IBM Business Analytics

IBM Software Group

# Big Data – Little Impact?

## Raising the bar for "Data Science"

Alexander Lang (alexlang@de.ibm.com)

## Let's set the record straight

- **Big Data is irrelevant**
- **Big Data Analytics is relevant**
- **Is the "Big" in "Big Data Analytics" *really* relevant?**

# Astron – Uses streaming analytics to deliver insights from the world's largest radio telescope

**99% faster identification** of relevant data and images, making information available to astronomers in minutes rather than several days

**Analyzes >1 exabyte** of data daily — twice the amount generated by global daily Internet traffic

**Integrates data from >3,000** dishes and antennas to form the largest and fastest radio telescope in the world

The transformation: Streaming analytics analyzes huge volumes of data-in-motion to gain insight from the world's largest Telescope.

**Solution components**

**Software**

- IBM® InfoSphere® Streams
- IBM SPSS® Modeler

ASTRON Group

# Vestas – Turns climate into capital with Big data

**97% decrease**
in response times for wind forecasting information

**Cuts cost per kilowatt hour from wind energy**
increasing customer's return on investment

**40% reduction**
in energy consumption, reducing IT footprint while increasing power

Solution components

Software
- IBM® InfoSphere® BigInsights™ Enterprise Edition

Hardware
- IBM System x® iDataPlex® dx360 M3
- IBM System Storage® DS5300

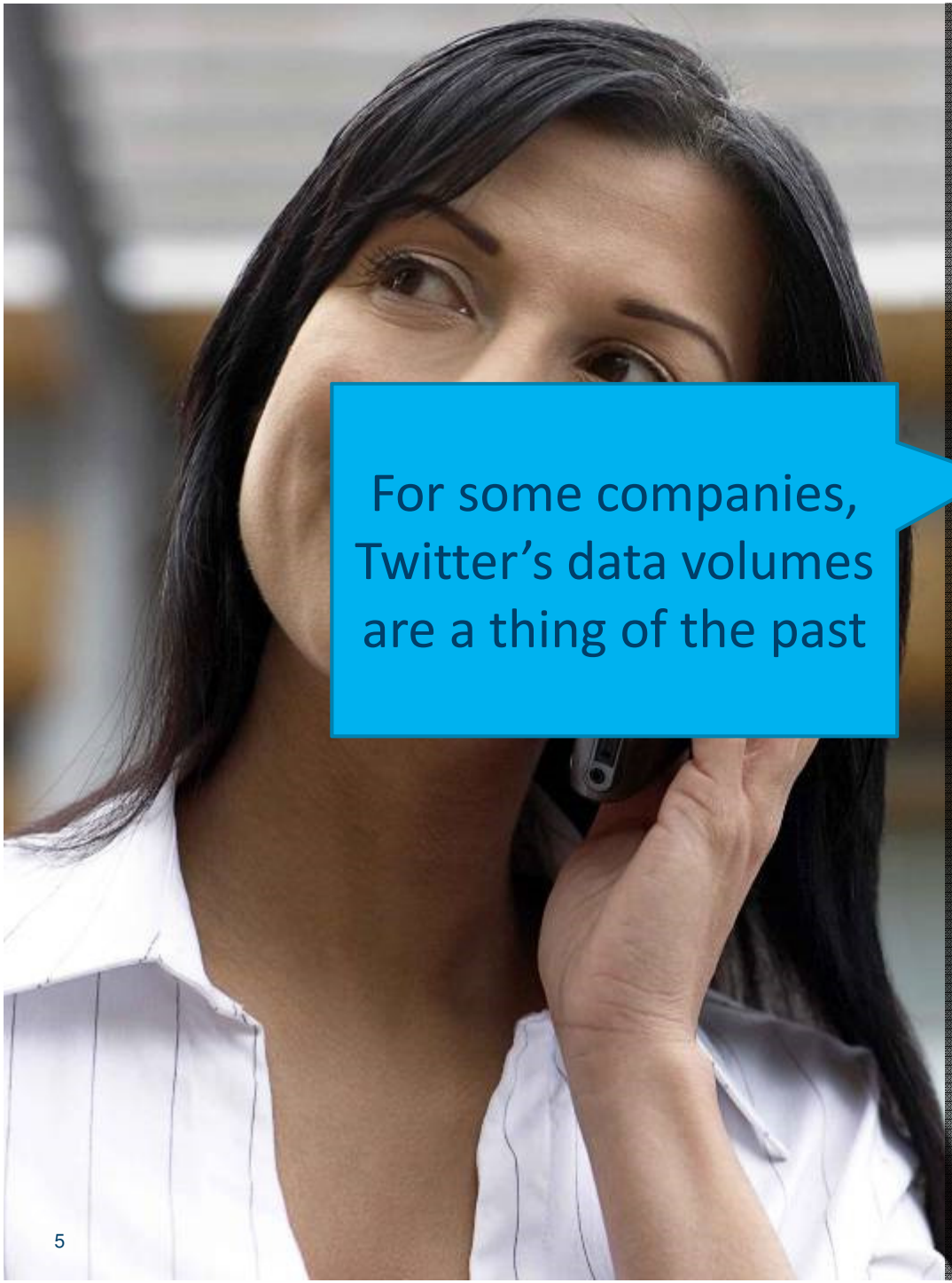
**The transformation:** Analyzing petabytes of wind data to pinpoint optimal turbine placement, maximizes power generation and reduces energy costs.

*"In our development strategy, we see growing our library in the range of 18 to 24 petabytes of data. And while **it's fairly easy to build that library**, we needed to make sure that we could **gain knowledge from that data**."*

*— Lars Christian Christensen, vice president, Vestas Wind Systems*

Vestas

IBM Software Group



For some companies, Twitter's data volumes are a thing of the past

# Asian Telco reduces billing costs and improves customer satisfaction

Real-time mediation and analysis of **5 Billion Call Detail Records per day**

Data processing time reduced from **12 hrs to 1 min**

**Hardware cost reduced to 1/8th**

Proactively address issues (e.g. dropped calls) impacting customer satisfaction.

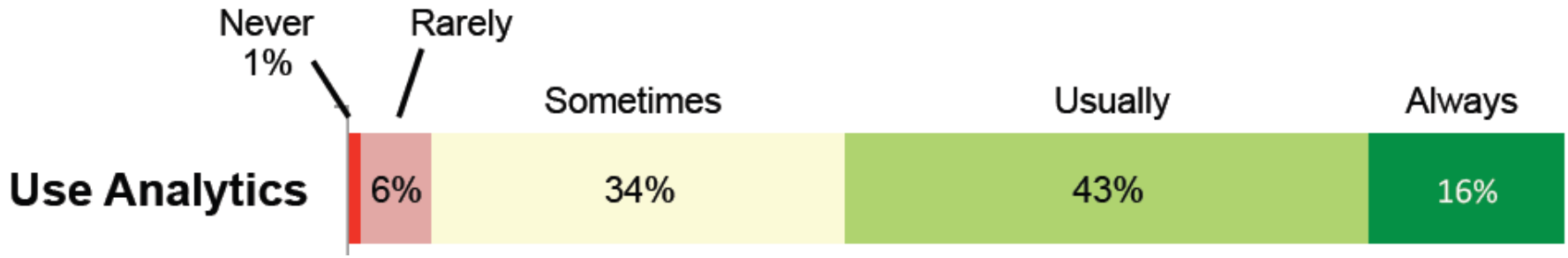
# Everything works.

# We're done.

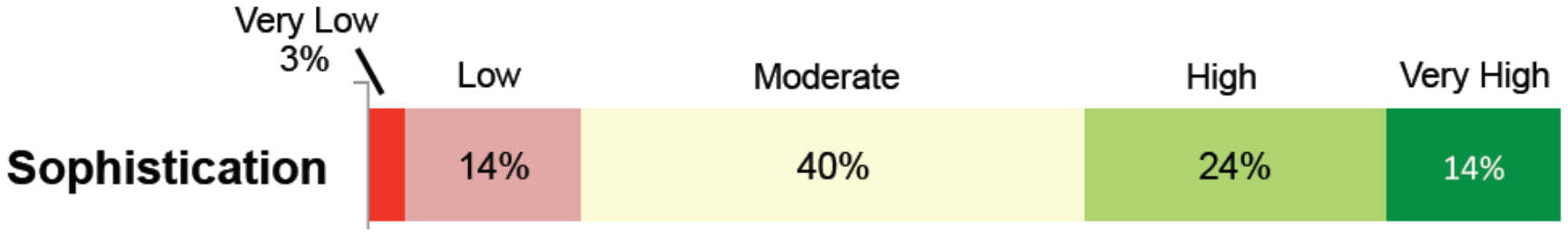
End of Story?

# Doing Analytics is NOT mainstream yet

(Source: Rexer Data Miner Survey 2013, http://rexeranalytics.com/Data-Miner-Survey-Results-2013.html)



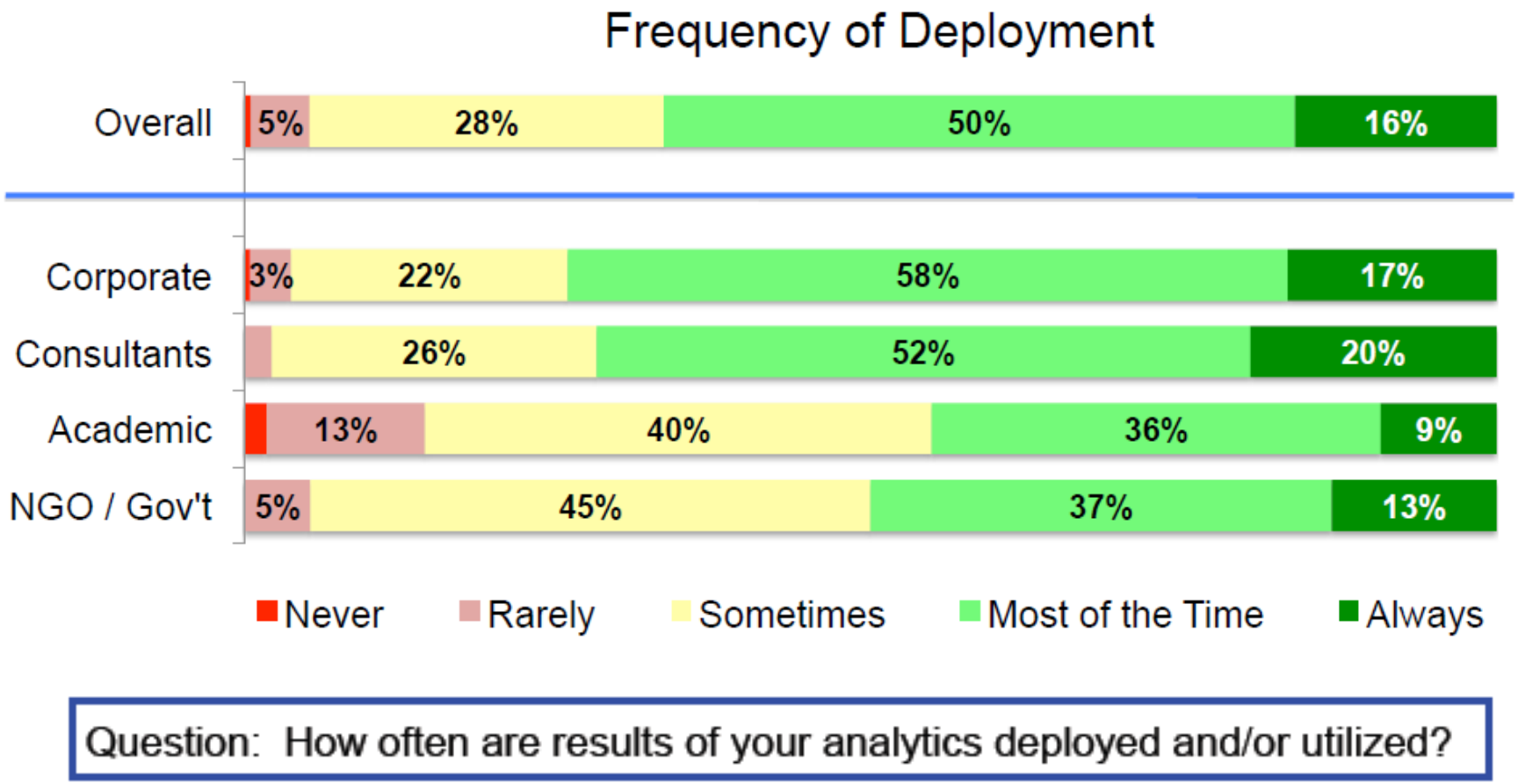

Question: When there are questions that can be addressed by analytics, how often does your company / organization use analytics to address them?




Question: In general, with what degree of sophistication does your company / organization approach analytic problems?

# Over 25% of analytic models are never or rarely deployed

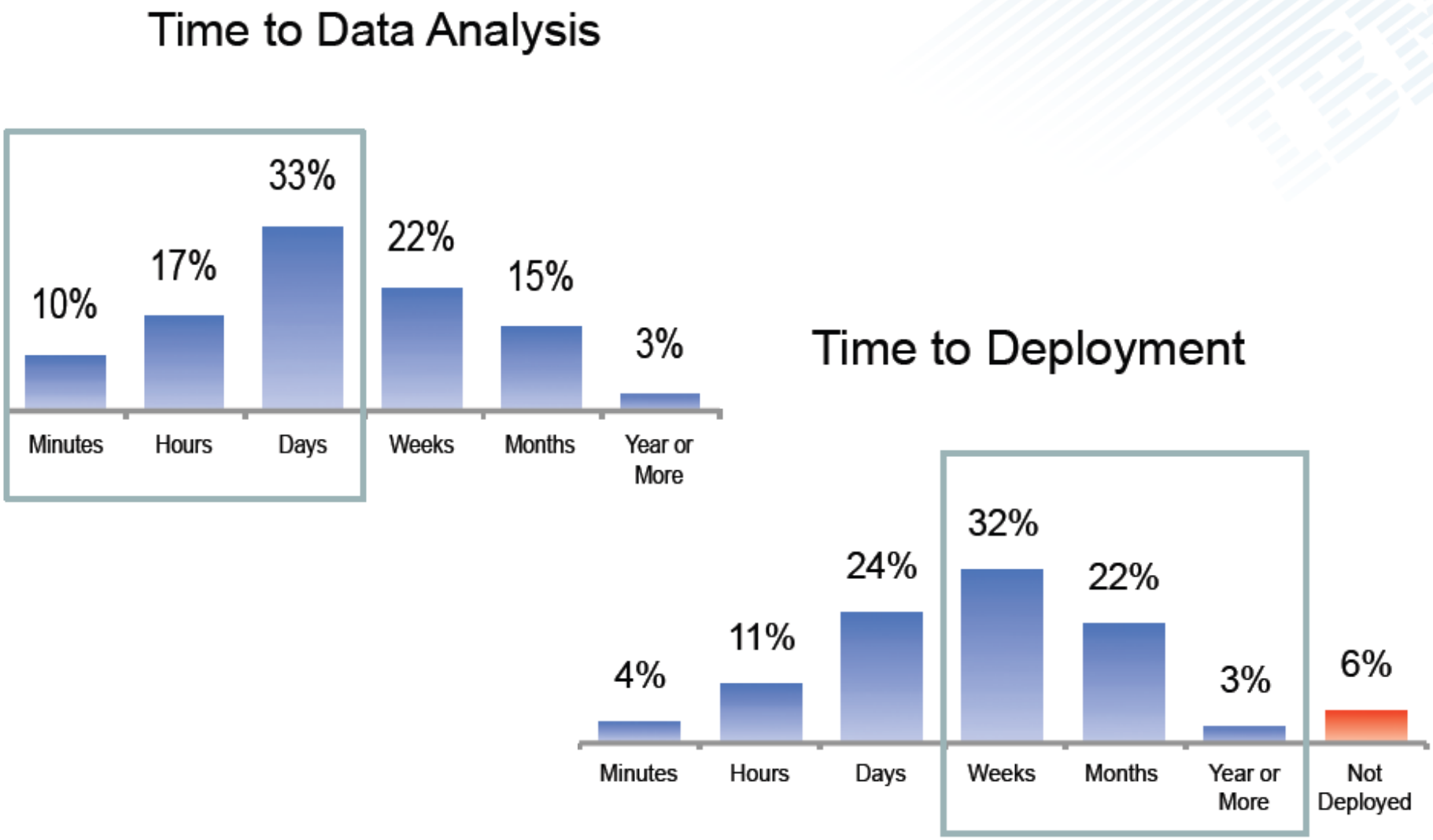
**(Source: Rexer Data Miner Survey 2013)**




Question: How often are results of your analytics deployed and/or utilized?

