DATA SOCIETY

Data Science for Managers Participant Guide

Table of Contents

Class activities

Activity: data governance assessment

The Basic Maturity Assessment is a condensed version of the Stanford Maturity Measurement Tool. It focuses both on foundational and project aspects of data governance and subdivides each of its six maturity components into the specific aspects of people, policies, and capabilities.

Whether your organization uses the Stanford Maturity Measurement Tool, the Basic Maturity Assessment, or something different, it is imperative that the maturity model you choose is finalized and adopted early in the rollout of the data governance program. Thoughtful input from across the organization will help assure the model's usefulness and long-term fitness.

Please take the assessment now, giving a rating between 1(low) and 3 (high):

Foundational Components

The foundational components (Awareness, Formalization, and Metadata) focus on measuring core data governance competencies and the development of critical program resources.

Awareness: The extent to which individuals within the organization have knowledge of the roles, rules and technologies associated with the data governance program.

	Objective	*Rating*
People	Are executives, employees, and stakeholders aware of the purpose or value of the DG program?	0
Policies	Are existing data policies documented, consistently maintained, and available to stakeholders?	0
Capabilities	Are stakeholders aware of the specific DG capabilities that are available at the organization?	0

Formalization: The extent to which roles are structured in an organization and the activities of the employees are governed by rules and procedures.

	Objective	*Rating*
People	Have DG roles and responsibilities been defined and vetted with program sponsors?	0
Policies	Are data policies around the governance of specific data defined as best practices?	0
Capabilities	Are classes of DG capabilities defined and is there an available solution?	0

Metadata: Technical metadata describes data elements and other IT assets as well as their use, representation, context and interrelations. Business metadata answers who, what, where, when why and how for users of the data and other IT assets.

	Objective	*Rating*
People	Do executives, employees or stakeholders have understanding of types and values of metadata?	0
Policies	Are metadata best practices produced and made available?	0
Capabilities	Is metadata consistently collected, consolidated, and available from a single portal?	0

Project Components

The project components (Stewardship, Data Quality, and Master Data) measure how effectively data governance concepts are applied during funded projects.

Stewardship: The formalization of accountability for the definition, usage and quality standards of specific data assets within a defined organizational scope.		
	Objective	*Rating*
People	Have DG or stewardship roles and responsibilities been defined within the organization?	0
Policies	Have policies around data stewardship been defined within a functional area?	0
Capabilities	Does a centralized location exist for consolidation of and/or access to stewardship related documentation?	0

Data Quality: The continuous process for defining the parameters for specifying acceptable levels of data quality to meet business needs, and for ensuring that data quality meets these levels.		
	Objective	*Rating*
People	Are people assigned to assess and ensure data quality within the scope of each project?	0
Policies	Have data quality best practices been defined and adopted as official organizational data policies?	0
Capabilities	Have basic data profiling tools been made available for use anywhere in the system development lifecycle?	0

Master Data: Business-critical data that is highly shared across the organization. Master data are often codified data, data describing the structure of the organization or key data entities.		
	Objective	*Rating*
People	Is there consistent understanding among stakeholders of the concepts and benefits of master data?	0
Policies	Are there formal policies that define what data are considered institutional master data?	0
Capabilities	Are master data identified, managed, and provisioned?	0

The average of your scores for each section have been calculated below. Review these scores and then take some time to think about and record your goals for the future.

Foundational components

	Average Score	Goals
Awareness	0	
Formalization	0	
Metadata	0	

Project components

	Average Score	Goals
Stewardship	0	
Data Quality	0	
Master Data	0	

Activity: data ethics

An agency that oversees administration of benefits collects large amounts of applicant data on a daily basis. To streamline the application process, the agency engaged an outside party to create an automated tool that makes decisions on applicant eligibility for the benefits program. The tool relies on models that gather data from different parts of the organization, including applicant employment and financial records, and analyzes the chances of applicant success within the program. In operation for over two years, the tool helps eliminate numerous manual processes, identifies potential fraud, and better deploys limited resources. During this time, the tool has made thousands of decisions on applicant eligibility for benefits, impacting countless lives.

