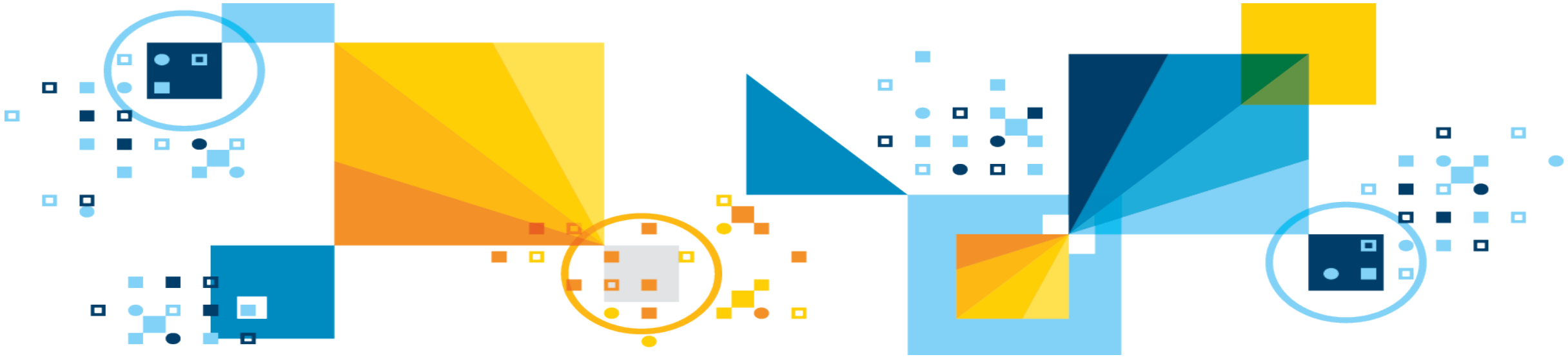
Data Science Workshop
with IBM SPSS Modeler

Add Date and Location and Presenter Name



Agenda

- Welcome and Introductions
- Overview of SPSS Data Science
- IBM SPSS Modeler Overview
- Exercise 1: Predictive in 20 Minutes
- IBM SPSS Modeler Basics
- Methodology and Applications
- **Break**
- Exercise 2: Find Patterns and Groups
- Exercise 3: Understand the Past, Predict the Future
- Exercise 4: Deploy Insights
- SPSS Data Science Platform Integrations
- Wrap up and Next Steps
- Resources

Introductions

- Name and Organization you represent
- Does your Organization have any data science initiatives?
- Do you have any experience with any IBM SPSS solutions?
  - SPSS Statistics
  - SPSS Modeler (formerly Clementine)
  - SPSS Text Analytics for Surveys (TAfS)
- Do you use Open Source?
- How did you hear about this workshop?
- What are your objectives for today?

Today's Presenter: Lorem Ipsem

XXXX@us.ibm.com

XXX-XXX-XXX

linkedin.com/in/xxxxxx

Project Experience

- *XXX*
- *YYY*
- *ZZZ*

Purpose Of The Workshop

- Introduction to Data Science
- Stimulate thinking about how SPSS Data Science would benefit your organization
- Demonstrate ease of use of powerful technology
- Get hands on experience with SPSS' Data Science platform
- Explore multiple Data Science techniques
- Understand the SPSS Data Science portfolio and its integrations across IBM

Data Science Is For Everyone

Data Science is the analysis of all kinds of data using sophisticated quantitative methods (i.e., statistics, descriptive and predictive data mining, simulation and optimization) to produce insights that traditional approaches to business intelligence (BI) — such as query and reporting — are unlikely to discover.

Why Is Data Science Essential?

Complexity

A single truck visiting 5 different locations

120 different routes

A single truck visiting 10 different locations

Over 3 million routes

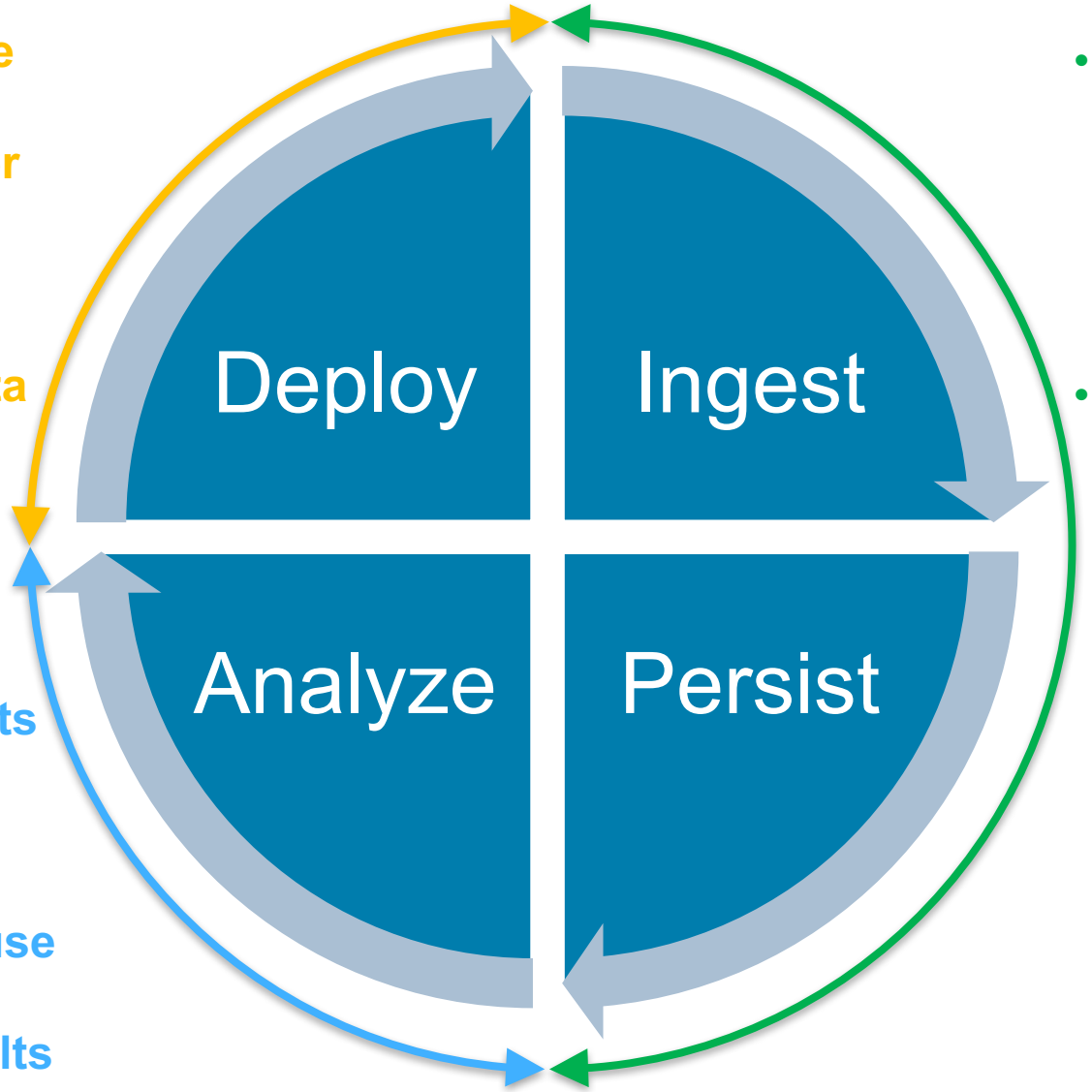
Real world problems contain

100s of trucks, 1000s of locations

There are more possibilities that the grains of sands in the world

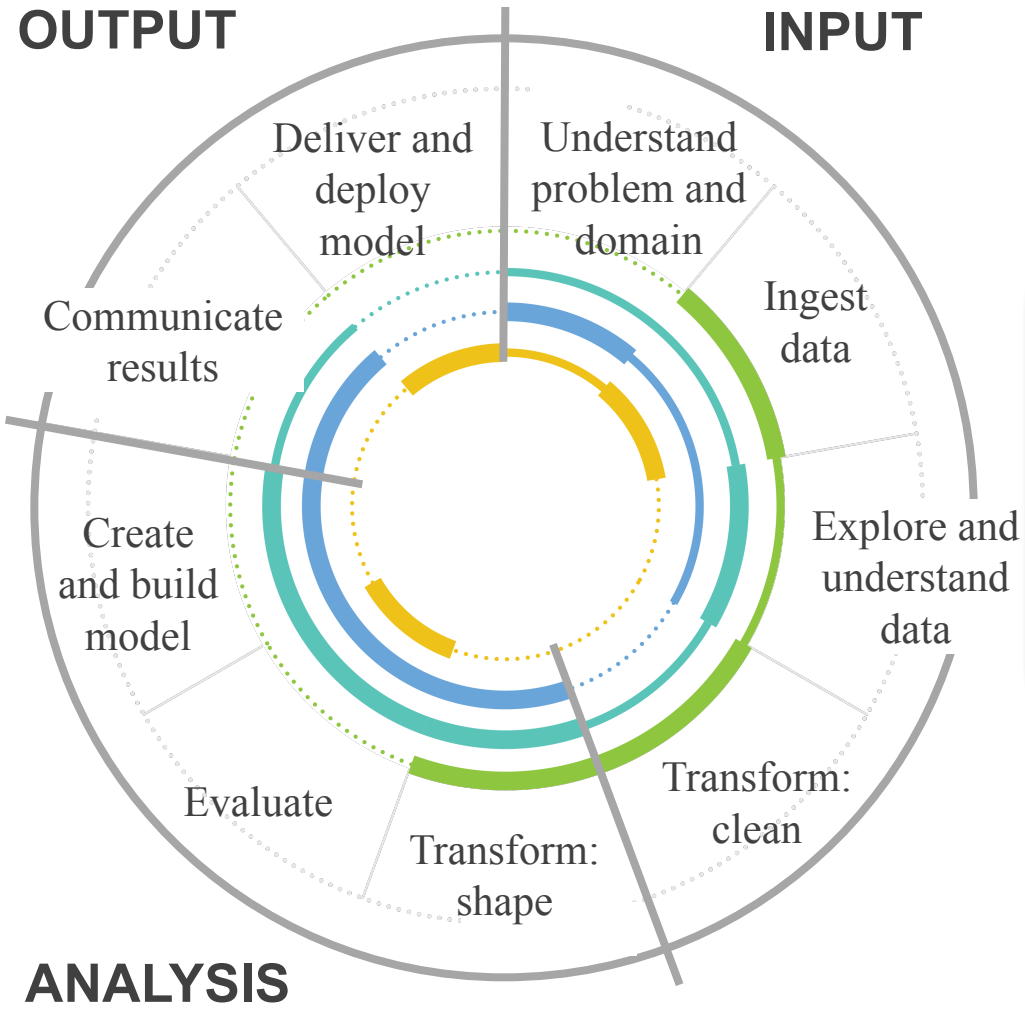
A Data Science Application

- User interface front end for business-user and end-user interaction
- Deployed Data Science models run here
- Data Scientists collaborate, explore data sets, create models and use algorithms to produce results



- Collect and save data from internal and external sources
- Manipulate, govern and expose data sets

SPSS Users



Data Engineer
Architects how data is organized & ensures operability

Data Scientist
Gets deep into the data to draw hidden insights for the business. You can distinguish a technical and citizen data scientist

Business Analyst
Works with data to apply insights to the business strategy

Developer
Plugs into data and models & writes code to build apps

SPSS Use Cases

Evidence-based medicine		• Improving patient care and satisfaction • Reducing costs through optimized allocation of resources • Measuring and improving patient outcomes
Human capital management		• Acquiring, growing and retaining employees • Helping ensure optimal staff levels • Increasing performance, efficiency and engagement
Crime prediction and prevention		• Identifying predictors of threat and fraud • Optimizing force deployment • Anticipating and visualizing crime hot spots
Supply chain management		• Increasing visibility into virtually all areas of the supply chain • Decreasing downtime and unpredictability • Improving customer satisfaction
Process optimization		• Improving accurate responses at the point of impact • Decreasing costs through operational efficiency • Transforming threat and fraud identification processes

Organizations Using SPSS See Results

IBM SPSS CUSTOMER TESTIMONIAL

AT&T Saves Money With IBM SPSS Modeler

> IBM SPSS Modeler has greatly enhanced our organization's ability to identify and address multiple cost saving issues.
>
> — Ed Clarke, Data Scientist, AT&T

Source: Ed Clarke, Data Scientist, AT&T

Validated Published: Sep. 12, 2016 TVID: 64A-606-767

IBM SPSS MODELER CUSTOMER STATISTIC

Provides competitive advantage

67% of IBM SPSS Modeler customers have realized at least a 26% improvement in competitive advantage from predictive analytics.