# …and when they're deployed, they're deployed late

(Source: Rexer Data Miner Survey 2013)

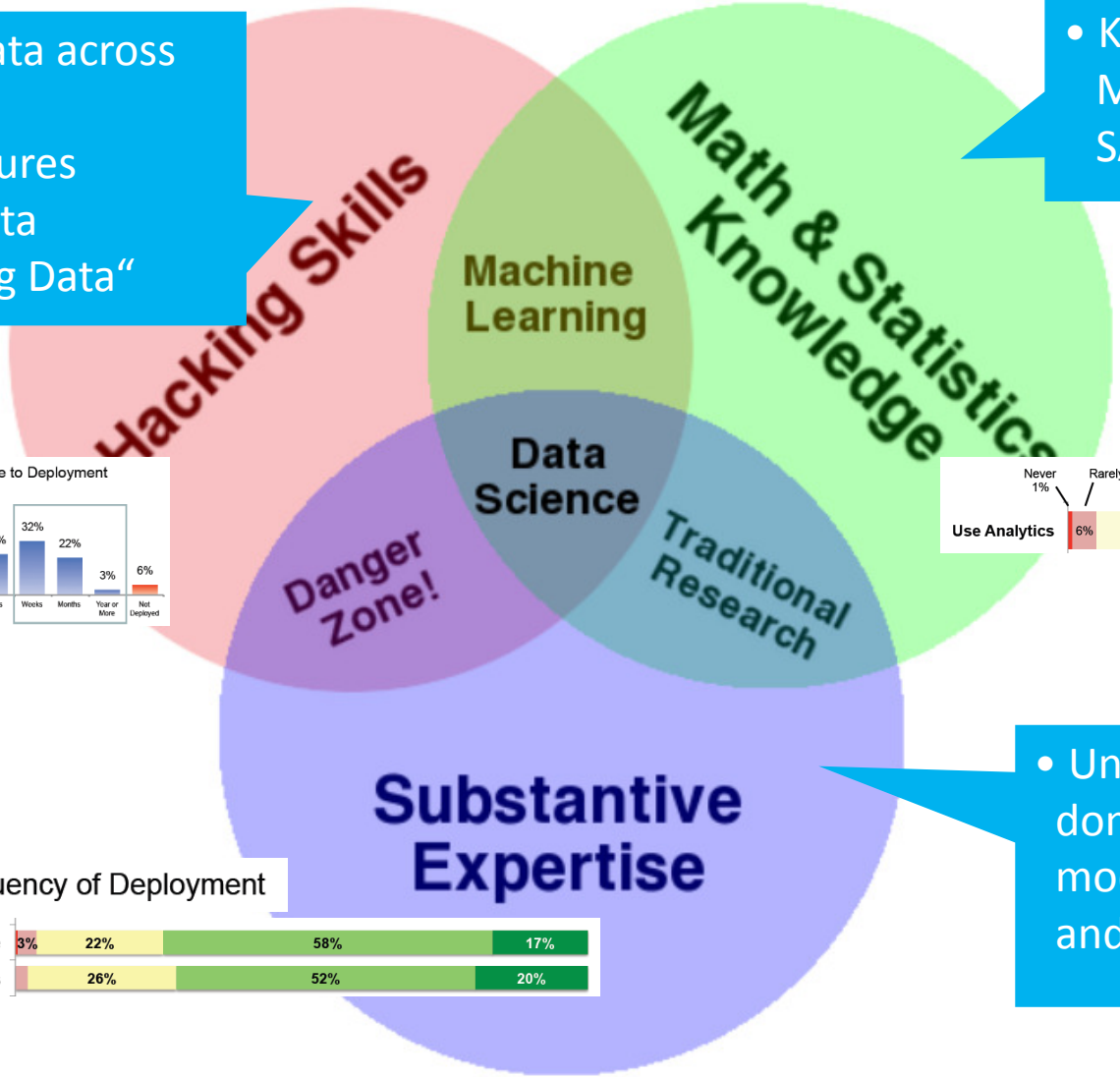


# It’s time to change that

# A new role emerges: The Data Scientist

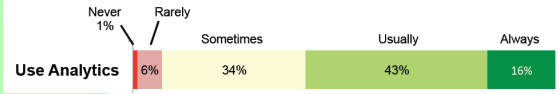
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

- Integrate data across sources
- Extract features
- Visualize data
- Deal with Big Data“




- Statistics Knowledge
- Knows Statistics / Mining-Tools (SPSS, SAS, R,...)







- Understands the domain to ensure that models are relevant and actionable

# Data Scientist: Job Descriptions (kaggle.com)

## Qualification:

*Ph.D. in Statistics, Machine Learning, Computer Science or other quantitative data science field*

*Master's or higher in computer science, artificial intelligence, statistics, or similar*

*Solid understanding of applied statistics, predictive modeling, data mining, machine learning, and other quantitative methodology*

Domain Expertise?
*Training on the job!*

## Tasks:

The Lead Data Scientist works **across the enterprise** to shape Bridgepoint's strategy by **analyzing, and interpreting complex data sets** and **building innovative products or services** that utilize big data.

Use analytics/statistics to answer business questions;
appropriately **selecting the relevant analytical technique**,
creating **meaningful data visualizations** and representations,
and effectively **communicating the data story** and resulting recommendations.

**Passion** for digging into large data sets and
extracting knowledge through **analysis and visualization**

# The Chief Data Officer: The executive for „Data Science“

http://www-935.ibm.com/services/c-suite/cdo/

## Data leverage

Find ways to use existing data assets

## Data enrichment

Augment data by combining internal and external data

## Data monetization

Find new avenues of earnings and revenue

## Data protection

Protect data as an asset

## Data upkeep

Manage the health of data undergovernance





Source: *The new hero of big data and analytics: The Chief Data Officer*, IBM Institute for Business Value. 

## Technology Imperatives

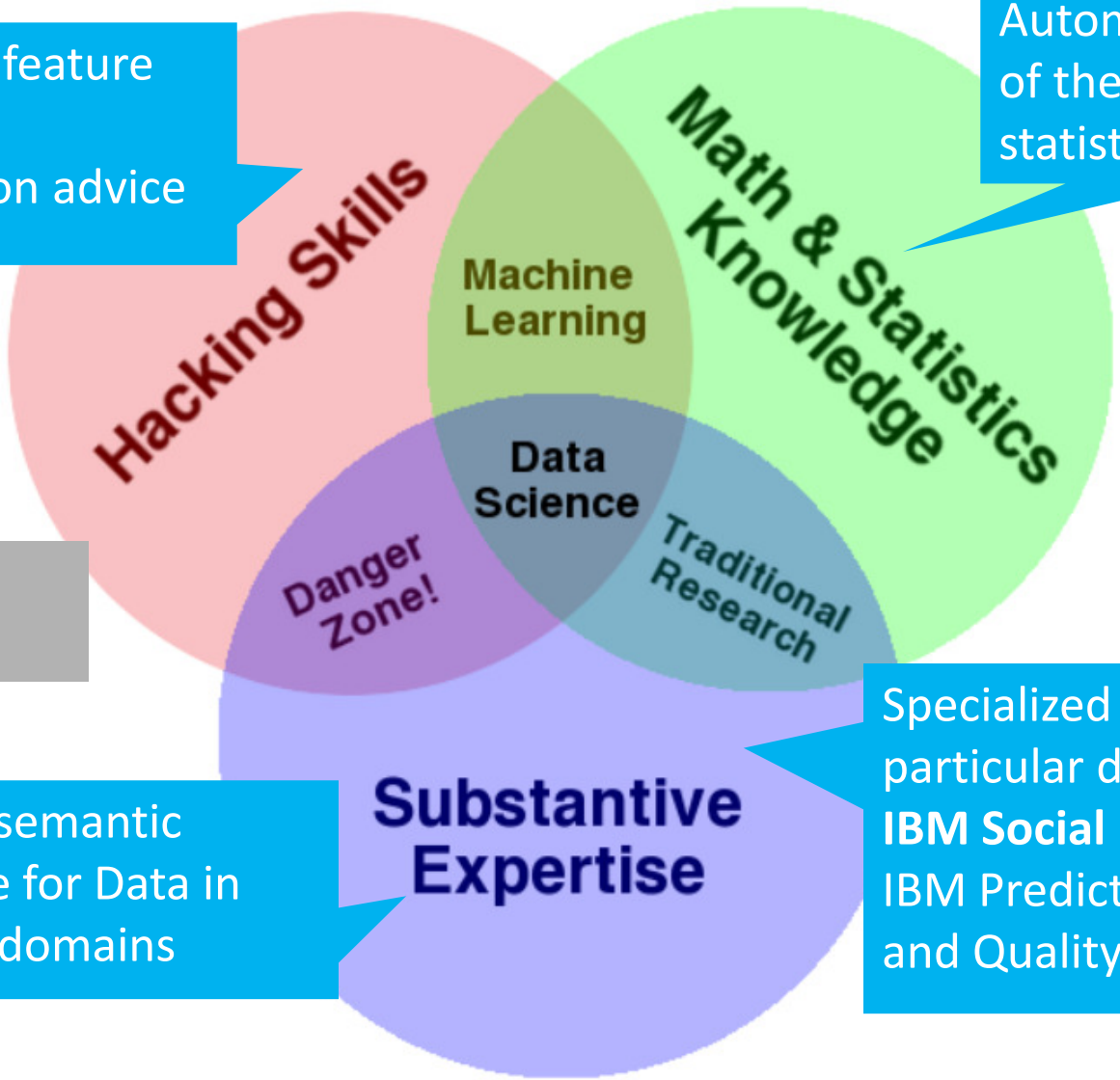
- **Make it easy for everybody to have a meaningful conversation with data**

- **Make it easy for Data Scientists to create Big Data Analytics applications**

# Data Scientists are hard to come by – how can Software help?

- Automatic feature extraction
- Visualization advice

Automated selection of the proper statistical approach

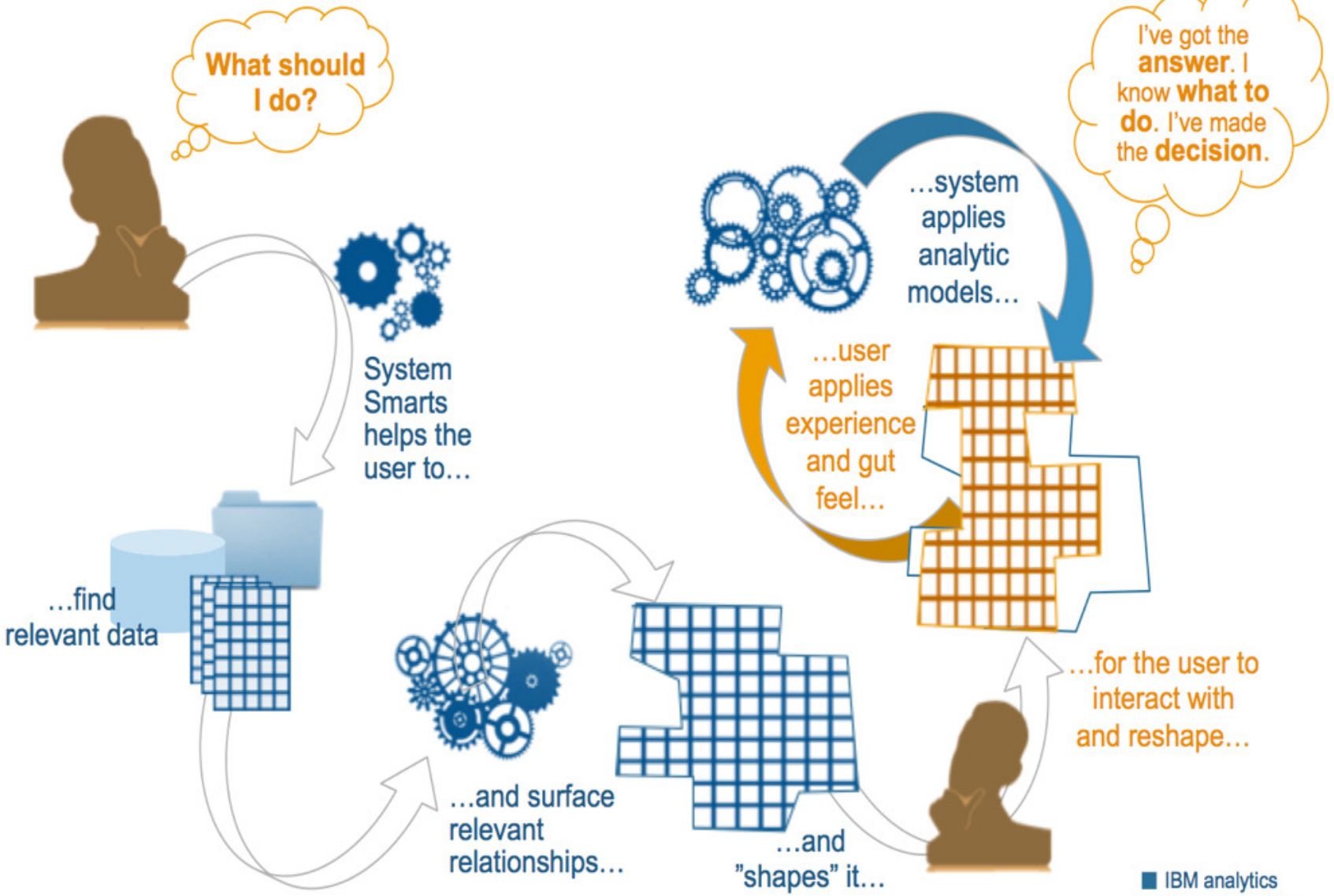



**IBM Watson Analytics**

„Built in“ semantic knowledge for Data in particular domains

Specialized Applications for particular domains:
**IBM Social Media Analytics**, IBM Predictive Maintenance and Quality,...

# End Goal: A meaningful conversation with data




Scientists

# Example: Improving responses to a campaign

- A regional insurance branch office has conducted a promotional campaign offering new insurance products to their existing customers.
- Customer responses have been recorded together with existing data on customer demographics and their insurance profile information.
- The branch manager wonders…
  - What are the key factors that influence whether customers respond to the campaign?
  - How can I improve my marketing, going forward?

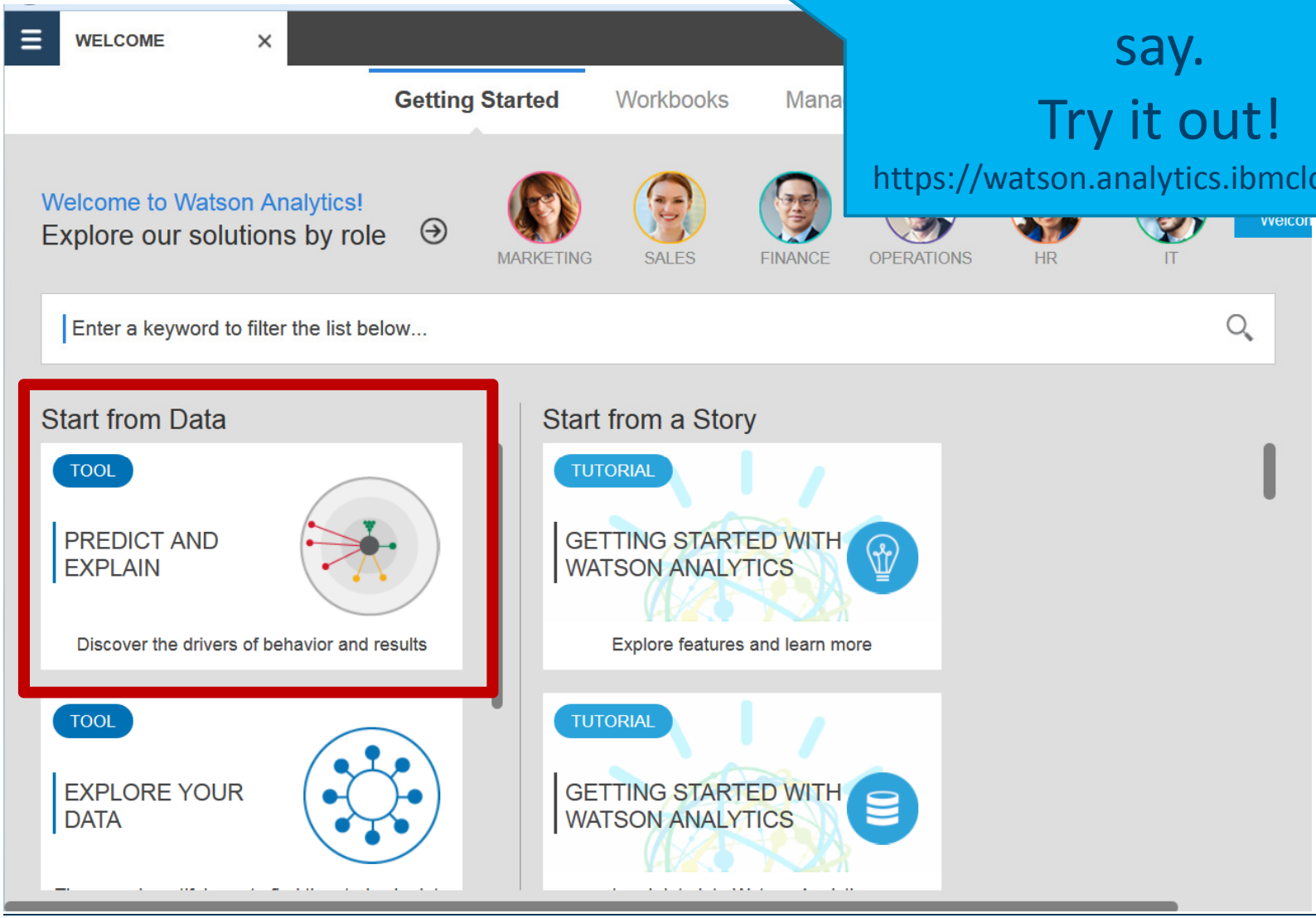
# The data set: your everyday "*wide* data" problem

- 9000 customers
- 48 attributes for each customer
- Some data may be wrong, some data is incomplete

| A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLTV | Response | AddDriverintoPolicy | AddressChangeCount | AgentId | AvgLength | AvgNoteA | ChangeAddressw | Collective | Coverage | Education | Effectivetc | Employ |
| 8.332.538.119 | no | 0 | 0 | - | 29 | 1 | 0 | 0 | Extended | high schoo | 1/27/2011 | Employ |
| 7.422.851.604 | no | 1 | 1 | - | 12 | 4 | 1 | 1 | Extended | high schoo | ######## | Unempl |
| 7.322.595.652 | no | 0 | 1 | - | 14 | 2 | 0 | 1 | Extended | bachelor | ######## | Employ |
| 679.072.705 | no | 0 | 1 | Agent-192 | 22 | 5 | 1 | 0 | Premium | bachelor | 1/13/2011 | Employ |
| 6.602.575.407 | no | 1 | 0 | - | 5 | 1 | 0 | 1 | Basic | bachelor | 2/13/2011 | Employ |
| 6.461.875.715 | no | 0 | 0 | Agent-26 | 9 | 2 | 0 | 0 | Extended | high schoo | ######## | Unempl |
| 6.185.018.803 | no | 0 | 2 | Agent-73 | 25 | 3 | 0 | 0 | Extended | college | ######## | Unempl |
| 6.113.468.307 | no | 0 | 2 | Agent-115 | 14 | 1 | 0 | 0 | Basic | college | 2/26/2011 | Unempl |

