MAS DSE 260: Capstone Project

İlkay ALTINTAŞ, Ph.D.

Lecture 1: Getting Started

Today's Topics

1. What is a capstone project?
   - Understanding class objectives
   - Setting expectations
   - Grading
2. Roadmap of our 10-step project
3. STEP I: Understanding the Challenge
4. Report I Format : DUE 1/20/21

What is a Capstone Project?

- **Objective:** To complete an end to end analysis of a large dataset with big data characteristics.
  - Includes
    - data collection,
    - data preparation,
    - exploratory analysis,
    - model building,
    - visualization, and
    - reporting.

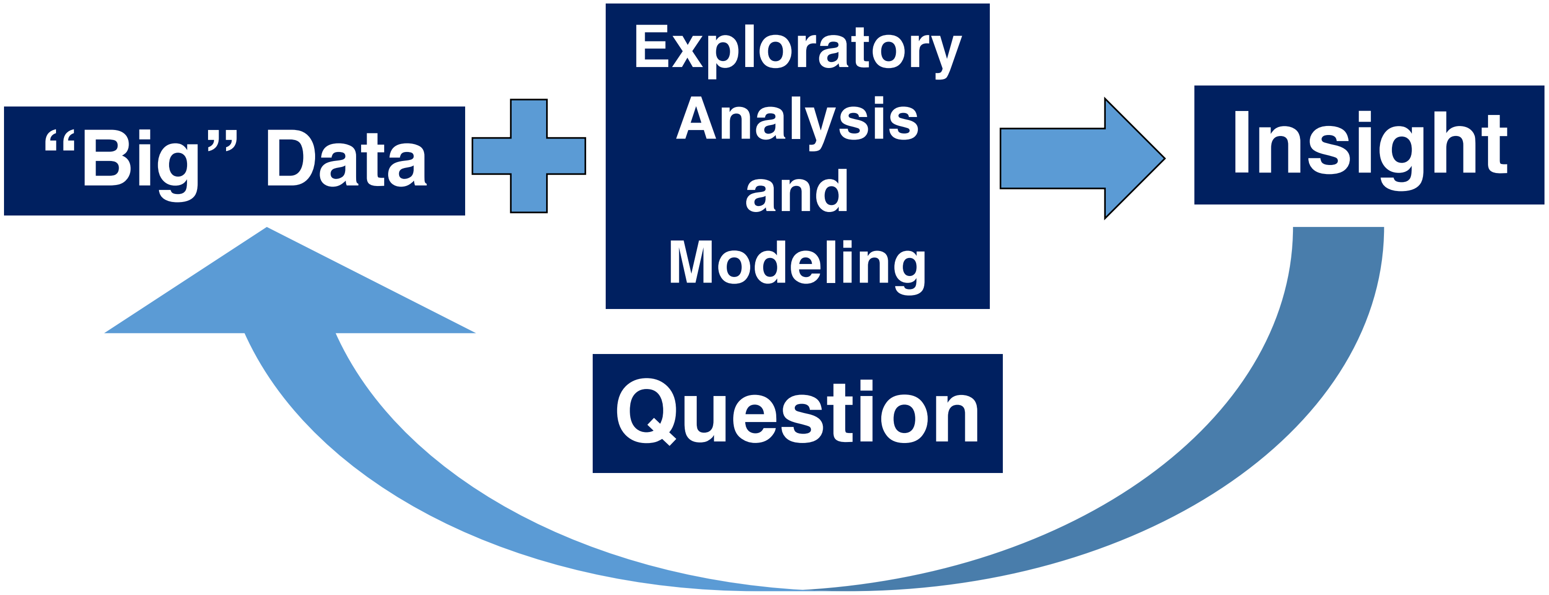
- **Products:**
  - Final report (preferred if publishable as a conference paper)
  - Output data products
  - Developed analytical tools/methods/workflows (if applicable)

Milestones for the Capstone Project

- **Second Year:**
  - Late Fall Quarter: Capstone project class is introduced. Advisors provide short summaries of projects so that students can identify who they want to work with. Students start to form teams, define project and find advisor.
  - Winter Quarter: Teams work on their projects and present progress reports. Suggested meeting schedule: once a month for 2 hours with advisor, twice a month with capstone faculty (i.e. Altintas).
  - Spring Quarter: Teams finalize their projects, including documentation and final report. Teams make open presentations to their peers, advisor and capstone faculty, and receive final grade.

