Data Science Academy

Academi Gwyddor Data

Exploring the Potentials of Data-Driven Discovery

DR OKTAY KARAKUS

BSC (HONS), MSC, PHD, FHEA, MIEEE

***LECTURER** IN COMPUTER SCIENCE AND INFORMATICS*

***DEPUTY DIRECTOR** OF DATA SCIENCE ACADEMY*

Outline

Part 1

- Introducing the Data Science Academy (DSA)

Part 2

- Project Showcases

Part 1.
Introducing the DSA

Data Science Academy (DSA)

Who we are?

- train highly-skilled and employable graduates
- run a range of postgraduate courses in multidisciplinary fields
- (DSA) is run by Cardiff University's
  - School of Computer Science and Informatics
  - in partnership with the School of Mathematics
  - with an industry advisory board

DSA

Mission

> We want to ensure that Wales produces highly-skilled and employable graduates in some of the fastest growing and in-demand areas, from data science and artificial intelligence to cybersecurity.

The DSA Model

- DSA offers several taught MSc programs in:
  - Artificial Intelligence
  - Data Science
  - Cybersecurity
  - Journalism
  - Business
  - Environmental Sustainability
- An interdisciplinary approach to teaching is used
- Necessary data science skills are combined with practical discipline specific skills
- Strive to ensure our teaching is informed by industry.

The DSA Model

MSc Programmes

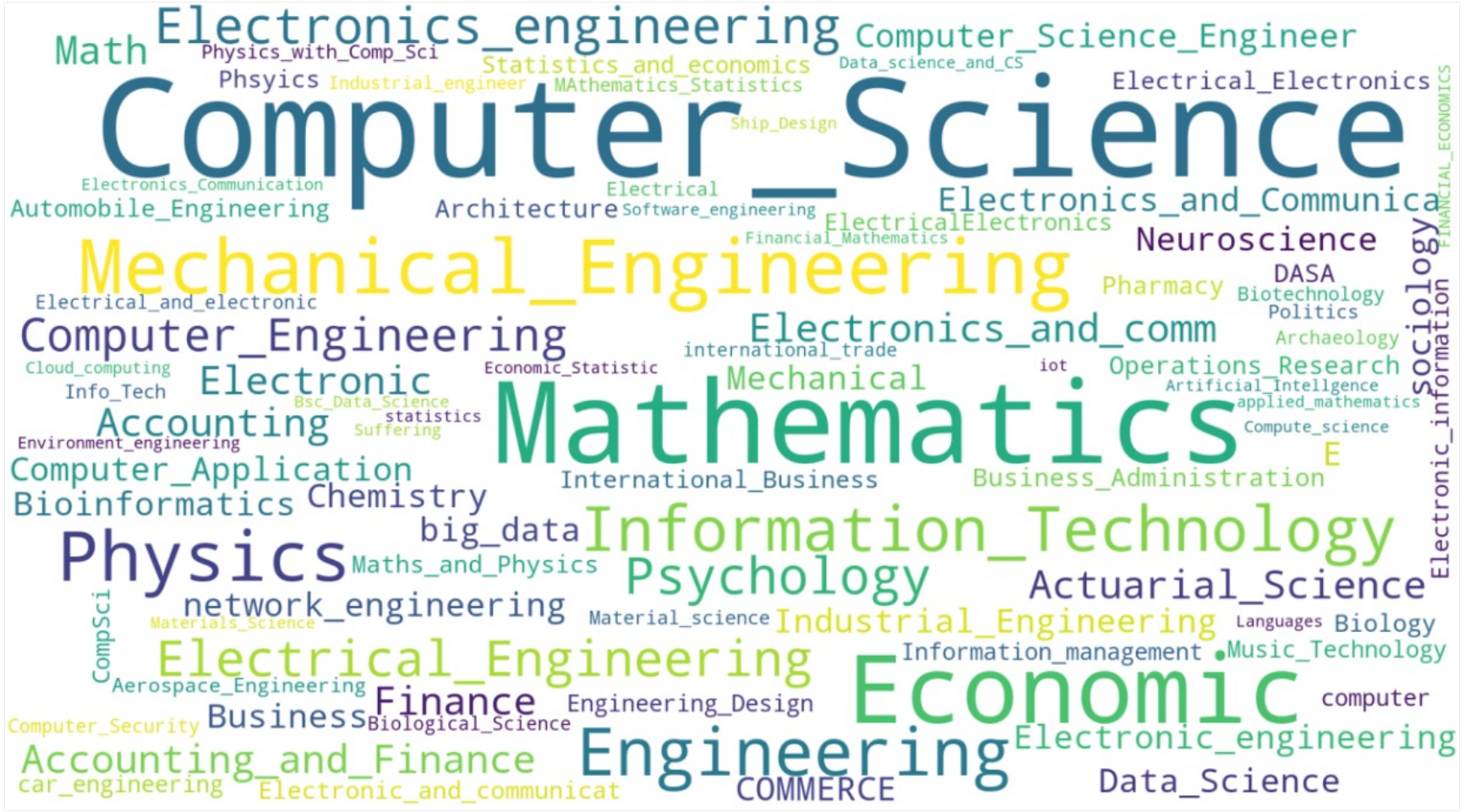
- MSc Artificial Intelligence
- MSc Data Science and Analytics
- MSc Cybersecurity
- MSc Cybersecurity & Technology (PwC)
- MSc Data Analytics for Government (ONS)
- MSc Computational Data Journalism
- MBA in AI
- MSc in NLP
- MSc in Data Science for Environmental Sustainability (**Sept 2025**)

The DSA Model

Industry & public sector engagement

- Student projects with real data/tasks
- Guest lectures/seminars or skills sessions
- Recruitment
- Internships
- Research Projects
- Apply for Funding with us (KTP, Innovate UK, etc.)

Diverse Student Skillset



DSA Industry Project

Requirements & Process

REQUIREMENTS

- Academic level:
  - Align with our curriculum and students' skill set
- Timeline:
  - Feasible to complete within 10 - 12 weeks.
- Supervision:
  - A supervisor from the organisation who must commit time to meet the student
- Resource availability:
  - Data (and any equipment) should be available from the start of the project

TIMEFRAME

- Early March
  - Submit project ideas to the DSA
- March-April
  - Selection of projects
- April-May
  - Notifications to organisations of projects chosen
- Early July
  - Projects start
- Late September
  - Projects finish

Some of Our Partners (Past & Current)

