MAS DSE 260: Capstone Project

İlkay ALTINTAŞ, Ph.D.

Lecture 4: Defining Your Hypothesis and Minimum Viable Modeling Product

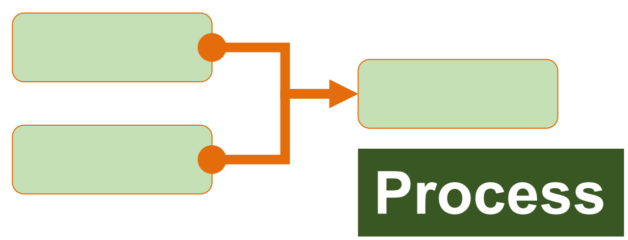
Today's Topics

1. Reviewing where we are
2. STEP IV: Exploring Data
3. Report IV Format : DUE 3/1/18

Process Roadmap (260 A)

- ✓ Step 1: Understanding the Challenge
  - ✓ REPORT 1: due 1/18
- ✓ Step 2: Designing the Data Acquisition and Preparation Pipelines
  - ✓ REPORT 2: due 2/1
- ✓ Step 3: Exploring Data
  - ✓ PRESENTATION 1: 2/3
  - ✓ REPORT 3: due 2/15
- Step 4: Defining Your Hypothesis and Minimum Viable Modeling Product
  - REPORT 4: due 3/1
- Step 5: Creating a Solution Architecture for Modeling and Optimization
  - PRESENTATION 2: 3/3
  - FINAL WINTER REPORT: due 3/16

Collaborative Data Science Process



Basic Steps in a Data Science Process

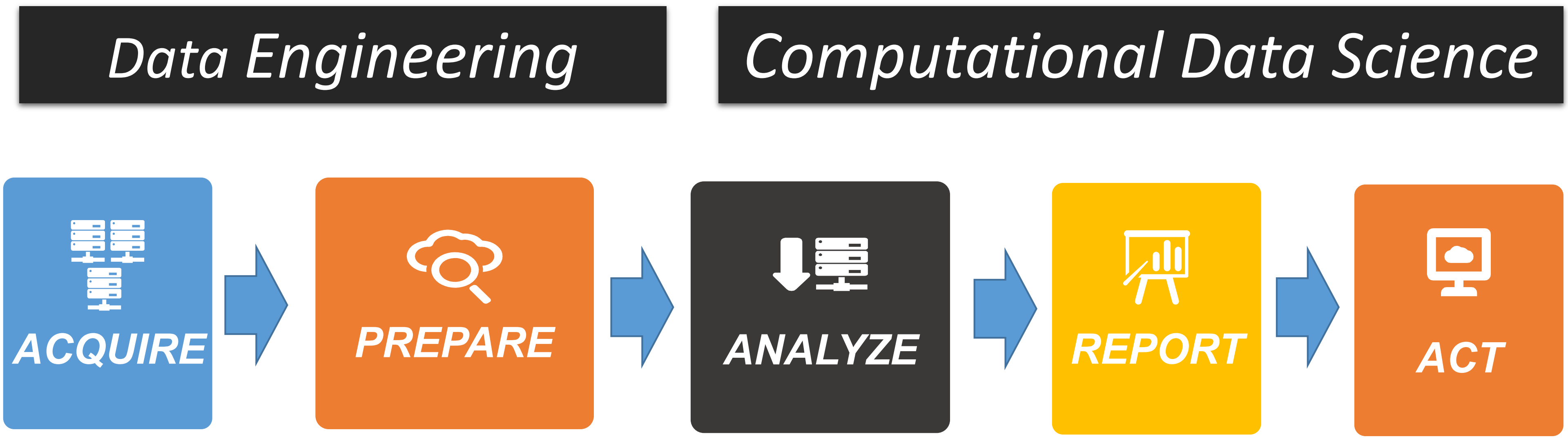
ACQUIRE
- Import raw dataset into your analytics platform

PREPARE
- Explore & Visualize
- Perform Data Cleaning

ANALYZE
- Feature Selection
- Model Selection
- Analyze the results

REPORT
- Present your findings

ACT
- Use them



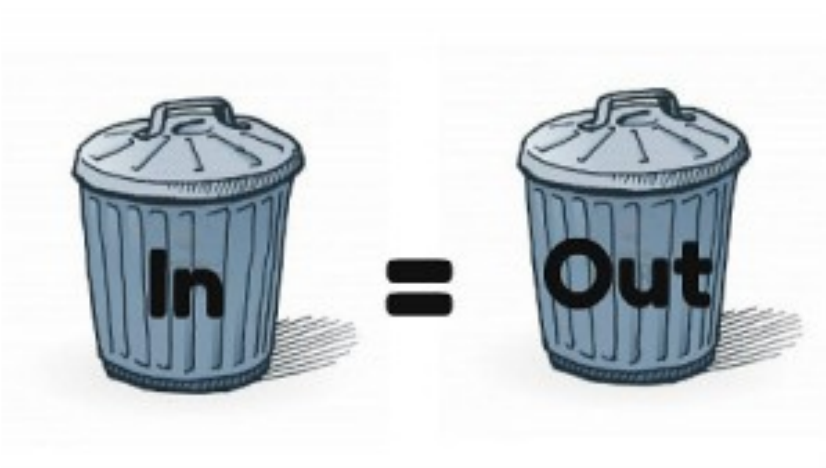
Many iterations and rollbacks between steps.

