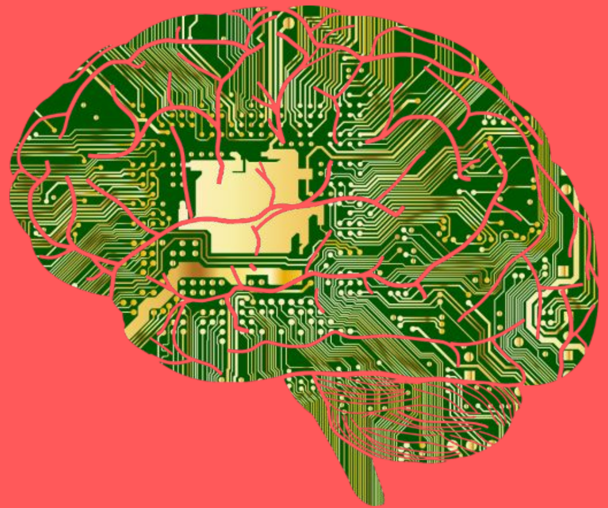
Data Science
vs
Engineering



Piero Cornice

SIGNAL

SIGNAL

MY BACKGROUND

Software Engineering, University of Bologna

> Delay-Tolerant Networks

Started working on **embedded software**...

SADEL Gruppo Almaviva — on-board information for Trenitalia

Provision communications — HD video coding/streaming over WiFi

TOMTOM — worked on the current sat-nav UI

...moved to **video streaming**

middleware for PS3, FireTV stick — amazon

on-demand TV, high-end STB — youview

...shifted to **backend** development...

...prototyped a **recommender system** using **NLP**...

...ended up doing working at SIGNAL

OUTLINE

1. WHAT WE DO
2. DATA SCIENTISTS and ENGINEERS

   (vs → ❤️)
3. WHAT WE DON'T DO

   (TO DO WHAT WE DO)
4. EPILOGUE

SIGNAL AI

WHAT WE DO

SIGNAL

WHAT WE DO

Empower decision making through media monitoring

Enable PR & Comms to **answer strategic questions**

AI = Augmented Intelligence

SIGNAL

WHAT WE DO

Founded by

David Benigson (CEO)
Miguel Martinez (Chief Data Scientist)

2013

Started life with three people in a garage

2020

150+ employees, three offices (London, NYC, HK)
20 engineers, 9 data science researcher (+ visiting researchers)

WHAT WE DO

We ingest **4+ million articles / day**

> currently 76+ billion articles indexed, 24+ TB of data

Real-time* **NLP** to extract information

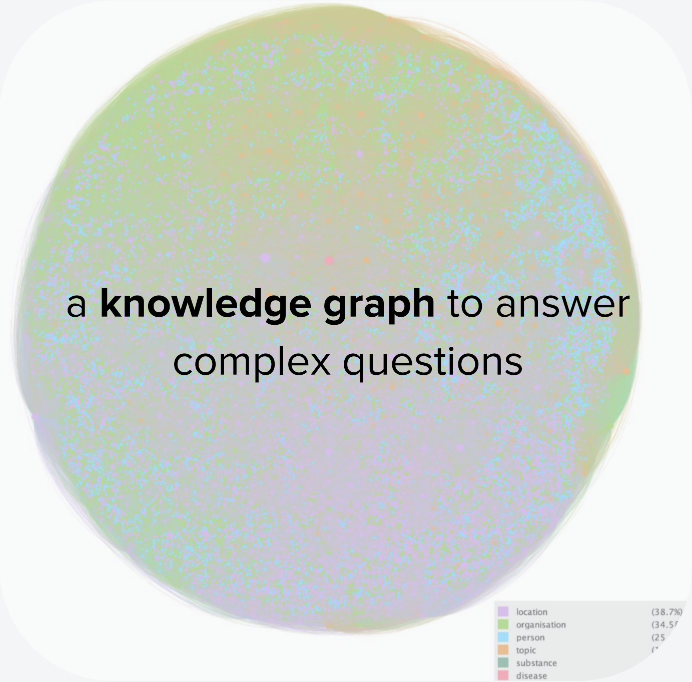
- > entities
- > topics
- > saliency
- > quotations
- > sentiment
- > …

Customers can search for articles in a **web application** and receive relevant **real-time alerts**.

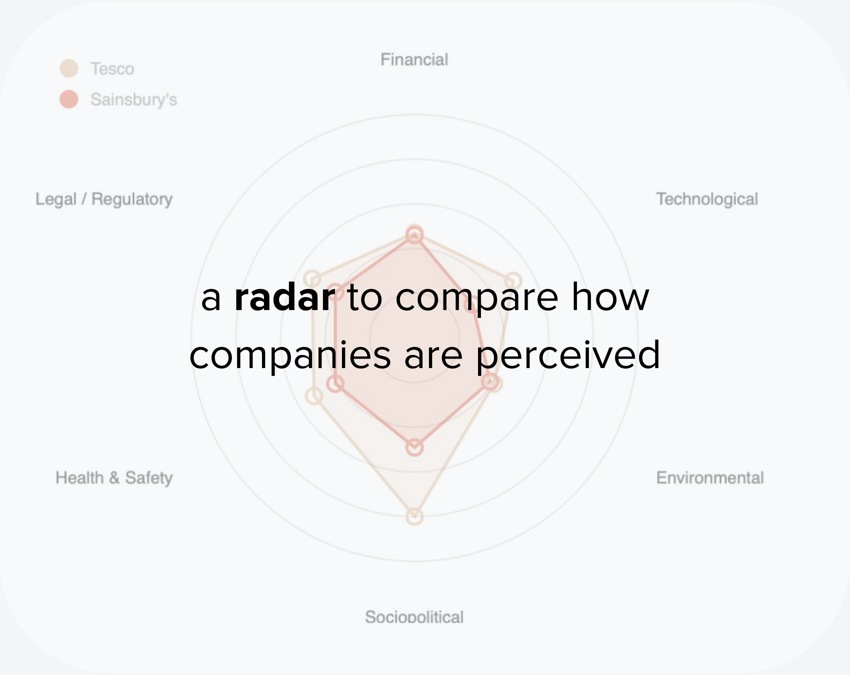
* *Real time = on average, pipeline lag is* ***< 1 minute*** *97% of the time.*

WHAT WE DO

Other things we’re working on…



a **knowledge graph** to answer complex questions



a **radar** to compare how companies are perceived

A selection of our 500+ world class **clients**.

