Basic Concepts in Big Data

ChengXiang (“Cheng”) Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
http://www.cs.uiuc.edu/homes/czhai
czhai@illinois.edu

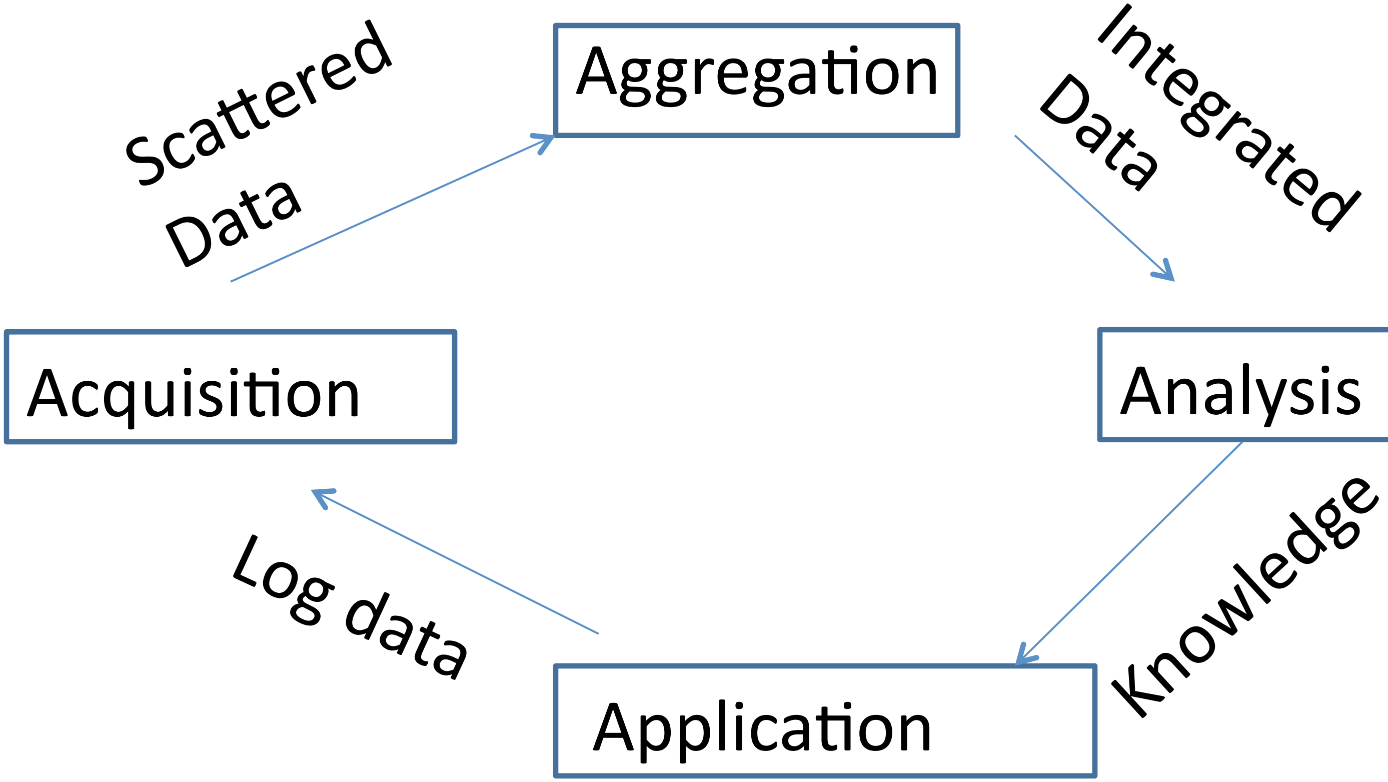
What is “big data”?

- "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization” (Gartner 2012)
- Complicated (intelligent) analysis of data may make a small data “appear” to be “big”
- Bottom line: Any data that exceeds our current capability of processing can be regarded as “big”

Why is “big data” a “big deal”?