# Enter: IBM Watson Analytics

Don't believe what I say.
Try it out!
https://watson.analytics.ibmcloud.com/

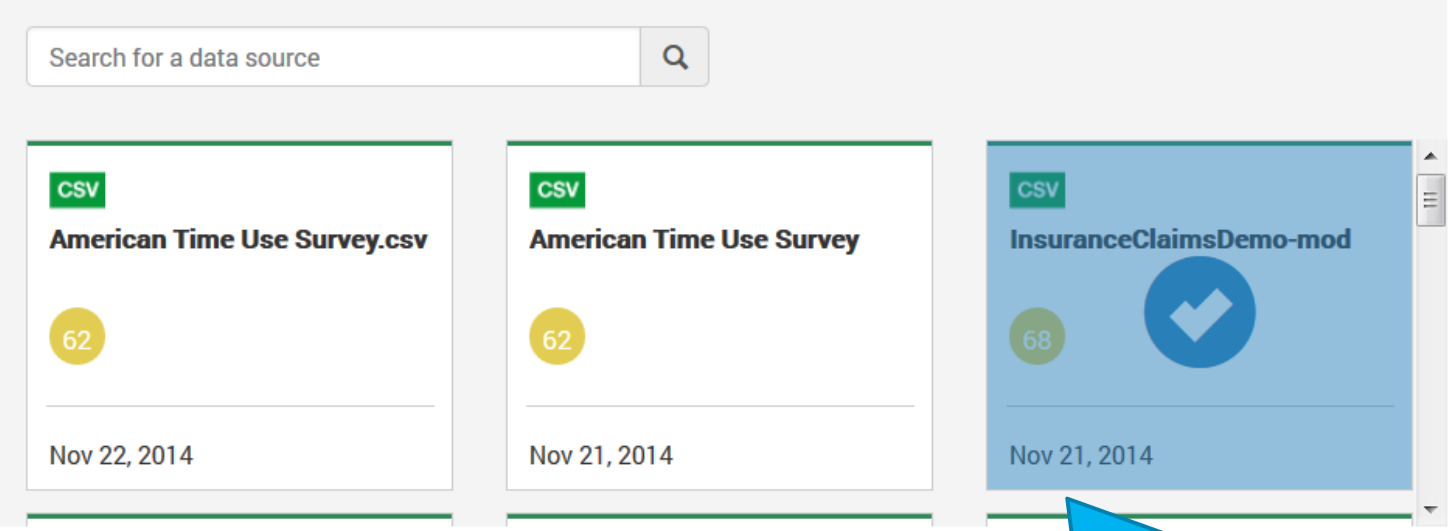


# Step 1 Create a Prediction

1. Name your workbook

Understand Insurance Campaign

2. Select a data source




Select the data set

3. Select target(s): up to 5 targets may be added per workbook

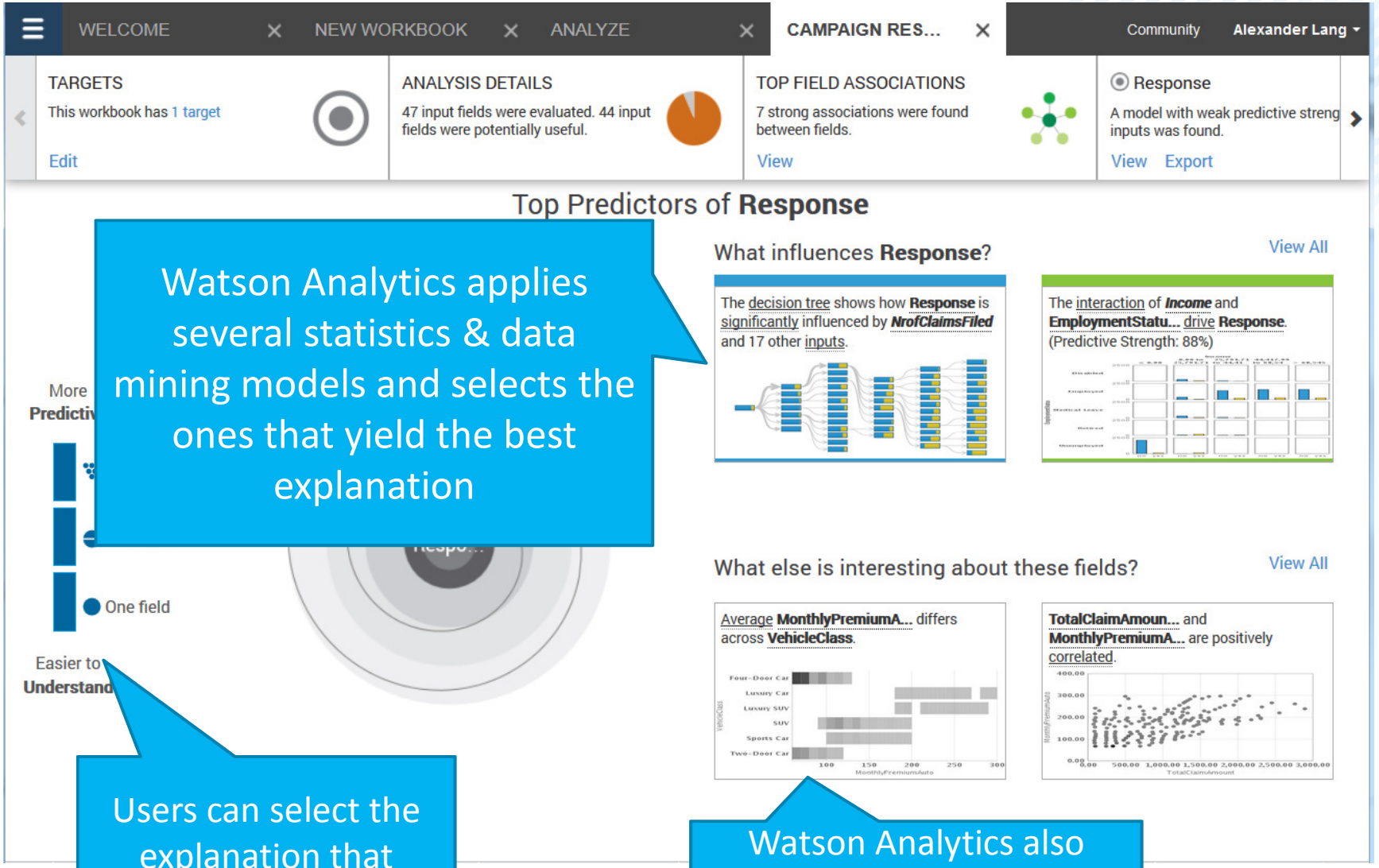
Select target | Response

Edit this workbook's field properties

Select the variable to explain

Create Workbook | Cancel

# Step 2: Watson Analytics identifies key influence factors

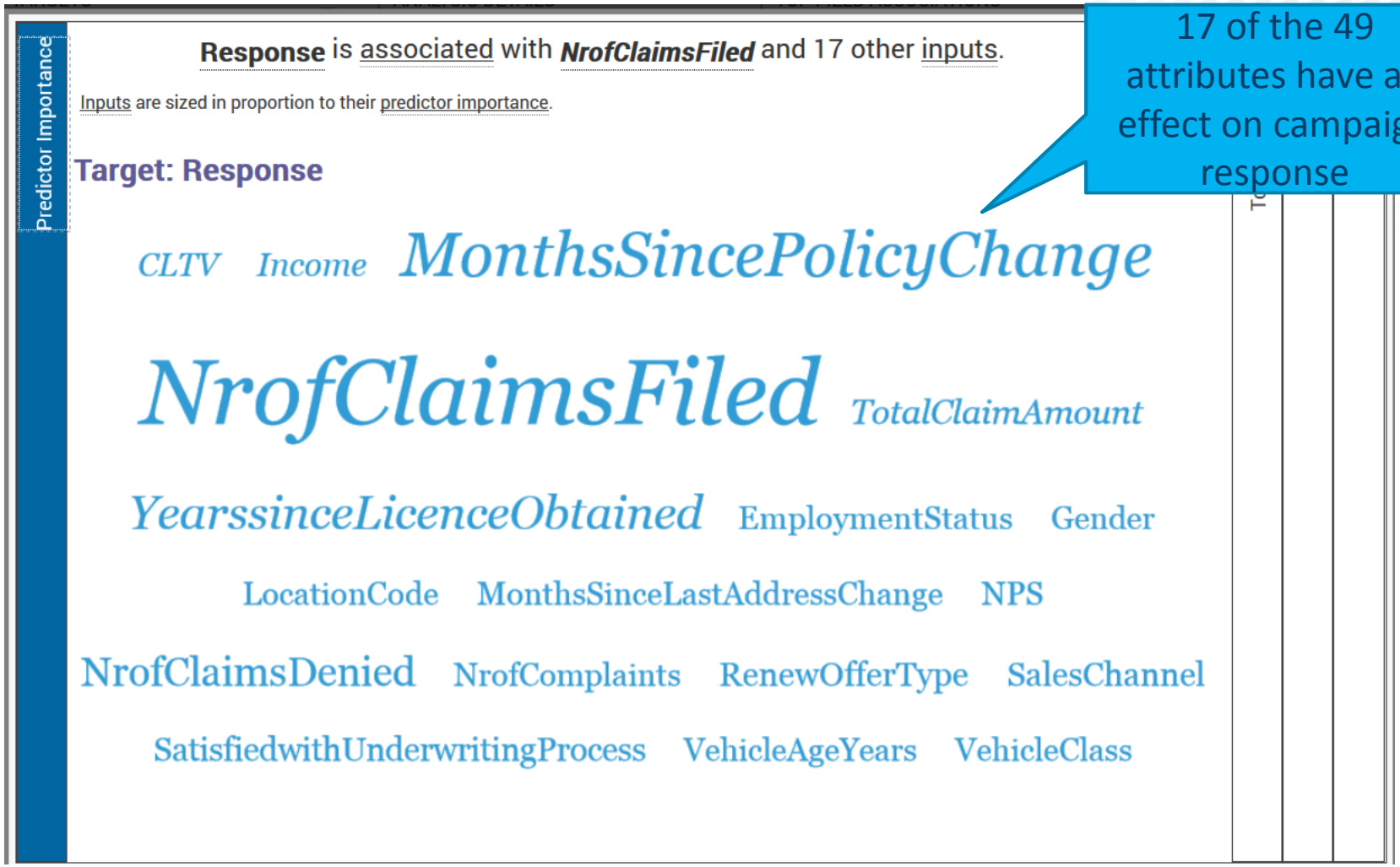



Watson Analytics applies several statistics & data mining models and selects the ones that yield the best explanation

Users can select the explanation that works best for them

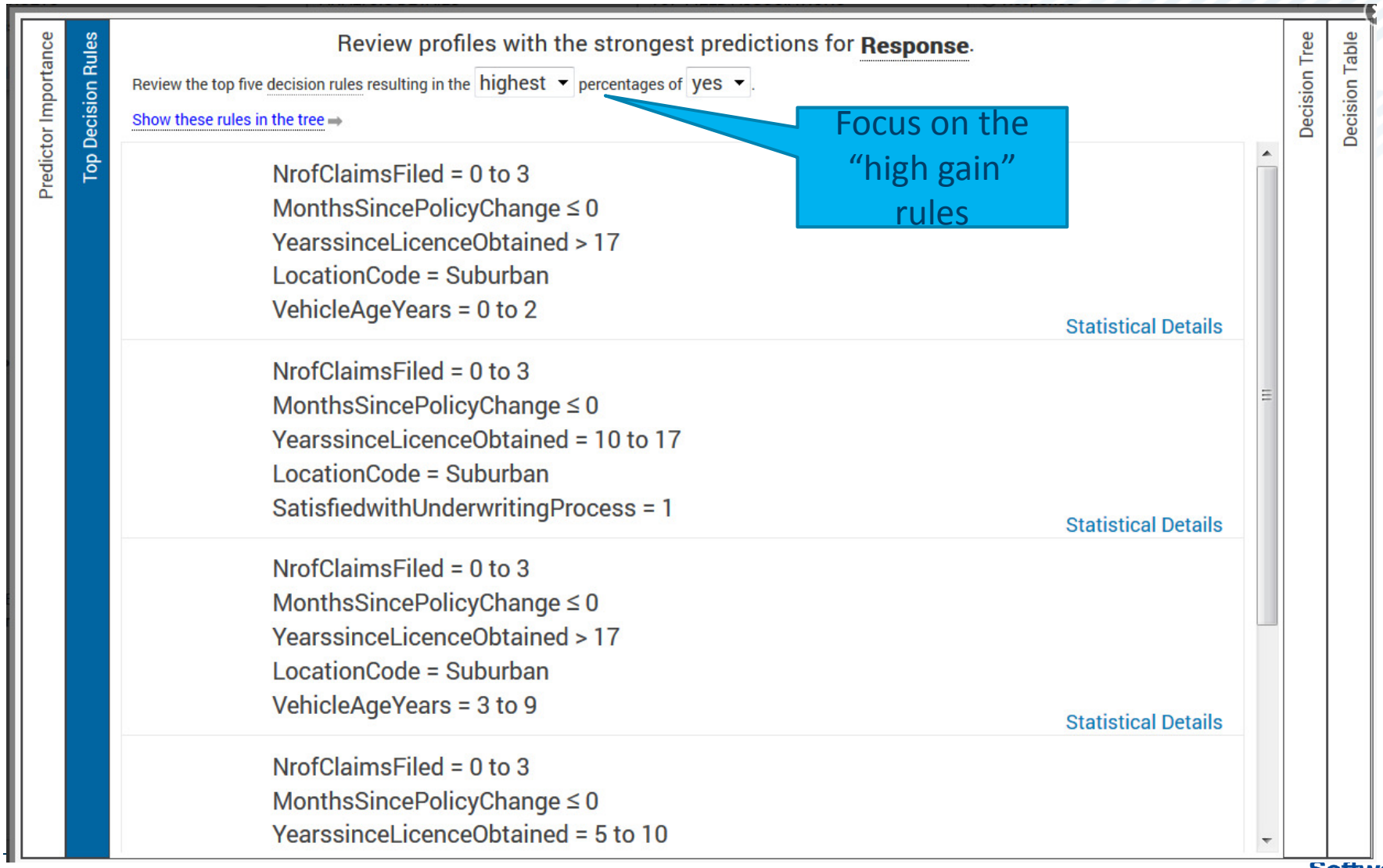
Watson Analytics also surfaces other insights about the data set

