The Center of Applied Data Science

Data Star March Cohort

Introduction to Data Science

March 2019

ABOUT TEAM CADS

Our Growth Story

2014

CADS was started with the goal to be the voice of Data Science & Analytics in ASEAN

2015 & 2016

Built Strategic Partnerships with world leading institutions to define and grow the Data Analytics industry

2017

Produced more than 600 data professionals through our enablement programs

Building Partnership with Coursera to strengthen scale & ease of Curriculum Delivery

Advised government bodies & corporate entities from various verticals on the strategy & implementation of ideal data solutions

WHAT WE OFFER

The Center of Applied Data Science

CADS TRAINING & EDUCATION

We focus on training, education, coaching, developing and mentoring the future data scientists of Malaysia

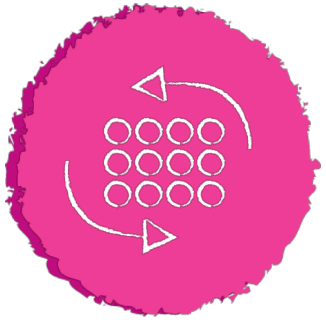
CADS ANALYTICS ADVISORY

Providing Data Analytics consulting services and solutions for businesses and governments

CADS TALENT DEVELOPMENT

We are focused on sourcing, developing and recruiting Data Science Talent for multiple industries

CADS DATA EXCHANGE

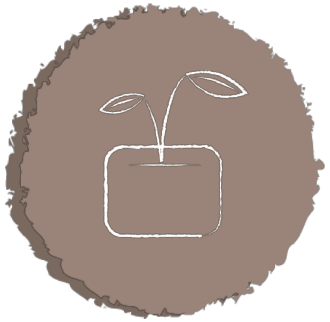


Managing and implementing Malaysia's first platform and exchange for Data Science initiated by MDEC (Malaysia Digital Economy Corporation)

CADS APP & TECHNOLOGY DEVELOPMENT

We look at innovations, technological developments and app development within the Data Science and Analytics space

CADS VENTURES



Identifying markets opportunities to invest in and incubate start ups

Darren Chong

Data Scientist

Functional Expertise	• Data Visualization • Data Storytelling • Python Programming • Data Analytics
Industry Expertise	• Financial Services • Shared Services • Management Consulting
Qualifications	▪ MSc. Management (Distinction), University of Warwick, UK ▪ BA (Hons) Business & Management, University of Exeter, UK

darren@thecads.com

Darren Chong

Darren is an MSc. Management graduate with Distinction from Warwick Business School, UK. He is a HRDF certified trainer with prior engagements in Financial Services and Corporate Training. At the Center of Applied Data Science Darren is interested in all aspects of the Data Science life cycle but is particularly focused on the extraction & communication of actionable insight from data to decision-makers. He is currently interested in Data Visualization & Data Storytelling and is an acting leader of the Tableau User Group Malaysia based in Kuala Lumpur.

Data Scientist, The Center of Applied Data Science

Led the development of the Enterprise Data Professional (EDP) Learning track targeted towards the development of data literacy in three areas: Excel Analytics, Data Storytelling & Data Visualization

Data Scientist Trainee, ASEAN Data Analytics Exchange

A 6-month paid finishing school of intensive data science enablement and mentorship with experienced Data Scientists. The course covered the fundamentals of Python, R, SQL, Big Data Technologies & Machine Learning.

Management Consultant Intern, EY Advisory Services

Delivered quantitative analysis to support strategic transformation of a large South East Asian government department. Winner and Team Leader of EY Annual Group Case Study Competition; awarded 'Best Presenter' of event.

Let’s get to know each other

Introductions

1. Educational Background
2. Prior work experience
3. Statistics & Programming (1-5)
4. Expectations for the Data Star Program

Agenda

Feel free to interrupt and ask questions at any point!

The Center of Applied Data Science

1. **Data Star Program**
2. **What is Data Science?**
3. **What are Data Scientist?**
4. **DDO Model**
5. **Introduction to Anaconda & Jupyter Notebooks**

Data Star Program

Begin your Data Analytics career

The Center of Applied Data Science

Extensive Data Science enablement training and mentorship with experienced Data Scientists.

01 Industry-centric Data Science enablement

02 Mentorship with experienced Data Scientists

03 Gain relevant industry experience

04 Opportunity to be part of a growing industry

05 Competitive income potential if hired

06 Earn allowance during industry placement

Upon completion participants undergo a 6 to 12 months placement with selected industry partners

Data Star Program

Three different learning paths

Data Analyst (DA)

The objective of the program is to equip the participants with the ability to set-up and run analysis using R, analyse data correctly, communicate and visualise findings well. Participants will also be taught to extract insights and knowledge from data.

Junior Data Scientist (JDS)

The objective of this program is to allow participants to expose the interdisciplinary field of data science. The program teaches participants how to conduct or approximate experiments, analyse data rigorously, develop predictive models and monitor model output.

Data Engineer (DE)

The objective of this program is to enable the participants to understand the importance of data management and to realize the value from an organization's data. The program teaches participants the key skills needed to understand and manage activities through the entire data life-cycle.

Data Star Program

Program Logistics

- Classes conducted 5 days a week
- Breakfast, Tea Breaks & Lunch provided
- Material is available on piazza. Save a copy of this material – you won't have access to piazza forever
- Class starts at 9am & ends at 5pm
  - Punctuality is expected
  - Attendance is compulsory – modules build on one another
- You are expected to review material after class to keep up (this is an intensive program)
- Trainer feedback forms to be completed at the end of every class

Data Star Program

Module Details (DS Only) – Dates available on Piazza (subject to change)

The Center of Applied Data Science

1. Introduction to Programming
2. Python Programming
3. R Programming
4. Linux, Python Scripting & Automation
5. Data Science Applications with Examples
6. Relational Databases & SQL
7. NoSQL Databases
8. Statistics
9. Data Visualization
10. Graph Theory
11. Big Data Technologies (Hadoop Ecosystem)
12. Graph Theory
13. Machine Learning (Supervised & Unsupervised)
14. Natural Language Processing (NLP)
15. Neural Networks

Data Star Program

Additional Questions?

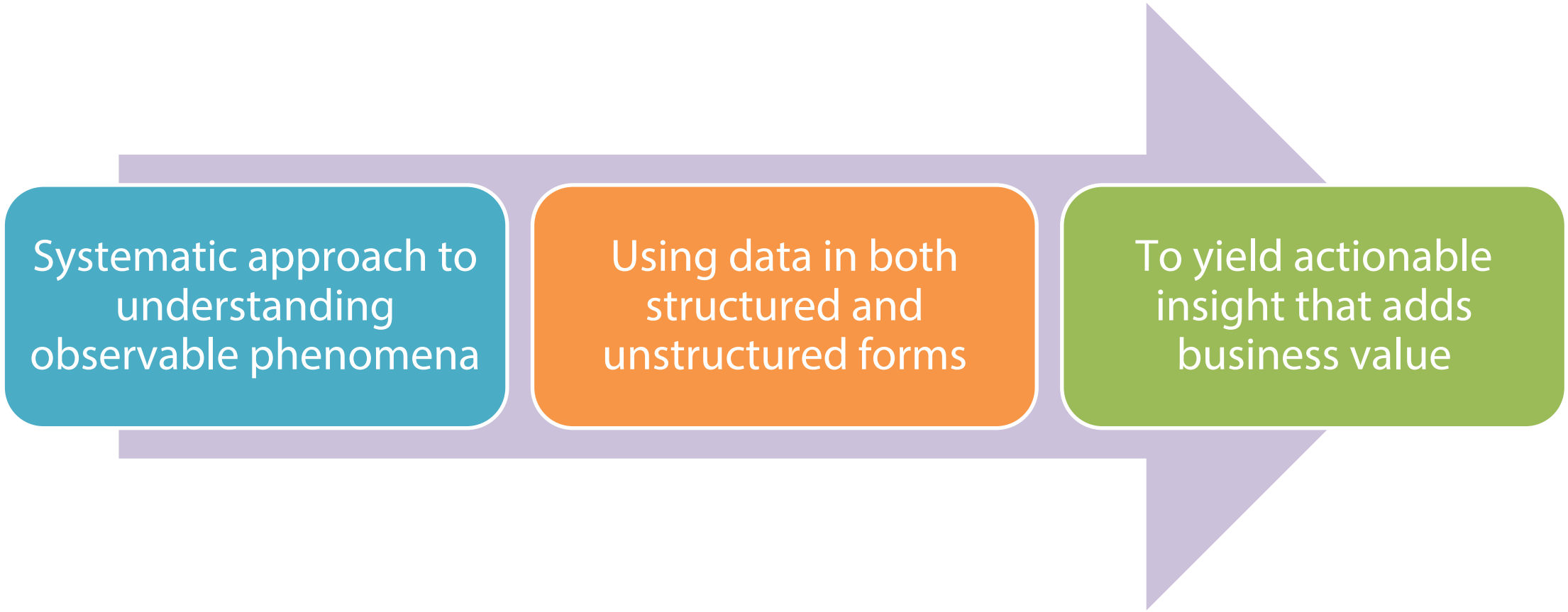
What is Data Science?

…and why is it so popular?

What is Data Science?

