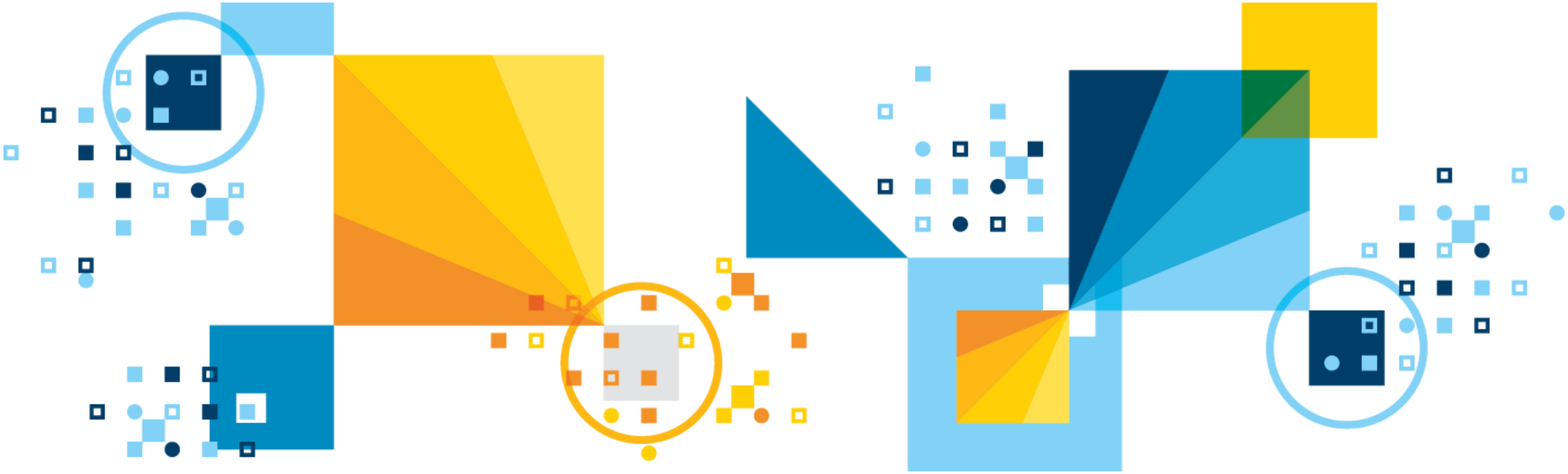
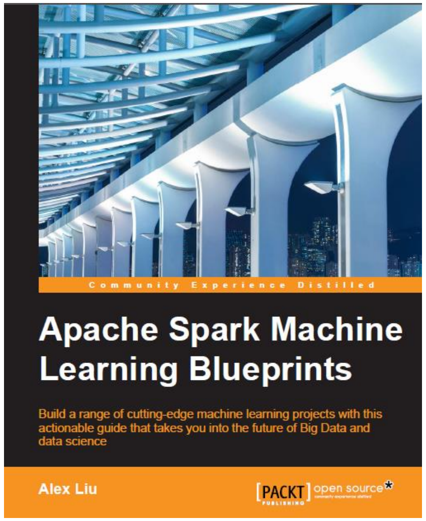
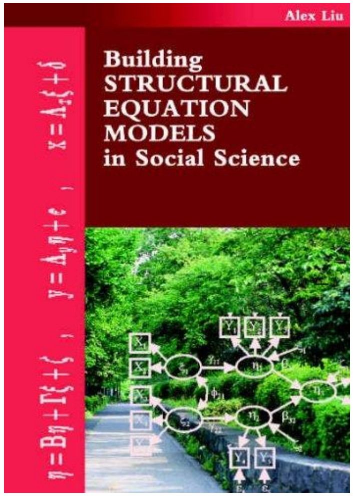
Turning Data to Value with Apache Spark and R on Big Data Platforms

Dr. Alex Liu, Chief Data Scientist, IBM Analytics Services



Alex Liu Introduction

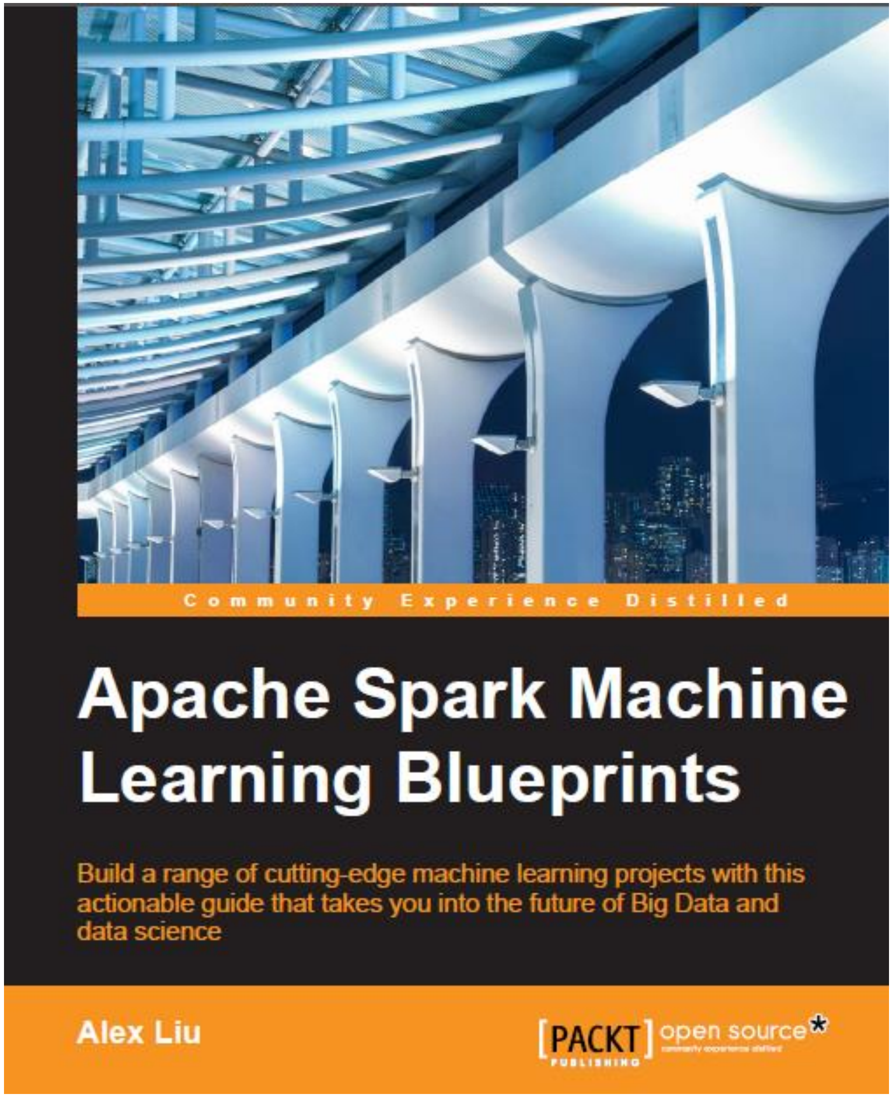

Community Experience Distilled
Apache Spark Machine Learning Blueprints
Build a range of cutting-edge machine learning projects with this actionable guide that takes you into the future of Big Data and data science
Alex Liu
[PACKT] open source*