Judy, who works for the agency's outreach department, has received an increasing volume of complaints from applicants in recent months. Applicants have consistently stated they are inappropriately screened and removed from the application process. Judy brings this issue to her management, which requests an internal review. The internal review finds that data sharing agreements required for the original, underlying data to be destroyed upon the tool's deployment on the agency's customer-facing website. As a result, the agency is unable to reproduce, assess, or scrutinize aspects of the tool and its supporting decision models.

https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf

Questions

1. Was the agency justified in using applicant data to improve its own processes via automated decision models?

2. What ethical considerations should have arisen during the design, development, and deployment of the automated tool?

Activity: data-driven culture assessment

Choose the answers that best apply to you and your organization then scroll to the next page to see your results.

I can easily access the data I need without asking others for help. *0 - Not at all* *1 - Only for some colleagues* *2 - Only for some teams* *3 - Organization-wide*	0
I can easily access the data I need in a timely manner. *0 - Not at all* *1 - Only for some data* *2 - Only for data in my team / related teams* *3 - Organization-wide*	0
Data is automatically collected and stored on a continuous basis. *0 - Not at all* *1 - Only at someone's request* *2 - Regularly, a few times a year* *3 - There is continuous data collection*	0
The data we have is accurate and good quality (few missing entries, few duplicates, accurate measurements). *0 - Not at all* *1 - Only for some data* *2 - Only for data in my team / related teams* *3 - Organization-wide*	0
Our data is stored securely either internally or offsite. *0 - Not at all* *1 - Only for some data* *2 - Only for data in my team / related teams* *3 - Organization-wide*	0

My company routinely offers data trainings and other educational opportunities. *0 - Not at all* *1 - Occasionally* *2 - Regularly, a few times a year* *3 - There are continuous learning opportunities*	0
Most of my colleagues understand the importance of data. *0 - Not at all* *1 - Only for some colleagues* *2 - Only for some teams* *3 - Organization-wide*	0
Our organization has a set of data standards that reviews how data should be collected, stored, and analyzed. *0 - Not at all* *1 - Only for some colleagues* *2 - Only for some teams* *3 - Company-wide*	0
My organization emphasizes the importance of using data to track initiatives. *0 - No one* *1 - A few people across the company* *2 - Some teams across the company* *3 - Organization-wide*	0
I am expected to present data metrics when I explain conclusions and decisions. *0 - Not at all* *1 - Only for some colleagues* *2 - Only for some teams* *3 - Organization-wide*	0

Results

Your data infrastructure score is 0 out of 15. This represents how well your organization does with data access, collection, and storage.

Your data literacy score is 0 out of 15. This represents how well your organization does with data knowledge, governance, and leadership.

Your overall data culture rating is based on these two scores. Check out the matrix below to see where you belong.

	- Data infrastructure +	
- Data literacy +	**Data literate** Data infrastructure 0-7 Data literacy 8-15 □	**Data driven** Data infrastructure 8-15 Data literacy 8-15 □
	Data nascent Data infrastructure 0-7 Data literacy 0-7 □	**Data prepared** Data infrastructure 8-15 Data literacy 0-7 □

