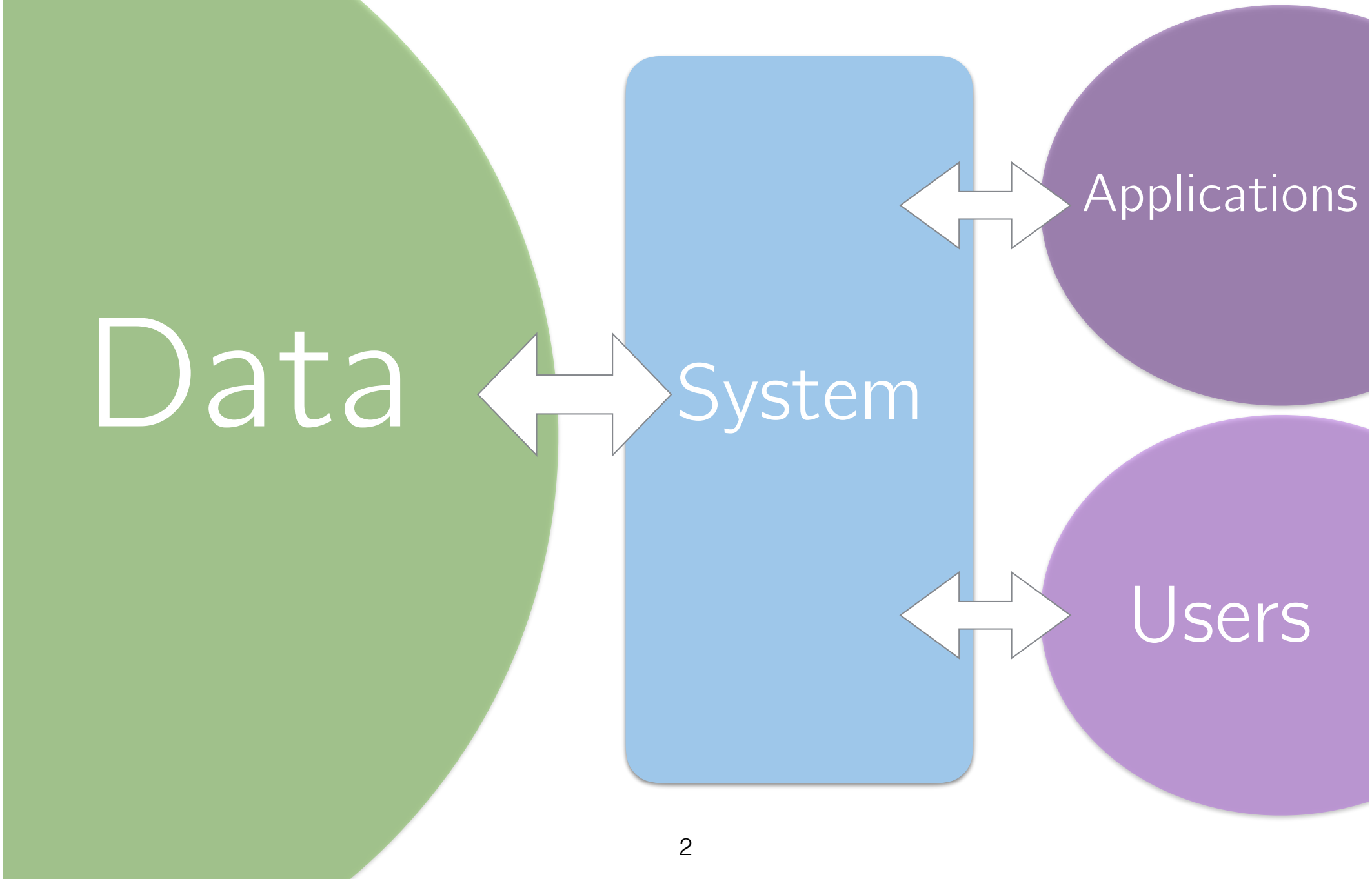
RWTH AACHEN UNIVERSITY

Big Data from a Database Theory Perspective