Source: TechValidate survey of 76 users of IBM SPSS Modeler

Validated Published: Dec. 26, 2016 TVID: B6E-CBF-E5D

IBM SPSS CUSTOMER SATISFACTION

86% of surveyed customers are likely to recommend IBM SPSS.*

86%

Source: TechValidate survey of 79 users

Validated Published: Dec. 7, 2016 TVID: 8AD-E6A-116

*Based on respondents who rated their likelihood to recommend as 7 or higher on a scale of 0 to 10.

IBM SPSS CUSTOMER TESTIMONIAL

IBM SPSS enables faster time to insight

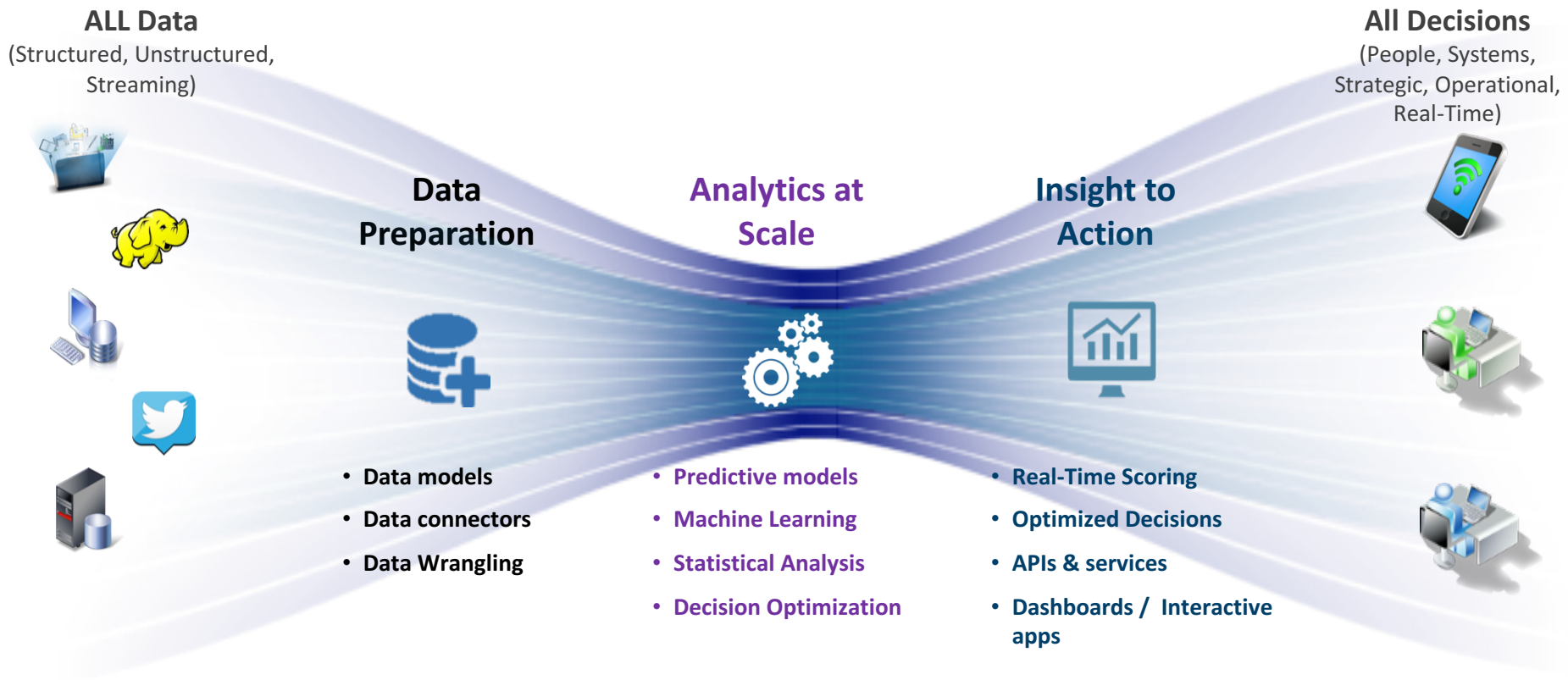
> IBM SPSS Modeler allows our analysts to provide insights at scale efficiently and pragmatically. While others are setting up infrastructure, we are providing insights.
>
> — Jahangeer Iqbal, Data Scientist, IBM Client Innovation Center

Source: Jahangeer Iqbal, Data Scientist, IBM Client Innovation Center

Validated Published: Nov. 14, 2016 TVID: 3E2-C1A-BB7

IBM SPSS Data Science Platform

How does it work?



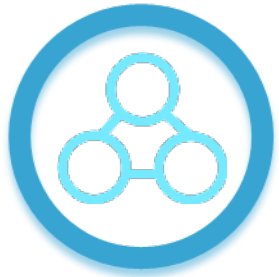
IBM SPSS Modeler – Code-Optional Data Science

Empower

every user

Open

and **extensible**



Deployment

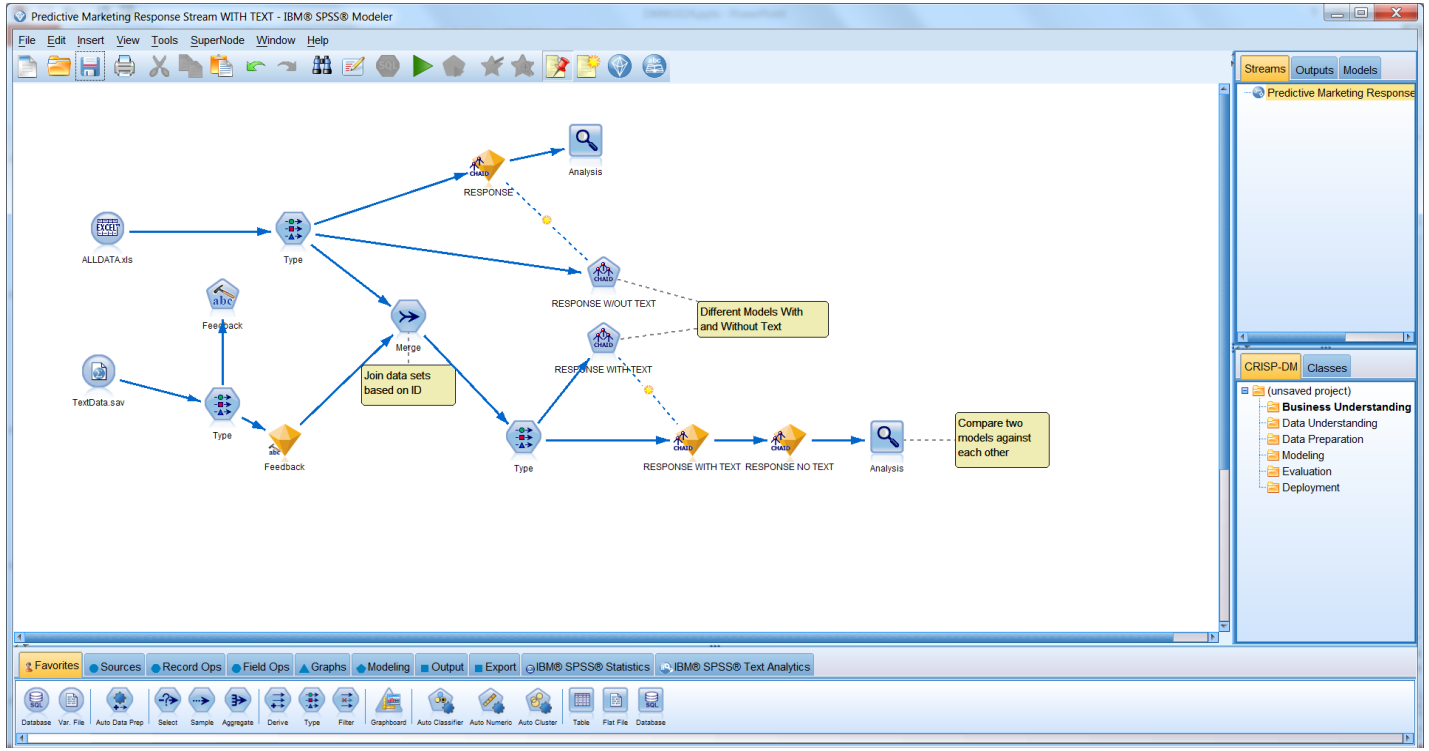
at **scale**

Simplicity without sacrifice

Code optional, open to open source

Predictive everywhere

IBM SPSS Modeler: At A Glance



Flexible adoption and configuration
On premises
Cloud
Hybrid
Bluemix

Scale from personal usage, point solution(s) to enterprise deployment

Comprehensive predictive analytics platform

Identify patterns to generate predictive intelligence and improve outcomes

Access a range of advanced analytics; decision management, text, entity, social network analysis and optimization

Exercise: Predictive In 20 Minutes

Goal:

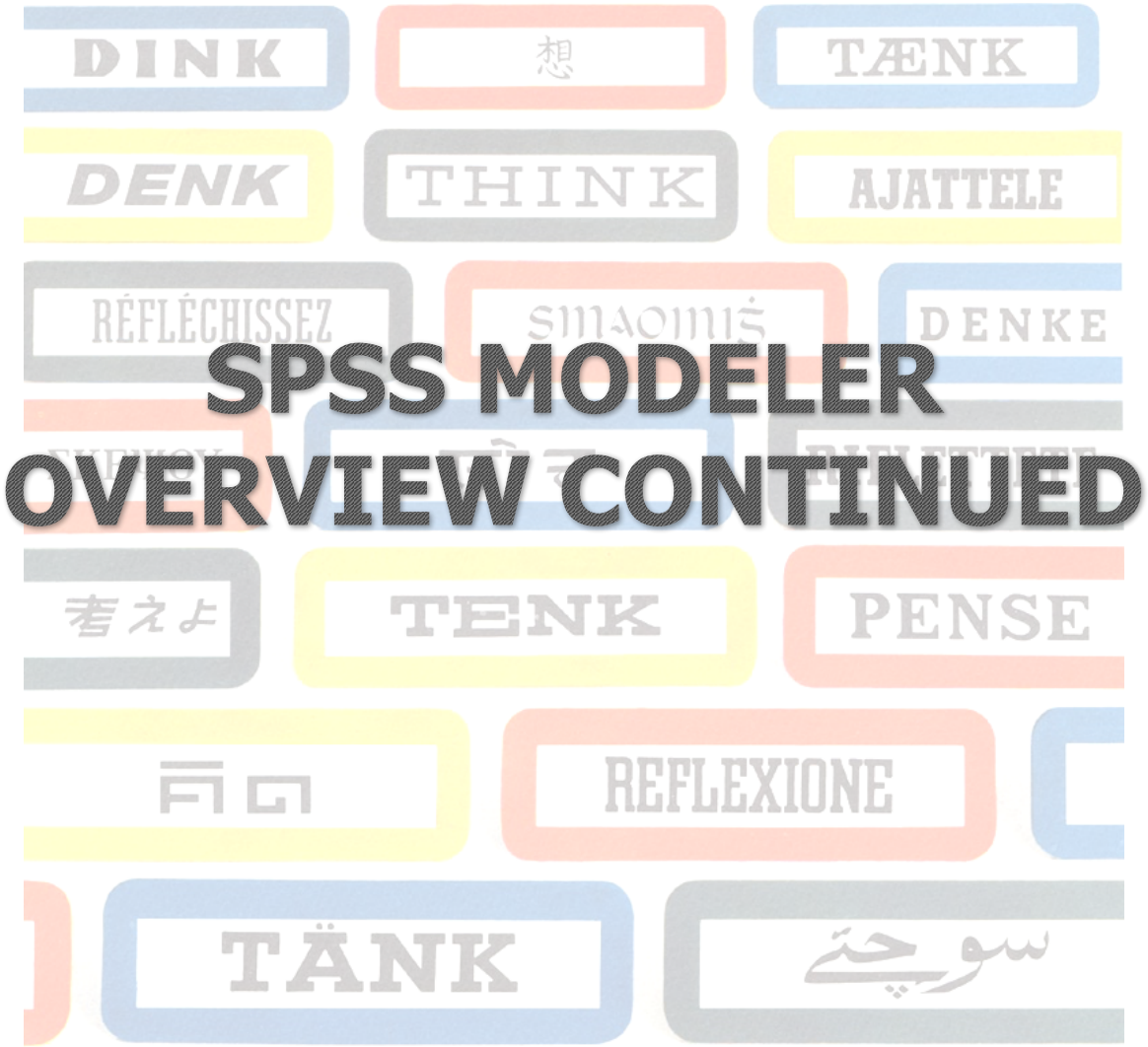
- **Identify who has responded to a marketing campaign**

Approach:

- Use a data extract from a CRM
- Prepare data for modeling
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who has responded
- Review results

Why?

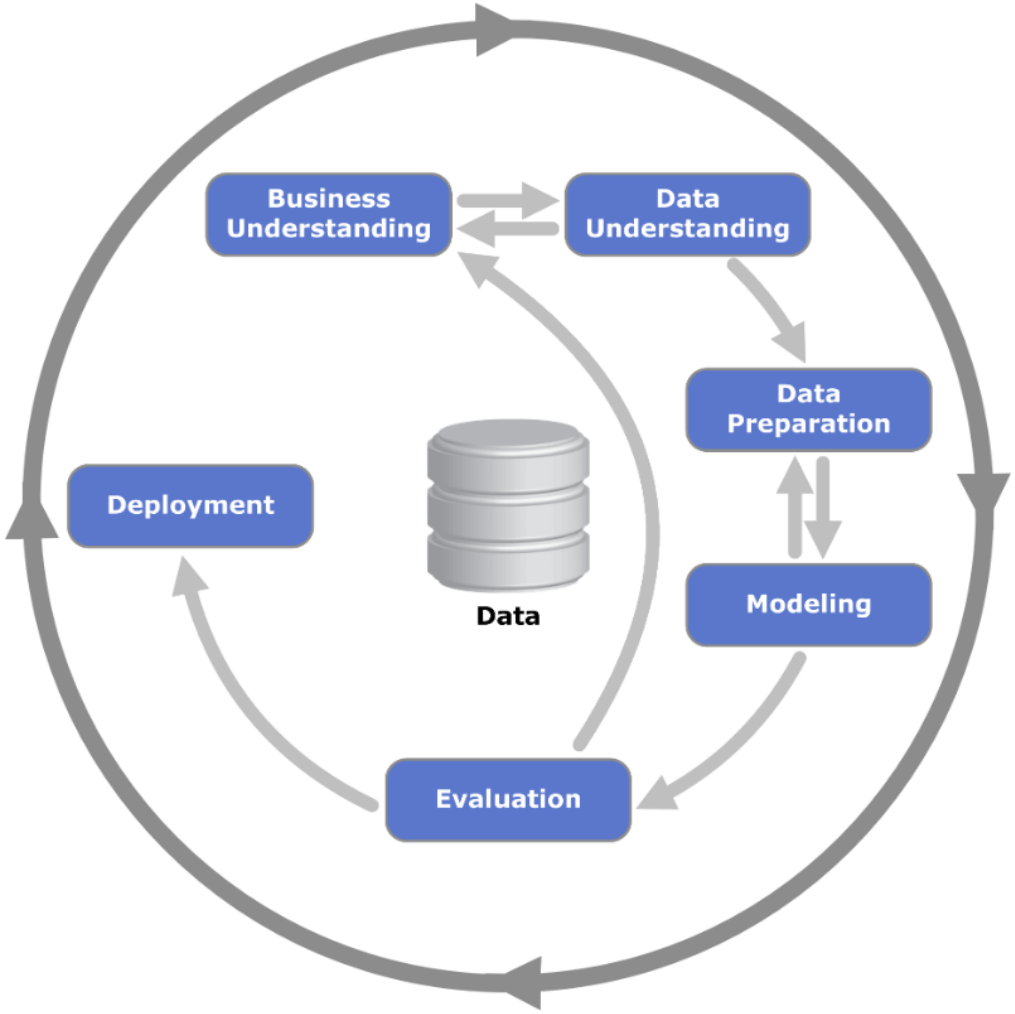
- To save marketing cost and increase marketing response, identify those likely to respond and focus marketing efforts on those prospects.



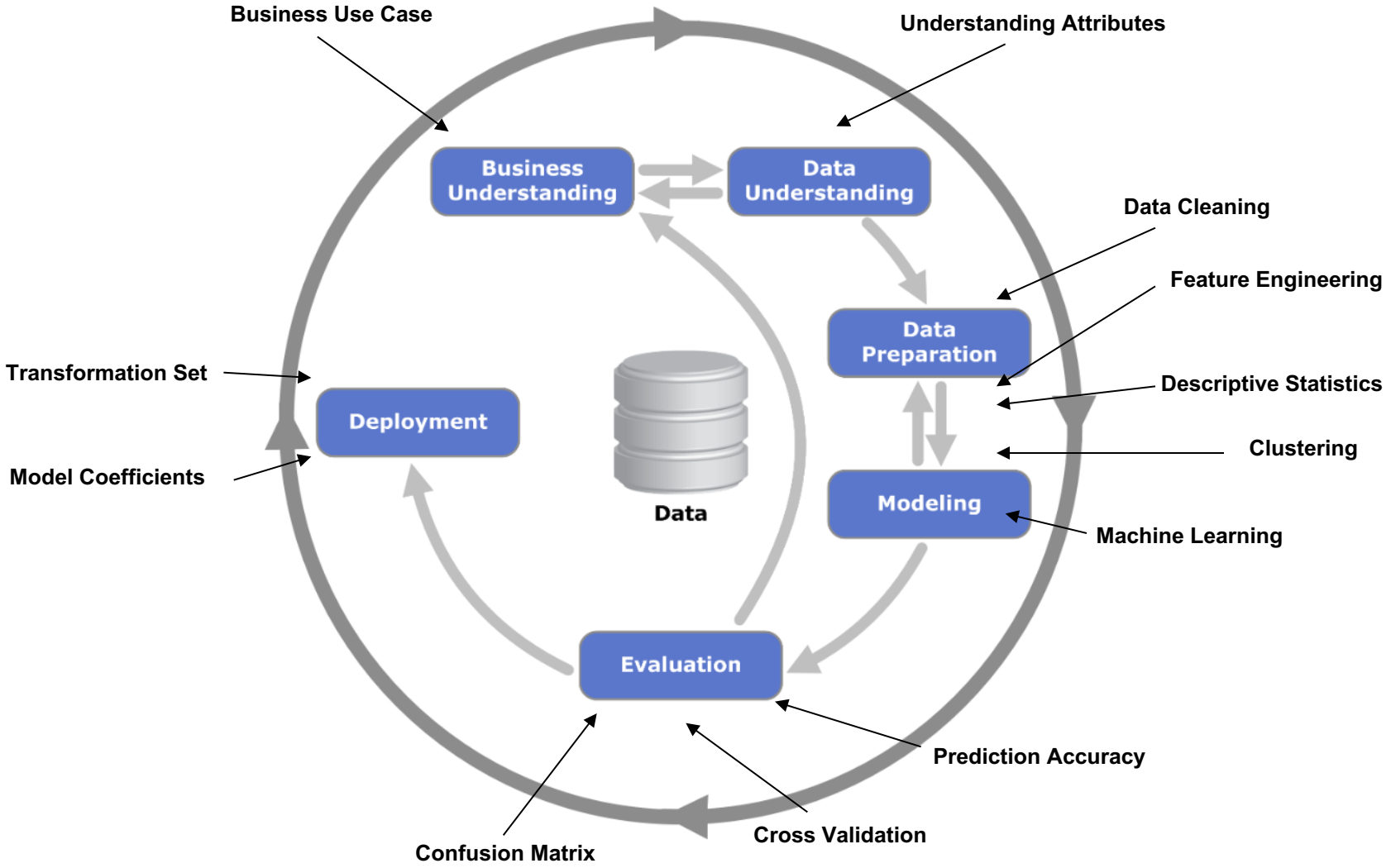
SPSS MODELER OVERVIEW CONTINUED

CRISP-DM: Cross Industry Standard Process for Data Mining

The Traditional Data Science Workflow



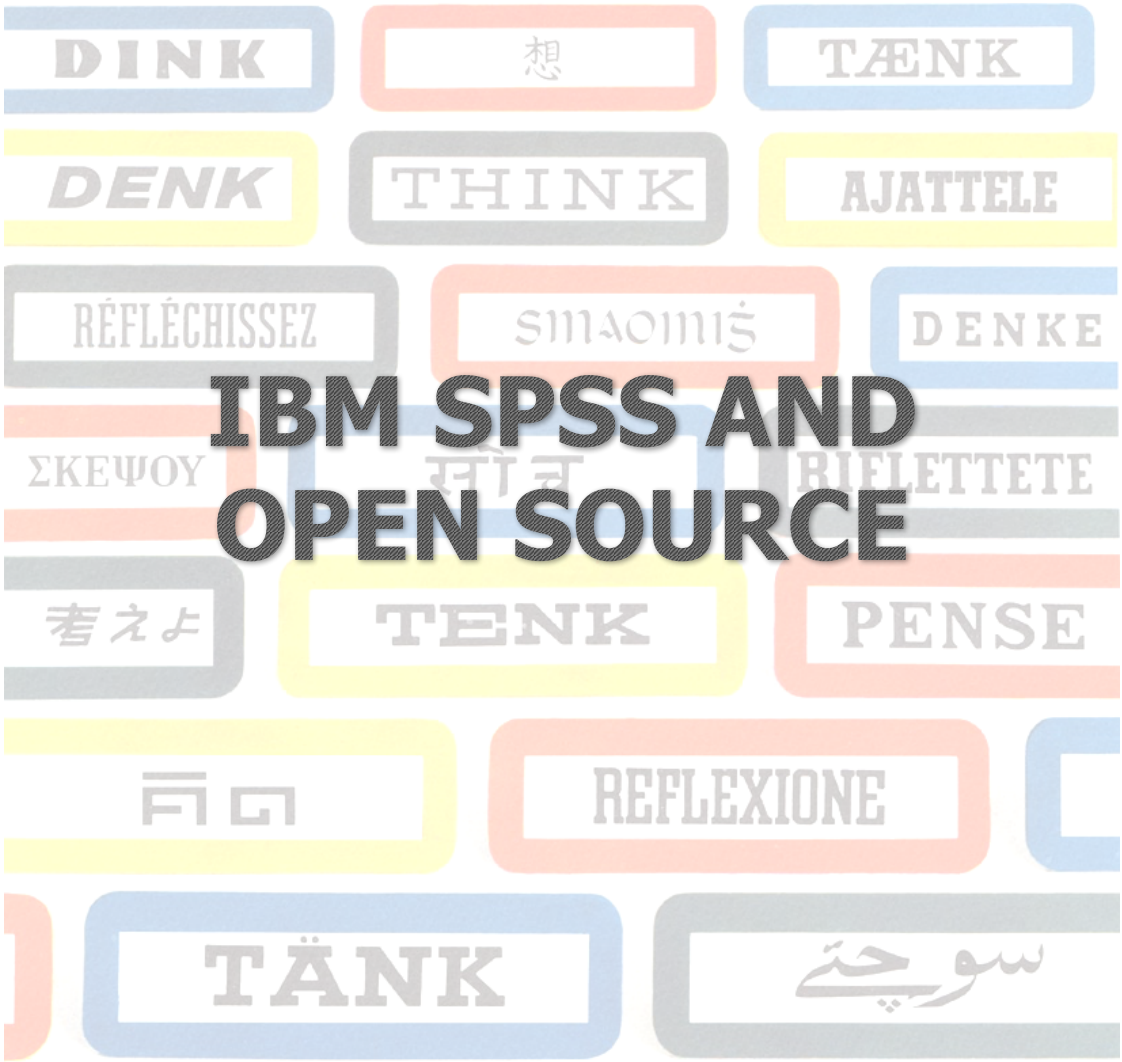
CRISP-DM: Cross Industry Standard Process for Data Mining
The Traditional Data Science Workflow



