DATA SOCIETY®

The premiere data science training for professionals

Day 2 recap

- Yesterday, we learned about:
  - How data science impacts agencies and organizations
  - Best practices for data visualization and data ethics
  - The value and prevalence of open data

How are you thinking about data differently?
Are there any projects you want to start developing using data that you have already?

Activation activity

- Read **sections 5 and 6** of "DATA SCIENCE AND THE USAF ISR ENTERPRISE"

- How is the Air Force thinking about using Data Science and data scientists? How are the scenarios described relevant to your line of work?

Activity time: 15 - 20 minutes

Outline for today

1. Data governance
2. Tools and technology
3. Building a data-driven culture

The biggest challenges of using data

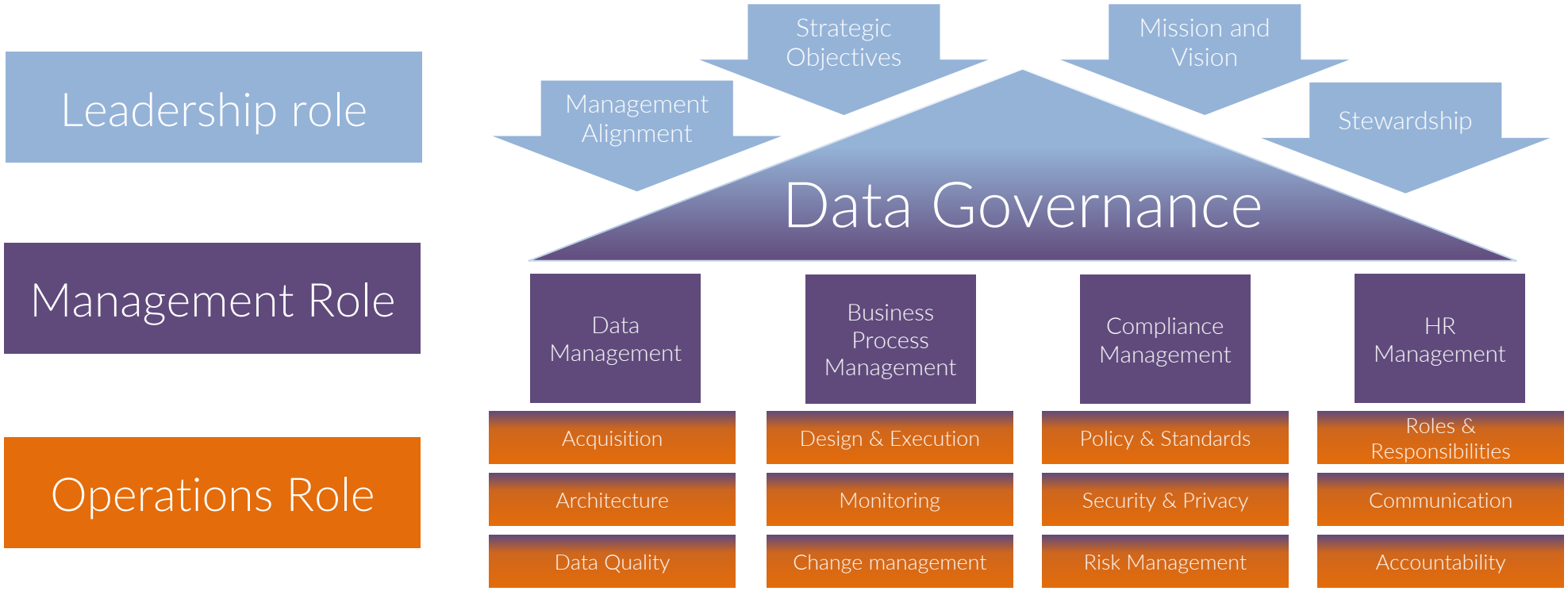
- These are the most common challenges that organizations encounter with their data:

1. Data is not collected or collected inconsistently
2. Data is not standardized and not formatted correctly
3. Data is missing values, contains duplicates, or is recorded incorrectly
4. Data cannot be accessed in a timely manner by the people who need it
5. Data is not secured properly
6. Data is not used to analyze and interpret the effectiveness of decisions

- The objective of a data governance structure is to avoid those issues

What is Data Governance?

- Data governance is oversight for all data management that leads to a set of processes and that the data is reliable and used effectively

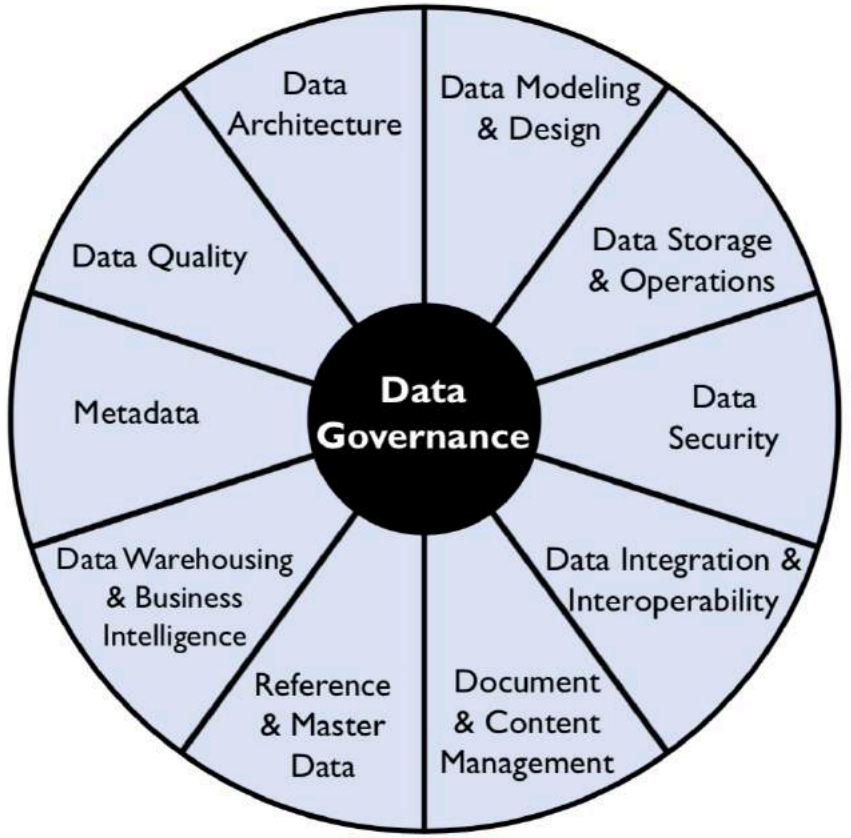


Why is data governance important?

1. Regulatory compliance – with increased regulation comes compliance that needs to be implemented and followed

2. Reduce risk – effective data governance enhances data security and privacy

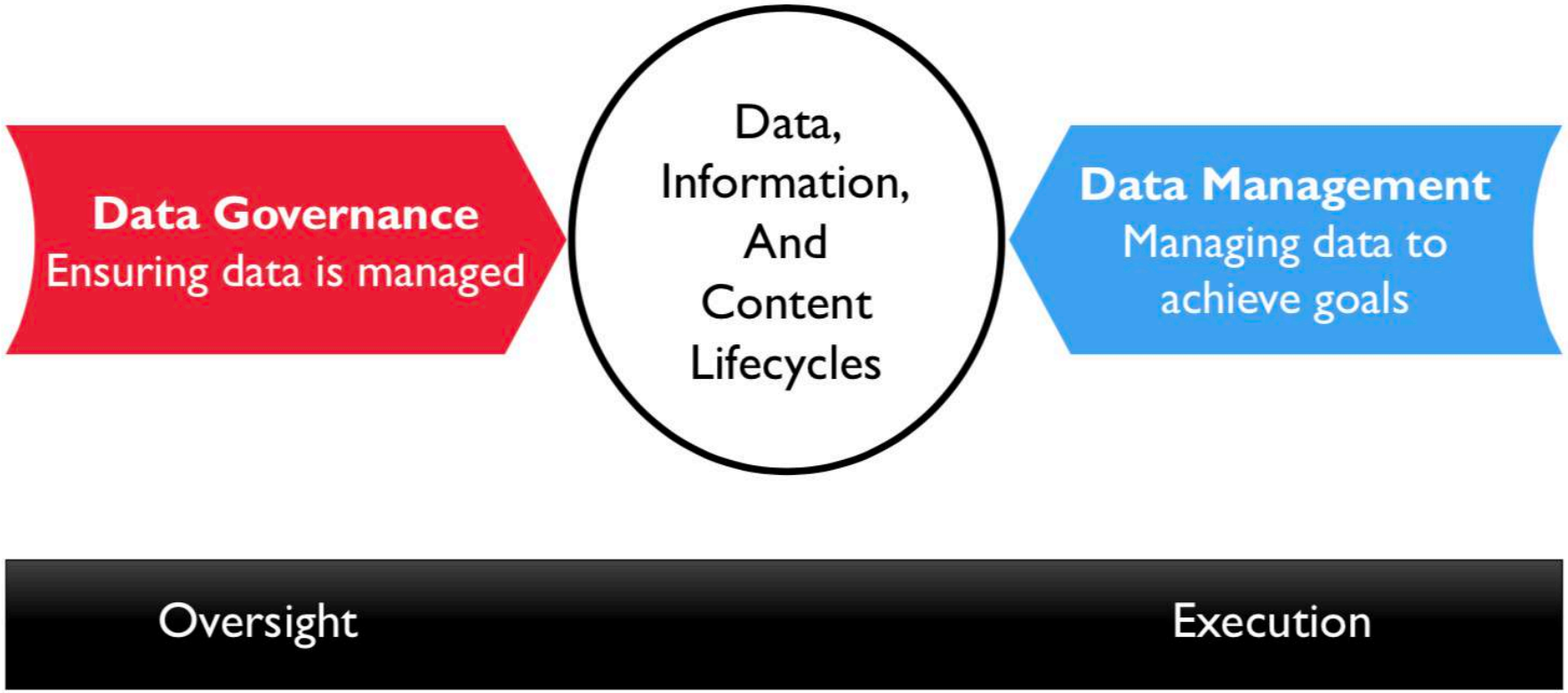
3. Improve processes – when everyone follows the same standards, projects and management become more efficient



DAMA-DMBOK2 Data Management Framework

Oversight vs execution

- Data governance and data management are closely interconnected



Data governance principles

Data governance should enable an organization to make data into an asset – it should be:

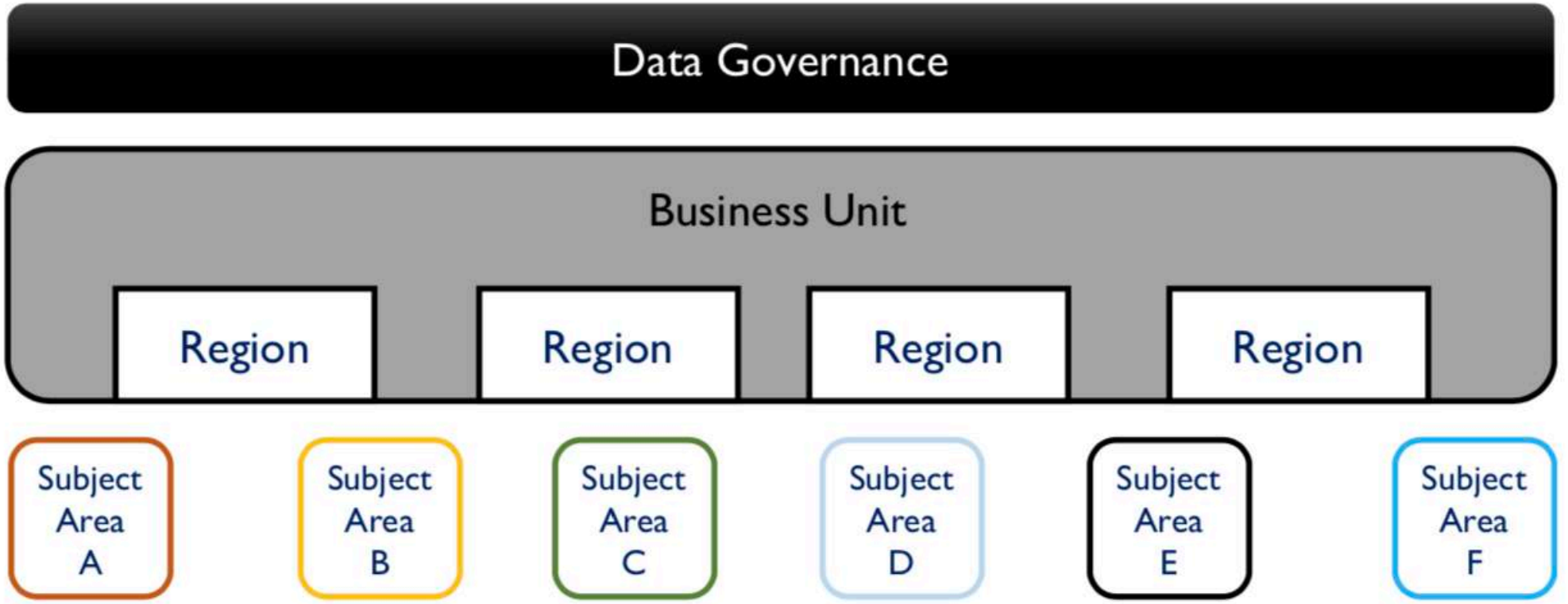
1. Sustainable – manage this change so that it survives beyond the initial implementation

