MAS DSE 260: Capstone Project

İlkay ALTINTAŞ, Ph.D.

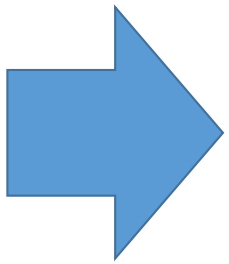
Lecture 4: Defining Your Hypothesis and Minimum Viable Modeling Product

Today's Topics

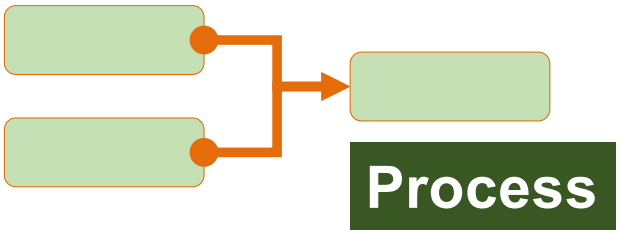
1. Reviewing where we are
2. STEP IV: Defining Your Hypothesis and Minimum Viable Modeling Product
3. Report IV Format : DUE 3/4/21
4. Presentation II Guidelines : DUE 3/5/22

Process Roadmap (260 A)

- ✓ Step 1: Understanding the Challenge
  - ✓ REPORT 1
- ✓ Step 2: Designing the Data Acquisition and Preparation Pipelines
  - ✓ REPORT 2
- Step 3: Exploring Data
  - ✓ PRESENTATION 1: 2/5
  - ✓ REPORT 3: due 2/18
- Step 4: Defining Your Hypothesis and Minimum Viable Modeling Product
  - REPORT 4: due 3/4
- Step 5: Creating a Solution Architecture for Modeling and Optimization
  - PRESENTATION 2: 3/5
  - FINAL WINTER REPORT: due 3/13



Collaborative Data Science Process



Basic Steps in a Data Science Process

ACQUIRE

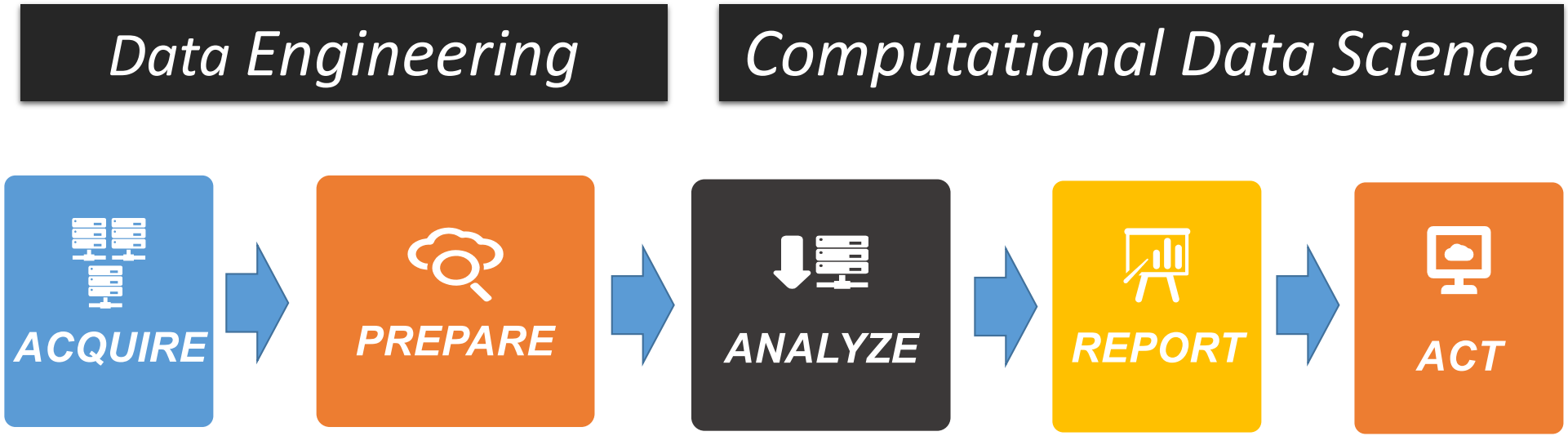
PREPARE

ANALYZE

REPORT

ACT

- Import raw dataset into your analytics platform
- Explore & Visualize
- Perform Data Cleaning
- Feature Selection
- Model Selection
- Analyze the results
- Present your findings
- Use them



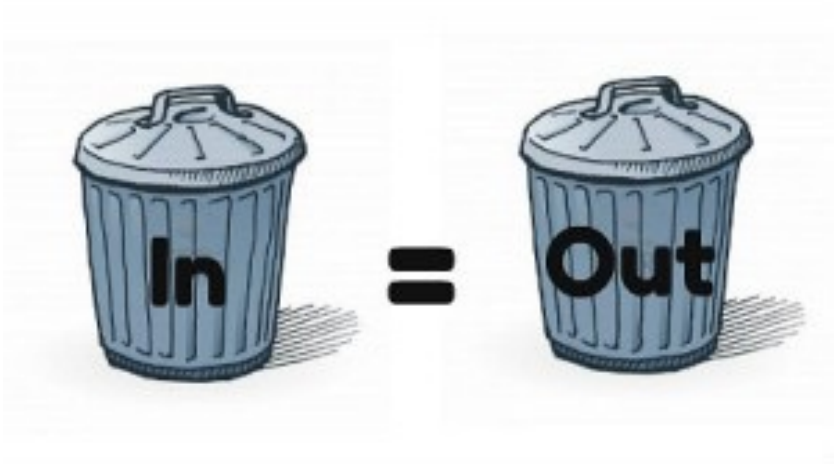
Many iterations and rollbacks between steps.

Process Roadmap

1. Understanding the Challenge
2. Designing the Data Acquisition and Preparation Pipelines
3. Exploring Data
4. Defining Your Hypothesis and Minimum Viable Modeling Product
5. Creating a Solution Architecture for Modeling and Optimization
6. Modeling and Visualization (Continued…)
7. Evaluating and Interpreting Modeling Results
8. Deploying a Robust and Scalable Solution
9. Developing a Communication Plan and Monitoring Dashboard
10. Business Integration and Optimization

Always Remember!

