DATA
SCIENCE

Data science in the real world

25th September 2018

Tom Begley

Objectives

Stimulate curiosity about the techniques and tools used in data science.

Acquire familiarity with some common tools and be able to try them out!

Objectives

Bonus: understand some of the jargon data scientists use against you!

"When we speak without jargon, it frees us from hiding behind knowledge we don't have. Big words and fluffy *business speak* cripples us from getting to the point and passing knowledge to others."

Outline

1. “Hang on… who are you exactly?”
2. What does a data science project look like?
3. Fundamental tools
4. Data science platforms

Demo

5. Data ingestion and exploration
6. Machine learning modelling
7. Model deployment
8. Q & A.

Who is ASI?

PEOPLE
FELLOWSHIP

EXPERTISE
CONSULTING

TECHNOLOGY
SHERLOCKML

Who is ASI?

BBC | tom | News | Sport | Weather | iPlayer | TV | Radio | More | Search

NEWS

Home | UK | World | Business | Politics | Tech | Science | Health | Family & Education | Entertainment & Arts | Stories | More

Technology

UK unveils extremism blocking tool

Dave Lee
North America technology reporter

13 February 2018

Share

WATCH: The BBC's Chris Foxx learns about the tool

The UK government has unveiled a tool it says can accurately detect jihadist content and block it from being viewed.

Top Stories

UK life expectancy progress 'has stopped'
Improvements have ground to a halt for the first time since records began, in 1982, ONS data shows.
2 hours ago

Pret baguette allergy alerts before death
1 minute ago

Beluga whale spotted in the Thames
8 minutes ago

Features

Debt killed my dad

CIO FROM IDG | POPULAR: | CIO 100 | CIO INTERVIEWS | CIO DIRECTORY | EVENTS | PARTNER ZONE | JOIN CIO | PLUS

Home > IT Business > IT Strategy

Amnesty International CIO John Gillespie using data science to track press and media monitoring

Amnesty International CIO John Gillespie on the charity's use of data science, AI and machine learning

By Hannah Williams February 7, 2018 CIO UK

Amnesty International CIO John Gillespie has partnered with a London startup to help use data science to measure sentiment analysis and improve media monitoring of the global human rights charity.

The CIO of London-based Amnesty International told *CIO UK* that AI and machine learning are two technologies the organisation brought together to help it quickly gauge how it is being represented in the media.

The charity receives a significant amount of coverage across a broad range of topics so Gillespie and his team looked towards emerging technologies to see how to improve Amnesty's data science capabilities.

"There are many media monitoring services available, and they are great at tracking sentiment and reporting how much is being written about an organisation," Gillespie said. "This is sufficient for a company that is sending out a handful of press releases each month, but when you are issuing four or five a day and you want to know the impact of each one individually, you need something more sophisticated."

Amnesty has already started using machine learning as a research tool, including to detect and classify violence and abuse against women on social media platforms. But spending resources on investigative work is a potentially costly risk for the charity, while tracking the sentiment analysis of Amnesty on each story and the effectiveness of its press campaigns was precisely the type of large-scale, complex process that advances in data science could help with - without having any kind of negative impact on its core mission.

The organisation turned to London-based startup ASI Data Science for assistance with the process. The startup believes that AI should be accessible for everyone and organises a 'Data Science Fellowship' that enables top PhD graduates and software engineers to go through a six-week programme covering data science, data engineering and applications in industry to work on real-world big data problems.

Most Popular

How UK CIOs manage vendor relationships

Good Energy CTO David Ivell plans to create a...

The CIO role in designing and delivering the...

Data science projects

Planning

- Select project according to business needs/user requirements.
- Assess technical feasibility.
- Estimate timelines and costs
- Get the data!

Exploration

- Assess data quality given the objectives.
- Get basic insights and verify simple assumptions.
- Define the modelling approach.

Modelling

- Prepare the data (link sources, clean data).
- Feature engineering.
- Develop a machine learning model.
- Evaluate model performance and optimize it.

Deployment

- Automate data access and model training.
- Wrap the model in a web service (API) and expose it to the internet.
- Build a user interface (web app).