- Chief Data Scientist – Analytics Services at IBM
- Chief Data Scientist for a few corporations before joined IBM
- Taught advanced data analytics for the University of South California and the University of California at Irvine
- M.S. and Ph.D. from Stanford University
- Started working on data analytics in ancient times

Apache Spark Machine Learning Blueprints

1. Spark for Machine Learning
2. Data Preparation for Spark ML
3. A Holistic View on Spark
4. Fraud Detection on Spark
5. Risk Scoring on Spark
6. Churn Prediction on Spark
7. Recommendation on Spark
8. Learning Analytics on Spark
9. City Analytics on Spark
10. Learning Telco Data on Spark
11. Modeling Open Data on Spark



Community Experience Distilled

Apache Spark Machine Learning Blueprints

Build a range of cutting-edge machine learning projects with this actionable guide that takes you into the future of Big Data and data science

Alex Liu

[PACKT] open source*

Today's Topics

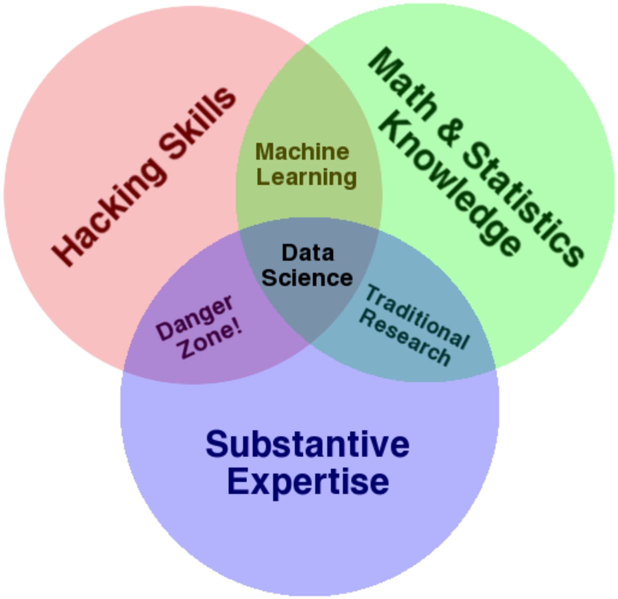
I. Big Data Driven Value

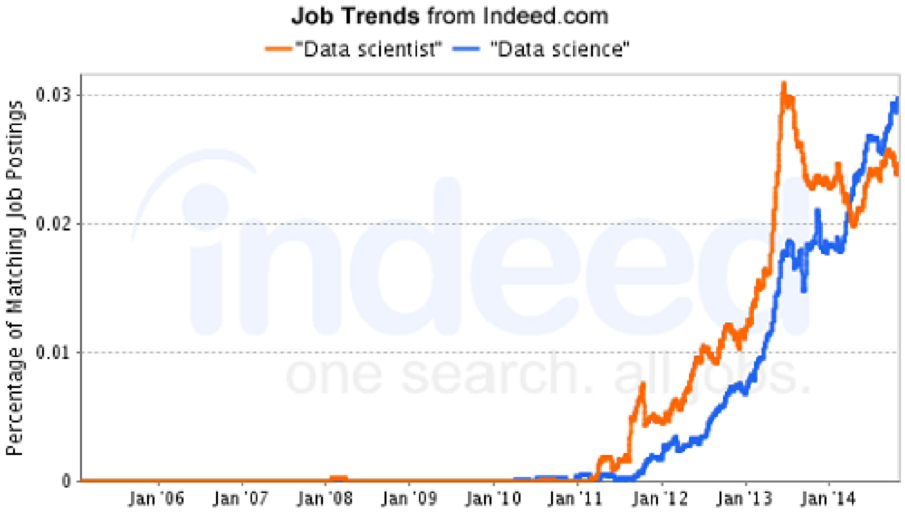
II. Data Science Challenges

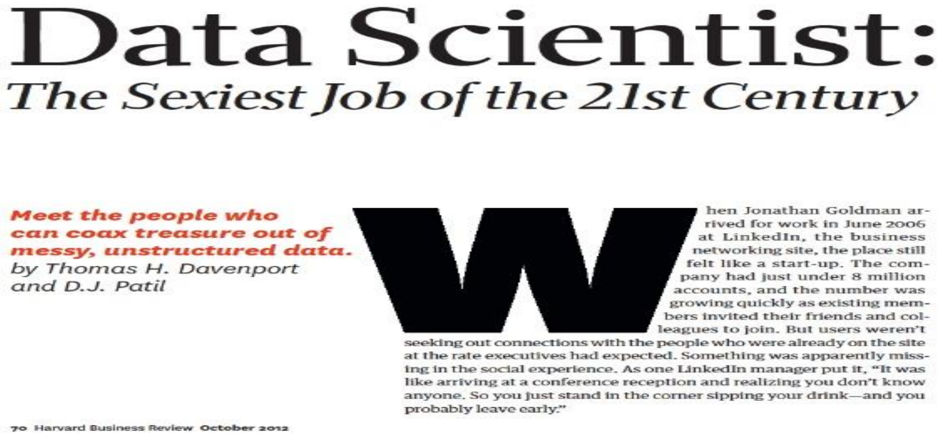
III. Apache Spark

IV. Integrated Approach

Data Scientist







Data Scientist: *The Sexiest Job of the 21st Century*

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

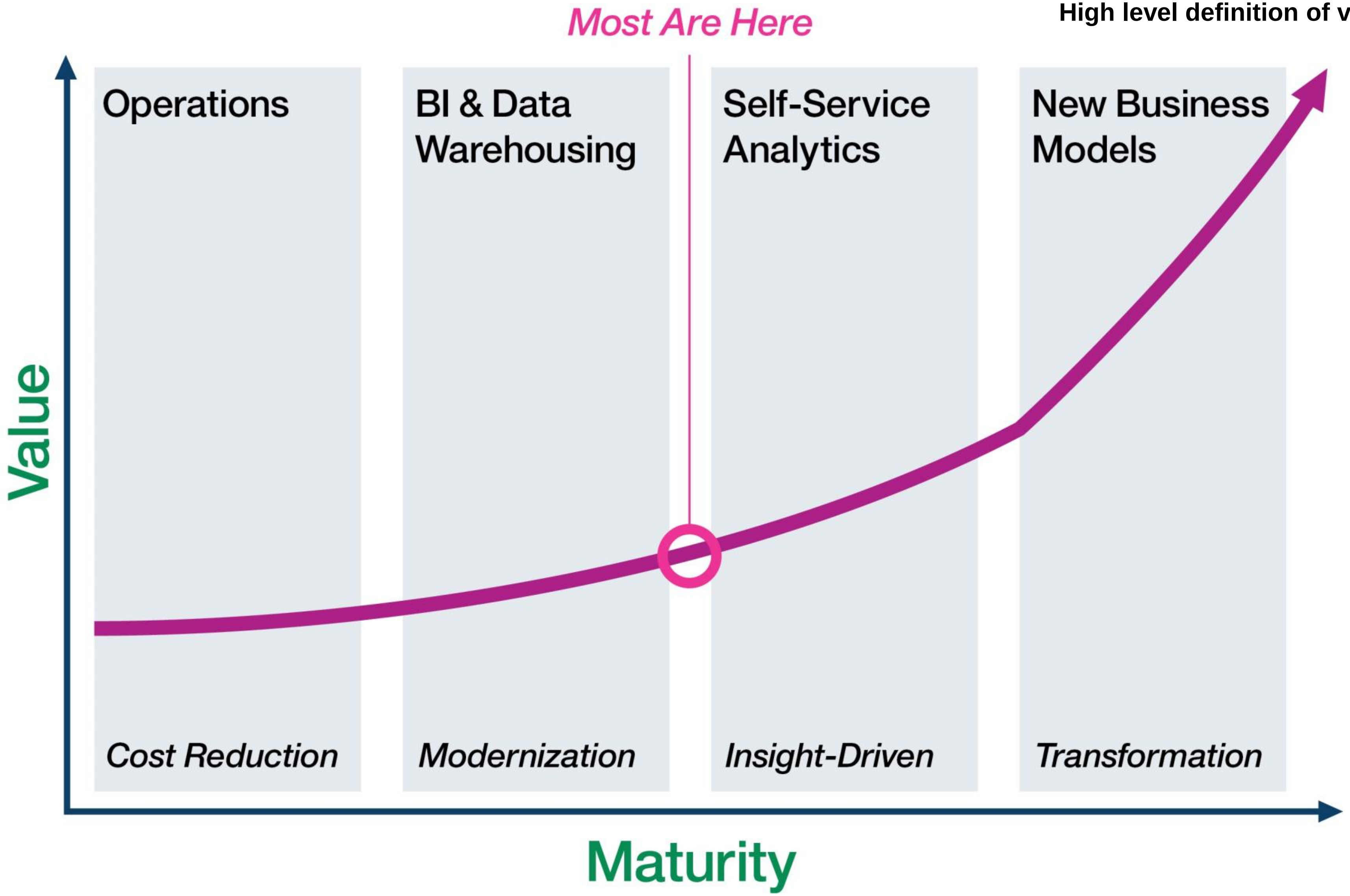
70 Harvard Business Review October 2012

Talk is over, and now

2017 – the year for VALUE.

Turning Data to Value (Data Science Maturity)

High level definition of value.

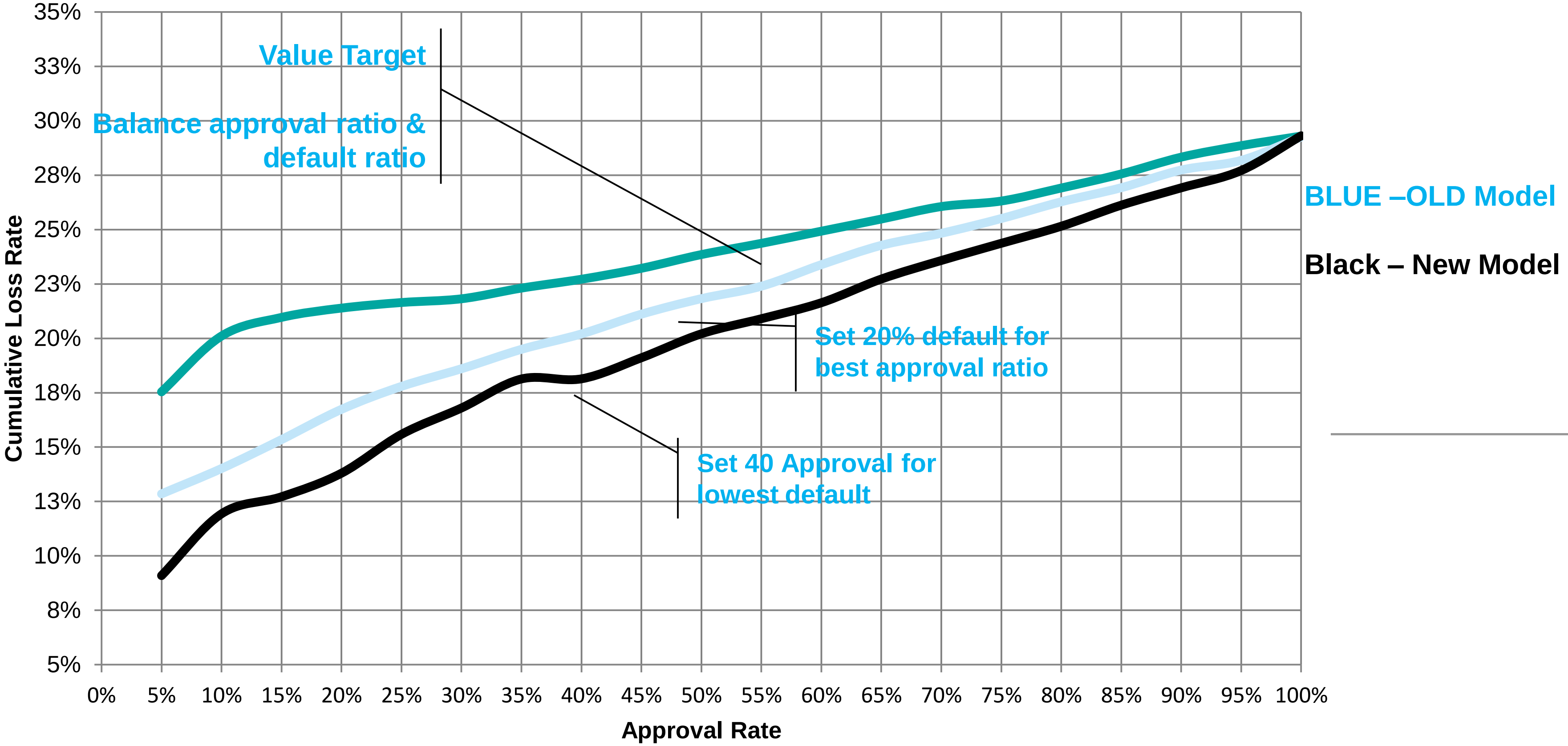


More info available here: https://ibm.biz/BdrxP7

Better models for more approvals with lower losses…

Specfic demand for value.

