HIGHER SCHOOL OF ECONOMICS

NATIONAL RESEARCH
UNIVERSITY

School of Data Analysis and Artificial Intelligence Department of Computer Science

DATA SCIENCE FOR BUSINESS

Lecture 1. Introduction to Data Science

Moscow, April 10th, 2020.

lzhukov@hse.ru

COURSE TECHNICALITIES

Lectures: Friday 18.10 - 19.30 , 10 lectures ZOOM: https://zoom.us/j/7723819319
Seminars: Friday 19.40 - 21.00, 10 seminars ZOOM: https://zoom.us/j/636910206

Class Website: http://www.leonidzhukov.net/hse/2020/datascience
Seminar Wiki: http://wiki.cs.hse.ru/Data_Science_for_Business_2020

Telegram Group: https://t.me/joinchat/ENzQEhr-hra2WhEjxvgayw

Modeling software: RapidMiner https://rapidminer.com

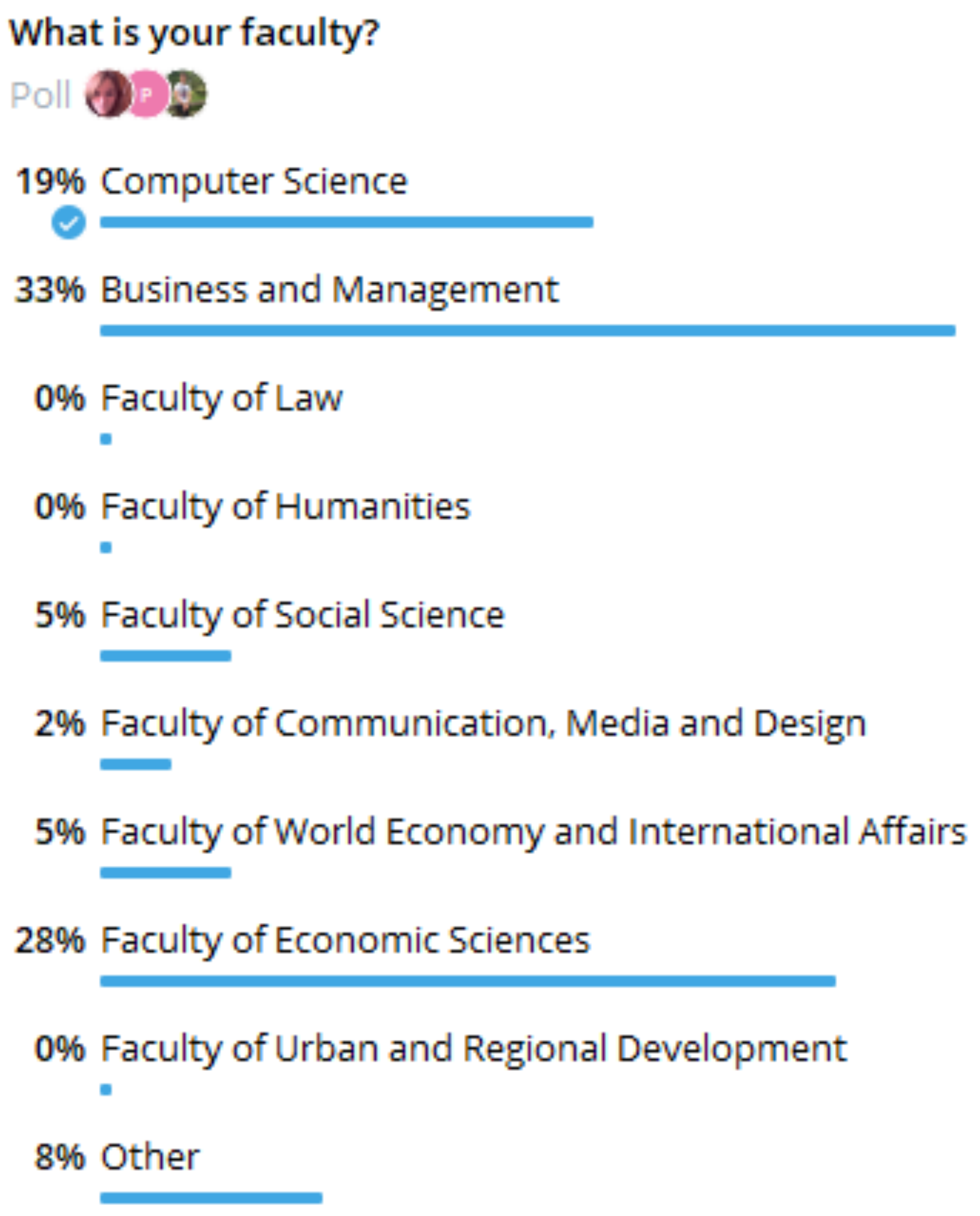
TEACHING TEAM

Prof. Leonid Zhukov
lzhukov@hse.ru

Anvar Kurmukov
kurmukovai@gmail.com

Ilya Makarov
iamakrov@hse.ru

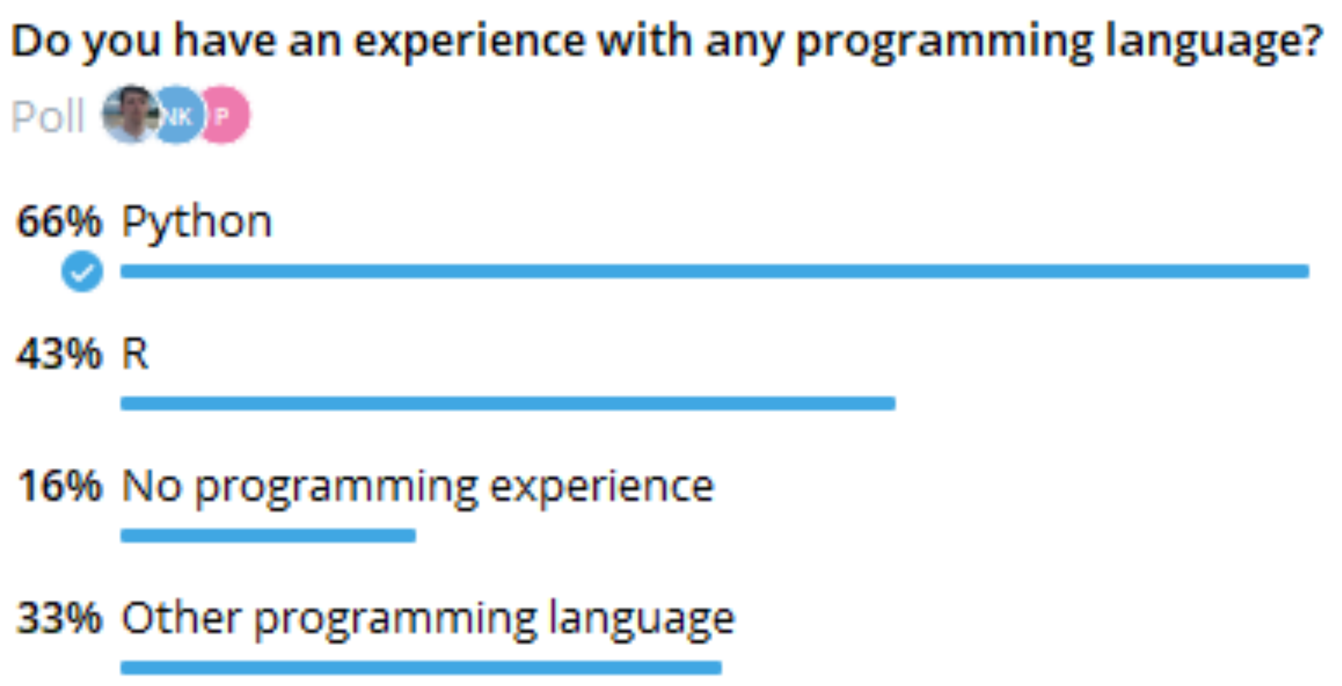
THE CLASS 2020



What is your faculty?