ASI

Fundamental tools

Programming languages	Libraries	Other tools
python™	R	Scala
• Scripting, • Machine learning • Software development	• Scripting, • Statistical analysis • Machine learning	• Distributed computing • Machine learning • “Big data”

Fundamental tools

Programming languages

Libraries

Other tools

NumPy

Numerical computation

pandas

$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$

Data cleaning, feature engineering

scikit learn

Machine learning

TensorFlow

Deep learning

Fundamental tools

Programming languages

Libraries

Other tools

jupyter

- Interactive programming environment
- Combines markdown with runnable code cells

- Text editors and IDEs
- Useful for production

Fundamental tools

How to get started with python :

- Interpreter
- Libraries
- Other tools

ANACONDA

- Open source distribution of lots of Python packages.
- Package manager.
- https://www.anaconda.com/download/

Fundamental tools

How to get started with R :

Console

Libraries

Other tools

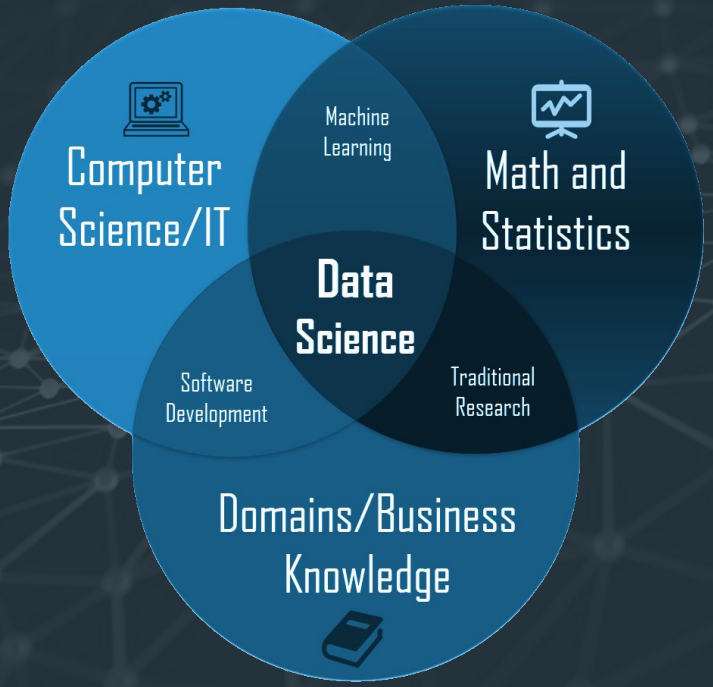
RStudio®

- Open source IDE for R
- Has popular libraries built in to the interface, e.g. knitr
- https://www.rstudio.com/products/rstudio/#Desktop

Open source tools

Let’s take a look!

Data science platforms



Challenges

- Data scientists are not (necessarily) **software developers**...
- ...nor **data engineers**...
- ...nor **system administrators**!
- What they can do depends heavily on background (academia?) and experience (in which type of company?).
- They might heavily rely on **support** for software development and infrastructure.
- **This may cause a lot of time and resources to get wasted!**

Data science platforms

Challenges → **Technology**

Data science platforms

Challenges → **Technology**

Collaboration

Data science platforms

Challenges → **Technology**

Infrastructure

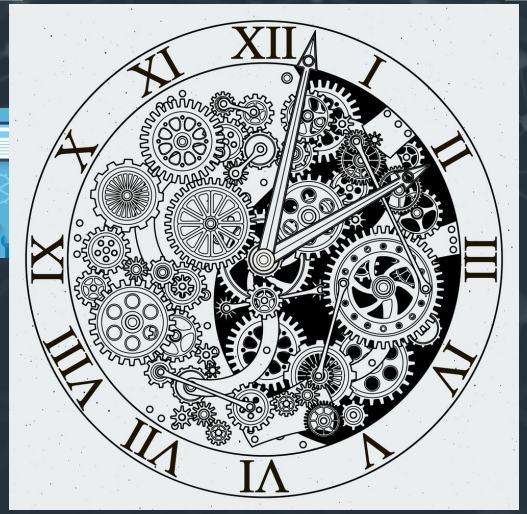
Data science platforms

Challenges → **Technology**

Tools

Data science platforms

Challenges → **Technology**



Configuration

Data science platforms

Challenges → **Technology**

Security

Data science platforms

Challenges → **Technology**





Data science platforms

Challenges → **Technology**





(platform!)

Data ingestion and exploration

Let’s log in!

