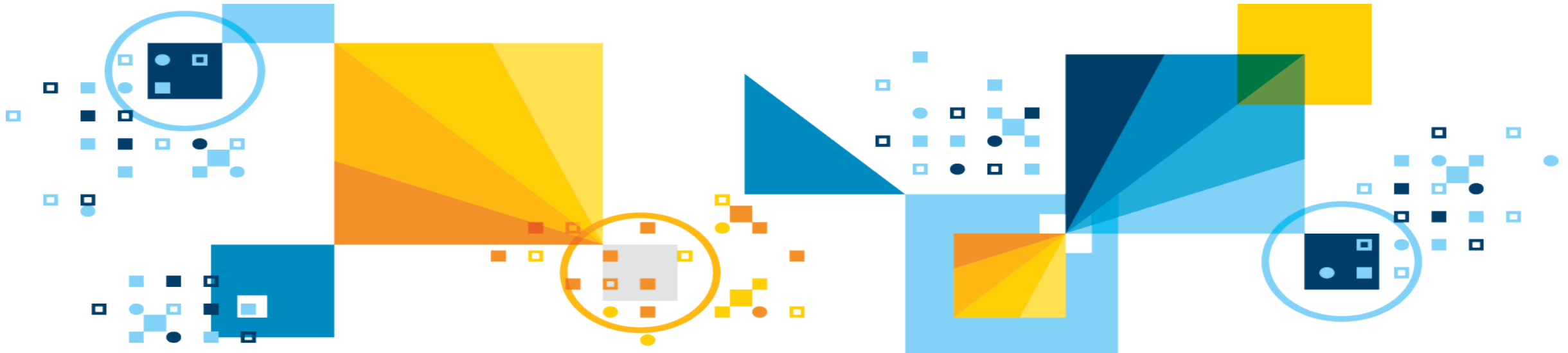


IBM Data Science

A Comprehensive Vision and Platform

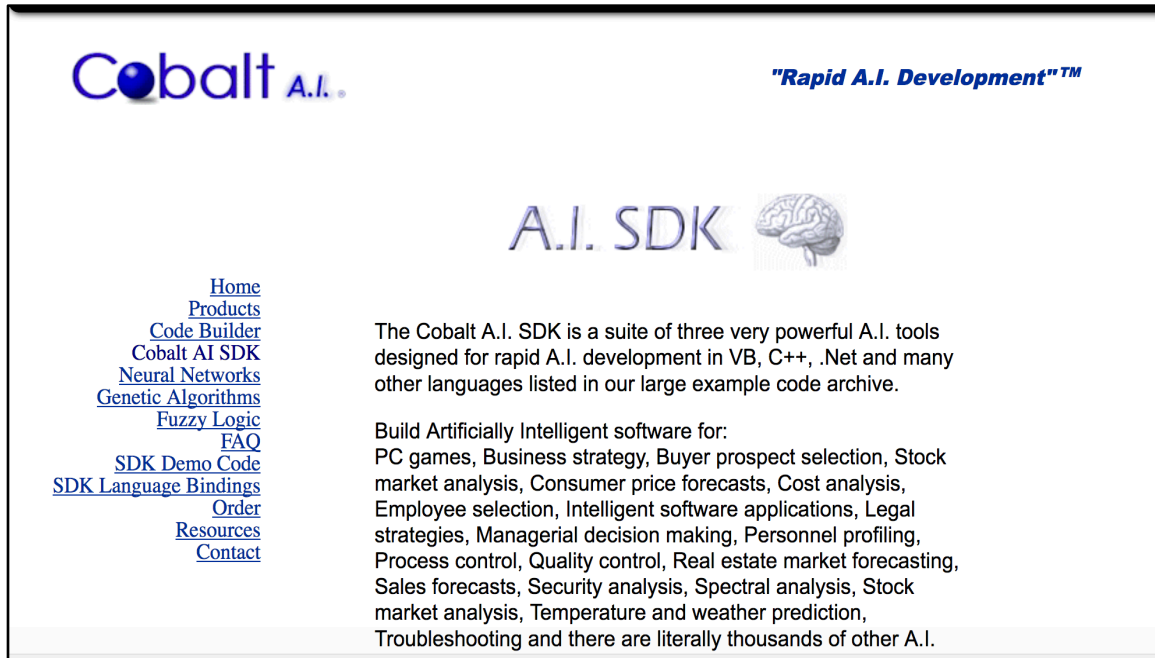


Steven Geringer
Data Science Solutions Architect
IBM Analytics
steve.geringer@ibm.com




■ About Me:

- I Love Algorithms
- Once owned CobaltAI.com (2002-2005)
- Kaggle “Competitions Expert” (Two Silver Medals, 2014-2015)



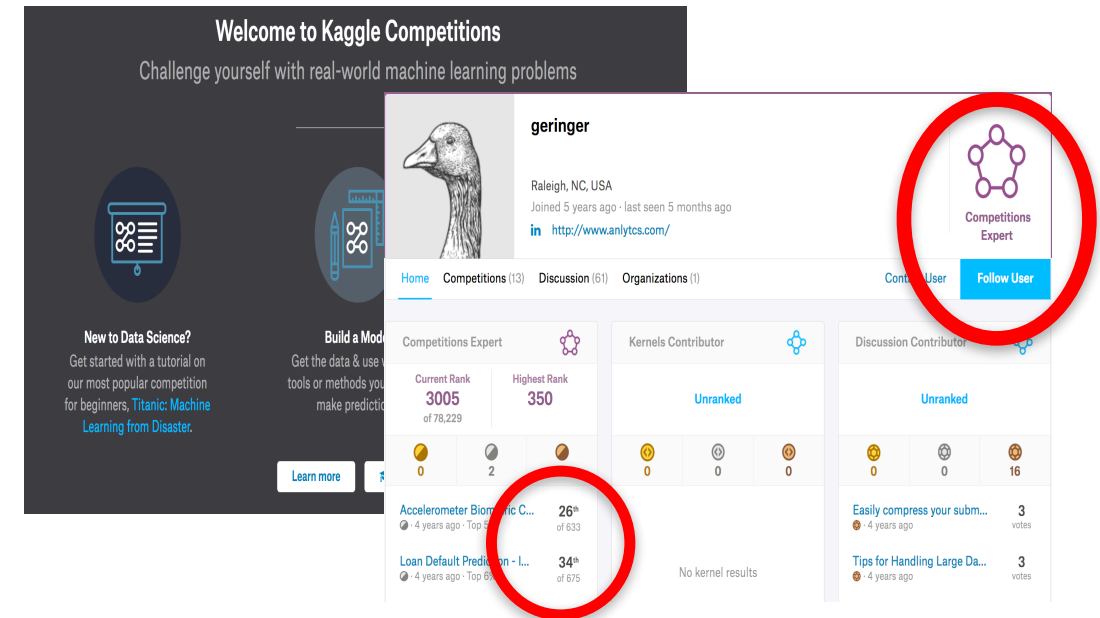
Cobalt A.I. *"Rapid A.I. Development"™*

A.I. SDK 

The Cobalt A.I. SDK is a suite of three very powerful A.I. tools designed for rapid A.I. development in VB, C++, .Net and many other languages listed in our large example code archive.

Build Artificially Intelligent software for:
 PC games, Business strategy, Buyer prospect selection, Stock market analysis, Consumer price forecasts, Cost analysis, Employee selection, Intelligent software applications, Legal strategies, Managerial decision making, Personnel profiling, Process control, Quality control, Real estate market forecasting, Sales forecasts, Security analysis, Spectral analysis, Stock market analysis, Temperature and weather prediction, Troubleshooting and there are literally thousands of other A.I.

[Home](#)
[Products](#)
[Code Builder](#)
[Cobalt AI SDK](#)
[Neural Networks](#)
[Genetic Algorithms](#)
[Fuzzy Logic](#)
[FAQ](#)
[SDK Demo Code](#)
[SDK Language Bindings](#)
[Order](#)
[Resources](#)
[Contact](#)



Welcome to Kaggle Competitions
 Challenge yourself with real-world machine learning problems

geringer
 Raleigh, NC, USA
 Joined 5 years ago · last seen 5 months ago
<http://www.anlytcs.com/>

Competitions Expert

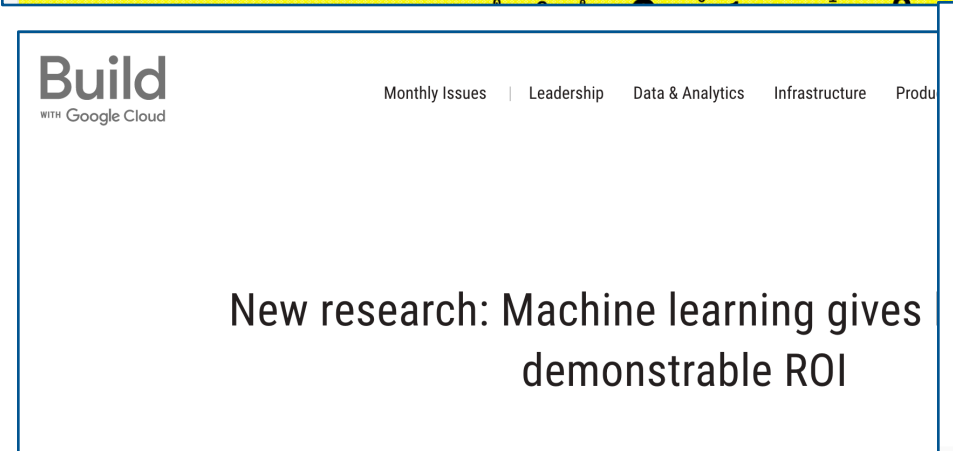
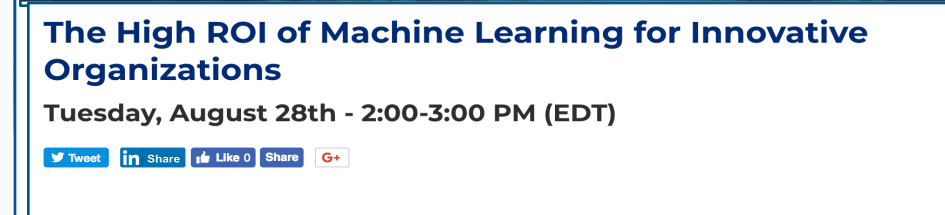
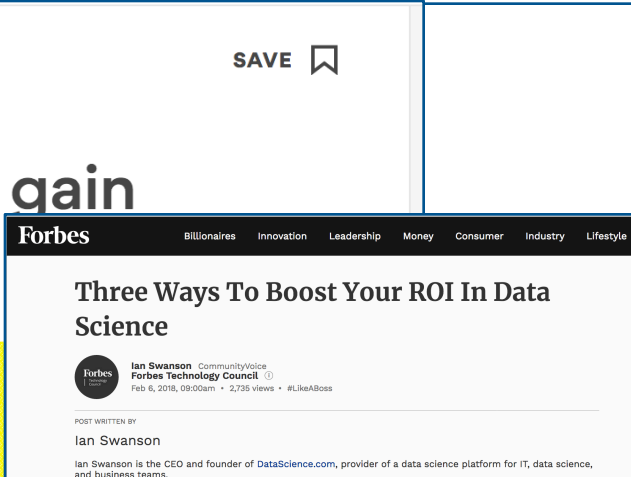
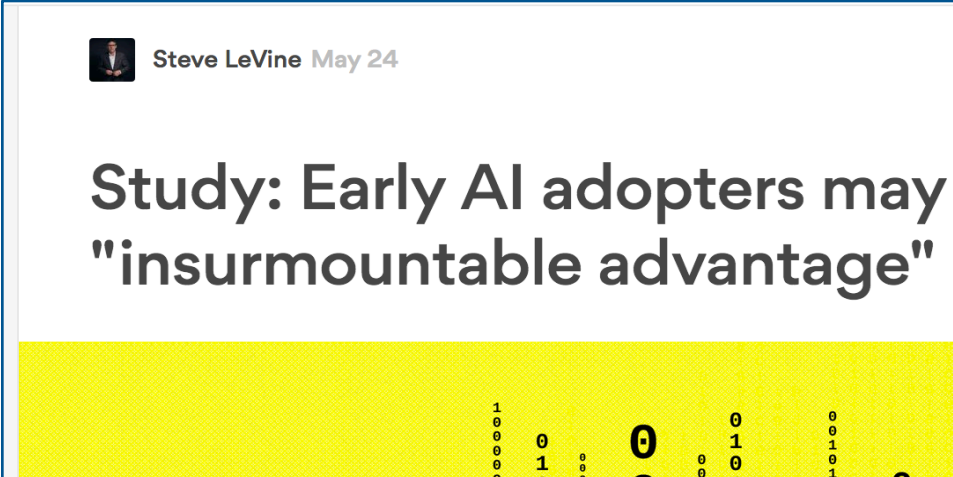
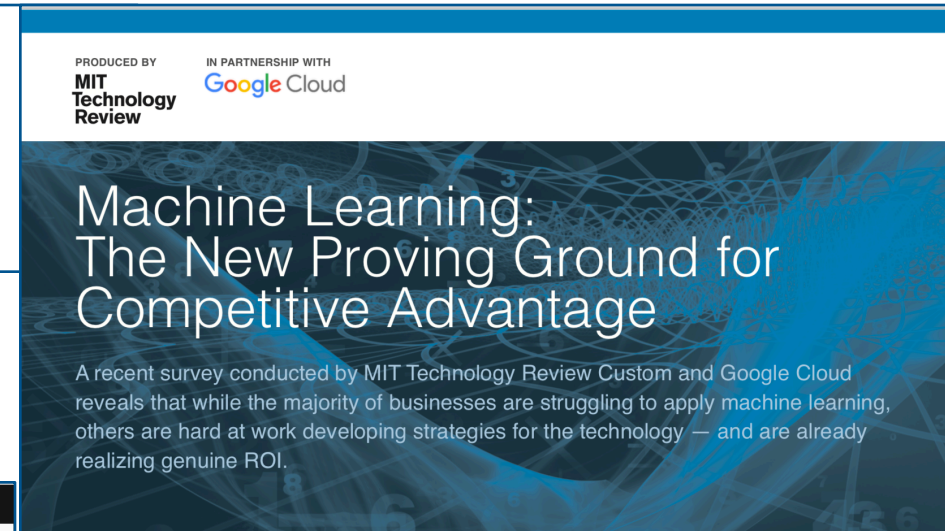
Competitions Expert	Kernels Contributor	Discussion Contributor
Current Rank: 3005 of 78,229 Highest Rank: 350	Unranked	Unranked
Accelerometer Biometric C... 26 th of 633 Loan Default Prediction - L... 34 th of 675	No kernel results	Easily compress your subm... 3 votes Tips for Handling Large Da... 3 votes

