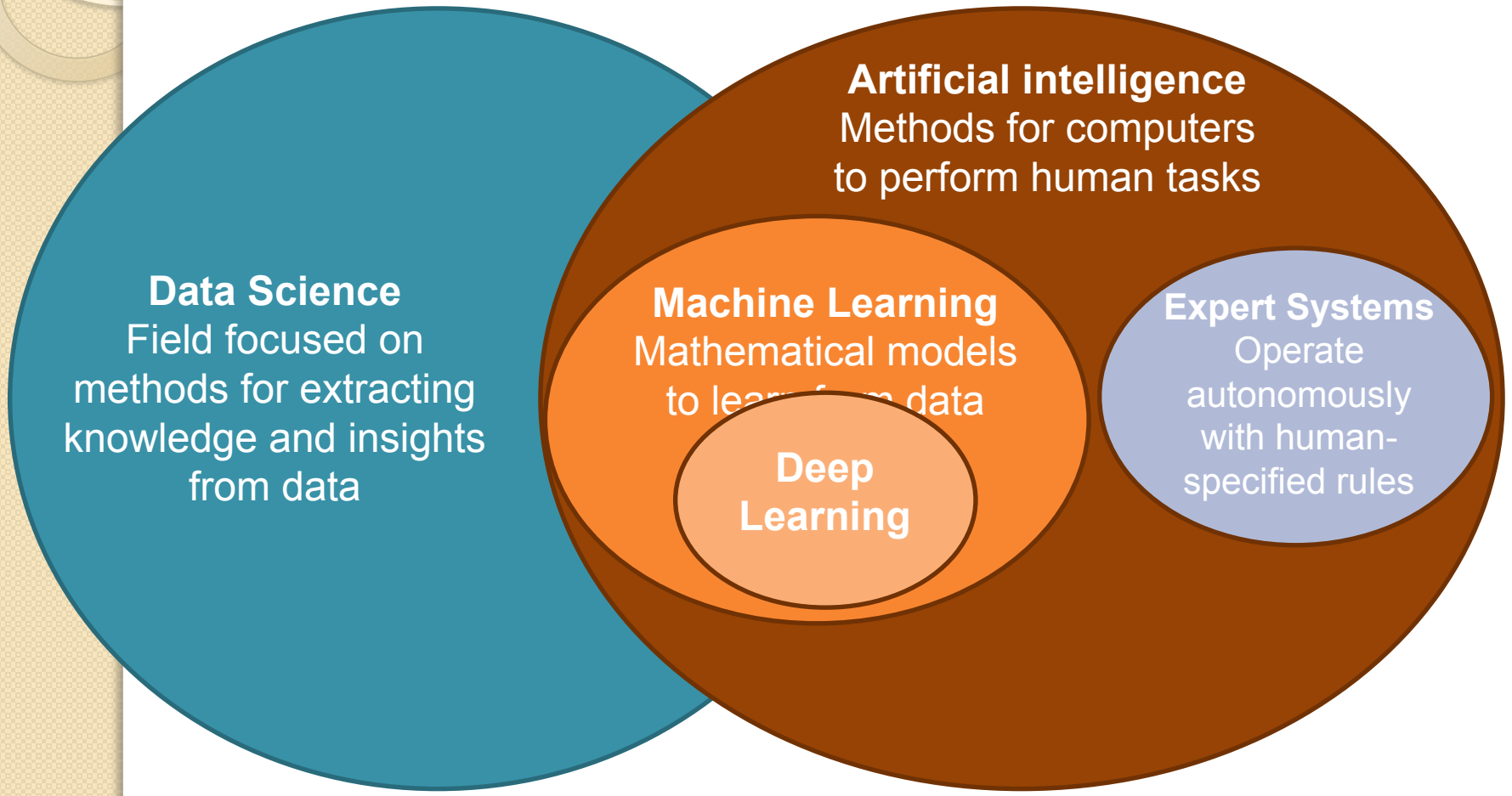
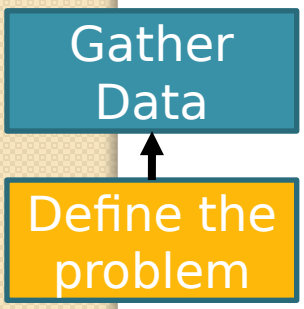
DSCI-560
Lecture 2: Data Science in Practice
Data Science Professional Practicum