Admiral

Office for National Statistics

GENLETICS

Principality Building Society
Cymdeithas Adeiladu

empirisys

WRU

Dŵr Cymru Welsh Water

Llywodraeth Cymru
Welsh Government

BT

How to propose a project

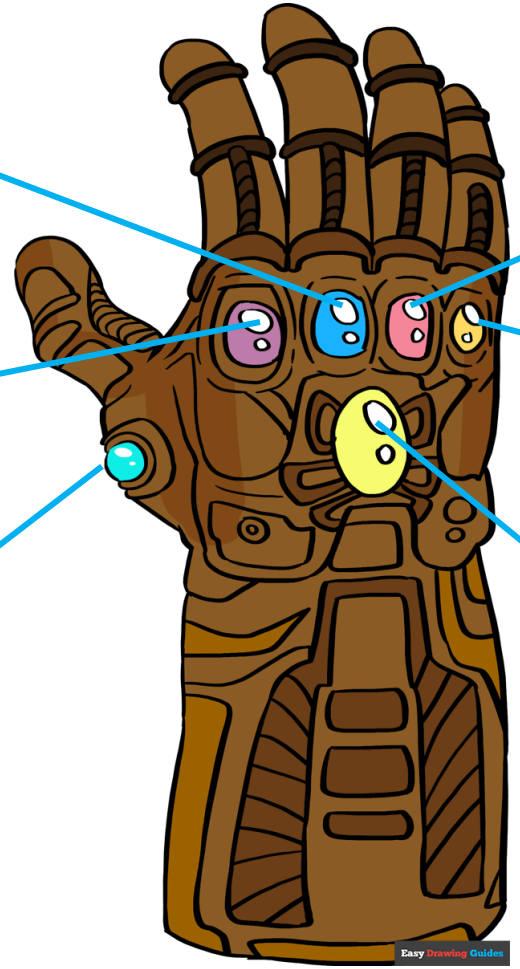
- Email us anytime!
  - dsa@cardiff.ac.uk
- ATTEND THE EVENT: Industry Projects and Engagement Session
  - Late January – Early February every year
- Submit a Data Science Project! (A submission portal will be active prior to deadlines)

PART 2.
PROJECT SHOWCASES

Why Example Projects for Data-driven Discovery?

- To show the power of
  - Data
  - Computing tools
  - Expert knowledge
  - Diverse student skills
  - Novel Technology
  - Academic Experience

Steps of Data-driven Discovery



Academic Experience

Expert knowledge

Novel Technology

Computing tools

Diverse personal skills

Data

2.1 Social Media Sentiment Analysis

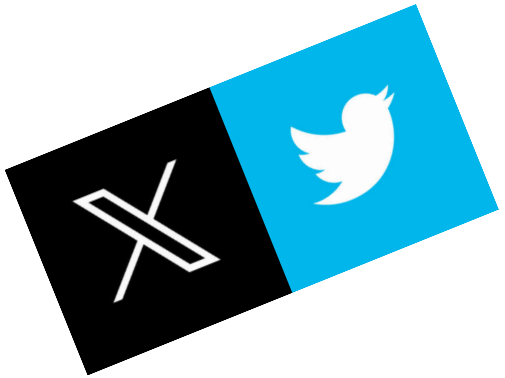
Project 1: Social Media Sentiment Analysis

QUESTION?

- **What can we learn from people's social media posts?**

ANSWER

- Consumer Preferences and Feedback
- Market Trends and Public Opinion
- Crisis Management and Public Relations
- Health Monitoring and Public Health Trends
- …



Social Media Sentiment Analysis

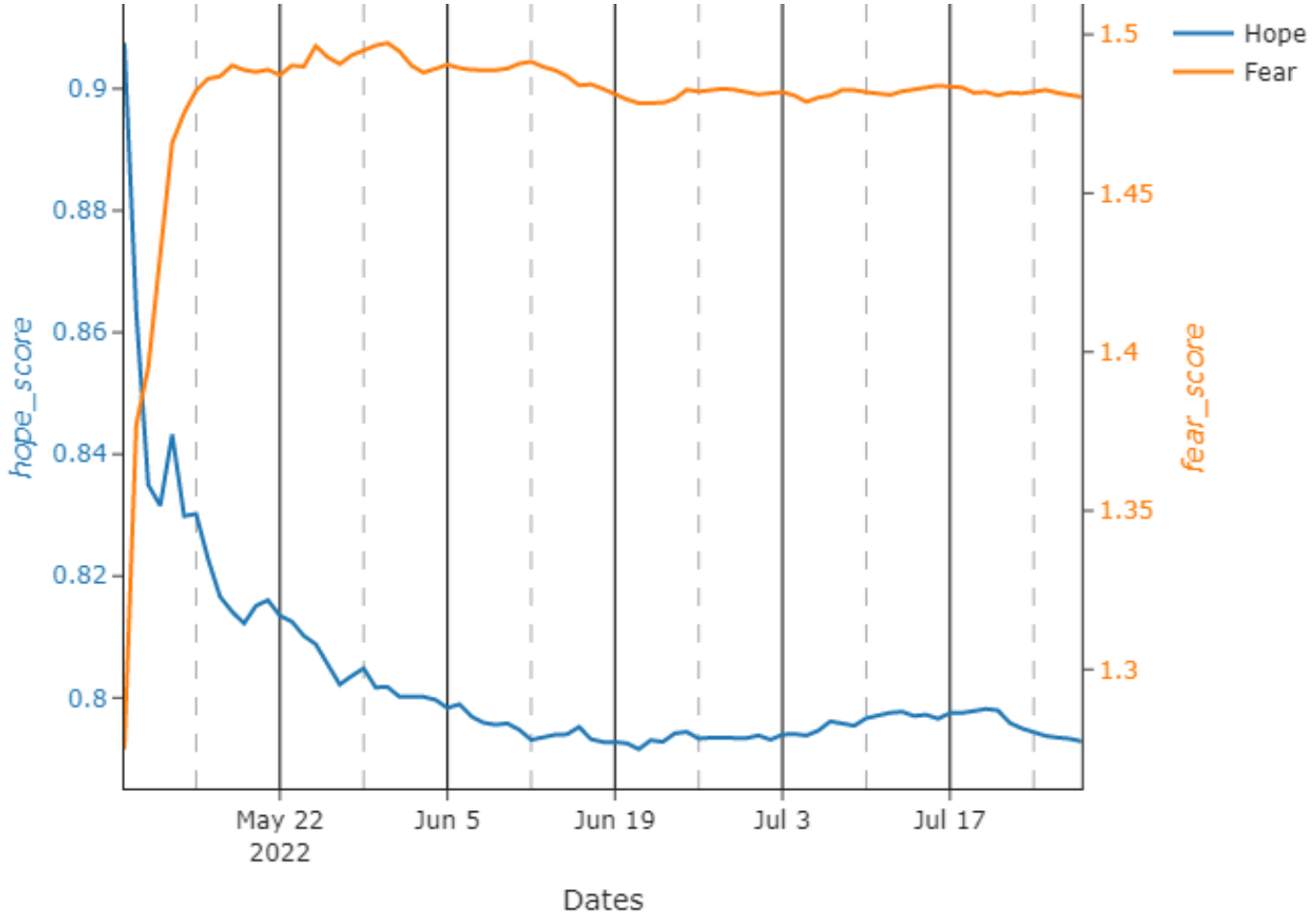
Our Question?