Martin Grohe

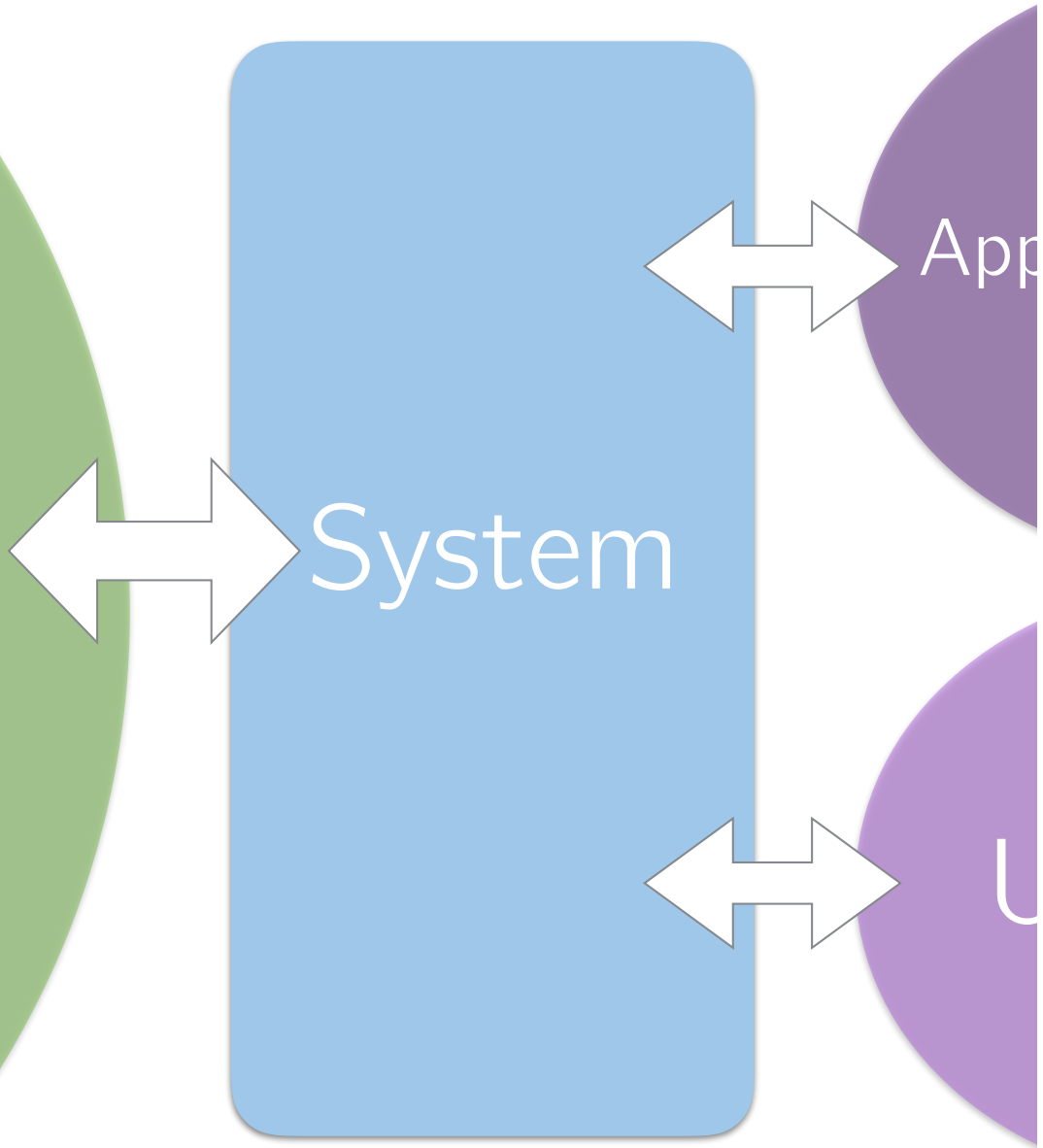
Lehrstuhl Informatik 7 - Logic and the Theory of Discrete Systems

