

The logo for 'celect' is displayed in white lowercase letters on a dark red rounded rectangular background.

celect

A Data Science Platform

Vishal Doshi
vishal@celect.com
@superdosh

Akshay Anand
Sai Kiran Burle
Tony Do
Ying-zong Huang
Peter Lin
Michael Lowney

Ritesh Madan
Balaji Rengarajan
Ben Schoener
Devavrat Shah
Jae Sim
Daniel Xu

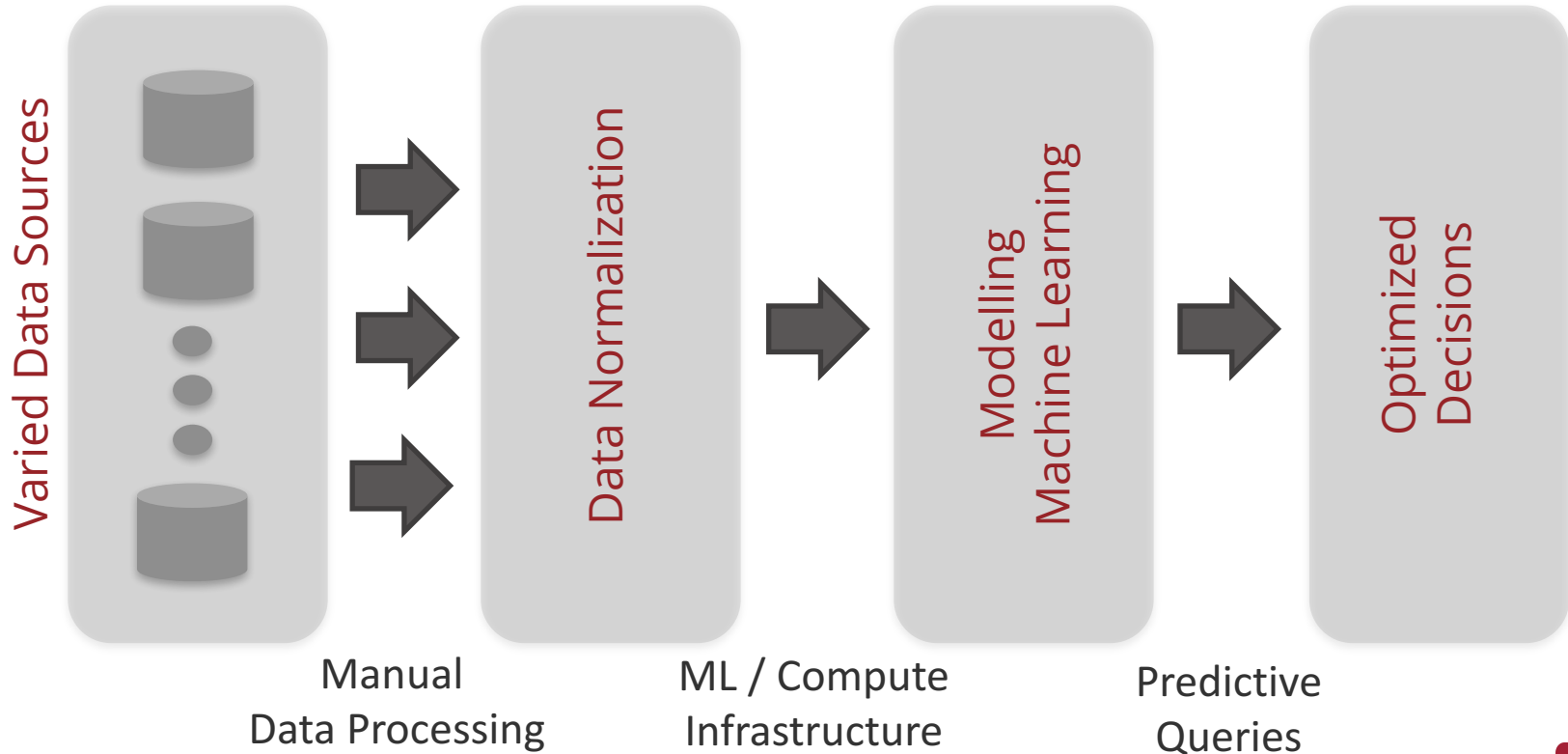
Prediction Problem Set

- Retail
 - Predict the demand for new products next quarter based on historical transactions, as well as product attributes
 - Predict the demand of products at stores that never carried them
