Please sit with your CTF teams and come up with a team name

Team 1

Correlate the Findings:
A Data Science CTF

What you'll take away

1. How Data Scientists interrogate a novel dataset
2. The *Exploratory Data Analysis* mindset
3. How to use your preferred toolset to quickly analyze data

Agenda

Part 1 (~20 minutes)

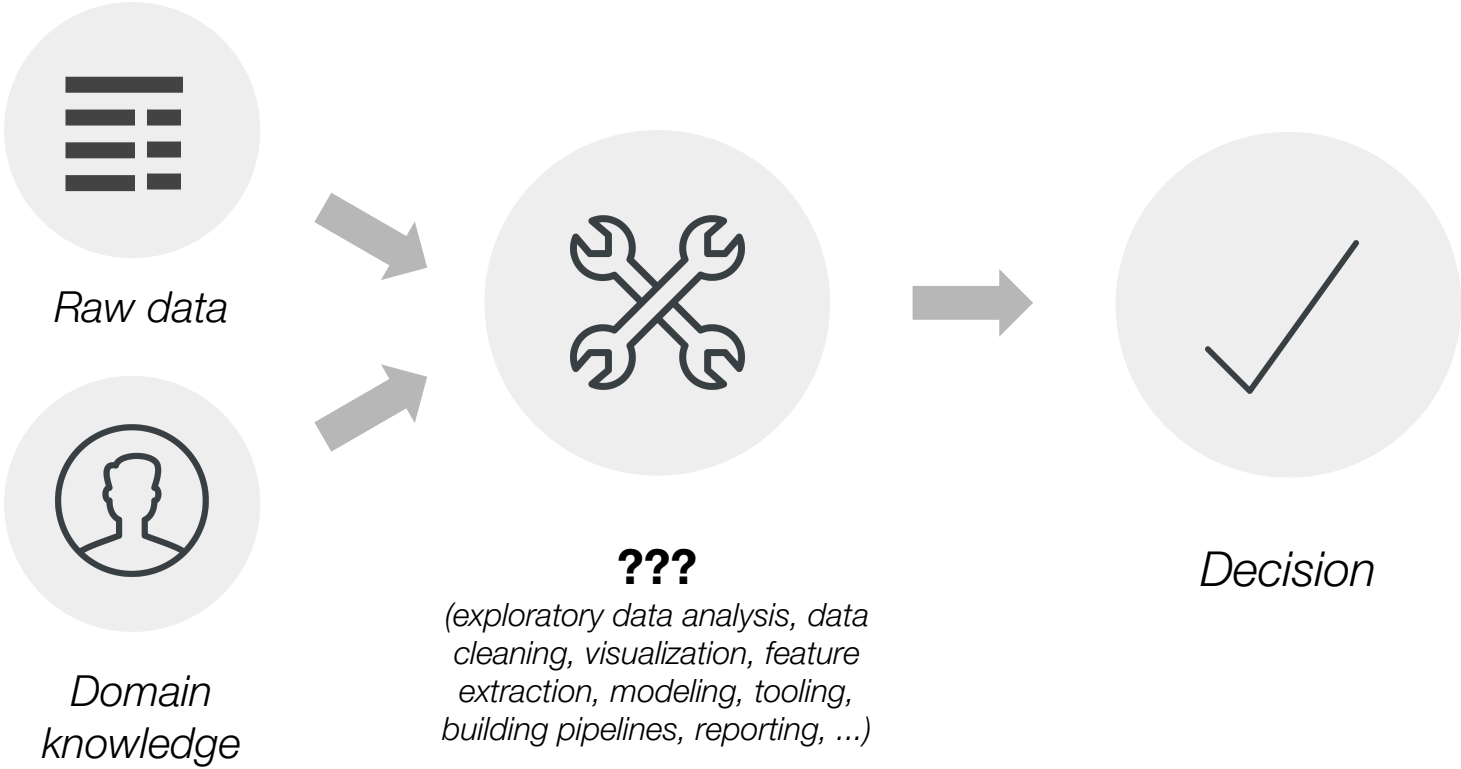
- Split into teams
- What is *Exploratory Data Analysis*?
- Introduce the CTF environment
- Discuss challenges

Part 2 (~50 minutes)

- *Let the challenge begin!* Teams work through data challenges to capture flags

Making data-informed decisions

Data to Decision



Data to Decision Examples

ENGINEER

Performance testing

▼

???