Young Cho
Department of Electrical Engineering
University of Southern California

What is Data Science

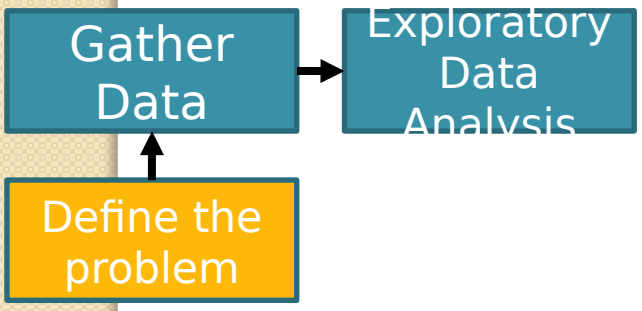


Data Science Pipeline

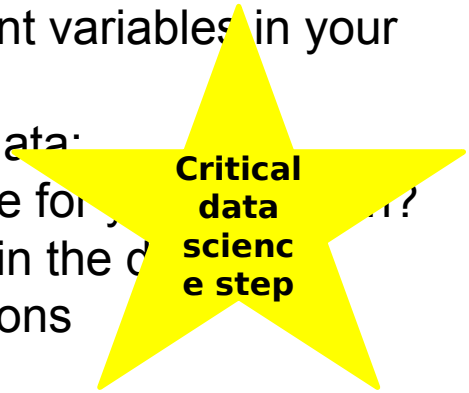


- Collect your own data
  - Surveys
  - Experiment
  - Theory-based model
- Search repositories