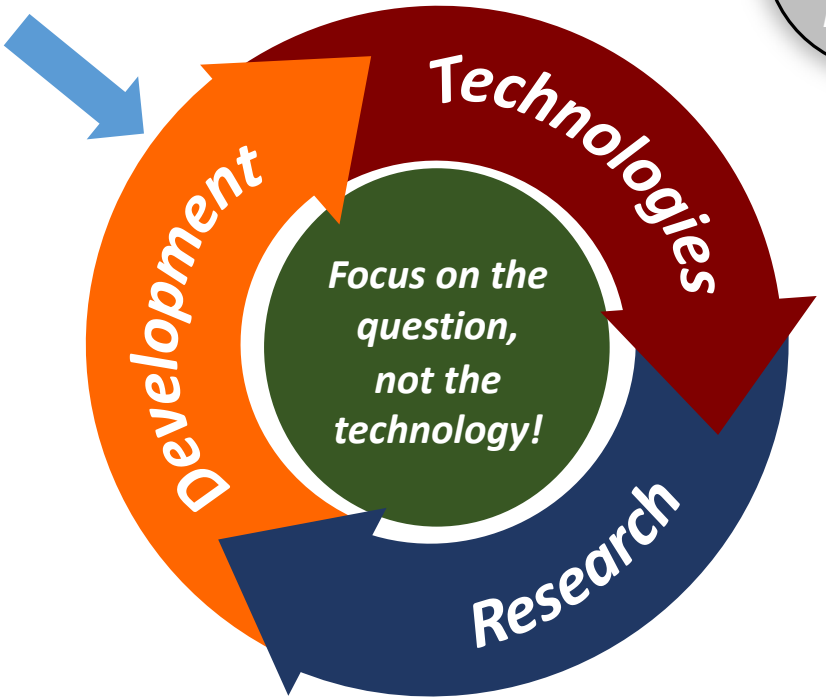
Garbage in = Garbage out

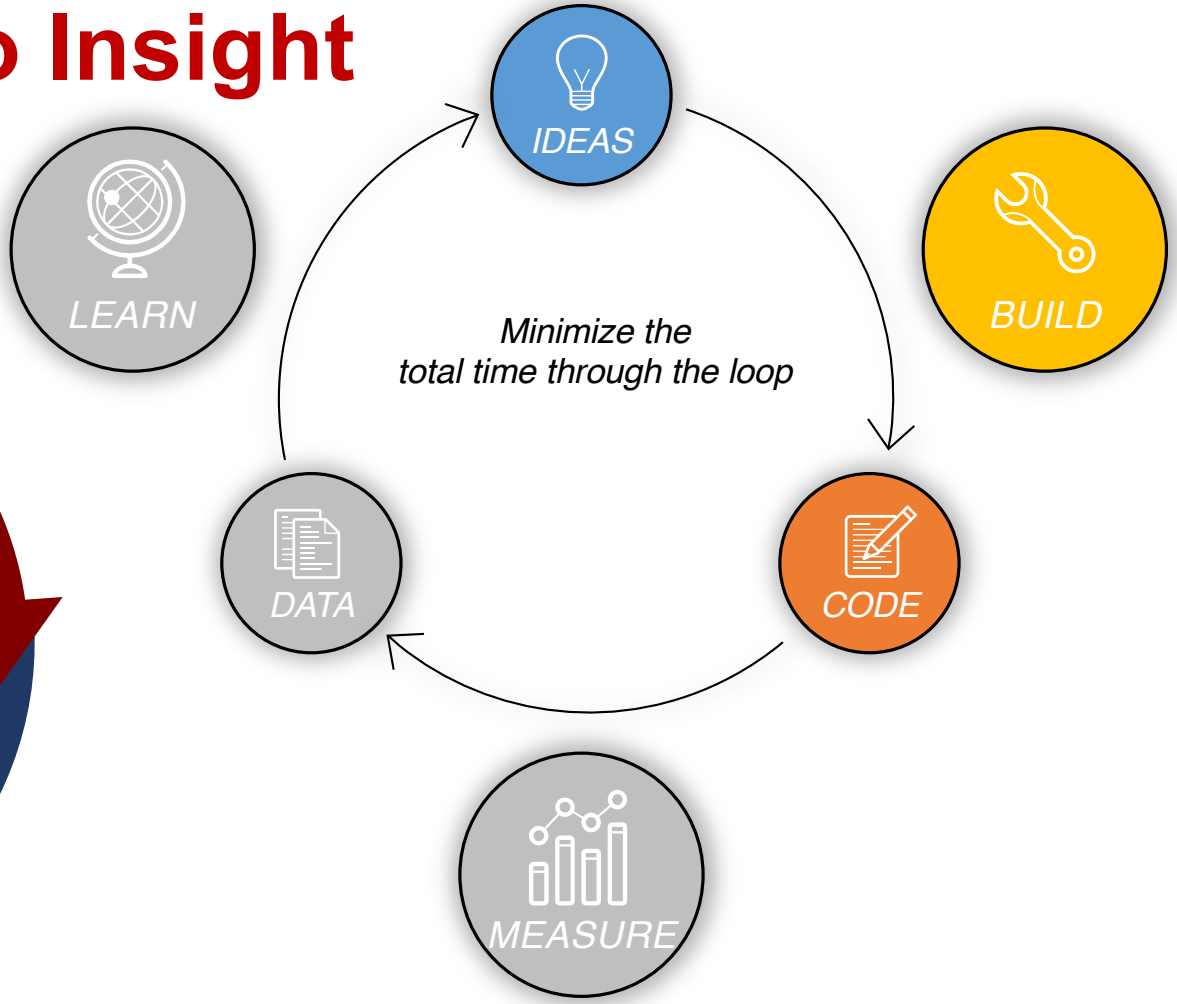


Data preparation is very important for meaningful analysis!

Purpose to Lead to Insight

Purpose ?





Defining Your Hypothesis

Hypothesis: There are three parts to it. Fill in the blanks.

