Traditional Data Science Pipelines Fail to Adapt
As soon as we change the initial dataset, the whole pipeline fails.

Traditional Data Science Pipeline
- Data Discovery - Relies on the user's knowledge and ability to search
- Data Preparation - Requires trial and error, manual fine-tuning
- Data Analysis - Fragile with respect to data distribution

Modern Adaptive Data Science Pipeline
- Data Discovery - Automated without user input
- Data Preparation - No fine tuning needed
- Data Analysis - Robust to distribution shifts

Leveraging Context to Achieve Adaptivity
Ways to capture context:-
- Goal-oriented data management - How dataset will be used can guide data discovery and prep
- Interactive data management - Users can provide additional knowledge
- Historical executions

Goal-oriented Data Discovery
- Discover new attributes
    - Augmentation through database joins

Challenges to automate data discovery:-
- Scale
- Heterogeneity of formats
- Presence of noise
- Missing schema and key information
- Most attributes are useless for the given problem

Traditional Approaches for Discovery
- Identify different ways to generate a robust search index
- Search using keywords, examples, natural language
- How to go over millions of datasets manually?

Traditional techniques ignore the objective. What if we perform feature selection?
- Identify a robust search index
- Add all attributes to the initial dataset
    - Adding millions of attributes is unscalable

    - Curse of dimensionality
- Perform feature selection

How to solve the problem?
- Exhaustive search
    - Sequentially calculate utility of every subset of the data and pick the best subset
    - Very high time complexity - n^k queries
- Clustering helps to diversify the search process
    - Similar datasets have similar utility
    - Approach 1: Diversify the search process
        - Cluster attributes by generating data properties to represent attributes
        - Goal: Minimize the distance between intra-cluster attributes
        - Solution: Bandit-based approach - $O(|C|^k)$
    - Approach 2: Leverage monotonicity of utility metric
        - Monotonicity: Easy to guarantee
        - How to make the utility submodular: Greedily choose the best augmentation
    - Final approach: Combination of both

Applications:-
- Scalable System Design - Allows users to interact whenever goal is not well defined
- Collaborations and Deployment - Data Discovery to discover causally related attributes, Semantic feature annotation and discovery, Attributes for cancer research

Summary: Model Training as Context
What we want: Find datasets to perform data analysis
Traditional Techniques:-
- Rely on user's knowledge and searchability
- Highly manual

Our approach:-
- Uses model training component to guide data discovery

Deployment: How to explain the output?
Option 1: Change the attribute - wrong approach
Option 2: Look at the data - Simpson's Paradox (Statistical phenomenon where an association between two variables in a population emerges, disappears, or reverses when the population is divided into subpopulations)

Need to capture causal dependencies - use a causal graph

Ladder of causation

Using Causal Reasoning for Responsible Analytics - Opaque algorithm: How do we evaluate fairness of this algorithm?

Key Takeaways

- A novel framework for data discovery
    - Using downstream goal to provide context
        - Does not need to search queries
        - Adapts to varied tasks
        - Handles millions of input datasets
- Explanation Framework
    - Using causal inference for reliable explanations
        - Generates counterfactual explanations
        - Captures causal dependencies between attributes
        - Efficient mechanism to generate explanation

Future Work

Pillars of research

- Effective
    - Correct and easy to use
    - Robust to noise
- Efficient
    - Low runtime
    - Easy to deploy and develop
- Equitable
    - Accessible: Adapts to users with varied domain knowledge and applications
    - Uncovers unwanted biases injected by the pipeline and allows users to ensure fairness
- Make use of multimodal data
- Optimizing data science pipelines