DATA SOCIETY®

The premiere data science training for professionals

Activation activity

- Present the articles that you read
- How is data analytics being applied in the Air Force and DoD? What are some of the challenges that were identified?

Activity time: 15 - 20 minutes

Outline for today

1. Data storytelling
2. Data ethics frameworks
3. Open data sources
4. Managing data science projects / teams

Why data storytelling?

Regardless of your role, you are a communicator first and foremost. Data is worthless if you don't communicate it properly. Great analysis must also have great storytelling.

Never assume that the results will speak for themselves. Stories always trump statistics alone, and communicating insights from data clearly, requires a structured approach.

Let's look at two frameworks you can use

Data storytelling - George Roumeliotis

- Current Airbnb data science manager and was head of a data science group at Intuit
- For projects, he developed a business story framework for communicating about each analysis:

1. My understanding of the business problem
2. How I will measure the business impact
3. What data is available
4. The initial solution hypothesis
5. The solution
6. The business impact of the solution

Note: he does not include details on statistical methods used, regression coefficients, etc. Focus on results and implications, start with what your audience needs to know, and add the methods to the appendix.

Step 1: Preparing key metrics

- Narrative framework graphic organizer

1. Before you begin, look at your data and determine:
- What is the message and who is your audience (what data will be most relevant to them)
- Identify your question or problem statement
- Know what your data is saying and choose the appropriate visualizations

2. Gain attention → 3. Content / body → 4. Call out or Call to Action → 5. Conclusion - keep it brief with key takeaways

Step 2: gain attention

- Headline / heading
- Pose a problem / ask a question
- Tell a story within your story

Step 3: content / body

- Stimulate prior knowledge
- Present content and appropriate data visuals
- Have planned questions and key points
- Compare, contrast and connect
- Add supporting evidences

Step 4: call out / call to action

- Propose a solution
- Describe benefits of your solution

Step 5: conclusion

1. Before you begin, look at your data and determine:
- What is the message and who is your audience (what data will be most relevant to them)
- Identify your question or problem statement
- Know what your data is saying and choose the appropriate visualizations

2. Gain attention → 3. Content / body → 4. Call out or Call to Action → 5. Conclusion - keep it brief with key takeaways

5 tips for your data presentation

1. Explain what the data axes mean (this is a part of orienting your audience)
2. Explain what the value of the data points mean
3. Explain the level of detail presented
4. Explain what data points they should be focusing on
5. As noted in the previous graphic organizer, always end with a key takeaway based on the visualization(s)

Keep in mind: 508 compliance

- Any outward facing data sharing or visualizations shared from a government agency must follow 508 Compliance requirements:
  - Don't rely on color as a differentiating factor
  - Use contextual and descriptive text for links and buttons
  - Use text, not images, in titles and navigational elements
  - Include text descriptions for all assets (variables, relationships, axes and CODE)

Keep in mind: 508 compliance

- Software considerations
  - Power BI has Tab menu, keyboard shortcuts and audio accessibility
  - STATA has been approved 508 compliant for government use
  - RStudio works with screen readers in most areas of the interface (with two key exceptions: the console and the editor)
  - Tableau has an “embedded view” feature that allows for modification and conforms to Web Content Accessibility guidelines (WCAG 2.0 AA)

Data storytelling example

- 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four (narrative framework)
- https://www.youtube.com/watch?v=jbkSRLYSojo

Recap

- What data storytelling elements did you notice?
- Was this a good example of data visualization and storytelling? Why or why not

Outline for today

1. Data storytelling
2. Data ethics frameworks
3. Open data sources
4. Managing data science projects / teams

What is data ethics?

Data ethics is a newer branch of ethics that studies and evaluates moral problems related to:

- Data (including generation, recording, curation, processing, dissemination, sharing and use)
- Algorithms (including artificial intelligence, artificial agents, machine learning and robots)
- Corresponding practices (including responsible innovation, programming, hacking and professional codes)

Source: University of Oxford

Why data ethics?

- Data science has huge opportunities, but those opportunities are accompanied by complex data ethical challenges
  - To formulate and support morally good solutions (e.g. right conducts or right values)
  - To maximize the value of data science for our societies, for all of us and for our environments

The best single thing you can do to further data ethics is to talk about data ethics!

Source: University of Oxford

Data ethics prep questions

- What data ethics guidelines do you currently have in place?
- What are some biases that you need to be aware of?
- Have you experienced bad visualizations or biases in your workplace?

Let's review some ethics frameworks & guidelines!

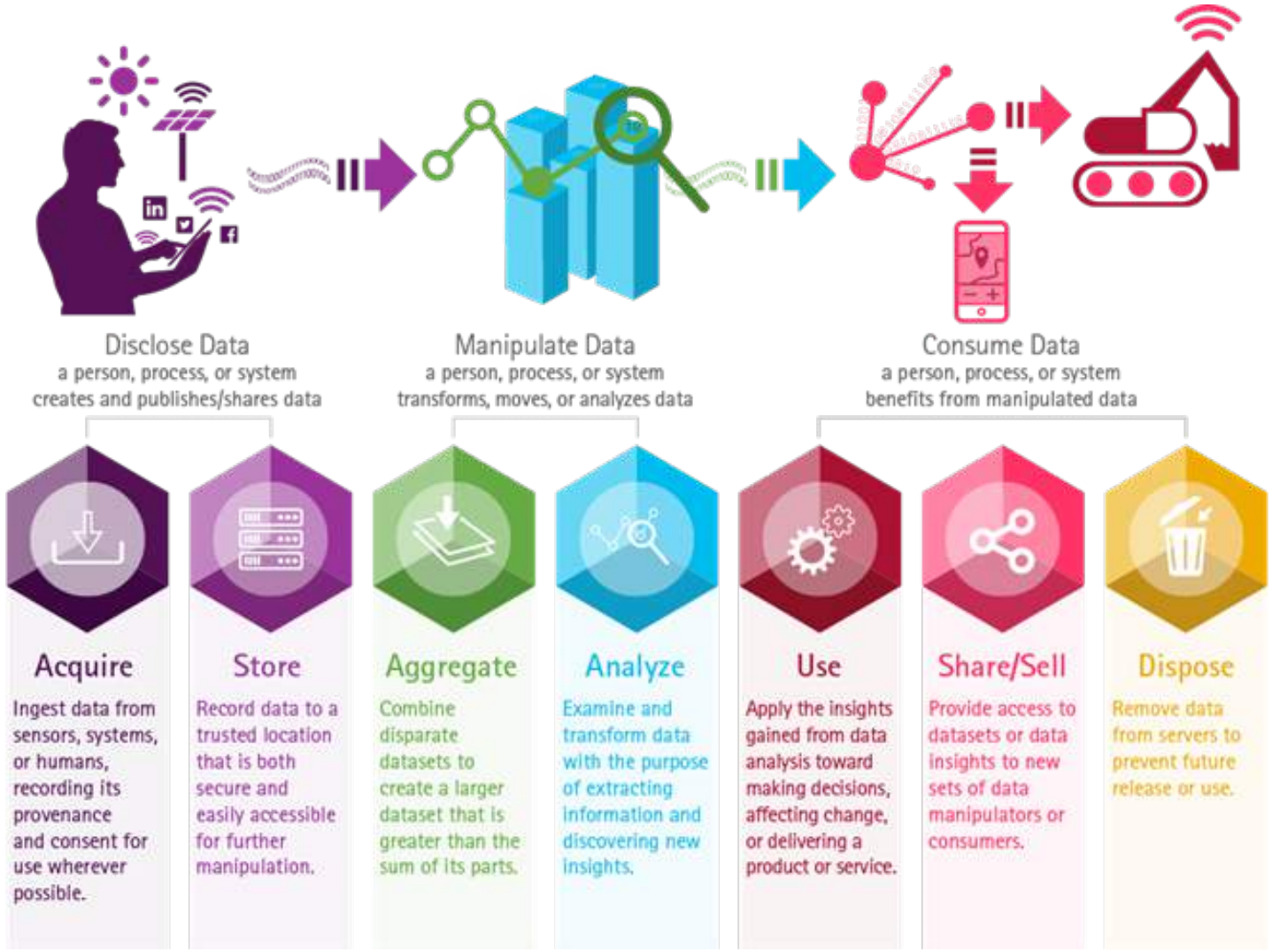
Guidelines – the 5 C's

- Five framing guidelines (the Five C's) for building data products:

1. Consent
2. Clarity
3. Consistency
4. Control
5. Consequences

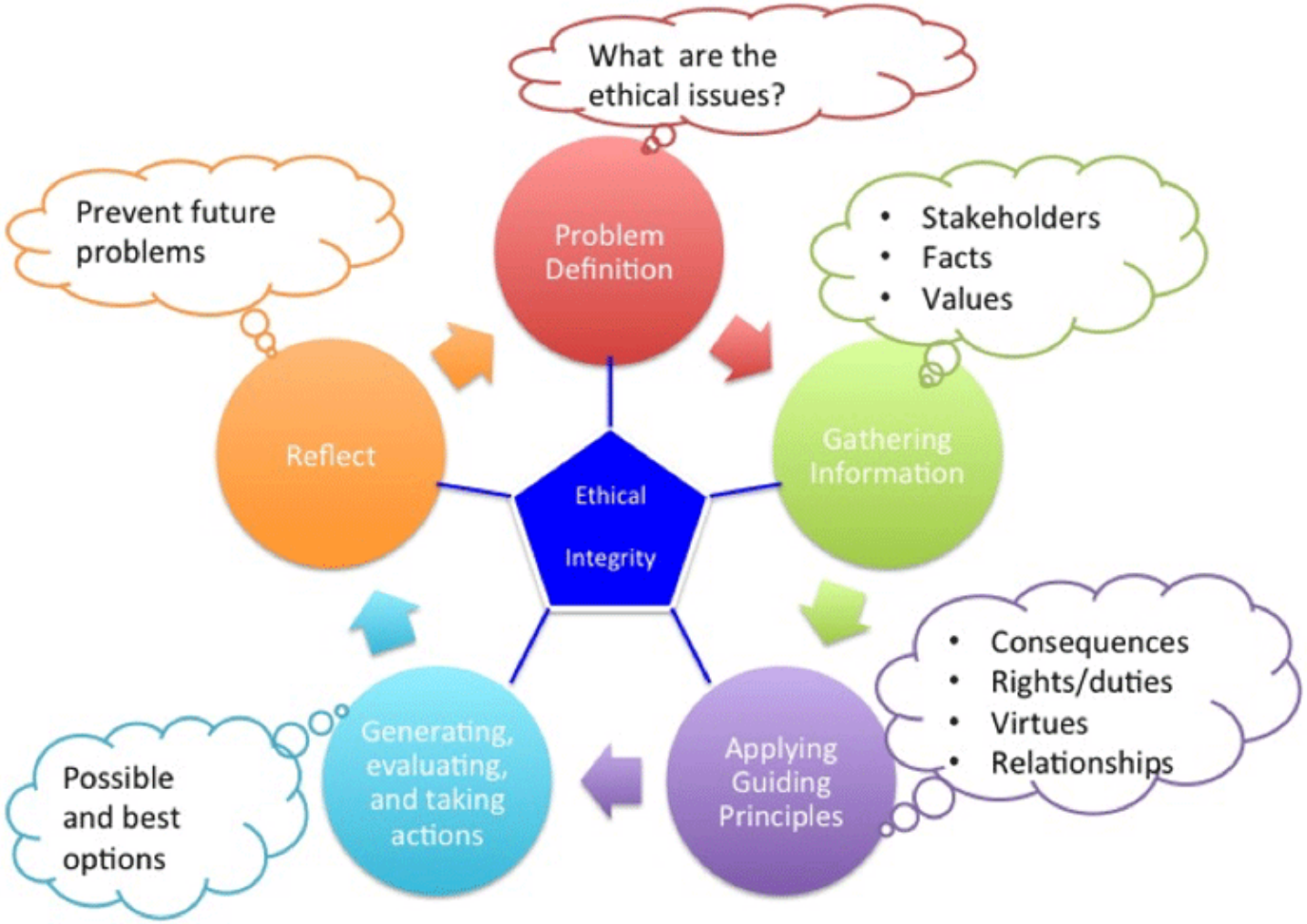
https://www.oreilly.com/library/view/ethics-and-data/9781492043898/

Accenture data ethics



Source: Accenture

Penn State framework



UK Govt. Data Ethics Framework

- Developed by the UK government, the framework is meant to guide companies and teams before they embark on a data project

1. Start with clear user need and public benefit
2. Be aware of relevant legislation and codes of practice
3. Use data that is proportionate to the user need
4. Understand the limitations of the data
5. Ensure robust practices and work within your skillset
6. Make your work transparent and be accountable
7. Embed data use responsibly

https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework

GDPR: new regulations

- Regulations developed in Europe to help individuals control their data

- It includes:
  - Breach notification
  - Right to access
  - Right to be forgotten
  - Data portability
  - Privacy by Design

https://eugdpr.org/

Data Society guidelines

1. Ownership: Who owns the data? Do you have the right to collect the data?
   - Google does not sell your data, but uses it to make money
2. History: How long can you store the data?
   - The legal system has long maintained historical data
3. Privacy: Who controls access to the data?
   - Those who have the ability have the responsibility
4. Uses: What kinds of inferences can you make?
5. Math (is dumb?): How do you prevent machine learning algorithms from learning the biases of the past? **Understanding how the math works is imperative for ethical data science!**

Data ethics: guiding light

"It's not hard to make decisions when you know what your values are."

– *Roy Disney (nephew of Walt Disney)*

Activity: evaluating ethics

- Turn to your participant guide to the Evaluating ethics page and read the accompanying article
- Use the guiding questions to jot down notes and ideas that you have about the ethical implications for building algorithms to detect terrorism
- Then, discuss your ideas with your group, and determine whether or not this solution follows the ethical guidelines discussed in the training

Activity time: 15 - 20 minutes

Outline for today

1. Data storytelling
2. Data ethics frameworks
3. Open data sources
4. Managing data science projects / teams

What is open data?

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

- Open Data International

(https://blog.okfn.org/2013/10/03/defining-open-data/)

What are the key features?

- In order for a dataset to be considered open, it has these 3 features:

Free access

Reuse and redistribution

Available to the public

What are some of its benefits?

- While it may seem daunting to publish internal data, there are tangible benefits:

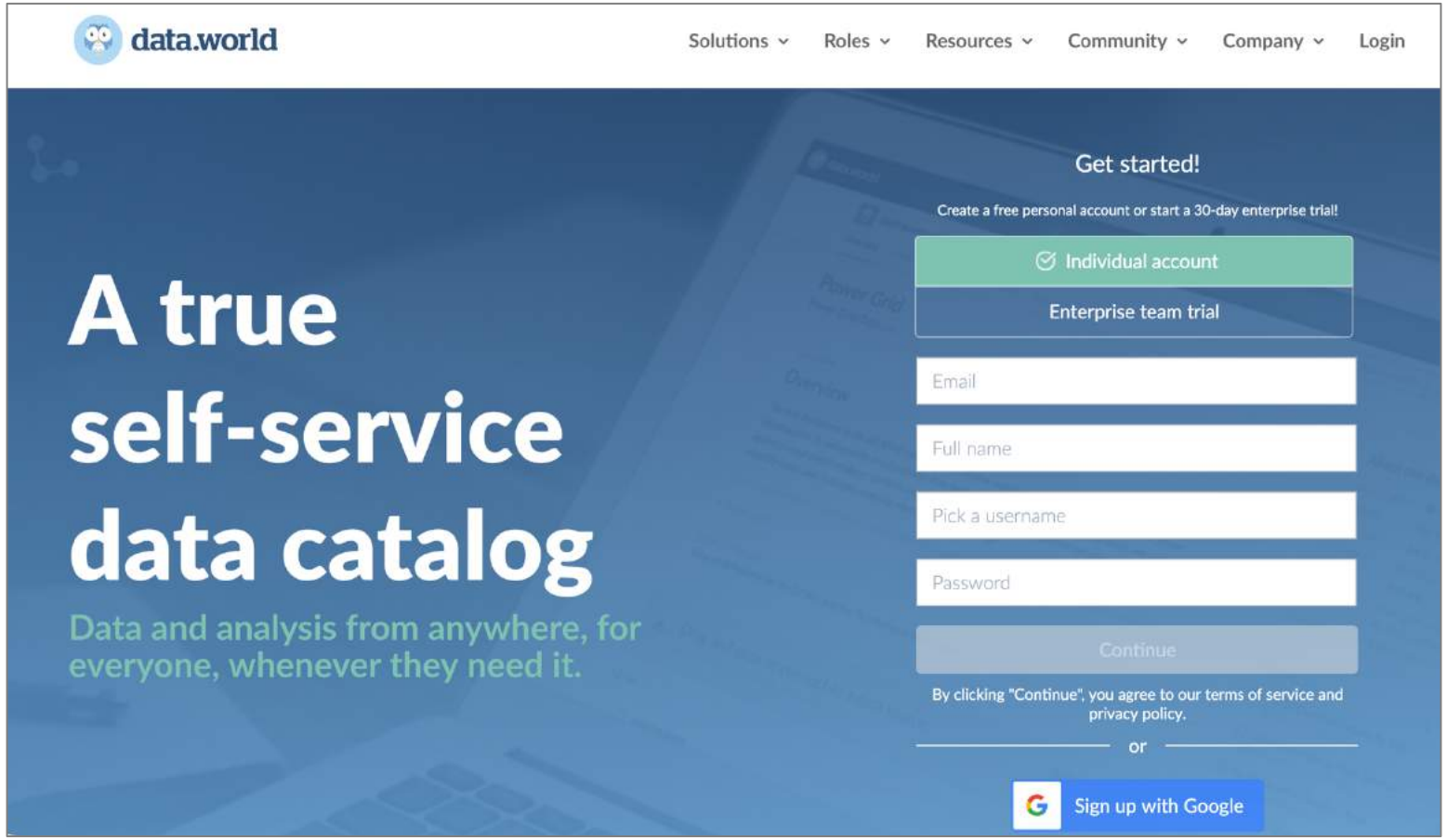
Government transparency

Public participation

New insights

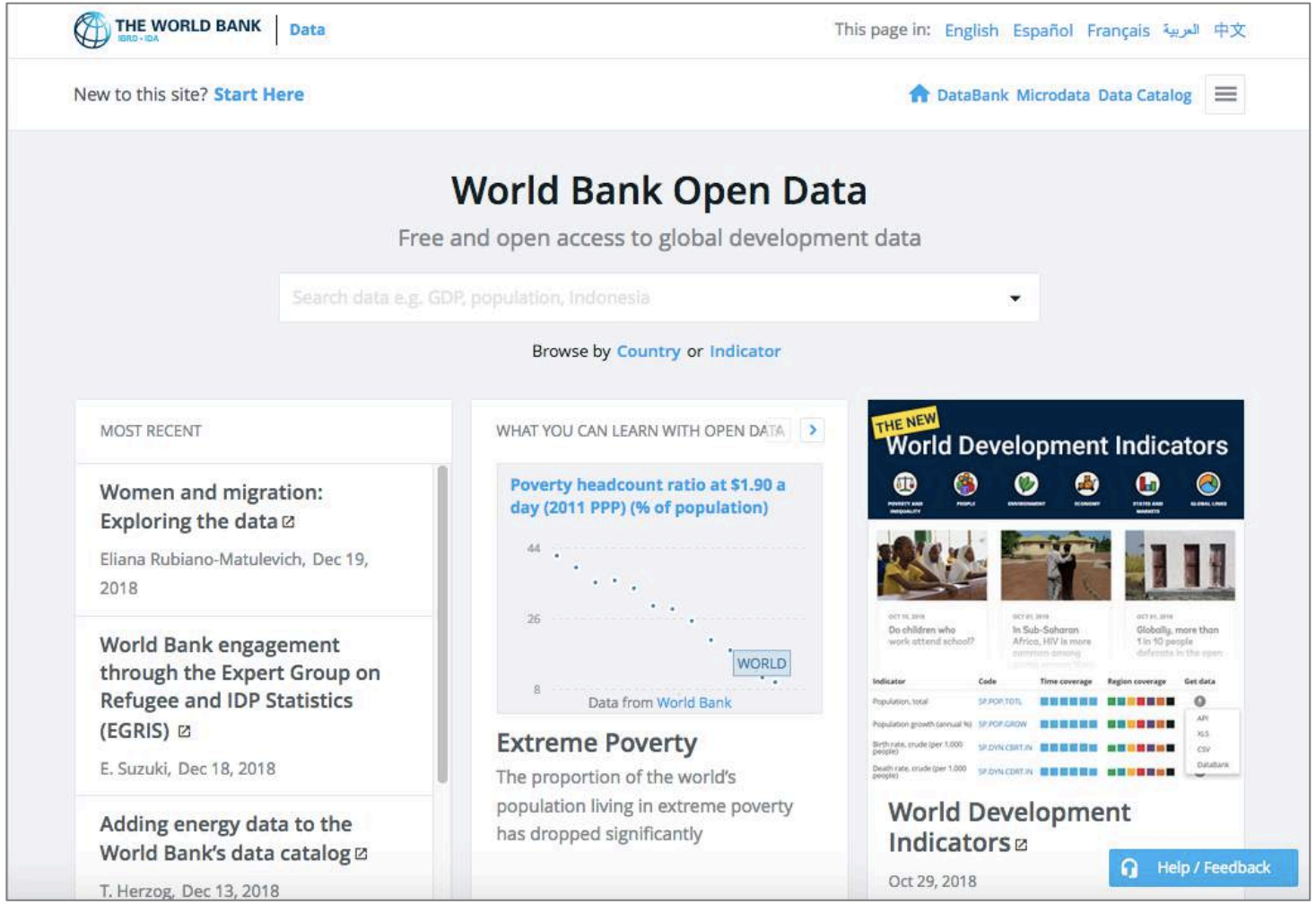
Open data sources

- https://data.world



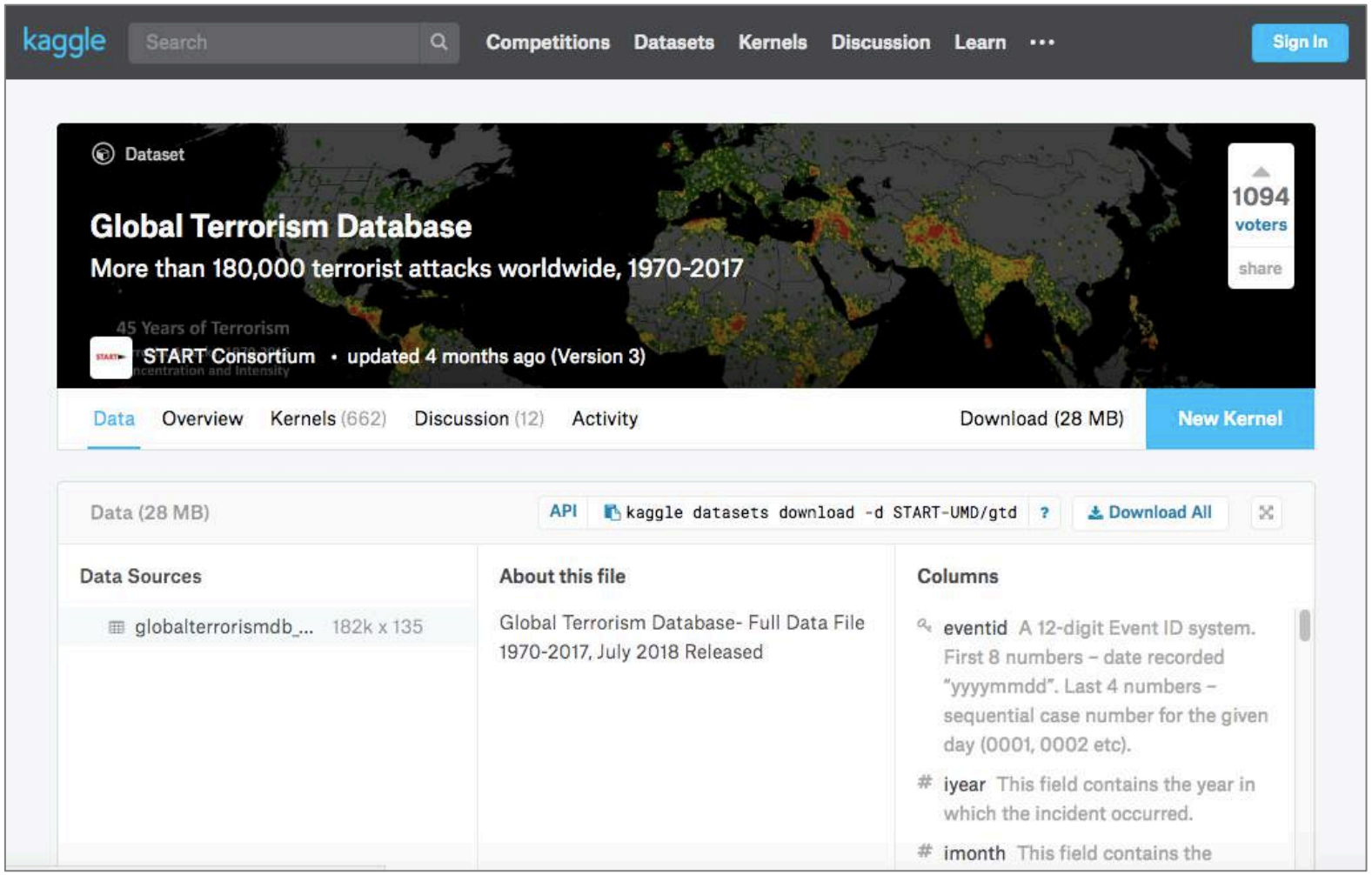
Open data sources

- World Bank



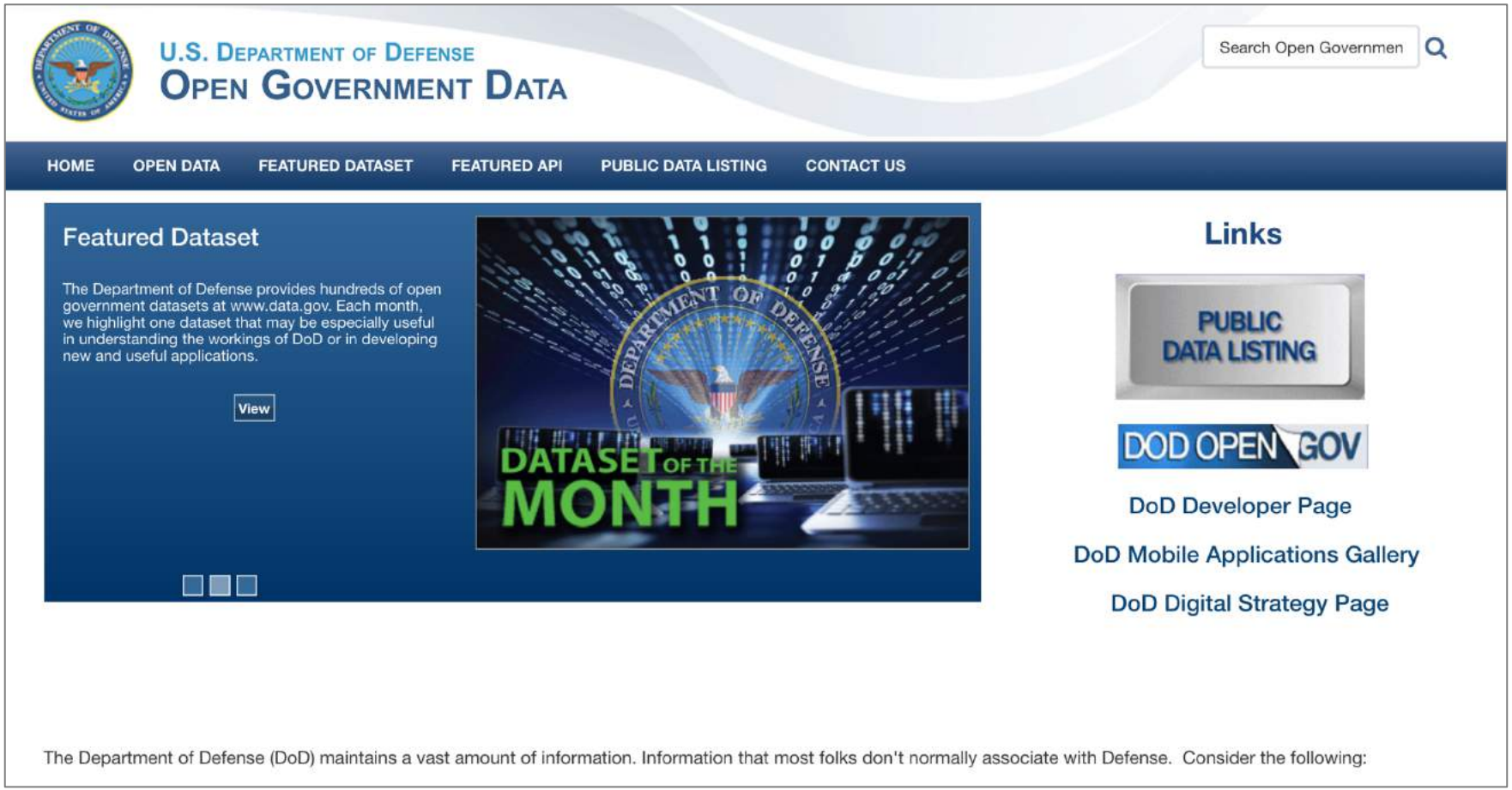
Open data sources

- Global Terrorism Data



Open data sources

- https://data.defense.gov/



Other data sources

Social Media APIs – Twitter, Telegram, Facebook, and others all offer APIs to developers which can also be used to scrape real-time data

News APIs – News aggregators such as Google news make news article metadata available

GDELT – A Google Jigsaw project which analyzes thousands of data sources, extracting key features such as location, actors, and type of incident. Updated every 15 minutes

ACLED – A curated dataset of dates, actors, types of violence, locations, and fatalities of all reported political violence and protest events across Africa, South Asia, South East Asia, the Middle East, Europe, and Latin America

Be careful!