Machine learning modelling

Let's train a **supervised** machine learning model!

Labelled data

$$(\vec{x}_i, y_i)$$

Target

Features

Machine learning modelling

Let's train a **supervised** machine learning model!

Labelled data

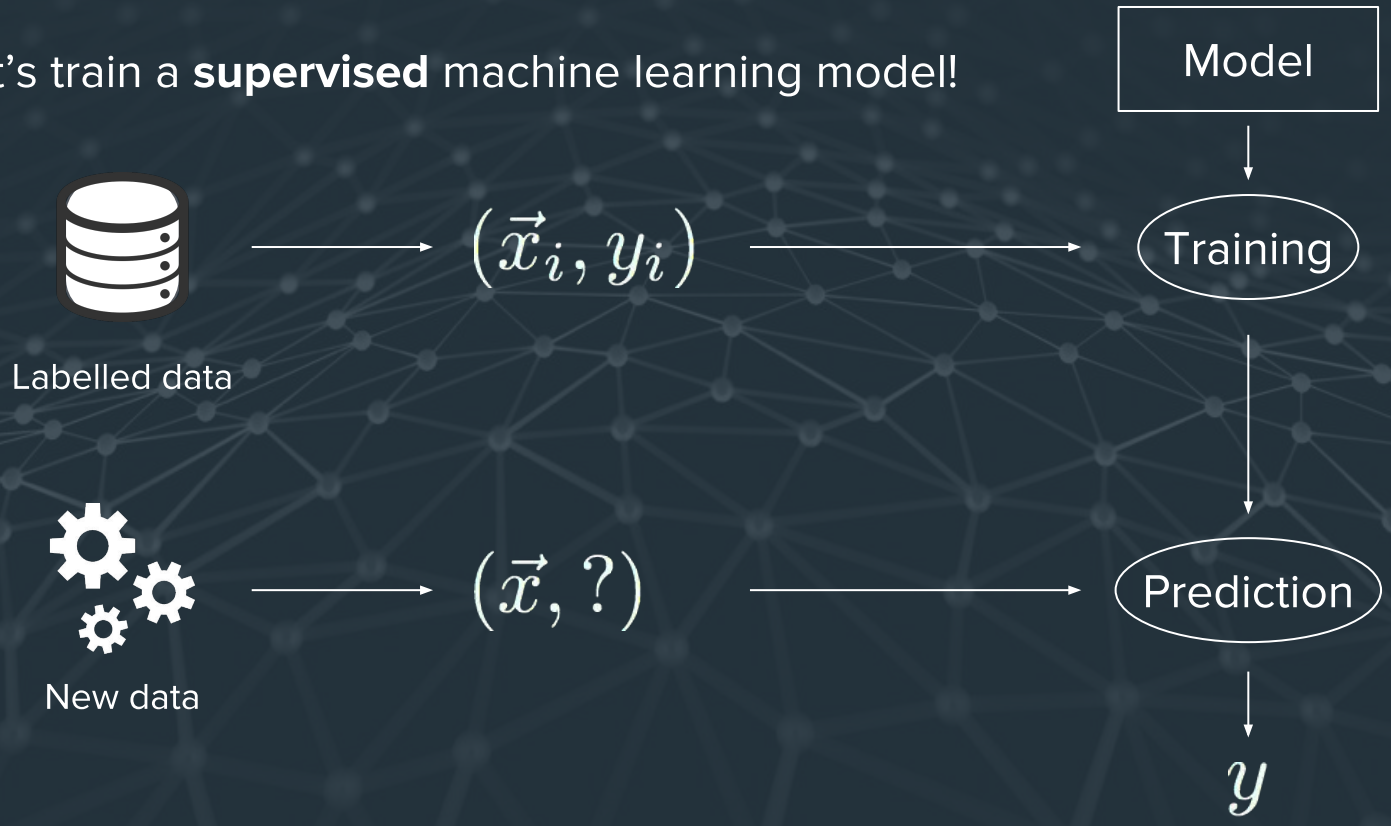
$$(\vec{x}_i, y_i)$$

New data

$$(\vec{x}, ?)$$

Machine learning modelling

Let's train a **supervised** machine learning model!