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

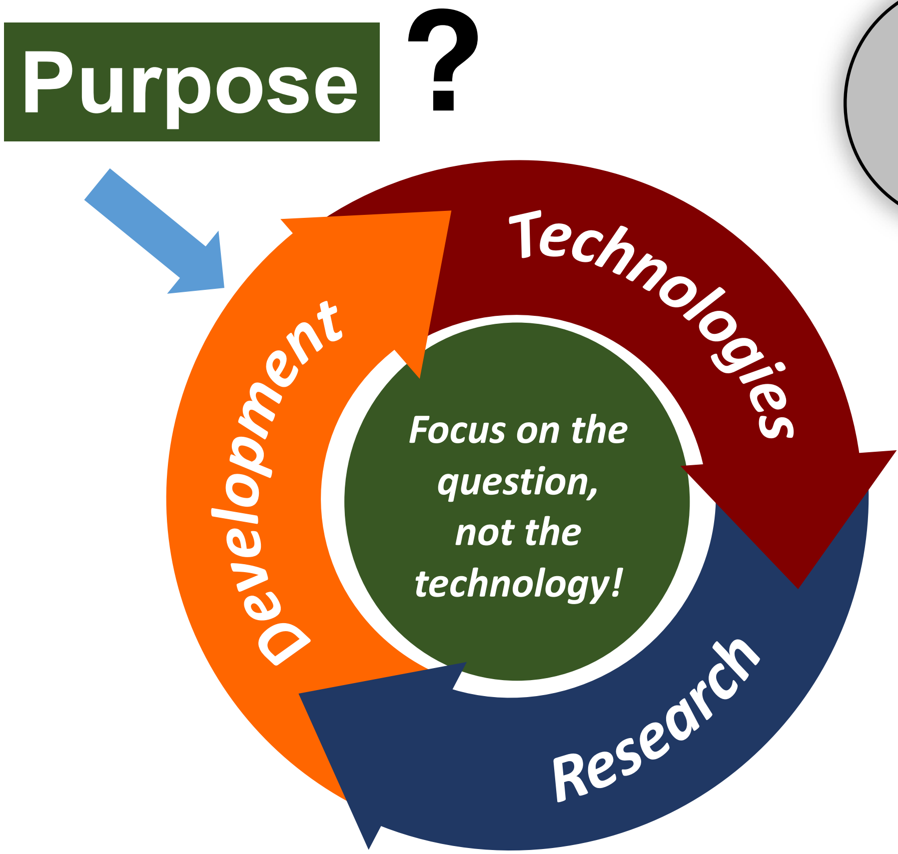
Always Remember!

