Lecture 12: Context + Efficiency

Intro to Data Science for Public Policy
Spring 2017

Jeff Chen + Dan Hammer

Roadmap

- Class roadmap
- Data science in context
- Data science: pipelines and architecture

Roadmap

- **Previous class (4/10)**
  - SQL + APIs
- **This class (4/24)**
  - Context, Pipelines + Architecture
  - Homework #5 due
- **Last class (5/1)**
  - A little Python, Clojure, and other technologies you should probably know
- Presentations (5/8)

Roadmap

- Class roadmap
- Data science in context
- Data science: pipelines and architecture

What we've done in this class

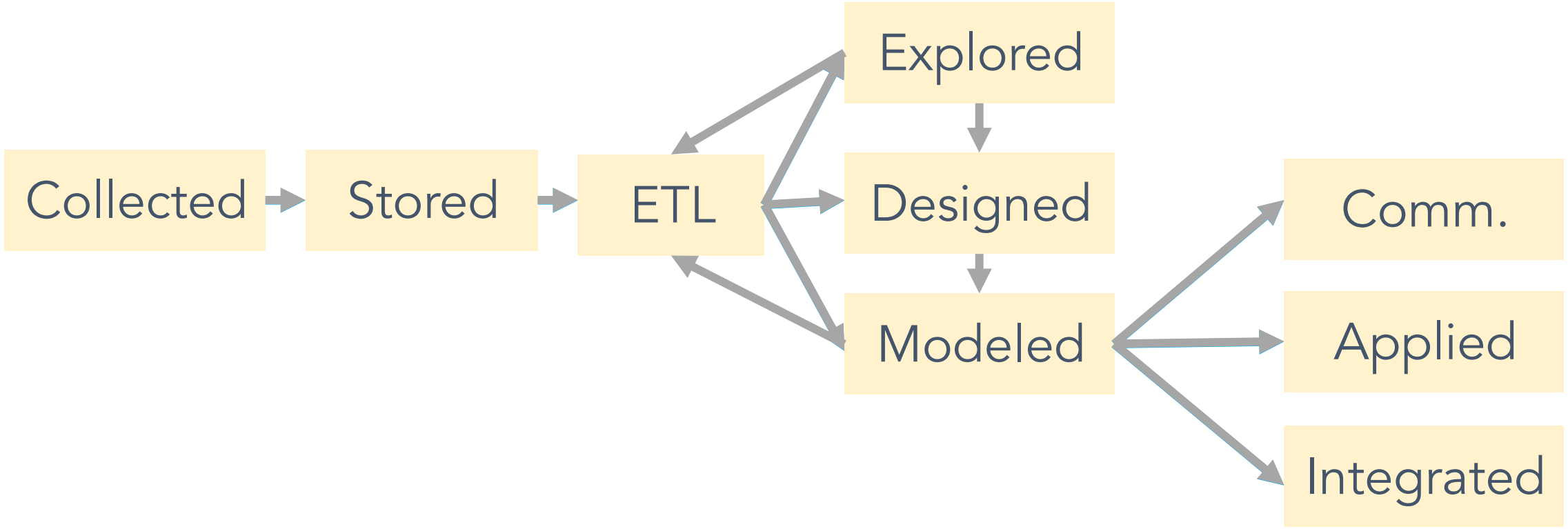
R programming / Loops / Parsing / Manipulation / Exploratory Data Analysis / Basic Visualization / Functions / Control Structures / OLS / Simulation / Supervised Learning / KNN / Cross Validation / Classifiers / TPR / Confusion Matrices / TNR / AUC / F1 / Logistic Regression / Random Forests / Scoring / Decision Trees / Support Vector Machines / Unsupervised Learning / K-Means / Hierarchical Clustering / SQL / Querying / Web Services / APIs / Spatial data

