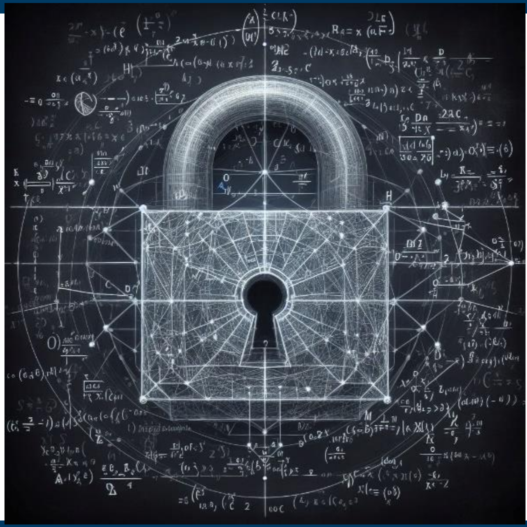
SMARTVISION Europe



Data Science in an hour

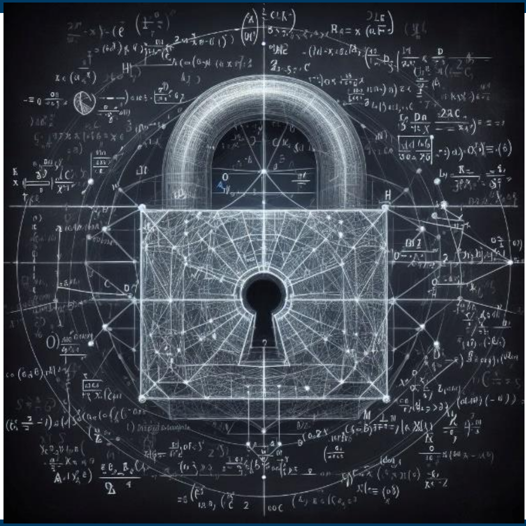
Jarlath Quinn – Analytics Consultant

www.sv-europe.com

A SELECT INTERNATIONAL COMPANY

SMARTVISION
Europe

Just waiting for all attendees to join…



Data Science in an hour

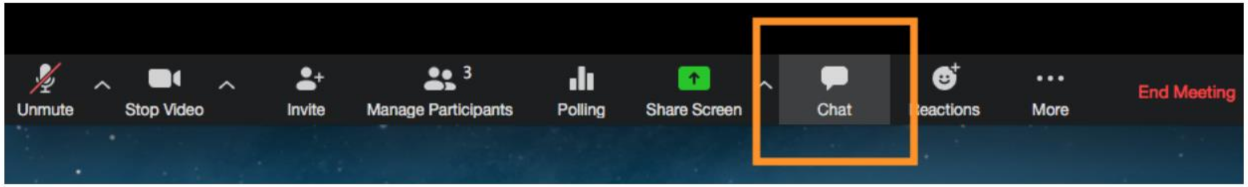
Jarlath Quinn – Analytics Consultant

www.sv-europe.com

A SELECT INTERNATIONAL COMPANY

FAQ’s

- Is this session being recorded? Yes
- Can I get a copy of the slides? Yes, we’ll email links to download materials after the session has ended.
- Can we arrange a re-run for colleagues? Yes, just ask us.
- How can I ask questions? All lines are muted so please use the chat panel – if we run out of time we will follow up with you.



SMARTVISION Europe

IBM® Gold Business Partner™

predictive SOLUTIONS

DataRobot

- Gold accredited partner to IBM, Predictive Solutions and DataRobot specialising in advanced analytics & big data technologies
- Work with open-source technologies (R, Python, Spark etc.)
- Team each has 15 to 30 years of experience working in the advanced and predictive analytics industry

- Deep experience of applied advanced analytics applications across sectors
  - Retail
  - Healthcare/Pharma
  - Finance/Insurance
  - Media/Telecoms
  - Utilities
  - FMCG
  - Charity/Housing/Government

How did we get here?

Statistical Analysis to AI

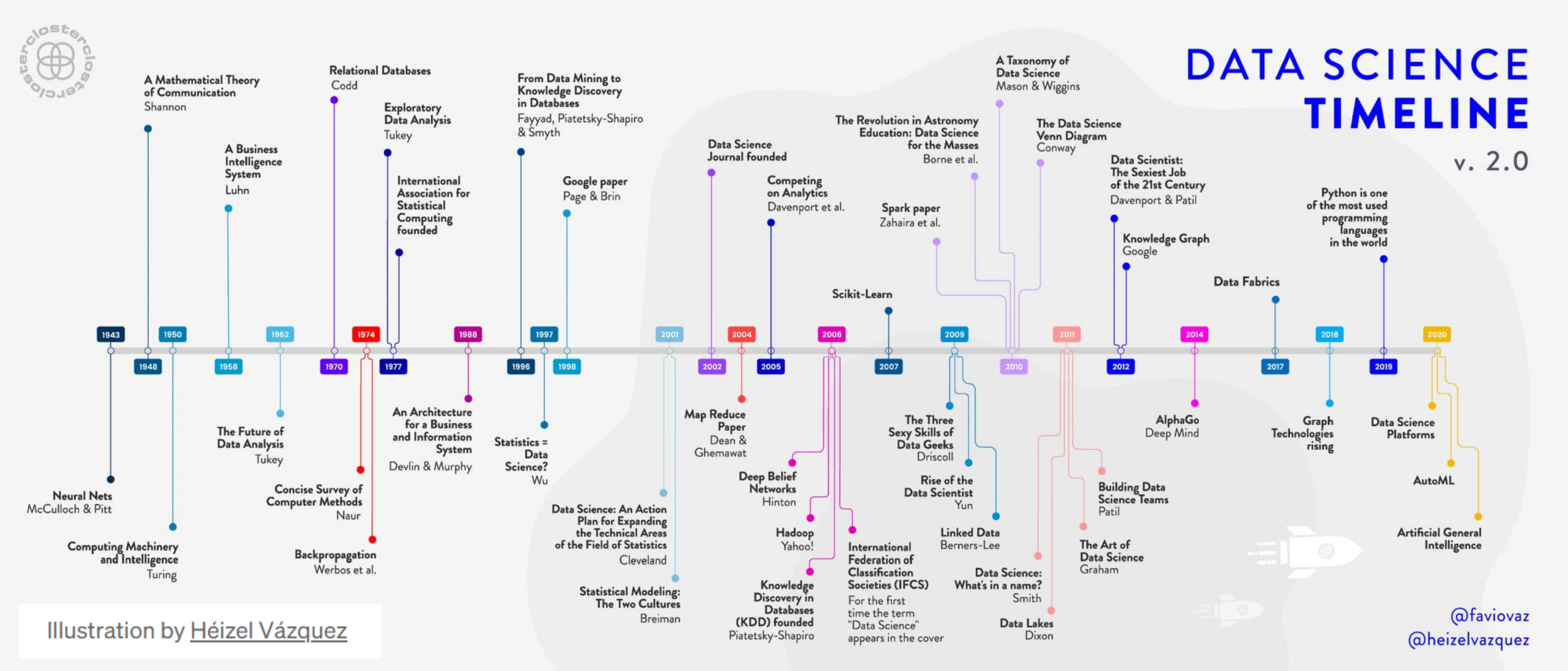
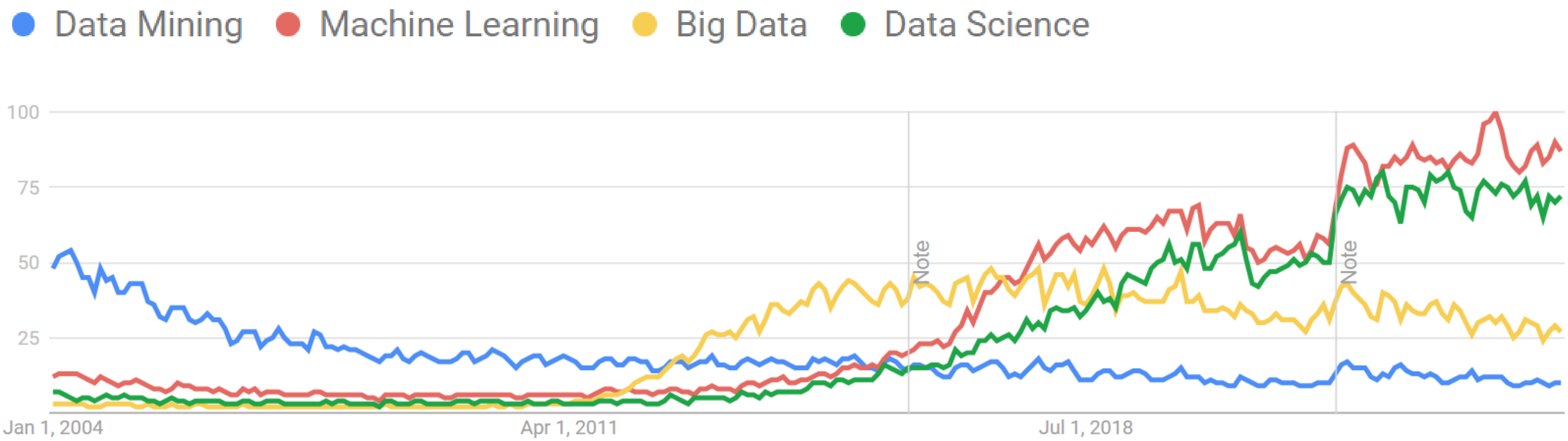

DATA SCIENCE TIMELINE
v. 2.0
1943
Neural Nets
McCulloch & Pitt
1948
A Mathematical Theory of Communication
Shannon
1950
Computing Machinery and Intelligence
Turing
1958
A Business Intelligence System
Luhn
1962
The Future of Data Analysis
Tukey
1970
Relational Databases
Codd
1974
Concise Survey of Computer Methods
Naur
Backpropagation
Werbos et al.
1977
Exploratory Data Analysis
Tukey
International Association for Statistical Computing founded
1988
An Architecture for a Business and Information System
Devlin & Murphy
1996
From Data Mining to Knowledge Discovery in Databases
Fayyad, Piatetsky-Shapiro & Smyth
1997
Statistics = Data Science?
Wu
1998
Google paper
Page & Brin
2001
Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics
Cleveland
Statistical Modeling: The Two Cultures
Breiman
2002
Data Science Journal founded
2004
Map Reduce Paper
Dean & Ghemawat
2005
Competing on Analytics
Davenport et al.
2006
Deep Belief Networks
Hinton
Hadoop
Yahoo!
Knowledge Discovery in Databases (KDD) founded
Piatetsky-Shapiro
International Federation of Classification Societies (IFCS)
For the first time the term "Data Science" appears in the cover
2007
Scikit-Learn
2009
The Three Sexy Skills of Data Geeks
Driscoll
Rise of the Data Scientist
Yun
Linked Data
Berners-Lee
2010
Spark paper
Zahaira et al.
The Revolution in Astronomy Education: Data Science for the Masses
Borne et al.
A Taxonomy of Data Science
Mason & Wiggins
The Data Science Venn Diagram
Conway
2011
Data Science: What's in a name?
Smith
Data Lakes
Dixon
Building Data Science Teams
Patil
The Art of Data Science
Graham
2012
Data Scientist: The Sexiest Job of the 21st Century
Davenport & Patil
Knowledge Graph
Google
2014
AlphaGo
Deep Mind
2017
Data Fabrics
2018
Graph Technologies rising
2019
Python is one of the most used programming languages in the world
2020
Data Science Platforms
AutoML
Artificial General Intelligence
@faviovaz
@heizelvazquez