 - Predict the daily demand of products at a store based on historical transactions
- Maritime
 - Predict likely paths for commercial shipping vessels based on AIS data
 - Predict the likely flag of a ship given its path
- Geopolitics
 - Predict future relations between countries along various axes based on an NLP-generated dataset (e.g. GDELT)

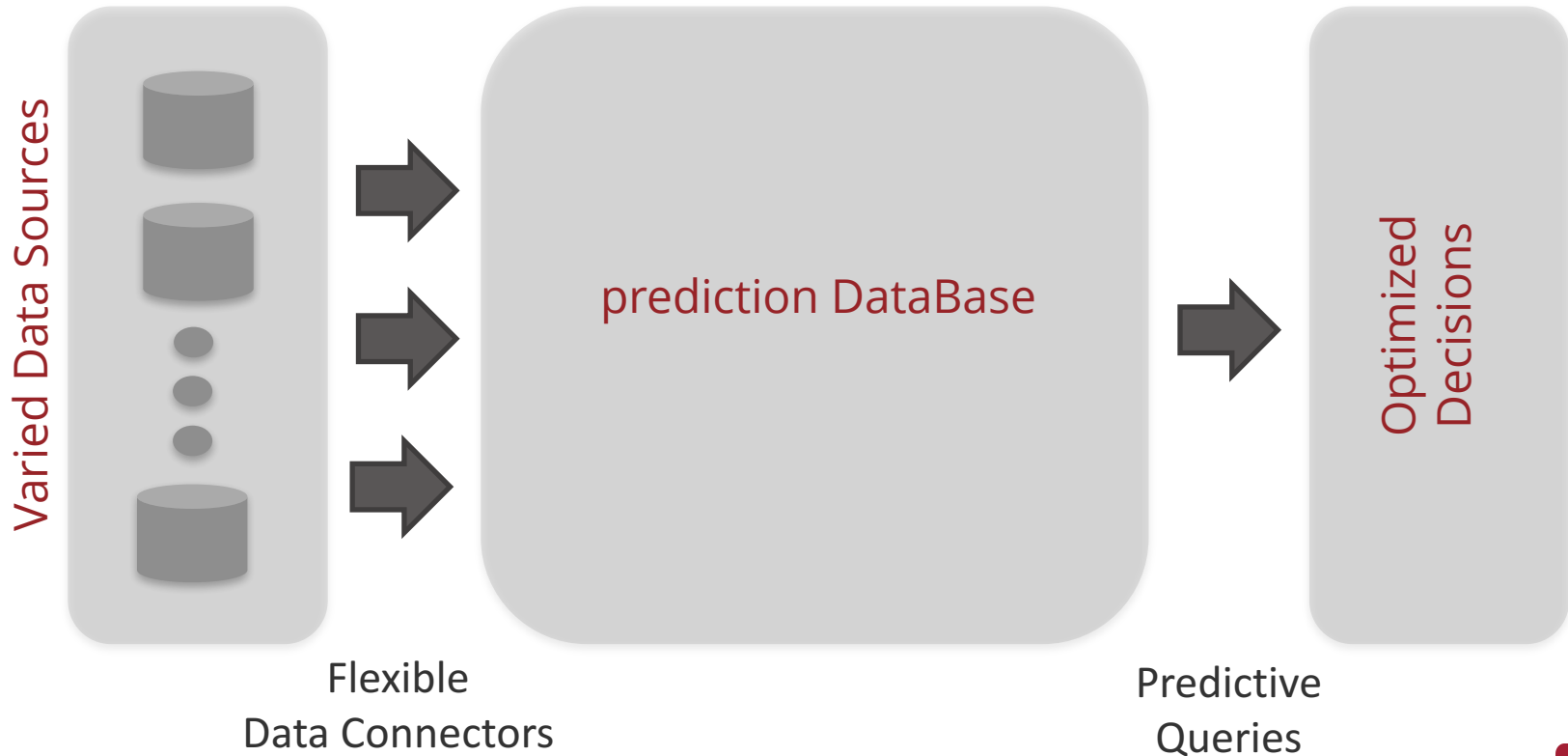
Approach

- High-level data-driven decision making
 - Step 0. Identify the problem
 - Step 1. Identify the available data
 - Step 2. Identify the “prediction” question (to solve)
 - Step 3. Make predictions to support decisions