Context

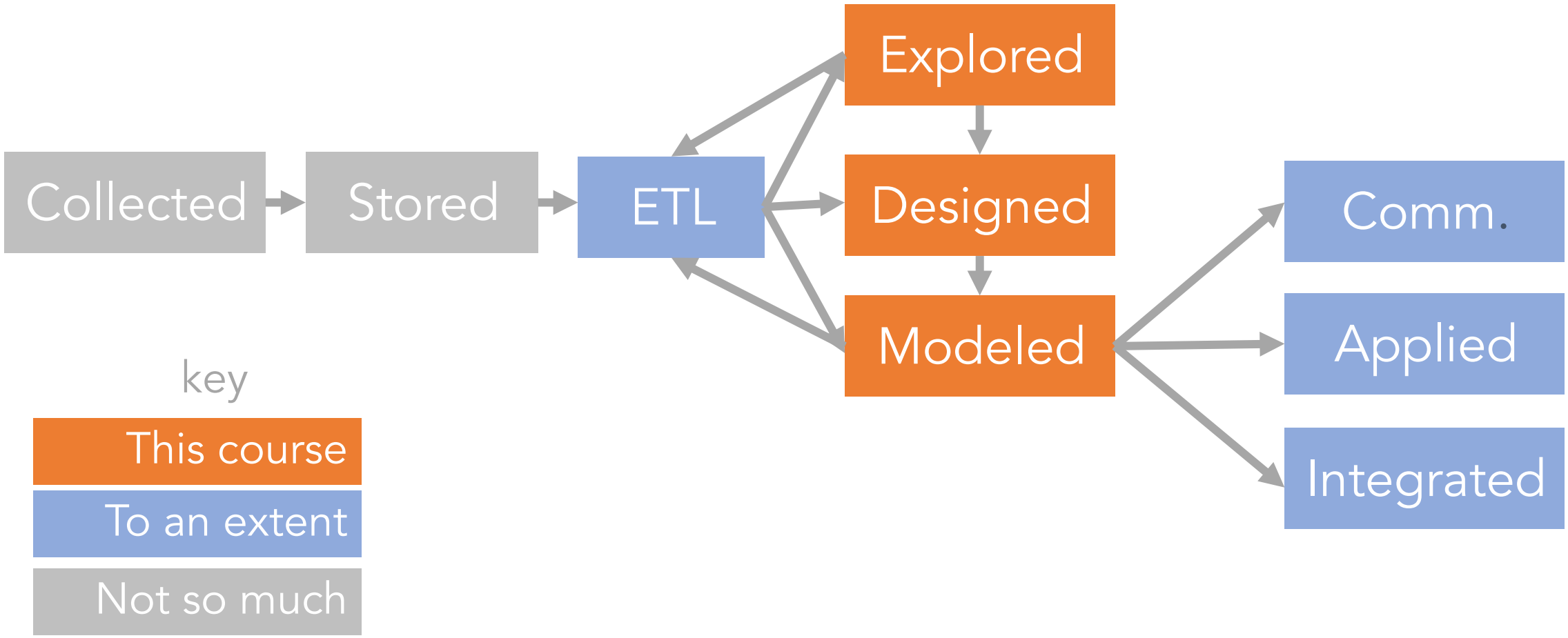
Survey of what goes on inside the data science process

R programming / Loops / Parsing / Manipulation / Exploratory Data Analysis / Basic Visualization / Functions / Control Structures / OLS / Simulation / Supervised Learning / KNN / Cross Validation / Classifiers / TPR / Confusion Matrices / TNR / AUC / F1 / Logistic Regression / Random Forests / Scoring / Decision Trees / Support Vector Machines / Unsupervised Learning / K-Means / Hierarchical Clustering / SQL / Querying / Web Services / APIs / Spatial data

Data is…

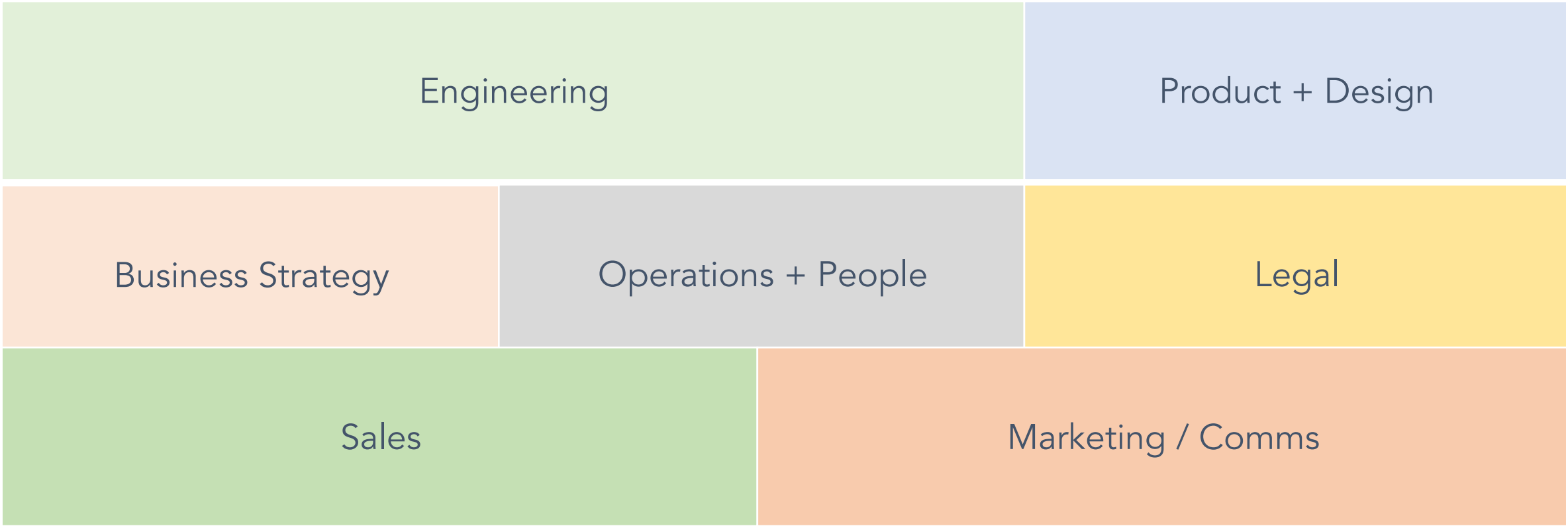

Collected
Stored
ETL
Explored
Designed
Modeled
Comm.
Applied
Integrated

What we covered



In practice, where does it get applied?