Agenda

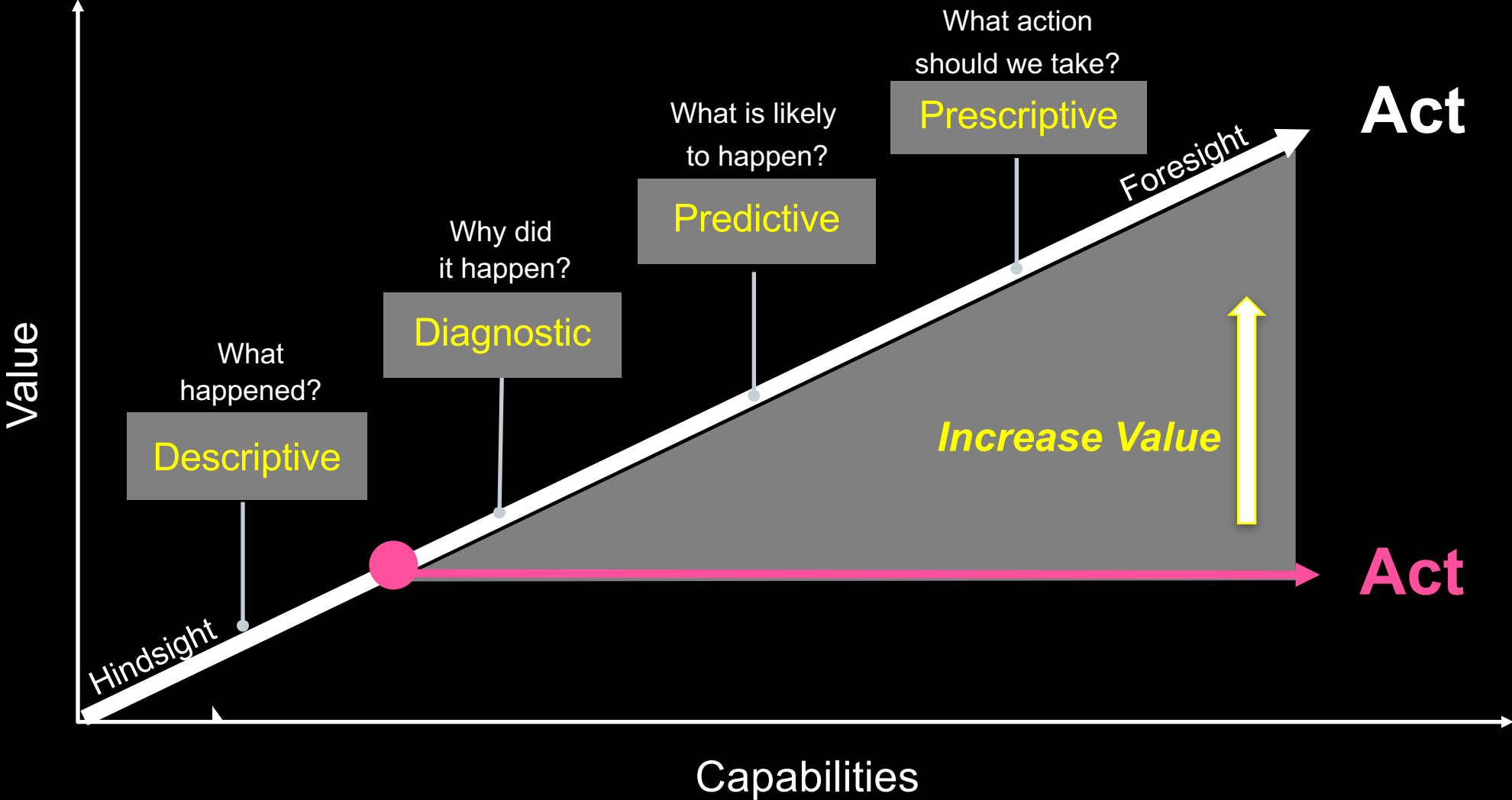
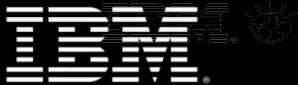
- **The Future of Data Science**
- **Analytics Portfolio Overview**
- **Data Science Experience Local Architecture**
- **Crash Course on Machine Learning**
- **DSXL Demo**

Benefits of Data Science

- **Competitive Advantage**
- **High ROI**



Analytics Maturity Curve



Picture the Future of Data Science

▪ Trends

- Data Is Growing Exponentially
- Data Science is Growing Linearly
 - Platform Market: CAGR 36.5%
 - Personnel: CAGR 28%
- Massive Expansion in:
 - Open Source
 - Data Science Use Cases
 - New Algorithms
 - Techniques
 - Data Types
- Microservices Replacing Monolithic Software

▪ Challenges and Pain Points

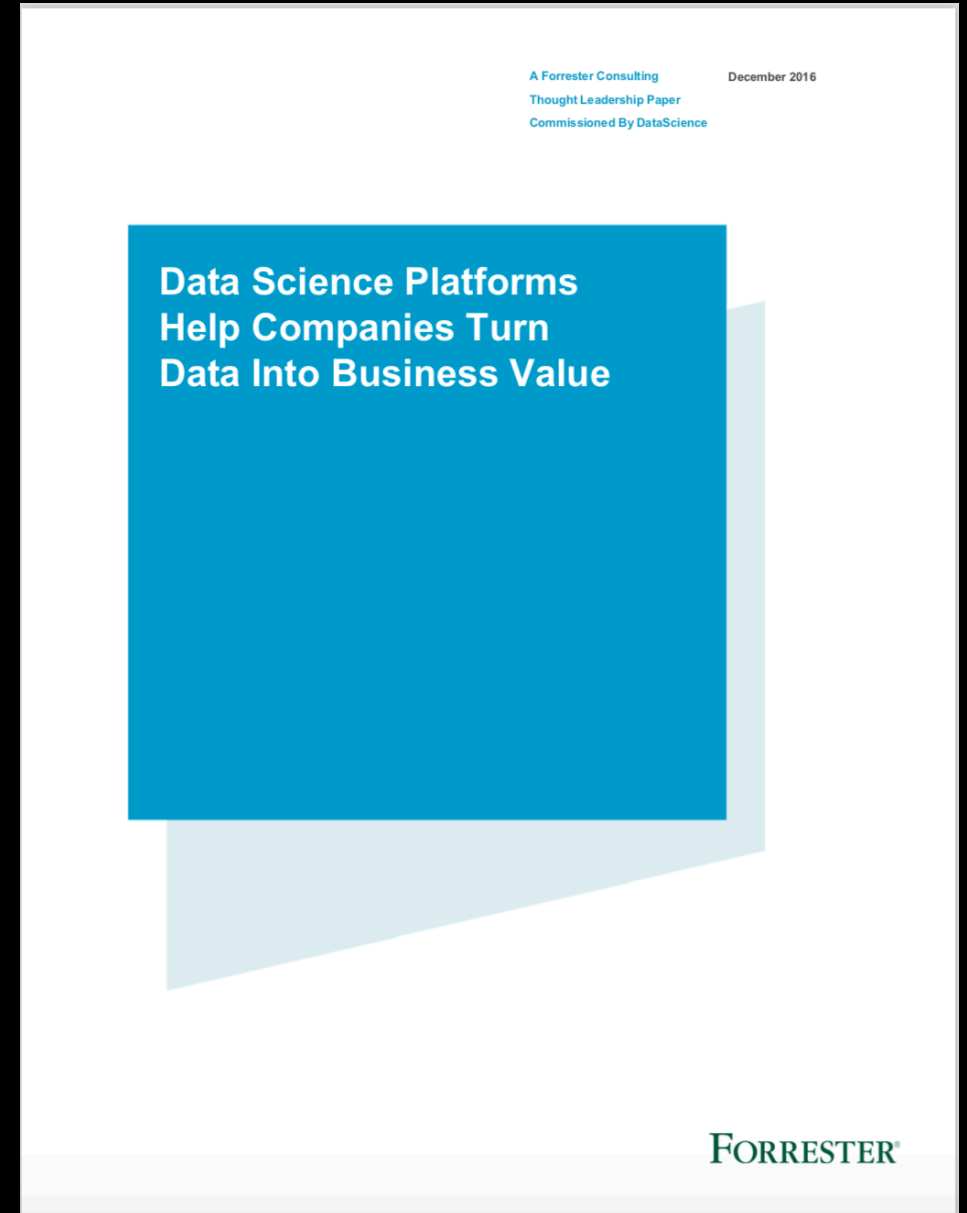
- Skill Shortages
- Job Hopping
- Low Productivity
- Disjointed Tools
- Data Integrity
- Cannot find or access to required data
- Organizations not using Insights from DS
- “Data Science Disillusionment”

Sources:

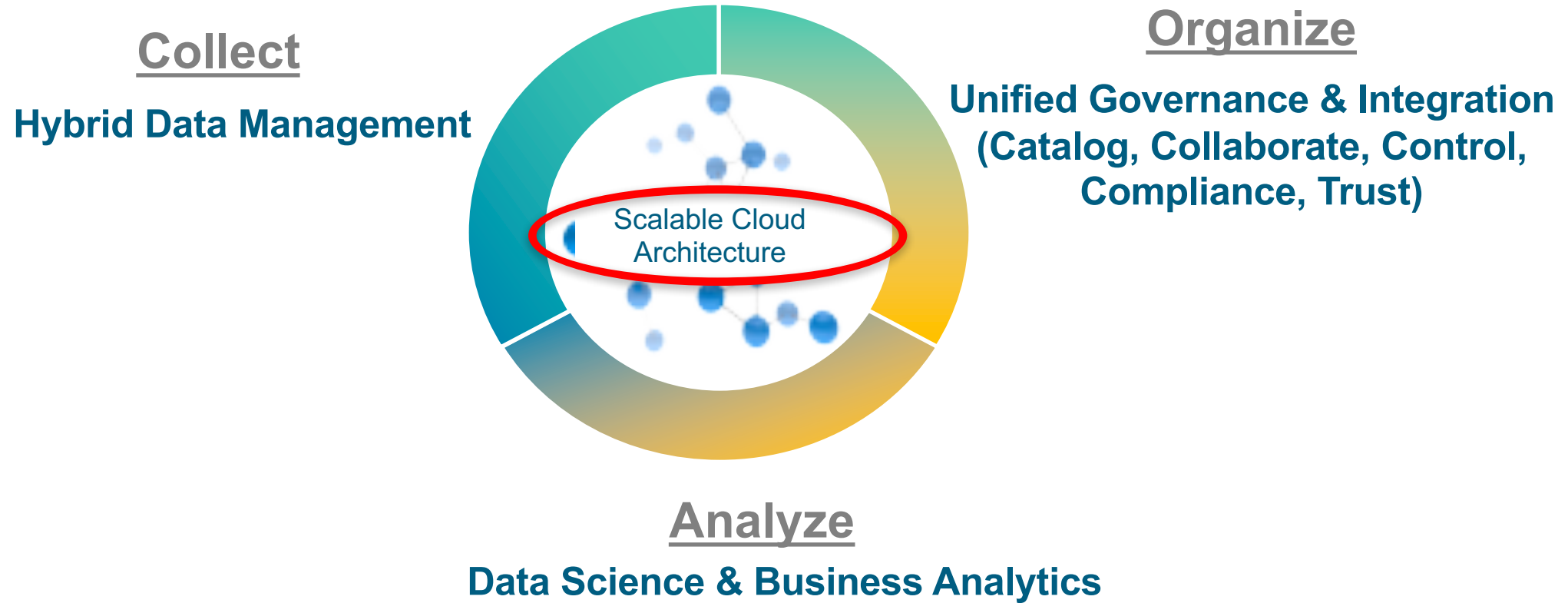
<https://www.forbes.com/sites/louiscolombus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#4a68516a7e3b>
<https://www.reuters.com/brandfeatures/venture-capital/article?id=5670>

