BUSA 603

Module 2: Marketing Analytics Tools & Data Platforms

Brian Weikel

brian.weikel@franklin.edu

Business Analytics
Franklin University

Spring 2024

Outline[1]

1. Spreadsheet Tools
2. Programming Tools
3. Variables and Variable Types
4. First, Second and Third Party Data
5. Data Management Platforms
6. Marketing Measurement and Optimization Solutions
7. Data Collection Laws
8. Appendix

[1] These lecture notes map to chapters 2 and 5 of Davis (2022).

Spreadsheet Tools

Excel

A **spreadsheet tool** is an interactive software application for structuring, transforming, analyzing and storing data in rows and columns. You will hear people say such data is in a *tabular format*.

Tabular format, or **tabular form**, is simply information presented in the form of a table with rows and columns.

With the proper permissions and syntax, Excel has the capabilities to read data from a host of sources or platforms.

Via the Excel ribbon, choose `Draw` and then `Get Data` to review potential sources and platforms.

Excel Data → Get Data → From File Options

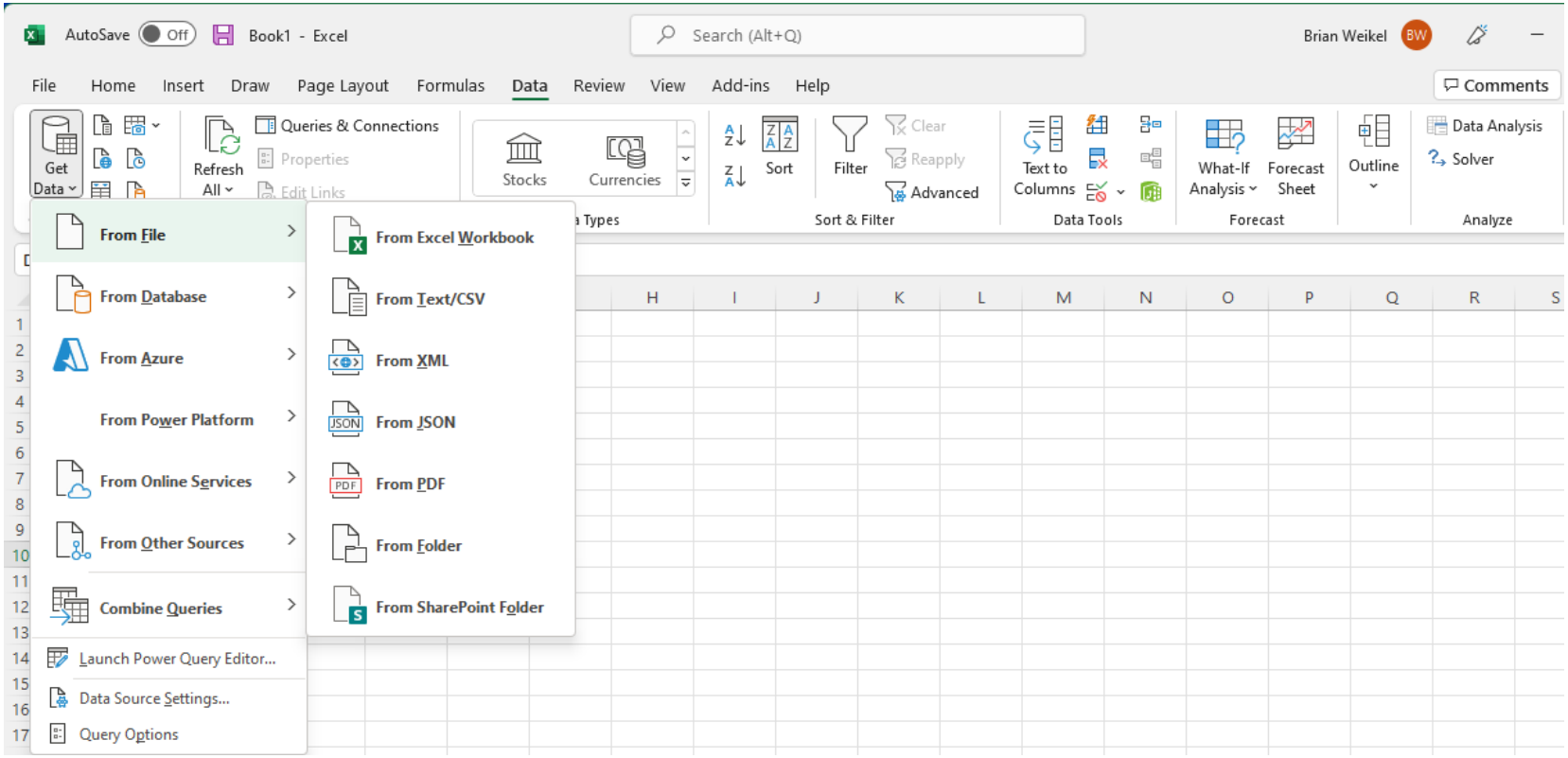

Figure 1: Excel `From File` Options

Options frequently used to read data into Excel include `Excel Workbooks`, `Text/CSV` files, and `JSON`.

Excel Data → Get Data → From Database Options

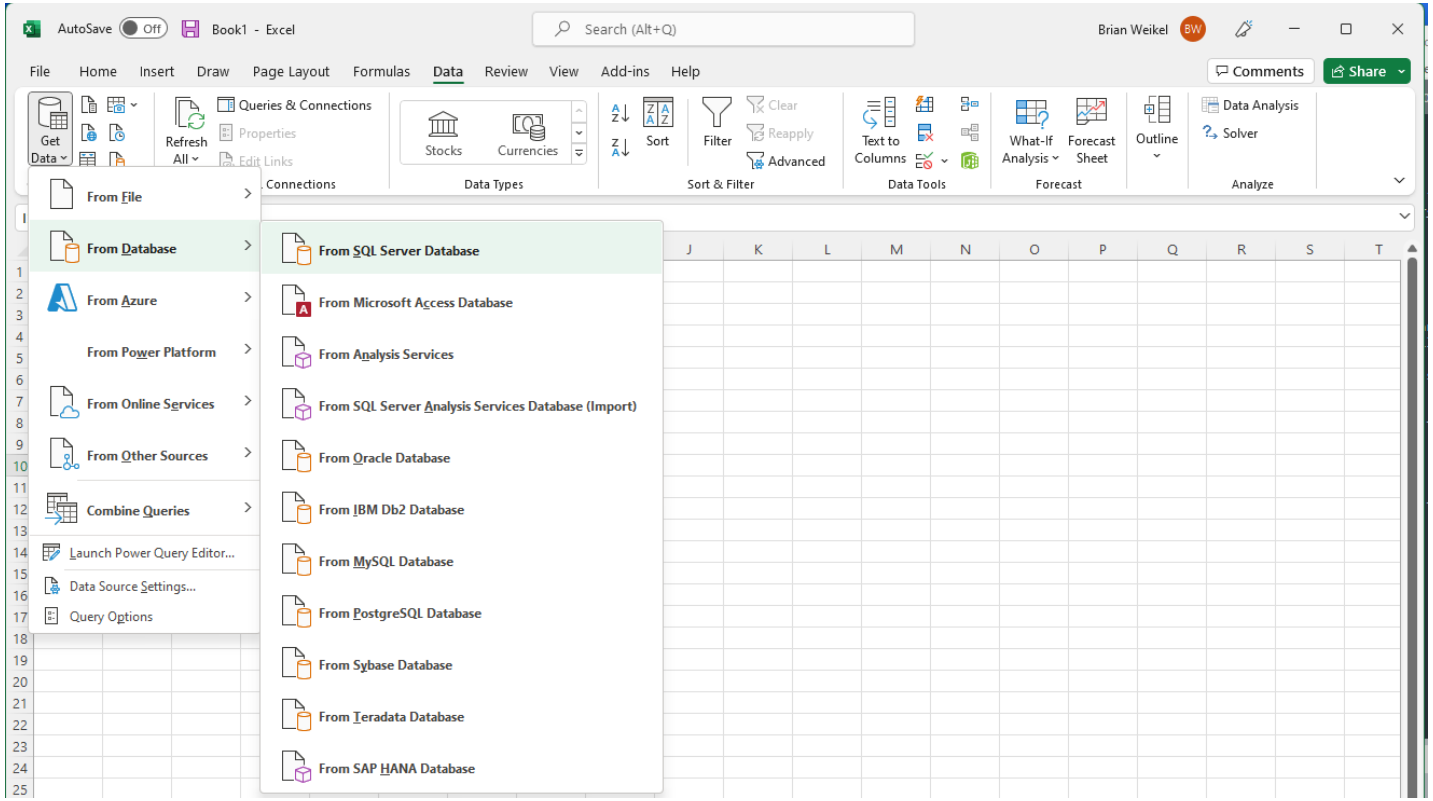

Figure 2: Excel From Database Options

Options include (unsurprisingly) SQL Server Database, (unsurprisingly) Microsoft Access Database, Oracle Database, IBM Db2 Database, Teradata Database, and SAP Hana Database.