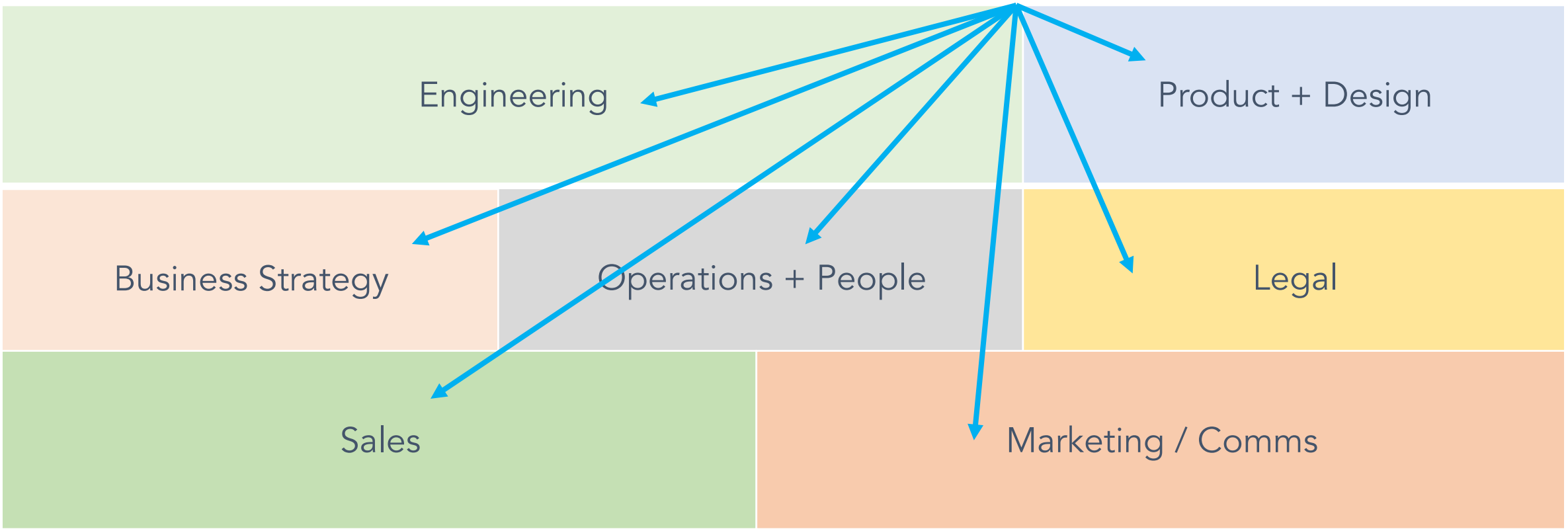
Team Layout
Product-Centric Organizations



Example 1

Where data science lives
Product-Centric Organizations

Data science skills may be applied in every team



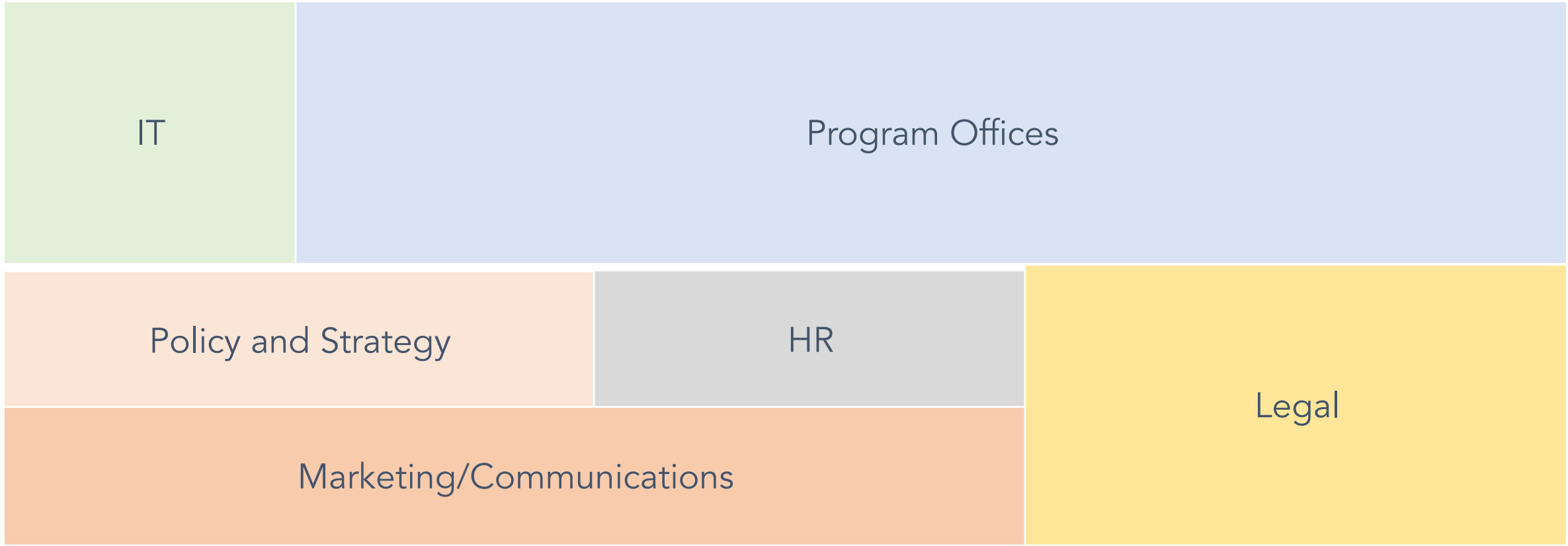
Example 1

What data scientists do
Product-Centric Organizations

Search engines, new features, Recommendation engines, cybersecurity		Product analytics, A/B testing, experiments
Pricing, competitive research, economics	Recruitment modeling, Fraud detection, payments	Risk management, privacy
Micro-targeting campaigns, lead generation	Micro-targeting campaigns, A/B, conversion experiments, social network analysis, user retention,	

