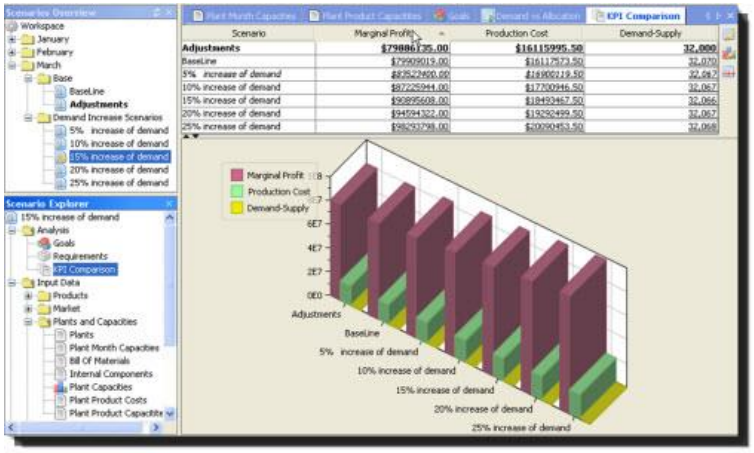


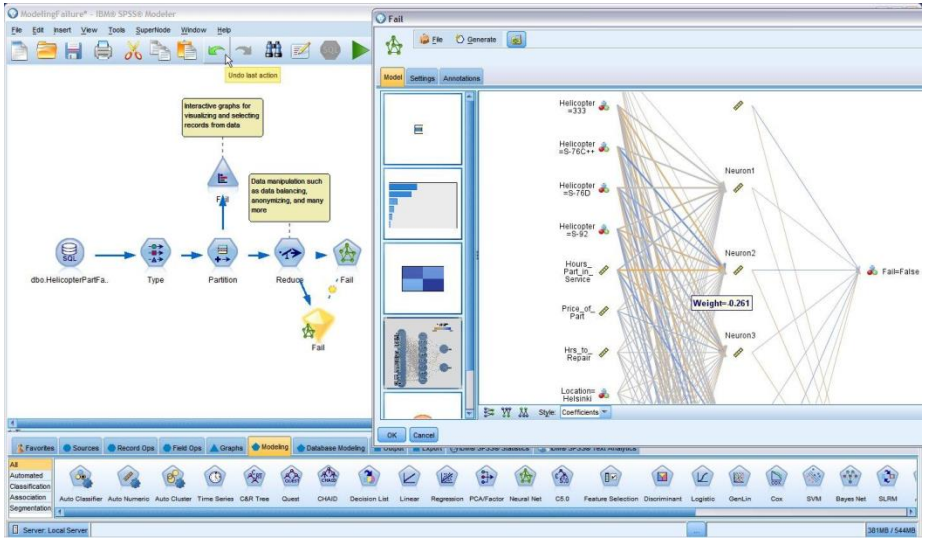
What could happen?