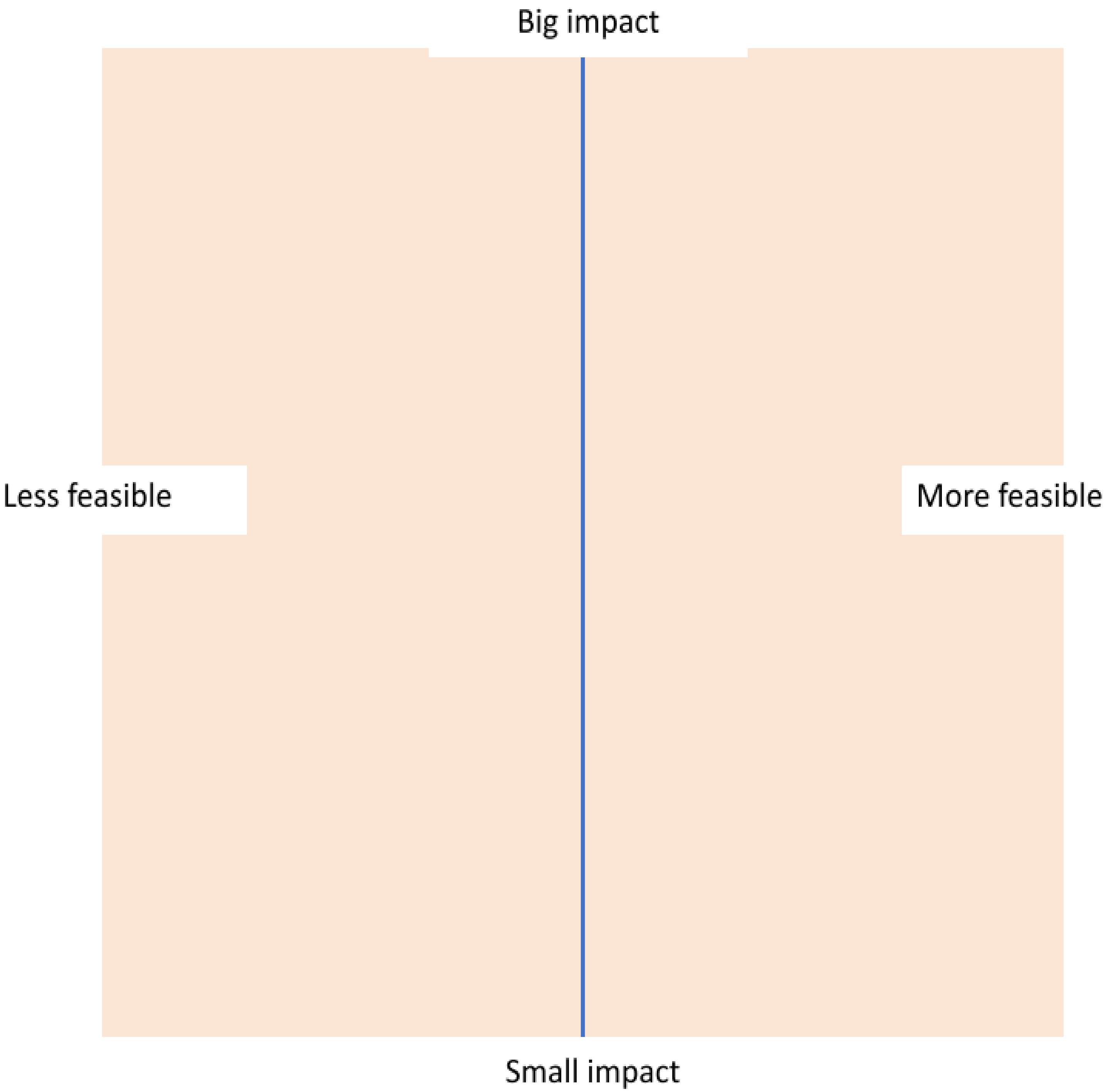
Activity: project brainstorm

Identify 3-5 ideas for leveraging data in your workplace and write them below.

1	
2	
3	

4	
5	

Then, on the next page, assess the feasibility and impact of your ideas and place them in the appropriate quadrant.


