FALMOUTH UNIVERSITY

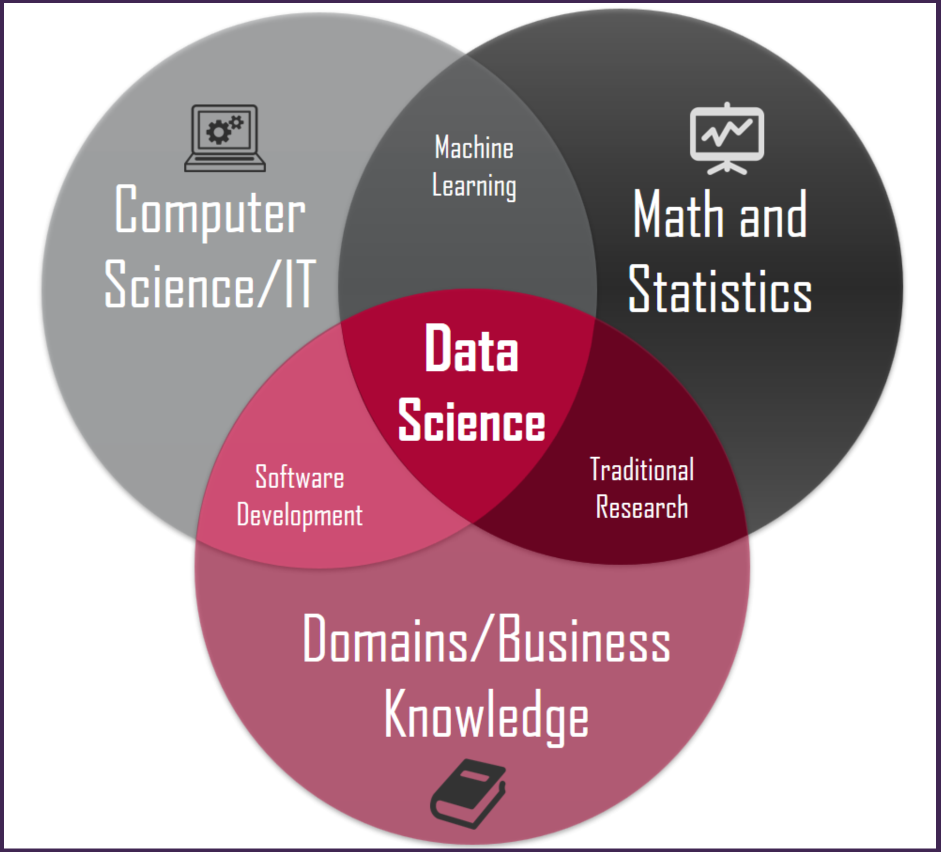
Lecture 2: Data Science pt. I

COMP704: Machine Learning
MSc Artificial Intelligence for Games

- Today’s session:
  - Dig into Data Science & Machine Learning
  - Look at data acquisition
  - Prepare for this week’s workshop

- Dig into Data Science & Machine Learning

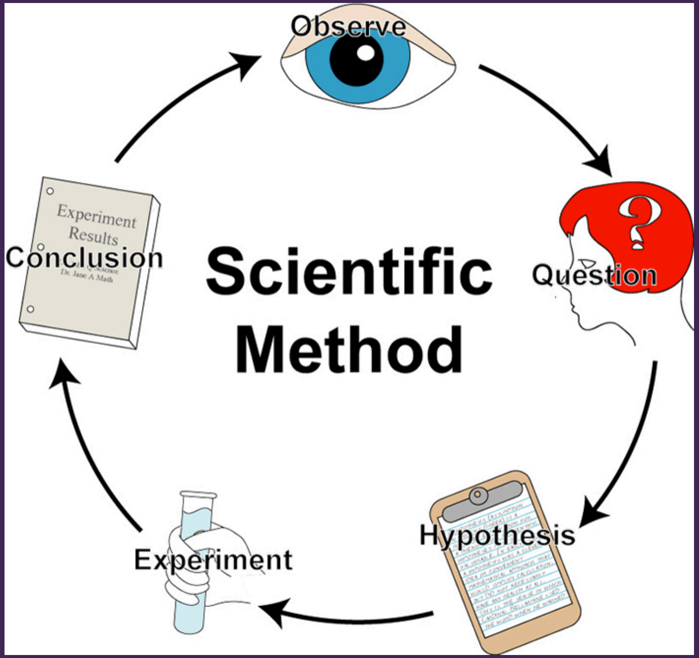
- Dig into Data Science & Machine Learning
  - Last week, we looked at DS & ML