- While open data is key to increasing public awareness and participation, you need to run through a series of checks and permissions to make sure:

1. You are not releasing any Personally Identifiable Information
2. You are not releasing any security-related data
3. The data you are releasing is accurate and can be used with confidence

Case study: fitness trackers

U.S. soldiers are revealing sensitive dangerous information by jogging

GPS tracking company Strava published an interactive map in Nov. 2017, showing where people have used fitness tracking devices. (Patrick Martin/The Washington Post)

https://www.washingtonpost.com/world/a-map-showing-the-users-of-fitness-devices-lets-the-world-see-where-us-soldiers-are-and-what-they-are-doing/2018/01/28/86915662-0441-11e8-aa61-f3391373867e_story.html

Activity: data ethics

- Within your group, answer the following questions:

  1. What are possible ethical considerations?

  2. What are possible additional open data sources?

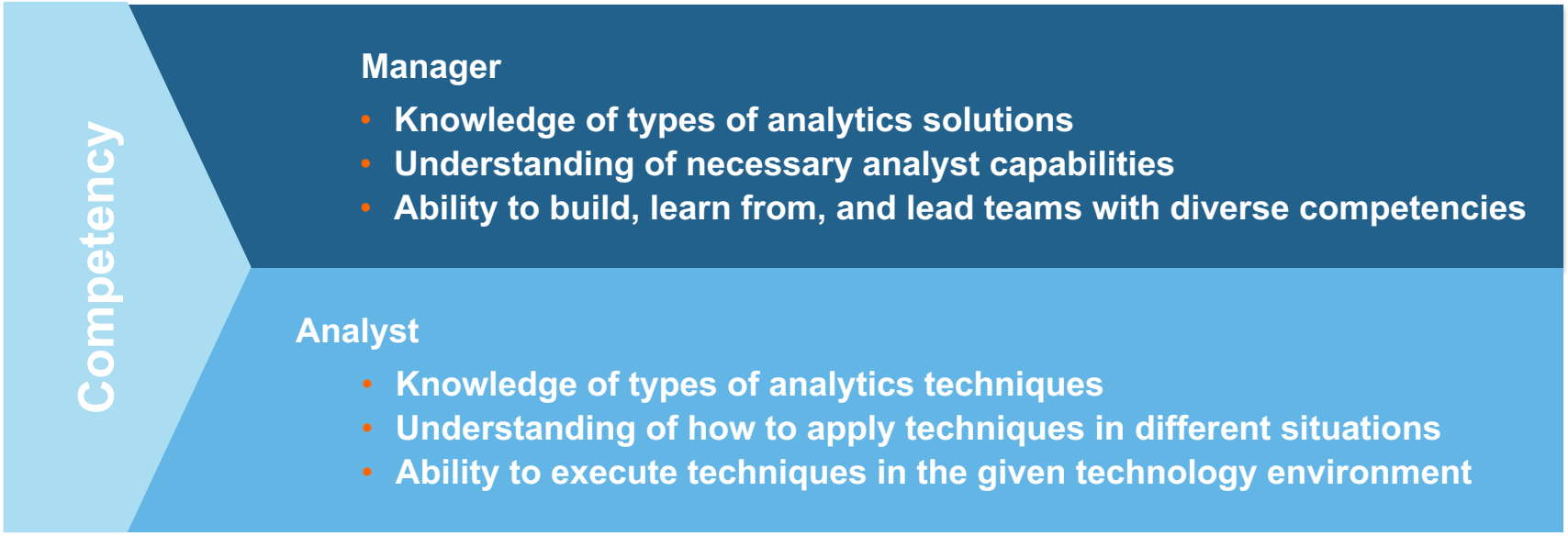
Activity time: 10 - 15 minutes

Outline for today

1. Data storytelling
2. Data ethics frameworks
3. Open data sources
4. Managing data science projects / teams

Management vs. analyst skill sets

- Managers and analysts play different roles but need to speak the same language



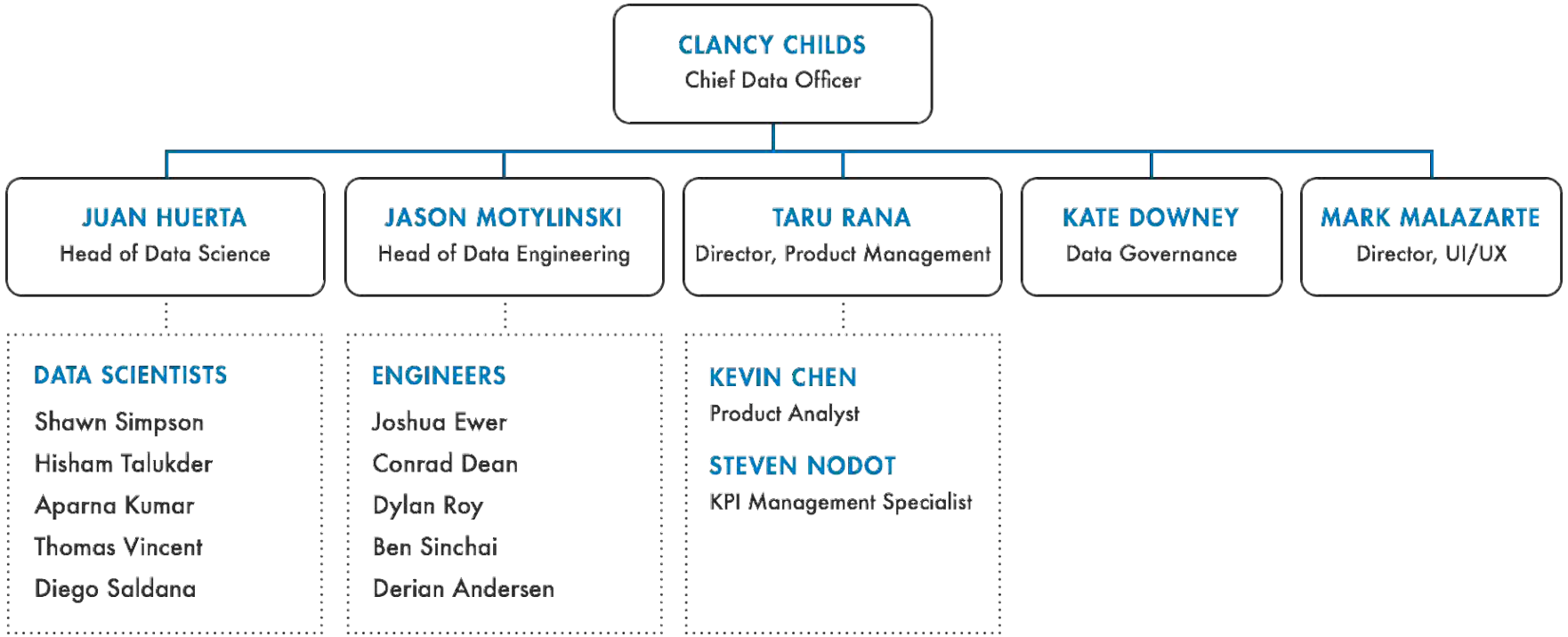
What kind of talent do you need?

- Accountability:
  - Make clear from the beginning for exactly where in the organizational chart the team will be located and who the main stakeholders will be
  - Who is in charge?
    - Chief Technology Officer (CTO)
    - Chief Information Officer (CIO)
    - Chief Data Scientist (CDS)

- Resources:
  - Technical talent in this area does not come at a low price, and is not easy to find!
  - Budgets are surprising ill-prepared
  - Make one wrong hire early and you could be in trouble!