# Step 3-1: Insurance Agent explores insights further




17 of the 49 attributes have an effect on campaign response

# Step 3-2: Key success factors for campaign response

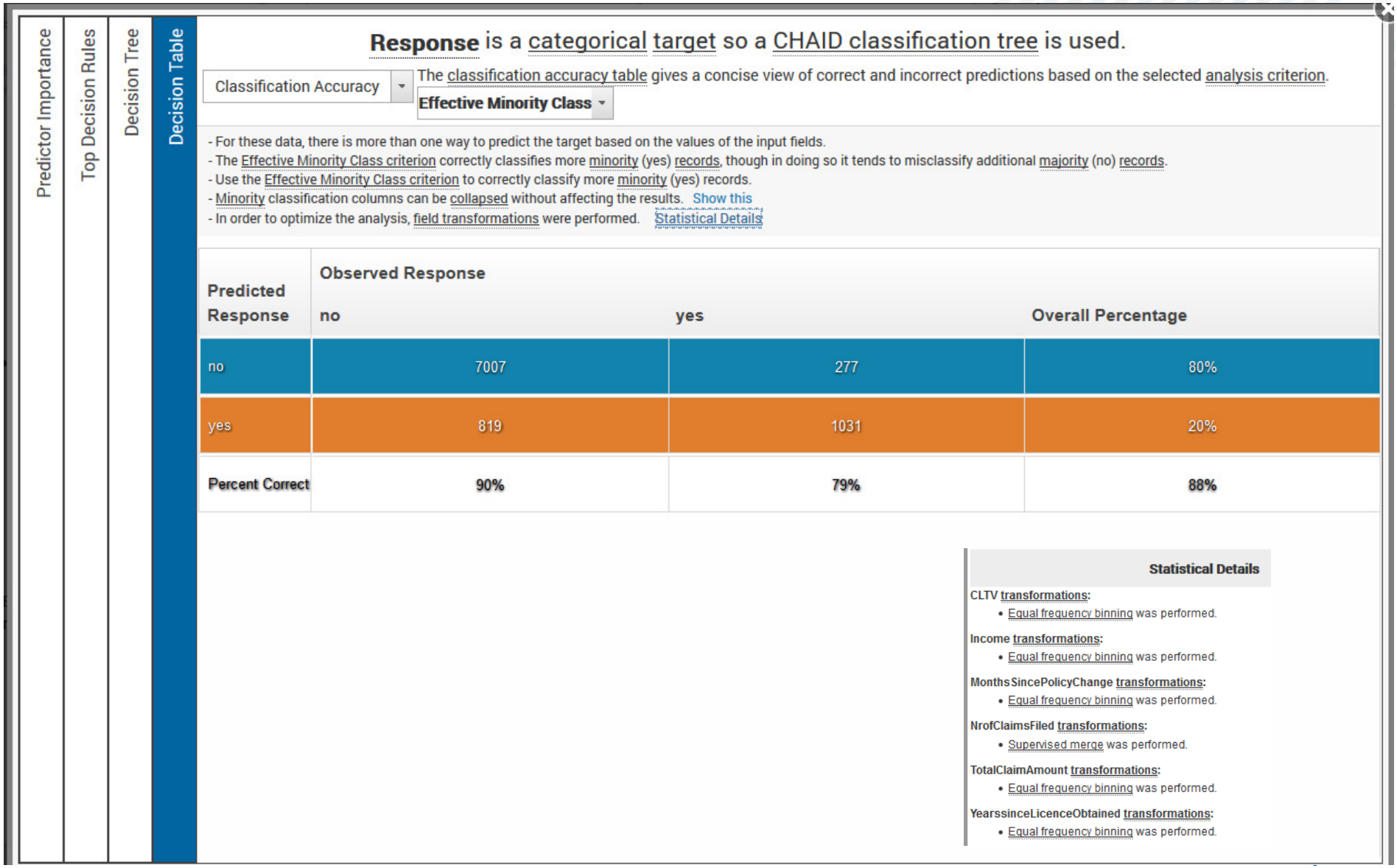


# The “Aha” moments….

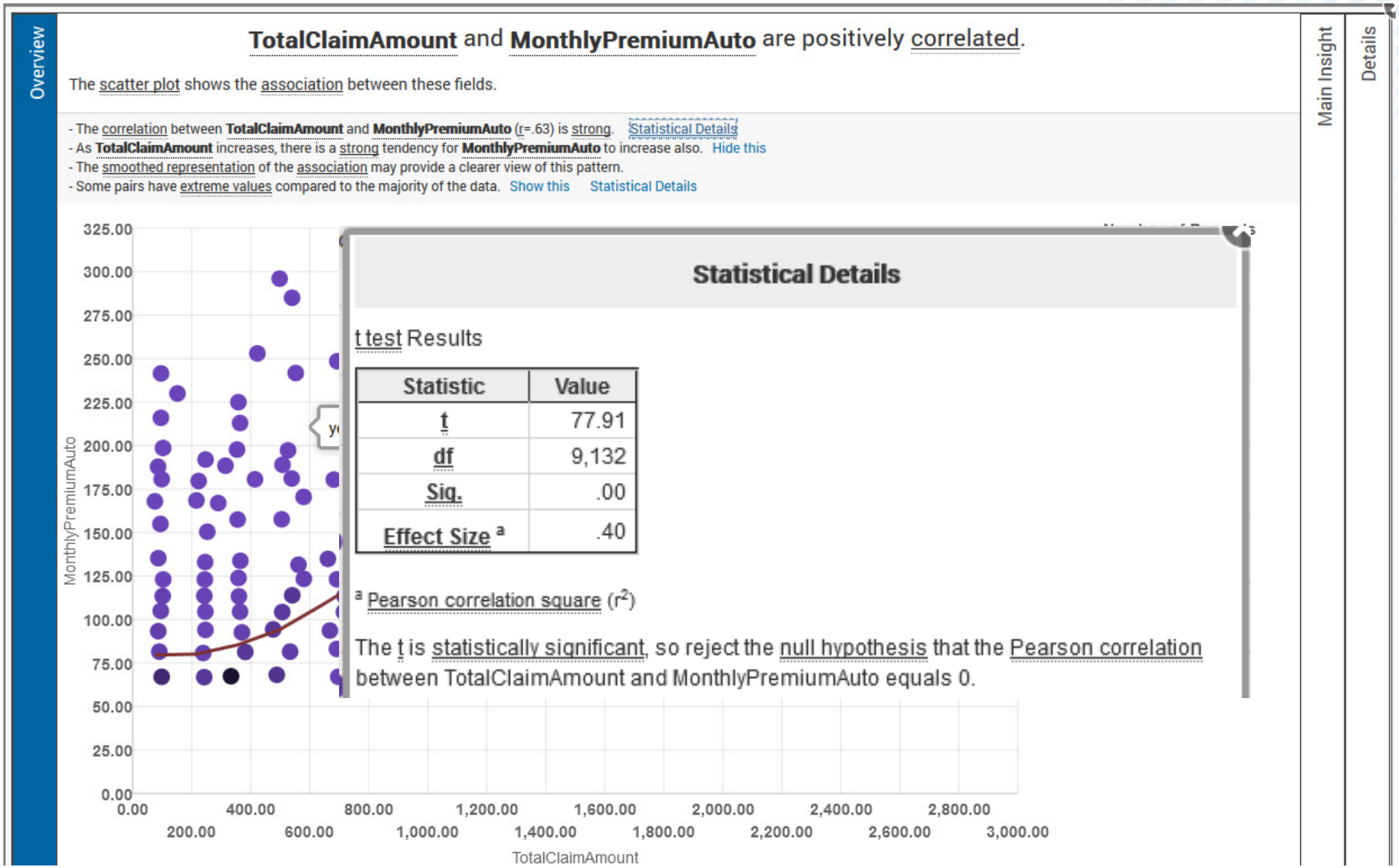
- The campaign mainly reasonated with costomers living in suburbs, not rural or metropolitan areas
- …and mostly with older customers (Years since License Obtained > 10)
- …and only with customers that don’t have many claims filed (0 to 3)

- Next time:
  - Get more info on what customers in rural areas need (a better deal on SUVs?)
  - Think twice before sending offers out to “frequent claimers”

# Step 3-3: Statistical details – for those who need it

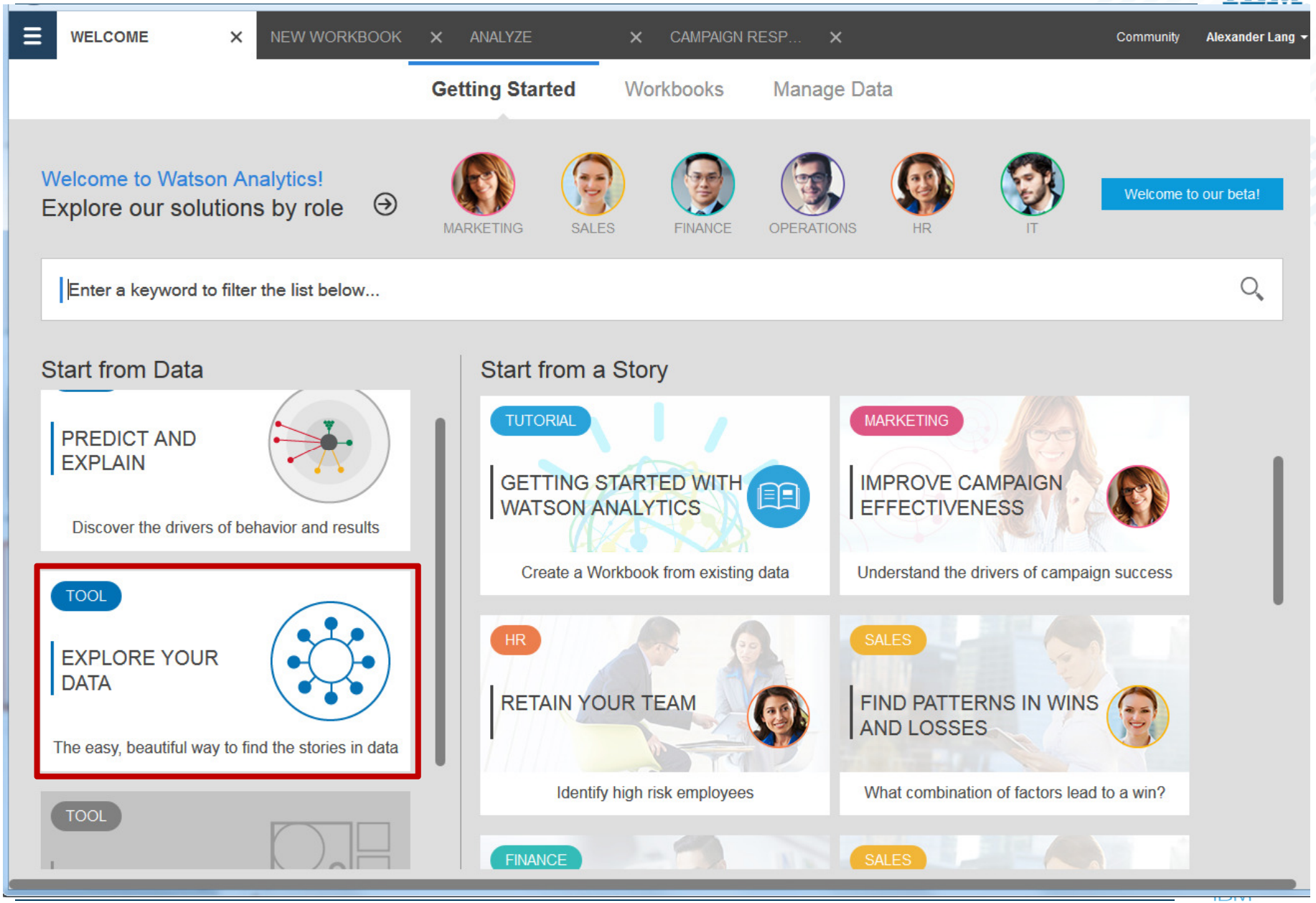


# Other insights: example



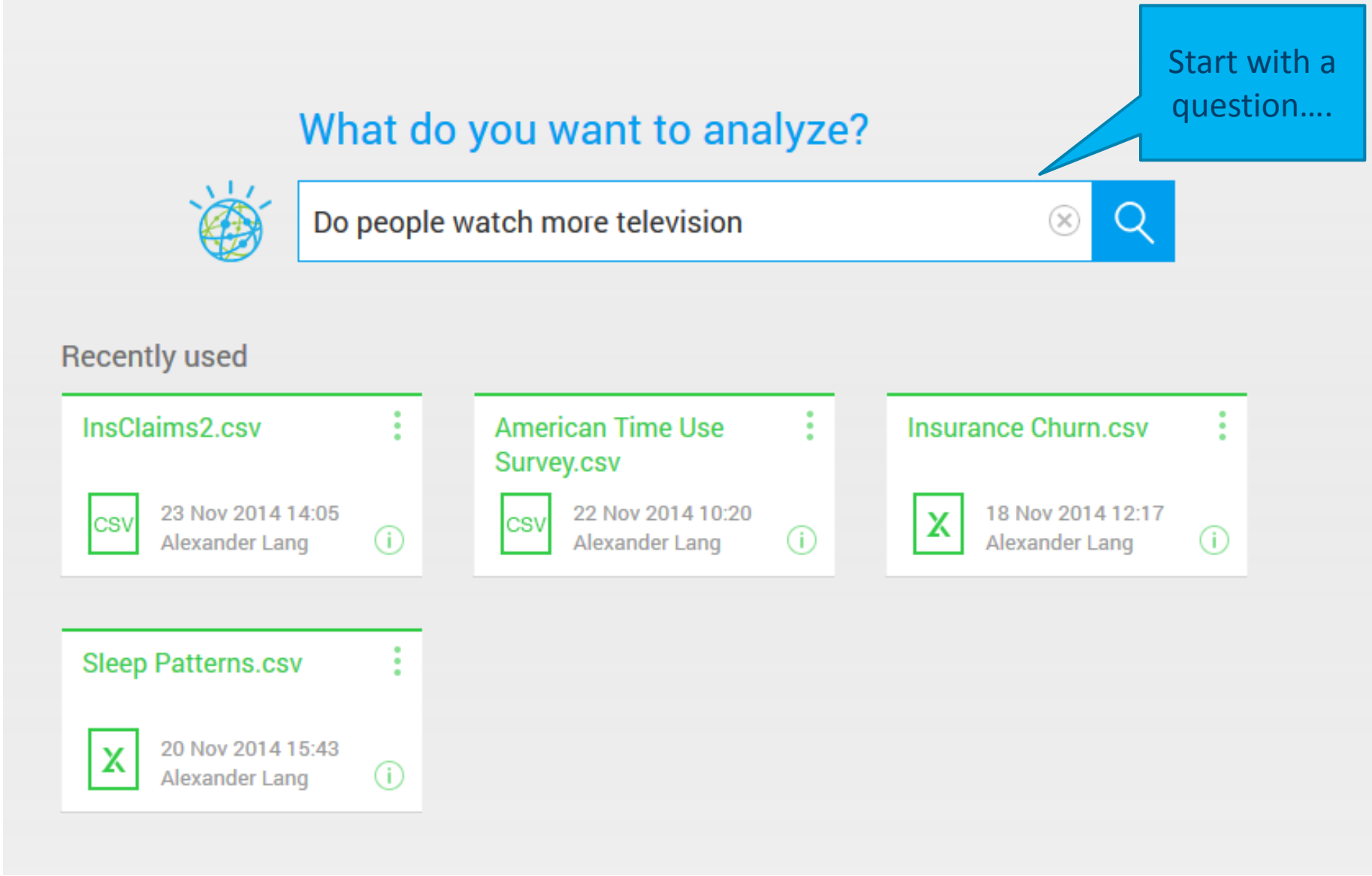
# But - wait a minute!

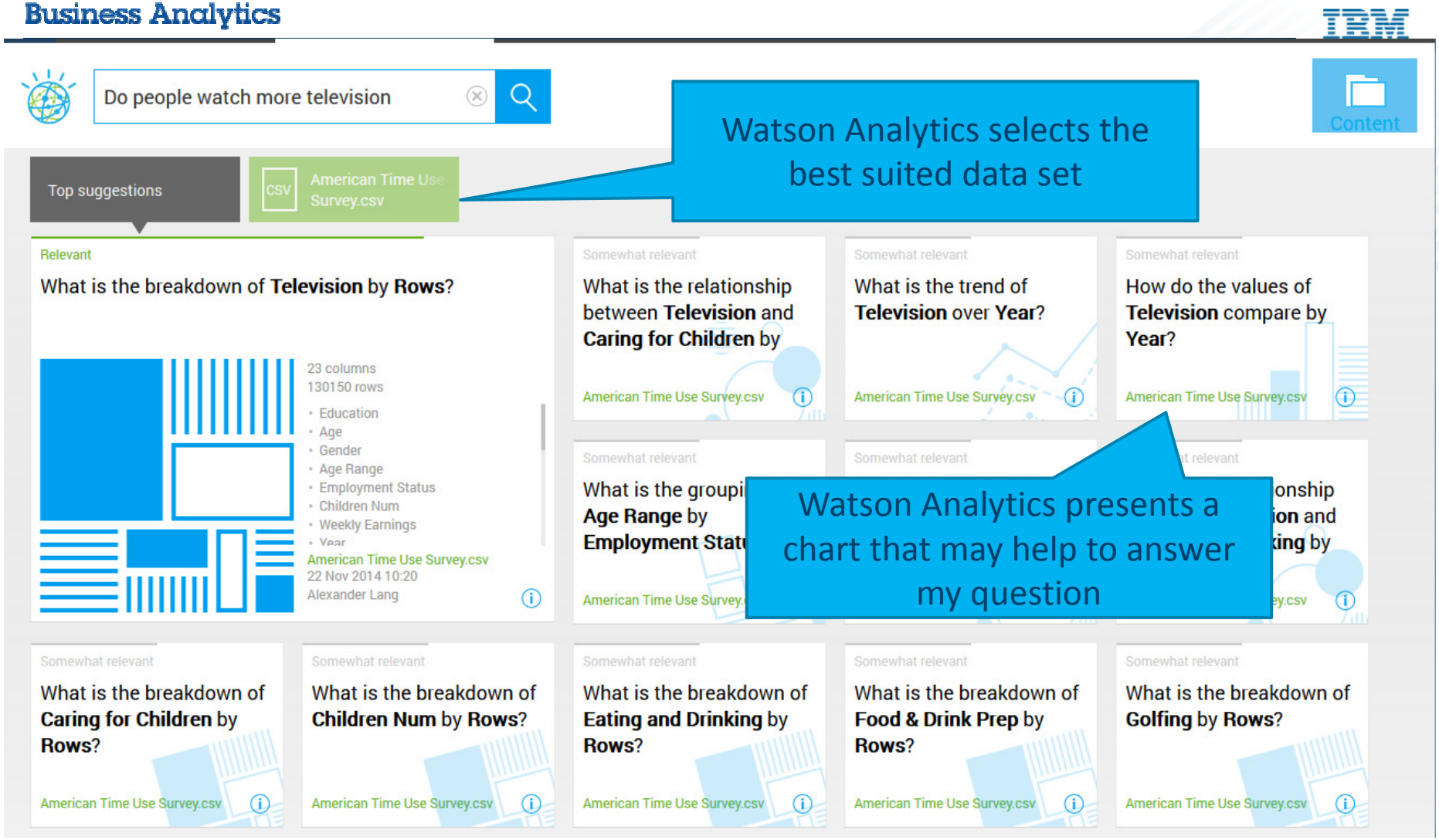
- Is this *really* the best predictive model that explains campaign responses?
- Probably not.
- It's the best model for an insurance branch manager, who has never heard the word "predictive" before
- It's waaay better than "Plan B": gut feel

