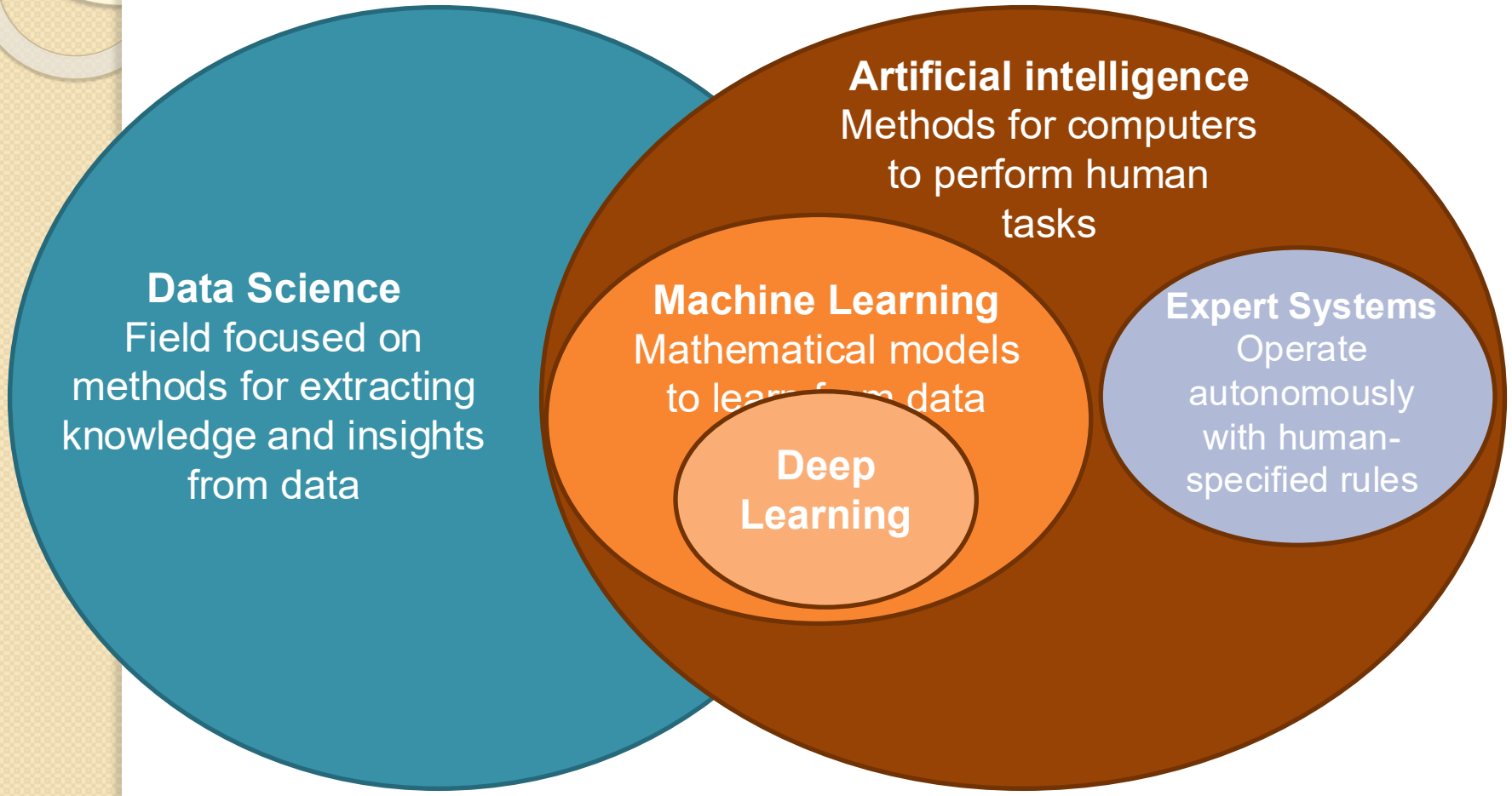
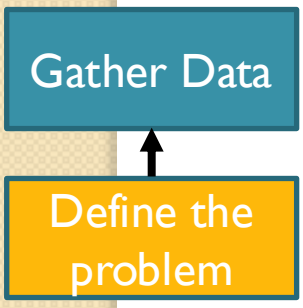
DSCI-560
Lecture 2: Data Science in Practice
Data Science Professional Practicum