Building your data team

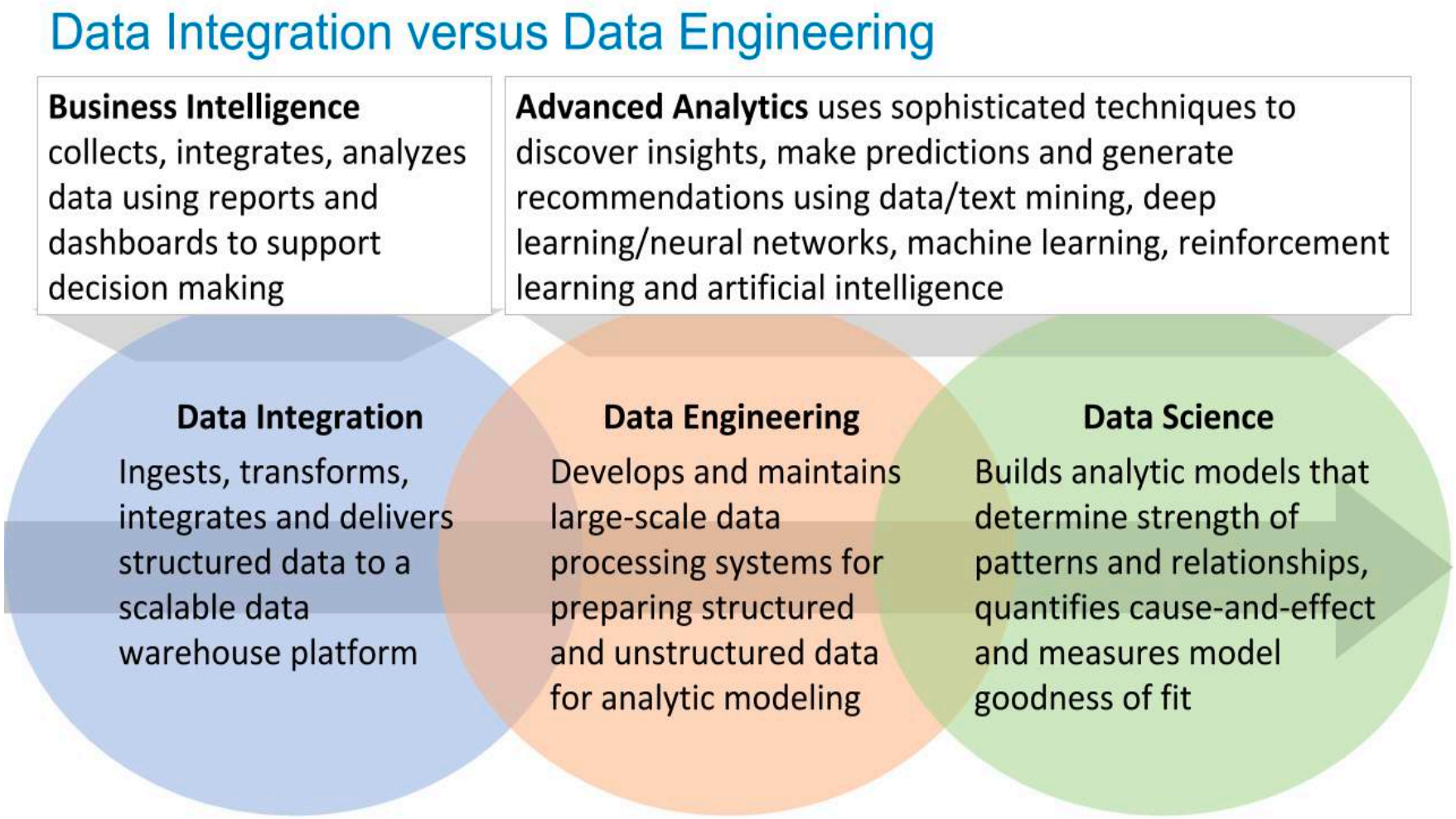
- Plan out your infrastructure before you start hiring – think through an org chart that makes sense for your needs



Sample org chart from Dow Jones

http://markmalazarte.com/projects/newscorp/dsehub/build/index.html

Building your data team



https://www.kdnuggets.com/2018/09/winning-game-plan-building-data-science-team.html

Data Analyst

- An analyst ensures that collected data is relevant and exhaustive while also interpreting the analytics results

- Main role and responsibilities include:
  - Wrangling the data
  - Managing the data
  - Creating basic analyses and visualizations

- Core skills to look out for: SQL, R / Python, Tableau / Power BI

Data Scientist

- A data scientist builds upon the analysts' data work to develop predictive models and complex algorithms

- Main role and responsibilities include:
  - Asking the right questions from the data
  - Building more complex predictive models
  - Interpreting the results critically and communicating them well

- Core skills to look out for: R, Python, Spark, Hadoop

Data Engineer

- A data engineer develops the infrastructure to house the data and maintains the structural components

- Main role and responsibilities:
  - Ensuring data integrity across different data sources
  - Building out additional data warehouses as needed
  - Maintaining data pipelines and access

- Core skills to look out for: AWS, MongoDB, MySQL, Hadoop, C++

Data Science Manager

- A data science manager oversees and directs data science teams and projects and is the bridge between data and non-data people

- Main role and key responsibilities include:
  - Planning out people and resources for projects
  - Communicating results to executives and stakeholders
  - Running the data science teams

- Core skills to look out for: management experience, programming skills (R / Python), strong communication

Setting the team up for success

1. Develop a data infrastructure and support – prioritize IT and data engineers before hiring a data scientist to set up the data pipelines and frameworks

2. Hire managers who understand the complexities of data analysis – data scientists need the guidance and understanding of their managers

3. Develop strong executive support – executives will be able to prioritize and allocate resources to infrastructure and data culture

4. Increase data literacy – the more people who know how to do basic data queries and check data quality, the better it is for data scientists

Do you need to hire?

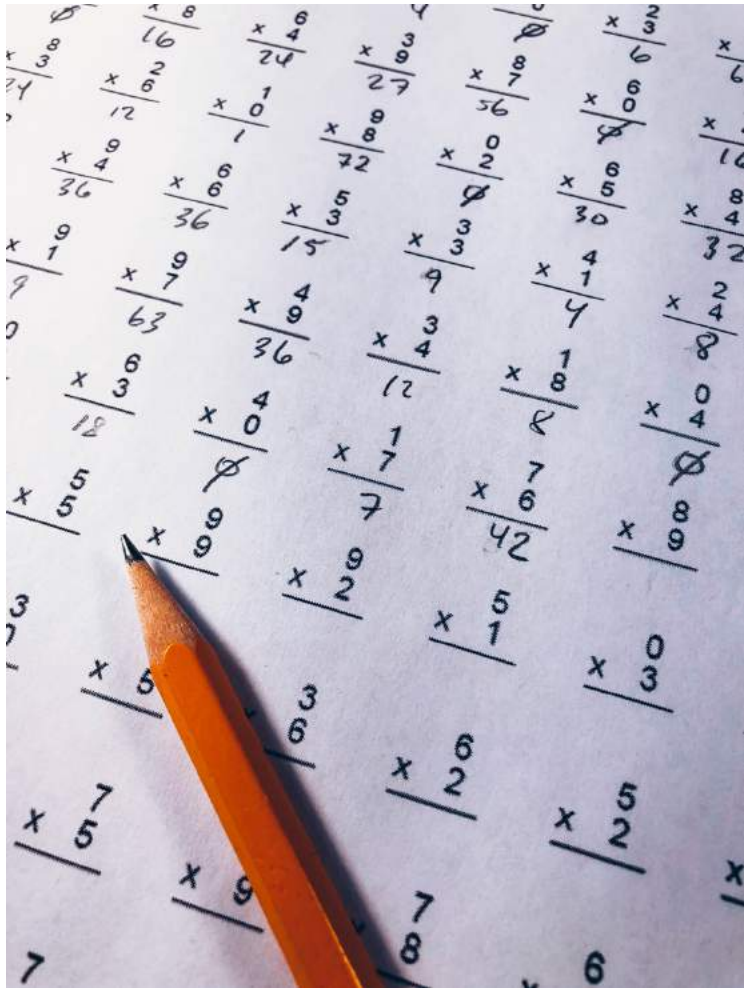
Build when you have the resources that provides what you need

Partner, when you don't know enough to make a decision and need help

Evaluating your current staff

- Do you even need to hire? Training up staff internally has several advantages:
  - They've proven themselves to be reliable
  - They already have institutional knowledge
  - Takes the same amount of time to train them as it would to hire someone
  - Demonstrates an investment in your employees

- There are many different ways to evaluate your staff, such as a technical questionnaire or data competition



