Data Science in Naimuri

A Postdoctoral Graduate's Tale

Data Science in Naimuri

Who we are

Software Technology Company, Focusing on Security, Law Enforcement, Intelligence & Home Office

Cloud Native Organisation - Deep Relationships with AWS & Azure

Operate from a Facility in Salford Quays, Manchester

Placing value on reducing complexity and promoting simplicity

Our Approach

DevSecOps delivery of custom designed innovative solutions meeting customer and user needs.

Working with a collaborative, Agile and One Team approach. Driven by a user journey centric design.

We use advanced building blocks to reduce the time to deliver value. From infrastructure to UI building blocks, we're rarely starting from scratch.

Delivering through iterative sprints, the customer has control of priorities and value is delivered regularly.

About Naimuri

Introduction to Naimuri

Our Mission is to *help make the UK a safer place*, and our Vision is to do this through *technological innovation*.
Historically this has focused on software, network security, search and encryption techniques, etc.
Today this means secure cloud and data as much as the above.

About Naimuri

What we do with data:

We are secure by default: security is intrinsic to our thinking, planning, designing, implementation and testing.

We are specialists in search, including semantic search with word and sentence embeddings.

We are specialists in ETL pipelines, including entity extraction.

Increasing work in:

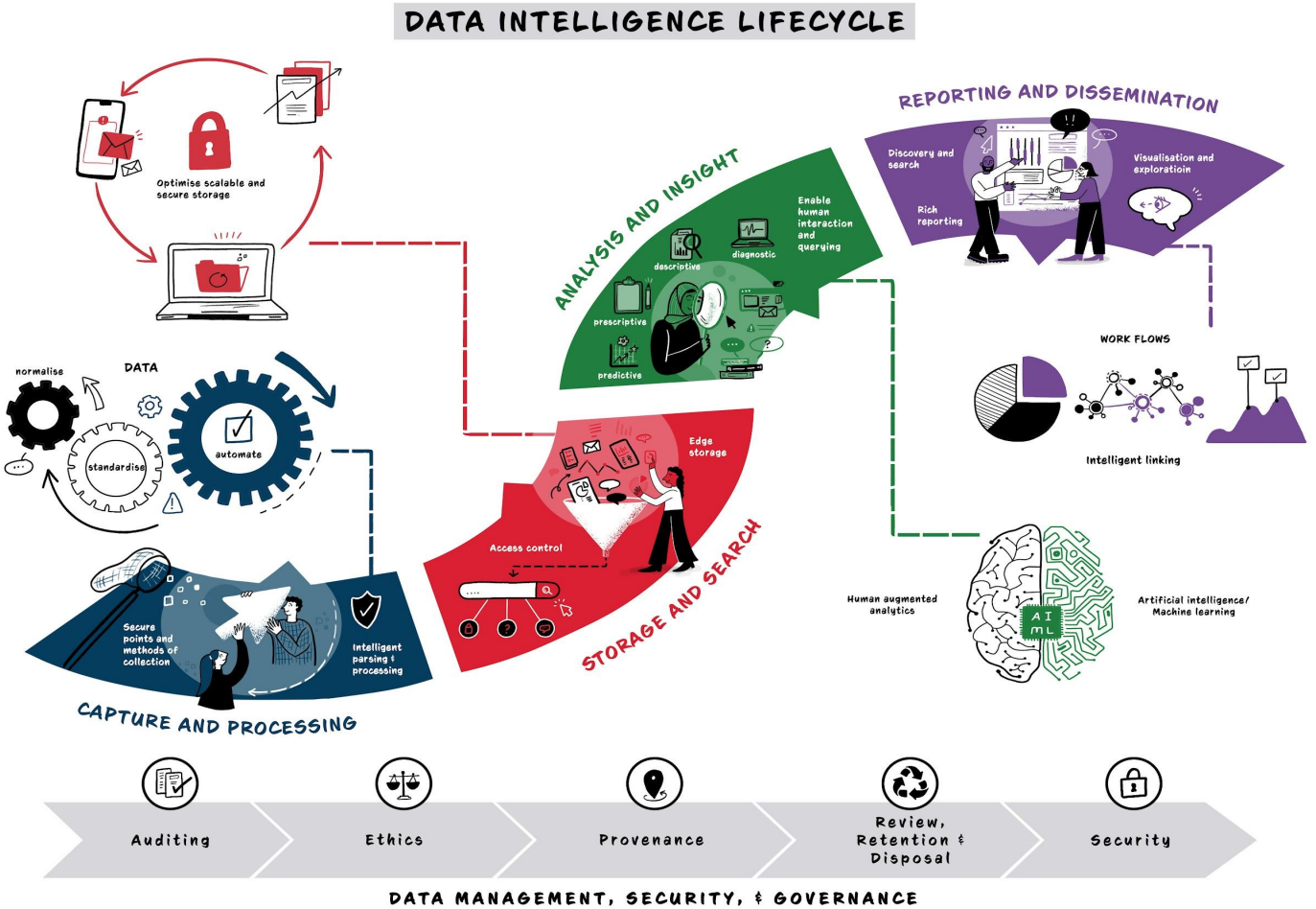
- recommendations engines
- computer vision (object detection in images and video)
- NLP: classification, entity extraction, relation extraction
- knowledge graphing and querying
- data synthesis: NLG, tabular data, image synthesis