Financial Services
BANK OF AMERICA
HSBC
BARCLAYS

Accounting
Grant Thornton
Deloitte.
EY

Legal
C/M/S/
SK STONE KING
CLIFFORD CHANCE

Agency
Hill+Knowlton Strategies
Edelman
MSL

Not for Profit
ICR The Institute of Cancer Research
DEC DISASTERS EMERGENCY COMMITTEE
THE WORLD BANK IBRD • IDA | WORLD BANK GROUP

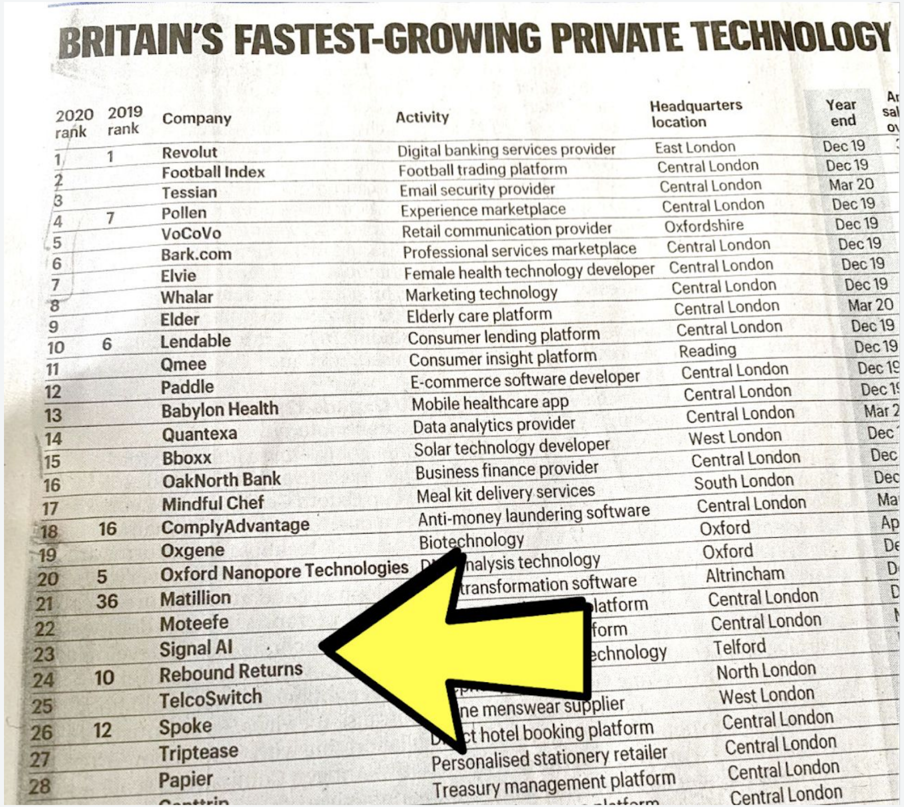
FMCG
L'ORÉAL
PRET ORGANIC COFFEE ★ NATURAL FOOD ★

Investment Houses
Invesco
LAZARD
Eaton Vance Investment Managers

Energy
centrica
Exxon
eDF

Tech
aws
Google
HUAWEI

WHAT WE DO



BRITAIN'S FASTEST-GROWING PRIVATE TECHNOLOGY

2020 rank	2019 rank	Company	Activity	Headquarters location	Year end
1	1	Revolut	Digital banking services provider	East London	Dec 19
2		Football Index	Football trading platform	Central London	Dec 19
3		Tessian	Email security provider	Central London	Mar 20
4	7	Pollen	Experience marketplace	Central London	Dec 19
5		VoCoVo	Retail communication provider	Oxfordshire	Dec 19
6		Bark.com	Professional services marketplace	Central London	Dec 19
7		Elvie	Female health technology developer	Central London	Dec 19
8		Whalar	Marketing technology	Central London	Dec 19
9		Elder	Elderly care platform	Central London	Mar 20
10	6	Lendable	Consumer lending platform	Central London	Dec 19
11		Qmee	Consumer insight platform	Reading	Dec 19
12		Paddle	E-commerce software developer	Central London	Dec 19
13		Babylon Health	Mobile healthcare app	Central London	Dec 1
14		Quantexa	Data analytics provider	Central London	Mar 2
15		Bboxx	Solar technology developer	West London	Dec
16		OakNorth Bank	Business finance provider	Central London	Dec
17		Mindful Chef	Meal kit delivery services	South London	Dec
18	16	ComplyAdvantage	Anti-money laundering software	Central London	Ma
19		Oxgene	Biotechnology	Oxford	Ap
20	5	Oxford Nanopore Technologies	D nalysis technology	Oxford	De
21	36	Matillion	transformation software	Altrincham	D
22		Moteefe	latform	Central London	D
23		Signal AI	form	Central London	N
24	10	Rebound Returns	echnology	Telford	
25		TelcoSwitch		North London	
26	12	Spoke	ne menswear supplier	West London	
27		Triptease	D ct hotel booking platform	Central London	
28		Papier	Personalised stationery retailer	Central London	
			Treasury management platform	Central London	
				Central London	

The Sunday Times Sage Tech Track 100

UK companies
with the fastest-growing sales
over their latest three years

September 2020

https://www.fasttrack.co.uk/league-tables/tech-track-100/

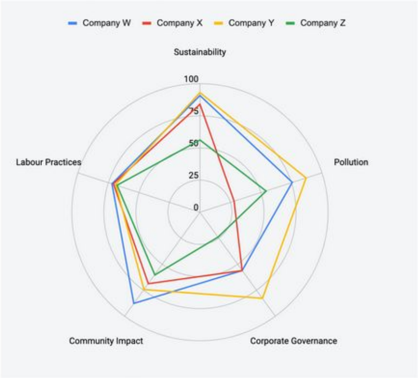
WHAT WE DO

NLP is used to automate quantification of a company's Environmental, Social and Governance (ESG) perception using the world's news

NLP can derive ESG perception scores by assessing the relationships and sentiments of products and companies with respect to client-specific ESG reputation pillars (e.g., environment, diversity, and more).