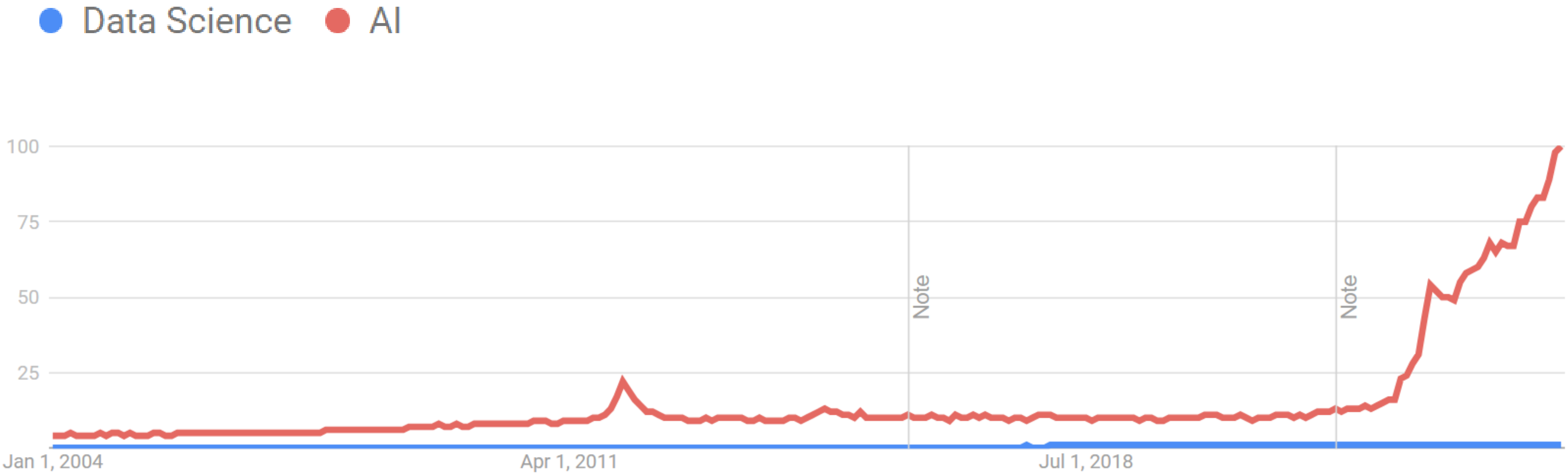
Illustration by Héizel Vázquez

Credit: https://medium.com/towards-data-science/the-roots-of-data-science-77c71115229

The new frontiers of data analysis



Not forgetting of course…the elephant in the room

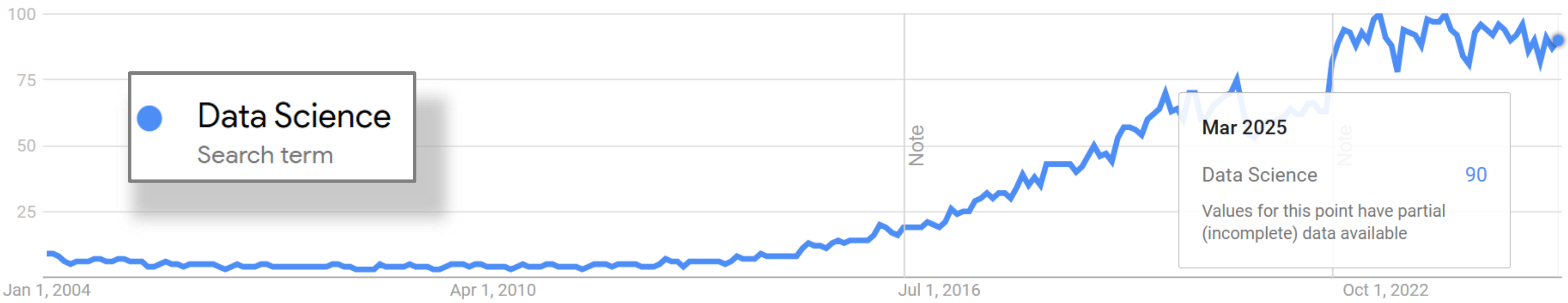


SMARTVISION Europe

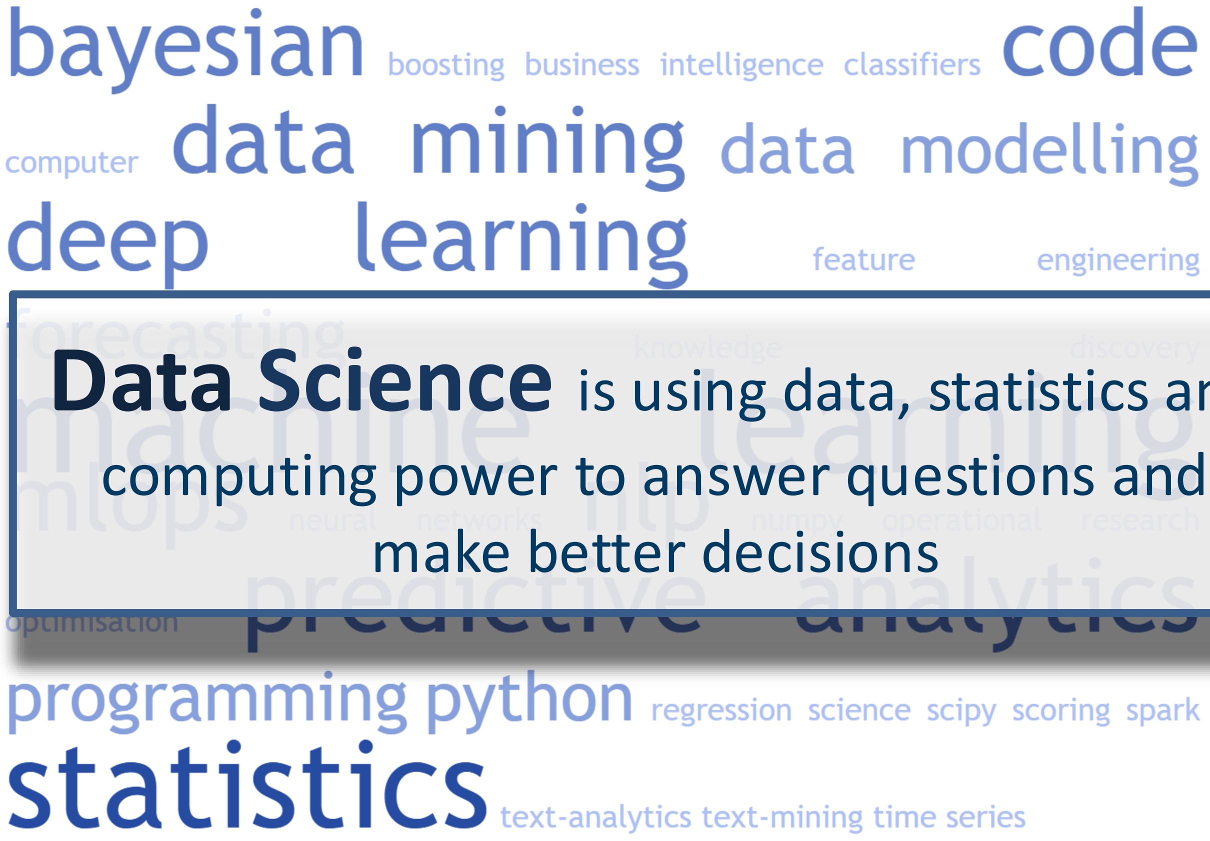
A SELECT INTERNATIONAL COMPANY

The term **Data Science** was first proposed by Peter Naur in 1974 as an alternative name for computer science.

But it wasn't until 2008 that Patil and Hammerbacher popularized the term **Data Scientist** to describe professionals who combine programming skills with statistical knowledge to extract insights from data.

