Finding Structure in Big Data

Ankur Moitra, IAS

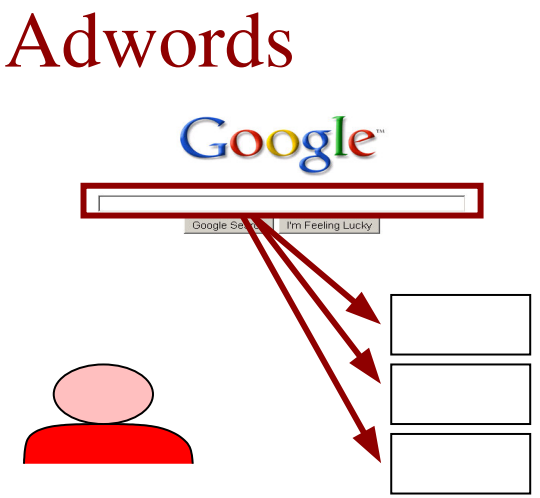
May 4, 2012

Recommendation Systems in Action

Some everyday examples of recommendations:

Recommendation Systems in Action

Some everyday examples of recommendations:



Recommendation Systems in Action

Some everyday examples of recommendations:

Recommendation Systems in Action

Some everyday examples of recommendations:

Also, Netflix Prize (catalyst for new research)

Recommendation Systems in Action

Some everyday examples of recommendations:

Also, Netflix Prize (catalyst for new research)

How good are these recommendations?

Recommendation Systems in Action

Some everyday examples of recommendations:

Also, Netflix Prize (catalyst for new research)

How good are these recommendations? What does **good** mean?

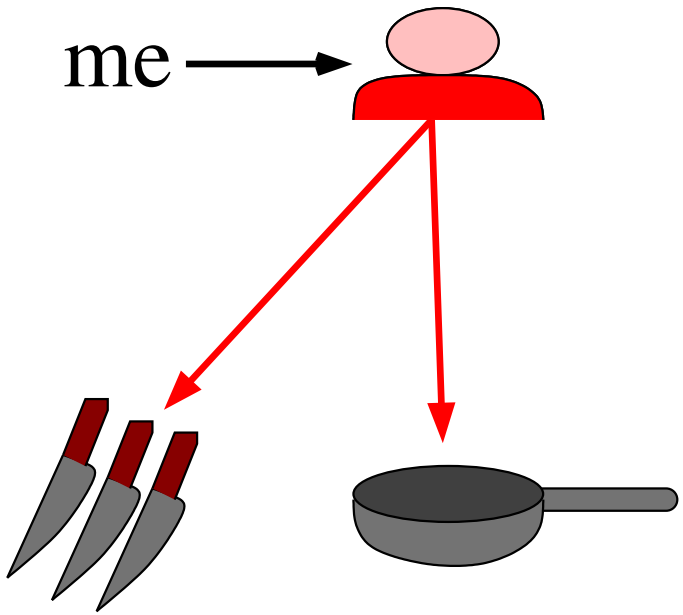
Recommendation Systems in Action

Some everyday examples of recommendations:

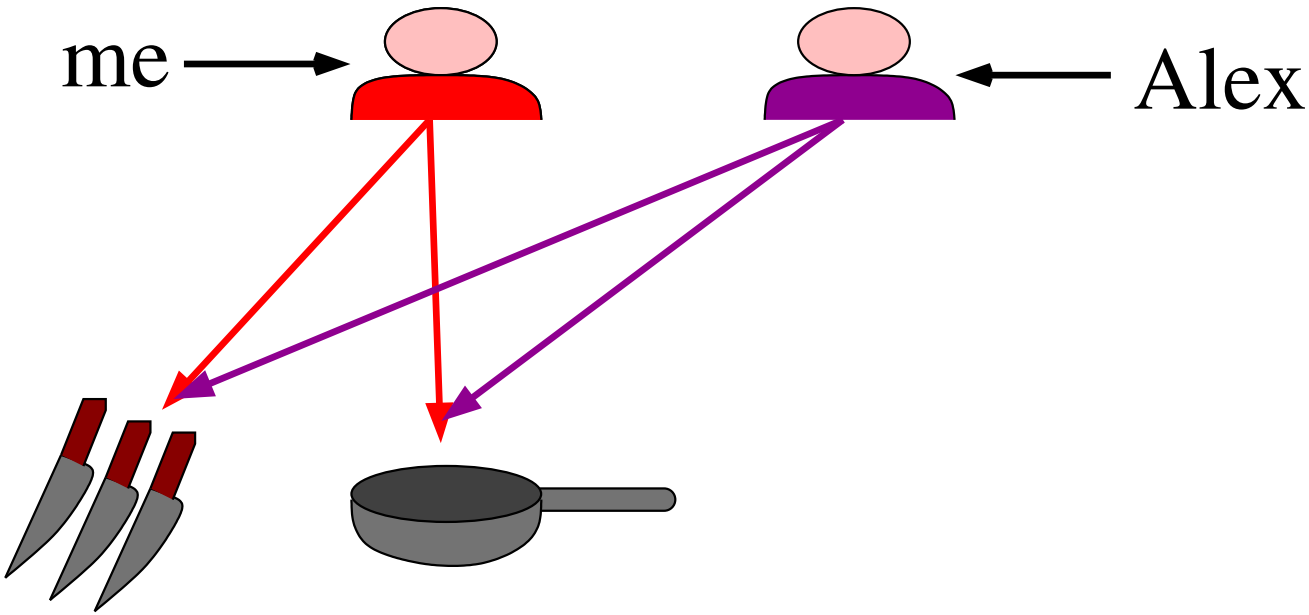
Also, Netflix Prize (catalyst for new research)

How good are these recommendations? What does **good** mean? And how do they do it?

