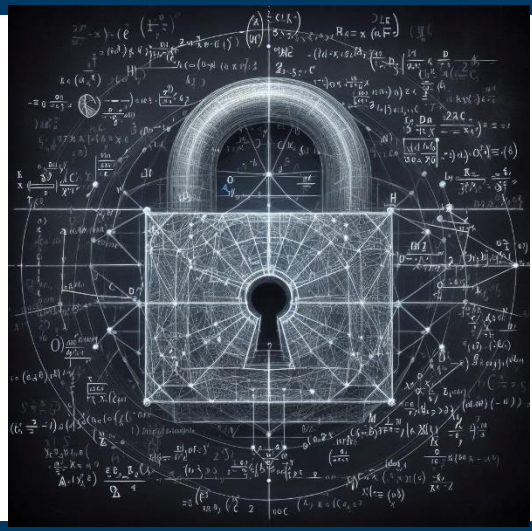
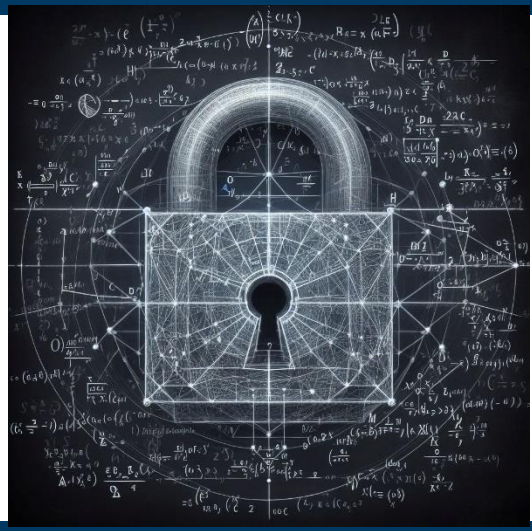


Data Science in an hour

Jarlath Quinn – Analytics Consultant

Just waiting for all attendees to join…

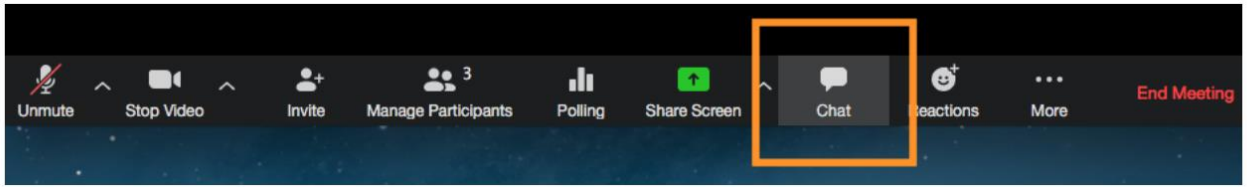


Data Science in an hour

Jarlath Quinn – Analytics Consultant

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.