Young Cho
Department of Electrical Engineering
University of Southern California

What is Data Science

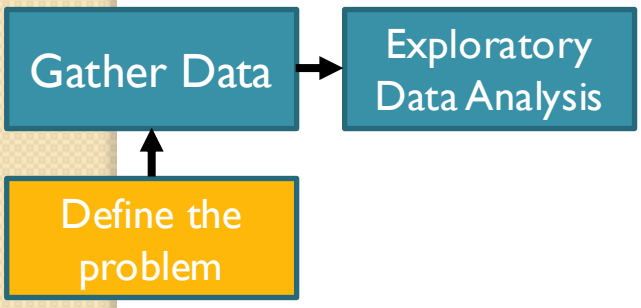


Data Science Pipeline

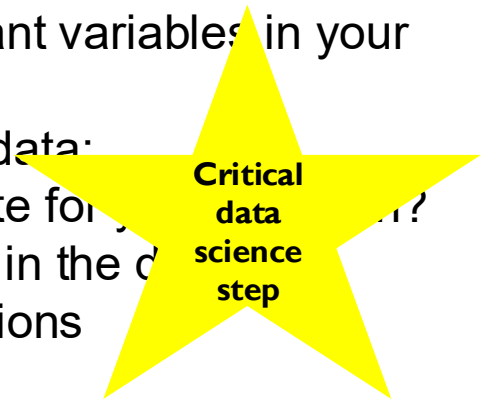


- Collect your own data
  - Surveys
  - Experiment
  - Theory-based model
- Search repositories

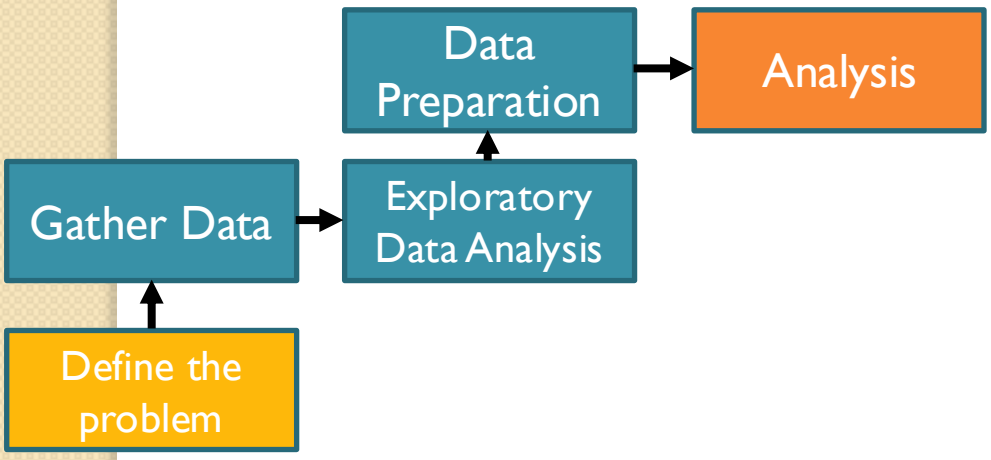
Data Science Pipeline



- Check for missing data and other mistakes
- Mapping and understanding the underlying structure of your data
- Identify the most important variables in your dataset
- Gain insight about your data:
  - Is the data appropriate for
  - Are there any biases in the d
- Often involves visualizations

