



exper^{teer}

Data Science for Jobboards

Or how machine learning helped us increase our processing capacity by 100x

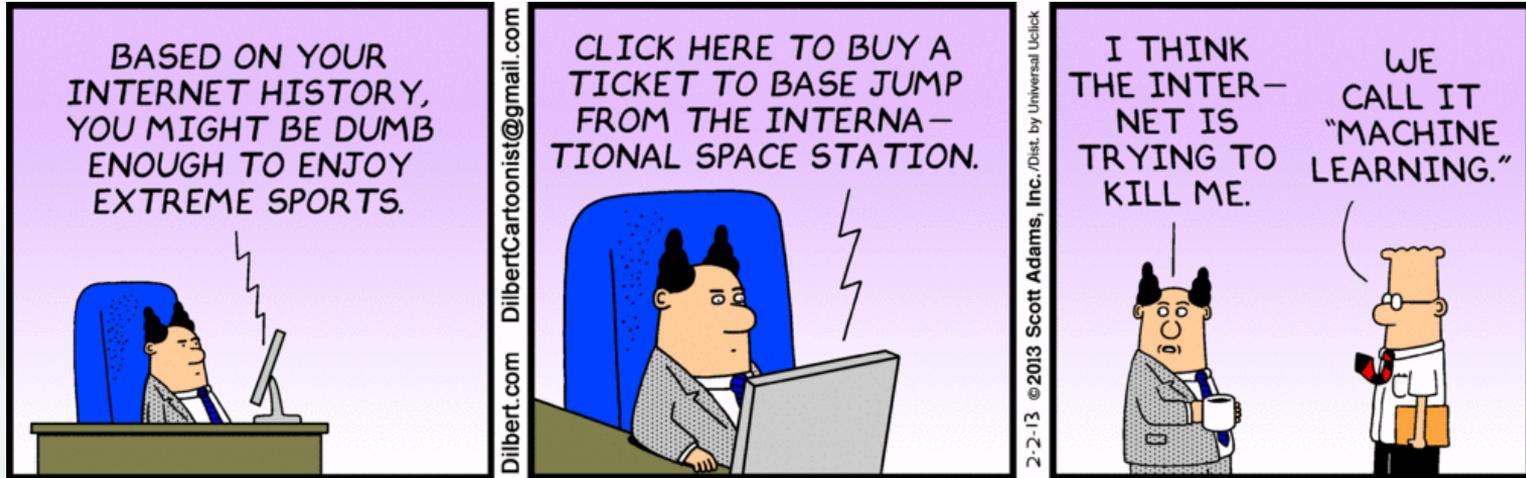
What are you going to learn today:

1. What is data science and machine learning;
2. What is Experteer and what do we do;
3. A basic overview of the US and German labor markets;
4. How we used **machine learning** to increase our data processing **capacity 100x**;
5. Data Science for Job Boards and what is in it for you and your business.

Extract knowledge or insights from data



What about machine learning?



Dilbert.com DilbertCartoonist@gmail.com

2-2-13 ©2013 Scott Adams, Inc./Dist. by Universal Uclick

What about machine learning?

Classical definition:

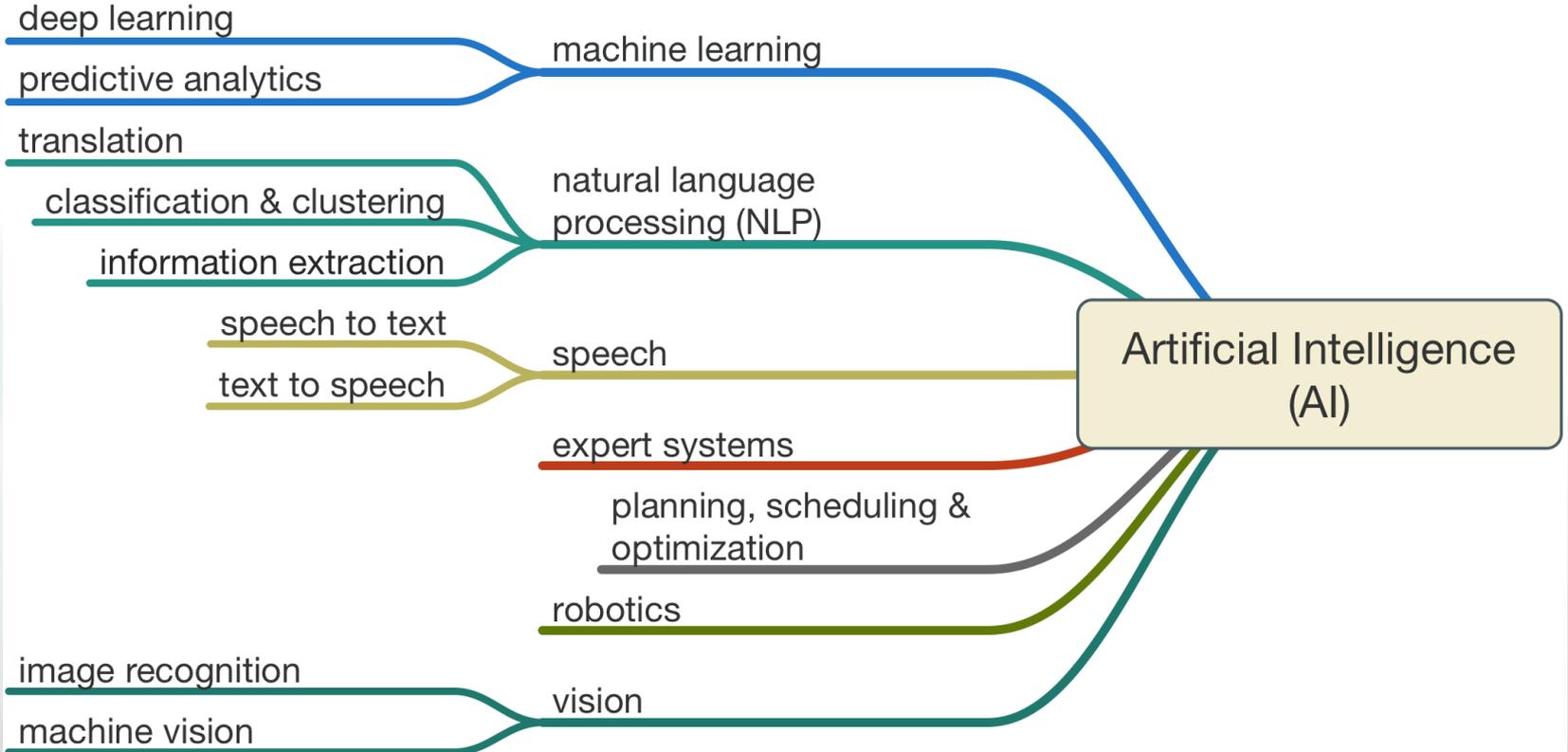
“Machine learning (ML) gives computers the ability to learn without being explicitly programmed” (Arthur Samuel, 1959)

Understandable definition:

ML works with data and processes to discover patterns that can be later applied to new sets of data so that a computer can make a decision.

What about machine learning?

One final overview:



Where are the limits?