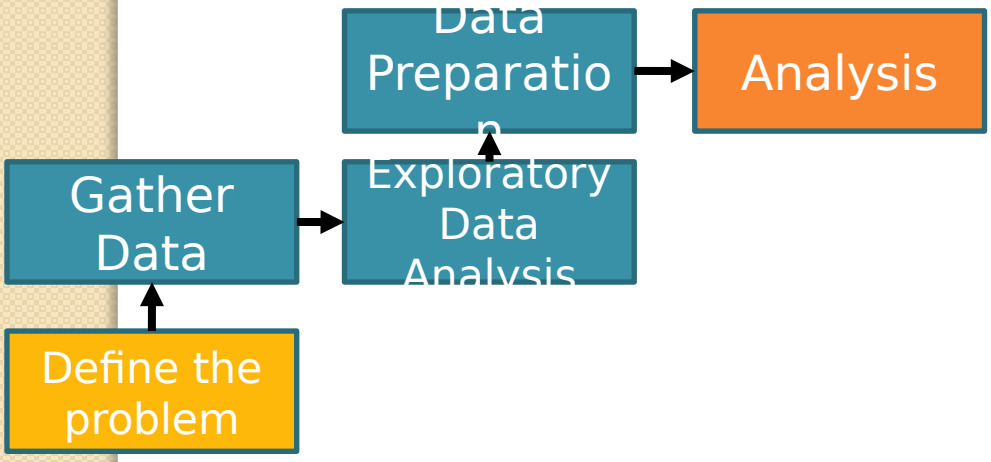
Data Science Pipeline



- 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

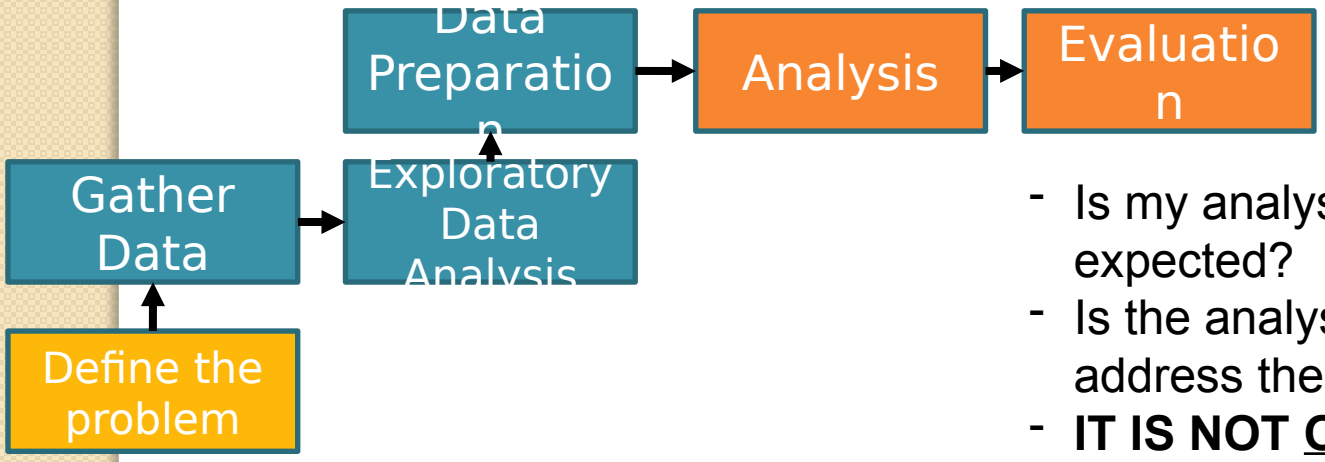


Data Science Pipeline



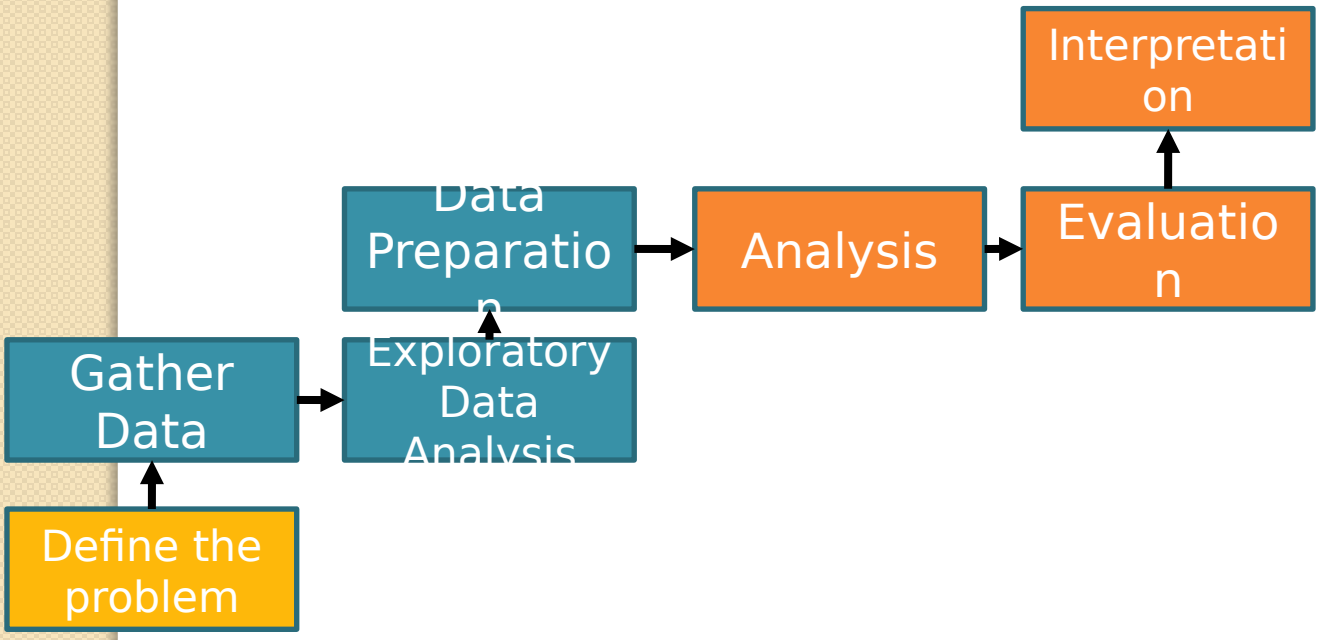
- Choose appropriate analysis for the question
- If using ML and a trained model, is the training data similar to the data to be analyzed?
- What are the pre-processing steps?