1. EDA shows there is a problem at _____.
2. We can help the problem with solution _____.
3. We will know if we are right if metric _____ changes.

Possible to have more than one hypothesis.
List them and prioritize.

Creating your ~~Minimum Viable Product (MVP)~~ Minimum Viable Modeling Product (MVMP)

MVMP: the least amount of work to be done to validate/invalidate a hypothesis.

Assumption: Up to this point,

- Challenge/purpose is defined,
- Questions iterated,
- EDA well underway, and
- Data pipelines are functional.

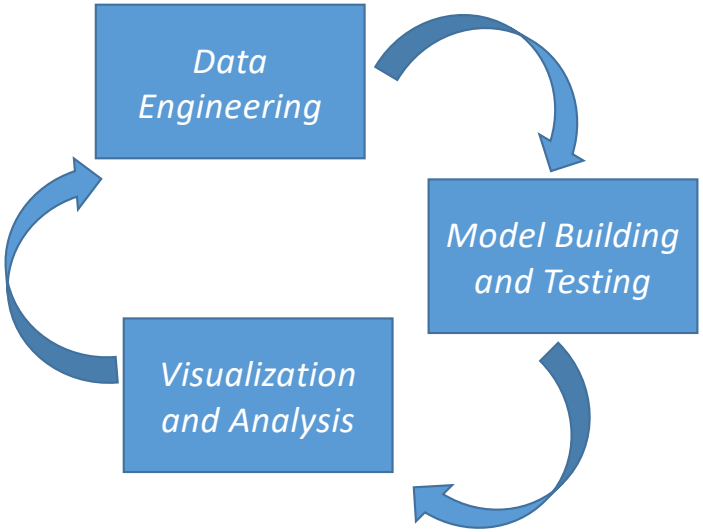
Apply design thinking to determine a hypothesis to test and MVMP to build.

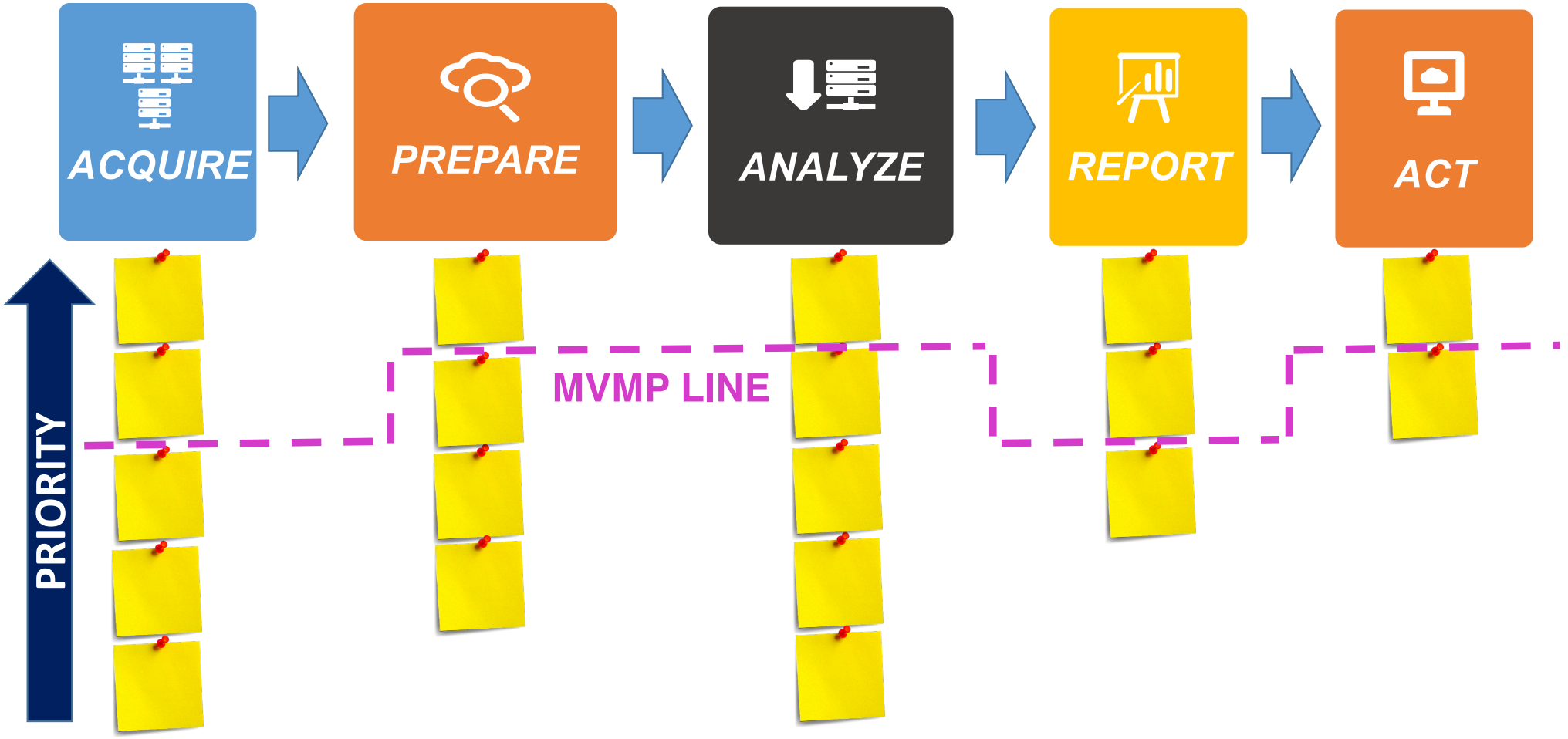
- Start with a small test case, reduce to a portion of the data or geospatial, etc.
- MVMP is not a proof of concept. It is a real product!