About Naimuri

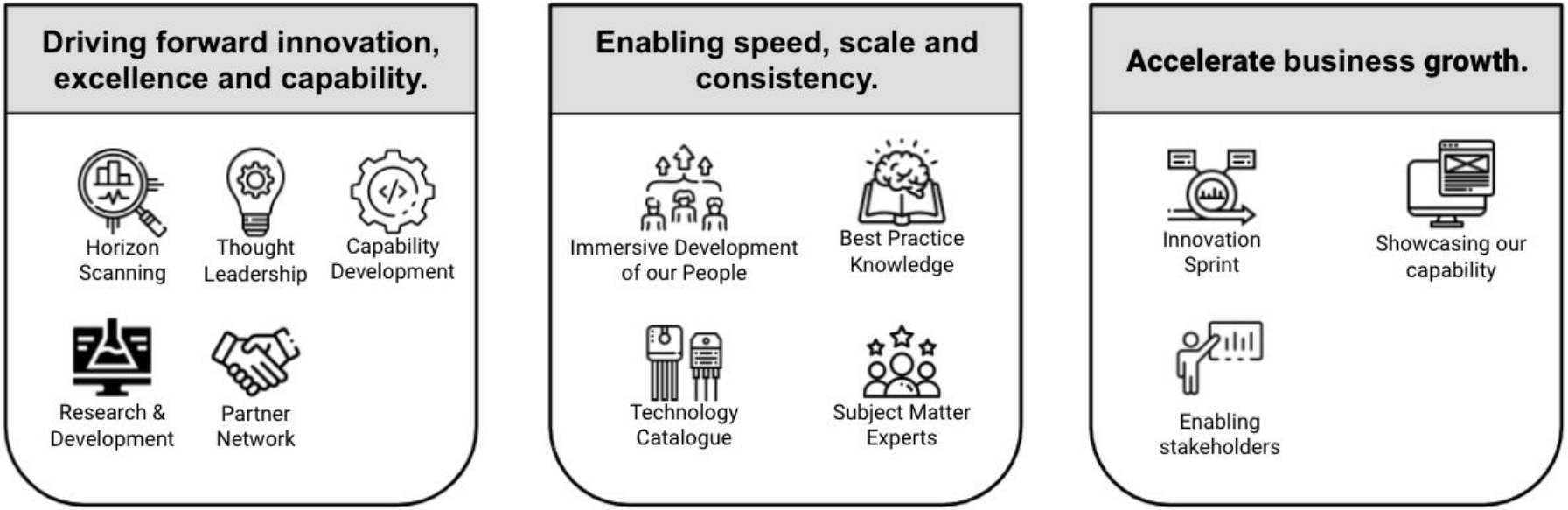
Types of deliverable

- Research
- Prototyping
- Minimum viable products
- Continuous development and support

Data Intelligence



Data Intelligence Centre of Excellence



"Joining the best minds in the North West to tackle Impactful Data Challenges"

naimuri

A QINETIQ company

Some of what we do (and who for)

GDPR

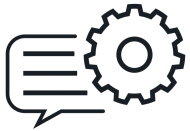
Reporting

Secure Cloud

Secure Storage

Dashboard

Search



NLP

AI/ML

Secure by Default

NCA
National Crime Agency

Home Office

Cabinet Office

NHS

TRANSFORMING FORENSICS Programme

COUNTER TERRORISM POLICING

Department for Transport

... and those are just some of the ones I can talk about

Who we partner with

naimuri

A QINETIQ company

Agile

Naimuri always works in an Agile way.
In fact, Agile is *why* we are called *Naimuri*.