- Investors are increasingly demanding evidence of ESG performance.
- This approach uses NLP to tag millions of news articles daily to identify and understand relevant coverage using entity linking, saliency and topic classification.

	Company W	Company X	Company Y	Company Z
Sustainability	91	84	93	56
Pollution	76	28	87	54
Corporate Governance	56	56	83	24
Community Impact	88	69	74	60
Labour Practices	72	71	69	68



State of AI Report 2020

October 2020

https://www.stateof.ai/

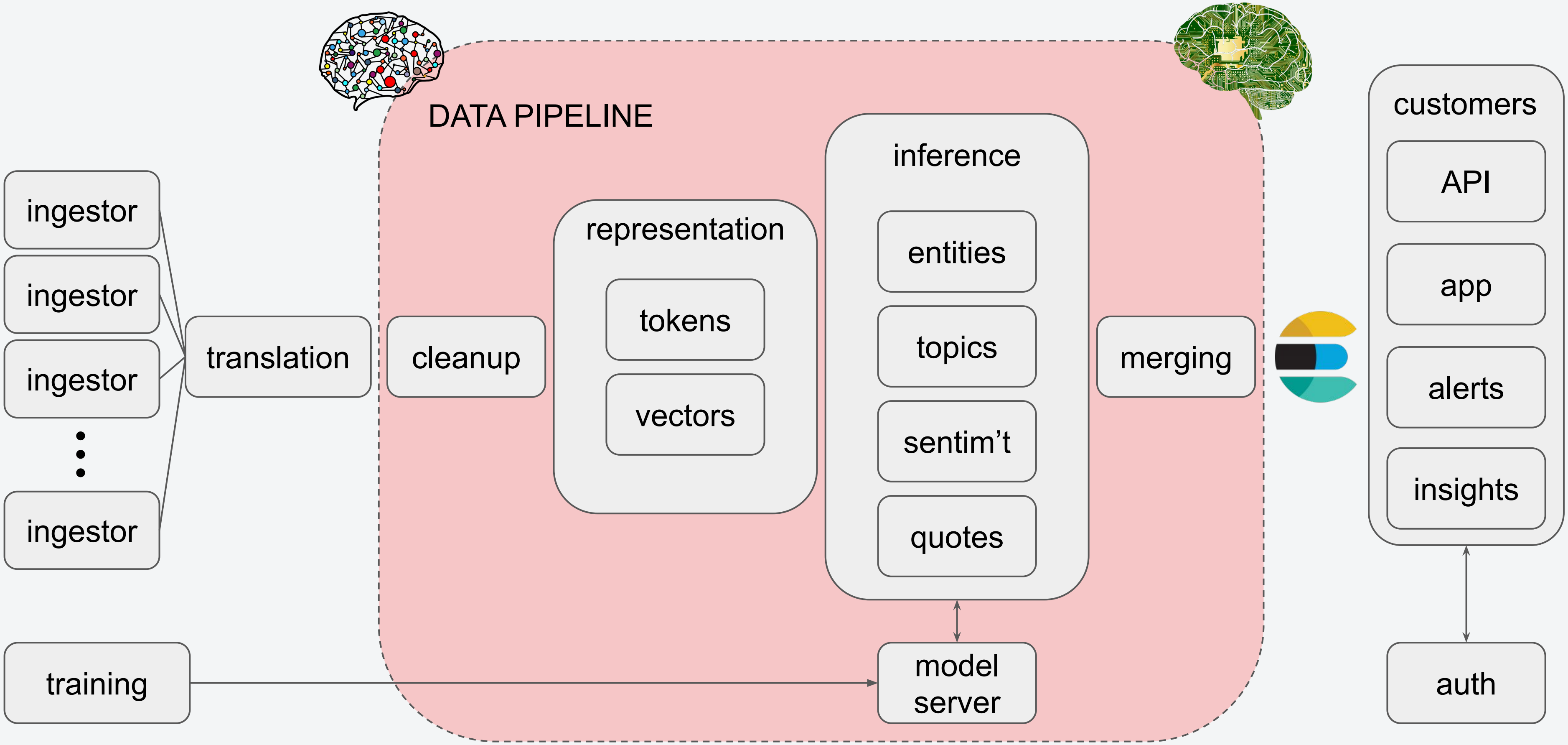
(slide 119)