- **Can we measure “Hope” and “Fear” from social media posts?**

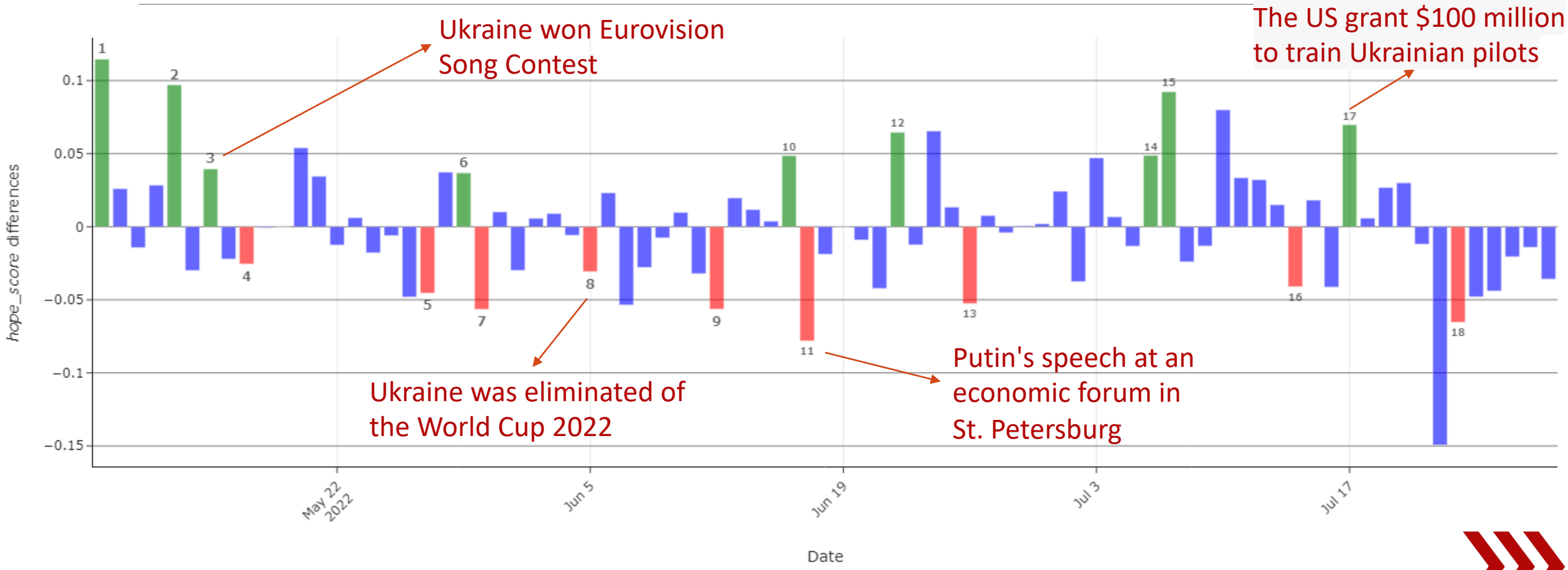
Our Answer:

- Case Study: Russo-Ukrainian Conflict
- Platform: Reddit
- Time frame: the first 3 months of the conflict
  - 10th May 2022 – 28th July 2022
- Around 1.2 Million posts

Social Media Sentiment Analysis



Social Media Sentiment Analysis



2.2 Football Analytics

Football Analytics

QUESTION?

- **Can football event data provide more than just statistics?**

ANSWER

- Player Performance Analysis
- Tactical Analysis
- Injury Prevention and Player Health
- Fan Engagement and Experience
- ...

Football Analytics

Story of Neal Maupay & Brentford

- Brentford promoted football analytics to find hidden gems!

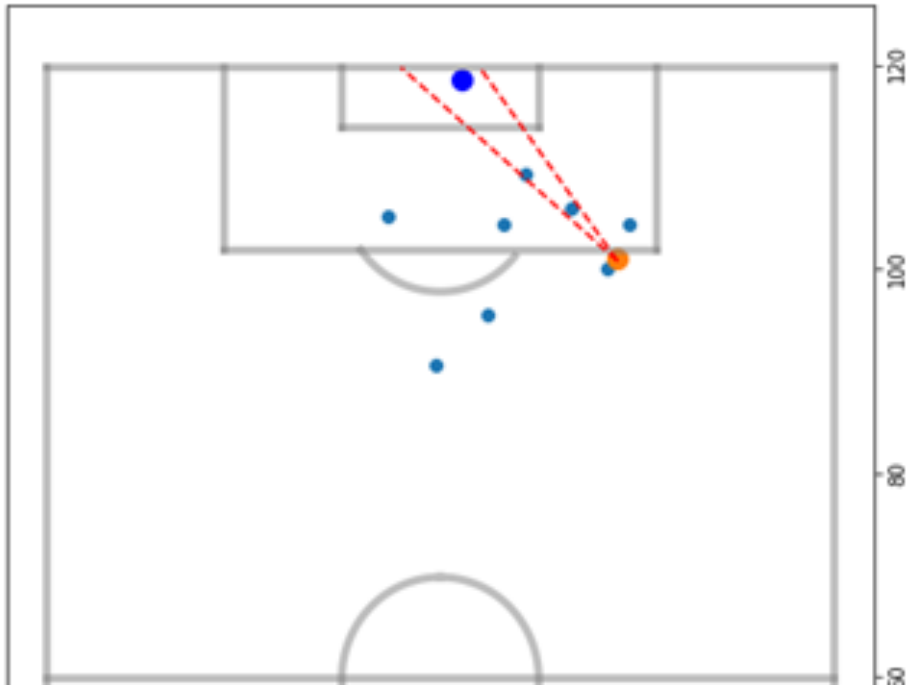
- 2017 → Brentford buys Neal Maupay
  - ~1.5 M GBP
  - Extra ordinary Expected Goals (xG) values
- 2017-2019
  - Played 85 games scored 37 goals
- 2019
  - Sold to Brighton for 20 M GBP

Football Analytics

- xG: the probability of a shot to be scored!
- Goal/xG → indicator of clinical finishers (> 1)
- Industry xG models assumes all players have the same probability for a given chance!

Our Question?

- **Is player and/or position-based correction possible for the xG?**



Football Analytics

Our Question?

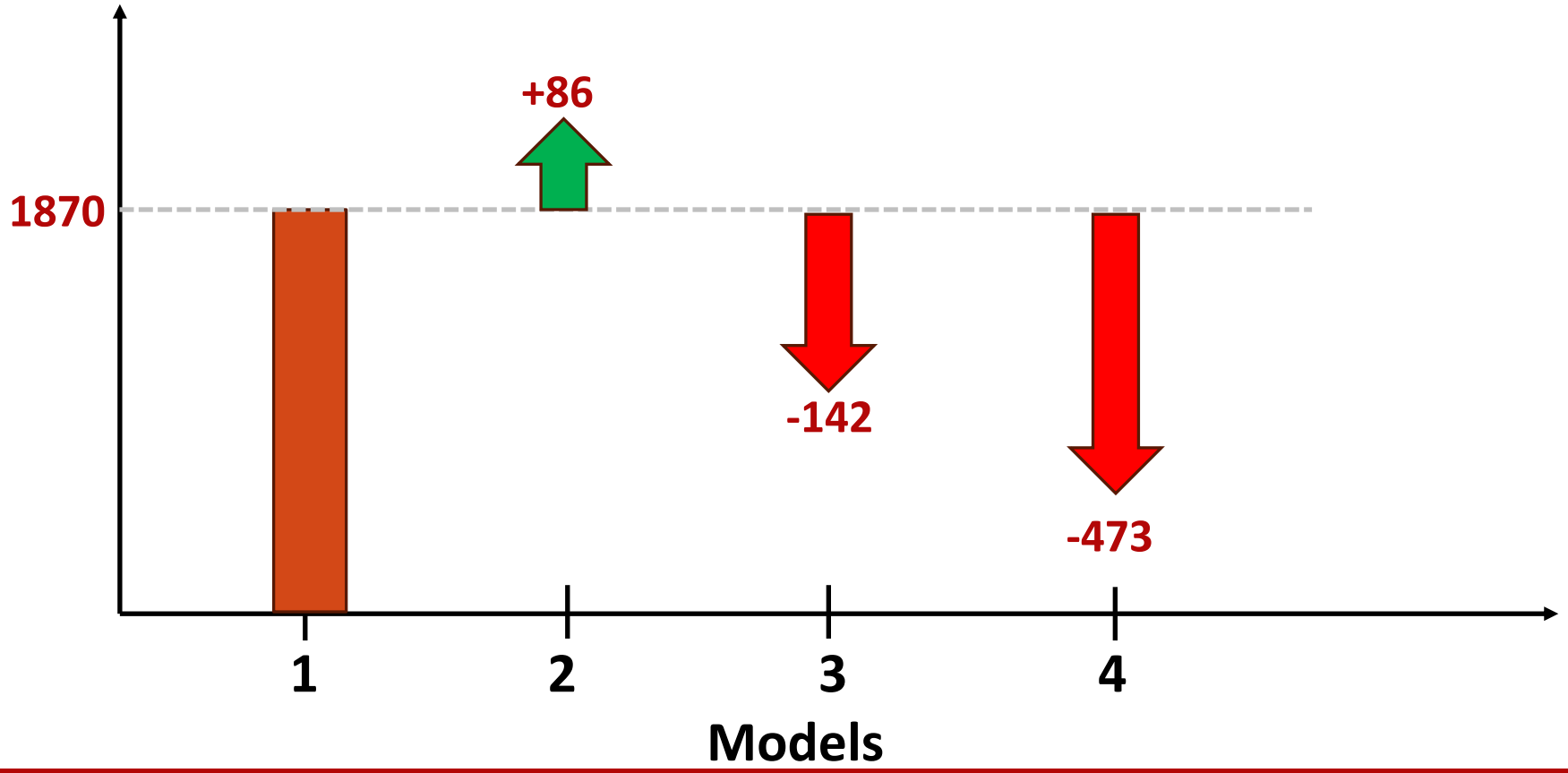
- **Is player and/or position-based correction possible for the xG?**

