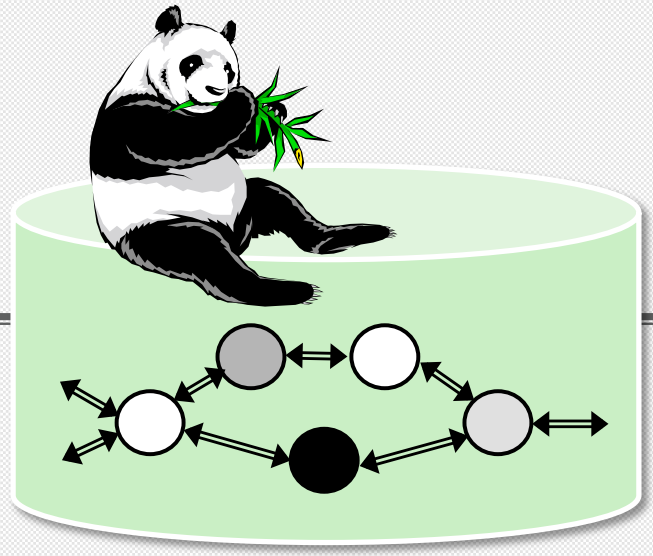
PANDA

A System for Provenance and Data

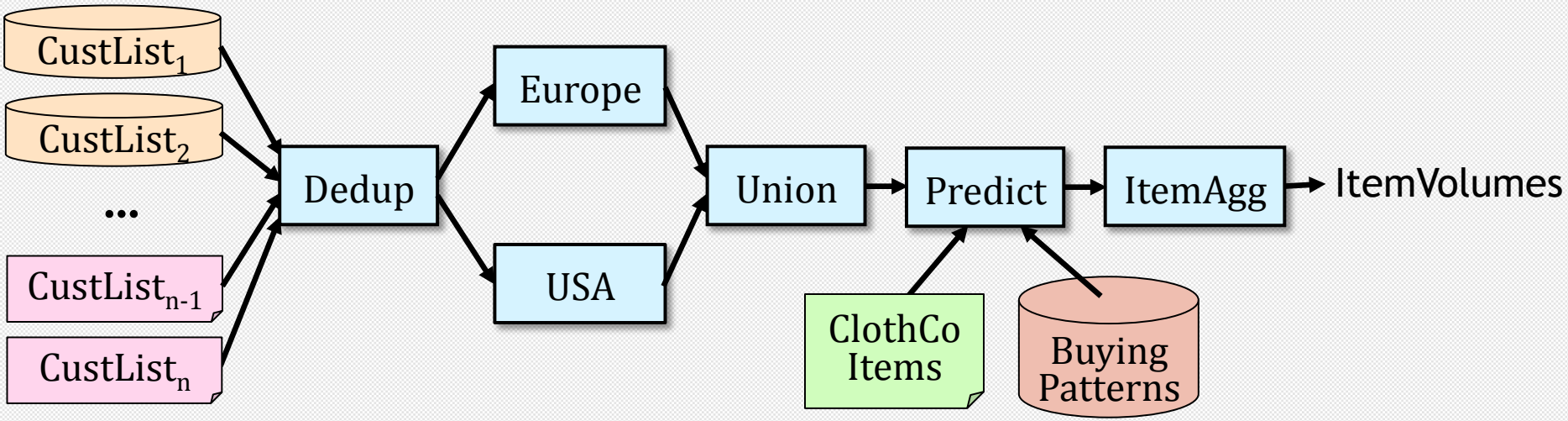


Robert Ikeda

Jennifer Widom

Stanford University

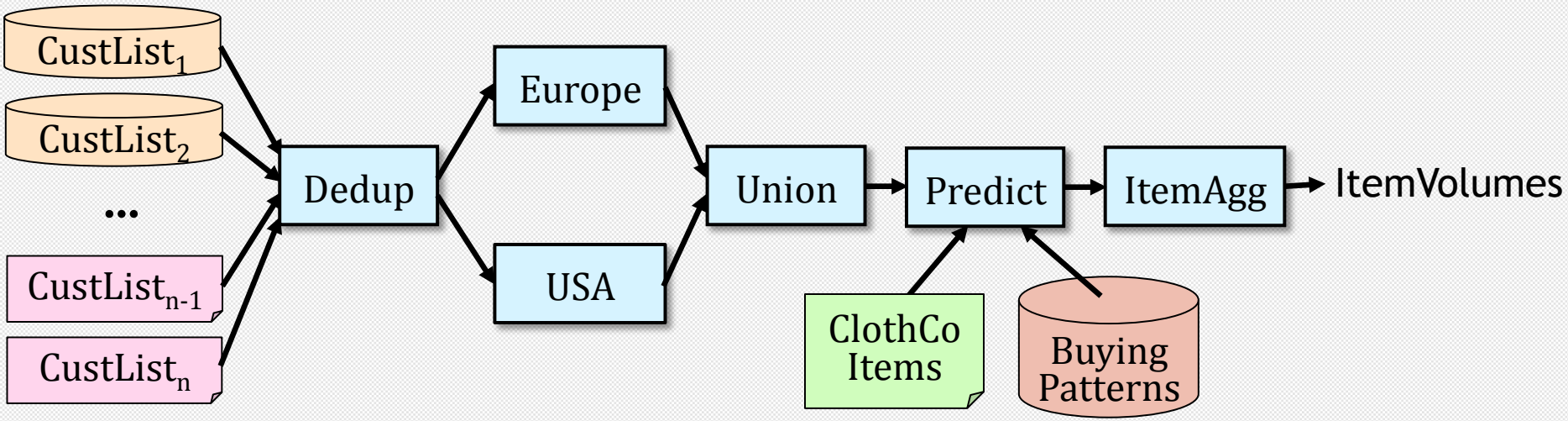
Example



Pipeline for sales predictions

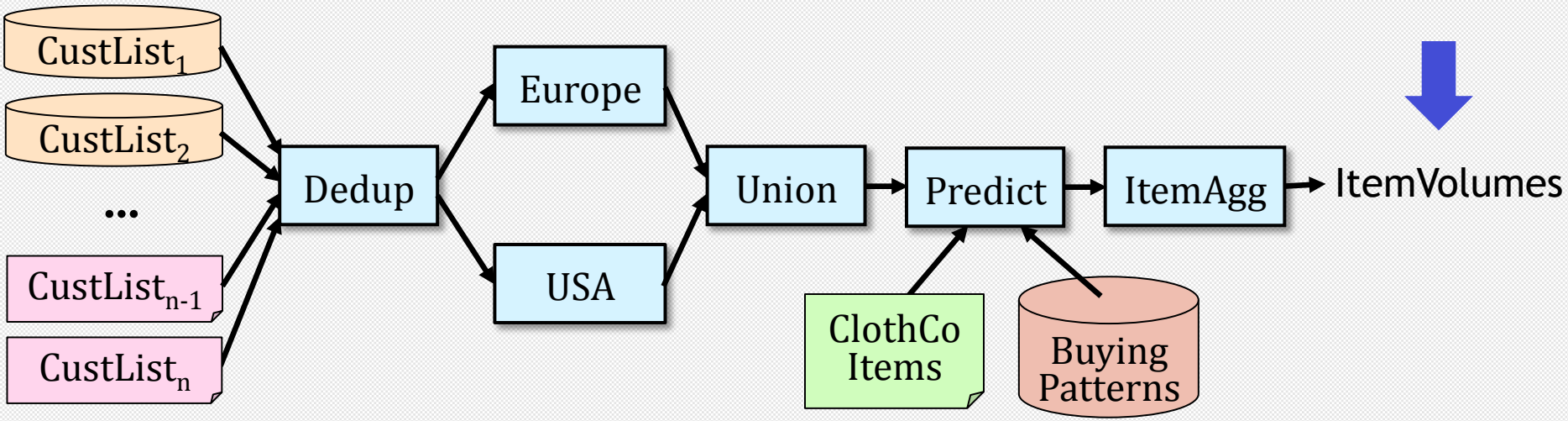
Robert Ikeda

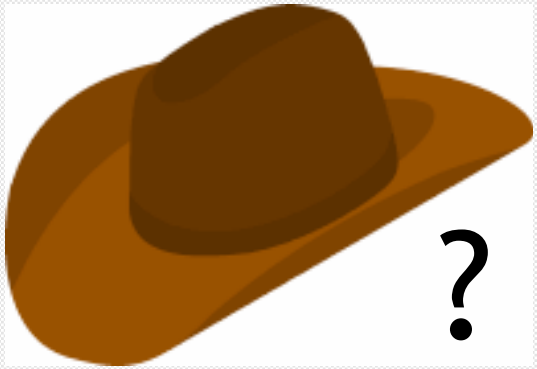
Example



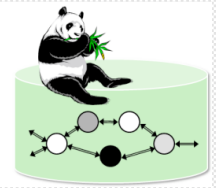
Robert Ikeda

Example



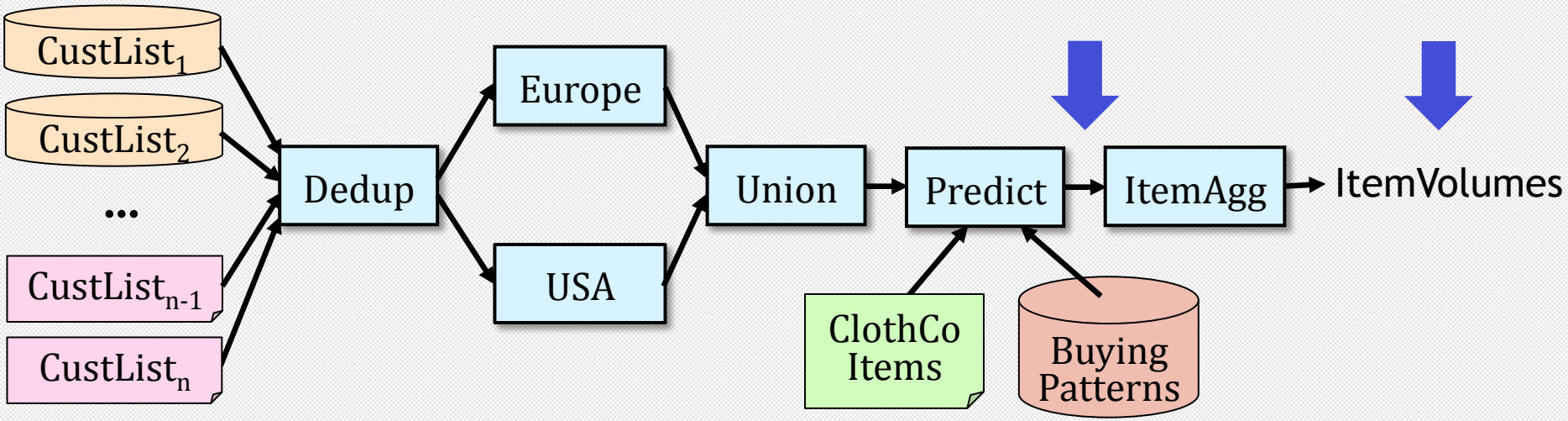


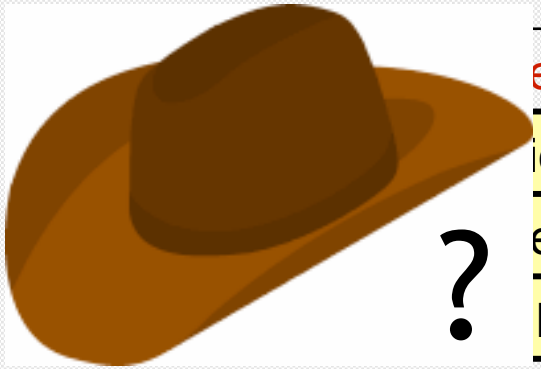
Item	Demand
Cowboy Hat	3



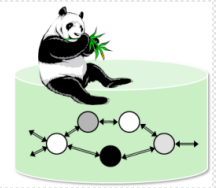
Robert Ikeda

Example



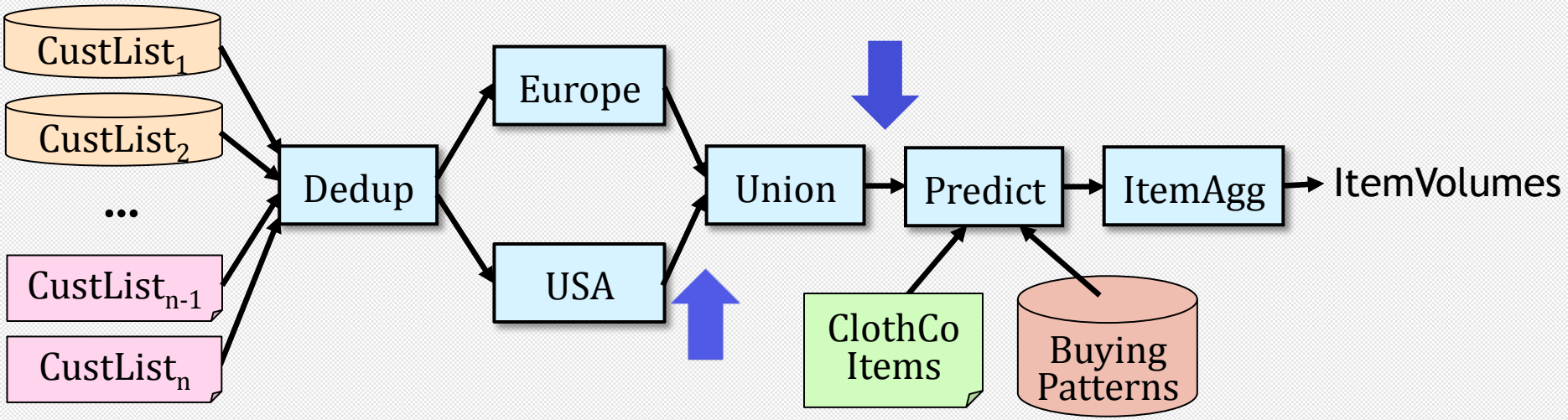


Item	Demand
Cowboy Hat	3
Cowboy Hat	
Cowboy Hat	

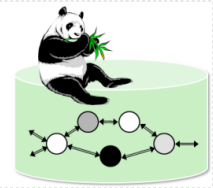


Robert Ikeda

Example

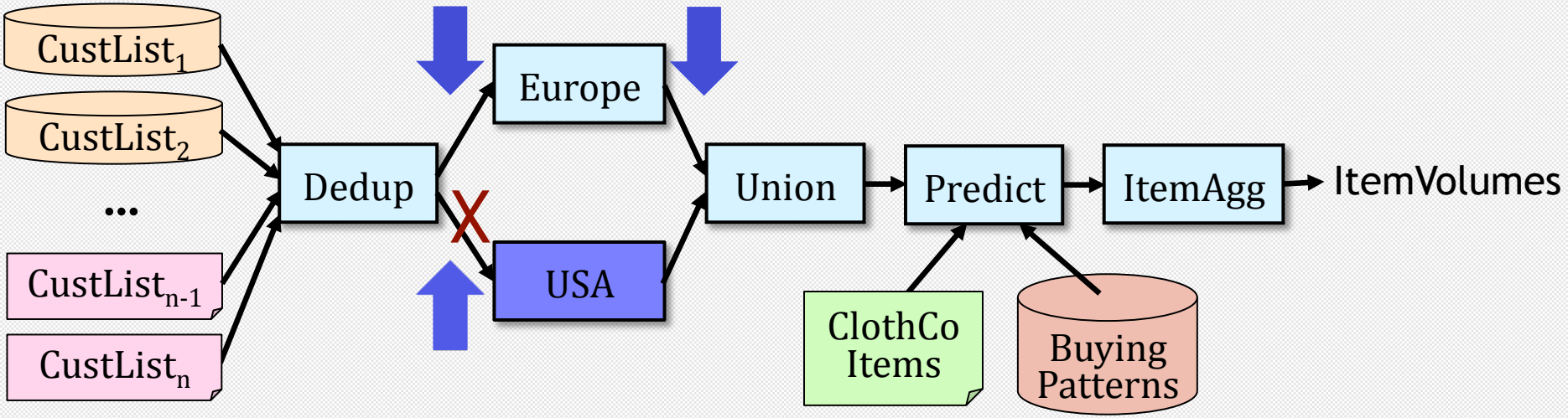


Name	Address
Amelie	…Paris, TX
Jacques	…Paris, TX
Isabelle	…Paris, TX