Excel Data → Get Data → From Azure Options

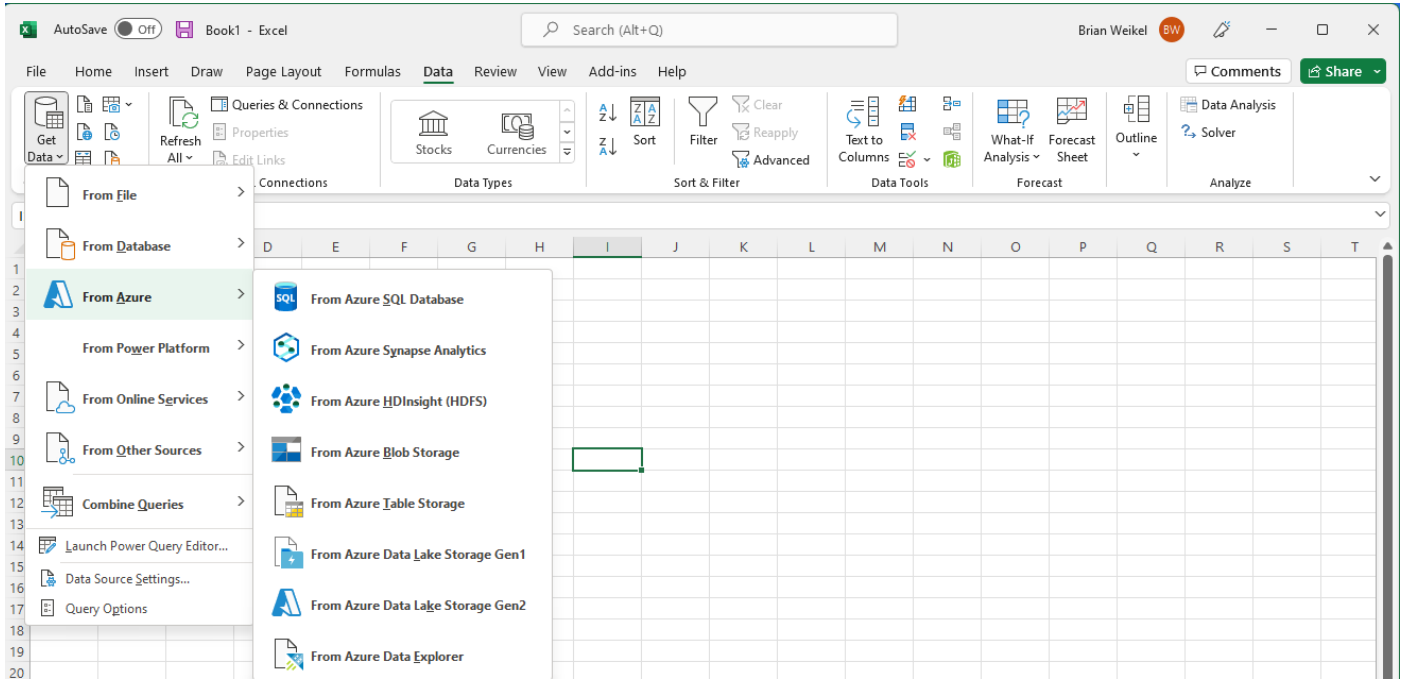

Figure 3: Excel From Azure Options

Options include Azure SQL Database, Azure Blob Storage, and several Azure Data Lake Storage generations. Though not ideal from an automated, scalable, and integrated pipeline process perspective, one can use Microsoft Query to import Amazon S3 data into Excel. See this page for an Open Database Connectivity (OCBC) approach.

Data Transformation

Spreadsheets, along with several programming language reviewed later in these notes, can be used to:[2]

- Impute missing data.
- Identify extreme outliers.
- Resolve erroneous value formats.
- Recognize impossible values.
- Map text to numbers.
- Conduct basic content analysis.
- Complete data integration.
- Remediate inconsistent variable values.
- Resolve duplicate records when appropriate.

[2] Additional information on appropriate transformations of data indifferent to the tool used may be found in Burke et al. (2021a).

Issues Resulting from Missing Data

Missing values means that some values exist and others do not exist for the *same row* of data.

Missing data may seriously compromise inference from randomized experimental designs and observational data, especially if the missing data are not handled appropriately.

The potential bias due to missingness depends on the mechanism causing the data to be missing, and the methods used to ameliorate the consequences of missing data.

Missing data also adversely affects supervised and unsupervised learners, primarily by reducing prediction and classification accuracy and/or precision.[3]

[3] For a review of metrics for multi-class classification, see Grandini et al. (2020).

Three Basic Options to Handle Missing Data

While one should attempt to identify the missingness mechanism since it will determine the preferred mitigation method, several rather simple options are frequently used in business.[4]

1. The simplest option is to nullify the cells of missing values. This would mean blanking out an Excel cell if the value is (say) "N/A", or simply keeping it blank if it is already blank.
2. The second option is to delete the entire record if there are missing values for the record.
3. The third option is **interpolation**, which is replacing the missing value with an estimate based on other records. One method is **hot-deck imputation**, where each missing value is replaced with an observed response from a "similar" record.[5]

[4] For an introduction to missing data mechanisms see Jakobsen et al. (2017).

[5] For additional information on hot-deck imputation see Andridge and Little (2010). The section *Missing Data* of Burke et al. (2021b) briefly reviews more advanced techniques to handle missingness.

A Simple Method to Access Missingness

To assess the impact of deleting records with missing data one may conduct a simple analysis of a variable with missingness by examining a summary statistic of a variable that does not have missingness.

For example, one may calculate the average of the variable with missingness for those records with missing values and calculate the average of the variable with missingness for those records without missing values, and then compare the averages.

- If the averages do not differ, then deleting the records with missing data may be acceptable.
- If the average values differ a lot, one should collect new data or use advanced imputation techniques rather than deleting missing data.

A Simple Method to Access Missingness Continued

Example: If one is concerned about whether records with missing gender versus records without missing gender have different purchase patterns, one can simply compare the average purchase patterns for missing versus non-missing gender records.

- If the averages are significantly different, then the missing data are concerning enough that one consider collecting new data or using more advanced imputation methods.
- If the averages are not significantly, then many would feel comfortable deleting the records with missing data since missing versus non-missing observations do not seem to matter in terms of purchases.

Outliers

Outliers are those data points that differ noticeably from the other data points.

Quartiles may be used to identify outliers.

1. The **first quartile**, typically denoted Q_1, separates the smallest 25% of the data from the largest 75%.
2. The **second quartile**, typically denoted Q_2, separates the smallest 50% of the data from the largest 50%. Q_2 is the same as the **median**.
3. The **third quartile**, typically denoted Q_3, separates the smallest 75% of the data from the largest 25%.

Interquartile Range Method for Outlier Detection

The **interquartile range** (IQR) is frequently used to identify outliers. IQR is equal to subtracting the first quartile from the third quartile.

The **IQR method** for detecting outliers is:

- Compute the lower outlier boundary: $O_L = Q_1 - 1.5 \times \text{IQR}$.
- Compute the upper outlier boundary: $O_U = Q_3 + 1.5 \times \text{IQR}$.
- Any data value that is less than O_L or greater than O_U is considered to be an outlier.

While the IQR method typically works well in practice, other methods exist.[6]

[6] For a survey of methods, see Aguinis et al. (2013).

IMDb Data

IMDb, launched online in 1990 and now a subsidiary of Amazon since 1998, is the world's most popular source for movie, TV, and celebrity content. It was designed to help fans explore the world of movies and shows, and decide what to watch.

IMDb offers subsets of their data to customers for personal and non-commercial use. While you can hold local copies of the data, it is subject to their terms and conditions.[7]

This page identifies data location, and the details of each provided data set. Information provided for each data set includes variable names, variable types (e.g., string), and variable descriptions.

[7] Given we will only be using the data for personal class purposes (i.e., not for commercial use), we are in compliance with their Non-Commercial Licensing and copyright/license agreements.

Excel Boxplot with Outlier Detection

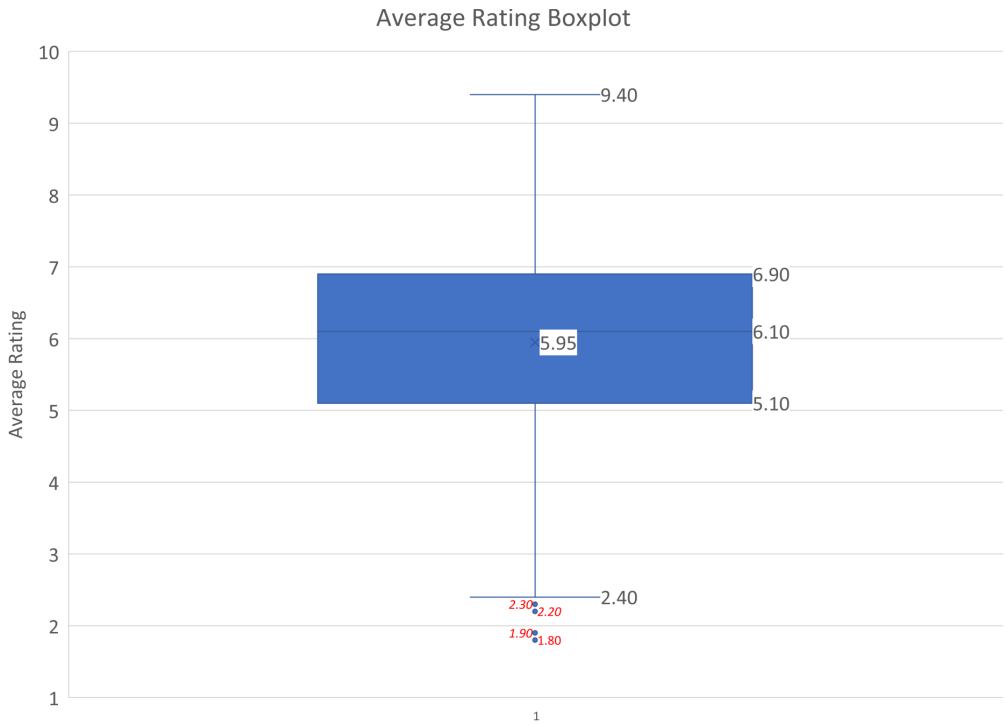

Figure 4: Average Rating Boxplot

In Figure 4, Q_1, Q_2, Q_3, the sample mean (i.e., 5.95), the maximum, and a local minimum are shown for IMDb Movie Average Ratings for movies released between 2017 and 2021. Any value less than 2.4 is considered an outlier by Excel; outliers are shown in *red font*. Dependent on the data, Excel may show a local maximum in lieu of the maximum.