Status Quo



pDB: prediction DataBase



Movie Recommendation System

- Building a Movie Recommendations System
 - Step 0. Recommend movies to users that s/he likes
 - Step 1. MovieLens dataset
 - Step 2. Predict what a user will rate a movie
 - Step 3. Build pipelines in pDB

MovieLens Data

- Basics:
 - 27K movies, 138K users
- Ratings: (userId, movieId, rating)
 - 20M ratings (0.53% density)
- Movies: (movieId, title, genre)
 - 27K movies
- Tags: (userId, movieId, tag)
 - 465K tags (free form text)

userId	movieId	rating	timestamp
1	2	3.5	1112486027
1	151	4	1094785734
91	3066	3	1111558027

MovieLens Data

- Basics:
 - 27K movies, 138K users

	userId	movieId	rating	timestamp
movieId	title		genres	
1	Toy Story (1995)		Adventure Animation Children Comedy Fantasy	
2	Jumanji (1995)		Adventure Children Fantasy	
3	Grumpier Old Men (1995)		Comedy Romance	
4	Waiting to Exhale (1995)		Comedy Drama Romance	
5	Father of the Bride Part II (1995)		Comedy	

- Tags: (userId, movieId, tag)
 - 465K tags (free form text)

MovieLens Data

- Basics:
 - 27K movies, 138K users

		userId	movieId	rating	timestamp
movieId	title	genres			
1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy			
2	Jumanji (1995)	userId	movieId	tag	timestamp
		18	4141	Mark Waters	1240597180
3	Grumpier Old Men (1995)	65	208	dark hero	1368150078
4	Waiting to Exhale (1995)	65	353	dark hero	1368150079
5	Father of the Bride Part II (1995)	65	521	noir thriller	1368149983
		65	592	dark hero	1368150078
		65	668	bollywood	1368149876
		65	898	screwball comedy	1368150160

- Tags: (userId, movieId, tag, timestamp)
 - 465K tags (free form text)

pDB Model of the World

pDB Language

(operation, (id1, id2)) : value

Prediction

(operation, (id1, id2)): ?

