Current Trends and Issues in Data Mining

Alexey Malashonok

Minsk, 2017

Solving a business problem…

Example: Small retail bank on emerging market

Clients do no pay debts (PD rises)

Bank is not able to return defaults (RR falls)

Customers are leaving the bank (Attrition)

…with quantitative research

Solution

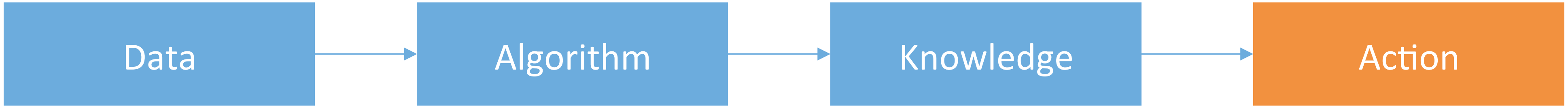
PD: adjust credit policy

RR: define effective collection strategy

Attrition: start marketing campaign

Common: predicting client's behavior

Data Mining in the Big Picture



Resources vs. model performance

Unclear, complex but relevant questions

Method usage is limited by data

Multiple parties involved

Cognitive errors and emotional biases

Data Mining is not…

Academic research

Supreme model accuracy

Data warehouse

Rocket science

Business reporting

Ethics matters

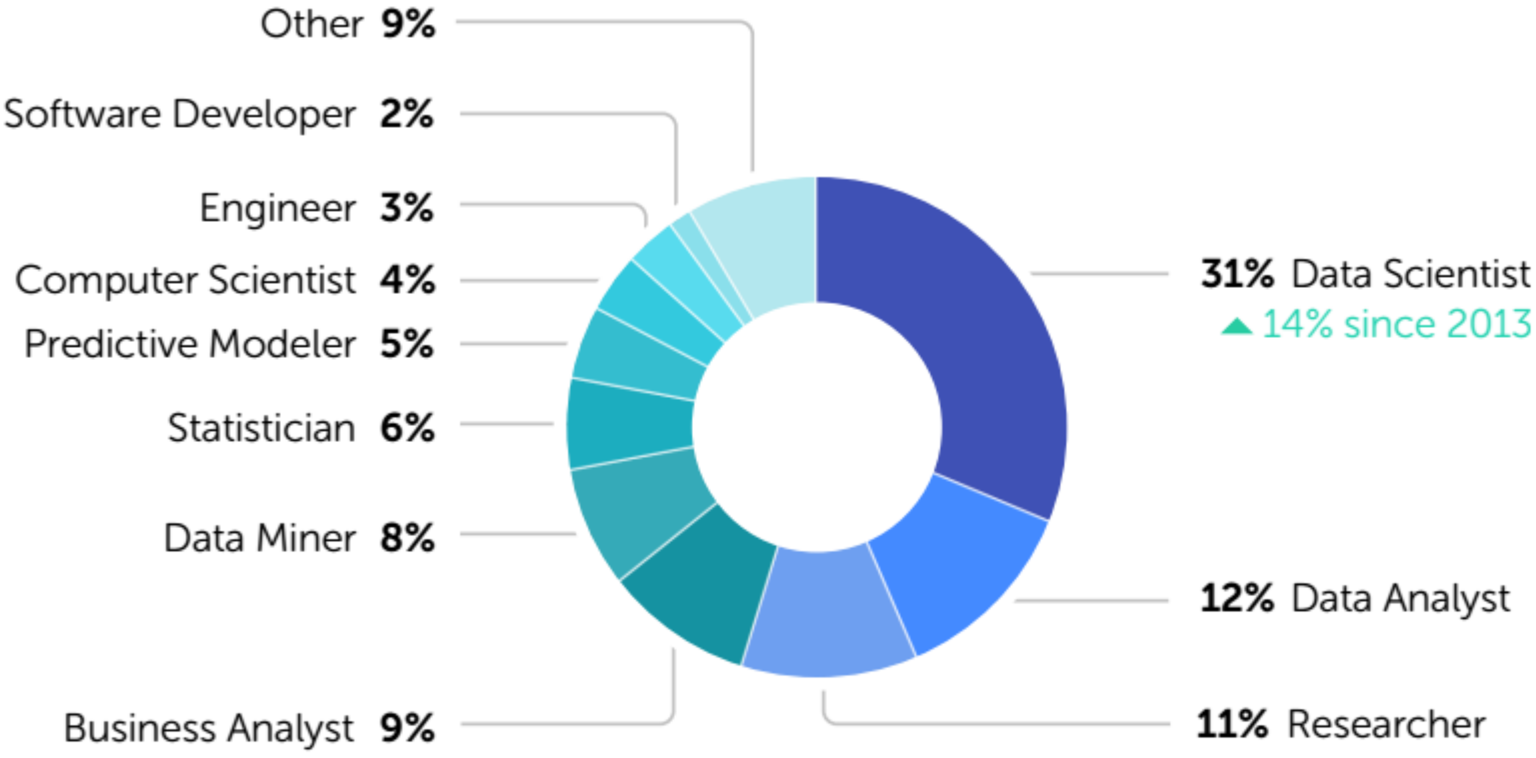
Confidentiality

Loyalty

Prudence and care

Does you company have code of professional standards?

Who is Data Miner or Scientist



Source: Rexer Analytics (2015)

Top 5 Mining Goals

Know your client

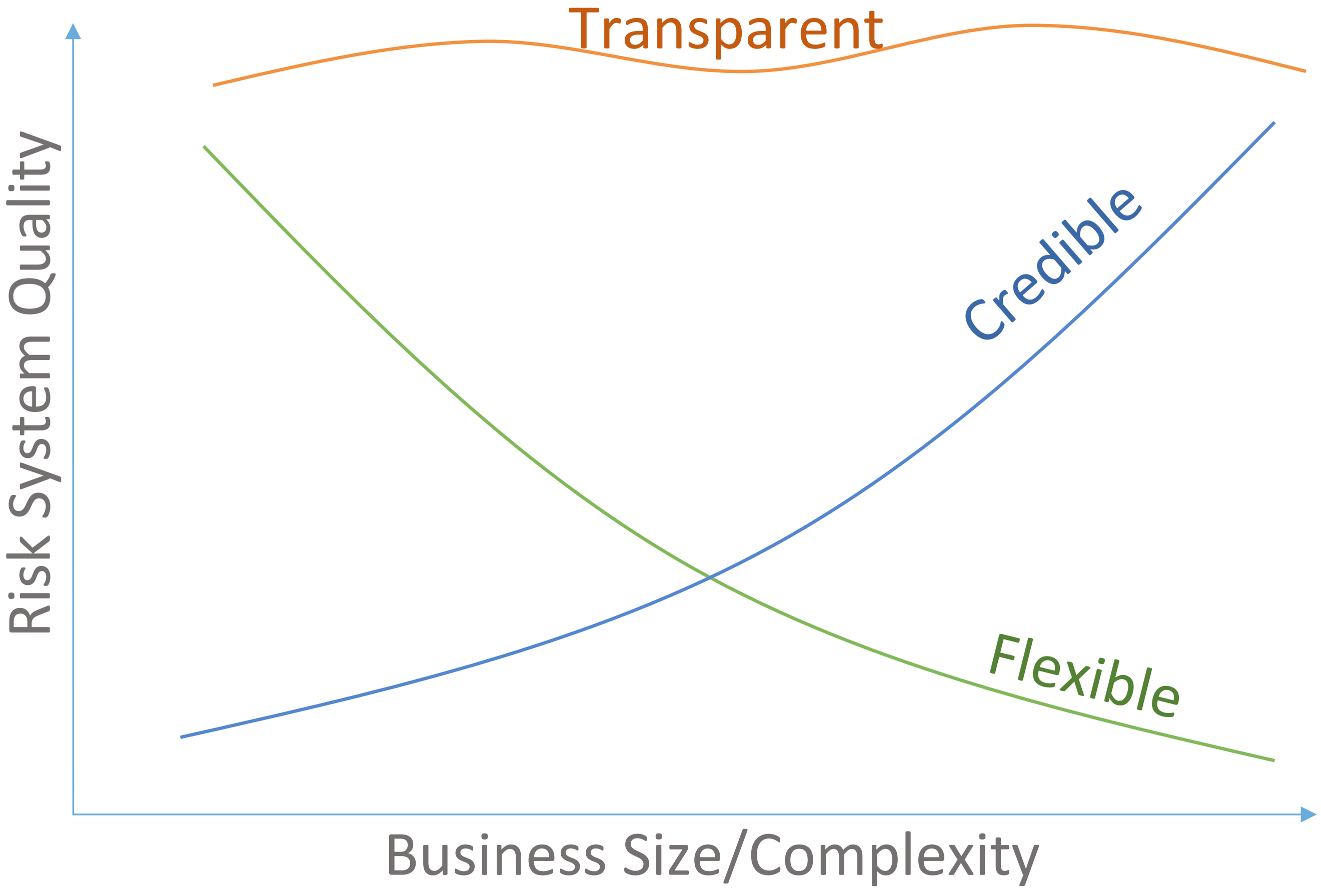
Market Research and Marketing

Sales Forecasting

Risk Management and Scoring

Manufacturing

Quality Tradeoffs in Data Mining

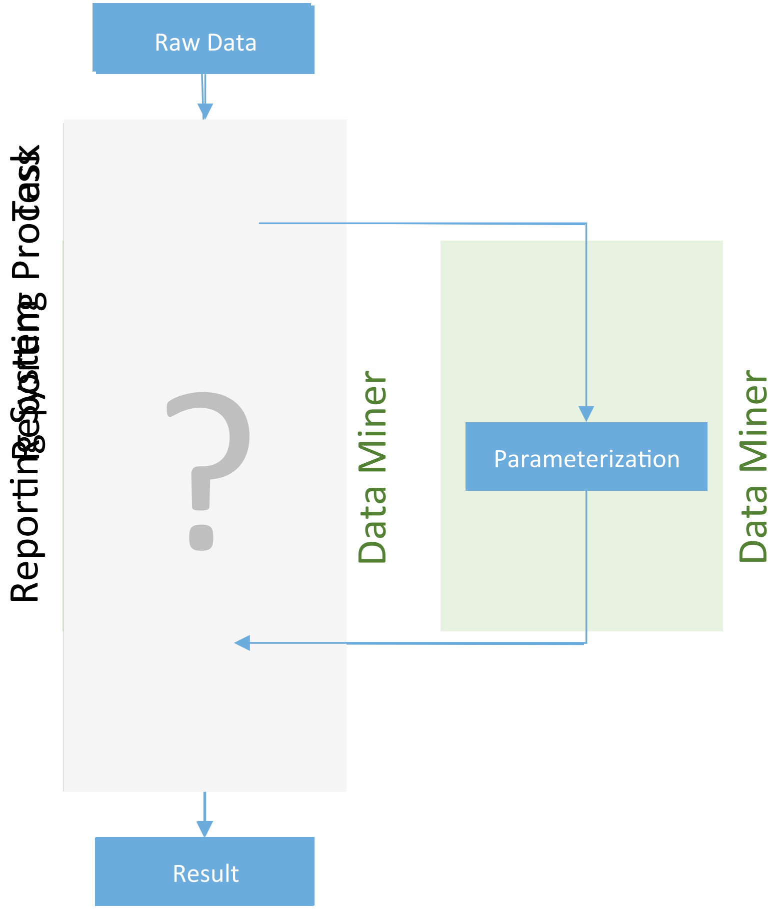


Data Mining Evolution

I. Task

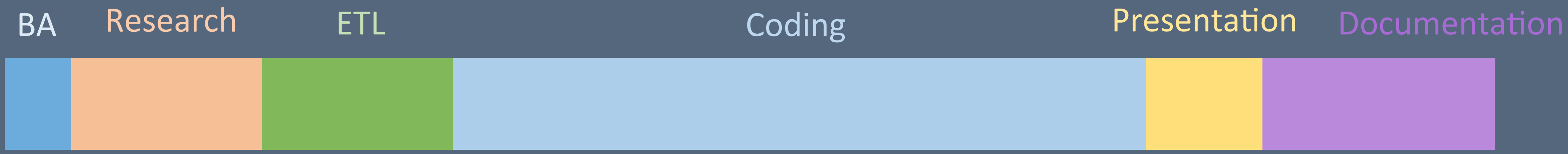
II. Process

III. System



Which stage is your company at?

Data Mining is not just Coding



Diligent research

Prudent presentation and communication

Care about interfaces and documentation

Big Dirty Data

Database complexity

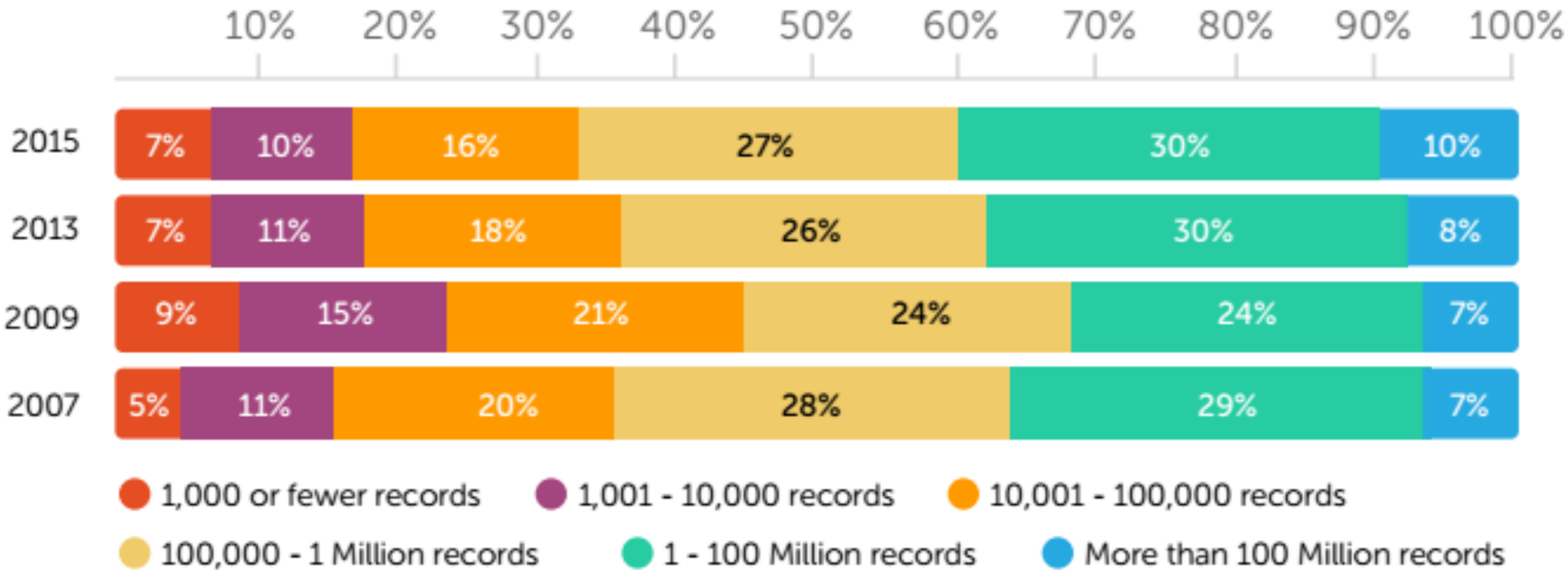
Need for Visualization

User Interfaces

Storage

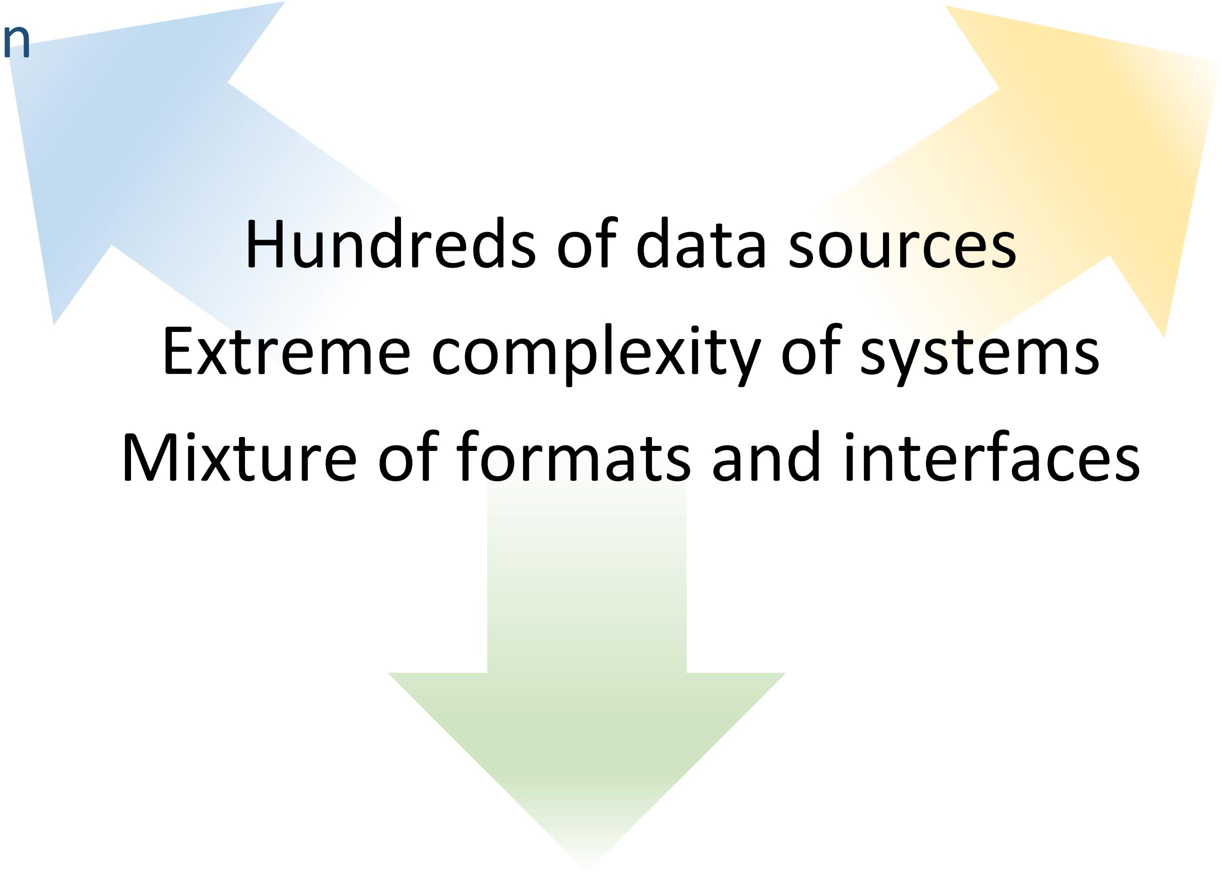
Searching for Data

Scales



Source: Rexer Analytics (2015)

Integrity Solutions



Industry solution

Customized solution

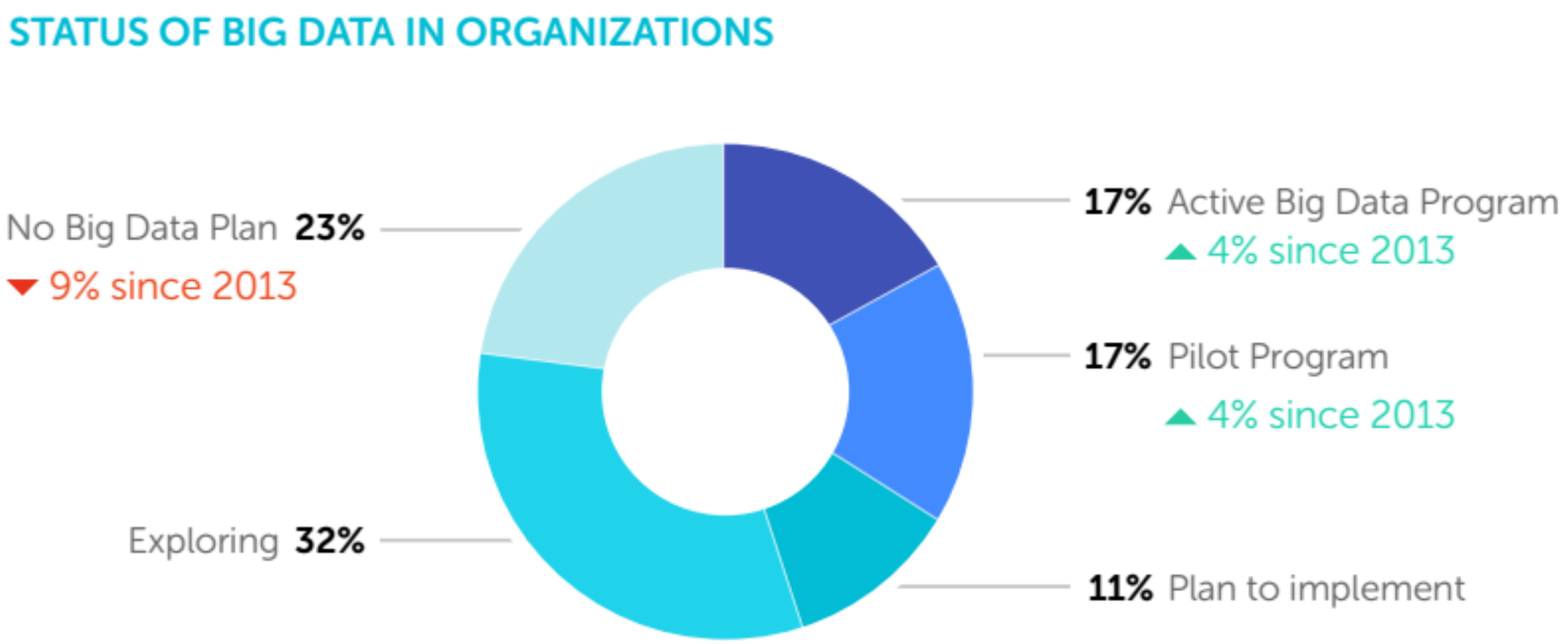
Hundreds of data sources

Extreme complexity of systems

Mixture of formats and interfaces

Own code

Is there any Progress?

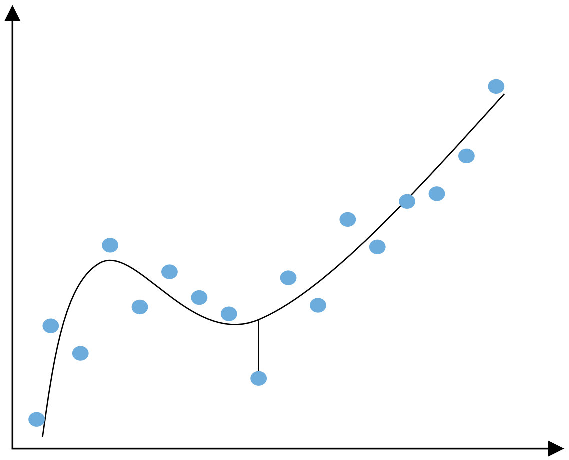


Source: Rexer Analytics (2015)

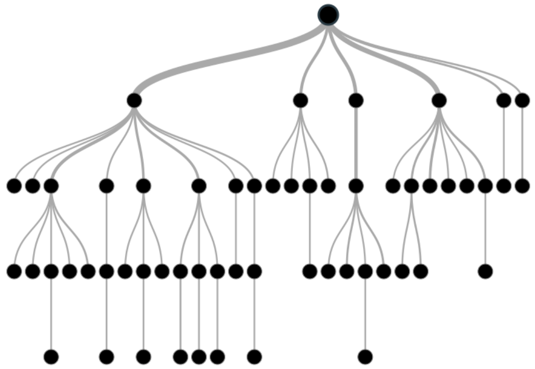
Primary Algorithms

More than one half of data miners use

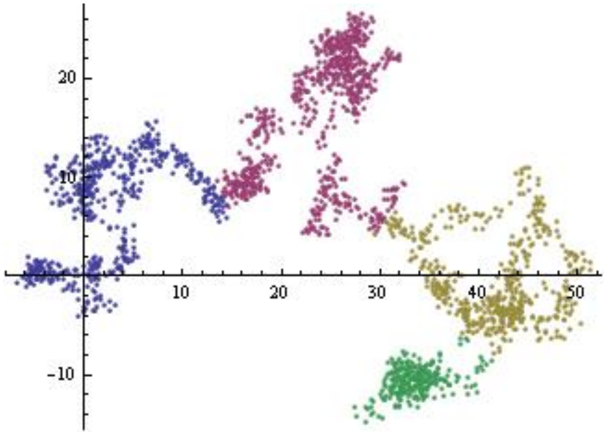
Regression



Decision Trees



Cluster Analysis



Communication: from Miner to Management

Suitable

Massage should suit the audience. But: maintain records!

Accurate

Use education and professional standards.

Complete

All relevant information included. Mind overconfidence bias!

How do you ensure that research is complete?

Communication: from Business to data Miner

Clear questions

Internal control

Use detailed plan

Do not use model performance to specify research target…

Seriously. Don't do that.

Case study: predicting provisions

A bank with USD 2 bln loan portfolio:

“We need a model which can predict monthly allocation for Loan Loss Provisions with +/-10k error”.

Historically monthly LLP was between 10k and 190k with standard deviation of 45k.

What is *expected* performance of the model?

Solution: Apply rule of thumb (2 sigma for 95% interval) to compute coefficient of determination

$$R^2 = 1 - RSS/TSS = 1 - (10/2)^2/(45)^2 = 0.98$$

Case study: predicting provisions (cont.)

1. Define target variable as Expected Loss = PD*EAD
2. Other studies say: factors that explain PD and EAD are different.
3. Test this hypothesis and estimate 2-stage model: Logistic regression for PD, and Lineal Regression for Exposure.
4. Literature provides industry averages of performance for similar models:
   - Logistic regression GINI 0.65-0.75
   - Exposure model R2 around 0.8

How do we resolve the issue with the manager?

What is Important for Programming Tool

Tool satisfaction

Interfacing

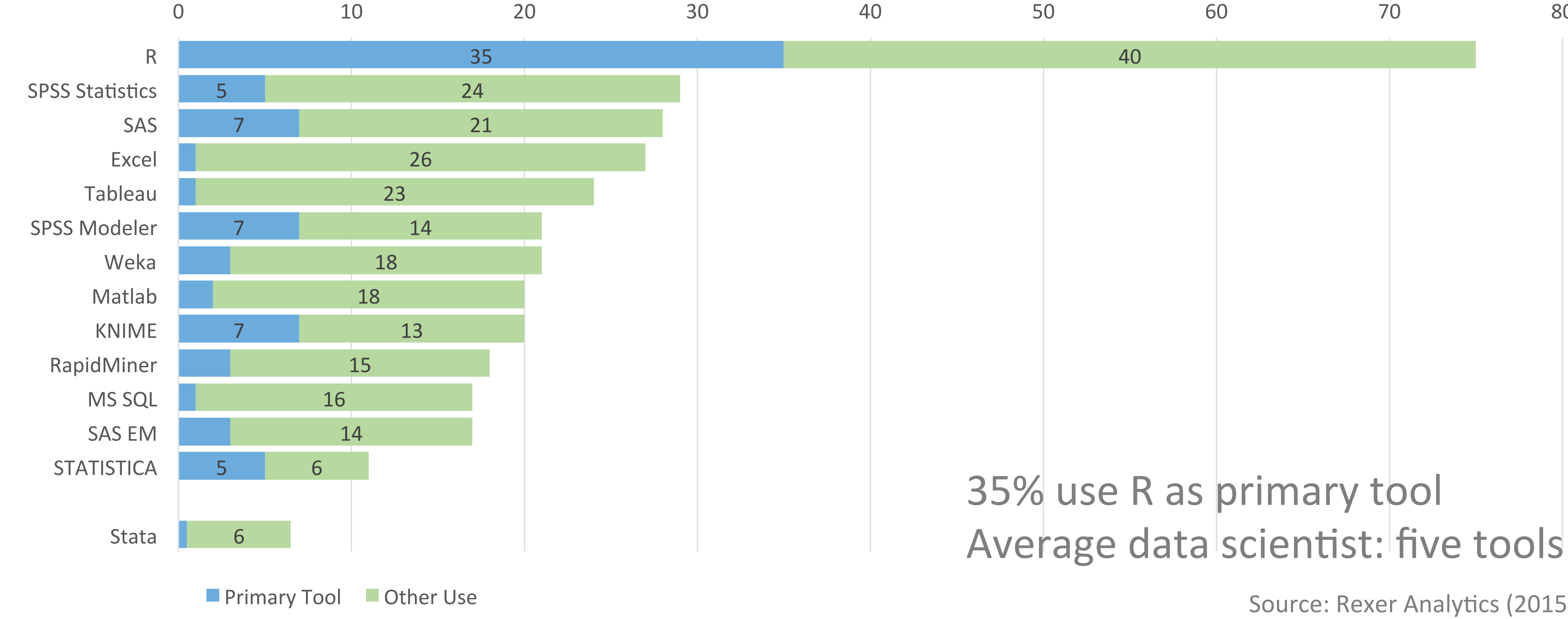
Deployment

Learning Curve

Stability and performance

The Rise of R

Language/platform usage (%, top 10)



35% use R as primary tool
Average data scientist: five tools

Source: Rexer Analytics (2015

R pros and cons

+

Zero Cost*

Writing own code

Automation

Visualization

Variety of algorithms

−

No code protection

Open Source

Not easy to use

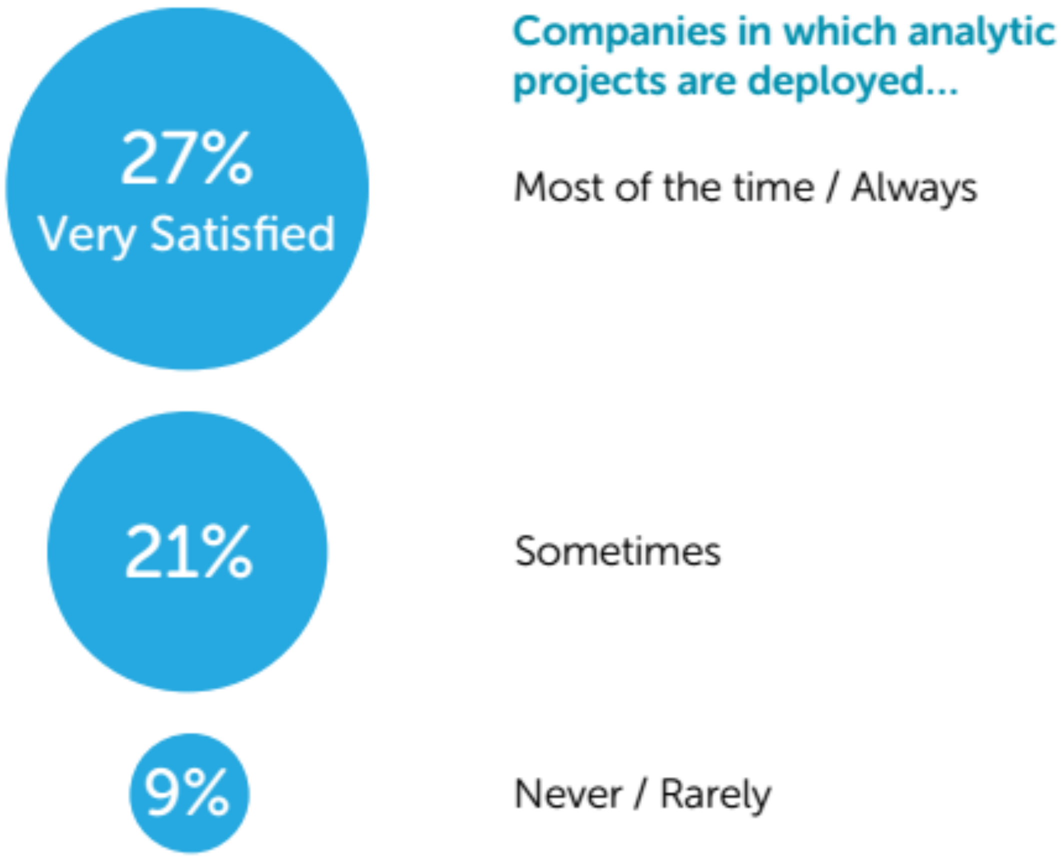
Speed and stability

Model Deployment

Coming back to IT and Infrastructure problem

Concentrate algorithms into one platform

Automate the process



Source: Rexer Analytics (2015)

Conclusion and Trends In Data Mining

High demand for analytics

Center of decision making

Skills, not models

Multiple languages

Ethics and confidentiality

Summary: Real Vacancies from Banks

Counterparty Credit Risk Modeler	Quantitative Risk Analyst - Model Validation
M.S. or Ph.D. in Finance, Quantitative Finance, Mathematical Finance or similar	M.S. degree in a quantitative field, preferably augmented by a PhD
2y+ work experience in risk modelling or pricing of derivatives	4y+ experience in a quantitative role in model development for derivatives, xVA or exposure
Solid understanding of financial markets, credit business, derivative products and risk modelling	In depth understanding of quantitative risk management, fin mathematics, stochastic calculus, numerical techniques such as MC/AMC
Programming experience, for example in C++, R , Matlab and VBA	Sound programming skills in C/C++/C#. Familiarity with LaTeX, Python, R is a plus