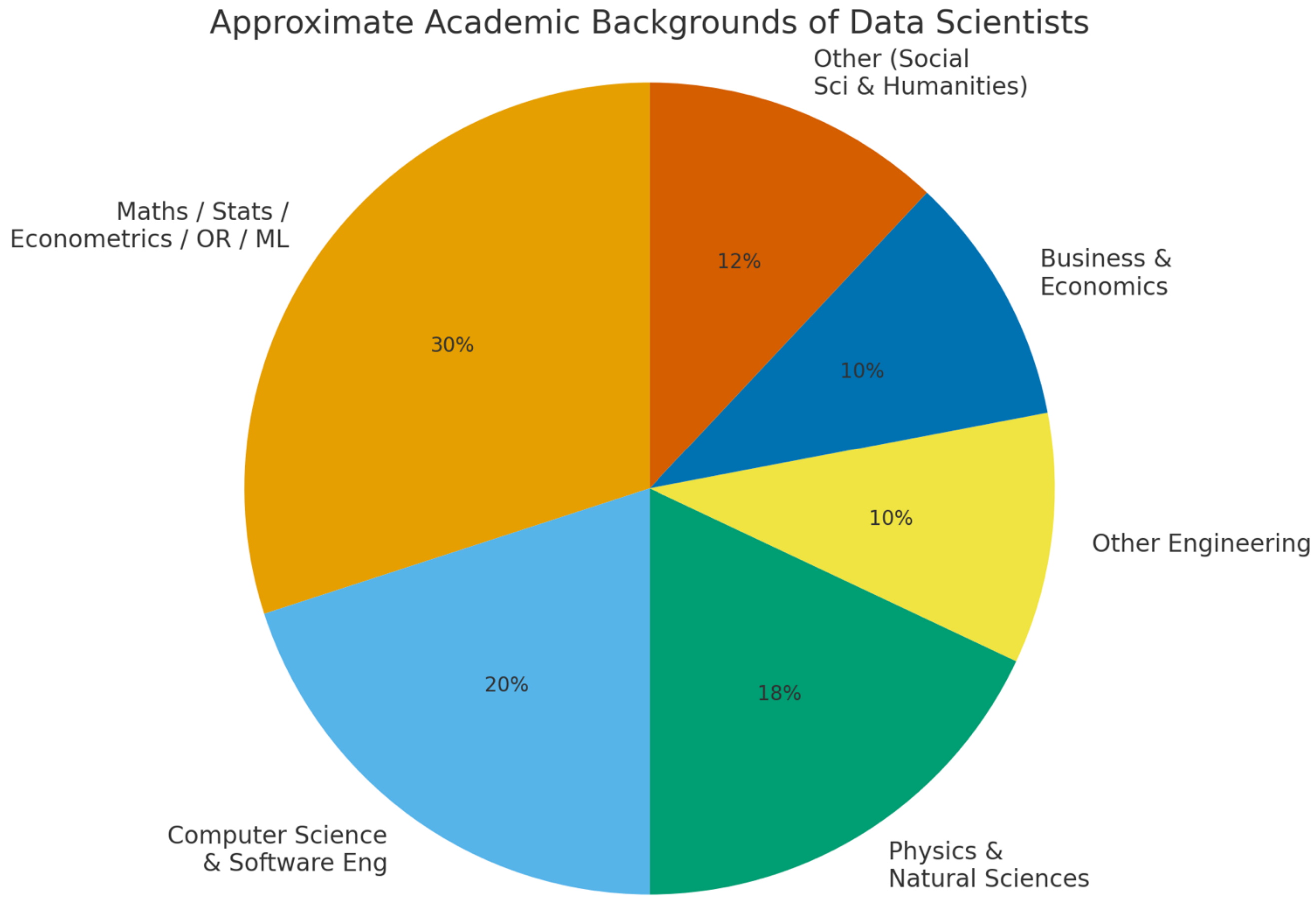


Values for this point have partial (incomplete) data available

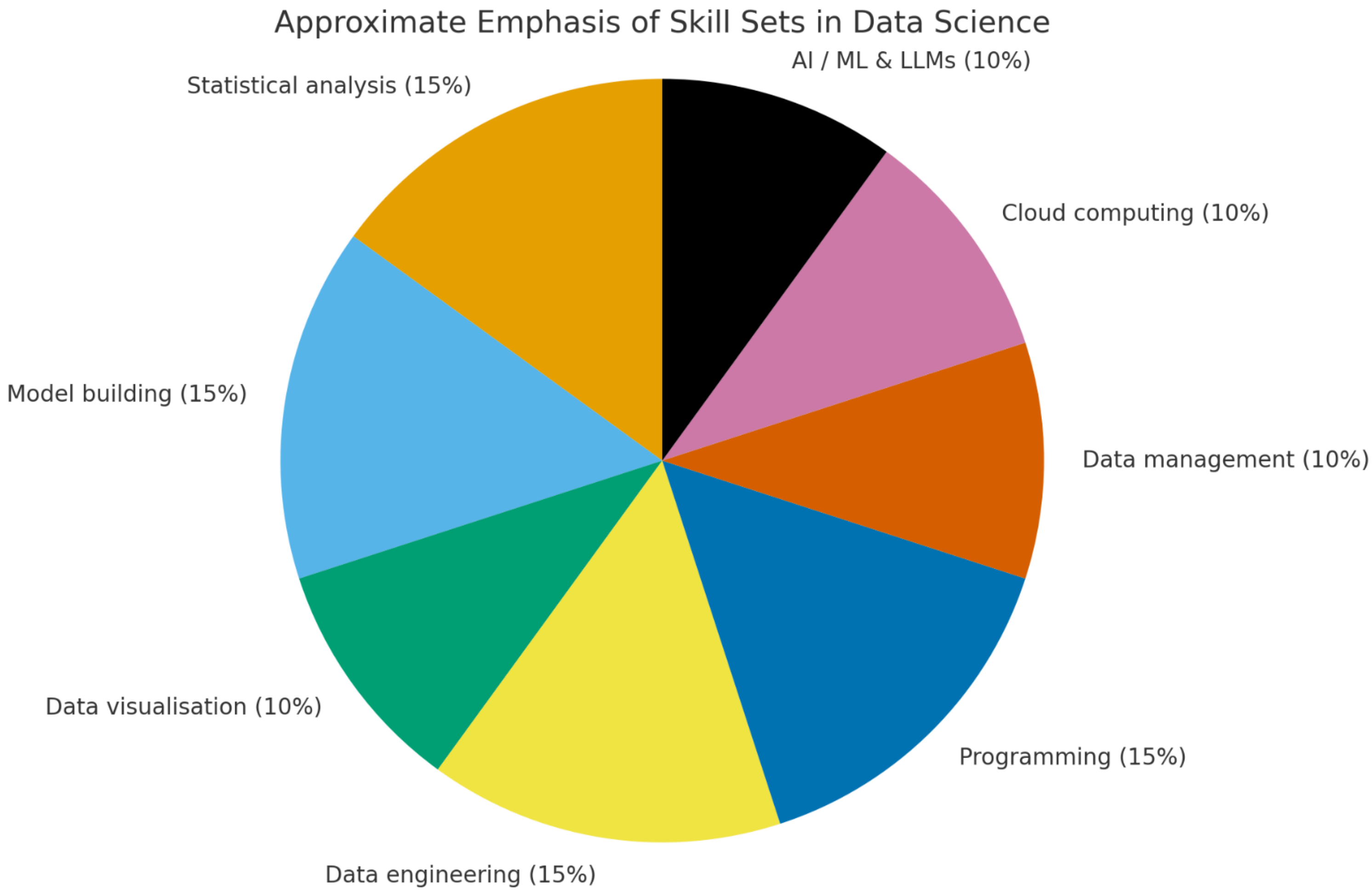


Data Science is using data, statistics and computing power to answer questions and make better decisions

