Analysing the World's News: Challenges and Learnings from Industry

Miguel Martinez, Co-founder / Chief Data Scientist, Signal AI
@miguelmalvarez
www.signal-ai.com / www.research.signal-ai.com

Before Starting... two questions for <u>you</u>

- Raise your hand if you are a CLEF lab organiser with industry co-organisation

Before Starting... two questions for you

- Raise your hand if you are a CLEF lab organiser with industry co-organisation
- Raise your hand if you, as a researcher, have involved industry partners in 2019

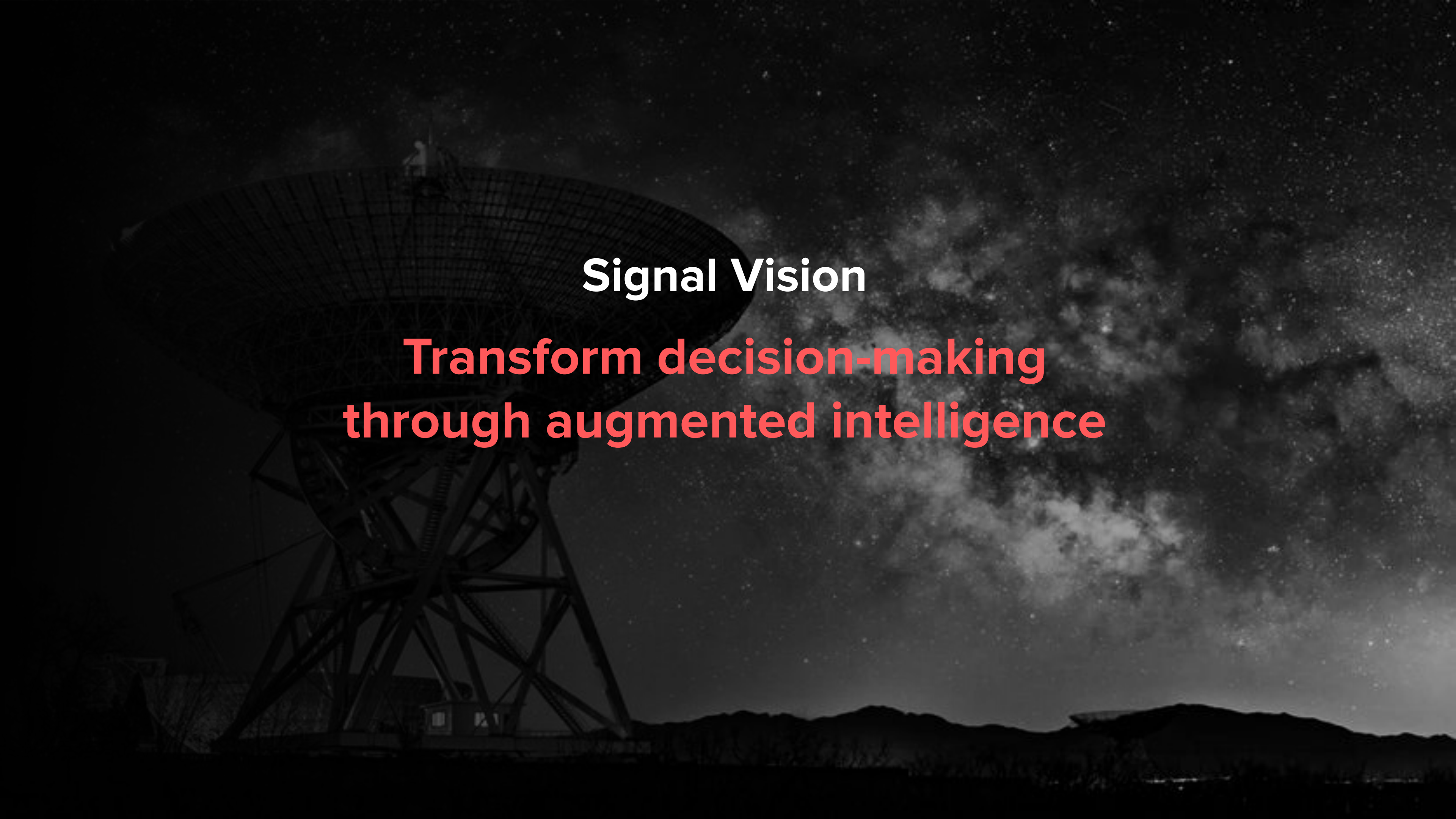
Today's agenda

- What is Signal AI?
- Connection to CLEF
- Signal AI's approach and lessons learnt
- Academic vs Commercial Research

Please ask questions on the fly!

- Signal AI is a 6 years old B2B company
- 150+ people in 3 continents (EMEA, NA, APAC)
- 100s of clients
- Academic and Practitioner Community involvement
- VC funded: $30M+ raised

- *Fun fact: 2/3 founders were “found” on meetup.com*


Signal Vision
Transform decision-making
through augmented intelligence

Use Cases

Opportunity

Leading Market Shift

Lead Generation

New markets

Reputation

Customer Feedback

Damaged Products

PR/Comms

Risk

Regulation

Competitors Initiatives

Similar Problems

Complex information access end-to-end tasks

Over multiple data Types

Changing over time

Some driven by industry needs (e.g. RepLab)

