ONLINE MASTERS IN DATA SCIENCE

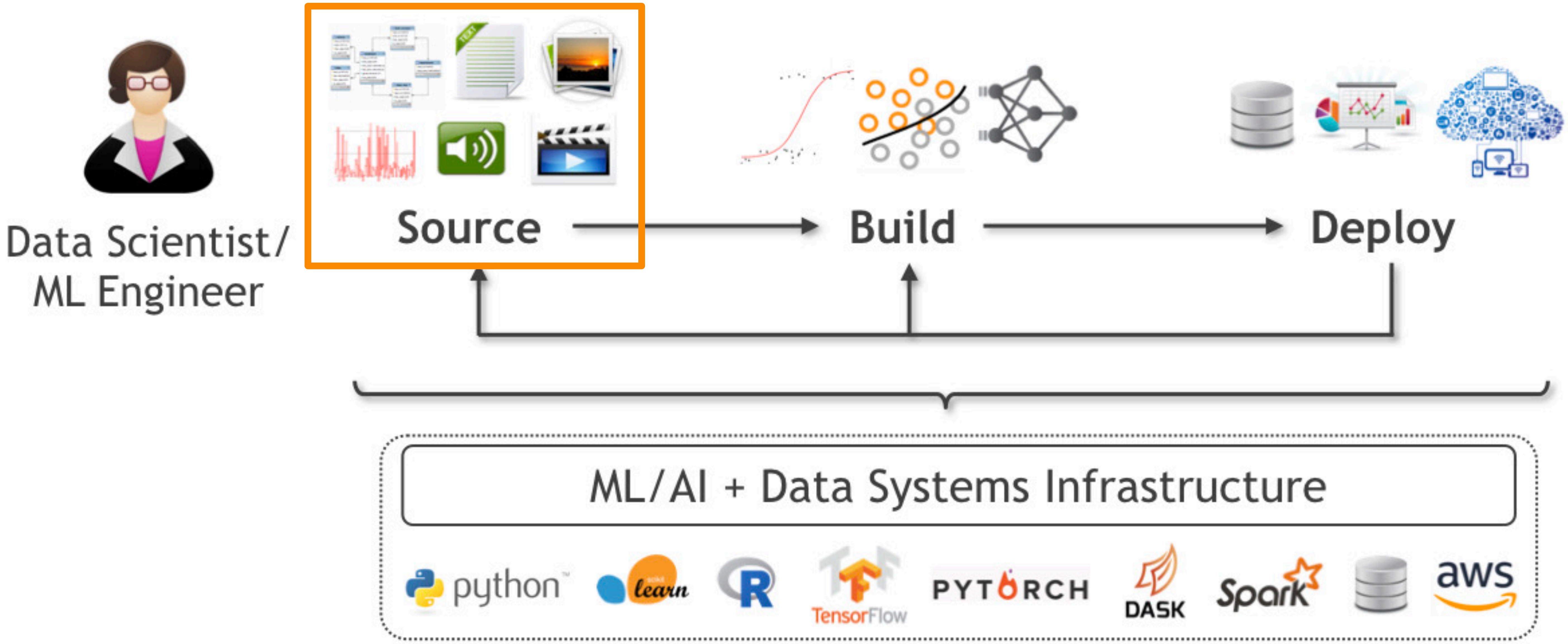
DSC 208R - Data Management for Analytics

