UWE Bristol | University of the West of England

Presentation by

Dr. Phil Legg

Associate Professor in Cyber Security

Security Data Analytics and Visualisation

2: Analytics Workflow

Date: Autumn 2019

Recap

Four key points from last week:

1. Why do we need analytics and visualisation in cyber security? Understanding of situational awareness, decision-making and rapid response.
2. What are the difficulties in analysis? Are we looking at the right thing, what's the story we are trying to tell
3. What are the difficulties in visualisation? Are we looking at the right data, are we telling our story clearer for the end-user
4. What kind of data may we be looking at? Network captures, but also computer interactions, written messages, user activity, etc.

UWE Bristol University of the West of England

A Data Analytics Pipeline

UWE Bristol University of the West of England

Data analytics pipeline

Data analytics pipeline

Where do we collect data from?

- Organisational data
- Online resources
- Sensors (e.g., IoT)
- Image / Video sensors

Data analytics pipeline

How do we store data?

- File formats (e.g., CSV, JSON)
- Databases
  - MySQL
  - MongoDB

Data analytics pipeline

How do we combine different data sources?
How do we transform data sources to tell our story?

- Aggregation
- Feature extraction
- Data Pre-processing

Data analytics pipeline

How do we analyse data?

- Classification of data
- Prediction or forecasting from data
- Clustering of similar data attributes
- Outlier / anomaly detection

