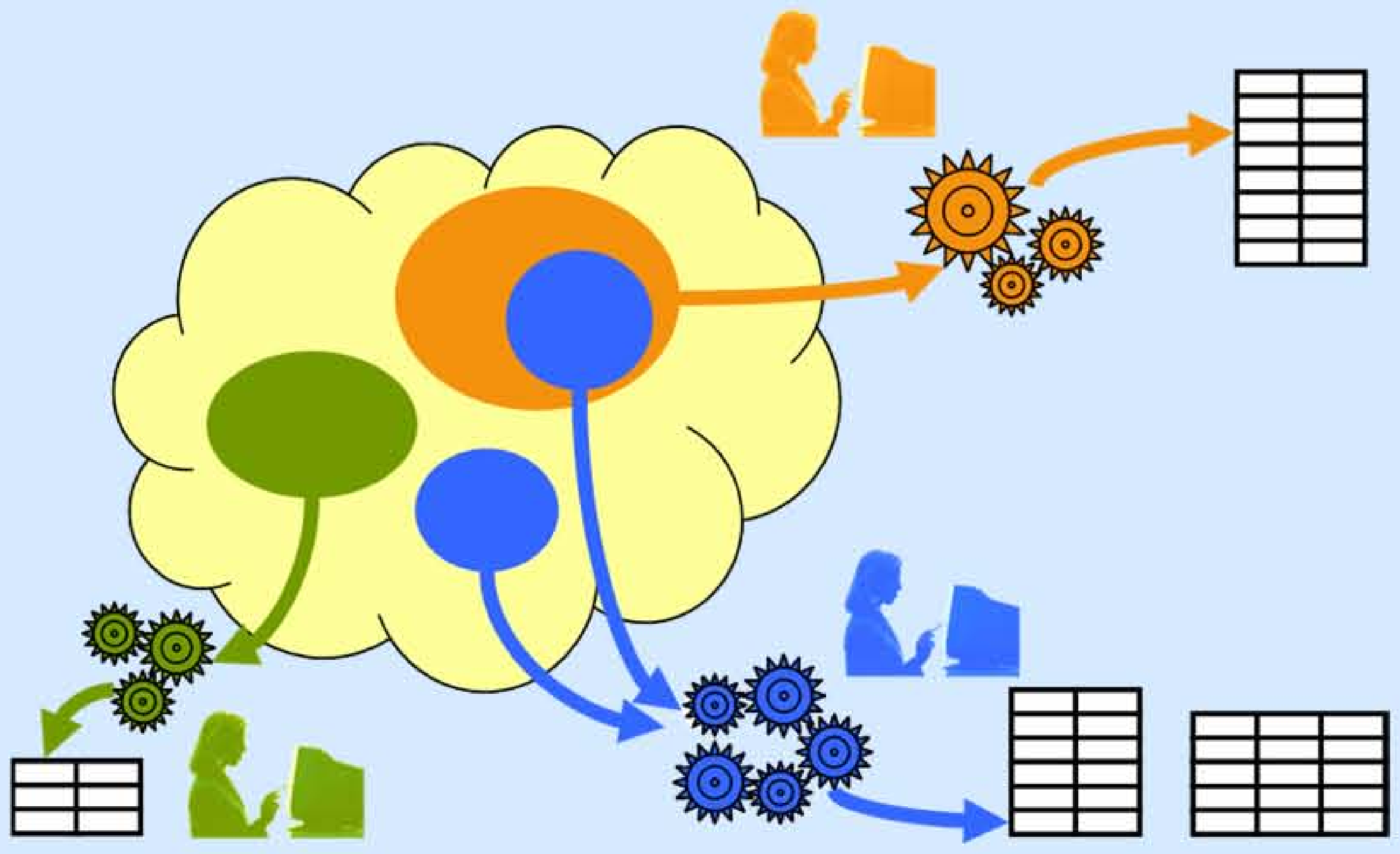
Collaborative Creation And Management Of Large High-Quality Data Sets