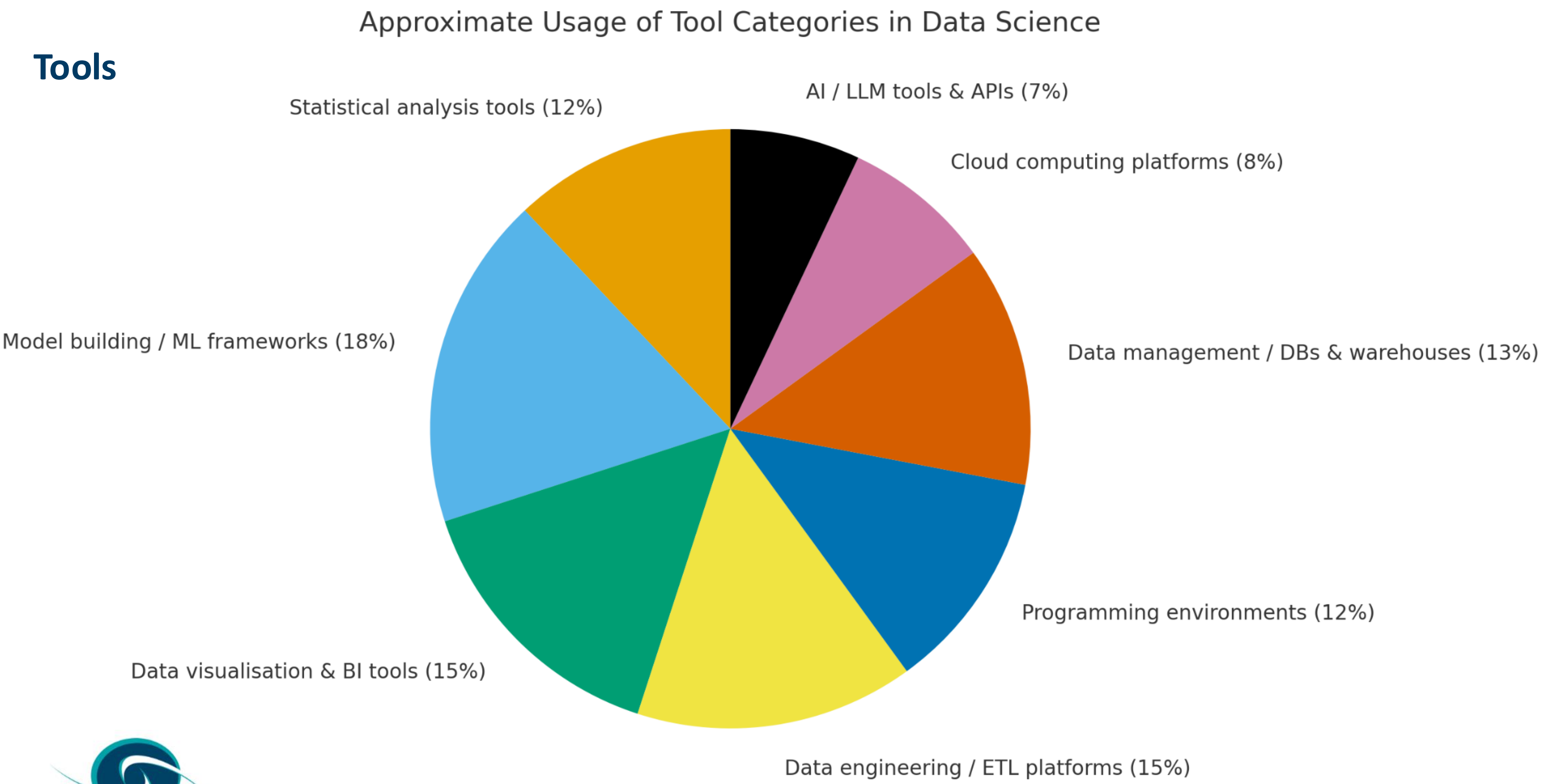
Disciplines



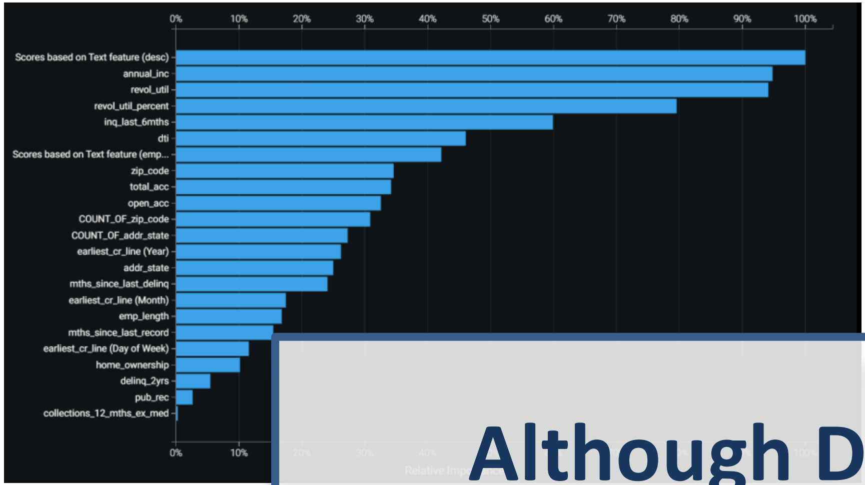
Skills

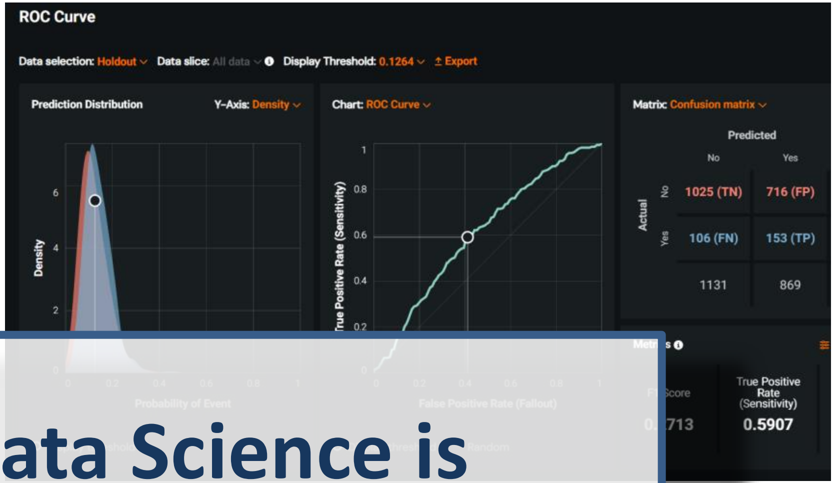

Approximate Emphasis of Skill Sets in Data Science
AI / ML & LLMs (10%)
Cloud computing (10%)
Data management (10%)
Programming (15%)
Data engineering (15%)
Data visualisation (10%)
Model building (15%)
Statistical analysis (15%)

Tools


Approximate Usage of Tool Categories in Data Science
AI / LLM tools & APIs (7%)
Cloud computing platforms (8%)
Data management / DBs & warehouses (13%)
Programming environments (12%)
Data engineering / ETL platforms (15%)
Data visualisation & BI tools (15%)
Model building / ML frameworks (18%)
Statistical analysis tools (12%)

What does Data Science actually produce?




ROC Curve

Data selection: Holdout Data slice: All data Display Threshold: 0.1264 Export

Prediction Distribution Y-Axis: Density

Chart: ROC Curve

Matrix: Confusion matrix

		Predicted	
		No	Yes
Actual	No	1025 (TN)	716 (FP)
	Yes	106 (FN)	153 (TP)
		1131	869

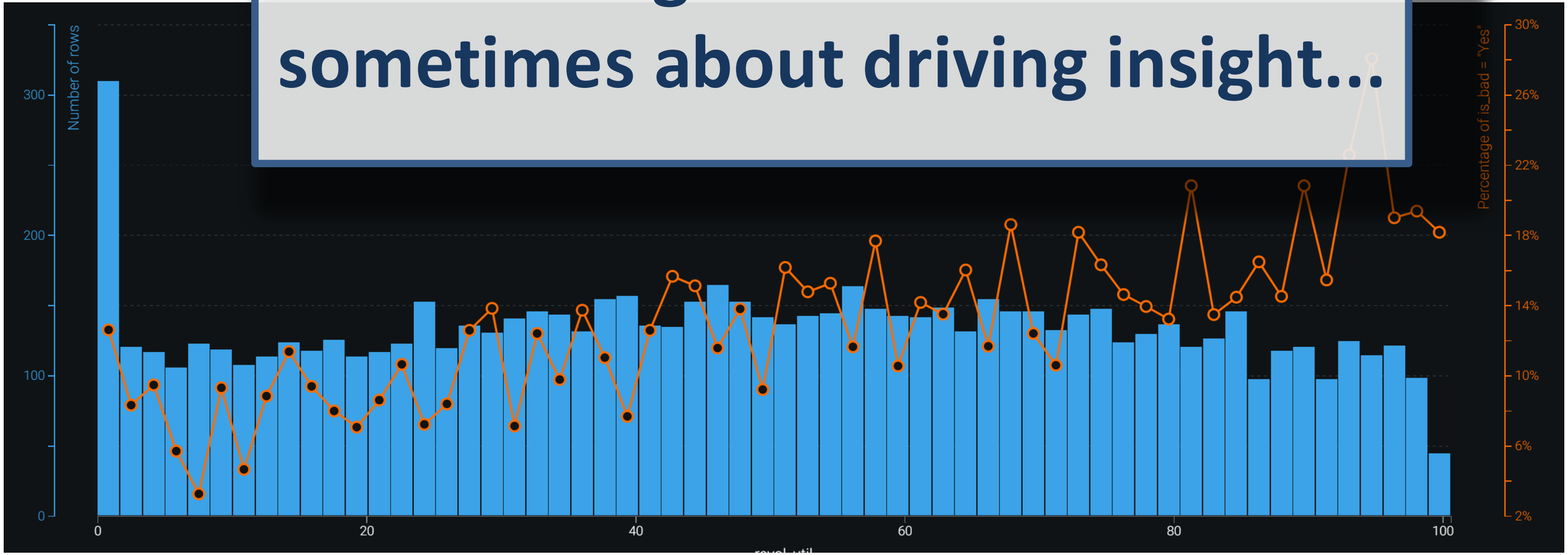
Metrics

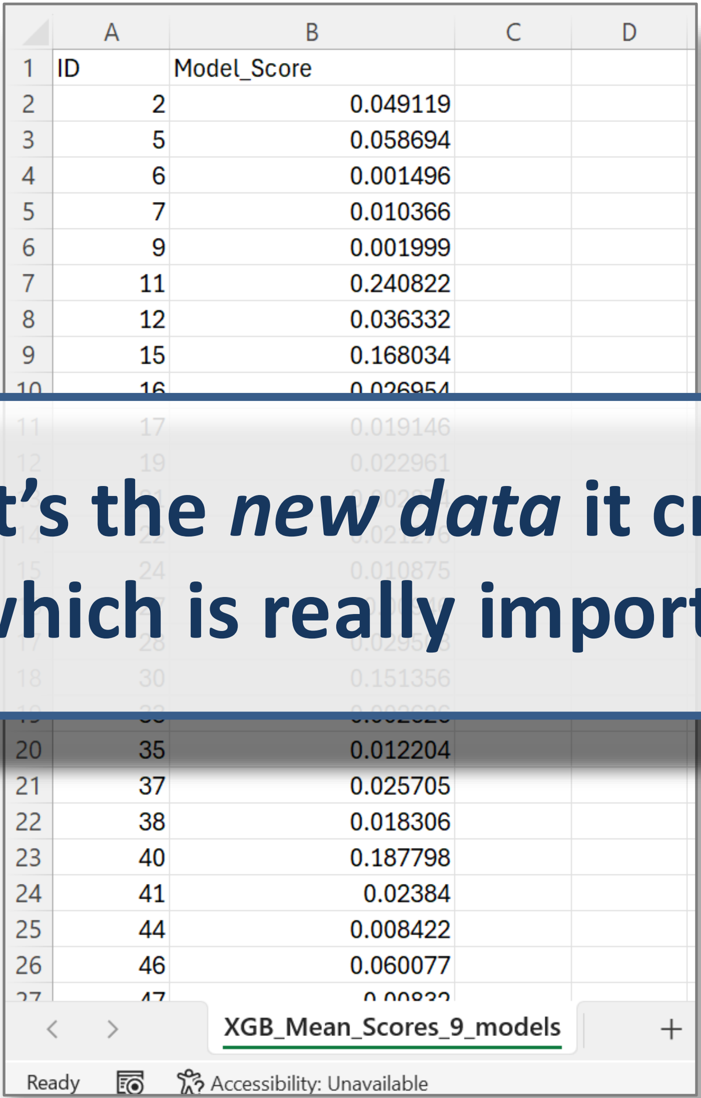
F1 Score 0.?713

True Positive Rate (Sensitivity) 0.5907

Density
6
4
2
0
0.2
0.4
0.6
0.8
1
Probability of Event
True Positive Rate (Sensitivity)
False Positive Rate (Fallout)
Random

Although Data Science is sometimes about driving insight…





	A	B	C	D
1	ID	Model_Score		
2	2	0.049119		
3	5	0.058694		
4	6	0.001496		
5	7	0.010366		
6	9	0.001999		
7	11	0.240822		
8	12	0.036332		
9	15	0.168034		
10	16	0.026954		
11	17	0.019146		
12	19	0.022961		
13	21	0.002974		
14	22	0.021276		
15	24	0.010875		
16	27	[illegible]		
17	28	[illegible]		
18	30	0.151356		
19	33	0.002626		
20	35	0.012204		
21	37	0.025705		
22	38	0.018306		
23	40	0.187798		
24	41	0.02384		
25	44	0.008422		
26	46	0.060077		
27	47	0.00832		

…it's the *new data* it creates which is really important.

SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

What do these new data represent?

- **Likelihoods**
  - recommend to a friend / complete a tv series / renew a subscription / click an offer / return to a store / make an insurance claim / choose a university / require maintenance / need a biopsy / make a complaint / fail a warranty / complete a course / return to hospital / fall into arrears / leave employment / defect to a competitor / commit fraud / show up for a flight / repay a loan / cause an accident / prevent infection / report a crime / vote for a party

What do these new data represent?

- **Estimates & Forecasts**
  - Student scores / regional sales / time to completion / blood pressure readings / pollution levels / website hits / survival times / growth rates / museum visits / medical costs / fuel consumption / crop yields / traffic volumes / causality patients / monthly expenditures / pupil numbers / power consumption / maintenance jobs / supply interruptions / flooding events / passenger volumes / property prices / infection rates / tickets sold

What do these new data represent?