- Government
  - Obama administration announced “big data” initiative
  - Many different big data programs launched
- Private Sector
  - Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data
  - Facebook handles 40 billion photos from its user base.
  - Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide
- Science
  - Large Synoptic Survey Telescope will generate 140 Terabyte of data every 5 days.
  - Biomedical computation like decoding human Genome & personalized medicine
  - Social science revolution
  - -…

Lifecycle of Data: 4 “A”s



Aggregation

Integrated Data

Analysis

Knowledge

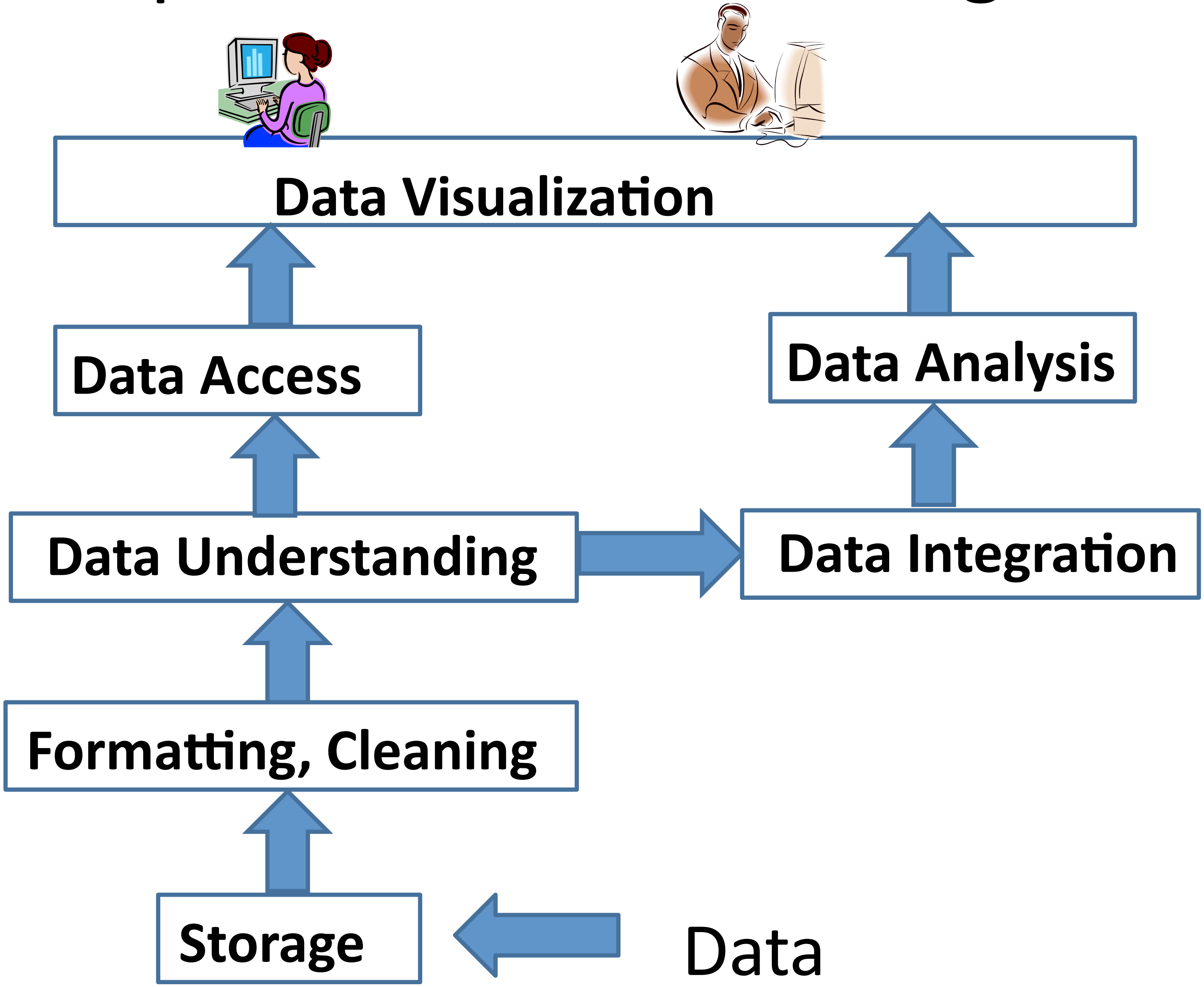
Application

Log data

Acquisition

Scattered Data

Computational View of Big Data



Data Visualization

Data Access

Data Analysis

Data Understanding

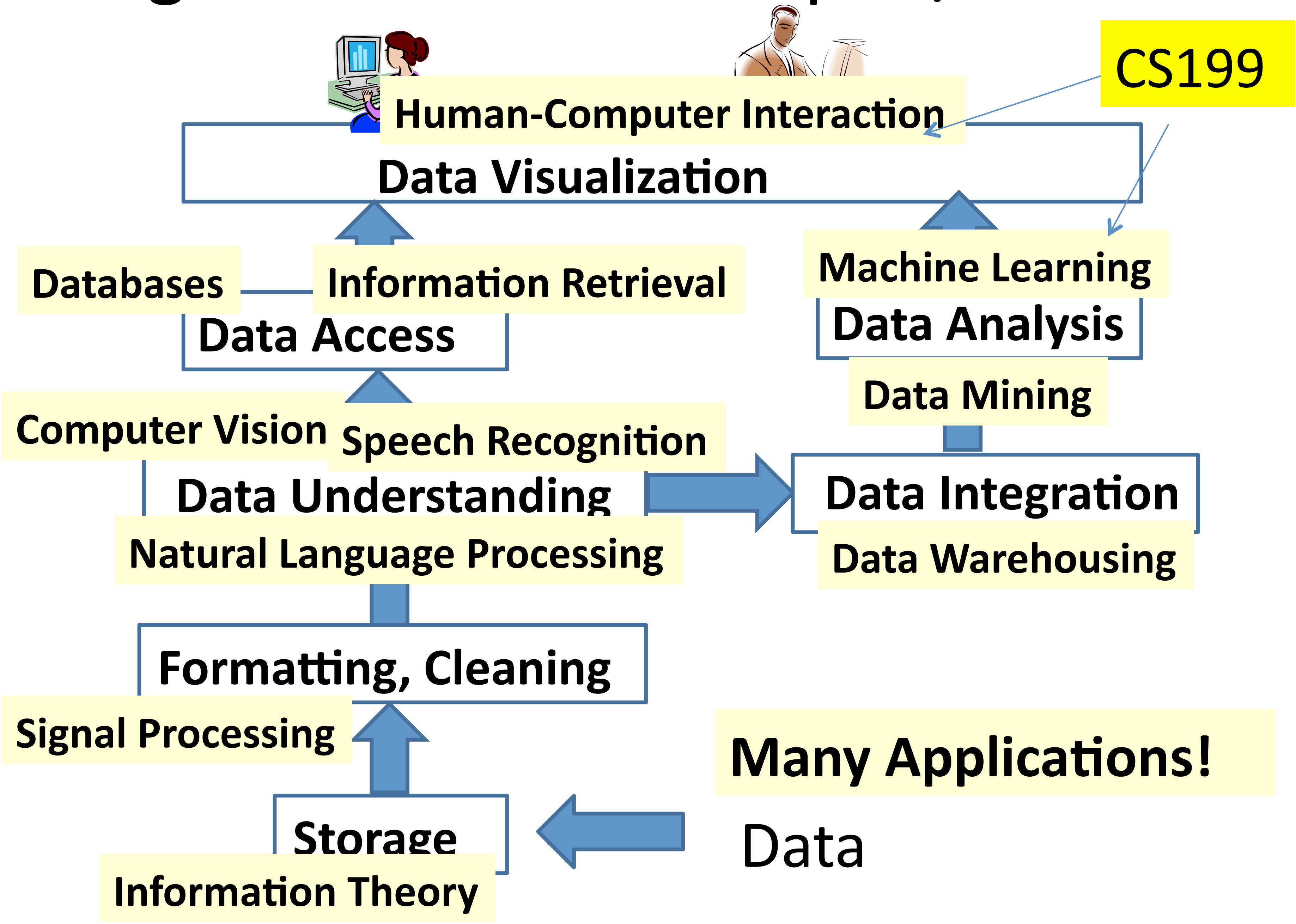
Data Integration

Formatting, Cleaning