Data Collection and Governance

Arun Kumar

UC San Diego
COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

Lifecycle of Real-World Data Science



Data acquisition
Data preparation
Data cleaning

Feature Engineering
Training & Inference
Model Selection

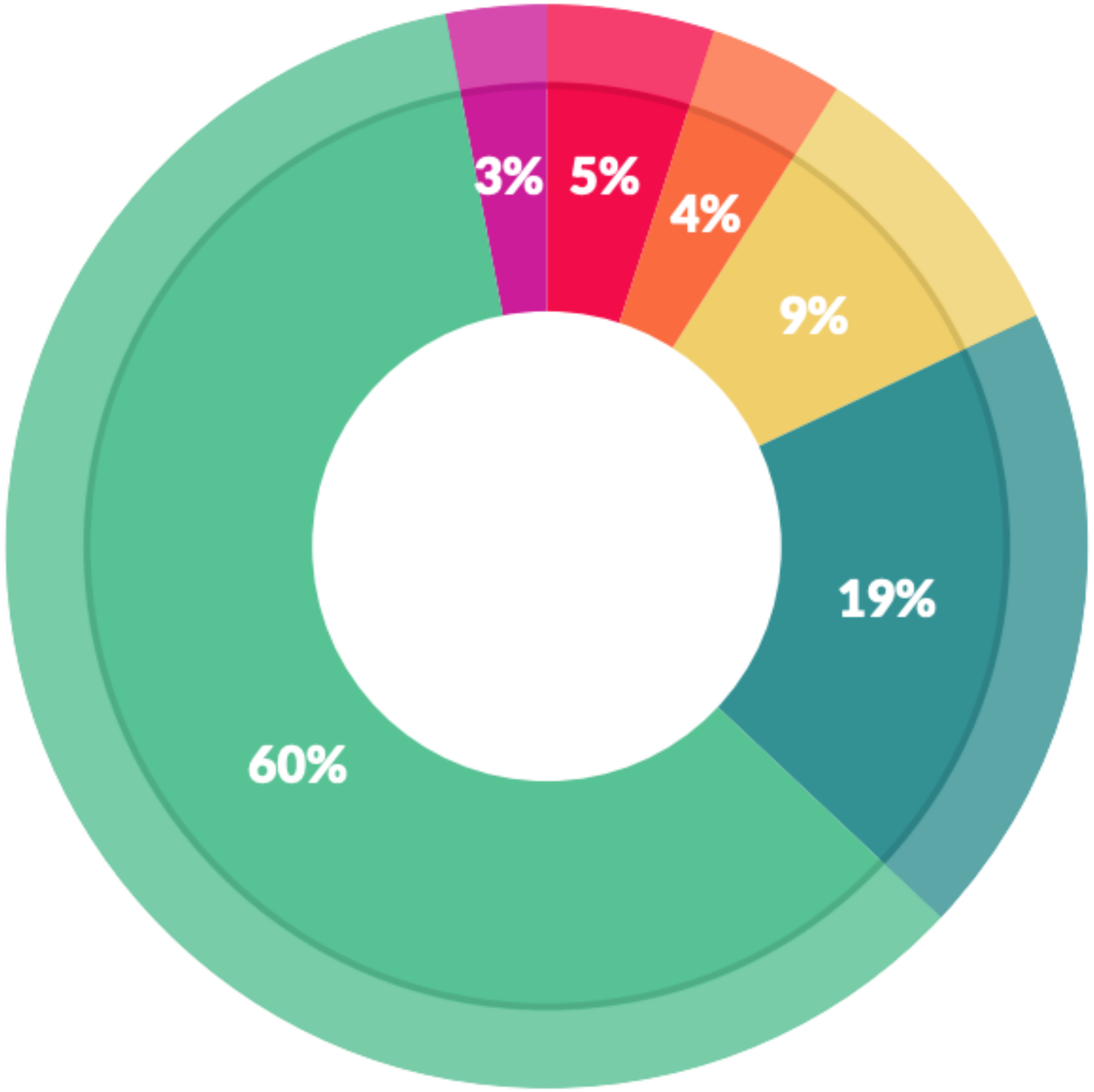
Serving
Monitoring

Outline

- **Overview**
- Data Organization and File Formats
- Data Acquisition
- Data Reorganization and Preparation
- Data Labeling and Amplification
- Data Governance and Privacy

Data Science in the Real World

Q: How do real-world data scientists spend their time?