Google AI invents its own cryptographic algorithm; no one knows how it works

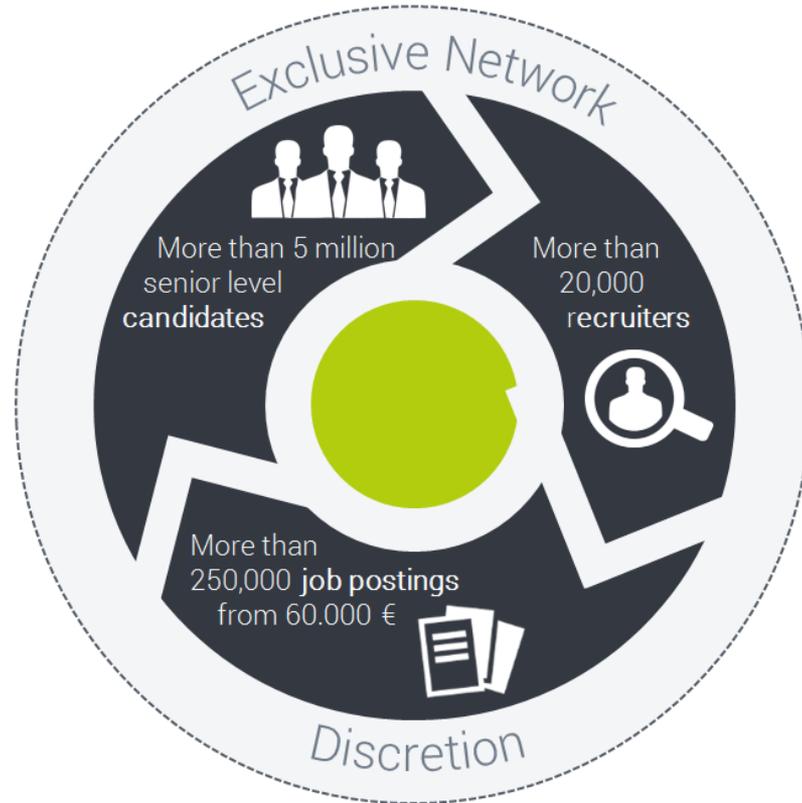
Neural networks seem good at devising crypto methods; less good at codebreaking.

SEBASTIAN ANTHONY - 28/10/2016, 14:26



Now that we covered the basics, let's go deeper and see how we used these marvellous technologies at Experteer...

About Experteer



We want to give job seekers the **most comprehensive** selection of curated executive level positions.

The high-level job market in US/DE

Country	Open Positions	High Level Segment	Relevant Jobs
DE	974,000	13% (above 60K € y.)	130,000
US	5,800,000	14% (above \$100K y.)	868,572

... and we want to have all of them (in all European countries and in the US).

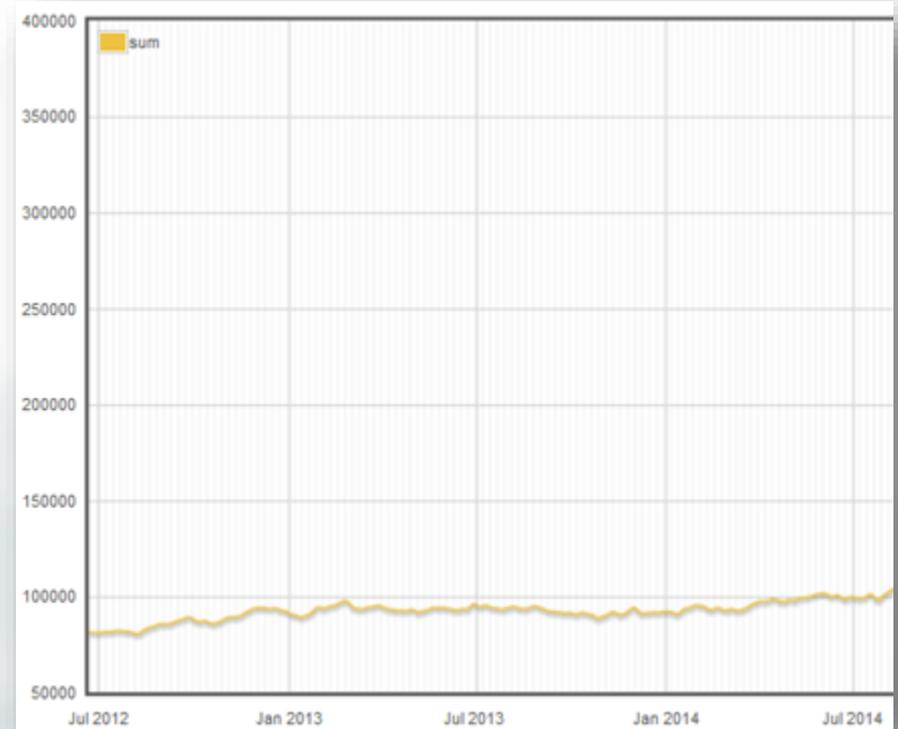
Our old job aggregation process:

- Team of 80 Researchers
 - Hand-picking and classifying jobs
 - Extensive Taxonomy
 - 19 Functions
 - 631 Industries
 - 8 Career Levels
 - Location, education, etc.
 - 8 Languages
- } **95,912 Combinations**

It was good while it lasted....