Forrester Research - December 2016

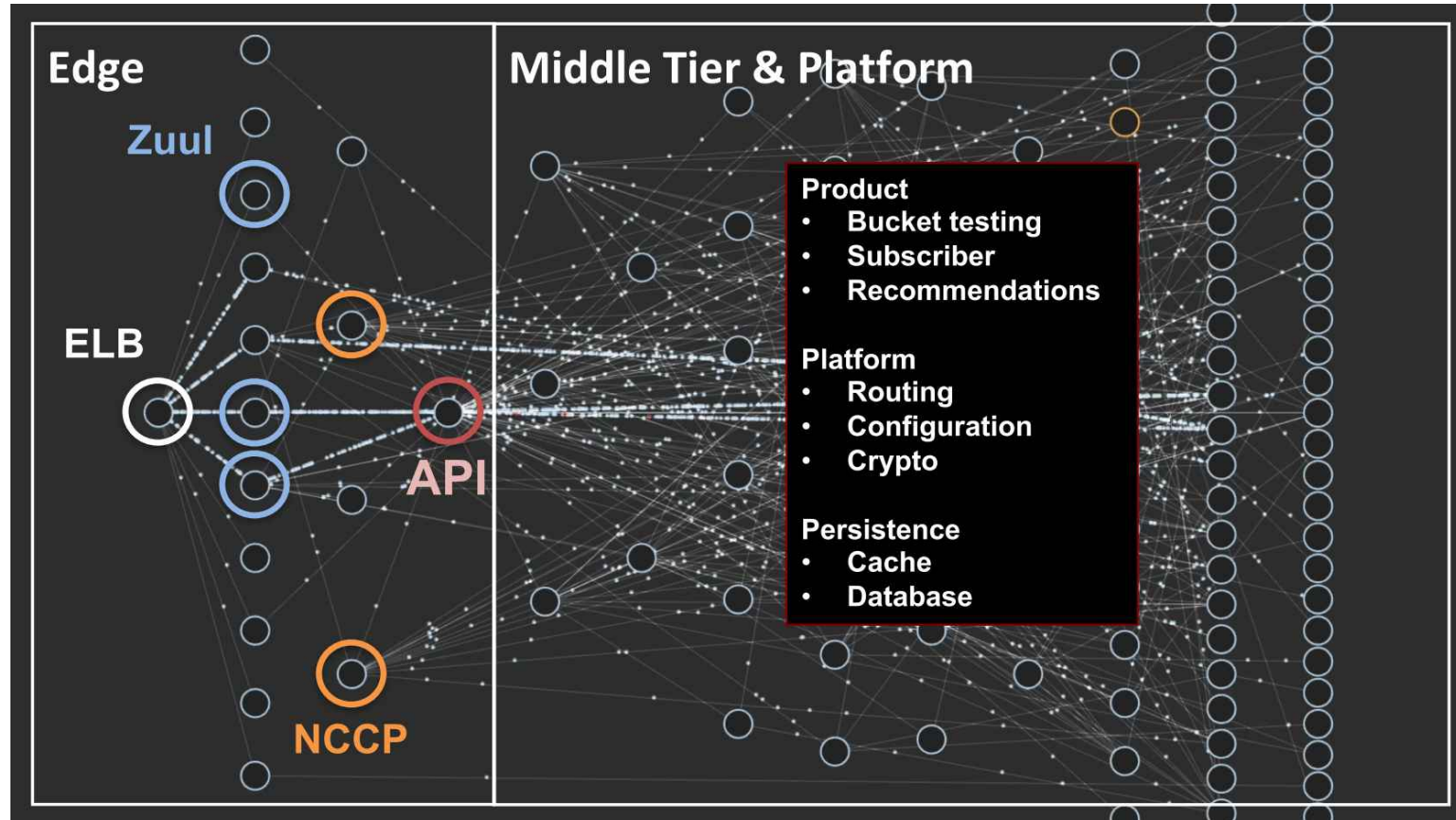
Data Science Platforms Lead to Better Data Science Results



IBM's Comprehensive Platform for Your Information Architecture



Benefits of Microservices vs. Monolithic Software



Netflix Microservice Architecture

Source: <https://www.infoq.com/presentations/netflix-chaos-microservices>

- **Scalability**
- **Agility**
 - Increase autonomy of teams
 - Enables continuous delivery
- **Manageability**
- **Encapsulation**
- **Innovation**
 - Polyglot Development
- **Quality**
 - Better fault isolation
- **Availability**

Data Science Is a Team Sport



Analytics Portfolio Overview

IBM Cognos Analytics (Descriptive)

- Integrated solution for managed reporting and business user self-service
- Designed for ease of use with a graduated user experience that enables analytic consumers to progress to access and model data, and create visual dashboards and stories on their own
- Smarter self service uses built in intelligence to guide data modeling and authoring based on intent
- Proven governed solution for performance, security and scalability

1
2



IBM Cognos Analytics



IBM Planning Analytics (Descriptive)

- Planning, budgeting, and forecasting for strategic, financial and operational needs
- Single platform for business planners to load data, model their business across multiple dimensions, and manage results in real-time
- Automates manual, disconnected spreadsheet-based planning with a more powerful and collaborative approach and customizable planning workspace

Modeling that Adapts to the needs of the Business

Intuitive, Collaborative Planning Processes

What-if Scenario Modeling & Analysis

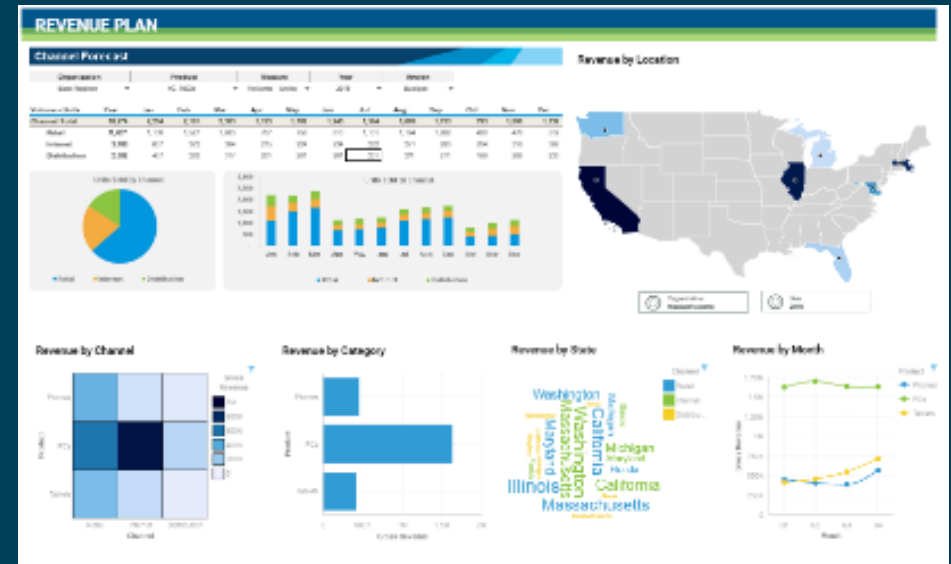
Native Excel Integration

Real-Time Answers To Complex Questions

KPIs, Dashboards, Scorecards, Reporting



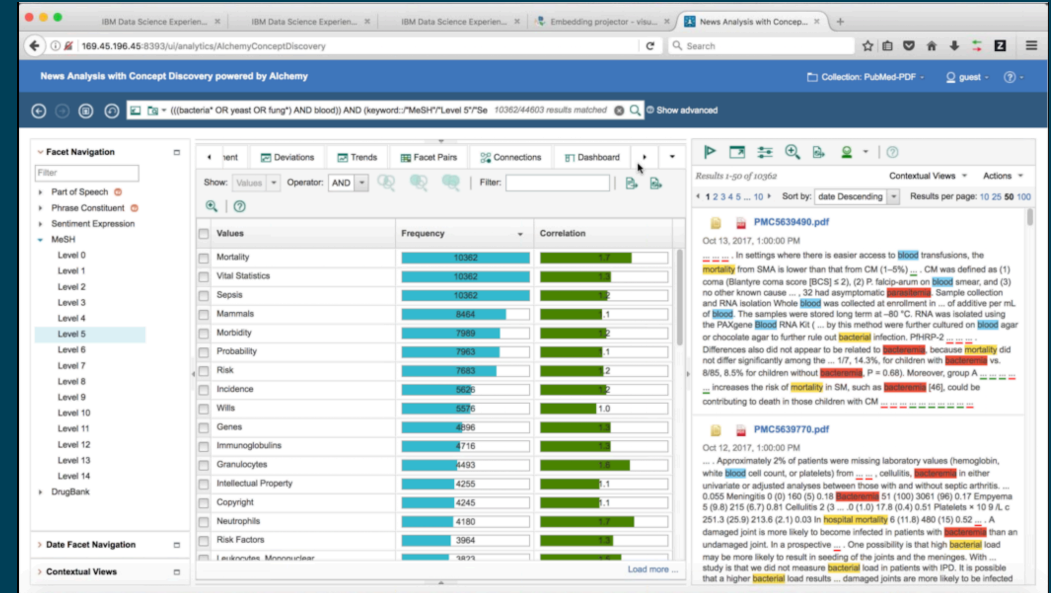
IBM Planning Analytics



IBM Watson Explorer (Diagnostic)

- Unlocks the Hidden Power of the your Unstructured Data
- Text Analytics, Natural Language Processing
- Gain deeper insights advanced content analytics
- Ability to handle all types of data
- Expert identification and location
- Team Collaboration
- Natural language query
- API Accessible

1
4



Watson Explorer (cont'd)

Extracts information from text comprehensively so text data can be handled as structured data

Computes statistical scores such as frequency and correlation of extracted keywords

Visually displays scores so users can understand characteristics of information in text data

Key information is identified and categorized through annotation

Key information is

Content mining: Annotations are stored in a text index. Statistical scores (correlation, trends, etc.) are computed and visualized

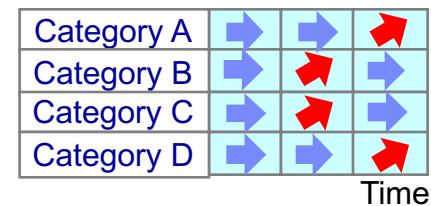
PC 143 (Hunter)
 15 June 2006 23:47
 Suspect identified himself as John Setsuko. Matched description given by night club doorman (IC1, Male, Ag 22-24 yrs, blue Everton shirt). Stopped whilst driving White Ford Mondeo, W563 WDL. Address given as 22 East Dene Ridge, Copdock, Ipswich. Searched at scene and found in possession of 1 oz Cannabis Resin and lockable pocket knife.



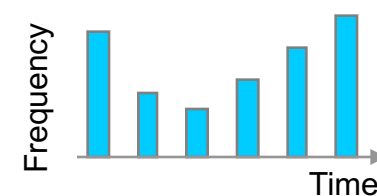
Arresting_Officer	PC 143
Arrest_Date_Time	15/06/2006 : 23:47
Suspect_Forename	John
Suspect_Surname	Setsuko
Suspect_VRN	W563WDL
Suspect_Vehicle_Color	White
Suspect_Vehicle_Make	Ford Mondeo
Suspect_Addr_Street	22 East Dene Ridge
Suspect_Addr_Town	Ipswich
Evidence_1_Description	1 oz Cannabis Resin
Classification	Drug possession



Trends Analysis



Time Series Analysis



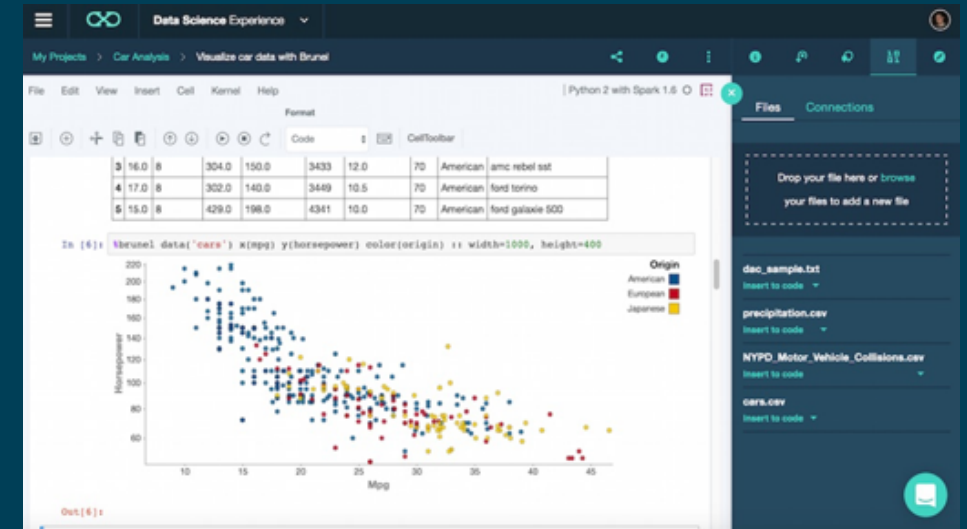
2D Map Analysis



IBM Data Science Experience (Predictive)

- Enterprise Grade Data Science platform
- Team Collaboration Features
- Curated Open Source Packages
- Multiple IDEs
 - Jupyter, Zeppelin, H2O Flow, RStudio
- GPU Support
- Scalable Microservices Cloud Architecture
- Community: tutorials, examples and experts
- Model Management and Deployment
- Integrates IBM's Data Science Portfolio

1
6



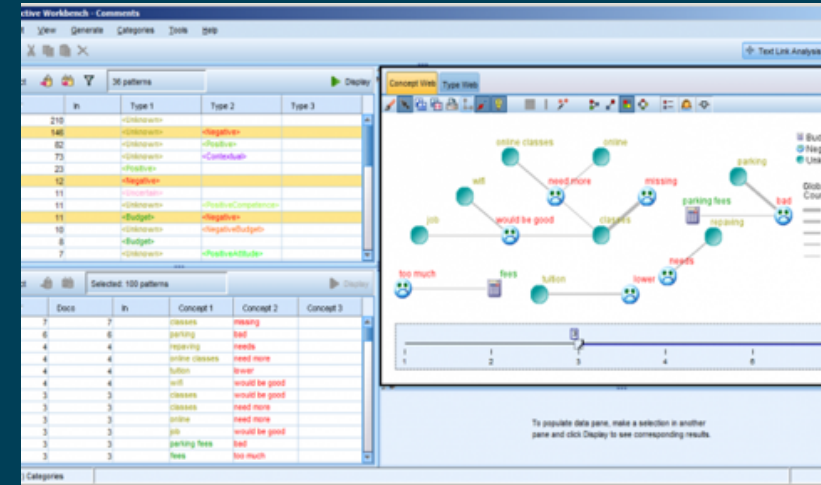
Enterprise Grade Data Science Platform

IBM SPSS Modeler (Predictive)

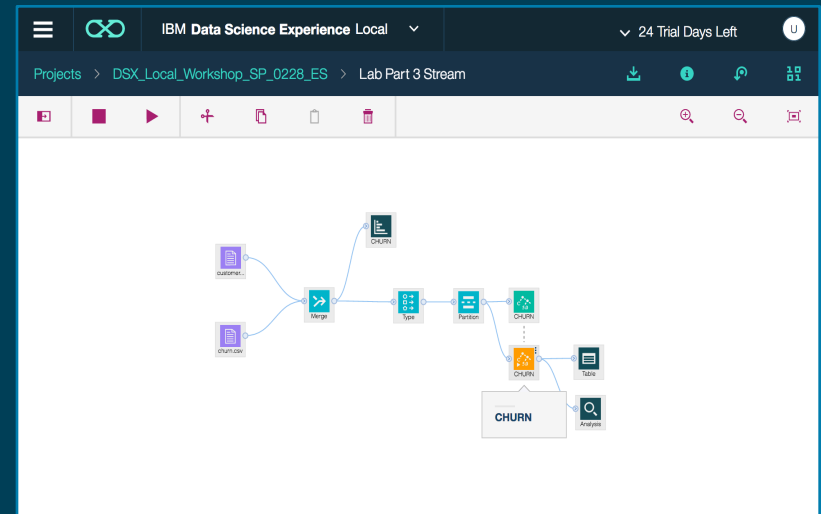
- “Drag and Drop Data Science”
- Accelerate time to value – from data discovery to machine learning and application development
- Powerful single environment for data, algorithms, model development and machine learning
- Out-of-box industry leading algorithms and capabilities
- Mission-critical deployment and scale

17

SPSS Modeler on Desktop



SPSS Modeler in DSXL



IBM Decision Optimization (Prescriptive)

State-of-the-art set of optimization modeling tools, APIs and algorithms.