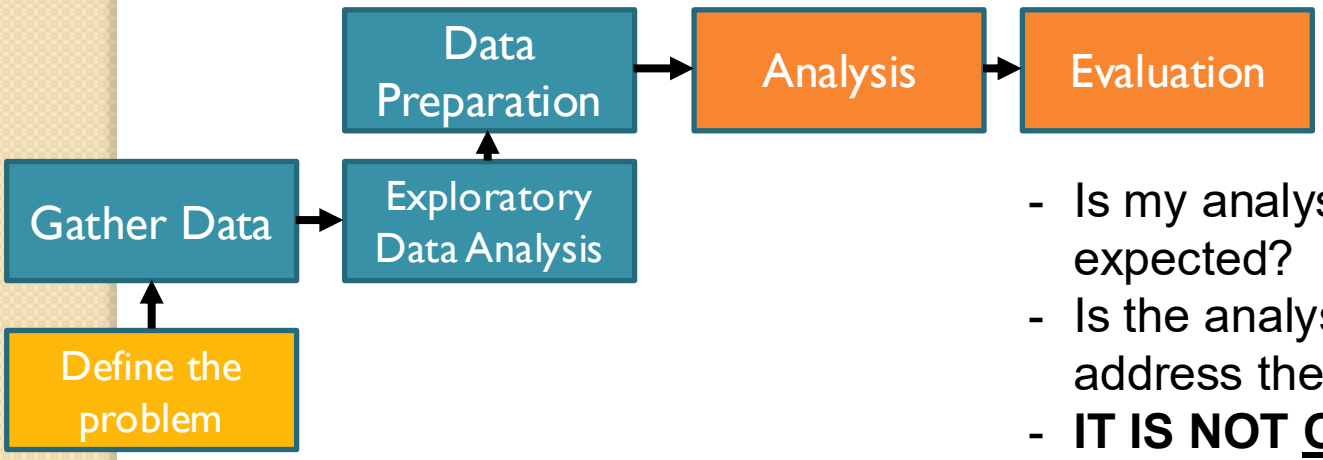


Data Science Pipeline



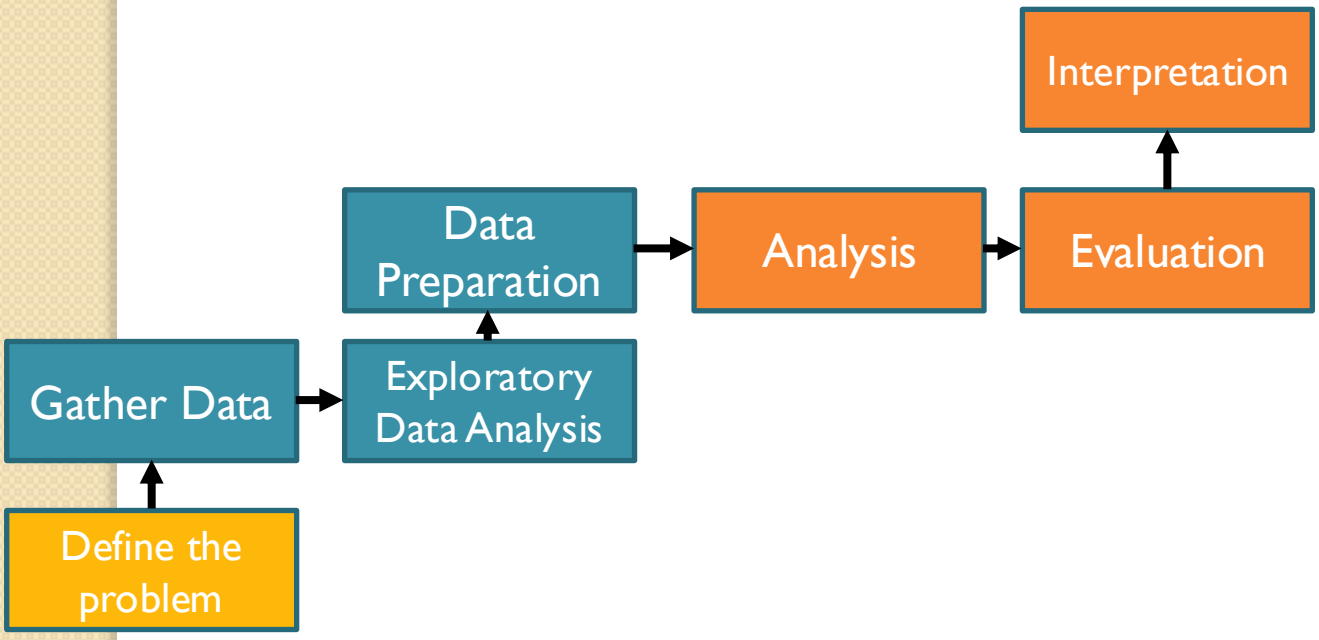
- Choose appropriate analysis for the question
- If using ML and a trained model, is the training data similar to the data to be analyzed?
- What are the pre-processing steps?

