Tips for switching to data science

Max Baak — 15 Nov 2021

ING

WHOLESALE BANKING
advanced
analytics

Max Baak - data scientist / statistician / former HEP physicist

Background in HEP (2002 - 2015)

- 2007: PhD @ Nikhef, on research at BaBar experiment (SLAC)
- 2007 - 2015: fellow, then staff researcher at CERN

KPMG Advanced Analytics & Big Data team (2015 - 2018)

- Data science consultancy
- First “Chief Data Scientist”, then team lead (2017).

ING Bank (2019 - now)

- Chapter lead data science (data scientist & manager of 18 data scientists)

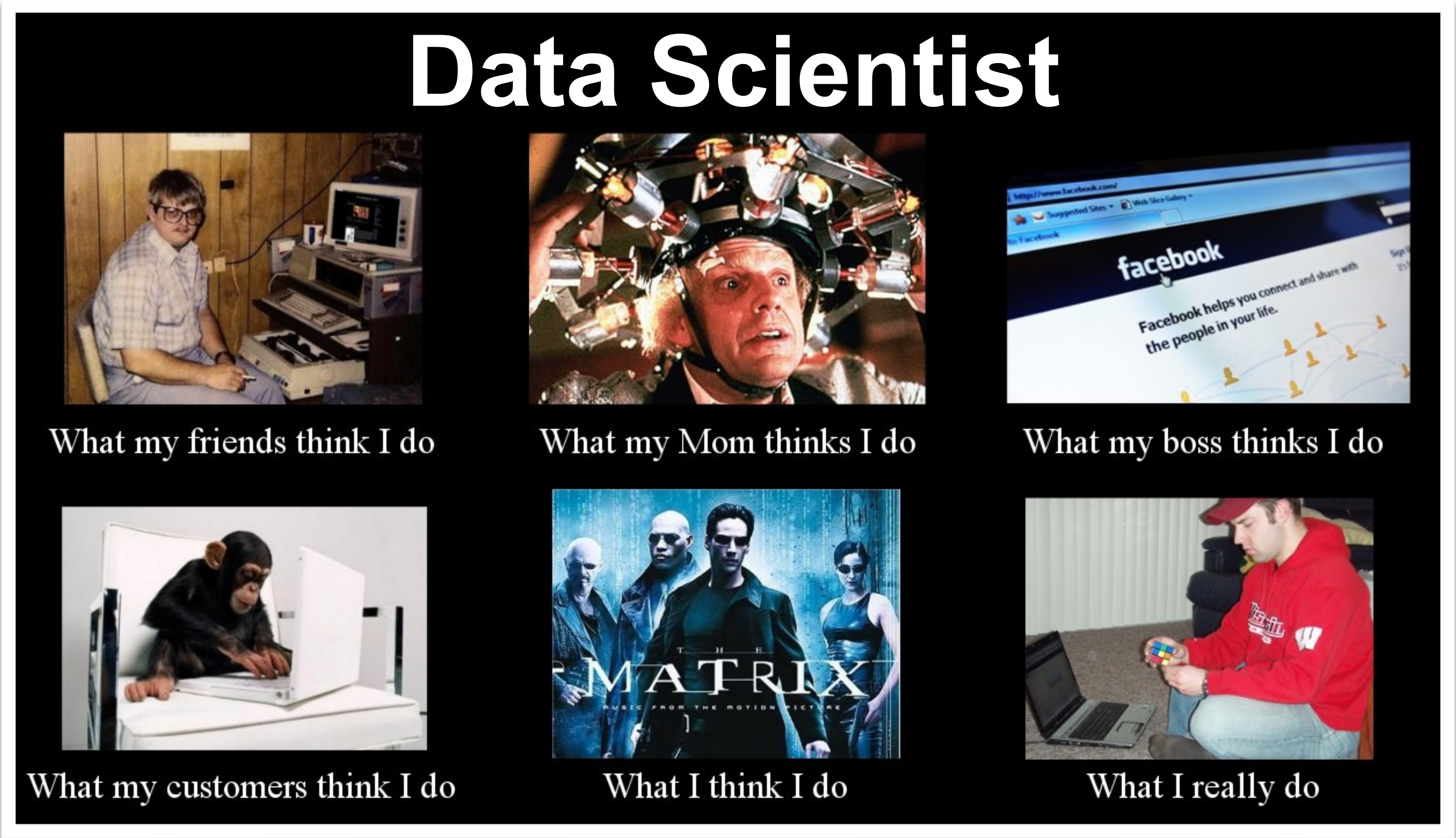
Guest scientist at AI department of UvA (2021 - now)

- Interest in Explainable AI: statistical methodology

Outline

- (Brief) project example.
- Tips for your Data Science job interview.