- **Categories & Recommendations**
  - Customer segments / fault causes / medical diagnoses / tumour classes / replacement parts / treatment risk groups / preferred movie genres / political affiliations / fashion preferences / mobile phone plans / satisfaction levels / recommended crop types / product assortments / suggested drug regimes / targeted advert recommendations / content filters / document categories / customer sentiments / image classifications / speech-emotion classes

Typical Data Science Applications

- Customised Offer Creation
- Subscriber Retention
- Drug Performance Prediction
- Patient Outcomes/Readmission
- Predictive Maintenance/Asset Failure
- Fraud Detection
- Loyalty Modelling
- Path to Purchase

- Capacity planning & scheduling
- Anomaly Detection
- Association Analysis
- Forecasting
- Chatbots and virtual assistants
- Text Classification
- Optimisation Engines

It’s important to understand that depending on the circumstances, some of these applications may be driven by old statistical methods whilst others rely on cutting edge AI algorithms

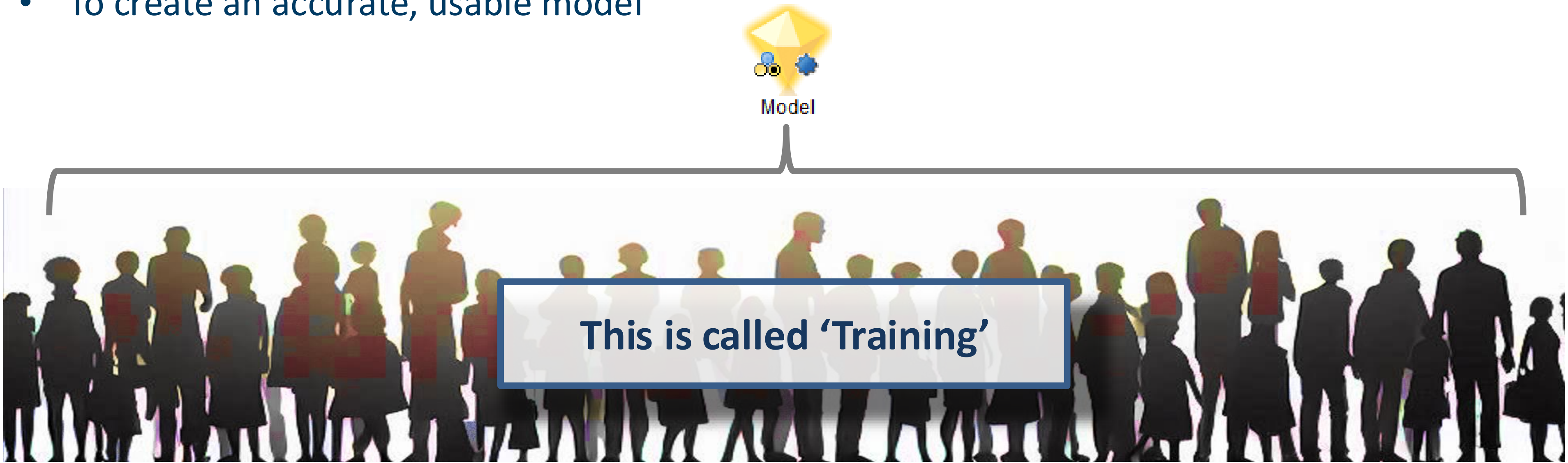
So where does AI like Chat GPT fit into Data Science?

- **Code Generation and Debugging**: Writing code snippets and fixing errors
- **Feature Engineering**: Helping to prepare the data for modelling
- **Data Transformation**: Turning unstructured data like text, images and audio into structured data like numbers
- **Exploratory Data Analysis (EDA):** Carrying out basic data analysis and generating summary reports and visualisations
- **Documentation and Explanation:** making projects more transparent and understandable for other team members

Building a Data Science Model

At the heart of a Data Science application is a model

- Typically uses historical data from many people/incidents/assets
- Age, Gender, Spending, Region, Tenure, Usage etc.
- With a known outcome/result
- Responded, upgraded, defaulted, recommended, cancelled, donated, failed, renewed etc.
- To create an accurate, usable model



This is called ‘Training’

At the heart of a Data Science application is a model

- We can take new data from new individuals or incidents…
- Age, Gender, Spending, Region, Tenure, Usage etc.
- Using a model based on the same information…
- Generate likelihood scores, estimates and classifications
- In other words,…..predictions

This is called ‘Scoring’

32% CHANCE OF CANCELLATION

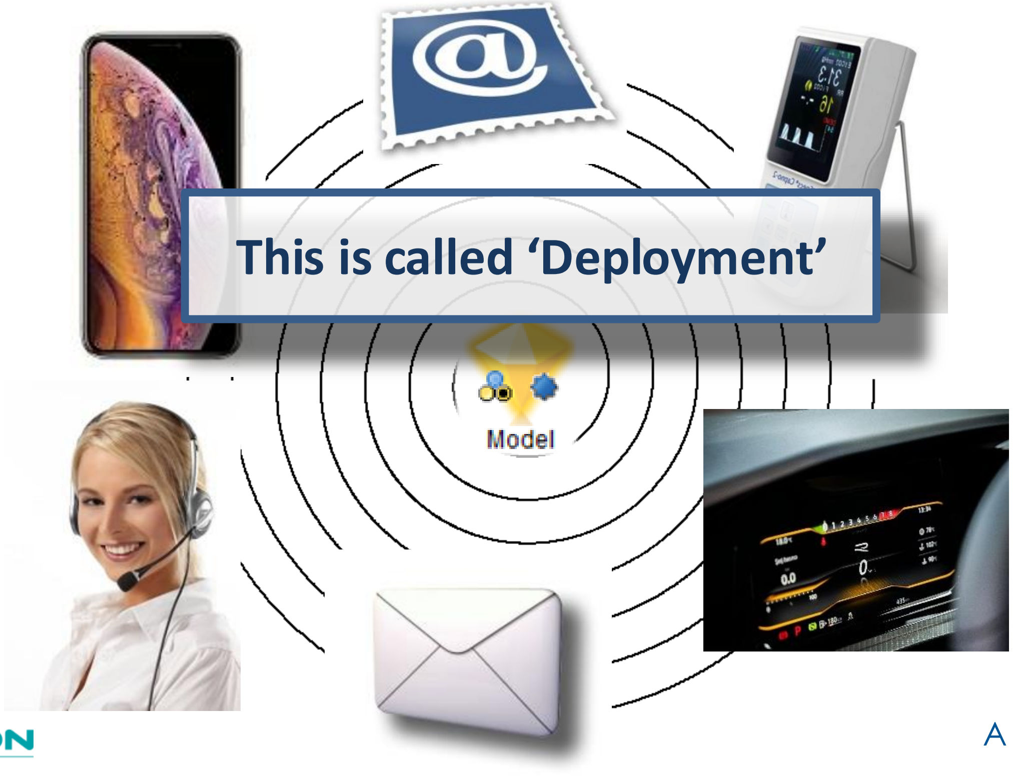
Predicted 12 month spend = £938

0.13 probability of defaulting

Recommended Genre = A12 – True Crime Drama

At the heart of a Data Science application is a model

- We can then send the model scores to different platforms to drive better outcomes



This is called 'Deployment'

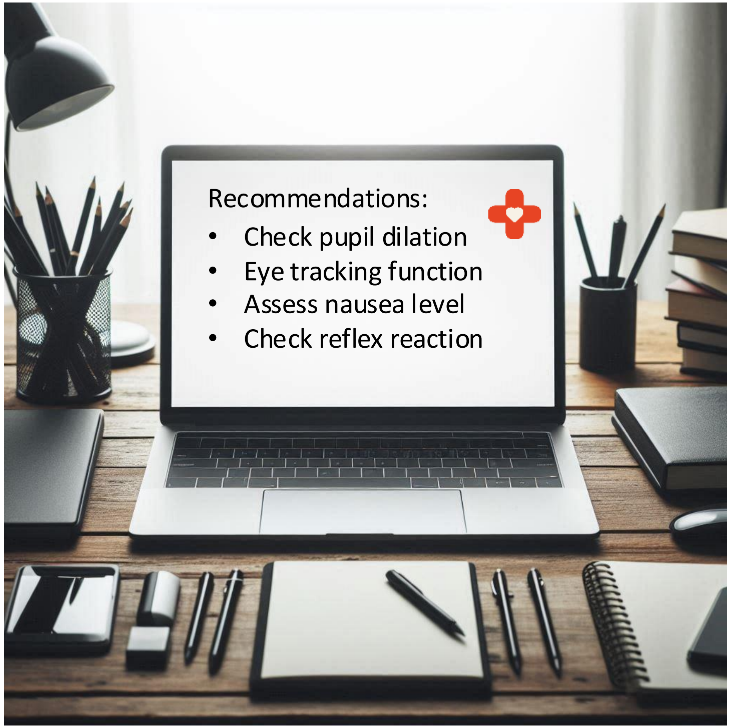
However, a Model is *not* an Application…

Until it is used in the real world to drive outcomes

$$
\begin{aligned}
\text{maximize } f(c_1 \ldots c_n) &= \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i (\varphi(\mathbf{x}_i)\cdot\varphi(\mathbf{x}_j)) y_j c_j \\
&= \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i k(\mathbf{x}_i,\mathbf{x}_j) y_j c_j
\end{aligned}
$$

$$
\text{subject to } \sum_{i=1}^{n} c_i y_i = 0, \text{ and } 0 \le c_i \le \frac{1}{2n\lambda} \text{ for all } i.
$$