Erroneous Value Formats and Impossible Values

Erroneous value formats are typically words that are meant to be numbers. For example the number of days since last purchase is recorded as "two hundred" in lieu of 200.

Impossible values are similar to outliers: they do not belong in the range of acceptable values. For example if the number of days since last purchase, relative to the current date, is 365,000 or -100 there is no need to use the IQR method to determine if the values are extreme; the data are obviously wrong.

Many times, data is not in a numerical format. For example the values for a column of data called Gender may be "Female", "Male", "Other", or "Unknown". Thus the variables would need to be mapped to categorical variables or indicator variables (e.g., dummy variables). As a case in point let the variable $X1$ be defined as 1 if Gender is Female, and 0 otherwise.

Content Analysis

Narrowly defined, **content analysis** is the examination of digital text, photos, audio, or visual formats of communication, which may or may not be extracted from social media. More generally, content analysis includes the careful examination of face-to-face human interactions, analysis of character portrayals in media venues ranging from novels to online videos, computer-driven analysis of word usage in news media and political speeches, advertising, and blogs, examination of interactive content such as video gaming and social media exchanges, and much more.[8]

Another way to deal with text is to count the number of times a word appears, a basic form of content analysis. This can be done as an analysis for the entire data spreadsheet; such as counting the unique instances of "Female", "FEMALE", "female", "Male", "MALE", "male", ... for the variable Gender.

[8] For an introduction to content analysis, see Krippendorff (2019).

Data Integration

Data transformation frequently includes integrating two sources of data. While we will soon cover more advanced ways of doing this task with programming languages, such as with Python or structured query language (SQL) for relational databases, Excel can be used to accomplish a few basic integration tasks.

- Excel offers a way to connect data in two spreadsheets with the VLOOKUP function.
- Another useful Excel data integration option is the MATCH function.

Programming Tools

Programming and Programming Language Defined

Programming is the process of solving a problem using computer algorithms. A **computer algorithm** is a well-defined procedure that allows a computer to solve a problem. **Statistical programming** is the process of solving *data-related* problems using executable computer algorithms.

A **programming language** is a formal set of instructions that can be used to produce various kinds of data output.

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for rapid application development, as well as for use as a scripting or glue language to connect existing components together.

R and SQL

R is a language and environment for statistical computing and graphics. It is a GNU project. R is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) in the 1970's by John Chambers and colleagues.

Structured Query Language, or simply **SQL** is a domain-specific language designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS).

While the most popular languages for marketing analytics and data science are Python and R, other languages one will encounter include Julia and Scala.

Defining Programming Tool, Code and Software

A **programming tool** is a software package that allows for the execution of a programming language. A **statistical programming tool** is a programming tool focused on statistical analyses of data.

Programming code are statements written in a particular programming language.

Programming software for data is a set of specialized computer programs for analysis using programming code.

Non-programming software for data is a set of specialized computer programs for analysis using a graphical interface.

Excel versus Programming Tools, like R and Python

FRANKLIN UNIVERSITY
Makes It Possible

Excel Advantages

- It is easier to get started and learning resources are rich.
- Has many built-in easy-to-use functions.
- Can assist in understanding many operations available in R and Python.

Excel Disadvantages

- Has limited advanced statistics (even with add-ins).
- To fully master Excel you need to learn VBA, which has more limited capabilities than R and Python.
- The Excel data file can only hold 1.08 million rows sans the aid of other tools, and its not suitable for processing large-scale data sets.
- Unlike R and Python, there is a fee for Excel.[9]

[9] Google Sheets does not charge a fee for tool use.

Comparing R and Python

1. Python is a general purpose language (GPL). R is a domain-specific language (DSL).
2. Python has more big-data supervised (e.g., deep learning), unsupervised learning, and scientific computation capabilities, and data handling modules: Numpy, SciPy, scikit-learn, SymPy, Pandas ....
3. R has more statistics and econometric packages available for immediate use.
4. While R and Python both have many graphical techniques, R graphics tend to be more elegant and visually appealing.

Comparing R and Python Continued

5. R has many packages, about 19,000, where some accomplish the same task. Python is designed on the philosophy that "there should be one and preferably only obvious way to do it."
6. Python is typically easier to learn than R. However, learning Python libraries can be a bit complex relative to learning R libraries and plots.
7. R is slower than Python, but not by much.
8. Python is more popular than R, particularly so in business software engineering (SE) departments.

Integrated Development Environments for Python

Popular integrated development environments (IDEs) used for Python include:

PyCharm provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems, and supports web development with Django. PyCharm is developed by the Czech company JetBrains.

The textbook mentions Rodeo, which is is an open-source IDE that is lightweight and customizable.

Caution is advised about using Rodeo as an IDE. The linked Github repository has not been updated in years. Thus the code base does not appear to be actively maintained.[10]

[10] Neither I nor anyone I know has ever used Rodeo.

Jupyter Products

The classic Jupyter Notebook is the original web application for creating and sharing computational documents. The open source web application offers a simple, streamlined, document-centric experience.

JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

The name "Jupyter" comes from the core supported programming languages that it supports: Julia, Python, and R.

Juypter Notebook

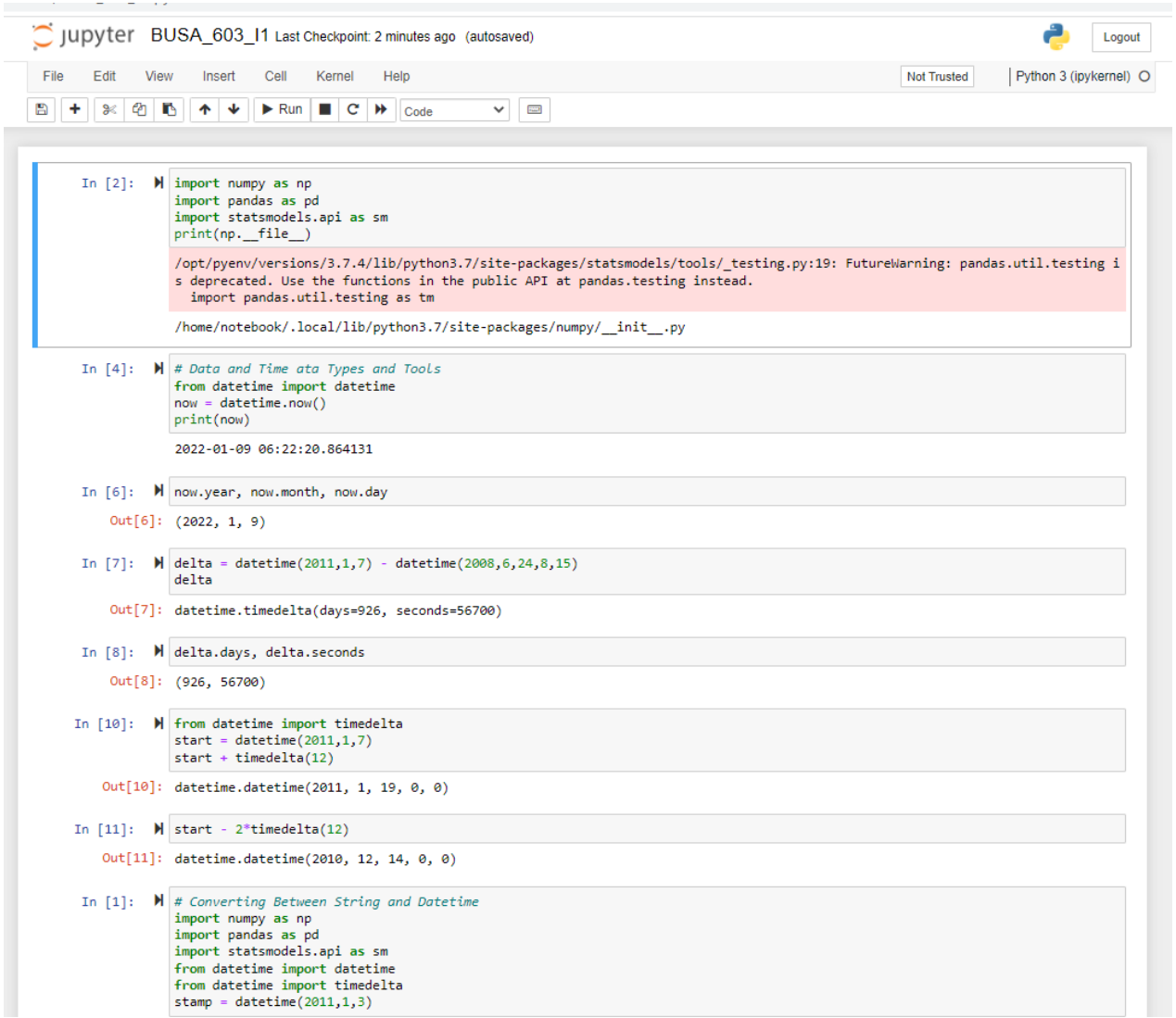

Figure 5: Juypter Notebook with Python Executable Code

An R IDE and the Sublime Text Editor

RStudio is an open source IDE for R. It is available in two formats: RStudio Desktop is a desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser.