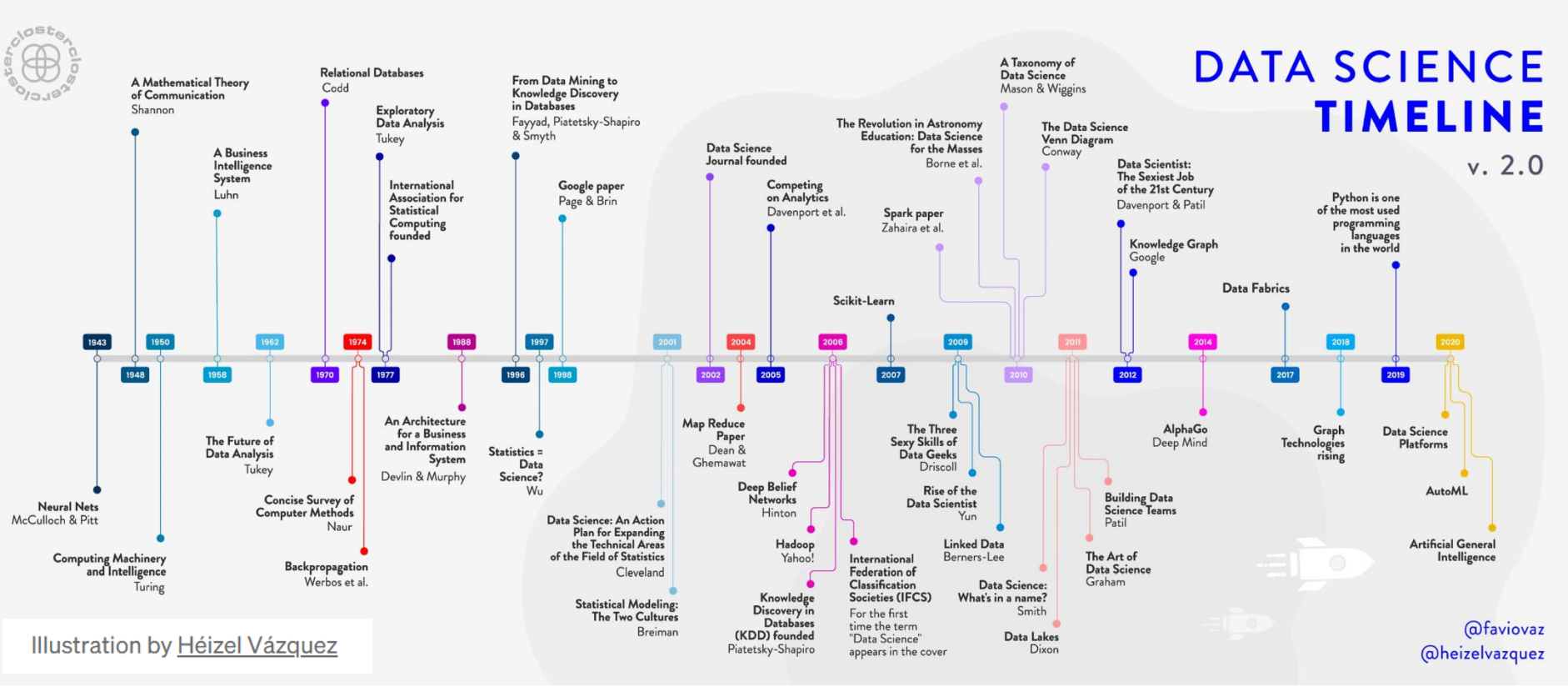


- Premier accredited partner to IBM, Predictive Solutions and DataRobot specialising in advanced analytics & big data technologies
- Work with open source technologies (R, Python, Spark etc.)
- Team each has 15 to 35 years of experience working in statistics and the advanced analytics industry

- Deep experience of applied advanced analytics applications across sectors
  - Retail
  - Gaming
  - Utilities
  - Insurance
  - Telecommunications
  - Media
  - FMCG

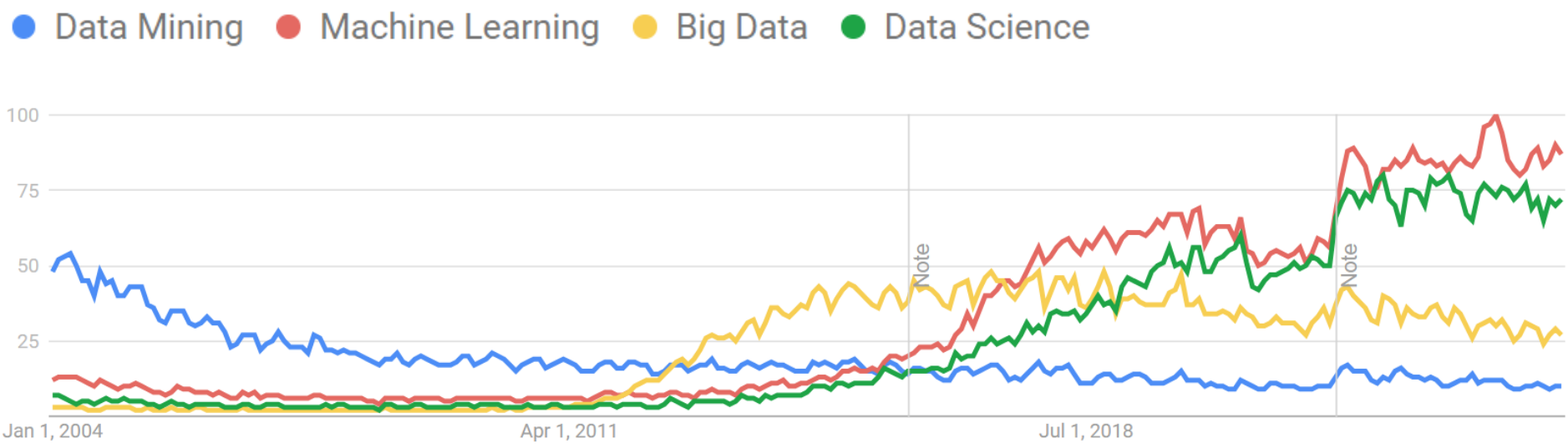
How did we get here?