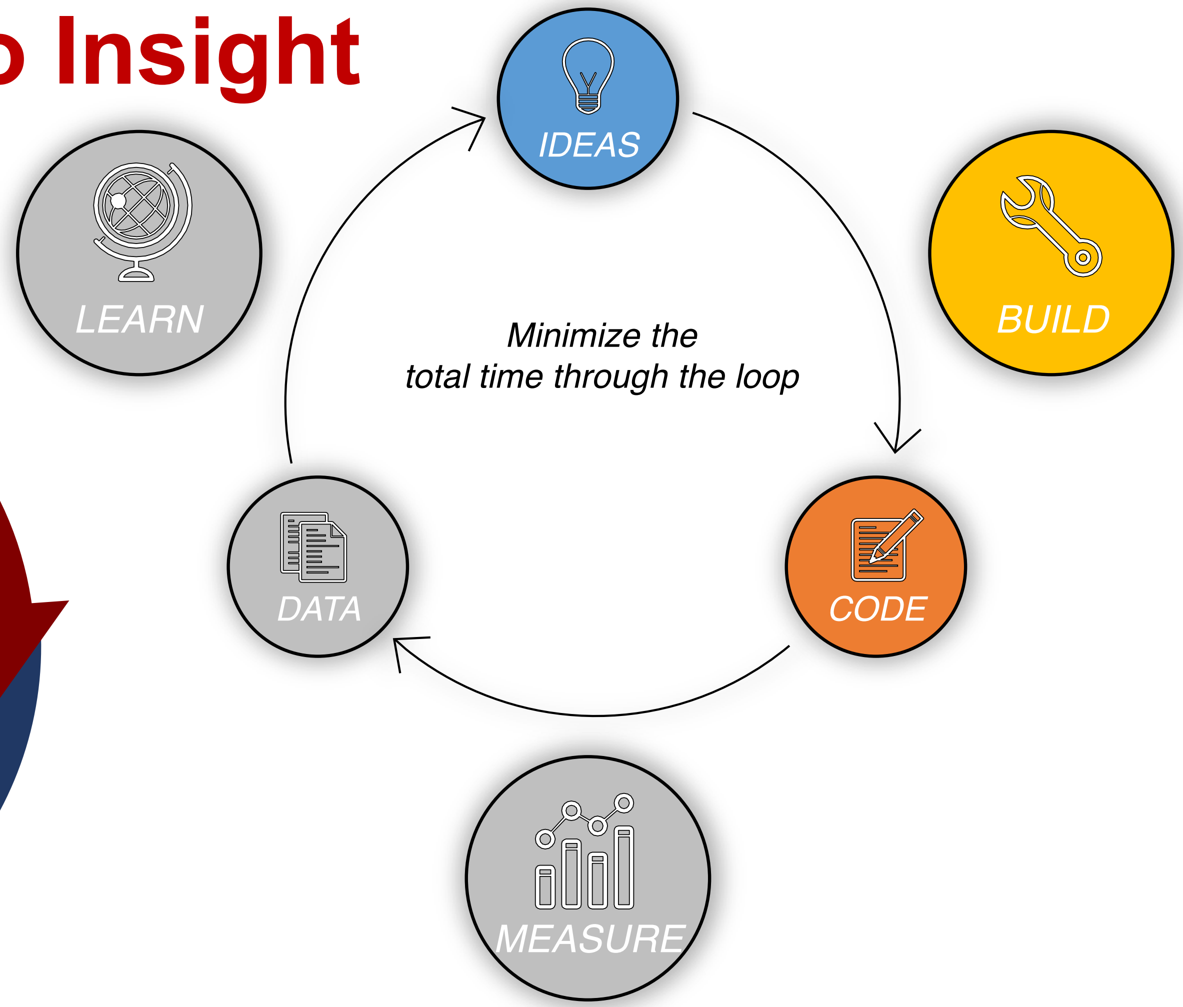
Garbage in = Garbage out



Data preparation is very important for meaningful analysis!

Purpose to Lead to Insight





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)

MVP: 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.

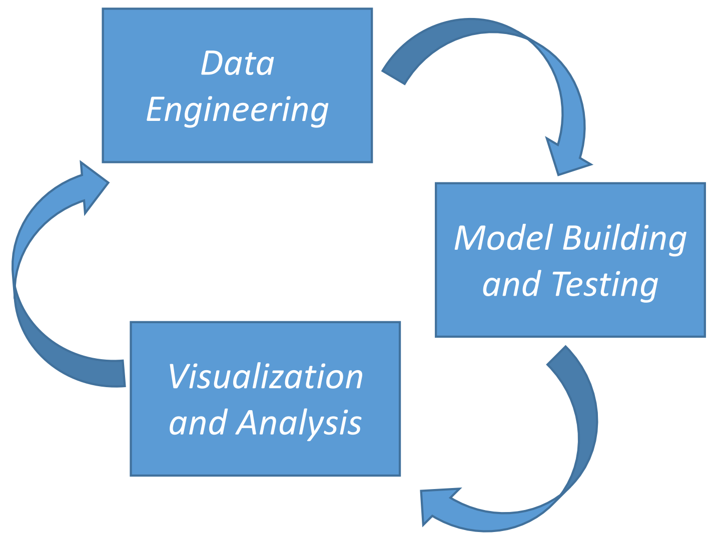
Apply design thinking to determine a hypothesis to test and MVP to build.

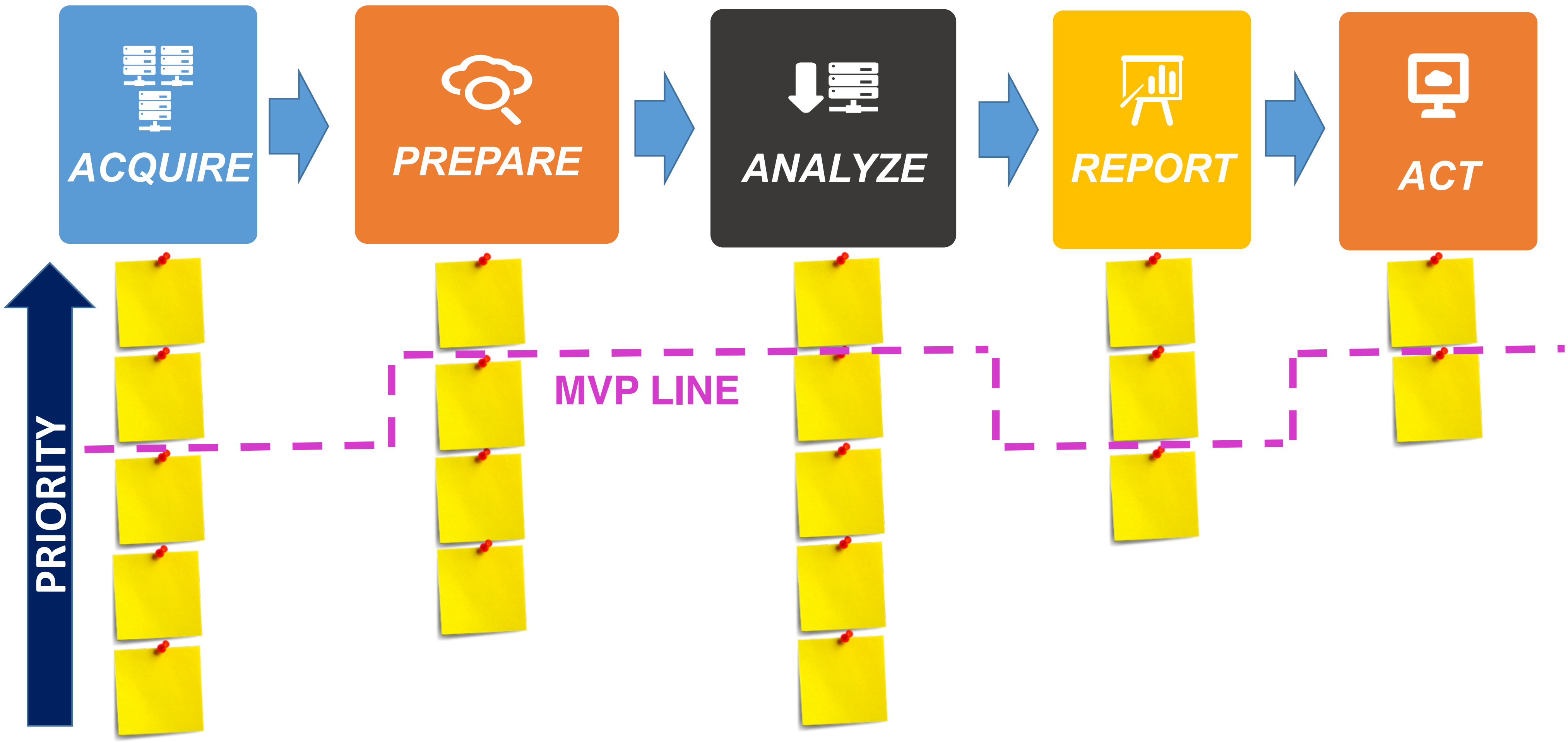
- Start with a small test case, reduce to a portion of the data or geospatial, etc.
- MVP is not a proof of concept. It is a real product!