Similar Perspective

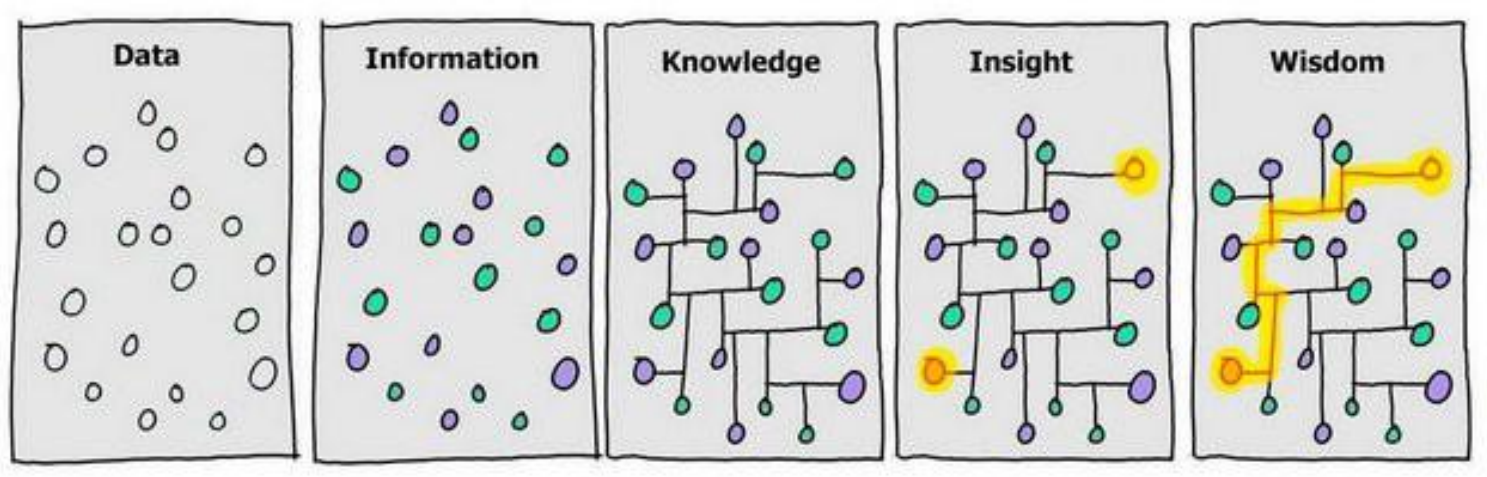
Realistic evaluation frameworks

Multi-field expertise

Open for collaborations

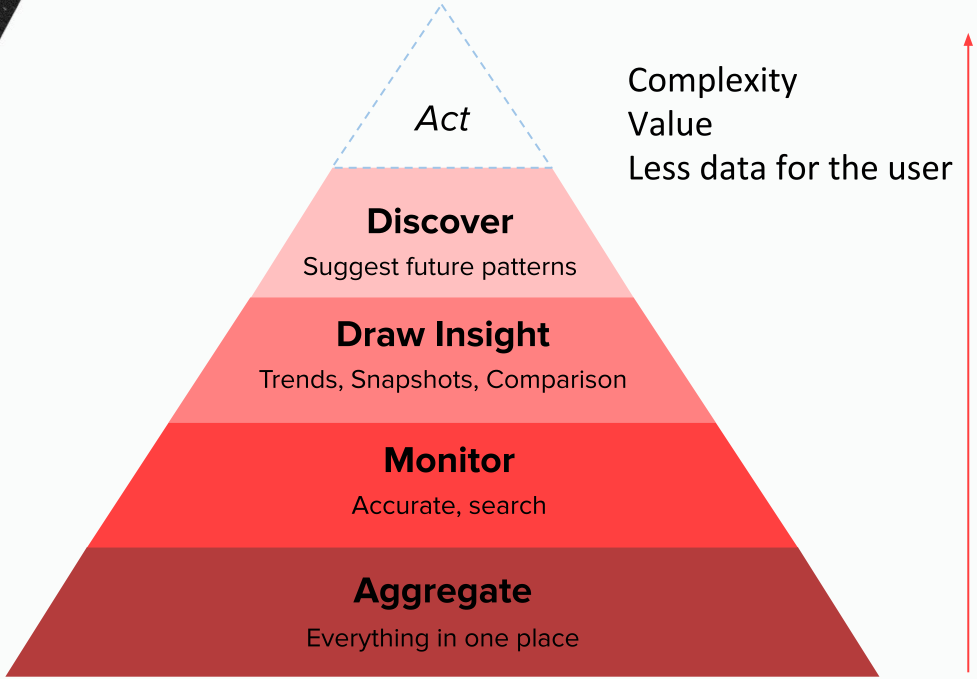
Adapting as new problems appear

PERSPECTIVE



Source: https://random-blather.com/2014/04/28/information-isnt-power/

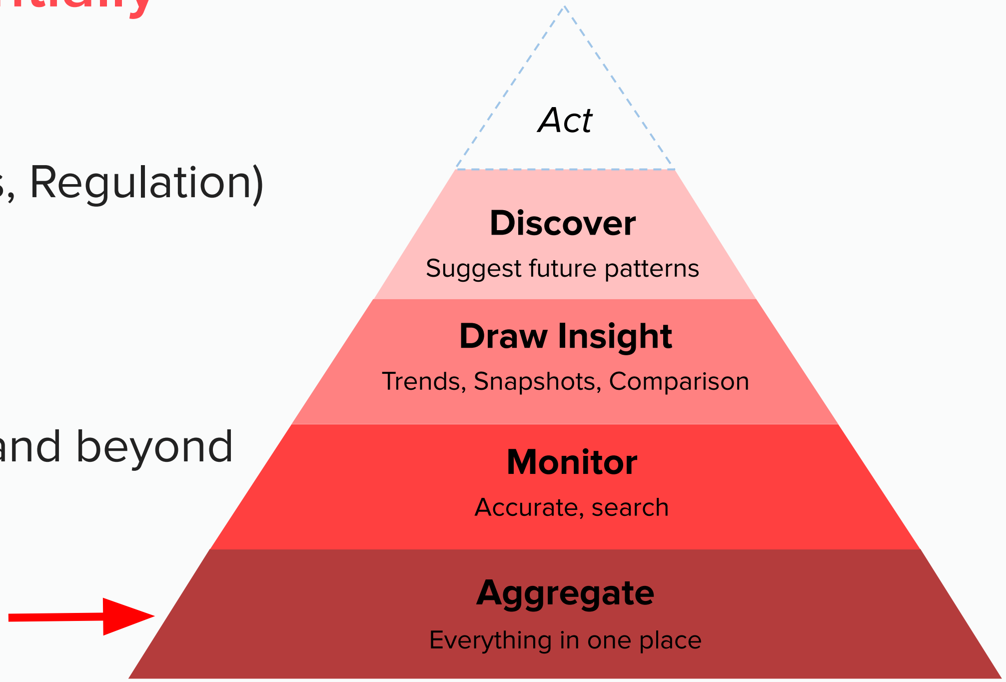
SIGNAL PERSPECTIVE



Complete coverage of potentially important sources