Harvard: the sexiest job of the 21st century


Data Scientist
What my friends think I do
What my Mom thinks I do
What my boss thinks I do
What my customers think I do
What I think I do
What I really do
facebook
Facebook helps you connect and share with the people in your life.
THE MATRIX

"We solve complex data puzzles"

ING bank

ING

- Wholesale Banking: caters to (big) companies.
  - We build data-driven “AI products”.
    - These benefit these clients and generate revenue / save costs for the bank.
- In terms of data science, our projects are quite diverse.
- Our team:
  - 120 people, 18 data scientists.
  - 11 projects.

WHOLESALE BANKING
advanced
analytics

The bond market

- Bonds are loans issued by companies. Bonds can be traded.
    - Every company can issue multiple bonds with different characteristics (like maturity, rate, seniority).
- The bond market is a traditional market
    - $700 billion traded daily, 3 times more than stocks
    - 50% of the value is voice traded.
    - Market can be inefficient and illiquid
- Bonds are grouped into “universes”.

Bond pair trading

- Bonds can move together.
    - Therefore market anomalies can appear between bonds.
- Look for a pair of bonds that are correlated and/or related to each other.
- Wait for an anomaly to appear, meaning:
  suddenly the difference between the two bonds increases.
- Assumption is that bonds will converge to each other again.
    - Process called: mean reversion.