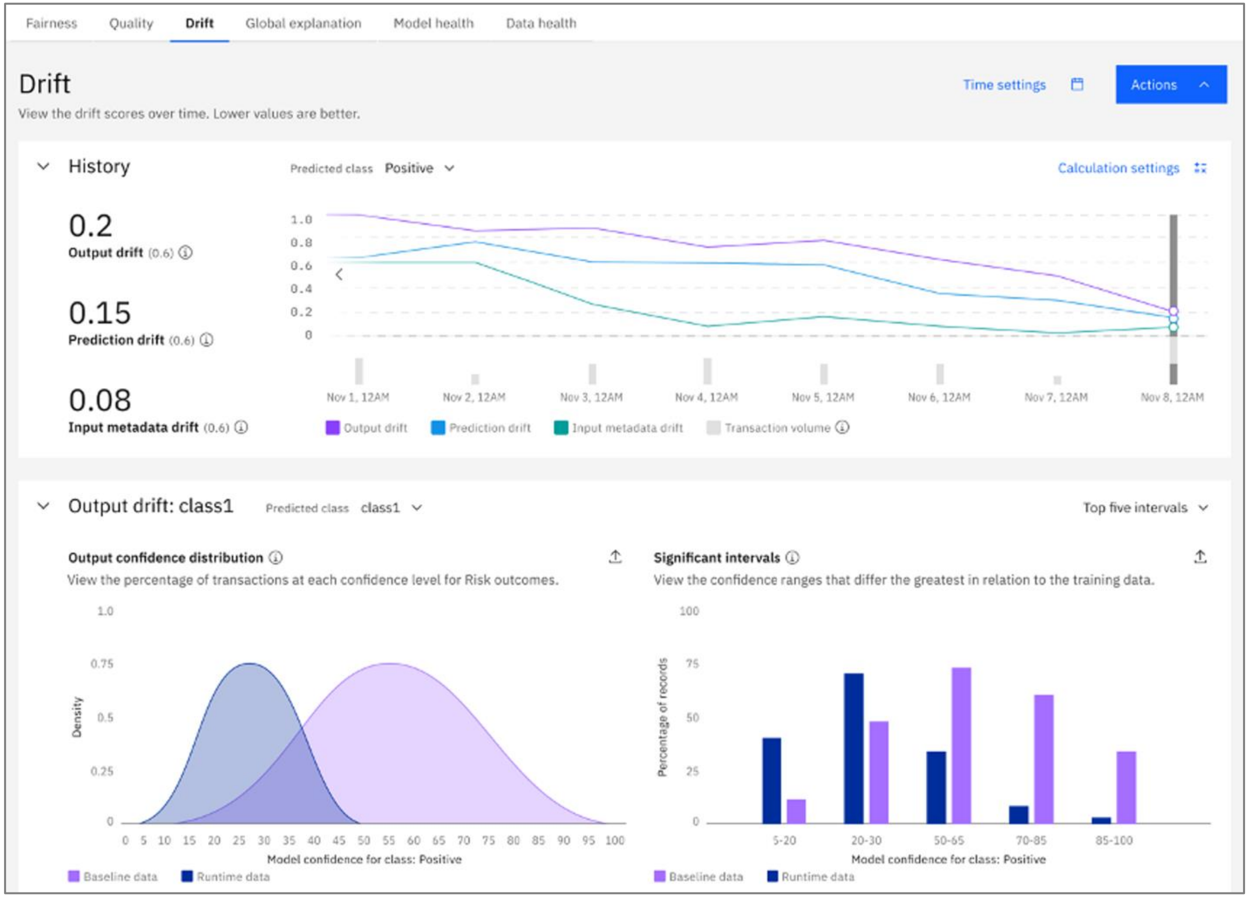
$\neq$



Recommendations:

- Check pupil dilation
- Eye tracking function
- Assess nausea level
- Check reflex reaction

And applications usually require Governance



Fairness Quality **Drift** Global explanation Model health Data health

Drift

View the drift scores over time. Lower values are better.

Time settings Actions

History Predicted class Positive Calculation settings

0.2
Output drift (0.6)

0.15
Prediction drift (0.6)

0.08
Input metadata drift (0.6)

Output drift: class1 Predicted class class1 Top five intervals

Output confidence distribution

View the percentage of transactions at each confidence level for Risk outcomes.

Significant intervals

View the confidence ranges that differ the greatest in relation to the training data.

And applications usually require Governance

- **Prevents harmful or biased outcomes:** Rules for data access, model use, and fairness checks, reducing the risk of discrimination, reputational damage.
- **Keeps models accurate over time (drift monitoring):** Monitoring picks up when data or behaviour changes (model drift), so you can retrain, recalibrate, or retire models before performance falls off a cliff.
- **Supports compliance and auditing:** Documentation and audit trails for models (who built it, what data, which checks) to help satisfy regulators, internal audit, and legal – especially in finance, healthcare, HR, and credit.
- **Improves reproducibility and handover:** Good practice standards for version controls, security, and model lifecycle ensure others can reproduce results, debug issues, and safely update or extend a model.
- **Builds trust with stakeholders:** Decision makers are reassured there's a formal process for reviewing, approving, and monitoring models in production, they're more willing to rely on data science for real decisions, not just "nice dashboards."

What are the real-world challenges with Data Science?

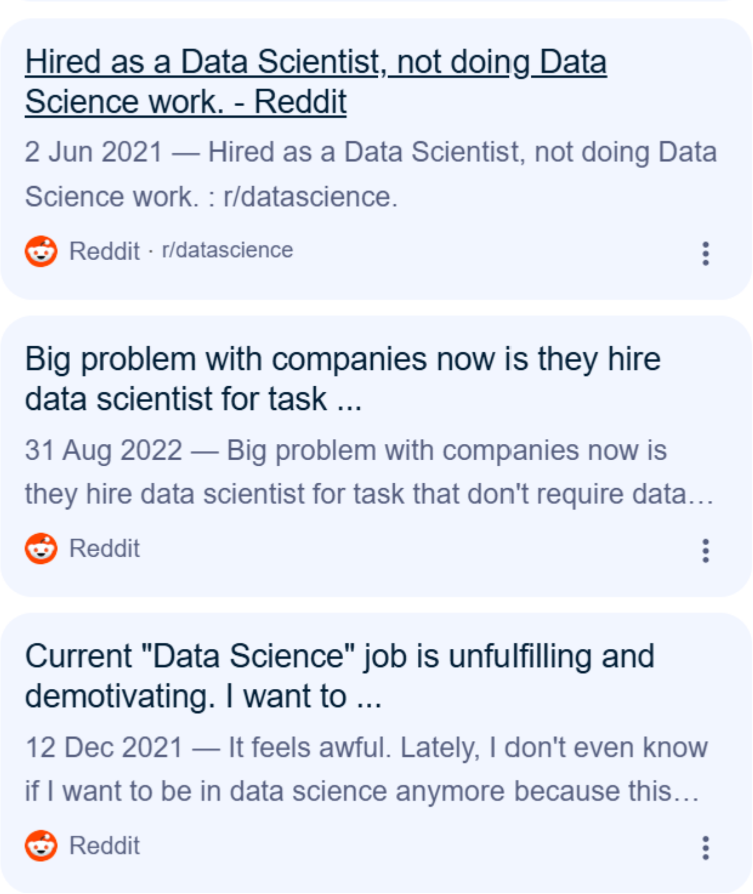
If you build it,
they will come

All the gear but no idea

- Even big companies make the mistake of thinking that Data Science/AI is all about having the right resources:
    - A new data science team
    - A cloud-based AI platform
    - Sophisticated data storage/process architecture

Do they have a use for Data Science or AI?

- A regular complaint among newly-hired but highly-qualified Data Scientists and AI specialists is that they find their roles consist of fairly basic analytical tasks such as running SQL queries or building dashboards
- Some companies may use the term "data scientist" as a buzzword to attract talent, without a clear understanding of what the role entails



Hired as a Data Scientist, not doing Data Science work. - Reddit

2 Jun 2021 — Hired as a Data Scientist, not doing Data Science work. : r/datascience.

Reddit · r/datascience

Big problem with companies now is they hire data scientist for task ...

31 Aug 2022 — Big problem with companies now is they hire data scientist for task that don't require data...

Reddit

Current "Data Science" job is unfulfilling and demotivating. I want to ...

12 Dec 2021 — It feels awful. Lately, I don't even know if I want to be in data science anymore because this...

Reddit

SMARTVISION Europe

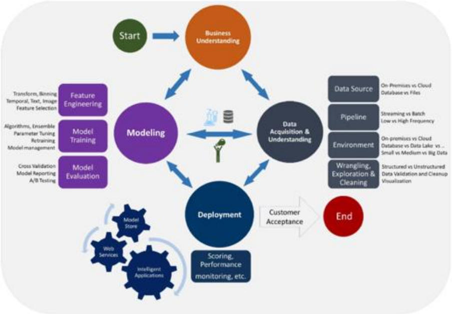
A SELECT INTERNATIONAL COMPANY

What are the biggest challenges in implementing Data Science?

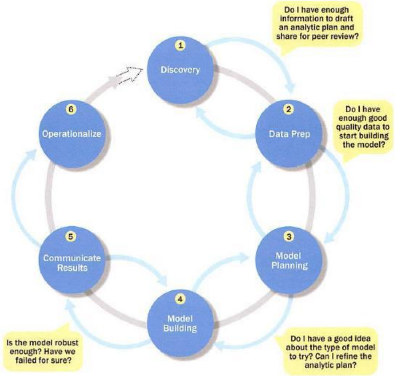
- The time and effort taken to consolidate, blend and prepare data so it can used effectively
- Coordination and communication across business units
- Matching the capabilities of data science to the needs of the organisation i.e. creating valuable applications
- Measuring the value of the application
- Creating a feedback cycle to manage things operationally

It’s helpful to know there are several methodologies dedicated Data Science

- Microsoft’s Team Data Science Process (TDSP)

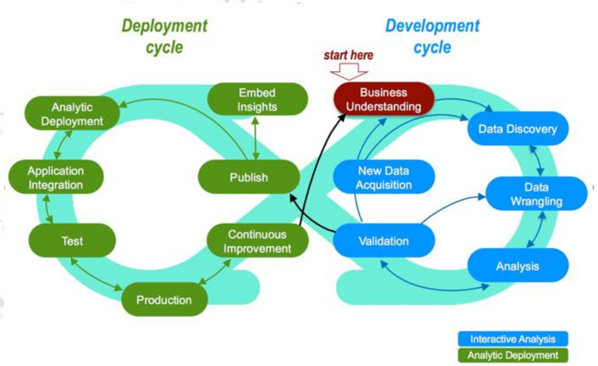


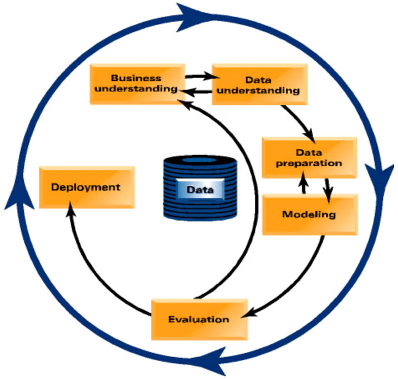
- EMC’s Data Analytics Lifecycle



Do I have enough information to draft an analytic plan and share for peer review?
1 Discovery
2 Data Prep
Do I have enough good quality data to start building the model?
3 Model Planning
4 Model Building
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
5 Communicate Results
Is the model robust enough? Have we failed for sure?
6 Operationalize