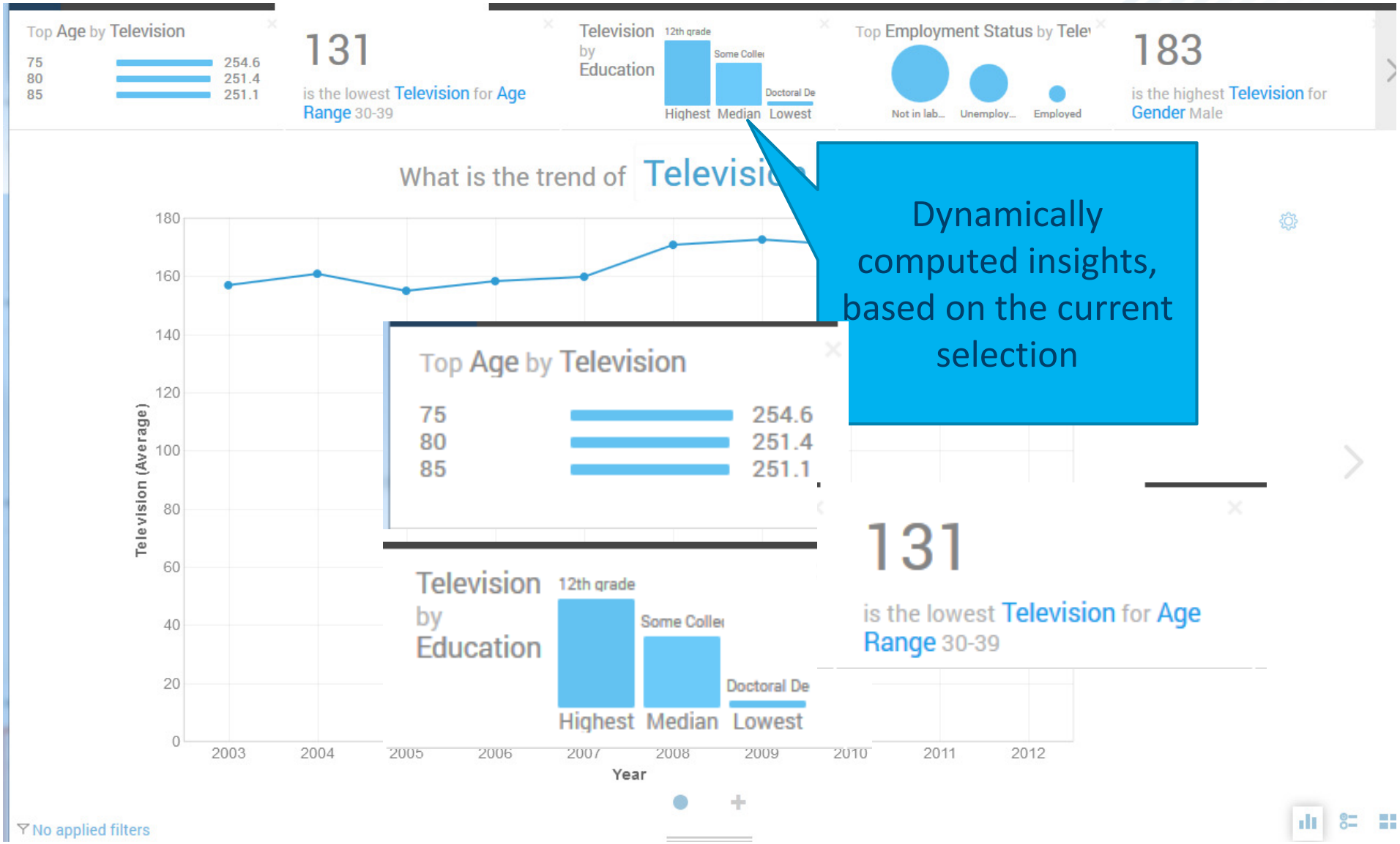

WELCOME
NEW WORKBOOK
ANALYZE
CAMPAIGN RESP...
Community
Alexander Lang
Getting Started
Workbooks
Manage Data
Welcome to Watson Analytics!
Explore our solutions by role
MARKETING
SALES
FINANCE
OPERATIONS
HR
IT
Welcome to our beta!
Enter a keyword to filter the list below...
Start from Data
PREDICT AND EXPLAIN
Discover the drivers of behavior and results
TOOL
EXPLORE YOUR DATA
The easy, beautiful way to find the stories in data
TOOL
Start from a Story
TUTORIAL
GETTING STARTED WITH WATSON ANALYTICS
Create a Workbook from existing data
MARKETING
IMPROVE CAMPAIGN EFFECTIVENESS
Understand the drivers of campaign success
HR
RETAIN YOUR TEAM
Identify high risk employees
SALES
FIND PATTERNS IN WINS AND LOSSES
What combination of factors lead to a win?
FINANCE
SALES

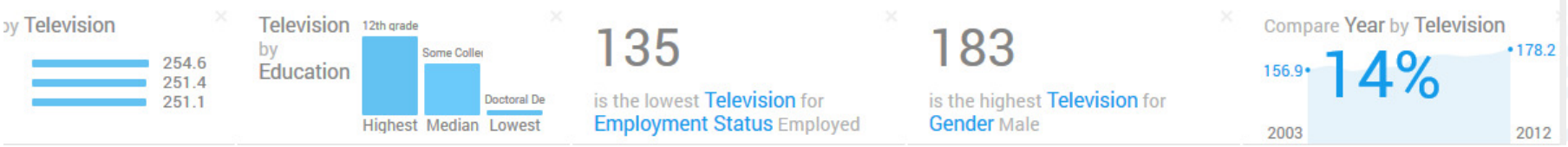
# The Data Set: The American Time Use Survey

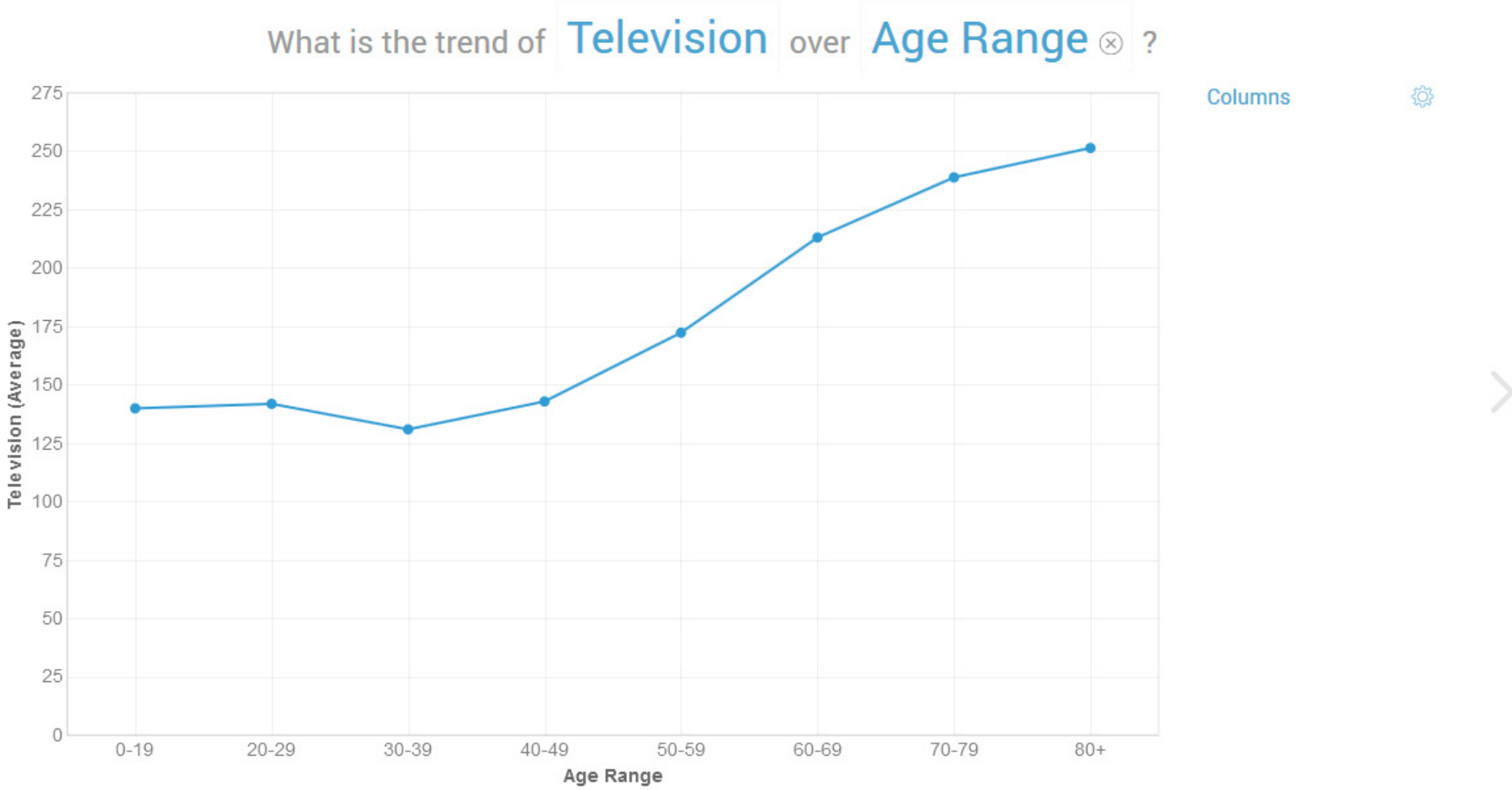
- **Available at http://www.bls.gov/tus/#tables**
- **Data on 130000 Americans, collected over 10 years, on how much time they spend in categories such as**
  - Sleeping
  - Housework
  - Food & Drink Prep
  - Caring for Children
  - Playing with Children
  - Shopping
  - Eating and Drinking
  - Socializing & Relaxing
  - Television


Start with a question....
What do you want to analyze?
Do people watch more television
Recently used
InsClaims2.csv
CSV
23 Nov 2014 14:05
Alexander Lang
American Time Use Survey.csv
CSV
22 Nov 2014 10:20
Alexander Lang
Insurance Churn.csv
X
18 Nov 2014 12:17
Alexander Lang
Sleep Patterns.csv
X
20 Nov 2014 15:43
Alexander Lang

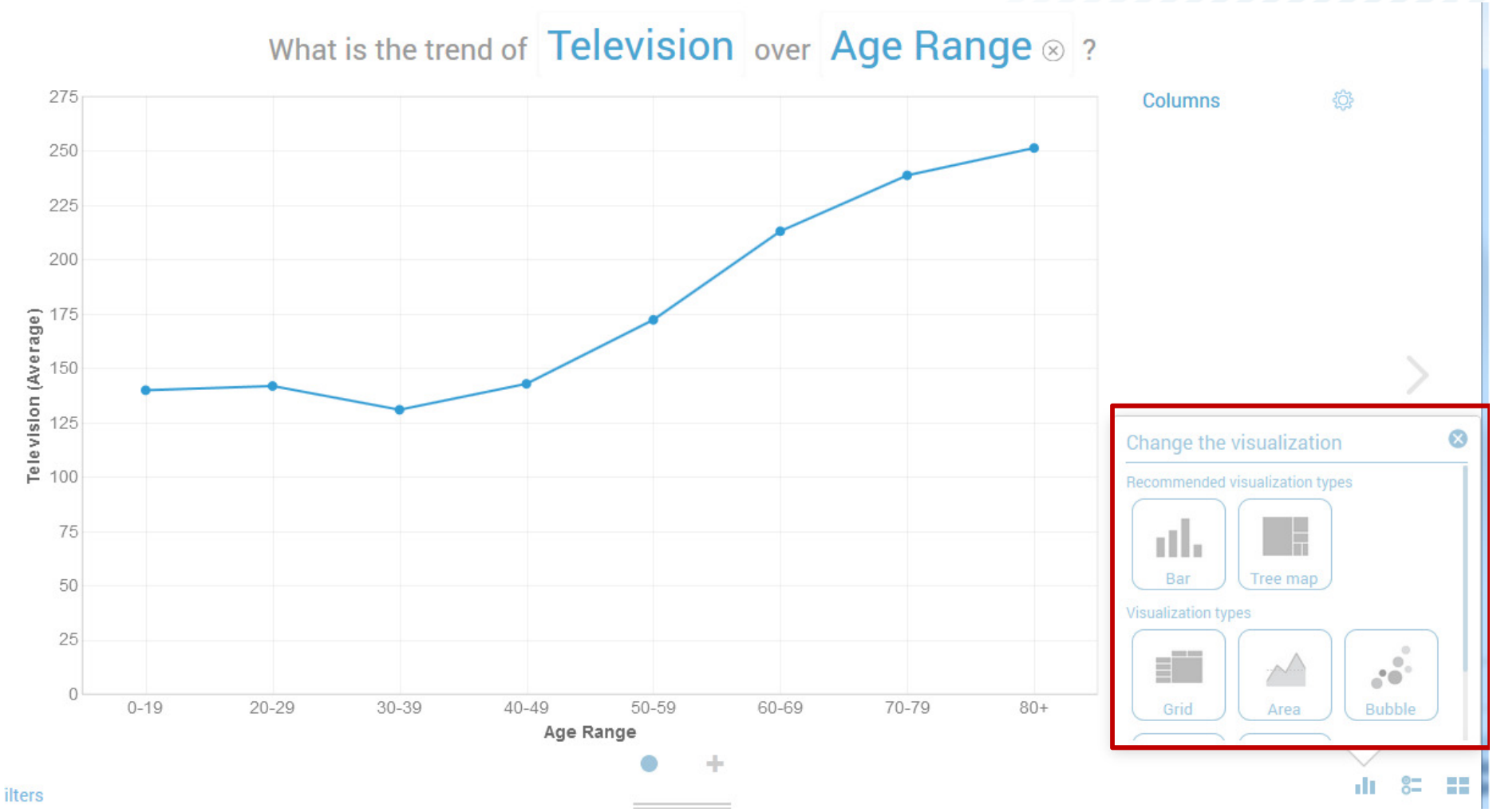

Do people watch more television
Content
Top suggestions
American Time Use Survey.csv
Watson Analytics selects the best suited data set
Relevant
What is the breakdown of Television by Rows?
23 columns
130150 rows
Education
Age
Gender
Age Range
Employment Status
Children Num
Weekly Earnings
Year
American Time Use Survey.csv
22 Nov 2014 10:20
Alexander Lang
Somewhat relevant
What is the relationship between Television and Caring for Children by
American Time Use Survey.csv
Somewhat relevant
What is the trend of Television over Year?
American Time Use Survey.csv
Somewhat relevant
How do the values of Television compare by Year?
American Time Use Survey.csv
Watson Analytics presents a chart that may help to answer my question
Somewhat relevant
What is the breakdown of Caring for Children by Rows?
American Time Use Survey.csv
Somewhat relevant
What is the breakdown of Children Num by Rows?
American Time Use Survey.csv
Somewhat relevant
What is the breakdown of Eating and Drinking by Rows?
American Time Use Survey.csv
Somewhat relevant
What is the breakdown of Food & Drink Prep by Rows?
American Time Use Survey.csv
Somewhat relevant
What is the breakdown of Golfing by Rows?
American Time Use Survey.csv

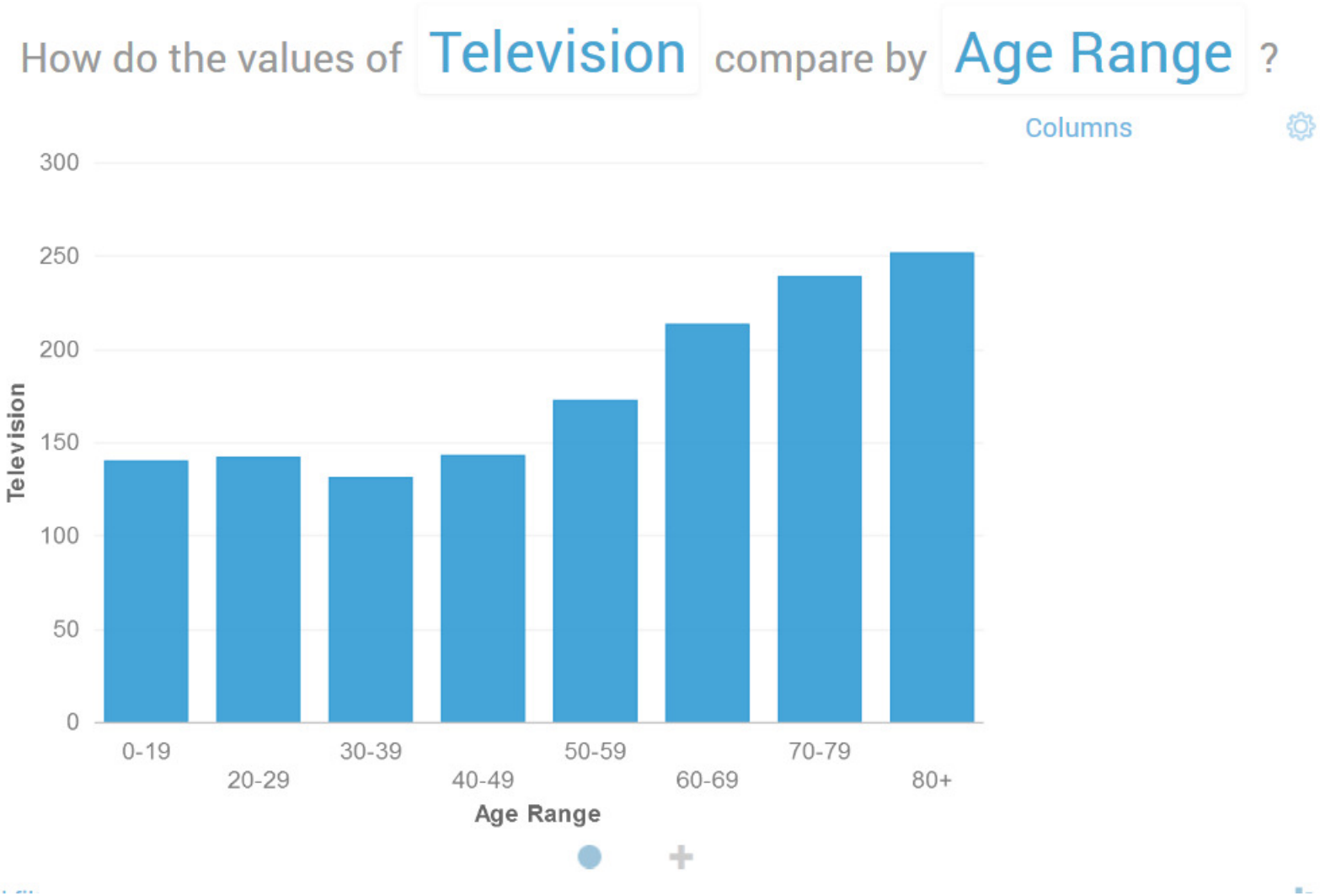

Top Age by Television
75 254.6
80 251.4
85 251.1
131
is the lowest Television for Age Range 30-39
Television by Education
12th grade
Some Colle
Doctoral De
Highest Median Lowest
Top Employment Status by Tele
Not in lab...
Unemploy...
Employed
183
is the highest Television for Gender Male
What is the trend of Televisi
Dynamically computed insights, based on the current selection
Television (Average)
Year
No applied filters


y Television
254.6
251.4
251.1
Television by Education
12th grade
Some Colle
Doctoral De
Highest
Median
Lowest
135
is the lowest Television for Employment Status Employed
183
is the highest Television for Gender Male
Compare Year by Television
156.9
14%
178.2
2003
2012