2. Embedded – DG should be present in all processes related to data

3. Measured – there should be some defined metrics to help demonstrate value to the organization

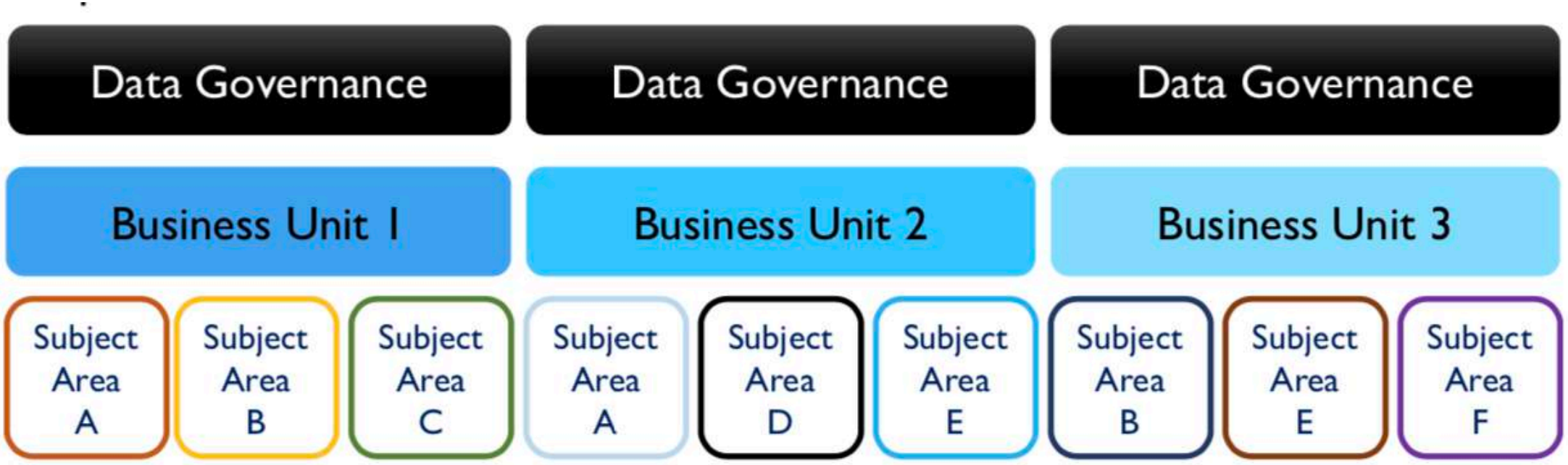
Operating models

1. Centralized – one overarching data governance organization applies to all sectors



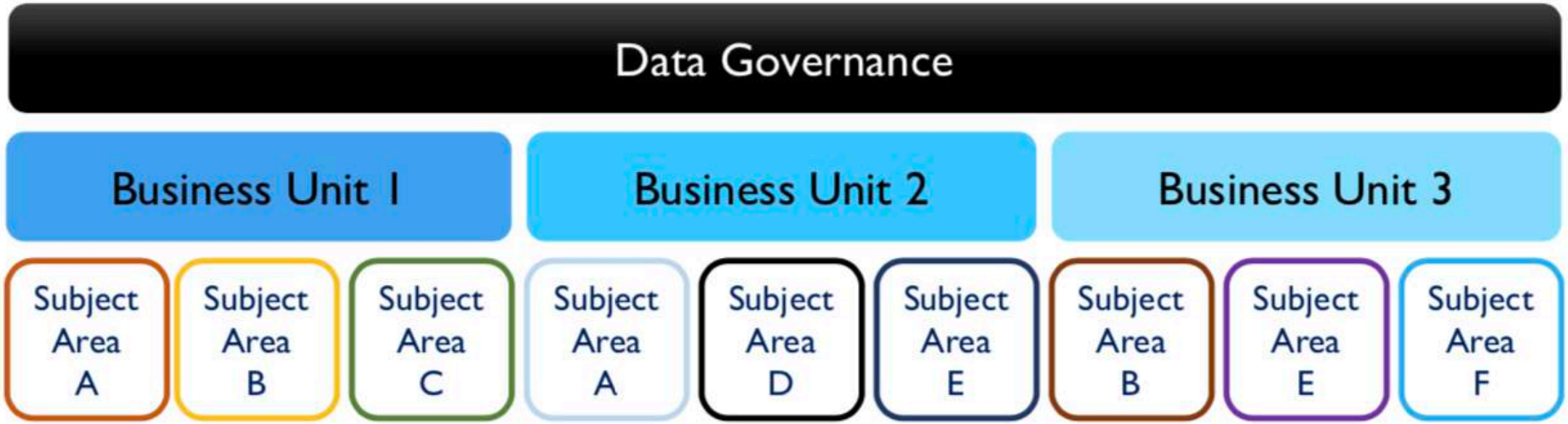
Operating models

2. Replicated – each data governance section is repeated across departments, but may have multiple governing bodies



Operating models

3. Federated – an overarching data governance organization works with multiple departments to maintain consistency



Assigning data stewards

- A data steward is someone who is accountable and responsible for the data processes – typically, it's someone who has already assumed some data management already

- Some of their responsibilities may include:
  - Creating and managing metadata
  - Documenting rules and standards
  - Managing data qualities
  - Executing operational data governance

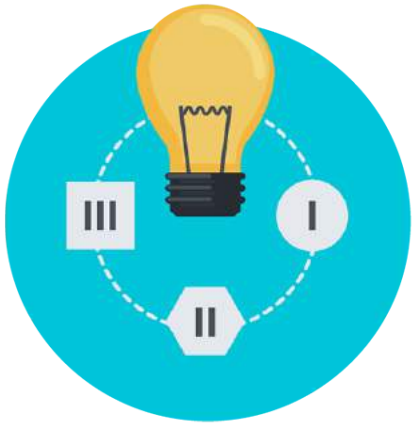
- This could be a CDO or someone in the data department

Developing your strategy

- Deliverables include:

Charter

Implementation roadmap



Operating framework / accountabilities

Plan for operational success

Data management knowledge areas

- When you develop your data governance documentation, you'll need to ensure that it lines up with your organization's objectives

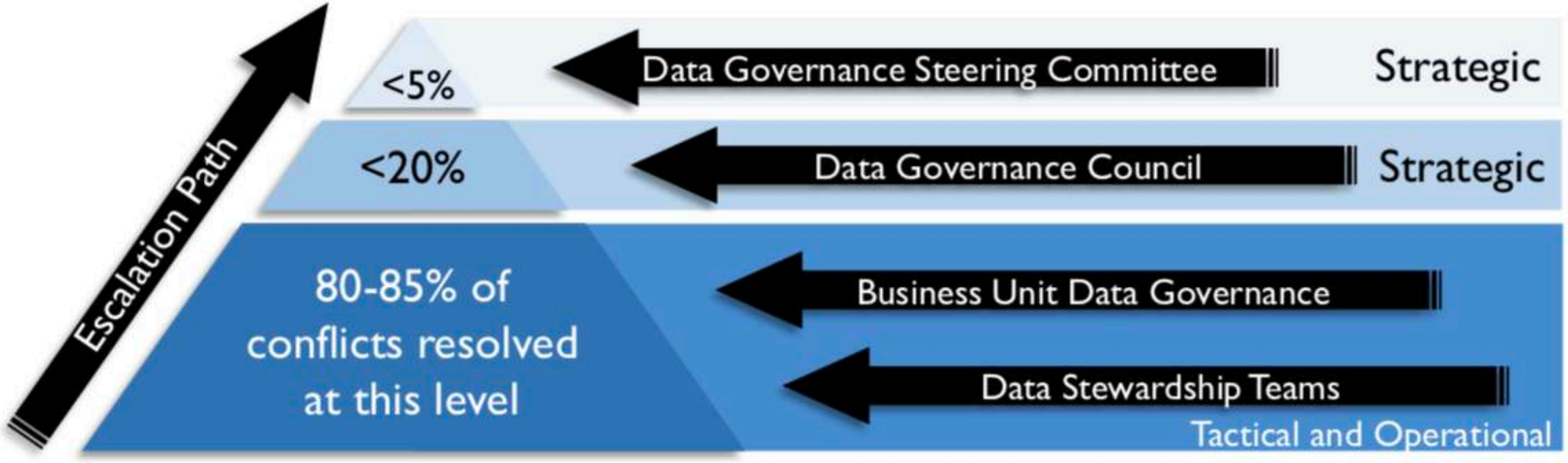
- Some typical areas covered include:
  - Data architecture
  - Data modeling
  - Data storage
  - Data security
  - Data integration

Measuring data governance

- The only way to develop a successful data governance program is to provide tangible metrics along the way
- The best metrics include ones that demonstrate the value of the data governance program towards the organization's objectives
- Sample metrics may include reduction of risk, speed of updated processes, conforming to procedures and compliance (such as with GDPR and other new data regulations)

Managing data issues

- As issues arise, it's important to define who is responsible for resolving them
- There might be different individuals for managing compliance issues vs data quality vs data security



A framework for data quality

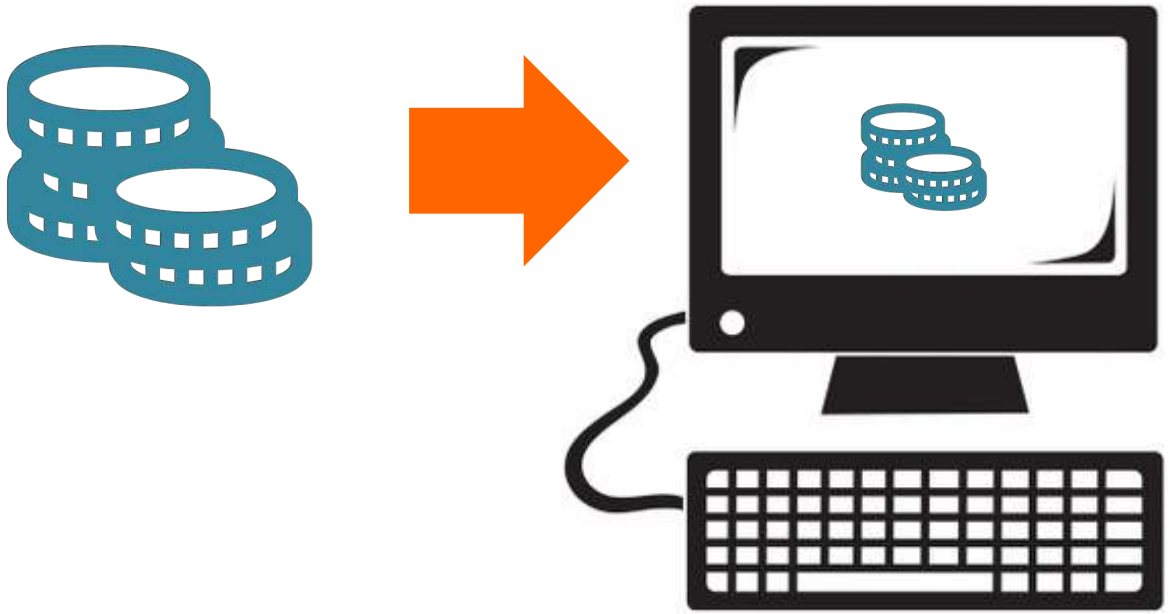
- Identify where human error occurs: keep track of data's travel path
- Redundancy for record-checking
- Organization-wide standards for:
  1. Data entry
  2. Data checking
  3. Records structure
  4. Setup for who owns the data
  5. Train analysts and data owners on data accuracy
  6. Set up regular checks to ascertain data accuracy
- Track who has access to the data and who has the permission to modify it
  - Security!
- Version control!
  - Backups and redundancy

Accuracy

Does the data reflect a real object of event?