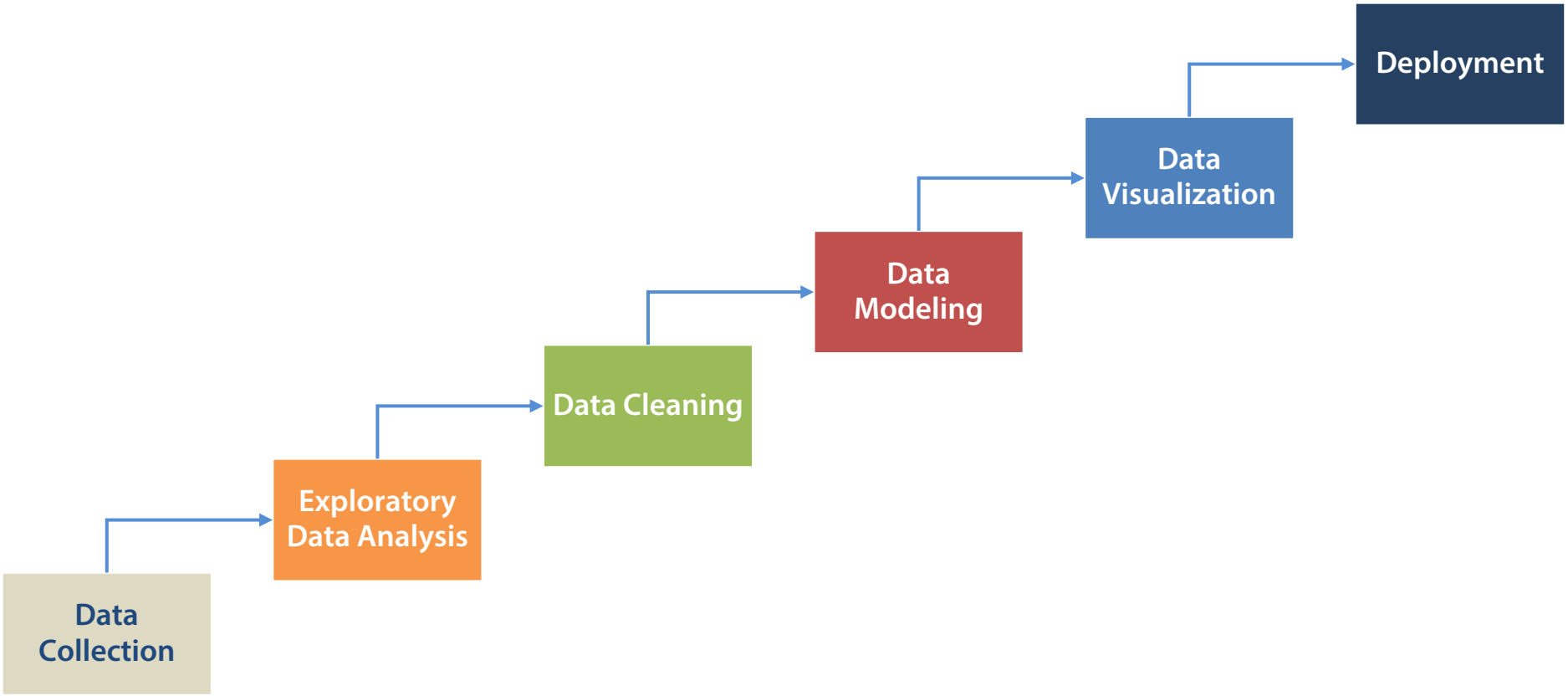
The 'new' hot multi-disciplinary field

The Center of Applied Data Science



Systematic Approach

Following a methodology for sense-making in uncertainty



Structured & Unstructured Data

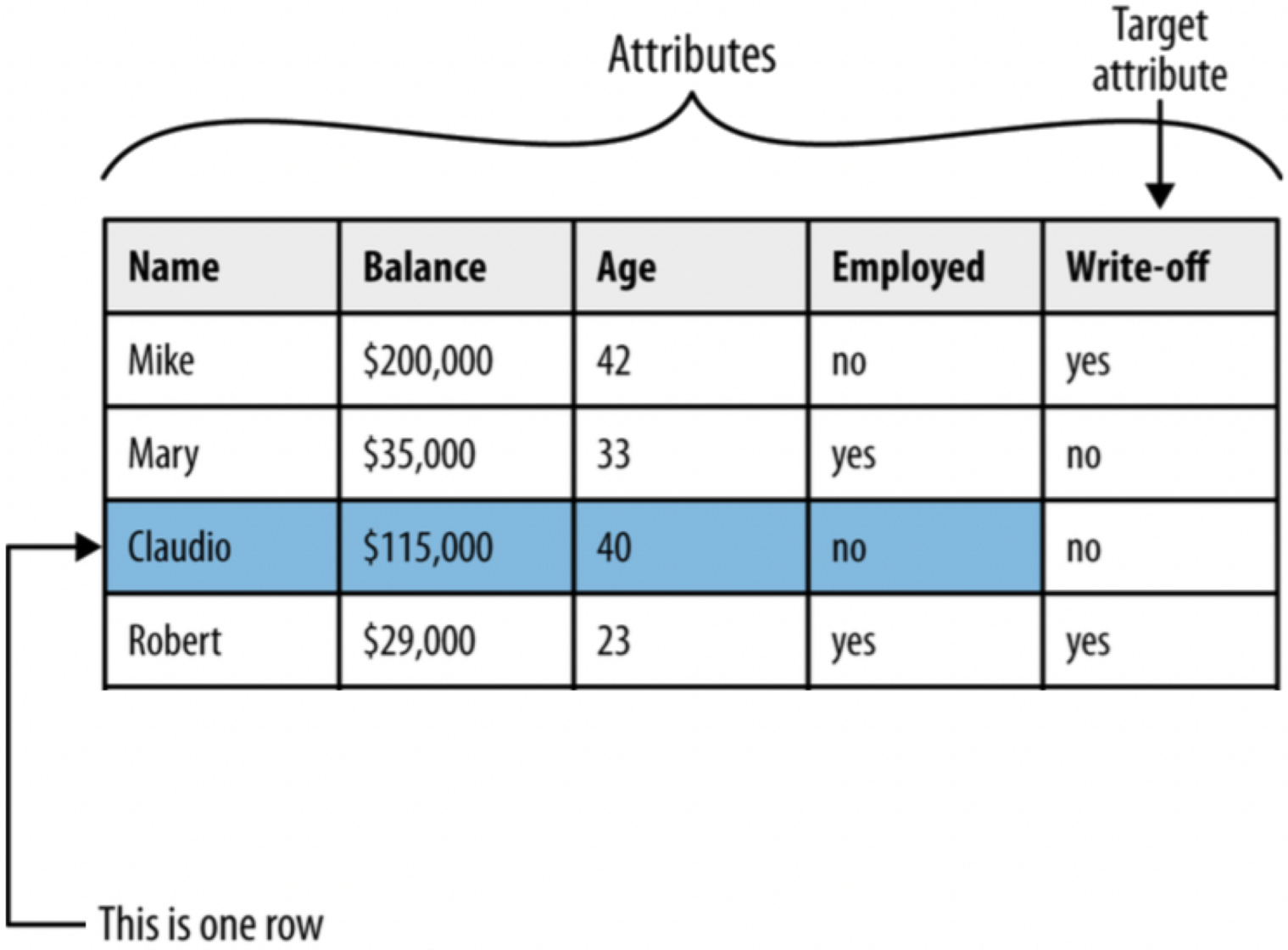
What's the difference?

The Center of Applied Data Science

Unstructured Data	Semi-Structured Data	Structured Data
E-Mails Images Videos Music	XML Metadata tags JSON files	Databases Spreadsheets Point-of-Sales Systems Web logs

Structured Data

You've seen this many times before (and will again in SQL)



Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes

Semi-structured data

Example: MongoDB document database

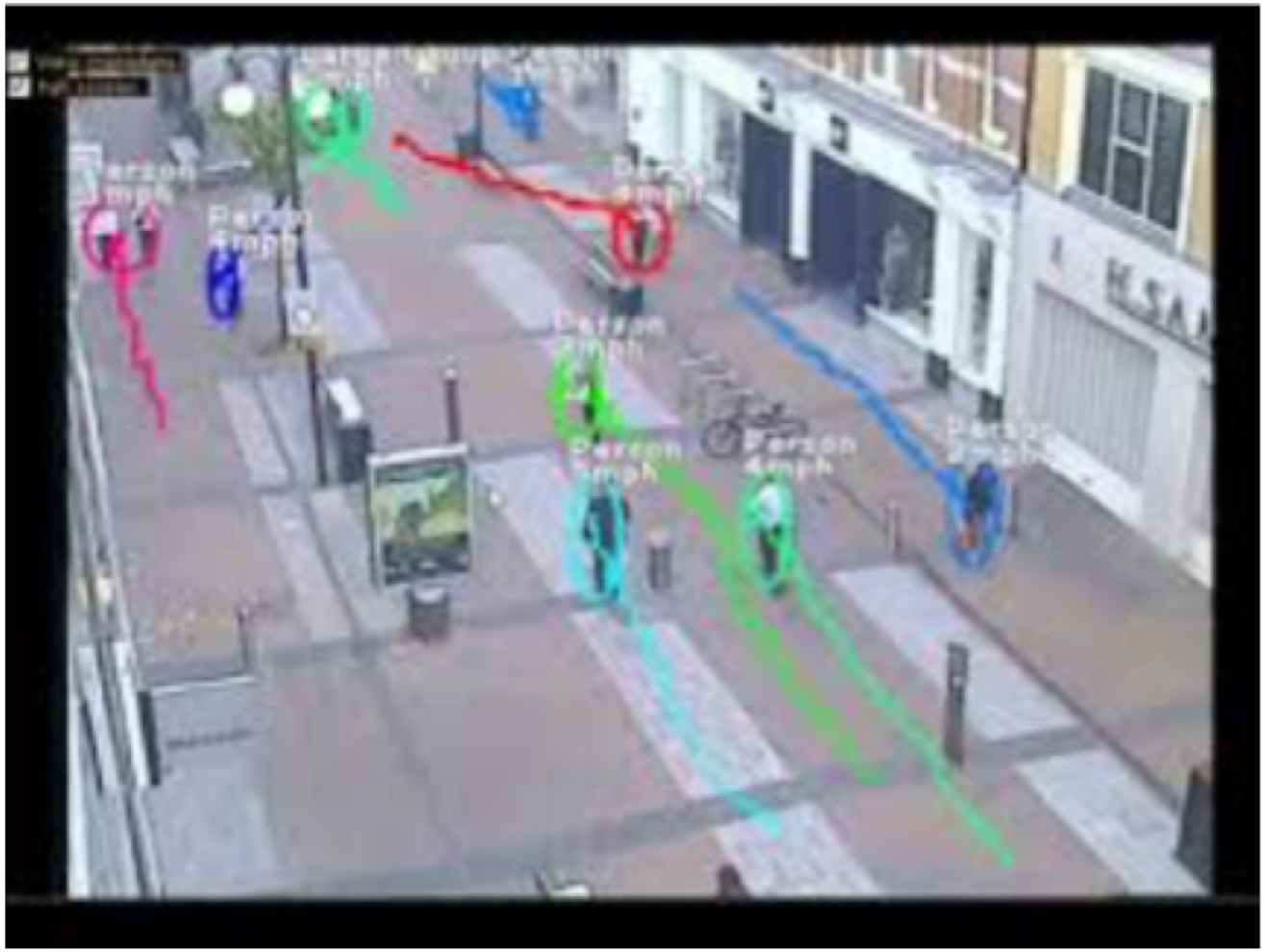
- Data is stored as nested dictionaries relating VALUES to KEYS
- Keys can be different for every data point
  - e.g. companies in a database may have different job descriptions

```
{
    "empid": "SJ011MS",
    "personal": {
        "name":"Smith Jones",
        "gender": "Male",
        "age": 28,
        "address":
        {
            "streetaddress": "7 24th Street",
            "city": "New York",
            "state": "NY",
            "postalcode": "10038"
        }
    },
    "profile": {
        "designation": "Deputy General",
        "department": "Finance"
    }
}
```