Data Science Pipeline

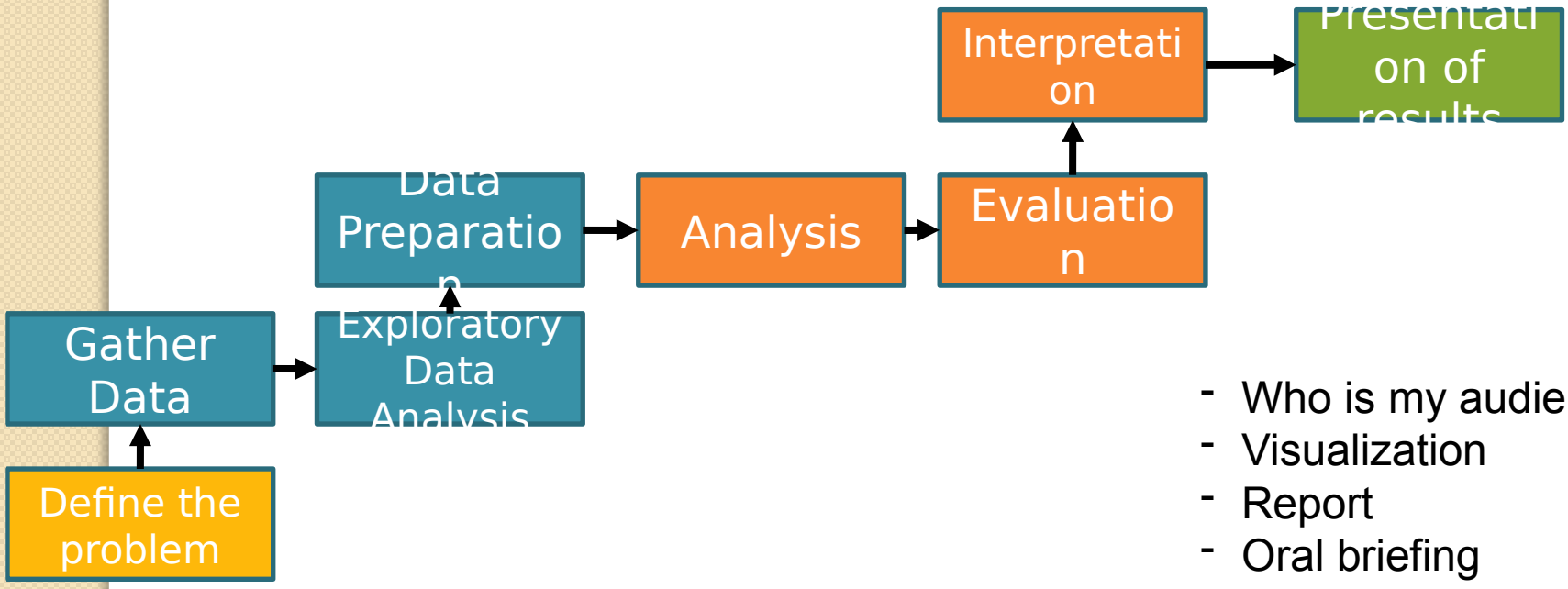


- Is my analysis performing as expected?
- Is the analysis allowing me to address the problem?
- **IT IS NOT <u>ONLY</u> ABOUT THE ACCURACY SCORE**