Data Science Pipeline

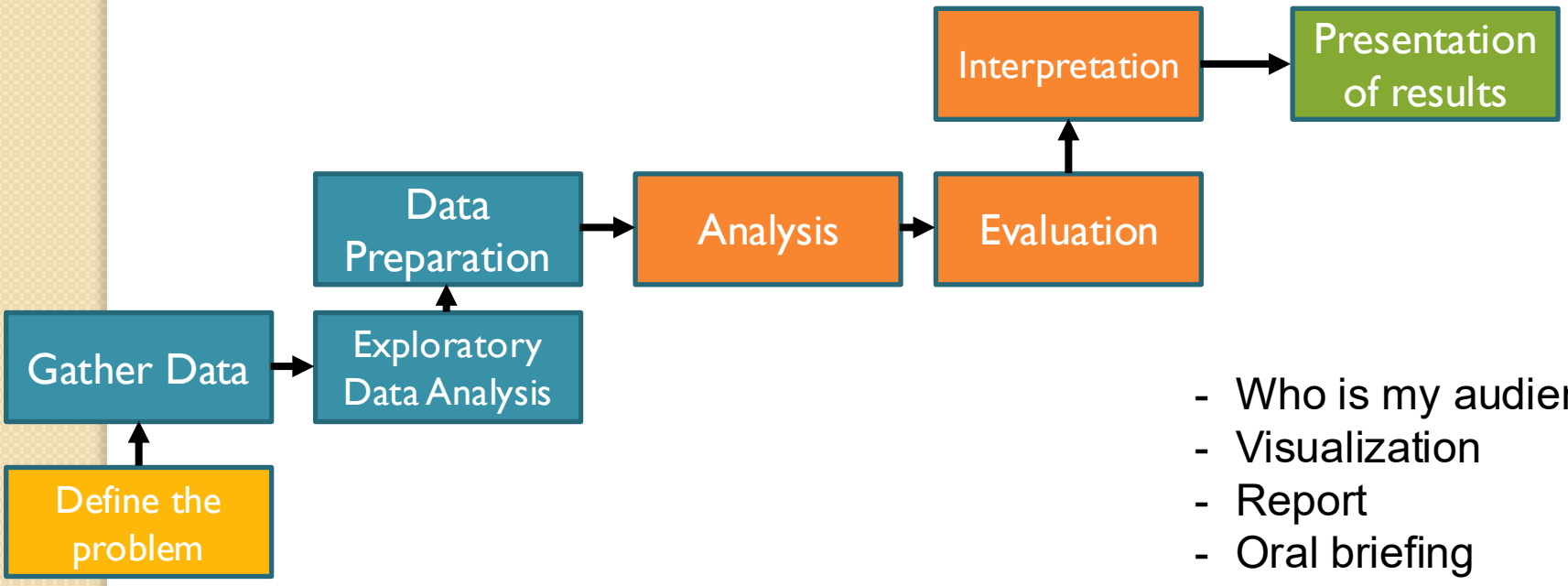


- Is my analysis performing as expected?
- Is the analysis allowing me to address the problem?
- **IT IS NOT <u>ONLY</u> ABOUT THE ACCURACY SCORE**