MVMP Development

MVMP development includes:

- More Data Engineering
- Modeling, Machine Learning and Visualization
- Evaluating and Interpreting Modeling Results




ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
PRIORITY
MVMP LINE

Step IV Report Guidelines

- Title, team members and advisor(s)
- Sections:
  - Hypothesis Definition
  - Analytic Approach for MVP
    - All possible inputs, targets and types of models -> Criteria for first cut of the product
  - Modeling
    - Models: Training and scoring, types of learners, learner parameterization, etc. as applicable
    - Results and Evaluation: Model validation, techniques used, Performance graphs, etc. as
    - Model Interpretation: Insights derived from results, significance of results, etc.
    - Next Steps for Modeling: What new features, datasets, techniques, etc. do you plan to add based on the results?
  - Bullets for each team member's individual contributions in Step 4
  - Any major updates to Steps 1 through 3 as a result of Step 4
- Keep it to 4-6 pages
- Due date: 3/4/2021 9am

Next Presentation (3/5/21)

- **Audience:** Data Science and Product Teams
- **Main points** to be made
  - How accurate/significant are the results?
  - What are the main insights so far?
  - What step in product design do you recommend based on these results?
  - How will this effect your data pipelines and solution architecture so far?
  - What are next steps for modeling based on the progress and why?
- **Don't forget** to include your team, problem definition and data definitions in the beginning of the presentation. Think story lines in the captions!

NEXT:

Think towards your Solution Architecture!

Designing Data Products

Data Products

"… a product that facilitates an end goal through the use of data"

-- DJ Patil, Former U.S. Chief Data Scientist in Data Jujitsu

What are data products?

Data Products are

systems models

that help us to

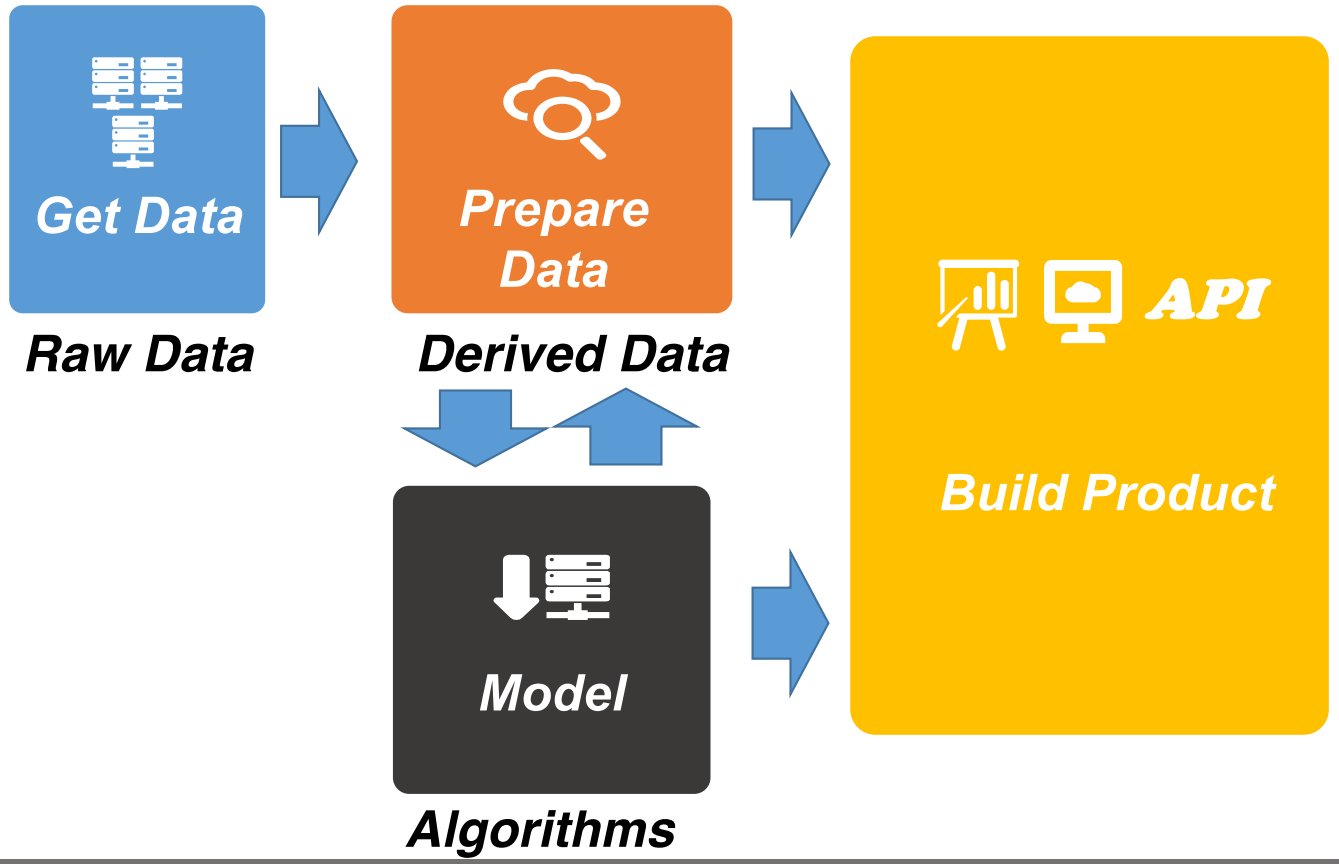
understand data

in order to

gain insights *and*

make predictions.

Going from raw data to a model using data science…



- *Visual Dashboards*
- *Web Interfaces*
- *Programming Interfaces*
- *Robotics platforms*

What are data products?

Data Products arise whenever we want to build systems that depend on ***predictive models.*** *For example:*

- *Predict users' future actions based on their past activities*
- *Recommend content to users that they are likely to consume*
- *Estimate demand for a product*

Examples of *Data Products* on the Web

(to name a few)

Examples – Recommender Systems

Modeling task: predict *what rating a person will give to an item*
e.g. rating(julian, Pitch Black) = ?

Data Product: *build a system to recommend products that people are interested in*


103 of 115 people found the following review helpful

★★★★★ **Excellent Sci-Fi**

Pitch Black was arguably one of the most overlooked films of the early year. Although the setting of the film could seem routine to a casual viewer(space travelers stranded and bickering on a hostile planet infested with alien nasties), director David Twohy's wonderful use of color and stylistic flourishes more than makes up for any trivial complaints. For...