- Accuracy
  - The data was recorded correctly
- Completeness
  - All relevant data was recorded
- Uniqueness
  - Entities are recorded once
- Timeliness
  - The data is kept up to date
- Consistency
  - The data agrees with itself

Data recorded should reflect real life

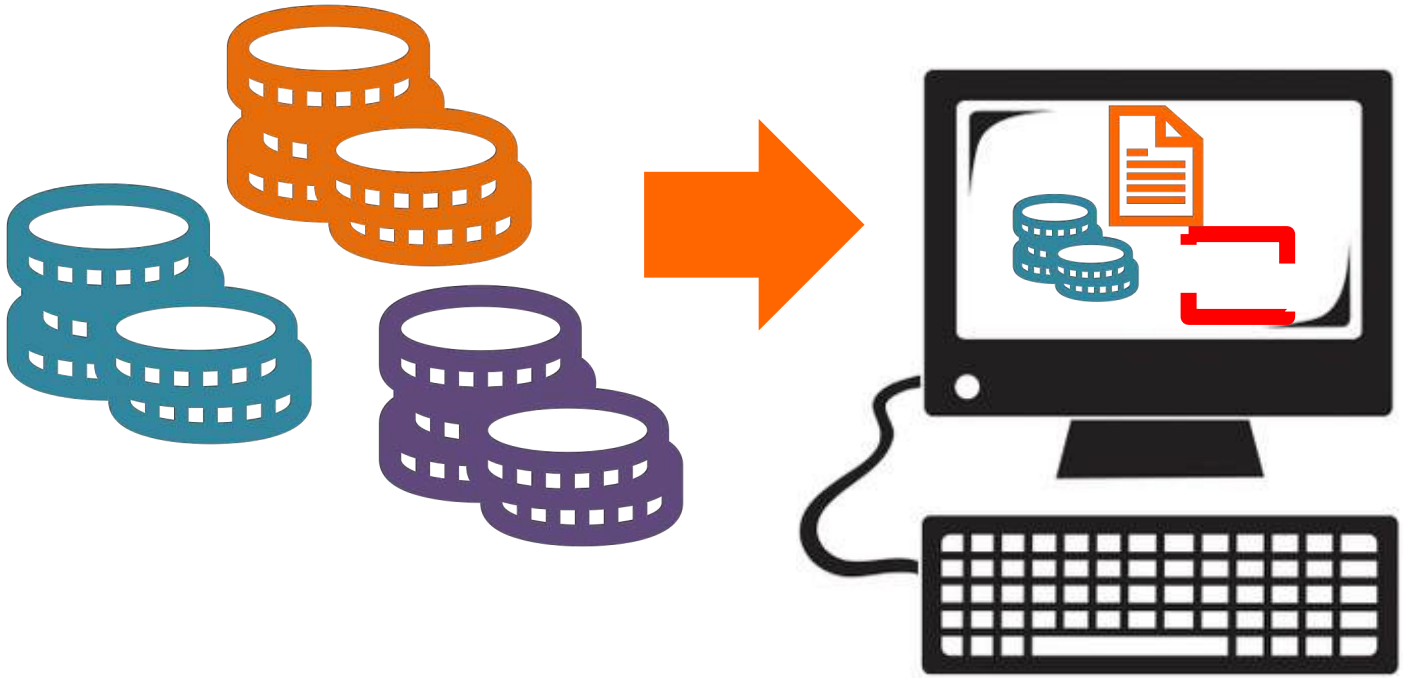


Completeness

Is the data available and populated where it's needed?

- Accuracy
  - The data was recorded correctly
- Completeness
  - All relevant data was recorded
- Uniqueness
  - Entities are recorded once
- Timeliness
  - The data is kept up to date
- Consistency
  - The data agrees with itself

Data recorded should represent the entire population of outcomes

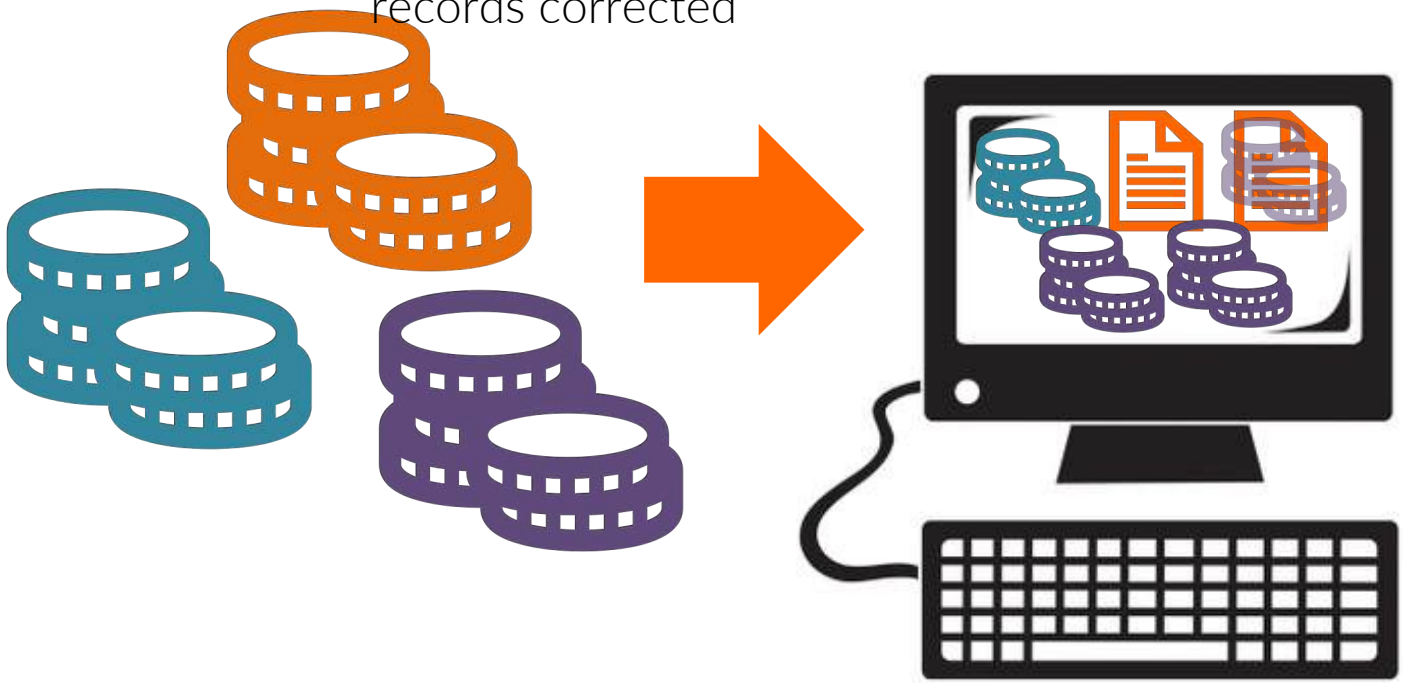


Uniqueness

Are there duplicate records? Can you differentiate entities?

- Accuracy
  - The data was recorded correctly
- Completeness
  - All relevant data was recorded
- Uniqueness
  - Entities are recorded once
- Timeliness
  - The data is kept up to date
- Consistency
  - The data agrees with itself

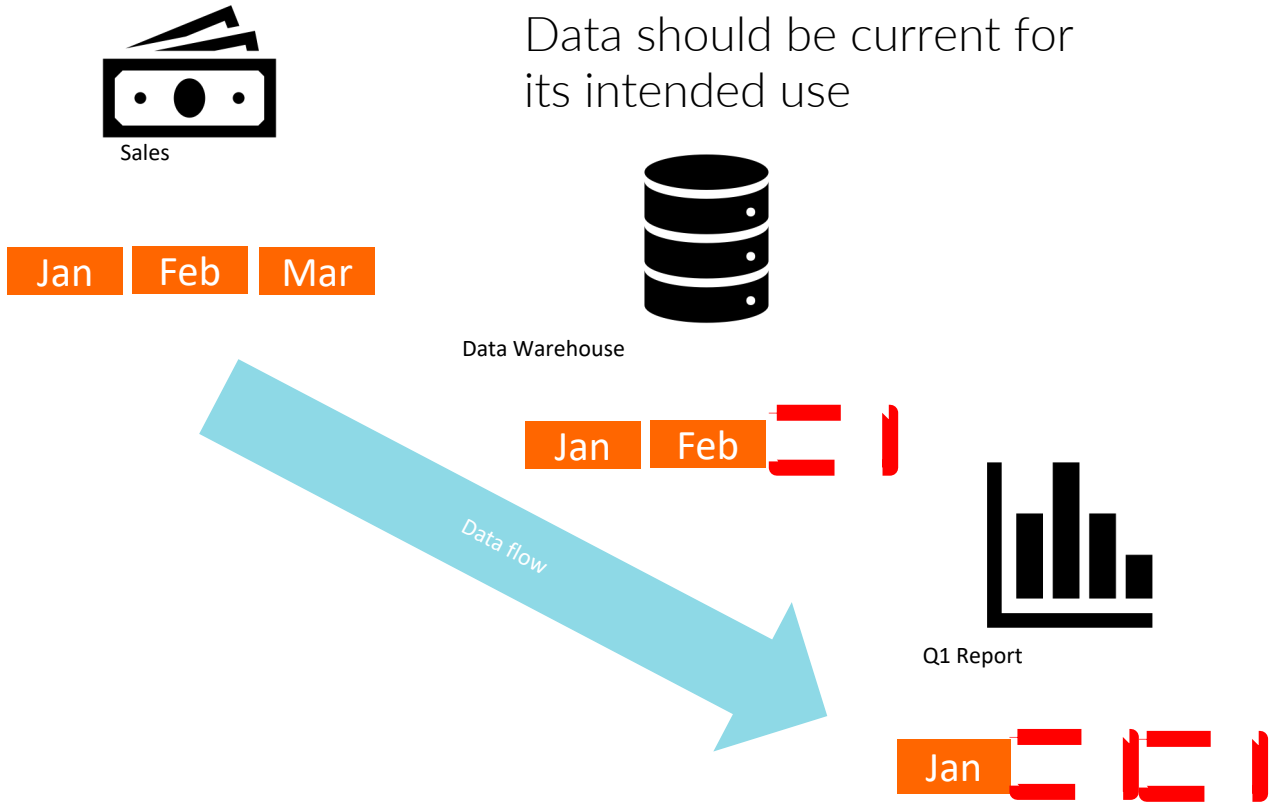
Duplicate records should be removed and indistinguishable records corrected



Timeliness

Is the data you need there when you need it?

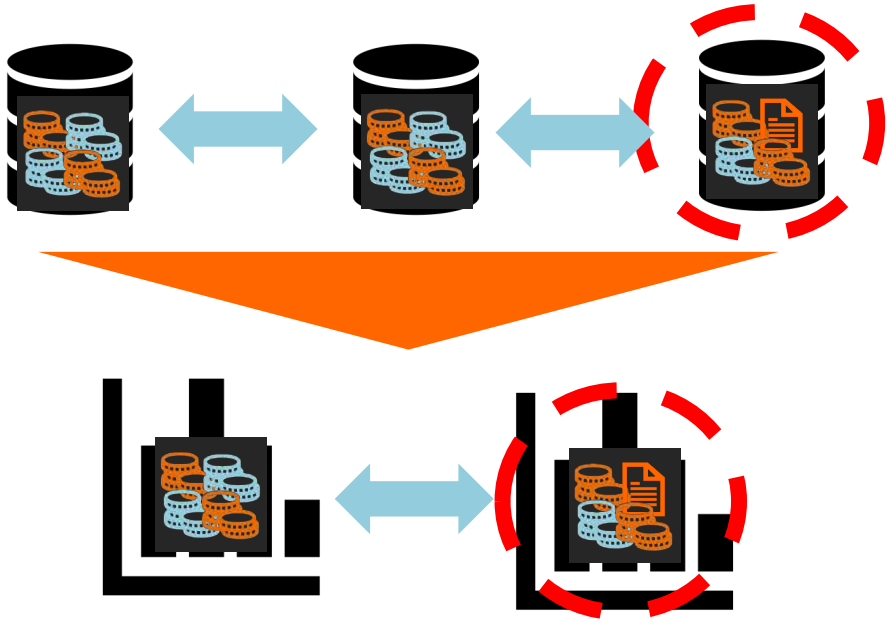
- Accuracy
  - The data was recorded correctly
- Completeness
  - All relevant data was recorded
- Uniqueness
  - Entities are recorded once
- Timeliness
  - The data is kept up to date
- Consistency
  - The data agrees with itself



Consistency

Does data about an entity stay the same across reports and systems?