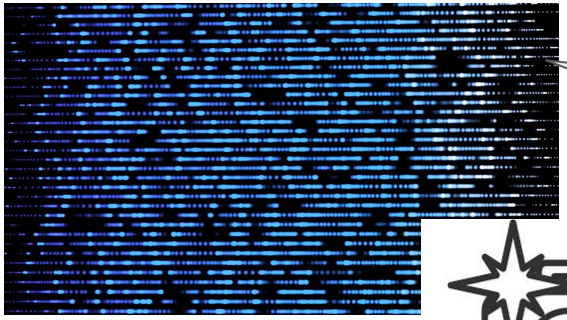
Naimuri: Case Study 1

Document Classification: Using machine learning methods to identify the themes of any document

Real data (10,000,000s documents) always moving

Naimuri: Case Study 1

Document Classification: Using machine learning methods to identify the themes of any document

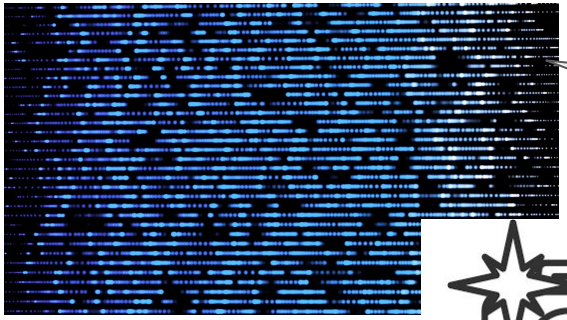


Real data (10,000,000s documents) always moving

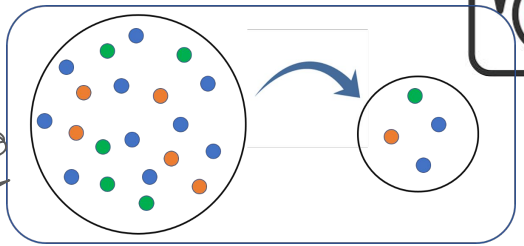
For repeatable experiments, data needs to be frozen

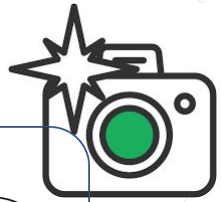
Naimuri: Case Study 1

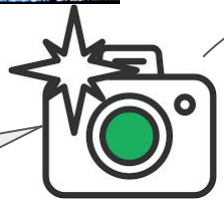
Document Classification: Using machine learning methods to identify the themes of any document



Real data (10,000,000s documents) always moving





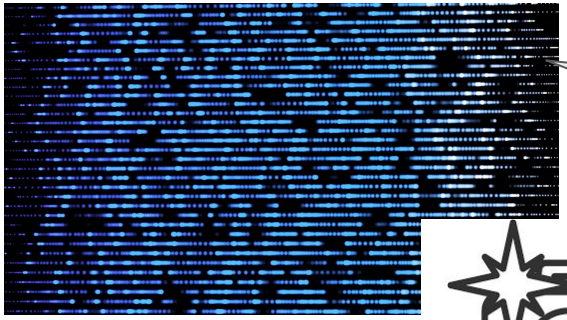


For repeatable experiments, data needs to be frozen

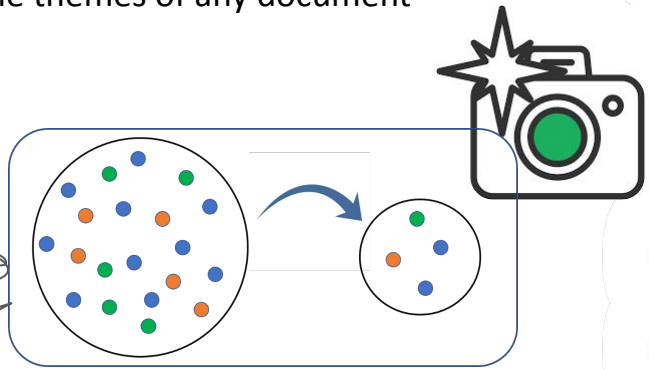
And so do our samples: training, testing and validation