Cumulative Loss Rate by Approval Rate



Value Target

Balance approval ratio & default ratio

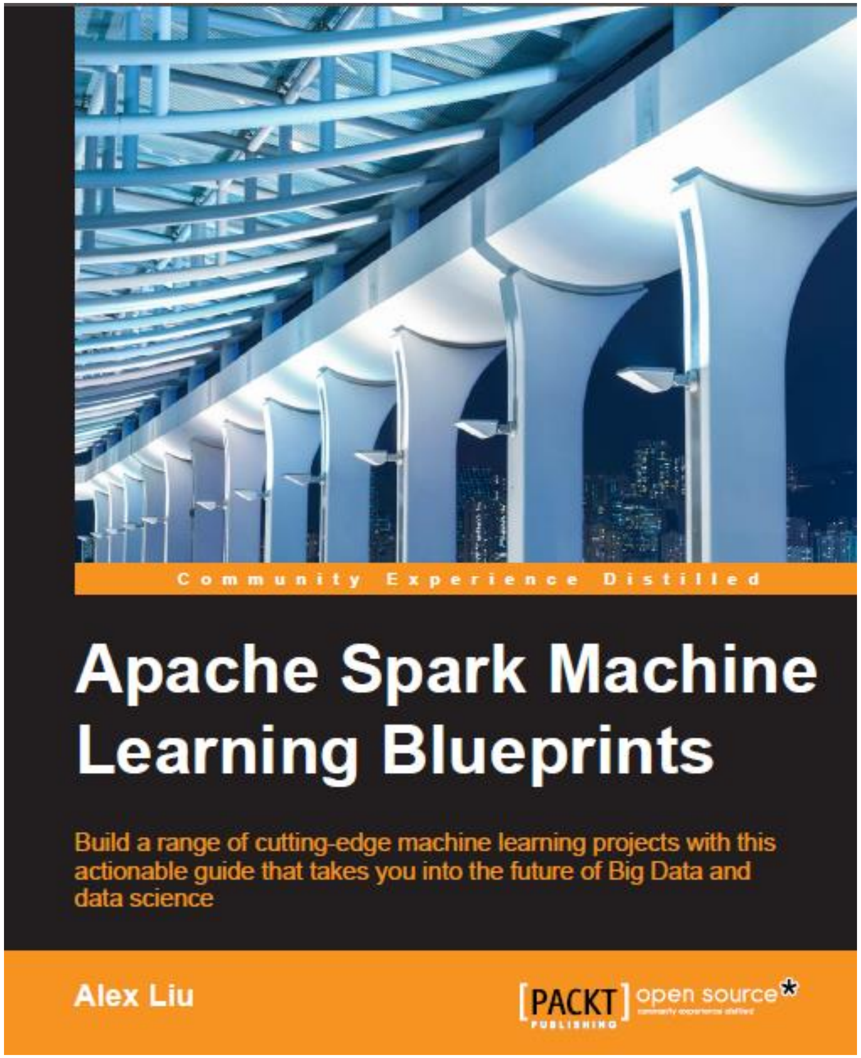
Set 20% default for best approval ratio

Set 40 Approval for lowest default

Confusion matrix for fraud detection, balance catch ratio and false positive ratio

BOOK - Apache Spark Machine Learning Blueprints
CHAPTER 5: Risk Scoring on Spark

- Spark for risk scoring
- Methods of risk scoring
- Data and feature preparation
- Model estimation
- Model evaluation
- Results explanation
- Deployment
- Summary



Community Experience Distilled

Apache Spark Machine Learning Blueprints

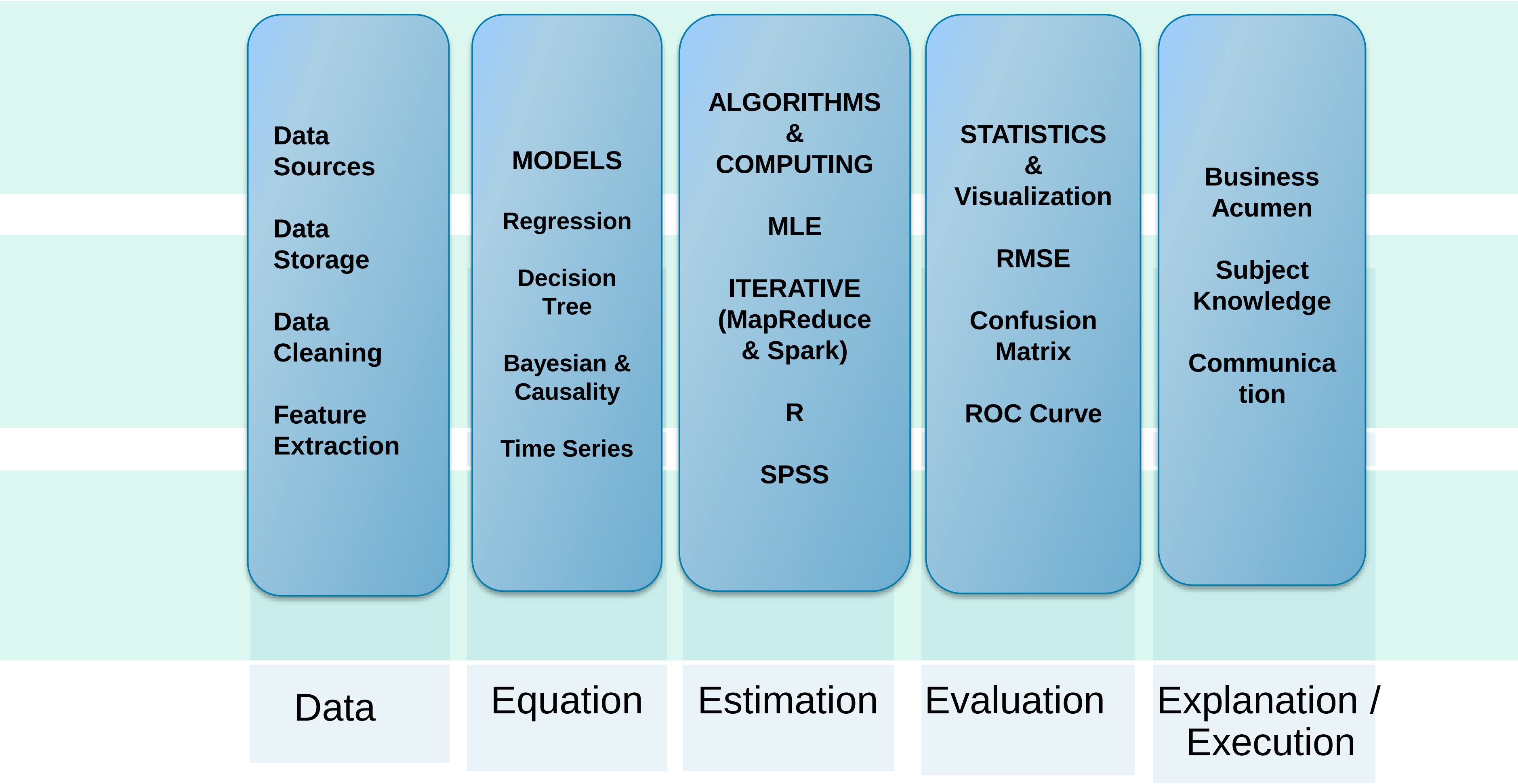
Build a range of cutting-edge machine learning projects with this actionable guide that takes you into the future of Big Data and data science