Data Science Pipeline

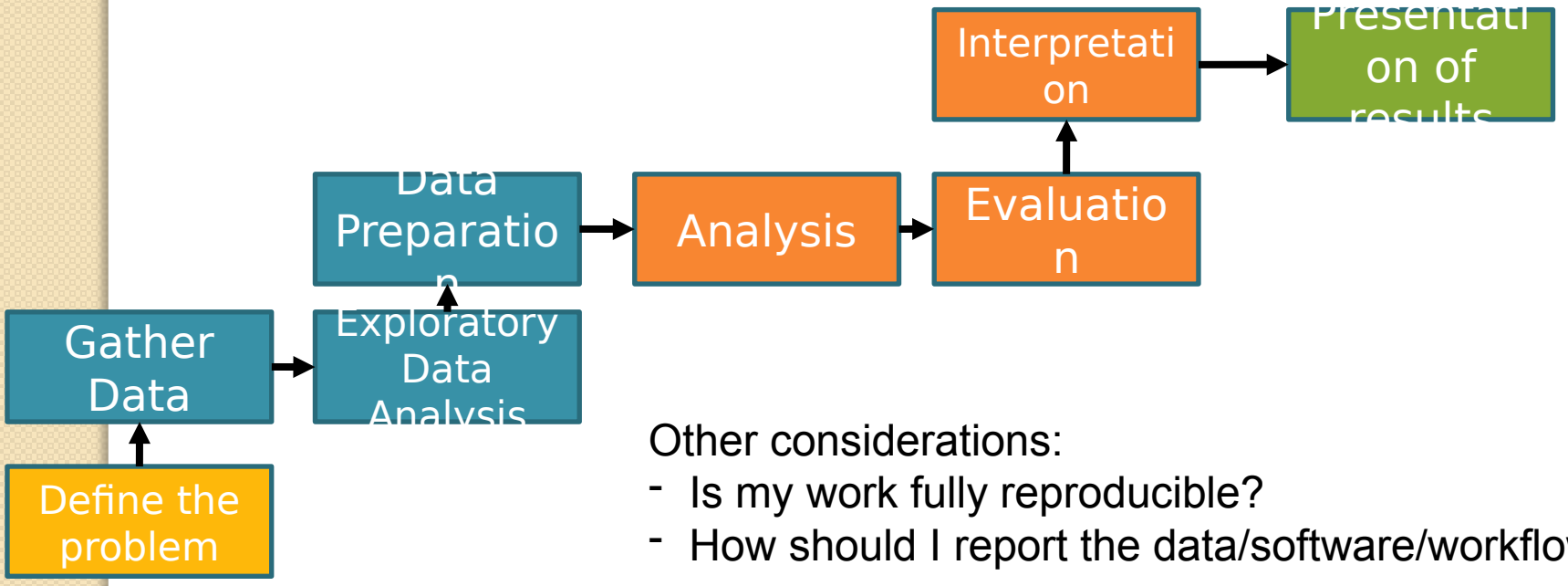


Data Science Pipeline



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

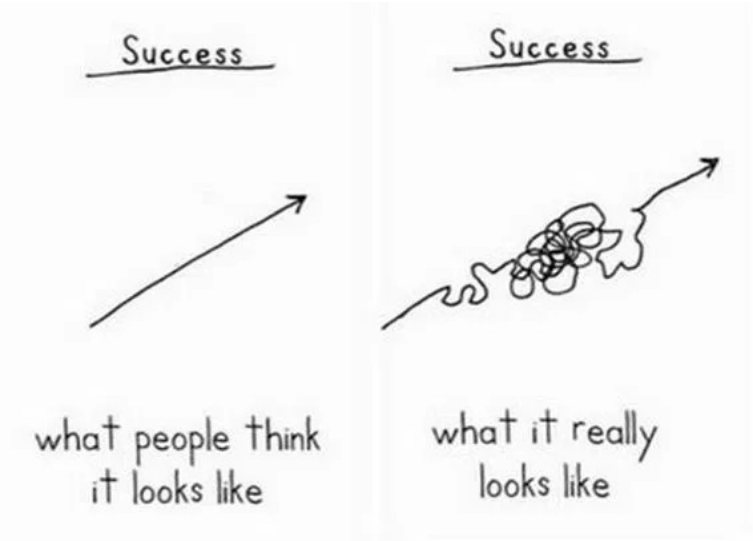
Data Science Pipeline



Other considerations:

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

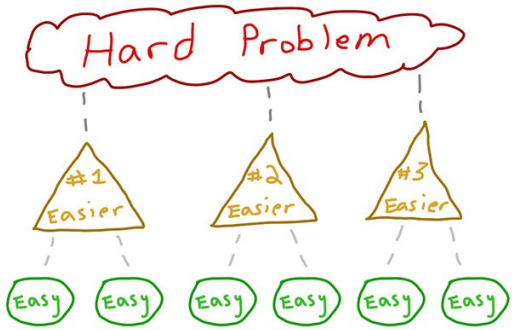
Failure is the Mother of Success

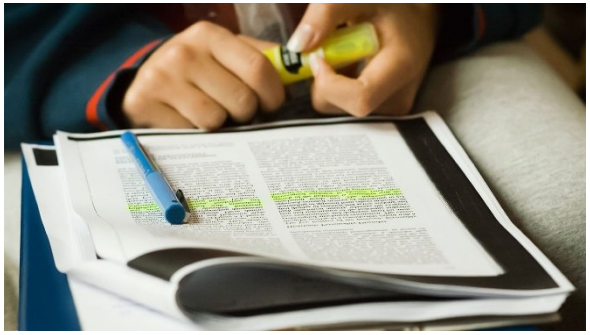


- Progress is rarely linear
- Missing/flawed data
- Problems need to be more refined
- Initial approach didn’t work
- Initial hypotheses invalidated
- **Your goal is to learn how to move forward**

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…
- **Research on your own**





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**

Product Problem

- Context
  - Company that has an on-line shopping site which would like to start to push products to customers as they browse the site
- Data
  - Data about the on-line purchases of customers for the last 5 years
  - Profile data for some customers who are repeat customers: address, credit card, shipping preferences
  - Data about customers who recommended products to their friends in order to get a discount

- Cost
  - When you push random products, 10% of customers do not like what is pushed to them and they leave the site
- Challenges
  - When you have a sale, many repeat customers buy many more items than usual
  - There is no profile data for many customers that pay through a third-party service

Bee Problem

- Context
  - Government of an island who would like to investigate how to reduce the bees so tourism can thrive again
- Data
  - Data available about the weekly water levels of all rivers and ponds for 30 years
  - Medical reports of bee bites and pollen allergies for the last 20 years
  - Climate data and population data, including rainfall and temperatures as well as pollen levels
  - A lot of data about population, pollution, pesticide use, and bird populations (bee predators)

- Challenges
  - Two bee experts in the island, but they do not know anything about data science.
  - What questions would you ask of them to help you figure out how to solve the problem?
  - May release pesticides on crops
    - Pesticides cost $1,000 per square kilometer
    - Pesticides reduce the bee population for 3 months
  - An environmental group that claims that the bee population can be reduced naturally by planting crops that have no flowers (eg corn, wheat, etc)

Fraud Problem

- Context
  - A bank, interested in detecting fraudulent activity in credit card customers
- Data
  - For each customer, there is detailed information from their card application about their address, salary and employment, and demographic
  - For each customer, there is a record of all their transactions (date, charges, and vendor) for the last 4 years
  - For 1% of customers, there is a flag that their credit card was reissued because of fraudulent use of their prior card
  - Additional data available with fee, like census data for any zip code

- Cost
  - When a fraud goes undetected, the average loss to the company is $3K
  - Reissuing a customer card costs $50
  - When card is reissued and there is no fraud, 0.5% of customers cancel their card
- Challenges
  - Volume of the data
    - There are 100M customers, with 30K transactions on average
  - Some credit cards were reissued but no fraud took place once investigated