Sublime Text is one of the most popular text editors in the world. It has many powerful features like multi-line editing, build systems for dozens of programming languages, regex find and replace, a Python API for developing plugins, and more. In addition, it is cross-platform (i.e., Mac, Windows, and Linux), and it is distributed as "shareware"; thus it is free to use with the occasional purchase pop-up.

RStudio

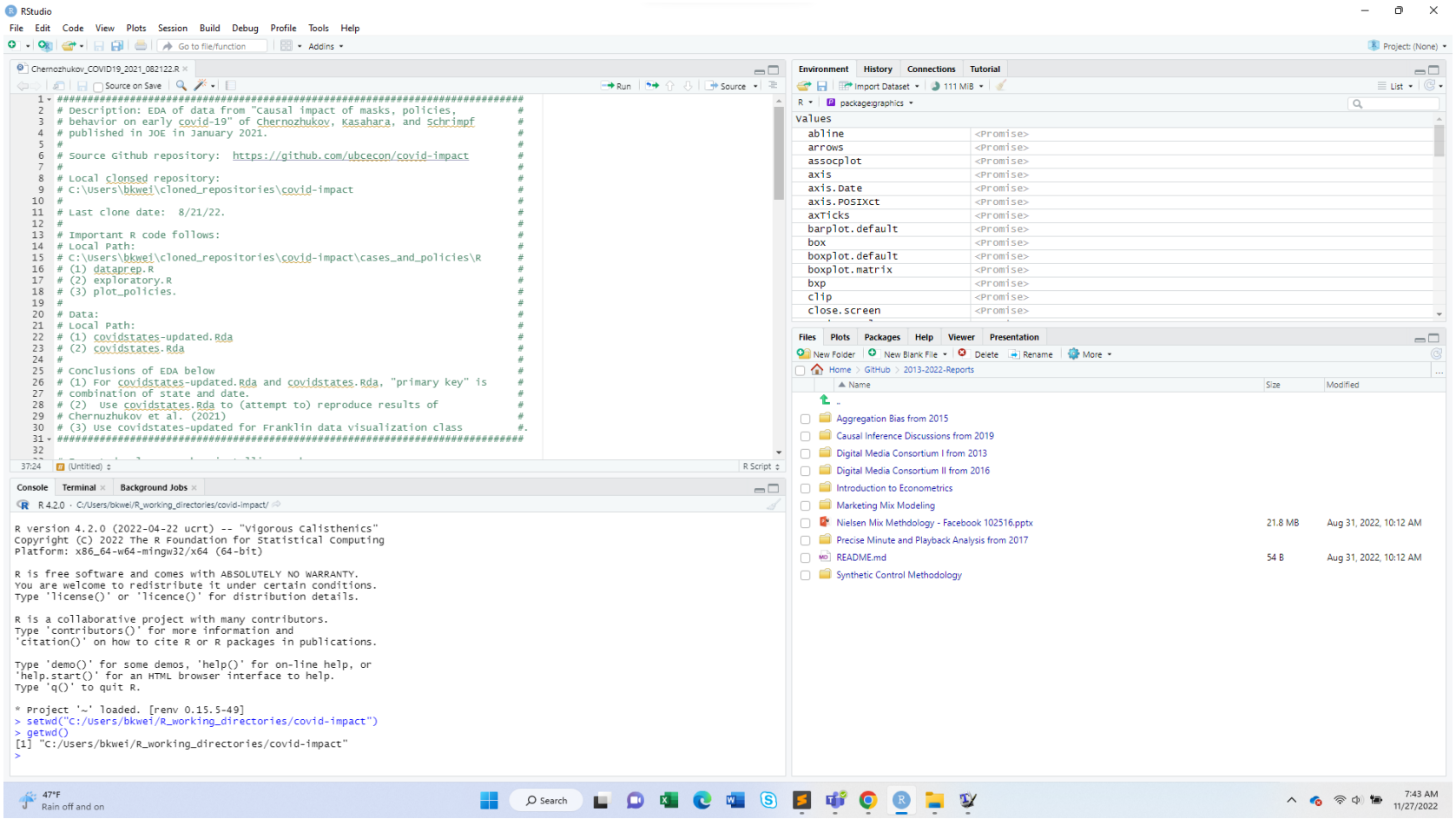

Figure 6: RStudio

Orchestration

Apache Airflow is an open-source platform for developing, scheduling, and monitoring batch-oriented workflows via a web interface, where its extensible Python framework enables one to build workflows connecting with virtually any technology.

Luigi is a Python (2.7, 3.6, 3.7 tested) package that helps one build pipelines of batch jobs. It handles dependency resolution, workflow management, visualization, handling failures, and command line integration.

Airflow tends to be the orchestrating workflow choice for many SE departments.[11]

[11] The Appendix provides additional information on the architecture of Airflow.

Repositories

GitHub is an online software development platform. It's used for storing, tracking, and collaborating on software projects. It provides the distributed version control of Git plus access control, bug tracking, software feature requests, task management, continuous integration, and wikis for every project.

GitLab is a DevOps software package that combines the ability to develop, secure, and operate software in a single application. Due to its ever expanding capabilities and support, it is typically favored by SE departments.

Variables and Variable Types

Variables and Variable Types

Excel, R, and Python use variables.

A **variable** is a storage mechanism for a particular identifier, which contains information referred to as a **value**.

A variable is like a column of data, and a variable's response values are like rows of data in an Excel spreadsheet, or a **data frame** in R.[12]

A **data set** is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer. A data frame and an Excel spreadsheet may both be considered a data set.

[12] A data frame is a generic R data object, which are used to store tabular data. One may use the Pandas package to create data frames in Python.

A Data Set Example

Name	HT	WT	Birth_Date	Birth_Place
Jet Greaves	6'0"	184	3/30/01	Cambridge, ON, CAN
Nolan Lalonde	6'1"	190	2/14/04	Kingston, ON, CAN
Elvis Merzlikins	6'3"	183	4/13/94	Riga, LVA
Daniil Tarasov	6'5"	196	3/27/99	Novokuznetsk, RUS

Table 1: Tender Employees

Some basic features are found in most structured data sets. Information is collected on individuals, households, persons, firms, animals, plants, subjects, items, objects, ...

In Table 1, the information collected is on employees. The variables of Table 1 `Name`, `HT`, `WT`, `Birth_Date`, and `Birth_Place`. The values of the variables are also called **data**.

You may hear some refer to the combination of attributes that result in unique identification as **records** or **observations**; The variable `Name` of Table 1 uniquely identifies an employee.

Common Data Types

- **Character**: the variable contains words that do not have order or numerical meaning. Also called a **string** or **text variable**. An example is `Name` of Table 1.
- **Numeric**: the variable contains numbers with decimal points. May also be called a **decimal**, **real**, or **float**. For example, `WT` of Table 1.
- **Integer**: the variable contains numbers without decimal points. For example, we could use `Birth_Date` of Table 1 to calculate an employee's age in years.
- **Logical**: the variable contains only two possible values, such as TRUE or FALSE. Also called a **boolean variable**, or an **indicator variable** or **dummy variable** when a variable is mapped to an integer. As an example, we may create a variable `NA_Birth` that equals 1 if Table 1's `Birth_Place` country for a `Name` is in North American, and 0 otherwise.

First, Second and Third Party Data

Classification of Data Sources

Marketing uses data from a variety of sources.

- **First party data** are a firm's own internal data, such as sales and customer information.
- **Second party data** are another firm's first party data, shared directly from the source.
- **Third party data**, also called syndicated data, are also data from other firms but without a direct relationship or agreement to the customer, such as data scraped from public websites or apps, or data purchased from a provider such as, IRI, Nielsen, NielsenIQ, NPD, or Spins.

The goal is to integrate first party data with second and third party data to optimize marketing outcomes.

Data Management Platforms

Inbound Marketing

Inbound marketing is a business methodology that attracts customers by creating valuable content and experiences tailored to them. It inspires long-term customer relationships.

An **inbound marketing tool** is marketing software designed to help marketers manage digital content and campaigns.