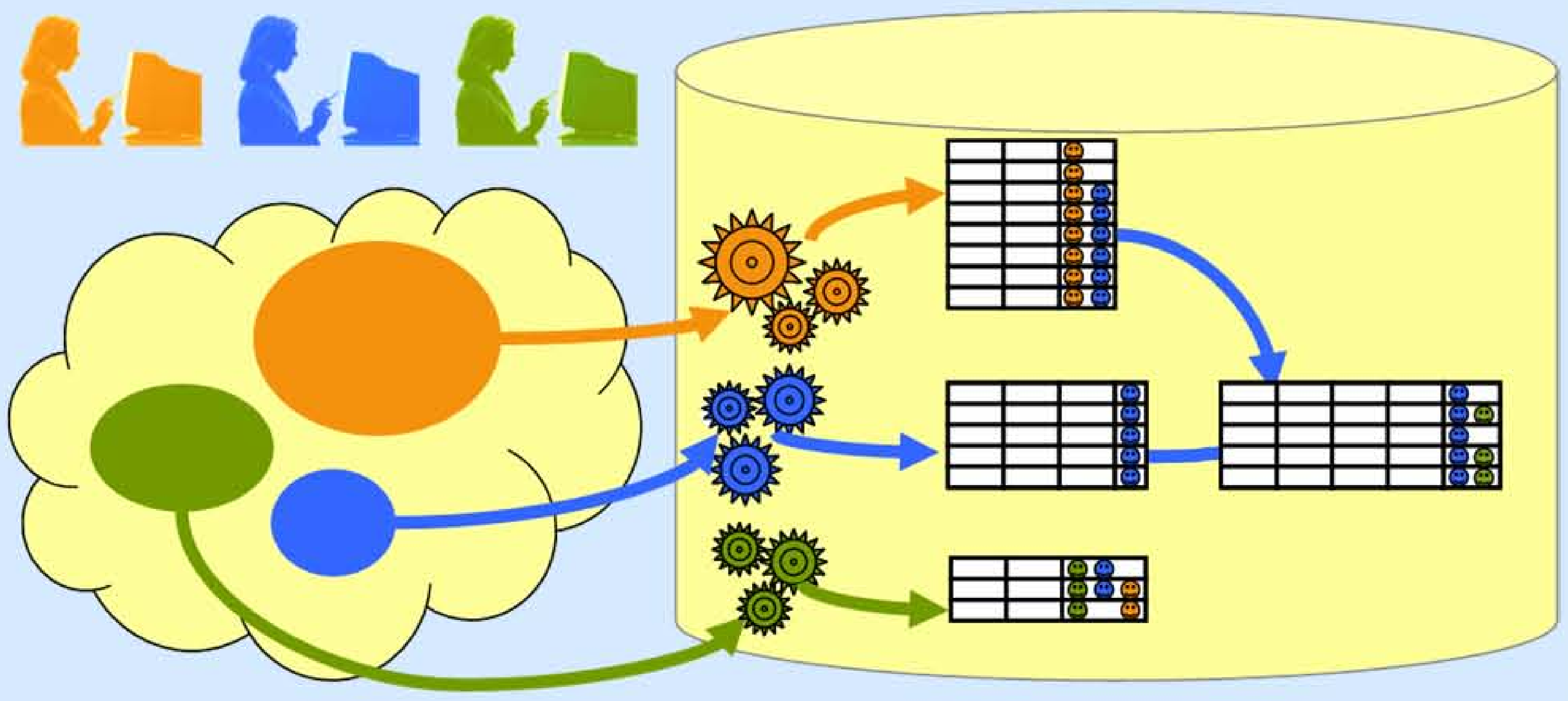
Felix Weigel, Biswanath Panda, Mirek Riedewald
Johannes Gehrke, Christoph Koch
Department of Computer Science, Cornell University, Ithaca, NY

Motivation



- Users with similar interests look for similar data
- Collaboration benefits a community
  - Less work per user
  - More knowledge for all

Collaborative Environment



- Availability is not enough

💡 **Reusability** is the key

Making Data Reusable

- Find relevant data
- Exploit compositionality
- Clean erroneous data
- Integrate views of different users

Use Case: Social Network Analysis In The Internet Archive

- The web has many social networks
  - Book reviewers at Amazon.com
  - Blogging communities
  - Citation network in DBLP
- Rich data source for large scale scientific analysis
  - Herd behavior
  - Network Dynamics
- Scientists use similar data sets

New Solutions To Old Problems

- Data extraction from unstructured, heterogeneous sources
- Data integration
- Data cleaning

💡 **Leverage user opinions in a collaborative environment**

Our Focus Areas

- Data model representing user opinions
- Resolving disagreement
- Finding database instances that reflect community knowledge
- System for collaboration on a web archive
  - Extraction tools for non-technical users
  - Search methods for finding useful information
  - Storing versions of data from different users