- Multiple Data Types (News, Blogs, Regulation)
- Real Time
- Multiple Languages

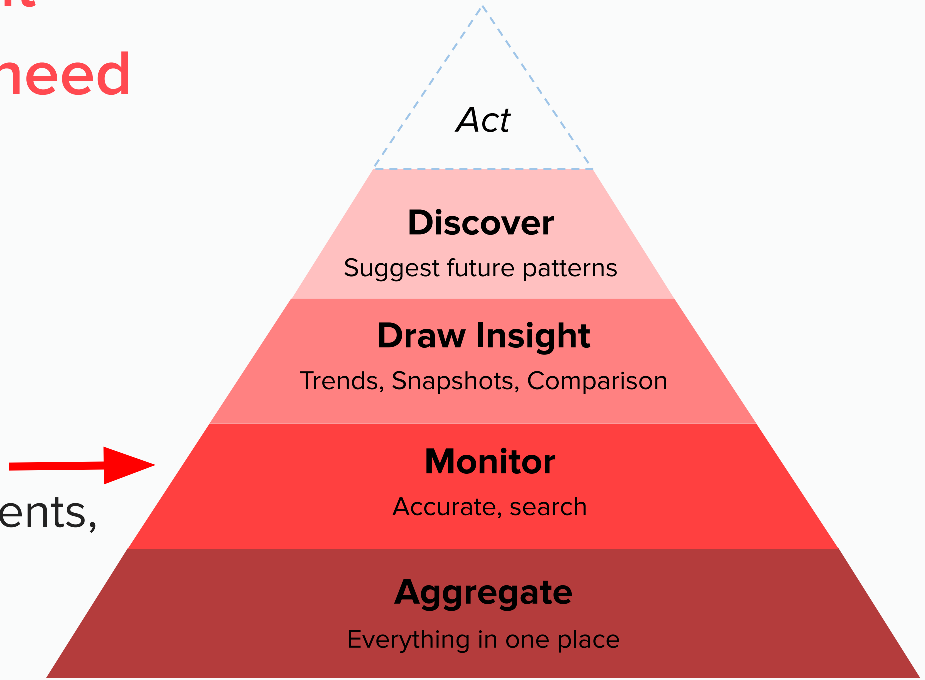
- Future: Images, Patents, Twitter, and beyond



Retrieving only, and all, relevant documents for an information need

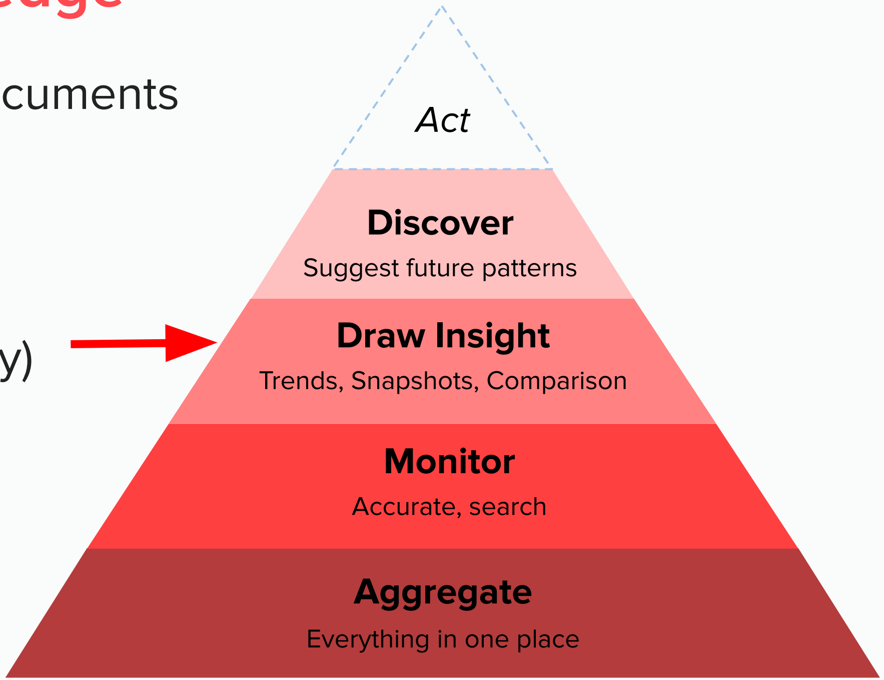
- Information Filtering (entities, topics, sources)
- Based on complex queries
- Document-focused