Ultimate Goal

Insight = Data Product

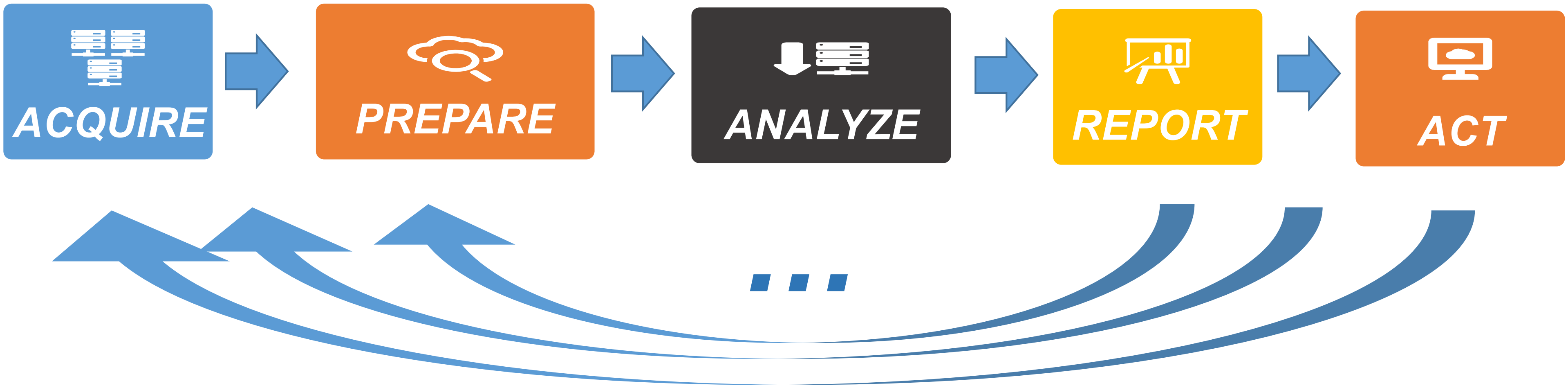


“Big” Data + Exploratory Analysis and Modeling → Insight

Question

We will do it through
10 deliverables and 5 presentations!

Approach: Focus on Process and Team Work

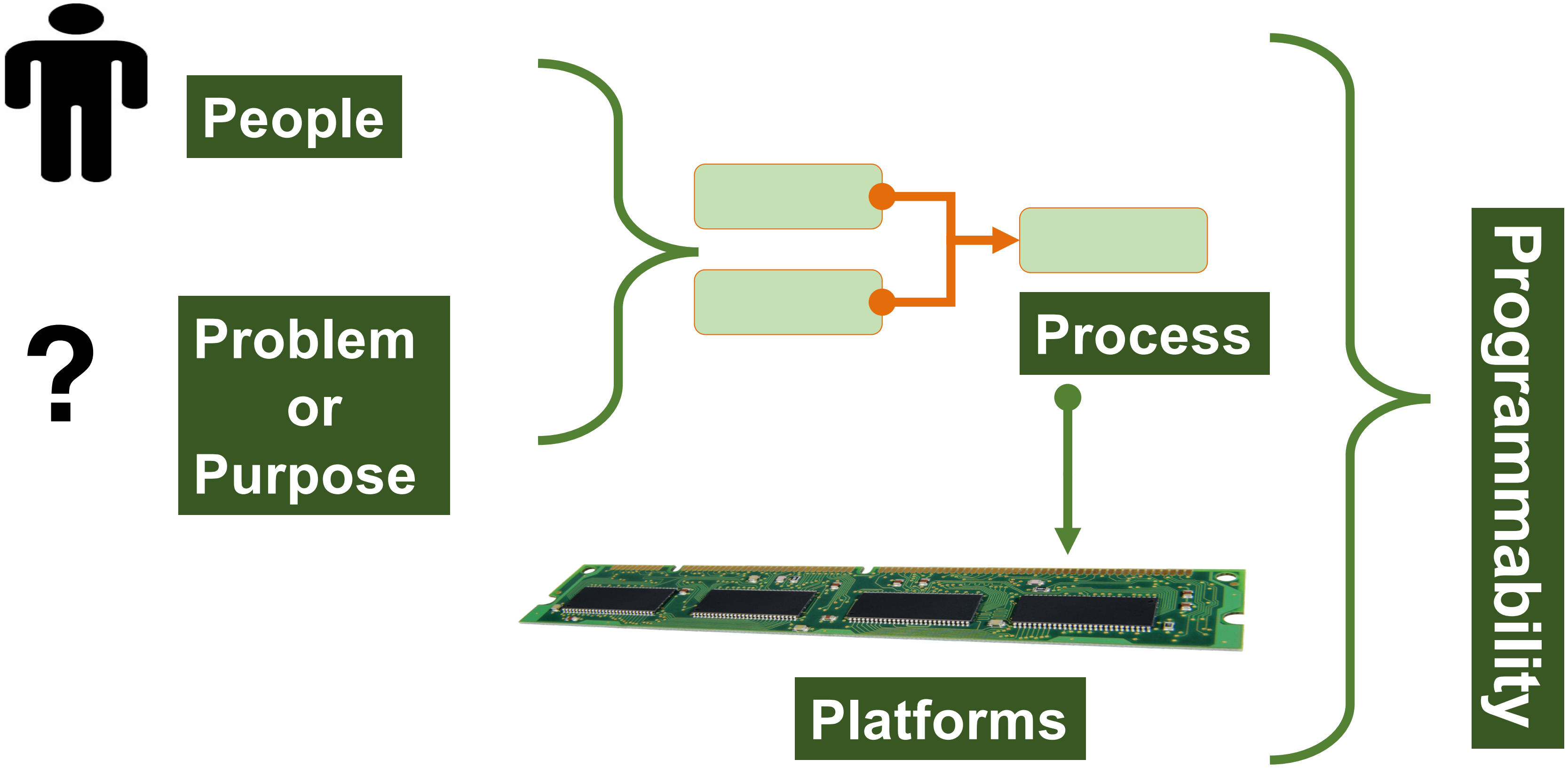


“The” Team

Expertise and skills often overlap, but nobody has it all!