Big impact
Less feasible
More feasible
Small impact

Activity: field trip

Visit https://quickdraw.withgoogle.com/

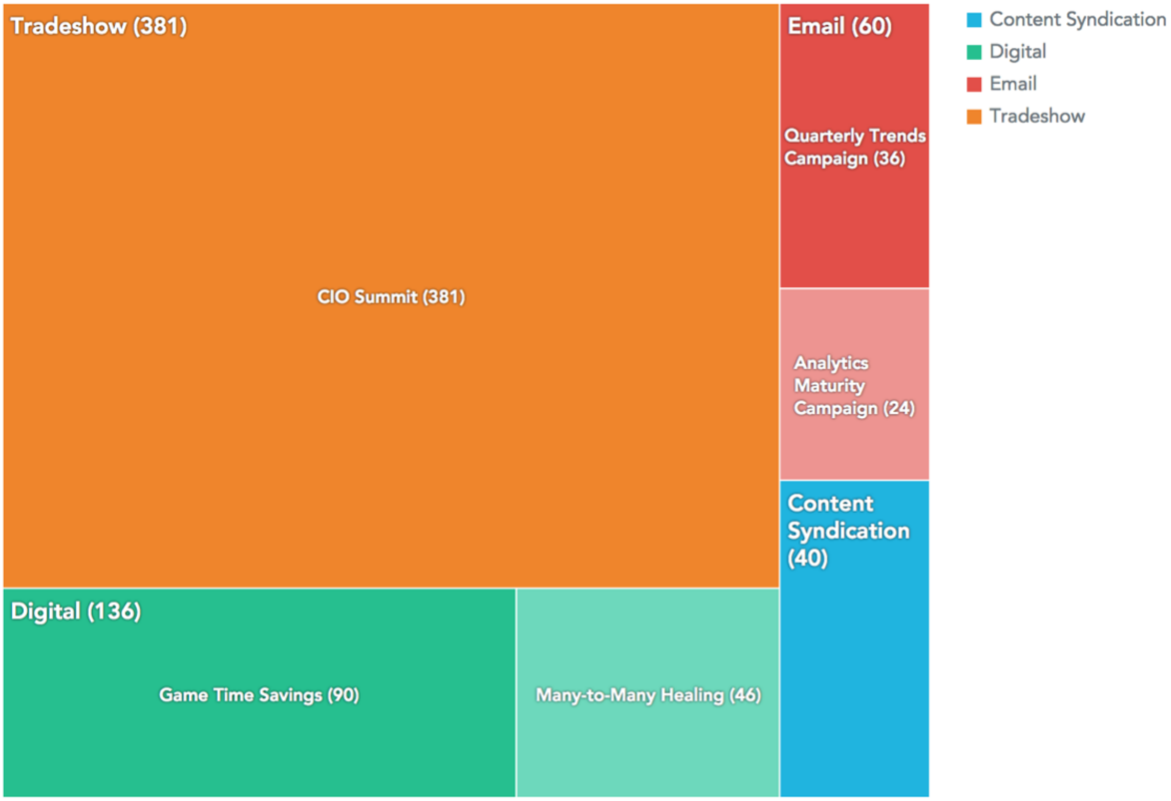
Click the "Let's Draw!" button and play a round (6 drawings).

At the end of the round, visit the data to see why guesses were made. Also, make a note of how many of your drawings were guessed correctly.