In contrast traditional marketing strategies, such as **outbound marketing**, attempt to push content to consumers to promote products and services.

Inbound Marketing and the Purchase Funnel

Inbound marketing attempts to make each element of the **purchase funnel** larger.

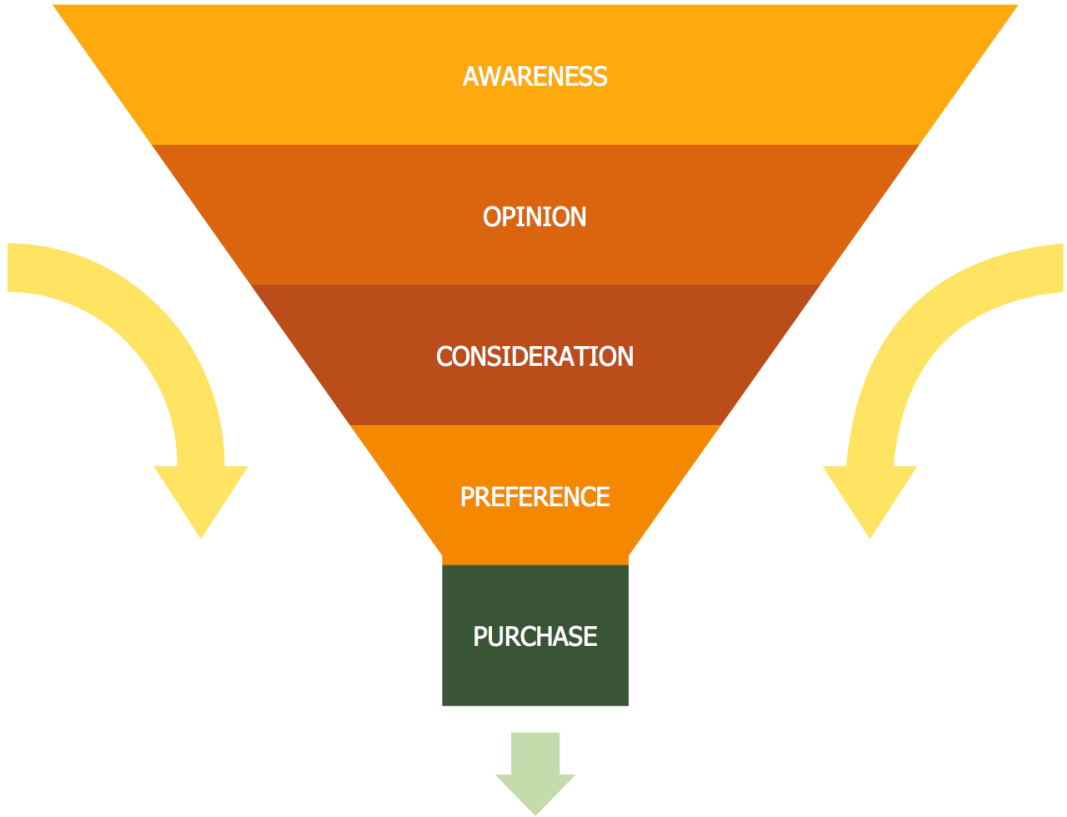

Figure 7: A Purchase Funnel

Inbound Marketing Media & Tools

Common type of inbound marketing media and/or content.

- Blog Posts
- Guides & E-books
- White Papers
- Infographics
- Search Engine Optimization (SEO)
- Inbound Email
- Webinars & Podcasts
- News Articles
- Direct Mailers
- Social Media
- Research Reports
- Slideshare
- Videos

Inbound marketing platforms manage digital marketing campaigns using just first party data. Inbound marketing tools include Ahrefs, Drift, Hubspot, Leadfeeder, Marketo, Salesforce, and SEMRush.

Data Management Platforms

Marketers need a system to store, organize, and analyze data before it becomes useful and easily understood. A **data management platform** (DMP) is such a solution. A DMP:

- Collects and organizes data so marketers may target specific audiences on **ad networks**.[13]
- Measures campaign performance across segments and channels to increase marketing outcomes, such as a consumer purchase.
- Integrates digital marketing data with non-digital data sources (i.e., second and third party data integration).

Notable DMPs include The ADEX, Adobe Cloud Experience, Lotame, Neustar, Nielsen Marketing Cloud, Oracle BlueKai, and Salesforce Audience Studio.

[13] Ad networks provide an outsourced sales capability for publishers and a means to aggregate inventory and audiences from numerous sources in a single buying opportunity for media buyers.

LUMA provides a set of online maps that categorizes different entities in the digital media and marketing ecosystems.