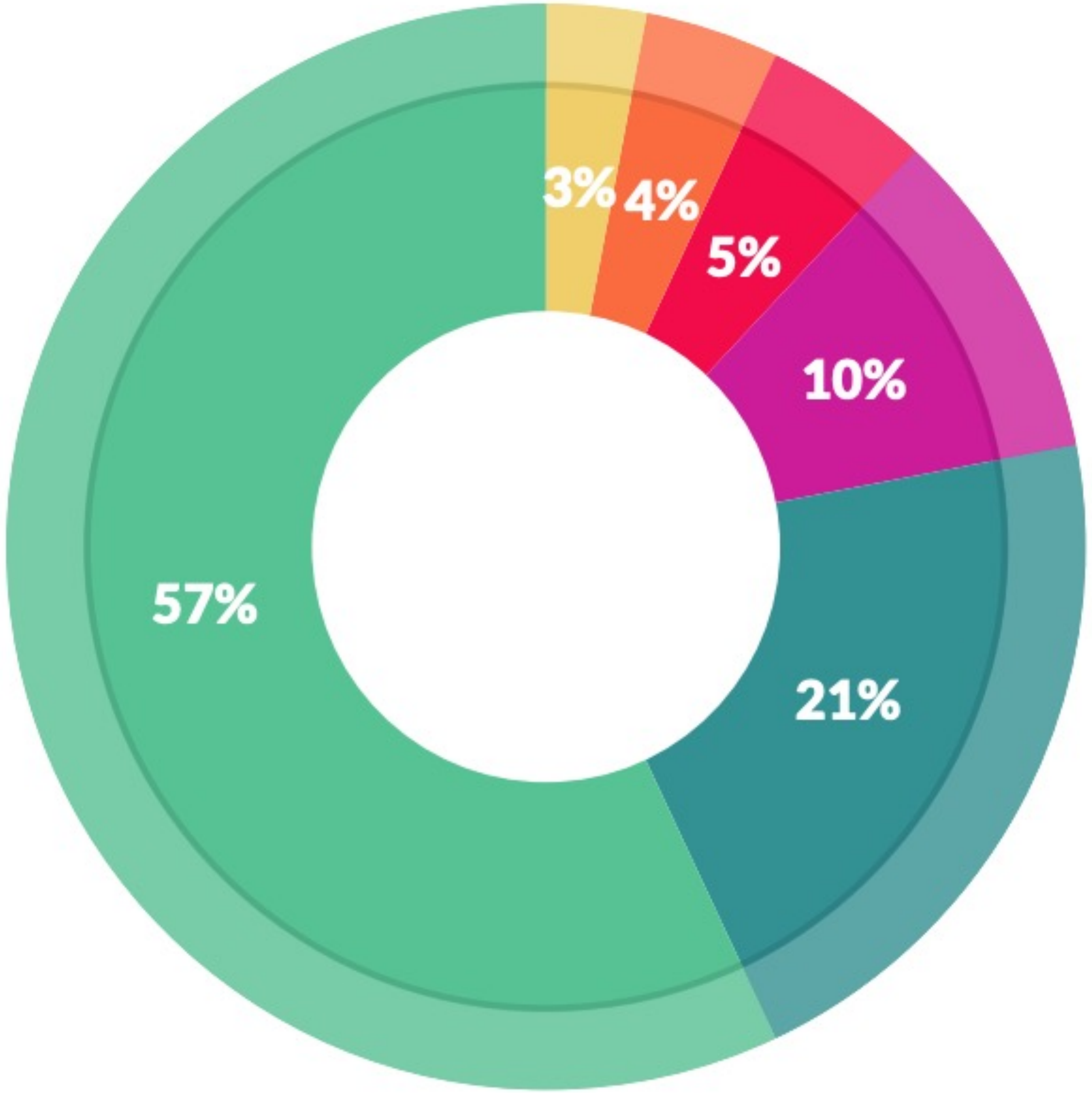
What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Data Science in the Real World

Q: How do real-world data scientists spend their time?