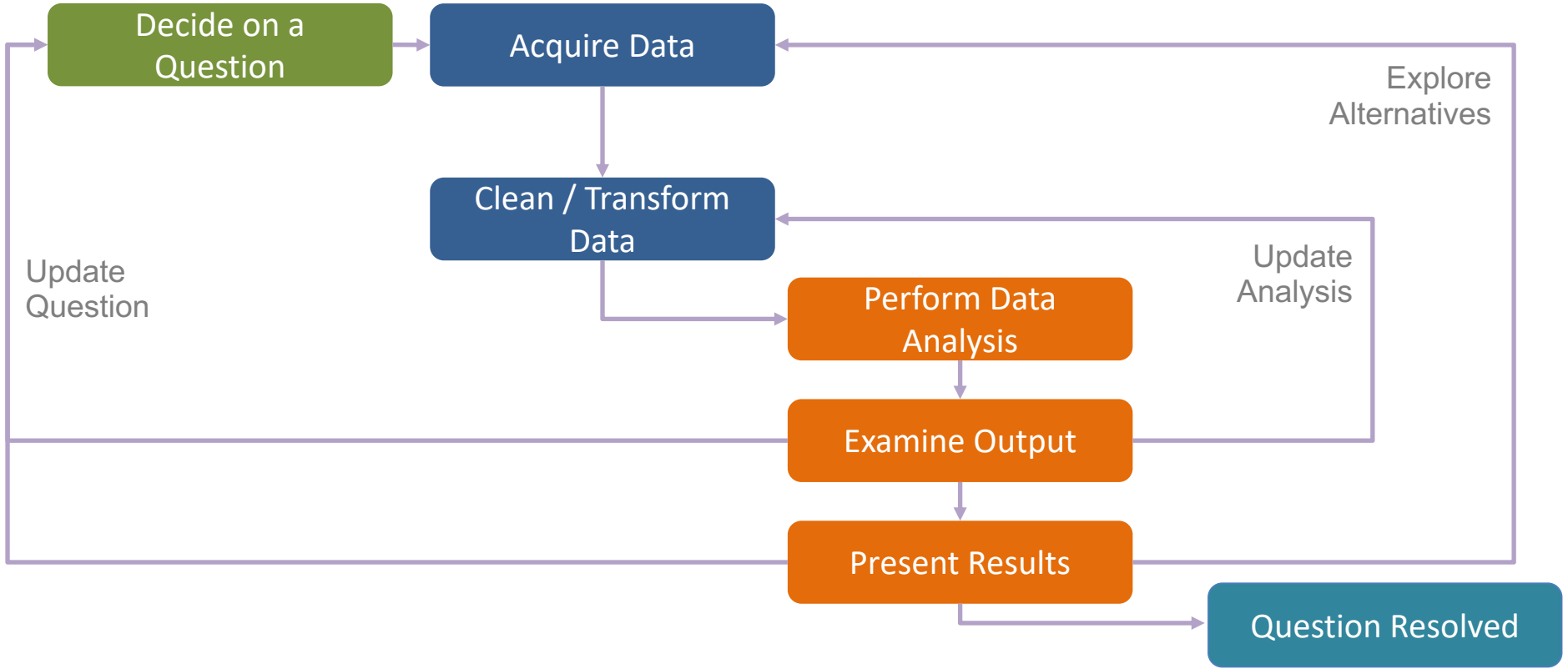
Data analytics pipeline

How do we visualize data?

- 2-dimensional charts and plots
- 3-dimensional data representations
- Focus-and-Context
- **Interaction**

Is it always that easy..?

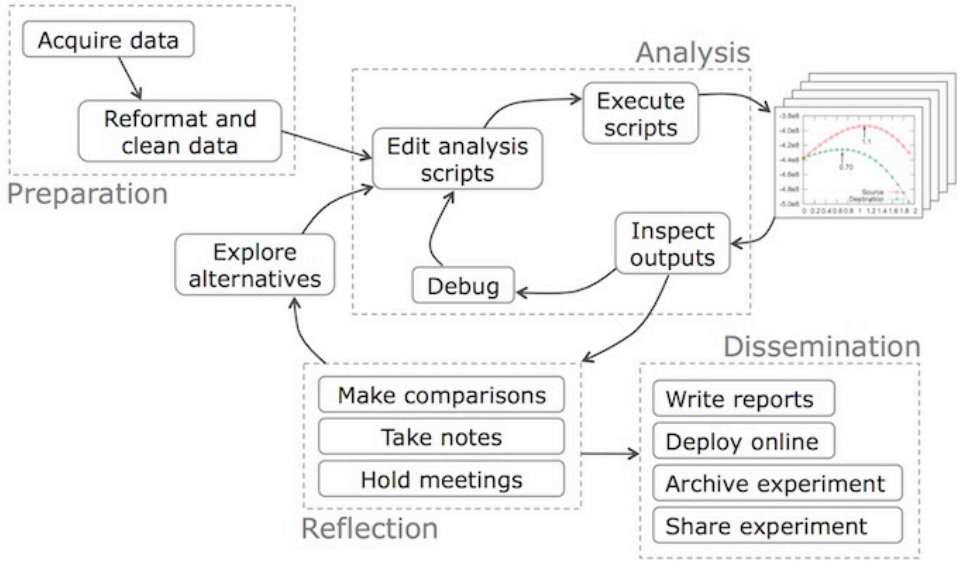
Data Science Workflow



Data Science Workflow

…yet another workflow!

The figure below shows the steps involved in a typical data science workflow. There are four main phases, shown in the dotted-line boxes: *preparation* of the data, alternating between running the *analysis* and *reflection* to interpret the outputs, and finally *dissemination* of results in the form of written reports and/or executable code.