- Accuracy
  - The data was recorded correctly
- Completeness
  - All relevant data was recorded
- Uniqueness
  - Entities are recorded once
- Timeliness
  - The data is kept up to date
- Consistency
  - The data agrees with itself



Databases and reports should reconcile

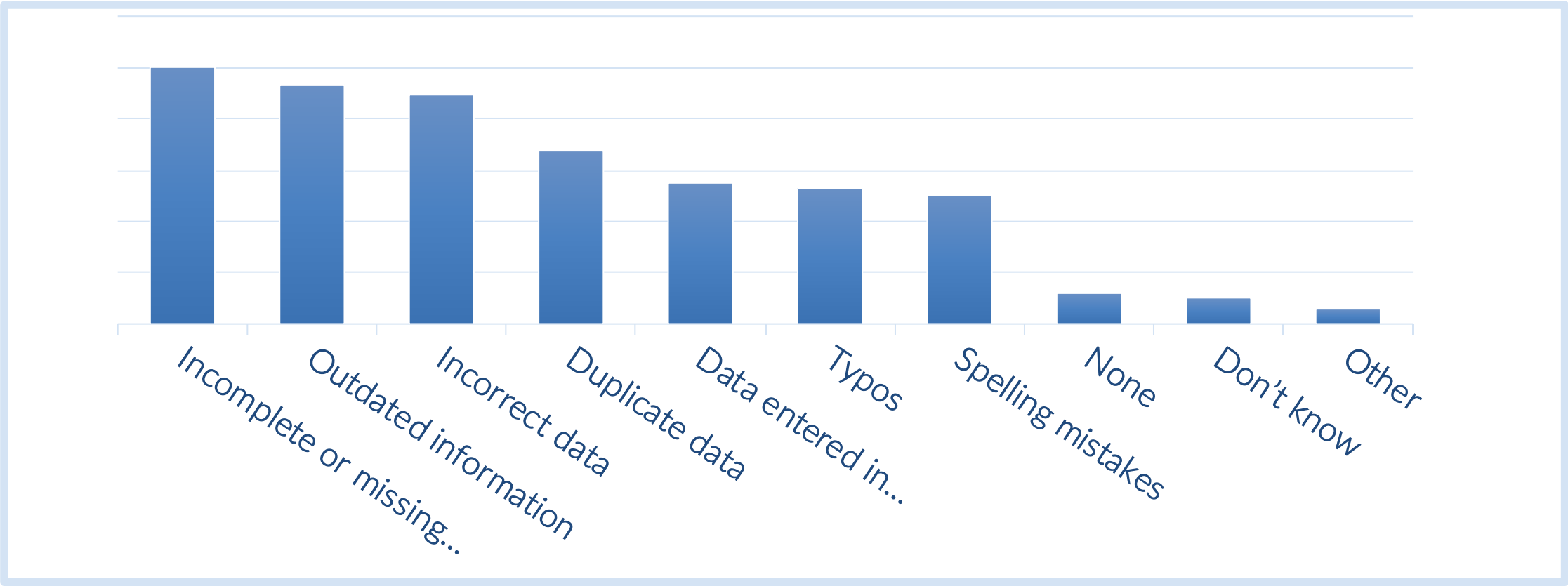
Basic data quality

Garbage in – garbage out!

- Accuracy
  - The data was recorded correctly
- Completeness
  - All relevant data was recorded
- Uniqueness
  - Entities are recorded once
- Timeliness
  - The data is kept up to date
- Consistency
  - The data agrees with itself

Acquiring accurate data is hard

- There are many data quality issues with many root causes
- Make sure your data quality processes check for and correct all material issues

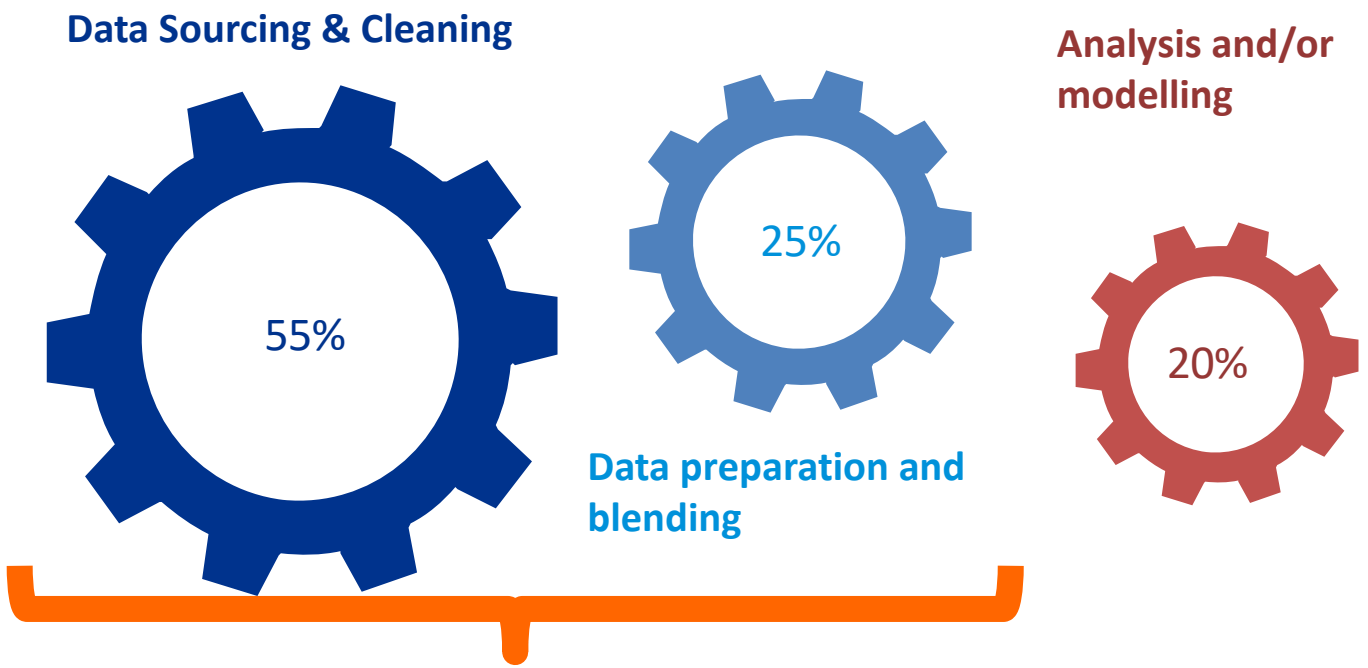


Data preparation

Bridging organizational silos requires a process for collaboration

- Often organizational and functional groups results in silos of information, knowledge, and communication
- Silos increase the time required to collaborate on data projects
- Effective governance standards are essential to mitigating the drain on resources created by silos

Time spent on data projects and processes



~80% of an analytics project spent data prep & cleaning

Data governance best practices

- In order to maximize your success, you should ensure to:

Involve leadership

Start with a manageable project

Iterate on the principles

Educate teams to build support

Activity: evaluate yourself!

- Turn to your participant guide to the **Data governance assessment** to see how far along you and your team are in the data governance cycle

- You'll measure the foundational components, such as **awareness**, **formalization**, and **metadata**, as well as the project components of **stewardship**, **data quality**, and **master data policies**

- Then, assess your progress and set goals for where you want your team to be and discuss with your group

Activity time: 15 - 20 minutes

Outline for today

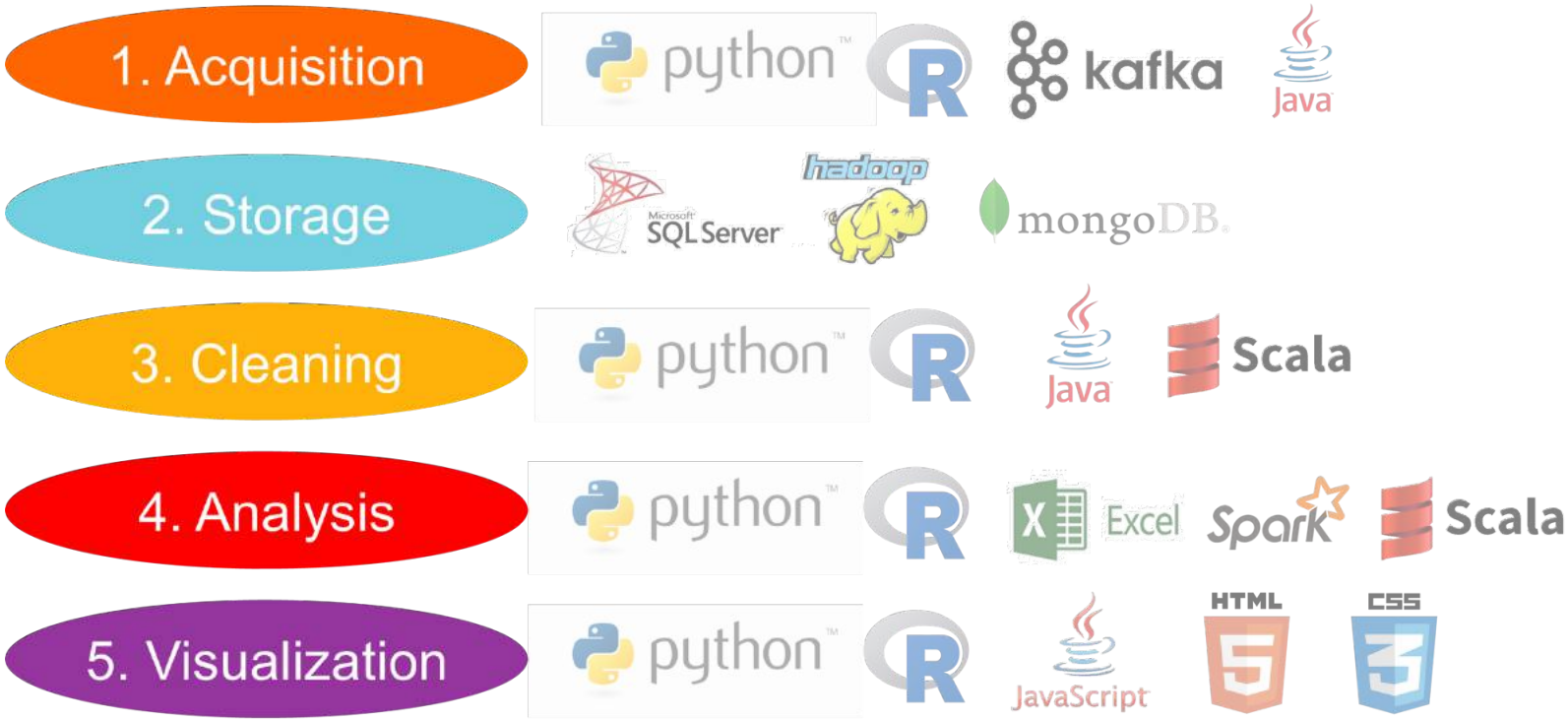
1. Data governance
2. Tools and technology
3. Building a data-driven culture

No shortage of tools

- If you only have a hammer, everything looks like a nail – most tools overlap in functionality
- Key issues many organizations encounter:
  - Only a handful of people know how to use a particular tool
  - Legacy tools can be slower and more expensive than newer open source tools
  - Tens of thousands of dollars are spent on licenses
  - The tool is more complex than necessary