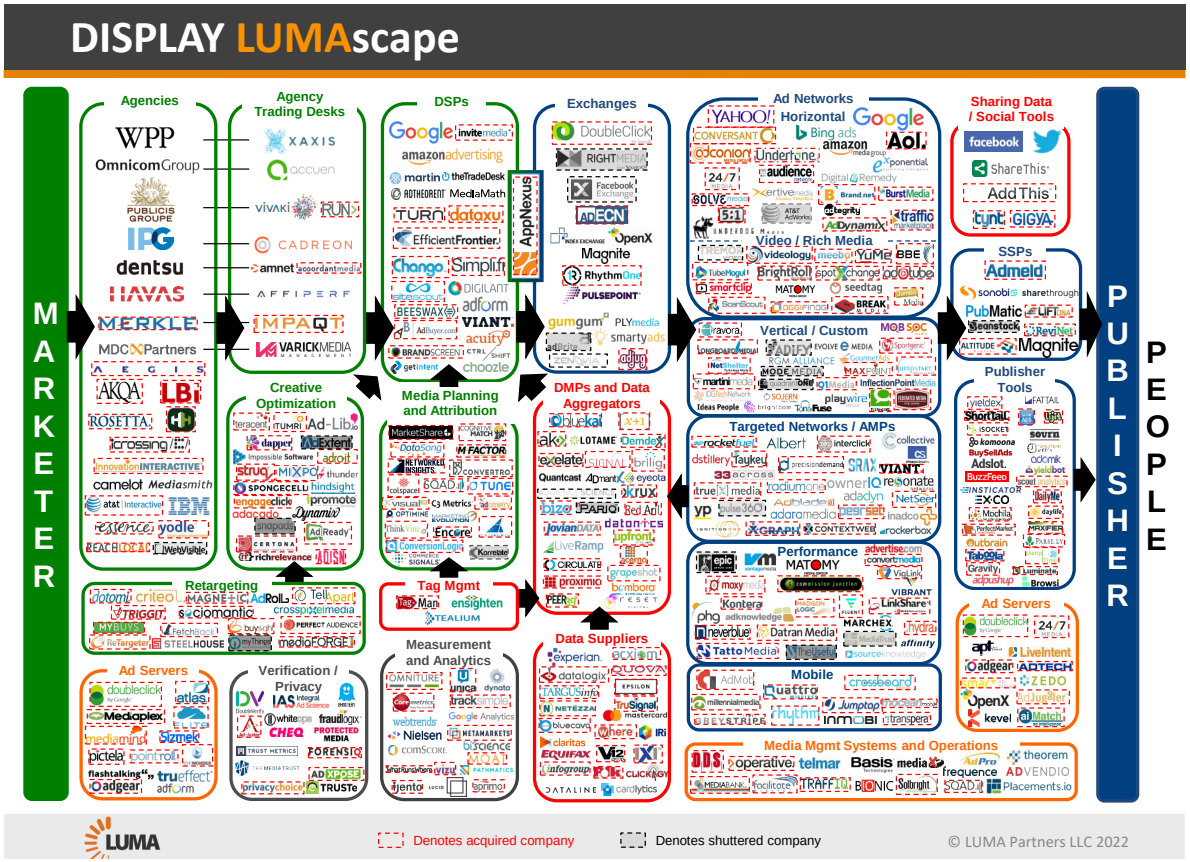

Figure 8: Display LUMAscape

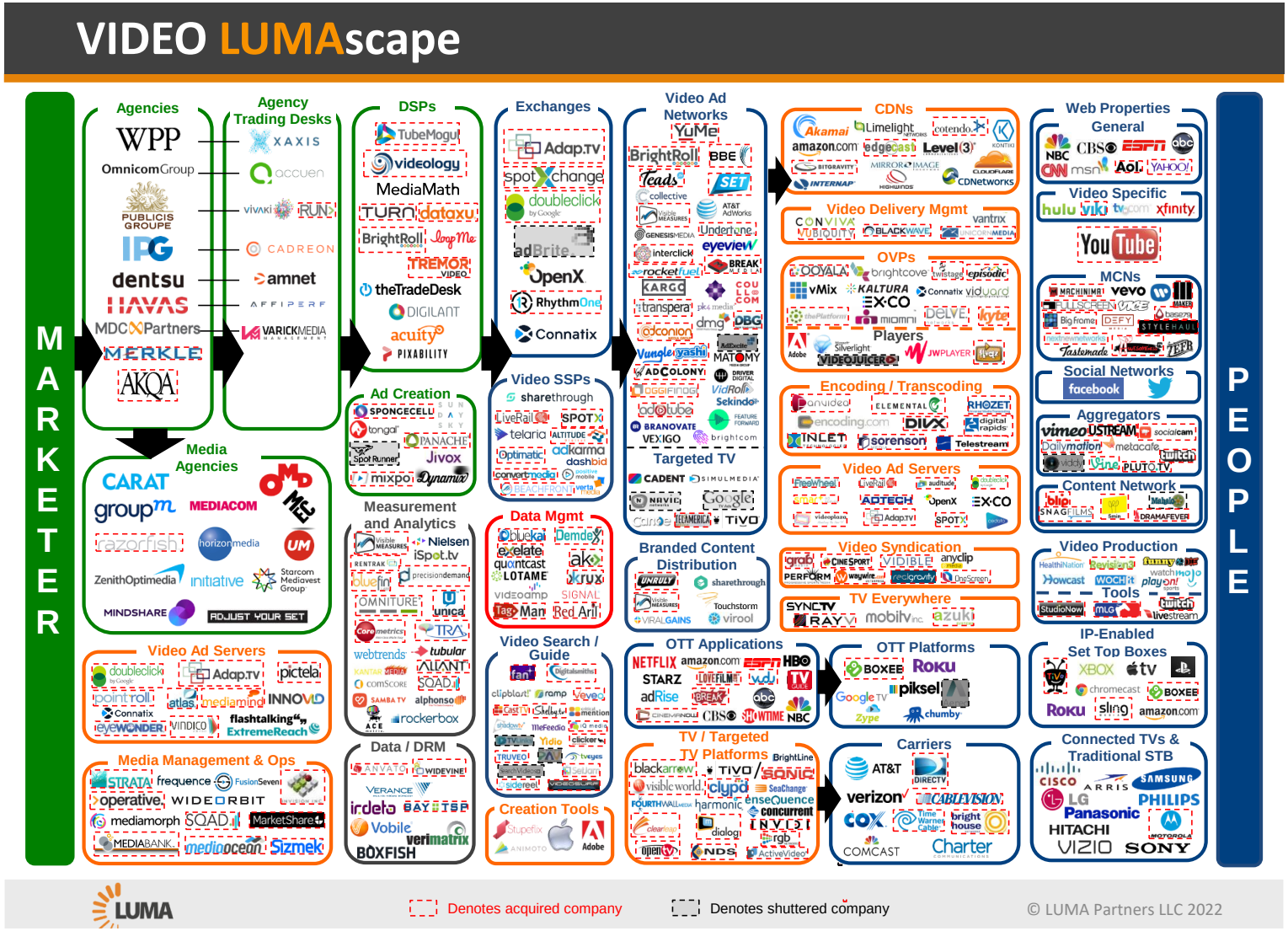

Figure 9: Video LUMAscape

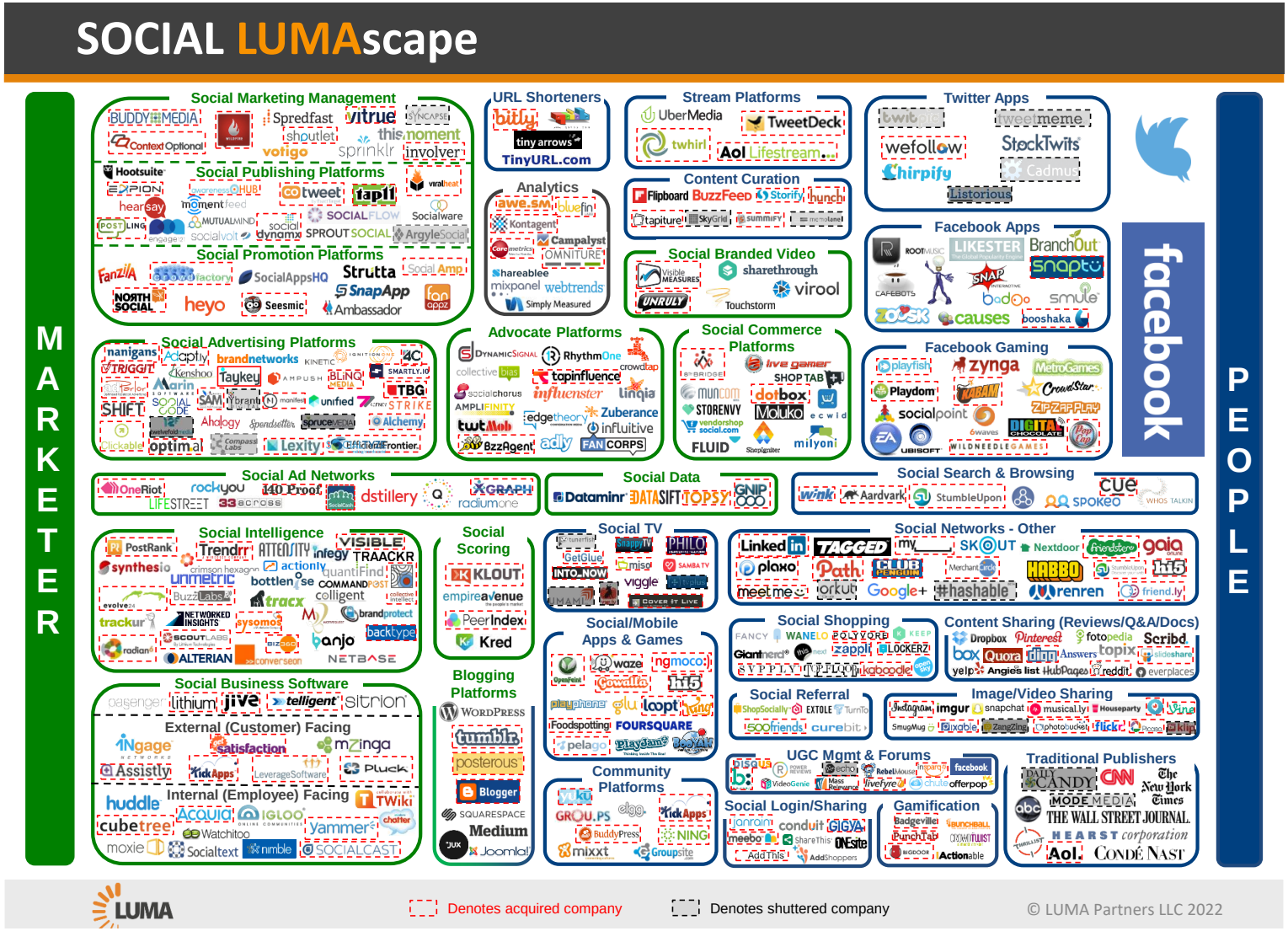

Figure 10: Social LUMAscape

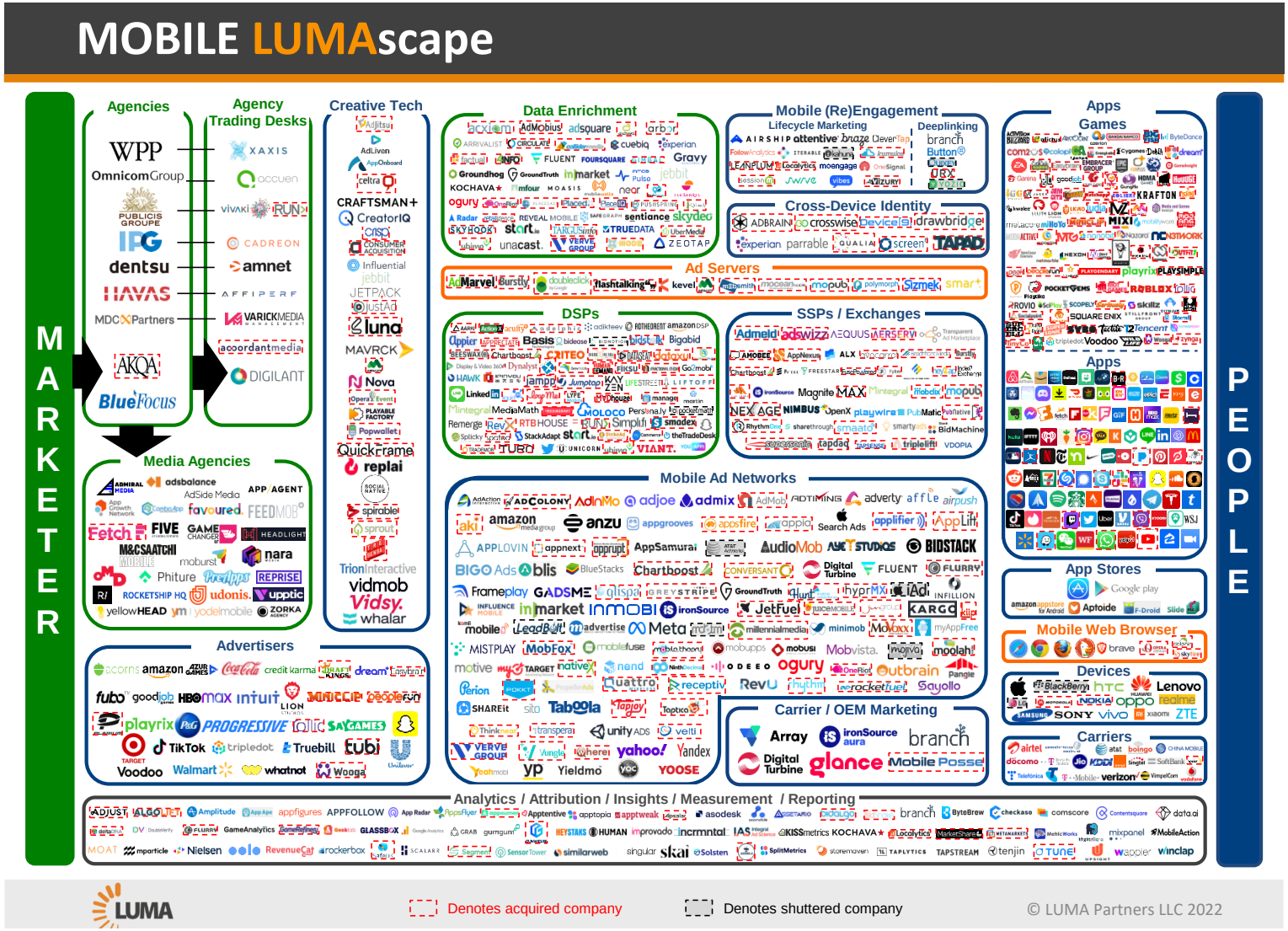

Figure 11: Mobile LUMAscape

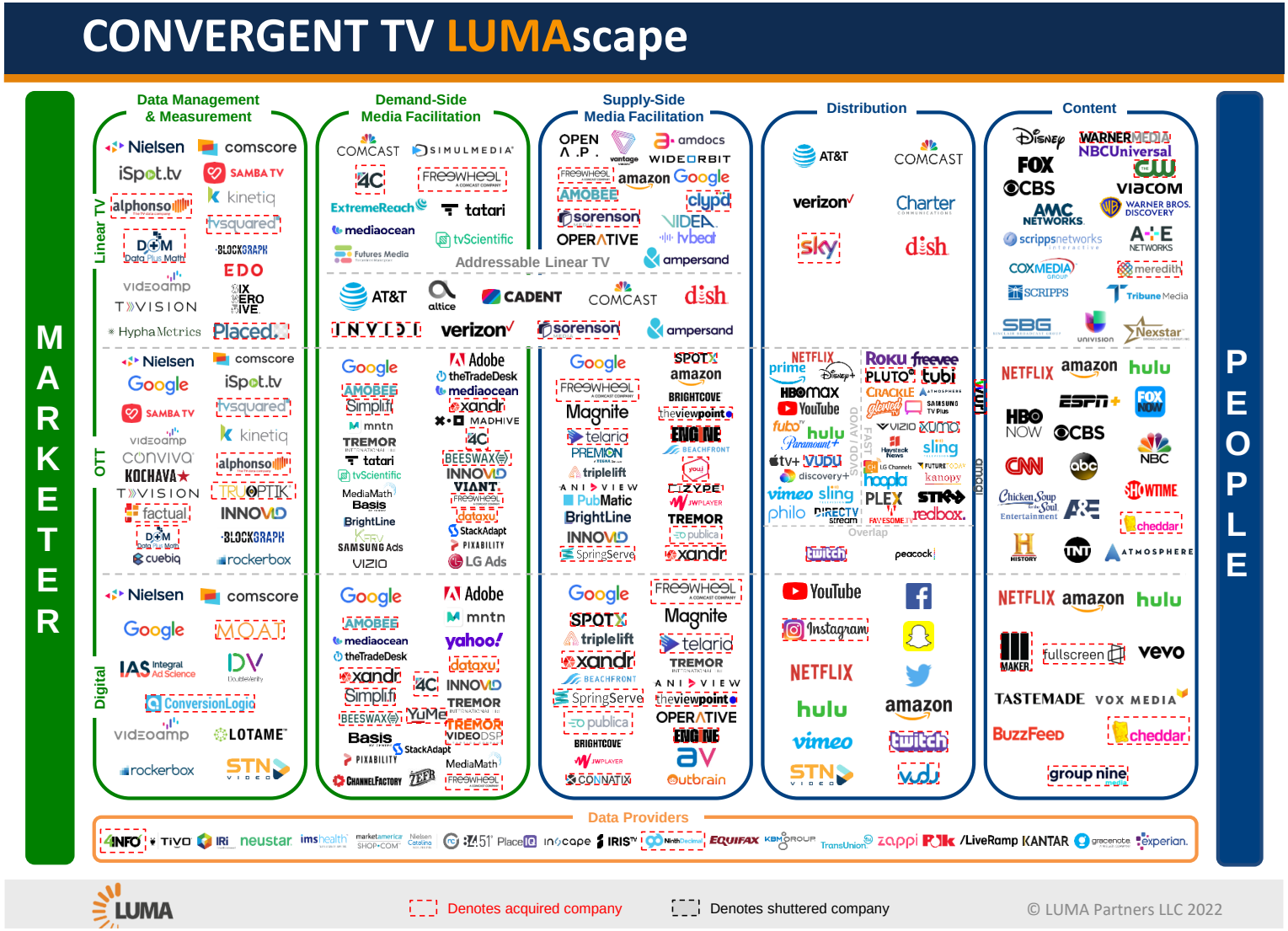

Figure 12: Convergent TV LUMAscape

Marketing Measurement and Optimization Solutions

Marketing Mix Models

The third type of marketing data platform is the **comprehensive marketing measurement and optimization solution**.

With both the ability to manage digital content like an inbound marketing tool and the ability to integrate digital and non-digital data like a DMP, comprehensive marketing measurement and optimization solutions also perform predictive analytics and recommend changes to a marketing plan in order to optimize ROI.

Marketing mix models use data to see what patterns in the present and past can best predict an optimal marketing mix for the future using advanced statistical techniques. We will review this methodology in detail later during the course.

The Forrester Wave

The Forrester Wave™ is a guide for buyers considering their purchasing options in a technology marketplace.

Results are based on the analysis and opinion of Forrester.

Forrester makes publicly available its methodology. They claim they apply this methodology consistently across all participating vendors.

Sometimes, vendors decide not to participate in the evaluation process. In these instances, Forrester evaluates the vendor according to The Forrester New Wave™ Vendor Participation Policy.

Gartner is a competitor to Forrester, and has like reports in select verticals.

Recent The Forrester Wave MM & OS Reports

The following pages present top-line Forrester Wave^{TM} results of their Marketing Measurement and Optimization Solutions (MM & OS) report.

- Q3 2023
- Q1 2022
- Q1 2020
- Q2 2018

Figure 1 of each report indicates substantial cross-time vendor performance variation. Some may believe that such short-term variation is odd.

Data Collection Laws

General Data Protection Regulation

The General Data Protection Regulation (GDPR) was the first standardized data collection law. It went into effect on May 25, 2018, in European Union (EU) countries.

- It protects a natural person's right to the protection of personal data.[14]
- It stipulates that the free movement of personal data within the EU shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data.
- For especially severe violations the fine framework can be up to 20 million euros, or in the case of an undertaking, up to 4% of their total global turnover (i.e., gross revenue) of the preceding fiscal year, whichever is larger.

[14] A natural person is a living, breathing human being.

California Consumer Privacy Act