SIGNAL AI

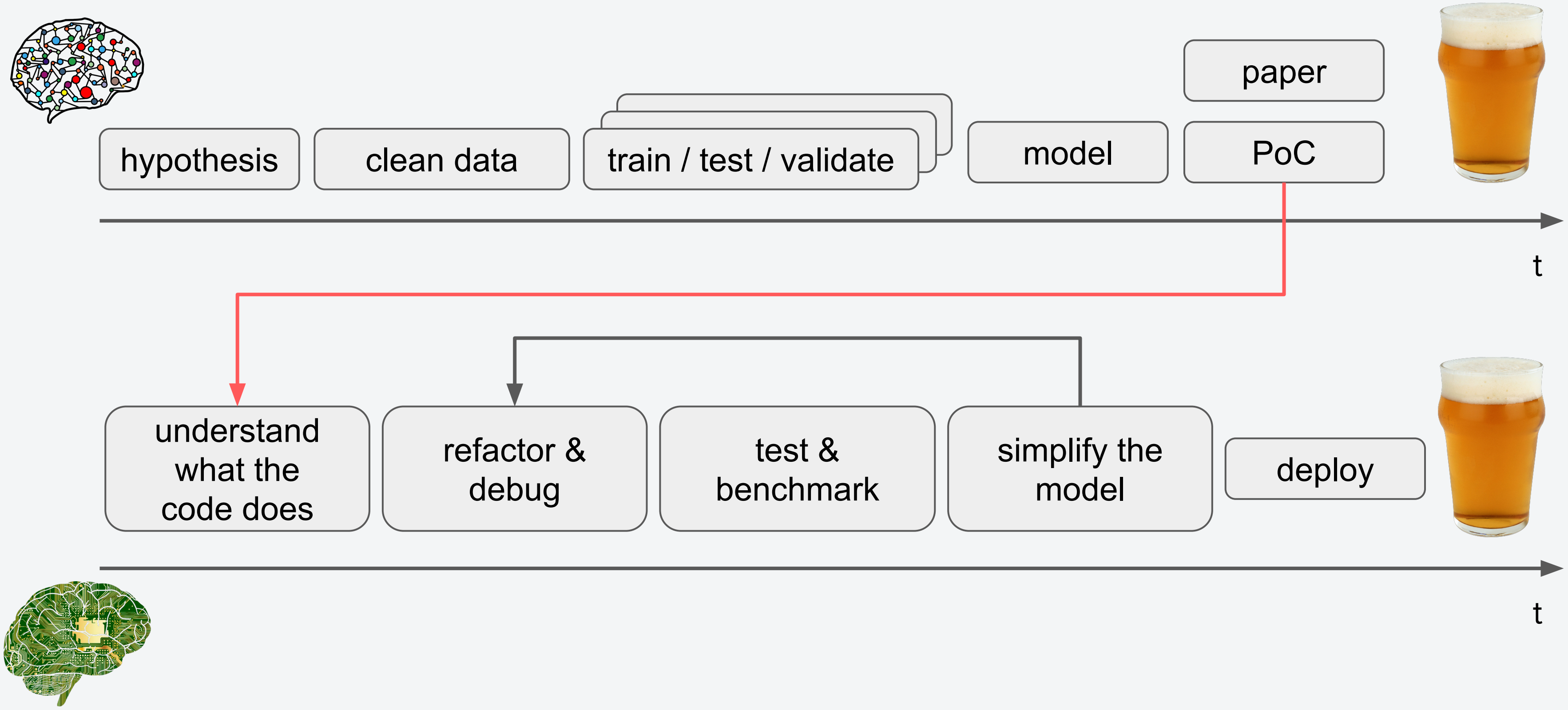
DATA SCIENTISTS
and
ENGINEERS
(vs ➔ ❤️)

SIGNAL

THE REAL-TIME PIPELINE, AS SEEN FROM SPACE



THE ~~WATER~~ BEER-FALL MODEL



RESEARCH vs ENGINEERING

Research work entails **uncertainty** and **long times**

> High cost if a line of research is not successful

> Data Scientists drink their beers while Engineers work

> > Engineers are not happy

Engineers **inherit** code that…

> has to be understood, optimised and tested

> once deployed, may give different results than what Data Scientists achieved

> > Data Scientists are not happy

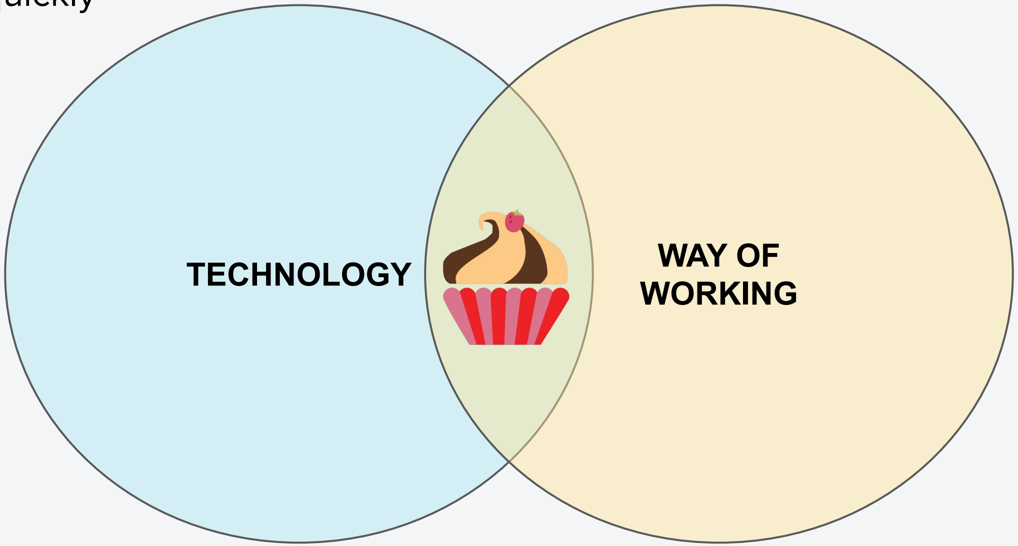
Long time to market

the **company** and its **customers** are not happy

CHASING THE SWEET SPOT

In order to stay competitive, at Signal we found a combination between **technological choices** and our **way of working**.

> **resolve the tension** between Data Science and Engineering
> **deliver value** quickly