Technologies often drive data quality

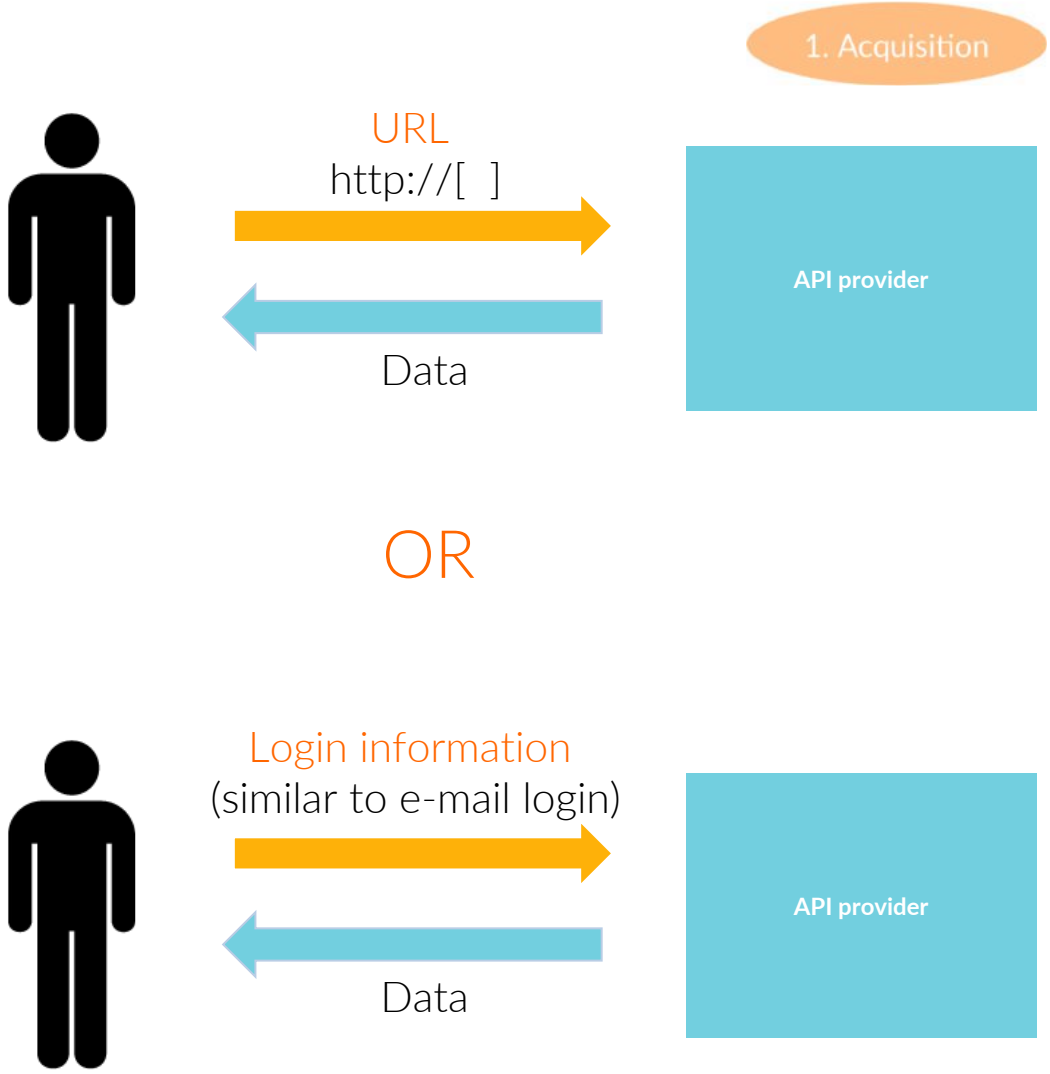
- The fewer tools and databases there are in the platform, the easier it is to avoid errors in the data



Data sources: API = a fire hose of data

What is an API?

- API stands for *application programming interface*
- APIs allow you to use the data of sites like Google Maps, Pinterest, Twitter, WalMart and Best Buy
- Allows you to download large amounts of data directly from the provider's repository
- APIs allow you to custom select the subset of data that you want to use



APIs: social media as a leading indicator

1. Acquisition

- Global Pulse and SAS International:
  - Social media conversations about work-related anxiety provided a 3-month early warning indicator about unemployment in Ireland

- A single API to access data from many social media sites https://gnip.com/

External data sources

1. Acquisition

- http://www.programmableweb.com offers one of the world's largest collections of publicly accessible APIs
- API.data.gov offers the open data available from Data.gov via API
- Large business are increasingly giving people access to their data:
  - WalMart, Best Buy, Trip Advisor, Expedia, Google, Spotify, the list grows every day!

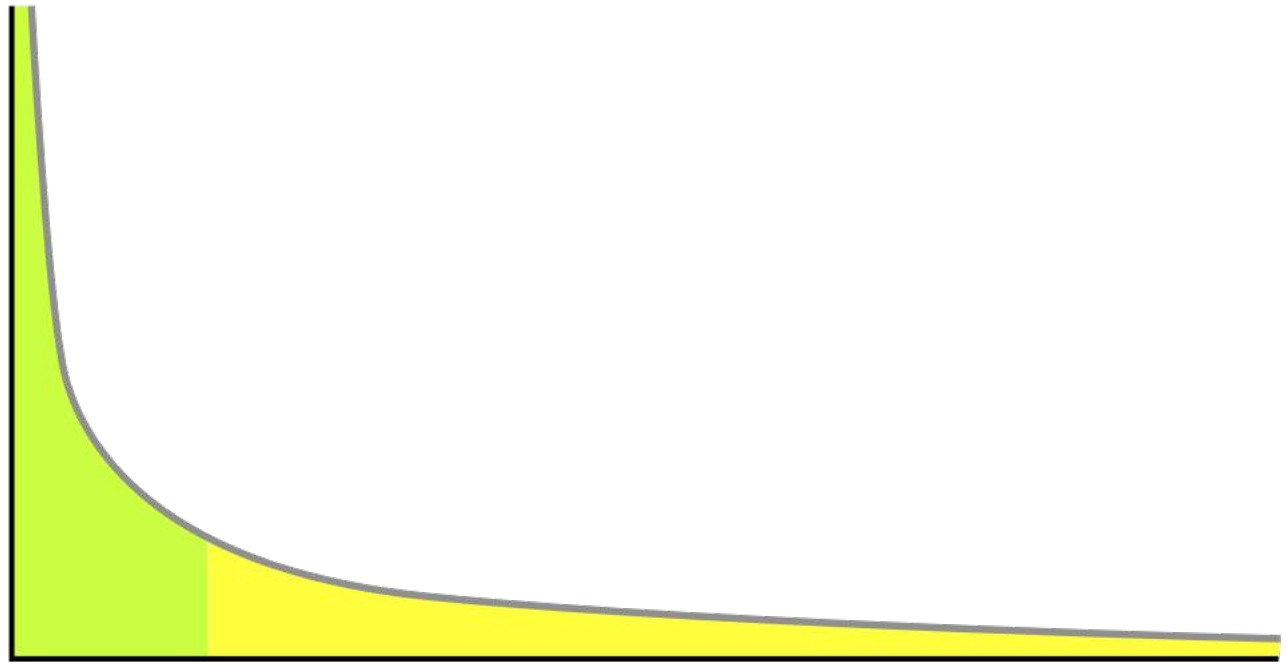
Questions:

- How many are buying?
- How many are traveling?
- How many are searching?
- How many are listening?

Internal data sources

1. Acquisition

- HR (performance data, salary/compensation, hiring, 360 view, etc.)
- Network Data (application logs, webserver logs, firewall alert logs, e-mails)
- Leads/sales (salesforce.com)
- Clickstream
- Webserver (travel management)
- Contracts/proposals/procurement
- The Long Tail of Big Data!



Different tools for different data

2. Storage

Structured

y1	x1	x2	x3

Semi-structured

```
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```

Quasi-structured

```
Sep 17 02:33:08.536 [debug]
connection_edge_process_relay
_cell(): Now seen 1802 relay
cells here (command 2, stream
5845).
Sep 17 02:33:08.536 [debug]
connection_edge_process_relay
_cell(): circ deliver_window
now 933.
```

Unstructured

- There is no "correct" way to parse unstructured data
- Data scientists can work with unstructured data, but we have to think harder about parsing, storing, and analyzing the results
  - This adds more time and more tracking back but has upside!
  - Data mining can unlock the information in unstructured data!

Relational databases: SQL

2. Storage

- Structured query language (SQL) is a programming language designed for accessing relational databases

- There are a variety of databases that work with SQL
  - SQL code often requires some modifications to be portable between databases

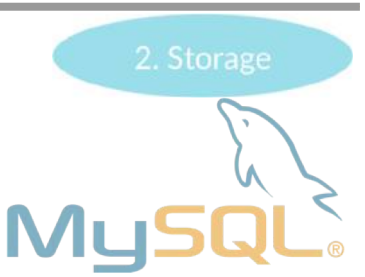
- Strengths of relational databases
  - Easy to store large amounts of tabular data

- Weaknesses of relational databases
  - Doesn't work as well with text and unstructured data

Relational database

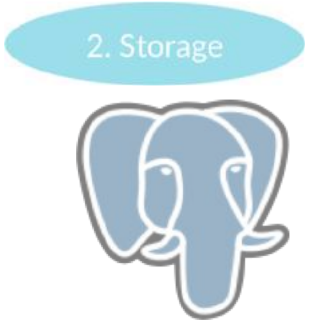
y1	x1	x2	x3
A	F	X	P
B	G	Y	Q
C	H	Z	W

Relational databases: MySQL



- Open-source relational database management system
  - One of the most popular in use today
- MySQL is the database of choice for many web applications including WordPress
- Command-line tool that requires SQL programming to access
- *Note: MariaDB is another relational database made by the original developers of MySQL and guaranteed to stay open source. Notable users include Wikipedia, WordPress.com and Google*

Relational databases: PostgreSQL



- Relational database management system
- Primary function: store data securely and allow access by other software applications
- Can handle large streaming applications with many simultaneous users
- Initial release: 1996
  - Widely used database system with a big user community

MongoDB & unstructured databases

2. Storage

- Free, open-source database best suited for storing document and text data
- Uses Javascript Object Notation (JSON)
  - A convenient way for storing semi-structured and unstructured data
- Databases for unstructured data are often called NoSQL databases