In 2014 we had slightly increasing cost and almost no growth.

Inventory Development 2012-2014



What we achieved so far 1/2

3x increase of inventory



What we achieved so far 2/2

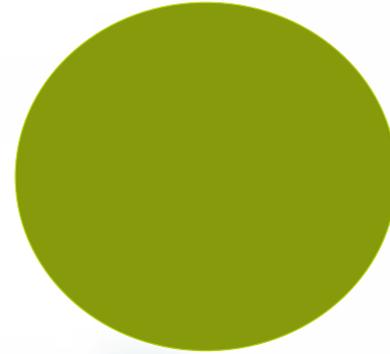
2014

2016

Jobs/day



10K jobs /day



1 mio jobs /day

Dept. cost



Steps we wanted to achieve:

1. Automatic sourcing of jobs (crawling) – *that's easy...*
2. Machine-Learning based classification of jobs in 8 languages – Done!
3. Smart filtering of high-level jobs from large datasets – Done!

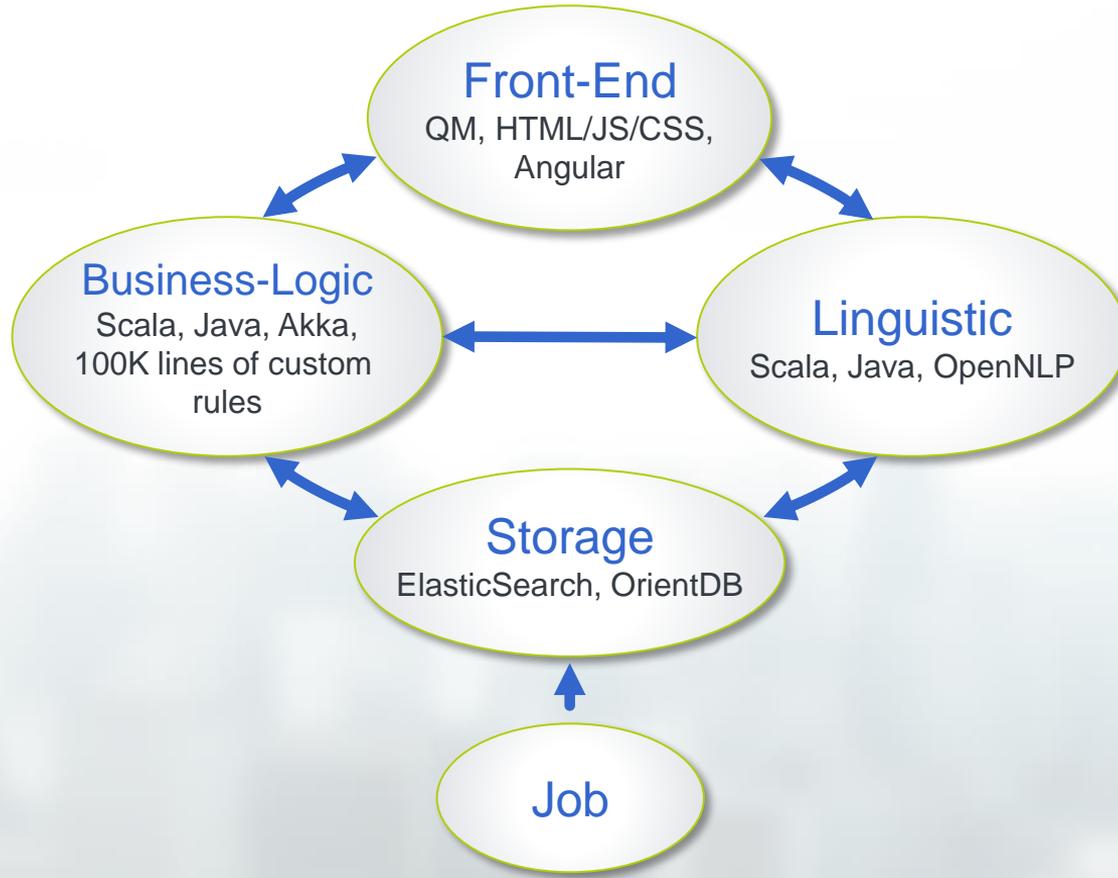
How classification works

1 **Director Regional**
4 **Sales Director Regional Sales**

The Regional Sales Director is responsible for all commercial personnel and 2 activities in a region for the Healthcare Market Division of (...). This includes managing six direct reports, approximately 60 people and over 3 \$300 million in sales. She/He is specifically accountable for the establishment of the company's sales plans and strategies, and for 6 attainment of market share goals, brand position,(...)(...). In addition, the Regional Sales Director will be responsible for leading professional and career development for commercial employees in the Region. Prefer to have candidate live in territory.. Minimum Requirements/Qualifications: Must have B.S./B.A. in business, 5 life sciences or equivalent. MBA is strongly preferred. Previous experience and responsibility for sales and marketing in the 5 laboratory diagnostic industry