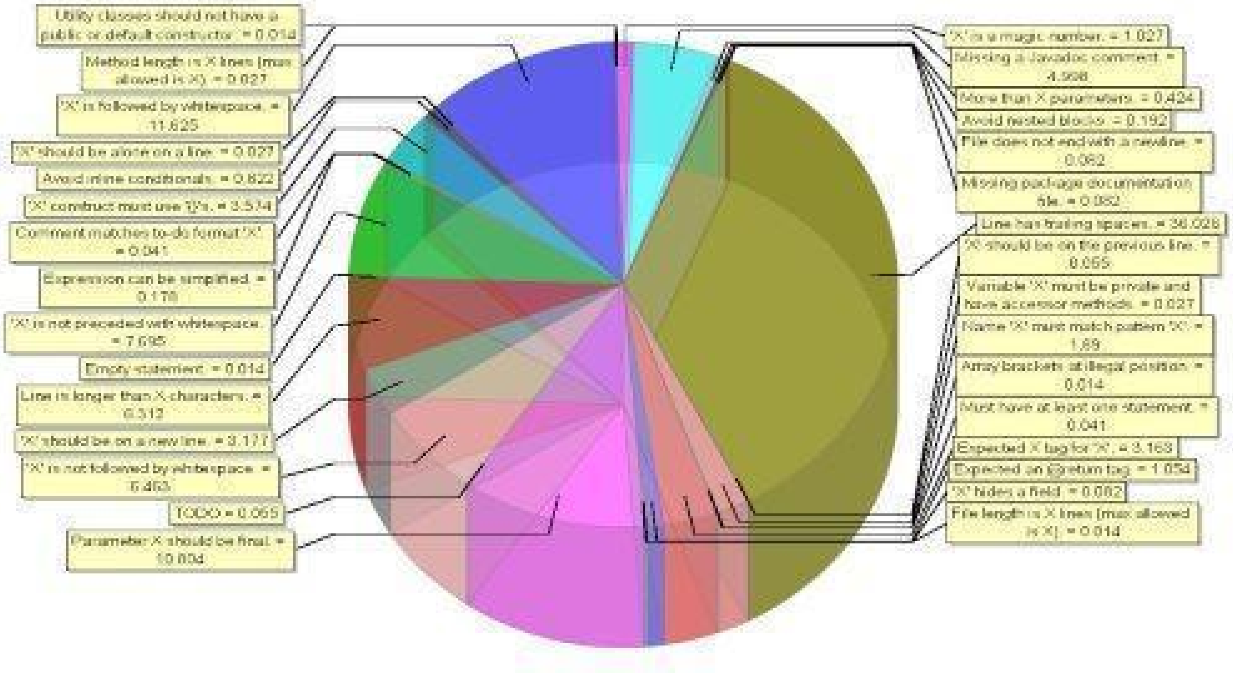
Activity: analyzing visualizations

Look at the four visualizations that follow on the next pages. How could they be improved?