Data Science Pipeline

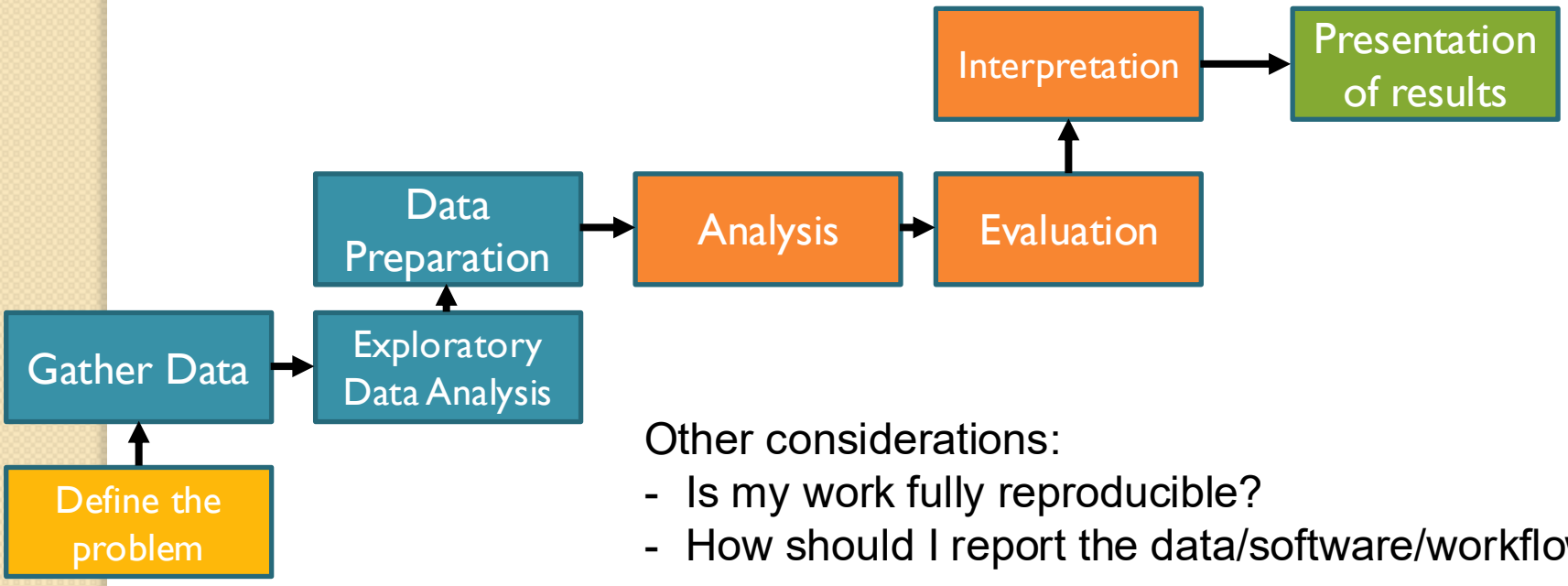


Data Science Pipeline



- Who is my audience?
- Visualization
- Report
- Oral briefing

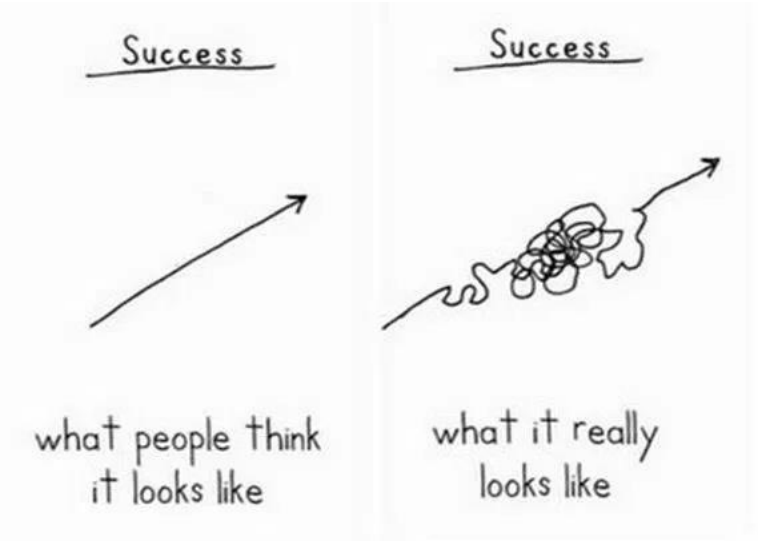
Data Science Pipeline



Other considerations:

- Is my work fully reproducible?
- How should I report the data/software/workflow?

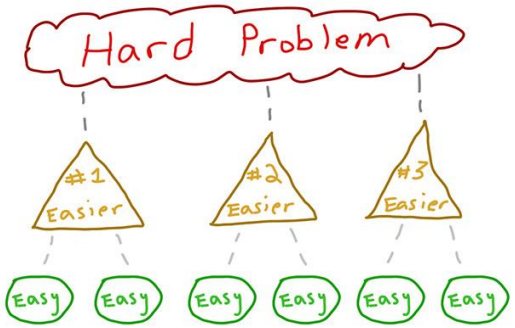
Failure is the Mother of Success



- Progress is rarely linear
- Missing/flawed data
- Problems need to be more refined
- Initial approach didn't work
- Initial hypotheses invalidated
- **Your goal is to learn how to move forward**

Debugging

- List all the components involved. Think about where the weak link might be
- Design a logical and simple troubleshooting process to find the problem
- Ask for help. Ask the internet, ask your peers…
- **<u>Research on your own</u>**