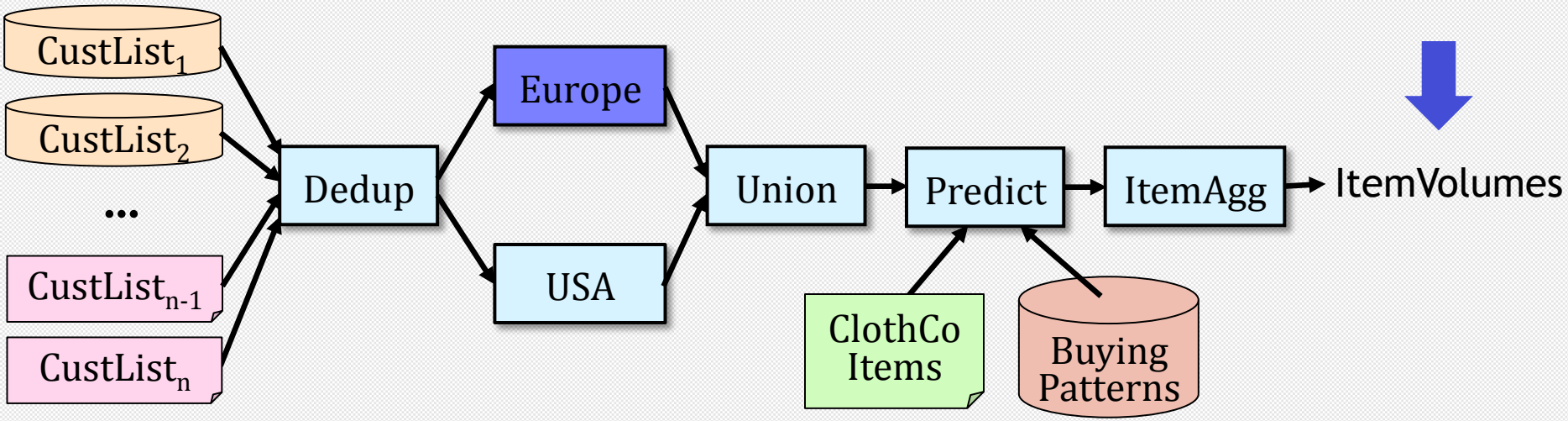
Robert Ikeda

Example

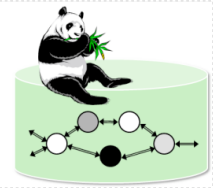


	Name Name	Address Address	
	Amelie	65, quai d'Orsay, Paris, France	
	Jacques	39, rue de Bretagne, Paris, France	
	Isabelle	20 Rue D'orsel, Paris, France	

Example



Item	Demand
Beret	3



Robert Ikeda

Panda

Past work tends to be…

Panda…

1. Either data-based or process-based

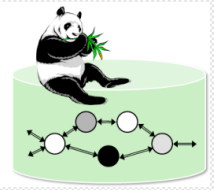
   Capture both — "data-oriented workflows"

2. Focused on modeling and capturing provenance

   Also provenance operators and queries

3. Specific application domains

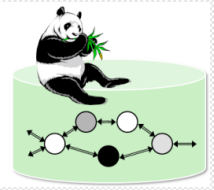
   General-purpose



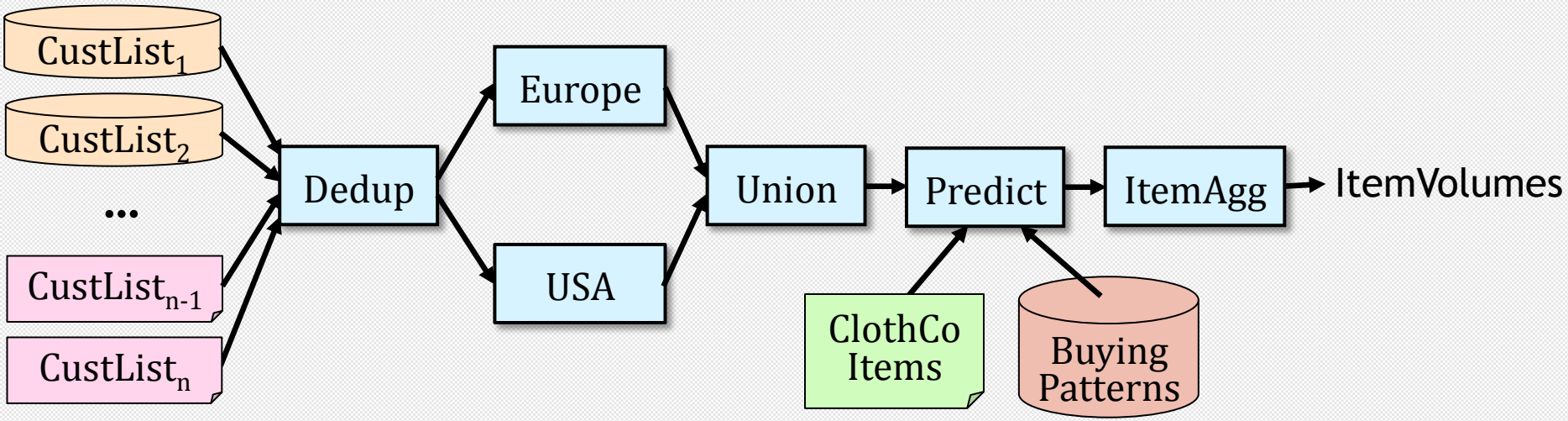
Robert Ikeda

Remainder of Talk

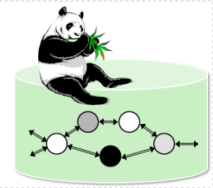
- Processing nodes and provenance capture
- Provenance operations
- Provenance queries
- System and other issues
- Current research