Unstructured data

New ways to capture value

- Data structure is too variable to be systematically explained
  - e.g. images may be different sizes, color or grayscale, contents are complex and can vary greatly



Actionable Insight for Business Value

Some tasks you can do with data

The Center of Applied Data Science

REPORT PATTERNS AND TRENDS

PREDICT

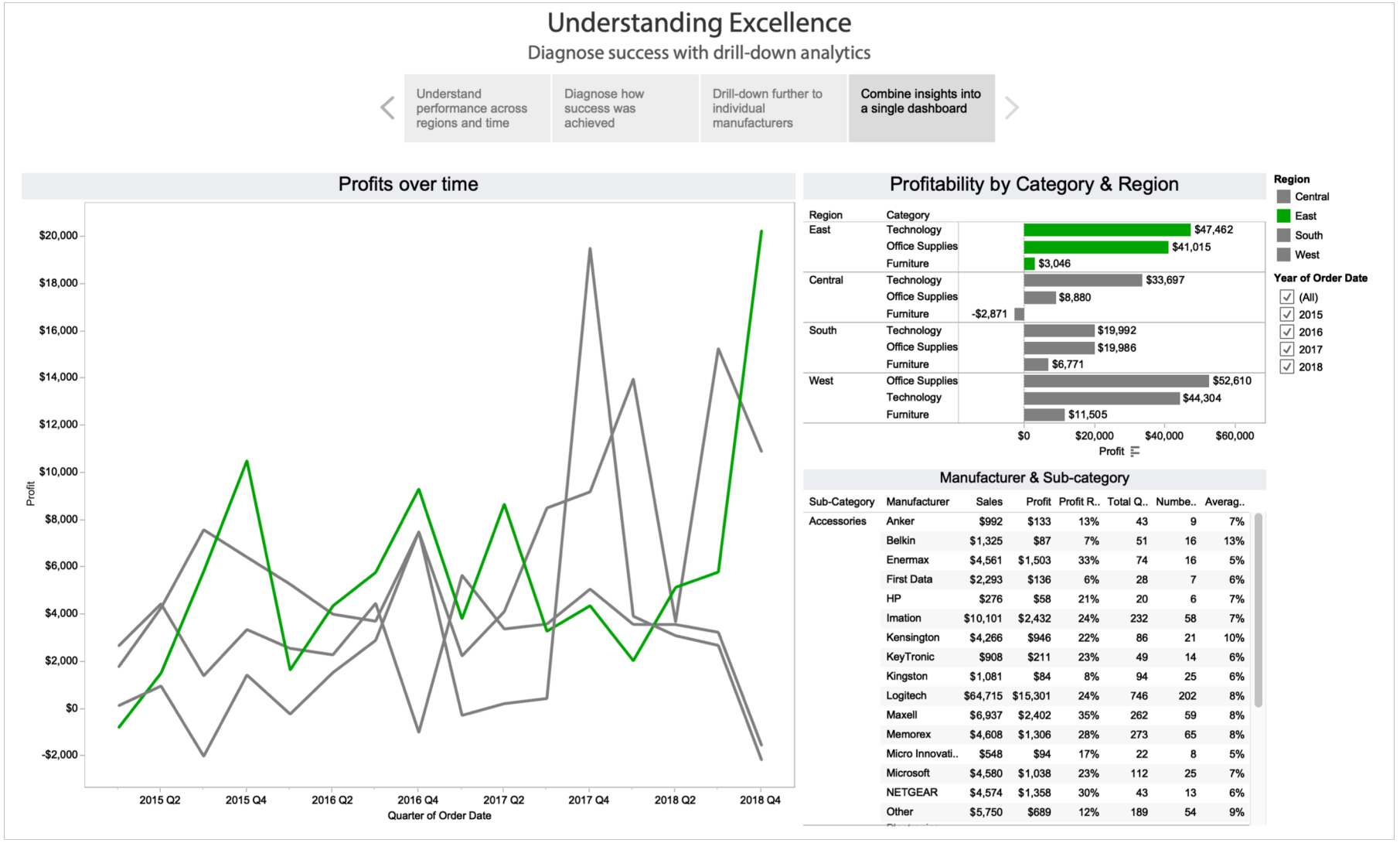
FINDING PATTERNS

FORECAST

INFER

Report Patterns & Trends

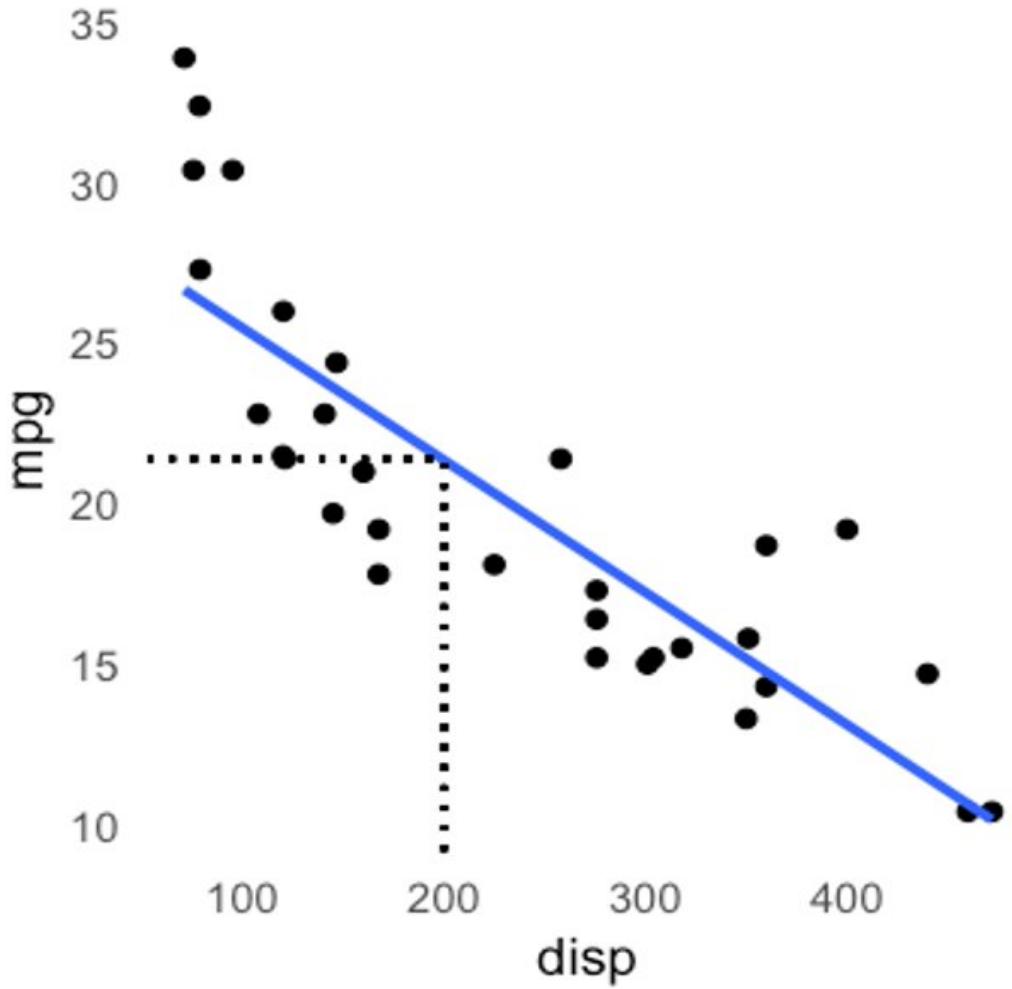
Dashboard Building



REPORT PATTERNS AND TRENDS

Predict

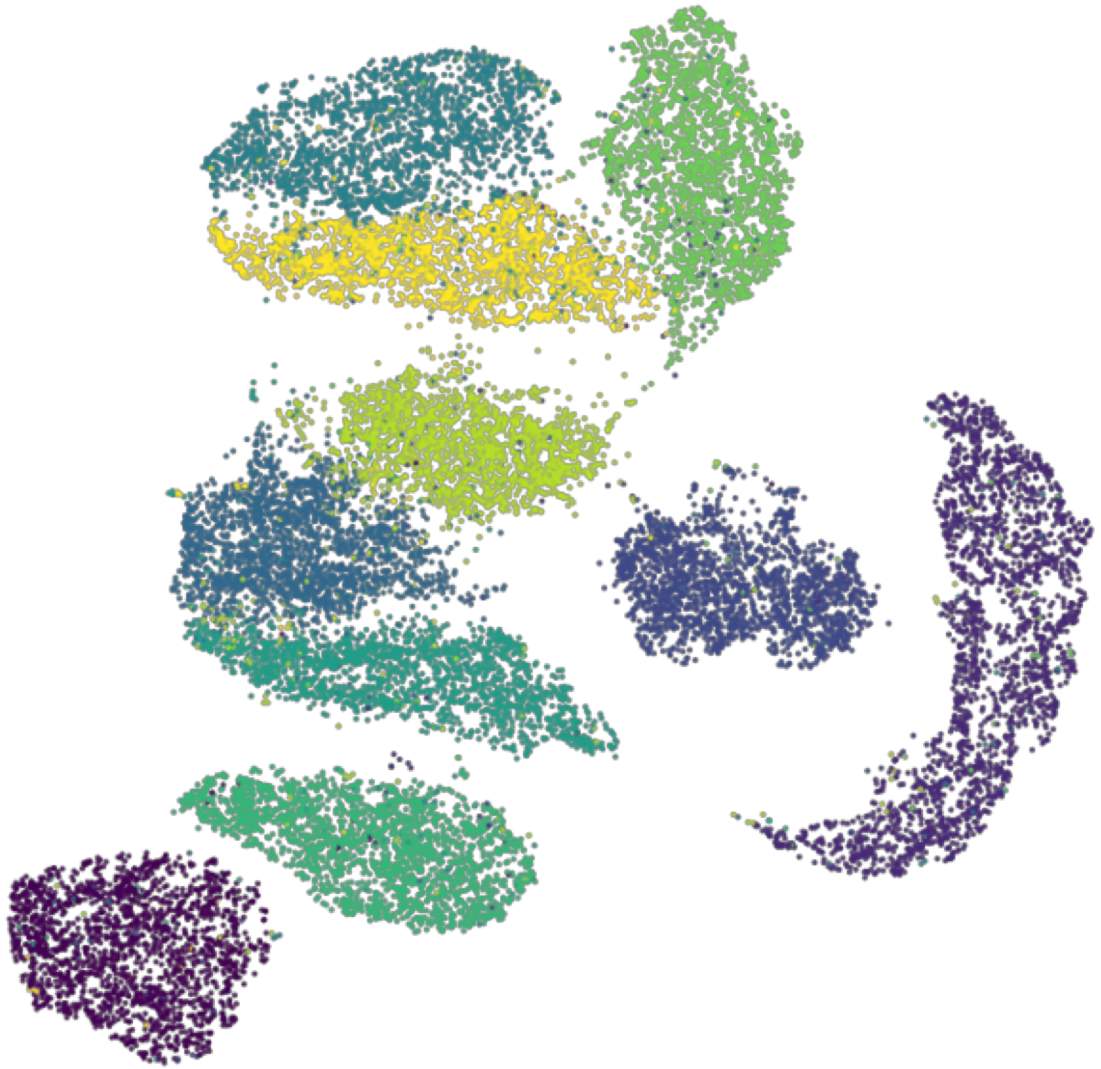
Perhaps more precisely – value estimation. Pictured – Linear Regression used to estimate fuel mileage