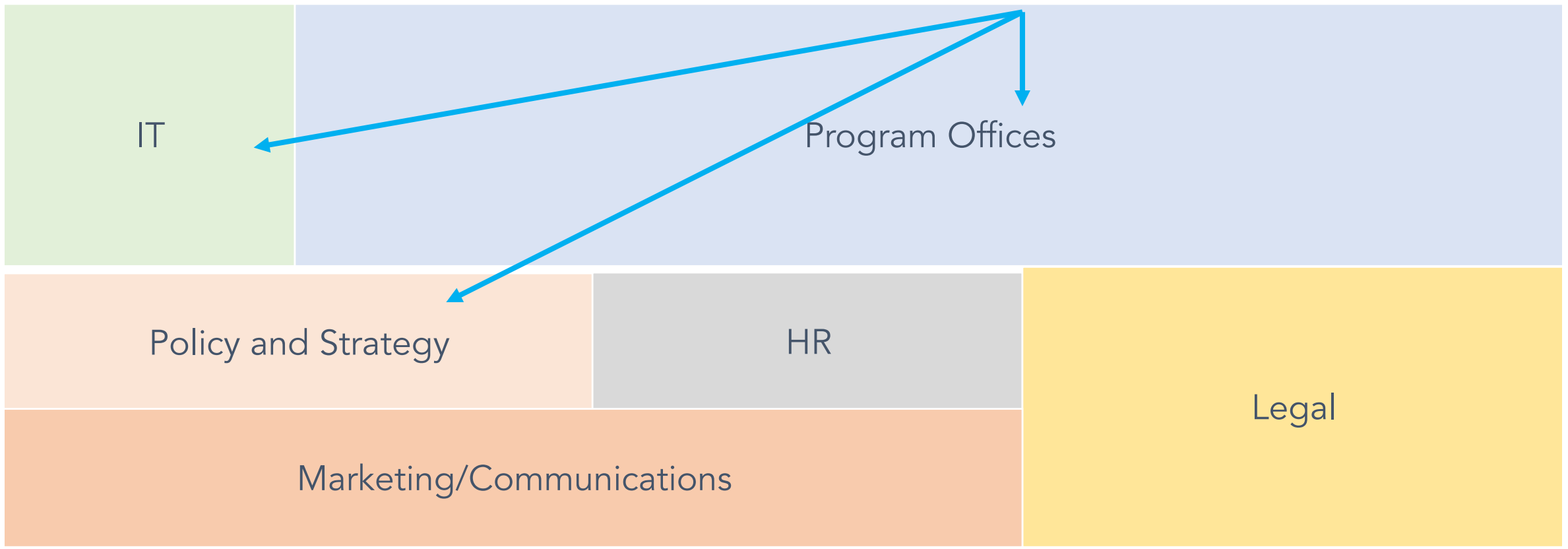
Example 1

Team Layout
Program-Centric Organizations



Example 2

Where data scientists live
Program-Centric Organizations



Example 2

What data scientists do
Program-Centric Organizations

Cybersecurity, etc

(Anything really..)
Competitive research, Micro-targeting campaigns, lead generation, computer vision, recommendation engine

Example 2

Types of data job titles

Data science (General)

- Data Scientist
- Data Analyst
- Statisticians
- Quantitative Analyst
- Research scientist

Strategy/Business

- Insights analyst
- Competitive insights analyst
- Pricing analyst

Engineering

- Search engineer
- Software engineer (machine learning)
- Business intelligence (BI) engineer)
- Big data engineer

Operations

- People analytics
- Recruitment insights analyst

Product + Design

- Product Analyst
- Data scientist, product

Marketing

- Marketing data analyst
- Ads Optimization Analyst

How do data scientists work?
(In Government)

Manage + Process data, engineering

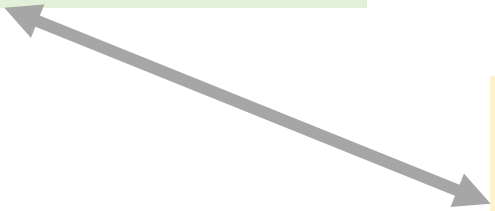
- Data Engineer
- Database Administrator (DBAs)
- Architects

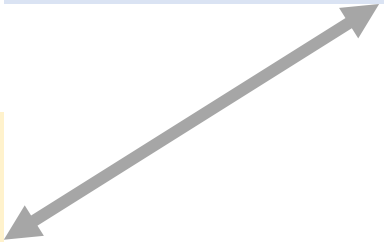
Conduct data science tasks

- Data Scientist
- Statisticians
- Quantitative Analyst

Identify new products + features, or works with business side

- Product Analyst
- Product Manager and Designers
- Business Analyst
- Policy Advisor (gov)





What do designers + product managers do?

(In Tech Sector)

Feature development

User interface design

Experiments to see if different UI designs yield different results

User experience (usability)

What do engineers do?

(In Tech Sector)

Search

Underlying IT infrastructure (e.g. servers, databases) and functionality of web page

What do data scientists do?