MVP Development

MVP development includes:

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




ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
PRIORITY
MVP 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: 3/1/2018 midnight

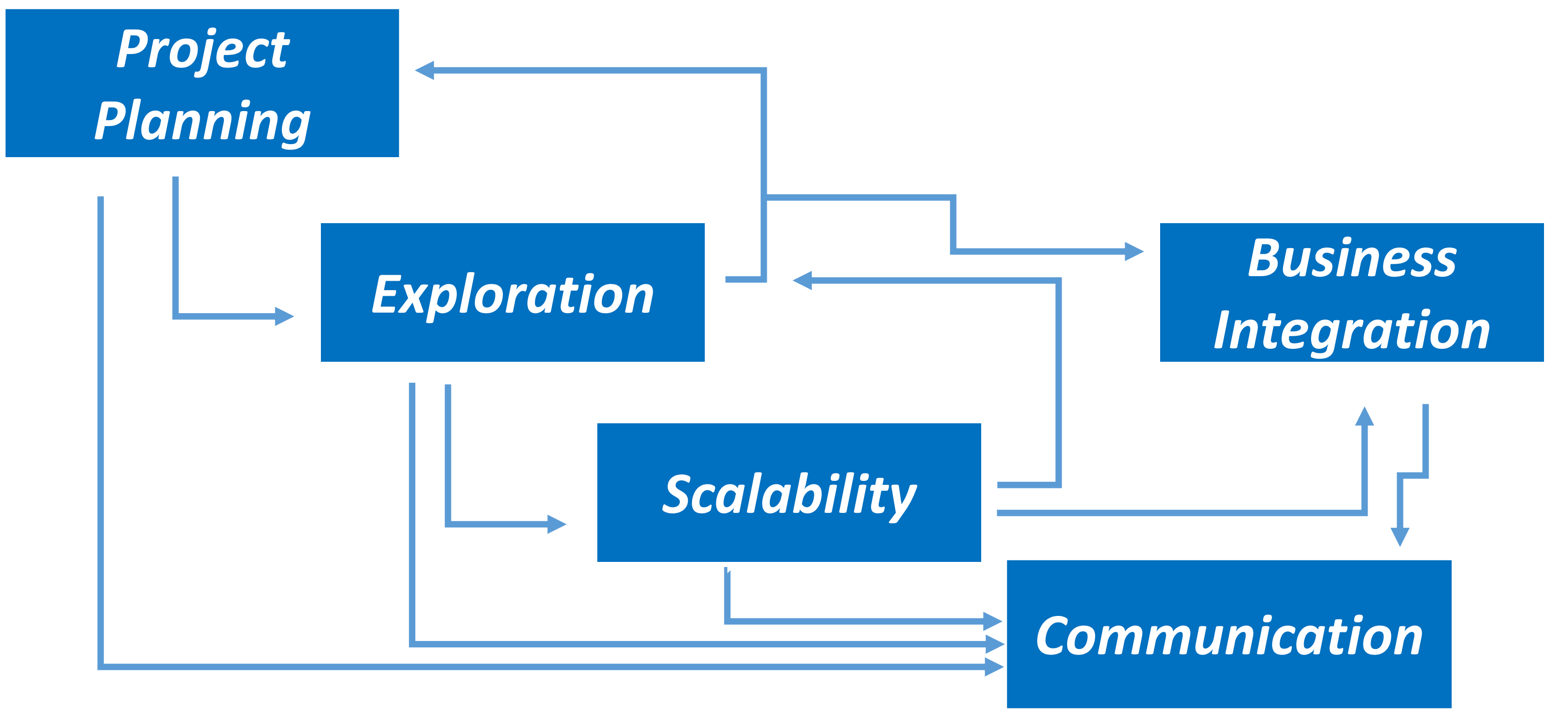
Next Presentation (3/3/18)

- **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!