Machine learning modelling

Model	• An hypothesis on how the **features** are related to the **target variables**. • A particular choice for the form of the function $F(x) = y$. • Contains **parameters** (weights) and **hyperparameters**.	Class (object)
Training	• Fit the model's parameters to the known, labelled data. • May need to set sensible hyperparameters also	`fit()` class method (function)
Prediction	• Use the model, with the values for the parameters learned in the training phase, to predict the target given a new datapoint.	`predict()` class method (function)

Machine learning modelling

Model

- An hypothesis on how the **features** are related to the **target variables**.
- A particular choice for the form of the function $F(x) = y$.
- Contains **parameters** (weights) and **hyperparameters**.

Class (object)

Class attribute

Training

- Fit the model's parameters to the known, labelled data.
- Nothing to say about the hyperparameters!

`fit()` class method (function)

Prediction

- Use the model, with the values for the parameters learned in the training phase, to predict the target given a new datapoint.

`predict()` class method (function)

Machine learning modelling

Model

A class is a **blueprint** for a custom type of variable: to get an actual model we have to "create a variable of that type", or **create an instance of the class**.

Class (object)

Training

Prediction

An instance of the Model class will possess all the methods (and attributes) that are defined within the class (all the functions an instance of that class can execute).

`fit()` class method (function)

`predict()` class method (function)

Machine learning modelling

Let’s train a model!

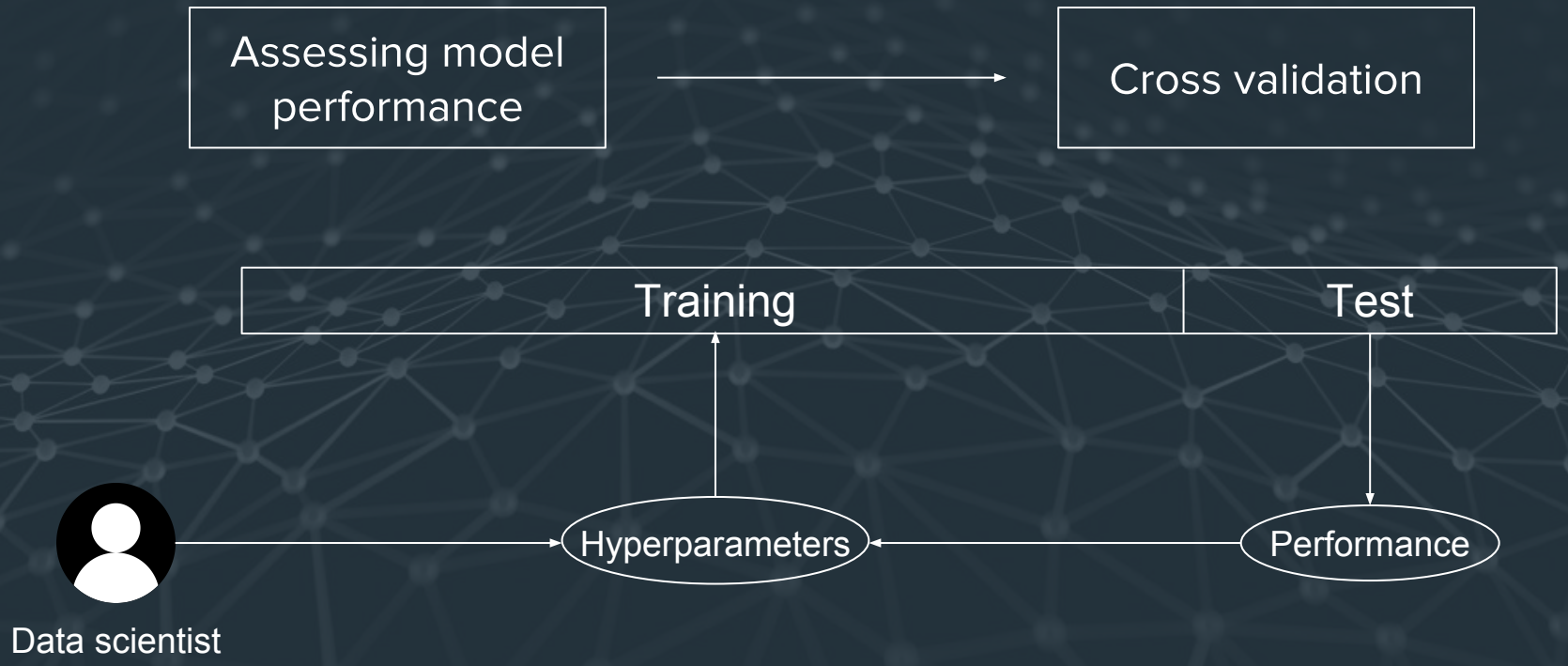
Machine learning modelling

Assessing model performance → Cross validation

- Split the labelled data into a **training** and a **test** dataset:

- Train the model on the training data.
- Get predictions for the test data and **compare** them with the true labels by choosing a performance metric.
- Repeat for different train/test splits and compute the average performance.
- Optimize the choice of hyperparameters.