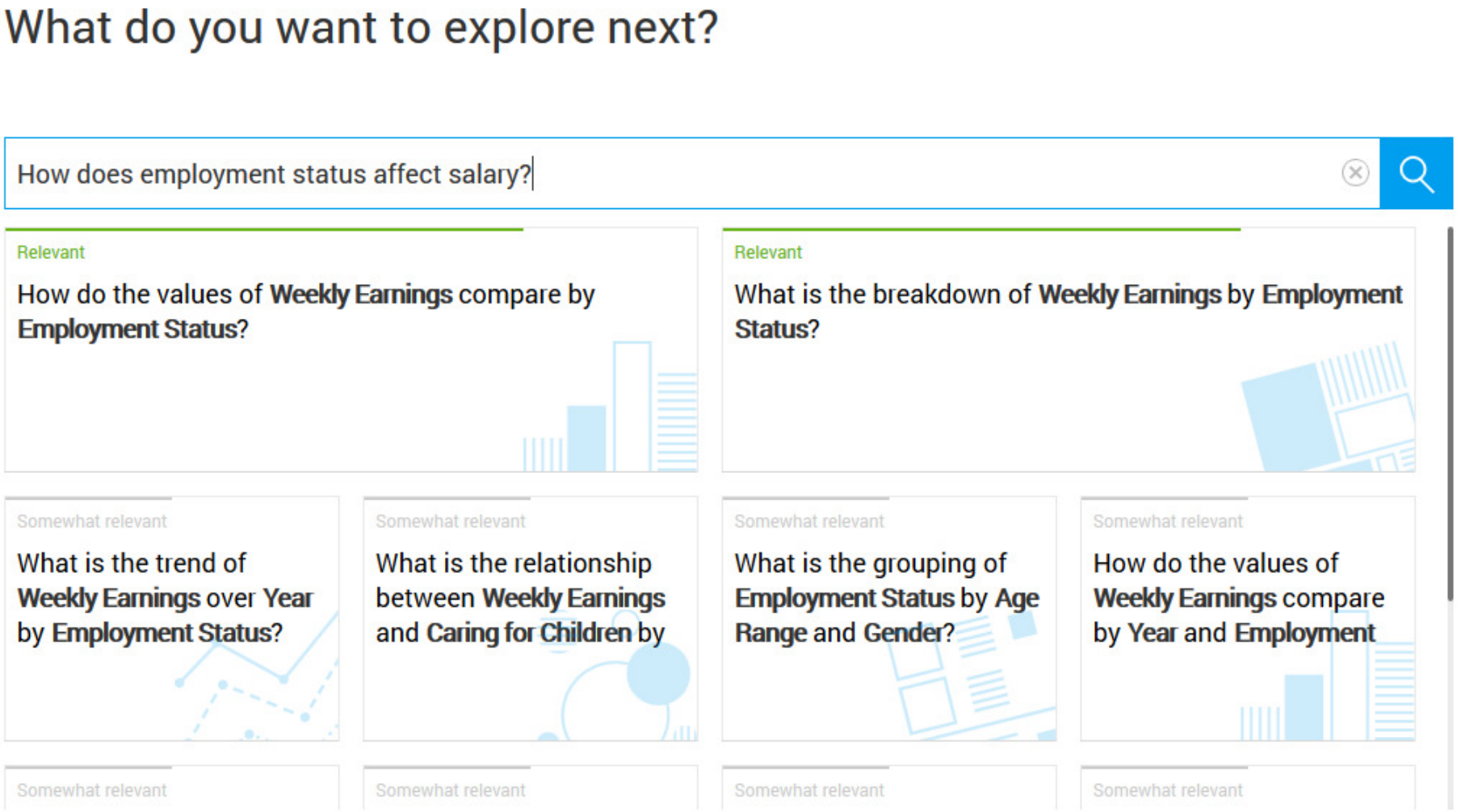

What is the trend of Television over Age Range ?
Columns
Television (Average)
275
250
225
200
175
150
125
100
75
50
25
0
0-19
20-29
30-39
40-49
50-59
60-69
70-79
80+
Age Range

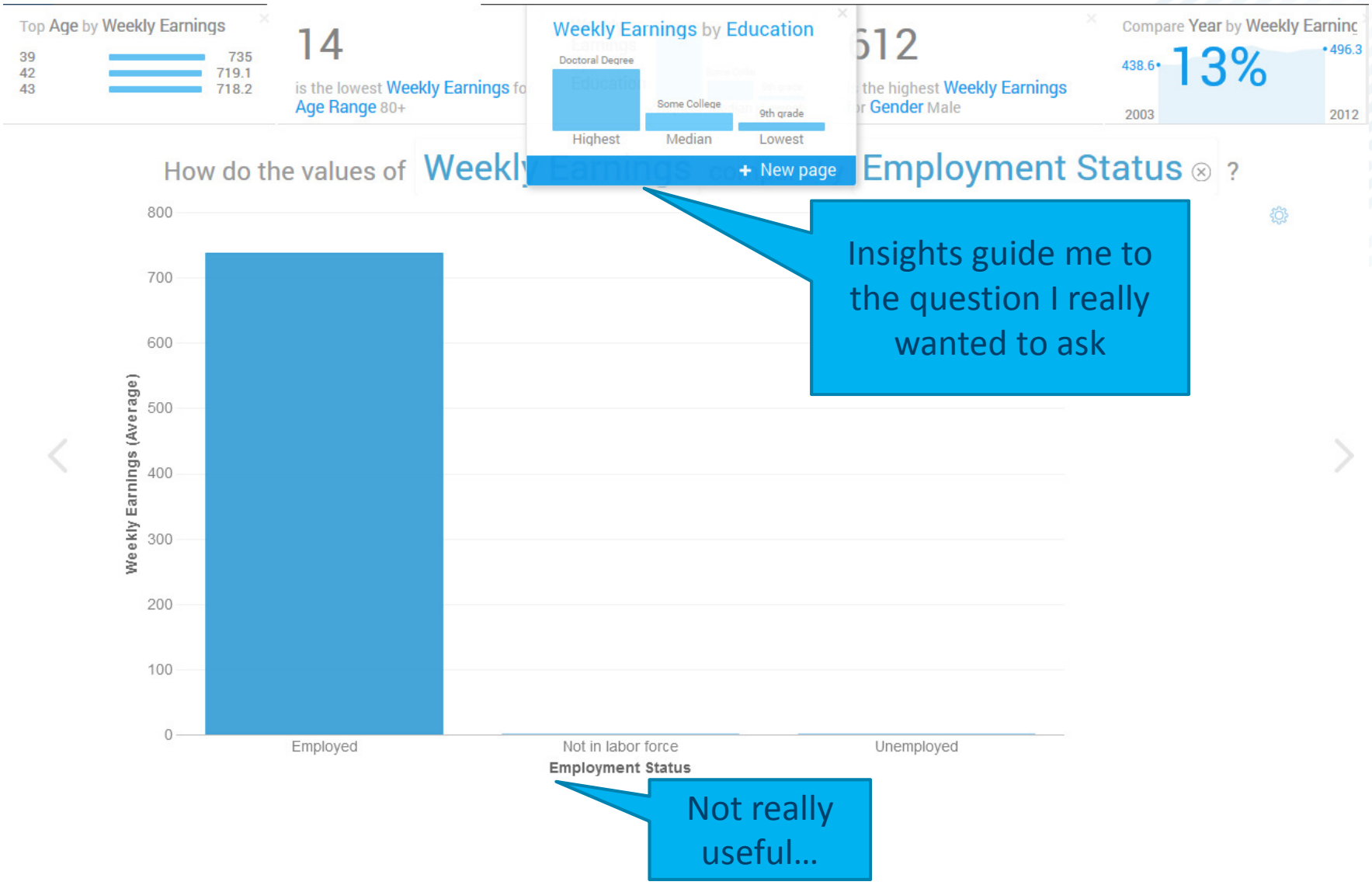
# Visualization Coach, suggesting a better visualization type



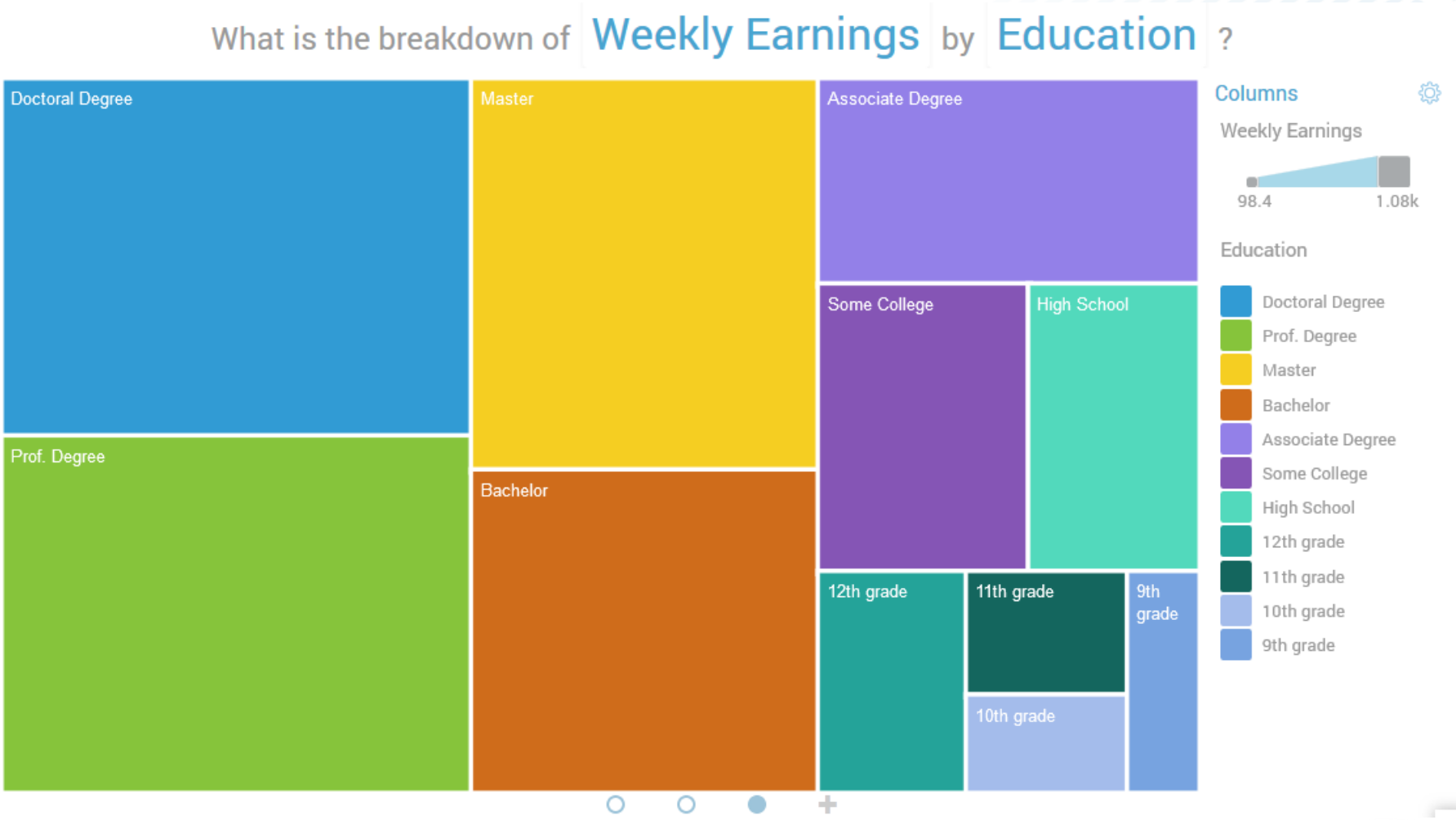


# Built-in data semantics: salary ~ weekly earnings



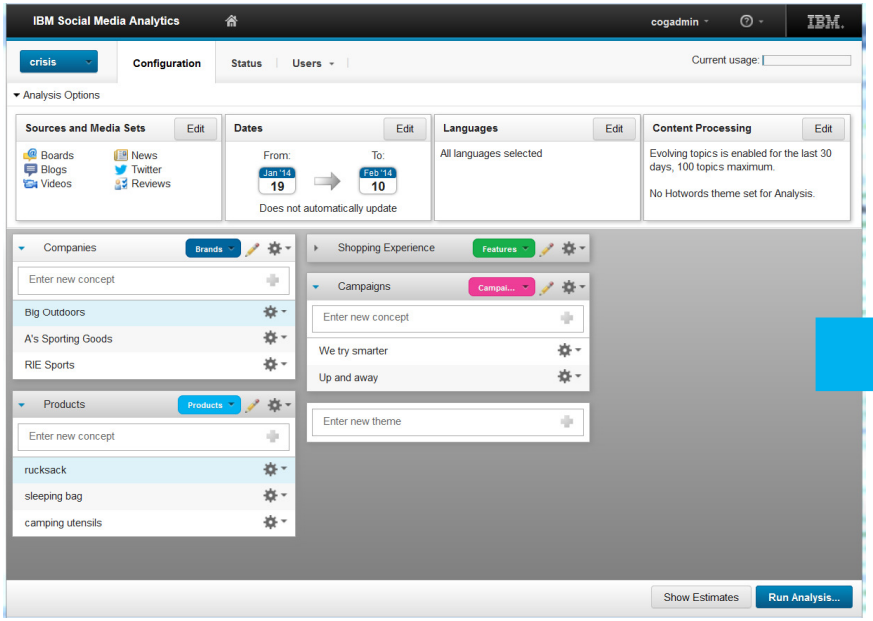

Top Age by Weekly Earnings
39 735
42 719.1
43 718.2
14
is the lowest Weekly Earnings fo
Age Range 80+
Weekly Earnings by Education
Doctoral Degree
Some College
9th grade
Highest
Median
Lowest
+ New page
612
the highest Weekly Earnings
Gender Male
Compare Year by Weekly Earnin
438.6
496.3
13%
2003
2012
How do the values of Weekly Earnings Employment Status ?
Weekly Earnings (Average)
800
700
600
500
400
300
200
100
0
Employed
Not in labor force
Unemployed
Employment Status
Insights guide me to the question I really wanted to ask
Not really useful…

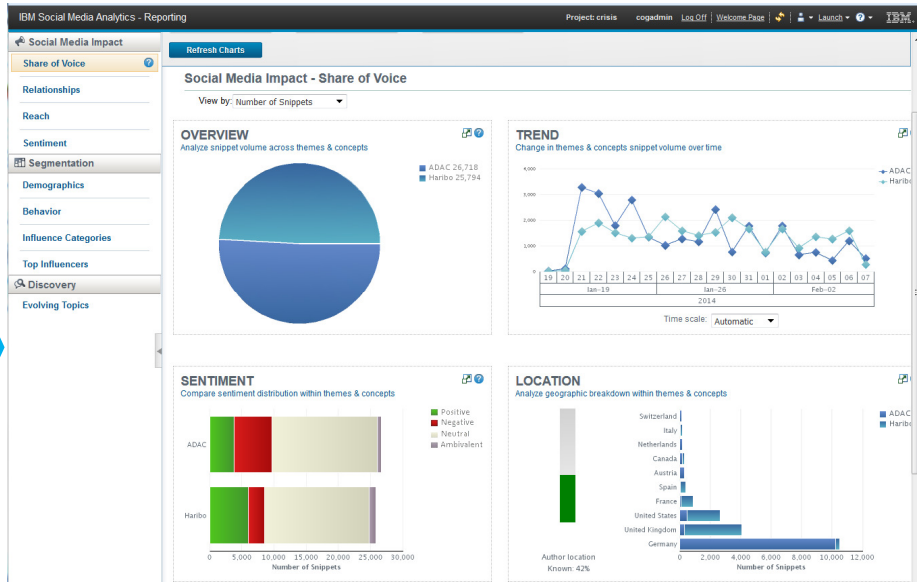
# Grad School still pays off – Yay!