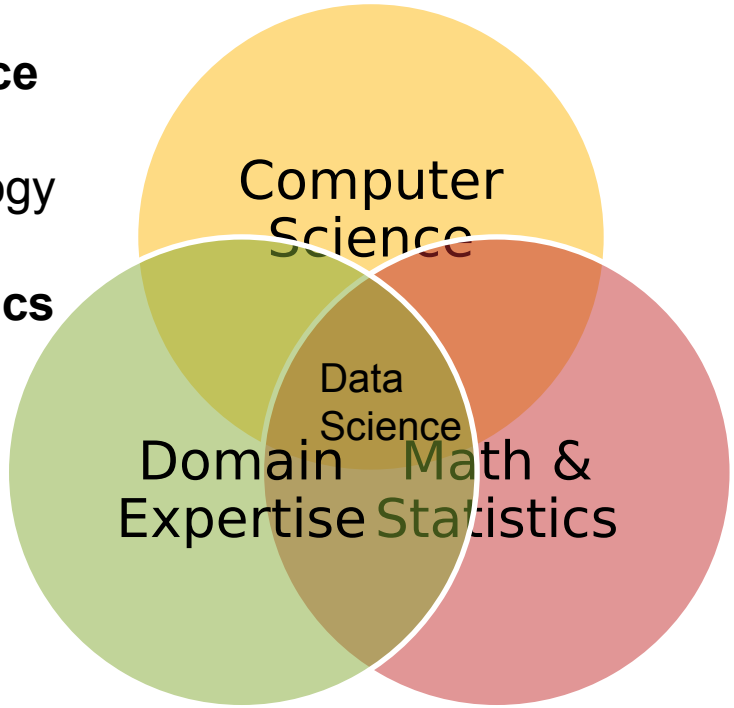
Disaster Relief Proble

- Context
  - A non-profit organization in a remote country who would like to understand where to send relief and in what form
- Data
  - Microblog data (eg twitter), where people are posting issues with bridges, roads, and general access to remote locations
  - Many hospitals are emailing hourly reports, with number of beds occupied and available, medical inventory status, and medical personnel
- Cost
  - A number of coders have contacted your headquarters to volunteer their time to help with data analysis and any data collection needed
- Challenges
  - The remote country's government seems open to take your advice for what roads need repair, what hospitals need more personnel, etc, but will ask for detailed justifications of all your recommendations