Planning and scheduling of scarce resources:

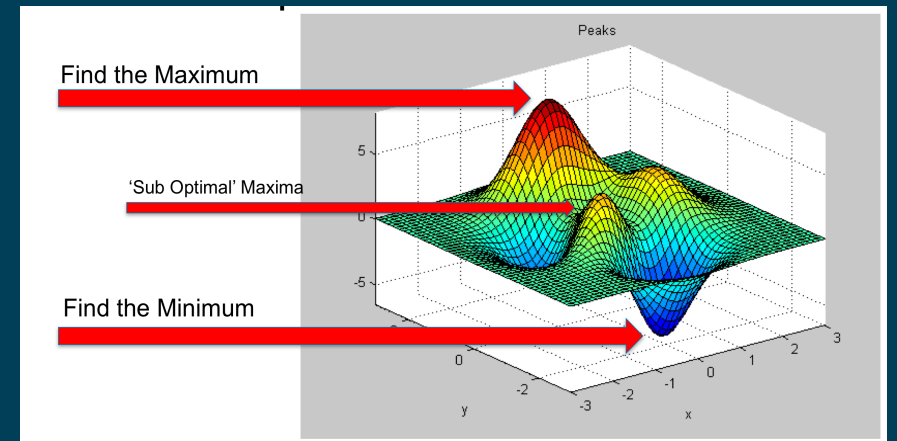
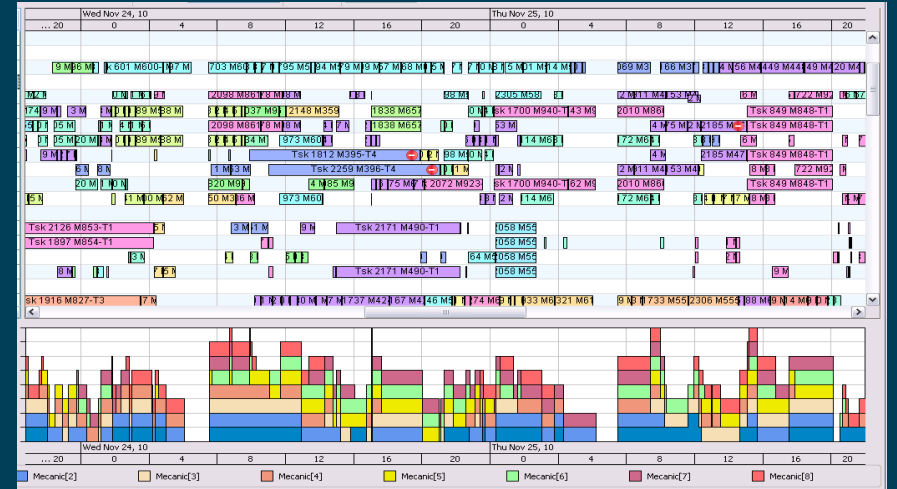
- Supply chain management
- Price optimization
- Product assortment
- Workforce planning.

Offerings:

- CPLEX Optimization Studio (COS)
- Decision Optimization on Cloud (DOcplexCloud)
- DSXL add-on component (DO4DSX)

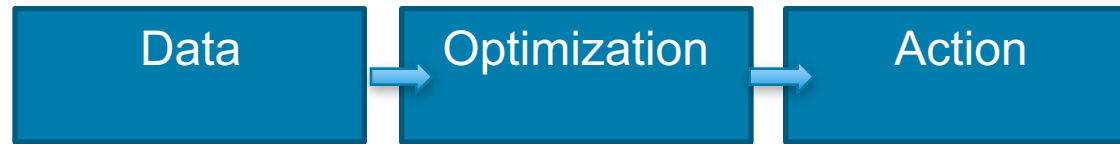


IBM Decision Optimization

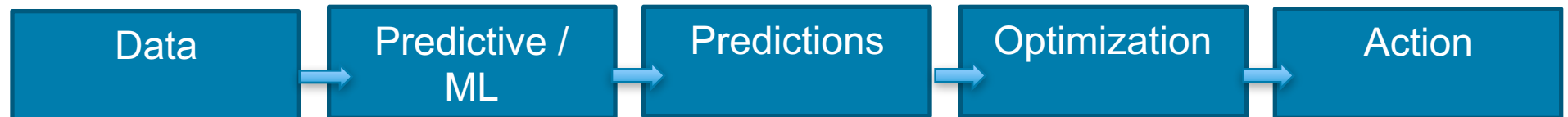


Moving From Traditional Optimization to Prescriptive Analytics

Old:
Optimization



New:
Prescriptive
Analytics



Example:

1. Predict Demand for Window Air Conditioners by Zip Code based on Weather Forecast
2. Optimize Quantity of Window Air Conditioners in Store to Maximize Sales
3. Ship x quantity to each store

DSXL Model Management and Deployment

■ **Features:**

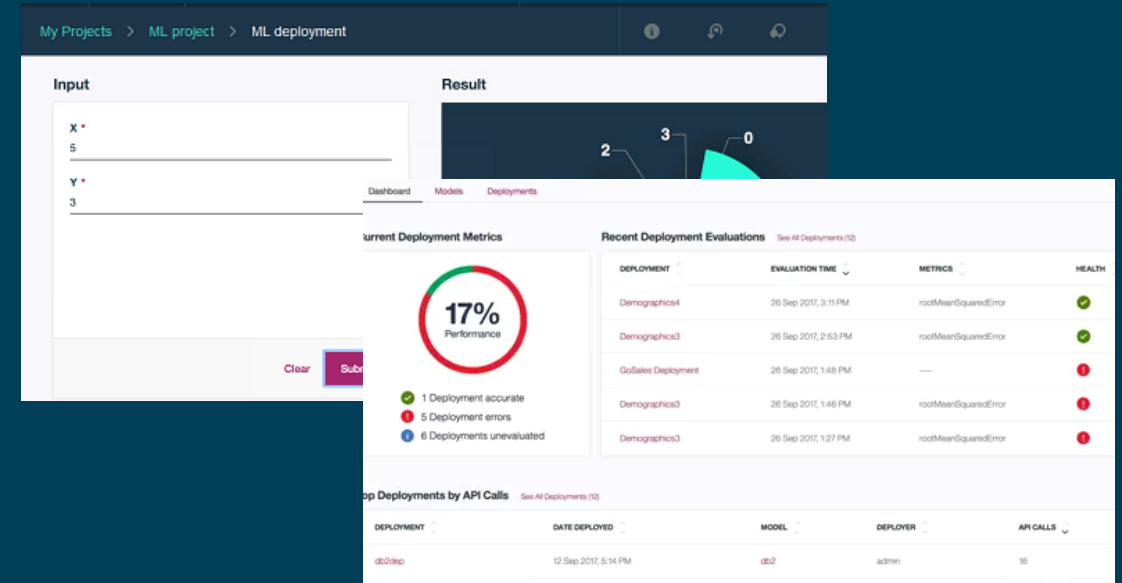
- **Deploy Python, R, & Spark Models online or batch**
- **Track model accuracy and schedule evaluations**
- **Load-balancing support**

■ **Pain Points:**

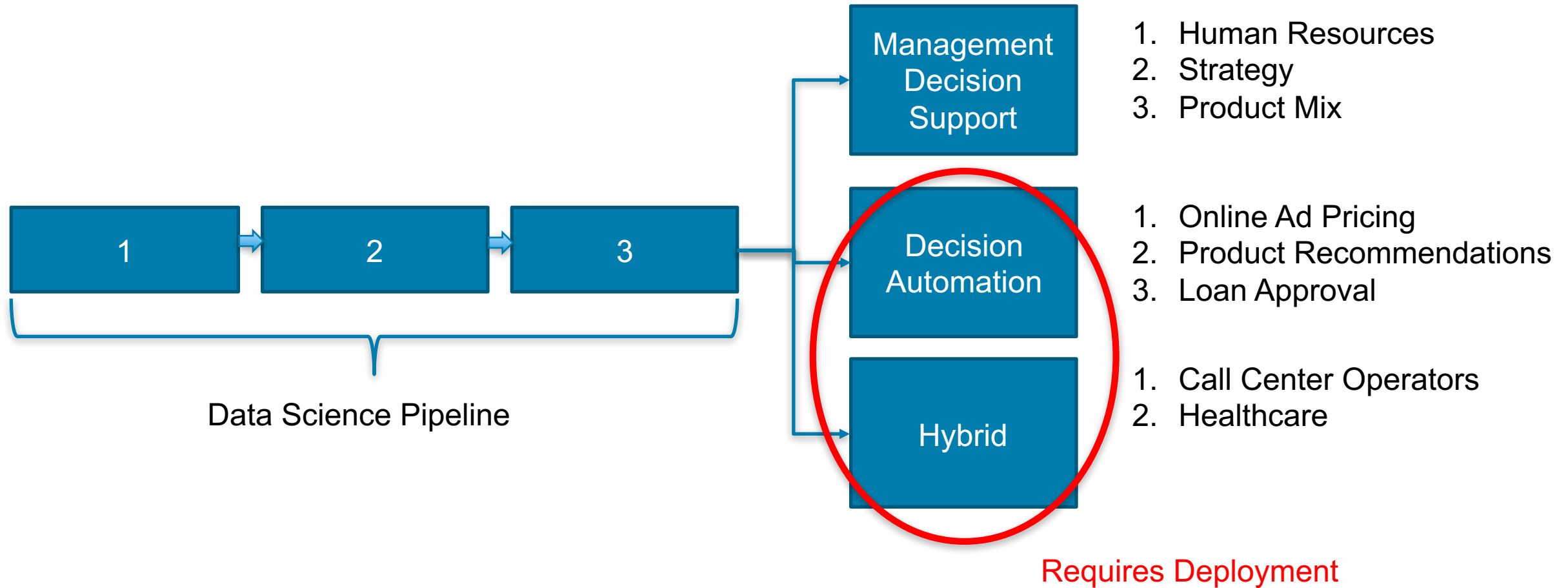
- **Lack of defined processes or standards for model deployment**
- **Handoffs from the teams responsible for development to production**
- **Difficult to compare and validate models value**

■ **Value Proposition:**

- **Get models into production faster**
- **Keep models performing at their best**
- **Securely manage the development to production workflow**