> infrastructure
> languages
> off-the-shelf / in-house

testing strategy <
deployment strategy <
x-team coordination <

SIGNAL

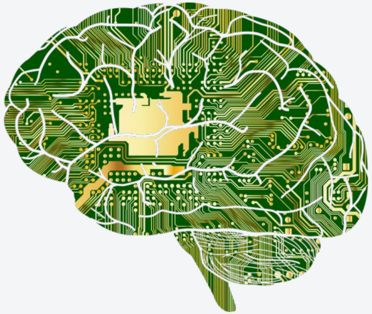
DATA SCIENTISTS vs ENGINEERS

Research implies **uncertainty** and **long times**

Engineers **inherit** code not ready for production

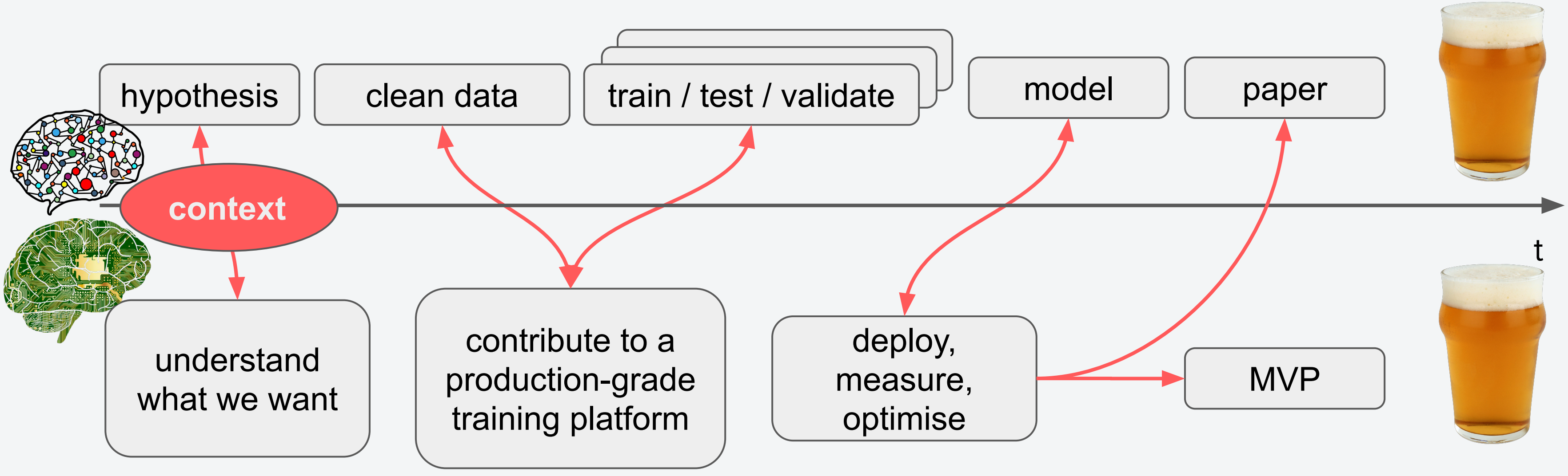
Does it have to be like this?

What if Data Scientists and Engineers could **pair from Day One**?



SIGNAL

DATA SCIENTISTS ❤️ ENGINEERS (A TALE OF TWO BEERS)



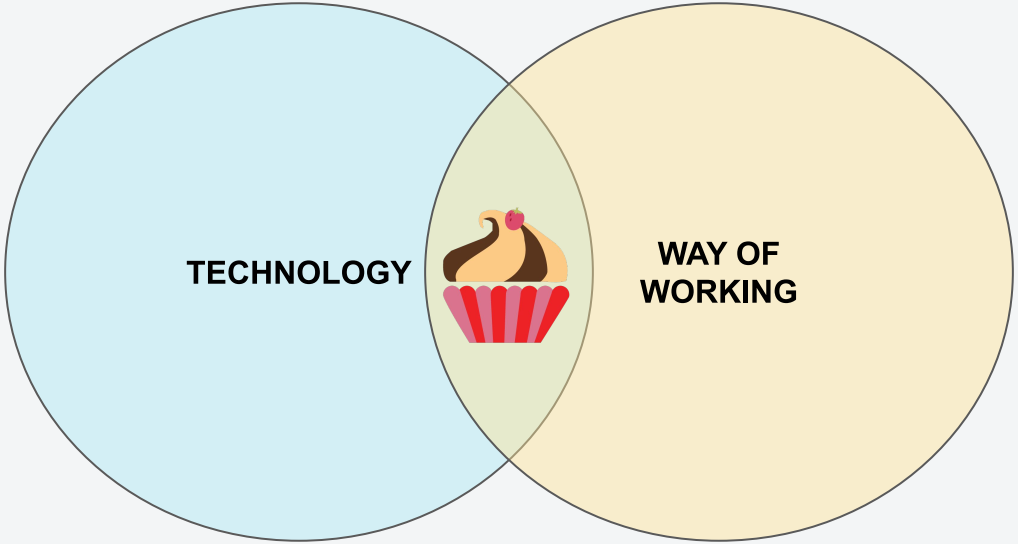
POC ➔ MVP (value)

WE MADE A MVP. NOW WHAT?

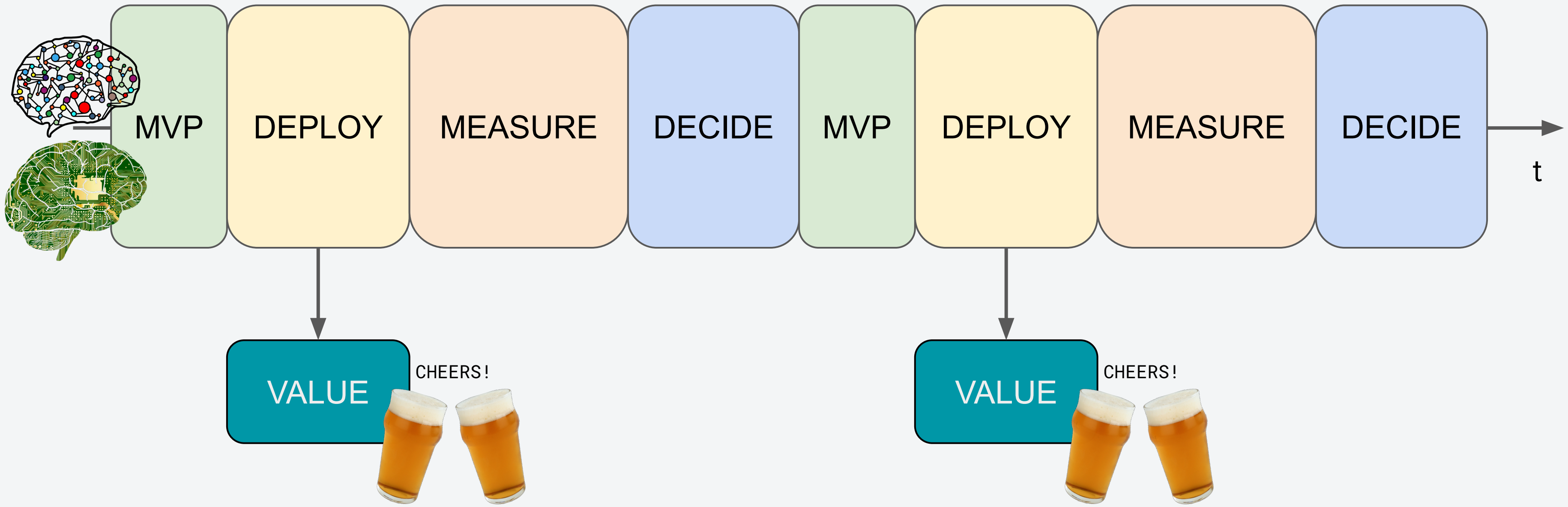
How to **evolve** a MVP

into a fully-fledged product?