- Data engineer
- Data analyst
- Methods expert
- Scalability and operations expert
- Business manager
- Business analyst
- Visualization and dashboard developer
- Solution architect
- Story teller/coordinator
- Project manager

PPoDS



People

?

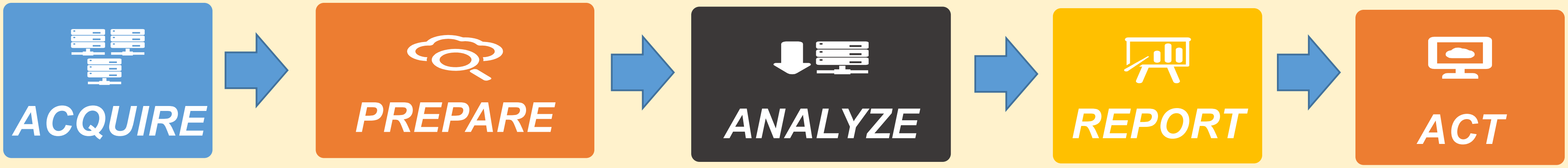
Problem or Purpose

Process

Platforms

Programmability

Create an Ecosystem that Enables Needs and Best Practices


ACQUIRE → PREPARE → ANALYZE → REPORT → ACT

- data-driven
- dynamic
- process-driven
- collaborative

- accountable
- reproducible
- interactive
- heterogeneous

Process Roadmap (260 A)

- Step 1: Understanding the Challenge
  - REPORT 1: due 1/20
- Step 2: Designing the Data Acquisition and Preparation Pipelines
  - REPORT 2: due 2/3
  - PRESENTATION 1: 2/5
- Step 3: Exploring Data
  - REPORT 3: due 2/17
- Step 4: Defining Your Hypothesis and Minimum Viable Modeling Product
  - REPORT 4: due 3/3
  - PRESENTATION 2: 3/5
- Step 5: Creating a Solution Architecture for Modeling and Optimization
  - REPORT 5: FINAL WINTER REPORT: due 3/12

Process Roadmap (260 B)

- Step 6: Modeling and Visualization
- Step 7: Evaluating and Interpreting Modeling Results
- Step 8: Deploying a Robust and Scalable Solution
- Step 9: Developing a Communication Plan and Monitoring Dashboard
- Step 10: Optimization

Grading

- Reports: 5% each, total 50% over two quarters
  - Along with the report, the following should be sent as a personal email:
    - Group member evaluation 1-5 for each report
    - Summary of personal contribution in the context of what was submitted
- Presentations: 5% each, total 20% over two quarters
- Final presentation and demo: 10%
- Final report: 5%
- Final poster: 5%
- Submission to the library: 2%
- Attendance during meetings and presentations: 3%
- Staying together as a group: 5%

STEP 1: Understanding the Challenge

(a.k.a. the PLANNING Phase)

Objectives

- Specify the key challenge that makes the project important
- Identify relevant data sources
- Distill specific and concise questions related to the challenge that can be solved using the identified datasets
- Define the project team responsibilities
- Define a baseline approach and success metrics

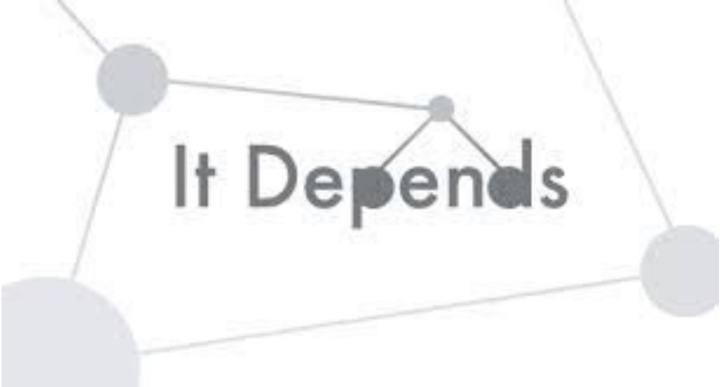
The project starts when a domain expert recognizes the opportunity and/or need.