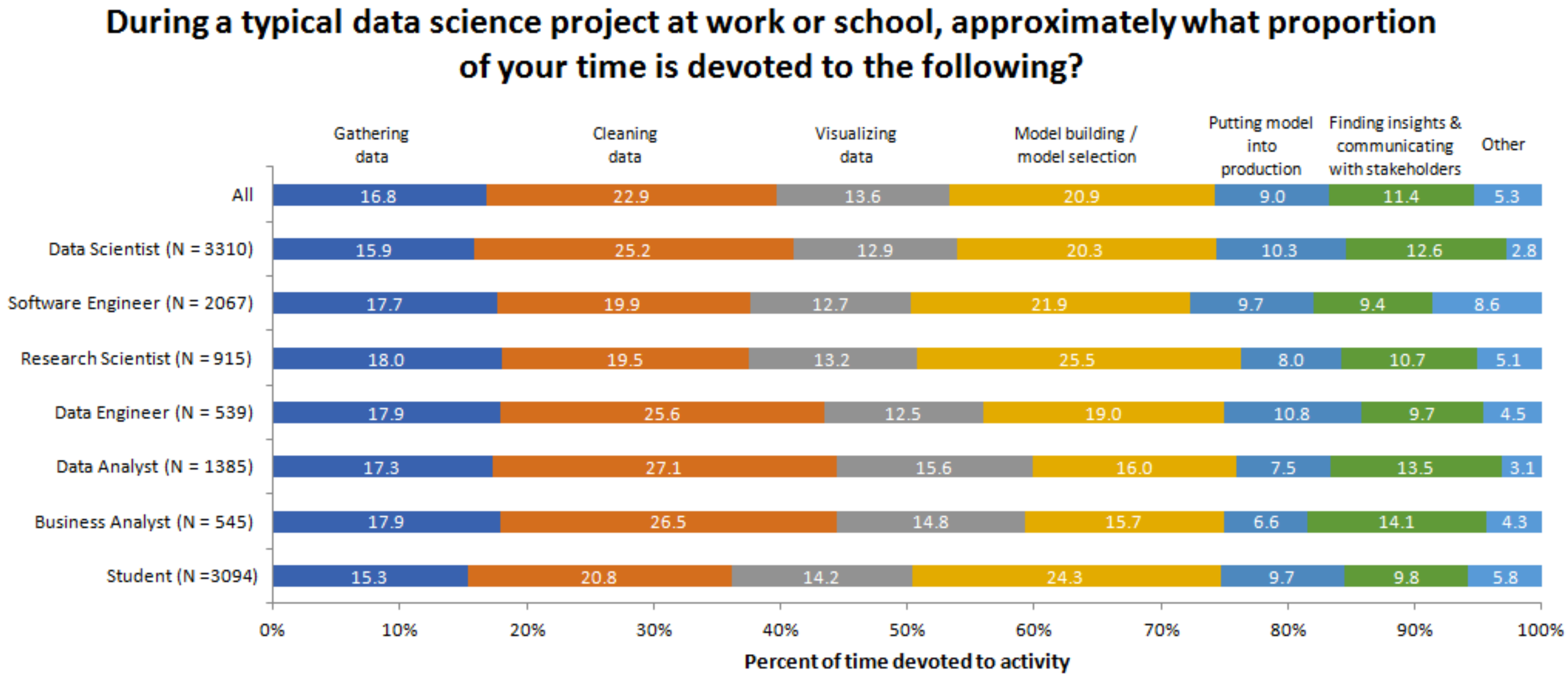
What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Data Science in the Real World

Q: How do real-world data scientists spend their time?



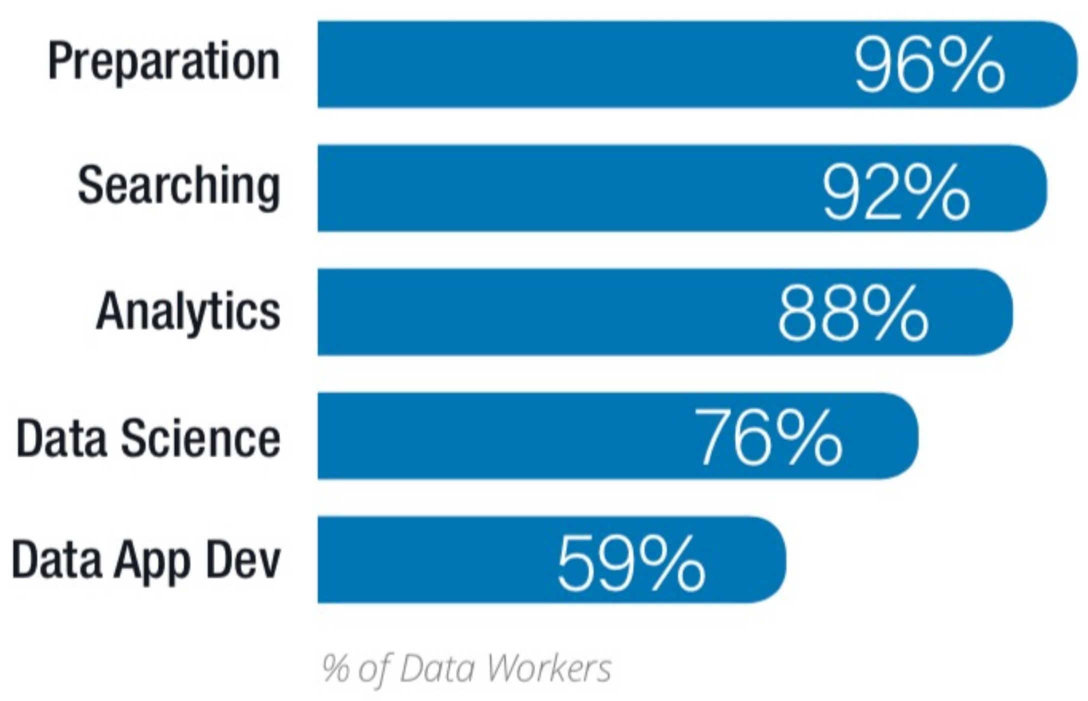
Kaggle State of ML and Data Science Survey 2018