PREDICT

Finding Patterns

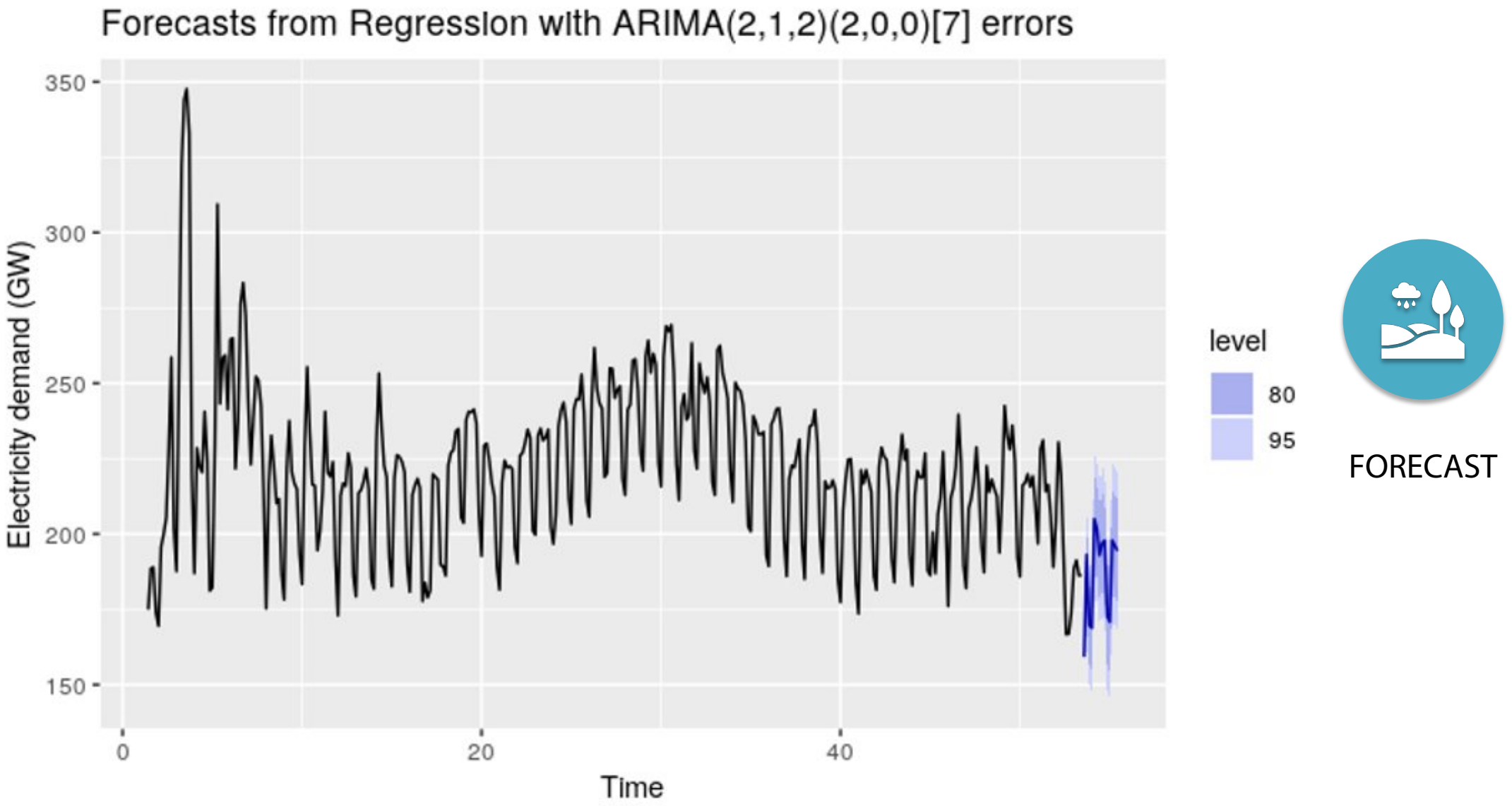
Clustering – Pictured is an example of DBSCAN on non-linearly separable data



FINDING PATTERNS

Forecasting

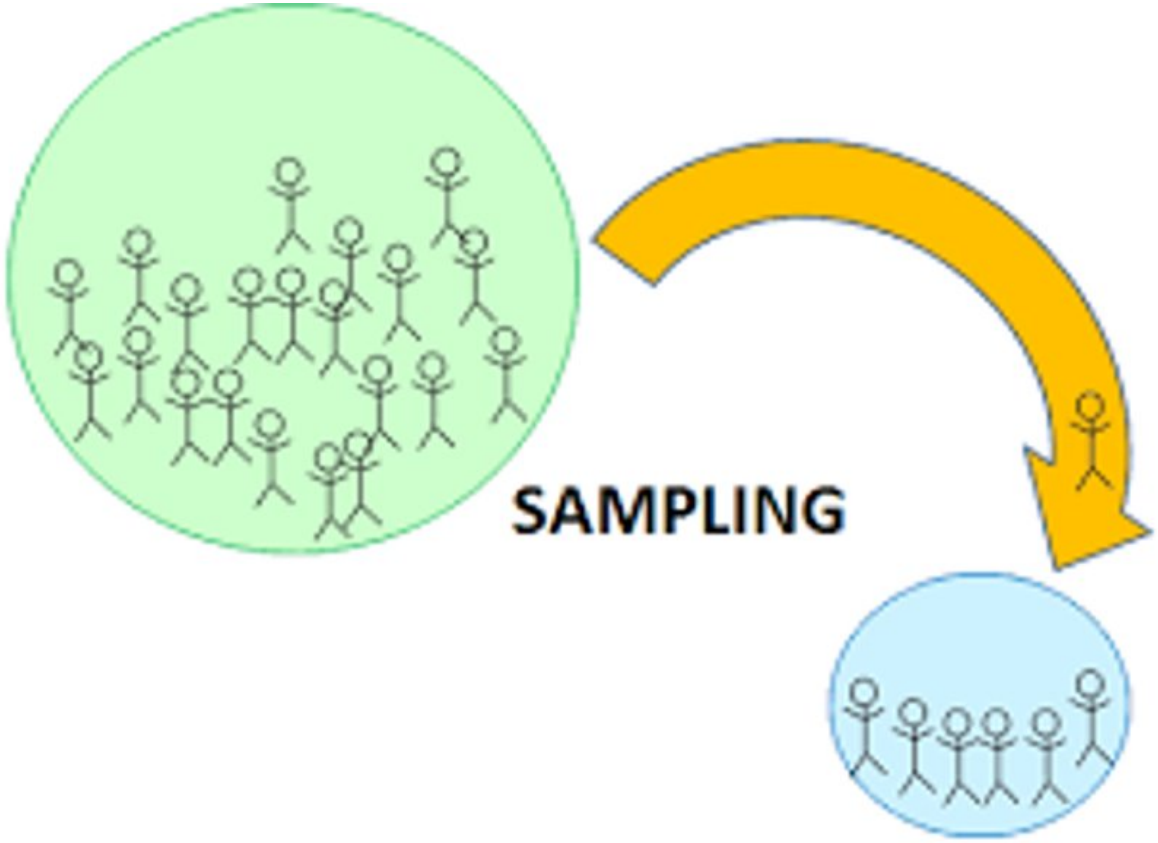
Pictured – Auto Regressive Integrated Moving Average model (ARIMA)