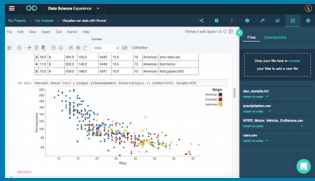
End Uses Of Data Science



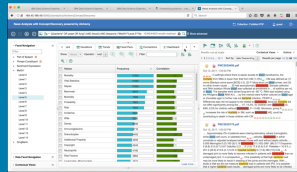
Data Science Experience

Putting It All Together

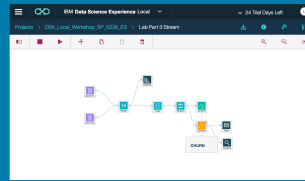
DSX IDEs



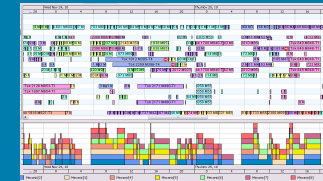
Watson Explorer for DSX



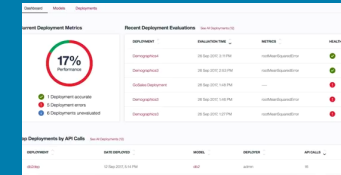
SPSS Modeler for DSX



Decision Optimization for DSX



Model Management and Deployment



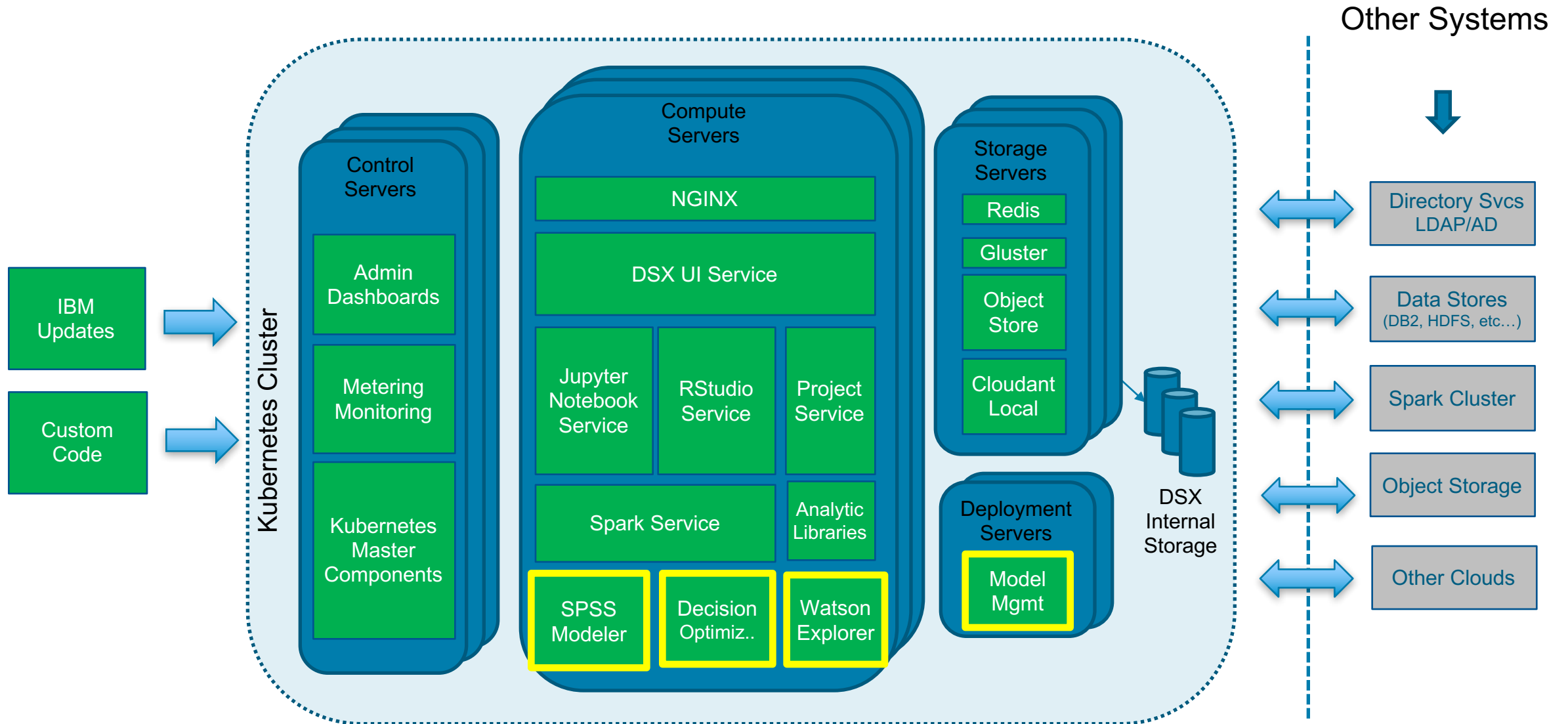
Data Science Experience



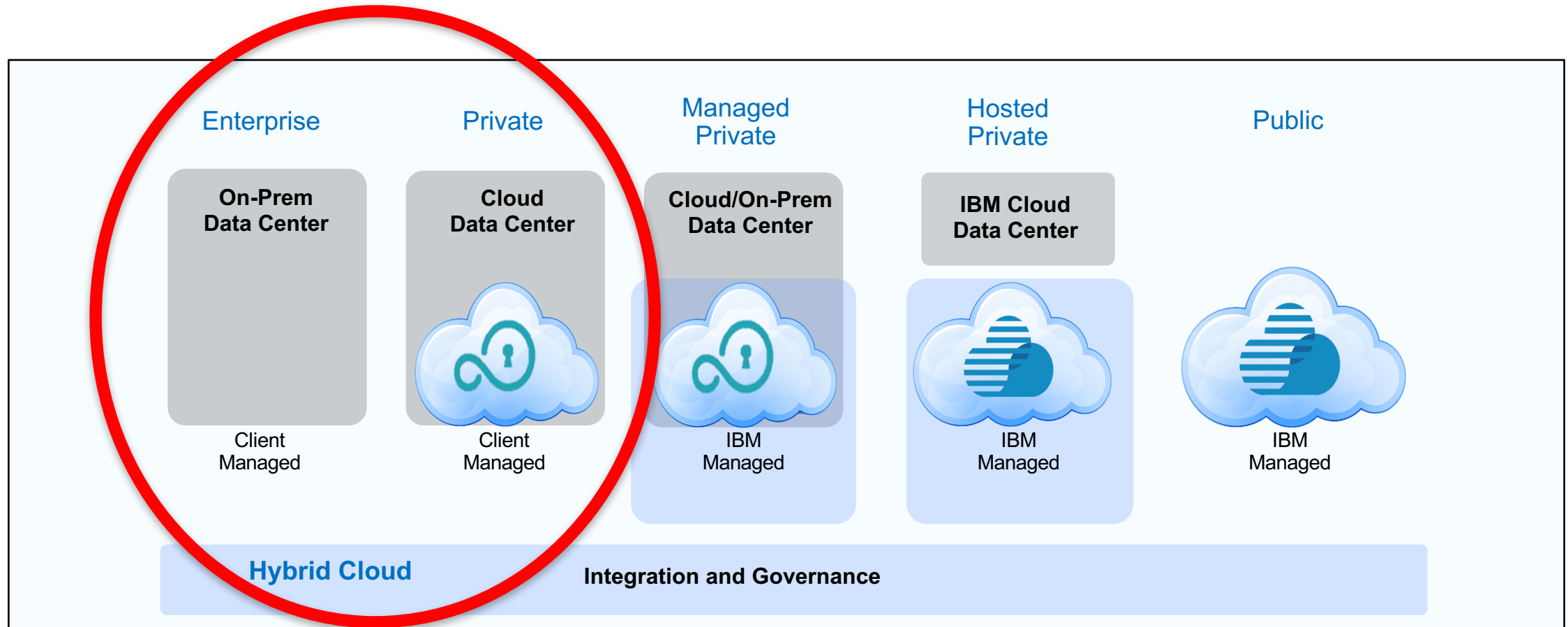
ICP4Data – More than Just DSX

	ICP for Data <i>End-to-end data management, governance, & data science</i>	DSX Local <i>standalone enterprise data science</i>
Analyze	<p>Data Science & ML Visualization & Reports etc.</p>	
Organize	<p>Data Catalog Data Integration / ETL Data Quality, MDM etc.</p>	
Collect	<p>IBM Data Stores (e.g. Db2) Non – IBM Data Stores Federation / virtualization etc.</p>	

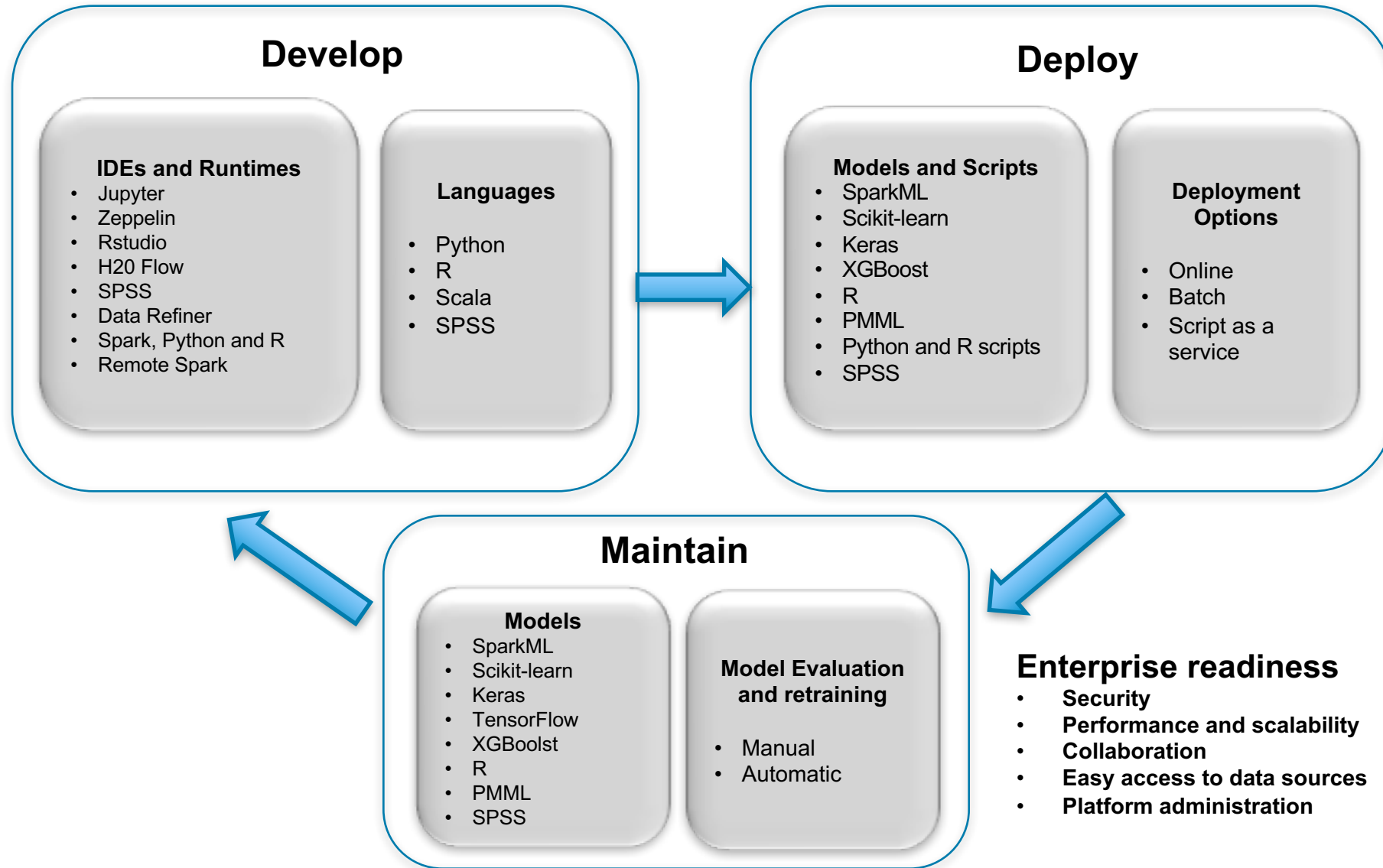
Data Science Experience Local - Architecture