WE DON'T



XP MODE = ON (BRING THE BEERS!)



SIGNAL

LESSONS LEARNED

Engineers have to accept failure

Data Scientists have to accept that quality without performance is a no-go

Shared understanding:

> Engineers know **what** Data Scientists want to achieve, and **how**

> Data Scientist know **what** the NFRs Engineers care about, and **adjust** accordingly

CHEERS!

ML code is tested for **quality** and **performance** as it's being developed

Teamwork! Sense of **ownership**, pride, and satisfaction

Engineers and Data Scientists can have beers together!

BEFORE WE MOVE ON...

QUESTIONS?

SIGNAL AI

WHAT WE DON'T DO
(TO DO WHAT WE DO)

SIGNAL

WHAT WE DON'T DO (TO DO WHAT WE DO)

Some common practices in the industry:

technology

> Kubernetes

> Kafka

> Imperative languages

WoW

> Automated tests

> Scrum

> Code reviews

We don't do

ANY

of those

SIGNAL

WHAT'S GOING ON???

BUT KUBERNETES IS AWESOME!

Introduces **complexity**

Docker

I. Turner-Trauring, “*Let’s use Kubernetes!” Now you have 8 problems*” (2020)
https://pythonspeed.com/articles/dont-need-kubernetes/

BUT KAFKA IS AWESOME!

Introduces **complexity**

SQS + **SNS**

We sometimes need **persistence**

> Being already locked-in to AWS, we use **Kinesis**.

FUNCTIONAL PROGRAMMING

Loved by **Engineers**

Clojure

small, pragmatic

G. Kim, "*Love Letter To Clojure (Part 1)*" (2019)
https://itrevolution.com/love-letter-to-clojure-part-1/

WHAT ABOUT PYTHON?

Loved by **Data Scientists**

GOOD: many libraries for Machine Learning

BAD: some libraries have side effects

UGLY: supports functional, yet it’s imperative

SIGNAL

LEAN PRINCIPLES

“Be lean, my friend”

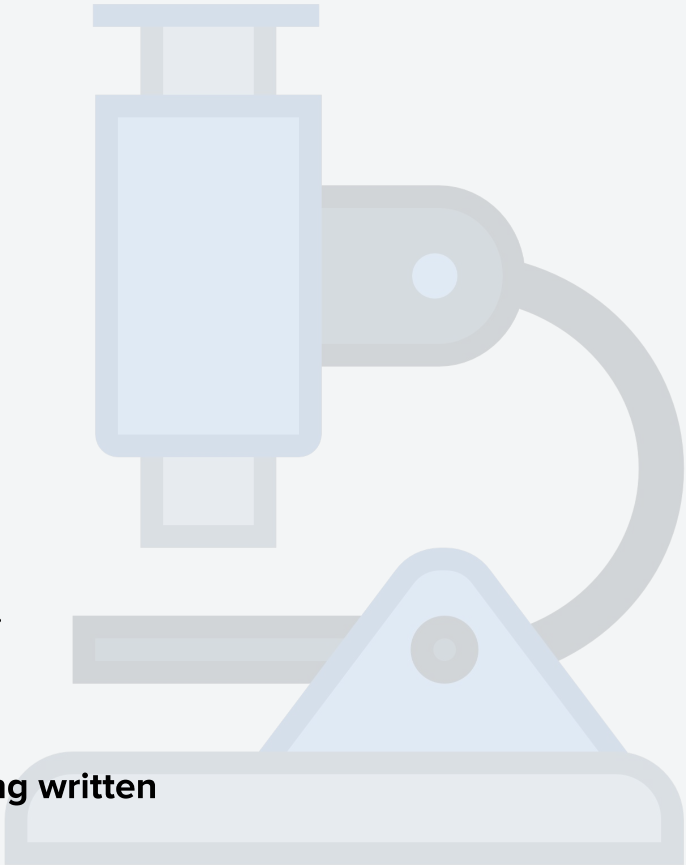
SCRUM: maximize the **team**'s ability to adapt

> roles (scrum master, …)
> tools (backlog)
> rituals (sprints, scrum of scrums, …)

LEAN: maximise the **flow** that generates value

> Eliminate **waste**
> Amplify **learning**
> **Decide** as late as possible
> **Deliver** as fast as possible
> **Empower** the team
> Build **integrity** in
> Optimize the **whole**

CAN YOU REVIEW MY CODE, PLEASE?



Code reviews are expensive.

> Do the reviewers have **enough context**?

> What if they **disagree** with the design?

Our recipe: Pair, pair, pair

> **Share context** and **agree** on design from the beginning.

> **Learn** from each other.

The code is reviewed as it is being written

TESTING IN PRODUCTION

The **real world** can only be found in **production**.

Our recipe: Frequent, incremental deployments

> Easy to **monitor**
> Easy to **rollback**
> Easy to **recover**

Strategic questions:

> What's the worst thing that can happen?
> Is it that bad?
> What can we do to mitigate it?

C. Sridharan, "*Testing in Production: the hard parts*" (2019)
https://medium.com/@copyconstruct/testing-in-production-the-hard-parts-3f06cefaf592

IT'S ALL DOWN!!