is required. (...) Must understand the key market drivers and dynamics of the US healthcare market. Must have 1 ten years management experience including prior P&L responsibility with emphasis on 6 strategic sales planning, resource utilization and development, customer relationships, large buyer initiatives, and managing large sales organizations.(...). Two plus years of 7 managing managers with similar scope of responsibility is strongly preferred. Non-Negotiable Hiring Criteria: Must have 8 excellent leadership and communication skills (...). Proven ability to assess, attract, retain and develop personnel

- 1 “Director” is a soft indication, has to be looked into context. It could indicate lots of responsibility, but it really depends on the rest of the responsibilities;
- 2 Looking for a manager with significant experience in people management – this is a strong indication for a people manager career level;
- 3 Large budget responsibility;
- 4 Name of the position – indicates the function;
- 5 Industry: “Life science”;
- 6 Responsible for the “Sales” area in the firm – strong indication for function again;
- 7 A manager that is responsible for managers – indicates a higher management level;
- 8 Soft skills are also a nice indication for career level.

The Stack



But classification is not the only problem...

Initially we thought that classifying the jobs is the only problem.

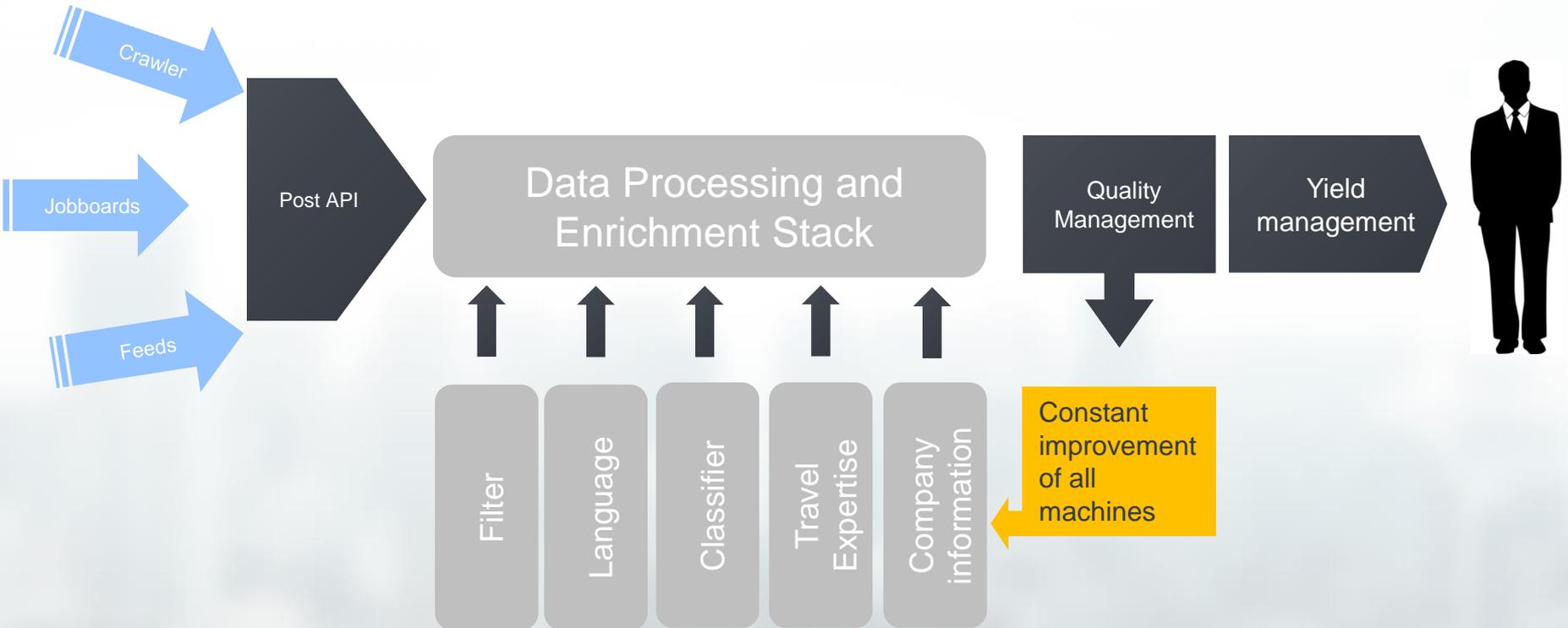
Turns out this was not even half of it problem...