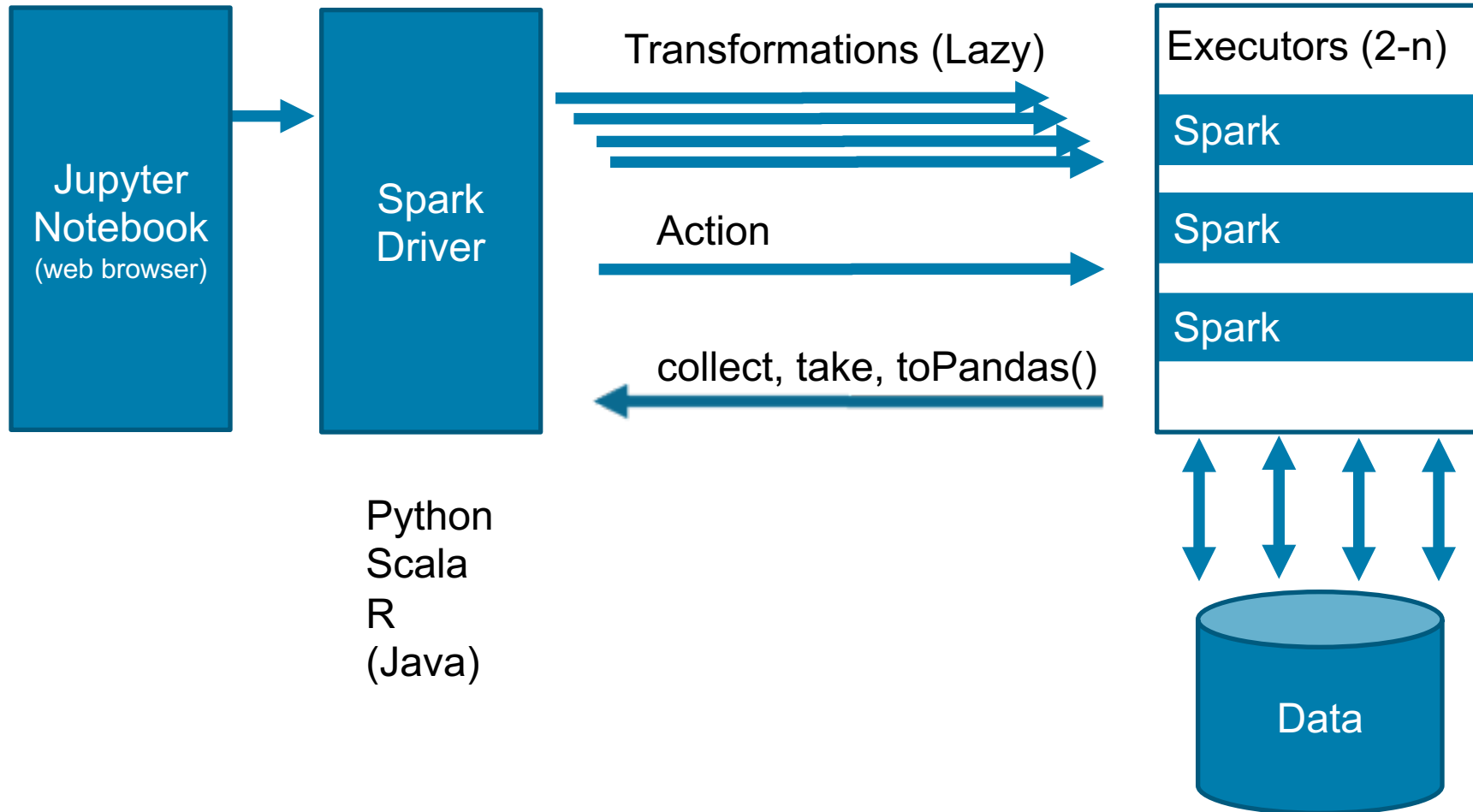
Installation For DSX Local



DSX Local



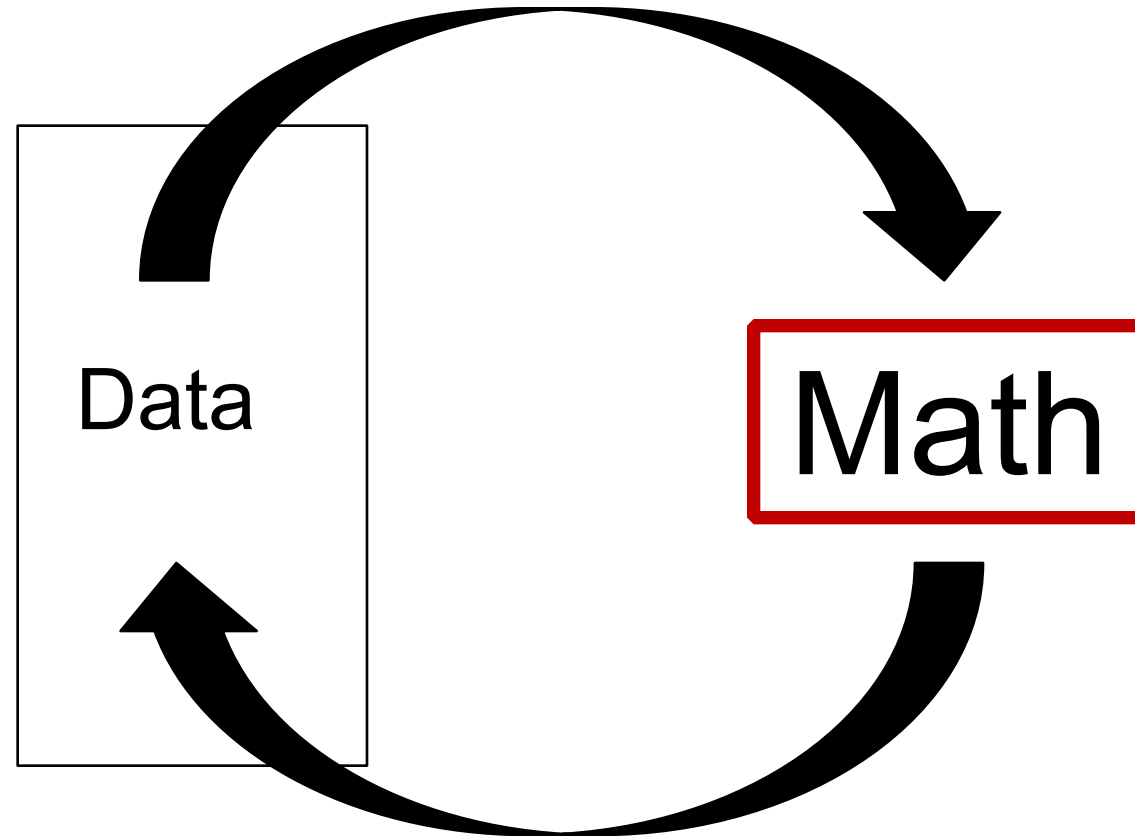
DSX - Spark Architecture



Crash Course on Machine Learning

Machine Learning

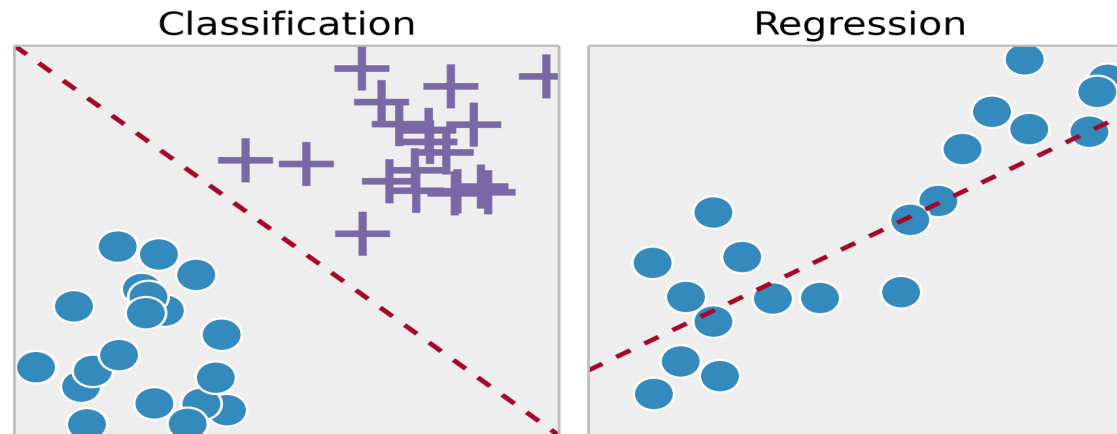
- Systems that learn from data using math



Supervised Learning

- Most Basic Type of Machine Learning
- Known 'Target' Values
- Either:

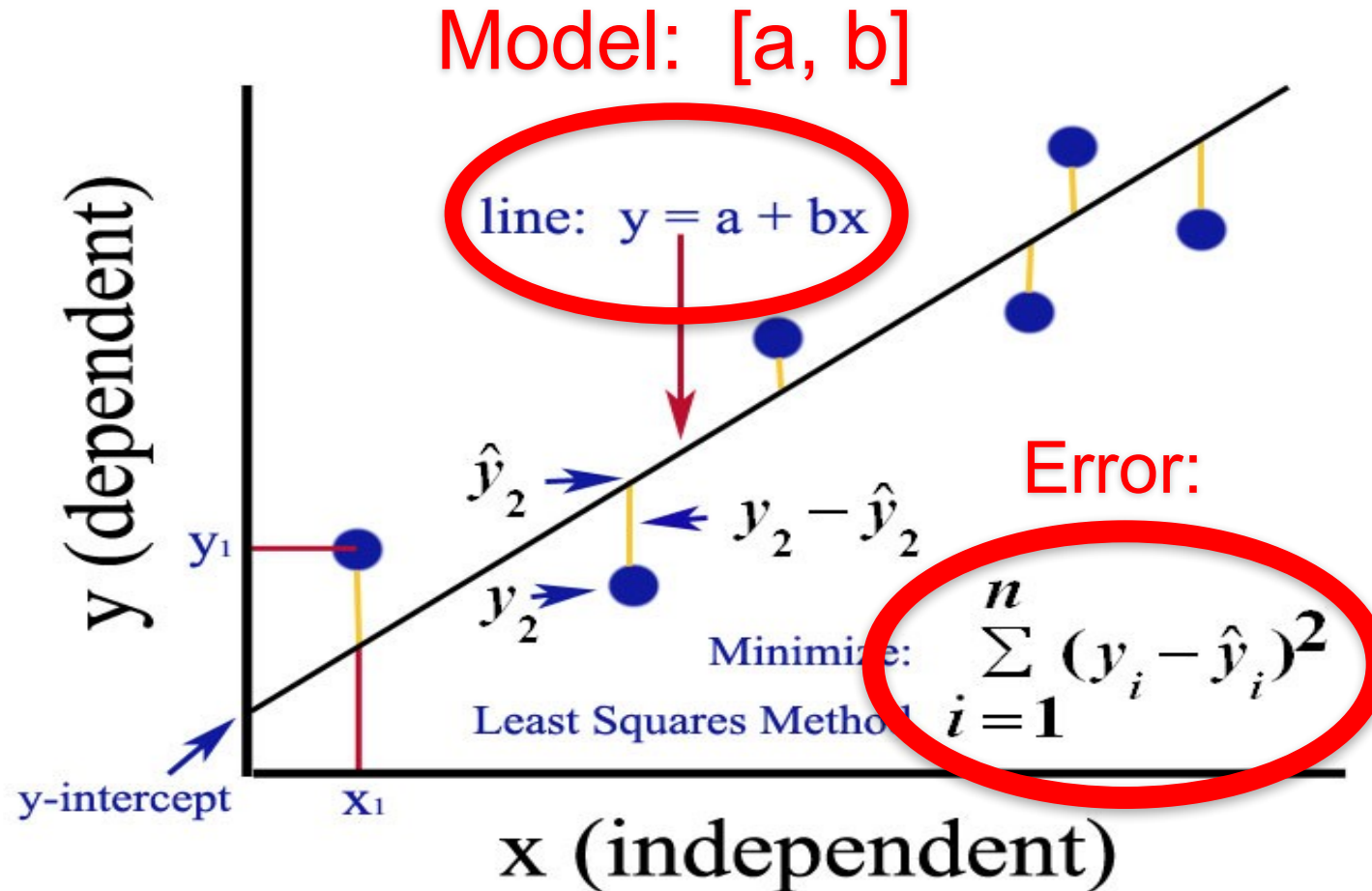
- Classification
- Regression



Source: <http://ipython-books.github.io/featured-04/>

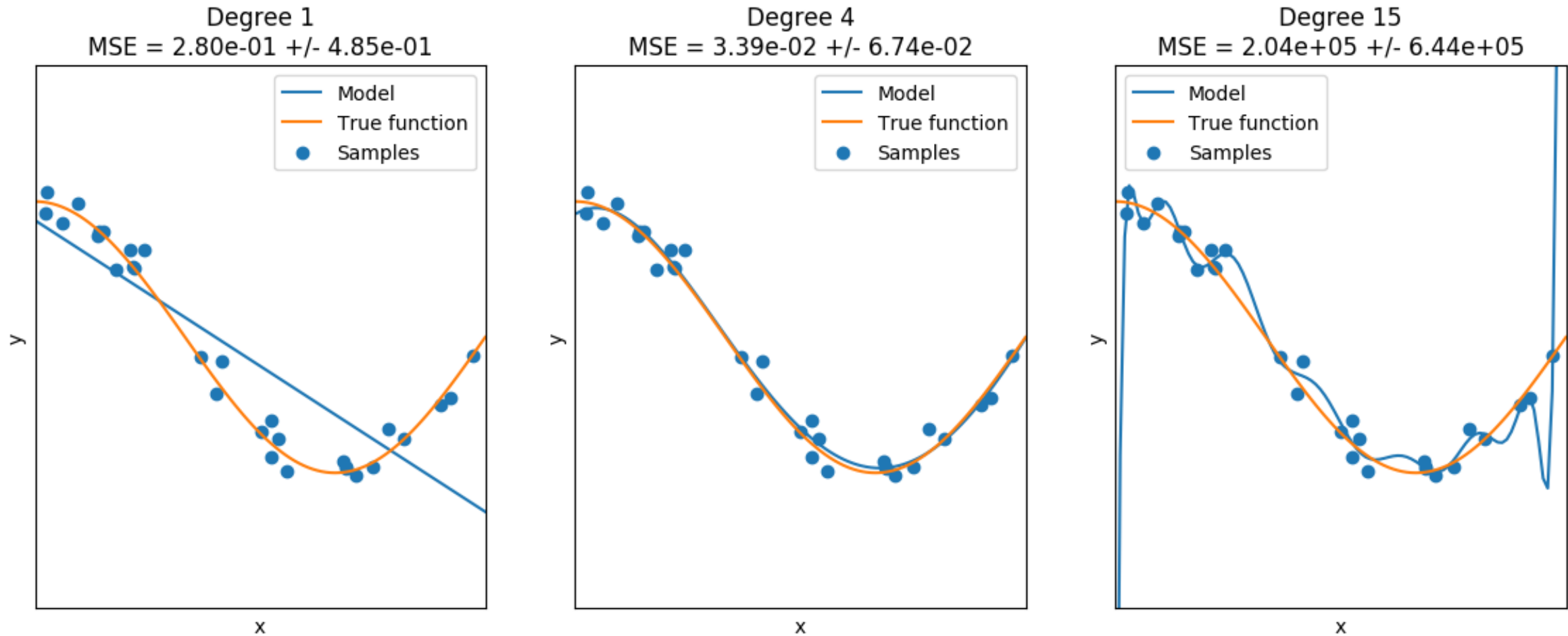
- Other types of ML include:
 - Unsupervised Learning, Recommender Systems, Anomaly Detection, Reinforcement Learning, Deep Learning, Cognitive Systems

What is an Algorithm ? – Example: Ordinary Least Squares



Source: https://bookdown.org/sbikienga/Intro_to_stat_book/

Underfitting vs Overfitting



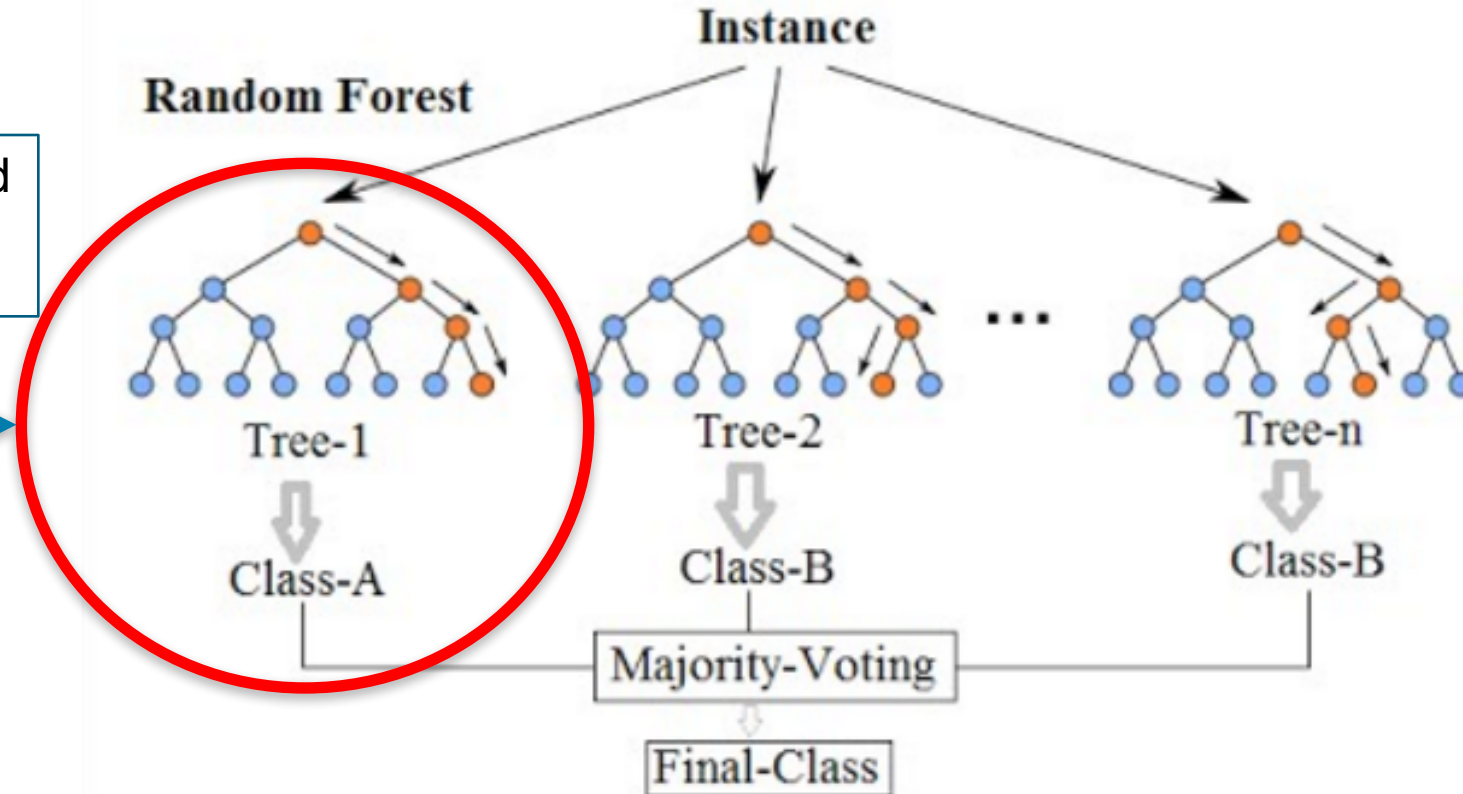
Source: http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