Knowledge Generation using Visual Analytics

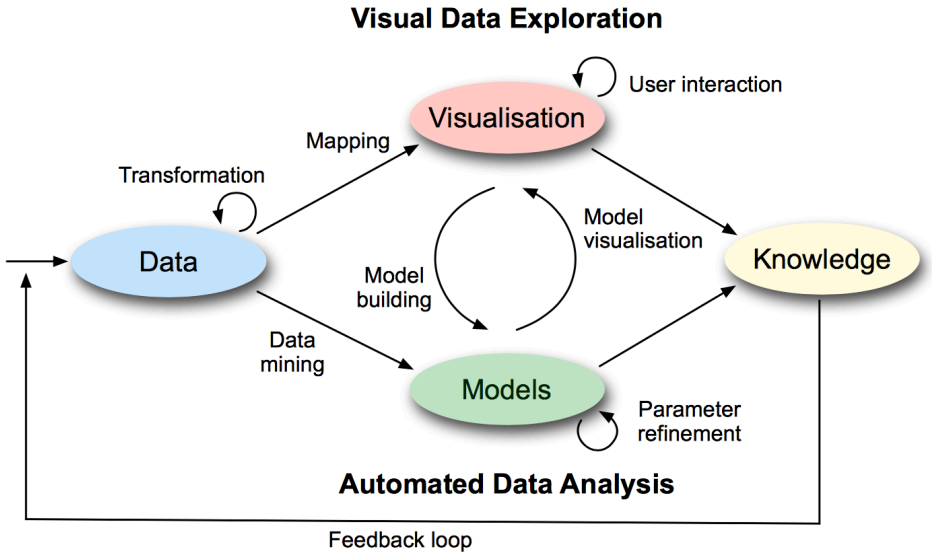


Figure 2.3: The visual analytics process is characterised through interaction between data, visualisations, models about the data, and the users in order to discover knowledge

D. Keim. Mastering the Information Age-Solving Problems with Visual Analytics, 2010.

UWE Bristol University of the West of England

... No.

Analytics and Visualisation is an iterative process

“Visual Analytics Loop”

UWE Bristol | University of the West of England

Noname manuscript No.
(will be inserted by the editor)

The Human is the Loop: New Directions for Visual Analytics

Alex Endert · M. Shahriar Hossain · Naren Ramakrishnan · Chris North · Patrick Fiaux · Christopher Andrews

Received: date / Accepted: date

Abstract Visual analytics is the science of marrying interactive visualizations and analytic algorithms to support exploratory knowledge discovery in large datasets. We argue for a shift from a ‘human in the loop’ philosophy for visual analytics to a ‘human is the loop’ viewpoint, where the focus is on recognizing analysts’ work processes, and seamlessly fitting analytics into that existing interactive process. We survey a range of projects that provide visual analytic support contextually in the sensemaking loop, and outline a research agenda along with future challenges.

A. Endert
Pacific Northwest National Laboratory
Richland, WA 99352, USA
E-mail: alex.endert@pnnl.gov

M. S. Hossain
Department of Mathematics and Computer Science
Virginia State University, Petersburg, VA 23806, USA
E-mail: mshossain@vsu.edu

N. Ramakrishnan
Department of Computer Science
Virginia Tech, Blacksburg, VA 24060, USA
E-mail: naren@cs.vt.edu

C. North
Department of Computer Science
Virginia Tech, Blacksburg, VA 24060, USA
E-mail: north@cs.vt.edu

P. Fiaux

1 Introduction

The emerging fiel
the needs of explo
The approach is t
bilities of analyti
ities of interactive
data is too large
the use of data pr
tasks are too exp
ods, requiring the
visualization as a
the data. This ap
understanding of
Thus, visual analy
ciplinary approac

Targeted task
are exploratory in
defined or unknov
available. Tasks a
translated into op
of the analysis. F
ness analysis, ana
amounts of textu
sense of. Stasko
and visualizations
mation, eventuall
stand the actual t

A Visual Analytics Loop for Supporting Model Development

Simon Walton* Oxford University
Eamonn Maguire† CERN
Min Chen‡ Oxford University

Abstract— Threats in cybersecurity come in a variety of forms, and combating such threats involves handling a huge amount of data from different sources. It is absolutely necessary to use algorithmic models to defend against these threats. However, all models are sensitive to deviation from the original contexts in which the models were developed. Hence, it is not really an overstatement to say that ‘all models are wrong’. In this paper, we propose a visual analytics loop for supporting the continuous development of models during their deployment. We describe the roles of three types of operators (monitors, analysts and modelers), present the visualization techniques used at different stages of model development, and demonstrate the utility of this approach in conjunction with a prototype software system for corporate insider threat detection. In many ways, our environment facilitates an agile approach to the development and deployment of models in cybersecurity.