Statistical Analysis to AI



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

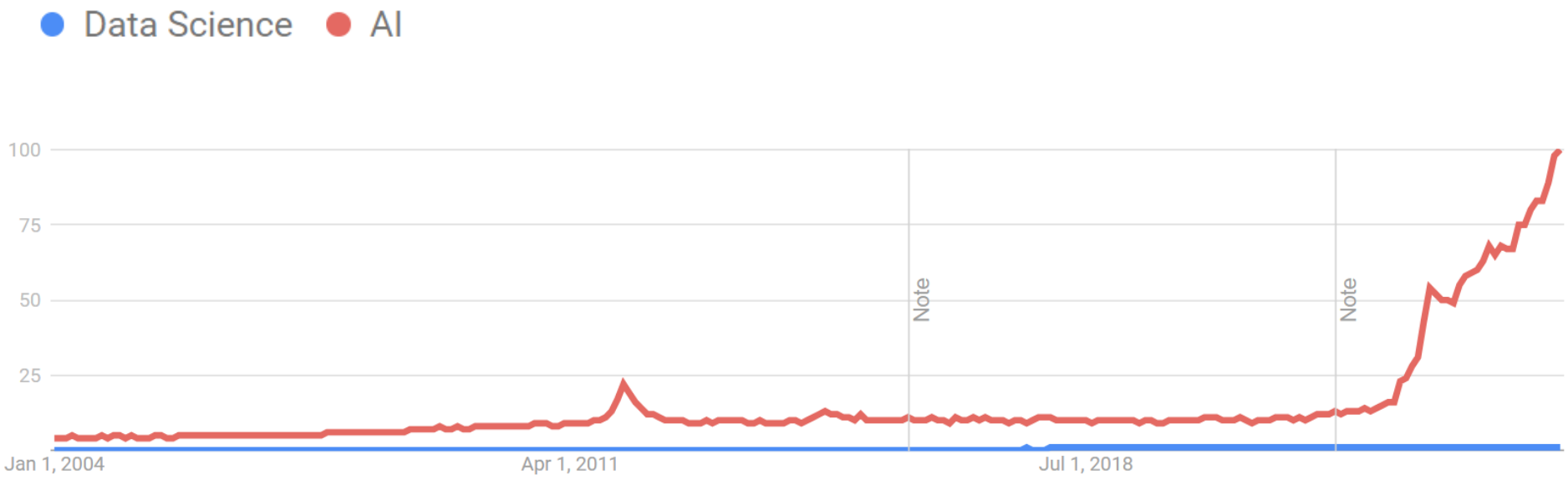
The new frontiers of data analysis



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Not forgetting of course…the elephant in the room

