GENERAL ASSEMBLY

INTRODUCTION TO DATA SCIENCE

Colin Crawford

INTRODUCTION TO DATA SCIENCE

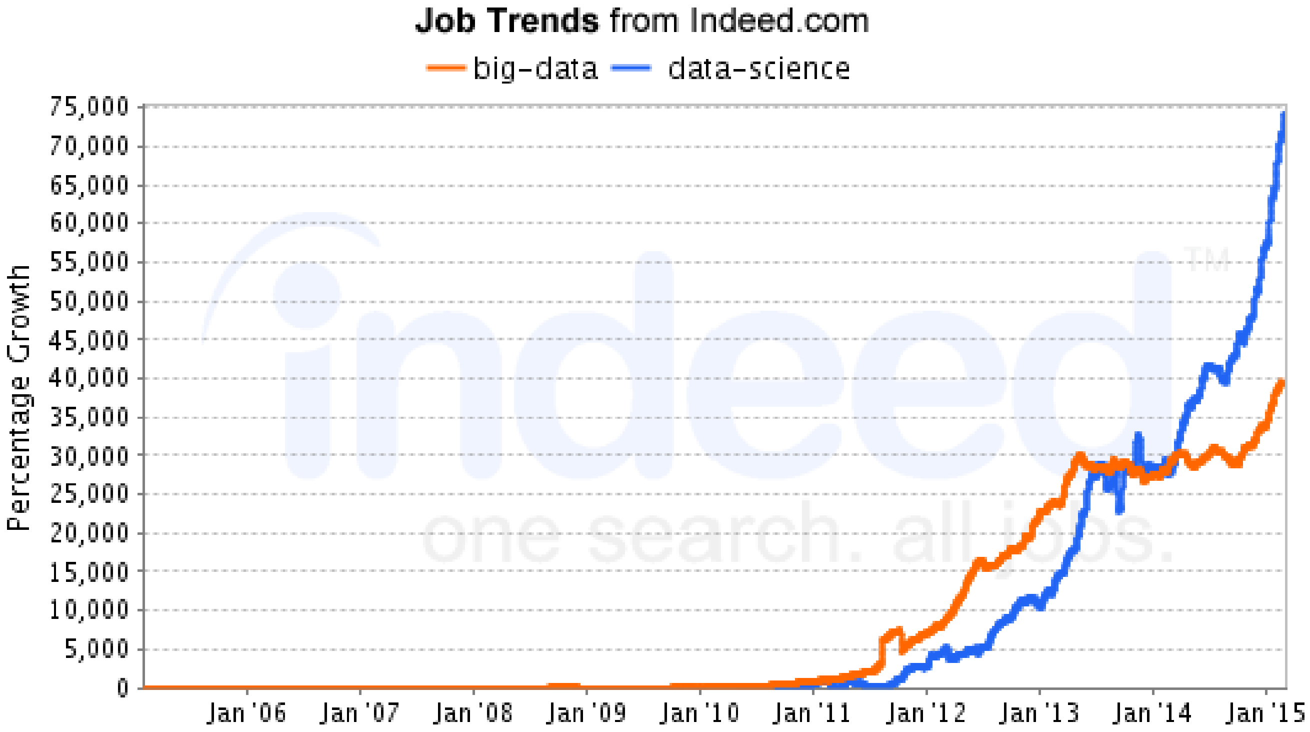
AGENDA

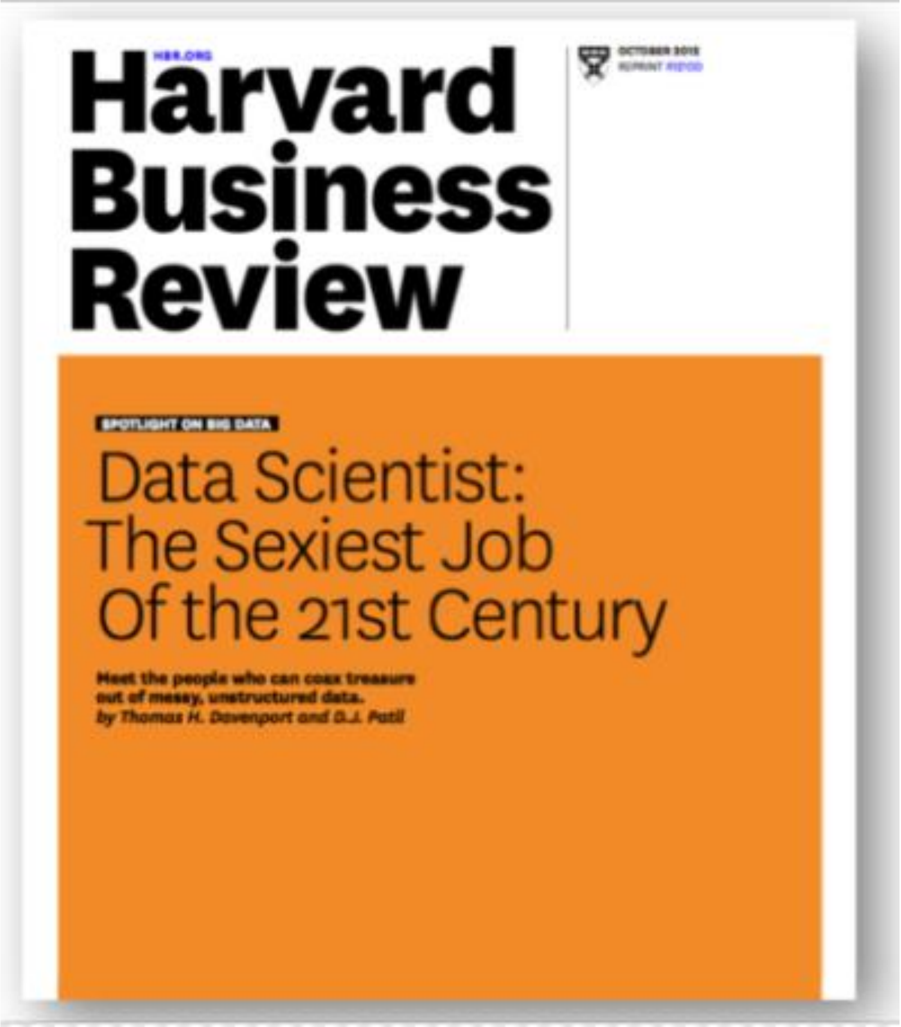
- What Is Data Science?
- How Is Data Science Used?
- What Does a Data Science Workflow Look Like?
- Q & A

INTRODUCTION TO DATA SCIENCE

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?