- Future: Factual vs non-factual documents, Reputation polarity



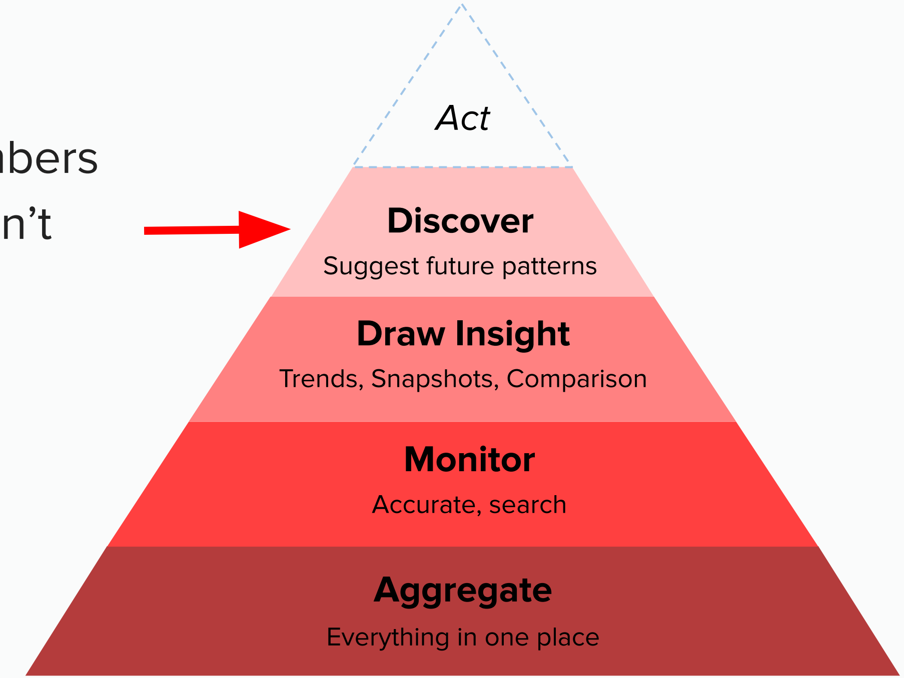
Distilling documents into knowledge

- Quick exploration of high-volume of documents
- Focused on sets of documents
- Trends and anomalies
- Time changes
- Data visualisation (visualisations are key)



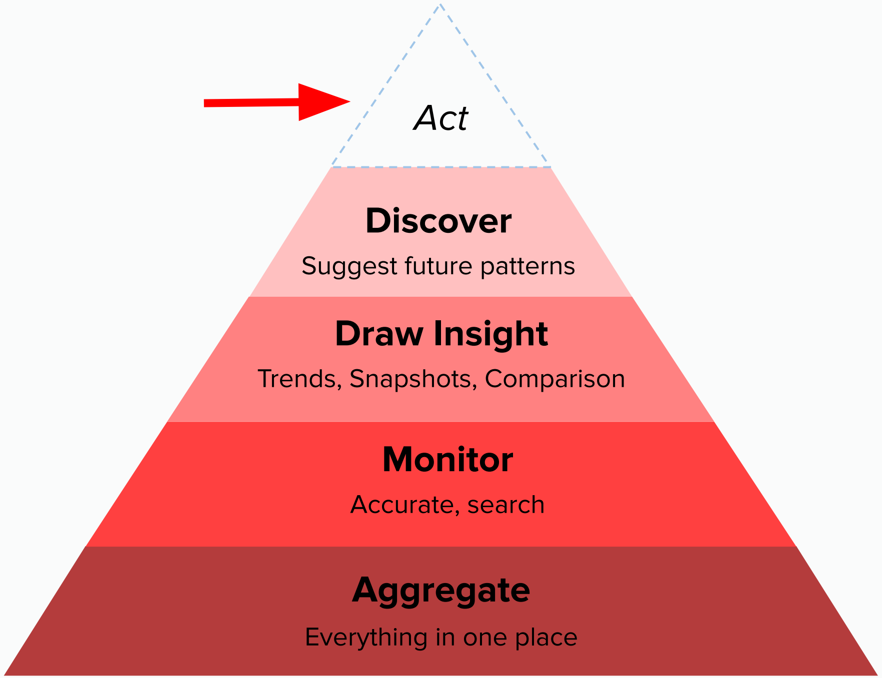
Unknown unknowns

- Horizon Scanning
- Recommendations outside echo chambers
- Factors you should care about but aren't aware of yet



Suggest actions and predict consequences

- Long-term ambition
- Predictive modelling and forecasting
- Digital *"Consiglieri"*

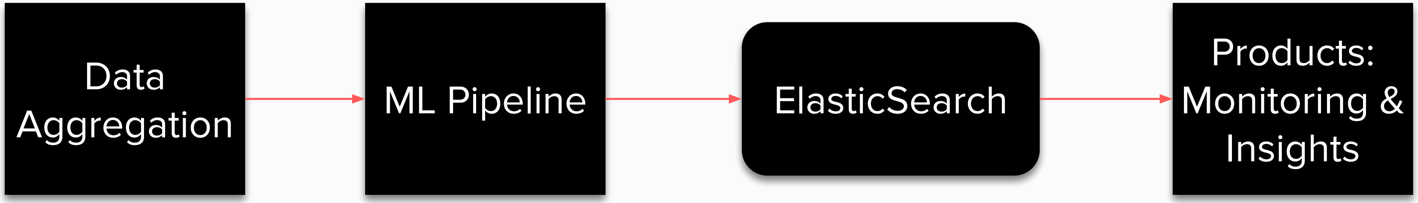


PROCESSING PIPELINE

- Process 3M+ documents daily
- Easy to add new components and amend
- Multiple types of textual data
- Multiple languages
- Reprocessing data with new models