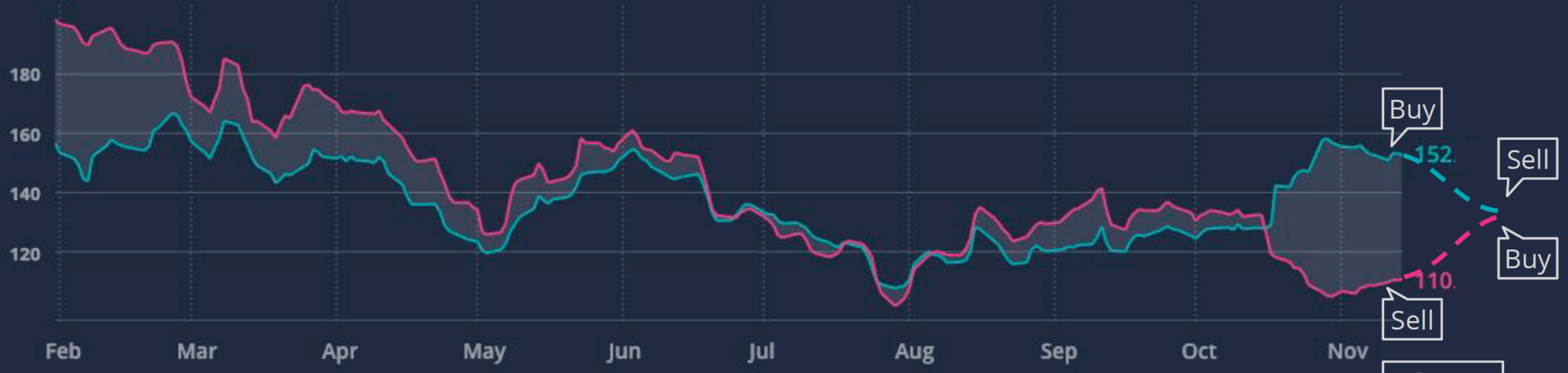
RENAUL 1 5/8 26/05/2026

FR0013334695 · France · BBB+

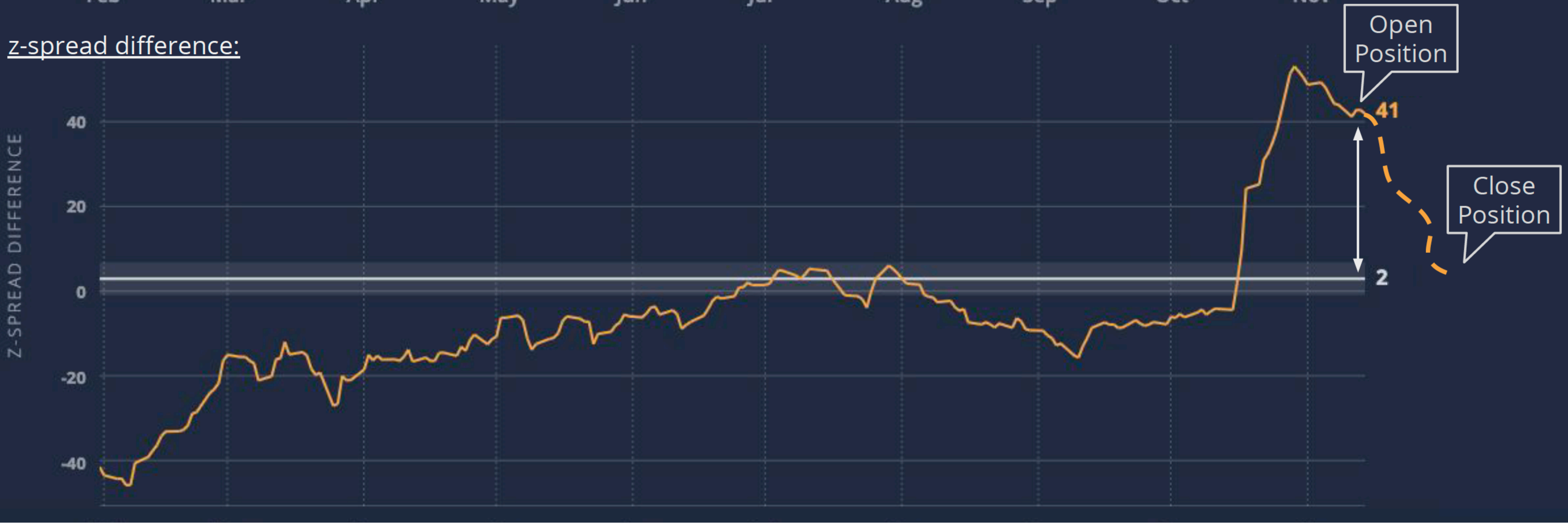
VW 2 1/2 31/07/2026

XS1944390597 · Germany · A-

z-spread of the 2 bonds:



z-spread difference:



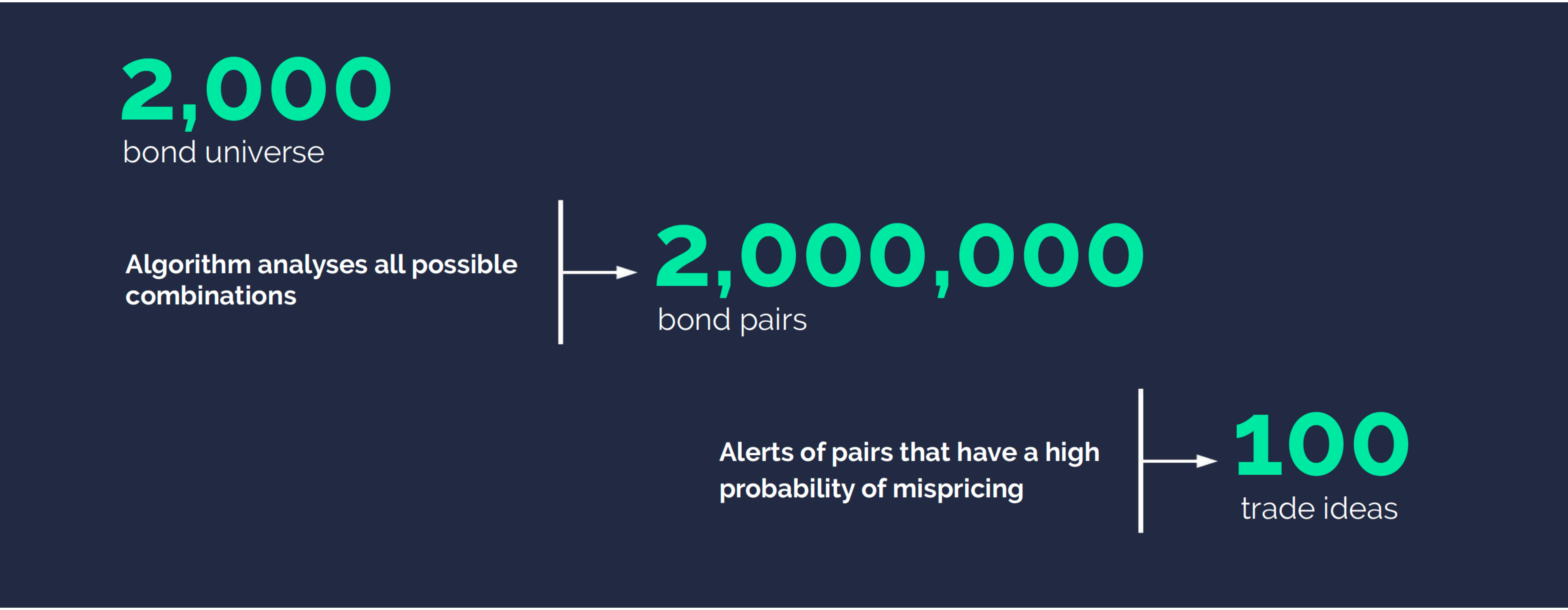
Detecting anomalies in relative value

- We built a product that gives trading suggestions to asset trader based on relative value of bond pairs.
    - Advice, not auto trading.

- There is trade-off between the number of alerts and profitability.

- The trick is: alerting enough in advance.

A filtering exercise



2,000

bond universe

Algorithm analyses all possible combinations

2,000,000

bond pairs

Alerts of pairs that have a high probability of mispricing

100

trade ideas

Data science problem break-down

How to perform:

1. Identifying relevant pairs of bonds
2. Detecting anomalies
3. When to open position
4. When to close position, after mean reversion happened

Tips for your job interview

Job market

- Market = Good.

Use your connections

- ex-High energy physicists are everywhere.
- Easiest way to open door to industry.

CERN

- CERN has a great reputation.
  - Everyone knows the Higgs particle.
- Use to your advantage.
- “At CERN I’ve been trained in big data and advanced analytics.”

Motivation for transition

- Everyone understands why you are leaving research.
  - No need to explain in detail.
- More important: explain what attracts you to working in (specific branch of) industry.
  - “I want to try something else.” (not really convincing enough)

Consultancy vs Industry

Data Science consultancy:

- Short projects: average 2-3 months. Really high pace!
- Crash course in DS & business.
- (I loved working with different data & algorithms all the time.)

Other industries:

- Longer projects. More time to develop proper models.

On experience with clients

- Potential worry of employer: no experience with clients.
- Emphasise your responsibilities to (big!) collaboration.
- Service work that was used by 3000+ researchers?
  - Those are your “clients”.

What makes you stand out

- … from other good data scientists?
- Make sure you have a convincing answer ready.

You know *a lot* about data

As high-energy physicist:

- Data acquisition, filtering, reconstruction, lineage, quality, running software in production, data monitoring, data exploration, data fixes, building analysis models, evidence collection, validation, error analysis, statistics, writing analysis documentation, defending your results, analysis review, etc etc.
  - >80% of the work of a data scientist!
- Way more than other, typical data scientist.

physicicist vs computer scientist

- Computer science: focus is often on highest performing algorithm.
- Important in physics: extract insights from (complex) data.
  - *Thorough: data understanding, covering systematic effects, completion of evidence gathering, defending your results.*
  - *My experience: strong background in statistics & methodology is very useful.*
- Complementary skills. Both are important!
  - Very useful to have ex-physicist(s) in DS team.
    - *Critical mindset: breaking down analytics problems.*

Machine learning

- Take ML courses - make sure you know the fundamentals.
  - Redo a few ML projects (Kaggle).
  - Know the popular ML libraries, algos in there, etc.
  - Make sure you can pick up new ones quickly.
  - Does CERN offer courses? Good book: Bishop.
- Good chance you have to do a technical exercise at interview.
- You’re expected to be able to use these techniques from day 1.

Scaling up Data Science

- CERN is an analytics factory.
- Many companies want to know how to scale up impact of data science.
    - E.g. reusable projects / analysis code / models in production.
    - (Employers love to hear successful examples.)

Agile / Scrum

- Agile / scrum way of working.
- Take a course, learn what it means, apply it. (CERN?)
- Actually, also good for physicists.