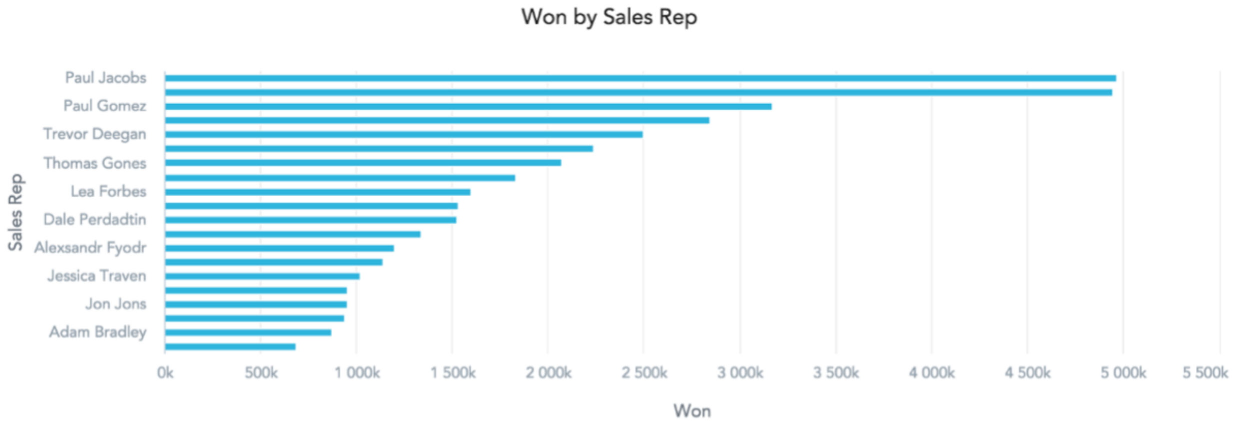
Visual 1



Visual 2



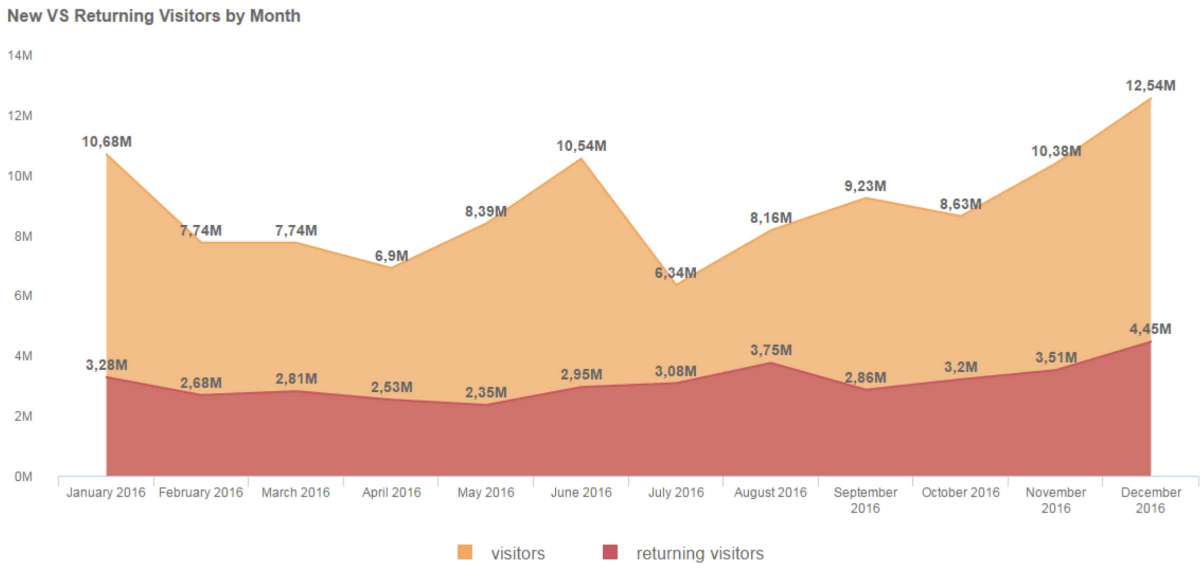
Visual 3



This analysis measures Won.

- Overall Won is 38,310,753.45 across all twenty sales rep.
- The distribution ranges from 684,740.93 (Dave Bostadt) to 4,968,600.86 (Paul Jacobs), a difference of 4,283,859.93.
- The top three sales rep (Paul Jacobs, Ravi Deetri and Paul Gomez) represent 34% of overall Won.
- The lowest three sales rep (Dave Bostadt, Adam Bradley and Cory Owens) represent 7% of overall Won.

Visual 4


New VS Returning Visitors by Month

Additional resources

Data science glossary

Algorithm – An algorithm is a series of steps to accomplish a task (a set of directions that gets you from point A to point B, if you will). It can be a thought exercise, a series of mathematical expressions, or a piece of code (or pseudo code). Algorithms are used in everyday life, and many industries. They are an essential building block of data science, analytics, and any other quantitative field.

API – An application programming interface (API) is an "entryway" to a computer system (such as a database) that allows you to access, retrieve, and edit its components. Most often APIs are used as data-communication mediums between applications. They are an essential part of scalable data-centric applications, research projects that are built around data, and anything else that requires automated data access.

Artificial Intelligence – Artificial intelligence (AI) is the apparent ability of machines to act "intelligently."

Bayes Theorem – Bayes' theorem is used to calculate conditional probability of an event given the knowledge of conditions that might be related to the event. Conditional probability is the probability of an event occurring given another related event has already occurred.

Example: We want to know the probability of A(ge), given the diagnosis of C(ancer). This quantity is unknown to us and hard to estimate. We could use Bayes Theorem to do that, because we have the probability of C(ancer) given the A(ge), the probability of A(ge), and the probability of C(ancer) like so:

P(Age given Cancer) = P(Cancer given Age) *P(Age)P(Cancer)

Big Data – Big Data is a term that describes a large volume of data—both structured and unstructured. It is produced in such quantities that it cannot be ingested or processed by a single machine at once (even a large machine like a supercomputer), so special tools and techniques are needed to process and store it.

Note: This term is often misused; most data that is called Big Data is not really that big. True Big Data is produced by things like Google Searches, satellite imagery of the Earth, climate simulations used by weather services, sensor data generated by cell towers, etc.

Classification – Classification is a supervised machine learning method where the output variable is a category, such as "yes" or "no."

Clustering – Clustering is an unsupervised machine learning method used to discover groupings that are inherent in the data.

Clickstream – A clickstream is a record of a user's activity on the Internet, including every Web site that the user visits, how long the user was on a page or site, in what order the pages were visited, and newsgroups that the user participates in and even the e-mail addresses of mail that the user sends and receives.

Data governance – The management of the overall quality, integrity, relevance, and security of available data.

Data Lake – A data lake is a system or repository of data stored in its natural/raw format, usually object blobs or files.

Note: Data lakes and data warehouses are both widely used for storing big data, but they are not interchangeable terms. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose. (Source: https://www.talend.com/resources/data-lake-vs-data warehouse/)

Data Mining – Data mining is a study of extracting information from structured/unstructured data taken from various sources.

Data Science – Data science is an interdisciplinary field that combines elements of mathematics, statistics, logical thinking, programming, and a range of domain knowledge from various fields, which is used to solve day-to-day problems using a combination of methods and tools from the above disciplines.

Data Warehouse – A data warehouse is electronic storage of a large amount of data, which is designed for query and analysis. It is typically used to connect and analyze data from heterogenous sources.