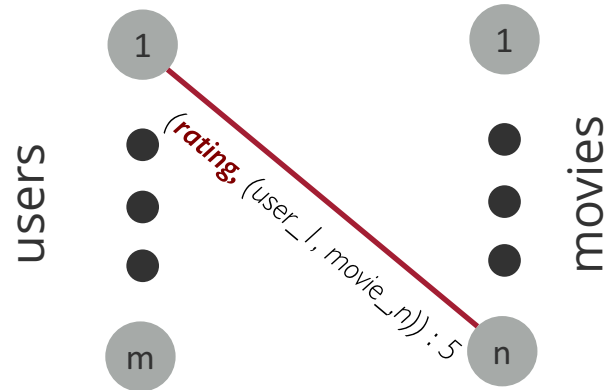
pDB Model of the World

pDB Language

$(rating, (user_1, movie_n)) : 5$

Prediction

$(rating, (user_1, movie_2)) : ?$



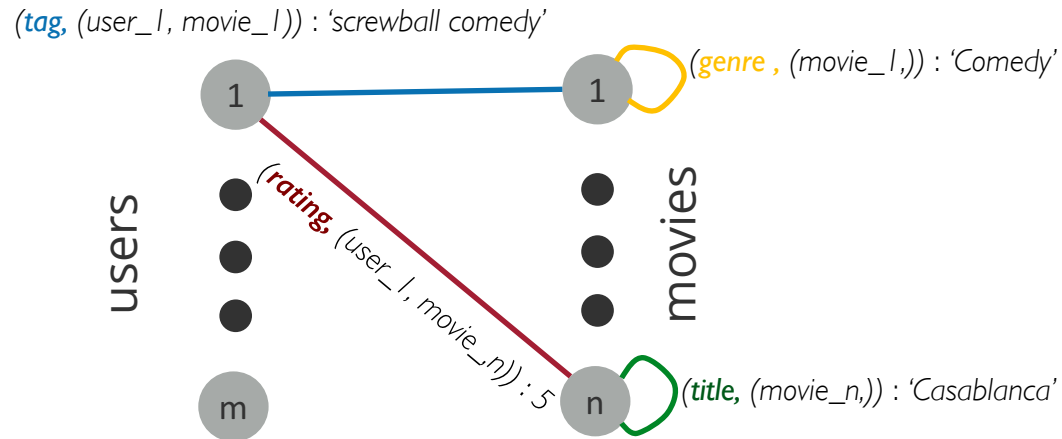
pDB Model of the World

pDB Language

$(rating, (user_1, movie_n)) : 5$

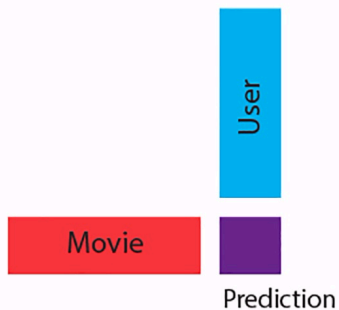
Prediction

$(rating, (user_1, movie_2)) : ?$



Crossvalidation

Spark
MLlib
Matrix Factorization

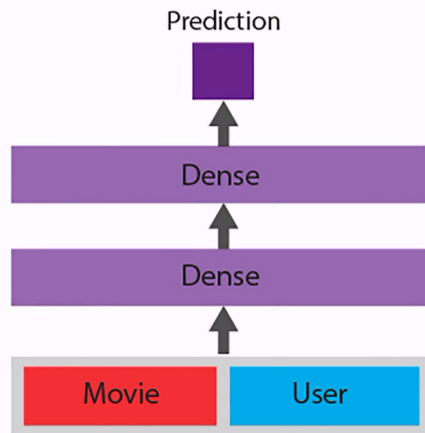


RMSE

0.860

APACHE
mxnetTM

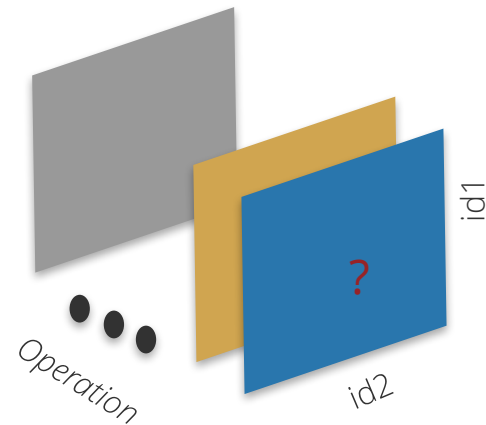
Deep Matrix Factorization



0.825

celect

prediction DataBase



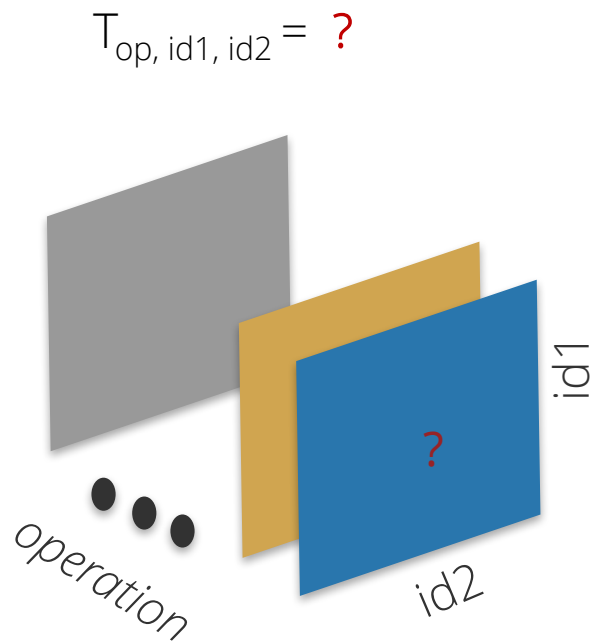
0.793

pDB Model of the World

- The language can simply express typical data science problems
 - Regression
 - Classification
 - Time Series (Interpolation, Forecasting and Multiple)
 - Matrix Completion
 - Tensor Completion

pDB Model of the World

- Prediction problem
 - 3-order tensor completion
 - with component being vector valued
- Non-parametric view
 - $T_{op, id1, id2} = f(x_{op}, y_{id1}, z_{id2})$
- Similarities through latent features
 - id1 vs id2 via y_{id1} vs z_{id2}



Solutions using pDB: Retail, Federal

Decision Making in Retail using pDB

Plan

Budget
Allocation

Am I investing in
the right areas?

Buy

Optimized
Assortments

Will I over-buy
or under-buy?

Allocate

Optimized
Allocation

Is the product in
the right stores?

Sell

Engaged
Customers

Will I maximize
full price sell-