Projects on github

- Side projects are good.
- Makes you stand out.
  - Shows your connection to DS, your coding skills, etc.

Retrospective

- I love my job. No regrets!
- I have very fond memories of my time in physics.
    - If you consider switching, do it before love for physics turns sour.

Go for it.

Best of luck!

Contact

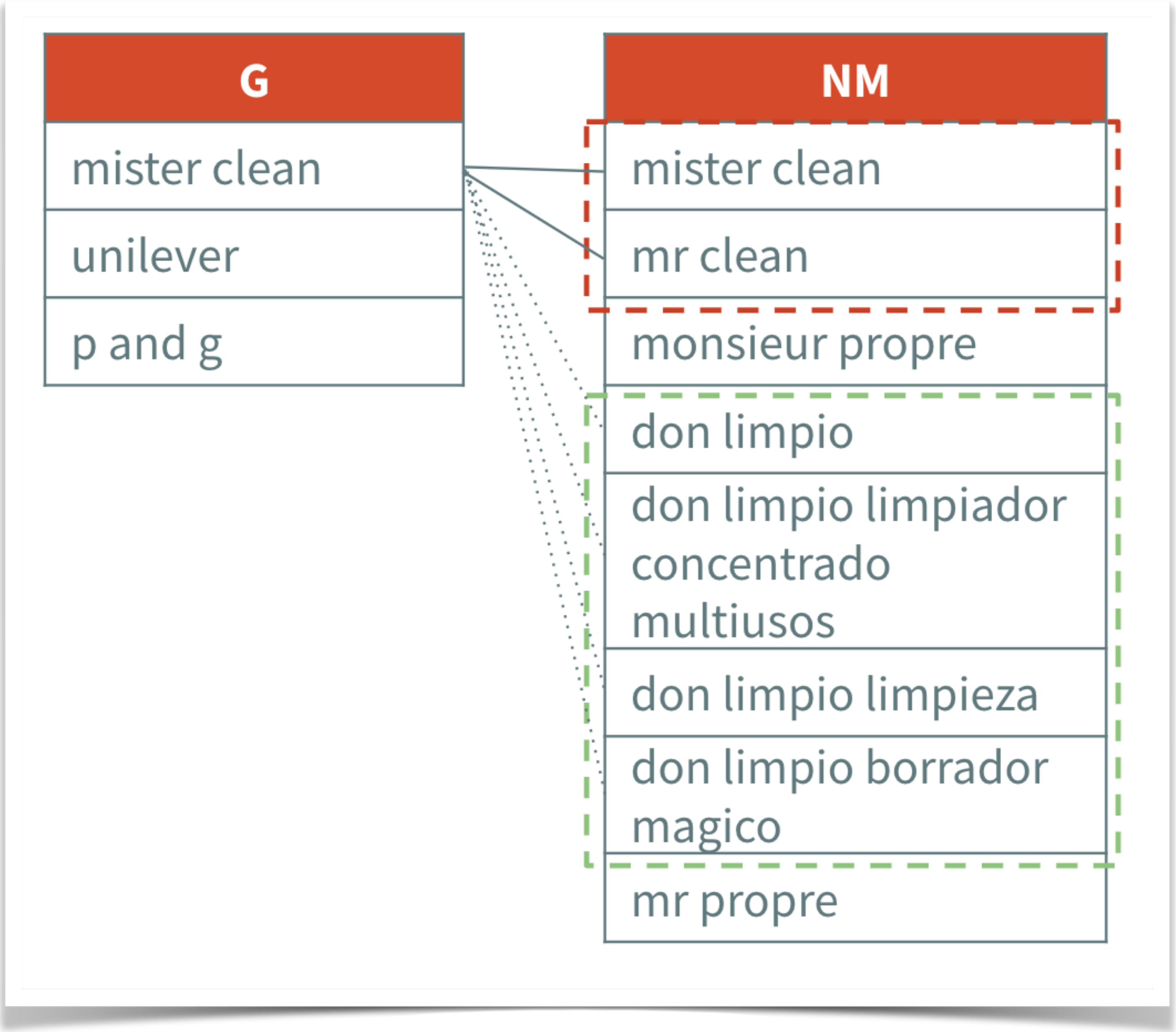
- max.baak@ing.com
- Or contact me on LinkedIn