# IBM Social Media Analytics (SMA)

- Analytics Application for Marketing, Sales, Brand Management, Data Scientists
- Analyzes Social Media Content (twitter, facebook, blogs, message boards, reviews, video comments, news) for share of voice, sentiment, customer demographics, behavior, evolving topics in multiple languages
- Both historical analysis as well as continuous updates (every 20 minutes)

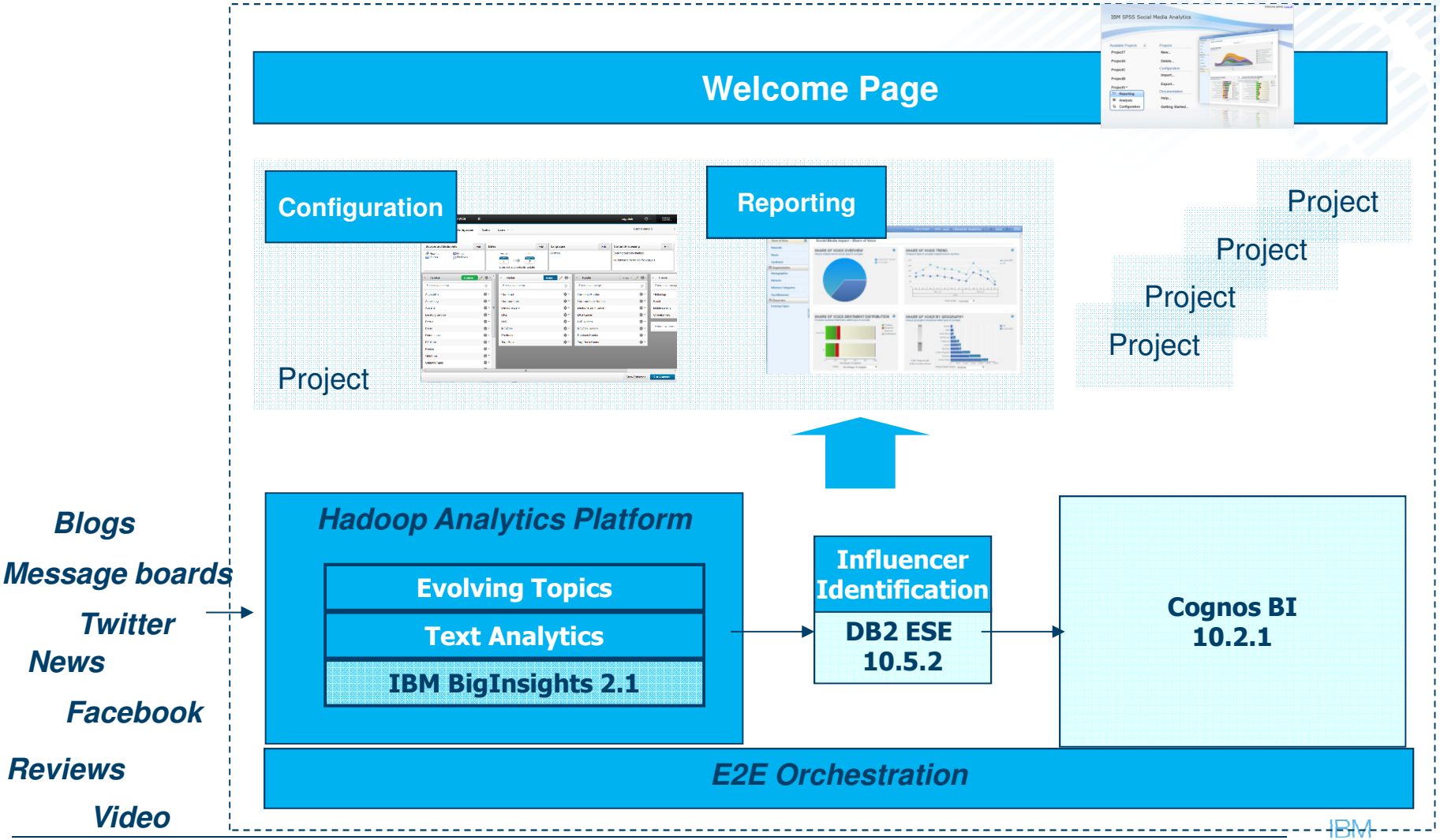



*Configuration* *Analysis*

IBM Software Group

# "Inside" IBM Social Media Analytics



# Hadoop / Warehouse / Big Data? All our users see is:


Application Users don‘t care about the processing platform

(but Data Scientists, Analysts do...)

# IBM Social Media Analytics: Identifying author demographics

- Gender
  - Identified through cues from the author‘s first name, the author‘s nickname and the author content

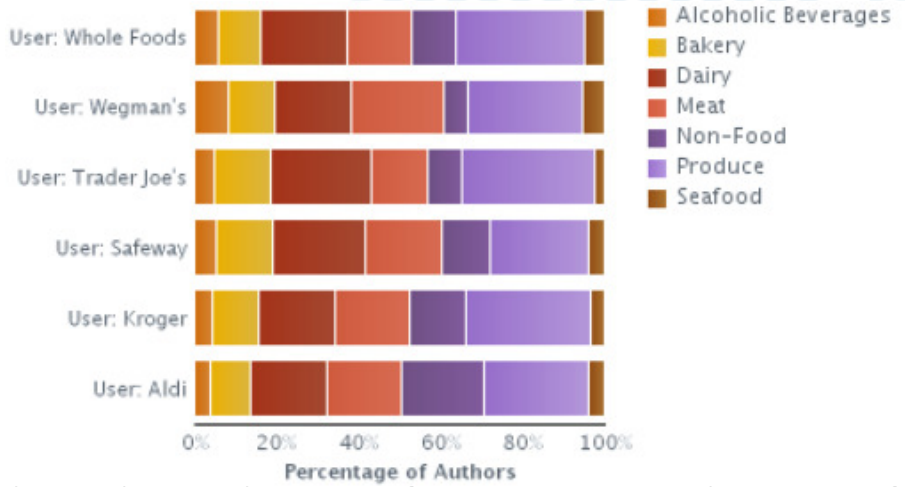
- Is author married or a parent
  - Identified in author content through trigger terms and text analysis rules

**Snippet:** Yes, Google owns a huge chunk of Motorola. This is precisely why my wife's Motorola Droid Razr MAXX is getting the new Android Jelly Bean update before my much more popular and better selling Samsung Galaxy s3

**Snippet:** Just waiting for OTA JB and just rock that. I recall you're on Speakout-my son is also with an unlocked Bell S3. I wonder if his S3 will get the OTA update through the Rogers network/Speakout?

# IBM Social Media Analytics: Identifying author behavior

- Users of a certain product or service
  - Authors mentioning „my X“ and other expressions
  - What product features are relevant for them?




*Author name*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | http://www.youtube.com/ | 1 | Unknown | Unknown | Unknown | not available | User: Whole Foods | purchased from Whole Foods |
| | http://www.chow.com | 2 | Unknown | Unknown | Unknown | not available | User: Whole Foods | my Whole Foods |
| | http://www.nasioc.com | 1 | Unknown | Unknown | Unknown | United States OH MILFORD | User: Whole Foods | our local Whole Foods |

- Recommenders / Detractors
  - „you should use X“ / „stay away from X“
- Prospective users
  - Potential sales leads for 1:1 engagement

# Big Data – Little Impact?

- The key "Big Data" impact: organizations finally realize that their data has untapped potential
  - "Big Data" can mean: many small data problems in parallel
- Data size does NOT matter – it's the *actionability* of the *analysis* that counts
- Big data will have a big impact *only* if we can make data analytics become mainstream

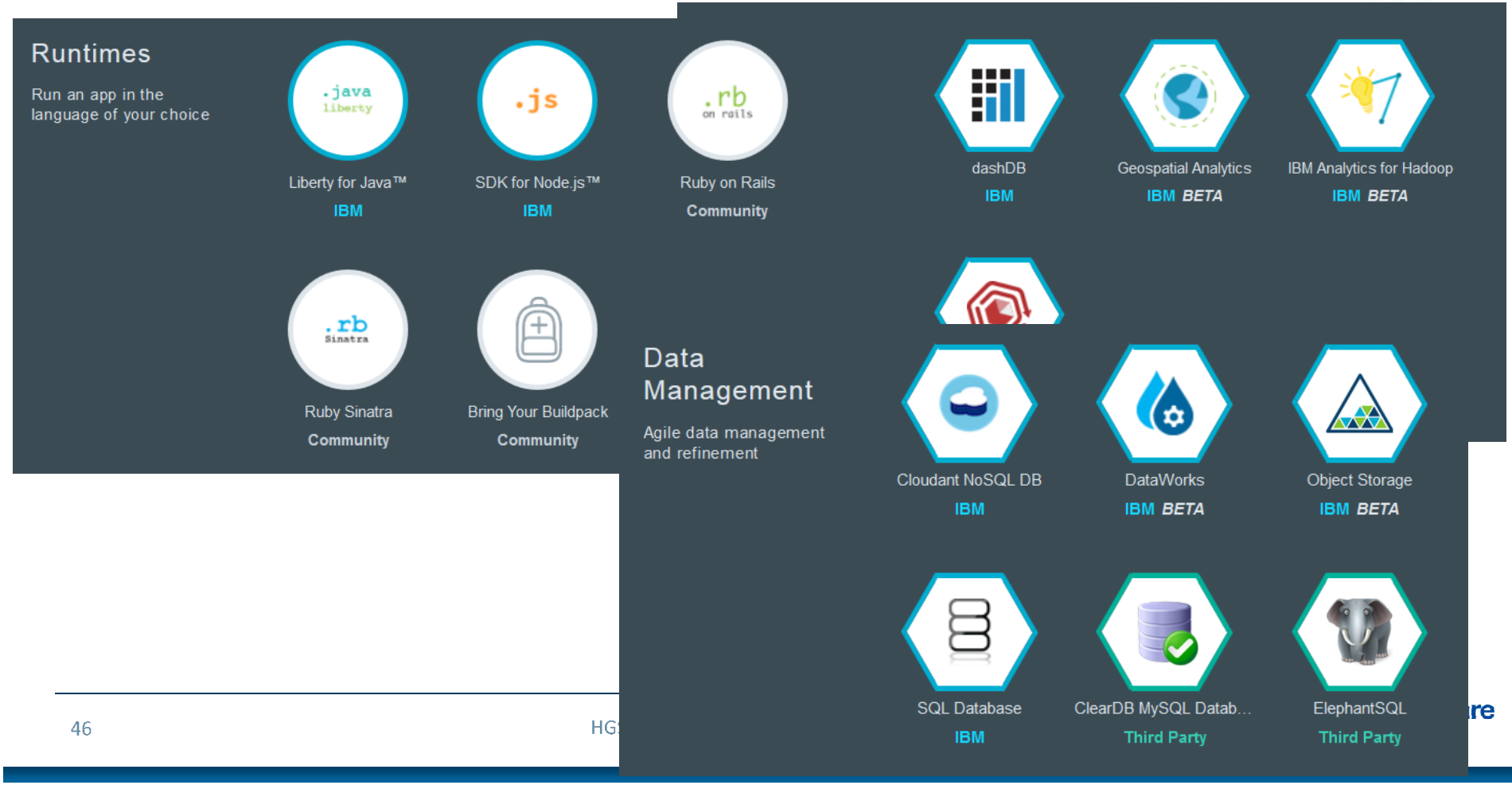
# Backup – the second piece of the technology equation…

- Make it easy for **everybody** to have a meaningful conversation with data

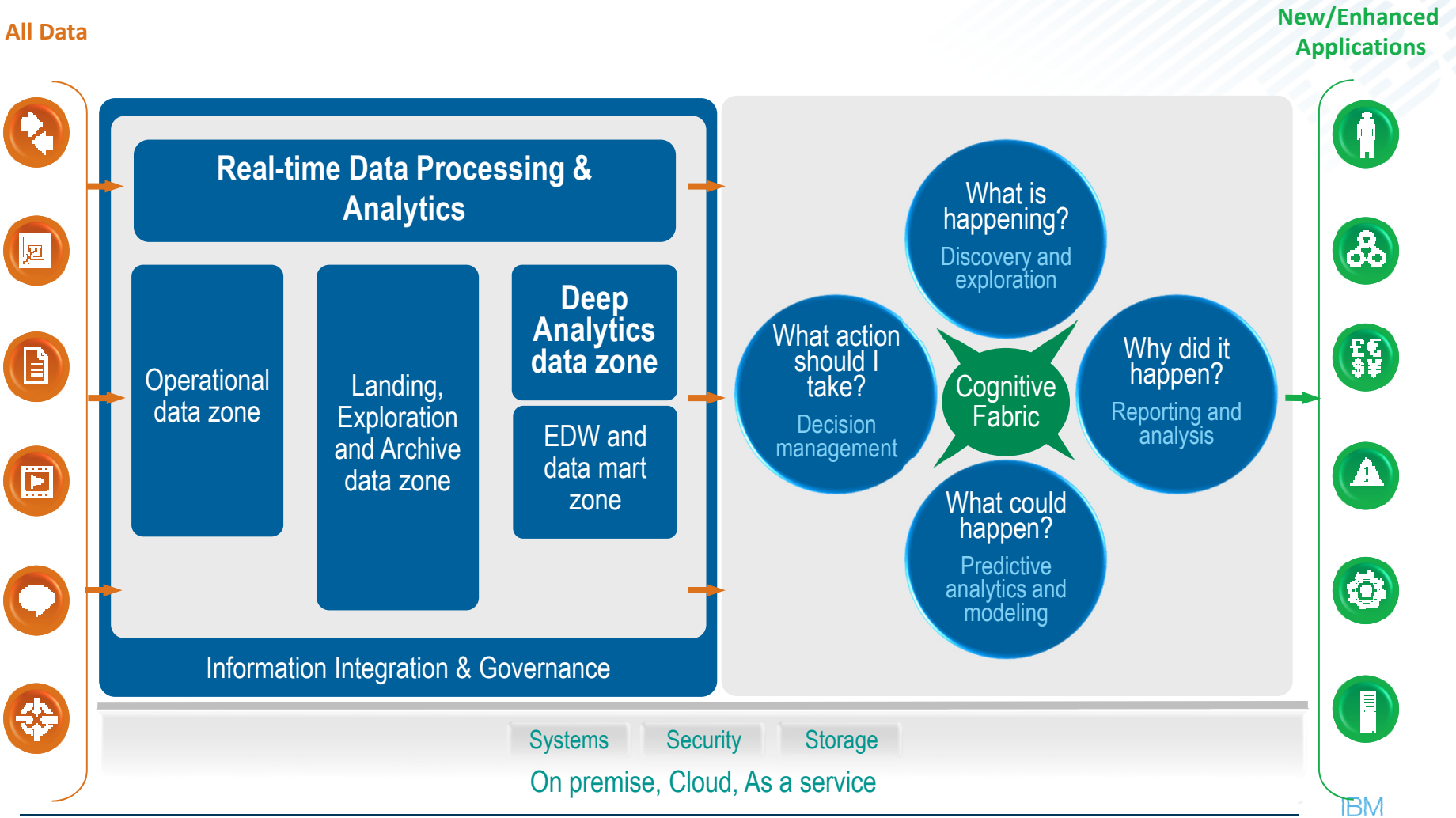
- **Make it easy for Data Scientists to create Big Data Analytics applications**

# IBM BlueMix - Sign Up at https://www.ng.bluemix.net

- Open-standards, cloud-based platform for building, managing, and running apps of all types – including analytic apps



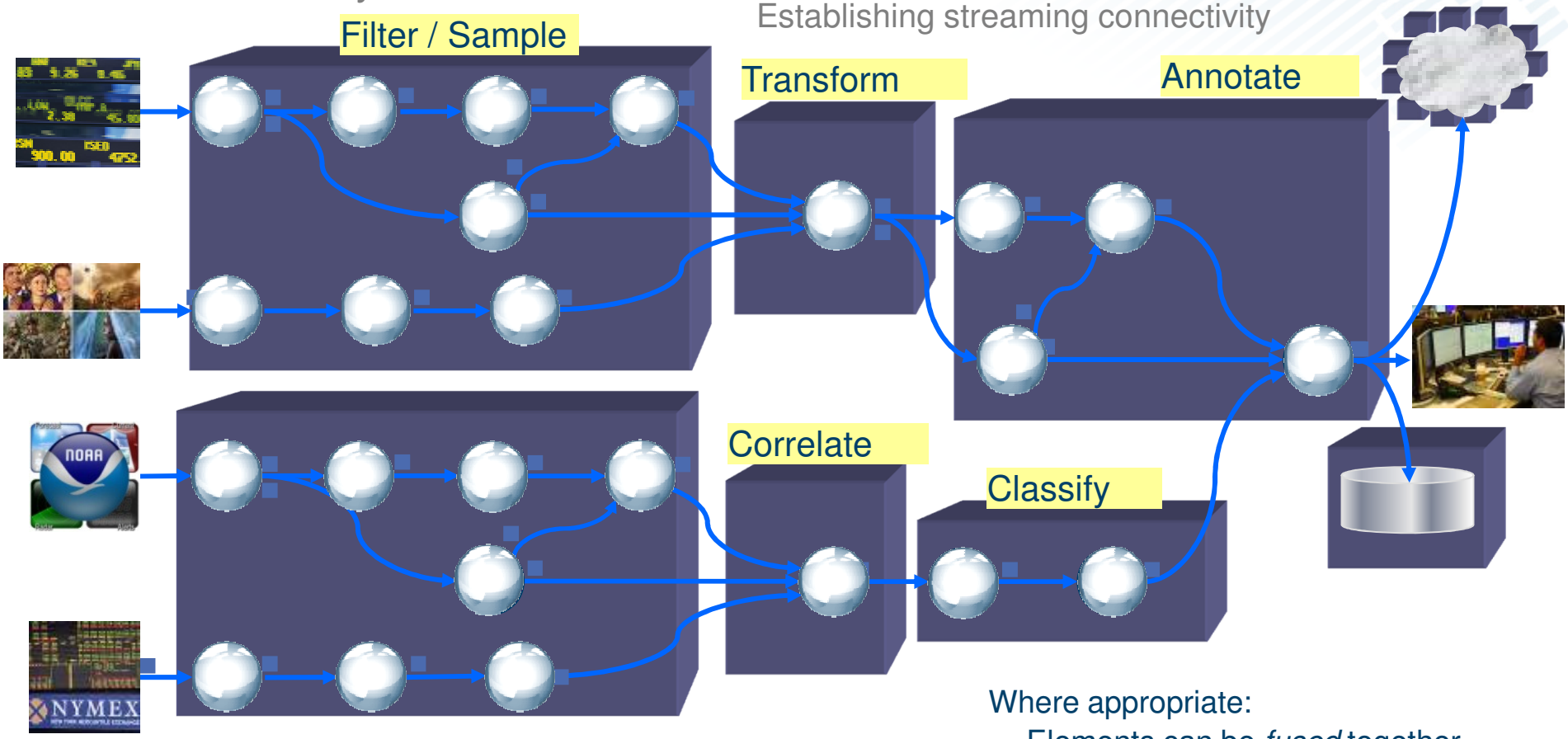
# IBM's big data platform



# InfoSphere Streams – Bringing the Data to the Analytics

→ Continuous ingestion
→ Continuous analysis

Infrastructure provides services for
Scheduling analytics across hardware hosts,
Establishing streaming connectivity