through?

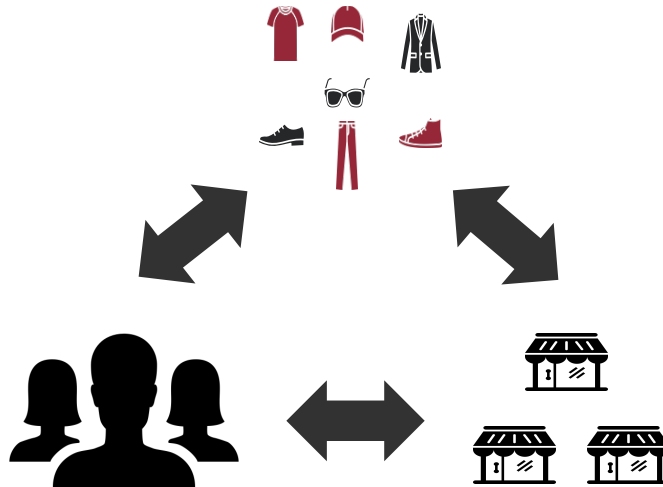
Liquidate

Optimized
Markdown

How much
margin will we
lose?

Decision Making in Retail using pDB

- Use pDB to *stitch* data across people, products, locations, time



Maritime Domain Awareness: Anomaly Detection

C AIS Application
Anomaly Detection Discover
Logout

Watchlist Items + ADD

- Antigua and Barbuda 5
- Canada 2
- China 2
- Hong Kong 2
- Taiwan 5
- North Korea 3
- France 5
- Germany 5
- India 3
- Iran 1
- Italy 1
- Japan 1
- Russia 2
- United Kingdom 2
- United States of America 1

Vessel movement over last 24 hours

Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

Alerts

WATCHLIST OVERALL

Confidence - 66% 4:41 PM

Cargo from China

The vessel seems to be going to a different destination than declared.

Confidence - 83% 4:41 PM

Cargo from Canada

The vessel seems to be going to a different destination than declared.

Confidence - 72% 4:41 PM

Cargo from Canada

The vessel seems to be more typical of vessels from another country.

Confidence - 70% 4:41 PM

Cargo from Antigua and Barbuda

The vessel seems to be taking longer (or shorter) than it typically does on this route.

Confidence - 78% 4:41 PM

Cargo from Antigua and Barbuda

The vessel seems to be more typical of vessels from another country.

Confidence - 87% 4:41 PM

Cargo from Antigua and Barbuda

The vessel seems to be taking longer (or shorter) than it typically does on this route.

Confidence - 75% 4:41 PM

Cargo from Antigua and Barbuda

Status	MMSI	Country	Ship Type	Distance Traveled	Anomaly Category	Anomalous Value	Expected Value	Action
		Filter Country				Filter Anomalous Vali		
	305654000	Antigua and Barbuda	Cargo - Hazard A	593.14	Country	Antigua and Barbuda	Germany	INVESTIGATE
	305977000	Antigua and Barbuda	Cargo	21.72	Country	Antigua and Barbuda	Germany	INVESTIGATE
	304726000	Antigua and Barbuda	Cargo	25.80	Country	Antigua and Barbuda	Germany	INVESTIGATE

Maritime Domain Awareness: Predicted Destination

AIS Application Anomaly Detection Discover Logout

Watchlist Items + AD

- Antigua and Barbuda
- Canada
- China
- Hong Kong
- Taiwan
- North Korea
- France
- Germany
- India
- Iran
- Italy
- Japan
- Russia
- United Kingdom
- United States of America

Investigate

- Current Path
- Projected Path
- Typical Paths

MMSI	Ship Type	Time Period	Anomaly
305654000	Cargo - Hazard A	Last 24 hours	The vessel seems to be more typical of vessels from another country.

Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

CLOSE

WATCHLIST OVERALL

- Confidence - 80% 4:42 PM
 Cargo - Hazard B from Taiwan
The vessel seems to be more typical of vessels from another country.
- Confidence - 88% 4:42 PM
 Cargo from Taiwan
The vessel seems to be more typical of vessels from another country.
- Confidence - 87% 4:42 PM
 Cargo from Taiwan
The vessel seems to be going to a different destination than declared.
- Confidence - 65% 4:42 PM
 Cargo from Hong Kong
The vessel seems to be more typical of vessels from another country.
- Confidence - 78% 4:42 PM
 Cargo from Hong Kong
The vessel seems to be taking longer (or shorter) than it typically does on this route.
- Confidence - 67% 4:41 PM
 Cargo from China
The vessel seems to be more typical of vessels from another country.
- Confidence - 66% 4:41 PM
 Cargo from China

Maritime Domain Awareness: Anomalous Behavior

Watchlist Items

- Antigua and Barbuda
- Canada
- China
- Hong Kong
- Taiwan
- North Korea
- France
- Germany
- India
- Iran
- Italy
- Japan
- Russia
- United Kingdom
- United States of America

Investigate

Current Path
 Projected Path
 Typical Paths

Country	Vessel Type
Germany	Cargo - Hazard A
Germany	Cargo
Germany	Cargo
Germany	Cargo

MMSI | **Ship Type** | **Time Period** | **Anomaly**

| 305654000 | Cargo - Hazard A | Last 24 hours | The vessel seems to be more typical of vessels from another country. |

WATCHLIST | **OVERALL**

- Confidence - 82% 4:43 PM
Other from France
The vessel seems to be more typical of vessels from another country.
- Confidence - 70% 4:42 PM
Cargo from North Korea
The vessel seems to be more typical of vessels from another country.
- Confidence - 80% 4:42 PM
Cargo from North Korea
The vessel seems to be more typical of vessels from another country.
- Confidence - 82% 4:42 PM
Cargo from North Korea
The vessel seems to be going to a different destination than declared.
- Confidence - 89% 4:42 PM
Cargo from Taiwan
The vessel seems to be going to a different destination than declared.
- Confidence - 69% 4:42 PM
Cargo from Taiwan
The vessel seems to be taking longer (or shorter) than it typically does on this route.
- Confidence - 80% 4:42 PM
Cargo - Hazard B from Taiwan

Enterprise Grade Architecture

- Micro-services architecture leveraging gRPC and protobuf
- Stateless services with end-to-end fault recovery
- Datastores, models and Spark clusters managed seamlessly
- pDB deployed as Docker containers via Kubernetes

pDB is versatile

- High bandwidth connectors integrate in existing environment
 - You *do not* need to *upload* your data
- pDB language allows for solving *any* predictive problem
 - Using non-parametric solution, at scale
- Unstructured data is fully utilized
 - In-built “feature extractors” for image, text, geo
- Predictive models are built using ALL the available data
 - Overcomes data sparsity challenge using non-parametric methods
- Provenance of predictions explains answer
 - System is *not* a black-box