Name matching

- Wholesale banking deals with corporate clients.
  - Bank wants a holistic view of these clients.
- **One way to enrich: look at names on in-/outgoing transactions**
- Two use cases:
  - Match external bank accounts to ING accounts
  - Match (high-risk) names to international watch list(s)

Reality

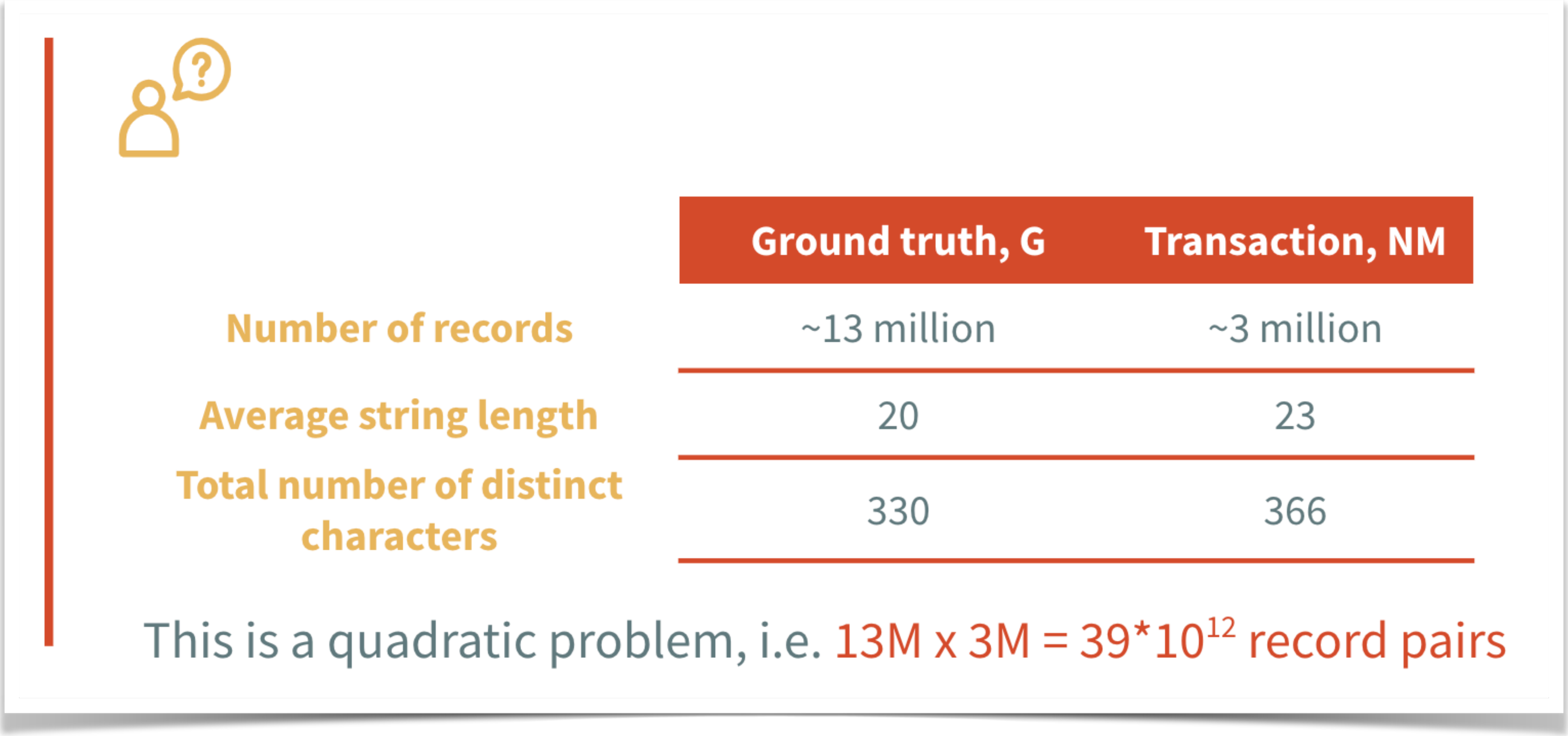
- Significantly different names used for same entity
  - company names / owner names
  - abbrev. / wrong field / generic name
  - entities in different countries
  - different companies under same entity
- Mismatch with ground-truth names when doing entity matching on full set.