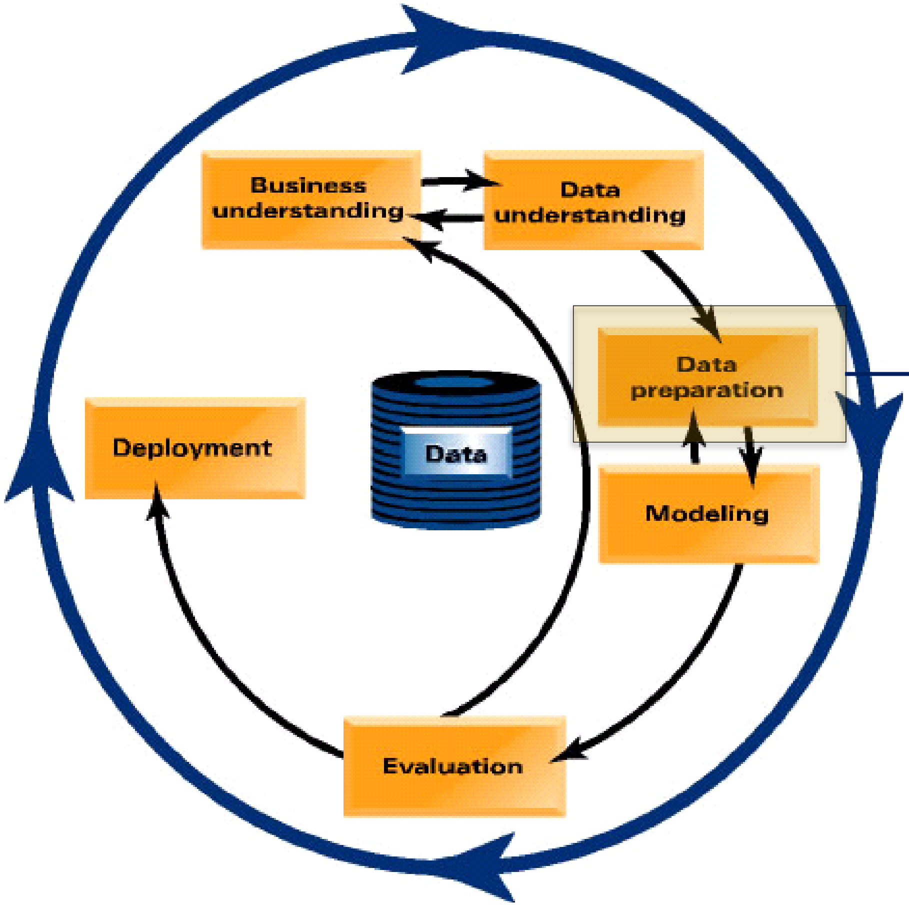
IBM’s Analytics Solution Unified Method (ASUM-DM)





- Cross-Industry Standard Process for Data Mining (CRISP-DM)

And they illustrate that it's not all just building models



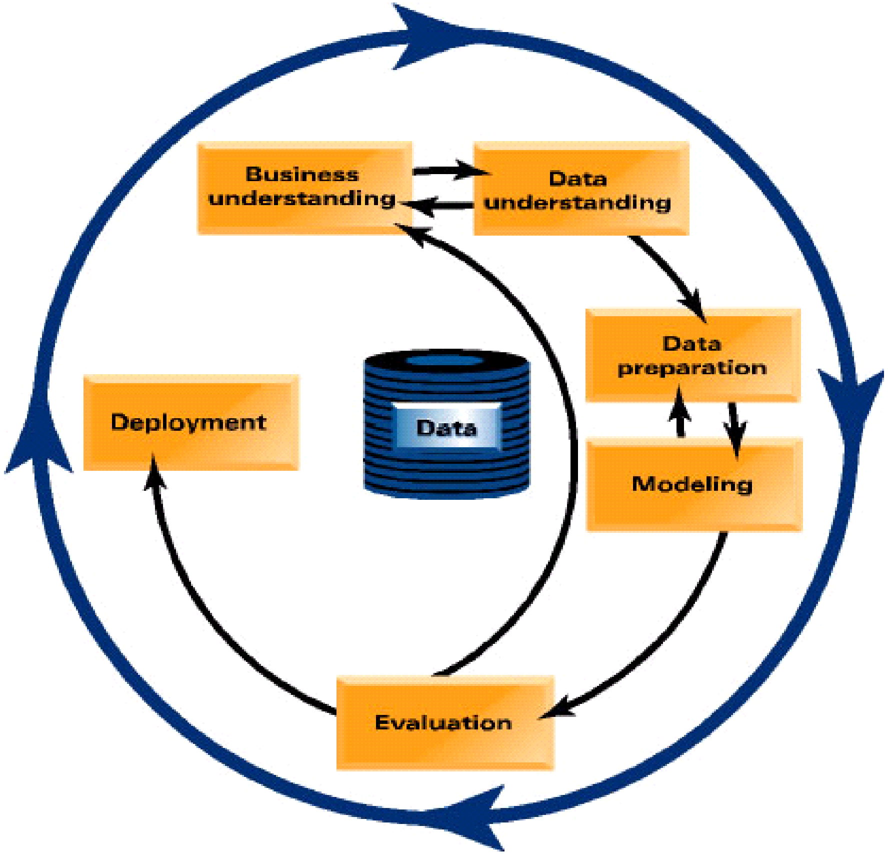
Often Data Scientists may spend 50 % to 70% of their time just wrangling and preparing the data when working on a new project

Important questions for intrepid adventurers

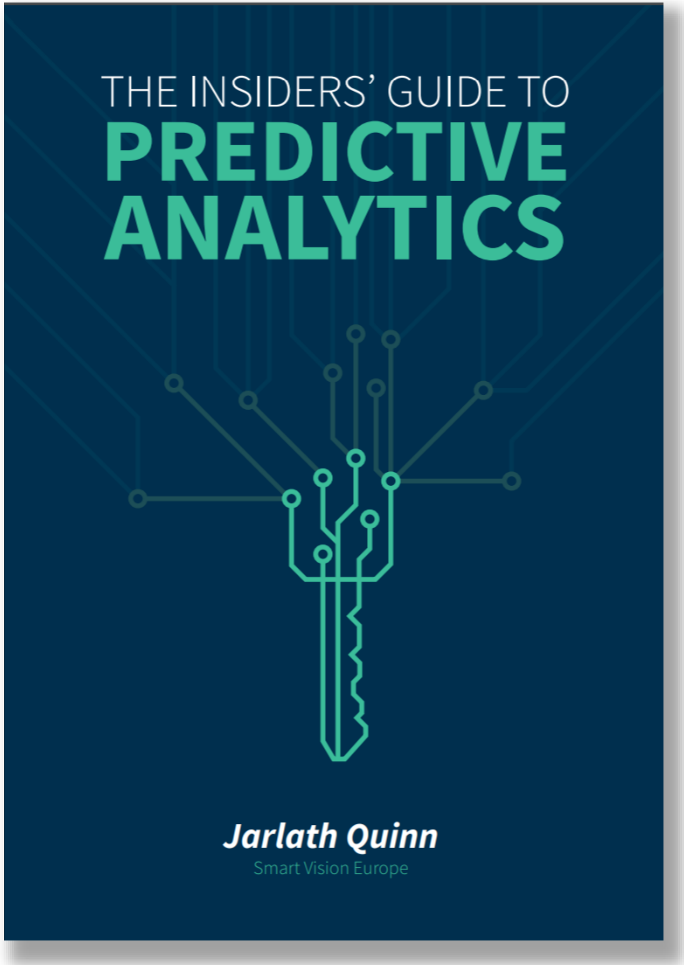
- What does ‘good’ look like?
- What will you do differently?
- How will you know it worked?
- Does everyone agree or understand?
- What methodology will you use?

Think of a methodology as a route map to successful deployment

- CRISP-DM: Cross-Industry Standard Process for Data Mining
- Each application can be developed and progressed through a series of key phases
- http://crisp-dm.eu/



Download our e-book for free



THE INSIDERS' GUIDE TO

PREDICTIVE ANALYTICS

Jarlath Quinn

Smart Vision Europe

The insider's guide to predictive analytics

£0.00

- 1 + Add to basket

Category: books

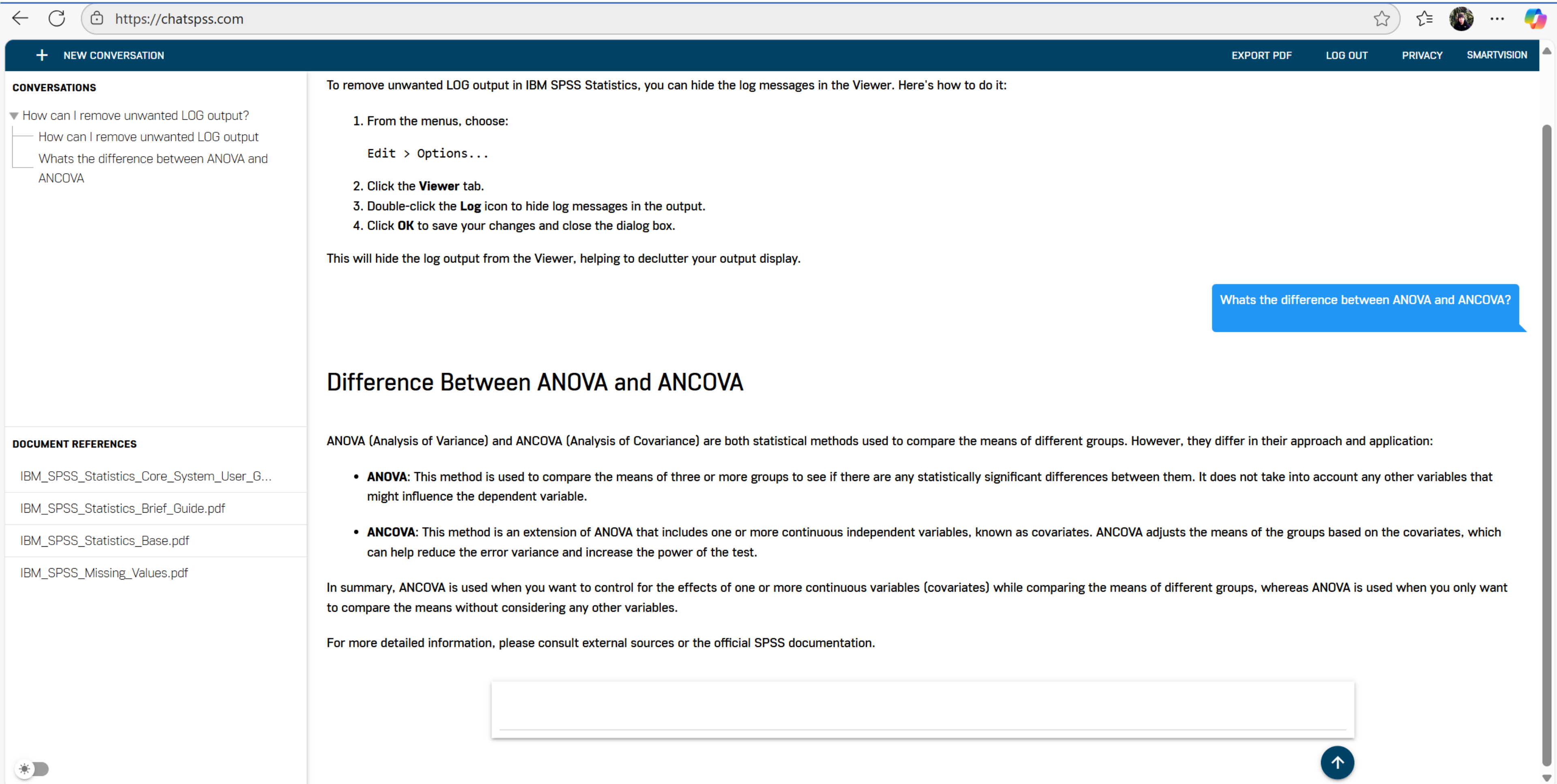
SMARTVISION

Europe

A SELECT INTERNATIONAL COMPANY

watsonx.ai™

ChatSPSS


https://chatspss.com
NEW CONVERSATION
EXPORT PDF
LOG OUT
PRIVACY
SMARTVISION
CONVERSATIONS
How can I remove unwanted LOG output?
How can I remove unwanted LOG output
Whats the difference between ANOVA and ANCOVA
DOCUMENT REFERENCES
IBM_SPSS_Statistics_Core_System_User_G...
IBM_SPSS_Statistics_Brief_Guide.pdf
IBM_SPSS_Statistics_Base.pdf
IBM_SPSS_Missing_Values.pdf

To remove unwanted LOG output in IBM SPSS Statistics, you can hide the log messages in the Viewer. Here's how to do it:

1. From the menus, choose:

   ```
   Edit > Options...
   ```

2. Click the **Viewer** tab.
3. Double-click the **Log** icon to hide log messages in the output.
4. Click **OK** to save your changes and close the dialog box.