FORECAST

Inference

Generalising results onto a population – Take caution



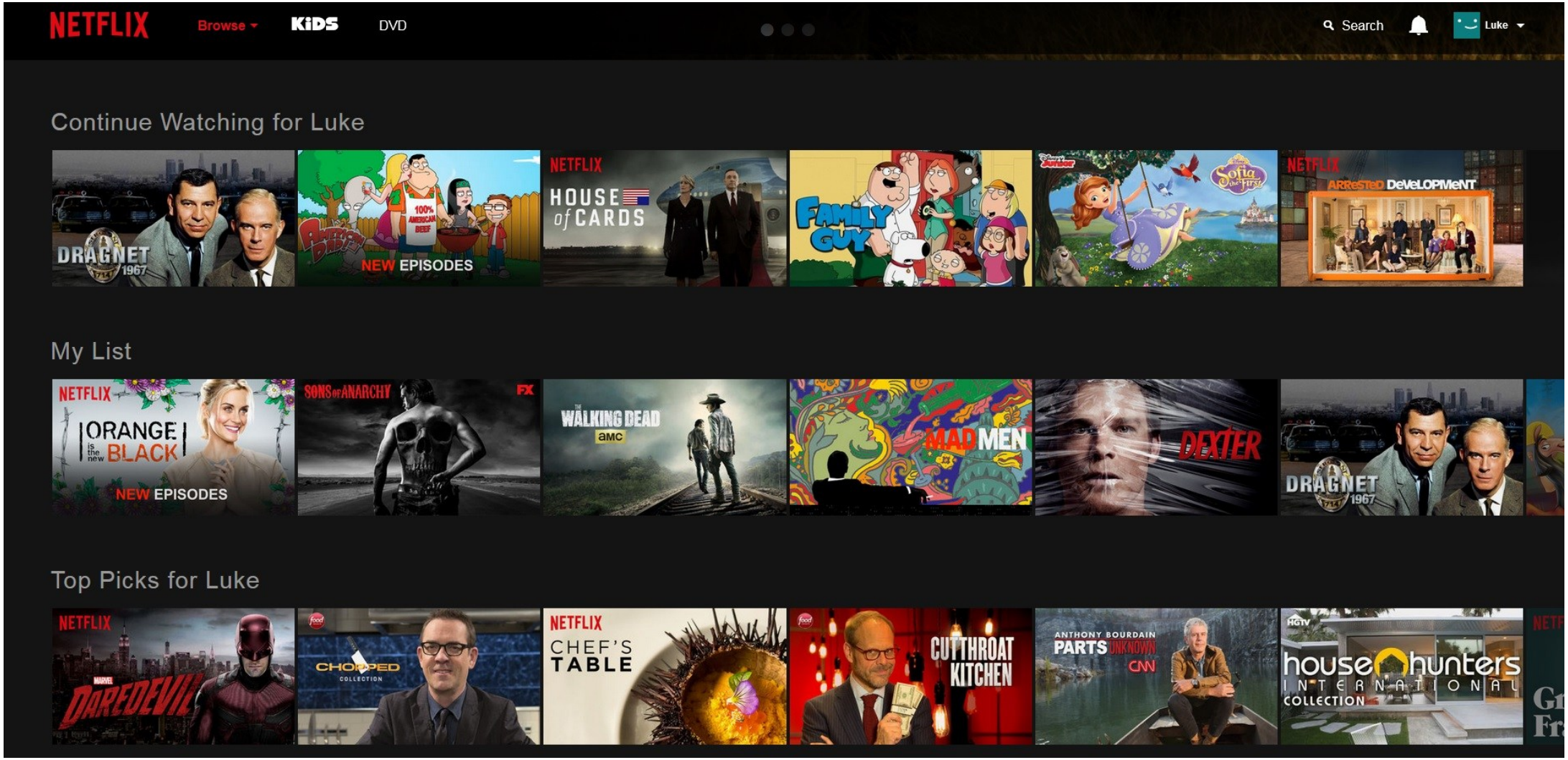
INFER

Data Analytics Use Cases

Popular examples all around us

Netflix

Data Science for content creation / curation

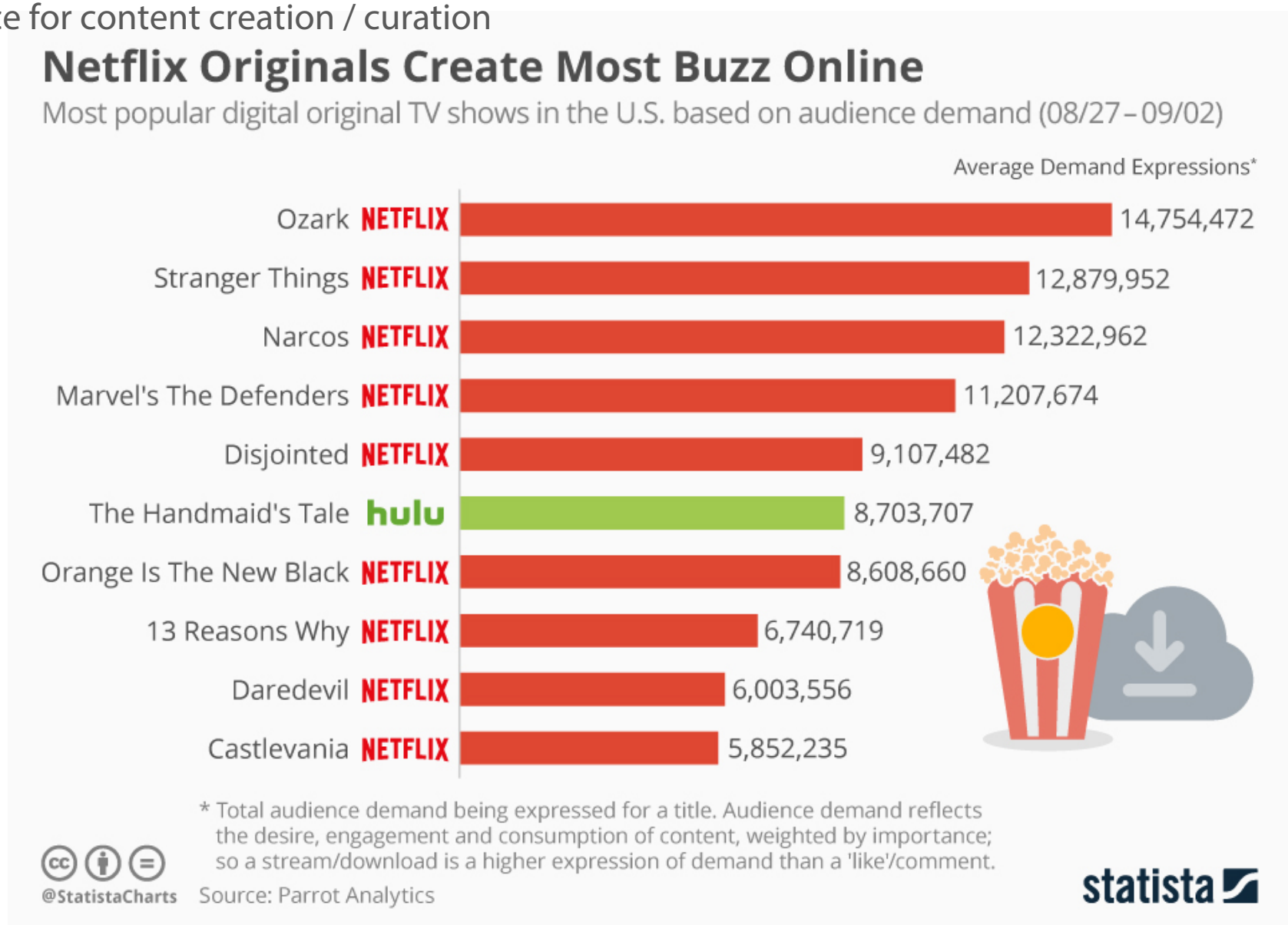


Netflix

Data Science for content creation / curation

Netflix

Data Science for content creation / curation



Source: https://www.statista.com/chart/9759/most-popular-streaming-series/

Netflix

Data Science for content creation / curation

The Center of Applied Data Science

What data was used?

- Time spent selecting movies, how many times playback is stopped, delays caused by buffering, bit rate (viewing experience)

What technologies were used?

- “Netflix data framework utilizes Hadoop, Hive, Pig as well as Teradata. It also includes our own open source applications like Lipstick and Genie. We are exploring Spark for machine learning,” Kurt Brown (Director, Data Platform)

How did the data create value?

- Much of the metadata is simple structured data but much more of this valuable data is ‘messy’ content through video and audio
- Netflix ‘unlocked’ this value by paying viewers to mark-up themes, motifs and patterns through
- Defined micro-genres such as ‘Historical genres featuring gay/lesbian themes’ or ‘Comedies featuring talking animals’ or ‘Brightly colored children’s cartoons’
- This is now being automated, computer vision takes a snapshot of the video as a JPEG, and processes them to determine what is happening on screen through facial recognition etc.

Walmart

Unexpected product & event interactions

- When Hurricane Sandy hit the US in 2004, they found unexpected insights can be found when data is studied as a whole rather than isolated sets
- Bad weather led to a surge in sales of flashlights and raincoats, but also the sale of strawberry popsicles
- Extra supplies of these were mobilized before another Hurricane in 2012, leading to a huge jump in sales of these products

Walmart
Save money. Live better.

Walmart

Unexpected product & event interactions

<u>What data was used?</u>

- 40 petabyte database of transactions from previous weeks. Stock data.

<u>How did they incorporate Big Data into their business structure?</u>

- They have a Data Analytics department (Data Café) that monitors real-time transaction data across all their outlets
- When they see a problem, they approach the respective business unit to devise a solution
- E.g.: One Halloween Walmart enjoyed their usual strong sales of novelty cookies, but the Data Café found that in some select outlets, none of the cookies were being bought at all. After approaching the business unit, they found that these cookies weren't even being displayed on the shelf.

<u>How did the data create value?</u>

- Amount of time between problem being recognized to a solution being proposed went down from 2-3 weeks to 20 minutes

Rolls-Royce

Data Science across entire business model: Design, Manufacture & After Sales Support

- Rolls Royce designs, manufactures and maintains aerospace engines for the likes of Boeing & Airbus
- They have **zero error tolerance** as human lives are at stake
  - 1% failure rate would still mean 1024 crashes a day!
- Big Data Analytics used in 3 areas:
  1. **Design**

     Fault reduction for cost savings down the road
  2. **Manufacture**

     Precision Quality Control / IoT Solutions
  3. **After Sales Support**

     Can avoid or mitigate what is likely to cause a problem through intervention (preventative maintenance) – in the future this intervention can become automated

Rolls-Royce

Data Science across entire business model: Design, Manufacture & After Sales Support