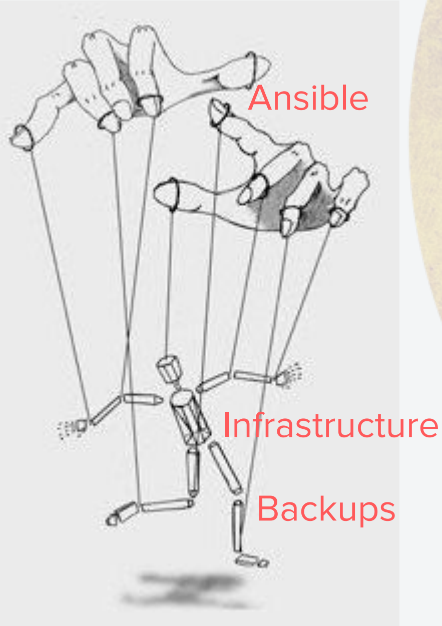
Disasters: they happen

If it hurts, do it more!

DISASTER RECOVERY GAME DAY

simulate that all the **production** environment is **compromised**

DISASTER RECOVERY GAME DAY



Full restoration of **primary** product functionality (with content)

> within 3 hours

Other **less-critical** components

> within 7.5 hours

Automation avoids wasting time on mechanical actions

> we use **Ansible** to control **Terraform** for infrastructure deployment

SIGNAL

SUCCESS STORY: SENTIMENT DETECTION

Problems

> Complex model (attention mechanism + Bayesian layer)

> Python's multiprocessing not efficient enough

Solution: **Ray*** + **improvements to the ML model**

Result: Computation time and costs went down ➜ we could release the product

CHEERS!

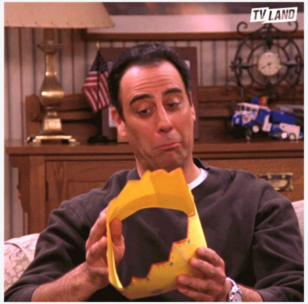
* *Ray is a Python framework for building and running distributed applications*