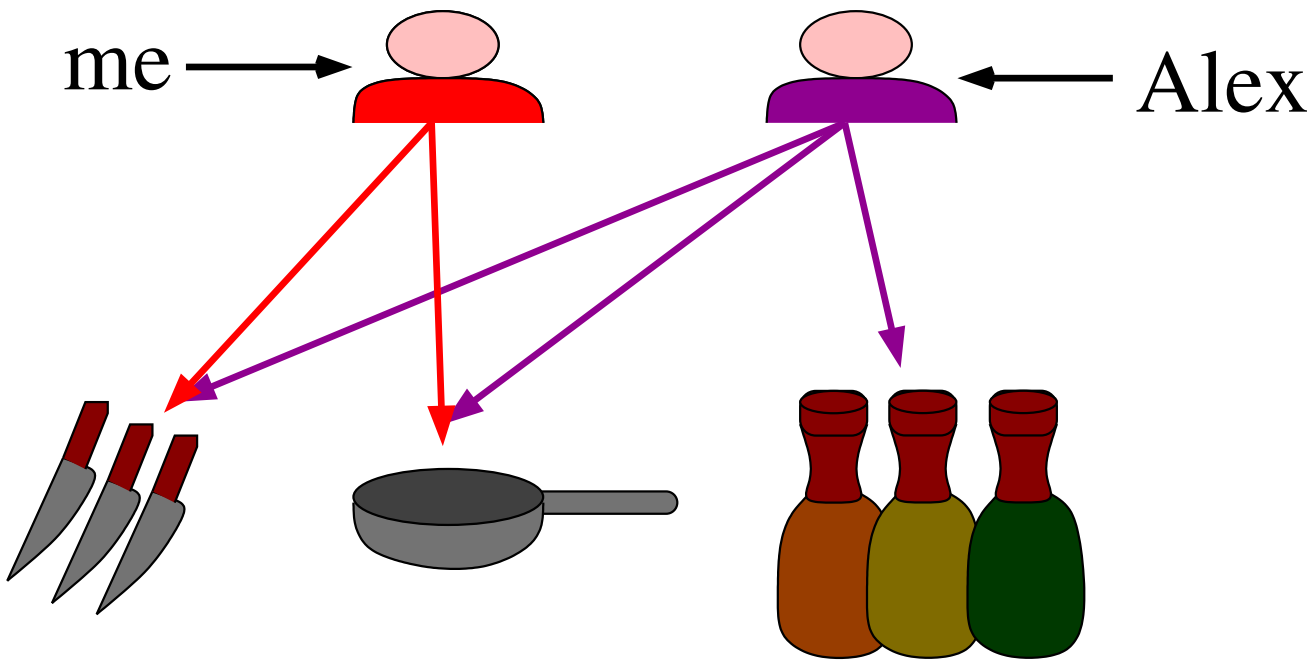
amazon.com®



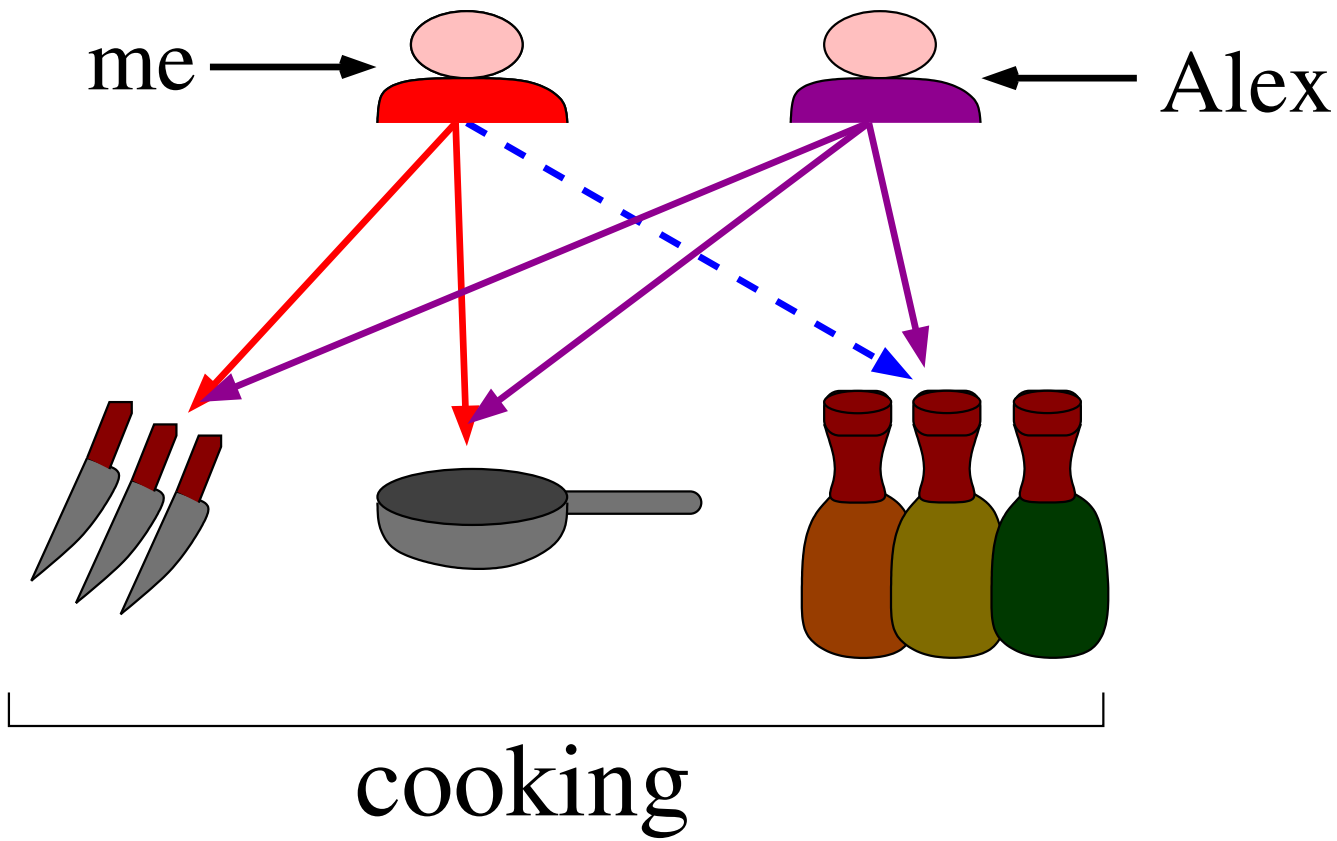
amazon.com®



amazon.com®