<u>What data is being used?</u>

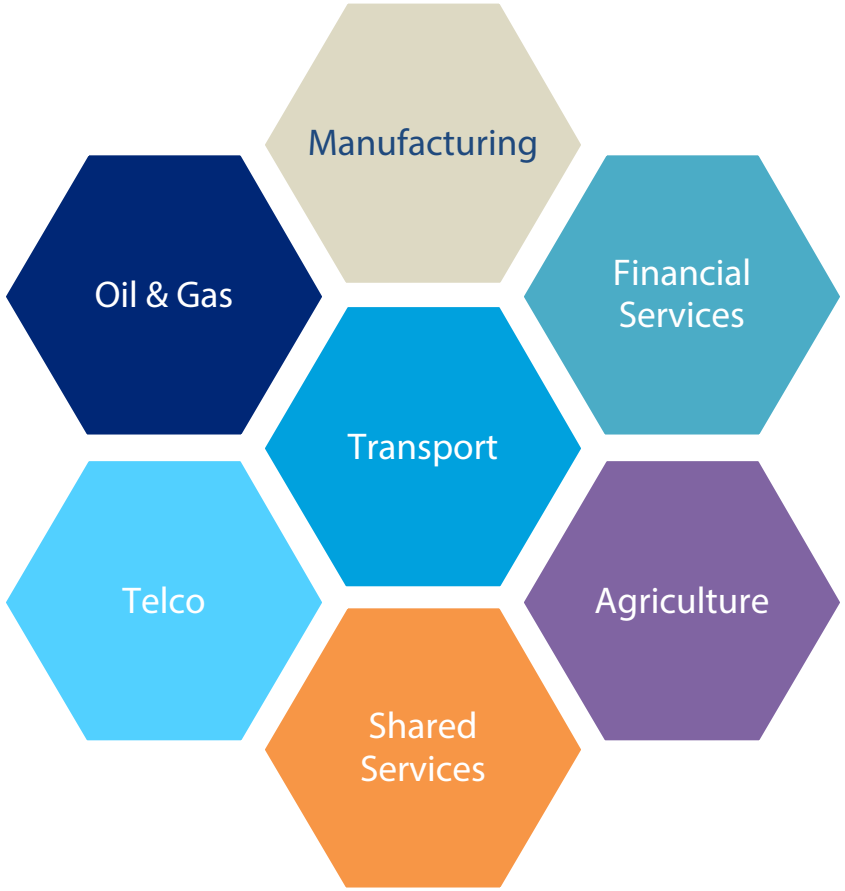
- Emphasis is on internal data, sensors fitted to the engines – wireless transmissions and 3G send snapshots of engine performance at key flight phases such as climb, cruise and take off

<u>What value has been created by Big Data?</u>

- All the information about why errors happen is fed back into the design process
- Product development time fell, quality and maintenance needs dropped, diagnosing faults has 'significantly' reduced costs
- This has also resulted in a new business model for the company – they are able to offer a new service 'total care' which is use per hour of the engine, with all the service costs borne by Rolls Royce

Group Activity (20 minutes)

Discuss the possible use cases you may have read / heard about in the following industries



What is a Data Scientist?

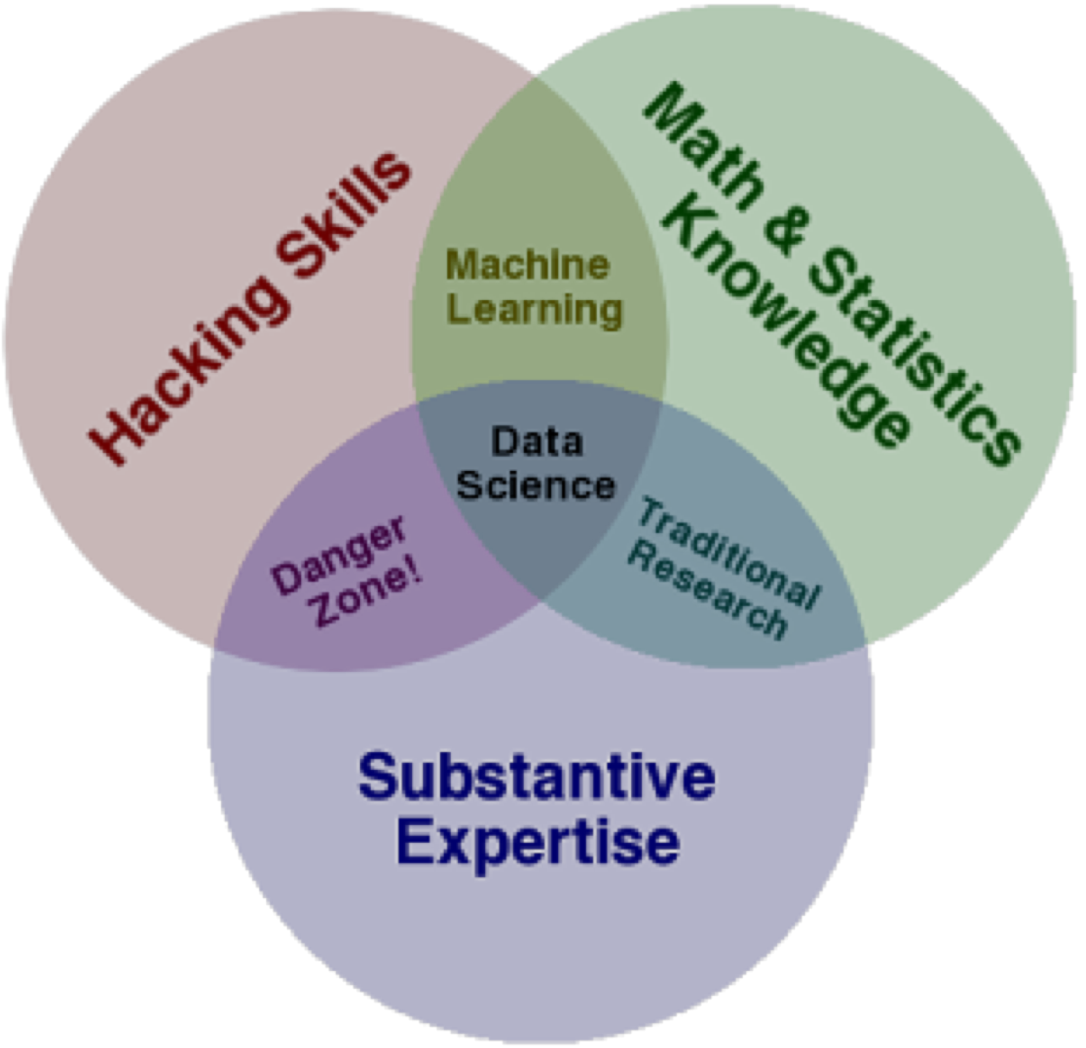
Demystifying the hype

Misconception about Data Scientists

We are not unicorns!

A melting pot of various skills

Jack of all trades (master of some?)



Inquisitive

Driven by a desire to know more

Creative

Tell engaging stories

Skeptical

Question everything

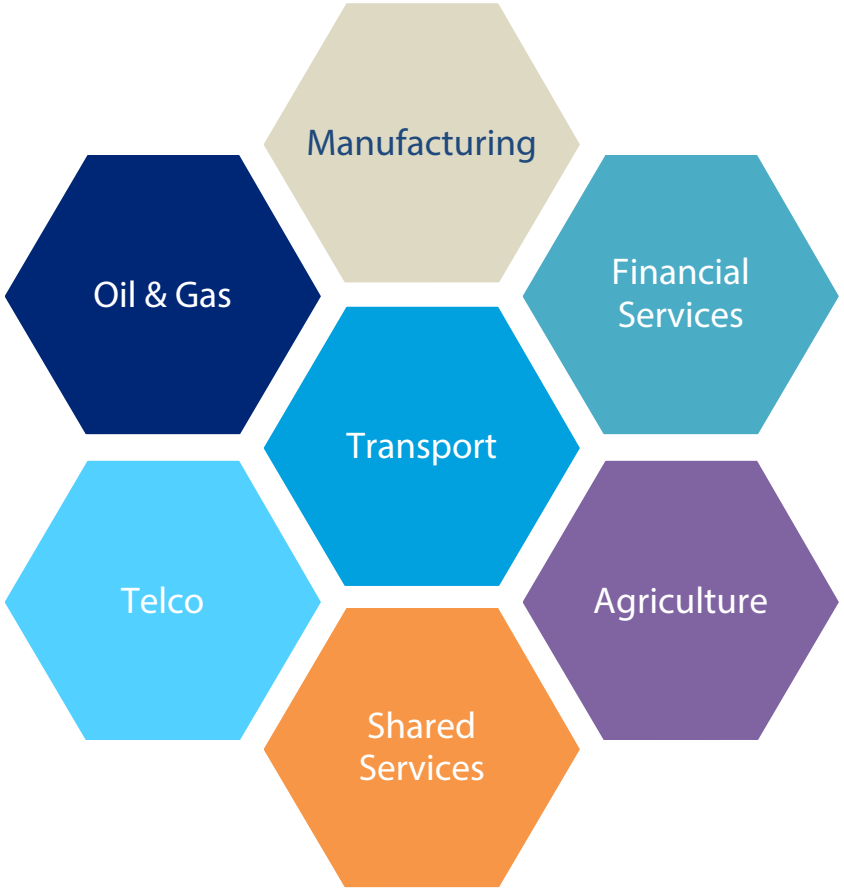
Technical

Strong background in mathematics/statistics and programming

Photo credit: Drew Conway

A Data Scientist's Toolbox

Domain Expertise



<u>Additional Skills (non-exhaustive)</u>

1. Budgeting/costing
2. Project management (Slack/Teams/Planner)
3. Governance/compliance
4. Product design/development
5. Stakeholder management
6. Negotiation & conflict resolution

A Data Scientist's Toolbox

Technical Skills (Computer Science & Mathematics)

Programming Languages	: Python / R / HTML
Machine Learning APIs	: scikitlearn / Keras / TensorFlow
Big Data Technologies	: Hadoop / Spark / MapR
Database Management	: SQL / MongoDB / Neo4j / Cassandra
Cloud Technologies	: AWS / Google Cloud / Microsoft Azure

Statistical Modelling	: ANOVA / Time-Series / Monte Carlo
Machine Learning	: Regression / Decision Trees / SVM / PCA / DBSCAN
Deep Learning	: Neural Networks / RNN / CNN
General Mathematics	: Linear Algebra / Calculus
General Statistics	: Hypothesis testing / Sampling / Simulation
Graph Theory	: Social Network Analysis / Connectivity

A Data Scientist's Toolbox

Communication Skills & Development Tools

Visualisation Tools	: Tableau / Microsoft PowerBI / QlikView / D3.js / ggplot2 / matplotlib
Data Storytelling	: Interpretable Visuals / Intuitive Story
Presentation Skills	: Public Speaking / Fielding Questions / Client Management

---

Version Control	: GitHub
Package Management	: Docker / Conda
File Management	: Linux Scripting
Programming IDE	: Jupyter Notebooks / PyCharm / R Studio