The California Consumer Privacy Act of 2018 (CCPA) gives consumers in California more control over the personal information that businesses collect about them. The law secures new privacy rights for California consumers, including:

- The right to know about the personal information a business collects about them and how it is used and shared.
- The right to delete personal information collected from them (with some exceptions).
- The right to opt-out of the sale of their personal information.
- The right of non-discrimination for exercising their CCPA rights.

Business Ramifications

As mentioned in the Appendix of Module 1, companies like Google and Apple responded respectively by phasing out the use of third-party cookies in Chrome browsers and prompting iPhone users to choose whether they would like to be tracked by each of their apps, respectively.

These decisions, and forthcoming decisions, affect marketing analytics. Beyond the alternatives that were mentioned in the Appendix of Module 1, which are now used or could be used, the following have been used or may be used.

1. A **walled garden** is an environment that controls the user's access to network-based content and services.[15]

[15] For example, Google has Account IDs that can be linked to first-party user IDs. Google shares some of its first-party data, usually via a data privacy mechanism, with the firm who has linked its IDs with the Google Account IDs.

Cooked Data and Topics

2. **Cooked data** are raw data that has been processed, which may include data that is inferred or imputed.[16]
3. Google has put forward Topics, a new Privacy Sandbox proposal for interest-based advertising.
   - Google has a list of 300 topics (for now) that they use to categorize the websites people visit. When a person visits a new website, Google will categorize it into whatever topic it fits best.
   - Topics will only show advertising partners three of your interests. It pulls one interest from each week to share with advertisers.

[16] For example, if a data set has geographical information about a person such as their zip code, it is possible to infer additional data like income, age or purchase habits, information that is no longer as pervasive due to the consumer protection laws.

Economic Consequences of GDPR

Goldberg et al. (2024) study the economic consequences of GDPR for a large and diverse collection of online firms. They examine the effect of GDPR on site traffic, a measure of site health and its capacity to generate advertising revenue, and site revenue arising from e-commerce sales.

Using Adobe Analytics website performance data of 1,084 firms, they show that relative to the prior year recorded page views decrease by 11.7% and e-commerce revenue decreased 13.3% from EU users after GDPR implementaion.