Processing Nodes



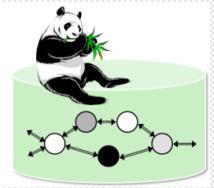
- Relational nodes: structured, well-understood operations
- Opaque nodes



Robert Ikeda

Provenance Capture

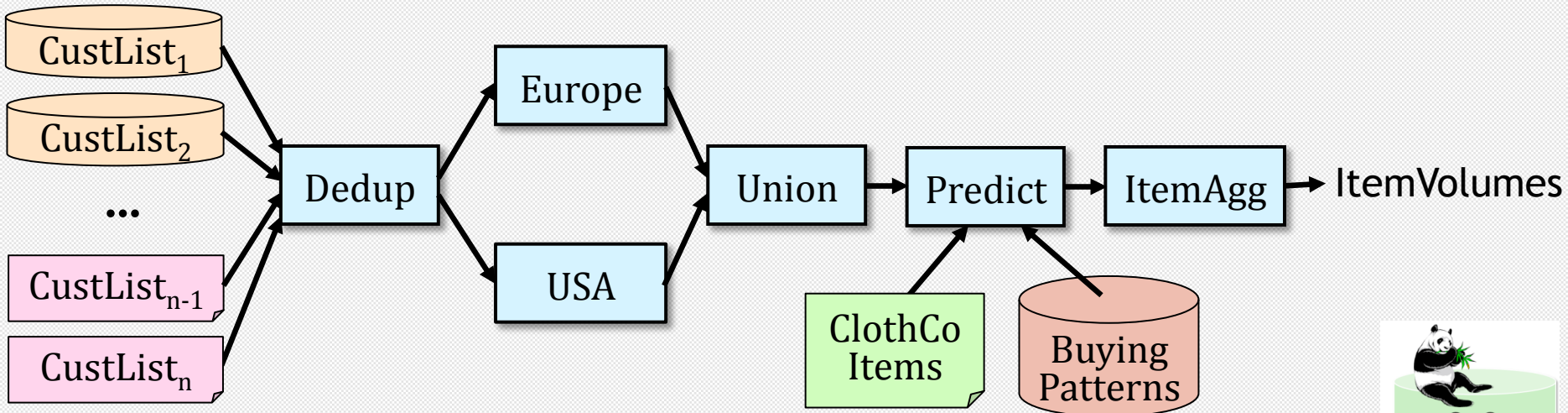
- Model
  - Likely to be similar to Open Provenance Model
  - Support provenance at a variety of granularities
- Interface
  - Allow processing nodes to create and manipulate provenance
  - For relational operations, can plug in existing provenance work



Robert Ikeda

Provenance Operations

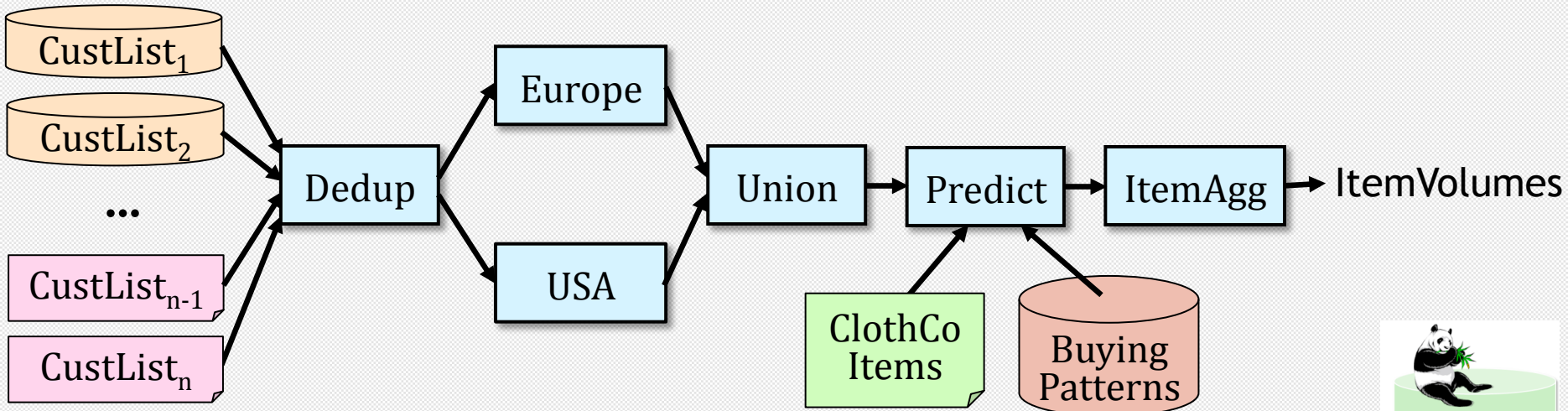
- Basic operations
  - Backward tracing
    - Where did the cowboy-hat record come from?
  - Forward tracing
    - Which predictions did this customer contribute to?



Robert Ikeda

Provenance Operations

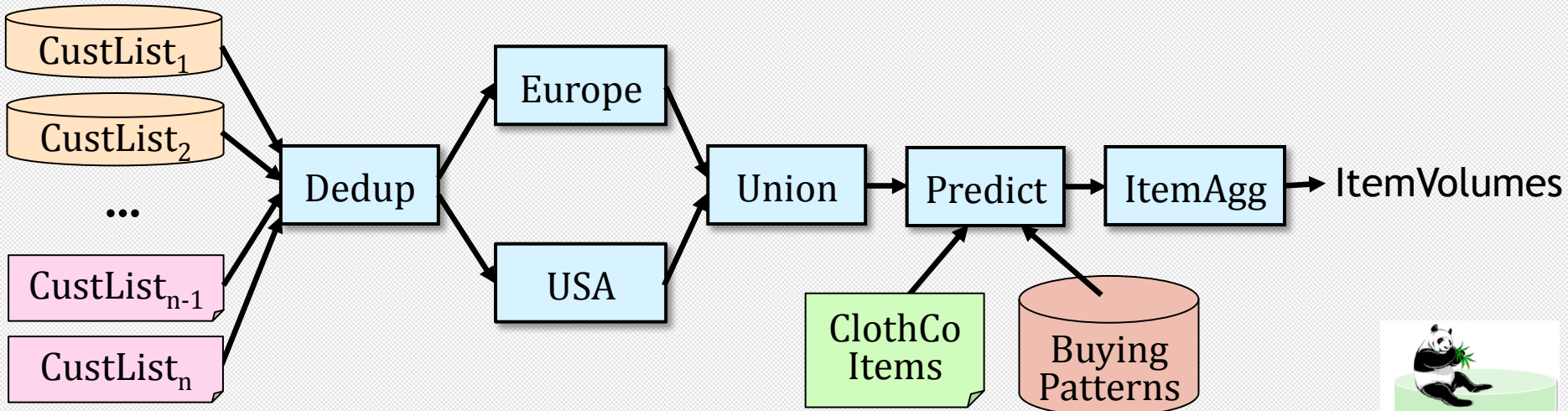
- Examples of additional functionality
  - Forward propagation
    - Update all affected predictions after customers have moved from France to Texas



Robert Ikeda

Provenance Operations

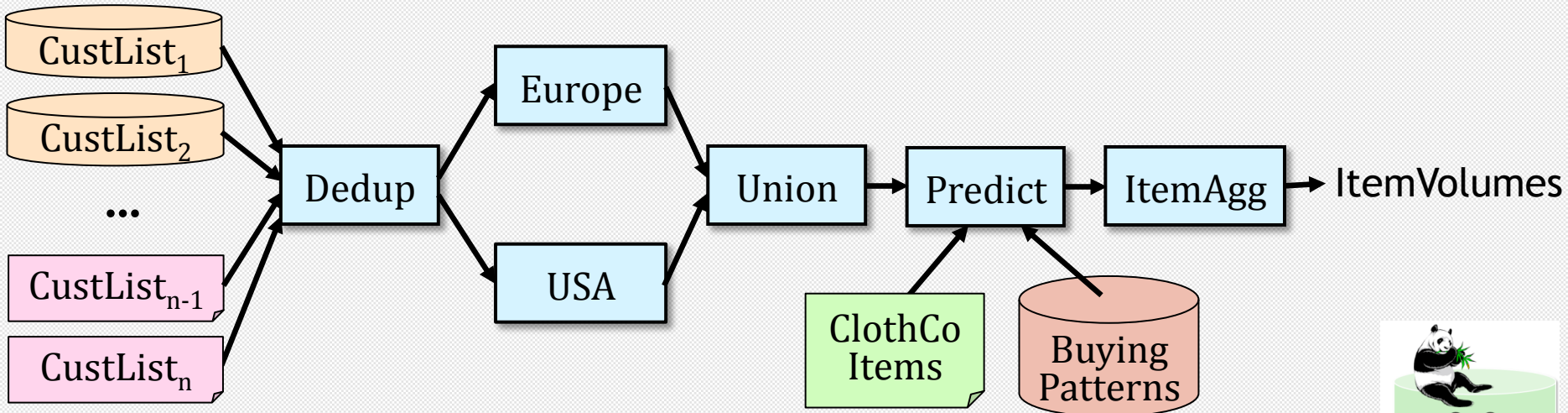
- Examples of additional functionality
  - Refresh ≈ Backward tracing + forward propagation
    - Get latest predicted volume for cowboy hat sales (only) using latest customer lists and buying patterns



Robert Ikeda

Provenance Queries

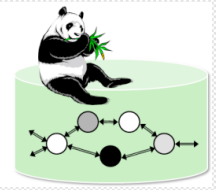
- Examples
  - How many people from each country contributed to the cowboy hat prediction?
  - Which customer list contributed the most to the top 100 predicted items?