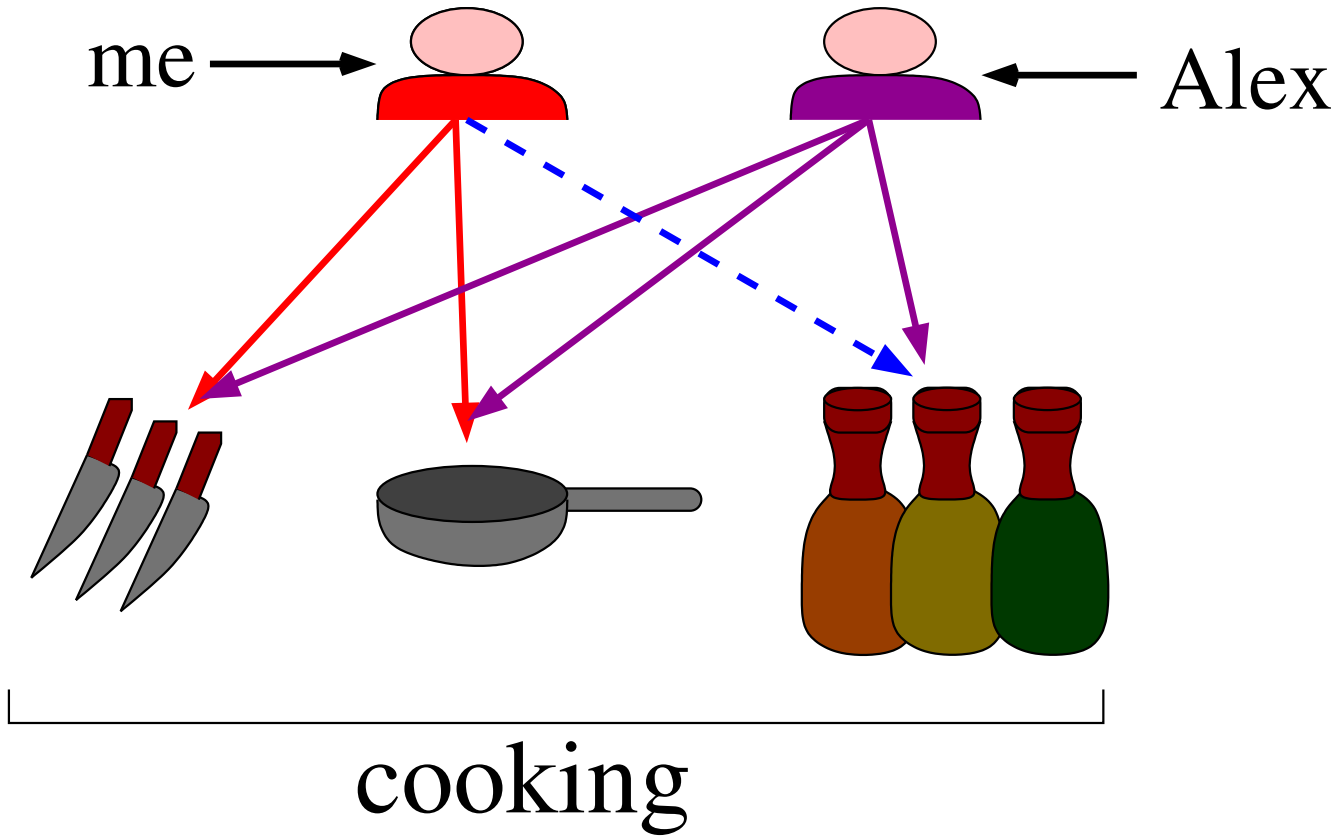


amazon.com®



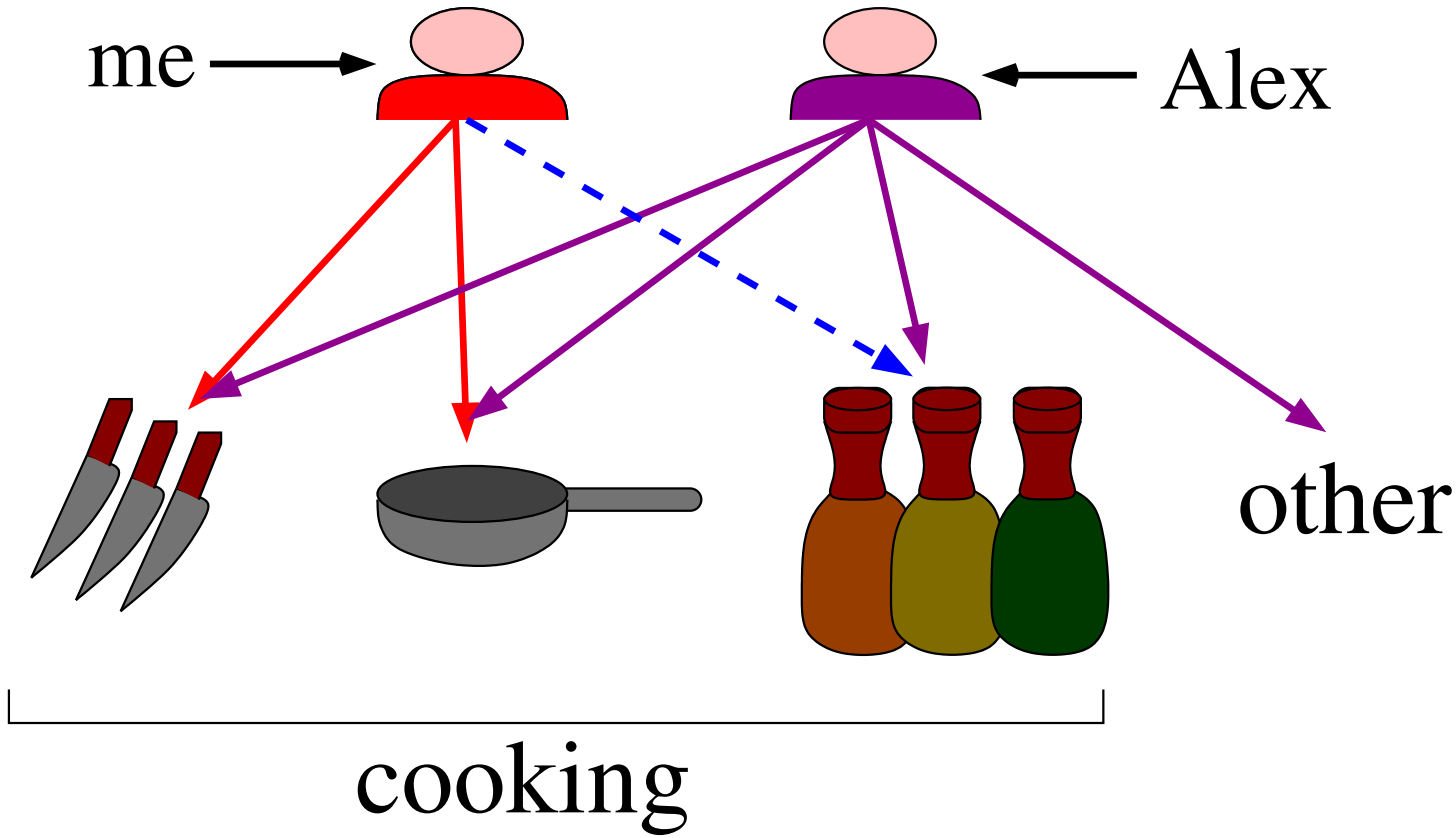
Lots of data (about similar