Analytical Techniques in IBM SPSS Modeler

Technique	Algorithms	Usage
Classification (Or Prediction)	Autoclassifiers, Decision Trees, Logistic, Support Vector Machines, Time Series	Predict Group Membership (e.g., Will This Employee Leave?) Or a Number (e.g., How Many Widgets Will I Sell?)
Segmentation	Autoclusters, K-Means, Anomaly Detection	Classify Data Points Into Groups That Are Internally Homogenous and Externally Heterogeneous, Identify Cases That Are Unusual
Association	Apriori, CARMA, Sequence	Find Events That Occur Together Or In a Sequence Market Basket
Geospatial	Space-Time Boxes	Ability to Improve Model Accuracy (for Any Model Type) By Including Inputs Derived From Geospatial Data Sources
Automated	Autoclassified, Autonumeric, Time Series, Clustering	Automatically Find the Right Algorithms Based On Data and Outcome to Create An Ensemble Model
Simulation	Monte Carlo	Run Different Scenarios to Identify Which Is Best From Historical Data Or Generated Data
Specialized	Text Analytics, Entity Analytics, Social Network Analysis	Improve Overall Model Accuracy
In-database	Netezza, DB2, Oracle, Microsoft	Provide User Friendly Interface On Top of Vendor Algorithms
Open Source	R/Spark/Python	Utilize Open Source Algorithms Within Modeler UI. Enhance By Easily Building Custom Dialogs (or downloading from community)

A single analysis project may include multiple techniques



IBM SPSS AND OPEN SOURCE

IBM SPSS and Open Source

First R, then Python,
now Spark

Make coding optional

Make it massively useful

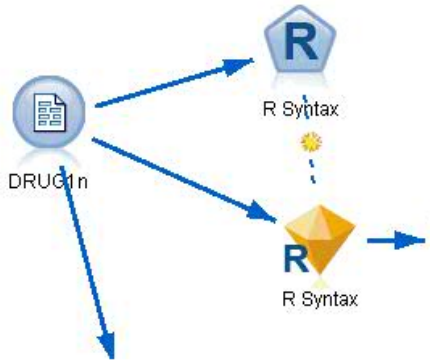
Extend Using Spark, Python Or R

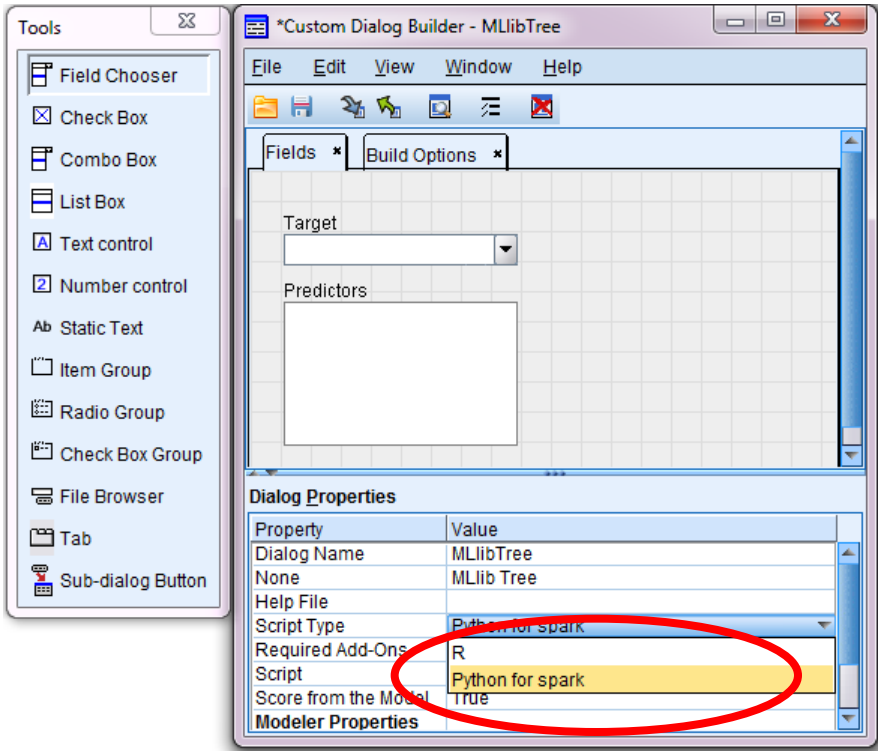
Use script within your modeling stream

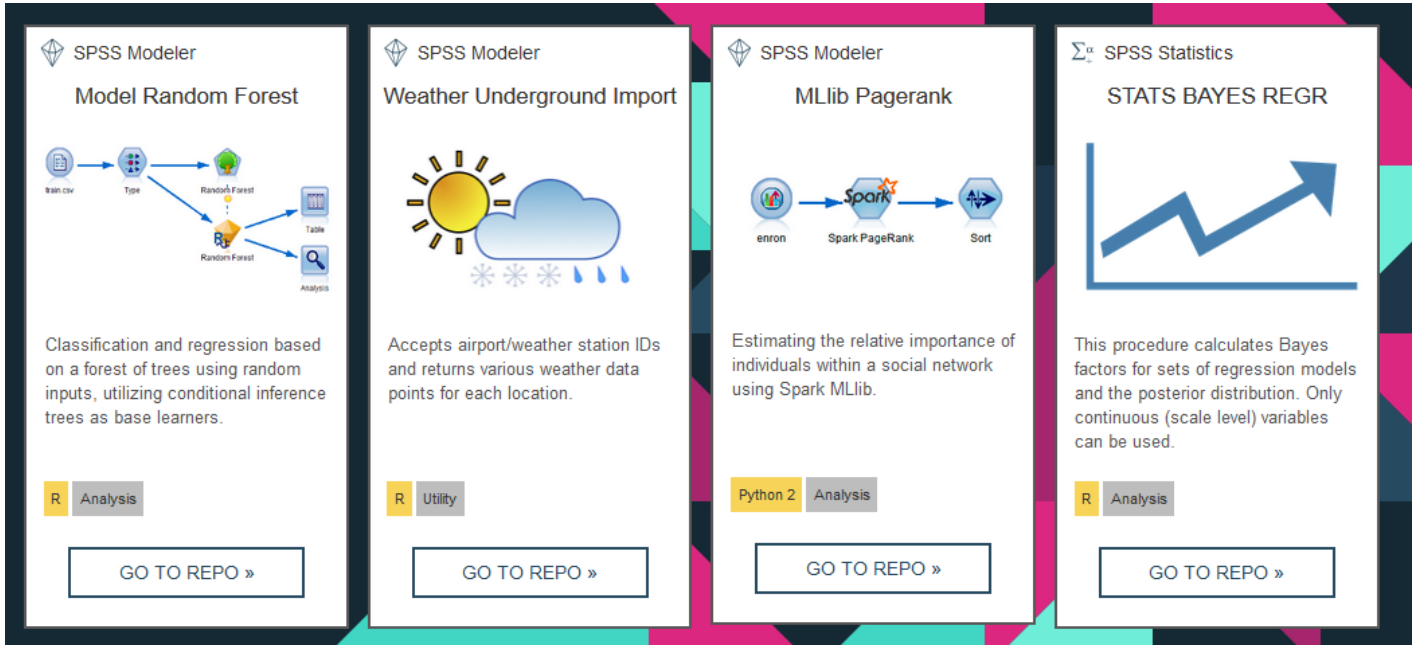
Create extensions for novice users to exploit

- R, Spark's machine learning library (Mllib)
- Other common Python libraries
  - e.g.: Numpy, Scipy, Scikit-learn, Pandas

Abstracting code behind a GUI makes code usable for non-programmers







Growing Catalogue of Predictive Extensions

https://developer.ibm.com/predictiveanalytics

Break - please return in 10 minutes

Exercise 2: Find Patterns And Groups

Goal:

- Create segments of customers

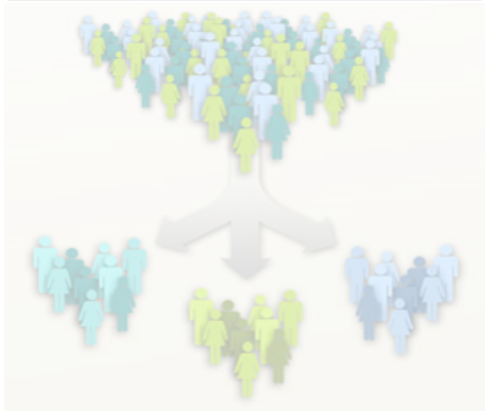
Approach:

- Merge disparate data sources
- Define which fields to use
- Automatically generate a model to group customers
- Apply business terms to new groups
- Export newly created groups to database

Why?