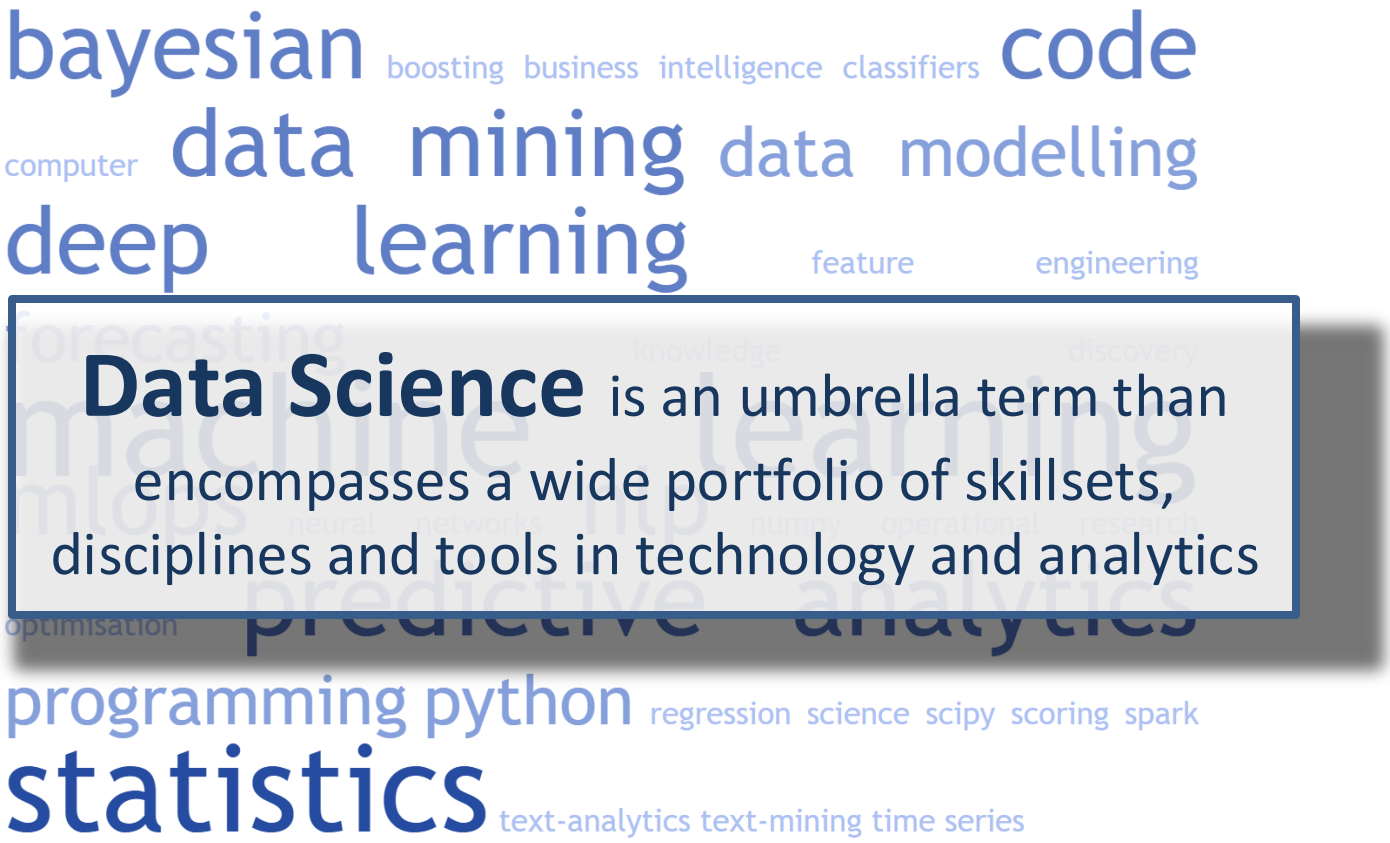


SMARTVISION Europe

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.