Predictive analytics



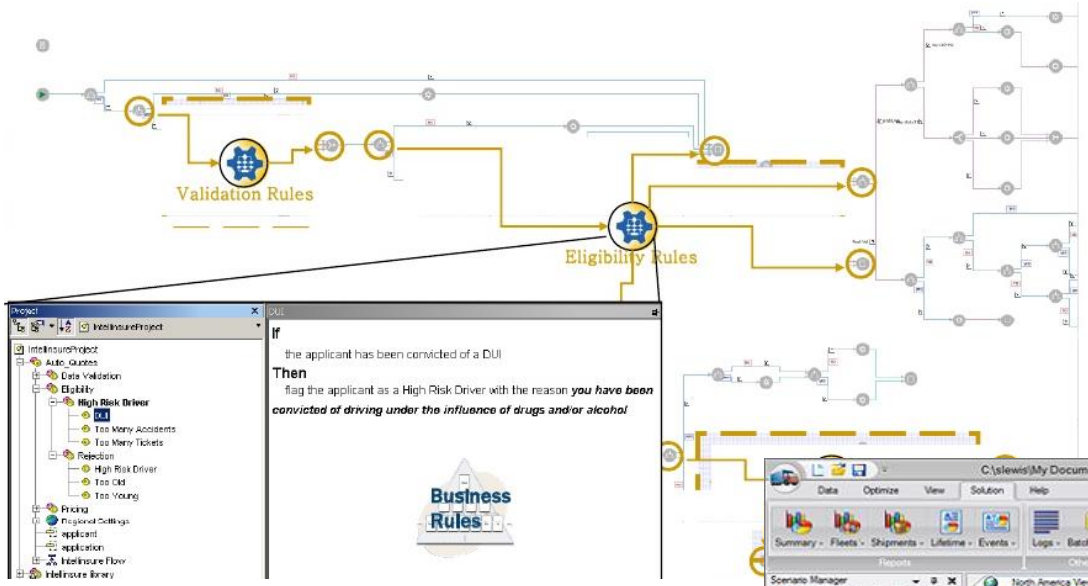






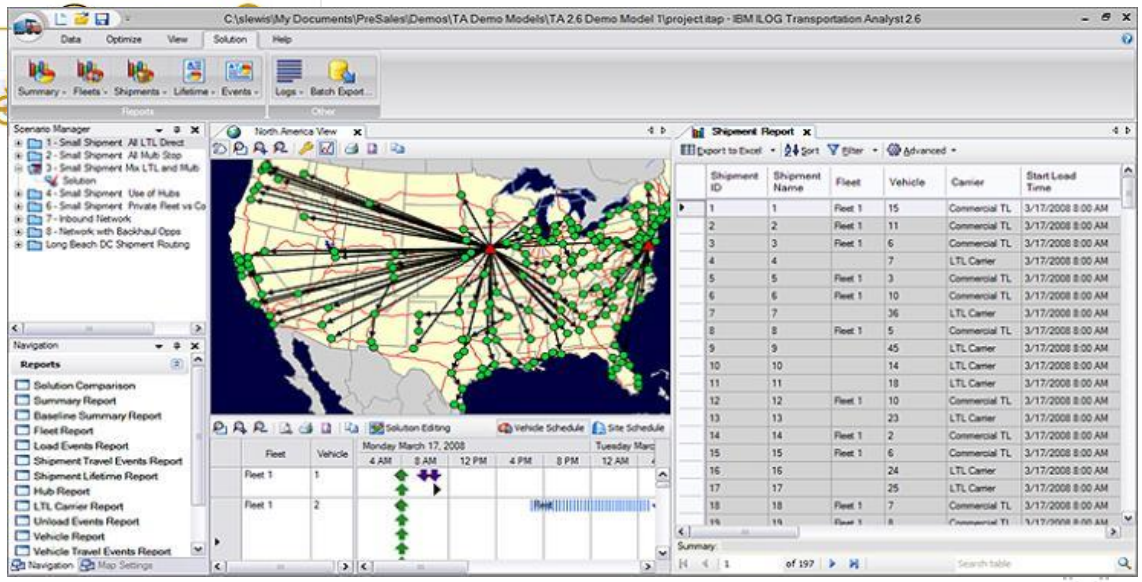
What should I do?