Our Answer:

- ~16K open play shots
- Lionel Messi as a test case!
  - Data sets with and without Messi
- Engineered 40 features!

Football Analytics

Positional Analysis



Model 1

- Trained /w **D+M+F**

Model 2

- Trained /w **F** only

Model 3

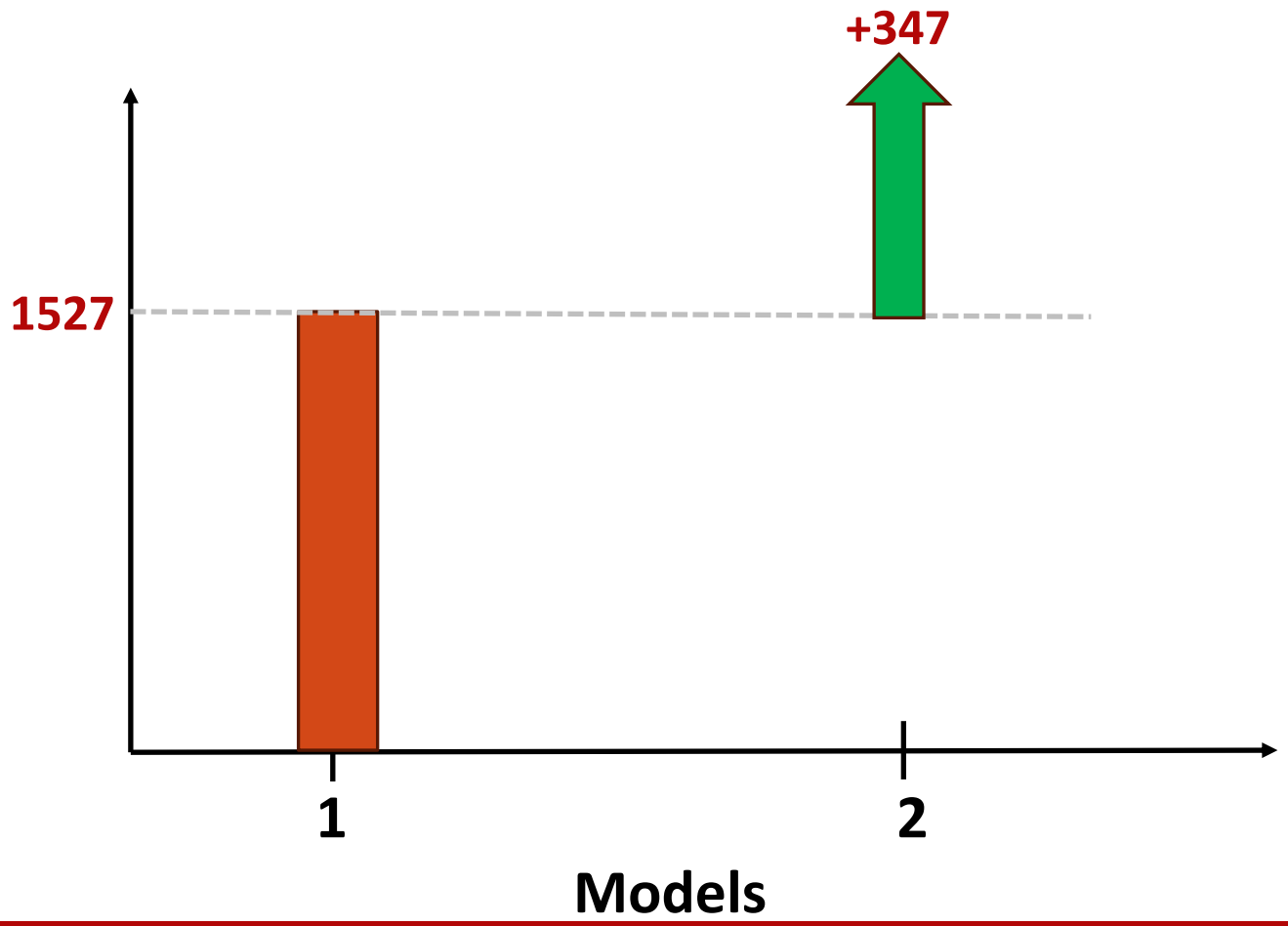
- Trained /w **M** only

Model 4

- Trained /w **D** only

Football Analytics

Player Correction



Model 1

- Trained /w original data set

Model 2

- Trained /w **Messi** only data

SPECIAL CASE

What happens if Lionel Messi shot the shoots Luis Suarez had?

*107.45 + **18.76** = 126.21*

*(Suarez + **Messi correction**)*

2.3 Marine Debris Monitoring

Marine Debris Monitoring

QUESTION?

- **Can satellite imagery help fight environmental problems?**

ANSWER

- Deforestation Monitoring and Prevention
- Climate Change and Carbon Emissions
- Disaster Management and Recovery
- Agricultural Management and Food Security
- …

Marine Debris Monitoring

Our Question?