```
{
    "firstName": "John",
    "lastName": "Smith",
    "age": 25,
    "address": {
        "streetAddress": "21 2nd Street",
        "city": "New York",
        "state": "NY",
        "postalCode": "10021"
    },
    "phoneNumber": [
        { "type": "home", "number": "212 555-1234" },
        { "type": "fax", "number": "646 555-4567" }
    ]
}
```

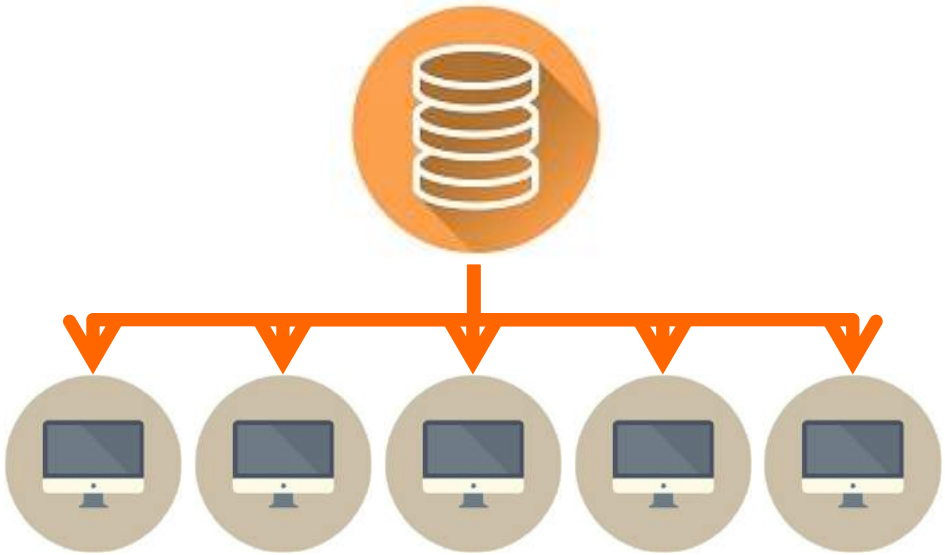
- As of July 2015, MongoDB was the 4th most widely mentioned database engine on the web, and the most popular for document stores

Source: Wikipedia.org

Distributed databases: Hadoop

2. Storage

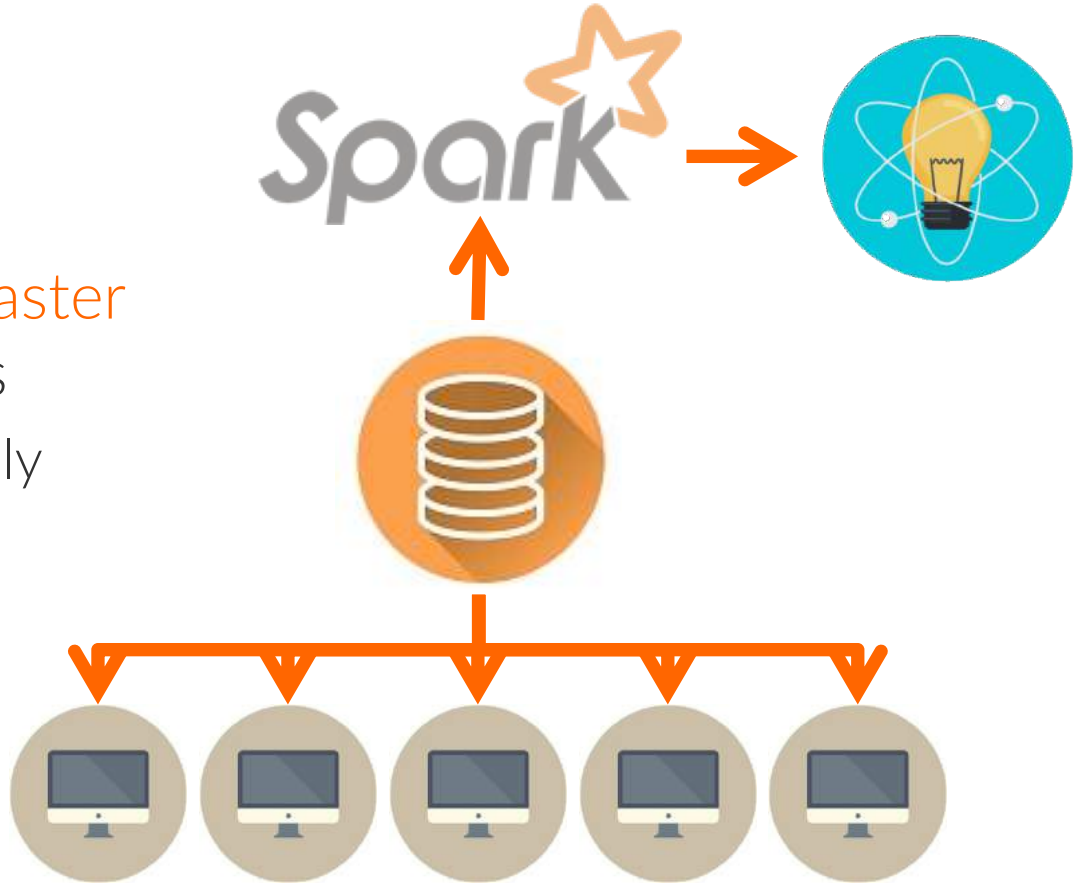
- Hadoop: free software for distributed data storage
- Two components:
  1. Hadoop Distributed File System (HDFS)
  2. MapReduce computational engine
- HDFS: a framework to store Big Data
- Lots of technologies available to query distributed data
- MapReduce is a framework to analyze data in a distributed fashion



Distributed, fast computation: Spark

2. Storage

- Spark is an alternative computation engine to MapReduce
- Spark is ~100x faster than MapReduce
- It features in-memory computations for faster processing and APIs in multiple languages
  - Accessible with Python, which is a more widely accessible skill set
- Spark is quickly becoming a new standard for Big Data analysis



Data warehouse: own vs. rent

2. Storage

- Amazon.com offers a suite of cloud storage and computational resources
  - Known as Amazon Web Services ("AWS")
- Amazon Simple Storage Service ("S3") is an inexpensive way of data outside of the organization
  - Analysis can be done on data stored remotely
  - Scale infrastructure up or down as needed
  - Decreases fixed costs – no need to build a "server farm"
  - Allows for operational flexibility
- GovCloud service for U.S. government clients
  - Designed to handle sensitive data and regulated workloads
- Socrata provides a central data platform for the government that allows for collaboration in the cloud

How to store the data?

2. Storage

Relational

- A government agency needs to store a large amount of tabular data for frequent access by analysts

Relational for live access

- A database needs to support continuous access to data by numerous users of an online tool

Text data repository

- A government agency needs to store a large amount of patent application data for easy search and analysis

Distributed data storage

- 1 terabyte of data about the usage of federal government resources needs to be stored for analysis of procurement and resource allocation

How to store the data?

2. Storage

Relational

- A government agency needs to store a large amount of tabular data for frequent access by analysts

Relational for live access

- A database needs to support continuous access to data by numerous users of an online tool

Text data repository

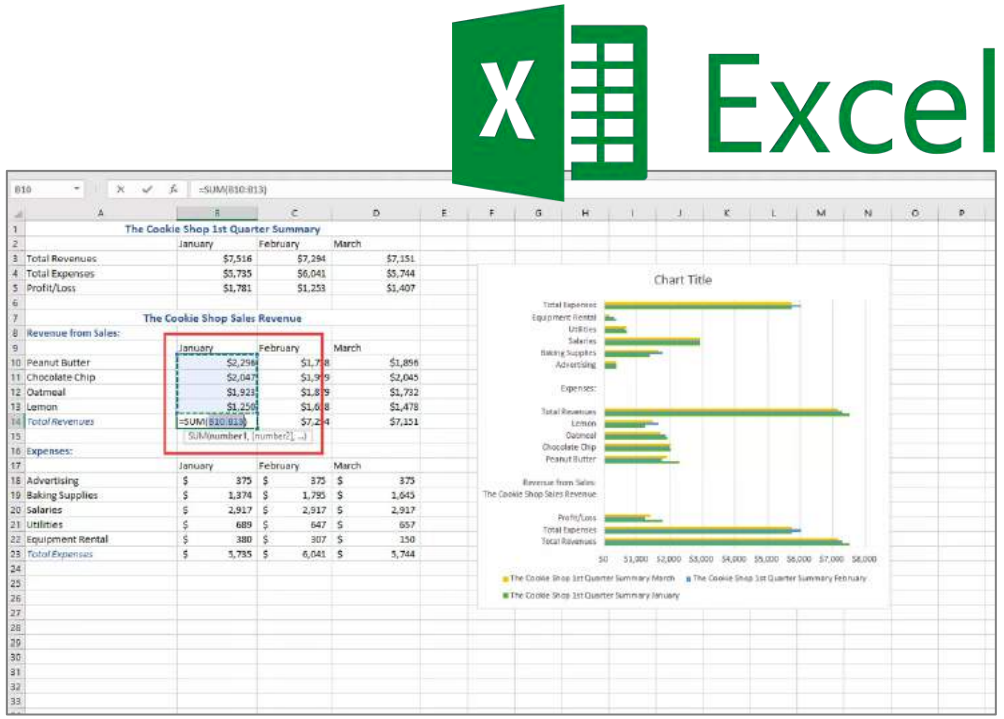
- A government agency needs to store a large amount of patent application data for easy search and analysis

Distributed data storage

- 1 terabyte of data about the usage of federal government resources needs to be stored for analysis of procurement and resource allocation

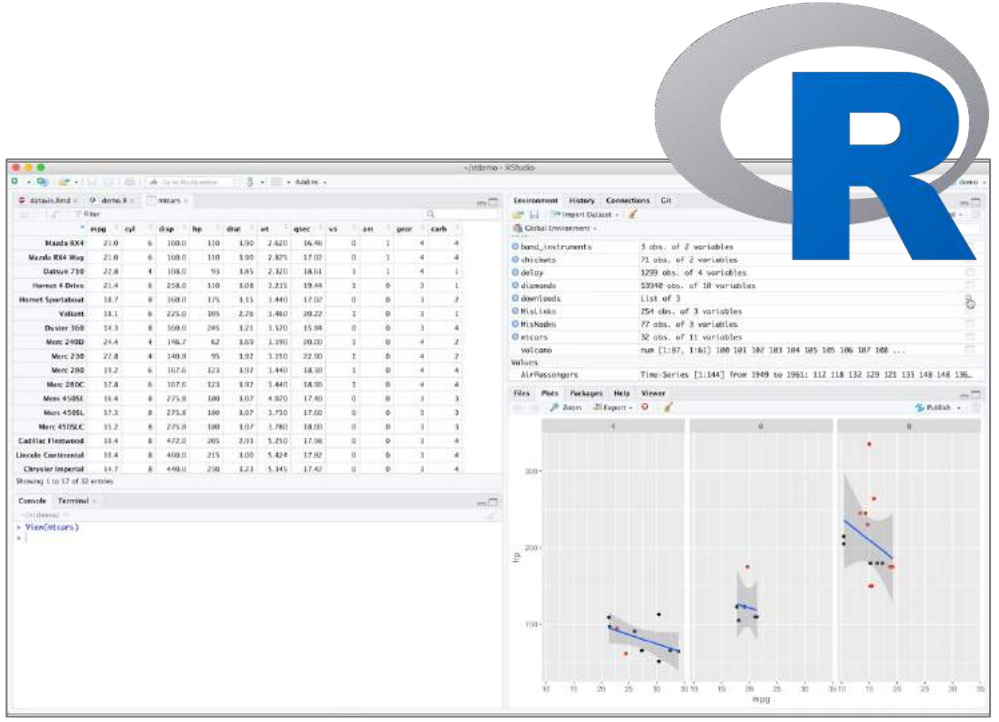
Data cleaning / analysis: Excel

- Ubiquitous to almost any office, it is a spreadsheet program used to store, organize, manipulate, and visualize data
- Limited by number of rows and columns in your data, useful for simple cleaning tasks and sharing spreadsheets
- Not as versatile for more advanced analytics or checking for errors in formulas



Data cleaning / analysis: R

- Built originally by statisticians to simplify data analyses, now used by data analysts and data scientists for complex algorithms and data visualizations



- Free and open source language, with over 10,000 packages contributed by millions across the globe

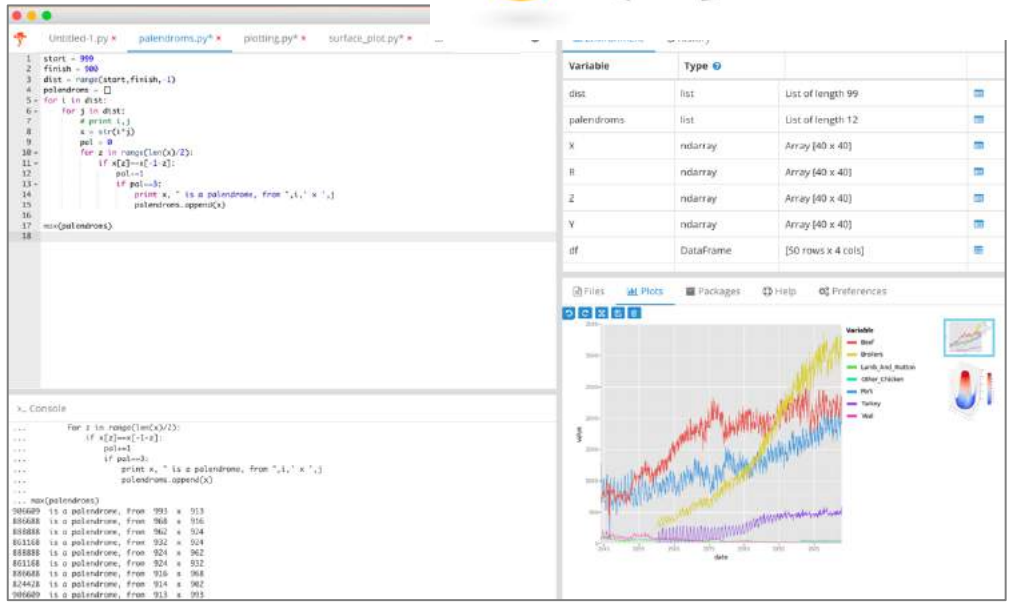
- Versatile coding scripts and reusable for different datasets

Data cleaning / analysis: Python

- Built originally by statisticians to simplify data analyses, now used by data analysts and data scientists for complex algorithms and data visualizations

- Free and open source language, with over 10,000 packages contributed by millions across the globe

- Versatile coding scripts and reusable for different datasets



Analysis & vis: Excel, R and Python

Great *Medium* *Limited*

4. Analysis

	Python	R	Excel
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

Analysis & vis: Excel, R and Python

Great *Medium* *Limited*

4. Analysis

	Python	R	Excel
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

Which tool to use?

4. Analysis

Consumer-facing software used by millions of people

- Claims handling engine for the Department of Veterans Affairs
- Built by programmers
- Optimized for large scale access
- Interfaces with many databases

Predictive analytics tool used by policy makers

- A team of analysts and policy makers needs on-demand predictive analysis to inform decision making
- Daily updates based on new data
- Make recommendations based on changing economic conditions

Record keeping for the accounting department

- A small firm needs to keep track of its expenditures on office supplies
- Anyone needs to be able to input data into the system
- Anyone can take over this responsibility with no training

Which tool to use?

4. Analysis

Consumer-facing software used by millions of people