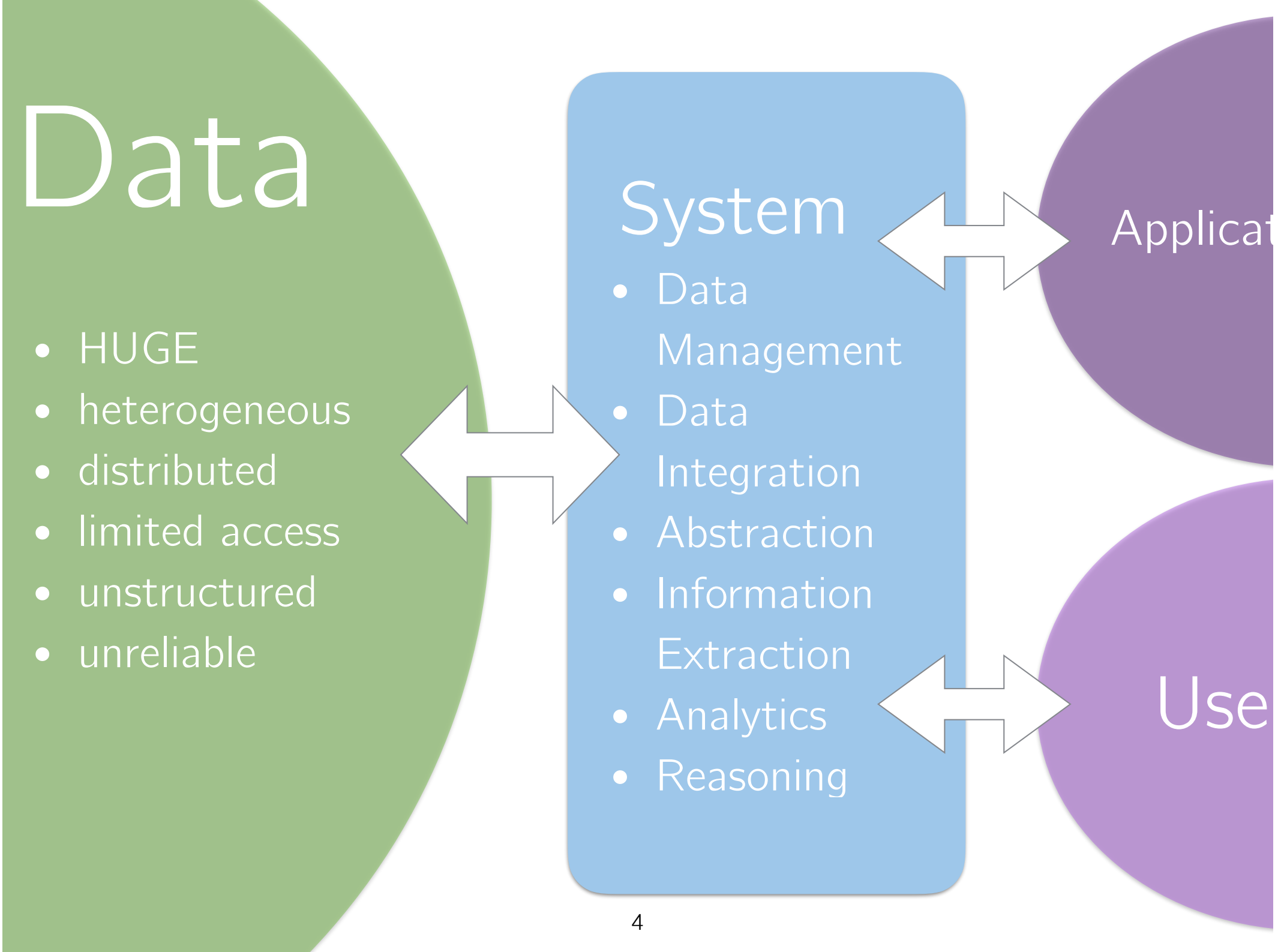
A CS View on Data Science

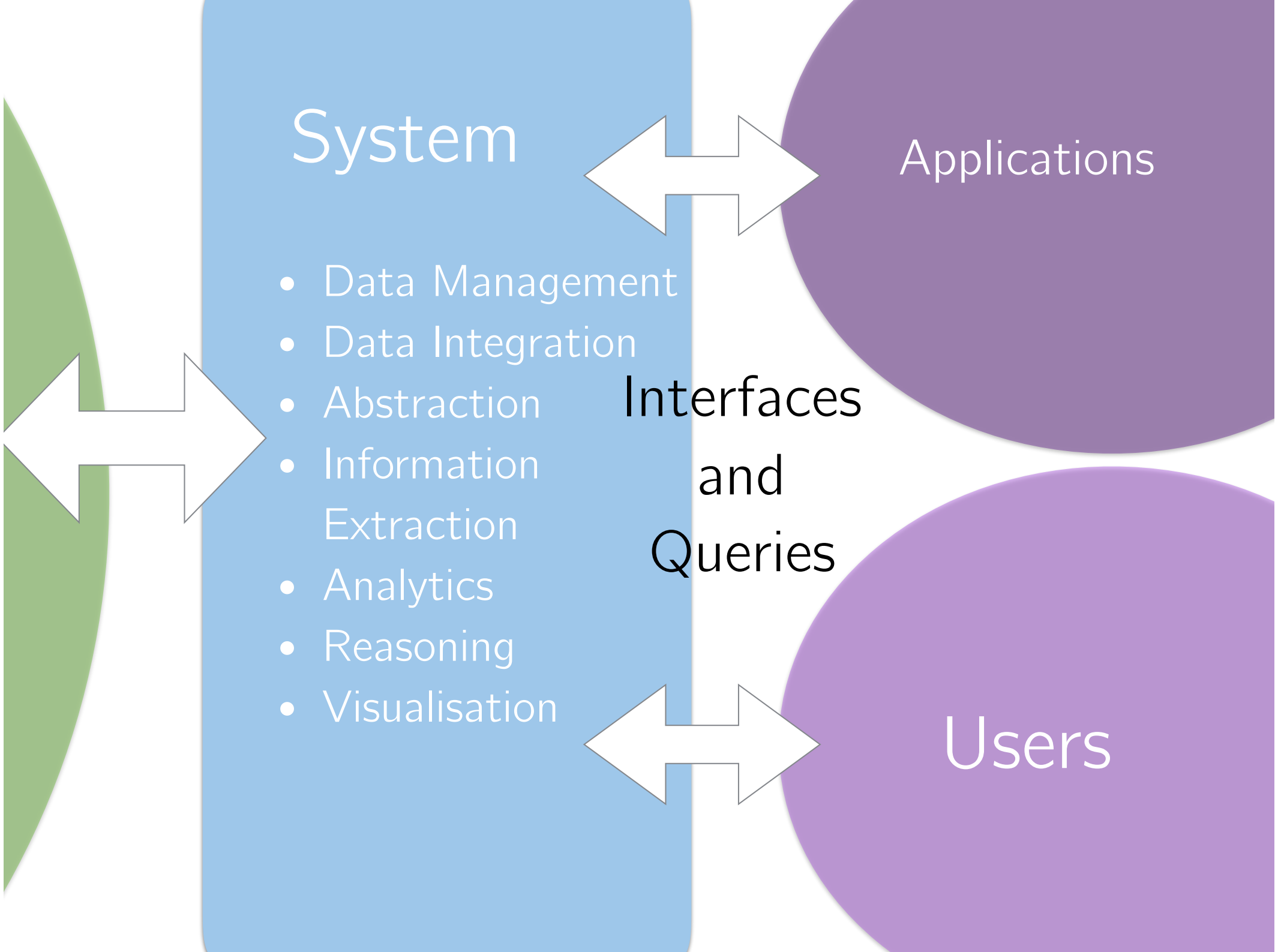


Data

- HUGE
- heterogeneous
- distributed
- limited access
- unstructured
- unreliable




Data
HUGE
heterogeneous
distributed
limited access
unstructured
unreliable
System
Data Management
Data Integration
Abstraction
Information Extraction
Analytics
Reasoning
Applicat
Use


System
Data Management
Data Integration
Abstraction
Information Extraction
Analytics
Reasoning
Visualisation
Interfaces and Queries
Applications
Users

Streaming Model

Data is not stored in main memory, but only read once as a data stream and immediately discarded. Goal is to design algorithms whose memory consumption is low compared to size of whole input.

- Efficient algorithms for many data analysis tasks
- Also strong impossibility results (lower bounds)

Massively Parallel Architectures

Map reduce is Google's computation framework for large scale data analysis tasks. Computation is carried out by several rounds of distributed local computations interleaved with communication rounds for data exchange between sites. (Hadoop similar open access framework.)

Again, both algorithmic and lower bound results are known.

Read-Write Streams

Idea

Allow data stream to be stored in external memory (say, on one or several disks), but limit random access to data.

Theorem (G., Schweikardt)

In the Read-Write Stream Model, Sorting data requires at least logarithmic internal memory space.

Theoretical Aspects (II)
Symmetry and Regularity

Efficient Reasoning

Reasoning and inference problems are very hard algorithmic problems. To be able to perform them, we need to exploit the structure of the data.

Two (somewhat complementary) approaches are:

- decompose data into more or less independent parts
- exploit symmetries and regularities in the data.

Lifted Inference

- technique for efficient reasoning in probabilistic models (graphical models)
- often, these models are large, but have many regularities
- Idea: use regularities to create smaller models and then „lift“ original inference task

Dimension Reduction for Linear Programs

Observation (Kersting, Mladenov)

Standard inference techniques in graphical models (loopy believe propagation) can be reduced to linear programming.

Colour Refinement

Simple and very efficient algorithm originally designed to classify vertices of a graph by similarity.

Theorem (G., Kersting, Mladenov, Selman)

Adaptation of color refinement to matrices can be used to efficiently transform linear program to an equivalent smaller linear program (obtained by averaging over similarity classes of rows and columns)

Theoretical Aspects (III)
Declarative Framework and Query Languages

Database Systems

Modern database systems provide declarative framework which allows users to specify what they want without specifying how it can be obtained.