G
mister clean
unilever
p and g

NM
mister clean
mr clean
monsieur propre
don limpio
don limpio limpiador concentrado multiusos
don limpio limpieza
don limpio borrador magico
mr propre

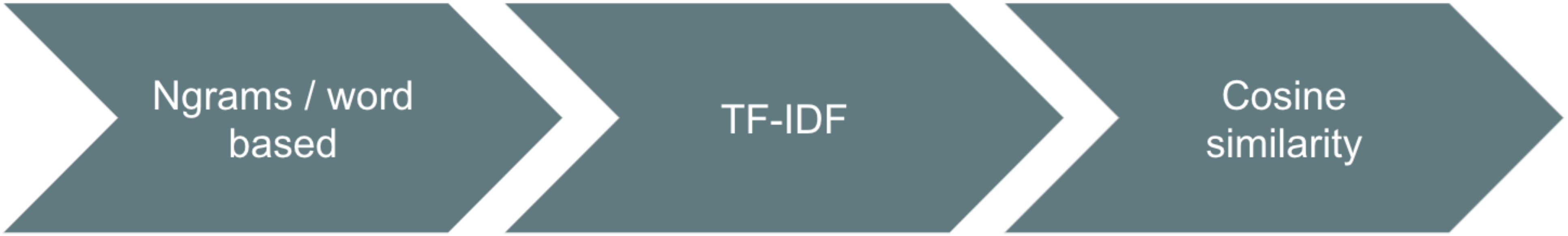
Why name matching at scale?



	Ground truth, G	Transaction, NM
Number of records	~13 million	~3 million
Average string length	20	23
Total number of distinct characters	330	366

This is a quadratic problem, i.e. 13M x 3M = $39*10^{12}$ record pairs

Cosine similarity metric



- Our string similarity distance is focussed on speed.
- Quadratic computational complexity.
  - 13M x 3M = 39*10^{12} name pairs
    - By hand (1 sec / pair): 1.2M years to sort through!
- Cosine similarity of any name-pair in range [0,1].
  - 1 = perfect match
- Alternatively: use classifier output.

https://github.com/ing-bank/sparse_dot_topn

Ranking the candidates

- Step 1: Candidate-pair generation
  - Name to match:
    - Lotuss watch
  - Ground truth candidates:
    - Lotus Watches
    - Lotus Toilette Paper Ltd.
    - Lotus Cars corp

- Step 2: We *rank* the candidates based on cosine similarity score:

Lotuss watch	-> 1. Lotus Watches	0.99	✔ correct match
	-> 2. Lotus Cars corp	0.90	✘ non-match
	-> 3. Lotus Toilette Paper Ltd.	0.80	✘ non-match

- Results in candidate of: rank 1, rank 2, etc. (We consider top-10 candidates only.)

Types of names to match

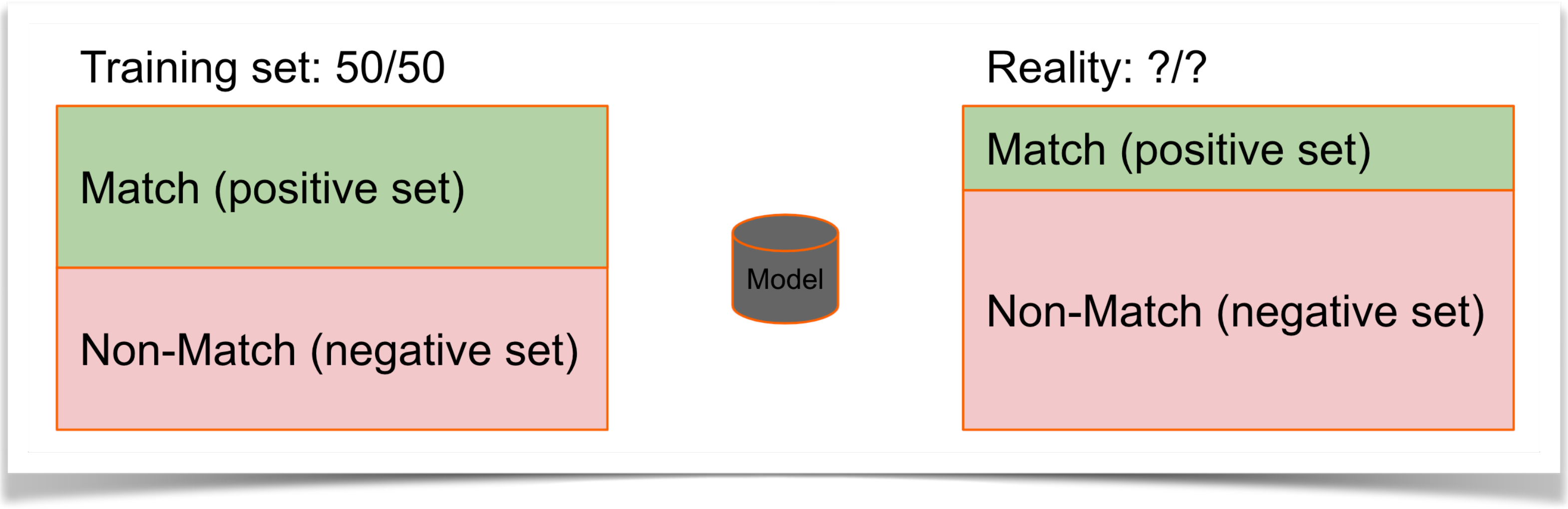
1. *Positive name:* The name-to-match belongs to a name in the ground truth