PIPELINE

HIGH-LEVEL

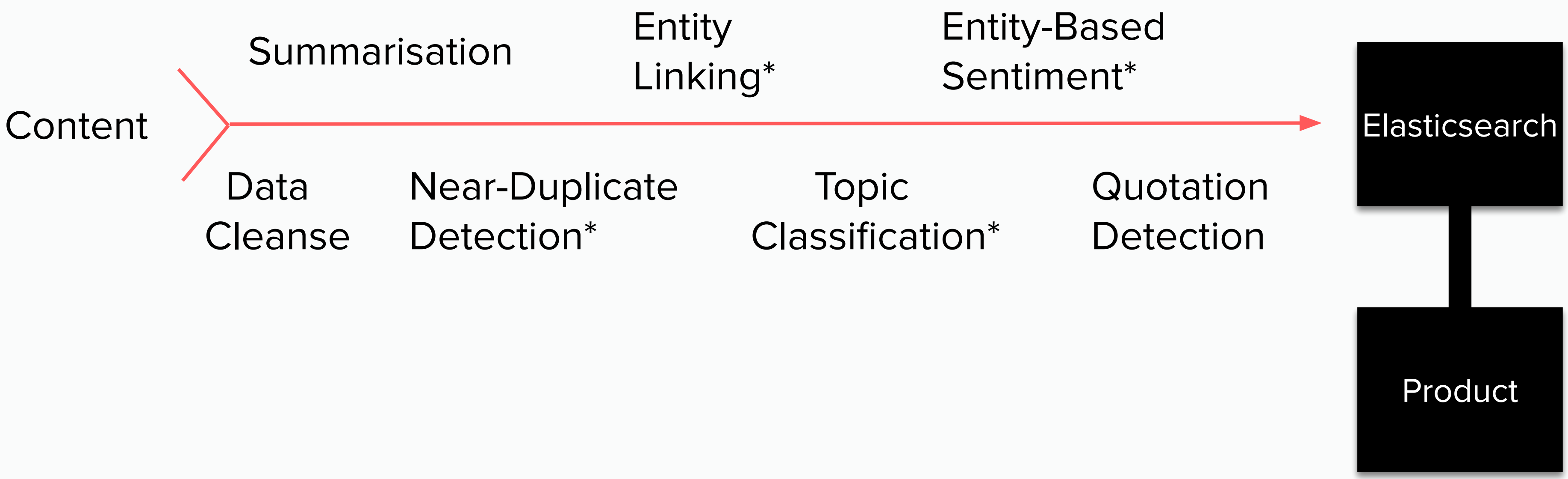


DATA AGGREGATION

- Multiple data types are crucial for many use-cases

- The line between different data types is blurring
  - Influencers using blogs or social media, like Twitter, are more impactful than some newspapers

ML PIPELINE


Content
Summarisation
Entity
Linking*
Entity-Based
Sentiment*
Data
Cleanse
Near-Duplicate
Detection*
Topic
Classification*
Quotation
Detection
Elasticsearch
Product

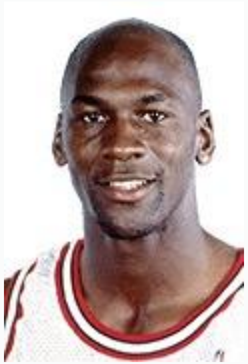
Isn’t it a solved problem?

- The problem is the sheer volume and velocity of data
- Hashing and dimensionality reduction models (LSH)

- Balancing <u>what is a duplicate</u> for different users is the main challenge
- Around half of our daily articles are duplicates

Disambiguating Mentions

- Name Entity Recognition (NER) is not meaningful for tracking relevancy
- Disambiguation to a known Knowledge Base is needed



<u>Michael Jordan</u> is great

<u>Michael Jordan</u> is a great researcher

How good is good enough?

Broad Coverage using Wikipedia:

- 100,000s of entities
- Close to 0.90 avg. F1 in Wikilinks EL dataset
- Much quicker than other (academic) implementations

GREAT NEWS... right?

How good is good enough?

- Broad Coverage using Wikipedia:
  - 100,000s of entities
  - Close to 0.90 F1 in Wikilinks EL dataset
  - Much quicker than other (academic) implementations