Alex Liu

[PACKT] PUBLISHING open source* community experience distilled

Data Science Challenges

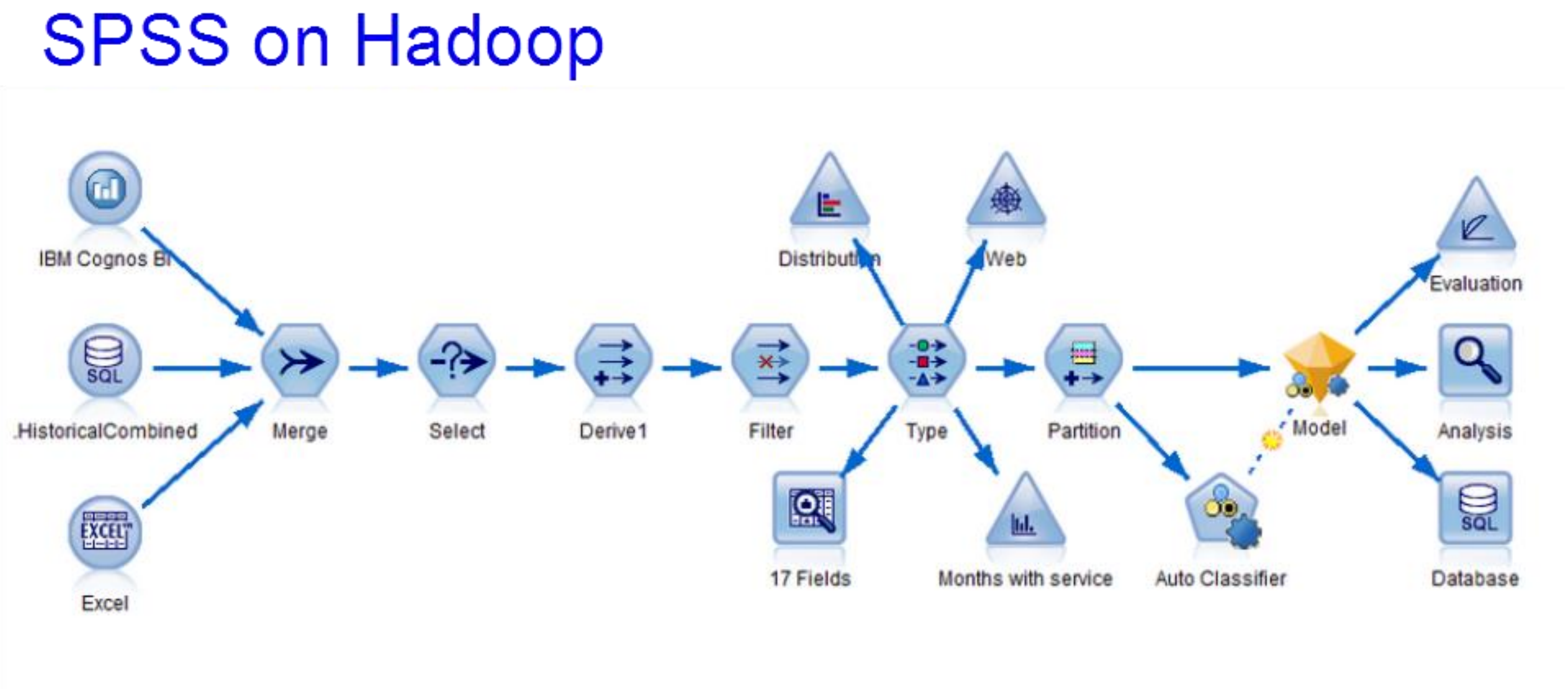
Data Science Framework – RM4Es Based Workflow



Data	Equation	Estimation	Evaluation	Explanation / Execution
Data Sources Data Storage Data Cleaning Feature Extraction	MODELS Regression Decision Tree Bayesian & Causality Time Series	ALGORITHMS & COMPUTING MLE ITERATIVE (MapReduce & Spark) R SPSS	STATISTICS & Visualization RMSE Confusion Matrix ROC Curve	Business Acumen Subject Knowledge Communication

Data science as workflows of data analytics tasks.