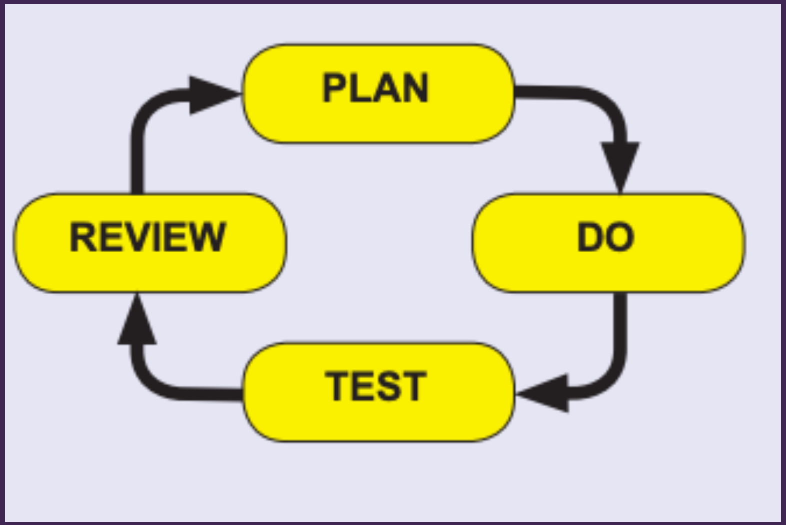


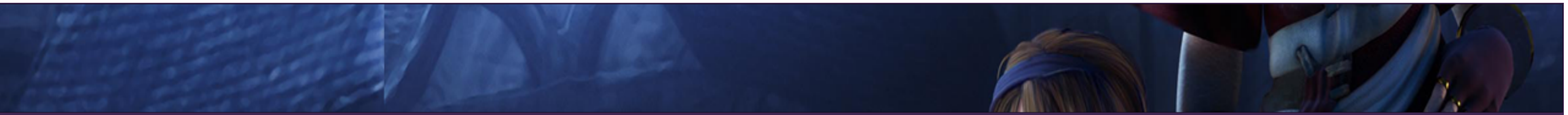
Data Science is the overall activity

Machine Learning is the fun part ;)

- Dig into Data Science & Machine Learning
  - Data Science as a process
    - Science is the key part of data science



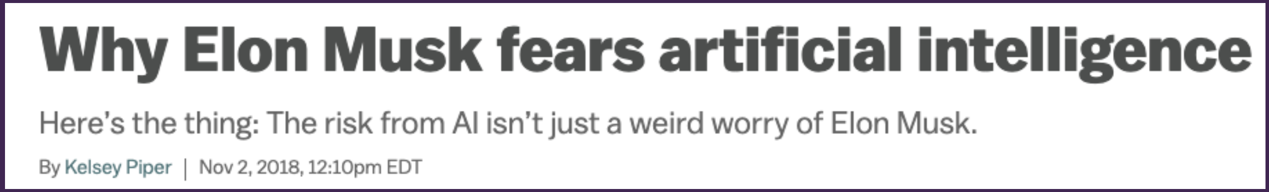




- Dig into Data Science & Machine Learning
  - Machine Learning as a process
    - Learning is the key part of Machine Learning
    - What does Searle tell us about learning in the Chinese Room?
      - Learning is ‘getting better’
      - Learning is not meaning
      - Learning is not knowledge
      - Often, a strong relationship between knowledge and introspection & reflective practice
        - » As much as I said domain experts will often use tactic or heuristic knowledge in their fields, they can normally go back to first principles to explain their processes (it may be painful)
      - Be mindful of domain specificity when talking about learning & knowledge ;)