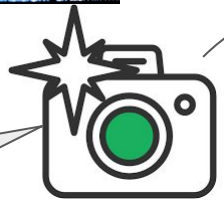
Naimuri: Case Study 1

Document Classification: Using machine learning methods to identify the themes of any document



Real data (10,000,000s documents) always moving

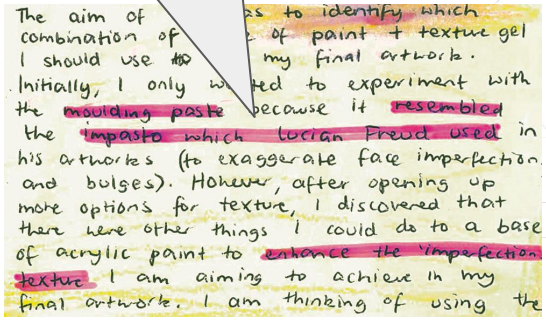




For repeatable experiments, data needs to be frozen

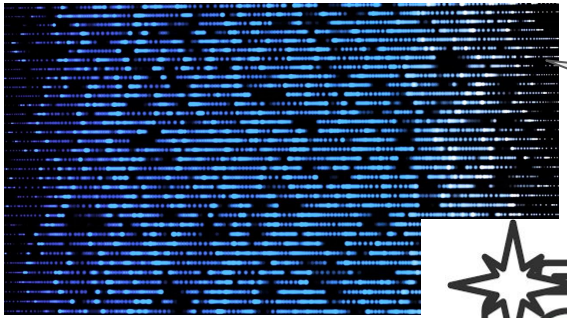
And so do our samples: training, testing and validation

A predictive model is only as good as its ground truth

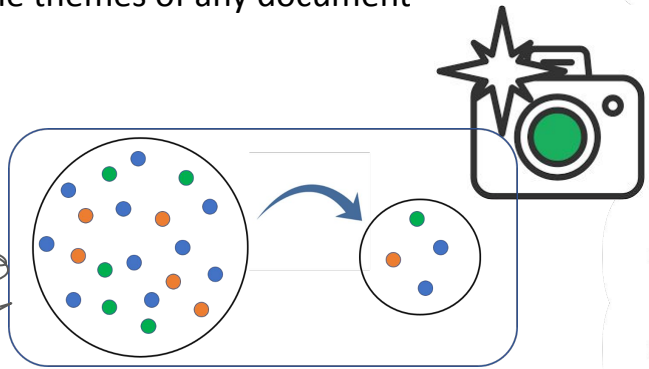


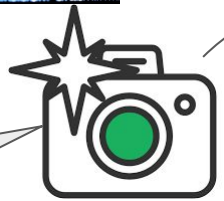
Naimuri: Case Study 1

Document Classification: Using machine learning methods to identify the themes of any document



Real data (10,000,000s documents) always moving

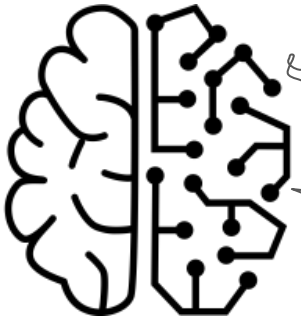




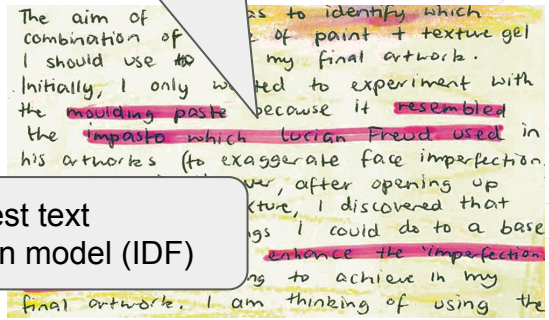
For repeatable experiments, data needs to be frozen

And so do our samples: training, testing and validation

A predictive model is only as good as its ground truth

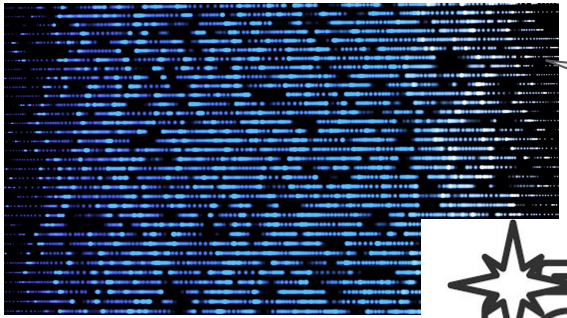


Train and test text classification model (IDF)