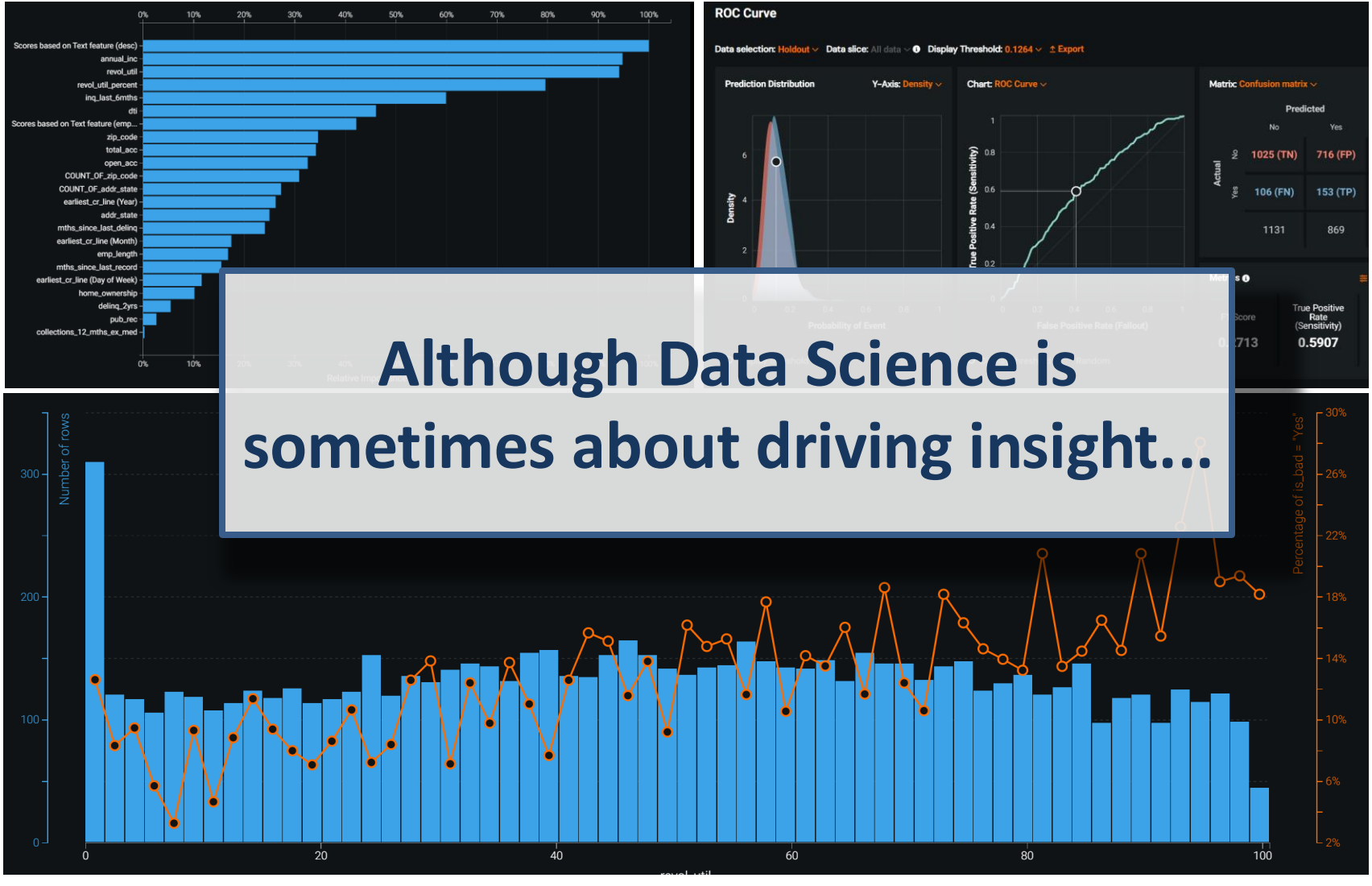


Data Science is an umbrella term than encompasses a wide portfolio of skillsets, disciplines and tools in technology and analytics

Data Science Expertise

- **Disciplines**
  - Statistics, machine learning, AI, physics, computer science, operational research, econometrics
- **Skills**
  - Statistical analysis, model building, data visualisation, data engineering, programming, data management, cloud computing, AI
- **Tools**
  - Python, R, Spark, SQL, Google Cloud, AWS, Databricks, IBM, Data Robot, Dataiku, Tensor Flow

What does Data Science actually produce?



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		
14	22	0.021[illegible]		
15	24	0.010875		
17	28	0.0256[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		

XGB_Mean_Scores_9_models

Ready Accessibility: Unavailable

...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 Outcome Prediction
- Predictive Maintenance
- Fraud Detection
- Loyalty Modelling
- Next-Best-Action

- Cluster Analysis
- Anomaly Detection
- Association Analysis
- Forecasting
- Text Analysis /Generative AI
- Video/Voice Analytics
- Optimisation Engines

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

Data Science Terminology

- Supervised vs Unsupervised Learning
- Structured vs Unstructured Data
- Loss function
- Deep Learning
- MLOPS (Machine Learning Operations)
- GitHub
- LLM (Large Language Model)

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

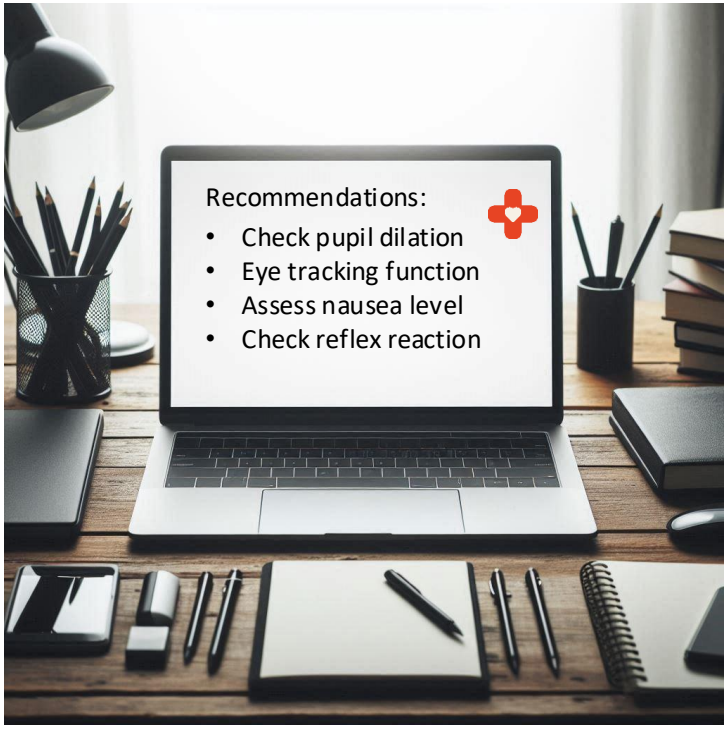
However, a Model is *not* an Application…

Until it is used in the real world to drive outcomes

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

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

$\neq$



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

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

Is there an actual need 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

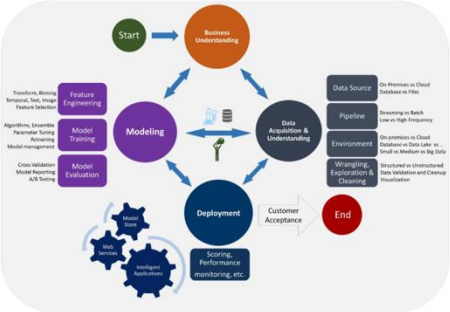
Current "Data Science" job is unfulfilling and demotivating. I want to ...