Robert Ikeda

Provenance Queries

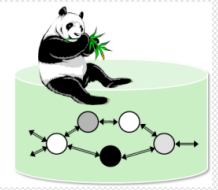
- Examples
  - How many people from each country contributed to the cowboy hat prediction?
  - Which customer list contributed the most to the top 100 predicted items?
- Seamlessly combine provenance and data
- Compact and intuitive language
- Amenable to optimization



Robert Ikeda

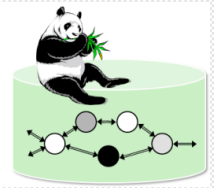
System and Other Issues

- Query-driven provenance capture
- Eager vs. lazy computation and storage
- Fine-grained vs. coarse-grained
- Approximate provenance



Current Research

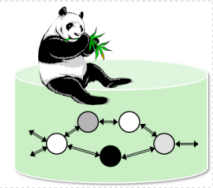
- Building up basic system infrastructure
- Refresh
  - Efficiently compute the up-to-date value of selected output elements
- Theoretical challenges
  - Optimizing provenance storage vs. recomputation



Robert Ikeda

System Infrastructure

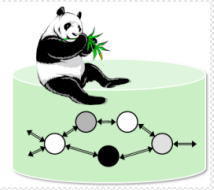
- Handles structured relational operations as well as arbitrary Python processing nodes
- Arbitrary acyclic transformation graphs
- Backward tracing and forward propagation



Robert Ikeda

Refresh

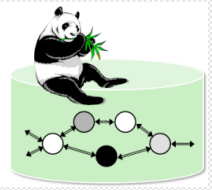
- Problem
  - Efficiently compute the up-to-date value of selected output elements
- Challenges
  - Formally defining the refresh problem
  - Understanding when refresh can be done efficiently
  - Supporting a wide class of transformations and workflows



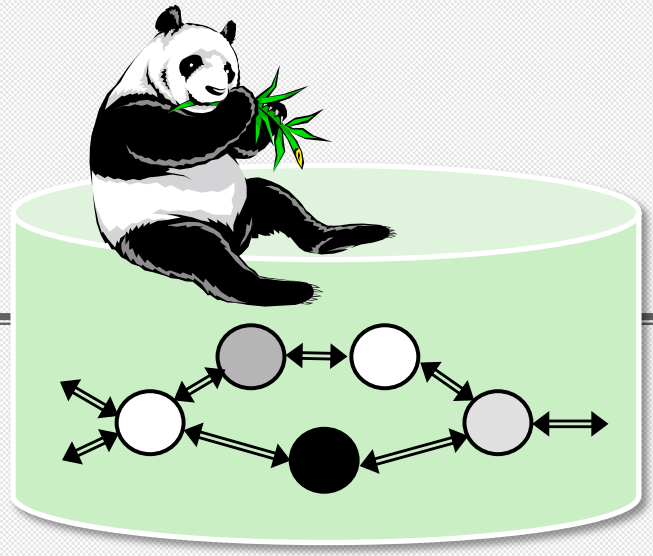
Robert Ikeda

Future Work

- Most everything in this talk ☺

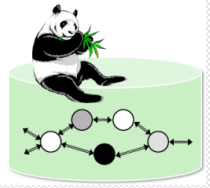


Thank You

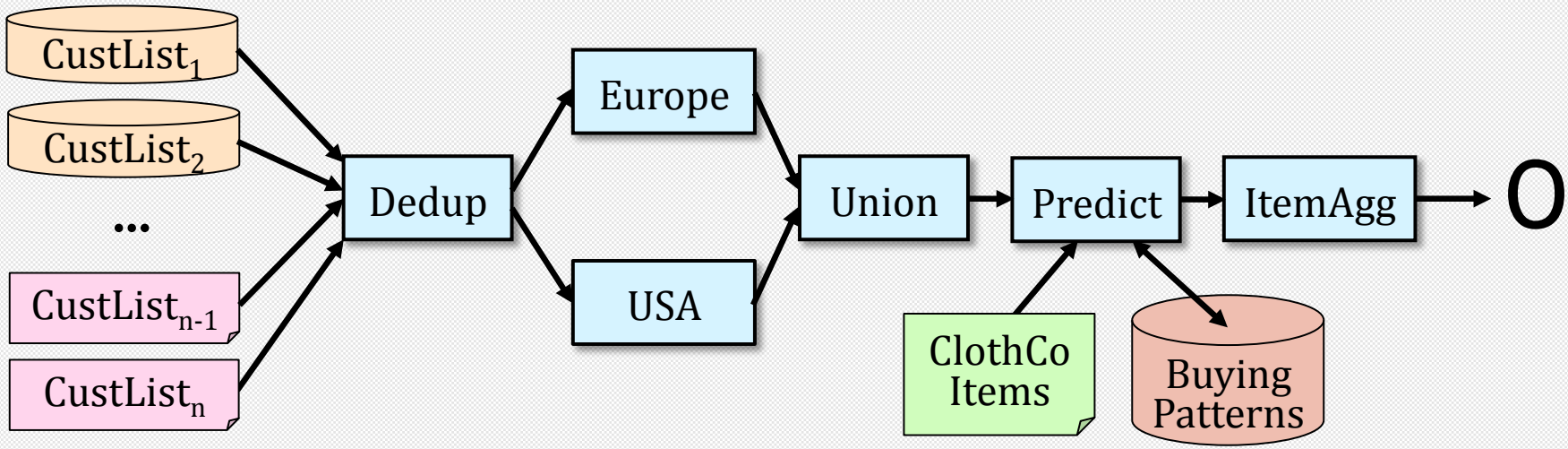


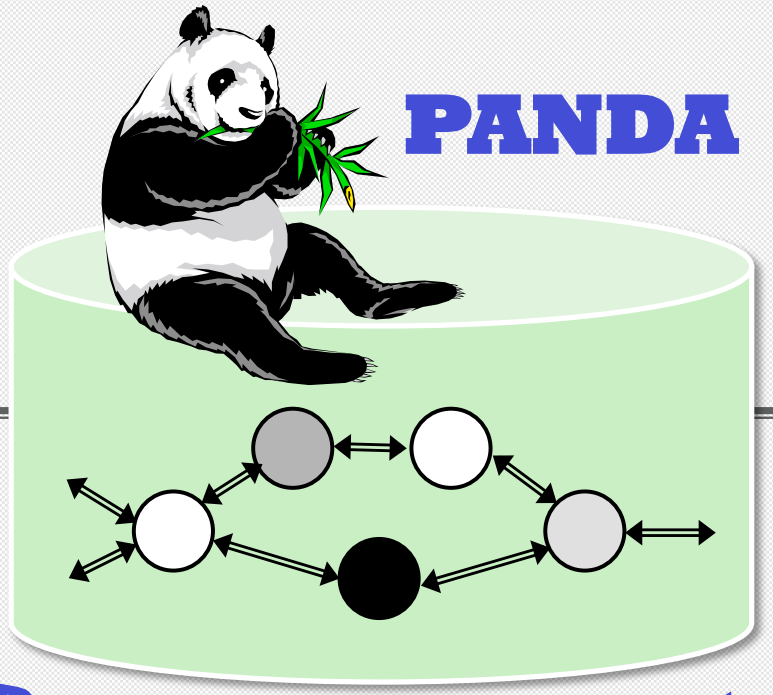
Parag Agrawal, Abhijeet Mohapatra,
Raghotham Murthy, Aditya Parameswaran,
Hyunjung Park, Alkis Polyzotis,
Semih Salihoglu

Extra Slides



Running Example

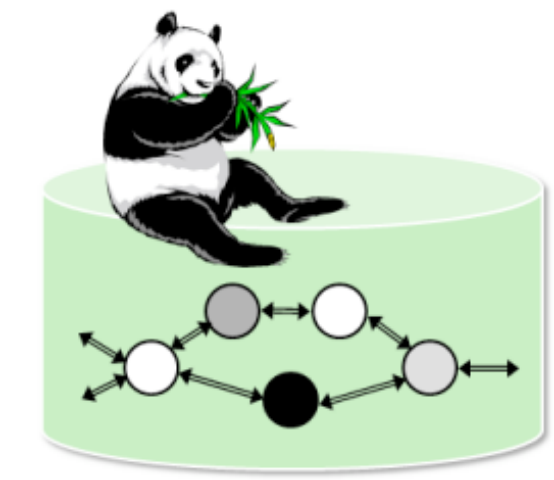




Provenance and Data

PANDA

A System for Provenance and Data



Robert Ikeda

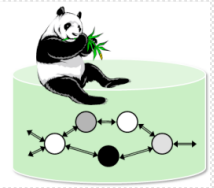
Jennifer Widom

Stanford University

Panda's Niche

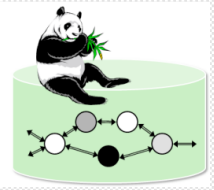
1. Data-based or process-based
2. Modeling and capturing provenance
3. Specific application domains

1. Merge data-based and process-based
2. Provenance operators and queries
3. General-purpose

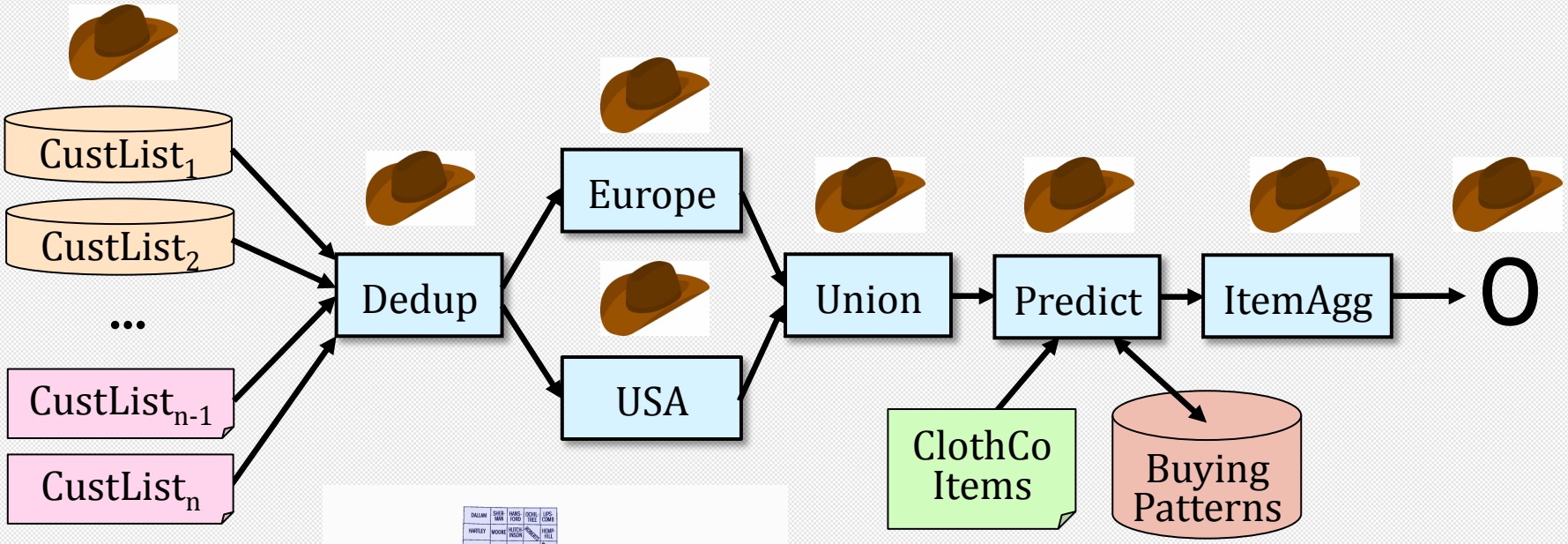


Overview of Past Work

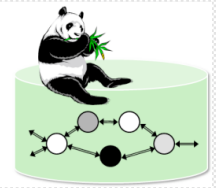
1. Data-based or process-based
2. Modeling and capturing provenance
3. Specific application domains