Poll

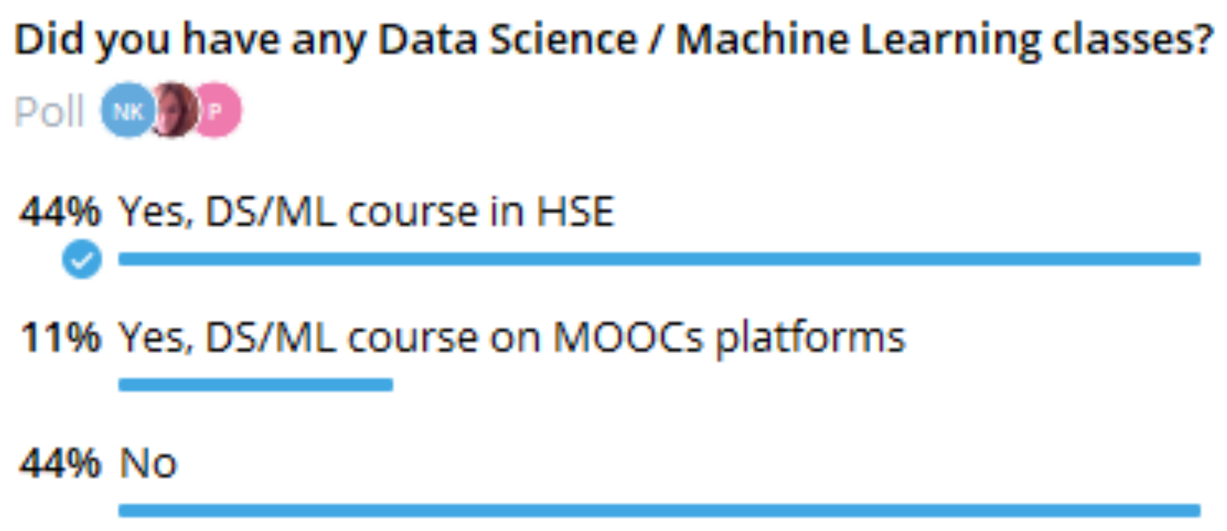
- 19% Computer Science ✓
- 33% Business and Management
- 0% Faculty of Law
- 0% Faculty of Humanities
- 5% Faculty of Social Science
- 2% Faculty of Communication, Media and Design
- 5% Faculty of World Economy and International Affairs
- 28% Faculty of Economic Sciences
- 0% Faculty of Urban and Regional Development
- 8% Other



Do you have an experience with any programming language?

Poll

- 66% Python ✓
- 43% R
- 16% No programming experience
- 33% Other programming language



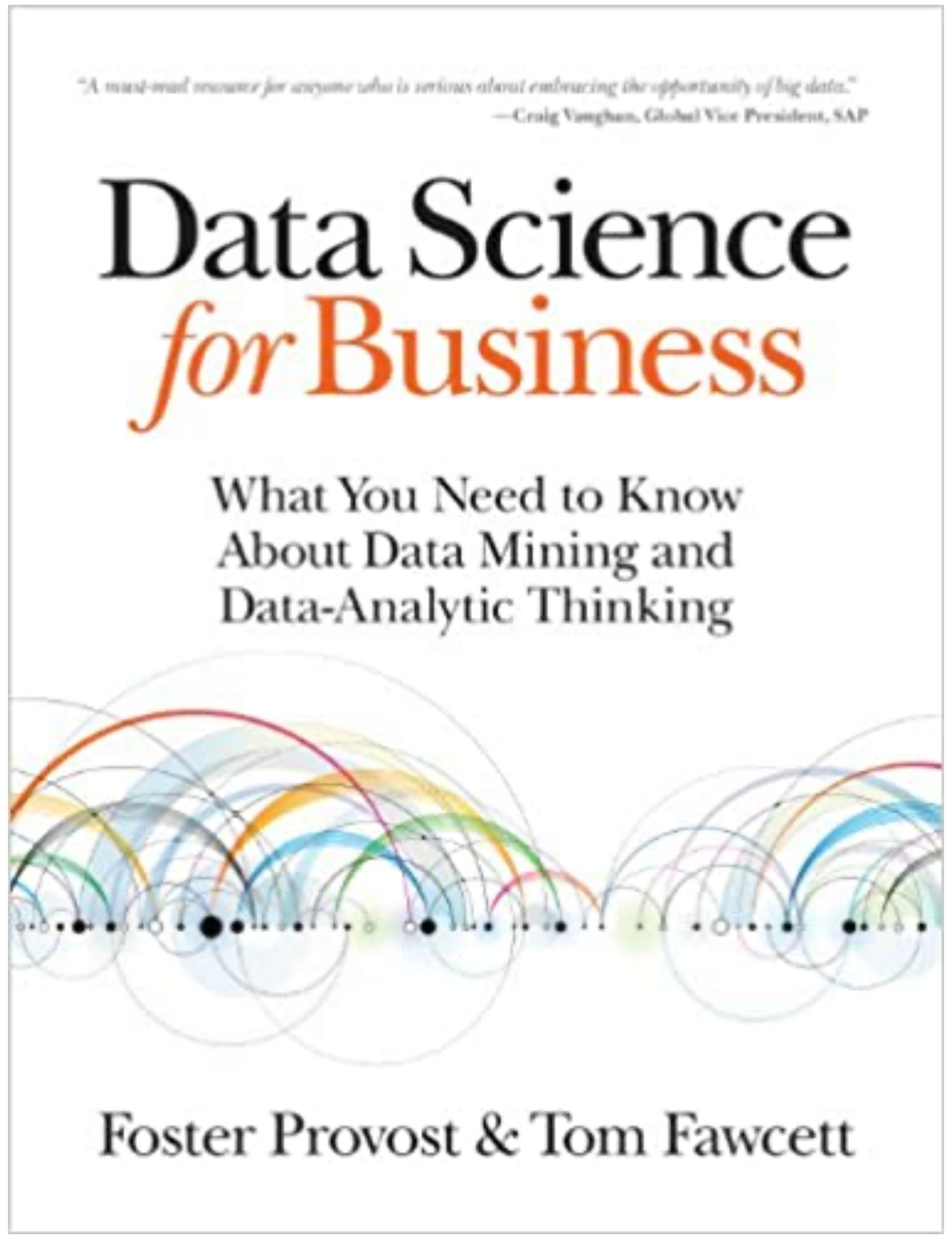
Did you have any Data Science / Machine Learning classes?

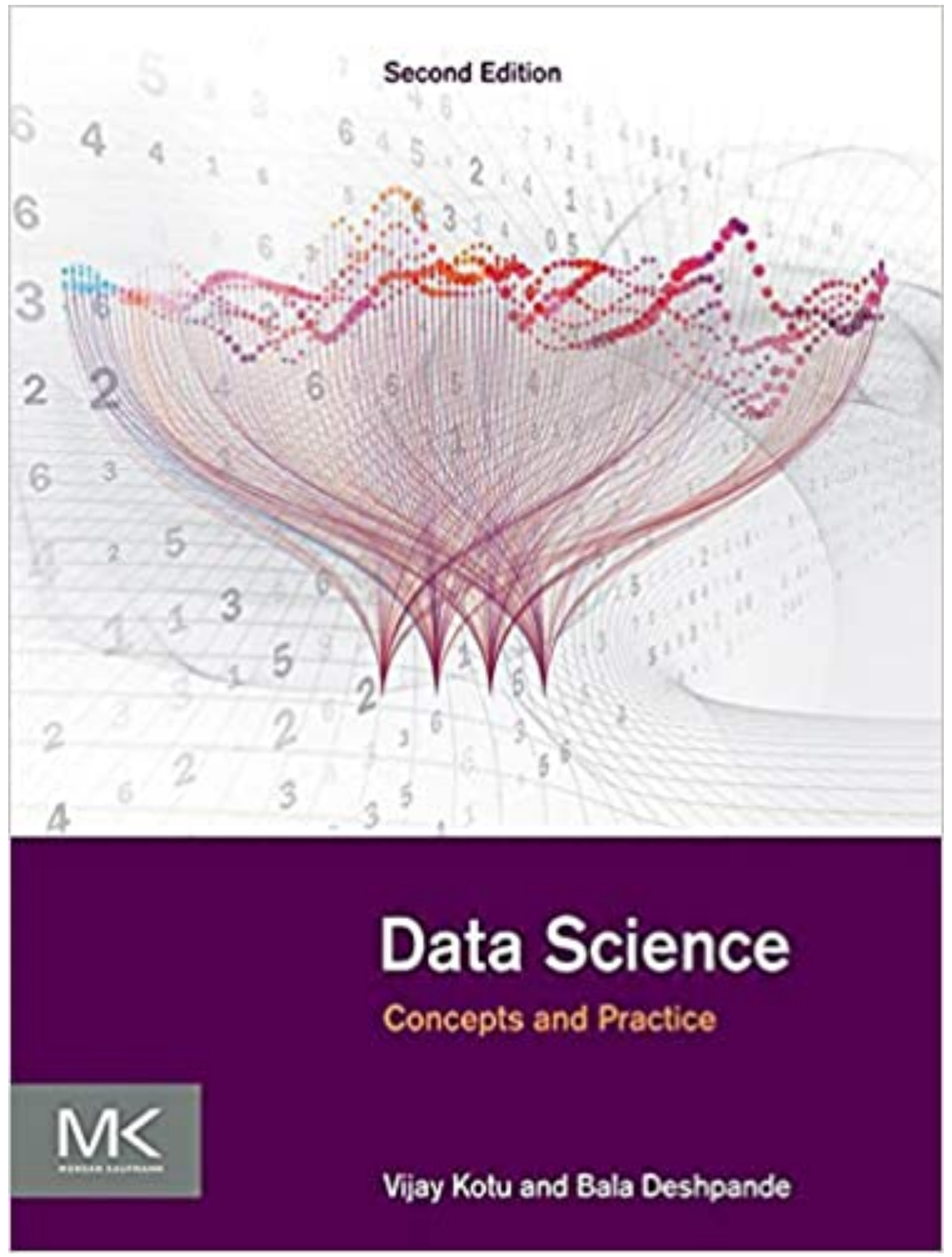
Poll

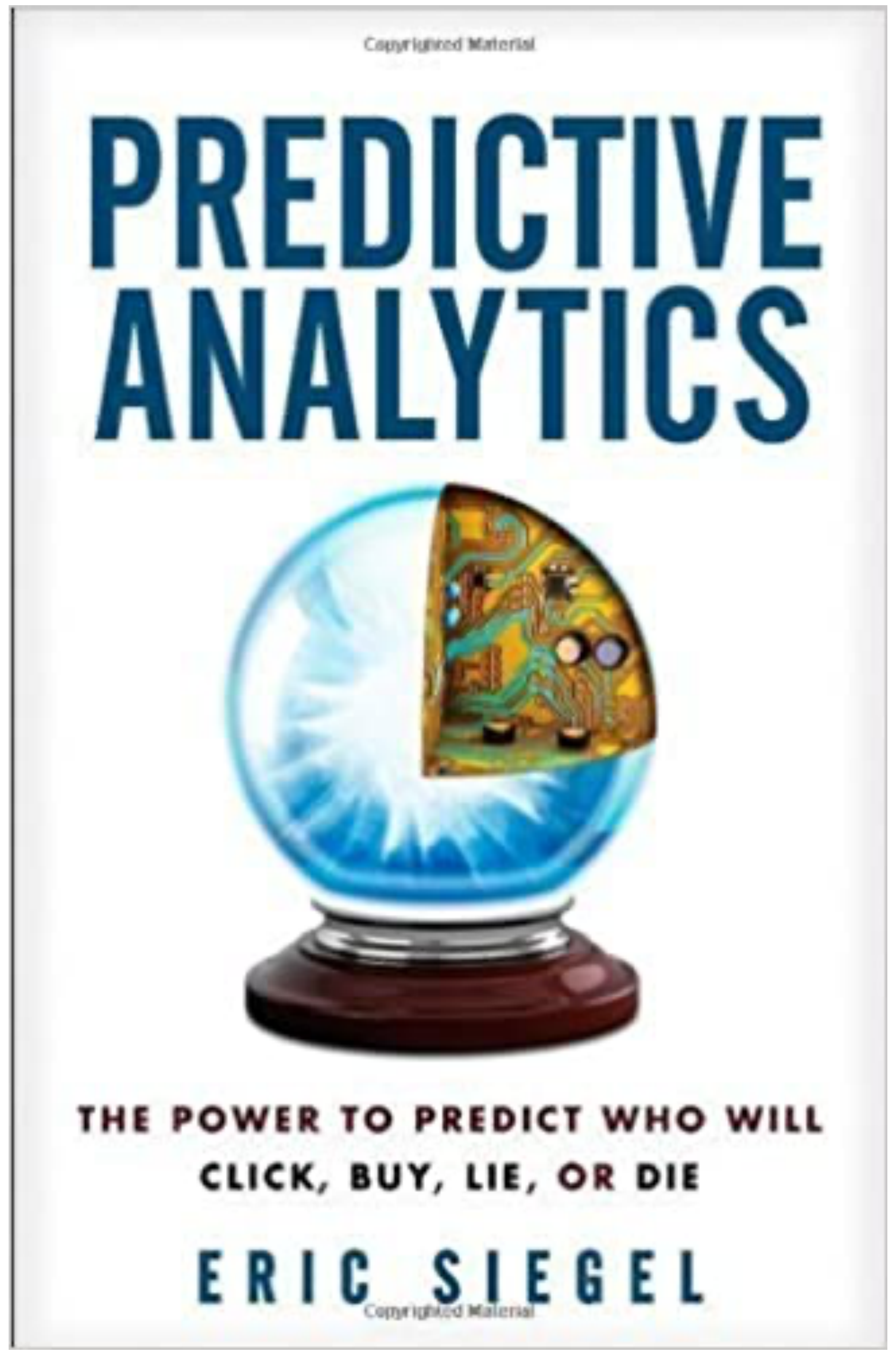
- 44% Yes, DS/ML course in HSE ✓
- 11% Yes, DS/ML course on MOOCs platforms
- 44% No

80-85 votes

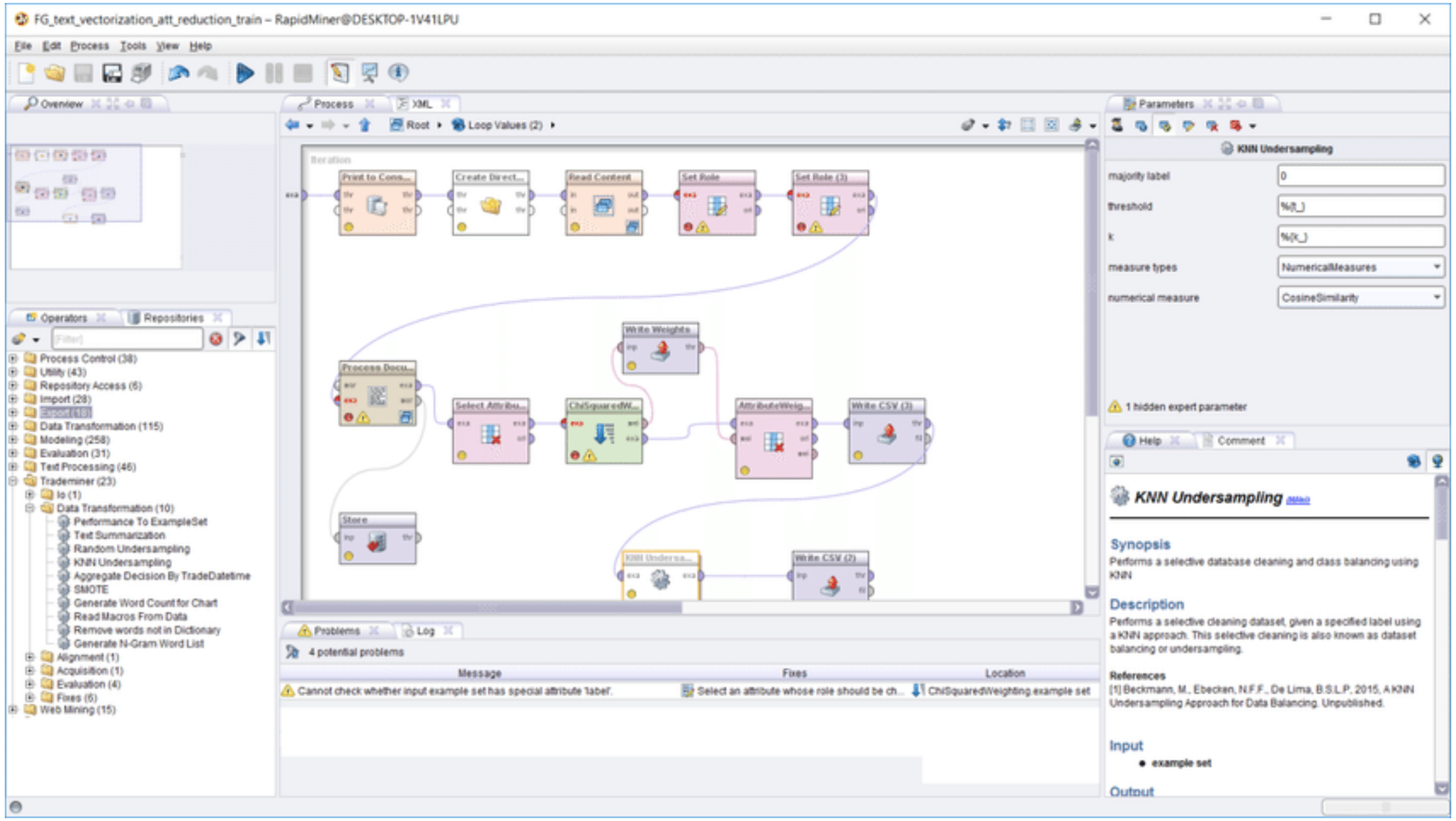
TEXTBOOKS FOR THE COURSE


"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP
Data Science for Business
What You Need to Know About Data Mining and Data-Analytic Thinking
Foster Provost & Tom Fawcett





RAPIDMINER MODELING SOFTWARE



KNN Undersampling

majority label	0
threshold	%{t_}
k	%{k_}
measure types	NumericalMeasures
numerical measure	CosineSimilarity

⚠ 1 hidden expert parameter

4 potential problems

Message	Fixes	Location
⚠ Cannot check whether input example set has special attribute 'label'.	Select an attribute whose role should be ch...	ChiSquaredWeighting.example set

KNN Undersampling (Mikio)

Synopsis

Performs a selective database cleaning and class balancing using KNN

Description

Performs a selective cleaning dataset, given a specified label using a KNN approach. This selective cleaning is also known as dataset balancing or undersampling.

References

[1] Beckmann, M., Ebecken, N.F.F., De Lima, B.S.L.P., 2015. A KNN Undersampling Approach for Data Balancing. Unpublished.

Input

- example set

Output

Do you have RapidMiner installed (with educational license)?

Poll NK P

52% I do ✓

48% I do not

https://rapidminer.com

COURSE SCHEDULE

Lecture topics

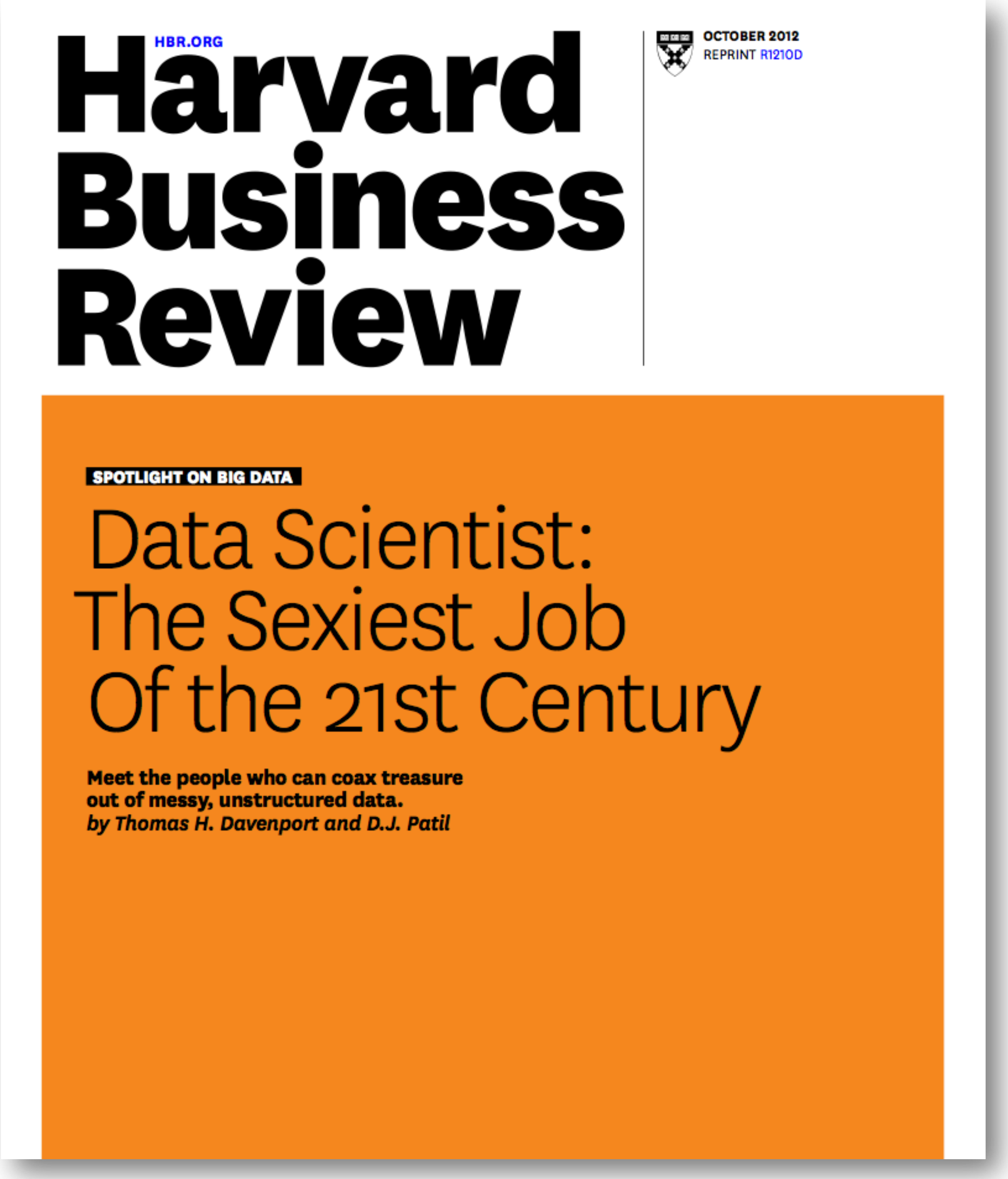
1. Introduction to data science.
2. Exploratory data analysis
3. Predictive analytics and machine learning
4. Case study 1: Retail pricing
5. Case study 2: Churn modeling
6. Case study 3: Customer segmentation
7. Case study 4: Personalization
8. Case study 5: Fraud detection
9. Case study 6. Demand forecasting
10. Impacting the business

Exam

Seminars - exercises

1. Data flow modeling and RapidMiner
2. Working with data, ETL process, data exploration
3. ML modeling pipeline
4. Regression
5. Classification
6. Clustering
7. Recommender systems
8. Anomaly detection
9. Time series forecasting
10. Problem solving

DATA SCIENCE



HBR.ORG

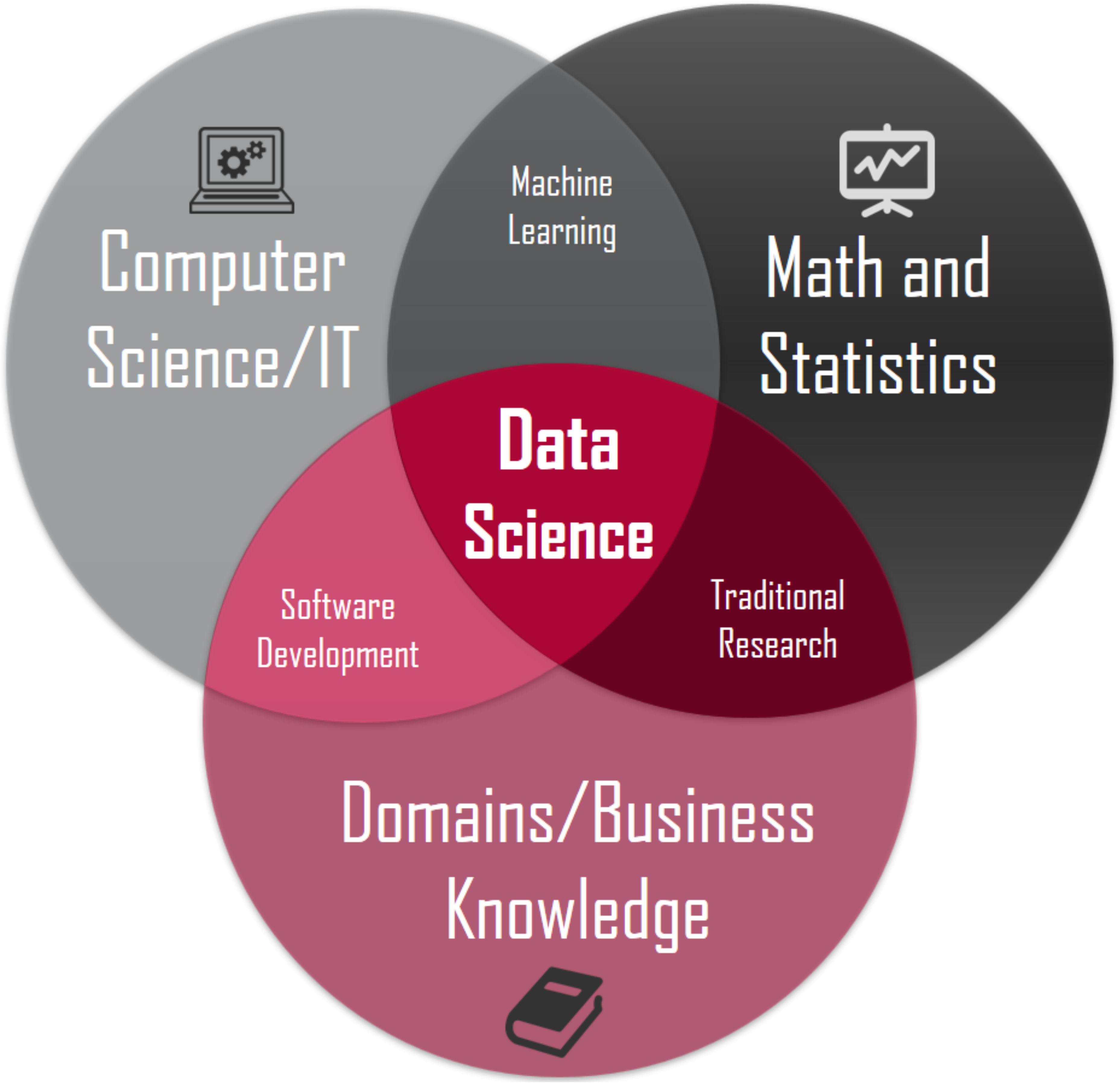
Harvard Business Review

OCTOBER 2012
REPRINT R1210D

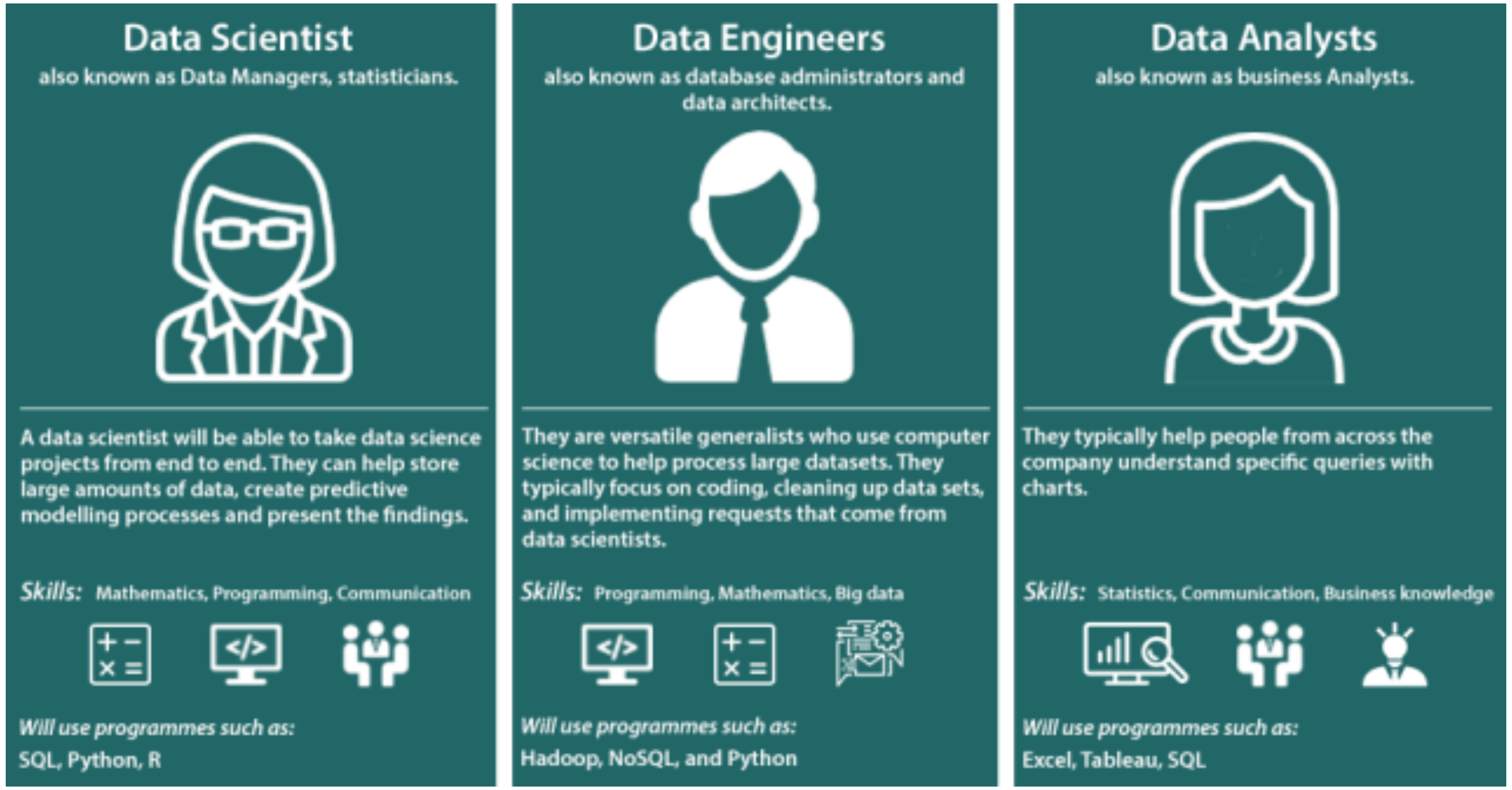
SPOTLIGHT ON BIG DATA

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



DATA SCIENTISTS



Data Scientist

also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:
SQL, Python, R

Data Engineers

also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:
Hadoop, NoSQL, and Python

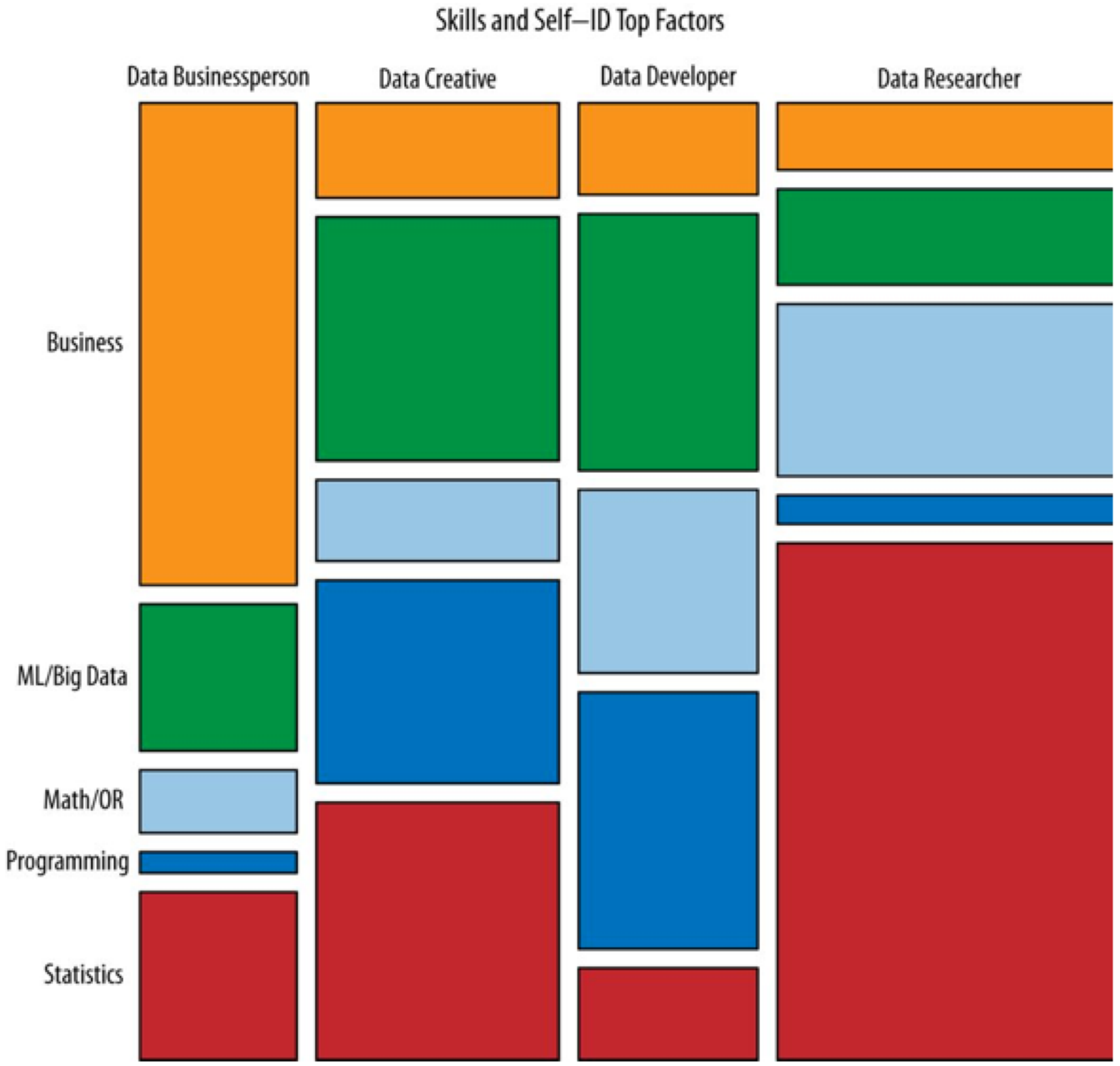
Data Analysts

also known as business Analysts.

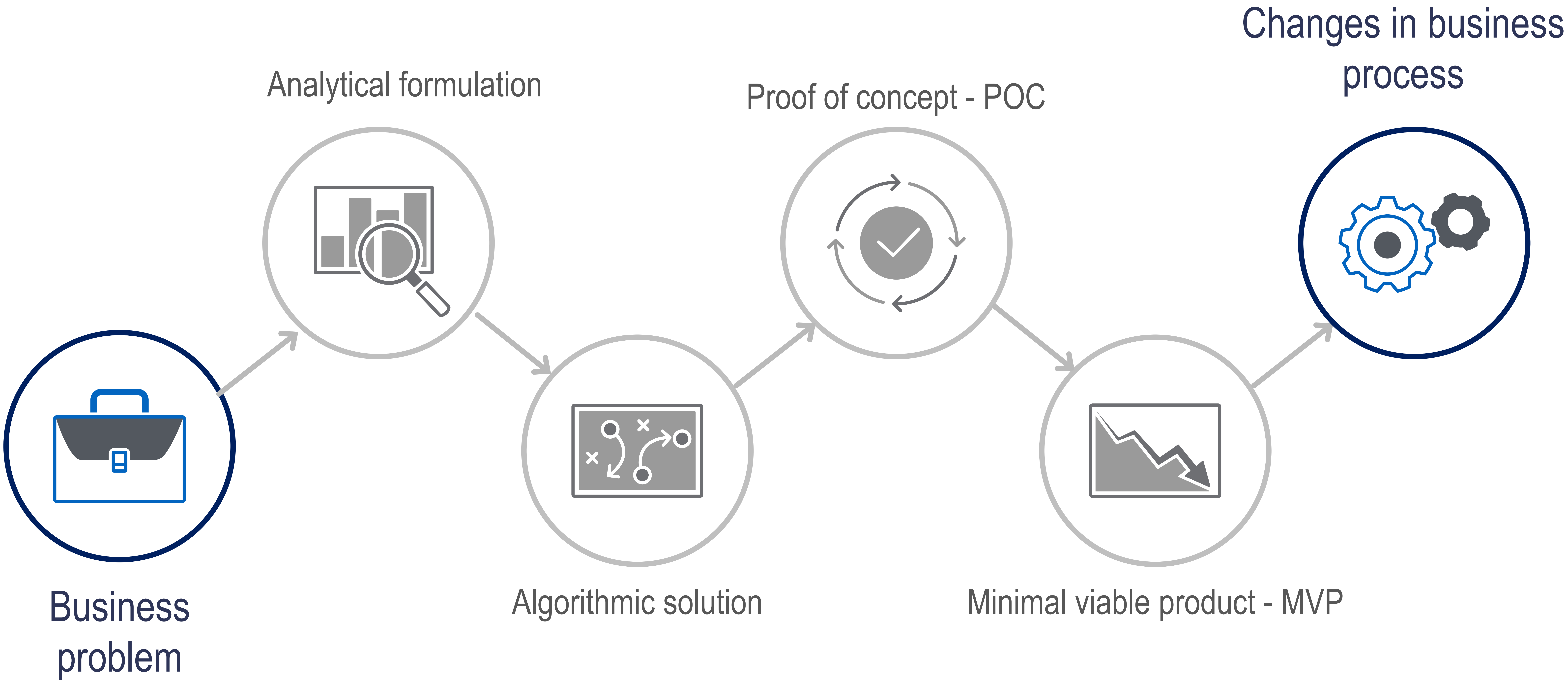
They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

Will use programmes such as:
Excel, Tableau, SQL



DATA SCIENCE BUSINESS PROCESS



Business problem

Analytical formulation

Algorithmic solution

Proof of concept - POC

Minimal viable product - MVP

Changes in business process

ANALYTICAL METHODS IN DATA SCIENCE

Predictive modeling
Machine learning
Data mining

Operations research
Optimization

Agent based modeling
Simulations

Geo analytics
Text analysis (NLP)
Computer vision

THREE MAIN REASONS TO USE ML IN BUSINESS

Detect

Predict

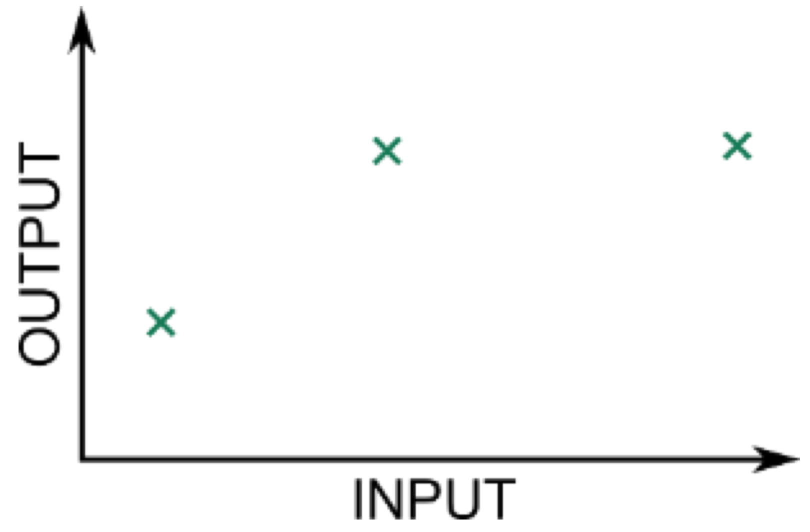
Explain

Optimise and improve

SIMPLE EXAMPLE

Statistical Analysis

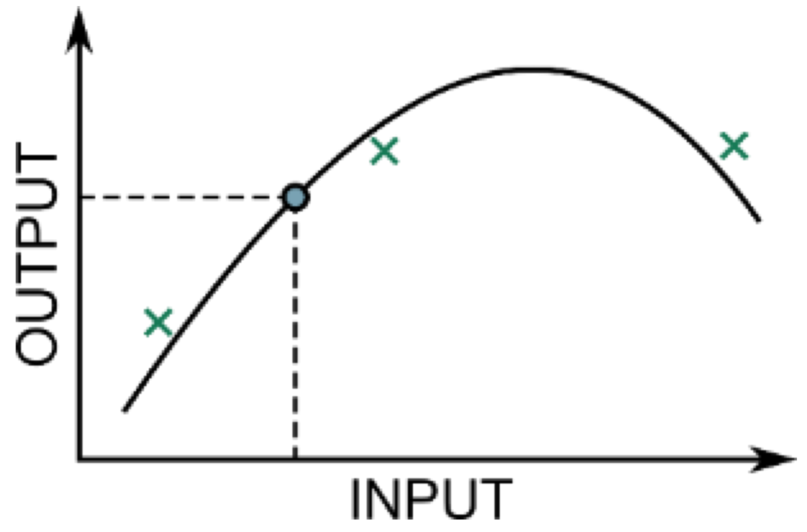
"Measure and understand"



Data exploration
Descriptive statistics

Predictive modeling

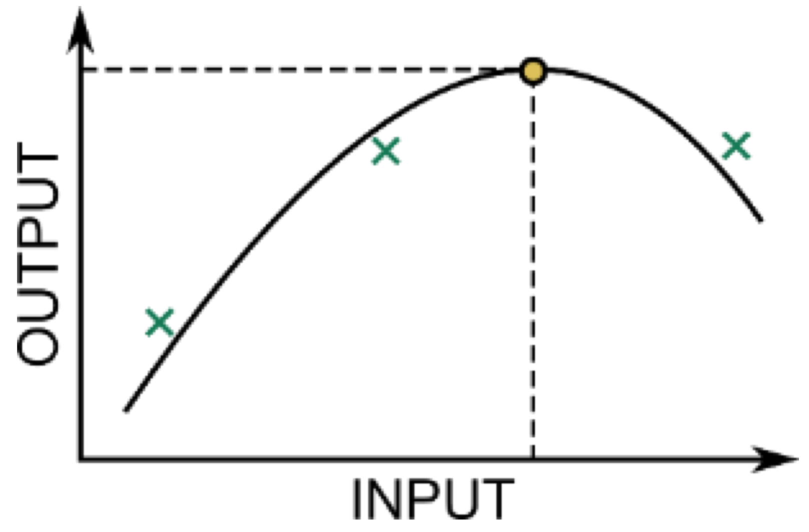
"ML predict outcomes"



Finding patterns
Predicting outcomes

Optimization

"Make optimal decisions"



Finding optimal values
Finding optimal parameters

BUSINESS USE CASES

Consumer Goods

- Demand forecast
- Marketing personalization
- Pricing and promo effectiveness
- Assortment optimization
- Cross sell and upsell

Telecoms

- Next best offer
- Churn and retention modeling
- Network optimization
- Infrastructure capacity and utilization

Banking/FI

- Credit risk assessment
- Fraud detection
- Claim management
- Churn and retention modeling
- Next best offer

Industrial Goods

- Manufacturing process optimization
- Predictive maintenance
- Demand and supply forecast
- Operations planning
- Energy efficiency

Energy

- Production optimization
- Predictive maintenance
- Logistics optimization
- Project risk management
- Robotics and automation

Enterprise

- Back office automation RPA
- Performance management
- Workforce planning
- Scenario simulations

THREE TYPES OF MACHINE LEARNING

Unsupervised Learning

Aim to discover structure:
no target variable known

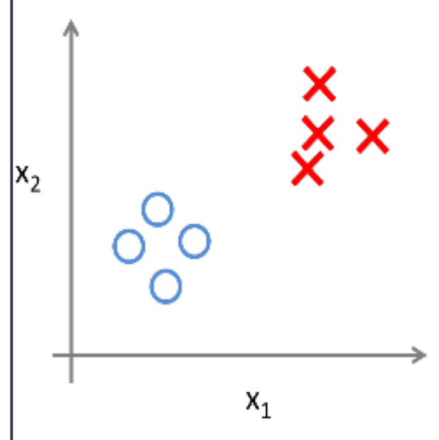
Supervised Learning

Aim to predict or model a
known target

Reinforcement Learning

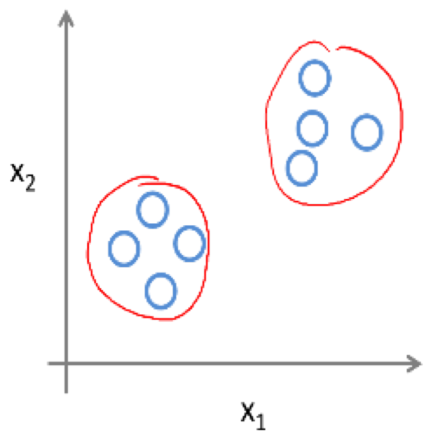
Optimise actions in a way that
maximises cumulative reward

Supervised Learning



Algorithms predict class of a new data point from a training set of previously correctly identified observations

Unsupervised Learning



Algorithms predict results without prior knowledge of the response

Types of algorithms

Algorithm examples

Classification

Given examples of classes, the model assigns new input data to classes

- **Decision trees,** k-nearest neighbors (kNN), Logistic regression
- **Random Forest,** Support Vector Machines (SVM),Gradient Boosted Decision trees (GBT)
- Neural networks + **Deep Learning**

Regression

Given several classes the model assigns input data to classes

- **Linear regression,** Elastic nets
- **Regression trees**

Ranking algorithms

Given ordered pairs examples the model ranks new data

Clustering

Divide the input data into groups with similar data points assigned to the same group

- **k-means, spectral**

Dimensionality reduction

Mapping input data in lower dimensional space

PCA

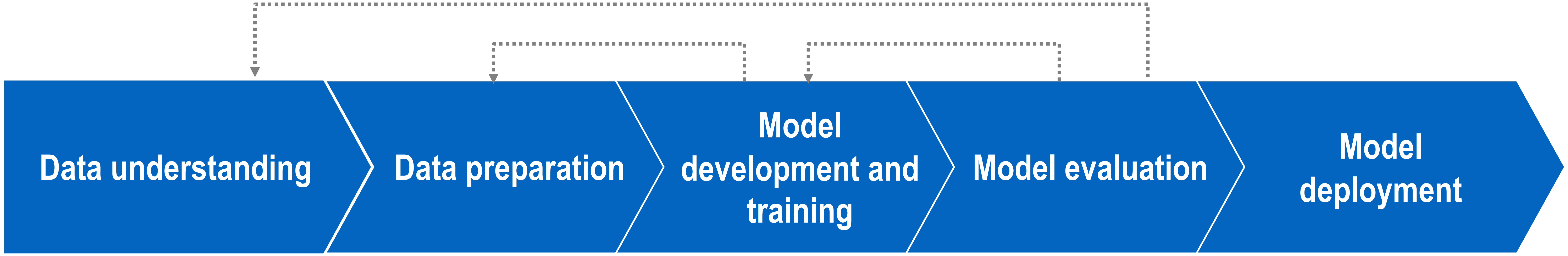
Density estimation

Estimates the probability distribution of values

Anomaly detection

Detecting outliers in data

MODELING PROCESS PIPELINE



Data understanding

Understanding the data at hand, checking the quality

Data preparation

ETL activities to get data into right format, cleaning, aggregation

Model development and training

Developing and training the model on given dataset

Model evaluation

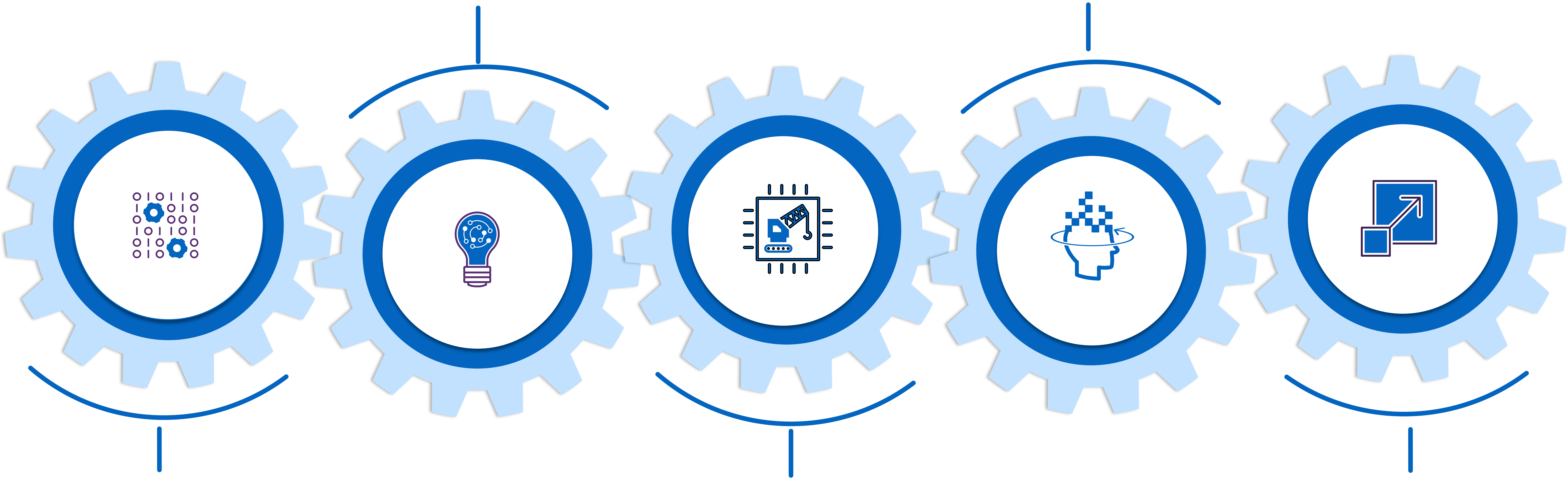
Validating model performance including business metrics on the holdout set