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

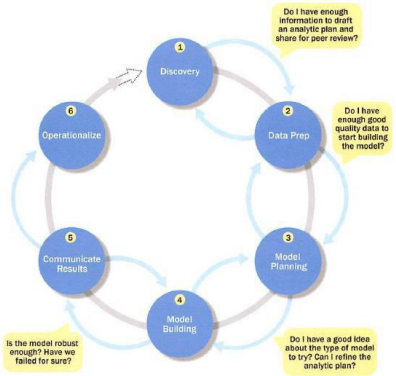
Reddit

It's useful to know there are several methodologies dedicated Data Science

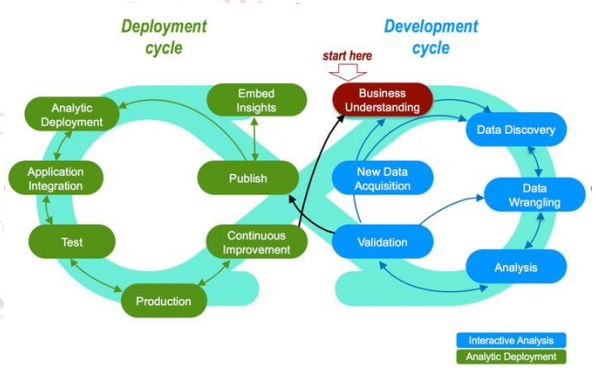
- Microsoft's Team Data Science Process (TDSP)

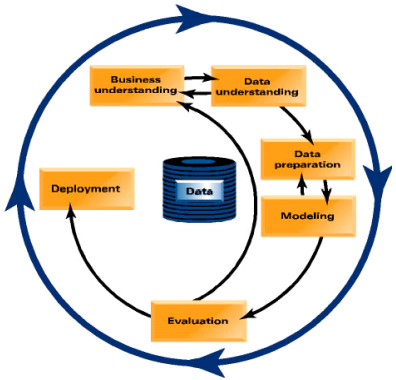


- EMC's Data Analytics Lifecycle



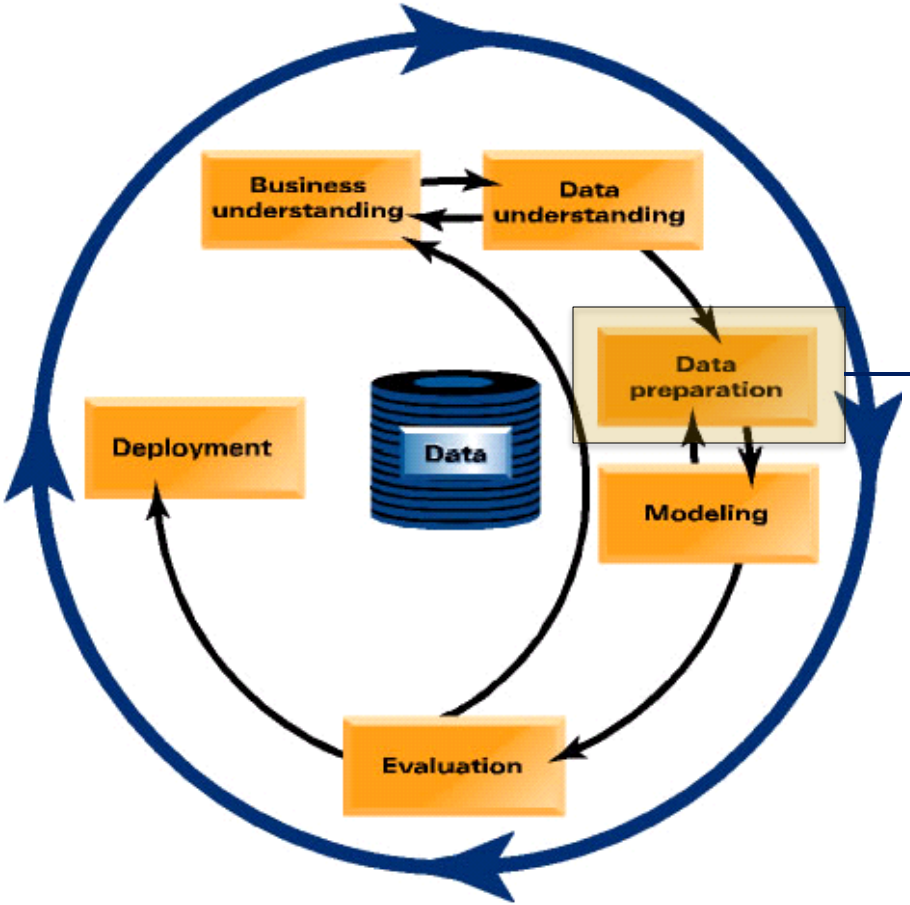
IBM's Analytics Solution Unified Method (ASUM-DM)