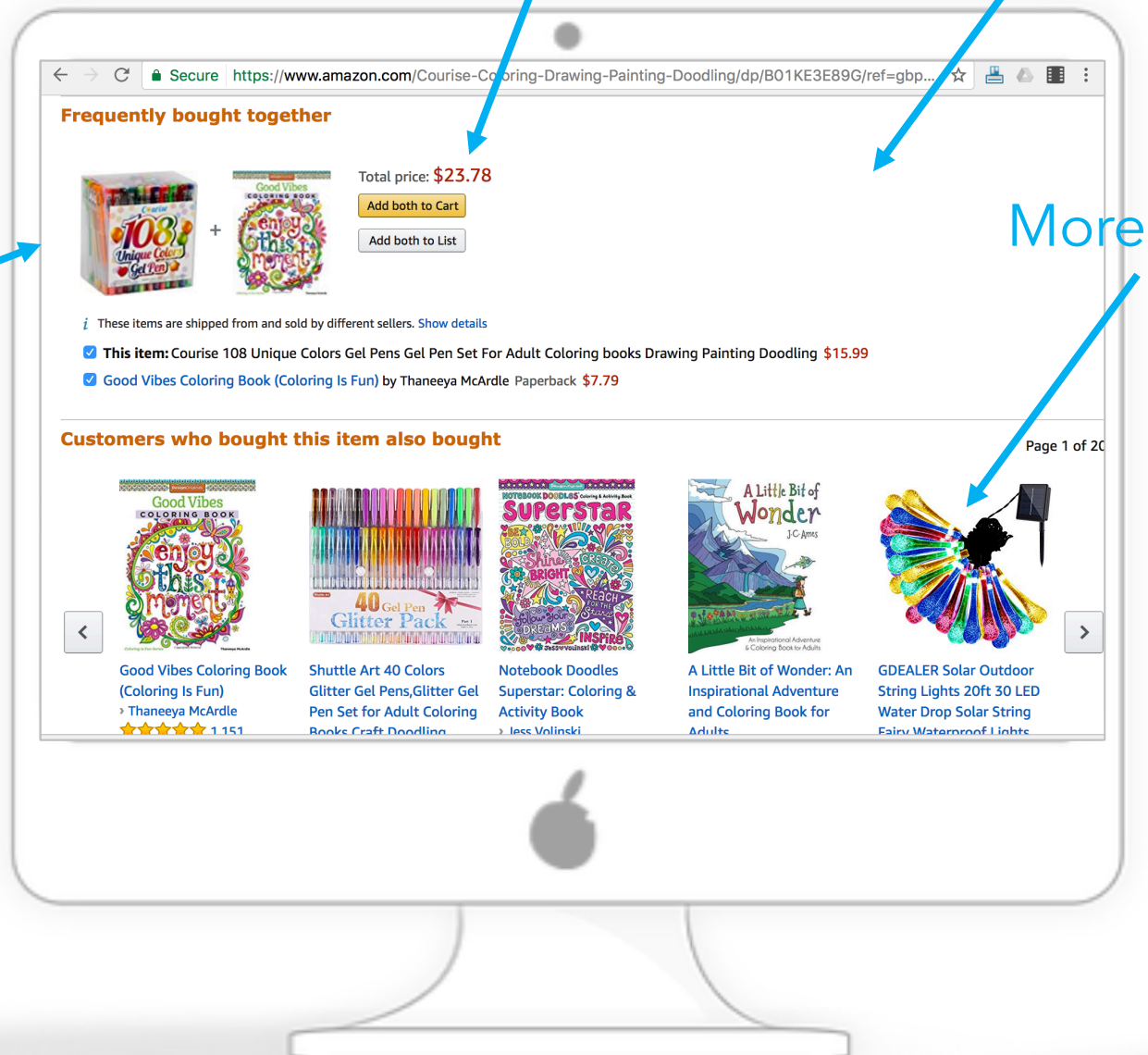
(In Tech Sector)

Automated Pricing

Sometimes search

Recommendations to the user

More recs



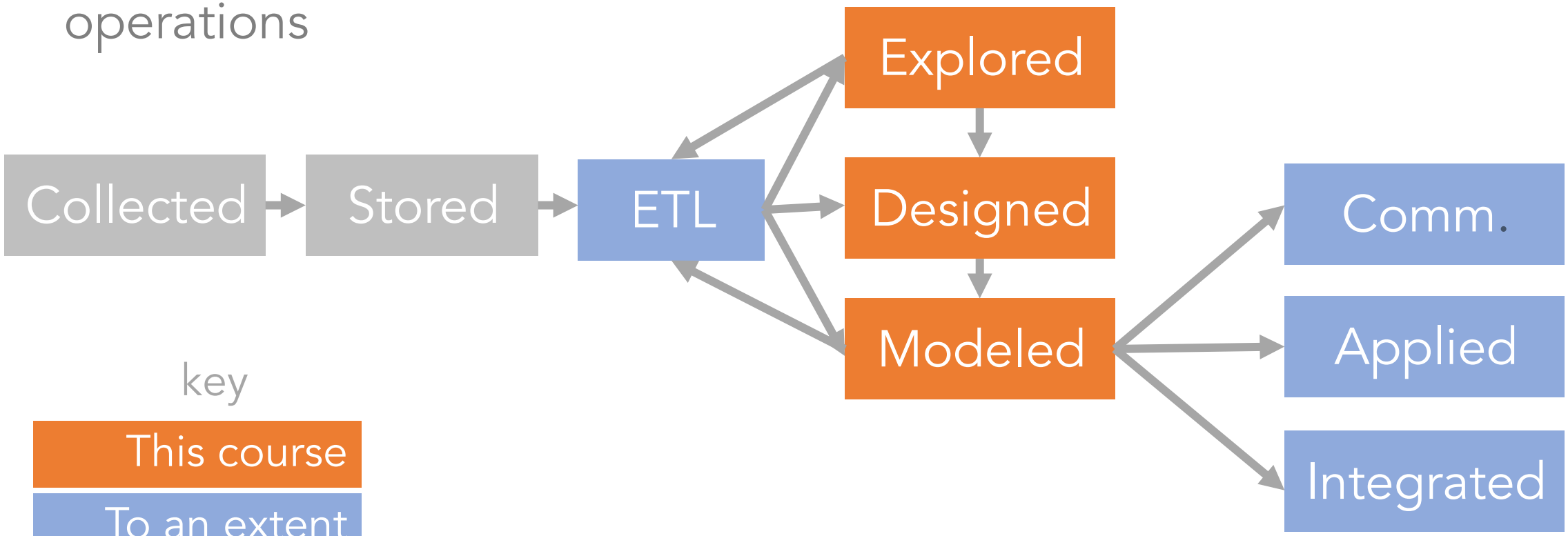
How do data scientists work?
(In Government)