HBR.ORG

Harvard Business Review

OCTOBER 2012
REPRINT R1210D

SPOTLIGHT ON BIG DATA

Data Scientist:
The Sexiest Job
Of the 21st Century

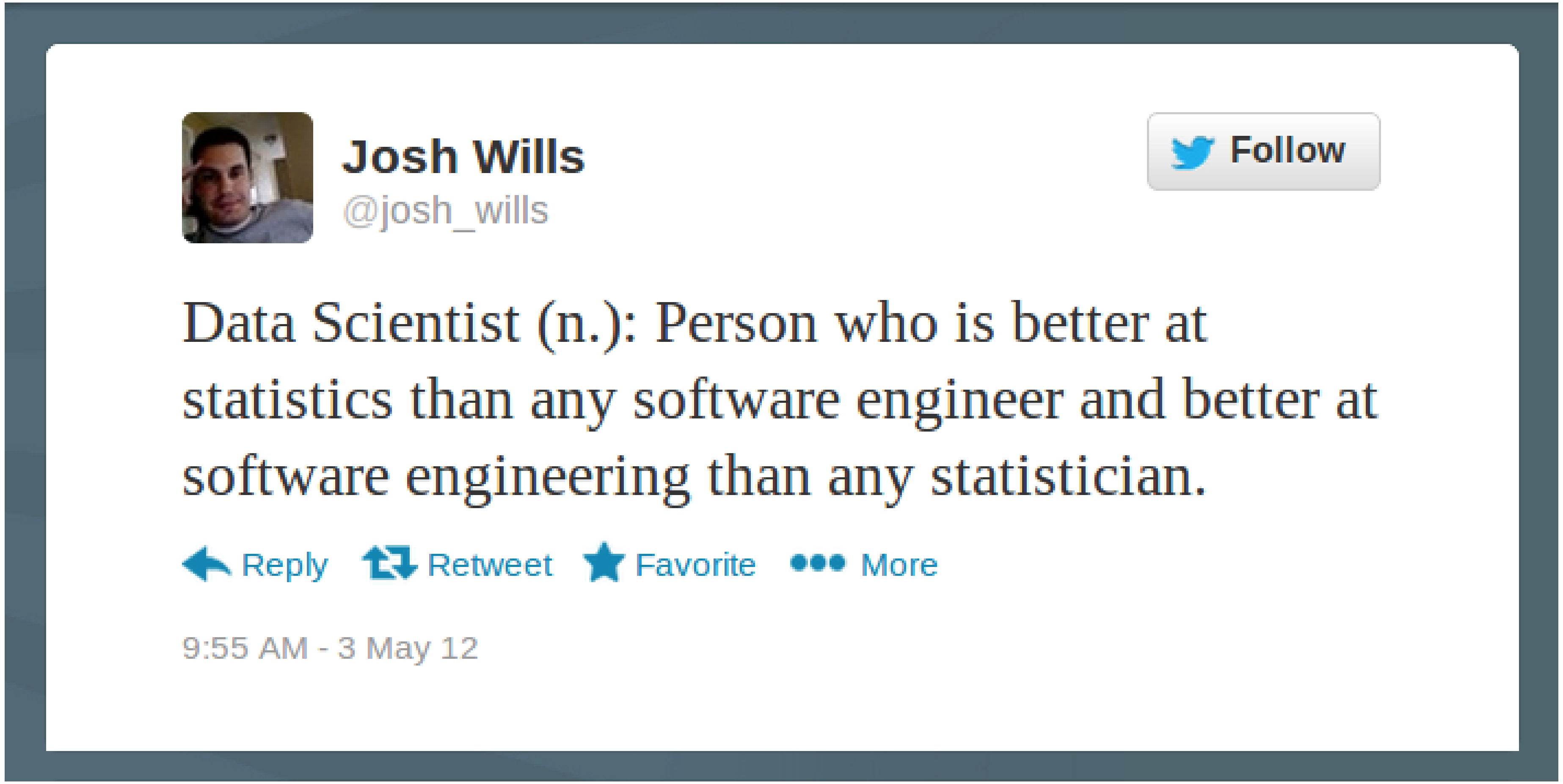
Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

WHAT IS DATA SCIENCE?

THE OFFICIAL ANSWER

- A set of tools and techniques used to extract useful information from data
- An interdisciplinary, problem-solving-oriented subject
- The application of scientific techniques to practical problems
- A rapidly-growing field

WHAT IS DATA SCIENCE?

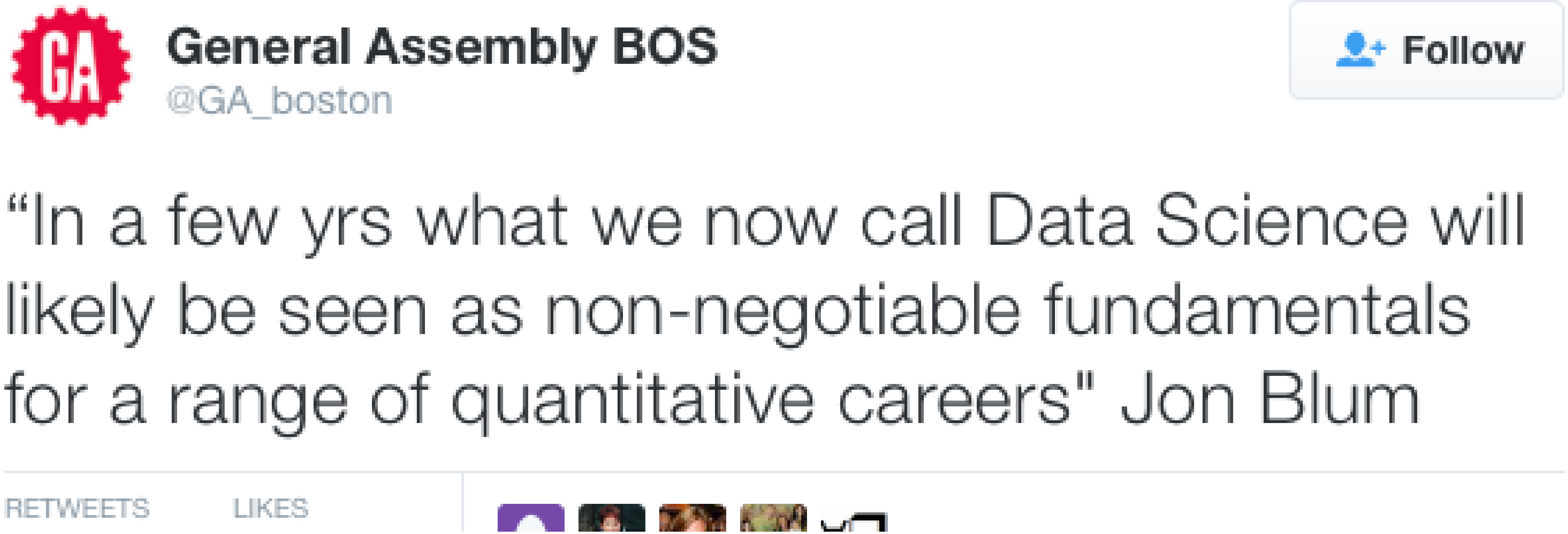


Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

9:55 AM - 3 May 12

WHAT IS DATA SCIENCE?