Data Science is a process



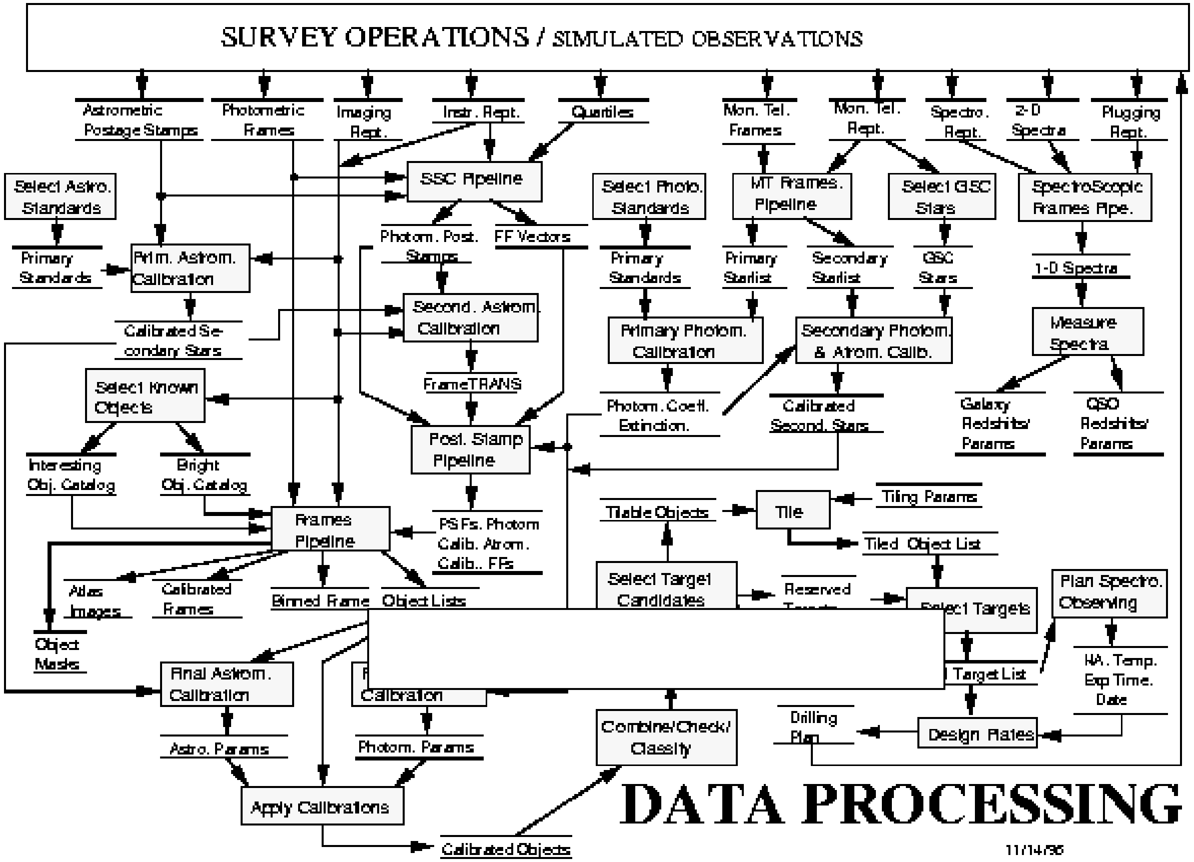
4Es – Equation – Estimation – Evaluation - Explanation

Research Become Very Complicated
- Research Flows Difficult to Manage

Number of models
to evaluate

Number of algorithms
to select

…



Challenges for Researchers

- Too much data to import
- Too much data cleaning to complete
- Too many analytical methods to select
- Too many algorithms to select
- Too many computing tools to select
- Too many IT systems to select/manage

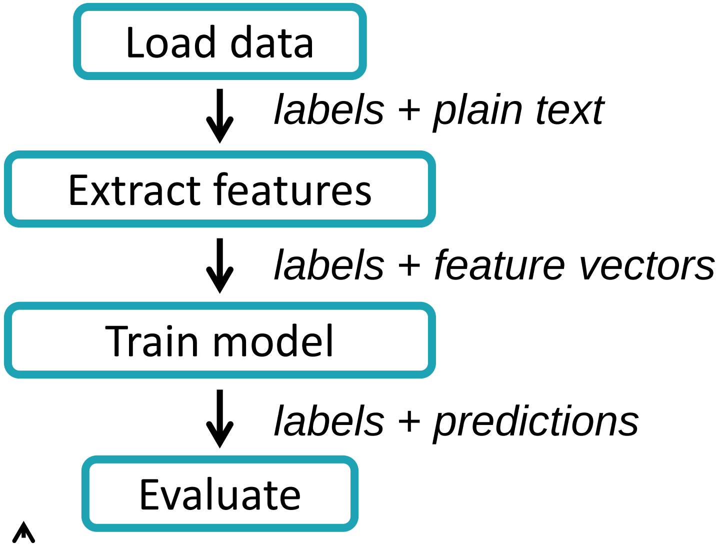
Where Apache Spark fits in, to help taking care of the above.

Apache Spark
- why data scientists like it

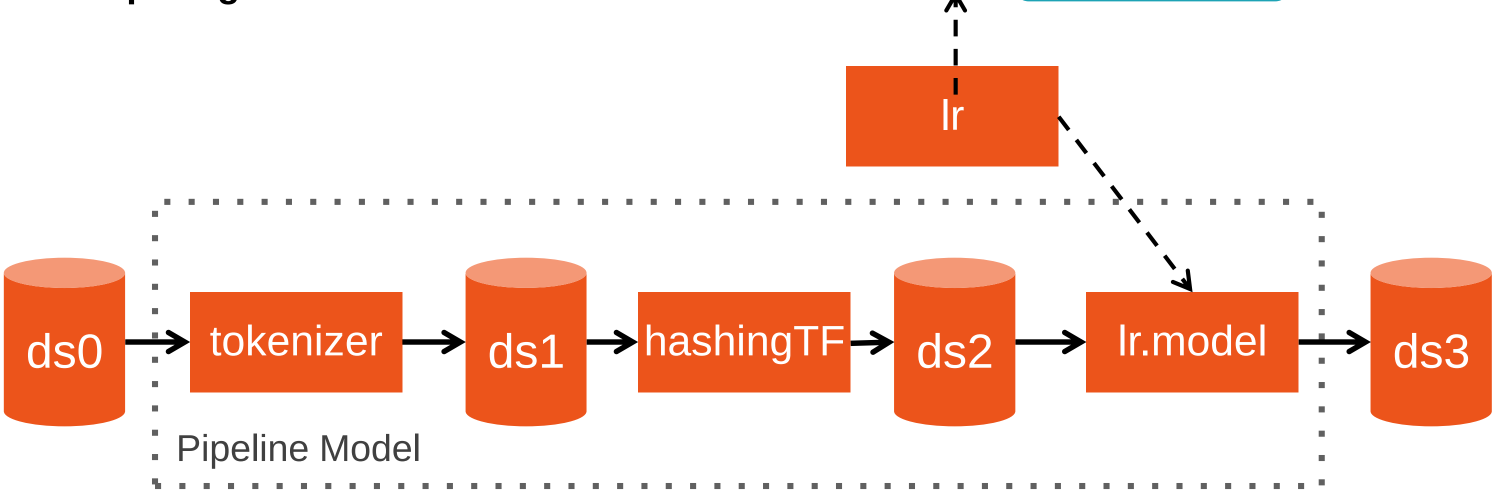
Spark MLlib Pipelines

```
tokenizer = Tokenizer(inputCol="text", outputCol="words”)
hashingTF = HashingTF(inputCol="words", outputCol="features”)
lr = LogisticRegression(maxIter=10, regParam=0.01)
pipeline = Pipeline(stages=[tokenizer, hashingTF, lr])

df = sqlCtx.load("/path/to/data")
model = pipeline.fit(df)
```



Apache Spark helps organize data science workflows.