Note: Data lakes and data warehouses are both widely used for storing big data, but they are not interchangeable terms. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose. (Source: https://www.talend.com/resources/data-lake-vs-data warehouse/)

Data Visualization – Any attempt to make data more easily digestible by rendering it in a visual context (e.g., charting, graphing, etc.).

Database – A database is a structured collection of data. The collected information is organized in a way such that it is easily accessible by the computer. Databases are built and managed by using database programming languages.

Dataset – A dataset is a collection of data, which is organized into some type of data structure. Several characteristics define a dataset's structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them, such as standard deviation.

Deep Learning – Deep learning is a branch of machine learning that uses deep artificial neural networks. Artificial neural networks are types of machine learning algorithms that are built based on the idea of how a brain works with its neurons

connected to each other and transmitting signals. Artificial neural networks are usually built in layers, those that have 3+ layers are considered "deep" and fall into the deep learning bucket.

Graph Analysis – Also known as **network analysis**. Graph analysis is an analysis of structures that model pairwise relationships between objects. The objects in graph analysis are called nodes. Their relationships (a.k.a. connections) are called edges. Graph analysis can be used across many industries, but the most common use case is analysis of social networks.

Machine Learning – Machine learning is the computational process wherein a machine "learns" and adjusts its behaviors based on feedback from data. Usually manifesting as an adaptable algorithm, machine learning helps computers predict outcomes without explicit human input.

Model – A model is a simplified replica of any real-life phenomenon / object at smaller scale. In quantitative disciplines, a model is usually a mathematical description of such a phenomenon / object.

Example: An example is a model to predict the level of education based on age. In a simplified way it looks like this:

Level of education = some quantity + some quantity * age (a linear model)

Although in real life there are a many more factors and variables, this model could potentially display a general trend.

Network Analysis – See graph analysis.

NoSQL Database – A NoSQL database provides storage and the ability to retrieve data that is modeled in means other than the tabular relations used in relational databases.

Open Data – Open data is data that can be freely used, shared and built-on by anyone, anywhere, for any purpose.

Outlier – An outlier is an observation that appears far away and diverges from an overall pattern in a sample.

Regression – Regression is a supervised learning method where the output variable is a real value, such as "amount" or "weight" and the input variable(s) is a real value, such as "size" or "height." The output variable depends on the input variable(s). The relationship between the input and the output are usually recorded in a mathematical model.

Structured Data – Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis.

Supervised Learning – Supervised learning is a type of learning in which we teach or train the machine using data that is well labeled, meaning some data is already tagged with the correct answer class (or group to which it belongs).

Text Mining – Text mining is the process of converting unstructured text data into meaningful and actionable information.

Unstructured Data – Unstructured data is information that either does not have a pre-defined data model or is not organized in a pre-defined manner.

Unsupervised Learning – Unsupervised learning is the training of machines using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

Comparison: popular analysis tools

Learning curve	Steeper learning curve for people without a programming background	Can be easy to learn for people without a programming background	Easy to learn for any analyst
Automation	Once you write commands you won't have to re do the work, just upload a new data set	Once you write commands you won't have to re-do the work, just upload a new data set	New data sets are not always plug and play with your analysis
Analyzing data	Lots of libraries contributed by a broad user community	Over 6,500 packages contributed by the community including top academics	Limited to any particular version
Speed	In-memory, only limited by your computer's RAM	In-memory, only limited by your computer's RAM	Can be slower unless an enterprise configuration is used
Type of data	Reads data of almost any type	Reads data of almost any type	Limited to xlsx and csv files unless macros are used

Compatibility	Compatible with almost any data output, storage or processing platform	Not as compatible as Python with some data architecture systems, may need custom-built interfaces	While macros enhance compatibility, Excel is comparatively limited

Data manipulation	Very flexible data manipulation	Very flexible data manipulation, augmented by numerous data processing and manipulation packages	Color-coded formulas can be easier to use but have a more limited functionality
Seeing data	Command line presentation unless visualization libraries are used	Spreadsheet-like view function that can be less intuitive to navigate	Easy to navigate spreadsheet
Graphics	Cutting edge graphics, however advanced coding and JavaScript knowledge may be necessary	Cutting edge graphics including dynamic visualizations, maps, network graphs, etc.	More limited options based on pre-set drop-down menus (unless macros are used)

