Structural Clustering

Analysts' Step 1
Approach to New Data

Dylan Hutchison <dhutchis@stevens.edu> and the Microsoft Research Tabular Team

Vision: Enable data analysts to run Probabilistic Machine Learning in Excel

Background: Tabular – an Excel-based DSL data analysts use to write generative models for probabilistic inference.

Tabular: A Schema-Driven Probabilistic Programming Language
A. Gordon, T. Graepel, N. Rolland, C. Russo et al
POPL 2014

Problem: *High entry barrier* for data analysts to write generative models.
Need expert domain knowledge, heavy time investment.

Solution Idea: Suggest a *default model* based on a dataset's structure and statistics.
Extends Singh and Graepel's InfernoDB in Tabular.

Automated Probabilistic Modelling for Relational Data
S. Singh, T. Graepel
CIKM 2013

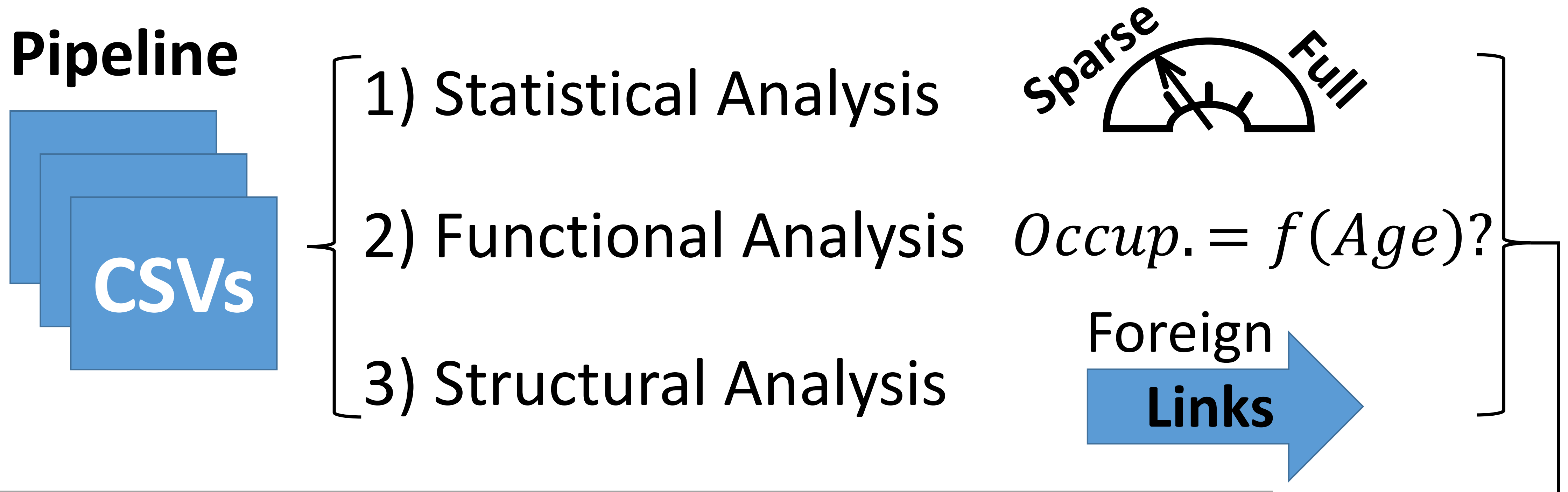
Evaluation: Compare model accuracy and fit with InfernoDB & other models.
Demonstrate value to data analysts via case studies.

MovieLens Example

<http://grouplens.org/datasets/movielens>

Application: Predicting users' ratings,
Suggesting movies to users

Pipeline



Users

User	Age	Gender	Occupation
1	24	M	technician
2	53	F	lawyer
3	23	M	writer
4	24	M	technician

Users

k	upto(Ku)	latent	CDiscrete(N=Ku)
Occupation	upto(Nocc)	output	CDiscrete(N=Nocc)[k]
Age	real	output	CPoisson(alpha=5.0,beta=5.0)[k]
Gender	bool	output	CBernoulli()[k]

Ratings

User	Movie	Rating
196	242	3
186	302	3
22	377	1
244	51	2

Movies

k	upto(Km)	latent	CDiscrete(N=Km)
Action	bool	output	CBernoulli()[k]
Adventure	bool	output	CBernoulli()[k]
Animation	bool	output	CBernoulli()[k]