SparklyR package in R, Rstudio

SparkR

Fast !
Statistics !
Scalable
Spark
+
Packages
Interactive Shell
Plots

SparkR reimplements **lapply** so that it works on RDDs, and implements other transformations on RDDs in R

http://files.meetup.com/3138542/SparkR-meetup.pdf

Overview by Shivaram Venkataraman & Zongheng Yang from AMPlab

Apache Spark integrates well with data scientists' favorite tool – R.

Key reasons for interest in Spark

High Performance

- In-memory architecture greatly reduces disk I/O
- Anywhere from **20-100x faster** for common tasks

Productive

- **Concise and expressive syntax**, especially compared to prior approaches (up to 5x less code)
- **Single programming model** across a range of use cases and steps in data lifecycle
- **Integrated with common programming languages** – Java, Python, Scala
- **New tools** continually reduce skill barrier for access (e.g. SQL for analysts)

Leverages existing investments

- Works well within **existing Hadoop ecosystem**

Improves continuously

- **Large and growing community** of contributors continuously improve full analytics stack and extend capabilities

Spark Philosophy by Databricks – Key reasons for Adopting Spark

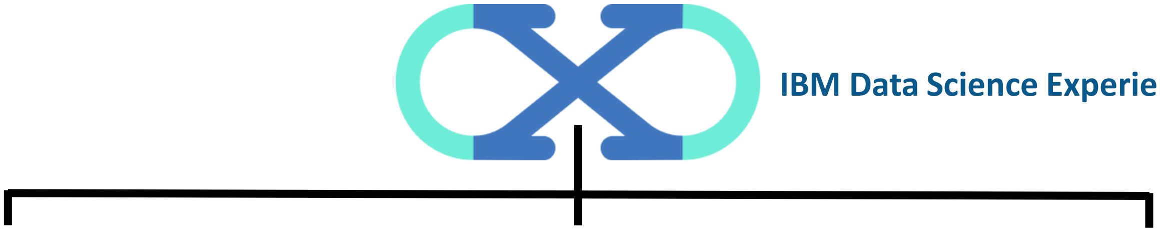
By Patrick Wendell's Spark 1.0 PPT

Make life easy and productive for data scientists

- Well documented, expressive API's
- Powerful domain specific libraries
- Easy integration with storage systems
- … and caching to avoid data movement
- Predictable releases, stable API's
-

An Integrated Approach – the direction to go

Core Attributes of the Data Scientist Experience



IBM Data Science Experience

Community

- Find tutorials and datasets
- Connect with data scientists
- Ask questions
- Read articles and papers
- Fork and share projects

Open Source

- Code in Scala/Python/R/SQL
- Jupyter and Zeppelin* Notebooks
- RStudio IDE and Shiny apps
- **Apache Spark**
- Your favorite libraries

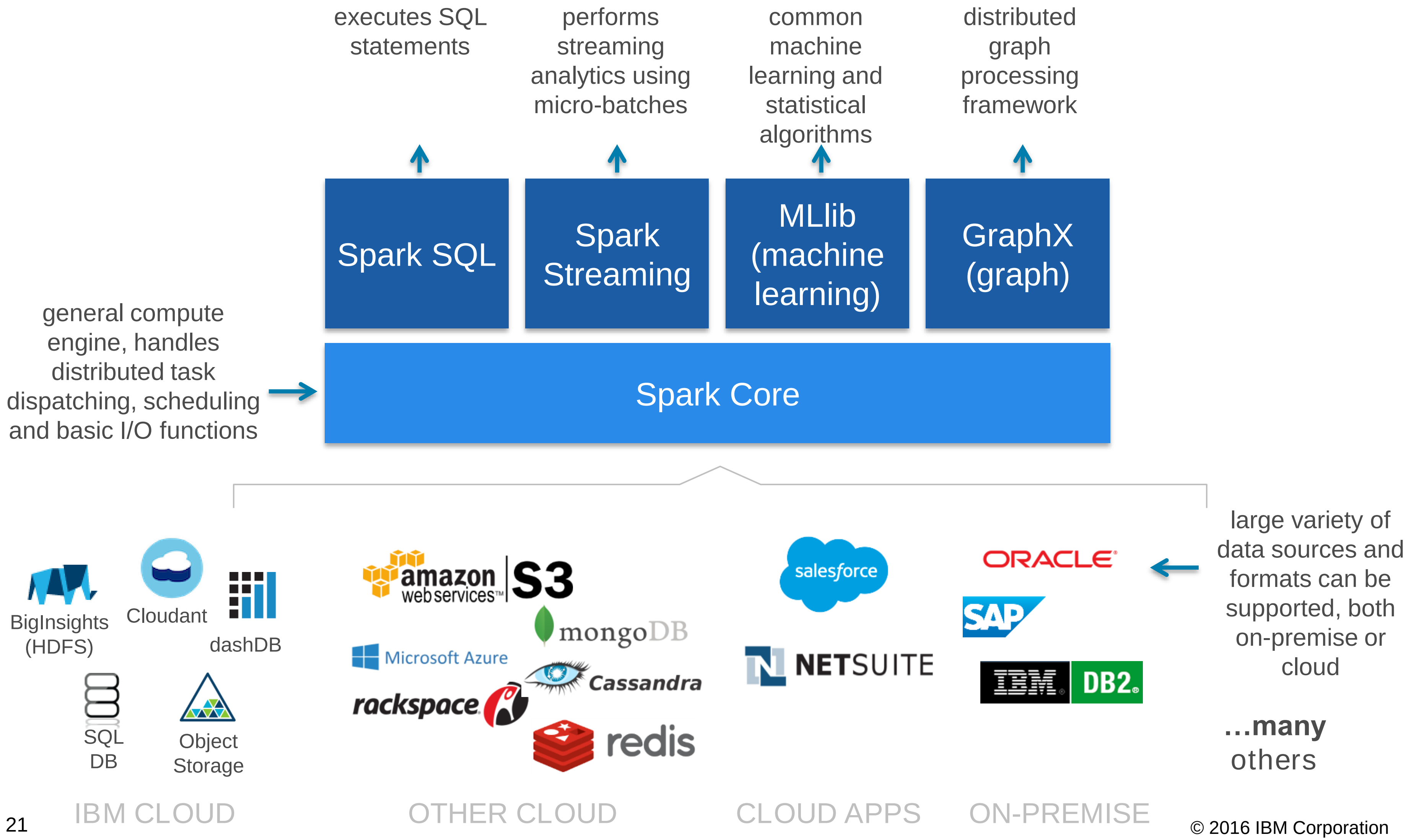
IBM Added Value

- Data Shaping/Pipeline UI *
- Auto-data preparation and modeling*
- Advanced Visualizations*
- Model management and deployment*
- Documented Model APIs*
- Spark as a Service