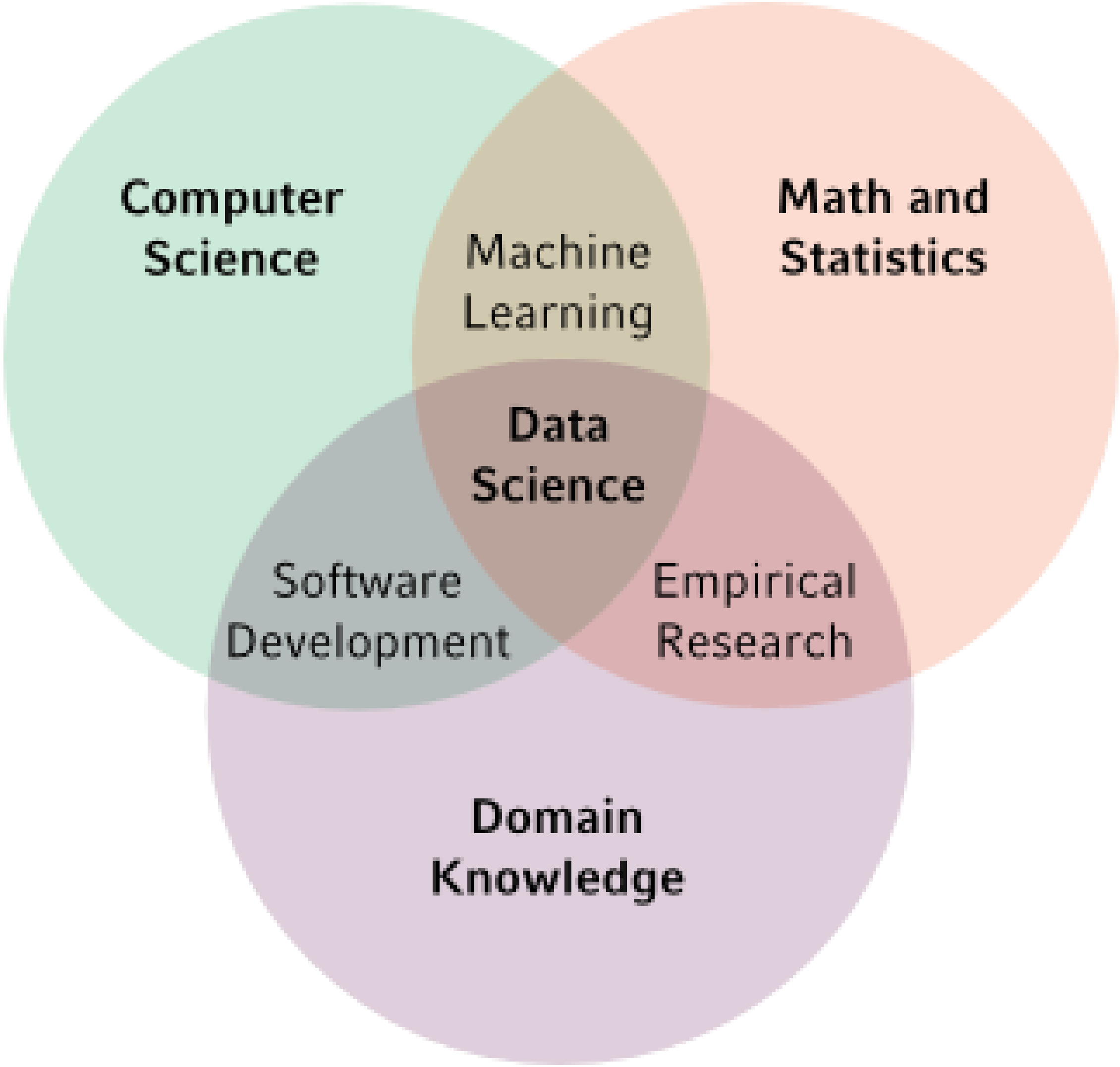
General Assembly BOS
@GA_boston

“In a few yrs what we now call Data Science will likely be seen as non-negotiable fundamentals for a range of quantitative careers" Jon Blum

RETWEETS LIKES

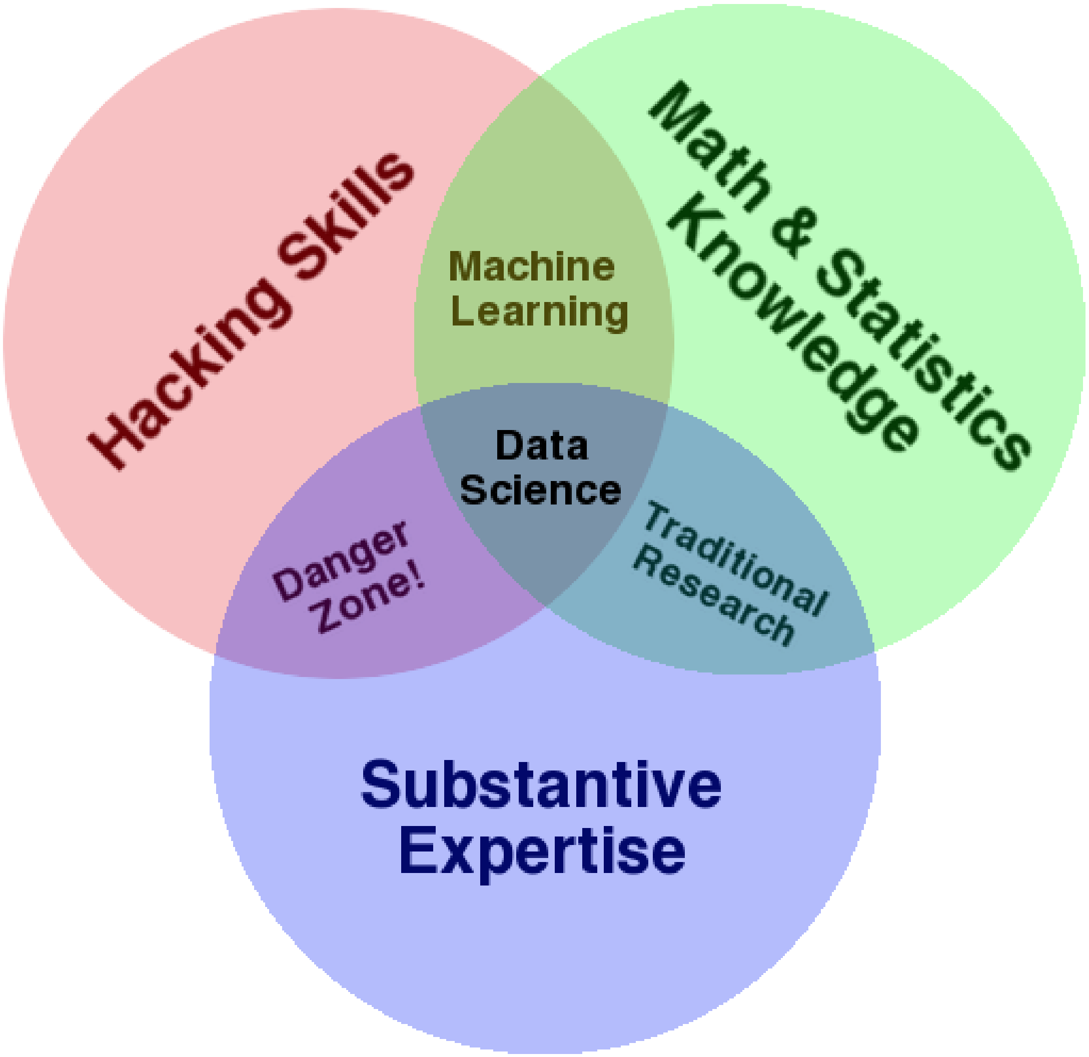
WHAT IS DATA SCIENCE?