- 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

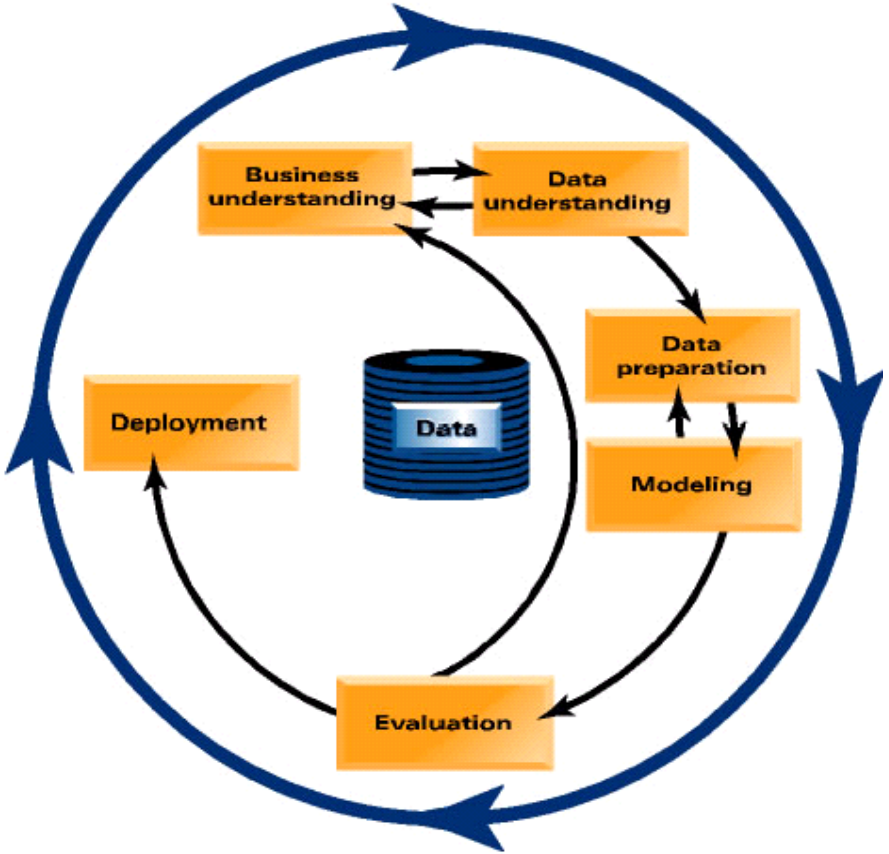
SMARTVISION
Europe

Important questions for intrepid adventurers

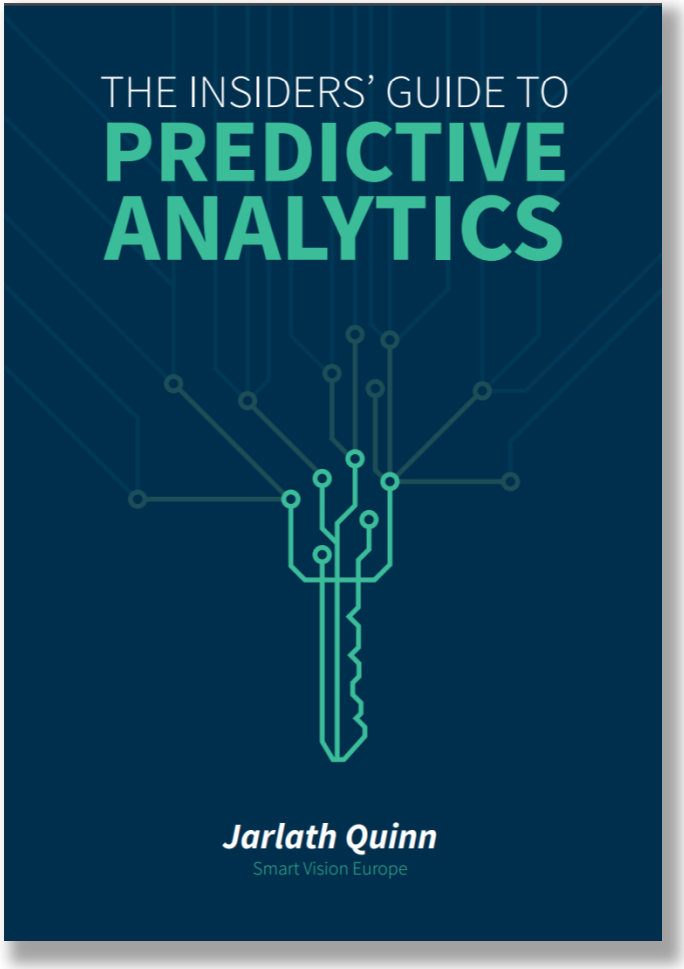
- What does ‘good’ look like?
- What will you do differently?
- How will you know it worked?
- Does everyone agree or understand?
- What methodology will you use?