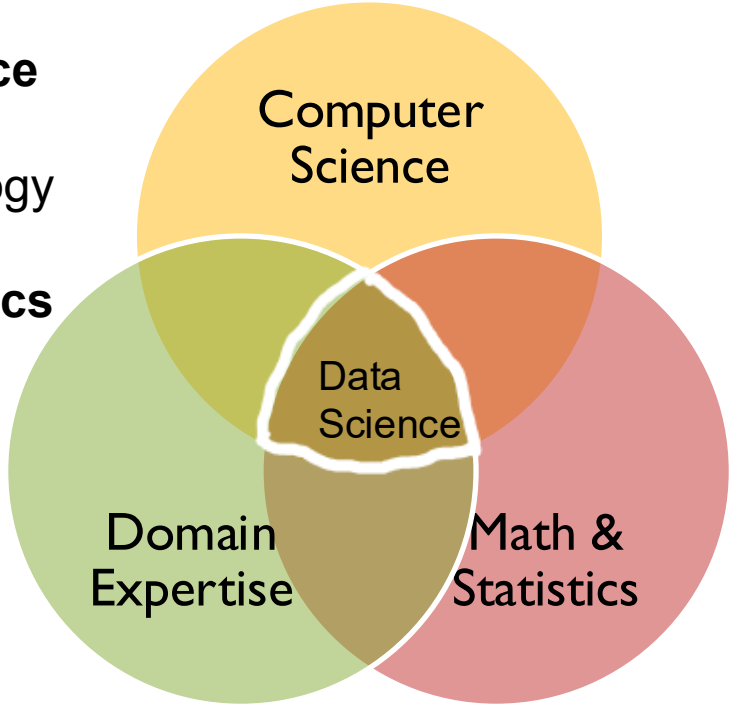
Data May Be the Problem

- Data Should be Representative of the Problem
- Data Should be Qualitatively the Same as the Training data
- Data Should be Without Bias, Sufficiently Large, and Spans All Interested Range
- **All Data Needs Cleaning before being Usable**

Data Science Team

Computer Science
Programming
Big Data Technology

Math and Statistics
Machine Learning
Multivariate
Calculus/Algebra



Domain Expertise
Expert systems
UI/UX
Visualization

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Team requirements

- Skills needed?
- Individuals identified?
- When are they needed?
- Where are they?
- Training needed?
- Interpersonal compatibility?

Project manager: role

- Focuses on a specific project objective
- Controls resources to best meet project objectives
- Manages the constraints (scope, schedule, cost and quality) of individual objective

Large projects may have several managers, each responsible for one part of the project.

Project integration is considered one of the objectives and requires its own manager

Manager may be involved in other parts of the project in a different role

Project manager: functions

- Define scope of the project
- Identify stakeholders and leadership (decision maker: client, organization, public…)
- Evaluate project requirements
- Develop a detailed task list
- Develop initial project management flow chart
- Estimate time requirements
- Identify cost estimation and budgets
- Identify required resources and evaluate risks

Project manager: functions

- Prepare contingency plan
- Identify interdependencies
- Identify and track critical milestones
- Secure needed resources, manpower
- Participate in project phase review
- Manage the change control process
- Report project status

Project manager: characteristics

- Knowledge
  - Must be well-versed with project management
- Performance
  - Application of project management knowledge
- Personal:
  - Effective
  - Attitude
  - Personality characteristics
  - Leadership, guidance to balance project constraints

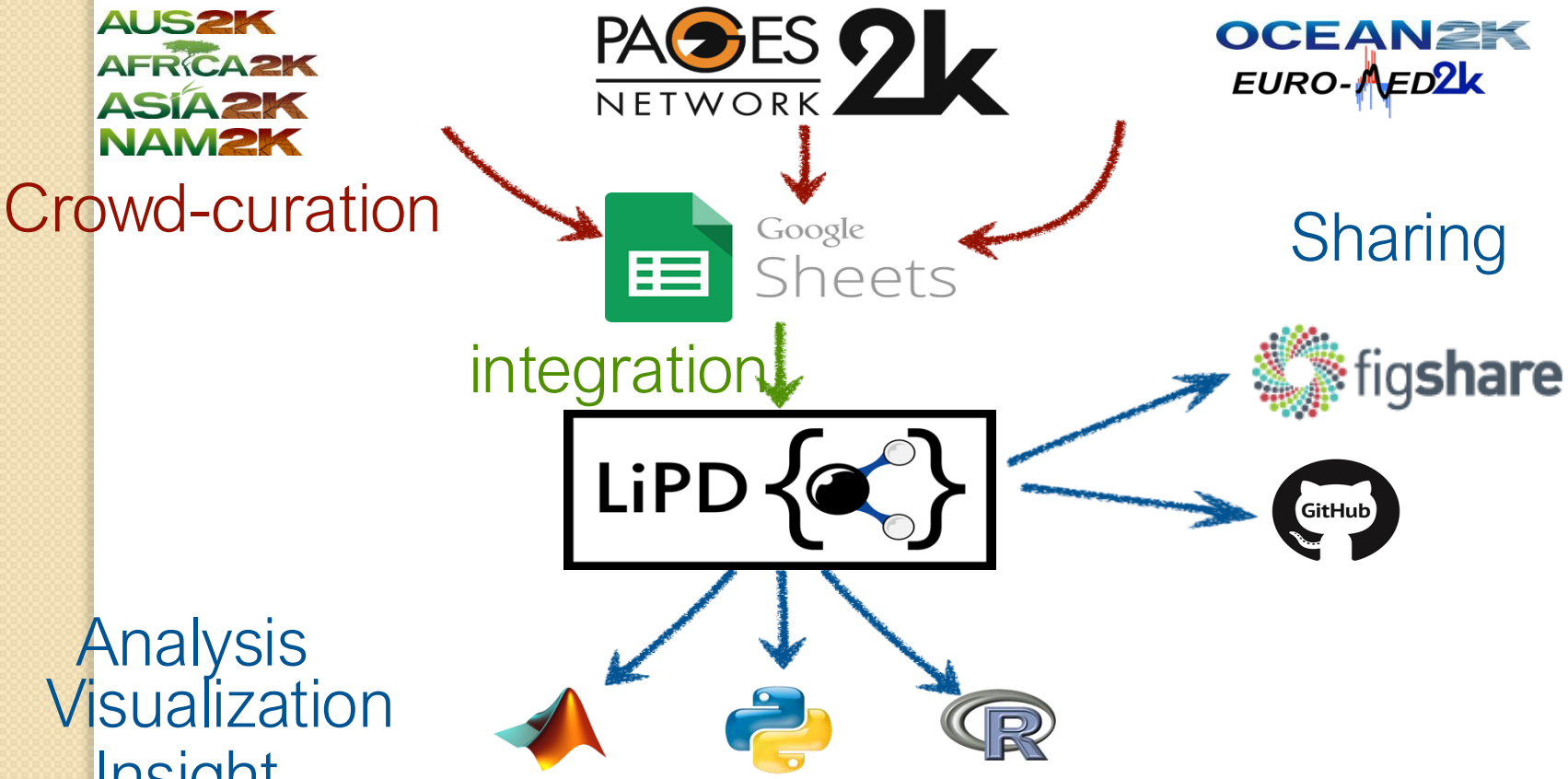
Examples of Systems

Example: Linked Earth

- Paleoclimate observations are crucial to assessing current climate change in the context of past variations. Yet, these observations come in very disparate formats, hindering their re-use and hence lowering their value to science and society.

Project Objectives

- **Data curation**: Build a platform to crowdsource the curation of paleoclimate data
- **Standards development**: Develop standards for how to store and share paleoclimate data
- **Analysis**: Craft tools that use these standards to do better science



McKay & Emile-Geay, Clim Past 2

Goal 1: Data Curation

Create a standard representation for the data	Develop a platform for curation of paleoclimate data
1. **LiPD:** born out of customer need to write a science paper 2. **LinkedEarth Ontology**: formal representation	**User requirements:** • Flexible to accommodate a large variety of data • Multiple users with multiple roles • Embargo on new datasets • Be able to download data in LiPD format • Support complex queries

Goal 2: Standard Development

Create a standard representation for the data	Create a standard vocabulary	Create a standard for reporting
1. **LiPD/LinkedEarth Ontology** 2. **LinkedEarth Ontology**: formal representation – linked.earth/ontology		**User requirements:** • Platform to discuss terms • Platform should allow for voting to reach rapid consensus • Need a mechanism to incorporate new terms in the ontology

Goal 3: Analysis

Craft tools for data analysis

User requirements:

- Use LiPD as input
- Automated data transformation
- Analysis workflows

Pyleoclim

GeoChronR