- Data scientists in government tend to be generalists as product teams are uncommon
- Often times serve as consulting teams working across organizations
- More self-reliant – less access to engineers and designers to round out capabilities

Roadmap

- Class roadmap
- Data science in context
- Data science: pipelines and architecture

Goal of course: deriving usable data insights and understanding fundamental operations



key

- This course
- To an extent
- Not so much

Volume
Velocity
Variety

Tech companies typically focus on the following V's (also known as the 3 V's of big data). Much of the rhetoric focuses on the value proposition of these data considerations.

Volume
Velocity
Variety

Notice that value or insight is not in the vector of concepts.

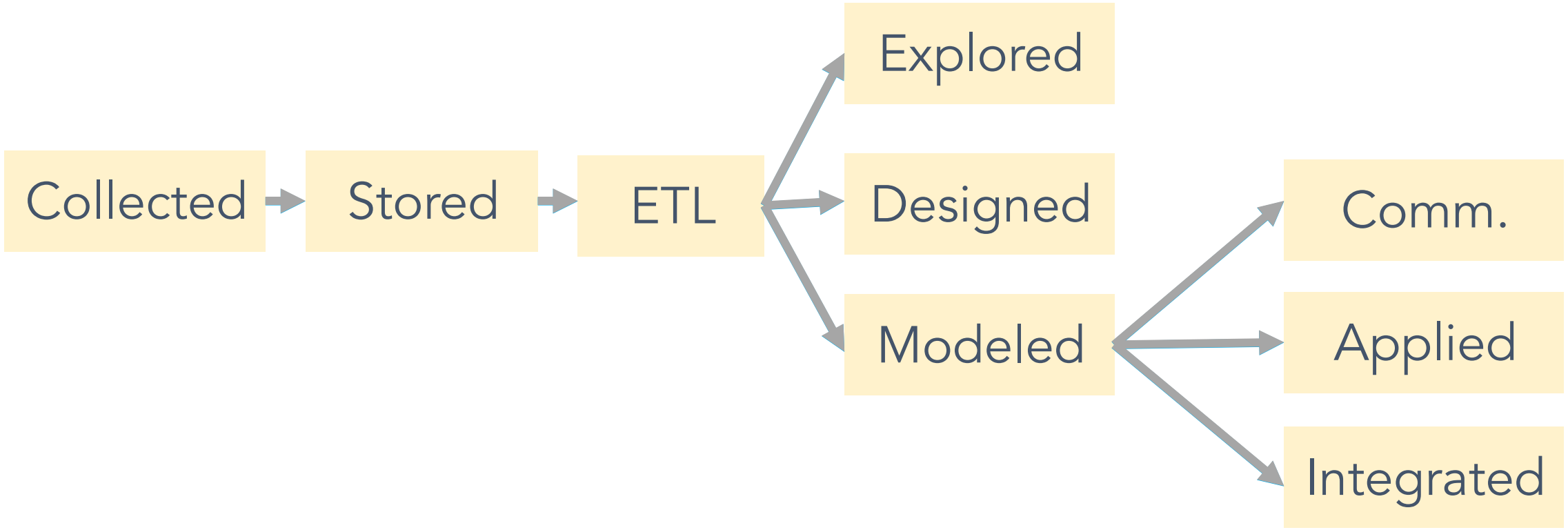
This is because value is dependent on data quality, research design and constituent needs – it's not the technology and it's more about the data itself.

Volume, velocity, variety without insight is noise.
Insight that is not timely or robust is old news.