Read the full review ›

Published on September 12, 2000 by Eric J. Pray

Modeling challenges: *how are opinions influenced by factors like time, gender, age, and location?*

Examples – Social Networks

Modeling task: predict *whether two users of a social network are likely to be friends*

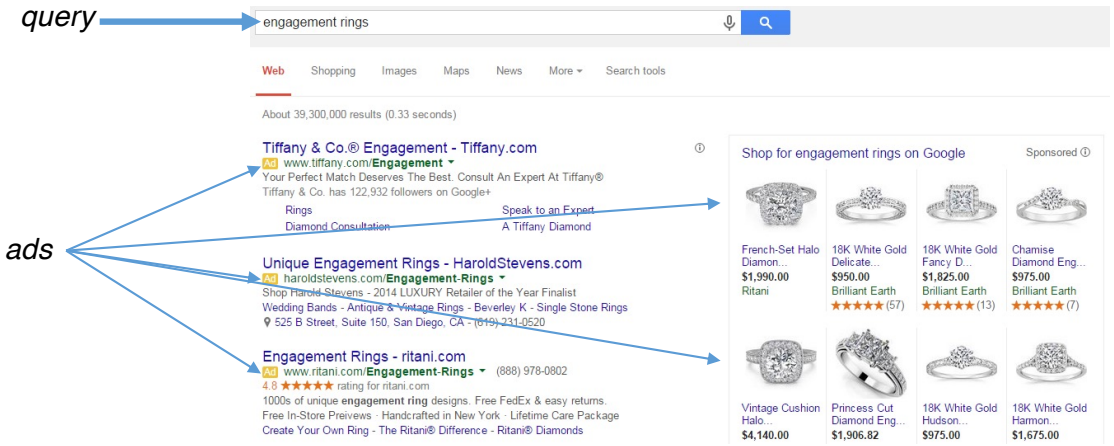
Data Product: *"people you may know" and friend recommendation systems*

Modeling challenges: *what are the features around which friendships form?*

Examples – Advertising

Modeling task: predict *whether I will click on an advertisement*

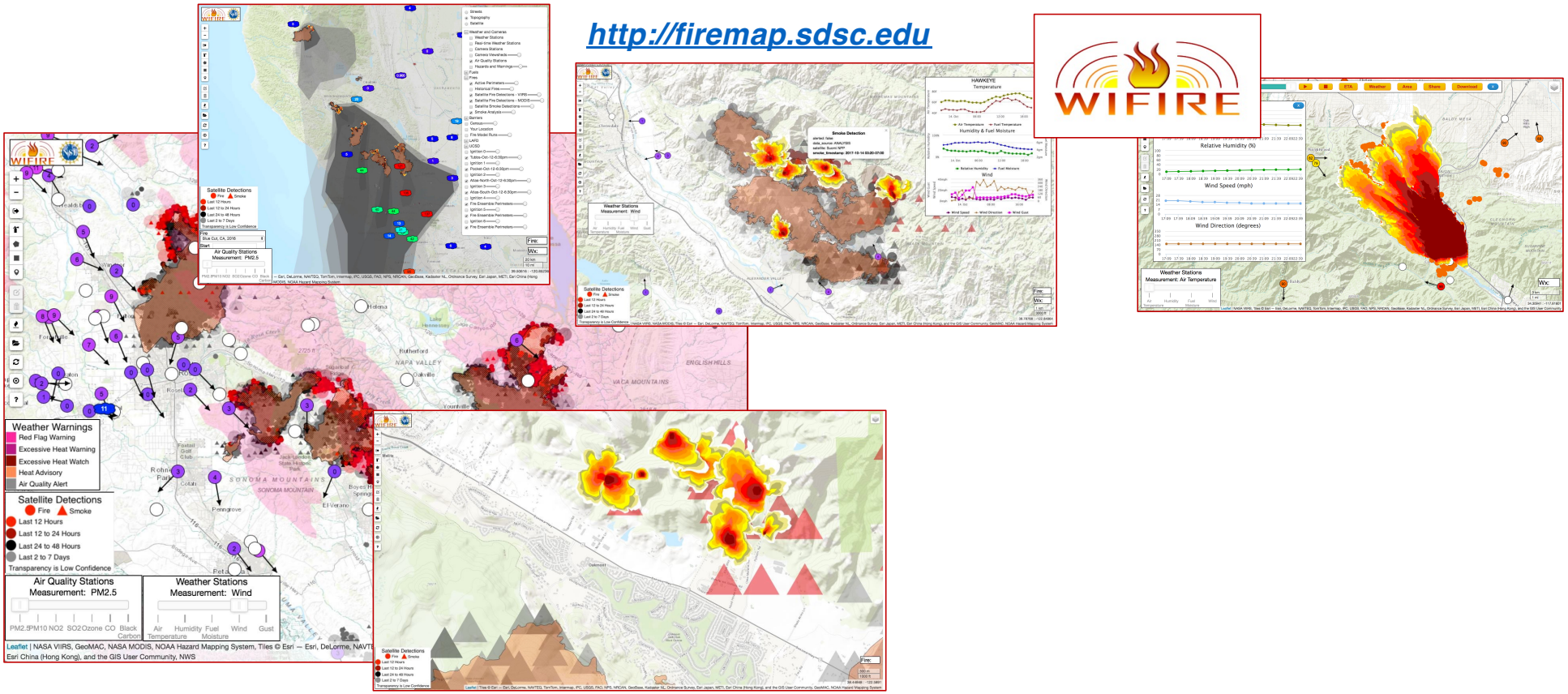
Data product: *Ad recommendation systems*



Modeling challenges: *what products tend to be purchased together, how do purchases change over time, etc.*