To begin with, models need a LOT of fine tuning...

Experteer focuses on the high-level job market so we need precise classification:

- Exclude blue-collar positions (title: Cleaner)
 - create a negative dictionary of over 5K words for job title;
- Salary data is given in the description
 - build a salary extractor for 8 languages;
- Low-level jobs can be identified only by content (title: manager; description: ... of cleaning staff)
 - Solution: a relevance filter based on job description;
- Company data in the description (You will work for our Subsidiary XXX)
 - Algorithm to recognize context

Event-driven Agile architecture based on microservices



We have some ambitious goals for 2017

- Get a 90% coverage of the high-level job market;
- Achieve 90% automatic classification;
- Extract more metadata from descriptions
 - Soft skills
 - Job Expertise
- Classification for companies

What can Data Science and ML do for job boards?

First, you need lots of data.

You have your data? Good. Then you can:

1. Automate any manual process for which you have lots of historic data
2. Predict different outcomes of events based on existing data

Some things that you could try:

- Build a predictive analytics model based on sales data and predict chance of success for a new b2b lead;
- Predict trending skills from job descriptions for a specific industry;
- Find out when a passive seeker is on the verge or becoming an active one and is most open to buy a paid subscription;
- Identify which skills are currently paying a significant premium on the market;
- Help candidates discover the next best step for their career.

Data science cloud-based tools make testing simple – you don't need a huge team of specialists or an expensive frastructure.

Starting has never been easier:



The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main workspace shows a workflow titled "Binary Classification: Direct marketing" in draft mode. The workflow consists of several modules connected in a sequence:

- Reader**: The starting point of the workflow.
- Metadata Editor**: Used to view and edit the data's metadata.
- Project Columns**: A module to remove columns that are part of the label.
- Split**: A module to split the data into training and testing sets.
- Two-Class Boosted Decision Tree**: A machine learning model (labeled with a '1') trained on the split data.
- Two-Class Support Vector Ma...**: Another machine learning model trained on the split data.
- Split**: A second split module, likely for further data partitioning or validation.
- Sweep Parameters**: Two modules used to tune the parameters of the models.
- Score Model**: Two modules used to evaluate the performance of the trained models.
- Evaluate Model**: The final module in the workflow, used to compare the results of the different models.

On the right side, the **Properties** pane is visible for the "Two-Class Boosted Decision Tree" model. It shows the following settings:

- Create trainer mode**: Single Parameter
- Maximum number of leav...**: 20
- Minimum number of sam...**: 10
- Learning rate**: 0.2
- Number of trees construct...**: 100
- Random number seed**: 0
- Allow unknown categ...**

At the bottom right, there is a **Quick Help** section with the text: "Creates a binary classifier using a boosted decision tree algorithm (more help...)"

What about the future



There are no limits...

THANK YOU!

Any questions?

If you have any sudden ideas and you want to know if they are technically possible, feel free to approach me afterwards.