Powered by IBM **Next Generation Platform** in the Cloud

* DSX product roadmap items

From a Notebook you can use IBM Analytics for Apache Spark to blend multiple data types, sources, and workloads



executes SQL statements

performs streaming analytics using micro-batches

common machine learning and statistical algorithms

distributed graph processing framework

Spark SQL

Spark Streaming

MLlib (machine learning)

GraphX (graph)

general compute engine, handles distributed task dispatching, scheduling and basic I/O functions

Spark Core

BigInsights (HDFS)

Cloudant

dashDB

SQL DB

Object Storage

large variety of data sources and formats can be supported, both on-premise or cloud

…many others

IBM CLOUD

OTHER CLOUD

CLOUD APPS

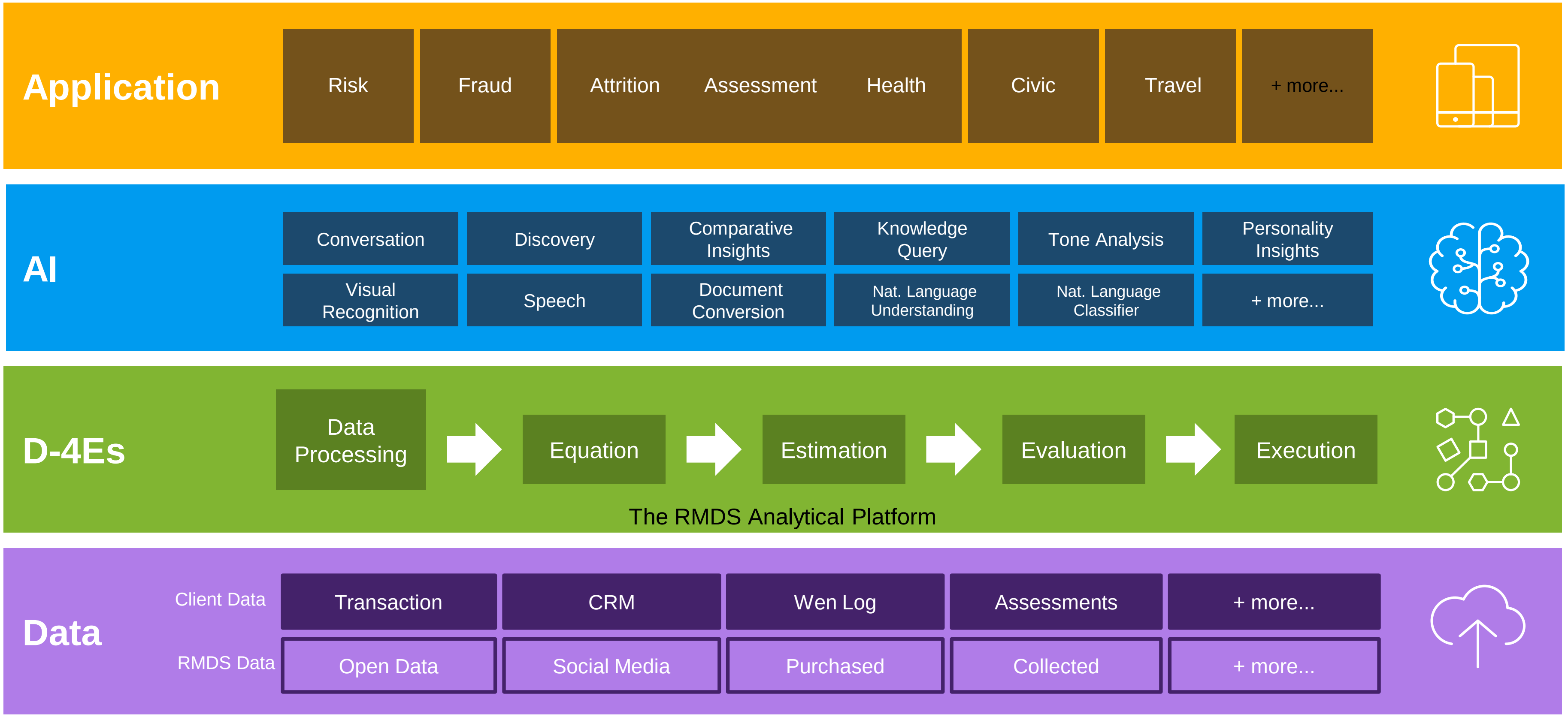
ON-PREMISE

A New Way to do Machine Learning Powered by Watson

IBM Watson Machine Learning

1. Machine Learning made Easy and Understandable
2. Full Machine Learning workflow as a service
3. Automation of the lifecycle
4. Train new Machine Learning Models with your own data: 27 data connectors and growing
5. APIs for developers to train and score Machine Learning models
6. Easily create apps powered by Machine Learning in your language of choice: Java, JavaScript, .Net, Swift, Ruby and more for the web or Android/iOS
7. Deploy in Batch/Streaming and Real-time
8. Generate billions of predictions in seconds
9. Scale the ML platform in 1-click
10. Collaborate with your team members and Learn from the Community

Data Science Platform for RMDS



Application

Risk | Fraud | Attrition | Assessment | Health | Civic | Travel | + more...

AI

Conversation | Discovery | Comparative Insights | Knowledge Query | Tone Analysis | Personality Insights

Visual Recognition | Speech | Document Conversion | Nat. Language Understanding | Nat. Language Classifier | + more...

D-4Es

Data Processing → Equation → Estimation → Evaluation → Execution

The RMDS Analytical Platform

Data

Client Data: Transaction | CRM | Wen Log | Assessments | + more...

RMDS Data: Open Data | Social Media | Purchased | Collected | + more...

рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
mersi
ko ora
barka
weialin
tack
spas
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
yinaka
спасибі
blagodaram
kiitos
dankie
nanni
nandri
dhanyavad
thank you
gracias
хвала
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gracie
hvala
mauruuru
köszönöm
dankon
aciu
akun
enkosi
djiere dieuf
tau
mochchakkeram
murakoze
дякую
mamnun
dziękuję
chnorakaloutioun
gratias ago
gracies
sulpay
taiku
go raibh maith agat
shukrane
sobodi
dėkuji
obrigado
sagolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
kam sah hamnida
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
merce
мерси
rahmat
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci