eoda
daten ▸ wissen ▸ nutzen

Data Science outside

Developing a Generic Scoring Algorithm for Customer Acquisition

Erik Barzagar-Nazari
Data Scientist

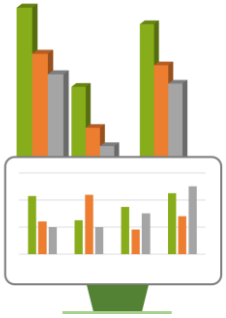
eRum 2016

About eoda

Interdisciplinary Team

Statisticians | Engineers | Economists | Sociologist | …

Based in Kassel - Germany



Data Science Consulting, Training, Support,
Software and Analytic Services with a focus on R

Aims of Today's Talk

I Present a real-world case study

II Discuss unique challenges

III Take a look into our solution

IV Reflect the benefits of using R

Our Client: databyte GmbH

Provides **business information**

Database of about **five million companies**

100 million pieces of information such as
sales, size, branches and many more

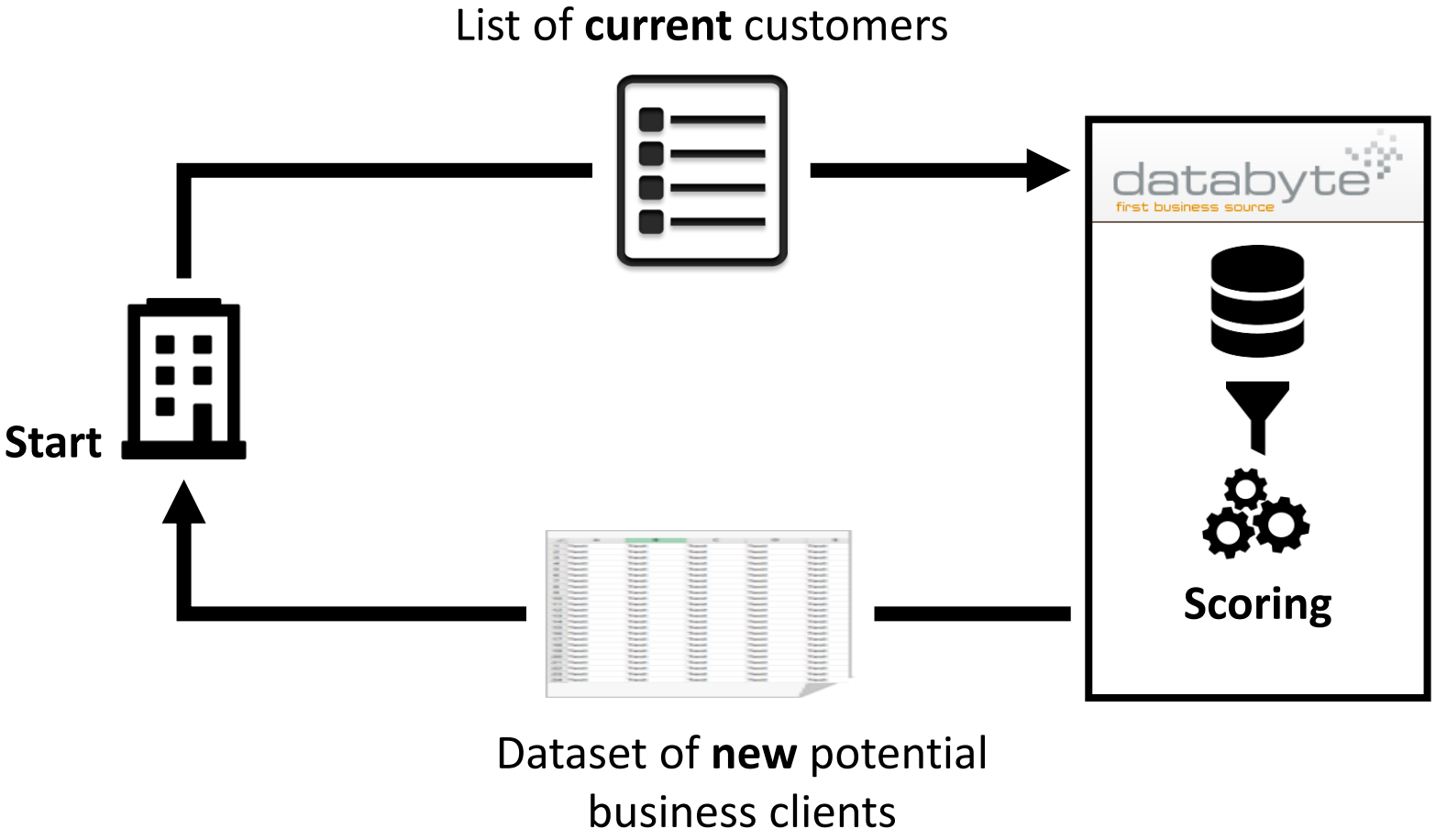
Daily updated!

Use Case: Customer Acquisition

databyte's clients are usually
businesses/organizations…

…looking for **new business clients**
(e.g. for direct marketing campaigns)

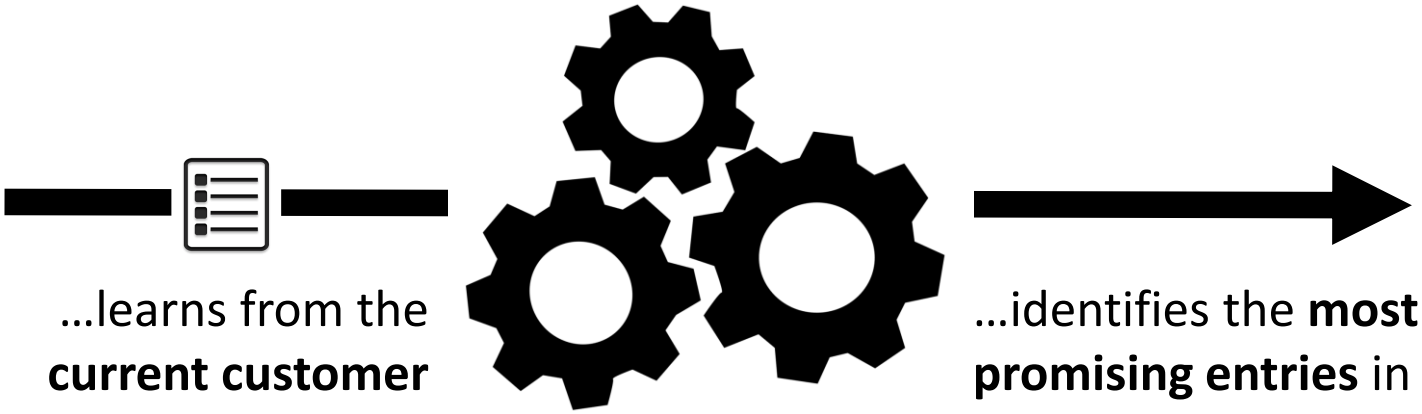
Use Case: Customer Acquisition



Case Study: Our Task

Main task

Develop a new **scoring algorithm**, that…



…learns from the **current customer** base &…

…identifies the **most promising entries** in databyte's database.

Challenges

Image source: http://vignette3.wikia.nocookie.net/simpsons/images/4/43/Daredevil_bart.jpg/revision/latest/scale-to-width-down/1000?cb=20160619043051

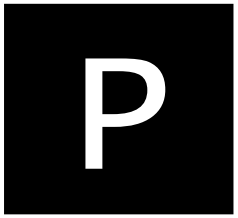
Challenges | Training on Customer Data

{0;1}

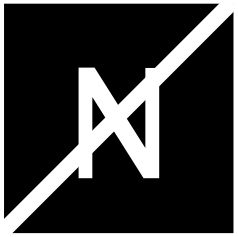
Standard approach
Train a **binary classifier** to distinguish between *non-customers & customers*

Bad News: Does not work in this case, because we only know the **positive data**.

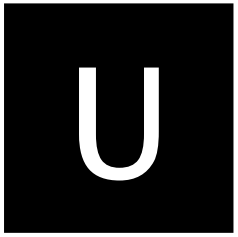
Challenges | Training on Customer Data



Positive Data = Customer Data
Already known customers of the client

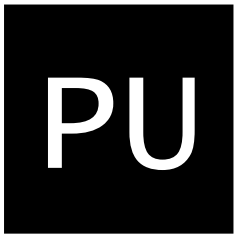


Negative Data = ?
Companies, that definitely do no fit into the client's customer base



Unlabeled Data = databyte's Database
Contains companies, that may fit into the clients customer base as well as companies that do not

Challenges | Training on Customer Data

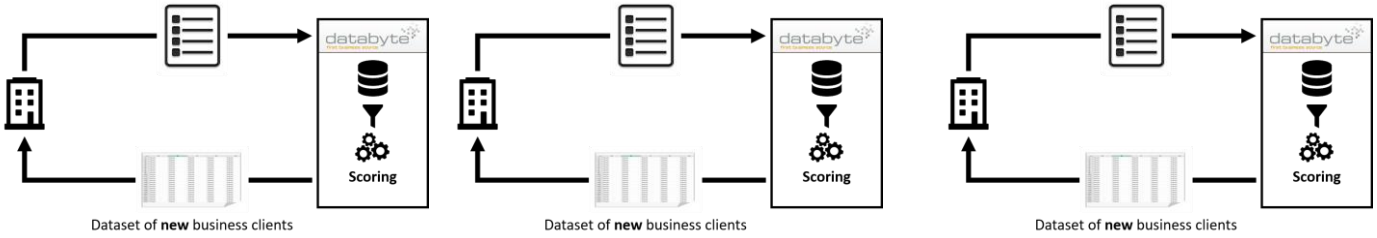


Positive-Unlabeled-Classification

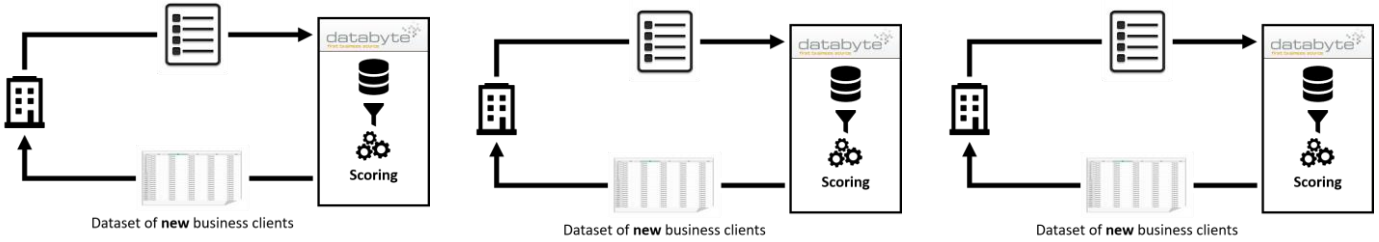
There are strategies to deal with PU-Problems, but…

*…there are no well established **best practices** yet*
*…strategies usually require **strong assumptions***
*…PU-Classifiers require **a lot of tuning**, and are quite **fragile***

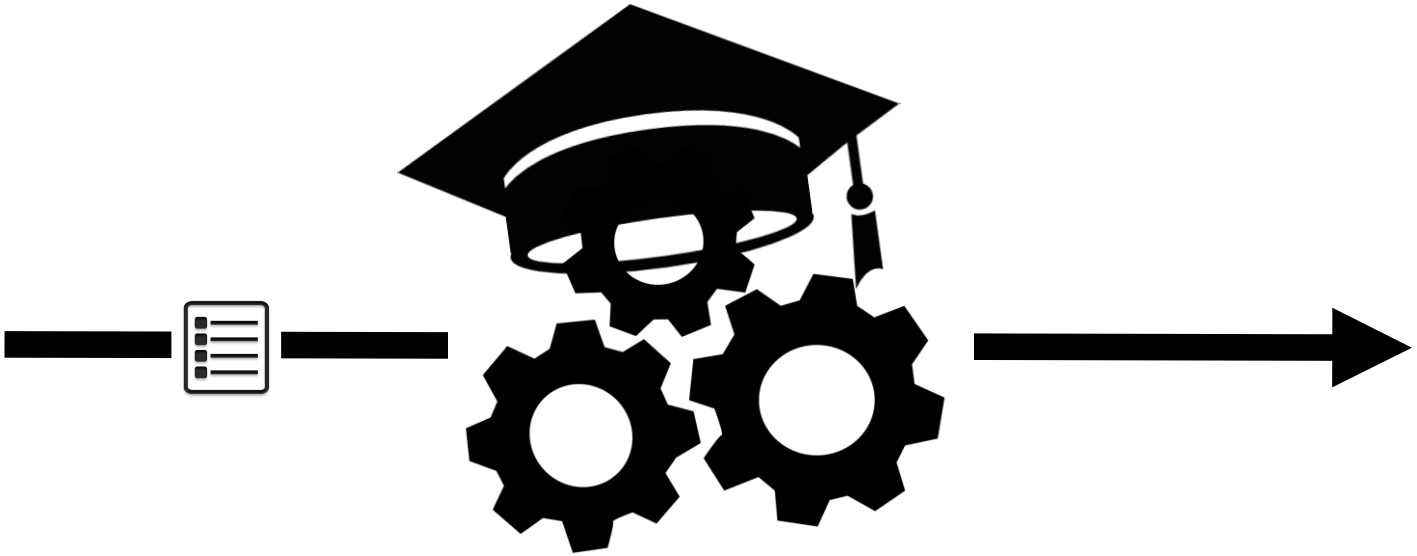
Challenges | Self-Training Algorithm



databyte has **many** clients

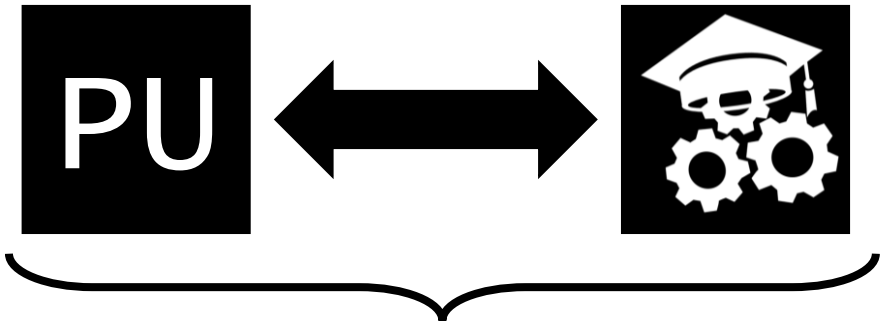


Challenges | Self-Training Algorithm



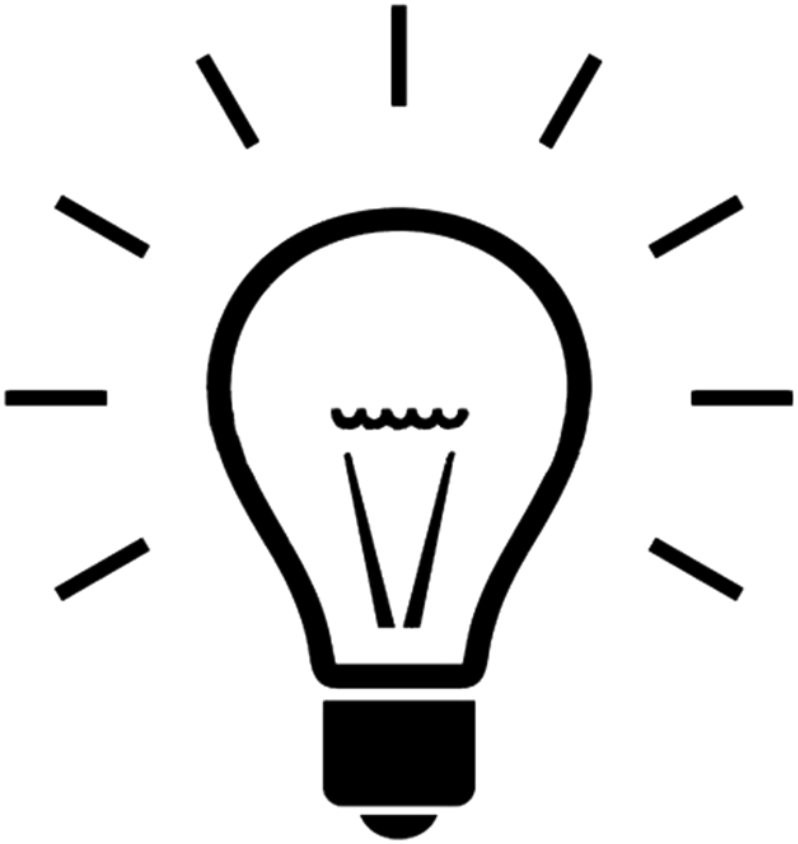
The scoring algorithm must be able to **train itself** based on **unseen training data** (= customer lists)!

Challenges | Conclusion



We have to get creative!

Solution



Solution | Basic Idea

Our approach is based on **similarities.**

Core concept:

1. *Cluster customer data and extract **medoids**, these are **representative customers***

2. *Calculate similarities between **database entries** and **medoids***

Solution | Basic Steps

Segmentation Step

Identify segments based on branches

Core concept:

1. *Cluster customer data and extract **medoids**, these are **representative customers***
2. *Calculate similarities between **database entries** and **medoids***

Weighting Step

Weight similarities based on the distribution of branches

Solution | Pros & Cons

It works and performs nicely!

Comprehensible approach, even for laymen.

Pro

Con

Similarity calculation is costly.

Lack of “rock-solid” theory.

Benefits of Using R

Benefits of Using R

`{data.table}`

Fast & efficient data handling

`{proxy}`

Library of distance and similarity measures

Allows calculation of cross-proximities

Many measures are implemented in C!

`fpc::pamk()`

Partitioning around medoids…

…with **estimation of number of clusters**

Thank you for your attention!

Any questions?

The Data Science Specialists.

eoda GmbH
Universitätsplatz 12
34127 Kassel - Germany

www.eoda.de/en
info@eoda.de
+49 561 202724-40

@eodaGmbH

blog.eoda.de

@eodaGmbH

eodaGmbH