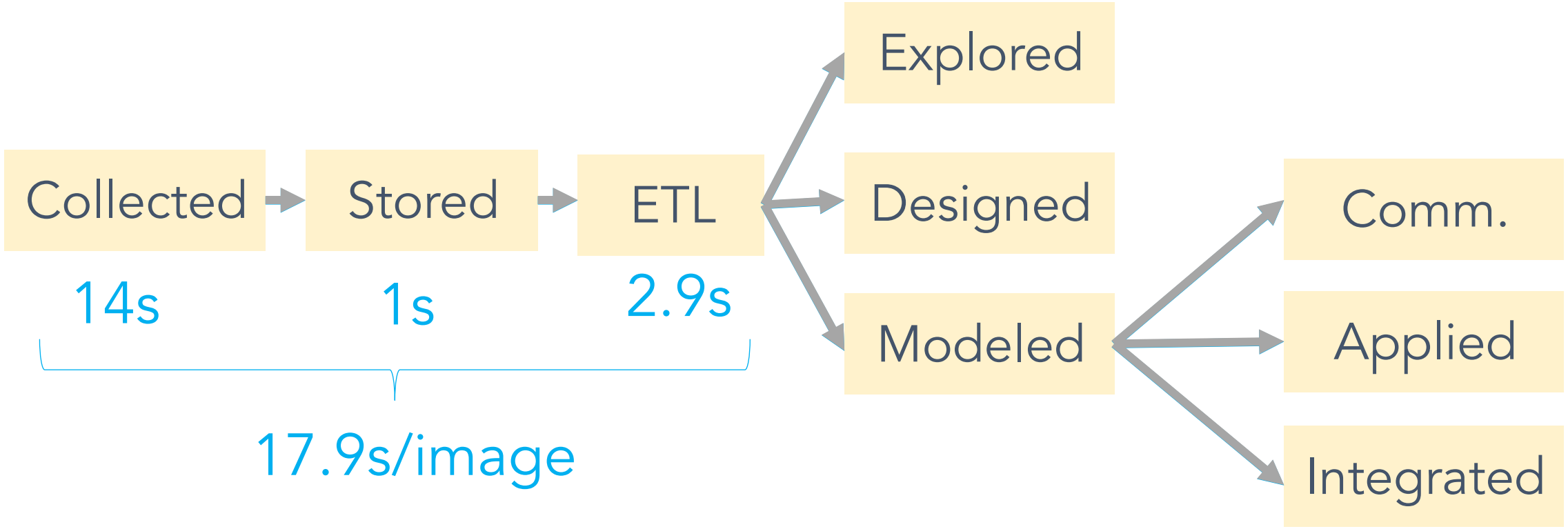
In Focus

- Benchmarking
- Efficient workflows
- Understanding architecture

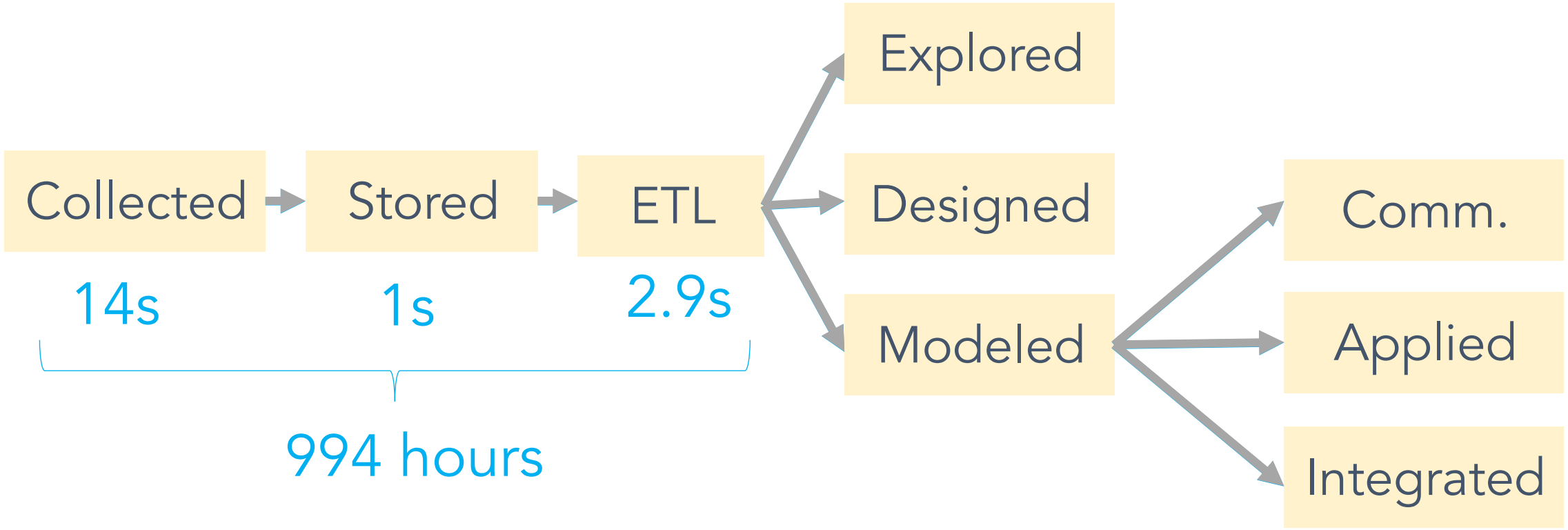
Data requires time



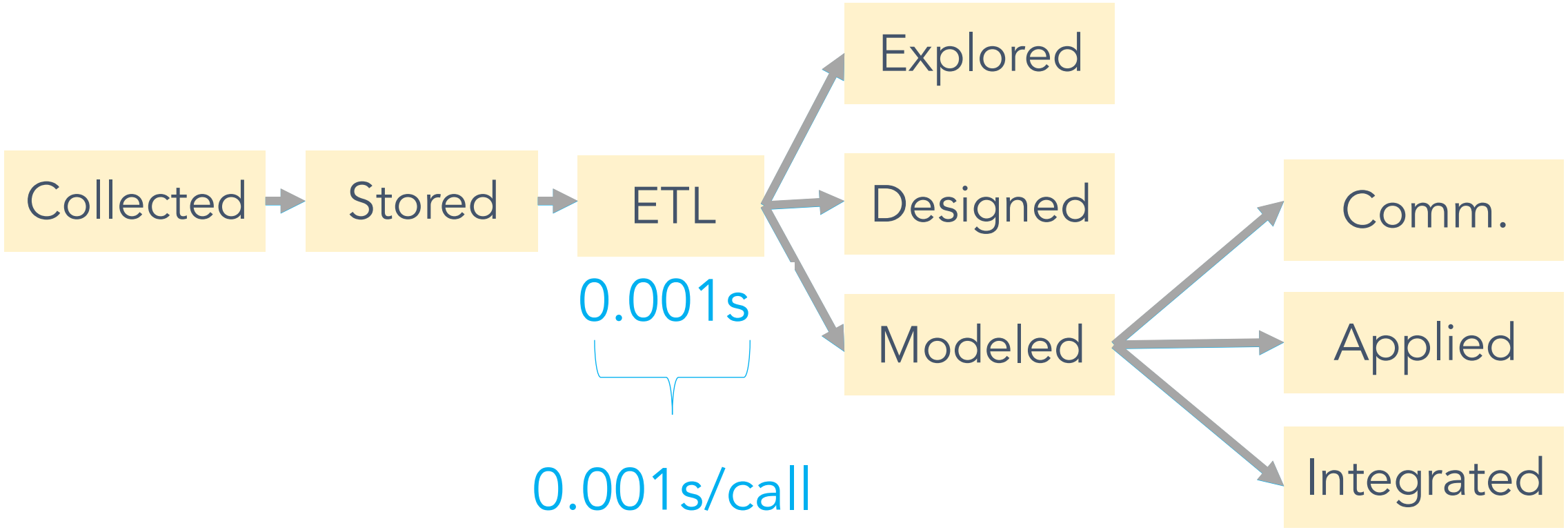
Example #1
Acquisition of one image



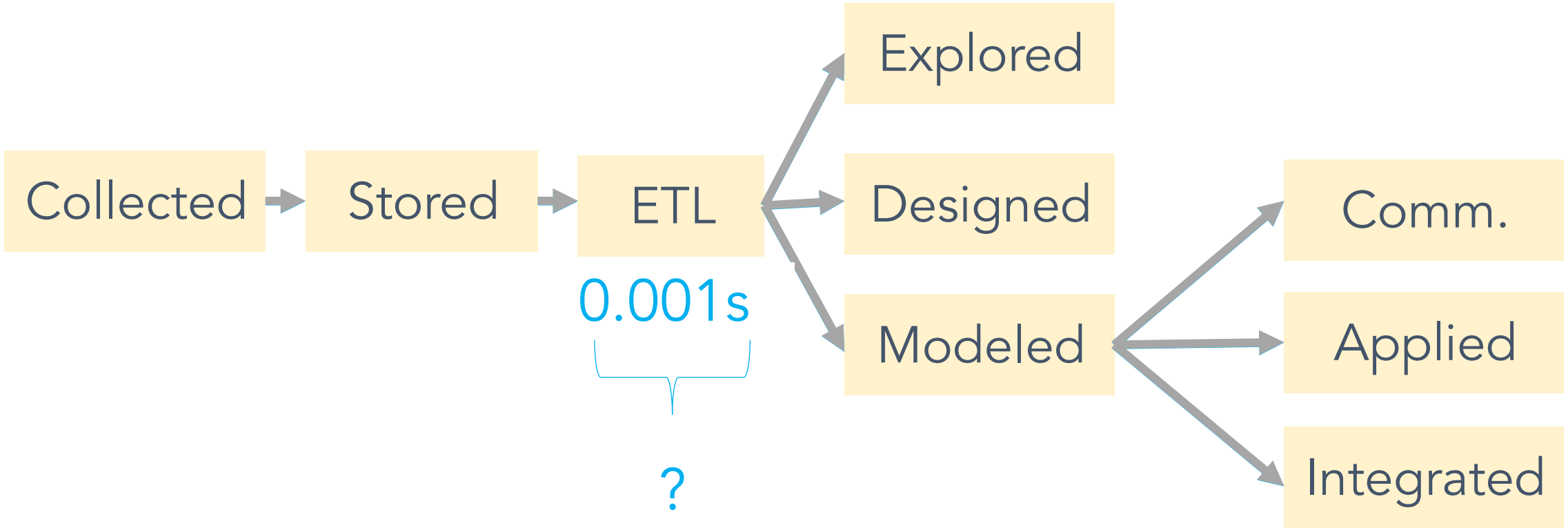
Example #1
Acquisition of 200k images



Example #2
Calculate a number



Example #2
Smooth 200k numbers



What's the point

- When writing code on scale (needs to be applied to a large volume), benchmark each step
- Reduce the number of lines of code
- Package redundant features into functions

<Code Time/>

Exercise 1: Parsing efficiency

Exercise 2: loops numbers

Exercise 3: loops-images

Intro to AWS