Renting vs. buying

Data scientists as contractors

- Strengths
  - Flexible cost structure can adapt to changing budgets
  - Easy to change staff if people don't work out
  - Quickly add staff with new skills
- Weaknesses
  - Internal know-how is not built up
  - Data science does not become an endemic capability
  - The organization becomes dependent on forces outside of its control

Hiring data scientists

- Strengths
  - Data science becomes an endemic capability – better decision making becomes part of the DNA
  - Internal know-how is developed and sustained – the analytics capability has a strong foundation
- Weaknesses
  - State-of-the-art capabilities may still need to be brought in from the outside ("rented")
  - Organizational challenge: data science must remain impartial to internal dynamics

Crowdsource data science talent

- www.kaggle.com
- Crowd-source your data scientists!
- Cost: as little as $500, as much as $1,000,000
- Over 50,000 registered competitors
- You're in good company:
- Platform for predictive modeling competitions where companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models

Build a team

- You design your analysis and maintain the staff to repeat the process
- Allows you and your processes to grow with the technology and evolve with your needs
- Downside:
  - Stakeholders may grow impatient with longer execution times
  - It can be time consuming and expensive

Sources of candidates

- Meetups
- LinkedIn groups
- Professional associations
- University career offices & data science departments
- Career websites & tools:

Job posting mistakes

1. Posting every tool under the sun
   - Example: "applicant must be fluent in R, Python, Java, Excel, PowerBI, C++, AWS, Hadoop, Hive, Tableau…"
   - Solution: be specific! Only list the key tools that the position will need to know – keep in mind that programmers can learn new languages fairly quickly

2. Posting basic responsibilities that wouldn't fall under the position
   - Example: "Conducting exploratory analysis and communicate results, including descriptive statistics, data visualizations, and diagnostics, to project teams"
   - Solution: make sure that the responsibilities match the job title

Key programming resume attributes

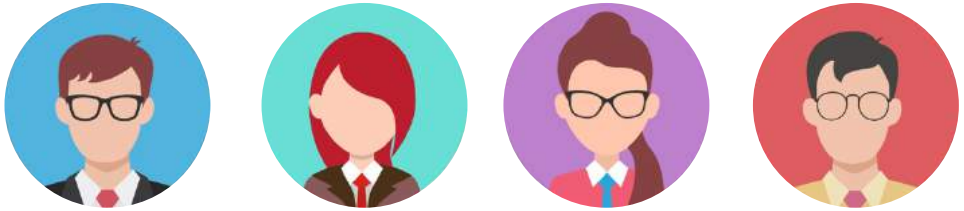
1. Experience optimizing code for run-time
2. Experience connecting multiple platforms
3. Experience with either back-end or front-end infrastructure
4. Intimate familiarity with a variety of algorithms and their implementation at scale
5. Fluency with a variety of core programming languages
6. Experience working with databases
7. Strong quantitative background

Have they built a working product before? Do they have a portfolio?

The interview process

You are looking for

- Someone who is ever learning and adapting



Inquisitive Storyteller Relentless Meticulous

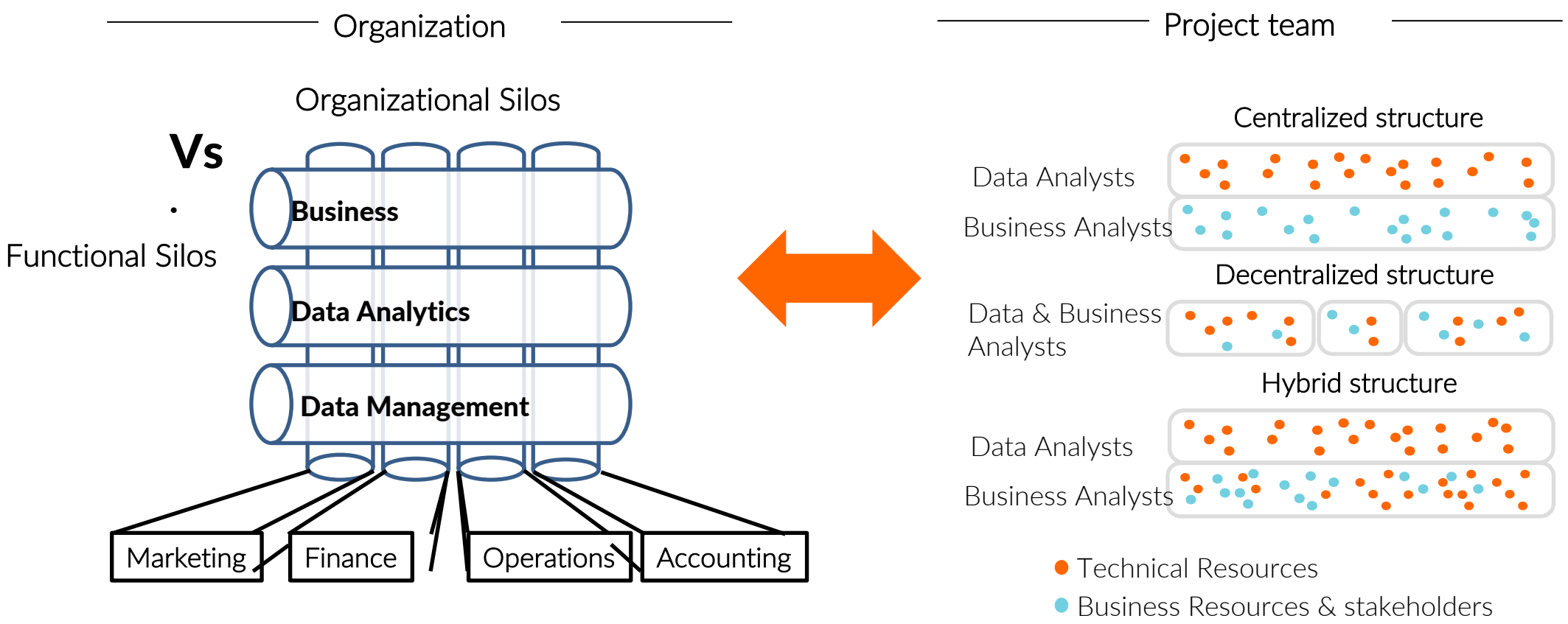
- Is a logical thinker
- Can stay calm under pressure
- Thinks critically
- Can work with technical and non-technical people

You need to check

- Their coding ability through a coding test
  - BUT: don't focus too much on any particular tool, a good programmer can always learn another language quickly
- Ask people to answer open-ended quantitative questions (i.e. How do you measure sentiment?; How do you measure the resilience of a cyber network?)
- Ask them to complete a mini-project and present the results to management
  - These steps can take 3 – 5 rounds

Project team structure: silos

- Like the stratification in an organization, data project teams also tend to have divided structures that can impede productivity

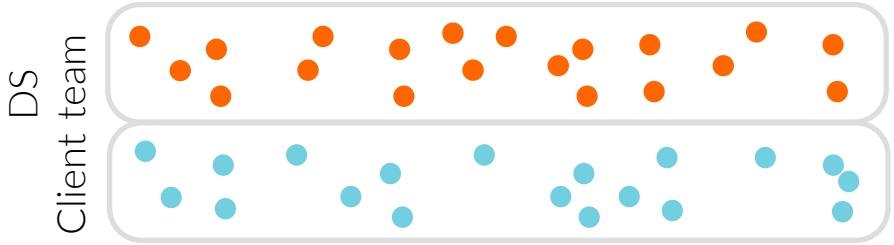


Data science team structure

- DS team
- Client

DS team has met the needs of functional areas!