- Better customer understanding (demographics, socio-economic etc)
- Tailored messages for each group/segment
- Personal and more relevant for consumers

Find Patterns and Groups



Classify customers into groups based on underlying characteristics

Understand the Past, Predict the Future

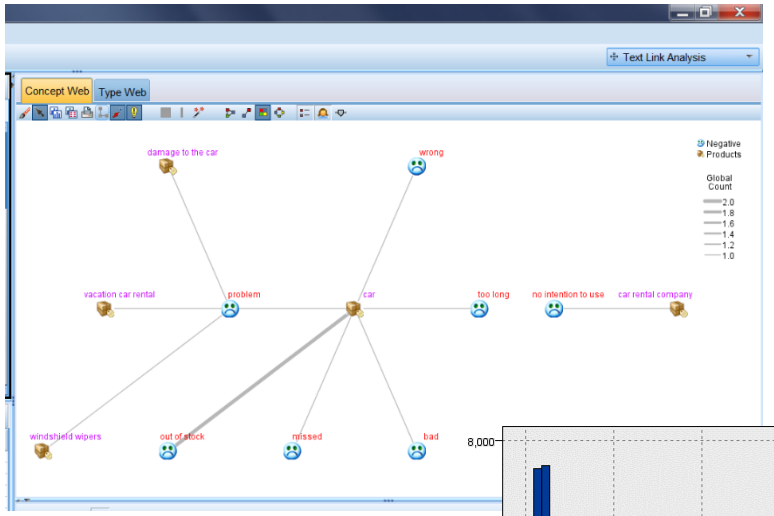
Model response to marketing offers using historical data

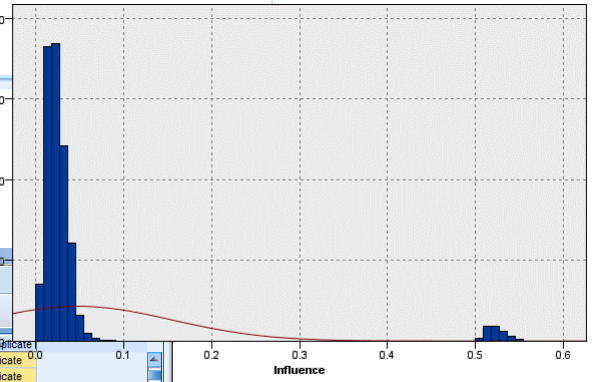
Deploy Insights on new projects

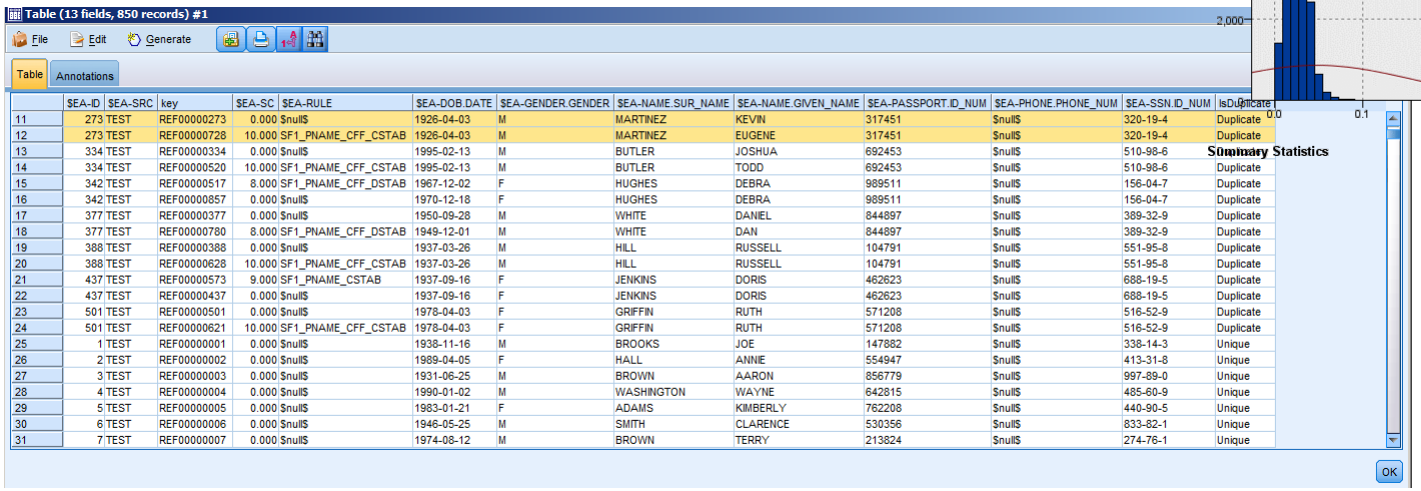
Leverage models to create insights on new customers

Uncovering Patterns In Unstructured Data

- Text Analytics
  - Natural Language
  - Sentiment Analysis
- Entity Analytics
  - Disambiguate identity
  - People, places, things
- Social Network Analysis
  - Uncover relationships
  - Find leaders and followers







	$EA-ID	$EA-SRC	key	$EA-SC	$EA-RULE	$EA-DOB.DATE	$EA-GENDER.GENDER	$EA-NAME.SUR_NAME	$EA-NAME.GIVEN_NAME	$EA-PASSPORT.ID_NUM	$EA-PHONE.PHONE_NUM	$EA-SSN.ID_NUM	isDuplicate
11	273	TEST	REF00000273	0.000	$null$	1926-04-03	M	MARTINEZ	KEVIN	317451	$null$	320-19-4	Duplicate
12	273	TEST	REF00000728	10.000	SF1_PNAME_CFF_CSTAB	1926-04-03	M	MARTINEZ	EUGENE	317451	$null$	320-19-4	Duplicate
13	334	TEST	REF00000334	0.000	$null$	1995-02-13	M	BUTLER	JOSHUA	692453	$null$	510-98-6	Duplicate
14	334	TEST	REF00000520	10.000	SF1_PNAME_CFF_CSTAB	1995-02-13	M	BUTLER	TODD	692453	$null$	510-98-6	Duplicate
15	342	TEST	REF00000517	8.000	SF1_PNAME_CFF_DSTAB	1967-12-02	F	HUGHES	DEBRA	989511	$null$	156-04-7	Duplicate
16	342	TEST	REF00000857	0.000	$null$	1970-12-18	F	HUGHES	DEBRA	989511	$null$	156-04-7	Duplicate
17	377	TEST	REF00000377	0.000	$null$	1950-09-28	M	WHITE	DANIEL	844897	$null$	389-32-9	Duplicate
18	377	TEST	REF00000780	8.000	SF1_PNAME_CFF_DSTAB	1949-12-01	M	WHITE	DAN	844897	$null$	389-32-9	Duplicate
19	388	TEST	REF00000388	0.000	$null$	1937-03-26	M	HILL	RUSSELL	104791	$null$	551-95-8	Duplicate
20	388	TEST	REF00000628	10.000	SF1_PNAME_CFF_CSTAB	1937-03-26	M	HILL	RUSSELL	104791	$null$	551-95-8	Duplicate
21	437	TEST	REF00000573	9.000	SF1_PNAME_CSTAB	1937-09-16	F	JENKINS	DORIS	462623	$null$	688-19-5	Duplicate
22	437	TEST	REF00000437	0.000	$null$	1937-09-16	F	JENKINS	DORIS	462623	$null$	688-19-5	Duplicate
23	501	TEST	REF00000501	0.000	$null$	1978-04-03	F	GRIFFIN	RUTH	571208	$null$	516-52-9	Duplicate
24	501	TEST	REF00000621	10.000	SF1_PNAME_CFF_CSTAB	1978-04-03	F	GRIFFIN	RUTH	571208	$null$	516-52-9	Duplicate
25	1	TEST	REF00000001	0.000	$null$	1938-11-16	M	BROOKS	JOE	147882	$null$	338-14-3	Unique
26	2	TEST	REF00000002	0.000	$null$	1989-04-05	F	HALL	ANNIE	554947	$null$	413-31-8	Unique
27	3	TEST	REF00000003	0.000	$null$	1931-06-25	M	BROWN	AARON	856779	$null$	997-89-0	Unique
28	4	TEST	REF00000004	0.000	$null$	1990-01-02	M	WASHINGTON	WAYNE	642815	$null$	485-60-9	Unique
29	5	TEST	REF00000005	0.000	$null$	1983-01-21	F	ADAMS	KIMBERLY	762208	$null$	440-90-5	Unique
30	6	TEST	REF00000006	0.000	$null$	1946-05-25	M	SMITH	CLARENCE	530356	$null$	833-82-1	Unique
31	7	TEST	REF00000007	0.000	$null$	1974-08-12	M	BROWN	TERRY	213824	$null$	274-76-1	Unique

The Importance Of Text

Because people communicate with words, not numbers, it has become critical to be able to mine text for its meaning and to sort, analyse, and understand it in the same way that data has been tamed. In fact, the two basic types of information complement each other, with data supplying the "what" and text supplying the "why".

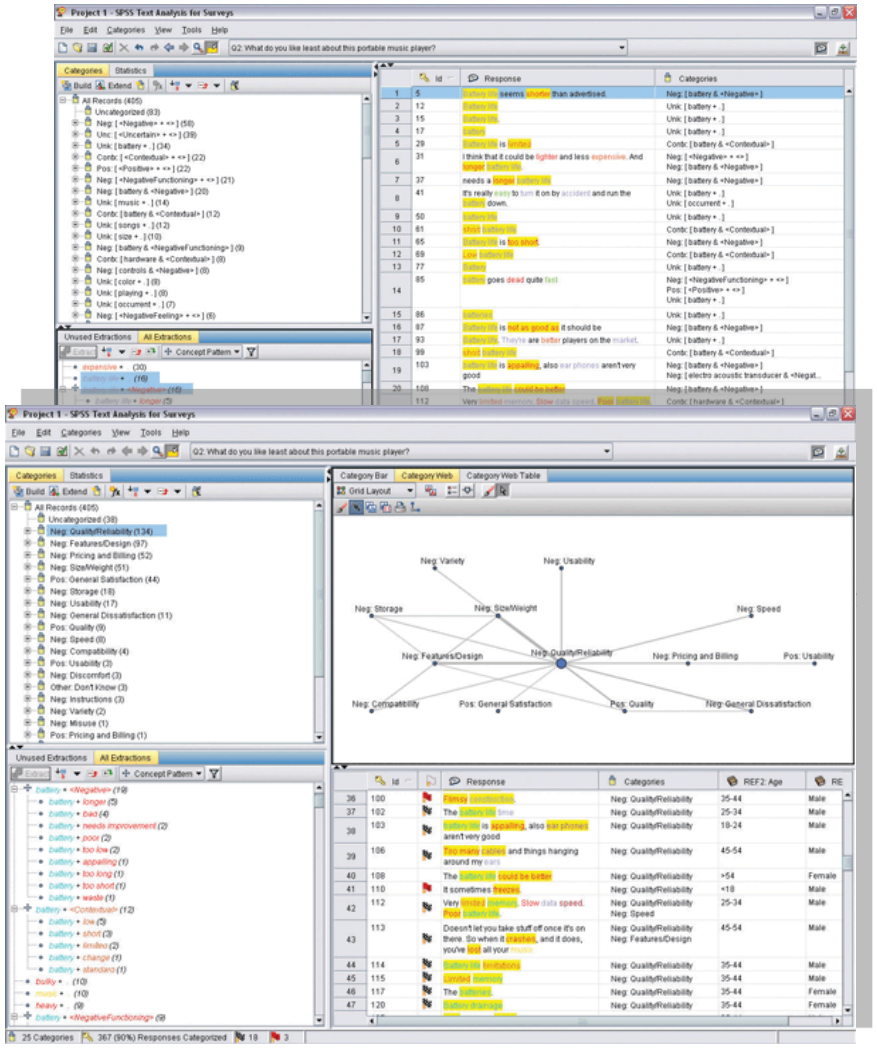
Source IDC: "Text Analytics: Software's Missing Piece?"

Text Analytics

Uses natural language processing heuristic rules and statistical techniques to reveal conceptual meaning in text

Extracts concepts from text and categorizes them

Makes unstructured qualitative data more quantifiable, enabling the discovery of key insights from sources such as:
-Documents
-Survey responses
-Call center Notes
-Social Media
-Web Pages



Exercise 3: Understand The Past, Predict The Future

Goal:

- Identify who is likely to respond to a marketing offer

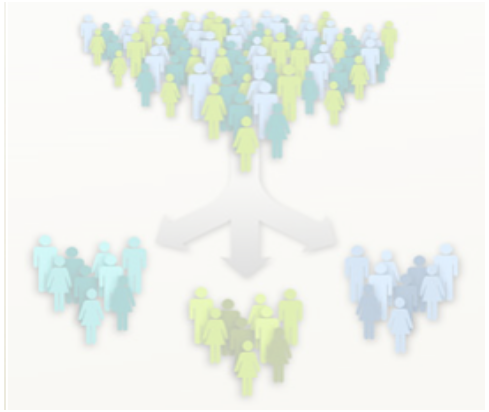
Approach:

- Use a data extract from a CRM
- Extract concepts from the open ended comments in a customer survey
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who is likely to respond
- Review results

Why?