Machine learning modelling



Machine learning modelling

Cross validation with Python

Model deployment

Development environment

Model

Model deployment

Development environment

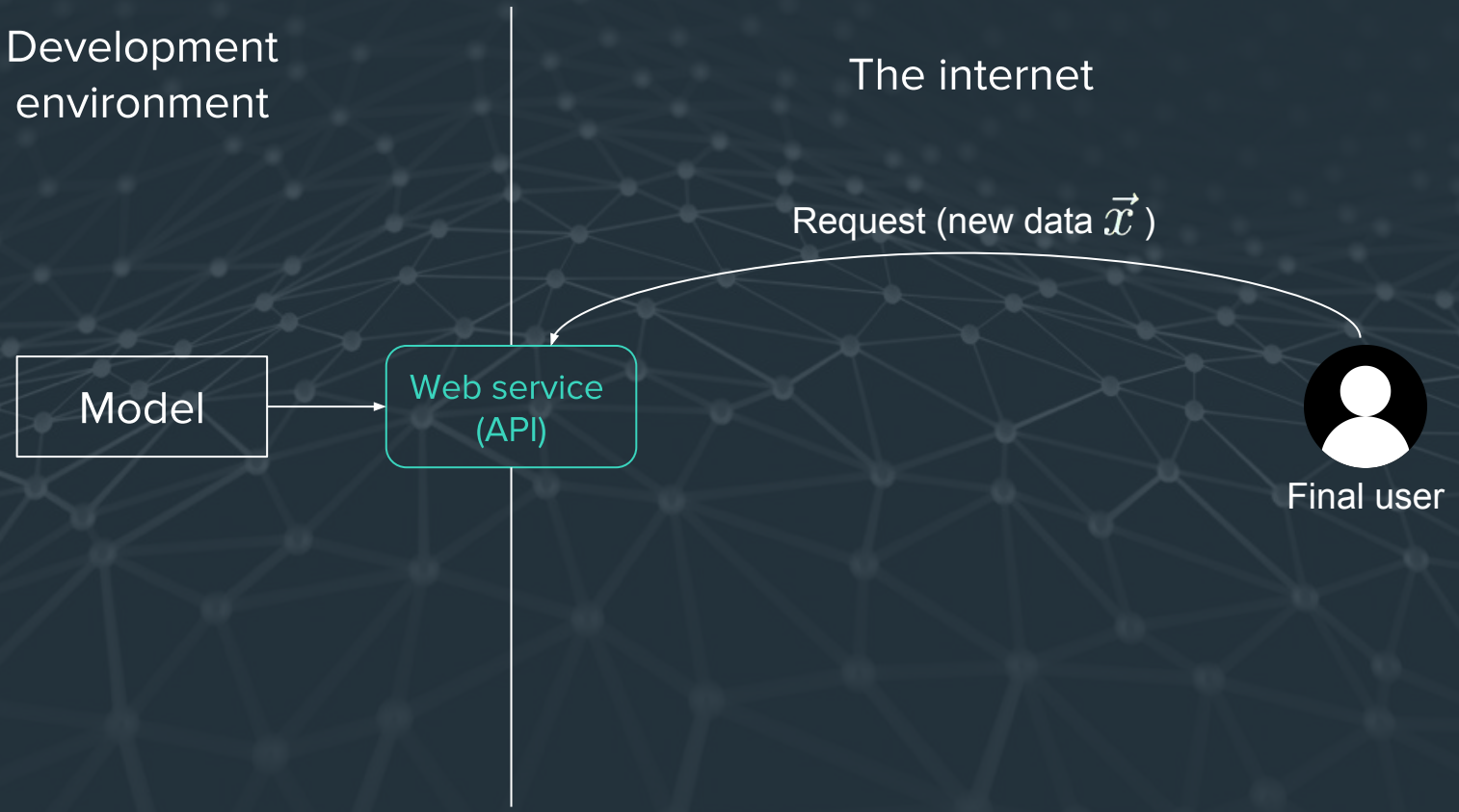
The internet