- Claims handling engine for the Department of Veterans Affairs
- Built by programmers
- Optimized for large scale access
- Interfaces with many databases

Predictive analytics tool used by policy makers

- A team of analysts and policy makers needs on-demand predictive analysis to inform decision making
- Daily updates based on new data
- Make recommendations based on changing economic conditions

Record keeping for the accounting department

- A small firm needs to keep track of its expenditures on office supplies
- Anyone needs to be able to input data into the system
- Anyone can take over this responsibility with no training

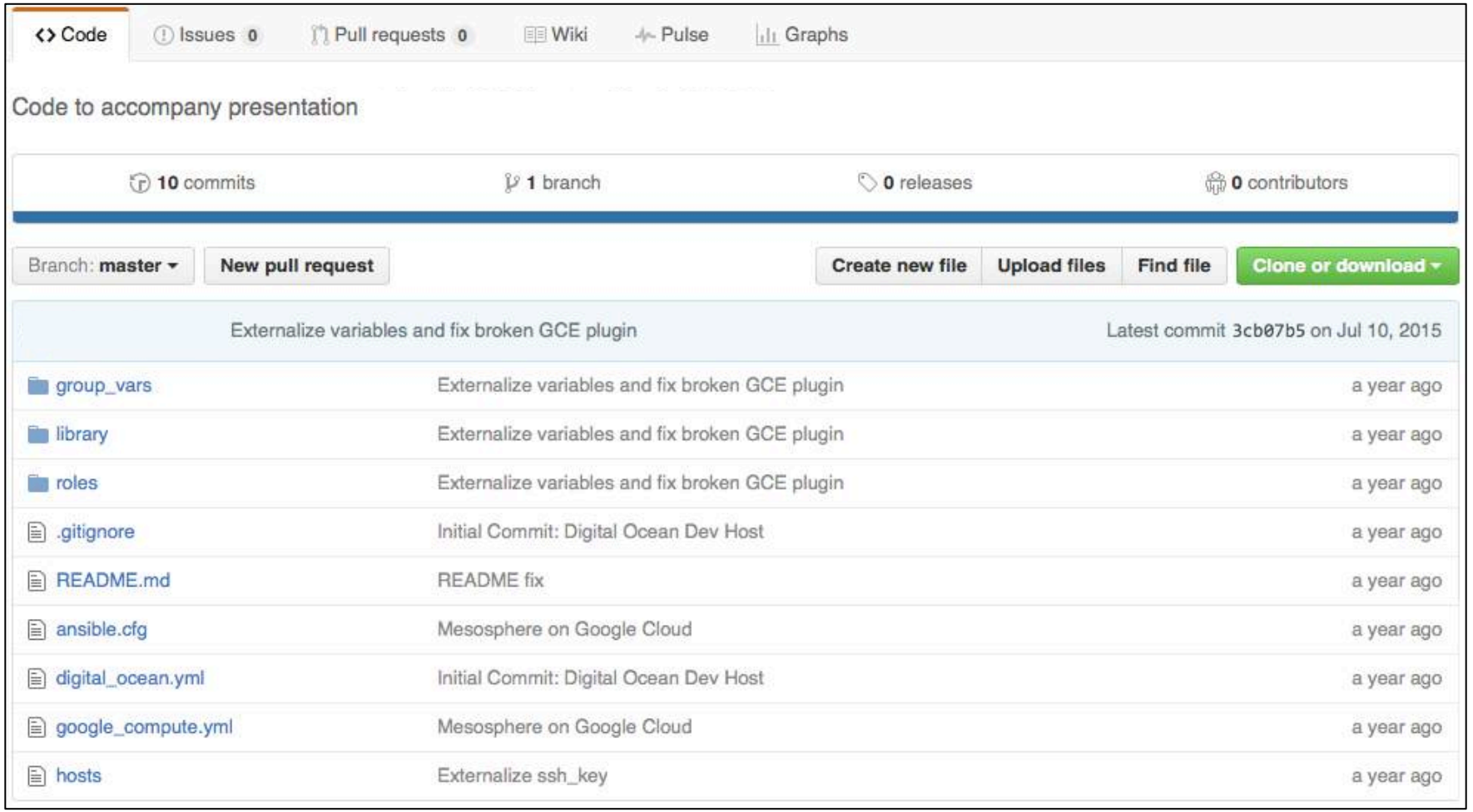
Collaboration tools: Git

- Git is a widely used version control system for software development
  - Now approved at the Department of Commerce
- The tool was launched in 2005 and has become a standard among software developers
- Git's strengths are:
  1. Speed
  2. Data integrity
  3. Distributed workflows (divide & conquer)
  4. Complete history with version tracking
- Allows the entire team to work together

Collaboration tools: GitHub

- GitHub is a free, web-based Git repository with over 14 million users
  - Largest host of source code in the world
- Multiple developers can work on a software product and have access to version control capability
  - Workflow
  - Bug tracking
  - Task management
- GitHub provides a user-friendly web-based interface
  - Can be accessed from the command line for easy workflow
- Free plans for open-source projects
- Paid private repositories

Collaboration tools: GitHub



Is open source secure?

- New libraries / function for open-source tools need to be downloaded from central repositories
- R, Python and other open-source tools have robust communities managed by sophisticated administrators to ensure integrity of the software
- Any organization may be vulnerable to cyber attacks
- Libraries of functions for Python and R may need to be pre-screened by IT
- Once loaded to a central in-house repository, libraries can be safely used by anyone in the organization

Data science tools: key questions

1. Which steps are required in the data pipeline from ingestion to analysis?
2. Which technologies are available for working with data at various stages of the data pipeline?
3. How do different tools and technologies for working with data compare in their functionality, strengths and weaknesses?
4. Do you have staff who can be trained or know how to use particular tools?
5. Do you have budget constraints you need to be mindful of?
6. Is it on the approved software list for the agency?

Remember!

- Many tools can achieve the same results – the key to selecting the right ones are dependent on the people on your team and the projects you need to complete

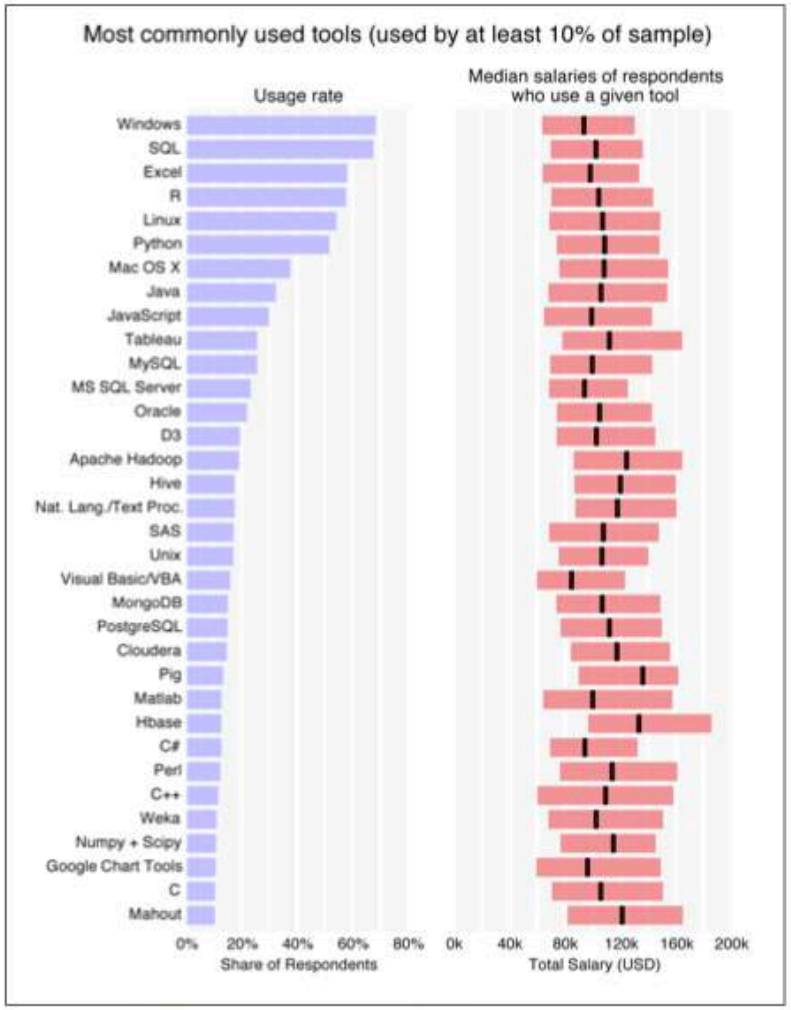


Figure 1-10. Most commonly used tools

Activity: evaluate current tools!

- Turn to your participant guide to the Data science tools page to evaluate your current tools and staff capacity to use them
- You'll list the tools you use for different data uses, and then identify the key people who use the tools, as well as the cost
- Then, assess your current setup – are there tools you are paying for that you're not using? Are there other tools you'd like to start using? Discuss with your group

Activity time: 15 - 20 minutes

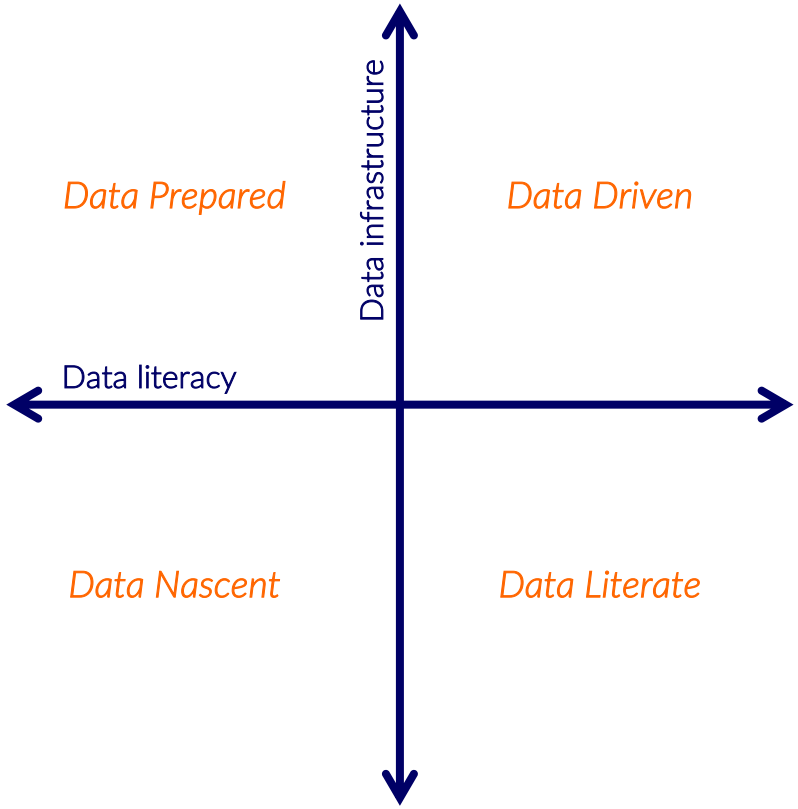
Outline for today

1. Data governance
2. Tools and technology
3. Building a data-driven culture

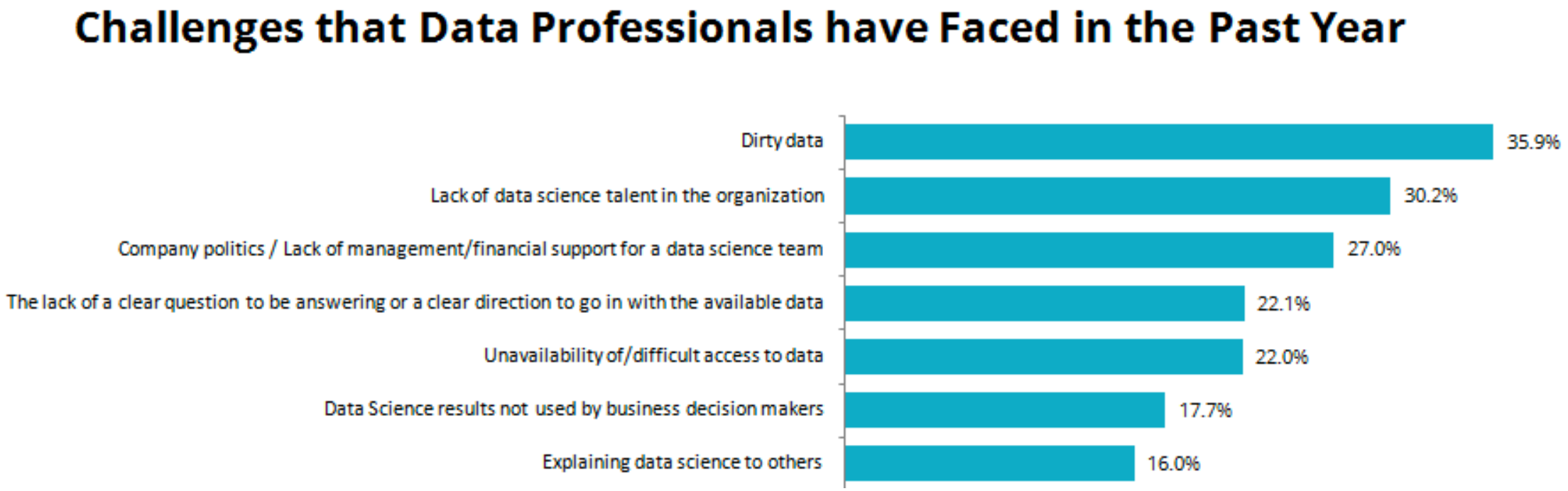
What is a data-driven culture?

- A data-driven culture incorporates data and analysis into its business decisions, systems, and processes

- Being data-driven should be the *default*, and can be separated into two main categories: data literacy and data infrastructure



Why do you need one?



The top responses all relate to a lack of a data driven culture and data governance!

Data are from the Kaggle 2017 The State of Data Science and Machine Learning study. You can learn more about the study and download the data here: https://www.kaggle.com/surveys/2017. Respondents were asked, "At work, which barriers or challenges have you faced this past year? (Select all that apply)." A total of 10153 respondents were asked this questions.

Data infrastructure

- Data infrastructure metrics include:

DATA ACCESS

Can staff access data easily and in a timely manner?

DATA STORAGE

Is the data stored securely with a backup?

DATA COLLECTION

Is data collected in a timely and clean way?

Data literacy

- Data literacy metrics include:

DATA LEADERSHIP

Do executives champion data usage?

DATA GOVERNANCE

Are staff aware of data standards and practices?

DATA KNOWLEDGE

Does staff understand how to ask questions of data?

Building awareness

- Step 1: find a champion (or be the champion!)

Building awareness

- Step 1: find a champion (or be the champion!)

- Step 2: identify a successful analytics project / team, and highlight their success through a newsletter, event, or lunch and learn

Building awareness

- Step 1: find a champion (or be the champion!)
- Step 2: identify a successful analytics project / team, and highlight their success through a newsletter, event, or lunch and learn
- Step 3: once there is more interest, offer additional data trainings (like this one!) to develop a common data vocabulary and empower staff