- Dig into Data Science & Machine Learning
  - Machine Learning in practice

- Dig into Data Science & Machine Learning
  - Types of Machine Learning
    - Given that ML is the intersection of compsci and stats, we need to think of learning as *error reduction*
    - Already said that ML is not knowledge-based so ‘self-awareness’ is unlikely to happen



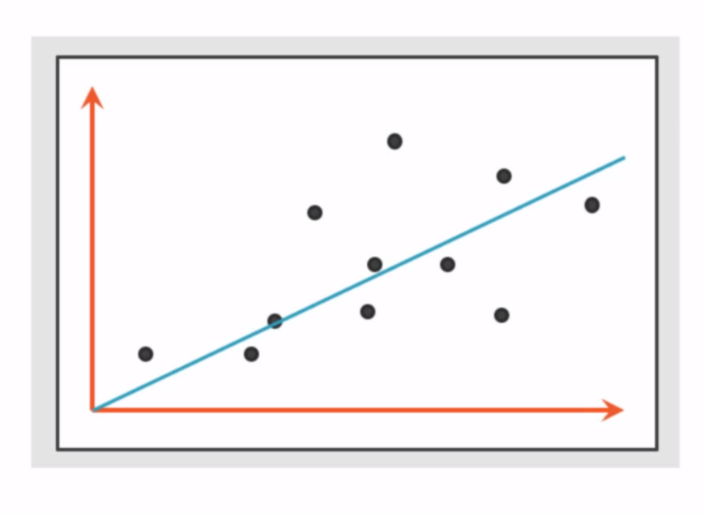
Why Elon Musk fears artificial intelligence

Here’s the thing: The risk from AI isn’t just a weird worry of Elon Musk.

By Kelsey Piper | Nov 2, 2018, 12:10pm EDT

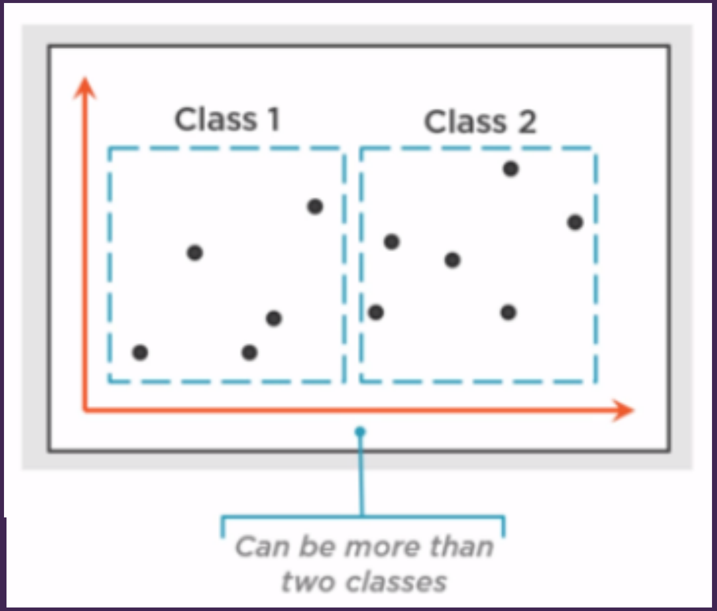
https://www.vox.com/future-perfect/2018/11/2/18053418/elon-musk-artificial-intelligence-google-deepmind-openai

- Dig into Data Science & Machine Learning
  - Types of Machine Learning
    - Regression
      - work out a continuous relationship between inputs and outputs



- Like standard regression, but for multiple dimensions
- E.g.:
  - Based on other examples, what is the value of something (house, car, watches, staff etc)
  - How much is a player likely to spend on my game through IAPs
  - When should I replace server equipment
- Typically, complex data is not ‘linear’, regression is generally split into buckets (short segments of linearity)

