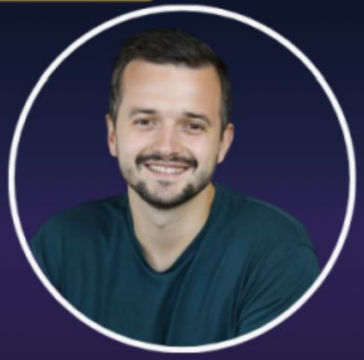
UKAP TRAİNİNG WEBİNAR

— UNLOCKİNG THE POWER OF DATA & ANALYTİCS —

9 July 2024
18:30—20:00 (GMT)

ALESIO DERVISHI
UCL Student
MODERATOR

SPEAKERS



LORIJENT LAMCE
Electronic Trading Compliance @ **Quadrature**

JONIDA BAJRAMI
Senior Content Manager, Data Solutions @ **FactSet**

DANIEL SAKO
Alternative Data Analyst @ **Balyasny Asset Management**

Q&A

- Any questions you might have for our panellists please drop them in the chat box below.
- We will try our best to go through them in this Q&A section.

What is Data Analytics?

Definition and Scope: Data analytics involves the systematic analysis of data to uncover, interpret, and communicate meaningful patterns. Originating from statistical analysis, it has expanded with advanced computational techniques and big data.

Real-World Applications:

- **Marketing**: Market analysis, customer behavior prediction.
- **Healthcare**: Enhancing patient care, predicting disease outbreaks.
- **Engineering**: Predicting equipment failures and maintenance needs
- **Finance**: Stock predictions, fraud detection, risk management.

Impact: Transforms raw data into actionable insights, enabling companies (and people) to make informed, strategic decisions that enhance efficiency, drive innovation, and provide a competitive edge.

When did Data Analytics become a thing?

… data analytics techniques and humankind go hand in hand …

Ancient Egyptian Grain Storage: The ancient Egyptians collected data on annual Nile floods and grain harvests to predict future crop yields and manage food storage and distribution. This early form of data analytics helped ensure food security and manage resources efficiently.

John Graunt's 17th Century Analysis: In 1662, John Graunt's systematic analysis of London mortality statistics in "Natural and Political Observations Made upon the Bills of Mortality" marked the first verified use of data analytics, laying the groundwork for demography and highlighting its potential in understanding population dynamics and public health.

John Snow's 1854 Cholera Investigation: By mapping and analyzing cholera cases in London, John Snow identified a contaminated water pump as the outbreak's source, demonstrating the power of data analytics in public health and laying the foundation for modern epidemiology.

Discovery of Linear Regression: Sir Francis Galton, in the late 19th century, pioneered the method of linear regression to study relationships such as parent-offspring height correlations, marking the foundational application of regression analysis in statistics and data science.

Why is it now a big deal?

Massive Data Generation: Exponential growth in data across sectors like social media and industry provides rich sources for insights and trends. From your phone to your travel card, almost everything now generates data!

Increased Computing Power: Advances in computing technology allow rapid and efficient analysis of large datasets, facilitating real-time decision-making. The Apollo Guidance Computer (AGC) had 2 kilobytes (KB) of RAM and 36 KB of read-only memory (ROM), which was the pinnacle of technology in its day. This would not be sufficient to hold a modern day gif.

Democratization of Computing Resources: Accessible cloud platforms and powerful tools empower companies (and people) to conduct sophisticated data analytics without extensive infrastructure. My small laptop can now connect to a supercomputer and crunch terabytes of data with ease (TARDIS).

Strategic Importance: Data analytics plays a pivotal role in deriving actionable insights that drive innovation, enhance competitiveness, and inform strategic decision-making in today's data-driven world. From assessing the odds of the next card being higher at a house party, to predicting the next market crash and making you millions, it can be used in all aspects of life.

What's the catch?

Data Quality: Data quality encompasses accuracy, completeness, reliability, and relevance of data used in analysis, crucial for deriving reliable insights and making informed decisions.

Garbage In, Garbage Out Principle: Ensuring accurate and error-free data inputs is essential as flawed data leads to flawed outputs, underscoring the importance of data quality in maintaining analytical integrity. We do not yet have (true) AI, and data analysis cannot verify whether the data you provide is good quality.

Role of Domain Knowledge: Domain expertise is critical in data analysis to identify anomalies or inconsistencies that may indicate data errors, enabling informed interpretation and enhancing the reliability of analytical outcomes. Getting the answer is one part, explaining the answer in the context of your business is also important, and that is where domain knowledge comes in.

What makes a good dataset?

A 'good' dataset typically has the following characteristics:

1. Relevance: The data should be relevant to the research questions or problems being addressed.
2. Completeness: The dataset should contain all necessary data points and variables needed for the analysis.
3. Accuracy: The data should be free from errors and accurately represent the real-world phenomena it is intended to measure.
4. Consistency: Data should be collected and recorded in a consistent manner, ensuring comparability across different observations.
5. Timeliness: The data should be up-to-date and collected within an appropriate time frame for the analysis.
6. Accessibility: The dataset should be easily accessible and available in a usable format.
7. Well-documented: The dataset should come with comprehensive metadata and documentation describing how the data was collected, processed, and any limitations or biases.
8. Integrity: The data should be collected and handled ethically, respecting privacy and confidentiality where applicable.

Overall, a 'good' dataset is one that is reliable, trustworthy, and suitable for answering the questions or solving the problems it is intended for.