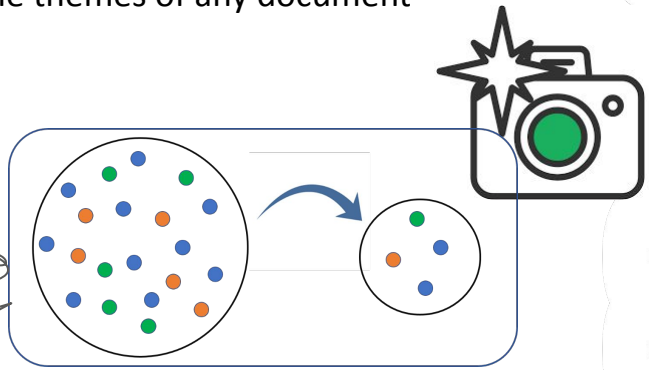


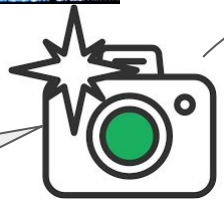
Naimuri: Case Study 1

Document Classification: Using machine learning methods to identify the themes of any document



Real data (10,000,000s documents) always moving

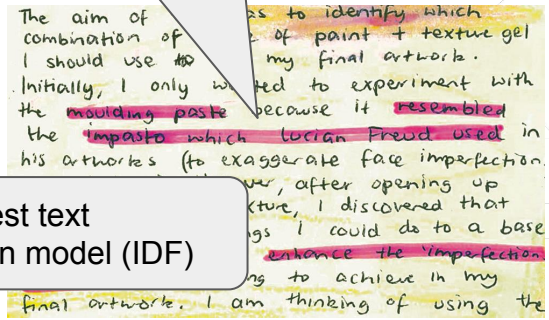


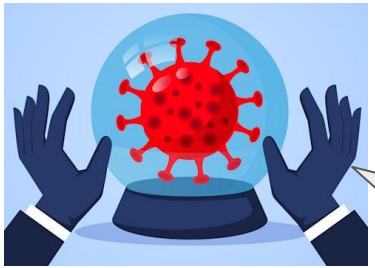


For repeatable experiments, data needs to be frozen

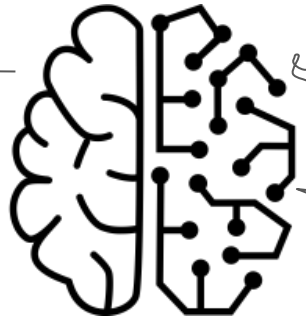
And so do our samples: training, testing and validation

A predictive model is only as good as its ground truth





Deploy & predict



Train and test text classification model (IDF)

Naimuri: Case Study 2

Semantic Search: Using neural networks to understand human queries

When dealing with 100,000s documents, there's always a gap between what you *want* and what you *get*

Naimuri: Case Study 2

Semantic Search: Using neural networks to understand human queries

When dealing with 100,000s documents, there's always a gap between what you *want* and what you *get*

Even fuzzy matching, synonym matching and stemming can't quite close the gap between meaning and querying

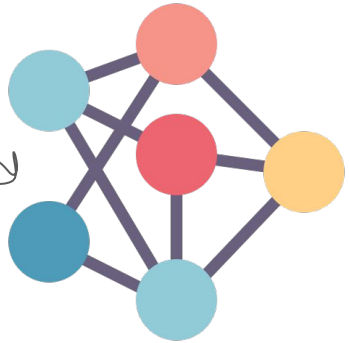
Naimuri: Case Study 2

Semantic Search: Using neural networks to understand human queries

When dealing with 100,000s documents, there's always a gap between what you *want* and what you *get*

Even fuzzy matching, synonym matching and stemming can't quite close the gap between meaning and querying

Transformers can be trained unsupervised to learn the semantic proximity of words