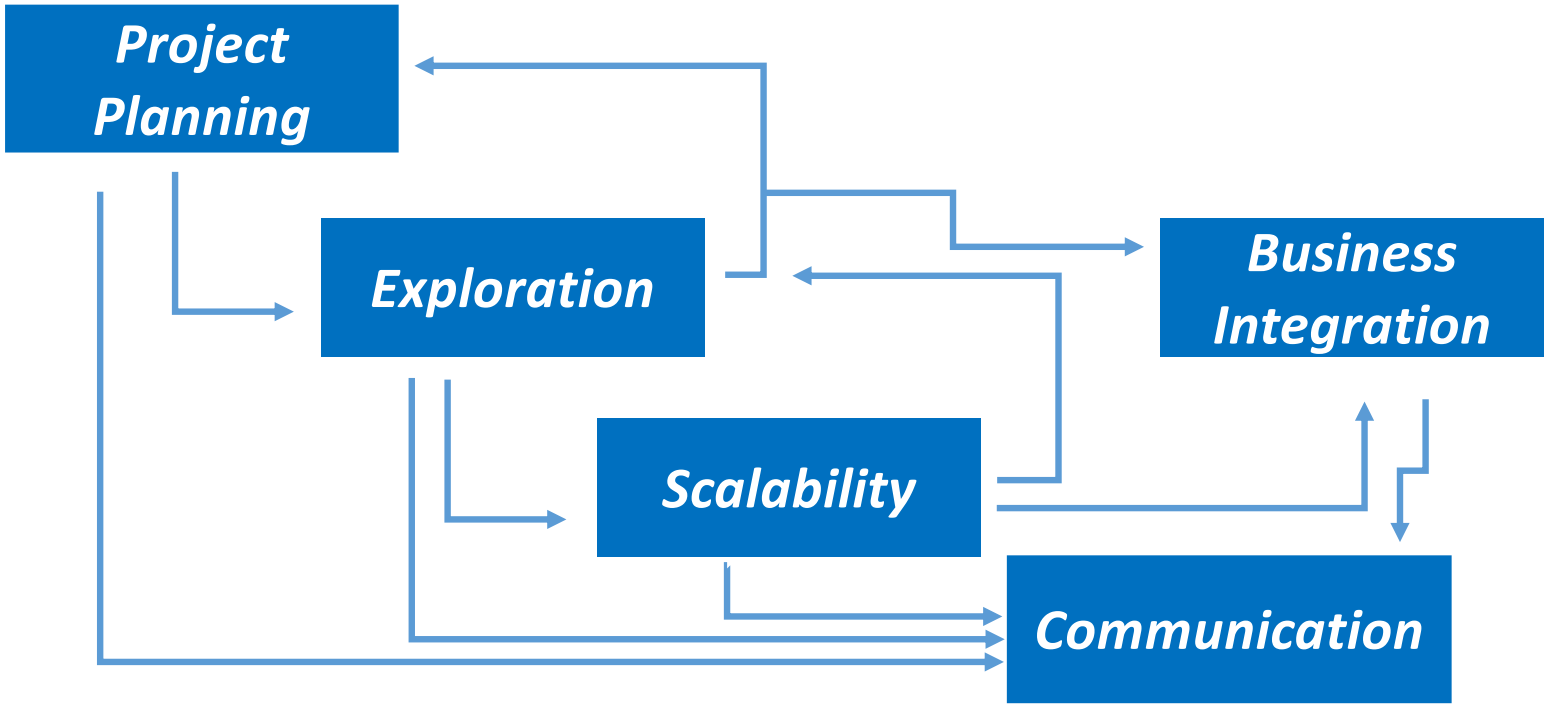
A Data Product for predictive wildfire modeling