Building awareness

- Step 1: find a champion (or be the champion!)
- Step 2: identify a successful analytics project / team, and highlight their success through a newsletter, event, or lunch and learn
- Step 3: once there is more interest, offer additional data trainings (like this one!) to develop a common data vocabulary and empower staff
- Step 4: build upon the community of practice that will develop from the trainings to bring wider awareness and more buy-in from executives and managers across the organization

Make it routine

- Lead by example! How can you adjust your practices now to reflect a data driven mindset?

1. Ask for the metrics / analysis summary behind conclusions and reports
2. Demonstrate data-driven decision making during meetings
3. Highlight data-driven team members or successful analyses

Example: data competitions

- The Inter-American Development Bank (IDB) hosted an internal data competition to try to build a more accurate financial forecast model
- They provided the dataset, guidelines, and timeline
- The winning team developed a model that was more accurate than what they were using

Outcomes: data competitions

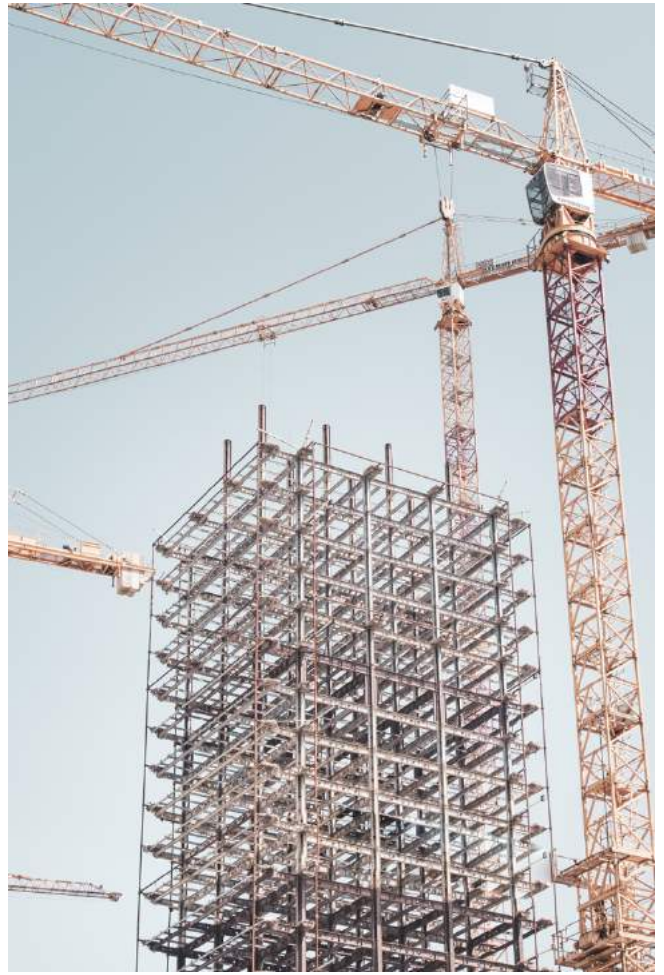
1. The organization found a better way to predict financial futures of countries they were working with

2. They identified top data talent within the IDB, as well as many others who expressed interest in building their skills

3. They boosted morale through recognizing the hard work of all participants and demonstrating how the organization valued data driven strategies

Steps to building a data competition

1. Set your goals and data problem clearly
2. Prepare dataset and provide support networks to participants
3. Decide on the prize
4. Set up the rules and name the judges
5. Set the criteria for evaluating the submissions
6. Develop a marketing plan
7. Measure and analyze the result of your competition
8. Celebrate!



Example: community of practice

- The Department of Health and Human Services (HHS) held an 8-week data science program - the HHS CoLab
- The participants shared their code on Github to update their projects
- Even after the program was over, a cohort of graduates continued to meet to problem solve data challenges

Outcomes: community of practice

1. The participants learned the most efficient way to share their projects and knowledge internally

2. They problem-solved similar issues together because they faced the same challenges within the agency

3. They saved the agency over half a million dollars through increased efficiencies and was able to provide better services to their constituents

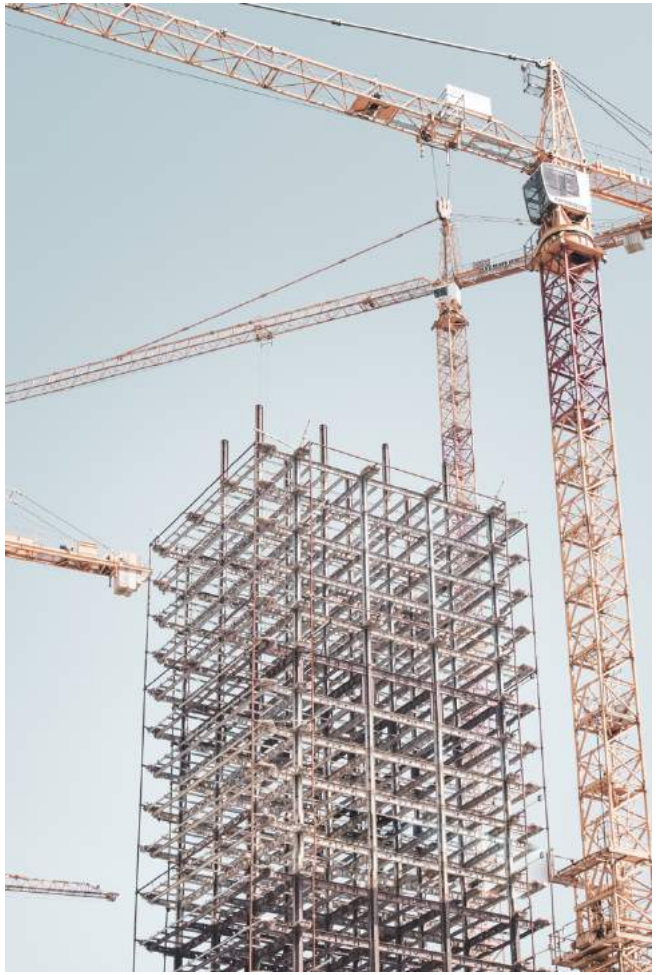
Steps to building a community of practice

1. Host an internal code sharing server (SharePoint, GitHub, BitBucket)

2. Identify a data-driven team and ask them to start sharing their code on the server

3. Highlight some of the code and projects on the shared server to garner interest

4. Champion a “data practitioner meetup”, where individuals can meet to discuss how to solve their data problems



What else can you do?

1. Bring in external / internal experts for “Lunch and learns”
2. Start a regular newsletter that highlights new tools / data successes / data teams, etc
3. Attend and send team members / colleagues to data conferences
4. Develop an internal Data Academy (like this one!) to make data education an integral part of the community
5. Develop and implement data governance principles and guidelines

Steps to making better decisions

Steps	Descriptions
Organizational analysis	1. Understand how your organization uses data and create a plan to improve your processes with data-driven decision making
Executive data science training	2. Train managers and executives to understand the benefits and capabilities of enhanced data analysis and data sources
Technical training	3. Train analysts on the skills necessary to leverage existing and new data resources
Infrastructure implementation	4. Setup the requisite infrastructure to source and analyze a variety of data types
Analytical insights	5. Deploy advanced machine learning algorithms to identify novel insights
Enhanced decision making	6. Leverage analytical insights as part of every day decision making to increase the efficiency of the organization

Remember!

- Give people the opportunity to fail
- This is an iterative process – it takes several tries to get it right
- Be flexible

Activity: plan a data event

- Turn to your participant guide to the **Data event plan** page to develop an event of your own

- You'll list the plan out the objective you have, and think through the resources you'll need in order to make the event happen

- Discuss your plan with your group – is anyone else planning something similar? Would you be able to collaborate to make it an inter-agency event?

Activity time: 15 - 20 minutes

Recap

- Today, we learned how to:

  1. Build up data standards and common vocabulary
  2. Describe common data science tools
  3. Implement events to improve data awareness and data-driven culture

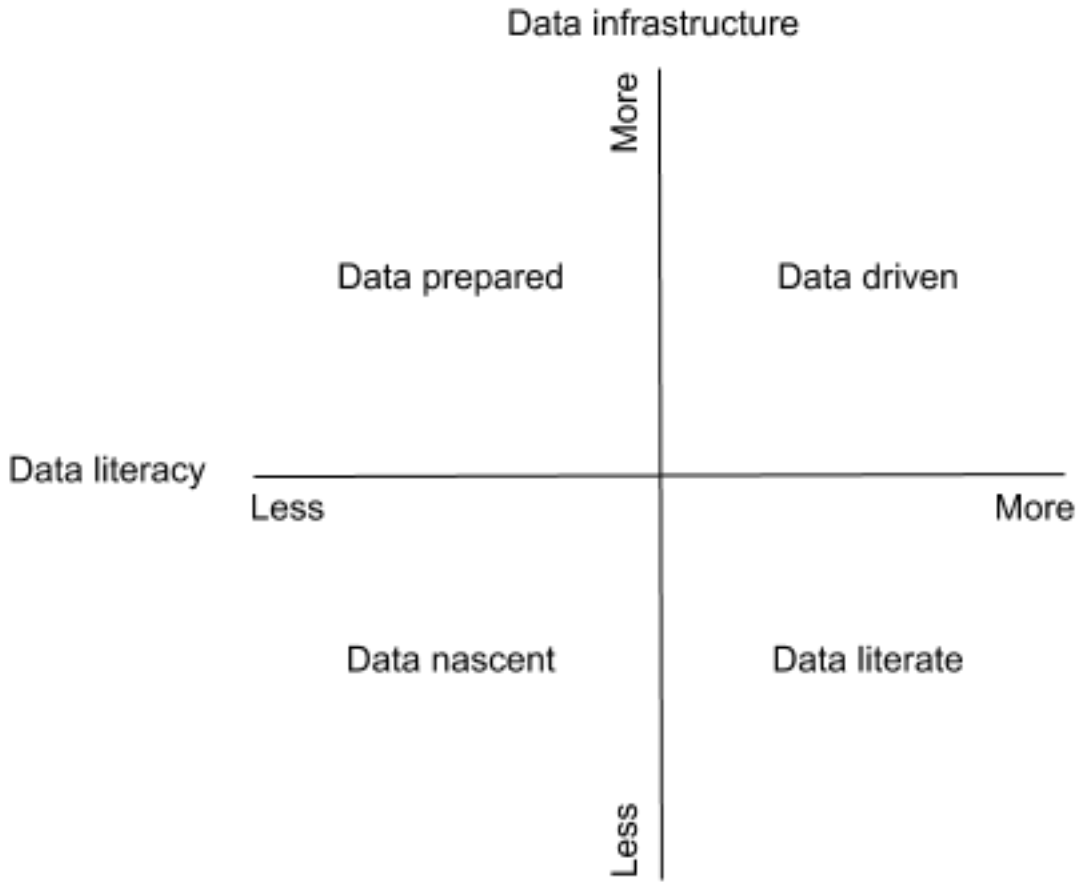
Recap: data-driven metrics

Data infrastructure metrics:

1. Data access
2. Data storage
3. Data collection

Data literacy metrics:

1. Data leadership
2. Data governance
3. Staff knowledge of data



Which metrics do you want to focus on first?