Running Example

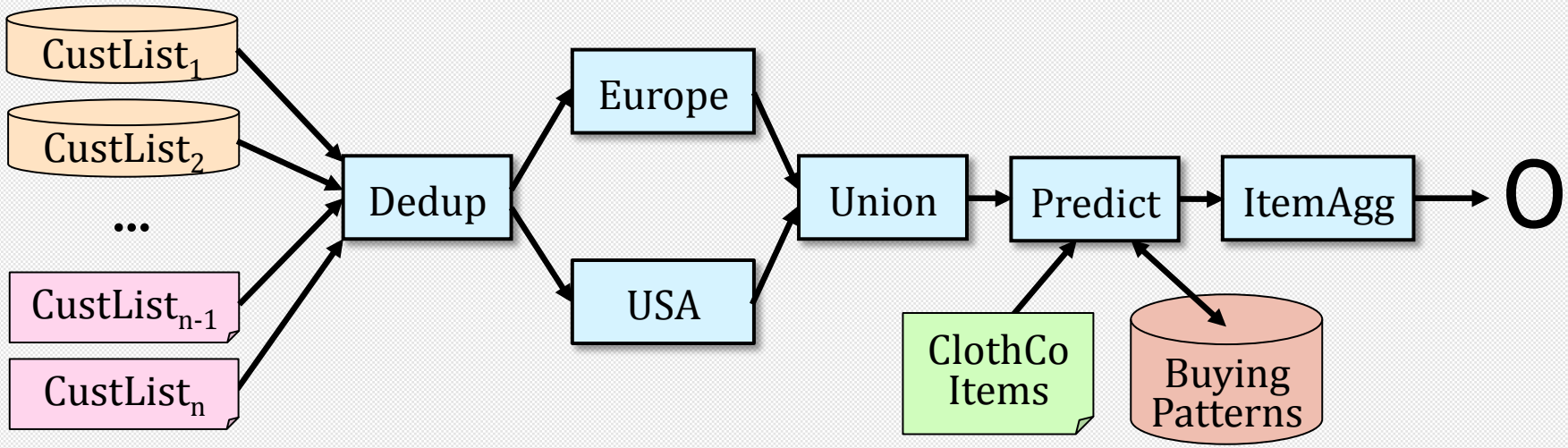


Paris, France ?

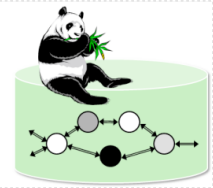


Robert Ikeda

Running Example

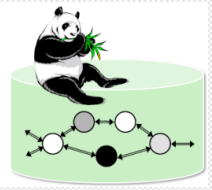


Pipeline for Sales Prediction



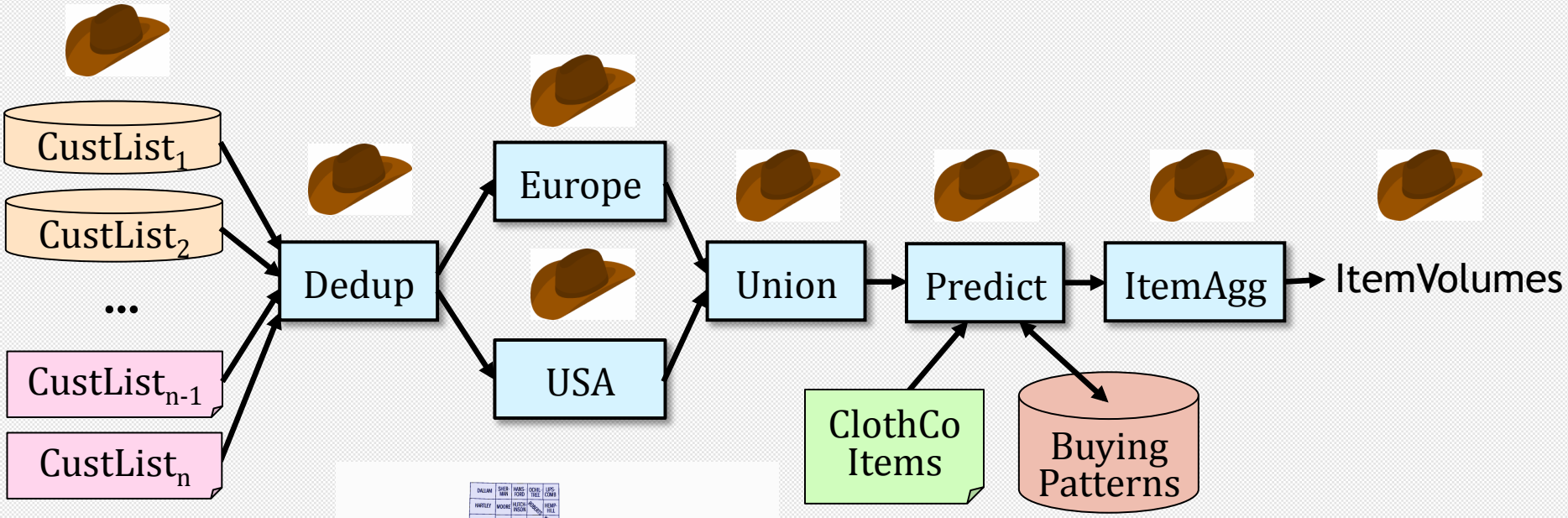
Provenance Capture

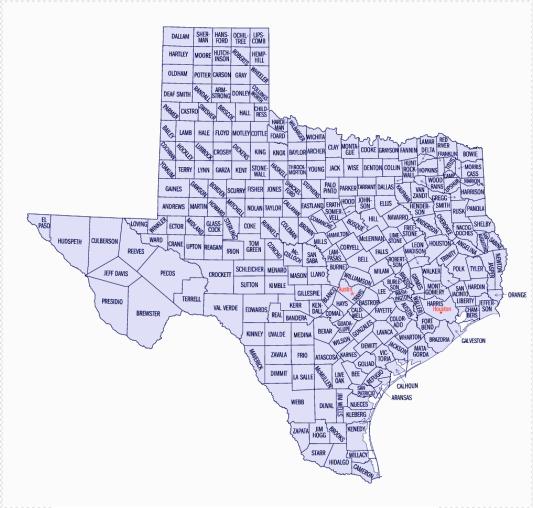
- Processing Nodes
  - Relational operations
  - Opaque processing
- Requirements
  - Interface
  - Model



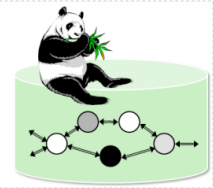
Robert Ikeda

Running Example





Paris, France ?

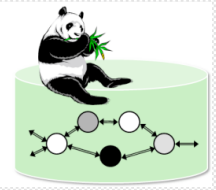


Robert Ikeda

Processing Nodes

- Relational Operations
  - Relational operations
  - Opaque processing
- Opaque Processing
  - Interface
  - Model

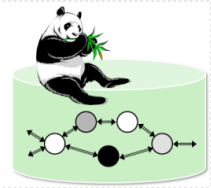
Robert Ikeda



Provenance Queries

- Operate over provenance and data
- Compact and intuitive
- Amenable to efficient planning

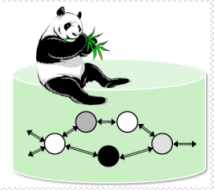
Considering only customers from a specific list, which items are in the highest demand?



Robert Ikeda

Provenance Queries

- Seamlessly combine provenance and data
- Compact and intuitive language
- Amenable to optimization

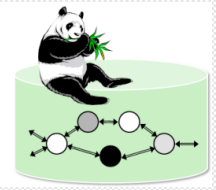


Robert Ikeda

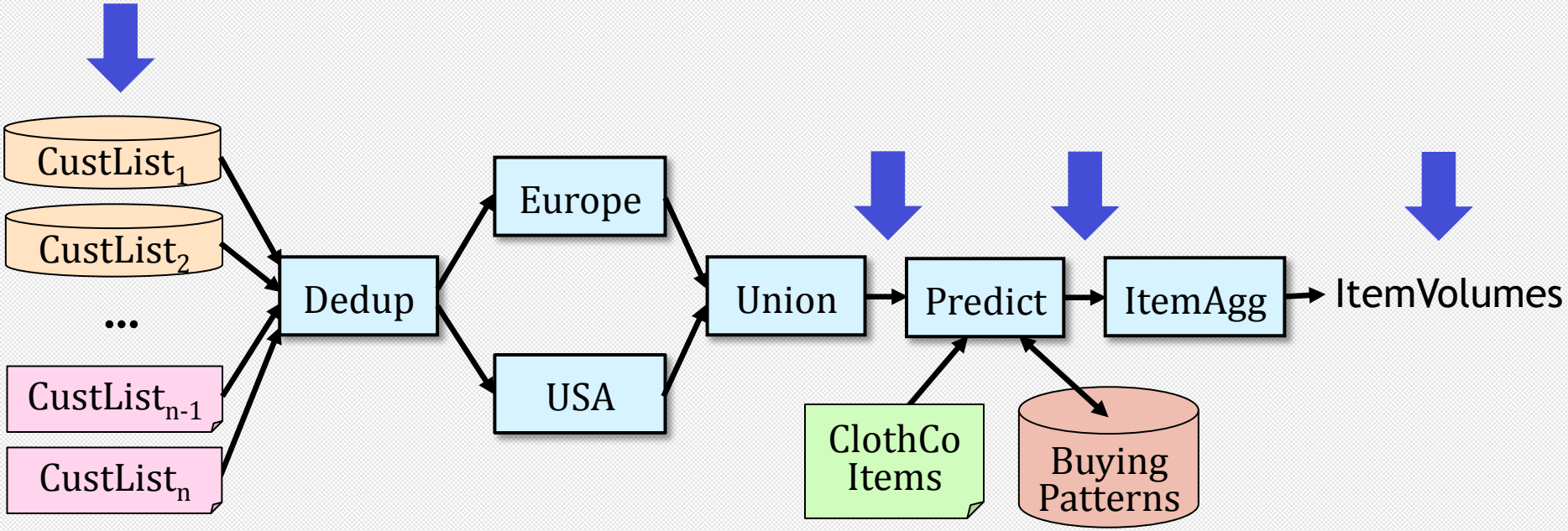
Provenance Query Examples

- How many people from each country contributed to the cowboy hat prediction?
- Which customer list contributed the most to the top 100 predicted items?

Robert Ikeda



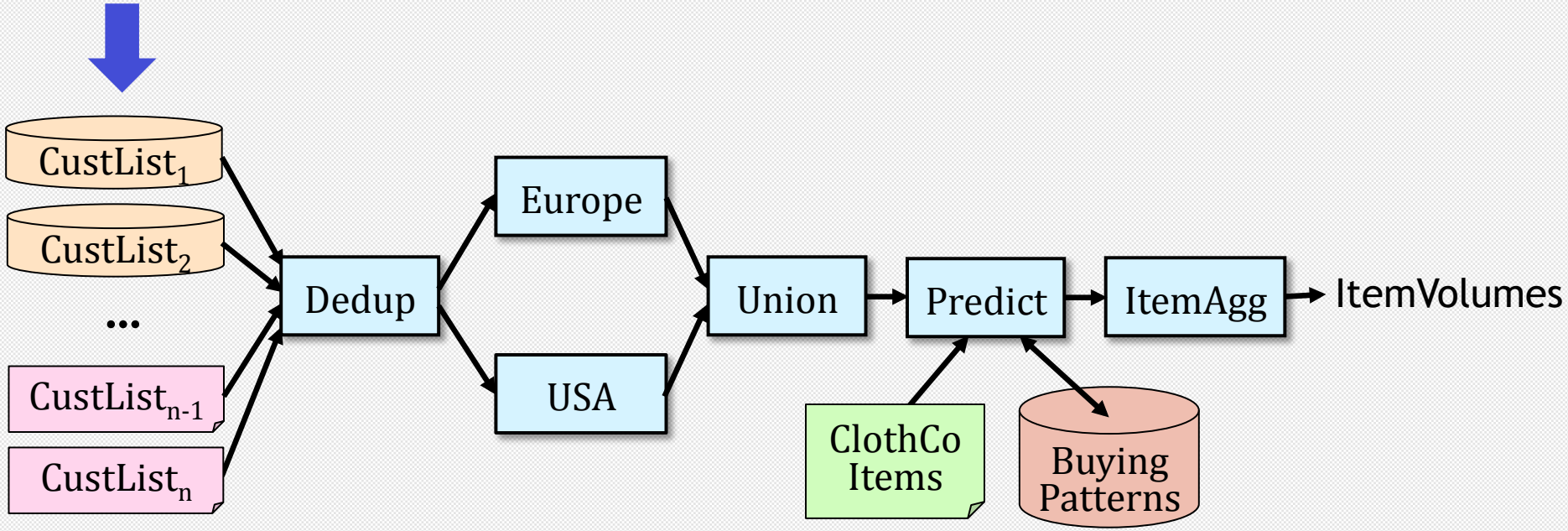
Running Example



Name	Address	Item	Demand
Amelie	65, quai d'Orsay, Paris	Hat	3
Jacques	39, rue de Bretagne, Paris	Hat	
Isabelle	20 Rue D'orsel, Paris	Hat	

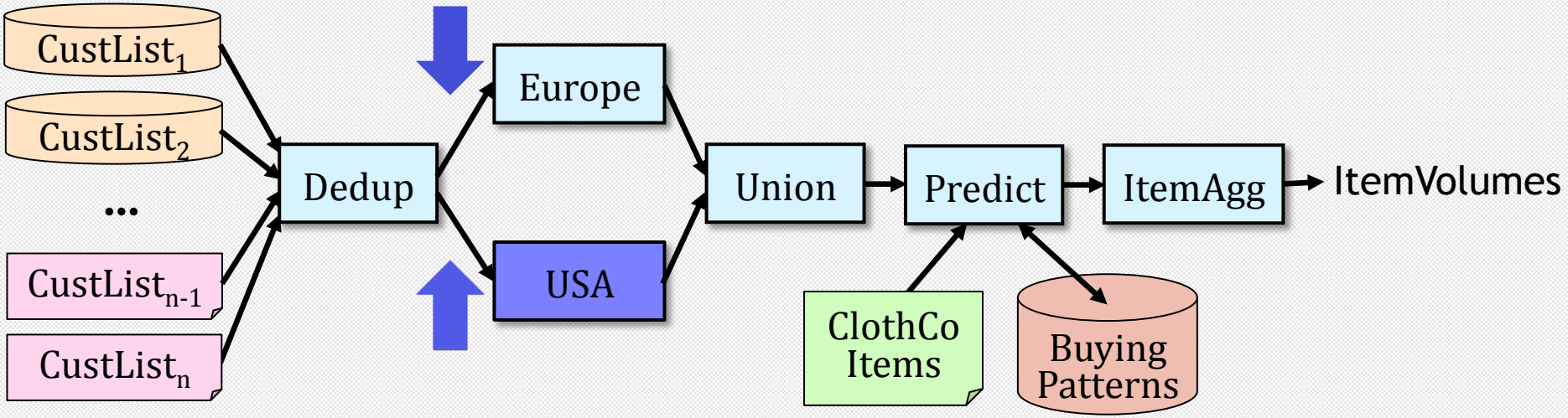
Robert Ikeda

Running Example

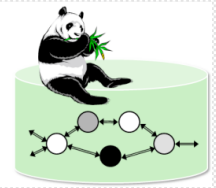


Name	Address
Amelie	65, quai d'Orsay, Paris
Jacques	39, rue de Bretagne, Paris
Isabelle	20 Rue D'orsel, Paris

Running Example

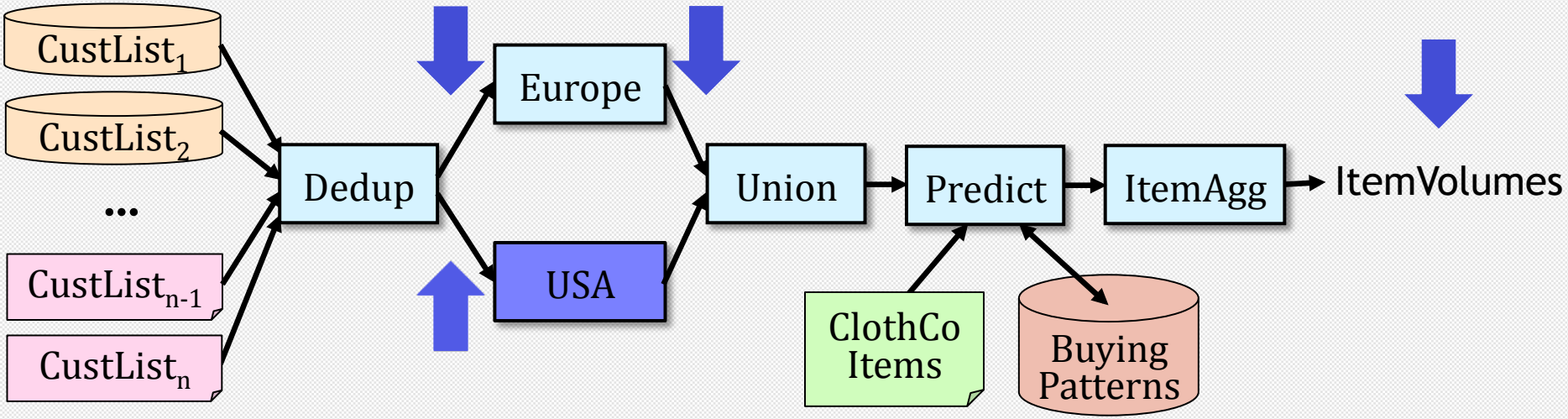


Name	Address
Amelie	65, quai d'Orsay, Paris
Jacques	39, rue de Bretagne, Paris
Isabelle	20 Rue D'orsel, Paris

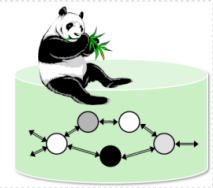


Robert Ikeda

Running Example

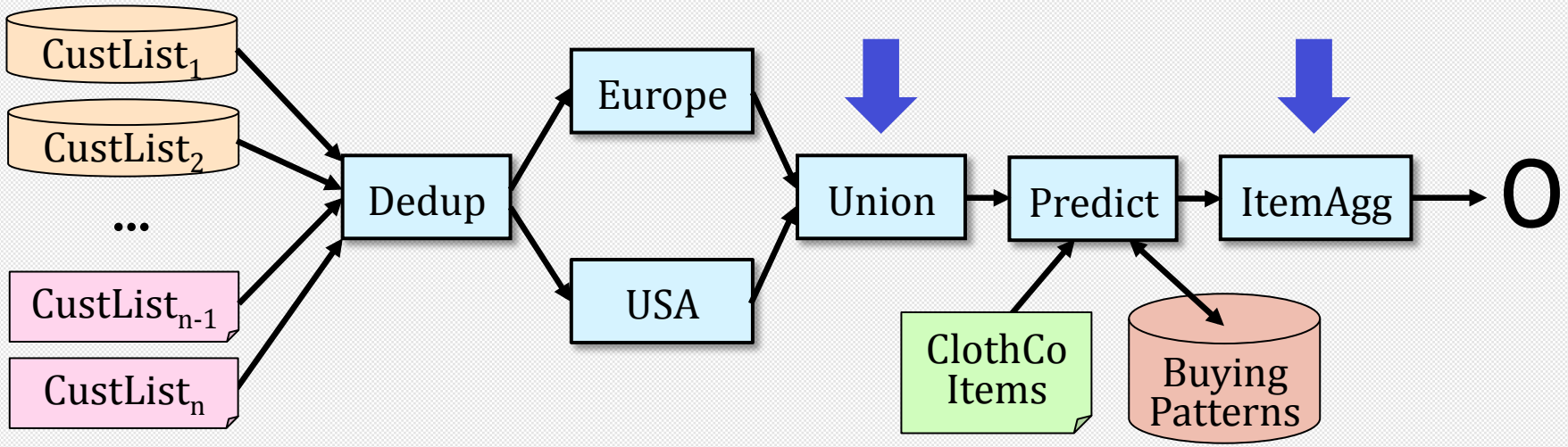


Item	Demand
Beret	3

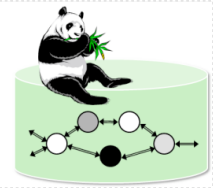


Robert Ikeda

Processing Nodes

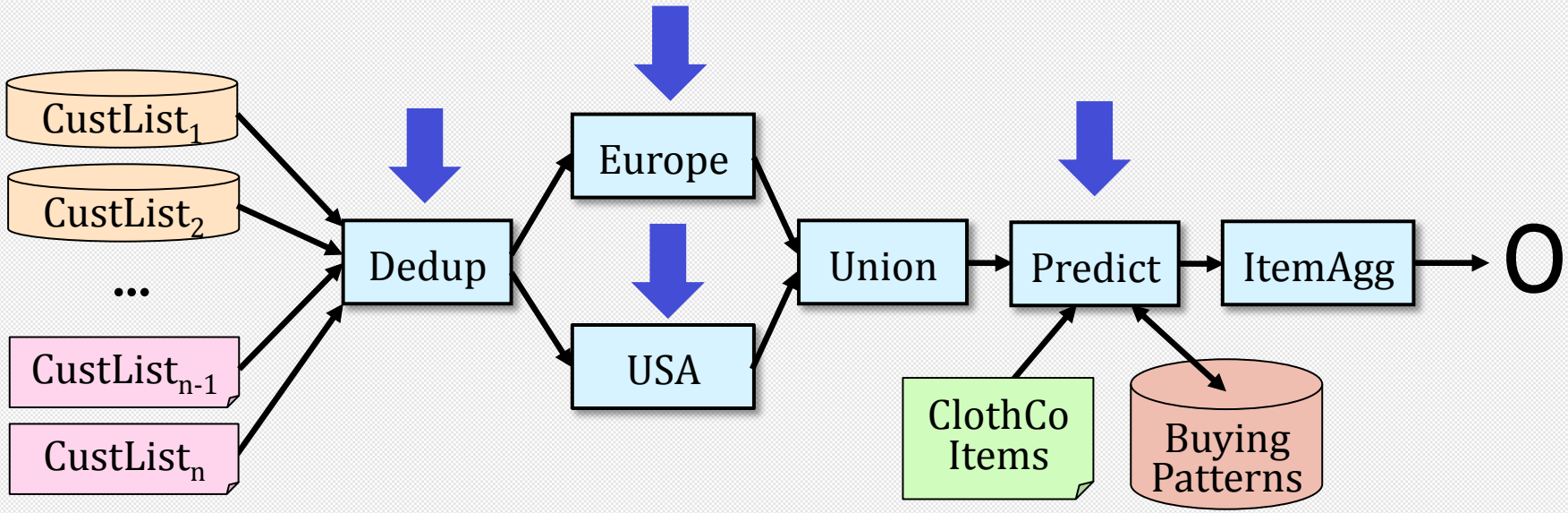


Relational Nodes: Structured, well-understood operations



Robert Ikeda

Processing Nodes

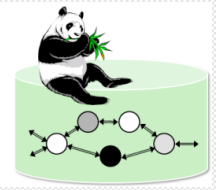


Opaque Nodes

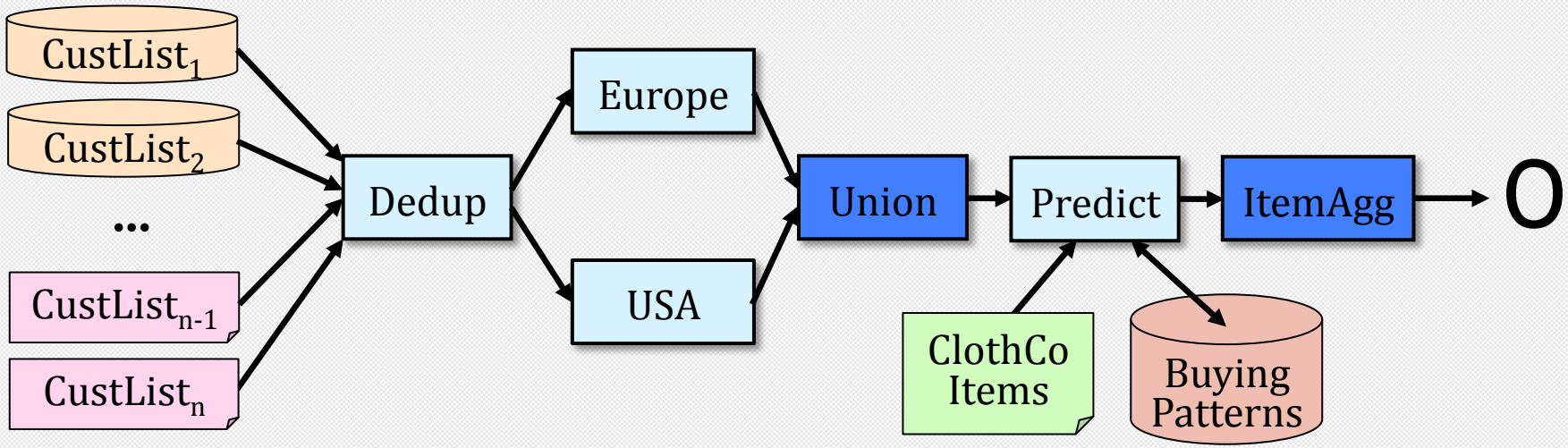
Robert Ikeda

Predicted Uses

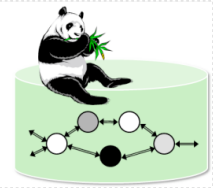
- Explanation
  - How was data derived?
- Verification
  - Is data erroneous or outdated?
- Recomputation
  - Can data be recomputed efficiently?



Processing Nodes

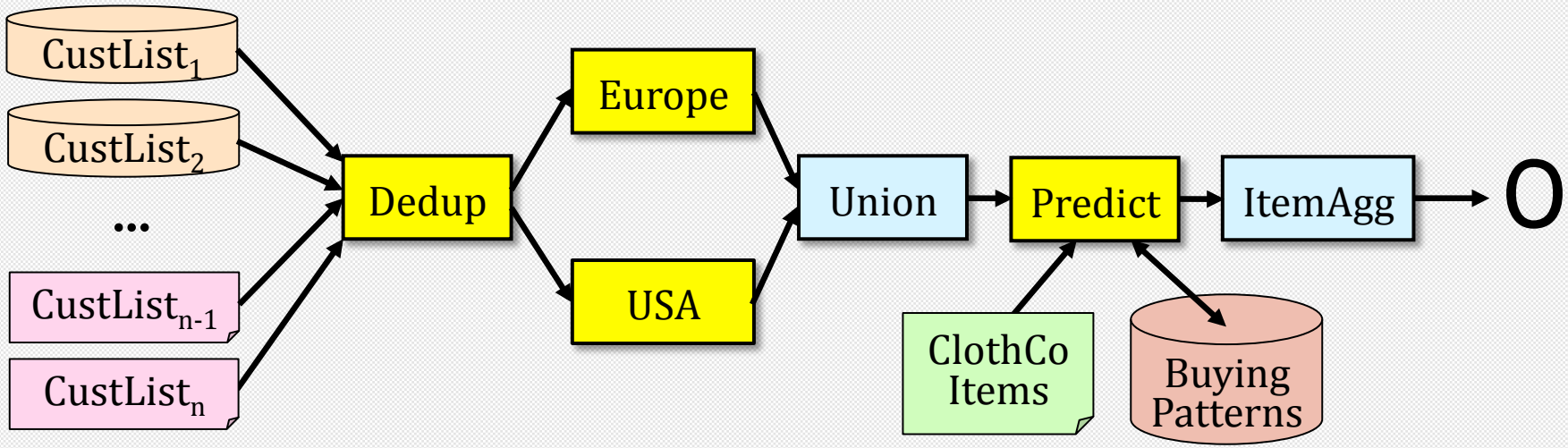


Relational nodes: structured, well-understood operations

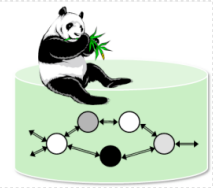


Robert Ikeda

Processing Nodes



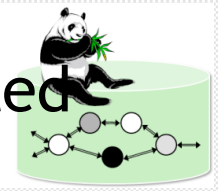
Opaque nodes



Robert Ikeda

Provenance Operations

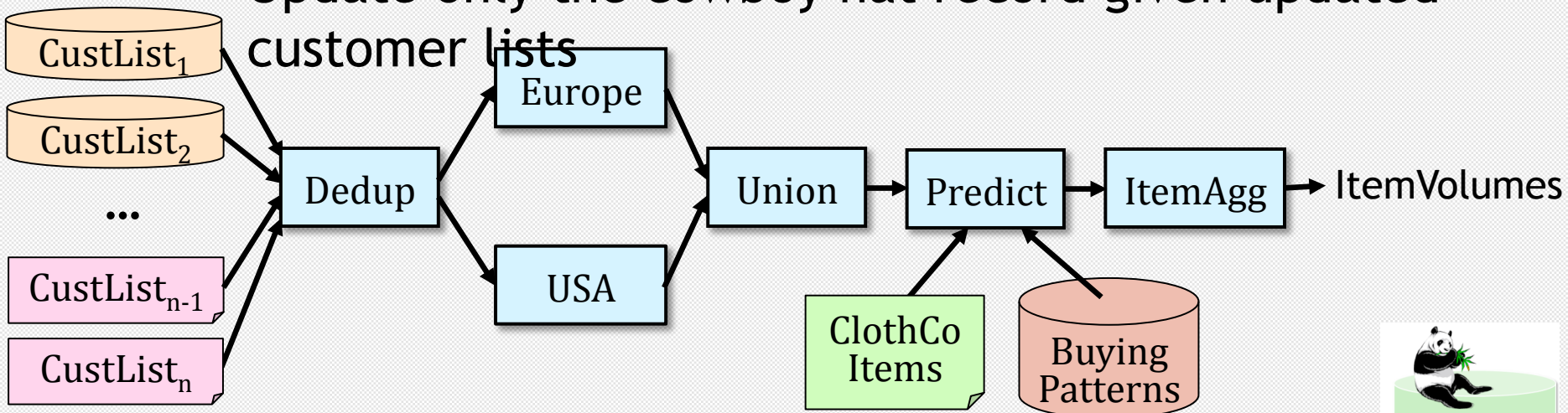
- Basic operations
  - Backward tracing
    - Where did the cowboy-hat record come from?
  - Forward tracing
    - Which predictions did this customer contribute to?
- Examples of additional functionality
  - Forward propagation
    - Update all affected predictions after customers move from France to Texas
  - Refresh ≈ Backward tracing + forward propagation
    - Update only the cowboy hat record given updated customer lists