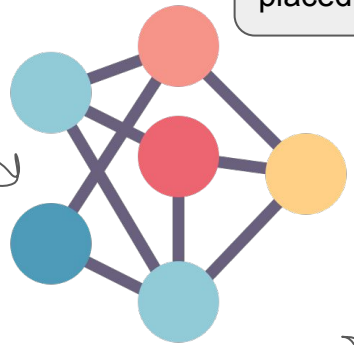
Naimuri: Case Study 2

Semantic Search: Using neural networks to understand human queries

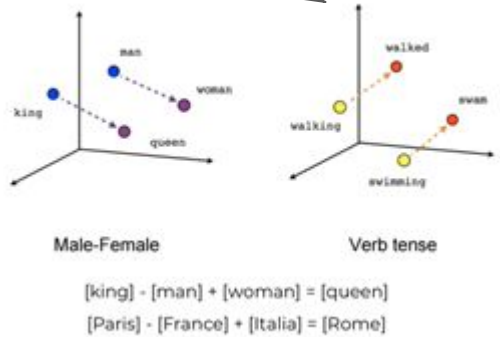
When dealing with 100,000s documents, there's always a gap between what you *want* and what you *get*

Even fuzzy matching, synonym matching and stemming can't quite close the gap between meaning and querying

Transformers can be trained unsupervised to learn the semantic proximity of words in context of sentences



Each sentence yields a vector placed in a vector store



Naimuri: Case Study 2

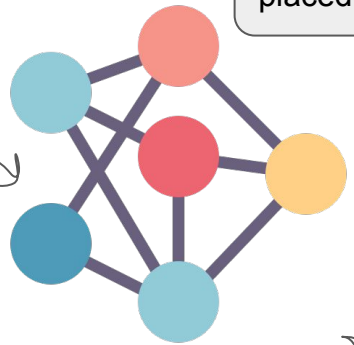
Semantic Search: Using neural networks to understand human queries

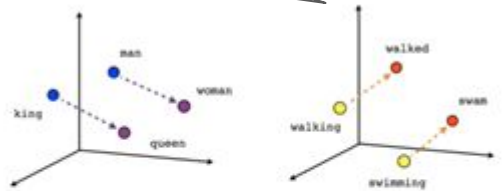
When dealing with 100,000s documents, there's always a gap between what you *want* and what you *get*

Even fuzzy matching, synonym matching and stemming can't quite close the gap between meaning and querying

Each sentence yields a vector placed in a vector store

Transformers can be trained unsupervised to learn the semantic proximity of words



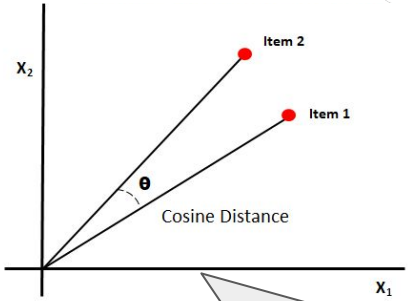


Male-Female

Verb tense

[king] - [man] + [woman] = [queen]

[Paris] - [France] + [Italia] = [Rome]



Search by cosine similarity

Naimuri: Case Study 2

Collaborative Filtering: Learning what similar users are interested in

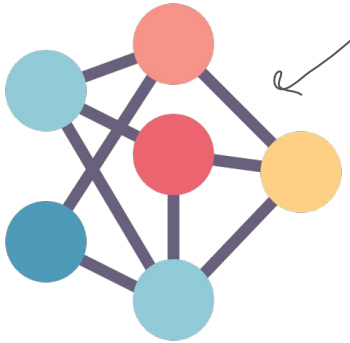
What different people click through once they've found something is insightful

Person A

Person A has a long history of interacting with the data.

Person B has a similar history, and has just made a query similar to one of Person A's.

Person B



We can train a neural network to learn how to score Person B's search results based on Person A's history (and vice versa).
Weights fitted on latent variables about users and documents (*a la* VAEs).

Dr. James Ramsden

James completed his PhD in Physics at the University of York in 2014 which a heavy focus on algorithm development and numerical optimisation. Following further postdoctoral work at the same university, he moved to Salford Quays to work apply his algorithm development skills in R&D as part of a startup, where he started working with predictive ML models. Salford Quays introduced him to Naimuri, the most exciting company in the area.