A vast ecosystem of tools & technologies

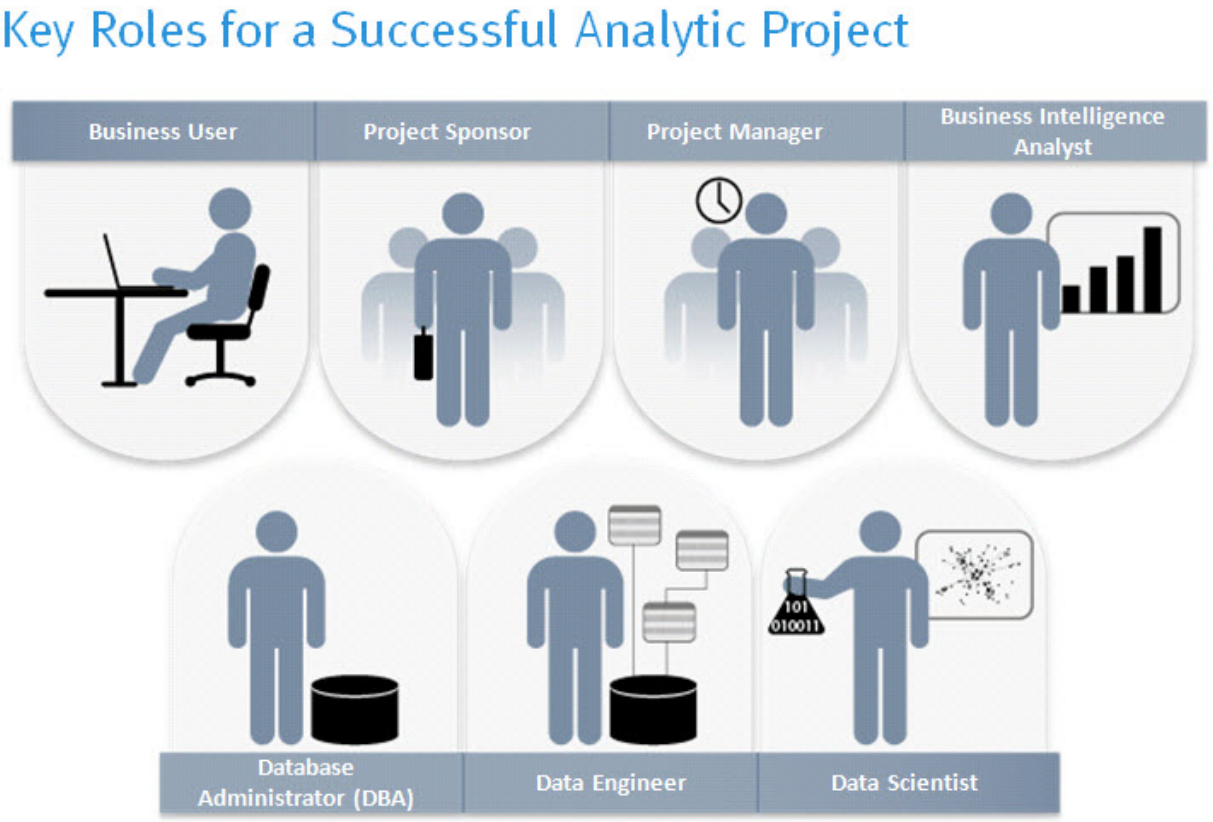
Which ones have you worked with before?

The Center of Applied Data Science

No Man Is An Island

You don't need to learn every single skill at one go!

- A successful data science venture is a team effort
- Very few individuals possess sufficient background in all the required fields



What really is a data scientist?

Putting the toolbox skills to use

The Center of Applied Data Science

"More generally, a data scientist is someone who knows how to **extract meaning from** and **interpret** data, which requires both tools and methods from **statistics** and **machine learning**, as well as being human.

She spends a lot of time in the process of **collecting, cleaning, and munging data**, because data is never clean. This process requires persistence, statistics, and software engineering skills—skills that are also necessary for understanding biases in the data, and for debugging logging output from code.

O'Neil, C., and Schutt, R. *Doing Data Science.* First edition. Via https://datasciencedegree.wisconsin.edu/data-science/what-do-data-scientists-do/

What really is a data scientist? (Continued)

Putting the toolbox skills to use

The Center of Applied Data Science

Once she gets the data into shape, a crucial part is exploratory data analysis, which combines visualization and data sense. She'll **find patterns, build models, and algorithms**—some with the intention of understanding product usage and the overall health of the product, and others to serve as **prototypes** that ultimately get baked back into the product.

She may **design experiments**, and she is a critical part of data-driven **decision making**. She'll **communicate** with team members, engineers, and leadership in clear language and with data visualizations so that even if her colleagues are not immersed in the data themselves, they will understand the implications."

O'Neil, C., and Schutt, R. *Doing Data Science.* First edition. Via https://datasciencedegree.wisconsin.edu/data-science/what-do-data-scientists-do/

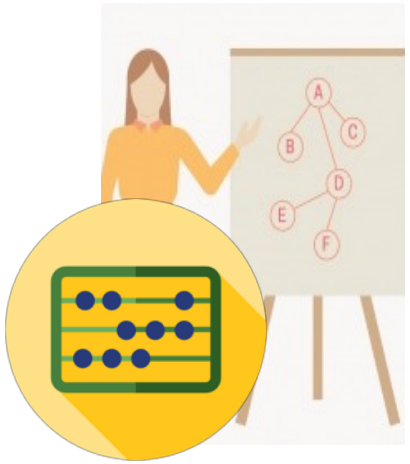
A new breed of Data Professionals

Data Scientists can hardly take all the credit

Analytics Manager

Data Engineer

Data Scientist



Data Analysts

CADS Data Driven Organization Model (DDO)

The Center of Applied Data Science

How would becoming a DDO help us and how to get there?

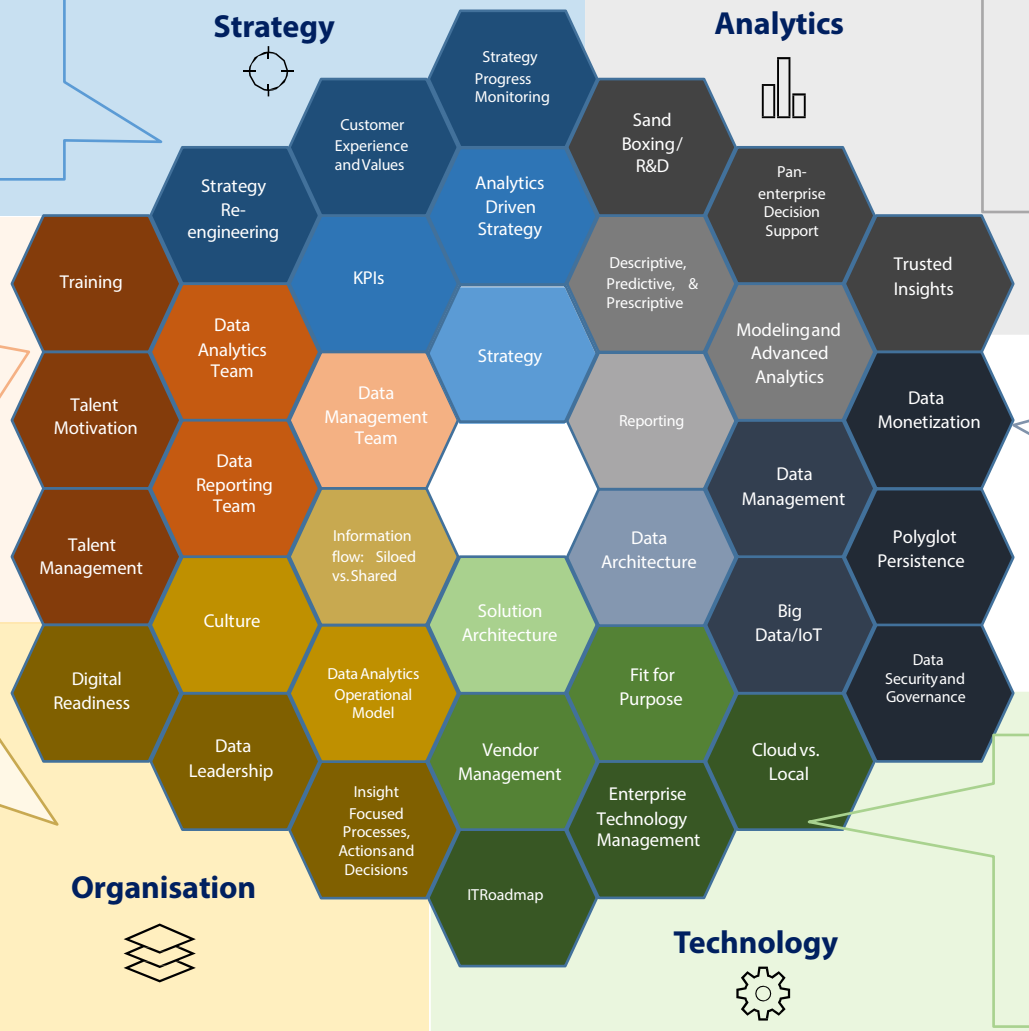
How capable is our analytics practice?

Do our people have the right skills and take the right actions?

How does our data support achieving our strategic objectives?

Do our culture and processes effectively support and control the delivery of insights?

Do we have an integrated scalable technology platform in place?