Example query: Is there a student who failed all exams?

Data Mining and Advanced Analytics

- no easy access, information only accessible for expert users
- access low-level operational: write a program to gather complex information
- design of systems that allow easier access is an important research goal
- What would be good query languages?

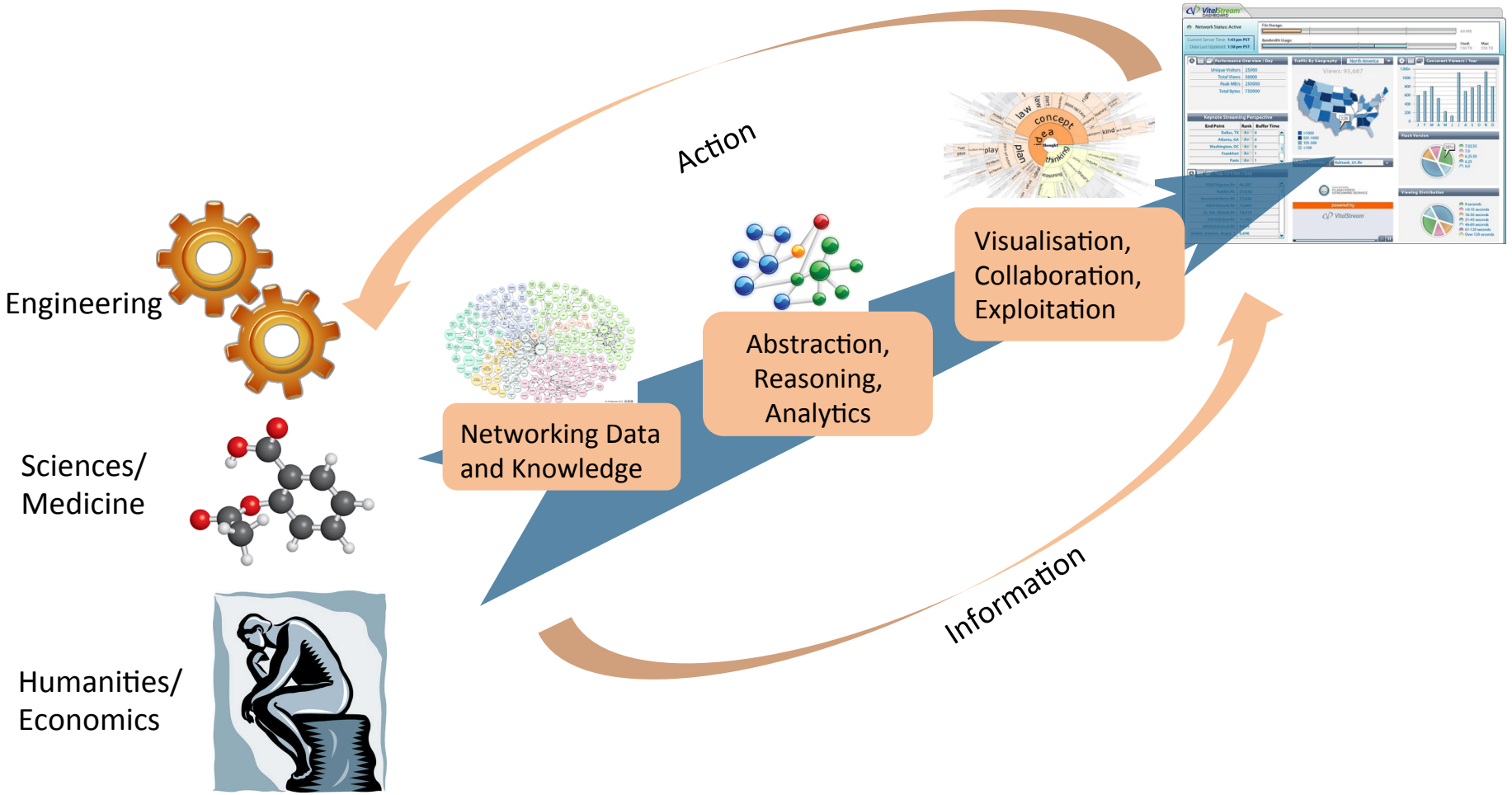
Example query: Are all clusters (in a social network) dominated by few people?

Learning Definable Concepts

Theorem (G., Turan)

Properties of certain graphs (such trees or planar graph) that can be specified in certain languages (monadic second-order logic or first-order logic) admit efficient learning algorithms.

Vision: A Knowledge Pipeline



Foundations of Data Science initiative in CS department
(Decker, Grohe, Jarke, Kobbelt, Leibe et al.)