Applying skills:

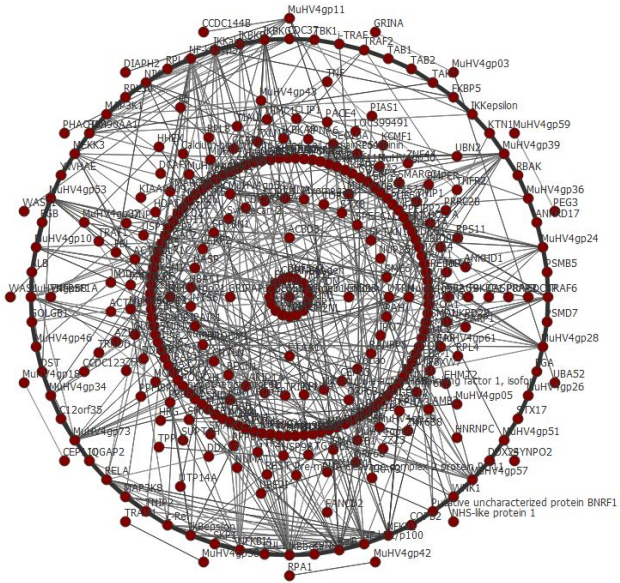
Numerical optimisation and algorithm development in Naimuri.
The relationship between numerical optimisation and other areas of data science.
Predictive ML models: training and inference.
Putting the **science** in **data science**.

Data Science in Naimuri

Dr. Phininder Balaghan (AKA Phini)

My PhD is titled *An Exploration of Graph Algorithms and Graph Databases.* It's a mixture between applied and theory.

- It discusses the *Finding the longest cycle in a graph* problem
  - Created an Integer Linear Problem (ILP) to solve the problem
  - Modified a simple DFS search to produce a heuristic
- Explored a number of computational graph algorithms, and how they can be applied in Graph Database Systems
  - Findings include the graph engines are optimised for certain features



I always say my "PhD is the worst thing I've written!"

Dr. James Ramsden: Doctoral Work

Properties of Exact Density Functionals for Electronic Quantum Transport
My research was in exact solutions for the exchange-correlation potential in time-dependent density functional theory, a reformulation of time-dependent many-body quantum mechanics.

QM laws are easy (ish) to write down but usually impossible to solve.
TD-DFT equations are trivial to solve, but exact laws are unknown (and, as was discovered, *unknowable*).

Skills acquired
Numerical optimisation
Algorithm development
Mathematical modelling
Scientific principles

Selected References
Exact Density-Functional Potentials for Time-Dependent Quasiparticles, J. D. Ramsden & R. W. Godby, *Phys. Rev. Lett.* **109** 3 (2012)
Intrinsic exchange-correlation magnetic fields in exact current density functional theory for degenerate systems, J. D. Ramsden & R. W. Godby, *Phys. Rev. B.* **88** 19 (2013)
Exact time-dependent density-functional potentials for strongly correlated tunneling electrons, M. J. P. Hodgson *et al, Phys. Rev. B.* **88** 24 (2013)
Origin of static and dynamic steps in exact Kohn-Sham potentials, M. J. P. Hodgson, J. D. Ramsden & R. W. Godby, *Phys. Rev. B.* **93** 15 (2016)

Dr. James Ramsden: Postdoctoral Life

Postdoctoral studies

Continued in physics research full-time for one more year.
Investigated methods for using wavefunction symmetries to solve the many-body Schrödinger equation without needing the full grid.
Implemented methods for solving the equation for periodic crystals.

R&D

Moved to the North West to join a startup.
R&D in mobile indoor location services.

- Numerical optimisation of probability clouds representing and exploring device state
- Algorithm development for matching device data to ambient magnetic fields
- Machine learning approaches to deriving latent variables (velocity) from device (accelerometer) data.

Able to track a mobile phone around large supermarkets for tens of minutes.

Didn't work in America. Buildings made of wood don't give off much in the way of magnetic fields.
Startup went bust.

Dr. James Ramsden: Naimuri Data Science

Data Analytics Team

Led the first Data Analytics team at Naimuri, using JuPyter Notebook-based ways of working to analyse customer data and solve customer data challenges. Typical methodology:

1. data cleansing and transformation;
2. exploratory data analysis;
3. hypothesis generation and testing;
4. presenting results using Notebook App Mode.

First customers in the Insurance sector.

Machine Learning for Document Classification

Applied Natural Language Processing (NLP) techniques to train document classification models for UK government customers.

Dr. James Ramsden: Naimuri Data Science

Data Science Team
In 2021, I became Chief Data Science Officer at Naimuri. In 2022, I became Data Science Capability Lead, wherein I work with the senior leadership team define the direction of data science capability development at Naimuri. That direction is selected according to customer need and high demands of ourselves.