customers) helps!

amazon.com

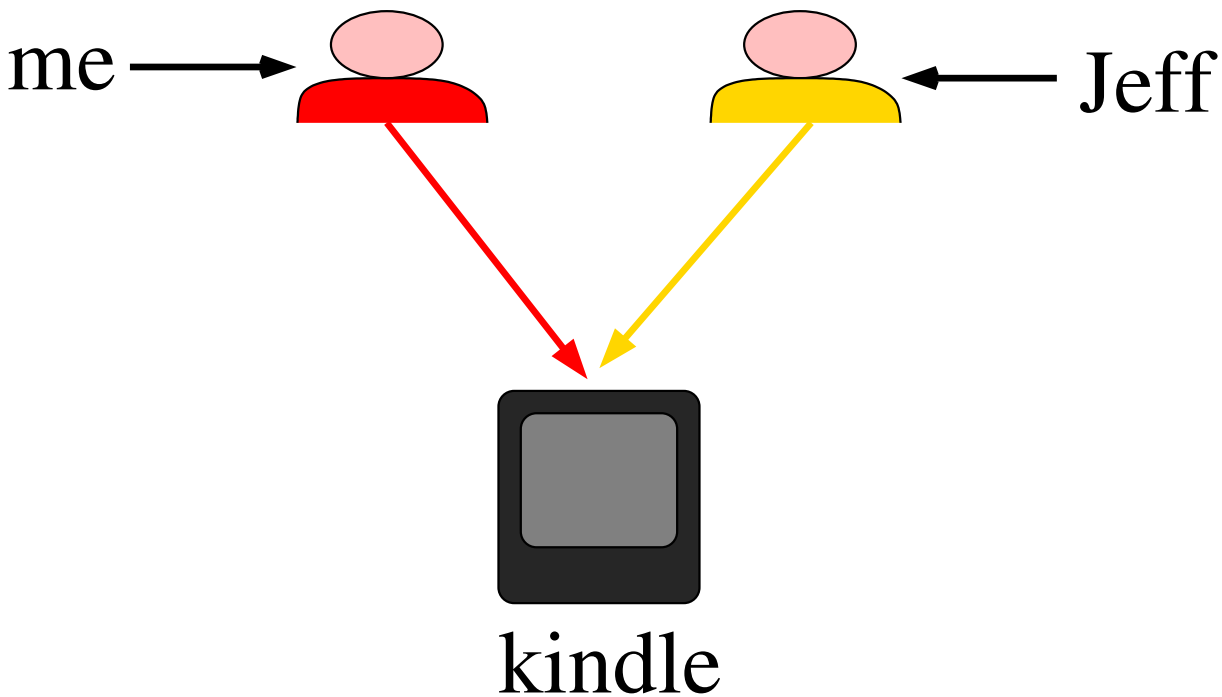


Lots of data (about similar customers) helps!