Characteristics of the maturity model

Maturity differs depending on the industry

The Center of Applied Data Science

Mature in data and technology

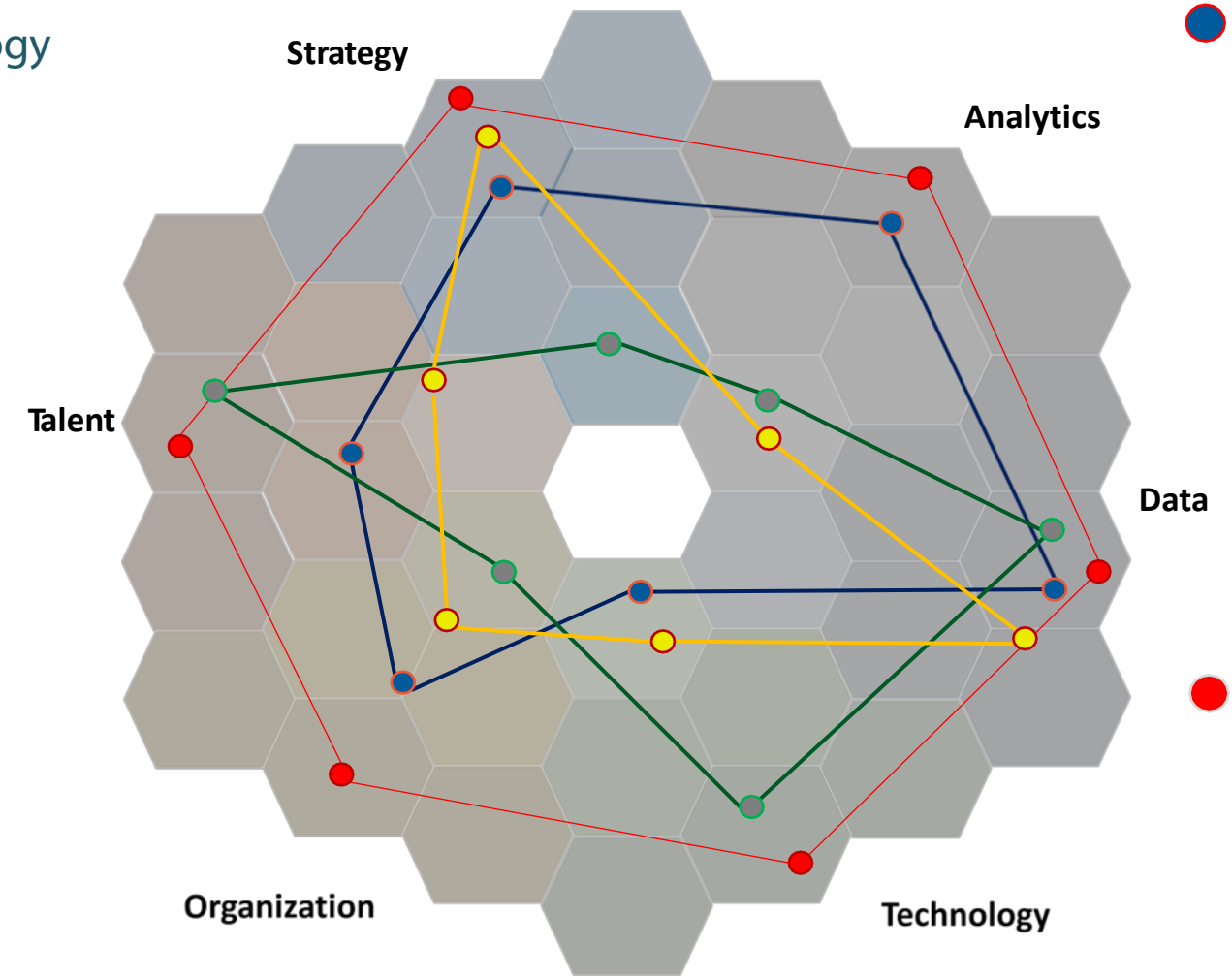
Typical sectors:

- Telco
- Media

Mature strategy

Typical sectors:

- Healthcare
- Newsprint



Mature data and strategy

Typical sectors:

- Retail banks
- Utilities

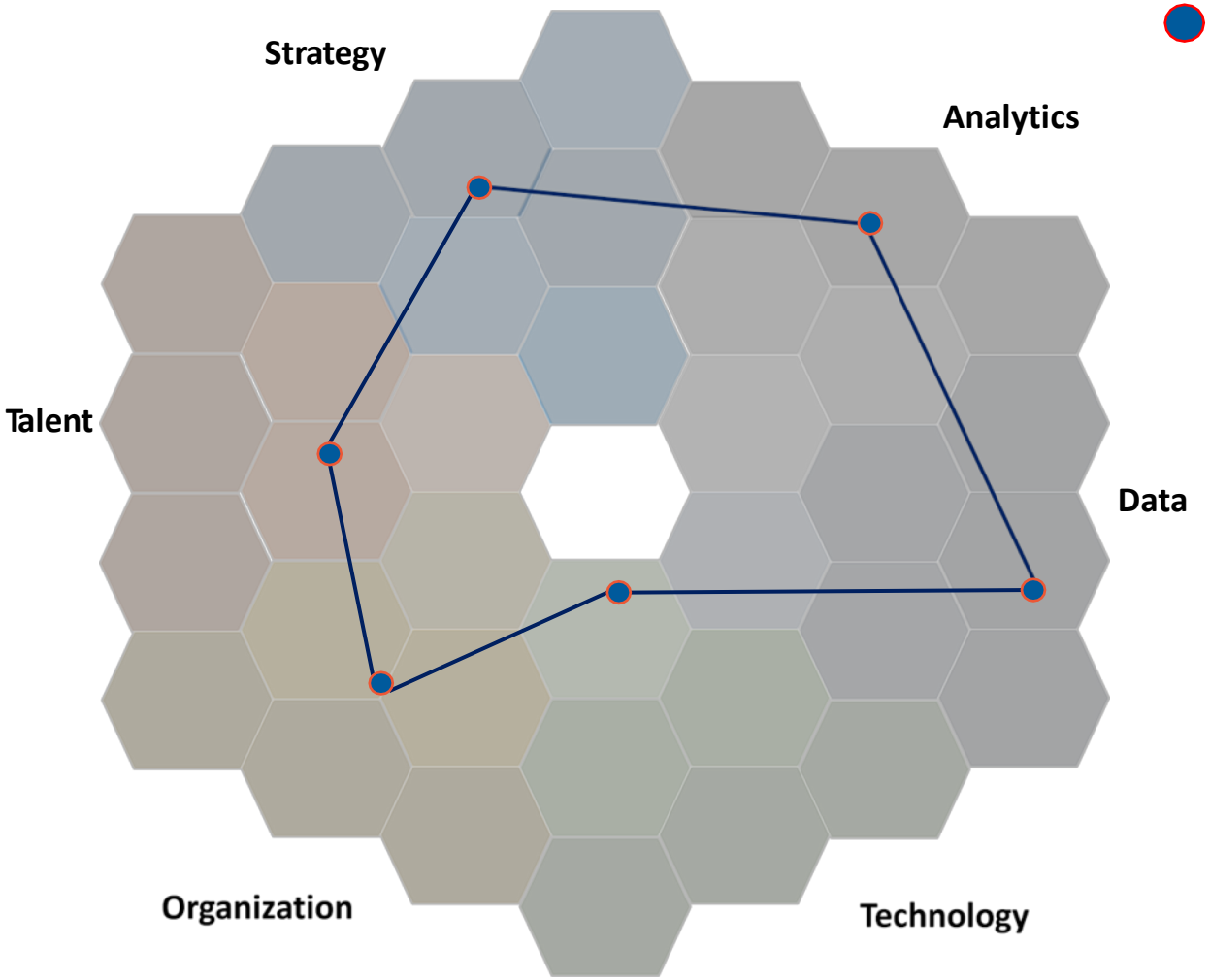
Mature across all areas

Typical sectors:

- Online gaming
- Consumer services

Characteristics of the maturity model

Maturity differs depending on the industry



- Mature data and strategy

Typical sectors:

- Retail banks
- Utilities

Anaconda – Package management made simple

Package & Library management made easy

The Center of Applied Data Science

What is Anaconda?

- Anaconda is a package manager software that simplifies the installation of Python packages
- Open source project developed by Continuum Analytics. Inc
- Available for Windows, Max OSX and Linux

What is Anaconda Navigator?

- Anaconda Navigator provides an alternative GUI method of interacting with conda instead of CLI

Anaconda – Package management made simple

What is a package & why do I care?

The Center of Applied Data Science

When programming in Python (and many other high-level programming languages) we use pre-built packages that perform specific tasks we need

- For example we call NumPy, Pandas, scitkit-learn packages to perform mathematical calculations, build data frames and then initiate machine learning models
- Using pre-built packages are a lot easier than building everything from scratch

ANACONDA

Anaconda – Package management made simple

What features are included in Anaconda?

- Doesn't require administrative privileges to install packages
- Installs non-Python library dependencies (MKL, HDF5, Boost)
- Provides 'virtual environment' capabilities
- Will 'solve environment' to ensure dependencies are installed
- Many channels exist that support additional packages
- Documentation here: http://conda.pydata.org.docs/

Anaconda – Package management made simple

What do you mean by “virtual environments”?

The Center of Applied Data Science

- You may not want ALL packages installed at the same time because of performance / dependencies.

- We can create different “environments” and switch between them depending on our work requirements

- Conda automatically ensured dependent packages are installed but minimizing number of packages ensure you can use more recently updated packages

ANACONDA

Anaconda – Package management made simple

Do I need to use Anaconda Navigator to interact with conda?

No. Installation of packages can be done on CLI

- e.g. conda install wxpython #installing a package
- e.g. conda install wxpython=3.0 #installing specific versions
- e.g. conda install –c vpython vpython #installing from different package channels (-c flag)
- e.g. pip install intervaltree #pip is a different package manager
- e.g. conda update wxpython
- e.g. conda remove wxpython

Anaconda – Package management made simple

Package & Library management made easy

The Center of Applied Data Science

How do I interact with environments?

- Create new environment called ”testenvironment” and install numpy version 1.7
  - conda create –-name testenvironment numpy=1.7
- Activate new environment on machine
  - Source activate testenvironment
  - Source deactivate

Do I need to worry about all this now?

- No. Most packages used in our classes are installed by default when we installed anaconda.

ANACONDA

Introduction to Jupyter Notebooks

Document your code easily

The Center of Applied Data Science

- Jupyter notebooks are documents that contain both code and rich text elements such as images, links & text elements
- These features lend well towards analysis description & display of results
  - A data scientist performing analysis using code can note down observations immediately
  - Others in the team can read the analysis and execute the code to perform the data-analysis themselves in real-time

jupyter

Introduction to Jupyter Notebooks

Launching Jupyter Notebook & Jupyter Lab

The Center of Applied Data Science

You can launch Jupyter in two ways:

- Anaconda Navigator
- Run the command “jupyter notebook” or “jupyterlab” in terminal

Jupyter is a server-client application (meaning that you edit and run your notebooks via a web browser).

- We find that Google Chrome runs the smoothest with Jupyter (set it to default for this course if possible)
- When you launch Jupyter for the first time it will prompt you to input a password

Introduction to Jupyter Notebooks

Components to Jupyter

Kernel

- The kernel is a program that runs and introspects the user's code.
- The Jupyter Notebook App has a kernel for Python code but there are other kernels available for other programming languages (you would have installed the R kernel already)

Dashboard

- The Dashboard shows us the notebooks documents you have made and provides us several options to manage kernels
- Choosing the kernel to run
- Shutting down running kernels

Jupyter Notebook Demonstration

Follow along or take notes if you haven't installed

The Center of Applied Data Science

E: info@thecads.com
W : www.thecads.com