▼

Ship feature?

SECURITY OPERATOR

Browser usage trends

▼

???

▼

Change policies?

DESIGNER

A/B User behavior data

▼

???

▼

Implement UI change?

SRE

Logging data

▼

???

▼

Change thresholds?

Exploratory Data Analysis (EDA)

R for Data Science, Grolemund & Wickham
Doing Data Science, O'Neil & Schutt, 2014
Principles and Procedures of Exploratory Data Analysis, Behrens, 1997

No Firm Hypothesis, No Model

- **Crucial time reserved for asking “What is going on here?”**
- Your starting point is usually a *loose understanding* of the process that generate the data. You don’t have a coherent hypothesis or model yet.
- During EDA, you begin to connect that understanding with the data by trial and error.

How do you EDA?

> *EDA is an iterative cycle. You:*
>
> 1. *Generate questions about your data.*
> 2. *Search for answers by visualising, transforming, and modelling your data.*
> 3. *Use what you learn to refine your questions and/or generate new questions.*
>
> *[...]* ***EDA is a state of mind.***

- R for Data Science (Grolemund & Wickham)

1.

Ask questions

2.

Understand, transform and visualize data

3.

Use your acquired insights to ask new questions

Check, Transform, Visualize

Sanity Checking

What does each variable represent?
How was the data captured?
Do all the values make sense?
Are there missing values?

Covariation

Are two or more variables increasing or decreasing together?
Is the covariation systematic, or dependent on other factors?

Distributions

How are variables distributed?
What is the most common value?
What is the typical range of values?
Are there anomalies?
Discrete values or continuous?

Patterns, Clusters, Models

Do different groups have different distributions or covariation between variables?
Does variables cluster in any way?
Do variables follow a simple mathematical model?

Tools of the trade

- **Curiosity and Skepticism**
- **Simple data visualizations**
  Bar charts, scatter plots, histograms, box-plots
- **Transformations**
  Operations on one or more columns that uncover information
- **Summaries**
  Compute quantiles, average, standard deviation, min, max, ...
- **Filtering and Grouping**
  Compare summaries and visualizations on different subsets of the data

Rules of thumb

- **Start small**
  Prefer simplest possible visualizations and data operations.
- **Keep it nimble**
  Resist the urge to polish. You are creating a report for *yourself*.
- **Document, document, document**
  Insights uncovered during EDA inform more rigorous work later.

Capturing Data Flags

EDA Capture the Flag Fun Time

Accelerate learning with:

Gamification

Teamwork

The Challenges

Challenge Datasets

- **Intro dataset** — Guided questions based on our tutorials.
- **Movie Ratings** — Fictional dataset of ratings for a superhero movie.
- **Challenge death** — Coroner's inquests from 18th century Westminster.
- **Jeopardy** — Game results from *Jeopardy!* episodes.

Logistics

Tools

Use any environment you're good at. We encourage the following:

Unfortunately we won't have time to teach a new environment from scratch. Use what helps you solve challenges the quickest!

Feel free to ask us for help at any point.

Introduce the scoring platform

Where are the goods?

Links to the scoring platform and datasets

Register an account and create your team

Teams

Team 1

Learning together

Coordinate with your teammates on how to collaborate!

Options:

- Each person can take a dataset and category of flags.
- Pair up to solve challenges together.

At least let your team know what you’re working on!

Let the challenge begin!

Start with the *Intro Dataset* category.