Model deployment

Running the pilot

rce: BCG

INTEGRATION OF ANALYTICS PROJECT INTO BUSINESS



Develop AI solution

Data science, machine learning algorithms and optimization

Embed in business process

Adjustment of ways of working and decision-making processes

Collect data

Historical data collection. Integration with data sources

Integrate in IT Ecosystem

Technology infrastructure enablement & data engineering

Deploy at Scale

Roll-out of solutions in additional geographies and/or business units

ANALYTICS PROJECTS SUCCESS FACTORS

Algorithms & Data

- Data analysis
- Algorithm development

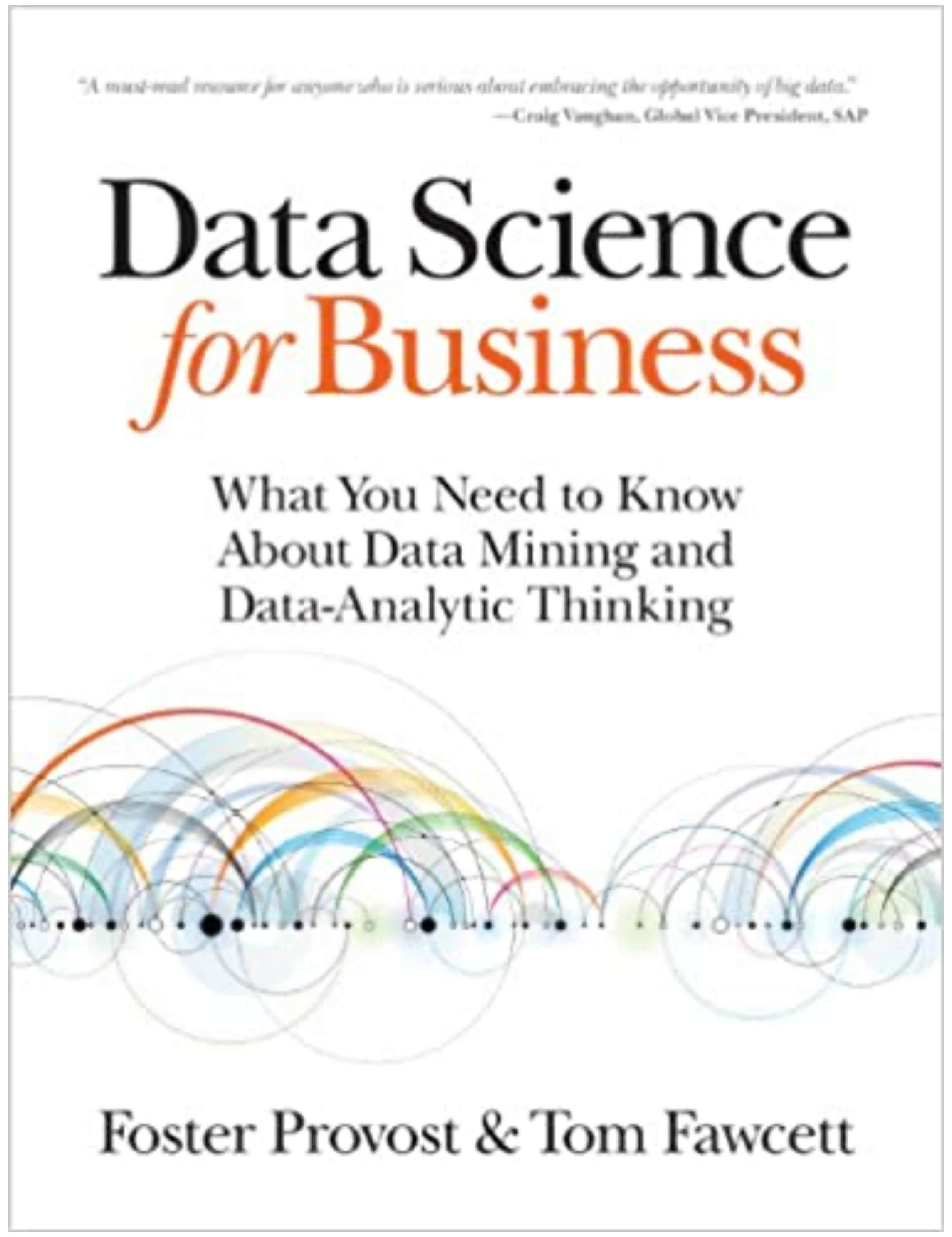
Technology/IT

- Algorithm industrialization
- Digital platforms development

Business transformation

- Business process redesign
- Enablement
- Change management

NEXT STEPS

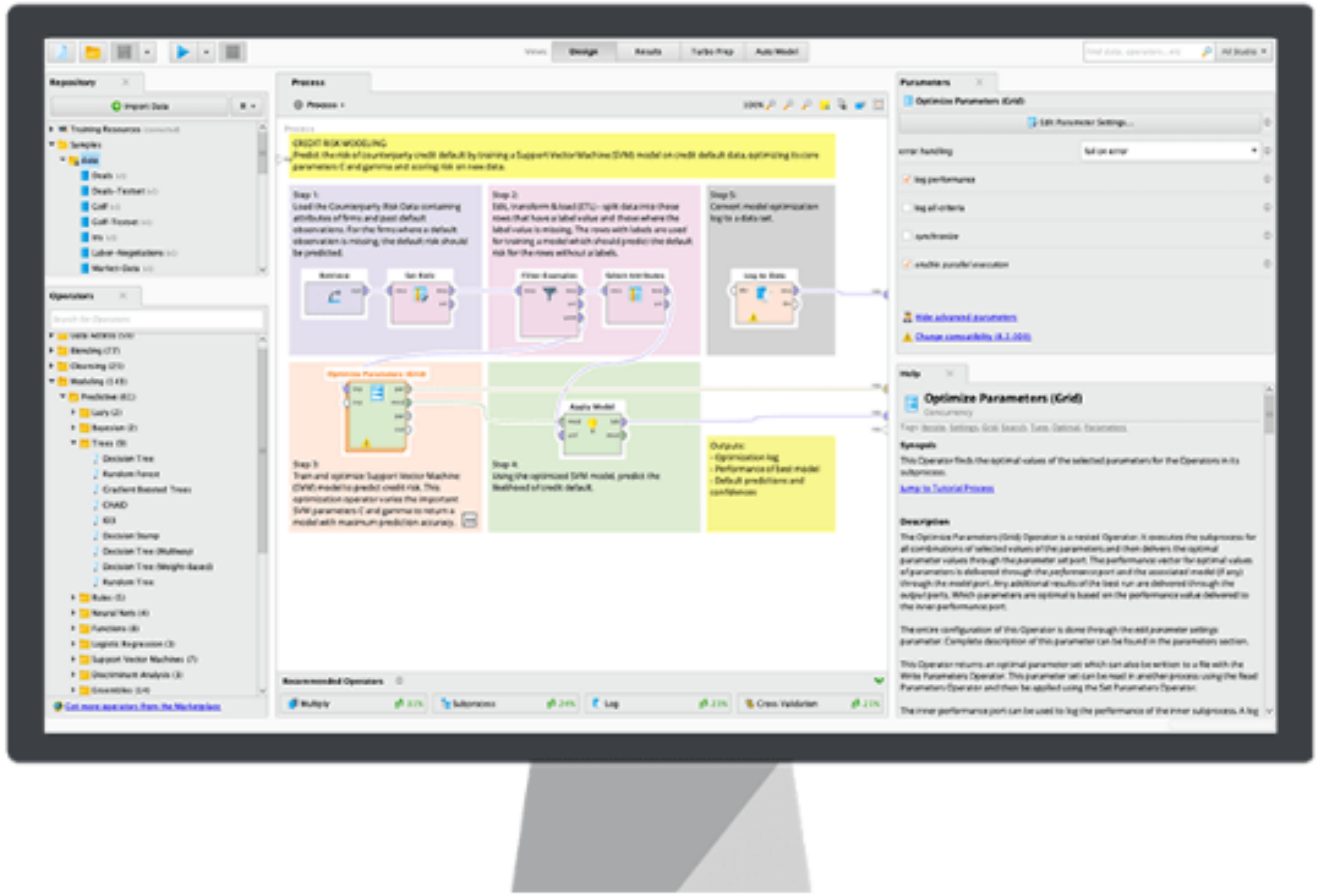


"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP

Data Science *for* Business

What You Need to Know About Data Mining and Data-Analytic Thinking

Foster Provost & Tom Fawcett



CREDIT RISK MODELING
Predict the risk of counterparty credit default by training a Support Vector Machine (SVM) model on credit default data, optimizing its core parameters C and gamma and scoring risk on new data.

Step 1:
Load the Counterparty Risk Data containing attributes of firms and past default observations. For the firms where a default observation is missing, the default risk should be predicted.

Step 2:
Edit, transform & load (ETL) - split data into those rows that have a label value and those where the label value is missing. The rows with labels are used for training a model which should predict the default risk for the rows without a labels.

Step 5:
Convert model optimization log to a data set.

Step 3:
Train and optimize Support Vector Machine (SVM) model to predict credit risk. This optimization operator varies the important SVM parameters C and gamma to return a model with maximum prediction accuracy.

Step 4:
Using the optimized SVM model, predict the likelihood of credit default.

Outputs:
- Optimization log
- Performance of best model
- Default predictions and confidences

HIGHER • SCHOOL • OF • ECONOMICS •

NATIONAL RESEARCH
UNIVERSITY