Prescriptive analytics



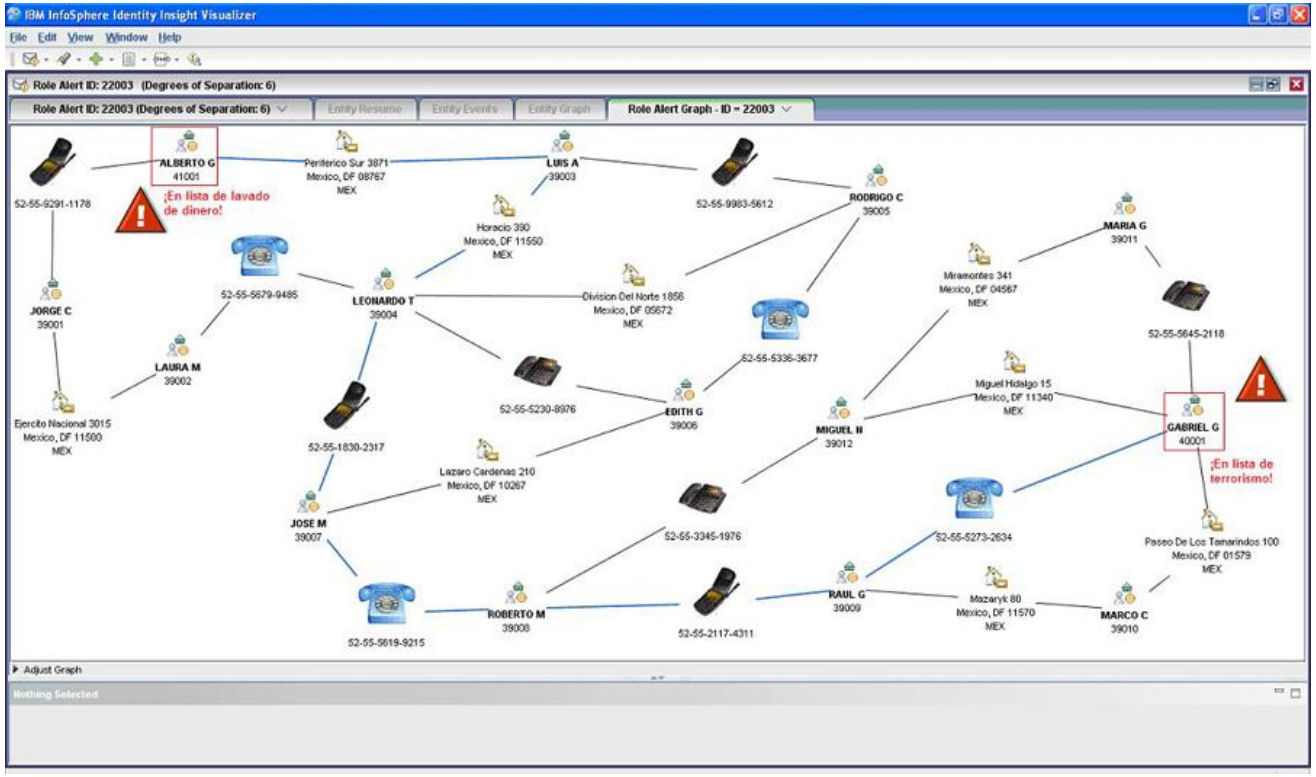
Business rules

Optimization



Who is who? Who knows who?

Entity analytics



IBM solutions for BD&A

Watson: state of the art in BD&A

What is Watson?

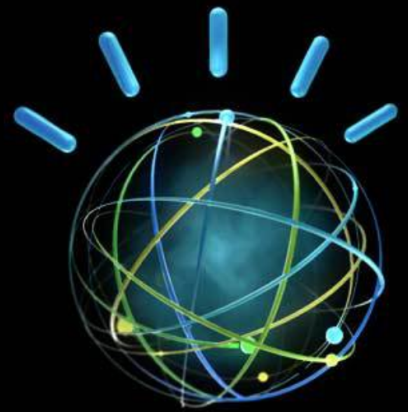
As a cognitive technology, Watson is a natural extension of what humans can do at their best.

Natural Language

Watson can read and understand natural language, important in analyzing unstructured data that make up as much as 80 percent of data today.

Hypothesis Generation

When asked a question, Watson relies on hypothesis generation and evaluation to rapidly parse relevant evidence and evaluate responses from disparate data.



Dynamic Learning

Through repeated use, Watson literally gets smarter by tracking feedback from its users and learning from both successes and failures.

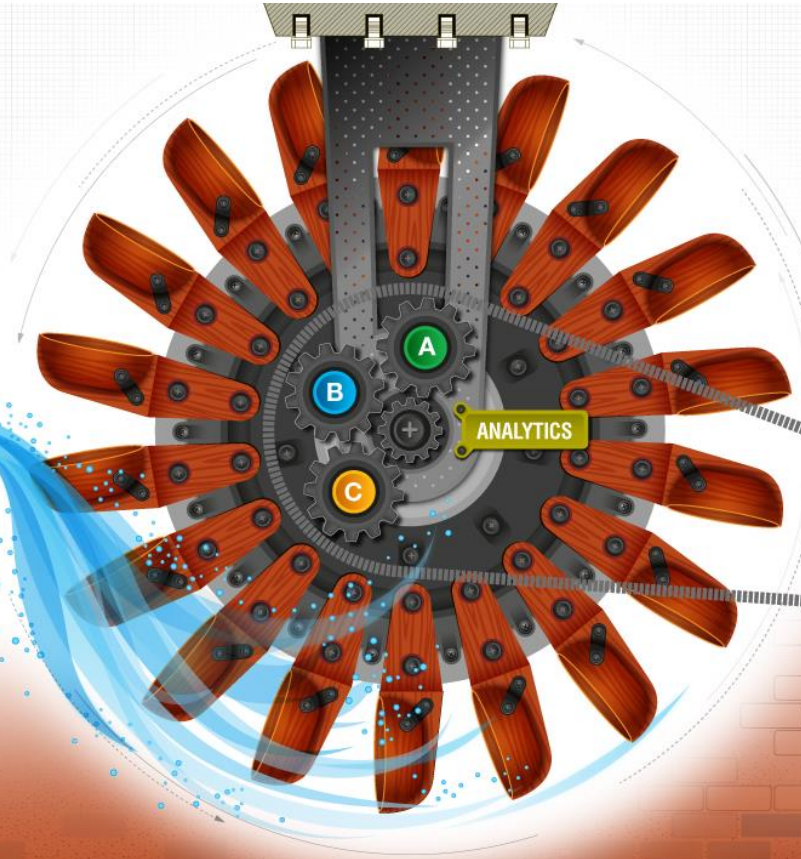
Watson Video

https://www.youtube.com/watch?v=P18EdAKuC1U#t=35

BIG DATA & ANALYTICS:

Transforming potential into action

By leveraging the natural fusion of big data and analytics, successful organizations can better transform unrefined information into new, actionable business insights. With these insights, companies have the power to embed and automate intelligent decisions into operational systems, driving substantial business outcomes and creating new opportunities for competitive advantage.



ORGANIZATIONAL COMPONENTS

PART:	DESCRIPTION:
A	**INFUSE ANALYTICS EVERYWHERE** across your organization, creating a culture that encourages employees to make big data central to business processes
B	**BE PROACTIVE ABOUT PRIVACY, SECURITY AND GOVERNANCE** to protect your customers, safeguard your data and foster trust in your company
C	**INVEST IN WATSON FOUNDATIONS AND IN BIG DATA AND ANALYTICS** to nimbly analyze and act on data in motion, while building out your master plan

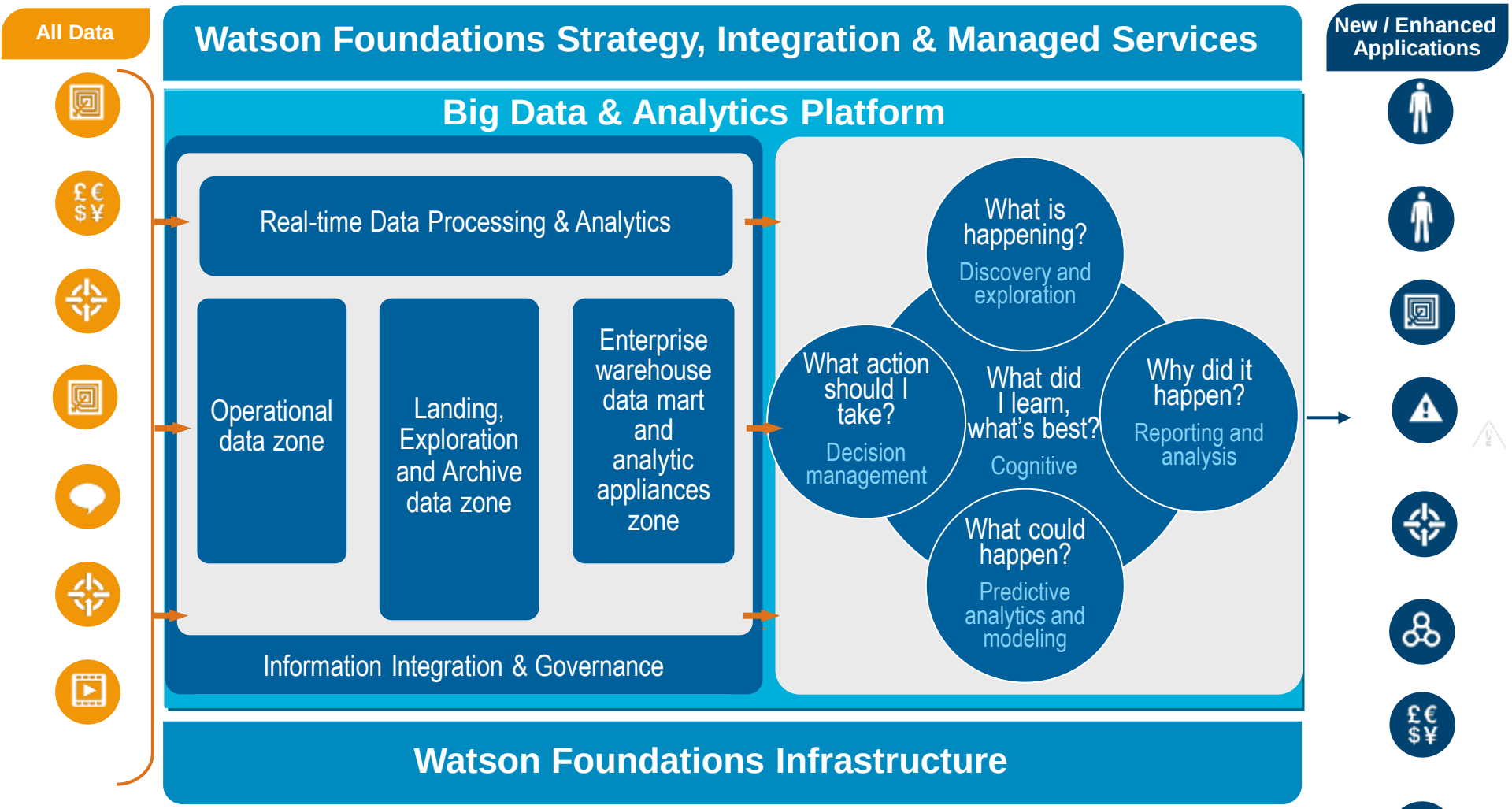
Companies harnessing the full power of Watson Foundations and of big data and analytics achieve results that can transform nearly every aspect of their business.

1. **ACQUIRE, GROW AND RETAIN CUSTOMERS**
   - **USD1 MILLION SAVED** when XO Communications enhanced its customer retention rates
2. **CREATE NEW BUSINESS MODELS**
   - **80% REDUCTION IN REVENUE LEAKAGE** when the Land Transport Authority of Singapore integrated transportation
3. **TRANSFORM MANAGEMENT PROCESSES**
   - **96% REDUCTION IN PROCESS TIME** when ANCAP gained a real-time view of its business
4. **MANAGE RISK**
   - **99% ACCELERATION IN REPORTING** when the Allianz Group built a centralized risk management platform
5. **OPTIMIZE OPERATIONS: COUNTER FRAUD AND THREATS**
   - **USD2.5 MILLION SAVED IN PAYOUTS** when Santam identified fraudulent customers
6. **MAXIMIZE INSIGHT, ENSURE TRUST AND IMPROVE IT ECONOMICS**
   - **3.5 BILLION EMAILS ANALYZED** when Constant Contact determined best practices for maximum response

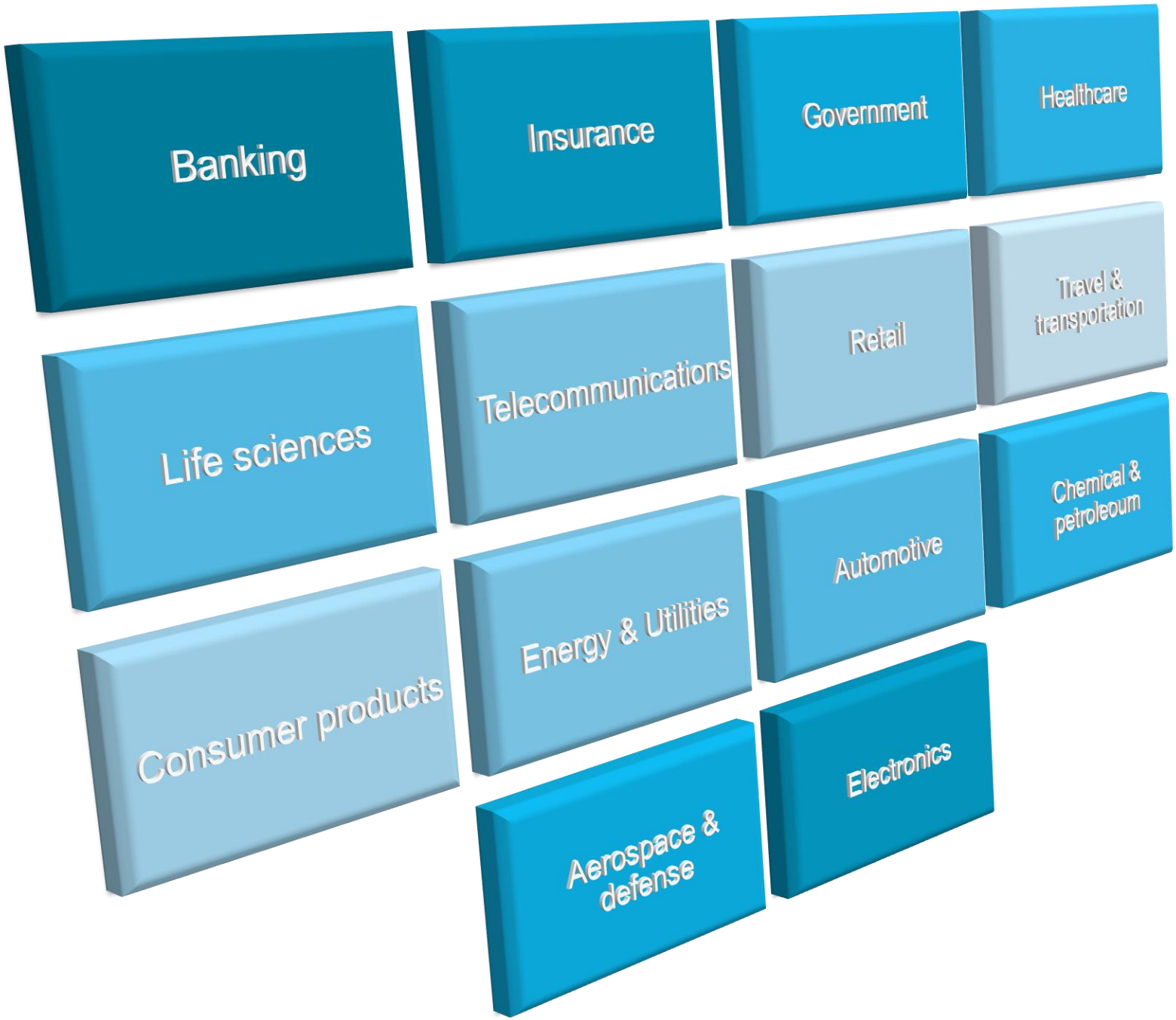
IBM, the IBM logo, ibm.com, and the planet icon are trademarks of IBM Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at ibm.com/legal/copytrade.shtml. © Copyright IBM Corporation 2014

Ready to get more value from your data?
Learn more about the transformative effect of IBM Watson Foundations and of big data and analytics solutions at **ibm.com**/bigdata-and-analytics

Watson Foundations



Industry cases using BD&A


Banking
Insurance
Government
Healthcare
Life sciences
Telecommunications
Retail
Travel & transportation
Consumer products
Energy & Utilities
Automotive
Chemical & petroleum
Aerospace & defense
Electronics

Banking

Big Data & Analytics Customer Case Studies

NYSE Euronext enhances trading insights and reporting with IBM PureData platform

Need

- Rapid analysis of trade data to facilitate compliance and regulatory reporting
- New functionality needed to track the value of a listed company, performing trend analysis, and searching for evidence of fraudulent activity
- The previous system trawled through large amounts of irrelevant information to complete searches

Benefits

- Reduced time to run market surveillance by 99%
- Improved ability to detect suspicious trading activity
- Enabled early investigative action, minimizing damage to investing public

NYSE Euronext

http://www.youtube.com/watch?v=J7KwuGFF9bU

Government

Big Data & Analytics Customer Case Studies

Real-Time Security

TerraEchos provides security and intelligence expansion with real-time analysis of streaming data

Need

- Technology solution that would detect, classify, locate and track potential threats to secure its perimeters and border areas

Benefits

- Only vendor providing a Big Data platform for real-time analysis from streaming data for security
- Unprecedented real-time facility surveillance with the ability to process 1.6GB of data per second for insight into any event
- Confidently identify and classify a potential security threat -the system is sensitive enough to distinguish between a bobcat and a bad guy

Retail

Big Data & Analytics Customer Case Studies

Constant Contact gains marketing insight from behavior pattern analysis

Need

- Improve their business value proposition and differentiate themselves against other companies that provide email marketing services

Benefits

- IBM is only vendor providing big data platform with integrated text analytics capability
- Constant Contact's combined analysis of time, date, recipient email addresses and email content resulted in up to 25% increase in email opening rate.

Constant Contact®

Life Sciences

Big Data & Analytics Customer Case Studies

North American Pharmaceutical leader achieves greater transparency, agility and productivity in R&D through secure access to huge volumes of valuable enterprise content

Need

- Executives wanted to transform information and data sharing capabilities by providing access to huge volumes of customer, patient, and research data stored in systems around the world

Benefits

- Improved R&D efficiency by 90 percent and reduced search time by 50 percent, saving millions of dollars in the first year alone
- Reduced new staffing requirements by 1.2 percent, saving USD13.4 million yearly
- Improved sales productivity to increase revenue by 1.4 percent

Thank you