Storage

Data

Big Data & Related Topics/Courses



CS199

Human-Computer Interaction

Data Visualization

Databases

Information Retrieval

Data Access

Machine Learning

Data Analysis

Data Mining

Computer Vision

Speech Recognition

Data Understanding

Natural Language Processing

Data Integration

Data Warehousing

Formatting, Cleaning

Signal Processing

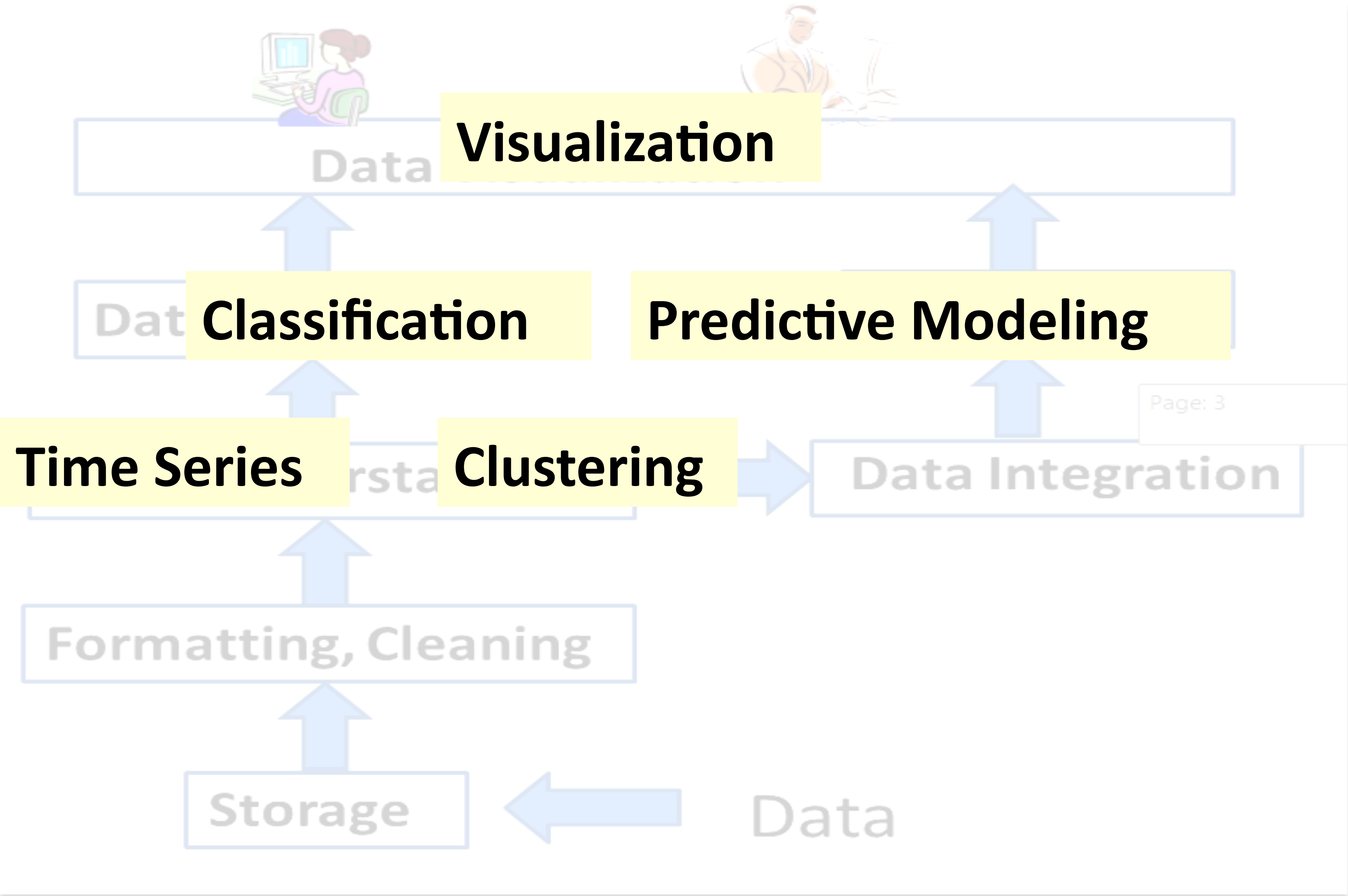
Many Applications!