> *efficient use of CPU cores*

> *fine-grained tuning of resources (took some time to get right)*

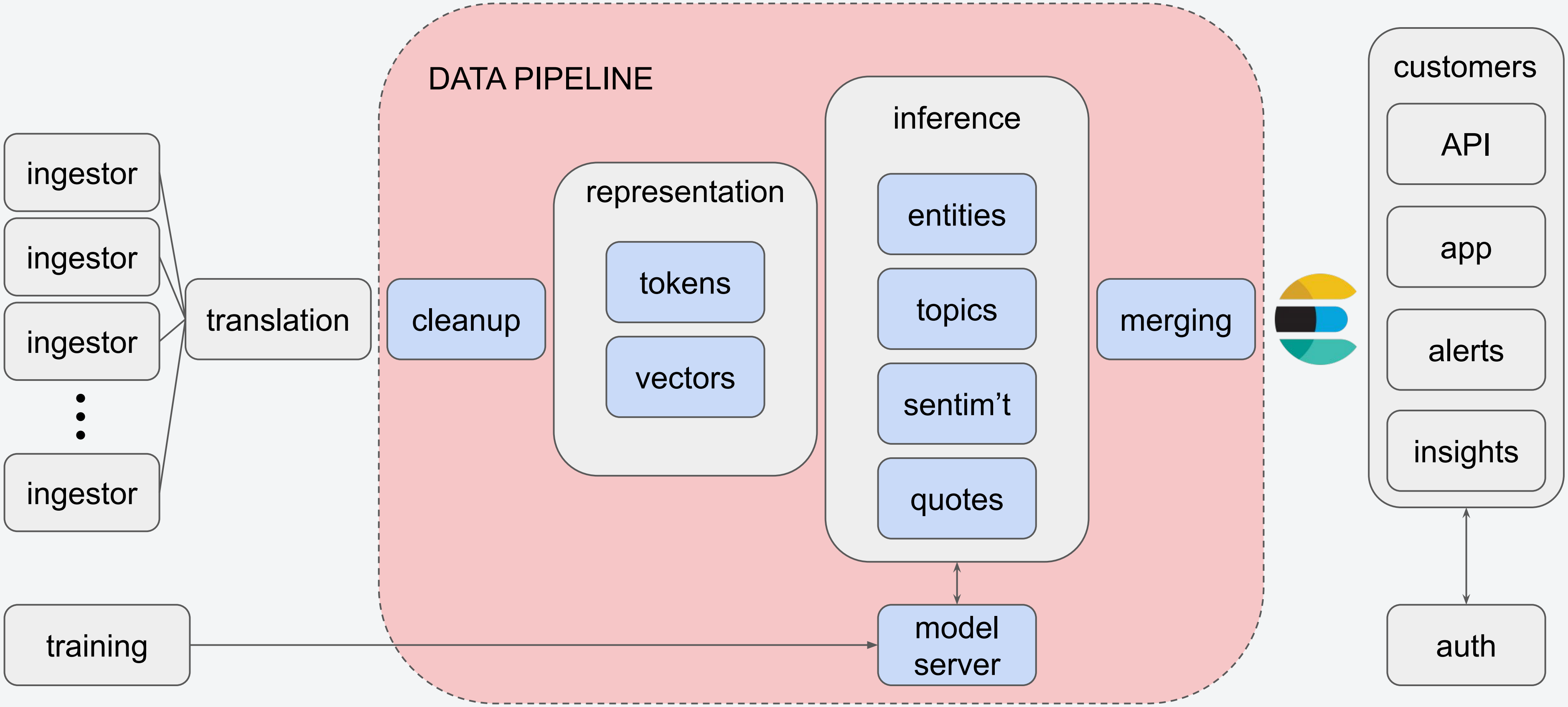
RAY https://github.com/ray-project/ray

SUCCESS STORY: SENTIMENT DETECTION

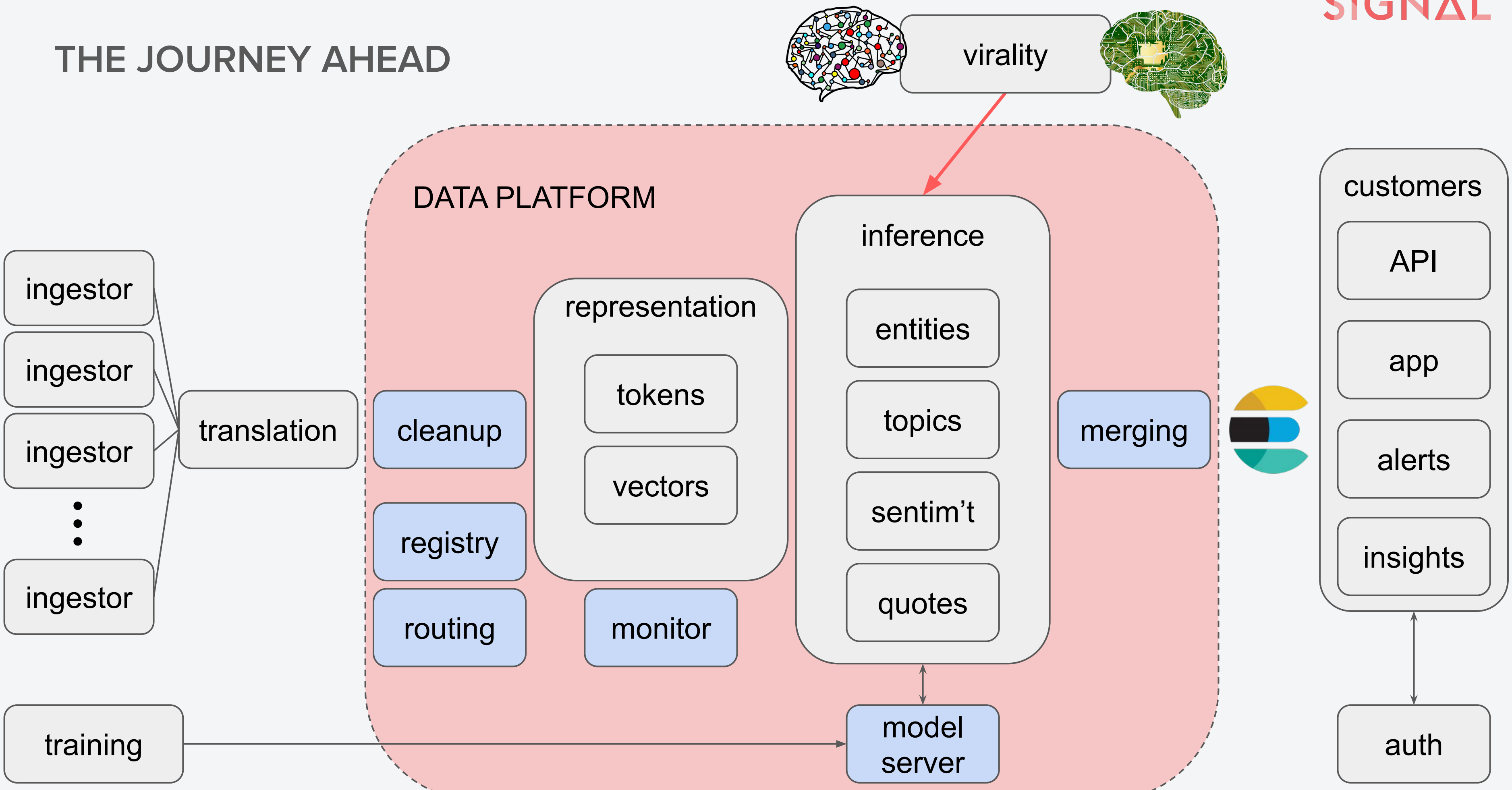


	Accuracy
SIGNAL	**65.45%**
G	53.77%
Azure	49.44%

THE JOURNEY AHEAD



THE JOURNEY AHEAD



SIGNAL

LESSONS LEARNED

Iterate fast by reducing waste

> adopt **simple technologies** when possible

> write code **together**, rather than asynchronous reviews

> no barriers to **continuous deployment**

LESSONS LEARNED

Robustness emerges from embracing uncertainty

> reduce surprises with **functional programming**

> rely on **real-time metrics** and alerts

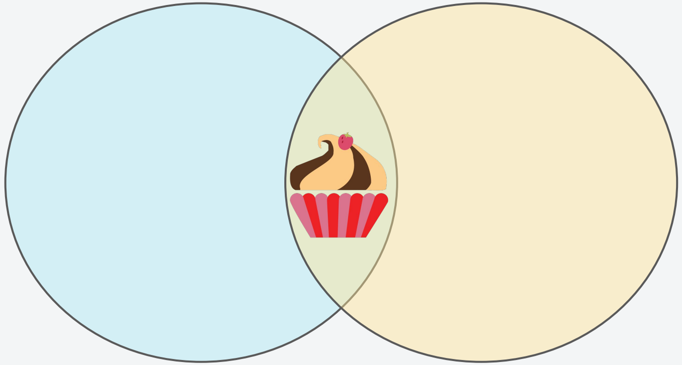
> **learn** from failure

SIGNAL AI

EPILOGUE

EPILOGUE

In order to stay competitive, at Signal we found a combination between **technological choices** and our **way of working**.



> **Pairing** between **Data Scientists** and **Engineers**
> **Challenge** common practices
> **Lean principles** all the way down
> **Embrace uncertainty** and failure to be robust

PS: The **recruitment process** is extremely important

> Don't focus on technologies, algorithms, data structures
> Focus on **solving business problems** and **hands-on pairing**

SIGNAL

SOME BOOKS THAT INSPIRED ME

> Optimising the flow

G. Kim, K. Behr, G. Spafford, “**The Phoenix Project**”, IT Revolution Press (2013)

> Integrating data science

G. Kim, “**The Unicorn Project**”, IT Revolution Press (2019)

> Empowering teams

D. Marquet, “**Turn the Ship Around!**”, Portfolio (2013)

> Designing for change

N. Ford, R. Parsons, P. Kua, “**Building Evolutionary Architectures**”, O’Reilly (2017)

SIGNAL

THANK YOU!

QUESTIONS?

@pierodactylus

www.linkedin.com/in/pierocornice