1 Introduction

Since the arrival of the Internet, algorithmic models have been widely used in a variety of applications such as virus detection, spam filtering and intrusion threat detection. For virus detection, models play an overwhelmingly dominant role in protecting systems across the globe. In some cases, such as spam filtering, models provide the first line of defense, but frequently rely on humans to correct false positives and negatives. In more complex cases, such as insider threat detection, human reasoning is absolutely essential as models are yet to reach the same level of utility as in virus and spam detection.

Axelsson [3] grouped intrusion detection models into the two categories of *anomaly detection* and *signature detection*. This categorization can be broadly applied to other detection problems. Anomaly detection models attempt to find deviations from the norm in a dataset; that is, data points that do not conform to what the model has been taught or instructed is ‘normality’. Such a deviation hints a possible threat that requires human operators’ attention. An exhaustive survey of anomaly detection techniques can be found in Chandola *et al.* [8]. Signature detection models detect patterns that are the same as or similar to those of known threats. Depending on the confidence level of detection, automated or semi-automated actions may be activated.

Recent news reports of large-scale intrusions into corporate networks have highlighted the requirement for real-time protective network monitoring, and the increasing scale of data necessitates a filtering or highlighting mechanism such as an intrusion detection model. The use of such models however is fraught with a number of problems, such as the possibility of being undermined [31], difficulties encoding scores’) or an aggregation scheme (e.g., ‘give me the mean of model’s intrusion scores’).

The literature on intrusion detection reveals that visualizati playing an active role in observing model outputs and also in obse and analyzing the performance of models under development. In many domain experts have commented that a combined approa visualization and detection provides a robust and scalable solutio protective monitoring and analysis [32, 14]. This is particularly when using models with either unknown performance or known performance[1].

Visual analytics provides a useful workflow for exploring the puts of models and refining those models based on the visualiza and the expertise of the human analysts. However, many workf deployed in practice suffer from a number of shortcomings. V models are less reliable, *(i)* model performance is usually not itored; *(ii)* analysts have no means to improve models as they developed by a remote third party; and *(iii)* analysts must either s a substantial amount of time dealing with false positives or ado attitude to disregard them. Within an organization, often no one countable for the failures of models, and hence the failures of intr detection. If models were humans, such a status quo would nev tolerated.

In this paper, we propose a scalable visual analytics loop for porting the continuous development of models in protective mo ing environments. Our loop defines the three distinct operator gories of *monitor*, *analyst* and *modeler*, and provides each with di

Transformation of an Uncertain Video Search Pipeline to a Sketch-Based Visual Analytics Loop

Philip A. Legg, David H.S. Chung, Matthew L. Parry, Rhodri Bown, Mark W. Jones, Iwan W. Griffiths, and Min Chen

Abstract—Traditional sketch-based image or video search systems rely on machine learning concepts as their core technology. However, in many applications, machine learning alone is impractical since videos may not be semantically annotated sufficiently, there may be a lack of suitable training data, and the search requirements of the user may frequently change for different tasks. In this work, we develop a visual analytics systems that overcomes the shortcomings of the traditional approach. We make use of a sketch-based interface to enable users to specify search requirement in a flexible manner without depending on semantic annotation. We employ active machine learning to train different analytical models for different types of search requirements. We use visualization to facilitate knowledge discovery at the different stages of visual analytics. This includes visualizing the parameter space of the trained model, visualizing the search space to support interactive browsing, visualizing candidature search results to support rapid interaction for active learning while minimizing watching videos, and visualizing aggregated information of the search results. We demonstrate the system for searching spatiotemporal attributes from sports video to identify key instances of the team and player performance.

Index Terms—Visual knowledge discovery, data clustering, machine learning, multimedia visualization

1 Introduction

With the advances in video capture and storage, many application areas have come to rely on collecting vast amounts of video data for the purpose of archive and review [16]. For instance, CCTV security systems can be capturing up to 24 hours of video each day from multiple camera sources. Likewise in sport, analysts will collect video data from multiple cameras for every match of a season. They may also be collecting and reviewing match videos from opposition teams to analyse traits in their performance that can then be exploited in their own strategy. This typically results in hundreds of hours of video footage (in our case 85TB and growing). The task of then reviewing this video content in order to identify interesting or important passages of play now becomes a significant challenge for an analyst to perform. Often the analysts will need to collect together a selection of example video clips that illustrate team strategy, to then analyse whether it resulted in positive or negative outcome for the team. However, identifying such clips from the video soon becomes a tedious and laborious task for the analysts to perform manually.