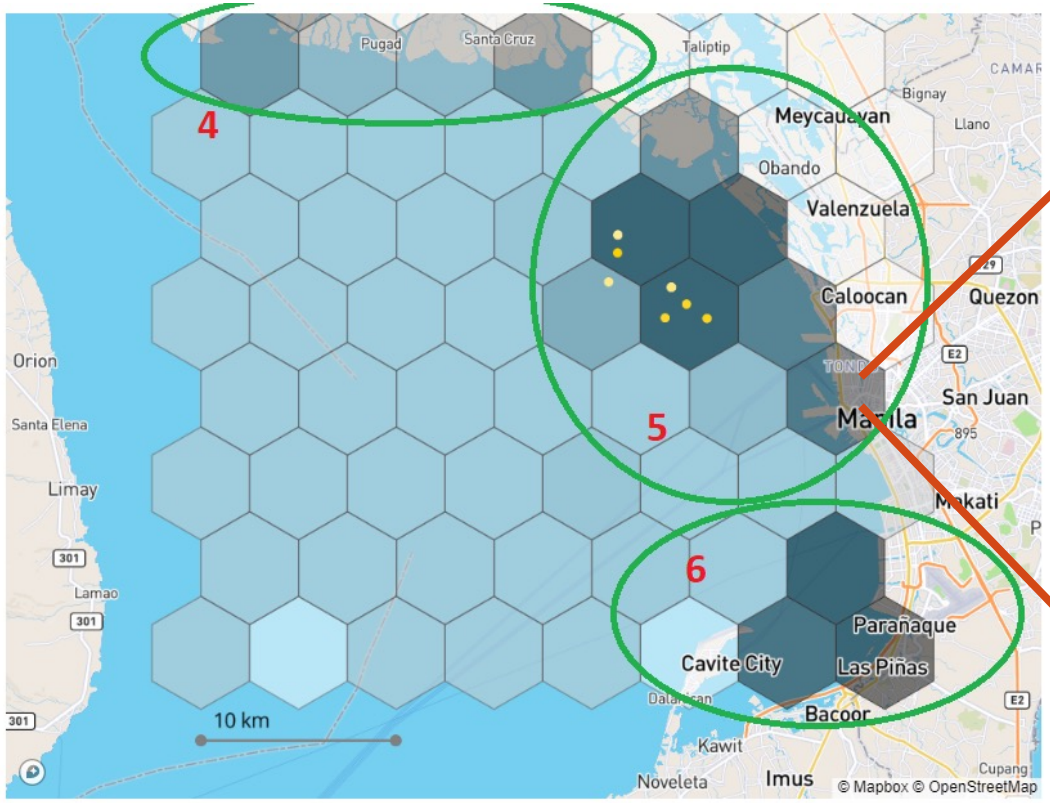
- **Can we develop a high-precision marine debris monitoring system by using satellite imagery?**

Our Answer

- Around 1K multispectral Satellite imagery in the data set.
- A lightweight ML model developed
- 95% precision reached!
- Unseen-data test with historically polluted regions!
  - Mumbai, Honduras, Manila

Marine Debris Monitoring

Pasig River, Manila, Philippines



2.4 Exploiting Large-Language Models

Exploiting Large-Language Models

QUESTION?

- **Can Large-language models (LLMs) be used to summarise and analyse long process reports?**

ANSWER

- Business Process Reports
- Scientific Research Reports
  - Literature reviews
- Engineering and Technical Reports
  - Design summaries
- Legal and Compliance Reports
  - Case summaries

Exploiting Large-Language Models

2.4 – Root Causes Extraction

Our Question?

- **Can LLMs help extract the key information from accident investigation reports?**

Our Answer:

- Project proposed by Empirisys
  - Provides Data Science solutions for high-risk industry
- harness different language processing models to identify the root cause of accidents
- 390 open access reports used.
- Developed un-supervised techniques based on BERTopic & GPT 3.5

Exploiting Large-Language Models

2.4 – Root Causes Extraction

PERFORMANCE?

- GPT3.5 based model reached 70% accuracy
- Distribution of incorrect classifications

GPT result \ Human annotation	Lapses	Organizational failure	No PIF	Total
Equipment failure	1	2	4	7
Lapses	-	12	1	13
Organizational failure	0	-	5	5
Procedures violation	3	5	4	12
Software failure	0	1	2	3
No PIF	0	2	-	2
Total	4	22	16	42

2.5 Power of Multi-modal Data

Power of Multi-modal Data

QUESTION?

- **Can data from different sources be used in novel computing tools? Can this bring advantageous results?**

ANSWER

- Healthcare
  - Electronic health records (EHRs), wearable devices, and genomic databases
- Smart Cities and Urban Planning
  - sensors, social media, and public records
- Agriculture and Precision Farming
  - satellite imagery, weather stations, and IoT devices
- Finance
  - financial data, market trends, and social media sentiment

Power of Multi-modal Data

2.5.1 – Air Pollution Mapping

Our Question?

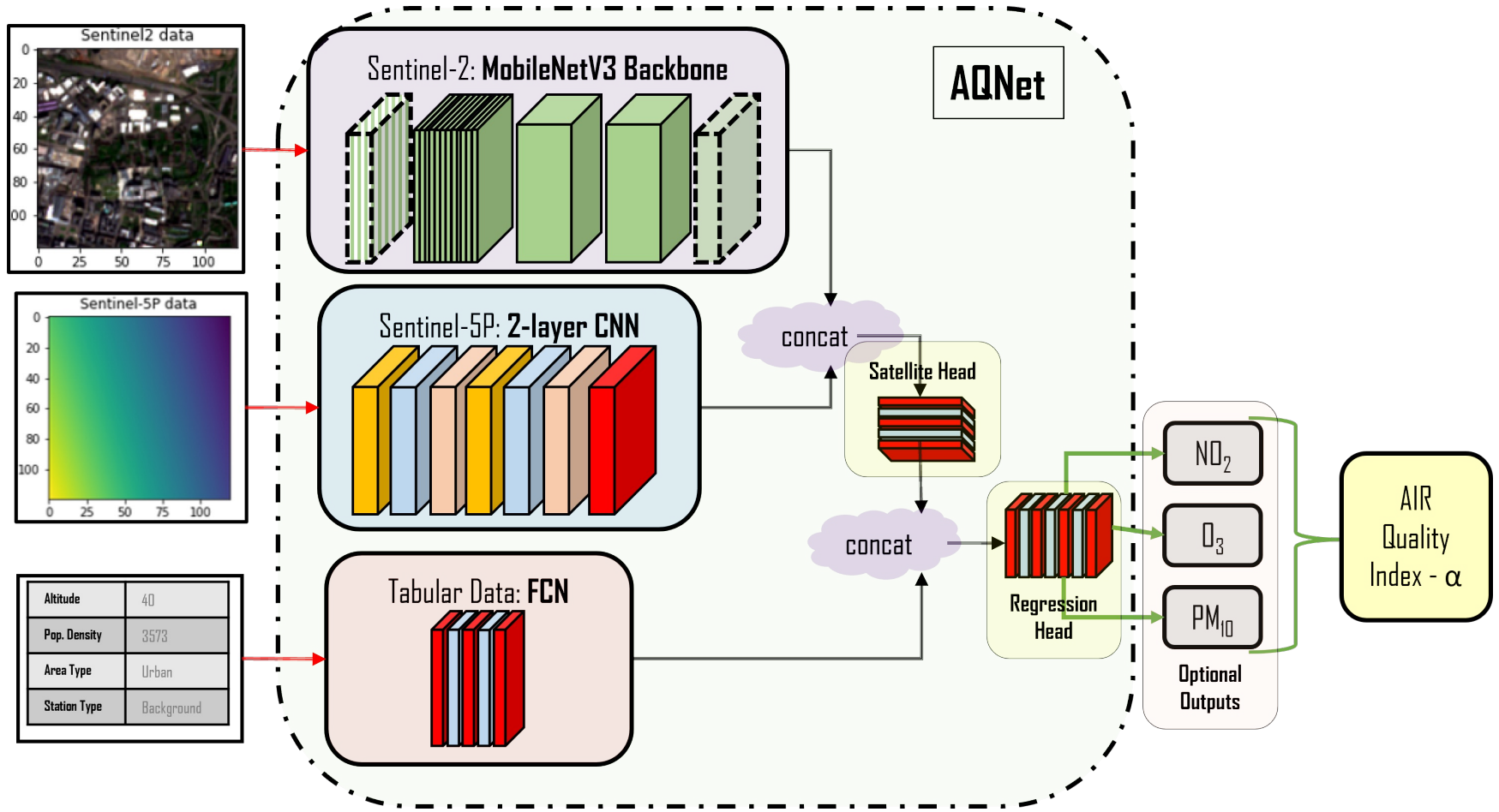
- **Can multi-modal data help improve air-pollution mapping?**