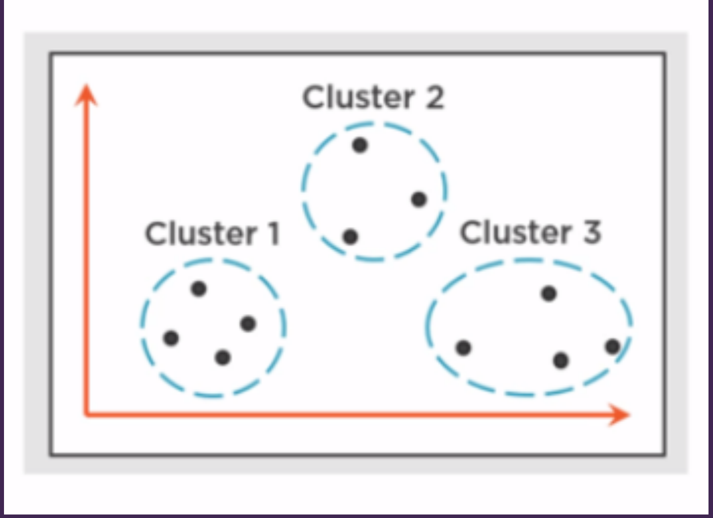
- Dig into Data Science & Machine Learning
  - Types of Machine Learning
    - Classification
      - Group inputs in to a single classification



- Data sorting
- E.g.
  - Mario example (classifications are which buttons should be pressed)
  - Is player behaviour normal?
  - Is a credit card transaction fraudulent?

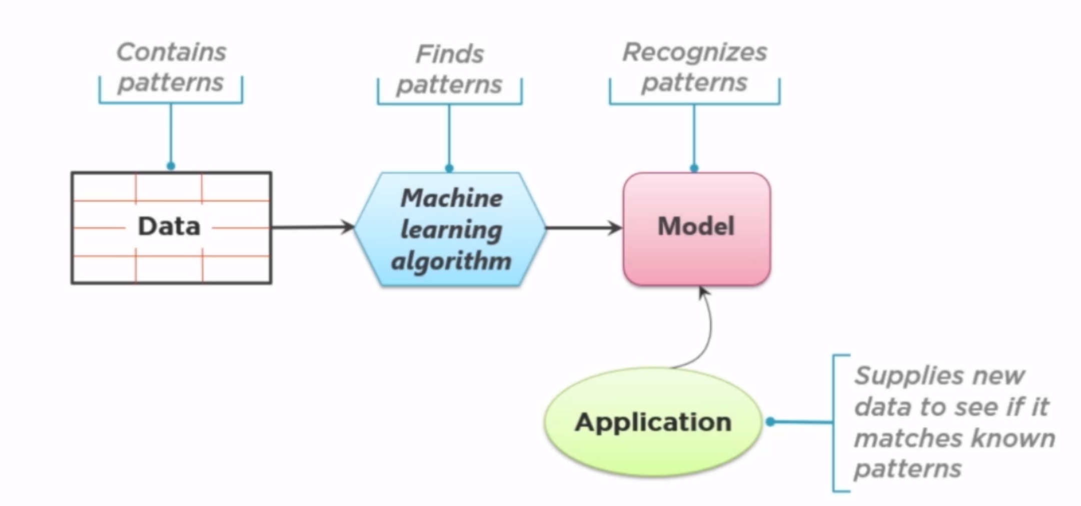
- This is typically 'classical machine learning'
  - Uses neural networks (back prop)
  - Or Marflow-style networks

- Dig into Data Science & Machine Learning
  - Types of Machine Learning
    - Clustering
      - Group *related* inputs in to distinct clusters / groups