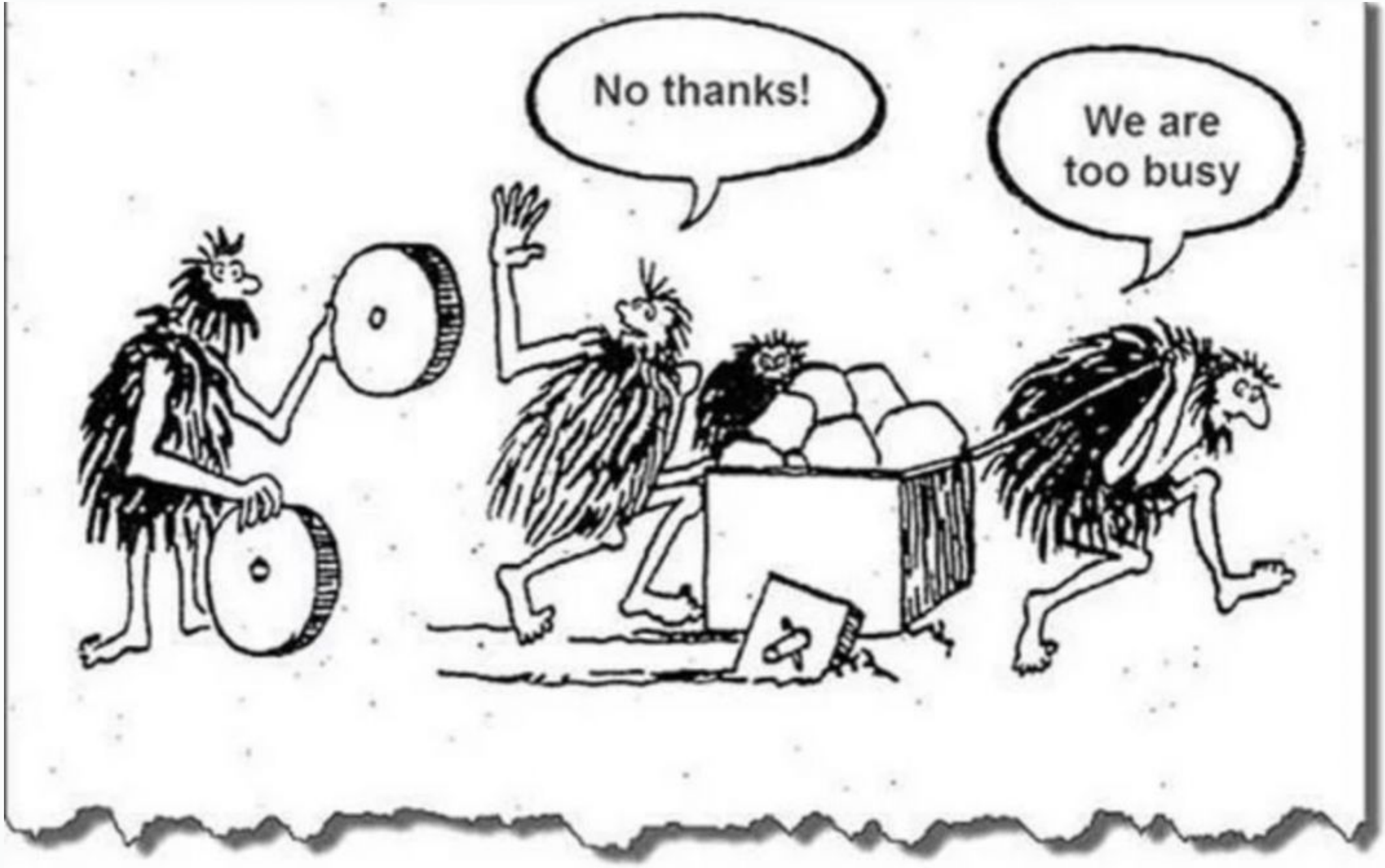
- **0.90 F1 is not useful and its variance is a problem**. Quality needs to be close to R:99/P90 (they of course ask for 100/100)
  - Supervised learning for client related entities (10K)
  - 0.98 avg F1 (with min. of 0.95) in internal datasets
  - Active learning, in-house labelling tool, quality estimation...

With respect to what/whom?

- Document based sentiment analysis has limited value
- We use Entity-based sentiment analysis
- What we will move towards is Reputation Polarity / Stance Detection
  - *“Lidl has fired 15,000 people and is closing in Germany”*
    - Neutral sentence
    - Very Negative Reputational Polarity
    - Positive for Lidl competitors

- Product Alignment and Value
- Data
- Evaluation
- Research vs Development Balance
- Organisational Structure

PRODUCT ALIGNMENT



PRODUCT ALIGNMENT

- Research should always bring value to the organisation
  - Always asking “why” and thinking about value
- Pareto rule and constant iteration
  - Strong baselines or simple models might be enough
  - Constant Prioritisation and Slicing with (many) competing lines of work
  - Different life-cycles from development and research
- Human + Algorithm collaboration is key
  - Talking to clients directly
  - Only build a ML system if needed. Rules are great for some problems

- In academia, focus on models
  - Given public datasets, how can I significantly improve quality?
  - Data is (usually) static and immutable
    - Leading to overfitting over years
  - Collections tend to be over-simplistic and/or “too clean”

- In industry, data is at the center
  - You can buy it, find it or create it
  - It changes over time (data and topic shift)
  - Bias in collecting labels
  - It is noisy (e.g., badly parsed articles)

- Evaluation is complicated, even more in industry
  - Users tend to ask for 100% accuracy

- What to measure?
  - Component vs User vs System based evaluation
  - What metric to use? F1? F0.5? P/R? ...
  - Some mistakes are worse than others
  - Quality just one aspect: model explainability, efficiency, consistency.

- <u>Evaluations to be linked to user value (even as proxy)</u>

EVALUATION (CONT)

- How to aggregate metrics
  - The academic community tends to show averages
  - Threshold quality more important than average in many cases
  - *All clients are equal but some are more equal than others*

- How to run evaluations
  - Unbalance problems
  - Biases in data collection
  - Labelling and evaluation steps related and dependent
  - Data and its distribution changes all the time:
  - Post-deployment monitoring and evaluation

RESEARCH AND DEVELOPMENT BALANCE

Source https://www.reddit.com/r/funny/comments/2a41i2/science_vs_engineering/

RESEARCH AND DEVELOPMENT BALANCE

- Flexibility and Adaptability
  - Researchers and Developers working together in production-code

- Pipeline Operations
  - Replicability and Reproducibility
  - Debuggable
  - Local vs Cloud Behaviour

- Scalability and Efficiency
  - How would it scale with 10x the volume? How quick is it?