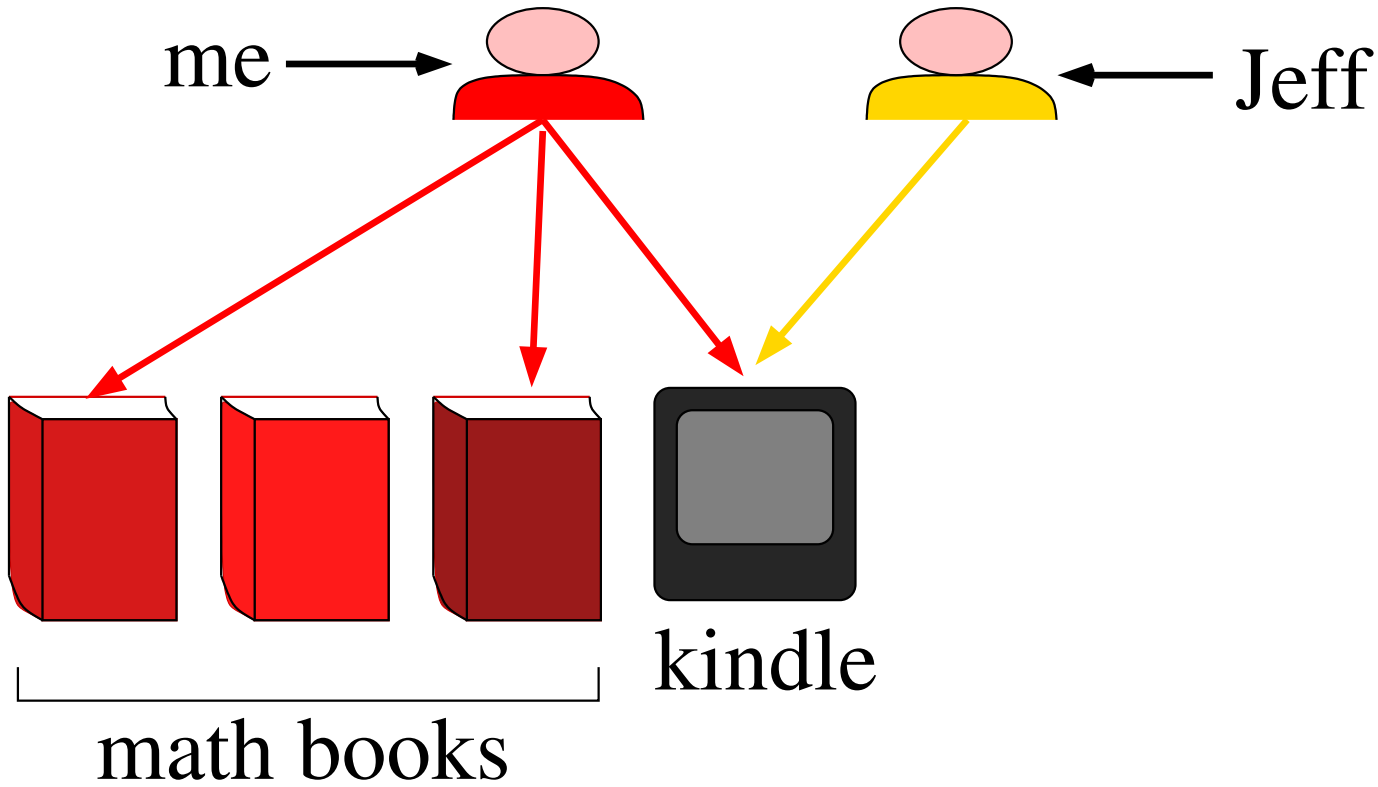
amazon.com



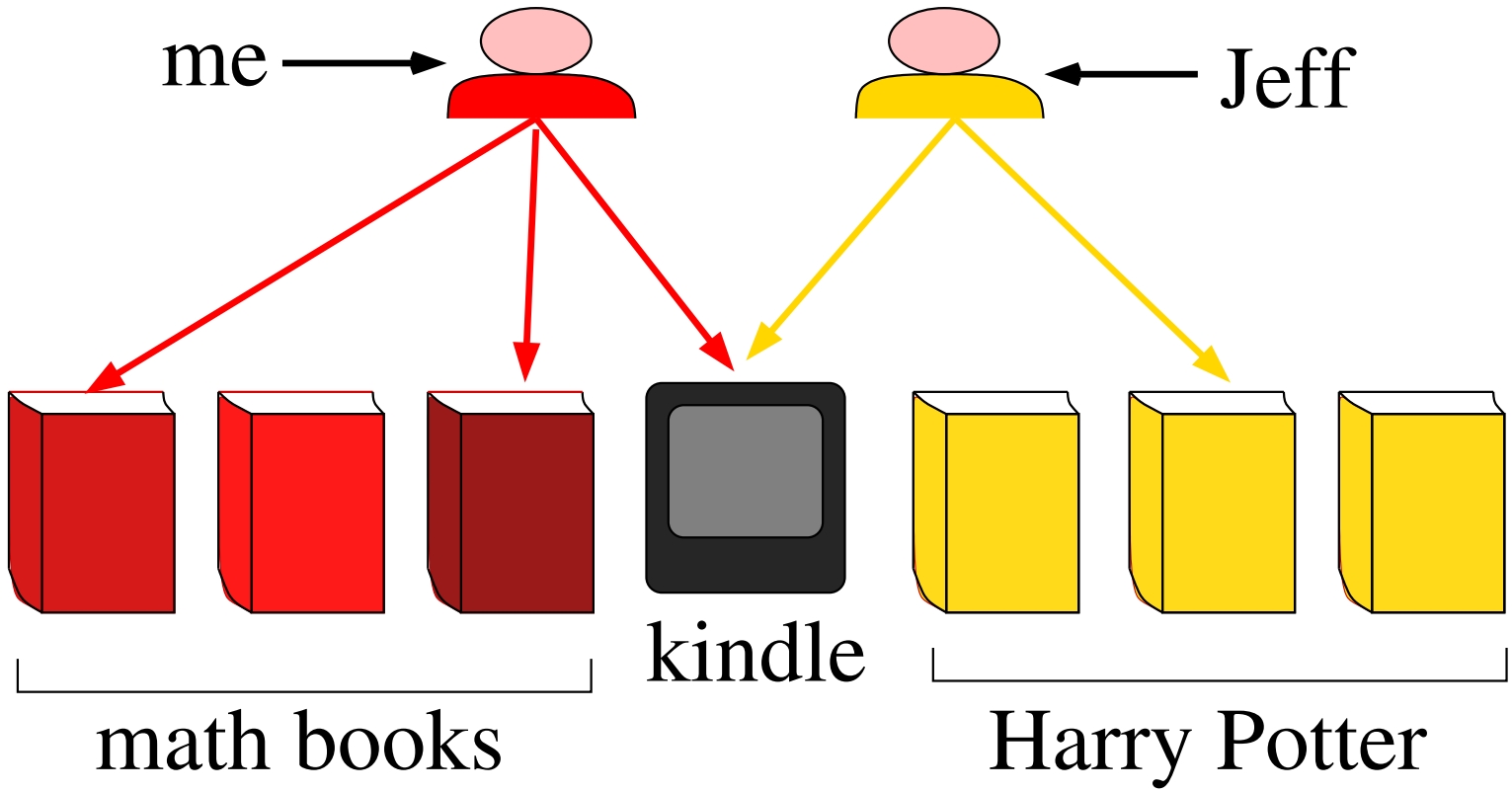
amazon.com®



amazon.com®



amazon.com®



How do we identify similar customers and products?