- Target those likely to respond to offers to increase revenue, cut costs
- Using unstructured data improves modeling accuracy and provides more insight

Find Patterns and Groups



Classify customers into groups based on underlying characteristics

Understand the Past, Predict the Future

Model response to marketing offers using historical data

Deploy Insights on new projects

Leverage models to create insights on new customers

Exercise 4: Deploy Insights On New Projects

Goal:

- Leverage predictive insights to score new consumers

Approach:

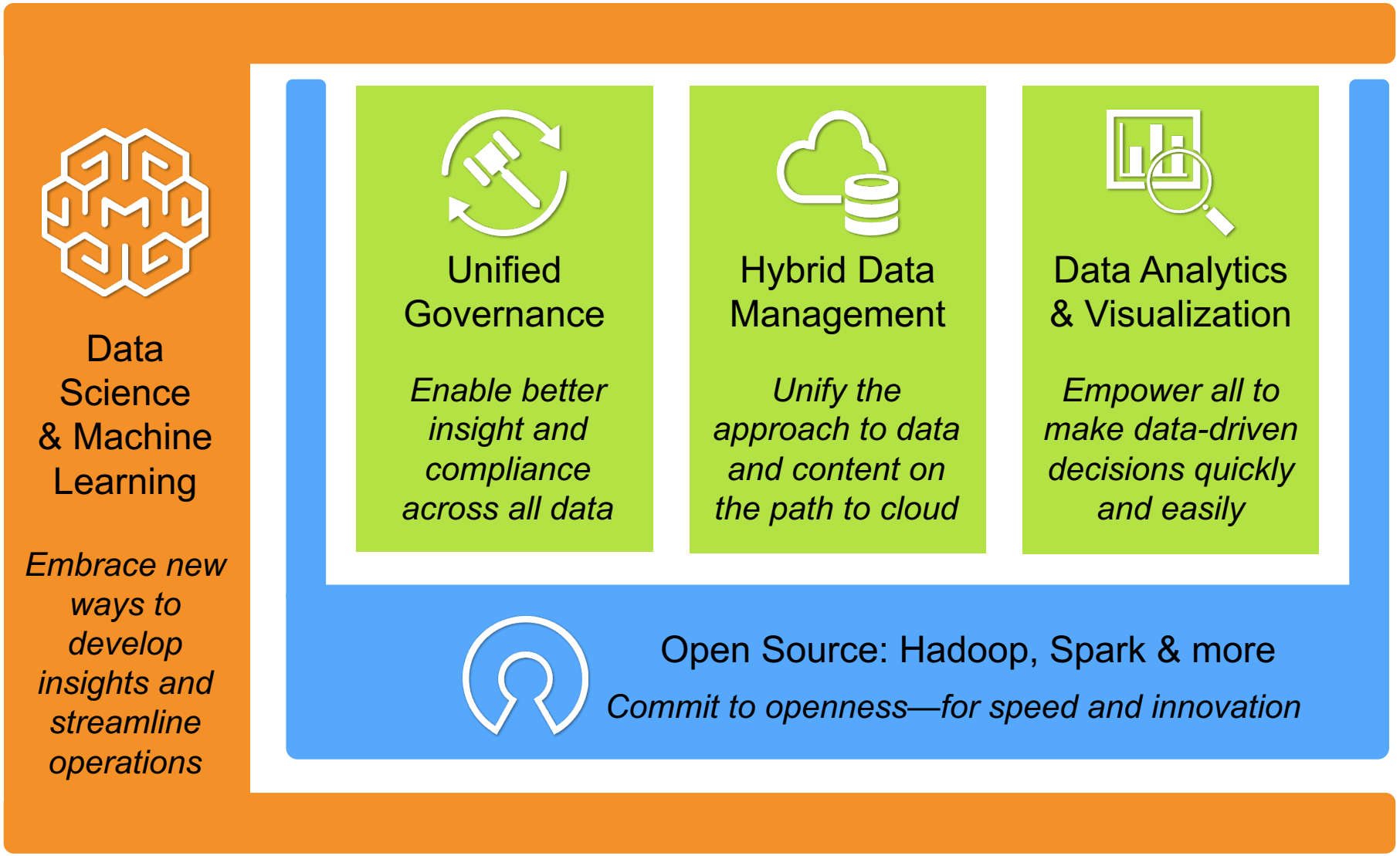
- Use new customer records who have never received an offer
- Leverage the text extraction and classification models used in exercise 3
- Automatically generate scores of who is likely to respond
- Review results
- Deploy results for use by marketing team

Why?

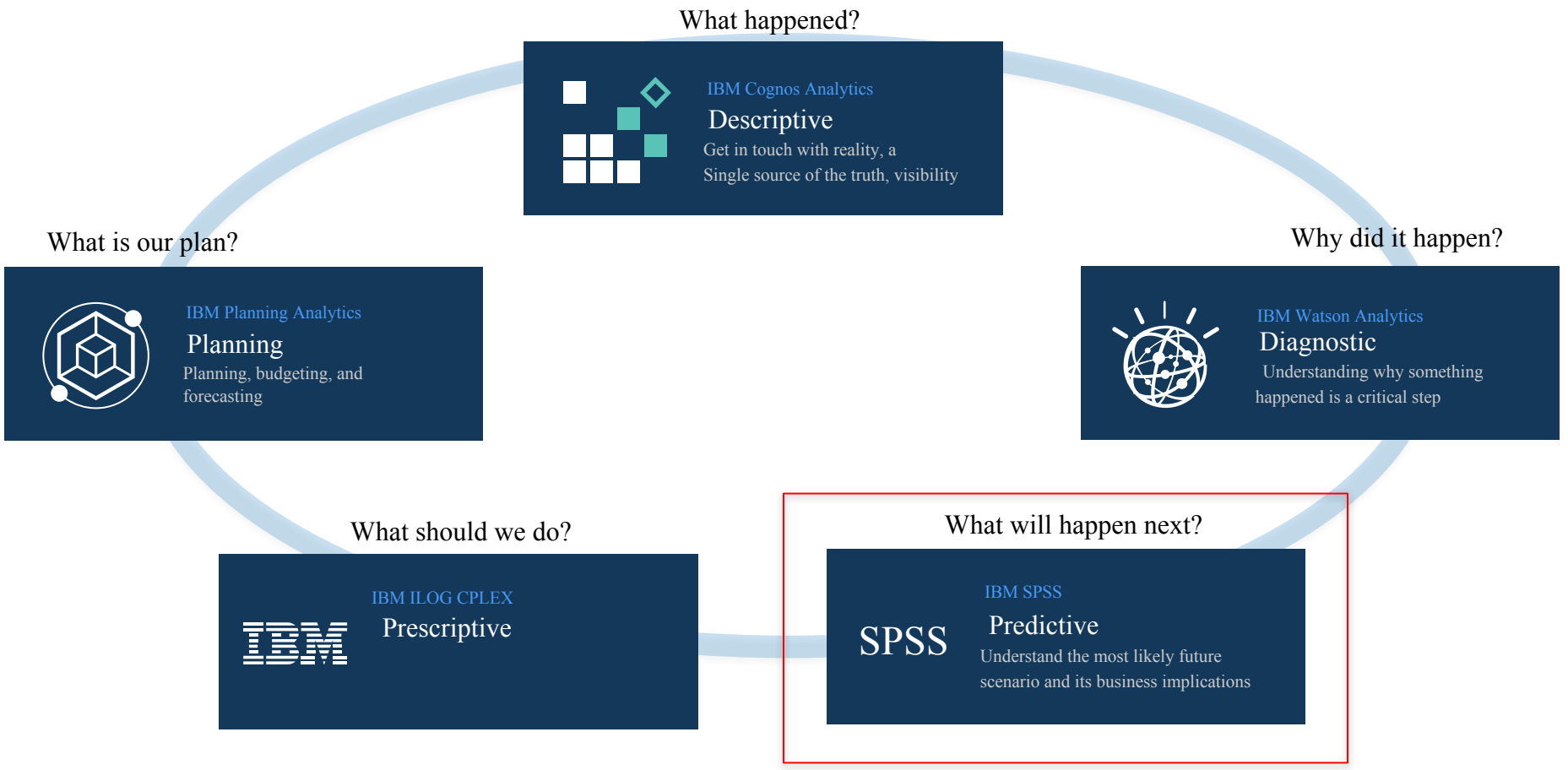
- Create models on large populations for robust, representative results, and easily score smaller record batches when necessary, on-demand.
- Apply consistent and seamless methodology over many model scoring iterations, ensuring quality control.

PORTFOLIO IMPLEMENTATION AND DEPLOYMENT

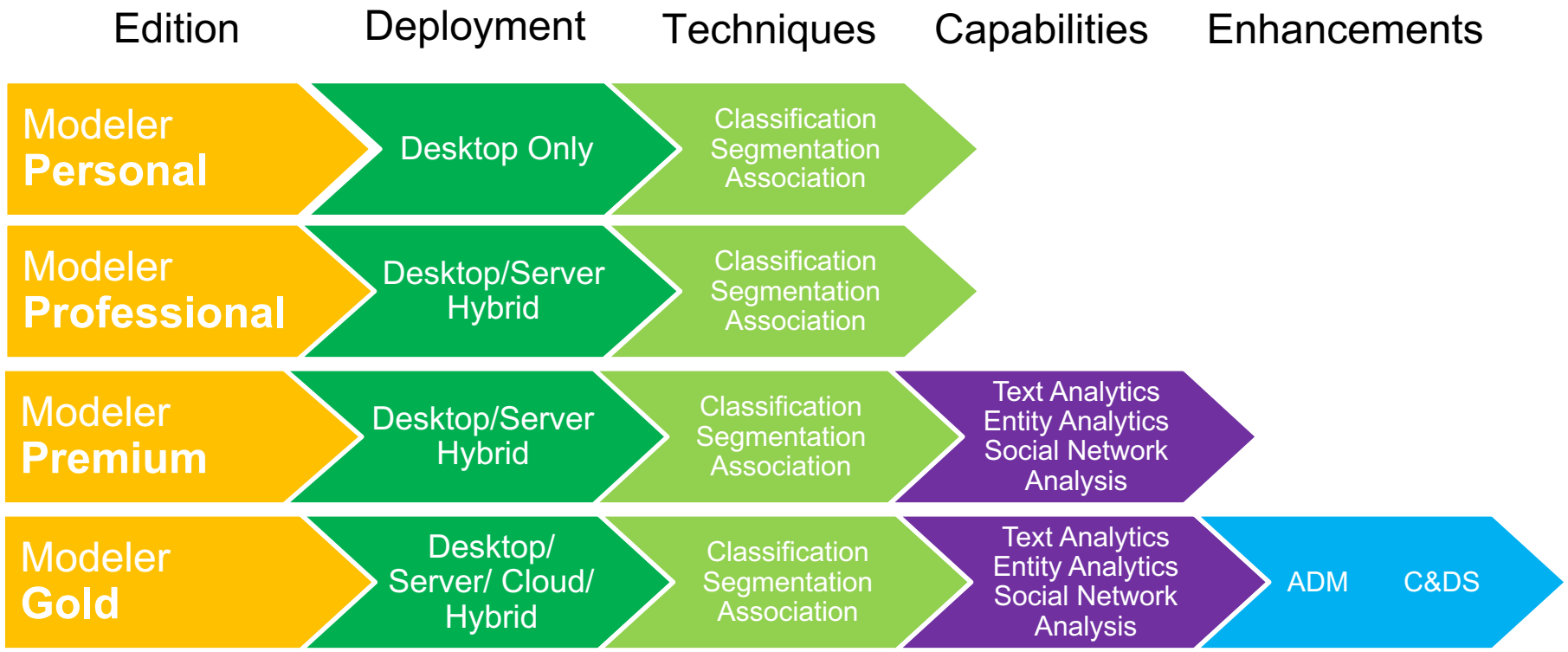
Our Portfolio: Power Behind And Across Your Firewall