Take Stock: Define vision and scope

- What is the exact need?
- What datasets are available?
- Who are the (current) stakeholders?
- What would you gain when the problem is solved?
- What are potential roadblocks? Think cultural, policy/privacy related, political, technical and data availability timelime.
- What is the timeline, resources and budget for the proposed project? e.g., 5 months , # of team members + $1000

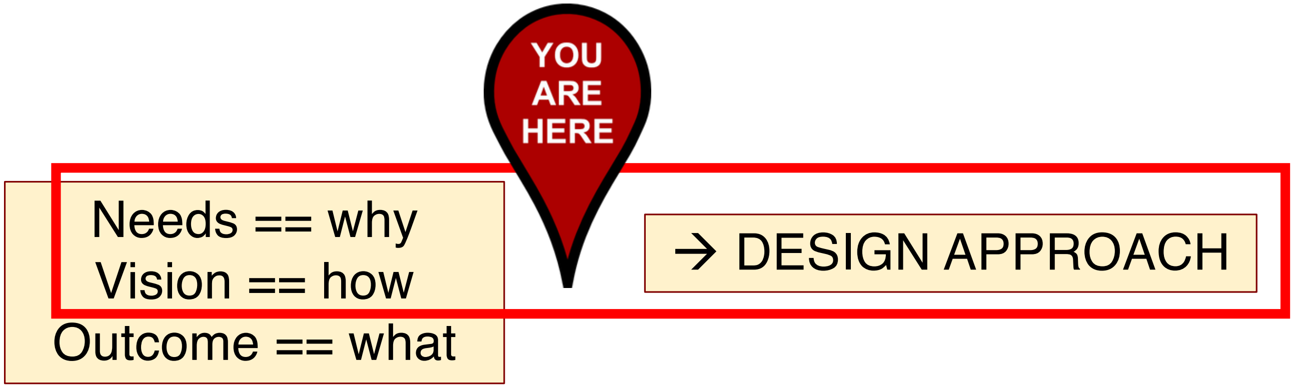
Start simple, iterate often, until a joint vision is defined!

What defines a data driven problem?



- When you know the question…
  - Look for factual answers
- When you have data or many potentially connected datasets…
  - Discover patterns
  - Look for known patterns
  - Explore connections and relationships
  - Derive questions

First focus on NEEDS to develop VISION!



- Think of the first step as a design effort
- Every discussion needs to have a purpose driven by needs
- Ask small concise and purposeful questions about the data entities to start exploring with data
- Do not focus on what yet!

Success-Oriented Design

- Vision for how success happens
- Design baseline success metrics
- Develop a data strategy based on vision and metrics

Domain needs + questions + data → Vision + metrics → Data strategy

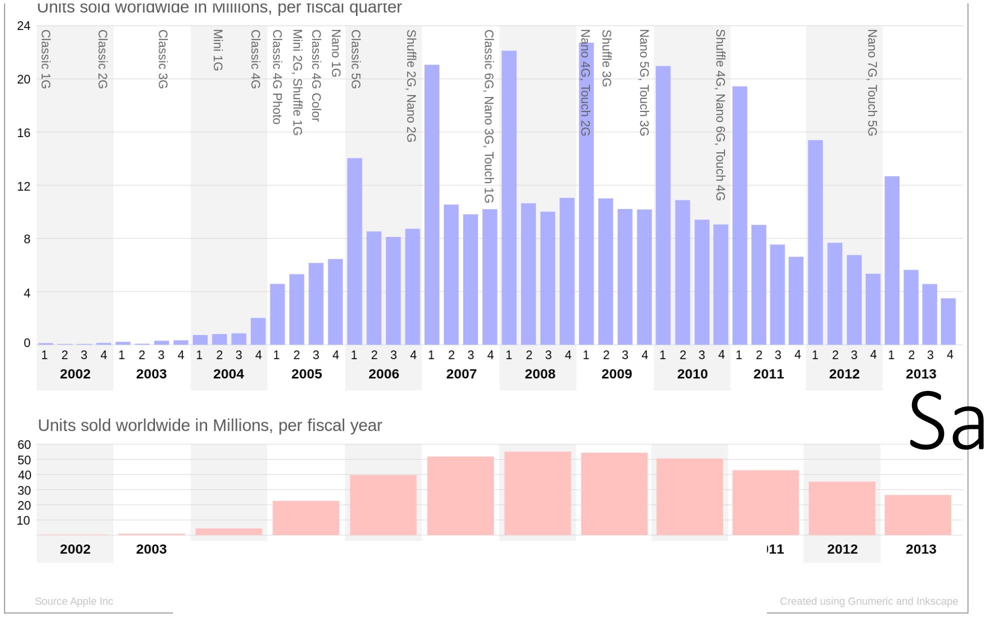
Division of Project Team Responsibilities

- Be flexible and ready to assume multiple roles
- Focus on your strengths, but also what you need to improve
- No bad tasks!
- Must assign:
  - a project coordinator/manager
  - a budget manager
  - a record keeper
- Each team member is expected to demonstrate both individual and collaborative work.