BY FIELD

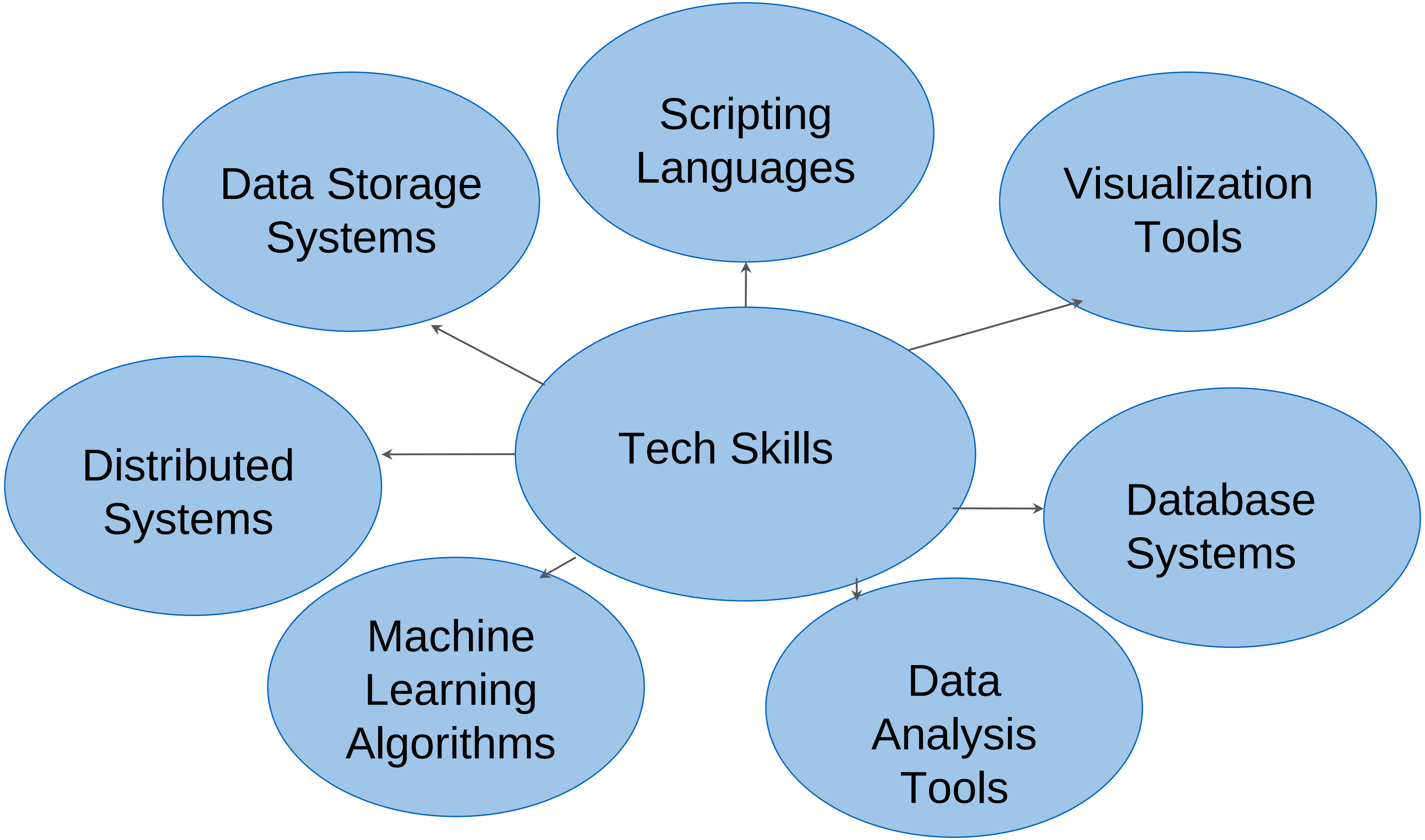


WHAT IS DATA SCIENCE?

BY SKILL



WHAT IS DATA SCIENCE?



Tech Skills

Scripting Languages

Data Storage Systems

Visualization Tools

Distributed Systems

Database Systems

Machine Learning Algorithms

Data Analysis Tools

HOW IS DATA SCIENCE USED?

HOW IS DATA SCIENCE USED?

"TARGET KNEW I WAS PREGNANT"

- DS team analyzed buying patterns of women on baby registries
- Trends emerged:

  Specific lotions and vitamins in 2nd trimester

  Switch to scent-free products in 3rd trimester

  Colored items signifying gender
- Marketing team used this data to send coupons
- Some family surprises…

HOW IS DATA SCIENCE USED?

NETFLIX PROGRAMMING DECISIONS

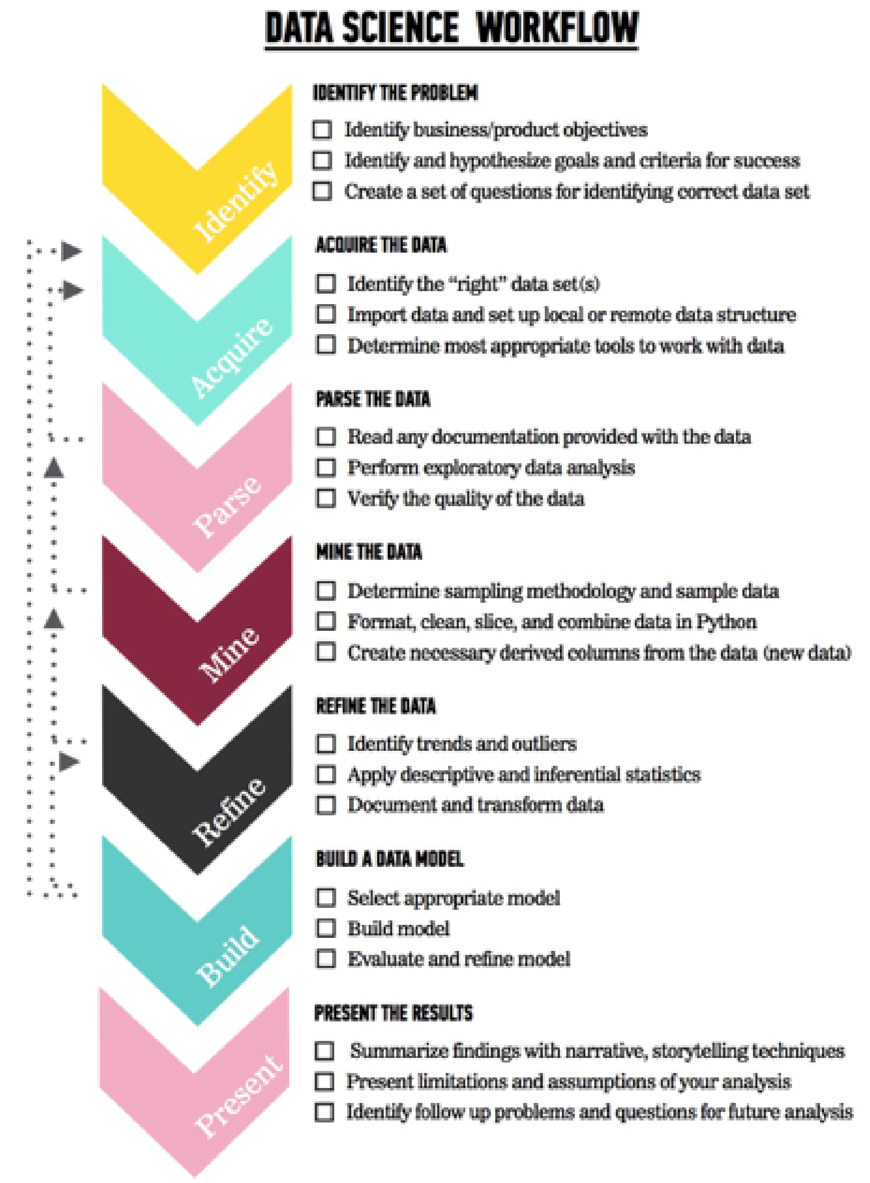
- Features:
  - When user watches and for how long
  - Where and on what device the user is watching
  - Rewinds, pauses, fast-forwards
- Same categorization for current programming used to determine new content
  - (Animated / Comedy / 18-29 / Cynical / 30 min / etc.)

DATA SCIENCE WORKFLOW

WHAT IS DATA SCIENCE?