IBM's Analytics Tools



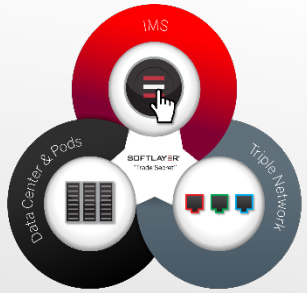
SPSS Modeler Editions



Flexible deployment

Flexible

Flexible Deployment Options, Including Cloud, Supported By An Infrastructure-agnostic Platform

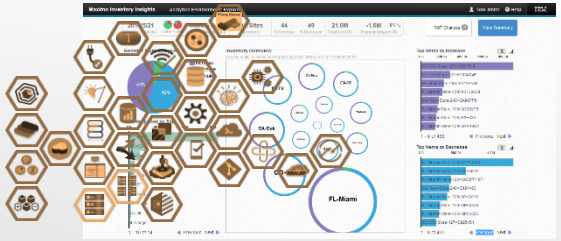


Integrated

Integrated With IBM's High-performance Systems Built for Big Data Analytics

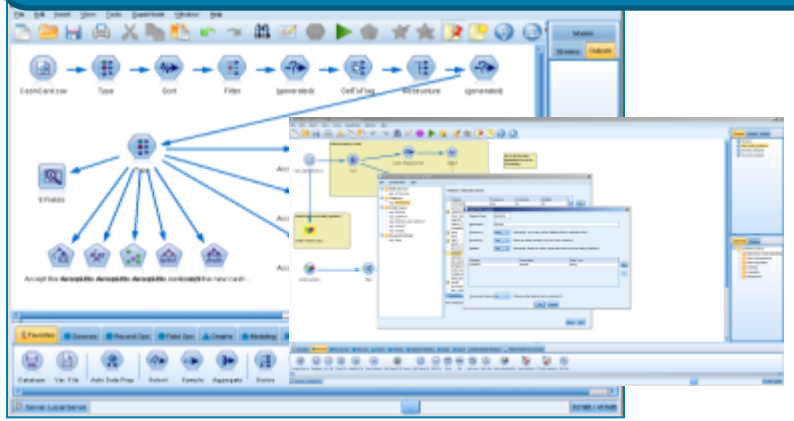
Embedded

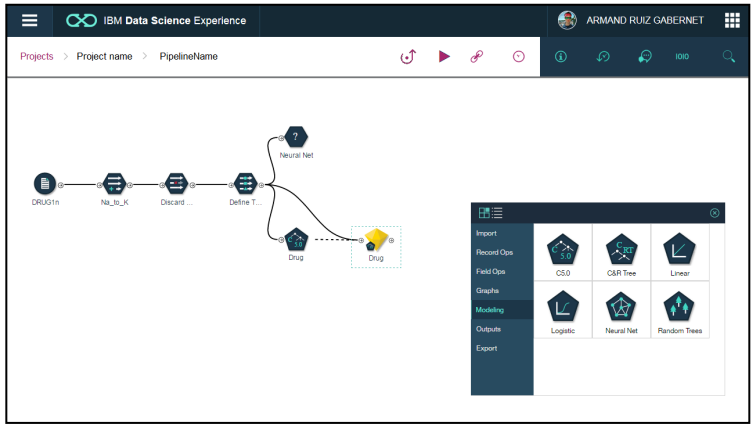
Embedded Into Operational, Mobile and/or Cloud-based Applications



IBM SPSS And The Cloud

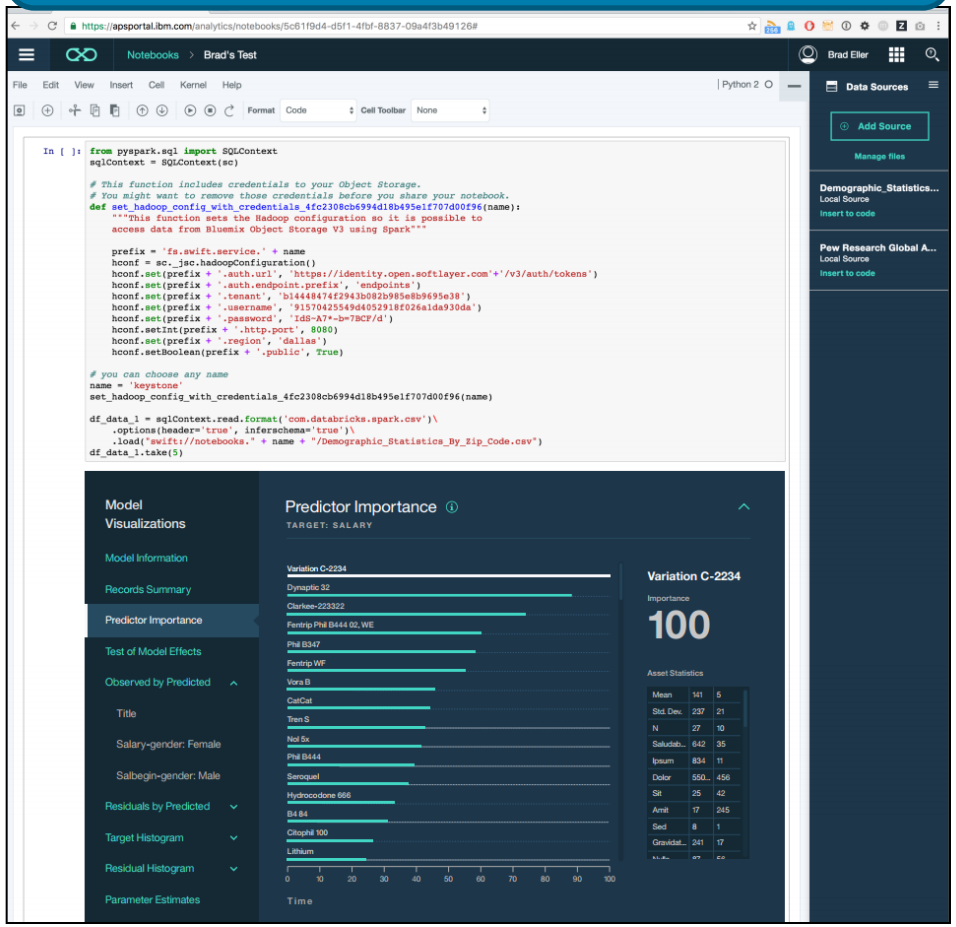
Powerful workbench for code-optional predictive analytics





Delivered in the **cloud** through IBM Data Science Experience

SPSS Algorithms in Python, R and Scala and Interactive Model Visualization through IBM Data Science Experience

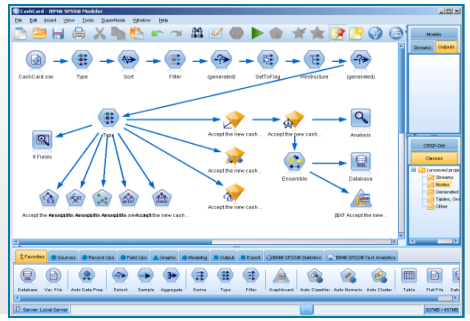


WRAP UP

Predictive Analytics For All

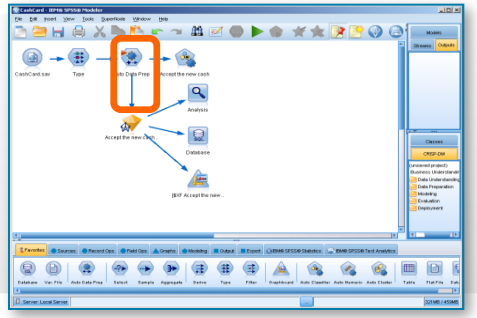
Expert User

Deep Predictive Capabilities With In-database, R and Python Integration, and Integrated Deployment



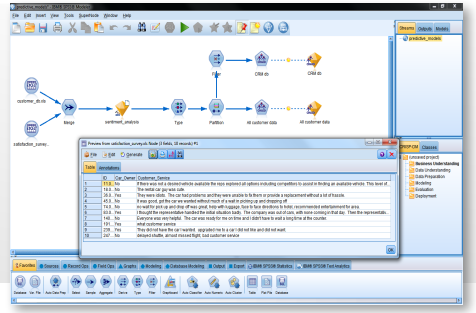
Intermediate User

Visual Workbench for Building Models of Any Complexity With Ability to Automate and Combine Tasks



Novice

Simple, Easy-to-Use, Visual, Guided Analytical Discovery, Intelligent Automation and Visual Storytelling



IBM Positioned In Leaders Quadrant Of The Gartner Magic Quadrant For Data Science Platforms

Figure 1. Magic Quadrant for Data Science Platforms

Source: Gartner (February 2017)

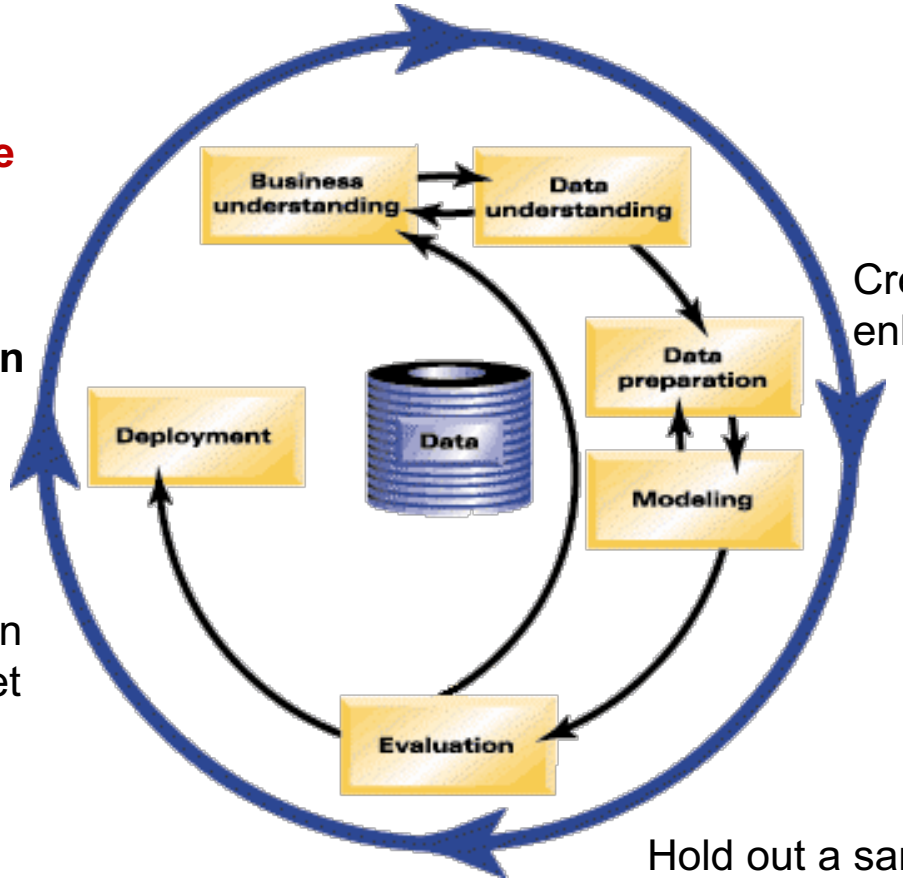
- "IBM has a vast customer base and remains committed to modernizing and extending its data science and machine-learning capabilities. DSX [Data Science Experience) is likely to be one of the most attractive platforms in the future — modern, open, flexible and suitable for a range of users, from expert data scientists to business people."
- "Customers choose IBM SPSS because of its ability to support a broad range of data types, including unstructured data, and its solid product quality."
- "Surveyed IBM customers rated SPSS's model management highly, with praise for its breadth of models, accuracy and transparency in workflows, model deployment, monitoring for degradation and automatic retuning. SPSS provides excellent features for analytics governance: versioning, metadata and audit capabilities."