Robert Ikeda

Provenance Operations

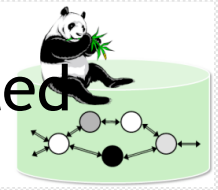
- Examples of additional functionality
  - Forward propagation
    - Update all affected predictions after customers move from France to Texas
  - Refresh ≈ Backward tracing + forward propagation
    - Update only the cowboy hat record given updated customer lists



Robert Ikeda

Provenance Operations

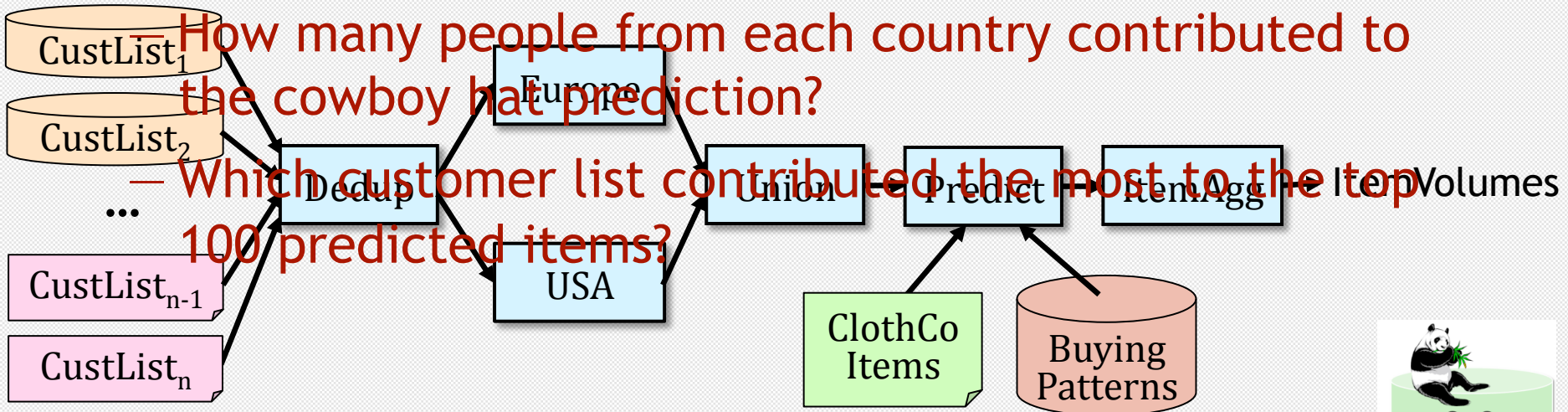
- Basic operations
  - Backward tracing
    - Where did the cowboy-hat record come from?
  - Forward tracing
    - Which predictions did this customer contribute to?
- Examples of additional functionality
  - Forward propagation
    - Update all affected predictions after customers move from France to Texas
  - Refresh ≈ Backward tracing + forward propagation
    - Update only the cowboy hat record given updated customer lists



Robert Ikeda

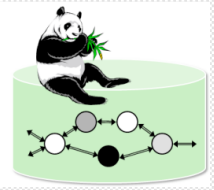
Provenance Queries

- Seamlessly combine provenance and data
- Compact and intuitive language
- Amenable to optimization
- Examples:
  - How many people from each country contributed to the cowboy hat prediction?
  - Which customer list contributed the most to the top 100 predicted items?



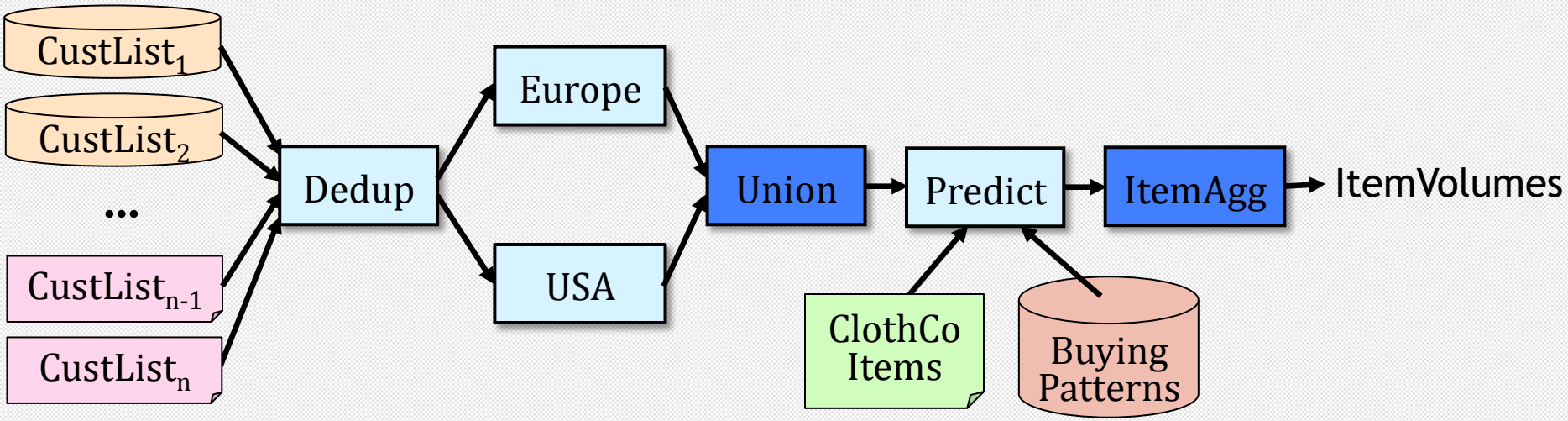
Provenance Queries

- Examples:
  - How many people from each country contributed to the cowboy hat prediction?
  - Which customer list contributed the most to the top 100 predicted items?
- Seamlessly combine provenance and data
- Compact and intuitive language
- Amenable to optimization

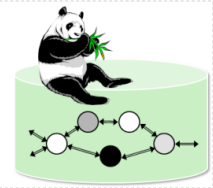


Robert Ikeda

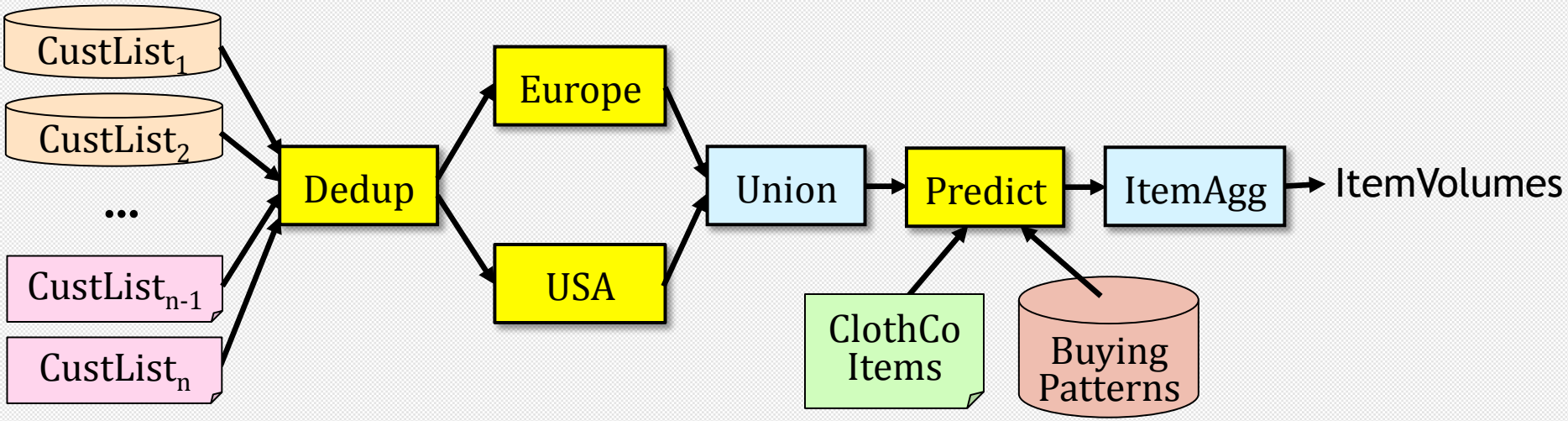
Processing Nodes



Relational nodes: structured, well-understood operations

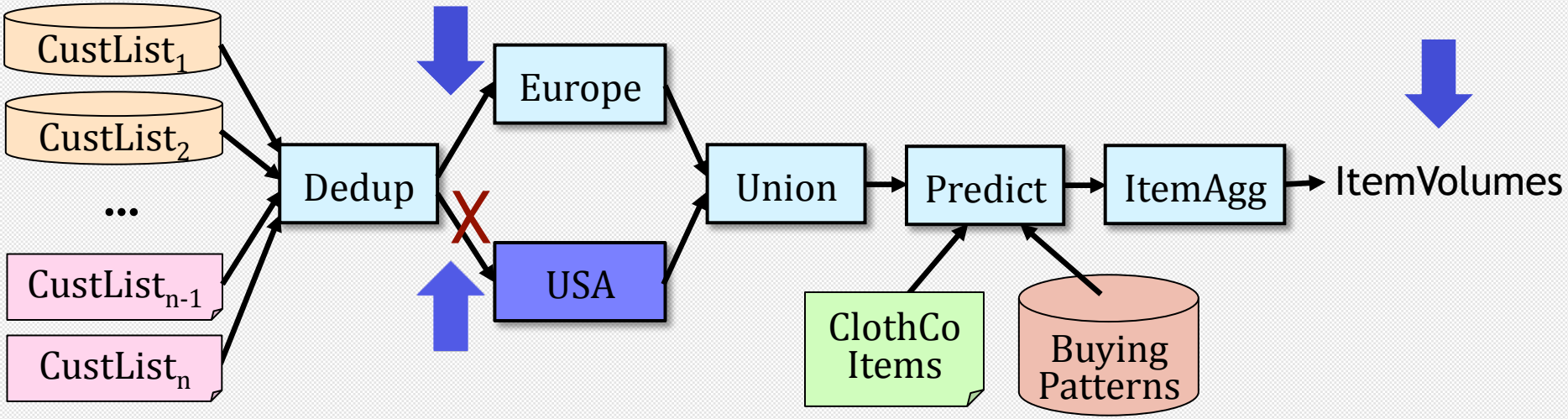


Processing Nodes



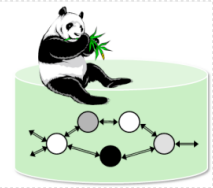
Opaque nodes

Example



Nam	ess
Amé	rsay,
Jacq	tagne, Paris
Isab	sel, Paris

Item	Demand
Beret	3



Robert Ikeda