Understanding Random Forests

Random Forest Simplified

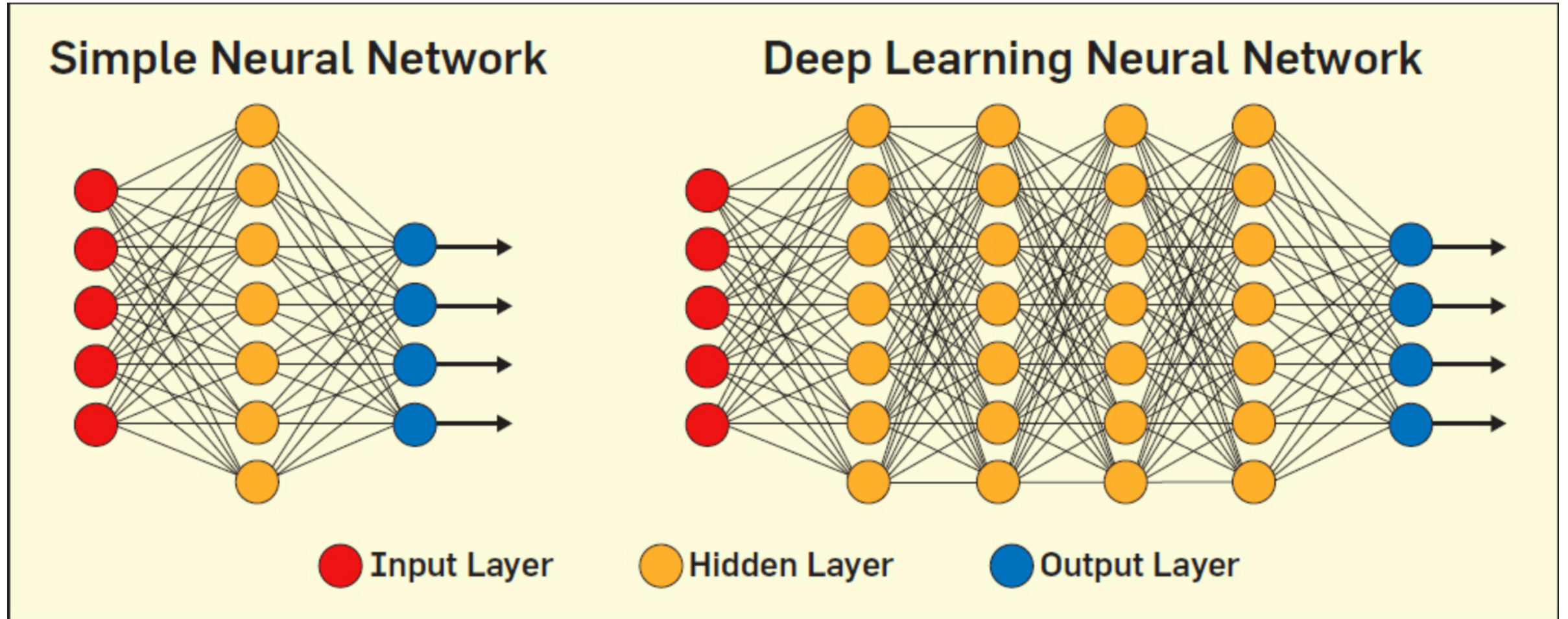
Each Decision Tree is Trained

- Random Set of Columns
- Random Set of Rows



Source: <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>

Understanding Neural Networks

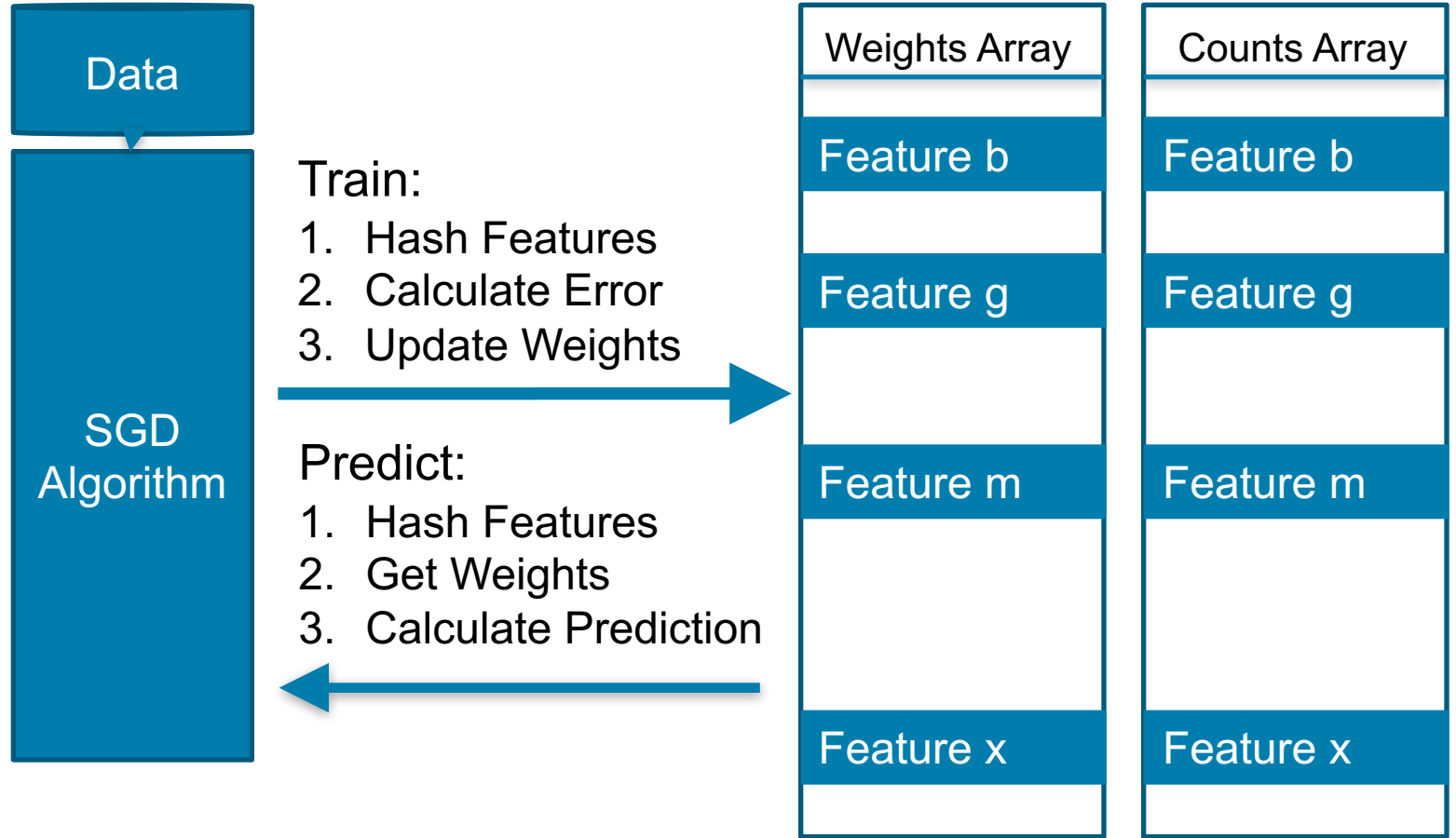


Source: <https://www.quora.com/What-is-the-difference-between-Neural-Networks-and-Deep-Learning>

Stochastic Gradient Descent for Online (Real Time) Learning

- Benefits:**
1. Fast
 2. Fixed space and time requirements, not dependent on data size
 3. Continuous Training
 4. Don't need to store historical data
 5. Works with both structured and unstructured data
 6. Tolerant of missing values
 7. Model adapts to changes
 8. Classification or Regression

- Drawbacks:**
1. Not as Accurate.
 2. Is essentially a linear model.



Fixed Size Array(s) – indexed by hash value

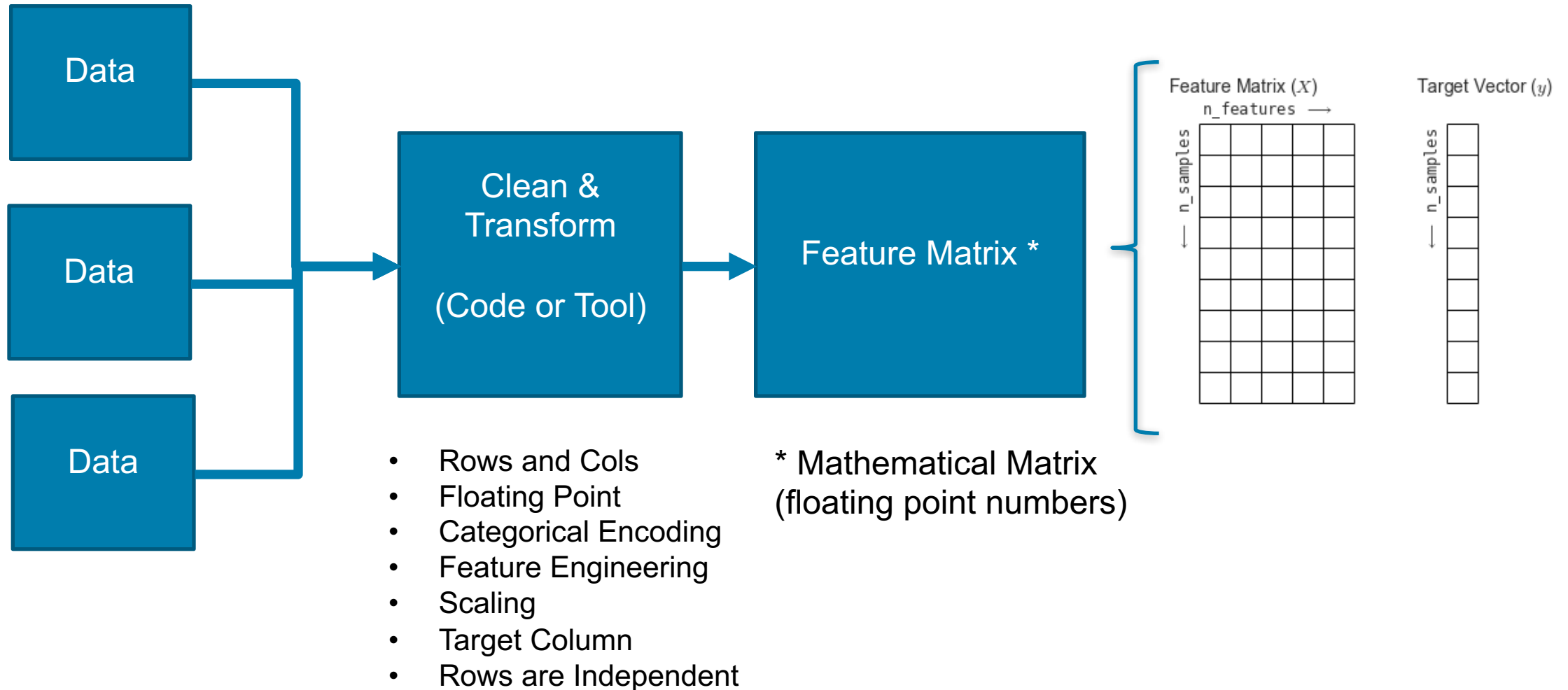
Step 1: Identify Business Objective

- What prediction would improve your business?
 - Be Realistic, ***This is not Magic, it is Math !***

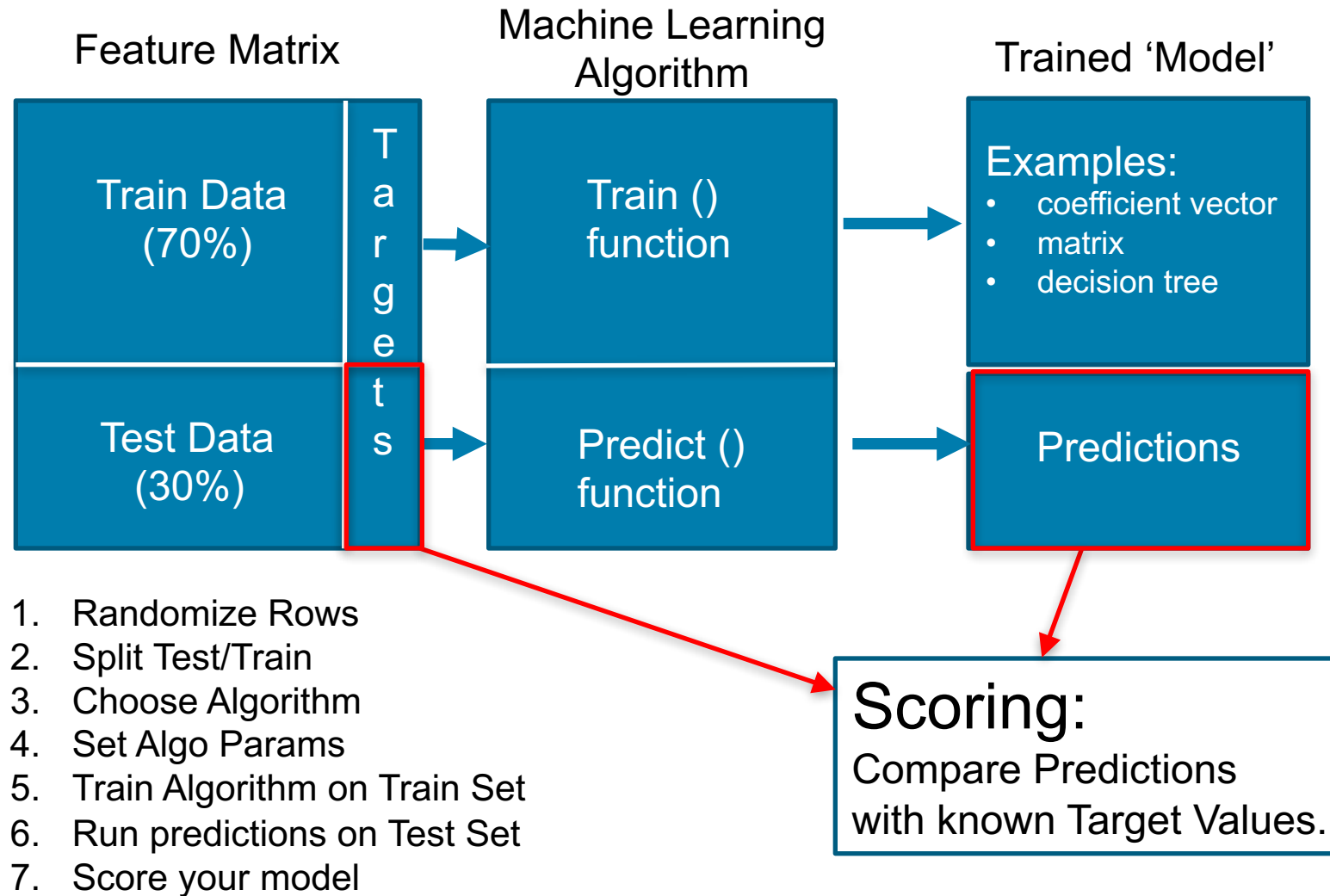
- What data do you have? Or can you get?