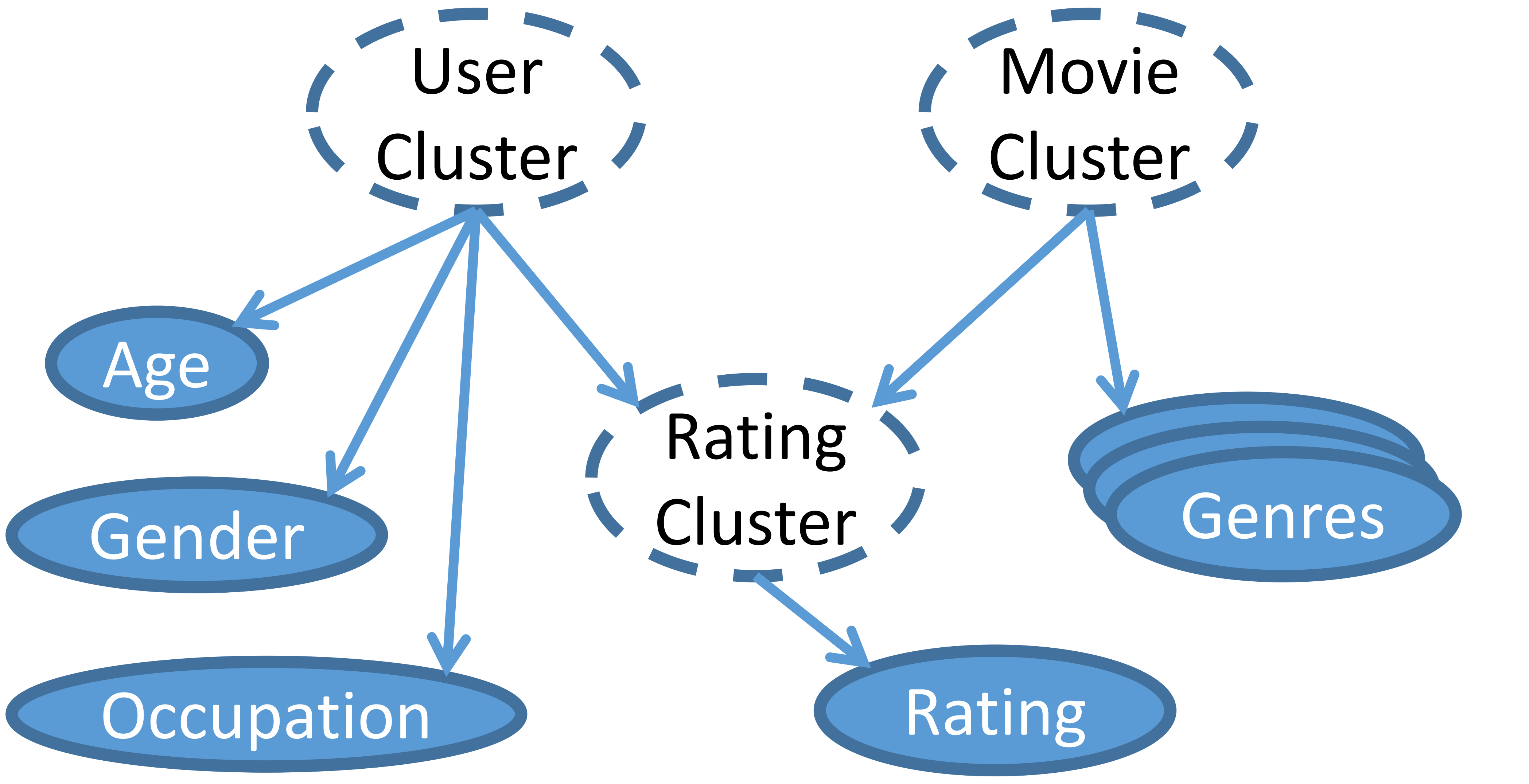
Generative Cluster Model

Ratings

User	link(T_User)	input	
Movie	link(T_Movie)	input	
k	upto(Kr)	latent	CDiscrete(N=Kr)[User.k][Movie.k]
Rating	upto(5)	output	CBinomial(N=5)[k]

Movies

Movie	Title	Action	Adventure	Animation
		Genres		
1	Toy Story (1995)	0	0	1
2	GoldenEye (1995)	1	1	0
3	Four Rooms (1995)	0	0	0
4	Get Shorty (1995)	1	0	0



User Cluster
Movie Cluster
Age
Gender
Occupation
Rating Cluster
Genres
Rating