Our Answer 1:

- A 3-level AI model (AQNet) developed for
  - Satellite imagery
  - Satellite pollution measurements
  - Region specific information, e.g. urban, suburban, population, elevation, etc.
- 10% improvement compared to single modality

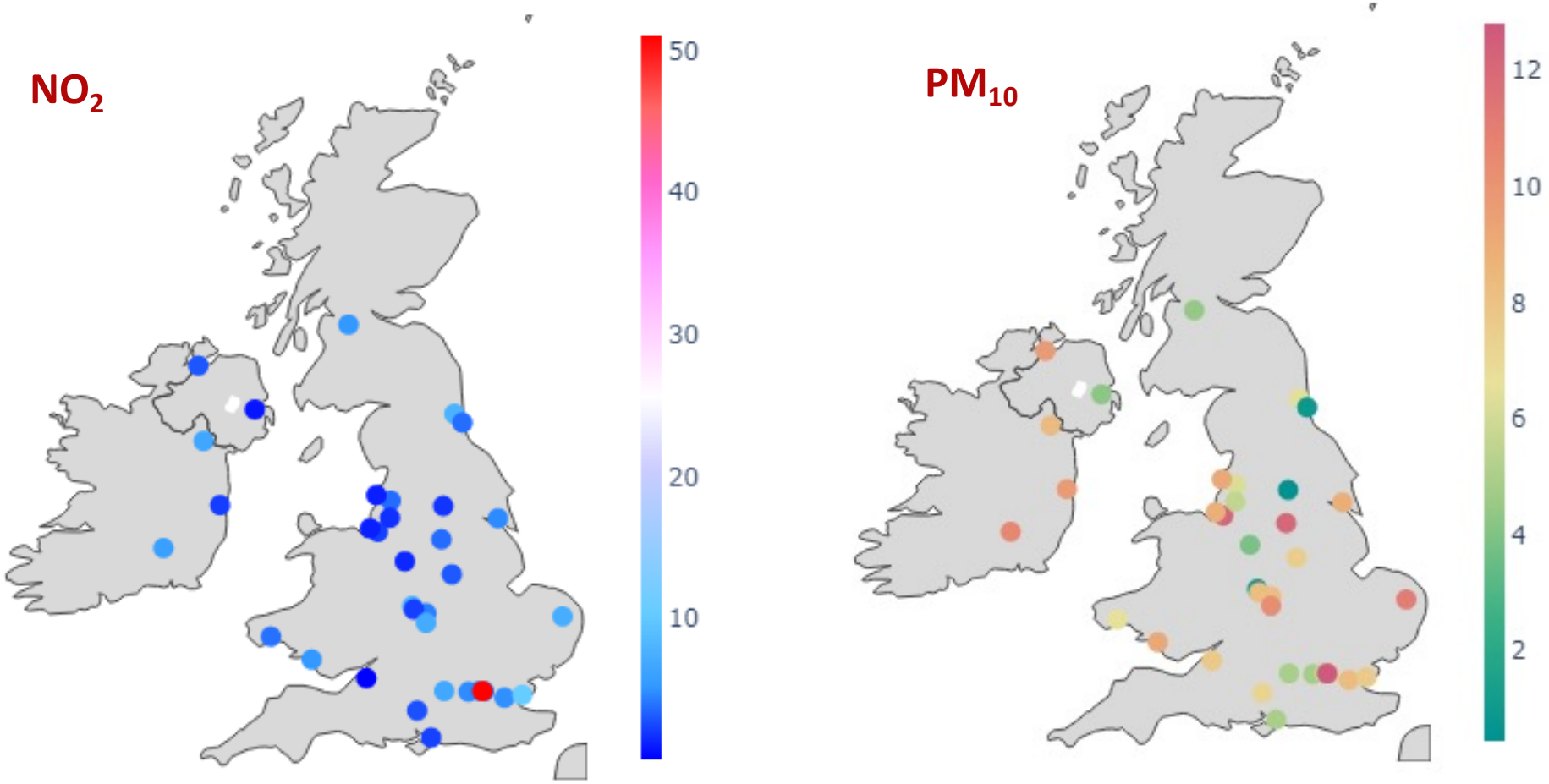
Power of Multi-modal Data

2.5.1 – Air Pollution Mapping



Power of Multi-modal Data

2.5.1 – Air Pollution Mapping



Power of Multi-modal Data

2.5.2 – Rice Crop Yield Prediction

Our Question?

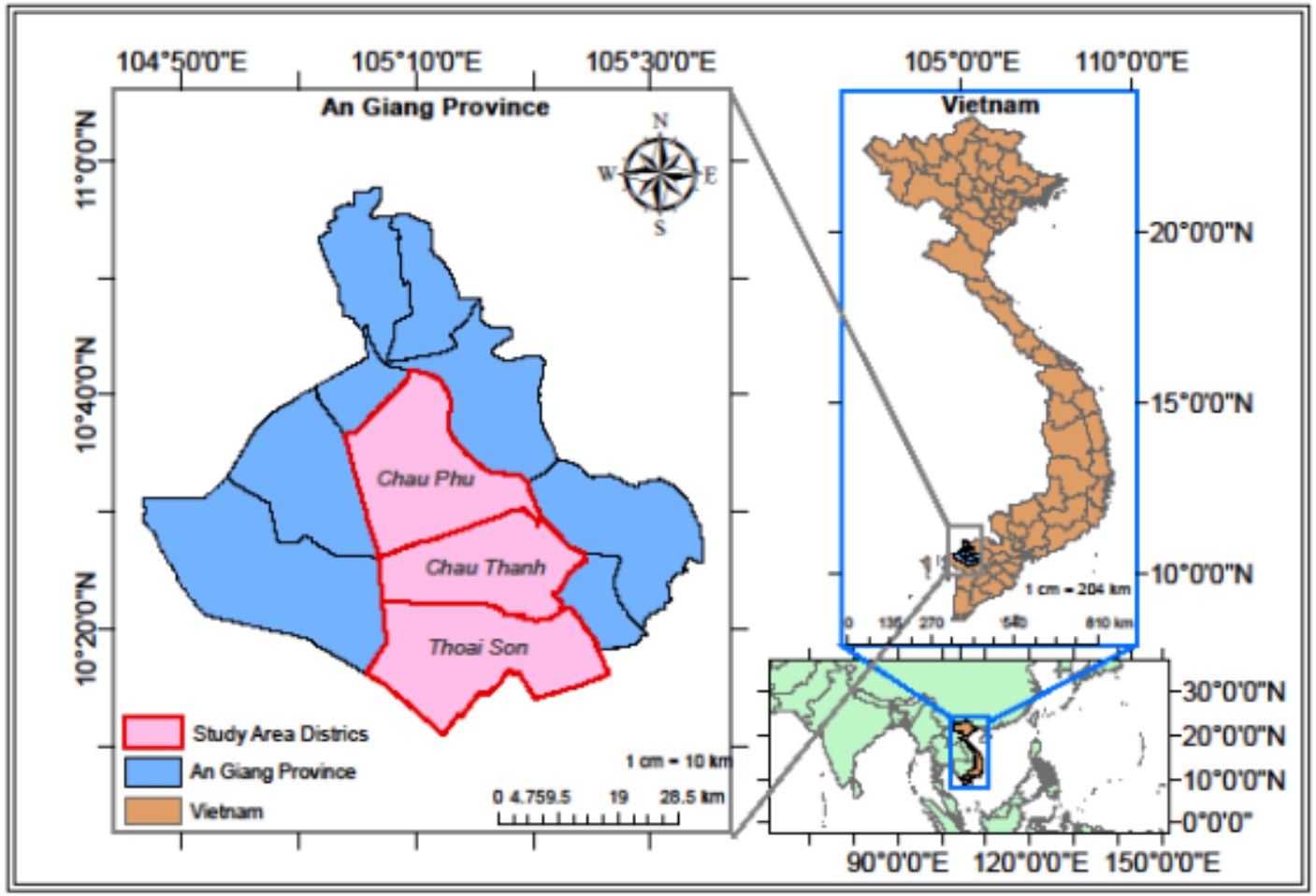
- **Can multi-modal data help improve prediction capacity of ML models?**