Model

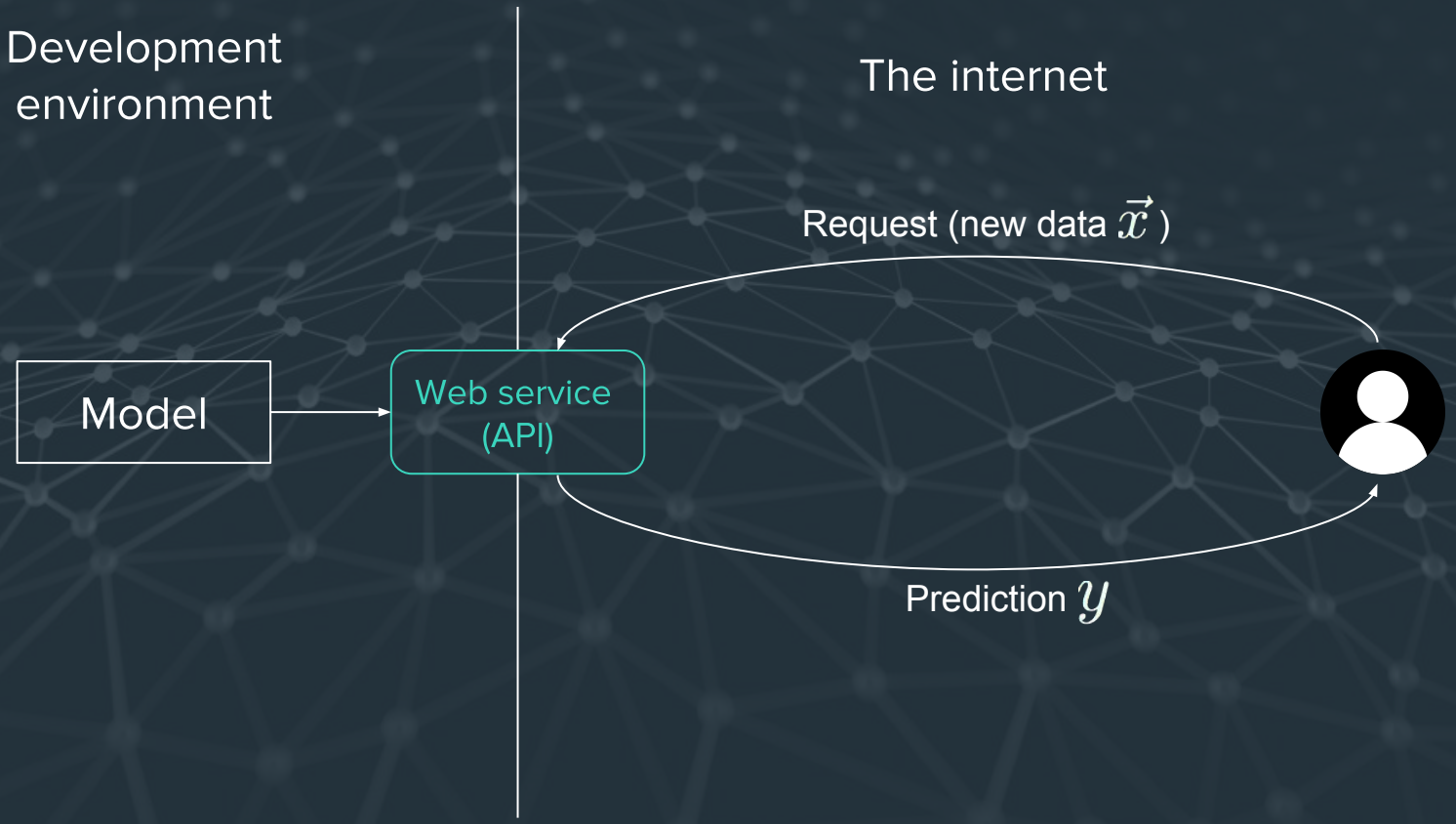
Final user

Model deployment

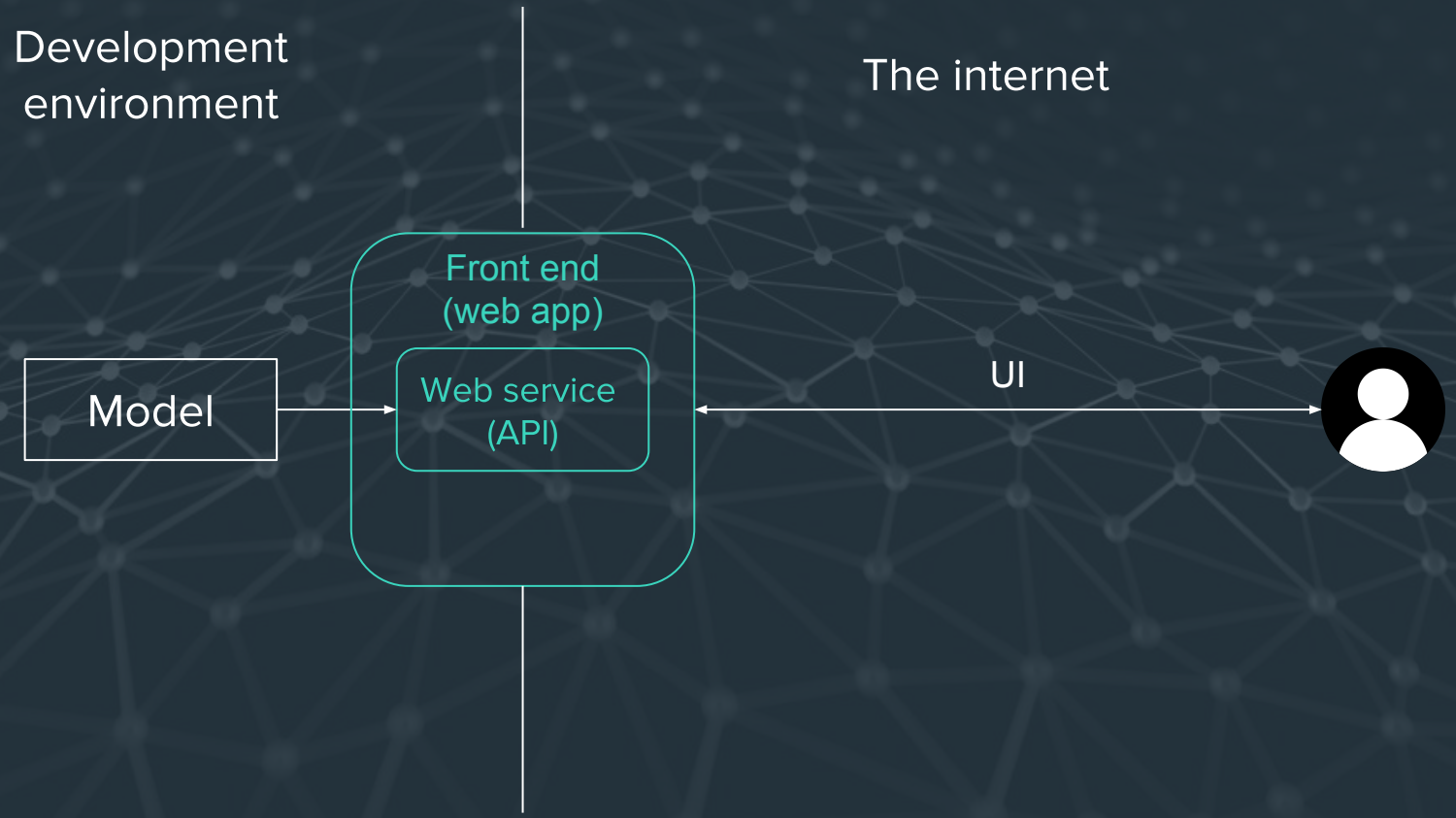
Model deployment



Model deployment



Model deployment



Model deployment

Software

Front end
(HTML, JavaScript,...)

Back end
(web framework)

Hardware

Infrastructure
(server)

URL
(www.myapp.com)

Plotly Dash

Flask, Django,
Tornado

EC2 machines
(Amazon), ...

Model deployment

Serve the trained model through a web app

Q & A