Semantic Search for NHSX
Using sentence embeddings to enable semantic federated search across multiple medical data sources.

Object Detection Models for UK Border Force
Training convolutional neural networks to locate smuggled humans in vehicles from blurry, warped, low-resolution black and white imagery, reaching an accuracy of 90%.

ML-Driven Data Synthesis for Counter Terrorism Policing

- tabular data synthesis using Variational Autoencoders (VAEs)
- image synthesis using generative adversarial networks (GANs)
- image-to-image translation
- NLG using Long Short-Term Memory (LSTM) models and OpenAI's GPT-2.

Building a Secure Data Science Environment for Counter Terrorism Policing

Dr. James Ramsden: What I've learned

Data Science is Two Key Words...

Data is paramount: how you get it, how you treat it, how you store it, how you sample it, how you think about it, what questions you ask of it.

Science is paramount: understanding the background, hypothesis generation and testing, statistical analysis and modelling, representative sampling, repeatability, prediction, how to validate results.

Research is essential: staying on top of developments in your field, both in academia and in white papers from major technology companies (Google, Microsoft, NVIDIA, IBM, etc.).

Data science is often done in the context of software development, but it's not like software development: it is much closer to other sciences in methodology.

However two aspects of software engineering have very close analogues in data science:

1. While the individual methods of **data and ML assurance** are data scientific, the philosophy behind the exacting standards expected of software systems applies to data science components;
2. Likewise **data and MLOps pipelines** are very different to traditional build and deployment pipelines, but the overall aim is the same.

Roles in Data Science

A lot of overlap in skills between roles, which conform more to *functions*.

Data Analytics
Data preparation
Mathematical modelling
Graph analytics
Hypothesis generation
Visualisation
Reporting

Data Engineering
Big data engineering
Test data engineering
Graph engineering
Analytical engineering

ML Engineering
NLP
Computer Vision
Forecasting
MLOps engineering

ML Assurance
Performance
Convergence, Regression
Ethics
Explainability
Security

Capability
Component development
Data pipeline development
Search
Capability assurance

Data Operations
Data architecture
Database administration
Data governance
Data assurance

Key Skills

Generic Skills

- Data Preparation: sampling, cleansing and transformation, how to handle missing data.

- Statistics & Exploratory Analysis: get to understand the data, its statistical properties, correlations, etc.
- Visualisation: plotting, dashboarding, reporting.

- Hypothesis Generation & Testing: understanding the data means having expectations of what you can learn from it, and knowing how to evaluate those expectations.
- Narrative building: akin to writing a paper, but not as hard.

- Feature Selection & Extraction: how to find/derive the right data features to answer questions.

- Agile: prioritise, fail fast, finish the task closest to completion soonest.
- Structure: the ability to break down larger problems or tasks into a logical sequence of smaller tasks.

- Staying on top of the latest developments in your field (e.g. deep learning)

- And TEST, TEST, TEST!!!

Key Skills

Task-Specific Skills

- Numerical Optimisation, machine learning, deep learning, hyperparameter optimisation.
  - Transformers & PyTorch
  - TensorFlow & Keras
- Training and inference:
  - NLP (document classification, entity and relationship extraction, NLG)
  - computer vision (image classification, object detection, image segmentation)

- Knowledge graphing, graph querying and analytics

- Component Development: Being able to take your work and turn it into something you can deliver.
  - object-oriented programming
  - test-driven development
  - incorporating best practise (linting, coverage tests, ...)

- Deployment Pipelines: DevOps & MLOps on the cloud (AWS, Azure), quality gating, Serverless

- And TEST, TEST, TEST!!!

How best to prepare for a career with us

- Practise!: There are great challenges you can get feedback on in e.g. Kaggle and Analytics Vidhya.
- In particular, practise data wrangling: sampling, cleansing, formatting, and feature extraction.
- Explore OO principles: they will always be useful.
- Familiarise yourself with Agile manifesto principles.
- Python is more widely used than R.
- Most work starts with a JuPyter Notebook.
- Subscribe to good blogs (e.g. Toward Data Science) and keep your eye on latest publications

Data Science in Naimuri

How to join us

Email: yourfuture@naimuri.com

Website: https://naimuri.com