Think 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 insider's guide to predictive analytics

£0.00

- 1 + Add to basket

Category: books

A SELECT INTERNATIONAL COMPANY

SMARTVISION Europe

Smart Vision Europe: Services and Expertise

Consultancy and Help

Guidance and support to help you get started. Embed our expertise alongside your team to help meet your objectives.

Training and Support

Educational support whether onsite or remote/virtual learning. Full day or half-day bite-size courses.

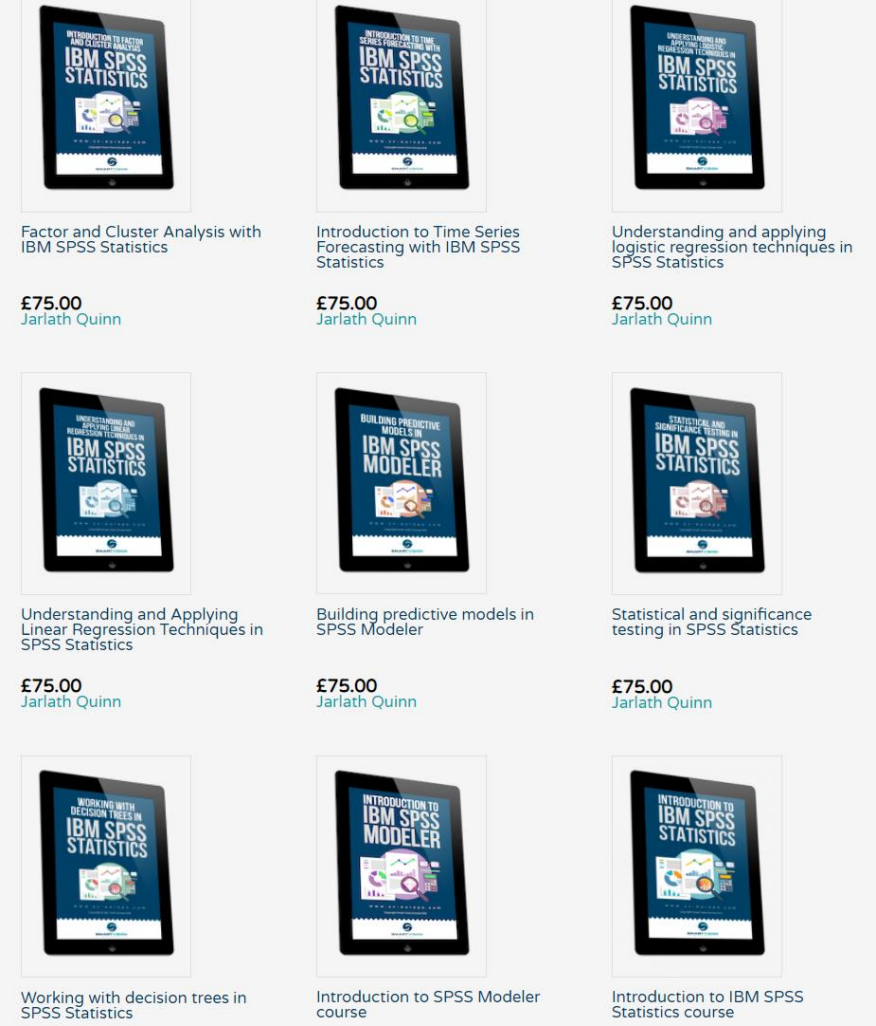
Data Science Recruitment

Help with staff recruitment utilising our technical expertise, extensive global network and decades of experience in the industry.

Smart Vision Europe: Services and Expertise

- **Software**
  - We can help you manage your existing SPSS licenses
  - You can buy your analytical software from us often with discounts
  - http://www.sv-europe.com/buy-spss-online/
- **Advice and Support**
  - We offer 'no strings attached' technical and business advice relating to analytical activities to anyone
  - Formal technical support services for the IBM SPSS product family
- **Access our online training catalogue**
  - Smart Vision Europe customers gain free access to our library of training materials and recorded self-paced training courses

Online training materials free to Smart Vision customers or available for purchase



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