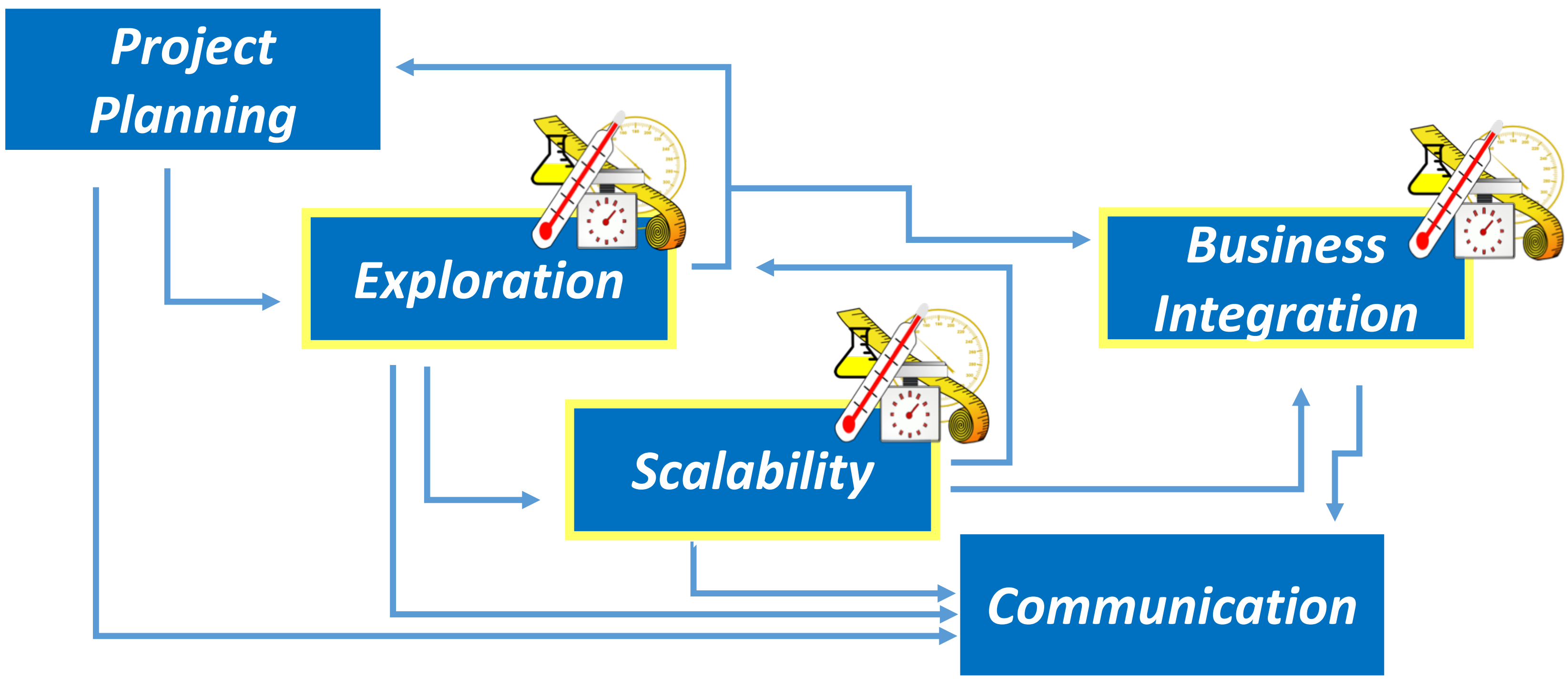
NEXT:

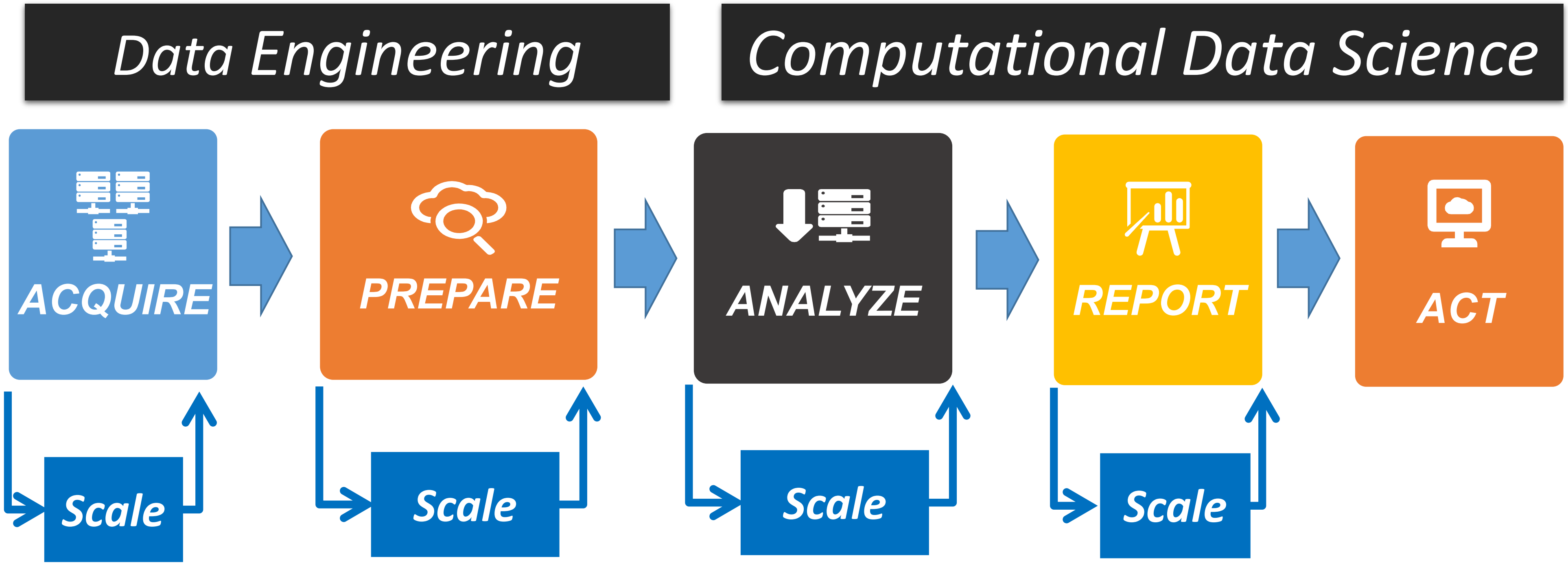
Think towards your Solution Architecture!

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

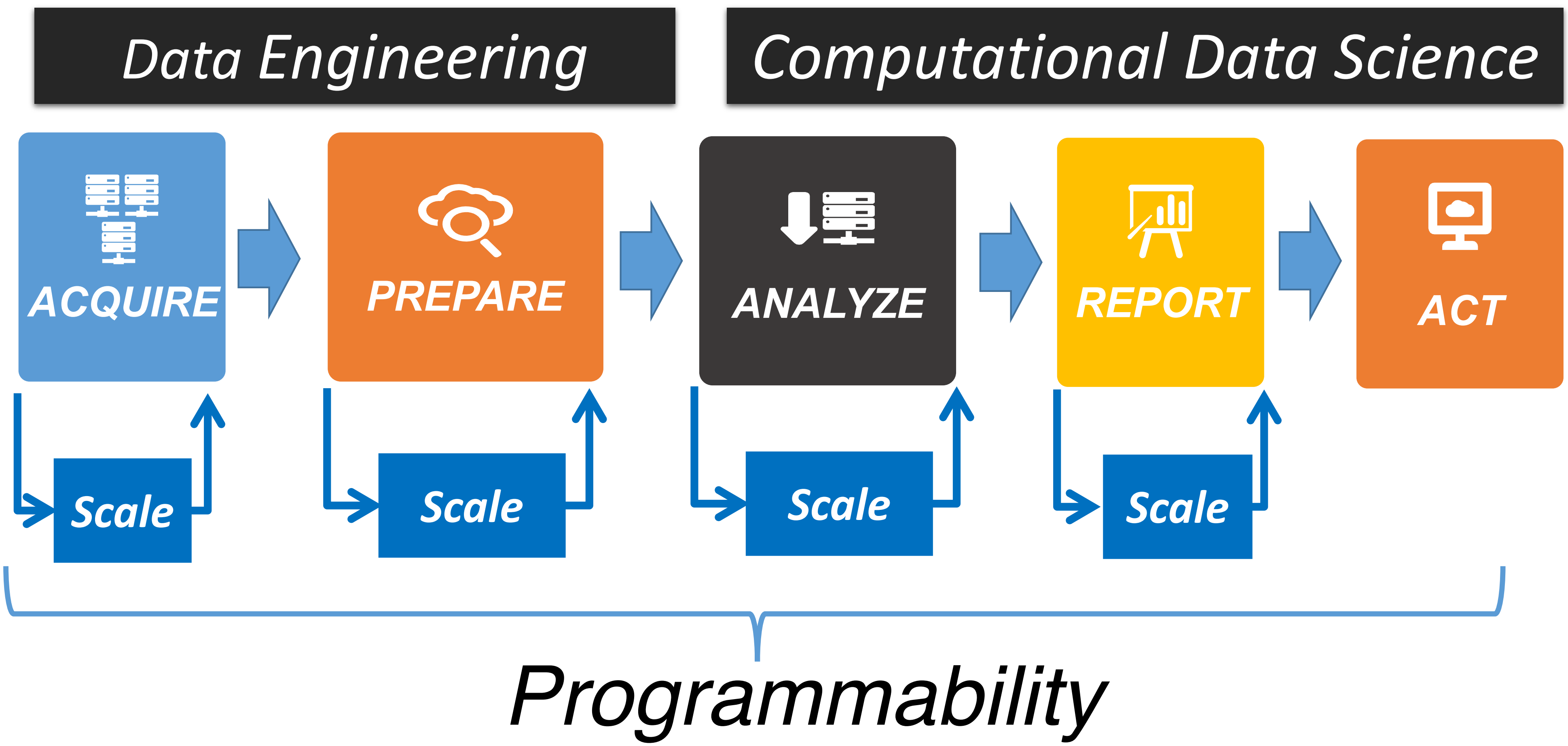


We need to measure metrics for each concern through the process.