Storage

Information Theory

Data

Some Data Analysis Techniques



Visualization

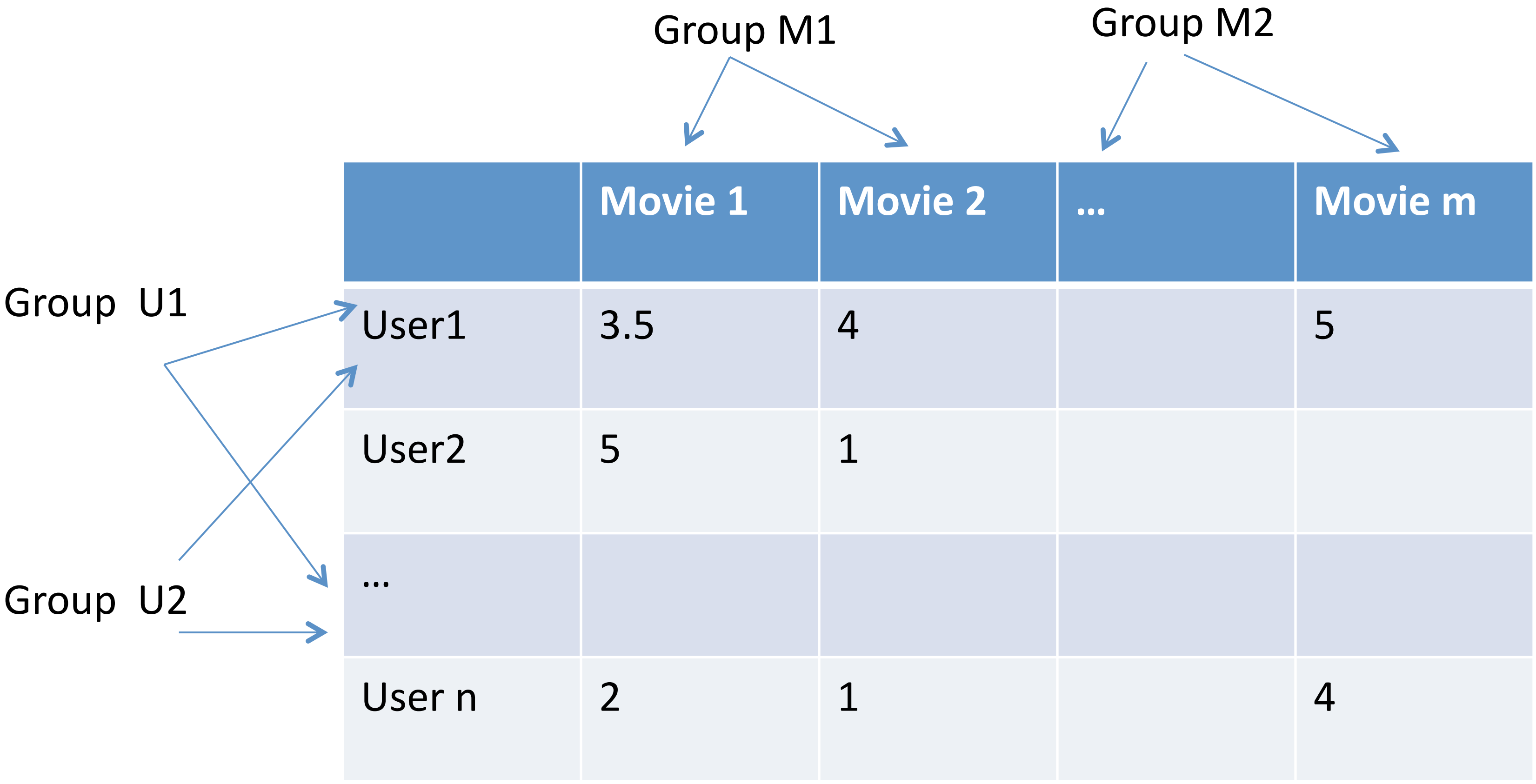
Classification

Predictive Modeling

Time Series

Clustering

Example of Analysis: Clustering & Latent Factor Analysis



Group M1

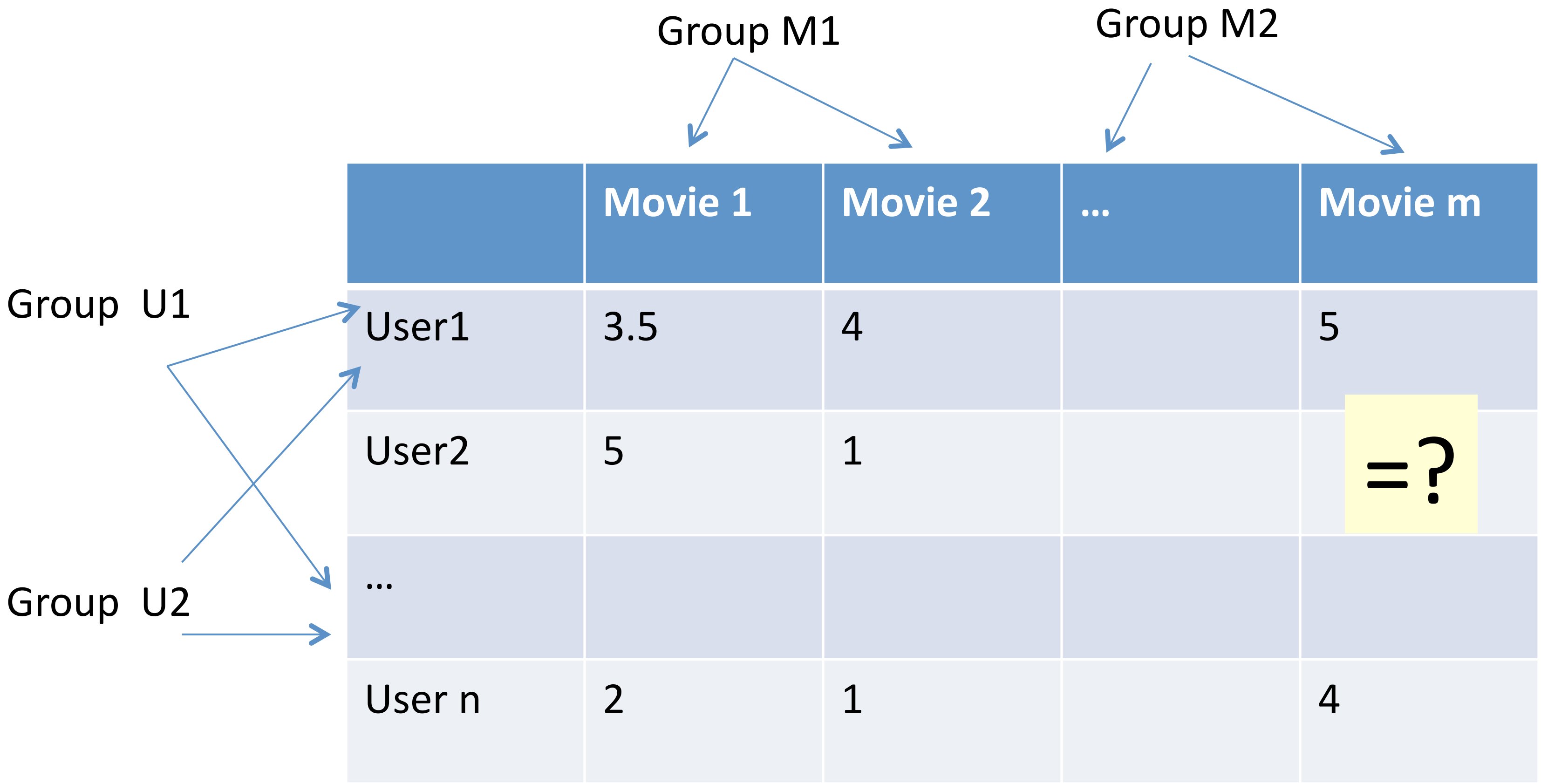
Group M2

Group U1

Group U2

	Movie 1	Movie 2	…	Movie m
User1	3.5	4		5
User2	5	1		
…				
User n	2	1		4

