

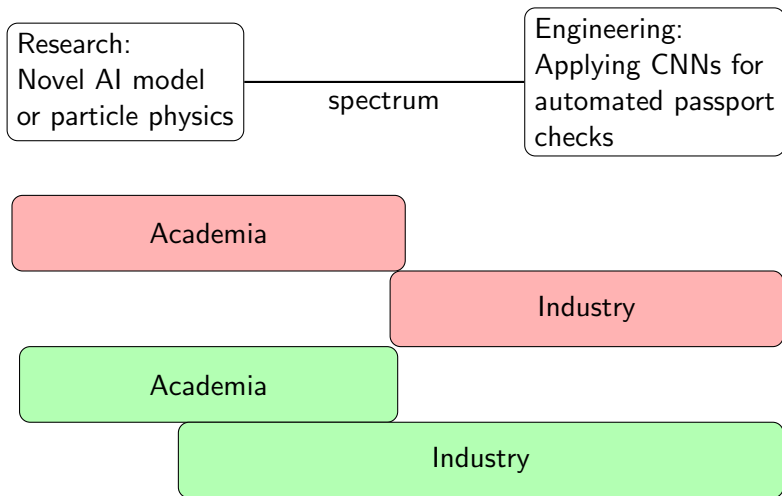
# Research in industry

Thomas Spriggs

Swansea University & AMPLYFI

October 31, 2022

# The spectrum of research



# The spectrum of research

Academia:

- Grants awarded
- Inspired by literature/benchmark
- Progress is rewarded
- Often feeds future engineers

In industry research is also done, but differently:

- Money must be made
- Work must be matched to a company problem
- Outputs are rewarded
- Problems influence academics\*
- Intellectual property restraints! :(

# My background

- 4<sup>th</sup> year PhD in theoretical physics and data science at Swansea University
  - ▶ Phase transitions in quantum field theories
  - ▶ Applications of machine learning in physics
- 2<sup>nd</sup> year machine learning engineer at AMPLYFI
  - ▶ Data extraction using natural language processing (NLP)
  - ▶ AI generated content creation

*My opinions do not necessarily reflect those of either organisation*

# AMPLIFYFI

Cardiff-based company with  $\sim 40$  employees.

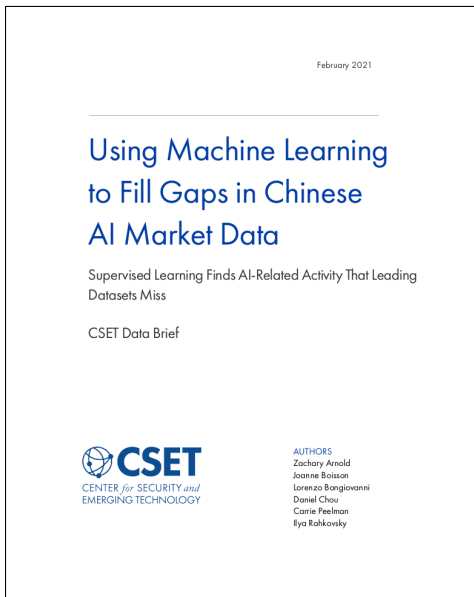
Created to automatically produce business insights:

- Search engine
- Insight alerting
- Business intelligence
- Bespoke AI-supported research projects

9 PhDs/PhD candidates from  $\sim 20$  engineers.

# Research projects

Bespoke research project.  
Collaboration with the  
Centre for Security and Emerging  
Technologies (CSET), Georgetown  
University, Washington DC.



## Research projects

CSET are a think-tank (research organisation) who advise the US Government on security and emerging technologies. They have previously research the state of the US AI market, and now wish to do the same for their largest competitor: China.

## Research projects

“Policymakers, market analysts, and academic researchers often use commercial databases to identify artificial intelligence-related companies and investments. High-quality commercial datasets have many advantages, but by design or by accident, they may overlook some AI-related companies ... We used machine learning (ML) models developed by Amplyfi Ltd. and Chinese-language web data to identify Chinese companies active in AI”

“We found that most of the companies identified by Amplyfi’s models were not labelled or described as AI-related in these databases.”

“using structured data alone ... will yield an incomplete picture of the Chinese AI landscape.”