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

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

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, ul qu du e

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

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

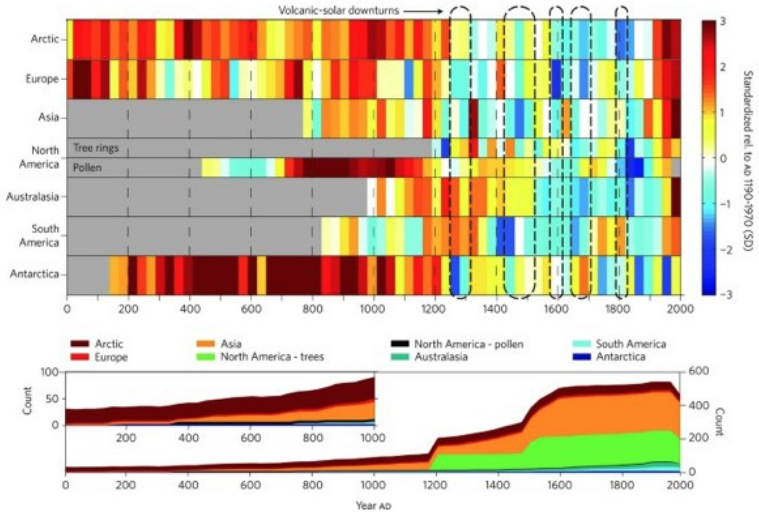
Published: 21 April 2013

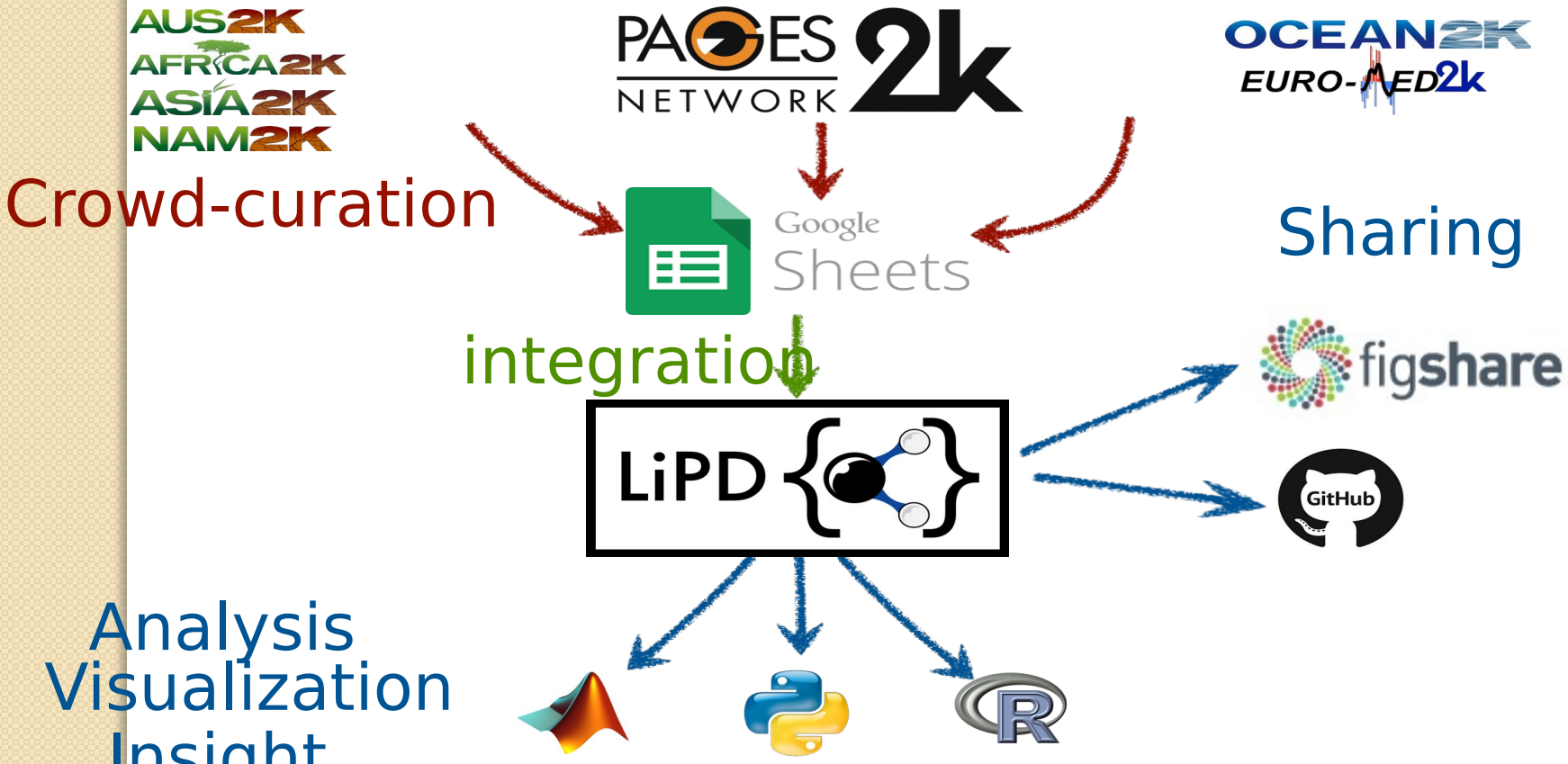
Continental-scale temperature variability during the past two millennia

PAGES 2k Consortium

Nature Geoscience **6**, 339–346(2013) | Cite this article

Figure 2: Continental-scale temperature reconstructions.





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