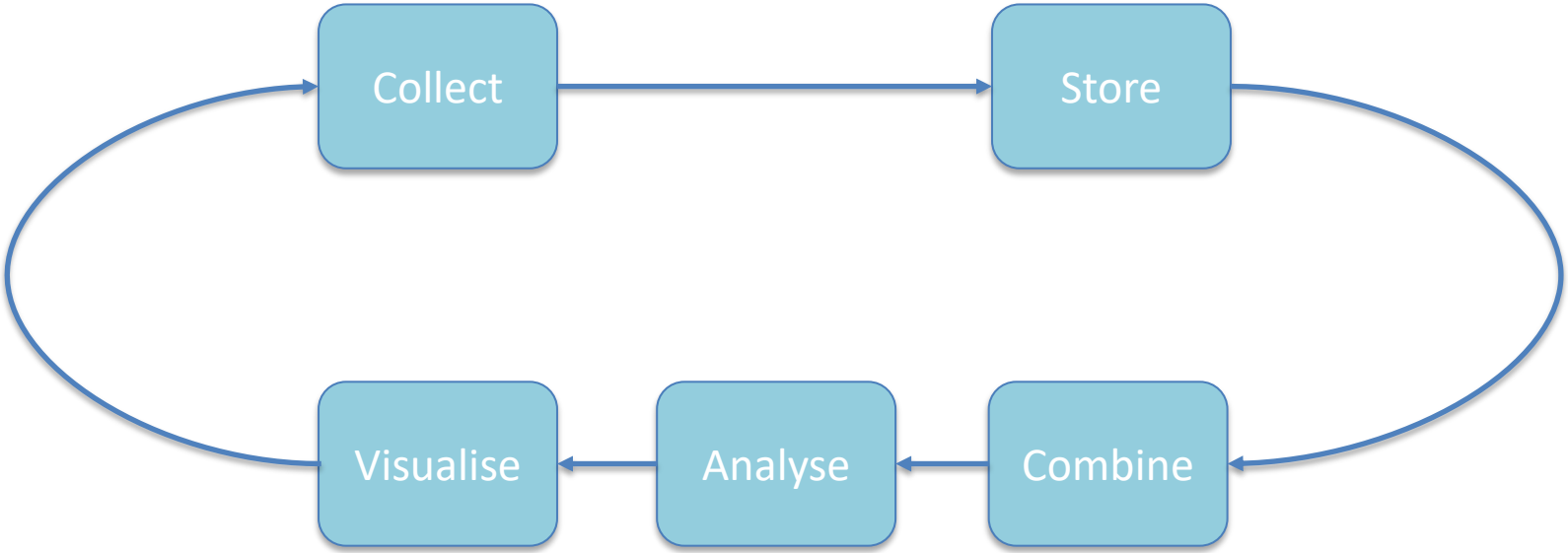
Many professional sports currently adopt an approach known as *notational analysis* [22]. Notational analysis is a process that allows for the occurrence of time-based events to be manually recorded in correspondence to a synchronised video capture that can then be reviewed this approach is widely used, it is a highly intense process for the analysts to conduct during the game when their expertise could be utilised much more effectively for crucial decision-making. Most importantly, notational analysis does not offer any means of directly searching the video content; it merely allows the user to search the tagged annotations that are associated with a video timestamp. Neither does it allow the user to obtain any more information than what was initially recorded in the first instance.