   A. *Positive correct:* matched to the right name

   B. *Positive incorrect:* matched to the incorrect name

2. *Negative name:* The name should NOT be matched to the ground truth

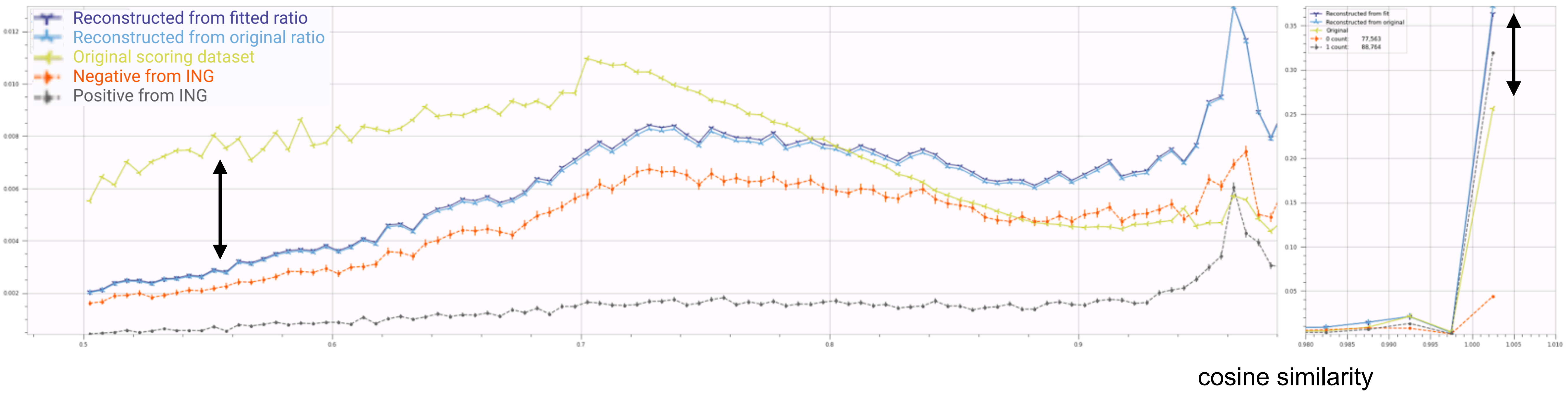
Adapting name matching to different name sets



- Existing model gives a score based on assumed ratio of positive/negative names.
- In “reality” we don’t know the negative fraction!
  - The correct value may be very big (even for ground truth with 9M names).
- We would like our model to give a calibrated probability that a name is a match or not.
- **How to estimate the negative fraction for scoring set?**

ING vs non-ING datasets behave differently

S. Collot



- The distributions are quite different between ING and non-ING names.
    - Both negative and positive name-pairs behave differently.
        - I.e. out-of-the-box name-matching is uncalibrated.

- Can one correct for these two types of dataset shift?
    - (Research I've been pursuing over the past half year.)

Data science problem statement

How to perform:

1. Preselection of relevant name pairs - lots of filtering.

2. Identifying *correct* name pairs.

3. Correct for differences between ING and non-ING data.