amazon.com

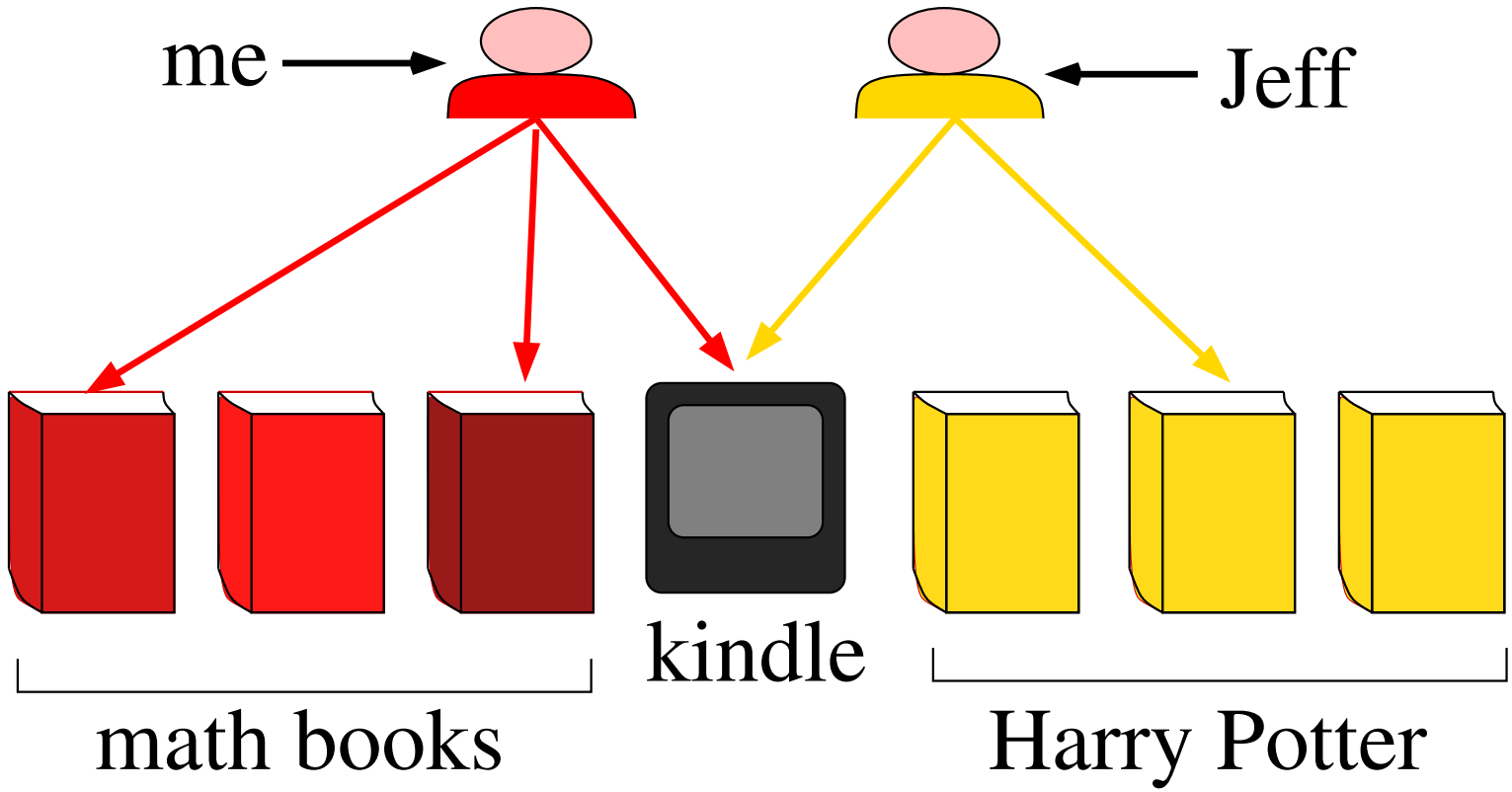


Table of which customers purchased which items:

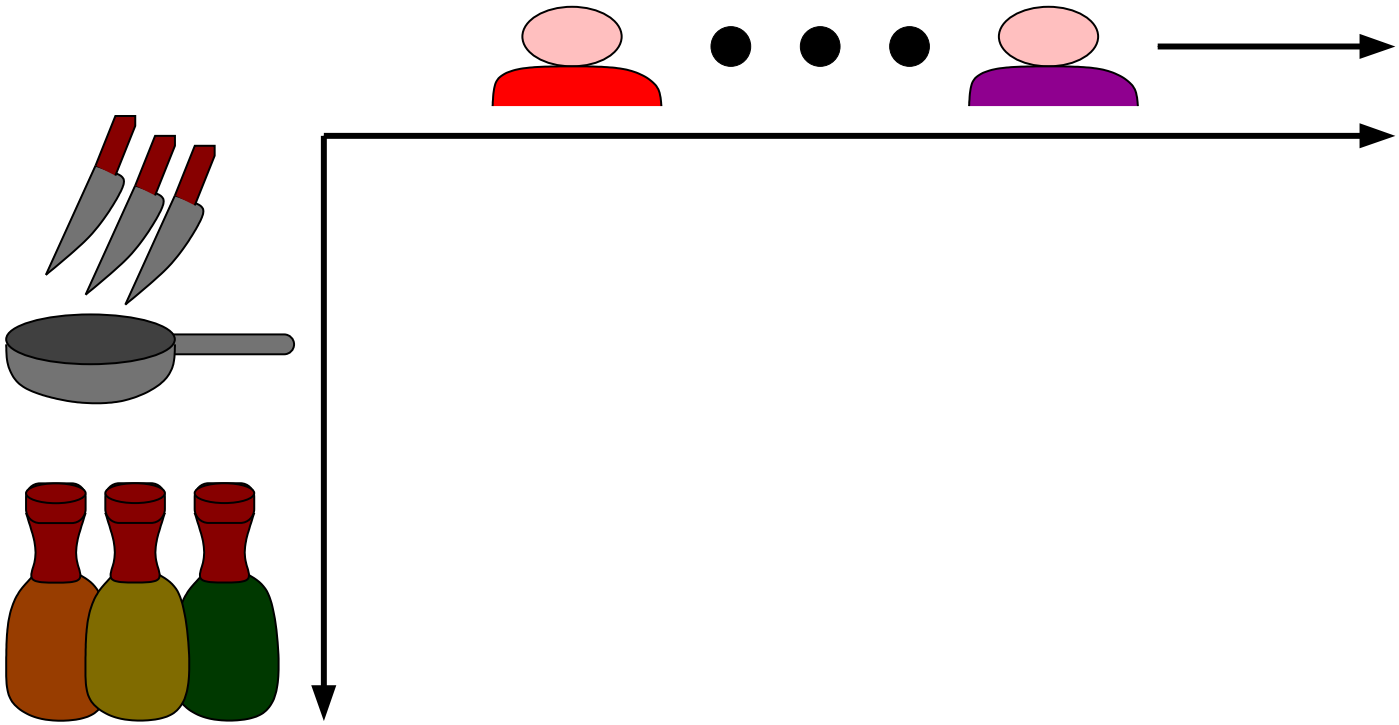
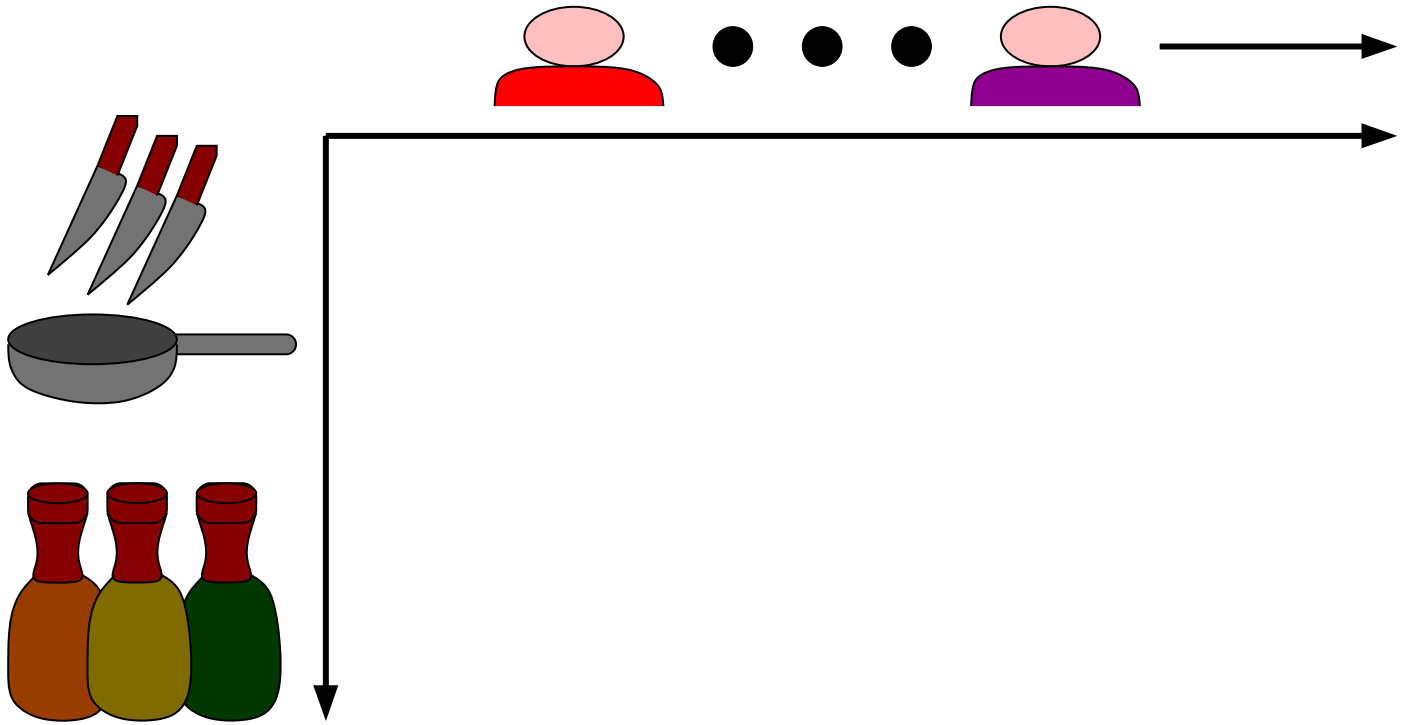