Video search enables finding segments from a collection of videos based on particular search criteria. One approach to video search is to use a sketch-based search, whereby through sketching the user can indicate a particular spatio-temporal pattern that can be associated with the video content. For instance, the user may sketch out a particular path of motion to find when people travel in that direction, or draw a region to query when ‘action’ occurs in that area. Developing a system that can support such open search queries poses some challenges. Firstly, the parameter space of possible sketches that the user could perform is significantly large. Secondly, the result expected by the user may differ from that returned due to other factors that can be difficult either for a user to encode them into the sketch, or for a system to interpret. Hence, the video search pipeline introduces a substantial

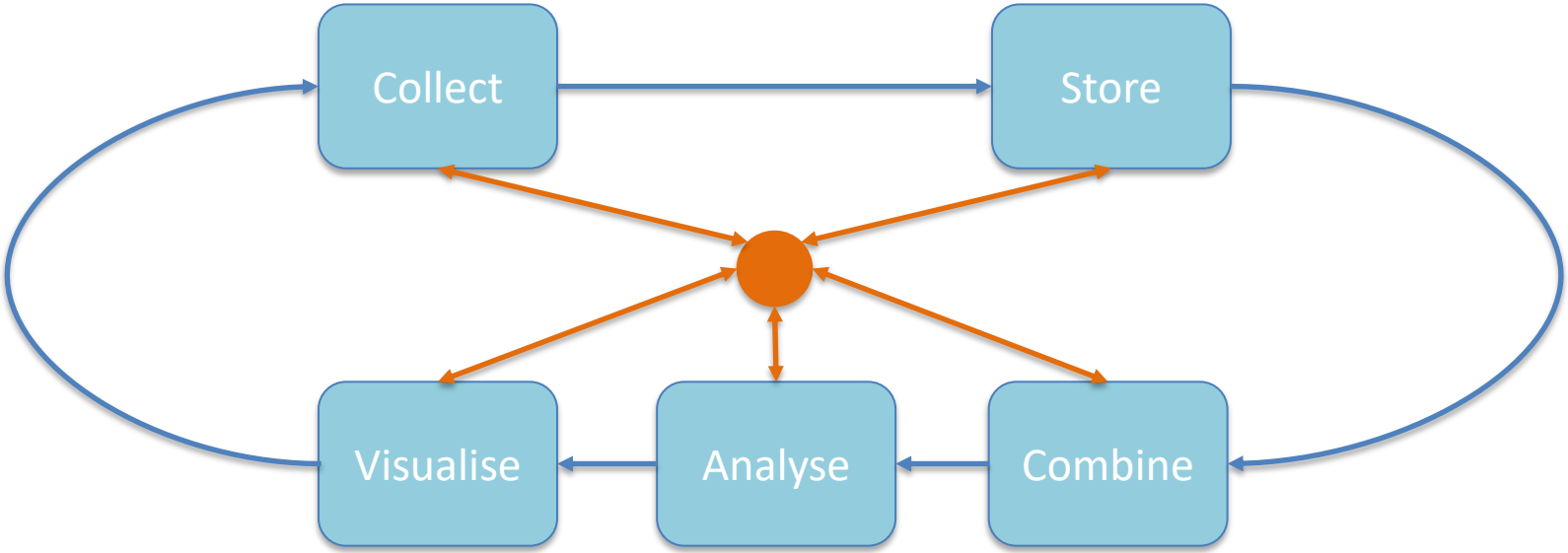
UWE Bristol University of the West of England

Data analytics pipeline

Data analytics loop



Data analytics loop



UWE Bristol University of the West of England

“Hello, Security Analytics”

What is ”Security Data”?

- Depends on who you are and what you want to achieve
  - Malware analysts -> process, memory and system binary dumps
  - Vulnerability researchers -> patch releases
  - Network analysts -> network traffic

- Other forms?
  - Cyber-psychologists may also be interested in human behaviour.
    - Movement data, written word (e.g., e-mail, social media)