DATA SCIENCE WORKFLO

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

MINE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

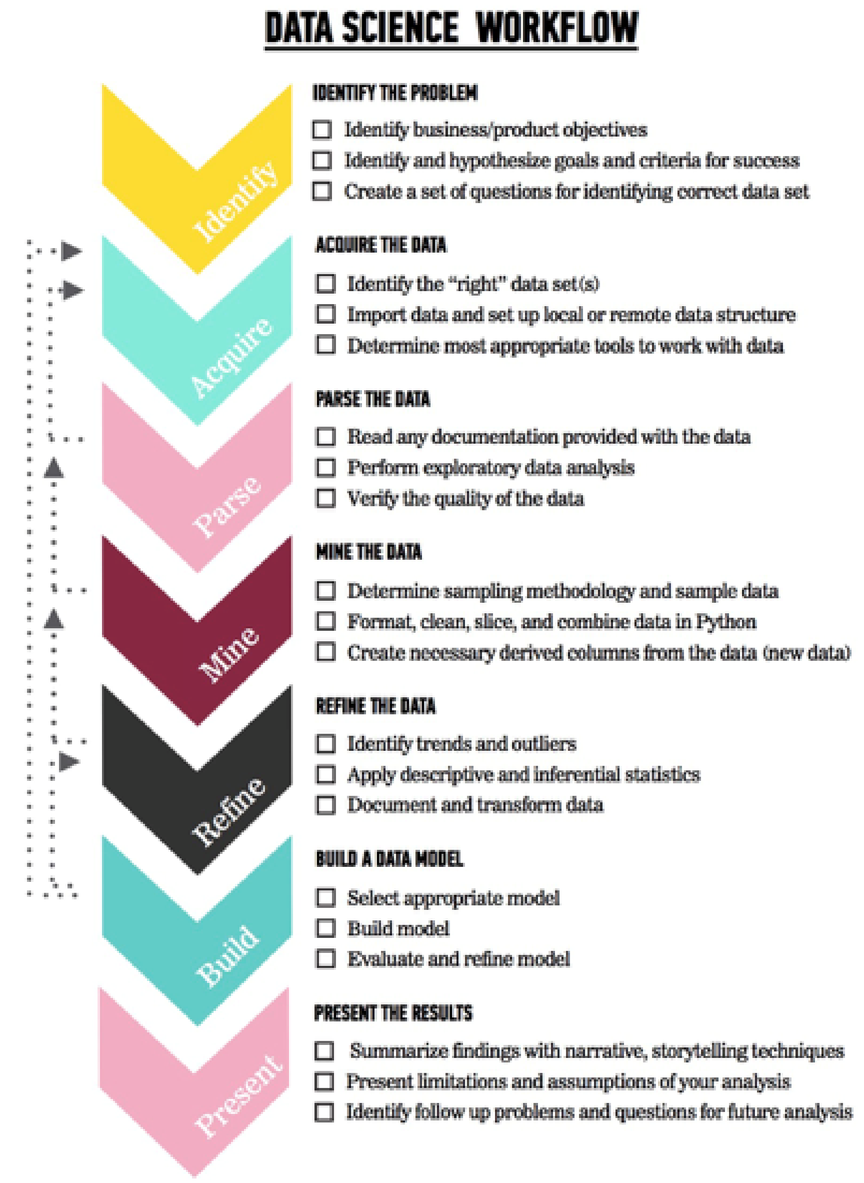
PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

WHAT IS DATA SCIENCE?

IDENTIFY

- Why are you doing this in the first place?
- Who are the stakeholders?
- What data will you need?
- How will you define success?



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

□ Identify business/product objectives
□ Identify and hypothesize goals and criteria for success
□ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

□ Identify the "right" data set(s)
□ Import data and set up local or remote data structure
□ Determine most appropriate tools to work with data

PARSE THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of the data

MINE THE DATA

□ Determine sampling methodology and sample data
□ Format, clean, slice, and combine data in Python
□ Create necessary derived columns from the data (new data)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

PRESENT THE RESULTS

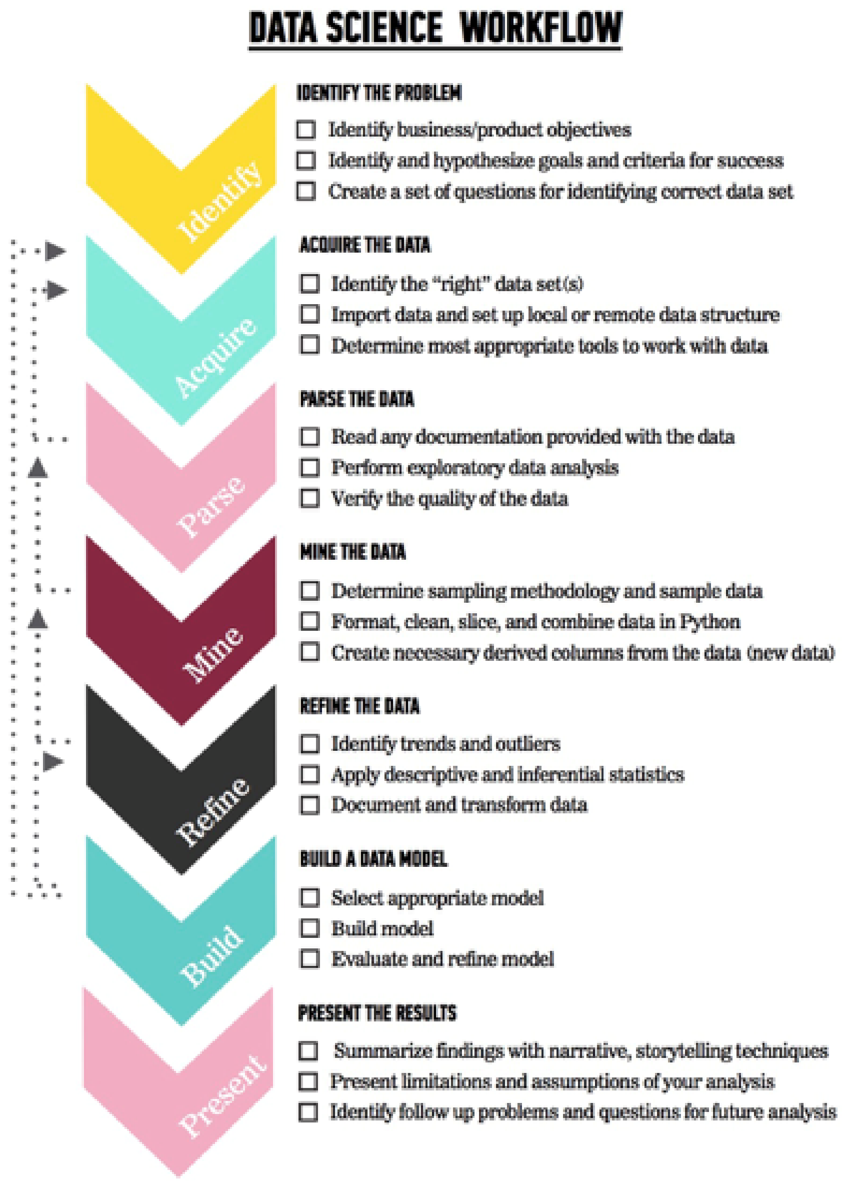
□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

WHAT IS DATA SCIENCE?

ACQUIRE

- Is the data you need available to you?
- How is it stored?

  Relational database, NoSQL datastore, files
- Is it public or proprietary?
- Can it be supplemented?
- What tools will you need to import it?
- What tools will you need to work with it?



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

□ Identify business/product objectives
□ Identify and hypothesize goals and criteria for success
□ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

□ Identify the "right" data set(s)
□ Import data and set up local or remote data structure
□ Determine most appropriate tools to work with data

PARSE THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of the data

MINE THE DATA

□ Determine sampling methodology and sample data
□ Format, clean, slice, and combine data in Python
□ Create necessary derived columns from the data (new data)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