Our Answer 2:

- Ernst & Young (E&Y) Data Science Contest 2023 data set used.
- 100+ data features from 5 distinct sources (modalities)
- Data Engineering applied → 15 best features selected
- A novel Deep-Ensemble Regression model → **RicEns-Net**
- **Not more than 10-12% error obtained.**

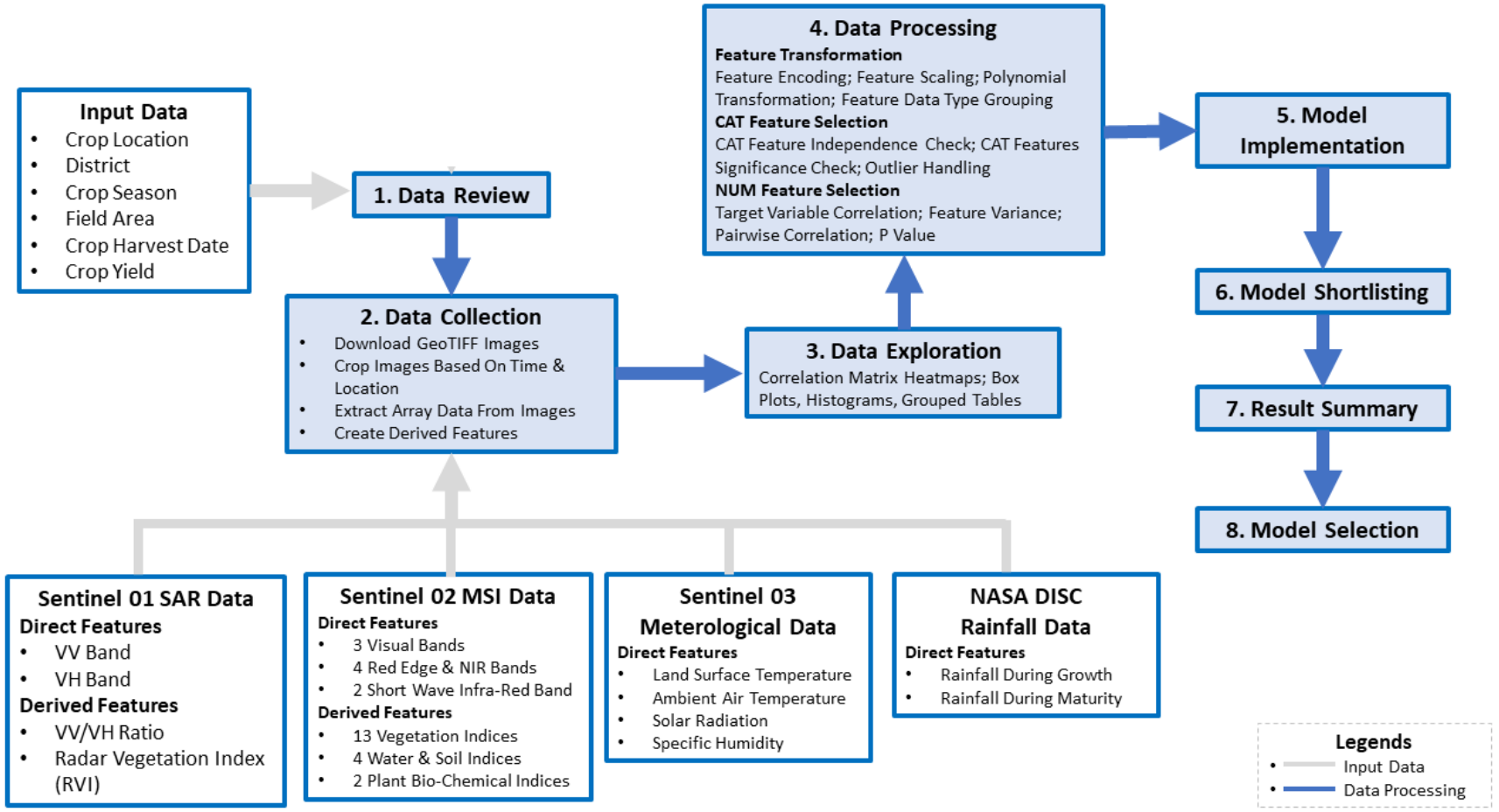
Power of Multi-modal Data

2.5.2 – Rice Crop Yield Prediction



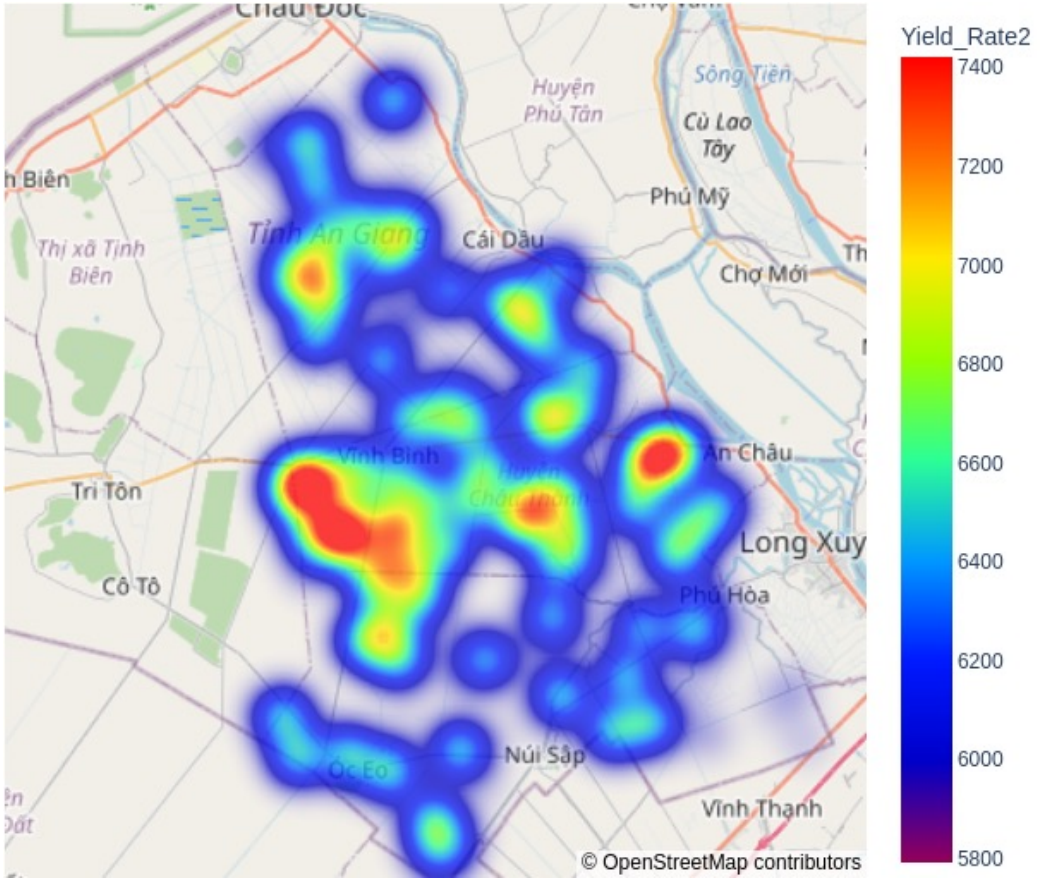
Power of Multi-modal Data

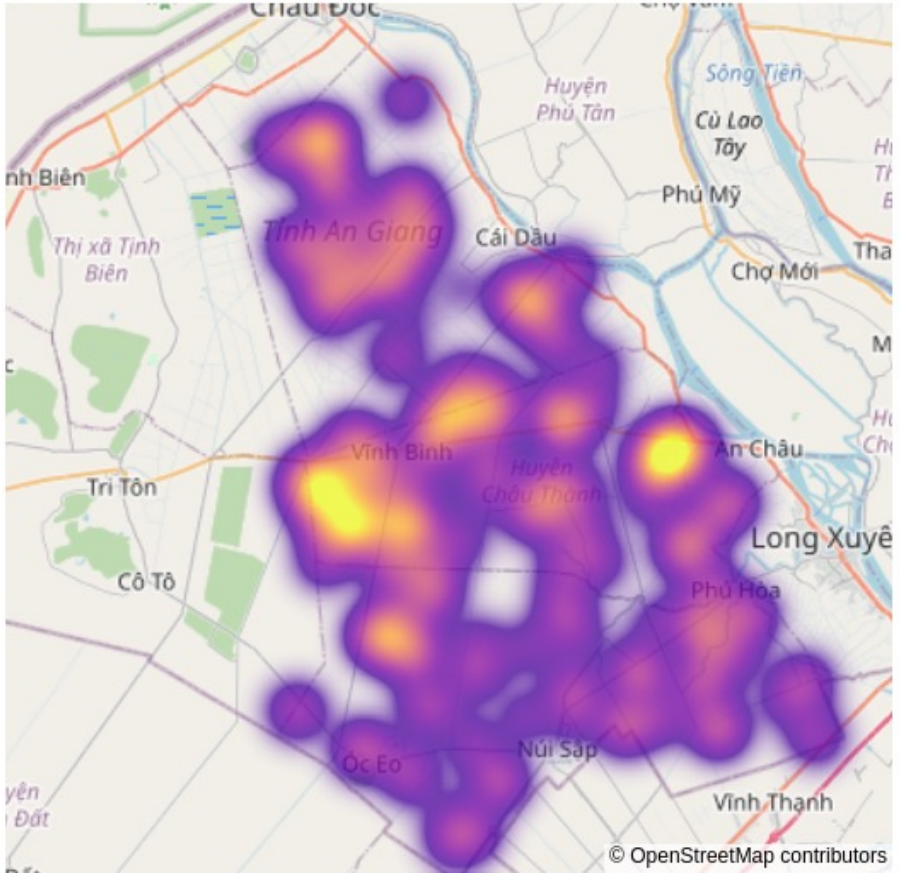
2.5.2 – Rice Crop Yield Prediction

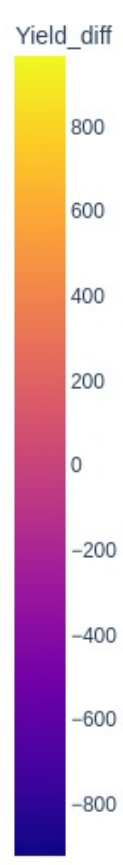


Power of Multi-modal Data

2.5.2 – Rice Crop Yield Prediction