Data Science in the Real World

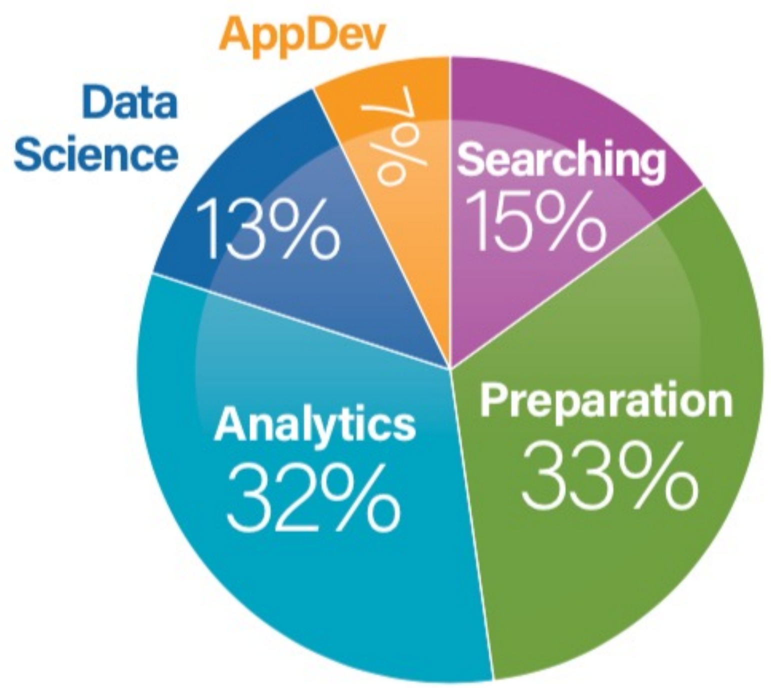
Q: How do real-world data scientists spend their time?

Data workers spend **90%** of their work week on data related activities

Activities Performed by Data Workers



Weekly Time by Activity



Searching for and preparing data are the most common activities regardless of role of data worker

IDC-Alteryx State of Data Science and Analytics Report 2019

Sourcing Stage of DS Lifecycle

- DS applications do not exist in a vacuum. They work with the *data-generating process* and prediction application.
- **Sourcing:**
  - The stage of where you go from raw datasets to "analytics/ML-ready" datasets
  - Rough end point: SQL analytics for BI, data/feature engineering for ML/AI analytics

Sourcing Stage of DS Lifecycle

Q: What makes Sourcing challenging?

- **Heterogeneity** of data modalities, file formats, sources
- Data **access**/availability constraints
- **Bespoke**/diverse kinds of prediction applications
- Unpredictable and continual **edits** to datasets
- **Messy**, incomplete, ambiguous, and/or erroneous data
- Large **scale** of data
- Poor data **governance** in organization

The Data-Centric AI "Movement"

Tools & methodologies for accelerating open-source dataset iteration:

- Tools that quantify and accelerate time to source and prepare high quality data
- Tools that ensure that the data is labeled consistently, such as label consensus
- Tools that make improving data quality more systematic
- Tools that automate the creation of high quality supervised learning training data from low quality resources, such as forced alignment in speech recognition
- Tools that produce consistent and low noise data samples, or remove labeling noise or inconsistencies from existing data
- Tools for controlling what goes into the dataset and for making high level edits efficiently to very large datasets, e.g. adding new words, languages, or accents to speech datasets with thousands of hours
- Search methods for finding suitably licensed datasets based on public resources
- Tools for creating training datasets for small data problems, or for rare classes in the long tail of big data problems
- Tools for timely incorporation of feedback from production systems into datasets
- Tools for understanding dataset coverage of important classes, and editing them to cover newly identified important cases
- Dataset importers that allow easy combination and composition of existing datasets
- Dataset exporters that make the data consumable for models and interface with model training and inference systems such as webdataset.
- System architectures and interfaces that enable composition of dataset tools such as, MLCube, Docker, Airflow

https://datacentricai.org/

Sourcing Stage of DS Lifestyle

- Sourcing involves 4 high-level groups of activities