PRESENT THE RESULTS

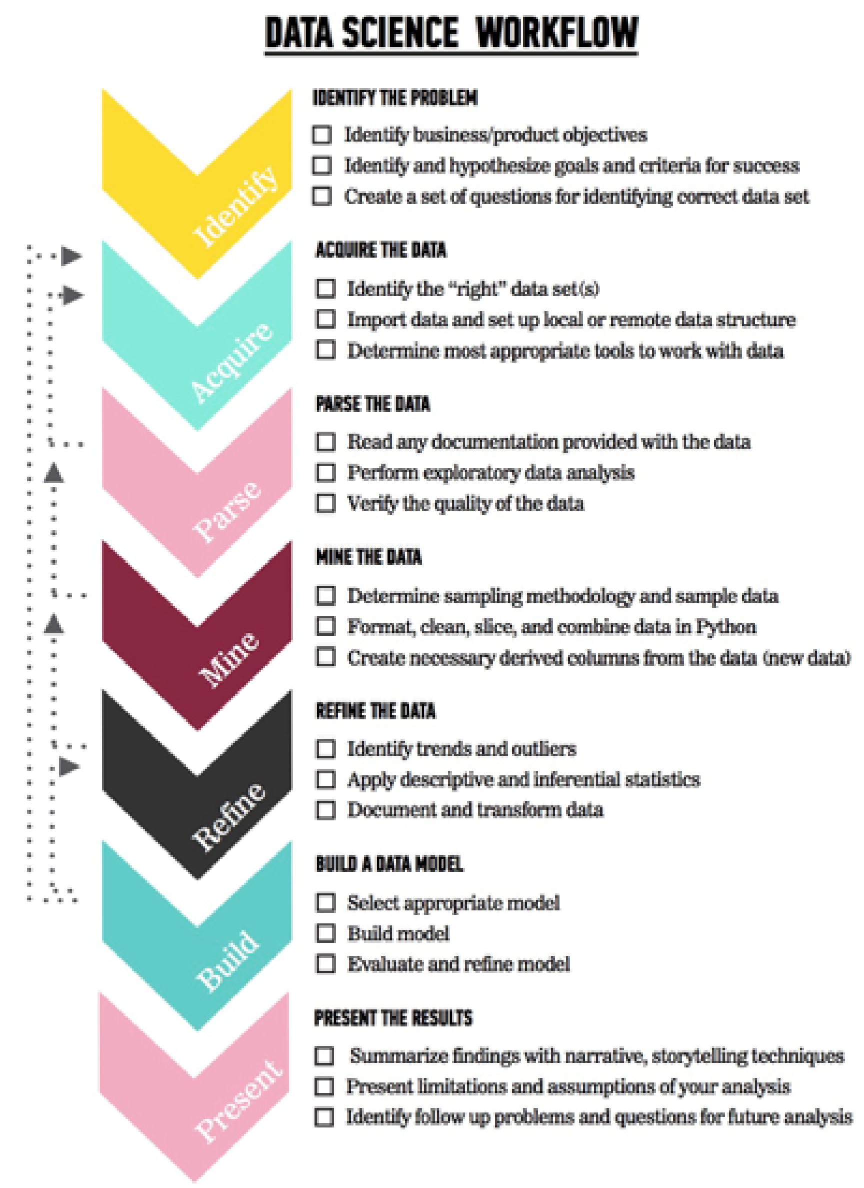
□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

WHAT IS DATA SCIENCE?

PARSE

- What documentation is available, if any?

  (People count as documentation.)
- Were you able to import the data?
- What did you learn from initial exploratory analysis?
- Is the data, in fact, sufficient?
- How much munging will it require?



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

- □ Identify business/product objectives
- □ Identify and hypothesize goals and criteria for success
- □ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

- □ Identify the "right" data set(s)
- □ Import data and set up local or remote data structure
- □ Determine most appropriate tools to work with data

PARSE THE DATA

- □ Read any documentation provided with the data
- □ Perform exploratory data analysis
- □ Verify the quality of the data

MINE THE DATA

- □ Determine sampling methodology and sample data
- □ Format, clean, slice, and combine data in Python
- □ Create necessary derived columns from the data (new data)

REFINE THE DATA

- □ Identify trends and outliers
- □ Apply descriptive and inferential statistics
- □ Document and transform data

BUILD A DATA MODEL

- □ Select appropriate model
- □ Build model
- □ Evaluate and refine model

PRESENT THE RESULTS

- □ Summarize findings with narrative, storytelling techniques
- □ Present limitations and assumptions of your analysis
- □ Identify follow up problems and questions for future analysis

WHAT IS DATA SCIENCE?

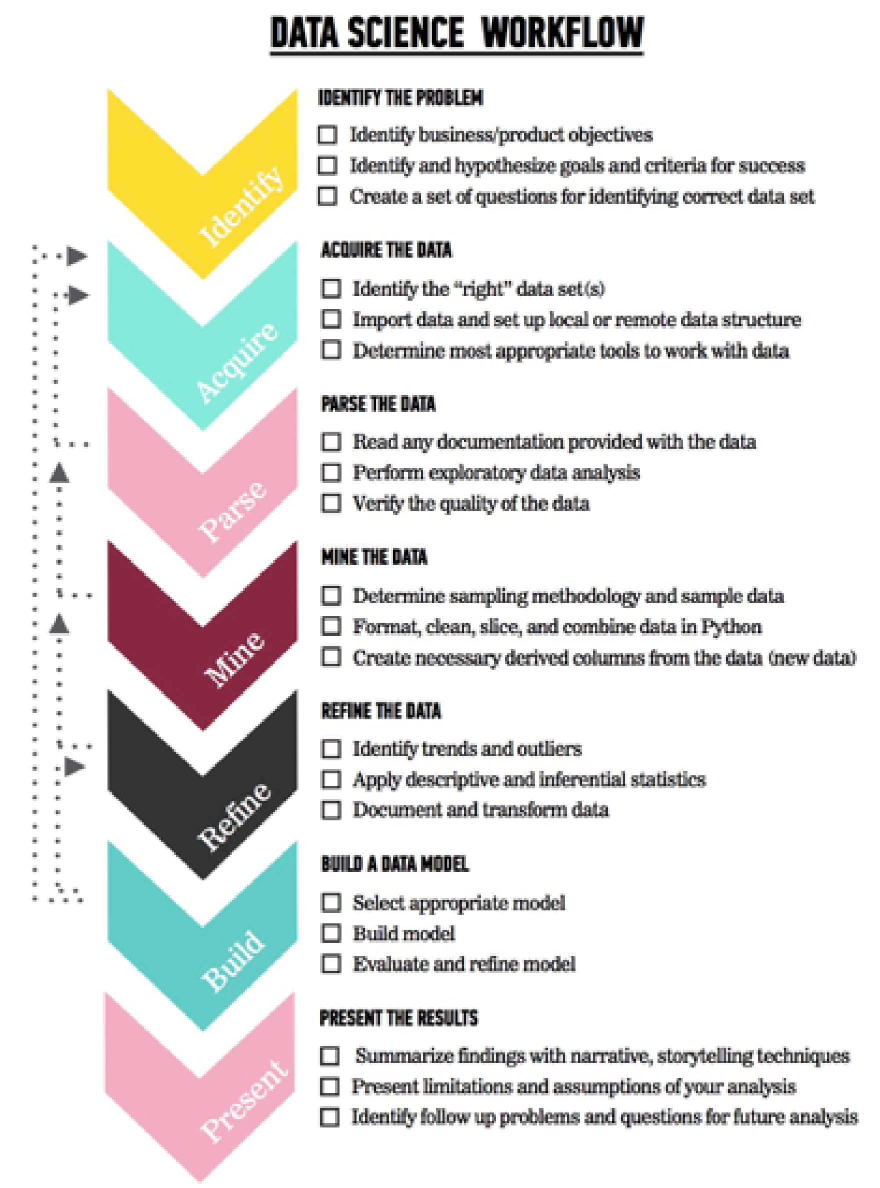
MINE

- What sampling methodology will you use?

WRANGLING/MUNGING/MINING

(…80% of work occurs here…)

- What secondary features need to be derived?