RESEARCH AND DEVELOPMENT BALANCE

- Flexibility and Adaptability
  - Researchers and Developers working together in <u>production-code</u>
- Pipeline Operations
  - Replicability and Reproducibility
  - Debuggable
  - Local vs Cloud Behaviour
- Scalability and Efficiency
  - How would it scale with 10x volume? How quick is it?

What's your ML Test Score? A rubric for ML production systems

Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, D. Sculley
Google, Inc.
{ebreck, cais, nielsene, msalib, dsculley}@google.com

Abstract

Using machine learning in real-world production systems is complicated by a host of issues not found in small toy examples or even large offline research experiments. Testing and monitoring are key considerations for assessing the production-readiness of an ML system. But how much testing and monitoring is enough? We present an ML Test Score rubric based on a set of actionable tests to help quantify these issues.

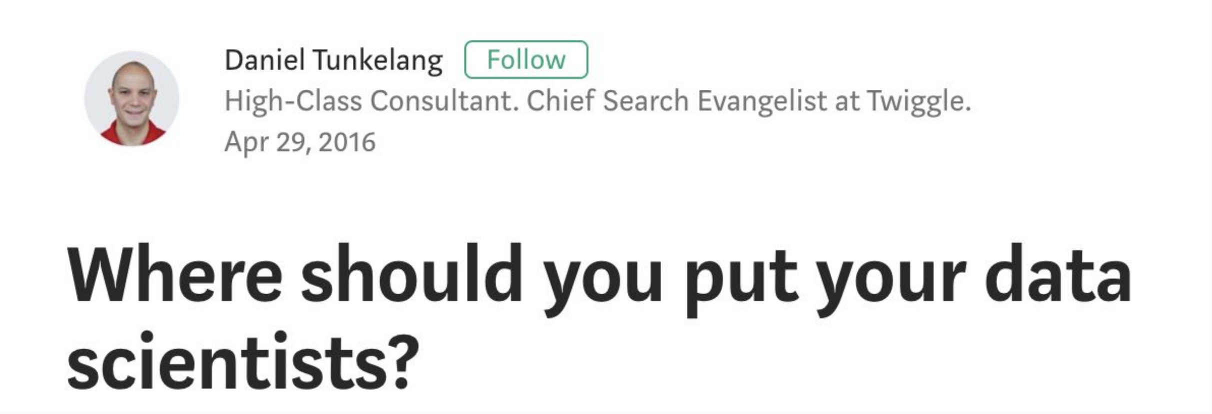
1 Introduction

Using machine learning in real-world software systems is complicated by a host of issues not found in small toy examples or even large offline experiments [1]. Based on years of prior experience using ML at Google, in systems such as ad click prediction [2] and the Sibyl ML platform [3], we have developed a set of best practices for using machine learning systems. We present these practices as a set of actionable tests, and offer a scoring system to measure how ready for production a given machine learning system is.

This rubric is intended to cover a range from a team just starting out with machine learning up through tests that even a well-established team may find difficult. We feel that presenting the entire list is useful to gauge a team's readiness to field a real-world ML system.

WHERE SHOULD YOU PUT YOUR RESEARCHERS?

- Research team (aka Research Lab)
- Integrated in product teams
- Embedded: “Rented” to teams



- Why collaborate with academia?
  - Influence focus of research
  - Hiring and retention
  - Company brand and reputation
  - Improve current/new services
  - Serendipity brainstorms

COLLABORATIONS

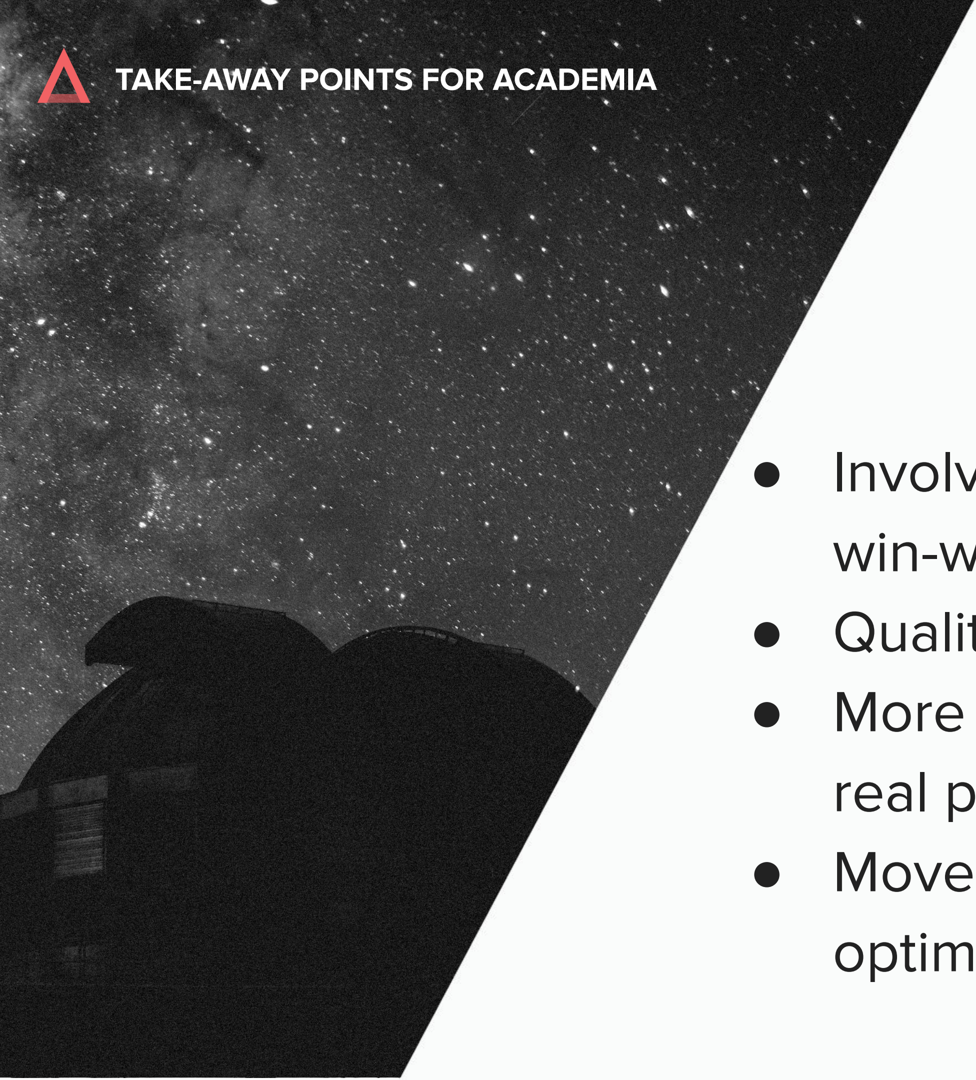
- MSc students
- Visiting researchers / interns
- Publications
- Grants
- Community involvement
- Industry Advisory Boards
- Invited speakers