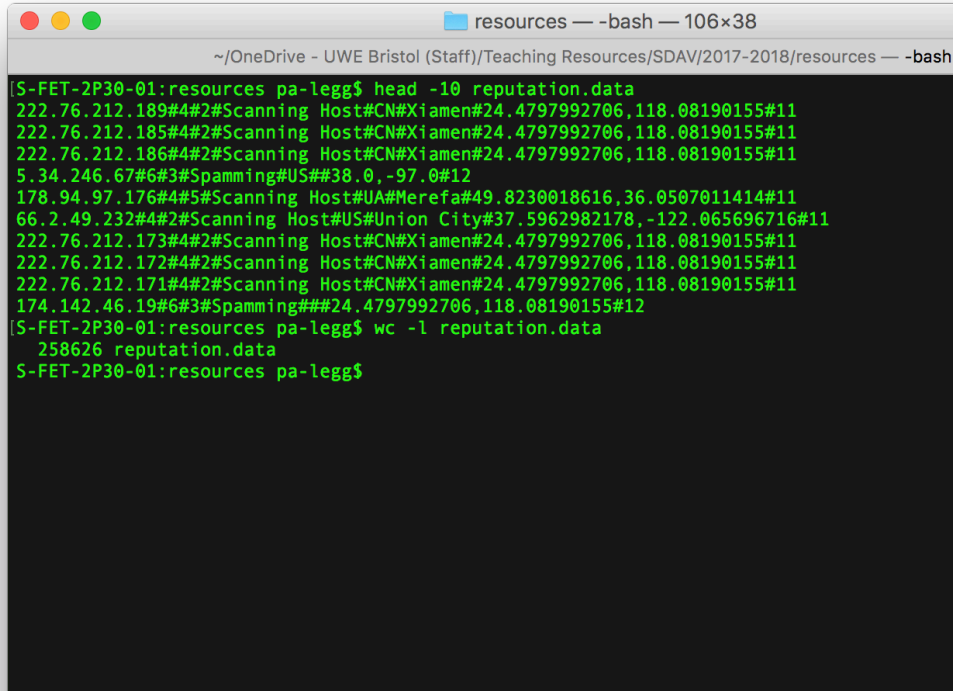
What's the Problem?

- For this "Hello World" example, we are working with a Security Operations Center (SOC). It seems the SOC analysts are becoming inundated with "trivial" alerts ever since a new data set of indicators was introduced into the Security Information and Event Management (SIEM) system. They have asked for our help in reducing the number of "trivial" alerts without sacrificing visibility.

- This is a good problem to tackle through data analysis, and we should be able to form a solid, practical question to ask after we perform some exploratory data analysis and hopefully arrive at an answer that helps out the SOC.