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

□ Identify business/product objectives
□ Identify and hypothesize goals and criteria for success
□ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

□ Identify the "right" data set(s)
□ Import data and set up local or remote data structure
□ Determine most appropriate tools to work with data

PARSE THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of the data

MINE THE DATA

□ Determine sampling methodology and sample data
□ Format, clean, slice, and combine data in Python
□ Create necessary derived columns from the data (new data)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

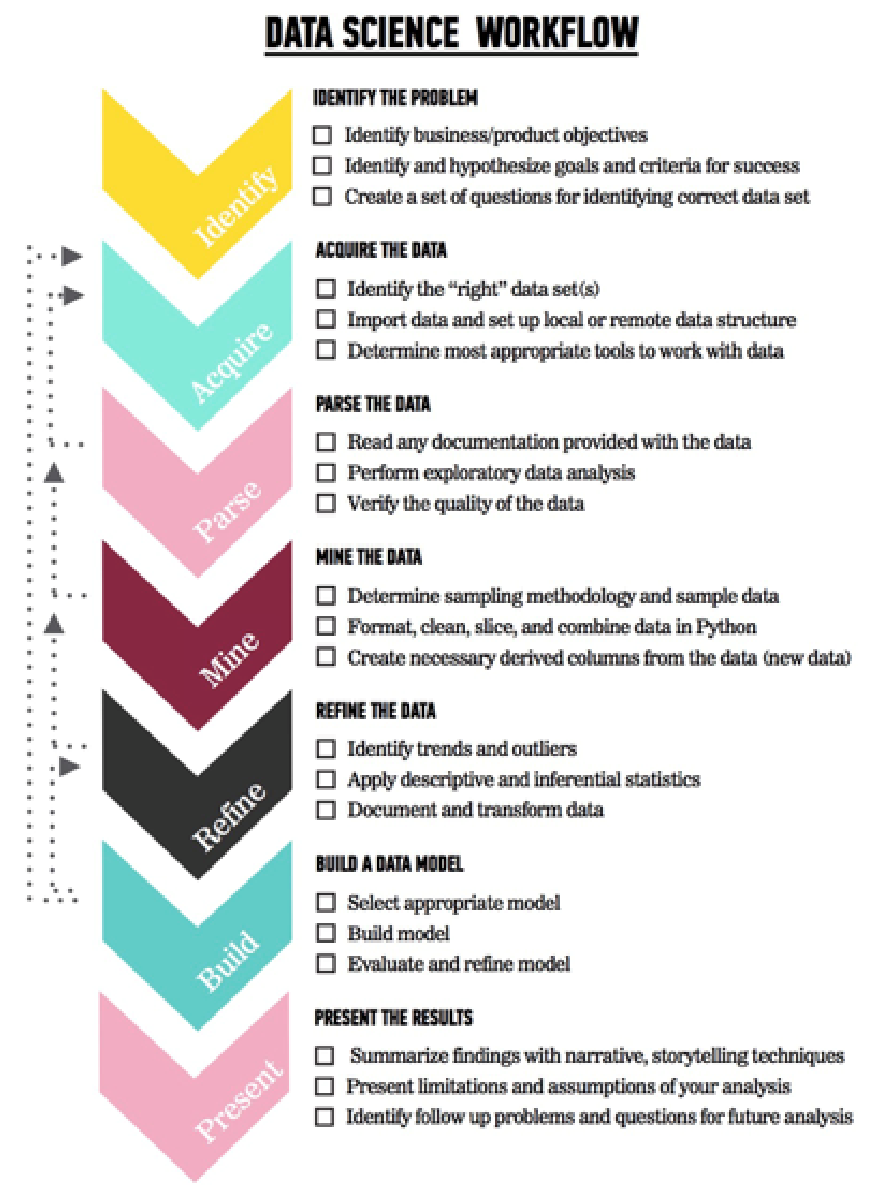
PRESENT THE RESULTS

□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

WHAT IS DATA SCIENCE?

REFINE

- What new trends appear?
- Any notable outliers?
- Notable results from applying derived features?
- Are you documenting your approach?



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

□ Identify business/product objectives
□ Identify and hypothesize goals and criteria for success
□ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

□ Identify the "right" data set(s)
□ Import data and set up local or remote data structure
□ Determine most appropriate tools to work with data

PARSE THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of the data

MINE THE DATA

□ Determine sampling methodology and sample data
□ Format, clean, slice, and combine data in Python
□ Create necessary derived columns from the data (new data)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

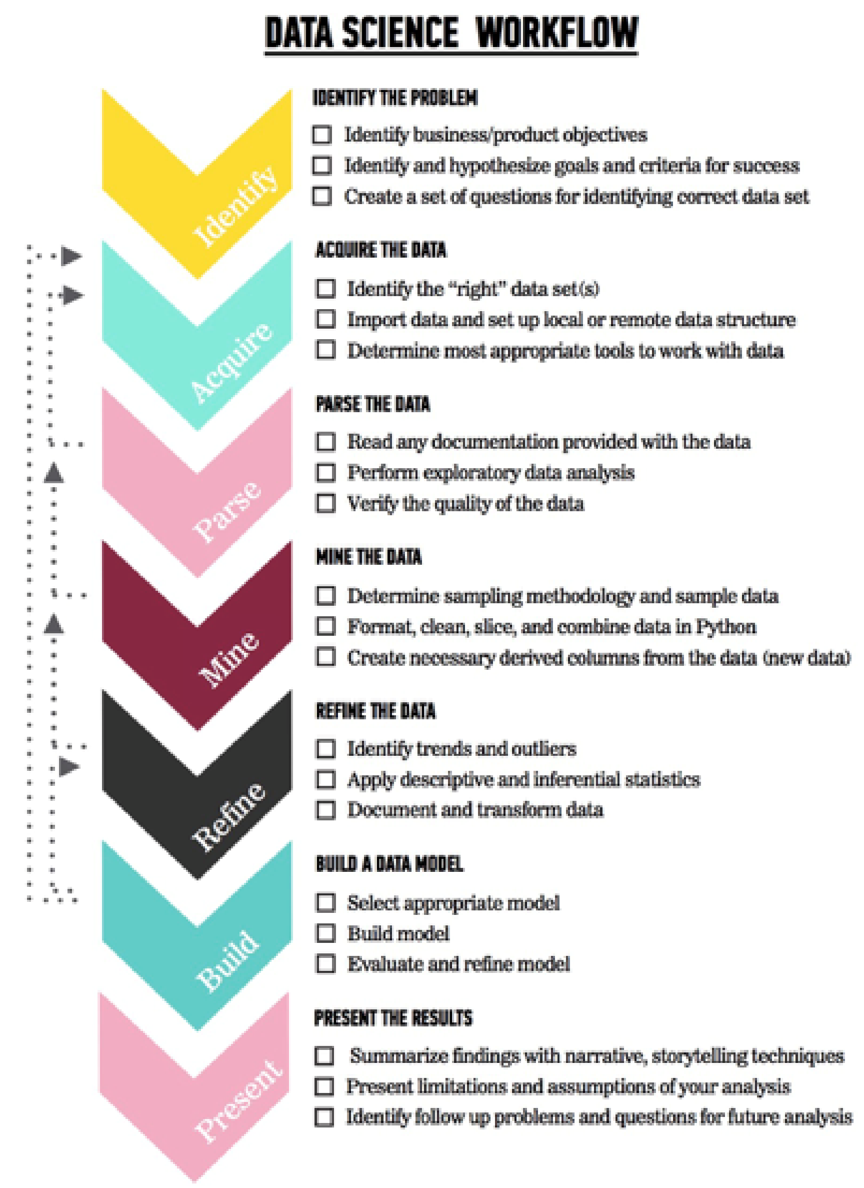
PRESENT THE RESULTS

□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

WHAT IS DATA SCIENCE?

MODEL

- What model or models are most appropriate for the data and the problem?
- Is the data in the appropriate format for the desired model(s)?
- How did the initial model perform?
- Any symptoms of over-fitting?