Best Practices

Relevance

- Objective Alignment
- Context Appropriateness
- Timeliness
- Granularity
- Scope and Coverage
- Source Credibility
- Consistency

Completeness

- Comprehensive Data Collection
- Handling Missing Values
- Robust Data Infrastructure
- Data Validation & Monitoring
- Quality Assurance
- Surveys & Data Collection Methodologies
- Post-Collection Analysis
- Documentation

Accuracy

- Precision in Data Collection
- Data Validation and Verification
- Standardization of Procedures
- Use of Accurate and Reliable Sources
- Error Identification and Correction
- Routine Updates and Maintenance
- Training and Awareness
- Documentation and Transparency.

Consistency

- Standardization of Data Formats and Values
- Use of Controlled Vocabularies and Enumerations
- Data Integration Protocol
- Database Constraints and Data Validation Rules
- Audit Trails and Change Management
- Consistency Checks and Periodic Reviews
- Education and Training.

Best Practices - Part 2

Timeliness:

- Define Timing Requirements
- Streamline Data Collection Processes
- Efficient Data Processing Pipelines
- Prioritization and Scheduling
- Monitoring and Alerts
- Data Storage and Access Optimization
- Communications Infrastructure
- Training and Organizational Processes
- Review and Continuous Improvement

Accessibility:

- Clearly Define Access Controls
- User-Friendly Interfaces and Tools
- Data Cataloguing and Metadata Management
- Documentation and Training
- Ensuring Data Quality
- Scalable and Reliable Infrastructure
- Compliance and Security Measures
- Mobile and Remote Accessibility
- Backup and Disaster Recovery
- Feedback Mechanism

Well-documented:

- Metadata Creation
- Data Dictionary Development
- Clear Versioning and Change
- Logs
- Usage Guidelines and Example
- Access Information
- Data Quality Reports
- Integration and Linkage Information
- Legal and Ethical Documentation
- Tools and Resources
- Feedback and Update Mechanisms

Integrity:

- Input Validation
- Data Quality Assurance
- Redundancy
- Backup Strategies
- Access Controls and Authorization
- Audit Trails and Logging
- Secure Data Transmission
- Database Constraints and Normalization
- Version Control
- Compliance and Regular Audits
- Education and Training

Incorporating data analytics into decision making - Part 1

- **Incorporating Data into Decision Making**:
  - **Data Collection and Analysis**: Gather relevant data through systematic collection methods.
  - **Data Interpretation**: Analyse data to uncover patterns, trends, and insights.
  - **Data Integration**: Integrate data-driven insights into decision-making processes to inform strategies and actions.

- **Dangers of Data-Driven Decision Making**:
  - **Bias and Misinterpretation**: Over-reliance on data without considering context or biases can lead to misinterpretation of findings. Overfitting is when you make the data tell you what you want to hear.
  - **Data Quality Issues**: Inaccurate or incomplete data can lead to flawed decisions and ineffective strategies.
  - **Lack of Human Judgment**: Ignoring qualitative factors or intuition in favor of purely data-driven insights may overlook critical aspects of decision making. This could be crucial in areas such as healthcare or policing.

Balancing data-driven insights with human judgment and considering the limitations and potential biases of data are crucial to effective decision making in any company.

Incorporating data analytics into decision making - Part 2

Incorporating Data into Decision Making

- **Building Trust & Value Perception:**
  - Stakeholders need to believe in the value of data analysis.
  - Data analysts should demonstrate how findings translate to concrete benefits (cost savings, increased revenue).
- **Defining a Clear Path to Monetization:**
  - Showing correlations between data and positive outcomes isn't enough.
  - Stakeholders need to see tangible benefits to fully embrace data analysis.

Dangers of Data-Driven Decision Making

- **Overpromising & Underdelivering:**
  - Overpromising the financial benefits of data analysis can lead to lost stakeholder faith if results fall short.
- **Misunderstanding Correlation vs Causation:**
  - Correlation doesn't equal causation. Basing decisions on correlations can lead to ineffective strategies.
- **Short-Termism & Neglecting Long-Term Value:**
  - Overemphasizing immediate monetization can neglect the long-term value of data (infrastructure, culture).
- **Ignoring Ethical Considerations:**
  - Focusing solely on financial gains can lead to overlooking ethical data collection and usage, damaging trust.

Tools and Techniques

Stage	Category	Description
Data Acquisition	Sources	Various data sources, APIs, web scraping tools, flat files
Data Acquisition	Tools	SQL clients, web scraping libraries, API connectors
Data Storage	Structured Data	Relational databases (SQL) (Postgre, MySQL)
Data Storage	Unstructured Data	NoSQL databases (MongoDb), data lakes (AWS S3)
Data Storage	Tools	Database management tools
Data Processing & Transformation	Languages	Python, R
Data Processing & Transformation	Libraries	Pandas, NumPy, scikit-learn, Matplotlib/Seaborn (Python), dplyr, tidyr, ggplot2 (R)
Data Analysis & Modeling	Tools	scikit-learn, TensorFlow, PyTorch (Python), caret, stats (R)
Data Visualization & Communication	Tools	Dash, Plotly (Python), Shiny (R), Tableau, PowerBI, Looker
Orchestration & Scheduling	Workflow Management	Apache Airflow
Orchestration & Scheduling	Real-time Processing	Apache Kafka

Demo

A review of corporate actions in Excel, Python and tableau

Finally …

Thank you for joining us today in UKAP’s first education online webinar!