- However, the data alone do not distinguish between the real and recording effects of the GDPR.
- They propose a model to separate GDPR's real effect on the volume of site visits and the GDPR's consent effect on the recording of site visit outcomes.

Economic Consequences of GDPR Continued

- Under the GDPR recorded economic outcomes may fall because some individuals do not consent to data sharing. Indeed, changes to when and what data are recorded is a primary goal of the GDPR.
- At the same time, a decline in recorded outcomes could reflect a decline in real economic outcomes, for instance because the regulation restricts personalized marketing.
- Privacy regulation thus creates an inference problem: data protection can both impact economic outcomes and obscure the observation of economic outcomes. Thus policymakers need to distinguish between the real and recording effects of privacy regulation in order to evaluate it.

Economic Consequences of GDPR Continued

- They conclude that consent accounts for at least 4.7% of the recorded page view estimate.
- They also provide conservative estimates for the contribution of GDPR's real effect on personalized marketing. The marketing effect alone represents 7.0% of the recorded page view estimate and 4.6% of the recorded revenue estimate.
- Despite concerns of consent fatigue, a substantial minority of EU users make the effort to register their nonconsent preferences. The authors provide evidence that smaller firms obtain lower consent rates, which suggests GDPR may have consequences for competition.

Appendix

Airflow Workflow

Airflow is a platform that lets one build and run workflows. A workflow is represented as a directed acyclic graph (DAG), and contains individual pieces of work called tasks, arranged with dependencies and data flows taken into account.

A DAG specifies the dependencies between asks, and the order in which to execute them and run retries; the tasks themselves describe what to do, be it fetching data, running analysis, triggering other systems, or more.

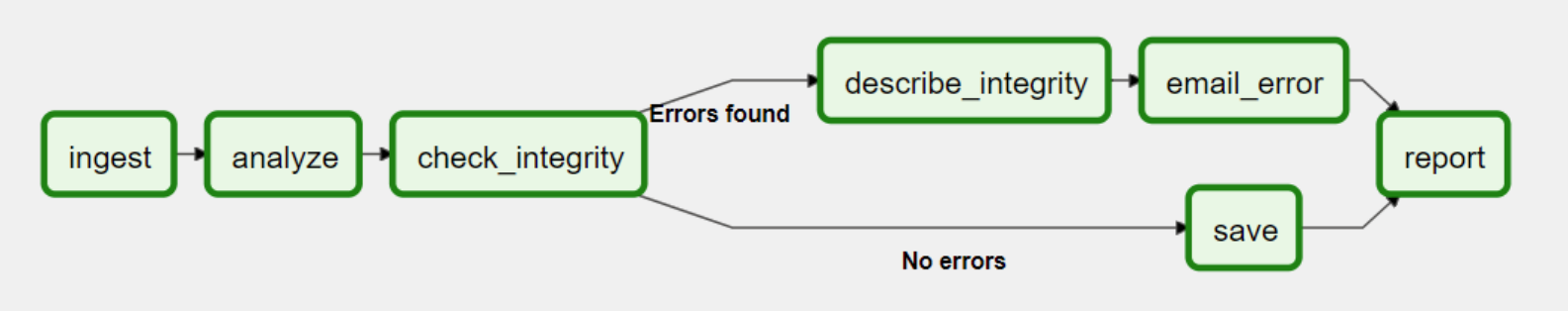

Figure 13: An Illustrative Airflow DAG

Introduction to the Airflow Architecture

An Airflow installation generally consists of the following components:

- A scheduler, which handles both triggering scheduled workflows, and submitting tasks to the executor to run.
- An executor, which handles running tasks. In the default Airflow installation, this runs everything inside the scheduler, but most production-suitable executors actually push task execution out to workers.
- A web server, which presents a handy user interface to inspect, trigger and debug the behavior of DAGs and tasks.
- A folder of DAG files, read by the scheduler and executor, and any workers the executor has.
- A metadata database, used by the scheduler, executor and web server to store the state of the process.

Airflow Architecture Visual

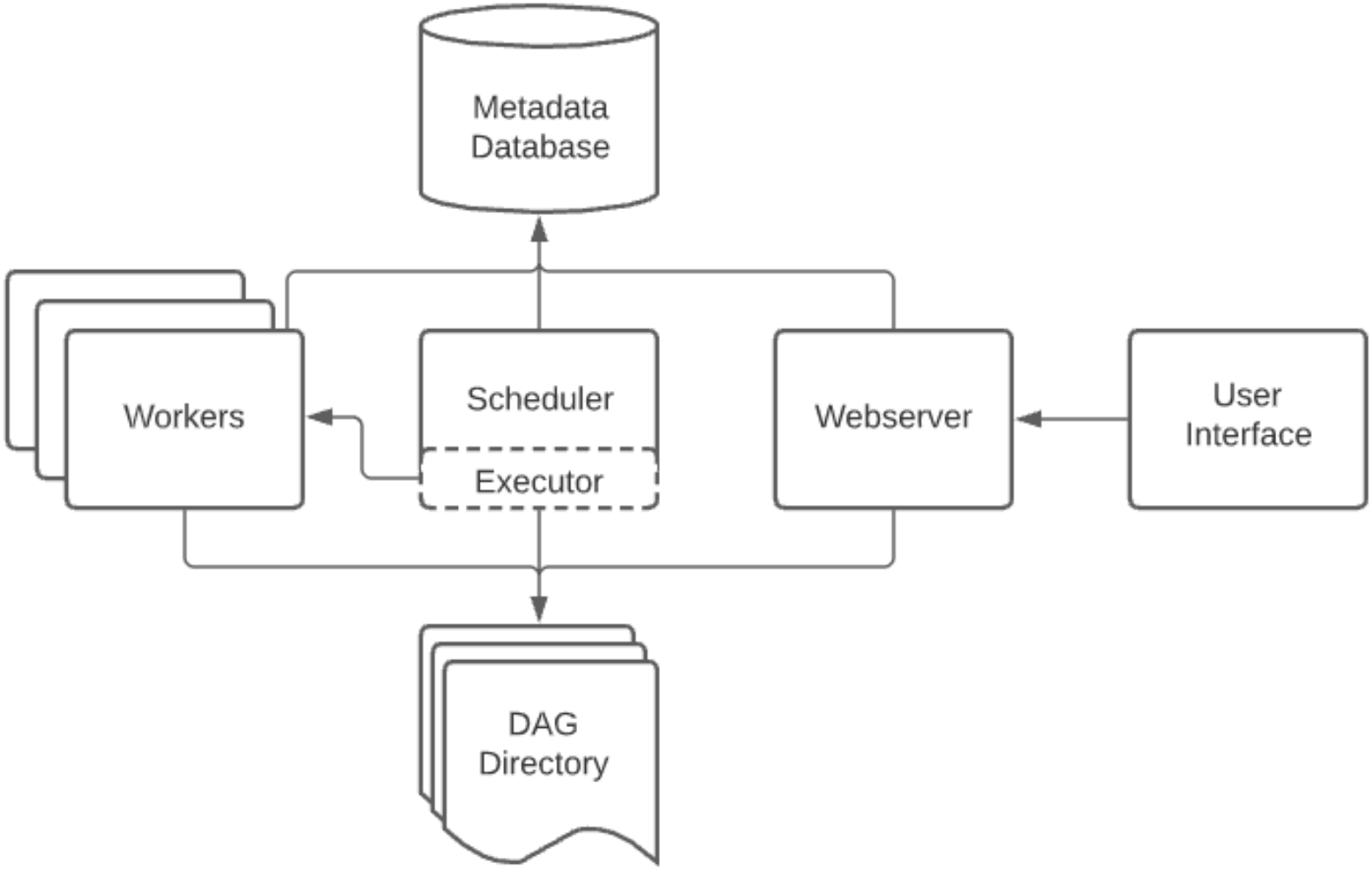

Figure 14: Airflow Architecture

References

Aguinis, H., Gottfredson, R. K., and Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2):270–301.

Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Burke, P., Hsu, K. Y., Karaoglan, L., Kodakkat, N., Lokhandwala, M., Lu, S., Mantha, S., Poortinga, V., Pourazarm, S., Saha, S. K., Wei, S., and Weikel, B. (2021a). Data structuring, governance, hygiene, and quality control: A review and new ideas. Product Leadership GRAD Technical Report, Nielsen.

Burke, P., Karaoglan, L., Lu, S., Mantha, S., Mut, M., Poortinga, V., Pourazarm, S., Saha, S. K., Shah, S., and Weikel, B. (2021b). Notes on causal inference challenges.

Davis, B. (2022). *Marketing Analytics*. Edify Publ., ISBN: 978-1-7346888-4-9.

Goldberg, S. G., Johnson, G. A., and Shriver, S. K. (2024). Regulating privacy online: An economic evaluation of the GDPR. *American Economic Journal: Economic Policy*, 16(1):325–358.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: An overview. https://arxiv.org/abs/2008.05756. [Online; accessed 26-Nov-2022].

Jakobsen, J. C., Gluud, C., Wetterslev, J., and Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – A practical guide with flowcharts. *BMC Medical Research Methodology*, 17(162):105–115.

Krippendorff, K. (2019). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc., Thousand Oaks, CA, 4th edition.