http://firemap.sdsc.edu

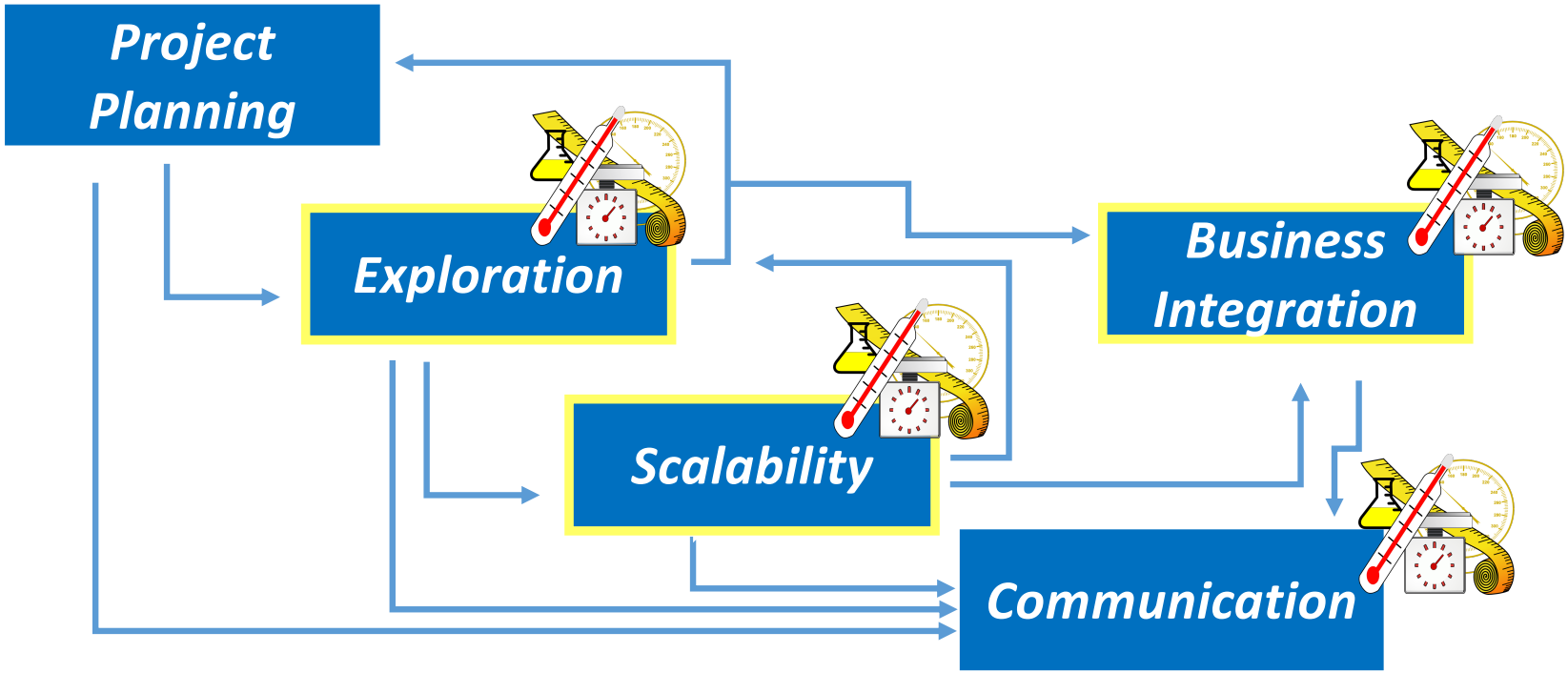


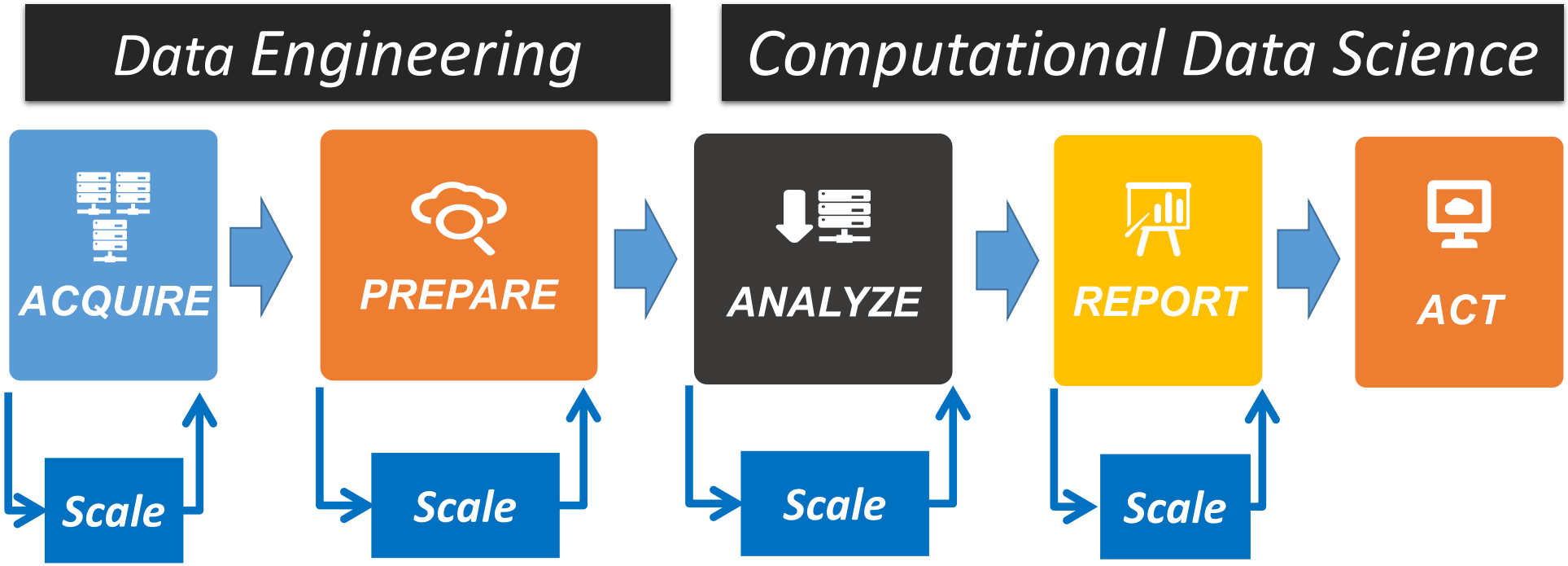
Designing an Engineering Process for Data Science

Good start, but the process starts even before acquiring data, involves scalability and constant iteration!