Asking the Right Question

> “A problem well defined is a problem half solved.”
>
> Charles F. Kettering

Define the Problem

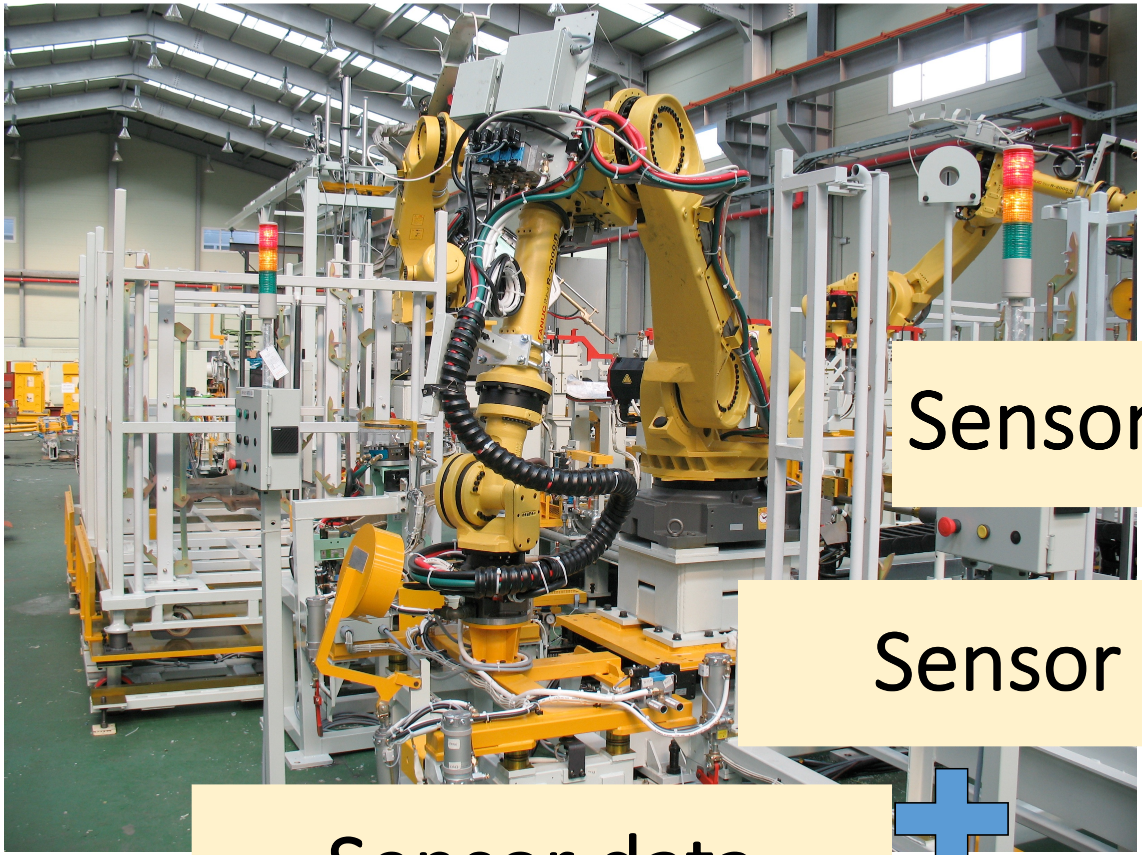


Evaluate a new product

Sales figures

Call center logs

Detect equipment failure



Sensor data

Sensor data

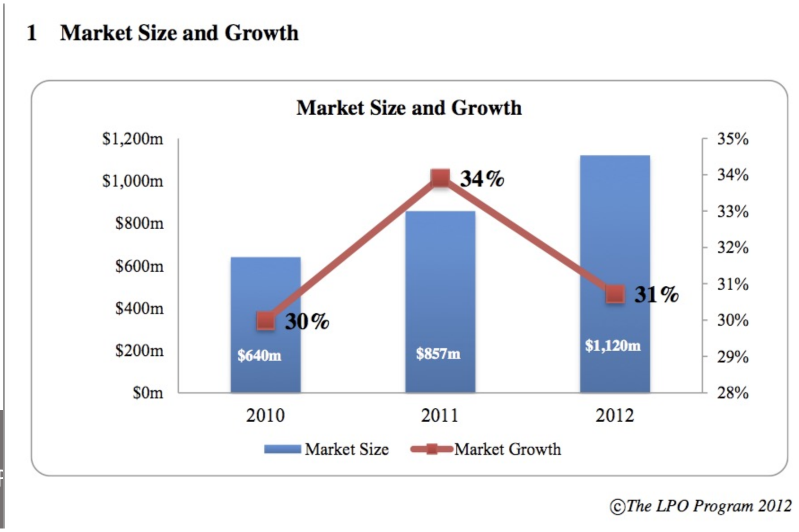
Sensor data

Customer data

+

Better targeted marketing

Marketing data



1 Market Size and Growth