Many iterations and rollbacks between steps.


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

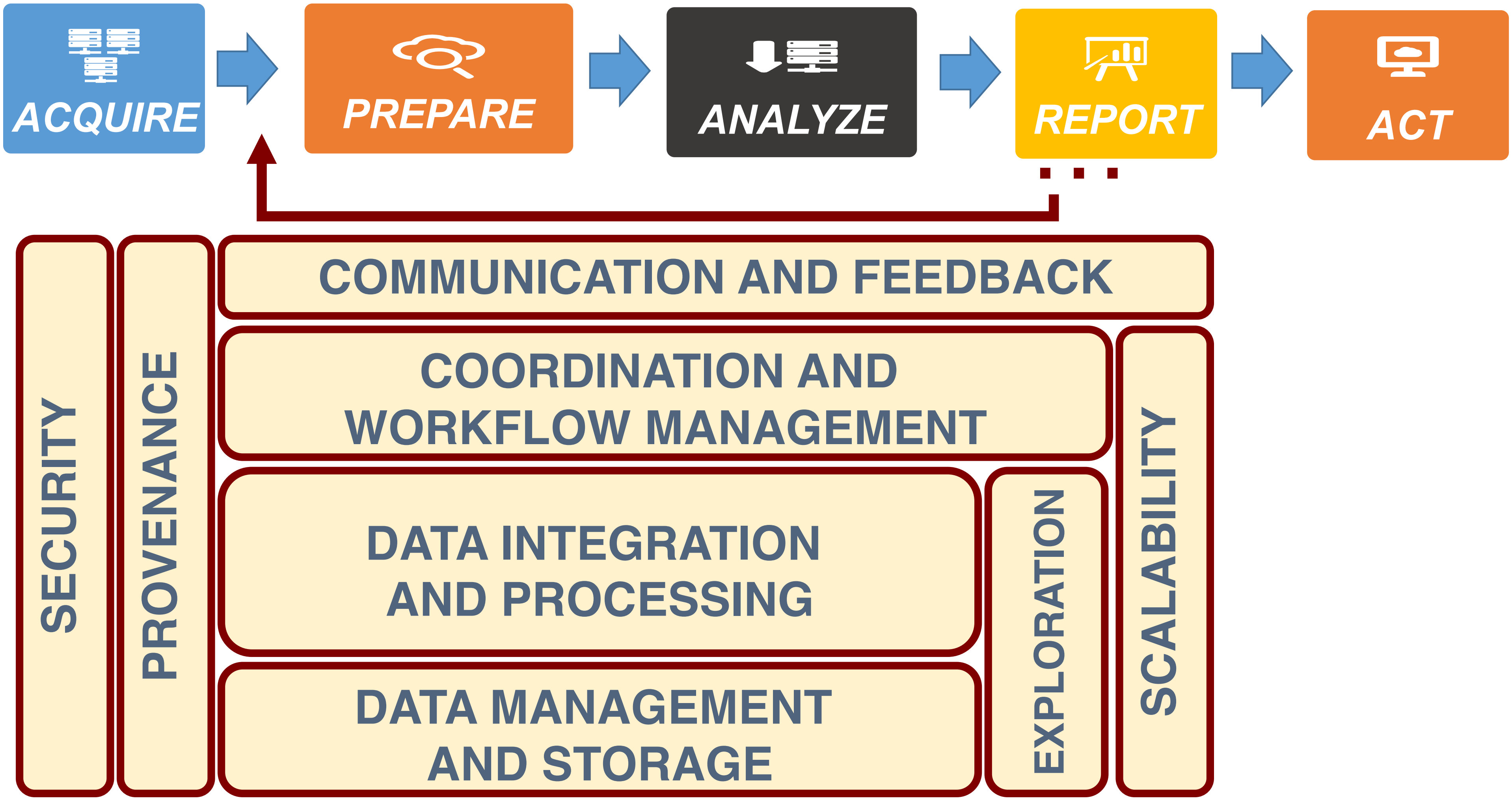
Creating A Solution Architecture

Process-driven Solution Architectures and the Role of Workflows

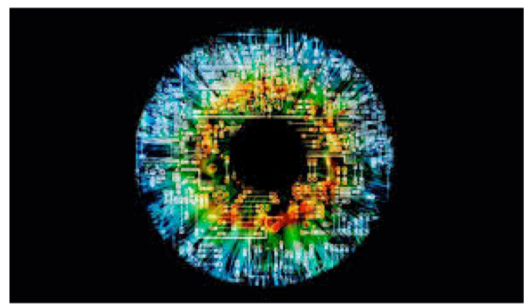
COORDINATION AND WORKFLOW MANAGEMENT

DATA INTEGRATION AND PROCESSING

DATA MANAGEMENT AND STORAGE


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

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



SECURITY

PROVENANCE

WORKFLOW MANAGEMENT

Application Integration, Coordination, Optimization, Communication, Reporting

COMPOSABLE DATA SERVICES