This will hide the log output from the Viewer, helping to declutter your output display.

Whats the difference between ANOVA and ANCOVA?

Difference Between ANOVA and ANCOVA

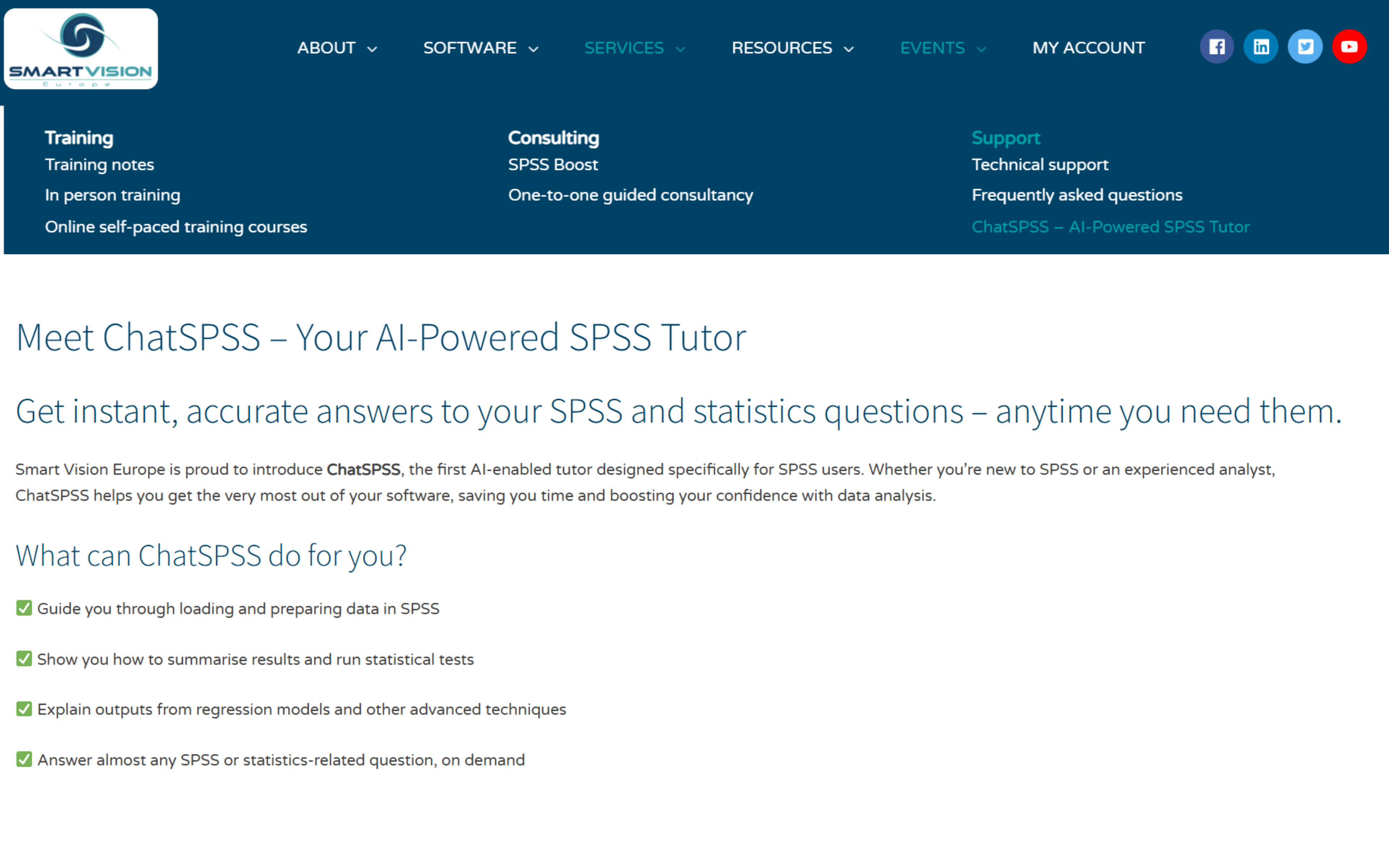
ANOVA (Analysis of Variance) and ANCOVA (Analysis of Covariance) are both statistical methods used to compare the means of different groups. However, they differ in their approach and application:

- **ANOVA**: This method is used to compare the means of three or more groups to see if there are any statistically significant differences between them. It does not take into account any other variables that might influence the dependent variable.

- **ANCOVA**: This method is an extension of ANOVA that includes one or more continuous independent variables, known as covariates. ANCOVA adjusts the means of the groups based on the covariates, which can help reduce the error variance and increase the power of the test.

In summary, ANCOVA is used when you want to control for the effects of one or more continuous variables (covariates) while comparing the means of different groups, whereas ANOVA is used when you only want to compare the means without considering any other variables.

For more detailed information, please consult external sources or the official SPSS documentation.



Meet ChatSPSS – Your AI-Powered SPSS Tutor

Get instant, accurate answers to your SPSS and statistics questions – anytime you need them.

Smart Vision Europe is proud to introduce **ChatSPSS**, the first AI-enabled tutor designed specifically for SPSS users. Whether you're new to SPSS or an experienced analyst, ChatSPSS helps you get the very most out of your software, saving you time and boosting your confidence with data analysis.

What can ChatSPSS do for you?

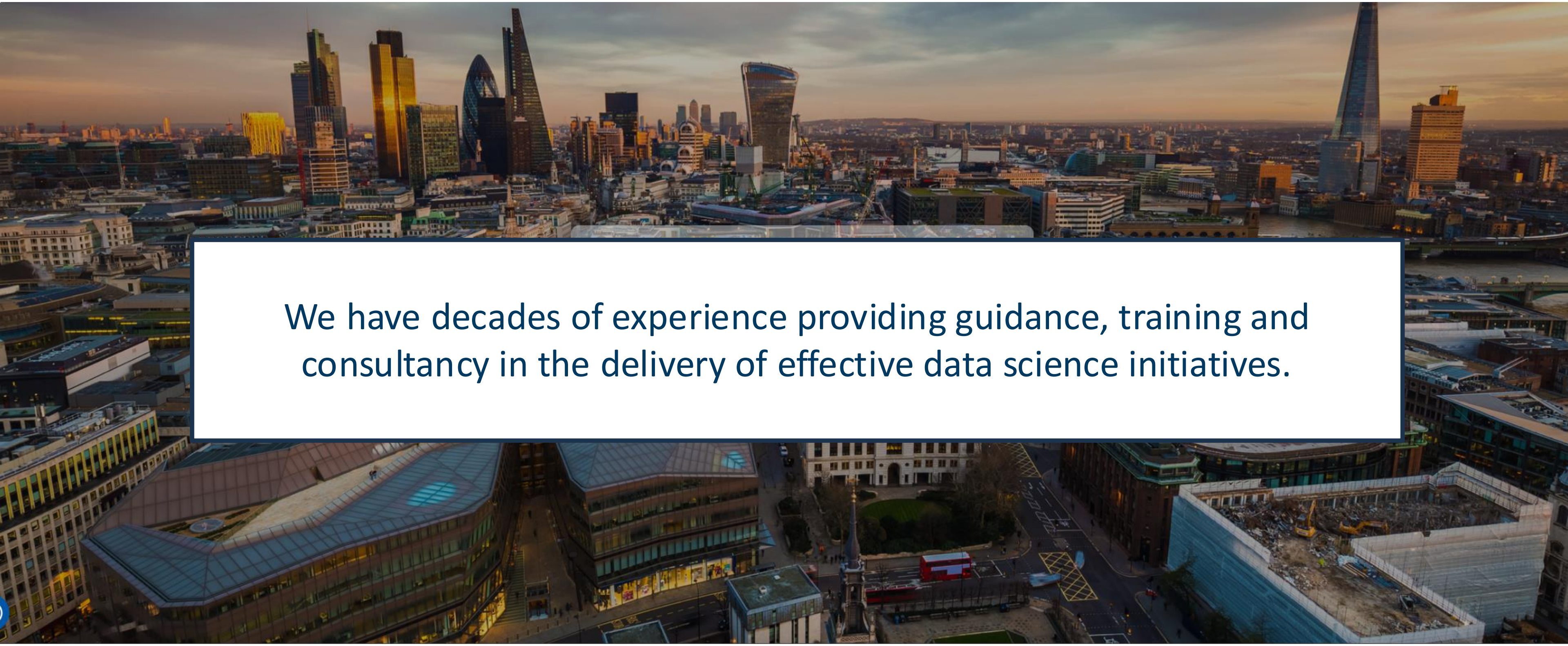
✅ Guide you through loading and preparing data in SPSS

✅ Show you how to summarise results and run statistical tests

✅ Explain outputs from regression models and other advanced techniques

✅ Answer almost any SPSS or statistics-related question, on demand

Smart Vision Europe: Services and Expertise



We have decades of experience providing guidance, training and consultancy in the delivery of effective data science initiatives.

A SELECT INTERNATIONAL COMPANY

SMARTVISION
Europe

Contact us:

+44 (0)207 786 3568
info@sv-europe.com
Twitter: @sveurope
Follow us on Linked In
Sign up for our Newsletter

Thank you