[Excerpt from CSET, *Using Machine Learning to Fill Gaps in Chinese AI Market Data*]



# Research projects

As an overview of some challenges


- China tends to use Chinese...
- Mix of technical backgrounds: computer scientists, economists and political experts
- Huge volume of data
- Not easily available, heavily unstructured

# AI used in CSET project

Named Entity Resolution (NER) is the task of extracting entities like companies, people, or locations from text. For example:

**Amazon**, and **Amazon Web Services**, have some servers in **Seoul**, **South Korea**.

But one step further would be to link Amazon Web Services to Amazon, or to uniquely tie Amazon to the company, and not the rainforest. This is called **entity linking**.

 Natural Language Processing

## Entity Linking

180 papers with code • 23 benchmarks • 30 datasets

# AI used in CSET project

We have three models for this. It is an unsolved task so we cannot\* just use something pre-made.

- **N gram for boundaries**
- BERT
- **Graph embedding**

## AI used in CSET project

Often, companies, especially in China, have systematic naming schemes.

Amazon Web Services

So we can train a model to tell us the most likely locations for a boundary between *N-grams*.

An N-gram is an N word phrase, for example:

South Korea is a 2-gram

The United Kingdom of Great Britain and Northern Ireland is a 9-gram\*

# AI used in CSET project

**Amazon**, and **Amazon Web Services**, have some servers in **Seoul**, **South Korea**.

Amazon ? Amazon ? Web ? Services ? have ? some ? servers ? Seoul ? South ? Korea.

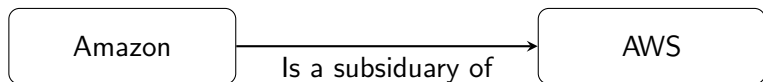
Amazon — Amazon Web Services — have — some — servers — Seoul — South Korea.

Long Short Term Memory (LSTM) architecture can be used here.

# AI used in CSET project

**Graphs/Networks** can store known relationships between entities.

- Use entity linking to create a graph
- Use a graph to help with entity linking
- Both!



Amazon — Amazon Web Services — have — some — servers — Seoul — South Korea.

# DeepInsight - an engineering project

DeepInsight is a product offered by AMPLFYI that allows companies to research at scales that humans cannot. At least 14,000,000 documents analysed by cutting edge AI models.

# DeepInsight - an engineering project

Project 1: How positively is an entity spoken about?

Lots of models that tell us the sentiment of a sentence, but not necessarily about the entity. Consider these two examples:

Apple are a great company, but their new phone is awful!

I hate big technology companies, but that new Apple phone is great.

But these can be very misleading for customers! So we have to be careful when we do this at scale.



## **TweetNLP: Cutting-Edge Natural Language Processing for Social Media**

**Jose Camacho-Collados<sup>1</sup> Kiamehr Rezaee<sup>1</sup> Talayeh Riahi<sup>1</sup> Asahi Ushio<sup>1</sup>  
Daniel Loureiro<sup>1</sup> Dimosthenis Antypas<sup>1</sup> Joanne Boisson<sup>1,6</sup> Luis Espinosa-Anke<sup>1,6</sup>  
Fangyu Liu<sup>2</sup> Eugenio Martínez-Cámara<sup>3</sup> Gonzalo Medina<sup>3</sup>  
Thomas Buhrmann<sup>4</sup> Leonardo Neves<sup>5</sup> Francesco Barbieri<sup>5</sup>**

<sup>1</sup>Cardiff NLP, Cardiff University, UK <sup>2</sup>LTL, University of Cambridge, UK

<sup>3</sup>DaSCI, University of Granada, Spain <sup>4</sup>Graphext, Spain <sup>5</sup>Snap Inc., USA <sup>6</sup>AMPLIFYFI, UK  
cardiffnlp.contact@gmail.com

# DeepInsight - an engineering project

Project 2: Technology phrase extraction.

What would do really well in a benchmark, could be very unhelpful for a customer. If we extract technologies that a company are involved in:

- Internet
- Data
- Digital service

These are all really common words, but of no actual insight. Imagine paying for research and finding that Amazon Web Services are involved in 'the internet'.

But these would all be considered positives in a 'standard benchmark'.

# General closing statements

So, in general:

- Research is done by industry
- Be adaptive, one worker covers many roles
- Goals strived for by academia and industry are not always aligned
- Novelty is inconvenient for industry
- Intellectual property concerns are prohibited

But the pay is better.

End

Cheers