- Identify Key Performance Indicator
 - Cost Reduction
 - Revenue Enhancement
 - Click Thru Rate
 - Save Lives

Step 2: Generate Feature Matrix From Historical Data

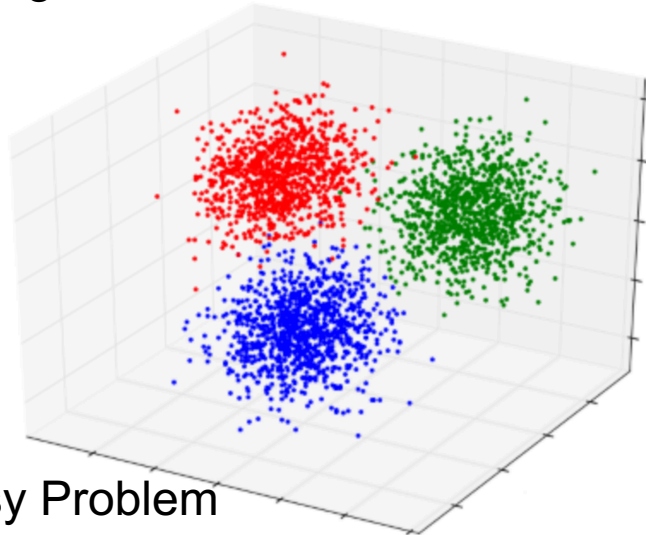


Step 3: Training and Scoring Process

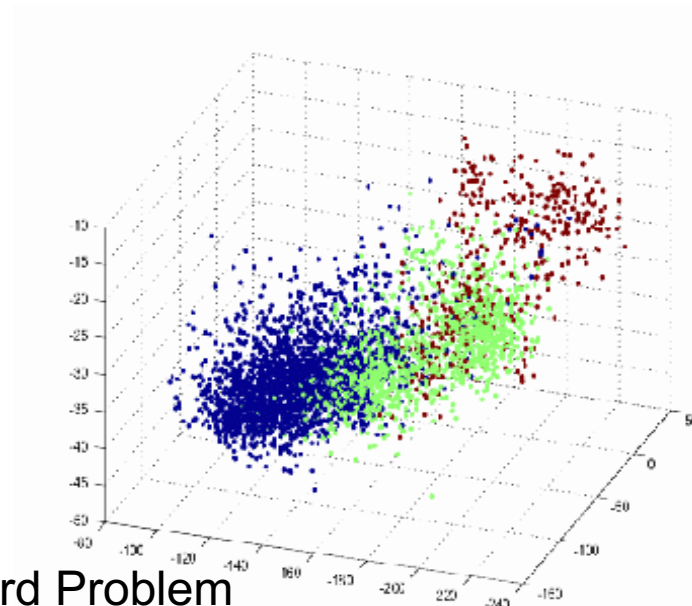


Step 3: Model Scoring

- *The true test of a machine learning model is making accurate predictions on unseen data.*
- Different Scoring Techniques for Different Problems
- Some Problems are Easier than Others
- **Experiment** with different techniques to improve score
 - Different ML Algorithms and Parameters
 - Different Mathematical Representations
 - Engineer New Features



Easy Problem



Hard Problem

Step 4: Model Deployment: Moving your machine learning model into production

Model Development

Model Deployment

Custom Coding

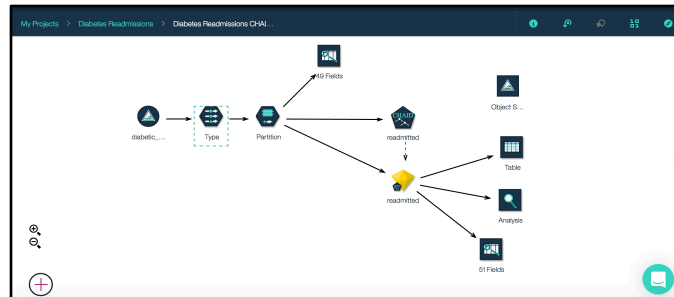
```

Almost ready to train. Let's define two common functions used in the training process
Here we declare two functions we'll re-use in our ML iterations.

In [21]:
# For ml test_classifier show what it sounds like
# Parameters are the instantiated classifier object and the 4 dataframes mentioned just above.
def test_ml_test_classifier(classifier, x_train, y_train, x_test, y_test):
    print
    print classifier
    # fit is the method to perform the training part of the process,
    # so provide both the input training matrix (x_train) and the target answer (y_train)
    classifier.fit(x_train, y_train)
    # the predict method returns predictions as 1 or 0
    y_pred = classifier.predict(x_test)
    # the predict_proba returns predictions as a probabilities matrix, that is, a probability vector for each row in the x_test matrix
    y_prob = classifier.predict_proba(x_test)
    # Return just the probability for Survived=1
    y_prob = y_prob[:,1] # if test all rows, but keep only second column, that is, the probability of survived = 1
    # Call the calc_classifier_stats to print basic statistics on the quality of our classifier for comparison
    calc_classifier_stats(x_test, y_pred, y_prob)
    print
    # Some classifier algorithms, also mentioned as "feature importances". Most columns are most useful in prediction.
    # Loop through the columns printing the name of the column and the importance value for that column
    print
    print "Feature Importances:"
    print
    print classifier.feature_importances_
    print
    print "Classification Feature Importances:"
    print
    print calc_classifier_stats(x_train, y_train, x_test, y_test)

```

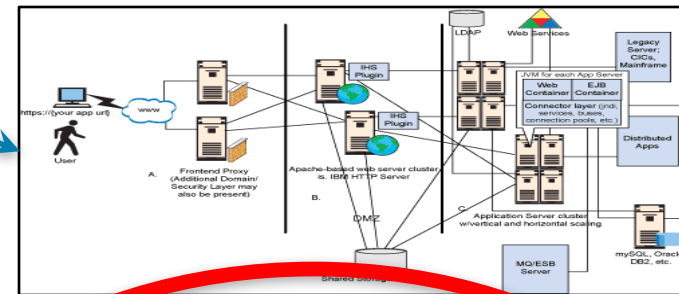
SPSS Modeler Flows



Automatic Model Builder

The screenshot shows the 'Automatic Model Builder' interface. It displays details for a model named 'Hospital Readmissions Automatic Model'. The model uses 'WML' (Watson Machine Learning) as the machine learning service. The label column is 'Readmitt_30days'. The training data scheme is 'View', the input data scheme is 'View', the runtime environment is 'spark-2.0', and the training date is '28 Jul 2017, 2:37 PM'. There is a 'Deployments' section at the bottom.

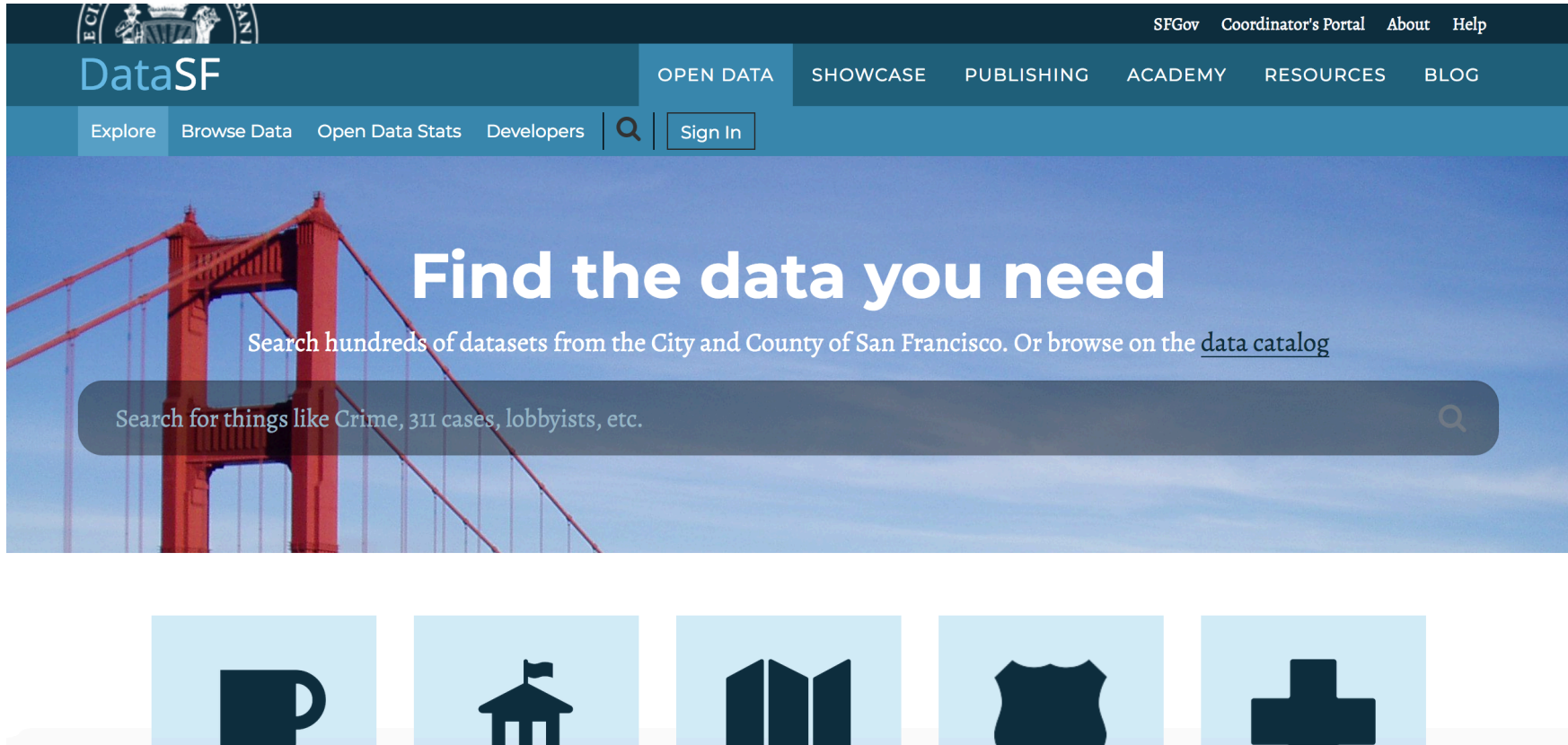
Custom built deployment

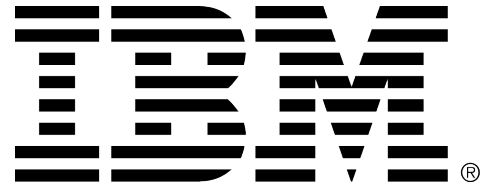


Deployment Server

The screenshot shows the IBM Bluemix Service Catalog interface. It displays two services: 'SPSS Streams Service' and 'Watson Machine Learning'. The 'SPSS Streams Service' section includes instructions: 'Create your data model using IBM SPSS Modeler (free Trial)', 'Upload your models to the Watson Machine Learning service', and 'Call the scoring API from your app'. The 'Watson Machine Learning' section includes instructions: 'Train, test models and score data using the powerful Spark MLlib or Python scikit-learn', 'Collaborate with Data Scientists using Data Science Experience', and 'Deploy and manage models as realtime REST APIs, batch jobs (beta), or stream processing pipelines (beta)'. Both services have a 'Launch Dashboard' button.

To The Demo – San Francisco Fire Department





Legal Disclaimer

- © IBM Corporation 2018. All Rights Reserved.
- The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.
- References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.
- If the text contains performance statistics or references to benchmarks, insert the following language; otherwise delete:
Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.
- If the text includes any customer examples, please confirm we have prior written approval from such customer and insert the following language; otherwise delete:
All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.
- Please review text for proper trademark attribution of IBM products. At first use, each product name must be the full name and include appropriate trademark symbols (e.g., IBM Lotus® Sametime® Unyte™). Subsequent references can drop "IBM" but should include the proper branding (e.g., Lotus Sametime Gateway, or WebSphere Application Server). Please refer to <http://www.ibm.com/legal/copytrade.shtml> for guidance on which trademarks require the ® or ™ symbol. Do not use abbreviations for IBM product names in your presentation. All product names must be used as adjectives rather than nouns. Please list all of the trademarks that you use in your presentation as follows; delete any not included in your presentation. IBM, the IBM logo, Lotus, Lotus Notes, Notes, Domino, Quickr, Sametime, WebSphere, UC2, PartnerWorld and Lotusphere are trademarks of International Business Machines Corporation in the United States, other countries, or both. Unyte is a trademark of WebDialogs, Inc., in the United States, other countries, or both.
- If you reference Adobe® in the text, please mark the first use and include the following; otherwise delete:
Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
- If you reference Java™ in the text, please mark the first use and include the following; otherwise delete:
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
- If you reference Microsoft® and/or Windows® in the text, please mark the first use and include the following, as applicable; otherwise delete:
Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.
- If you reference Intel® and/or any of the following Intel products in the text, please mark the first use and include those that you use as follows; otherwise delete:
Intel, Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- If you reference UNIX® in the text, please mark the first use and include the following; otherwise delete:
UNIX is a registered trademark of The Open Group in the United States and other countries.
- If you reference Linux® in your presentation, please mark the first use and include the following; otherwise delete:
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both. Other company, product, or service names may be trademarks or service marks of others.
- If the text/graphics include screenshots, no actual IBM employee names may be used (even your own), if your screenshots include fictitious company names (e.g., Renovations, Zeta Bank, Acme) please update and insert the following; otherwise delete: All references to [insert fictitious company name] refer to a fictitious company and are used for illustration purposes only.