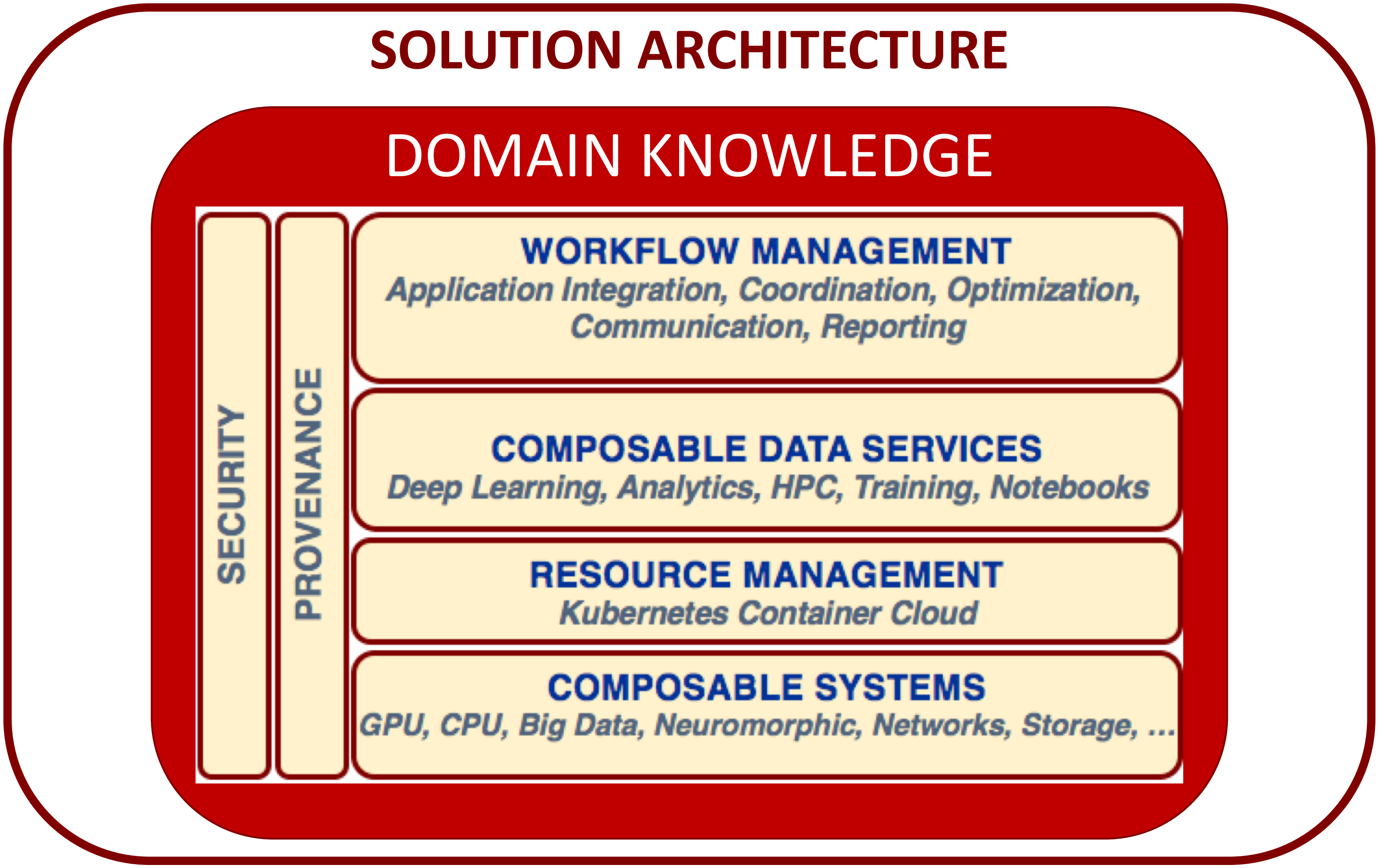
Deep Learning, Analytics, HPC, Training, Notebooks

RESOURCE MANAGEMENT

Kubernetes Container Cloud

COMPOSABLE SYSTEMS

GPU, CPU, Big Data, Neuromorphic, Networks, Storage, …


SOLUTION ARCHITECTURE
DOMAIN KNOWLEDGE
SECURITY
PROVENANCE
WORKFLOW MANAGEMENT
Application Integration, Coordination, Optimization, Communication, Reporting
COMPOSABLE DATA SERVICES
Deep Learning, Analytics, HPC, Training, Notebooks
RESOURCE MANAGEMENT
Kubernetes Container Cloud
COMPOSABLE SYSTEMS
GPU, CPU, Big Data, Neuromorphic, Networks, Storage, …

Questions?

ILkay Altintas, Ph.D.
Email: ialtintas@ucsd.edu