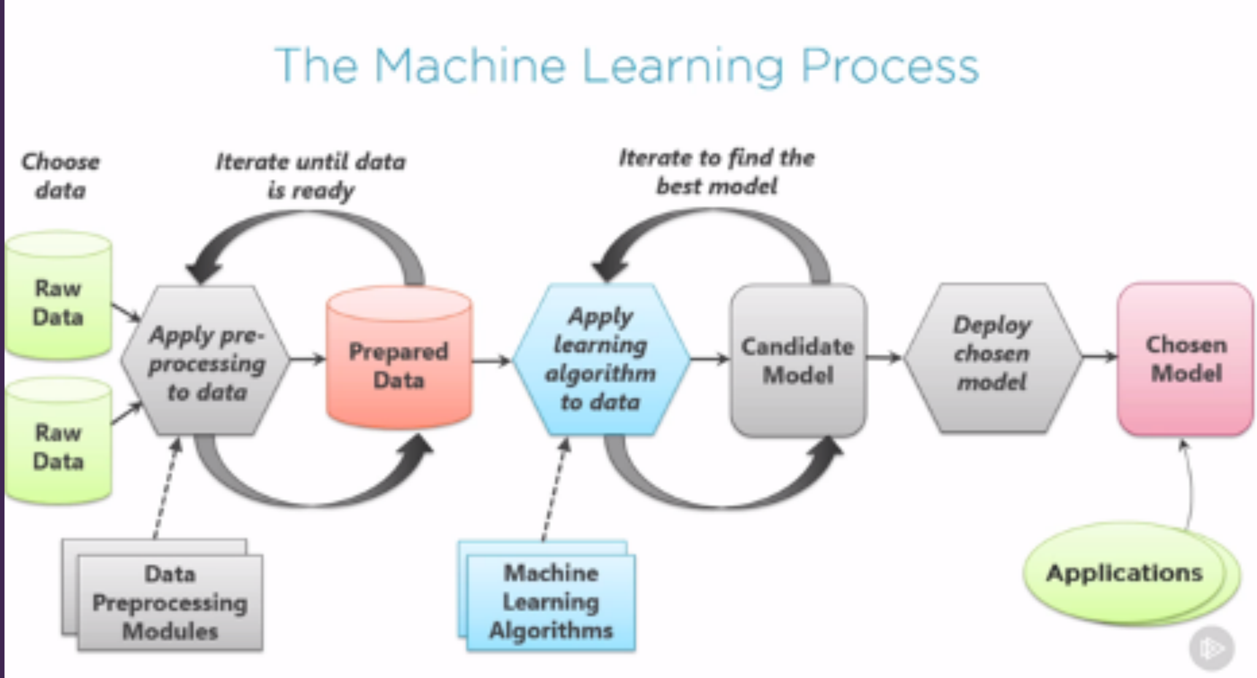


- (More) Data sorting
  - But the algorithms look to determine clusters without being told what they should be
  - Ideal for emergent clusters
    - Cooccurrence
    - Causality
- E.g.
  - What play styles do our players have (Bartle’s Player Types)

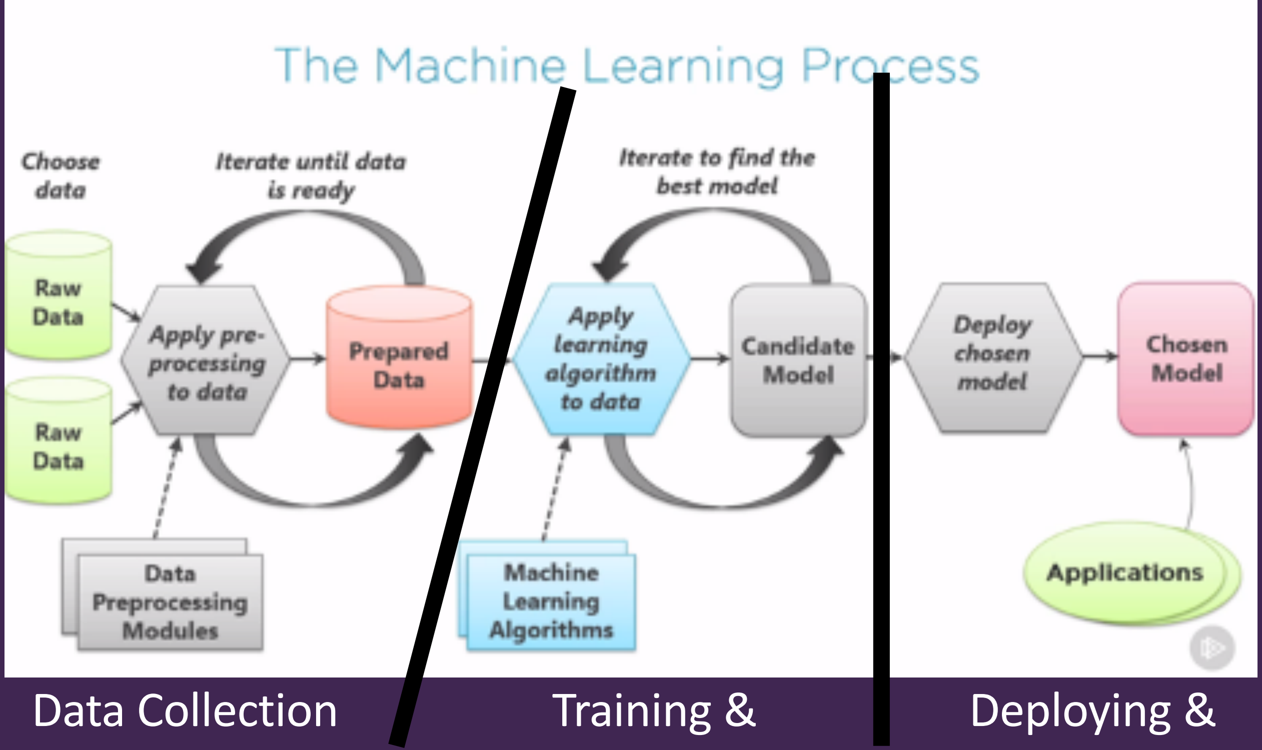
- Dig into Data Science & Machine Learning
  - Process of Machine Learning


Contains patterns
Finds patterns
Recognizes patterns
Data
Machine learning algorithm
Model
Application
Supplies new data to see if it matches known patterns

- Dig into Data Science & Machine Learning
  - Process of Machine Learning


The Machine Learning Process
Choose data
Iterate until data is ready
Iterate to find the best model
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model
Data Preprocessing Modules
Machine Learning Algorithms
Applications

- Dig into Data Science & Machine Learning
  - Process of Machine Learning


The Machine Learning Process
Choose data
Iterate until data is ready
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Data Preprocessing Modules
Iterate to find the best model
Apply learning algorithm to data
Candidate Model
Machine Learning Algorithms
Deploy chosen model
Chosen Model
Applications

Data Collection & Processing

Training & Learning

Deploying & Using

Note all the scope for iterating

- Dig into Data Science & Machine Learning
  - Process of Machine Learning
    - Data collection & processing
      - Collect enough suitable data for training & testing
      - Typically, this will be ‘live’ data that can come from a multitude of sources
        - Scope to use historic data too
      - Need to process the data into forms that make sense for training
      - This will all depend on type of ML used and algorithms used within that type

- Dig into Data Science & Machine Learning
  - Process of Machine Learning
    - Training & learning
      - ML describes two broad learning paradigms
        - » Supervised
          - Training is done with data + expected outcome (like classroom training)
          - Learning algorithm will look to minimise error between expected outcome and current outcome
          - Typically
            - Regression & classification
        - » Unsupervised
          - Training is done with data but no expected outcomes as algorithm will self-organise representations
          - Typically
            - Clustering

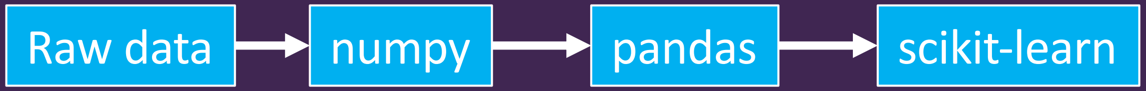
- Dig into Data Science & Machine Learning
  - – Process of Machine Learning
    - Training & learning
      - – Typically, data is split into two sets:
        - » Training data
          - In supervised learning
            - Algorithm is trained with training data to achieve desired / best (minimum) errors
          - In unsupervised learning
            - Algorithm is just presented with training data
        - » Test data
          - Novel test data is presented to the algorithm to assess performance with new cases
          - Performance can be assess quantitatively with a confusion matrix

- Dig into Data Science & Machine Learning
    - Process of Machine Learning
        - Iterating data collection & processing and Training & learning phases
            - Typically, testing an ML solution will result in issues
                - » Over-fitting
                    - training has been so heavily geared to training data, solution doesn’t perform well with novel data
                        - Learning the data and not the trends
                - » Under-fitting
                    - both training data and novel data produce poor results
                        - Algorithm may not be a good fit for data
                        - May not have enough data to train with

- Dig into Data Science & Machine Learning
  - Process of Machine Learning
    - Iterating data collection & processing and Training & learning phases
      - May require iteration in training and learning
        - » Change algorithm parameters
        - » Change algorithm
      - May require iteration in data collection & processing
        - » Process existing data differently
        - » Collect different data

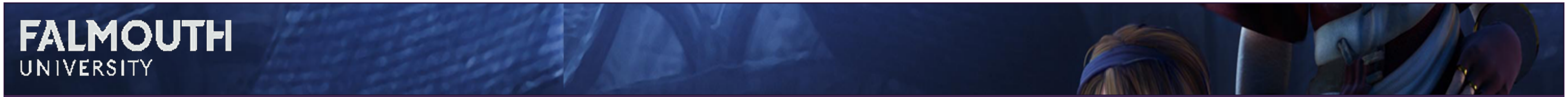
- Dig into Data Science & Machine Learning
  - Process of Machine Learning
    - Deploying and using
      - Once an algorithm has trained the data to acceptable performance (accuracy) it can be packaged and put into use
      - Take algorithm data and package into application
        - » In games and most applications, as a black box AI component

- Dig into Data Science & Machine Learning
  - Machine Learning in Python
    - Python has a lot of support for industrial and academic ML
      - Lots of package support
    - The pipeline we will use



Raw data → numpy → pandas → scikit-learn

- Other pipelines & parts are available
  - TensorFlow
  - ML.net
  - etc



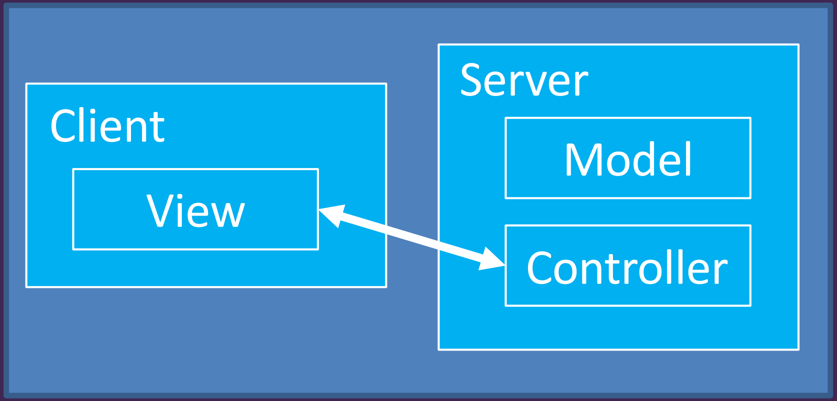
- Dig into Data Science & Machine Learning
  - Machine Learning in Python
    - Numpy
      - A library for ‘array & linear algebra’ for ‘large datasets’
    - Pandas
      - PANel DAta
      - Process data ‘like excel in code’
    - Scikit-learn
      - Library of ML algorithms

- Look at data acquisition

- Look at data acquisition
  - Two broad approaches to data acquisition
    - Live data (take from running services)
    - Historical data (take from offline sources)
  - Both require data processing (dependent on the data & solution requirements)

- Look at data acquisition
  - Live Data (analytics / metrics)
    - This is a common activity for GaaS and is a core part of their business
      - Regression
        - » Customer spend
      - Classification & clustering
        - » Customer behaviours
    - Relatively easy for GaaS, not so for traditional games

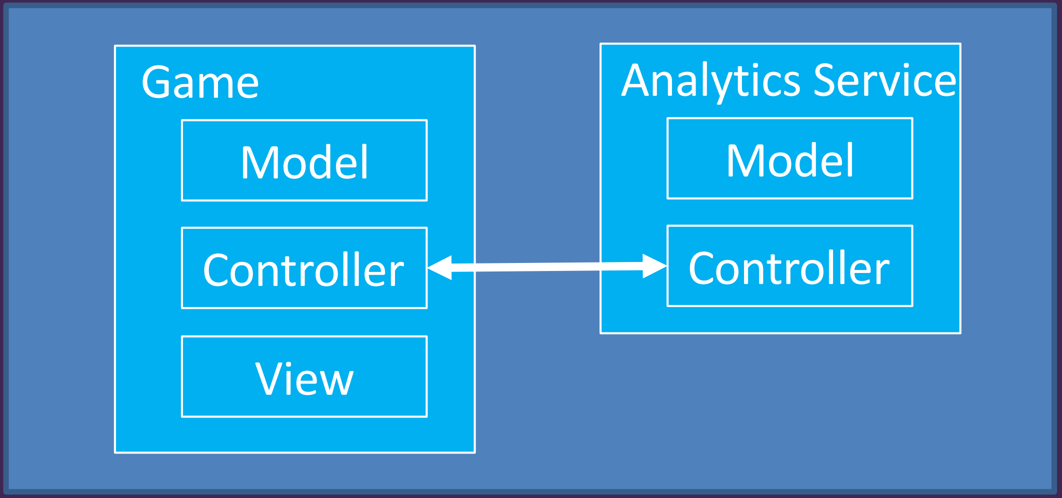
- Look at data acquisition
  - – Live Data (analytics / metrics)
    - Game As a Service (thin client)



– Typically, a thin client GaaS will have all the ‘important processing done on the server and just take user input and drawing on the client

» Easy to collect server-side data & instrument & re-instrument server code to collect different data

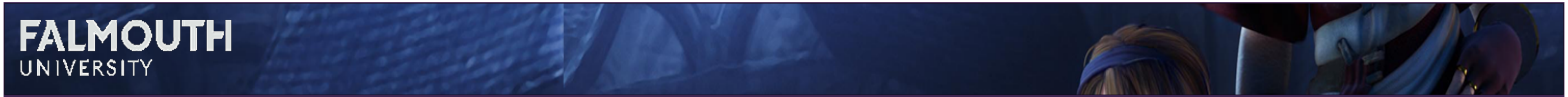
- Look at data acquisition
  - Live Data (analytics / metrics)
    - Standalone Game with analytics support



- Typically, standalone game will be a black box
  - » Only updated through patching
- (Potentially) costly to send data to analytics server
  - » Need to think about data packaging

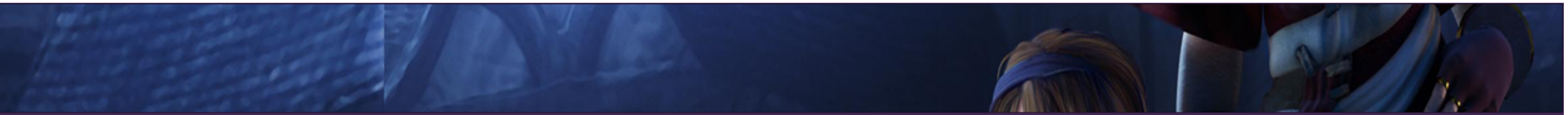
- Look at data acquisition
  - Approaches:
    - Save data to a local text file and process
      - Save data as csv, xls, xml, json etc
    - Send data to a server using HTTP (or other protocols)
    - Manage data as flat files
    - Manage data through xls, sql
      - Remember, ML has a tendency to create large amounts of data, so it needs to be stored carefully.

- Look at data acquisition
  - What to save?
    - All the keystrokes / game events
      - Can generate lots of data that may not process well
    - Do sessional processing locally & send results
      - Treat each play-through of game as a ‘session’ and send key results
      - Can work but relies on you having the correct data
    - Label game events and send
      - Again, relies on you having the correct data



- Look at data acquisition
  - We can see that ML & DS move the burden away from writing algorithms to working with data
    - What to capture
    - How to capture it
    - How to process it
  - Then worry about training

- Prepare for this week’s workshop



- Prepare for this week’s workshop
  - For this weeks’ workshop, we will look at different ways of collecting data from a game
    - Flat files
    - Json
    - Openpyxl
    - HTTP
  - And how we can store data from multiple players & sessions
    - Excel
    - Sql

- Do you have any questions for me