Cost & platform	Free, any platform	Free, any platform	Hundreds of dollars, functionality on a Mac does not always mimic a PC

Comparison: popular visualization tools

Microsoft Excel	Create basic chart types such as pie, line, bar, scatter and more Charts created in Excel can easily be ported to PowerPoint and Word
Google Charts	Free and open source Has more options than Excel (e.g., interactive, animated, and geospatial graphics) Includes a rich gallery, fully customizable controls and dashboards, and HTML5
Tableau	Tool for creating powerful and insightful visuals No programming required; drag and drop Share and collaborate on premise or in the cloud Platform can be used department or organization wide
R and R Studio	Free and open source Programming tool Mainly used for statistical analysis, but has sophisticated packages (code contributed by users) to create interactive dashboards as well
Python	Free and open source Programming tool You'll find libraries for practically every data visualization need
Power BI	Includes interactive visualizations and business intelligence capabilities Simple interface Create visualizations and dashboards

Data visualization checklist

This checklist draws on theory such as:

- the building blocks of visual design described by the Interaction Design Foundation
- the four categories of preattentive visual attributes described in Colin Ware's book, *Information Visualization: Perception for Design*
- the Gestalt Principles of visual perception, which describe how people group similar elements, recognize patterns, and simplify complex images when we perceive objects

It is meant to be used as a guide for the development of high impact visualizations; however, the guidelines may be broken intentionally to make a point.

1. Overall Message

☐ The type of graph is appropriate for the data
☐ Graph highlights the significant findings or conclusions and gives the audience the appropriate takeaway message

2. Text

☐ Text size is hierarchical (e.g., titles are larger than labels, which are larger than axis labels, which are larger than source information)
☐ Text size is readable for the selected format (e.g., paper, screen) ☐
Labels have been placed directly next to the relevant data ☐ There
are no extraneous or redundant data labels

3. Color

☐ The background is minimal and not distracting
☐ Visualization uses colors that are consistent with the rest of materials in which it is used
☐ Special effects have been removed or reduced (e.g., bolding, shading) ☐
Color is used effectively to make the audience focus on the most important
pieces of information
☐ Color has been used consistently (i.e., changes in color are used to reinforce a change in topic or tone)

☐ Boldness, saturation, and/or brightness have been varied to distinguish colors for colorblind users

☐ Colors chosen are appropriate for the emotion you want to arouse in the audience

☐ If applicable, brand colors have been applied

4. Lines

☐ Visualization is free of unneeded borders

☐ Unnecessary gridlines have been deleted and necessary gridlines have been lightened

☐ Axes do not have unnecessary tick marks or axis lines

5. Size / Arrangement

☐ Proportions are depicted accurately

☐ Items of almost equal importance are sized similarly

☐ If there's one really important piece of information, it is BIG ☐ Bars/columns are ordered by descending or ascending value (unless trying to show trend over time)

☐ In dashboards, charts that are closer to one another actually more related than other charts that are not as close together

☐ In dashboards, the most important pieces of information are placed at the top of the page

☐ In dashboards, elements are placed in a way that feels natural

Additional reading & reference

Analyzing the Analyzers by Harlan Harris, Sean Murphy and Marck

Vaisman *Doing Data Science* by Cathy O'Neil & Rachel Schutt

Data Science for Business by Foster Provost & Tom

Fawcett *Data Smart* by John W. Foreman

Information Visualization: Perception for Design by Colin

Ware *Mining the Social Web* by Matthew A. Russell

Predictive Analytics by Eric Siegel

The Building Blocks of Visual Design by Teo Yu Siang, for the Interaction Design Foundation

When Numbers Mislead by Stephanie Coontz, in The New York

Times ***Use cases***

BNY Mellon advances artificial intelligence tech across operations

Combining Satellite Imagery and Machine Learning to Predict

Poverty New York City uses "nudges" to reduce missed court dates

Proof of concept: Using predictive modeling to prioritize building

inspections Recruiting Chatbots in 2021: In-Depth Guide

Tactical Institute: Protecting people and communities with pre-emptive social media threat analytics

What Wal-Mart Knows About Customers' Habits