Centralized structure

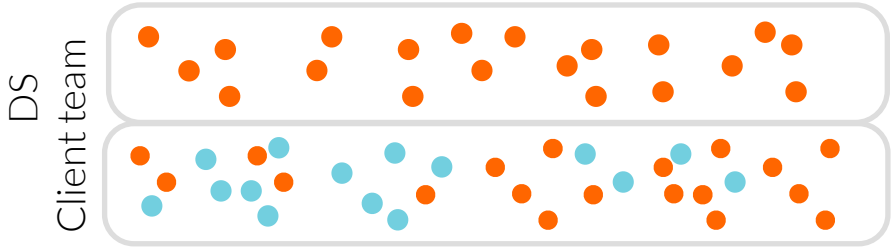


+ 1 Standardized processes
+ 1 Strategic goals/vision met
- 1 Client goals not met

Decentralized structure

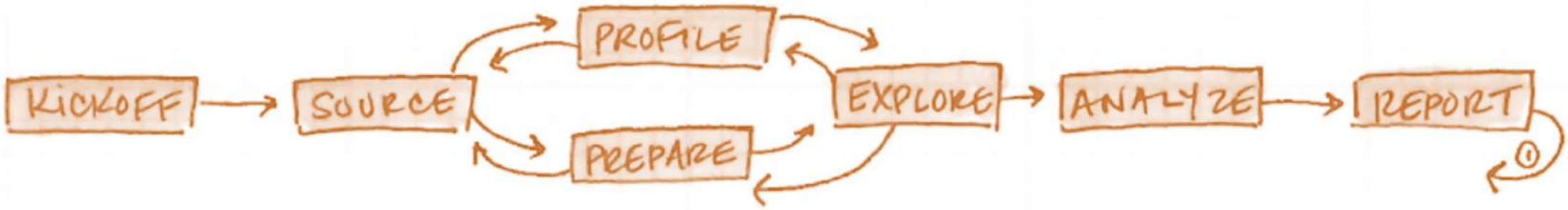
+ 1 Client goals met
- 1 NO strategic goals/vision met
- 1 Inconsistent & redundant

Hybrid structure



+ 1 Standardized processes
+ 1 Strategic goals/vision met
+ 1 Client goals met

Data project workflow



KICKOFF

Formulate your question / problem statement

SOURCE

Gather data, documentation, and team

PROFILE

Define the data's shape, features, and any limitations

PREPARE

Transform the data into a usable form

EXPLORE

Understand the data better through visualization / inspection

ANALYZE

Build models and test for validity / accuracy

REPORT

Communicate your results and plan for next steps

Kickoff

- During the kickoff, you will probably have a few rounds of refining your questions – spend time on this phase as it sets the direction for the project!

Formulate the initial question(s)

Refine according to stakeholder priorities / needs

Refine according to metrics and time constraints

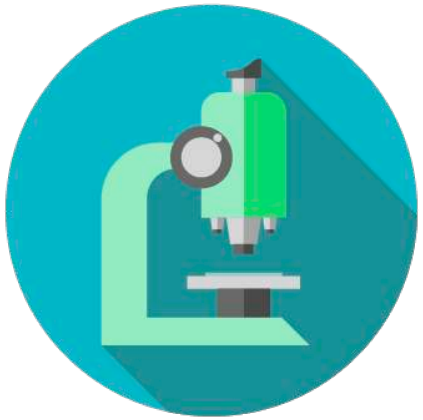
Source

- Now that you've identified the metrics you want, gather the data you need to answer the question

Use internal existing datasets

Find open data

Purchase from external sources



Collect your own data

Profile / prepare

- Once the data is collected, it's imperative that it's formatted and validated for analysis – this can take up to 70-80% of any data project!

Clean the data for analysis

Validate the variables and metrics

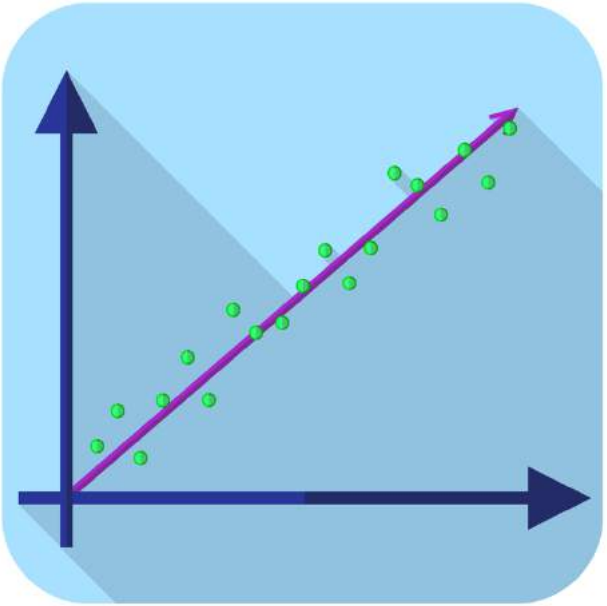
Scale the data as needed

Confirm how the data was collected

Data exploration

- Explore the variables and their interactions through visualizations and sampling

Univariate – explore the distribution of individual variables

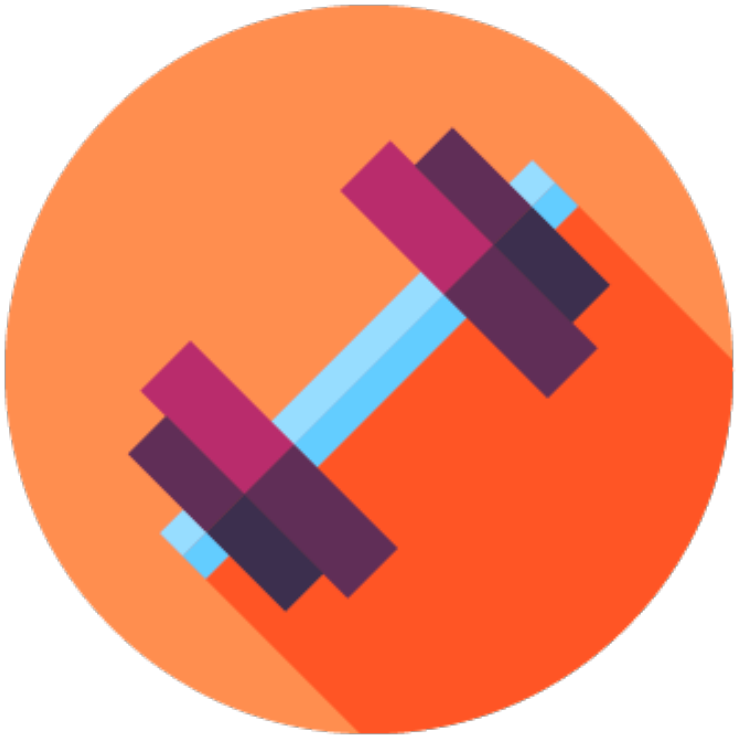


Bivariate – explore how variables relate to each other

Analysis

- Now that we have the dataset cleaned and refined, it's time for analysis – remember, your question may have changed as you explored the data

- You should have mapped your analysis out as you were setting the questions and the metrics to measure – refer back to Day 2 materials for the methods that answer a particular question



Report

- Refer back to the Data Storytelling module from Day 1 about how to build the right report and visualization – remember to:

Anticipate questions

Tailor the presentation to the stakeholder

Build clear visualizations

Post-mortem

- After the project has been presented and evaluated, do a debrief with your team about what went right and what could be improved for next time

What went well? What could be improved?

Did you reach your project goals?

Can it be repeated / built upon by others?

Activity: evaluate current skills

- Turn to your participant guide to the Internal data science capacity to see what skills you currently have and what skills you need or may need in the future
- You'll think through your current staff and identify skills gaps that you have
- Then, map out an organizational chart that is either your current team or, if you identified skills gaps, the team you would need to have to address these gaps
- Discuss your findings and ideas with your group

Activity time: 15 - 20 minutes

Congratulations!

1. Data storytelling
2. Data ethics frameworks
3. Open data sources
4. Managing data science projects / teams

Tomorrow, you'll learn

1. How to build a more data-driven mindset in your agencies
2. How to use and apply common data science tools
3. How to implement events to improve data awareness and data-driven culture