Many more …

Risk Profiling

IDENTIFYING THREAT INTELLIGENCE

Investigating Human Security Behaviours

COVID-19 VIRTUAL ASSISTANT

Identifying Repetition Patterns

Forensic Analysis

Queuing Analysis

Financial Forecasting

Emission Modelling

Road Traffic analysis

Conclusions

Future steps to Data-driven Discovery

Gather to Develop Clear Policies

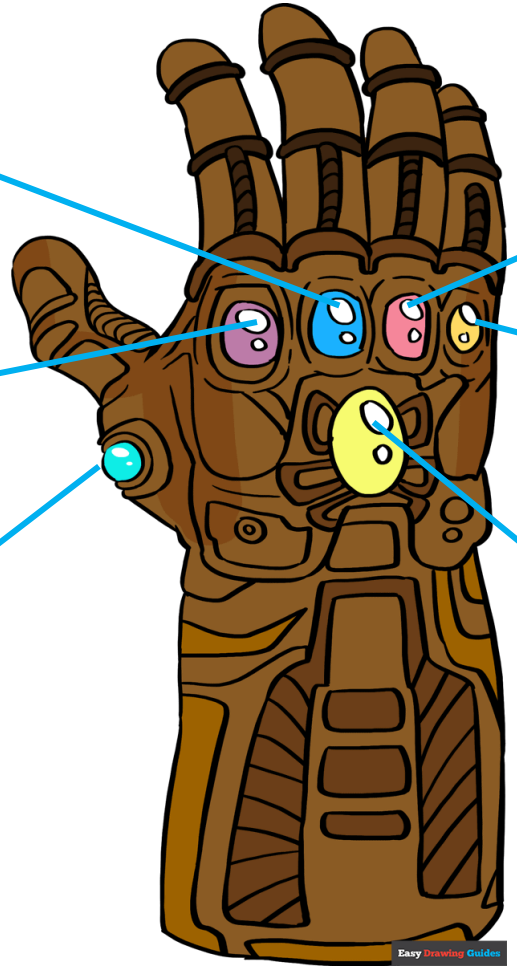
Increase Academy & Industry Collaboration

More Council & Governmental Support

Support Open Access Software

Embrace the Dynamic Nature

Make Open Access Data Available



Thanks for your Attention!

Ready for Questions

?

FC: CartoonStock