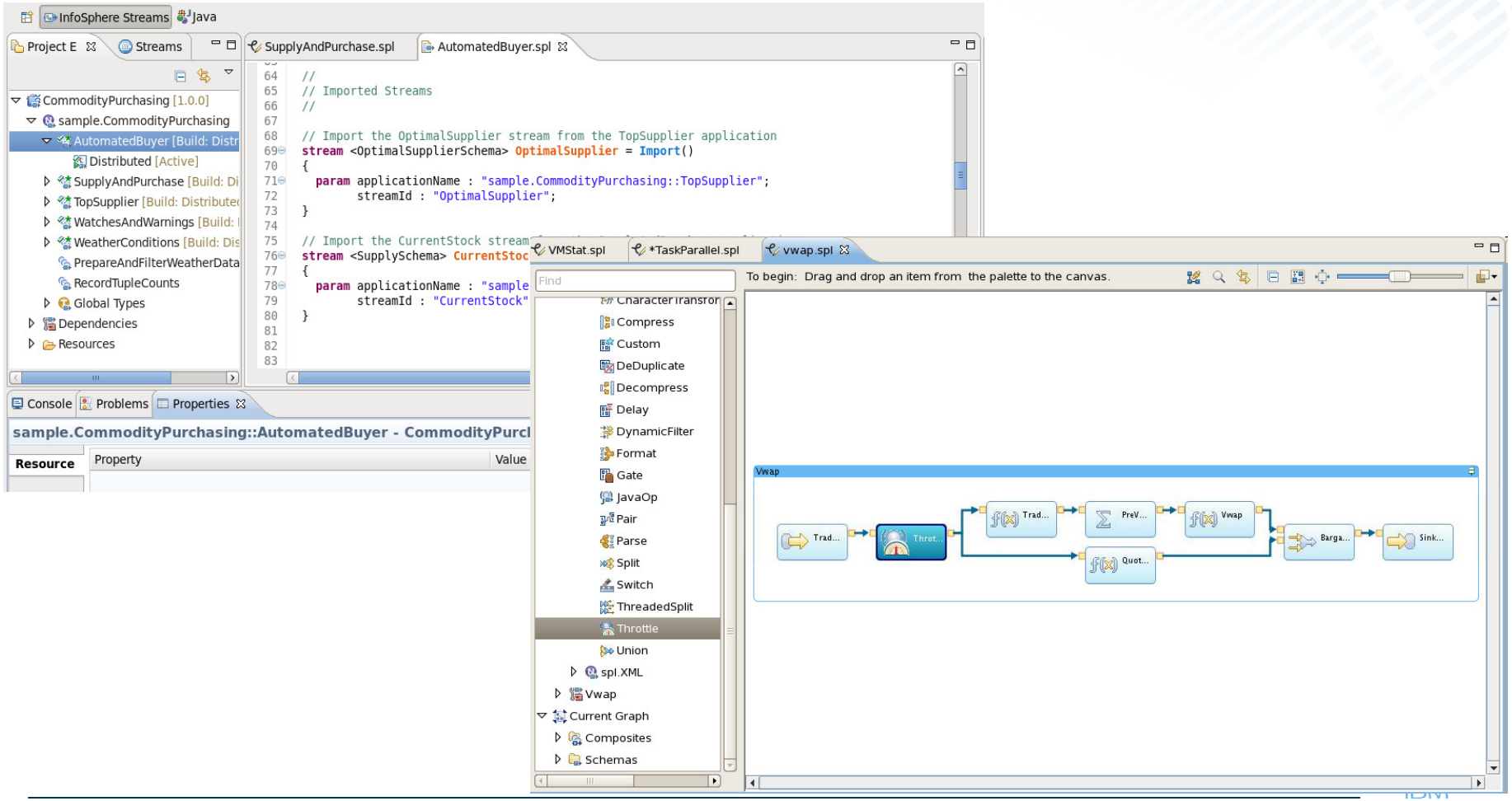



Achieve scale:
By partitioning applications into software components
By distributing across stream-connected hardware hosts

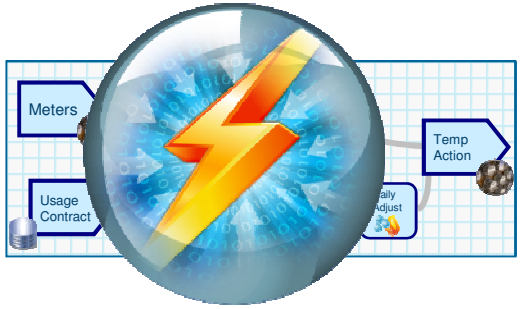
Where appropriate:
Elements can be *fused* together
for lower communication latency

# InfoSphere Streams Studio

- **Eclipse-based tool that enables developers to create stream applications**

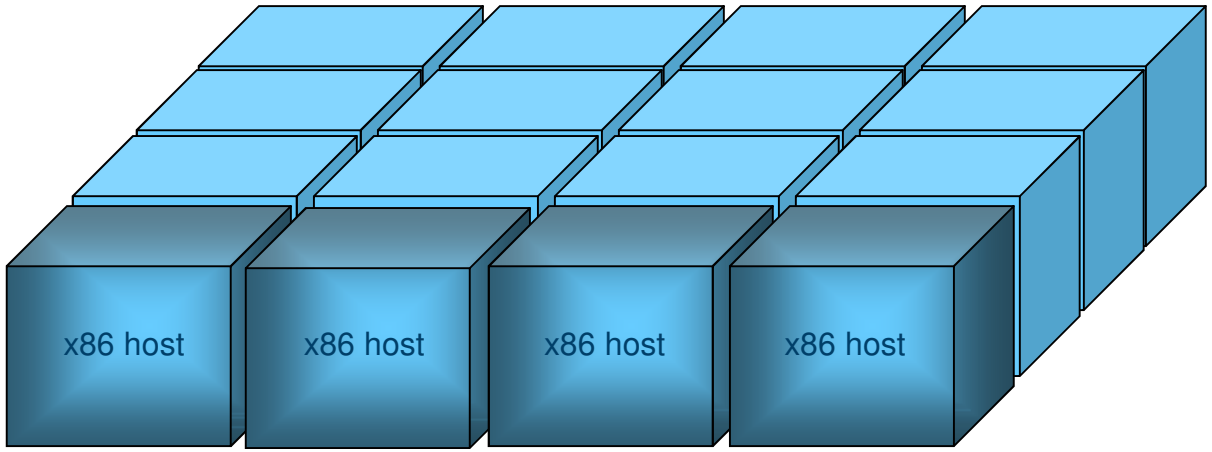


# InfoSphere Streams – Runtime




Optimizing scheduler assigns jobs to hosts, and continually manages resource allocation

Commodity hardware – laptop, blades or high performance clusters



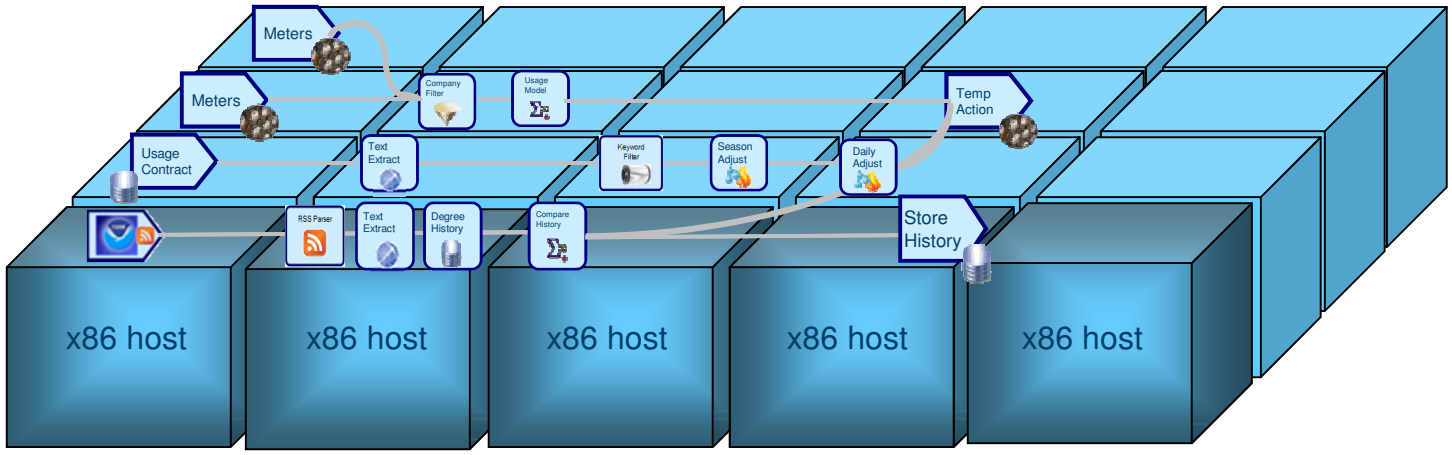
# InfoSphere Streams – Runtime

Optimizing scheduler assigns PEs to hosts, and continually manages resource allocation

Dynamically add hosts and jobs

Commodity hardware – laptop, blades or high performance clusters

New jobs work with existing jobs




# University of Ontario Institute of Technology (UOIT) Detects Neonatal Patient Symptoms Sooner

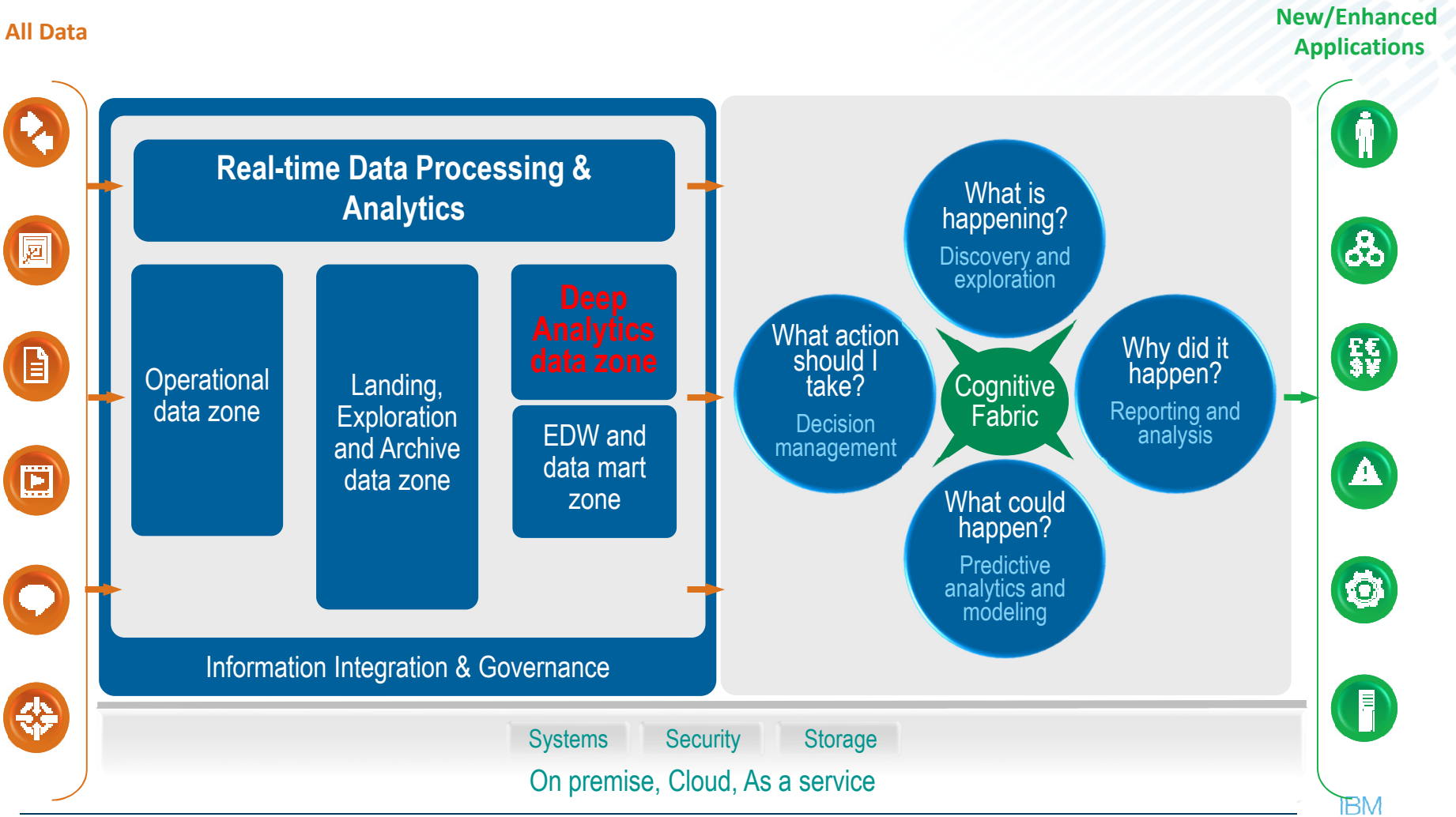
- Performing real-time analytics using physiological data from neonatal babies
- Continuously correlates data from medical monitors to detect subtle changes and alert hospital staff sooner
- Early warning gives caregivers the ability to proactively deal with complications

## Significant benefits:

- Helps detect life threatening conditions up to 24 hours sooner
- Lower morbidity and improved patient care

UOIT
CHALLENGE INNOVATE CONNECT

# IBM's view of a big data platform – Deep Analytics Zone



# IBM InfoSphere BigInsights: 100% Open Source Hadoop + Value-Adds


- **Scalable**
  - New nodes can be added on the fly
- **Affordable**
  - Massively parallel computing on commodity servers
- **Flexible**
  - Hadoop is schema-less, and can absorb any type of data
- **Fault Tolerant**
  - Through MapReduce software framework

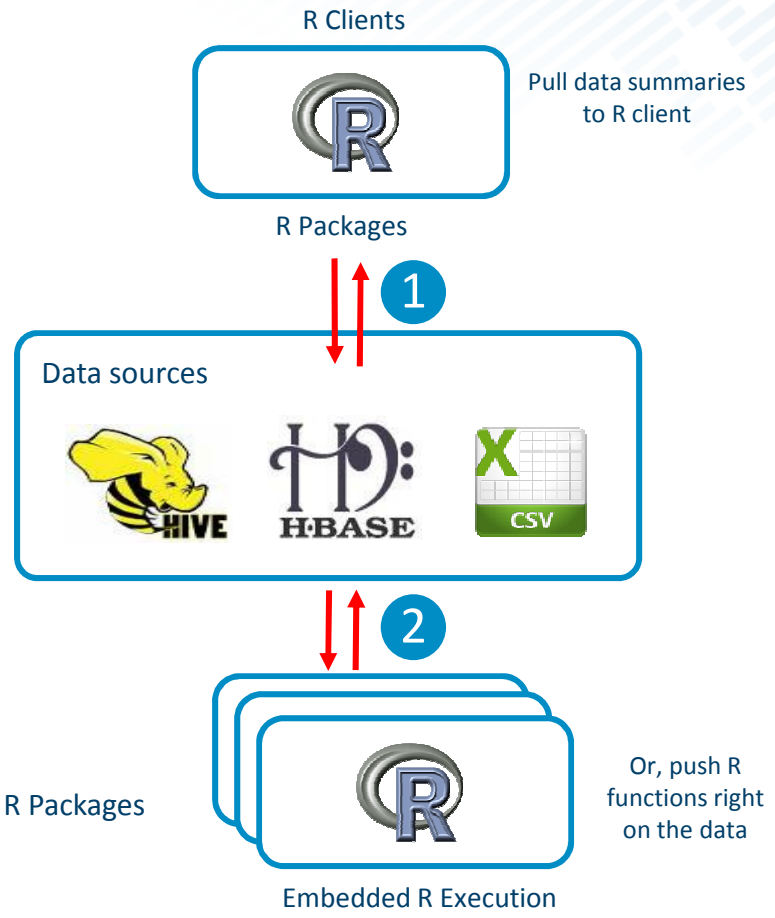
+

IBM Innovation

- **Performance & Reliability**
  - Compression, Flexible Scheduler
- **Enterprise Hardening of Hadoop**
  - Adaptive MapReduce
  - GPFS
- **Deep Analytics**
  - Text Analytics
  - Big R
  - Big SQL
- **Enterprise Integration**
  - Connectors to other Enterprise SW

# IBM InfoSphere BigInsights – Big R

- Explore, visualize, transform, and model big data using familiar R syntax and paradigm
- Scale out R
  - Partitioning of large data ("divide")
  - Parallel cluster execution of pushed down R code ("conquer")
  - All of this from within the R environment (Map/Reduce is hidden from you)
  - Almost any R package can run in this environment



# Big R Machine Learning API

| Big R functions | Inspired by R's | Algorithm |
|---|---|---|
| `bigr.lm()` | `lm()` | Linear regression |
| `bigr.glm()` | `glm()` | Generalized Linear Models |
| `bigr.logistic.regression()` | `glm()` | Multi-class Logistic Regression |
| `bigr.kmeans()` | `kmeans()` | K-means clustering |
| `bigr.naive.bayes()` | `naiveBayes()` | Naïve Bayes classifier |
| `bigr.svm()` | `svm()` | Support Vector Machine classifier |
| `bigr.univariateStats()` | - | Central tendency, dispersion, skewness, kurtosis… |
| `bigr.bivariateStats()` | - | Pearson's correlation, F-test…. |
| `bigr.sample()` | `sample()` | Uniform sample by percentage, exact number of samples, or partitioned sampling. |
| `bigr.transform()` | - | Recoding, dummy-coding, binning, scaling, and null value imputation |

* Machine learning components based on System ML accessible as Big R function in current beta release

# Don't believe me, try it out for yourself!
# BigInsights on IBM BlueMix

- Aimed at developers - No cluster management expertise required
- IBM Analytics for Hadoop in minutes on the Cloud

IBM Analytics for Hadoop

IBM

PUBLISH DATE
8/22/2014

TYPE
Service

VIEW DOCS

BETA

Analyze and visualize Big Data on Hadoop without having to configure or administer clusters.

- **Immediately build Big Data applications!**

  This service provides an easy way to access data on Hadoop clusters, build applications, and analyze structured or unstructured data. Visualize your findings in charts and graphs. You can bring your data into Hadoop for analysis without worrying about setting up or configuring Hadoop.

- **Built on open source technology**

  This service is powered by InfoSphere® BigInsights™, which is based on open source Hadoop. It provides open source capabilities of Hive, MapReduce, Pig and others. In addition, you can access enterprise capabilities from BigInsights including Big SQL, BigSheets, Text Analytics, and more.