Example of Analysis: Predictive Modeling



Group M1

Group M2

Group U1

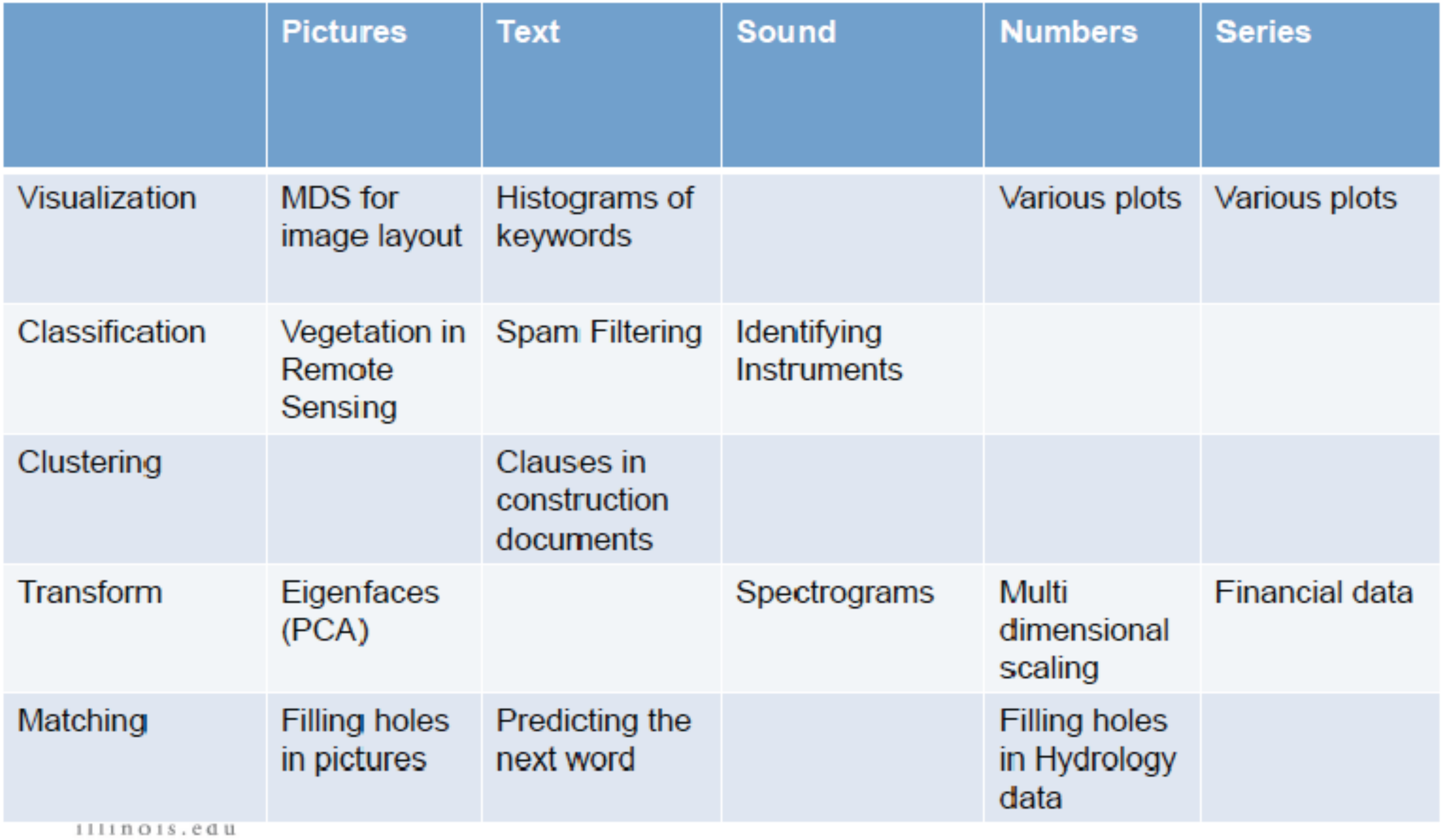
Group U2

	Movie 1	**Movie 2**	**…**	**Movie m**
User1	3.5	4		5
User2	5	1		=?
…				
User n	2	1		4

Does user2 like movie m? **(Binary) Classification**

What rating is user2 likely going to give movie m? **Regression**

Some topics we'll cover



	Pictures	Text	Sound	Numbers	Series
Visualization	MDS for image layout	Histograms of keywords		Various plots	Various plots
Classification	Vegetation in Remote Sensing	Spam Filtering	Identifying Instruments		
Clustering		Clauses in construction documents			
Transform	Eigenfaces (PCA)		Spectrograms	Multi dimensional scaling	Financial data
Matching	Filling holes in pictures	Predicting the next word		Filling holes in Hydrology data	

illinois.edu