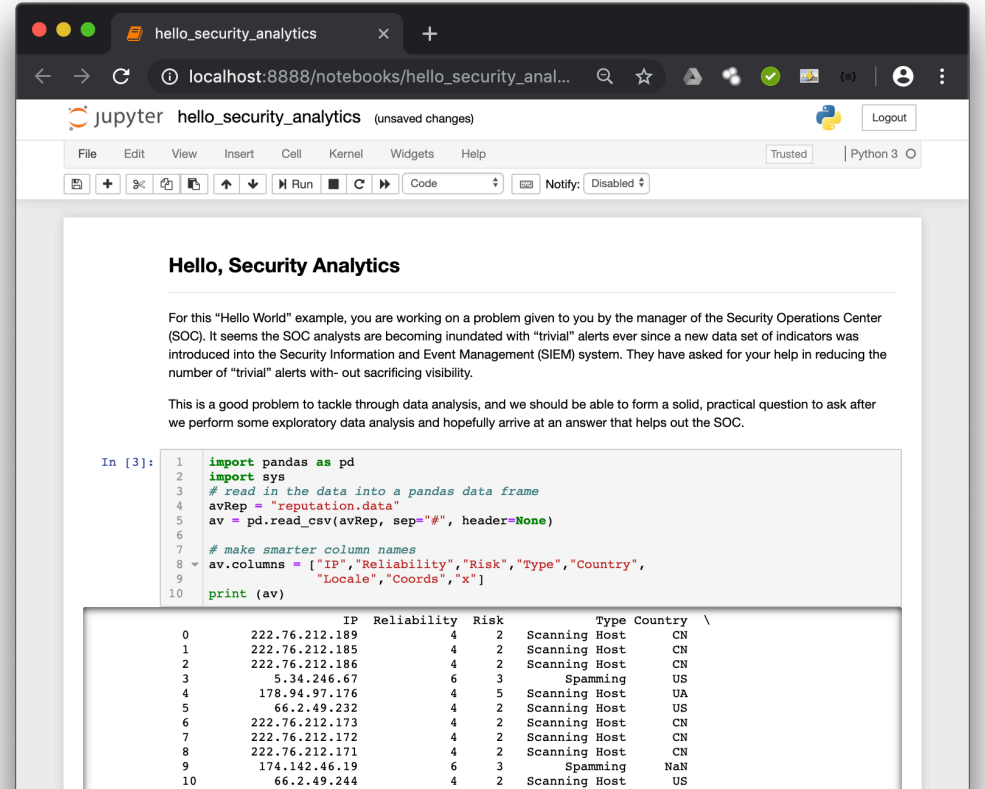
Attributes of the Data

- "head -10 reputation.data"
  - # See what the first 10 lines look like
- "wc -l reputation.data"
  - # Count number of lines in file (258626 lines)



Introducing Jupyter

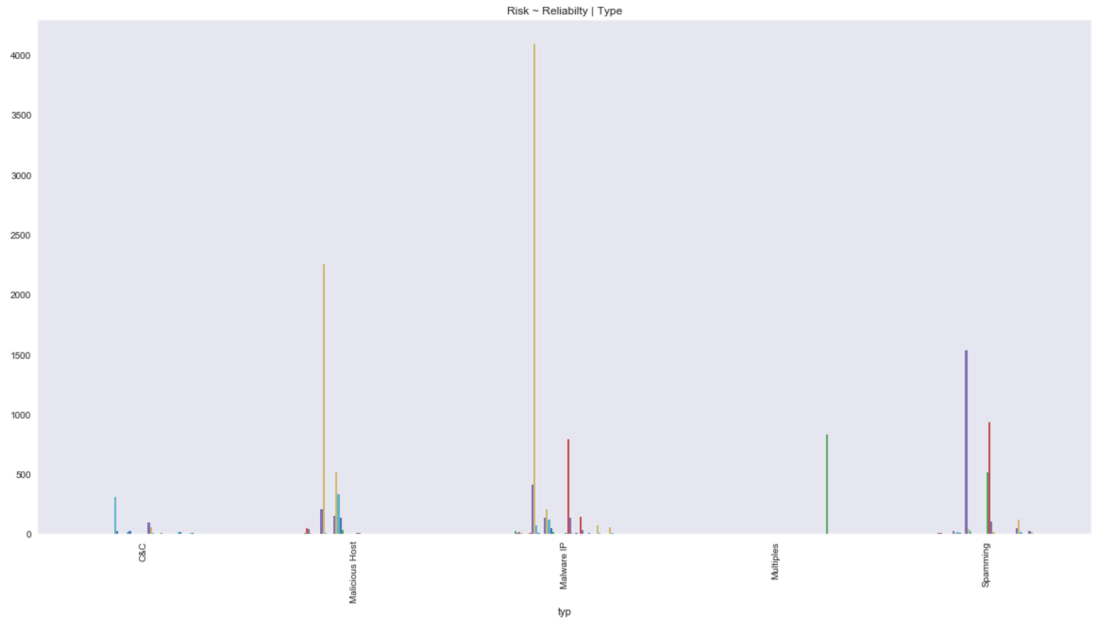
- Jupyter Notebook
  - Interactive Python notebook environment for combining code, figure, output and documentation
- `pip3 install jupyter-notebook`
- `jupyter notebook`
- Look at "*hellosecurityanalytics*" on Blackboard



Analysis

We have filtered the data to less than 6% of the original size.

We have identified the number of occurrences of risk/reliability across the different ‘types’.



Takeaway

- What are the different stages of a data analytics pipeline?
- What are the different stages of a data science workflow?
  - Iterative processes between machine and user
  - "Human-in-the-Loop"
    - How do we make best use of the machine **and** the best use of the human?

- Using the Jupyter notebook environment and trying out a sample data exploration

- **Worksheet 1 available on Blackboard for lab sessions**