Source: Gartner "Magic Quadrant for Data Science Platforms", Alexander Linden, Peter Krensky, Jim Hare, Carlie Idoine, Svetlana Sicular, Shubhangi Vashisth 14 Feb. 2017

Getting Started

- What do you want to achieve?
- Short project with the biggest impact
- ROI and Success Criteria are key to identifying early

- **Measure results**
- **Quantify your ROI**
- **97% of customers are able to pay for their investment within 11 months!**
- **Use return to invest in the next project**



Create new variables to enhance data

- Embed in an application
- Export to a spreadsheet

Hold out a sample of the data to test before deploying

Workshop Takeaways

Easy to use, visual interface

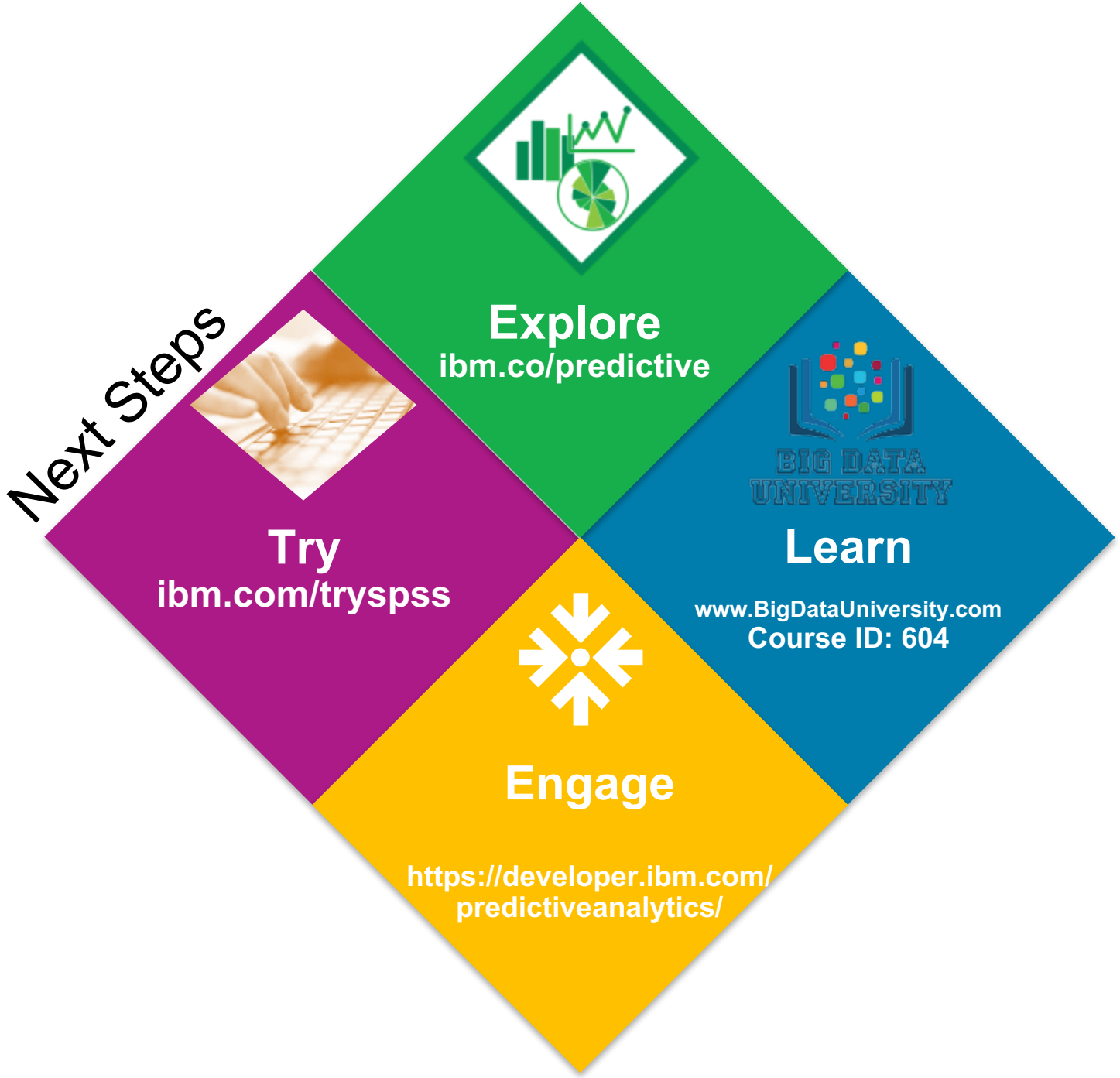
- Short timeframe to be productive with actionable results
- Does not require knowledge of programming language
- No proprietary data formats
- Open architecture

Business results focused

- Leverages the investments already made in technology
- Cost effective solution that delivers powerful results across organization
- Full range of algorithms for your business problems
- Big Data enabled (Hadoop, SQL Pushback)

End-to-end solution

- Data preparation through real time interactions
- Use structured, unstructured and semi-structured data
- Integrated portfolio for business analytics
- Scales from a single desktop to an enterprise deployments



44

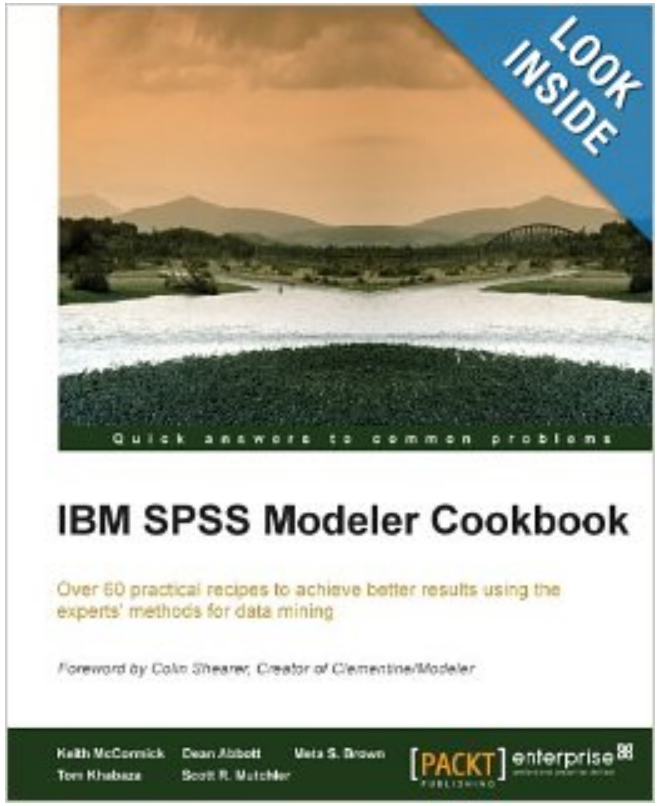
TELL ME HOW YOU CAN APPLY THIS TO YOUR BUSINESS

Appendix

Suggested Books

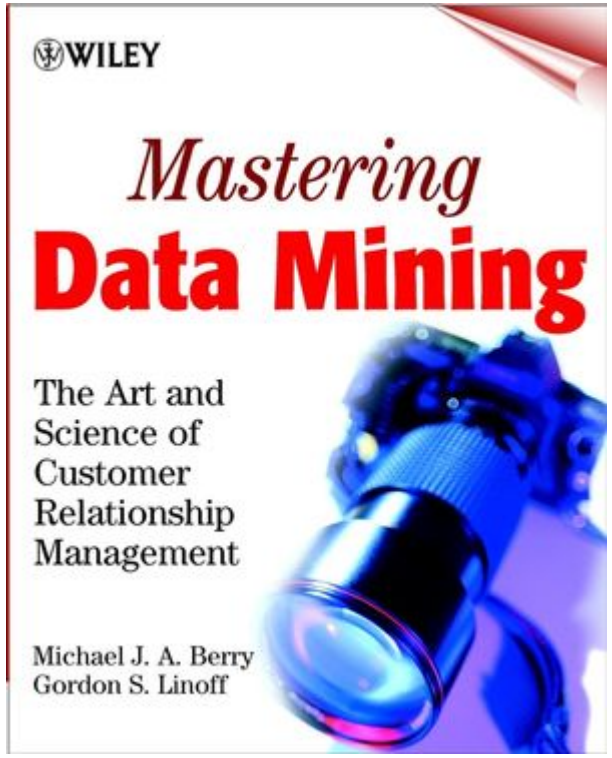
IBM SPSS Modeler Cookbook

- **From Amazon.com**
  - IBM SPSS Modeler Cookbook
  - by Keith McCormick, Dean Abbott, Meta S. Brown, Tom Khabaza, Scott R. Mutchler
  - Paperback - 382 pages (October 2013) (also on Kindle)
  - ISBN : 1849685460
- **Written by those who teach and have been working with IBM SPSS Modeler since its beginnings. Full of practical examples that span the full gamut of capabilities.**

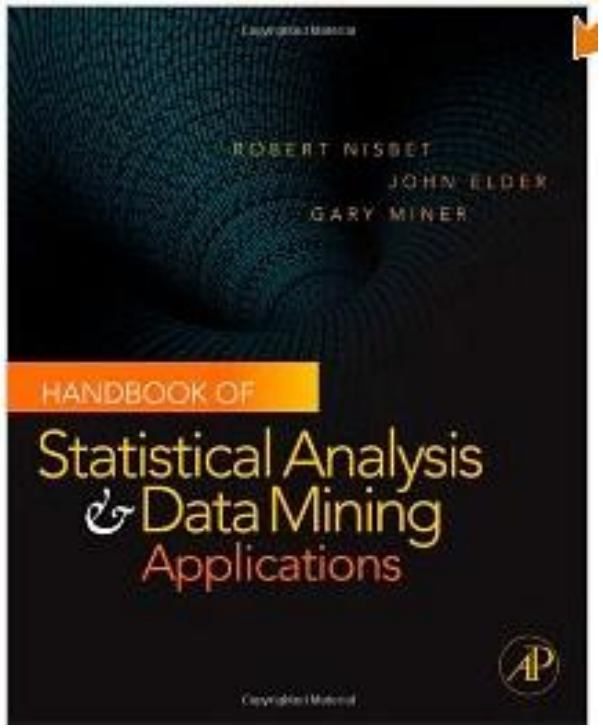


Data Mining Overview

- **From Amazon.com**
  - **Paperback:** 512 pages
  - **Publisher:** Wiley; 1 edition (December 28, 1999)
  - **Language:** English
  - **ISBN-10:** 0471331236
  - **ISBN-13:** 978-0471331230 ;
- **Good introductory text on data mining for marketing from two top communicators in the field**



Handbook of Statistical Analysis and Data Mining Applications



- Handbook of Statistical Analysis and Data Mining Applications
- Robert Nisbet, John Elder IV, and Gary Miner
- Academic Press (2009)
- ISBN-10: 0123747651

- An excellent guide to many aspects of data mining including Text mining.

Data Mining Algorithms

- **From Amazon.com**
  - Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations
  - by Eibe Frank, Ian H. Witten
  - Paperback - 416 pages (October 13, 1999)
  - Morgan Kaufmann Publishers;
  - ISBN: 1558605525;
- **Best book I've found in between highly technical and introductory books. Good coverage of topics, especially trees and rules, but no neural networks.**