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

MINE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

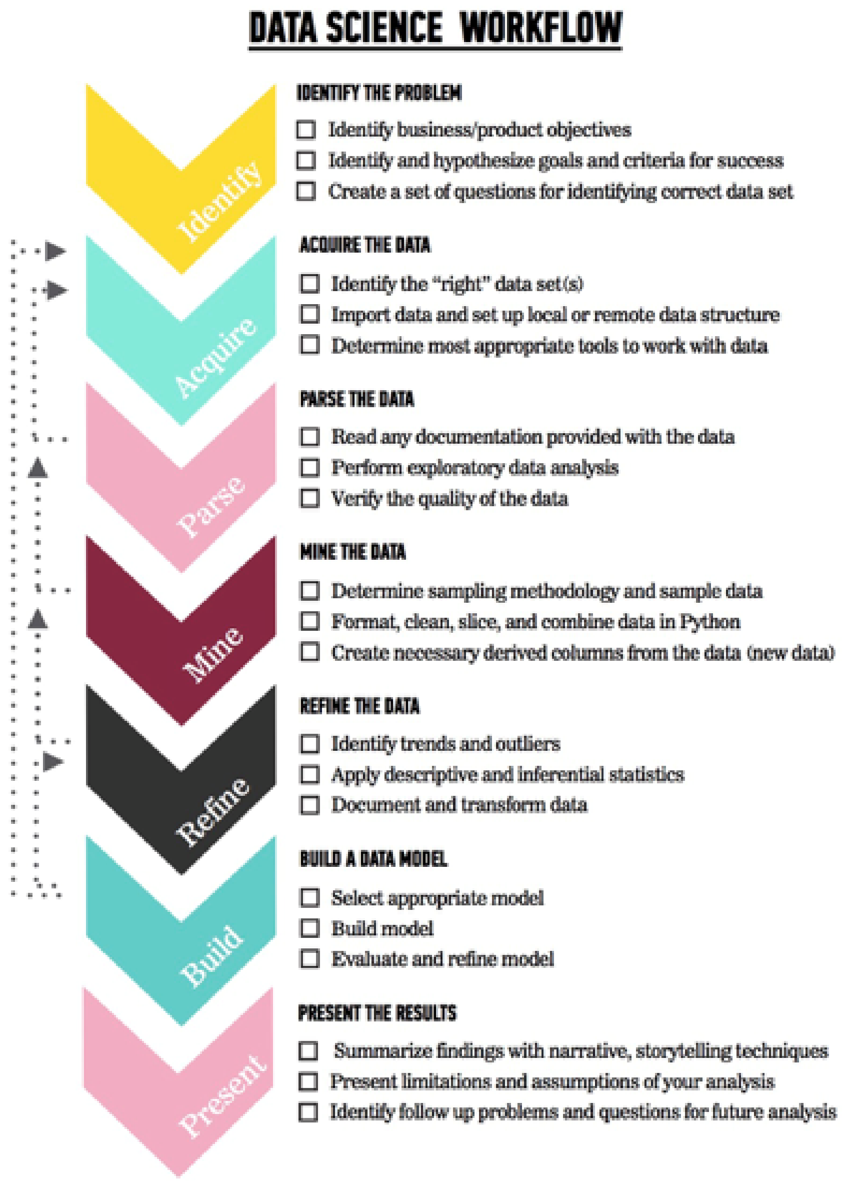
PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

WHAT IS DATA SCIENCE?

PRESENT

- What narrative do I want to tell?
- What assumptions are we making?
- What inherent limitations should be disclosed?
- Were the criteria for success met?
- What are the next steps?



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM
- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

ACQUIRE THE DATA
- ☐ Identify the “right” data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

PARSE THE DATA
- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

MINE THE DATA
- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

REFINE THE DATA
- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

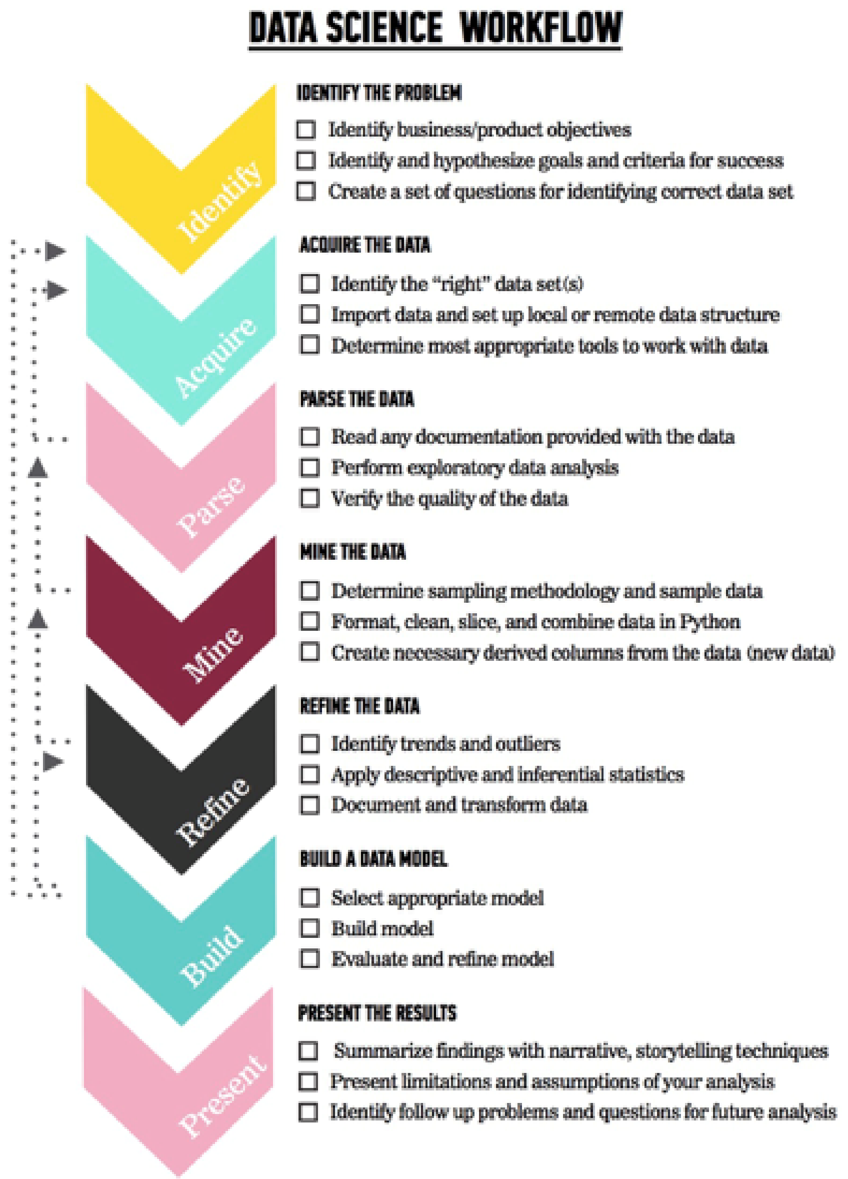
BUILD A DATA MODEL
- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

PRESENT THE RESULTS
- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

WHAT IS DATA SCIENCE?

(DEPLOY)

- What changes need to be made to run the model in production?
- Can the model handle real-time data?
- Will the model need to be retrained? How often?
- Does the deployment team have the context to satisfy the original intent?



DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM

□ Identify business/product objectives
□ Identify and hypothesize goals and criteria for success
□ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

□ Identify the "right" data set(s)
□ Import data and set up local or remote data structure
□ Determine most appropriate tools to work with data

PARSE THE DATA

□ Read any documentation provided with the data
□ Perform exploratory data analysis
□ Verify the quality of the data

MINE THE DATA

□ Determine sampling methodology and sample data
□ Format, clean, slice, and combine data in Python
□ Create necessary derived columns from the data (new data)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

PRESENT THE RESULTS

□ Summarize findings with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up problems and questions for future analysis

Q&A

Other Resources

- **An Introduction to Statistical Learning: with Applications in R**
- General Assembly Courses
- Coursera
- Udemy
- Youtube lectures (Stanford NLP)
- Dataquest / Code-academy
- Data Science Workflows / Competitions - Kaggle
- General CS - Medium
- Statistics - Khan Academy
- Podcasts - The Data Skeptic