©The LPO Program 2012

Assess the Situation

Risks

Benefits

Contingencies

Regulations

Resources

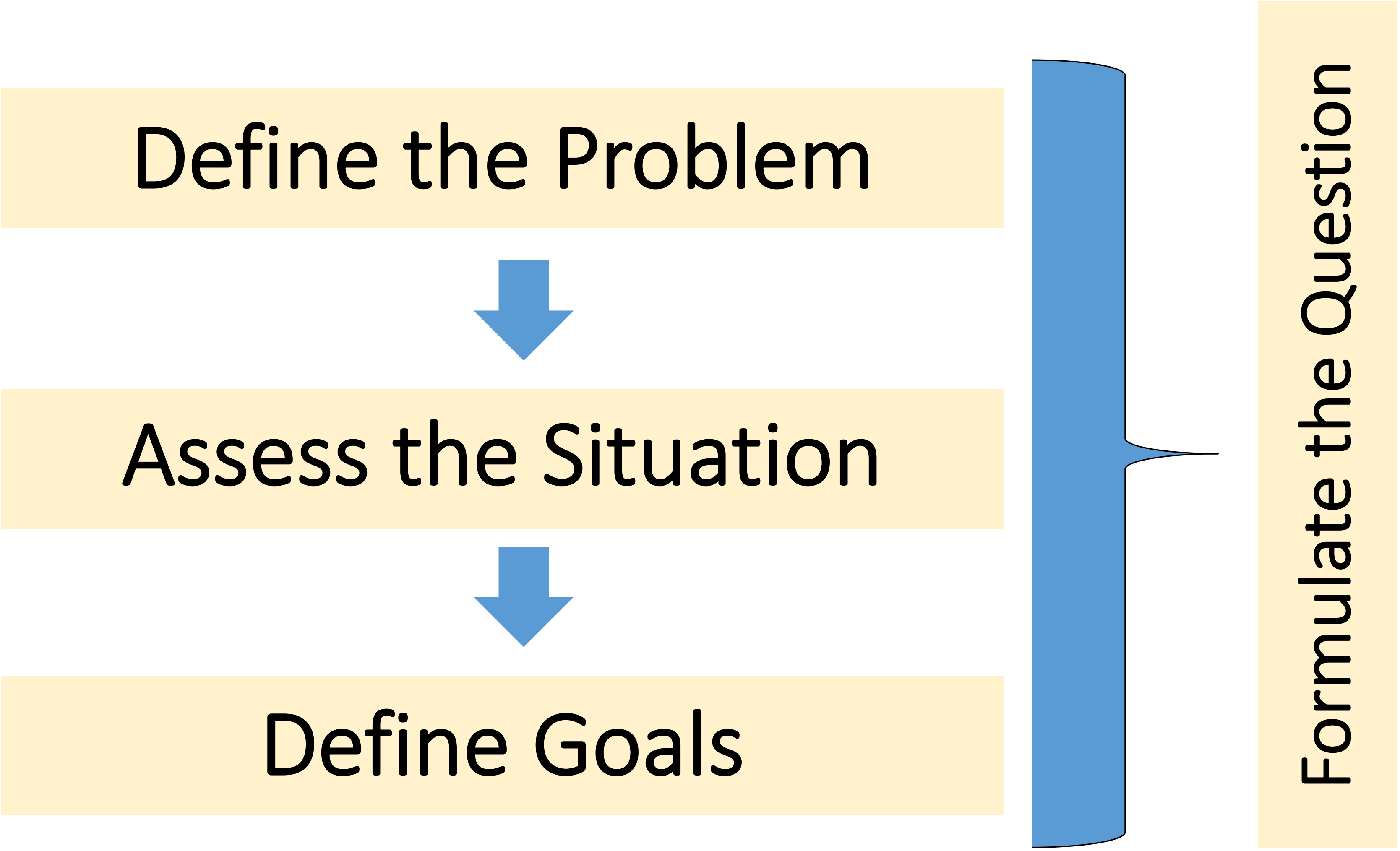
Requirements

Assess the Situation

Define Goals

Objectives

Criteria



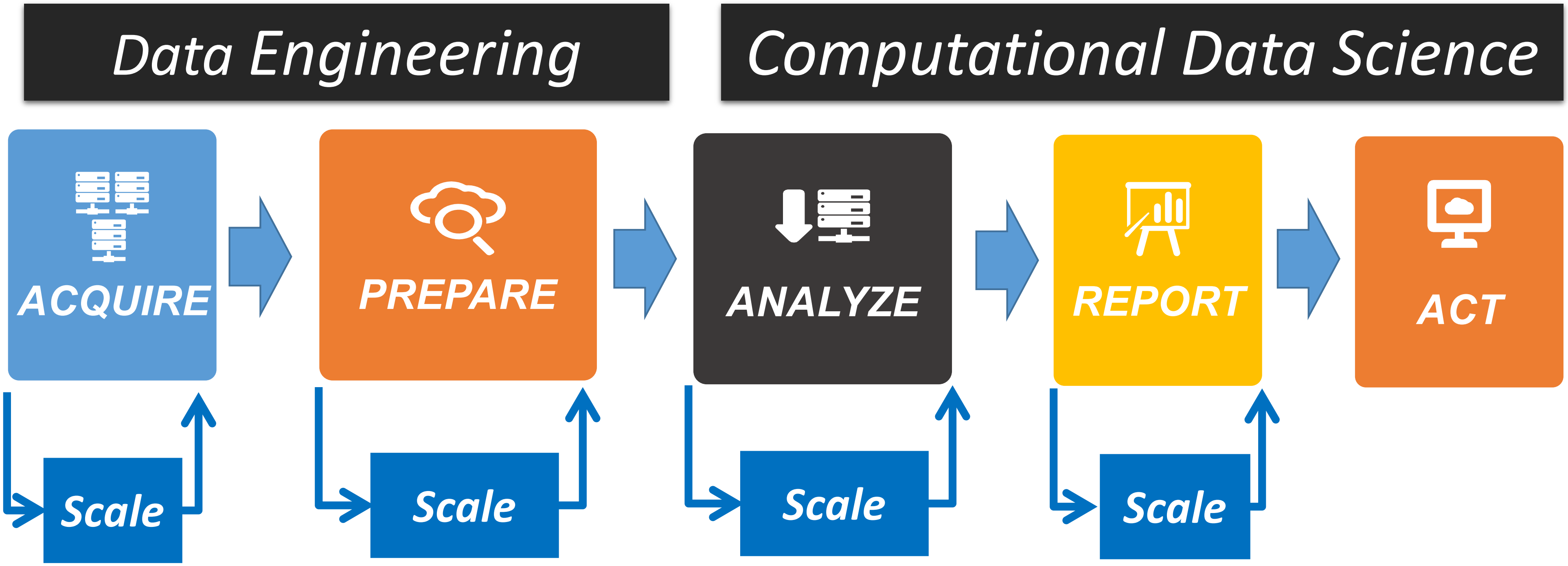
Define the Problem

Assess the Situation

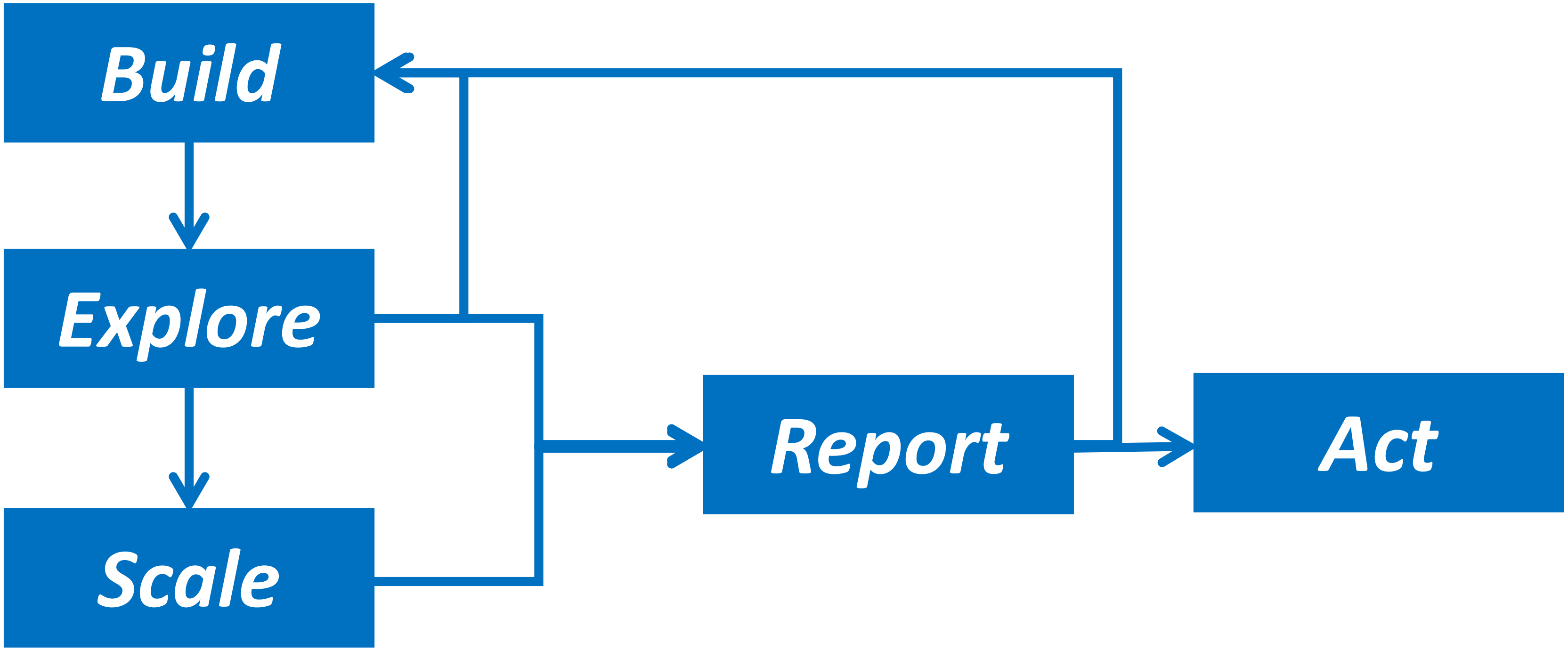
Define Goals

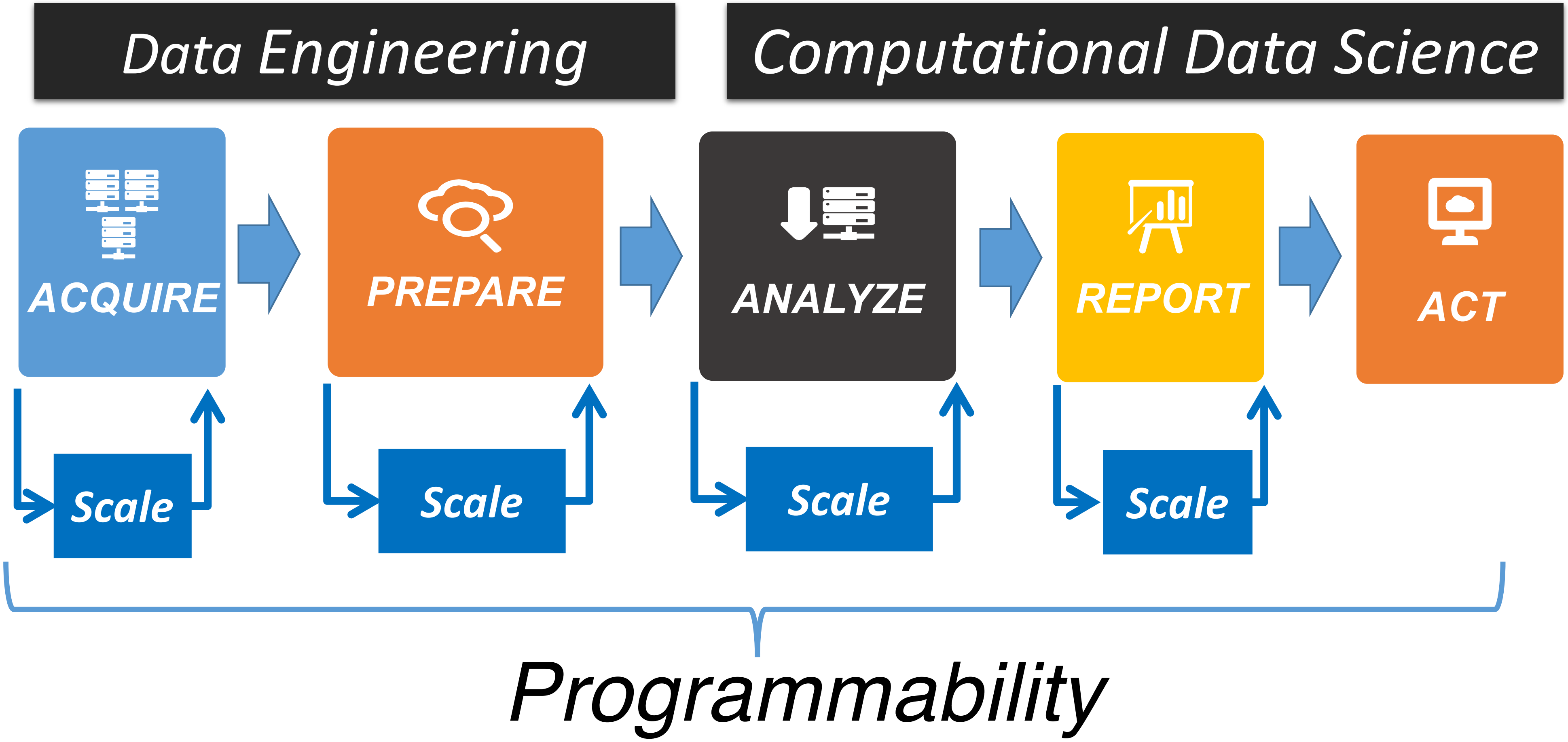
Formulate the Question

Data Science Process



Many iterations and rollbacks between steps.


Build
Explore
Scale
Report
Act


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

Report I Guidelines

- Upload to Canvas as a group submission
- A PDF document with the following
  - Title, team members and advisor(s)
  - Section Titles:
    - Challenge
    - Opportunities as a set of questions
    - Data sources
    - Approach
    - Team Roles and Responsibilities
    - Project Coordination and Communication Plan
    - Bullets for each team member's individual contributions in Step 1
- Keep it to 4-6 pages
- Due date: 1/20/21 midnight

Next… setting up your data process

1. STEP II: Designing the Data Acquisition and Preparation Pipelines
2. Report II: DUE 2/3/21

Questions?

ILKAY ALTINTAS, Ph.D.
Email: ialtintas@ucsd.edu