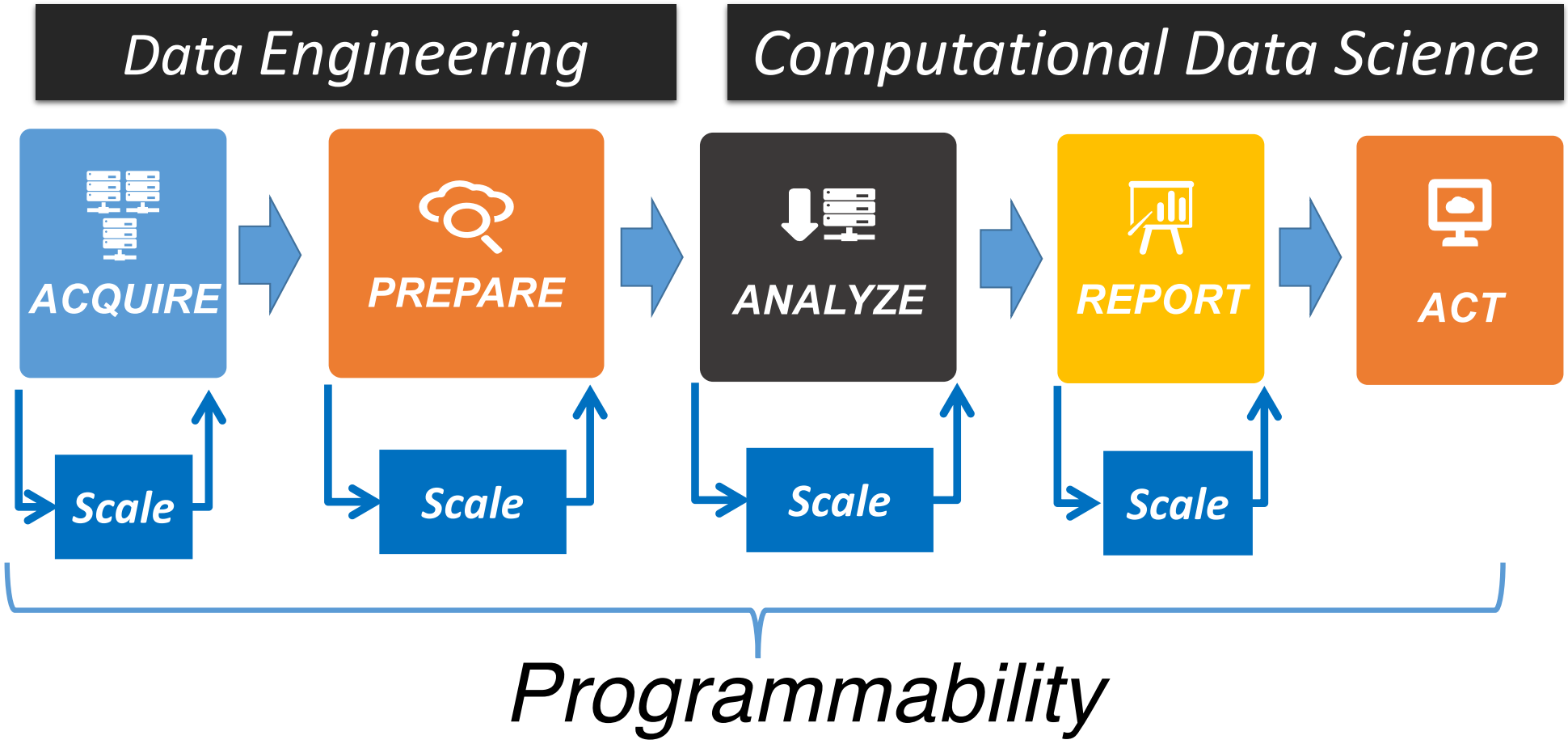


We need to measure metrics for each concern through the process.



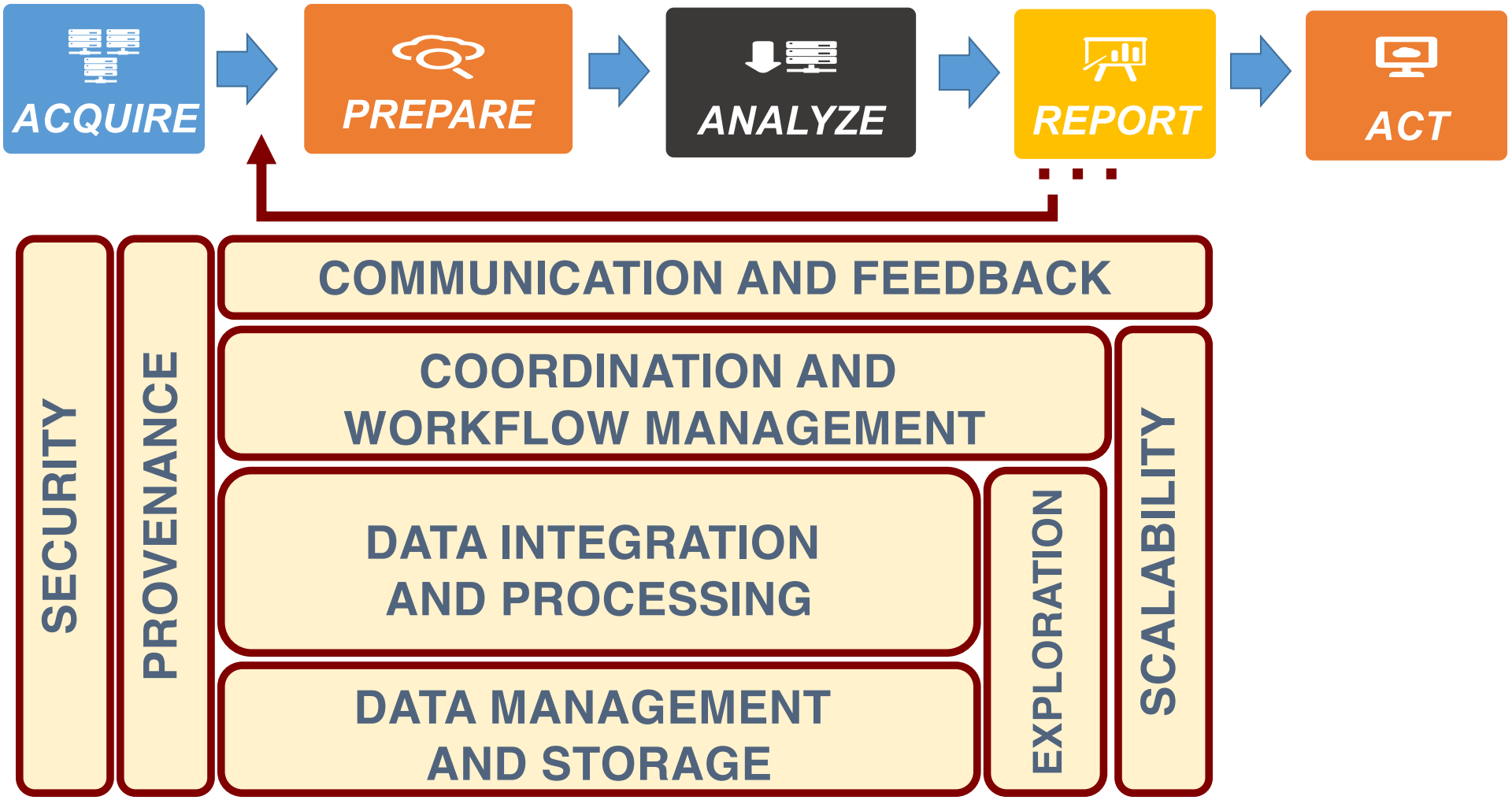


Many iterations and rollbacks between steps.


Data Engineering
Computational Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Scale
Scale
Scale
Scale
Programmability

Creating A Solution Architecture

Process-driven Solution Architectures and the Role of Workflows


ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
COMMUNICATION AND FEEDBACK
COORDINATION AND WORKFLOW MANAGEMENT
DATA INTEGRATION AND PROCESSING
DATA MANAGEMENT AND STORAGE
SECURITY
PROVENANCE
EXPLORATION
SCALABILITY

How do we make the data science process more dynamic and automatable?