ON THE SHOULDERS OF GIANTS

Dr. Daniel Gayo-Avello
Assoc. Professor
University of Oviedo

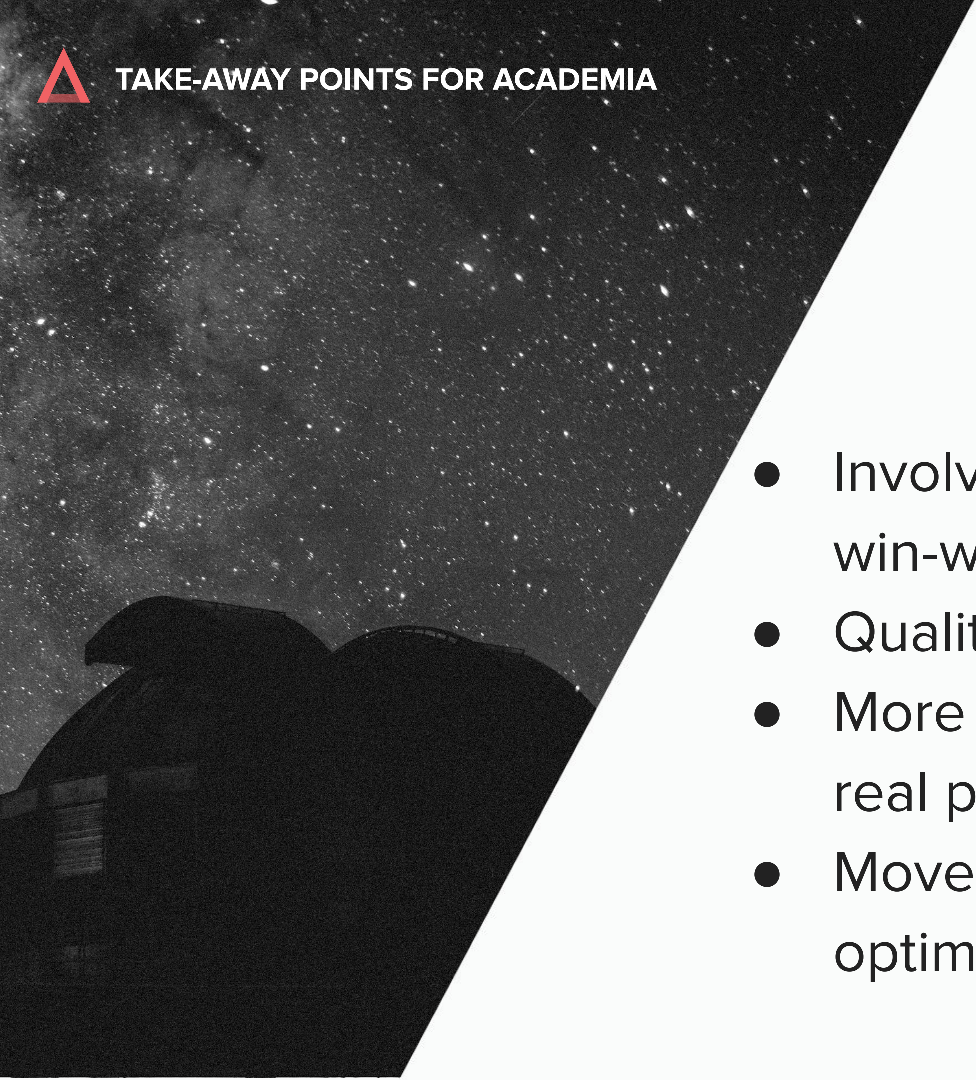
Dr. Thomas Roelleke
Lecturer
Queen Mary University of London

Dr. Udo Kruschwitz
Professor
University of Regensburg



TAKE-AWAY POINTS FOR ACADEMIA

- Involving industry in academia is a win-win and builds relationships!
- Quality is not everything
- More end-to-end evaluation focused on real problems to be solved
- Move away from static collections we optimise for decades



TAKE-AWAY POINTS FOR ACADEMIA

Questions?

- Involving industry in academia is a win-win and builds relationships!
- Quality is not everything
- More end-to-end evaluation focused on real problems to be solved
- Move away from static collections we optimise for decades