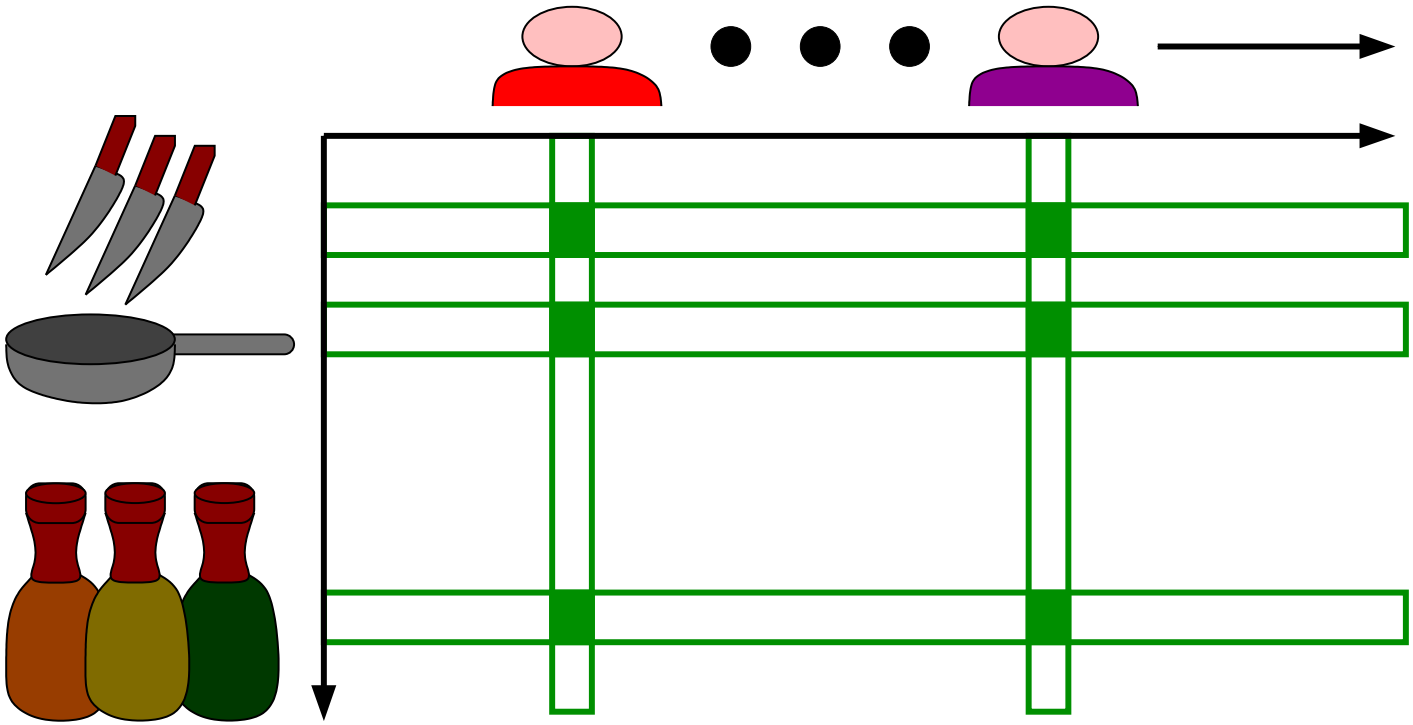


Table of which customers purchased which items:



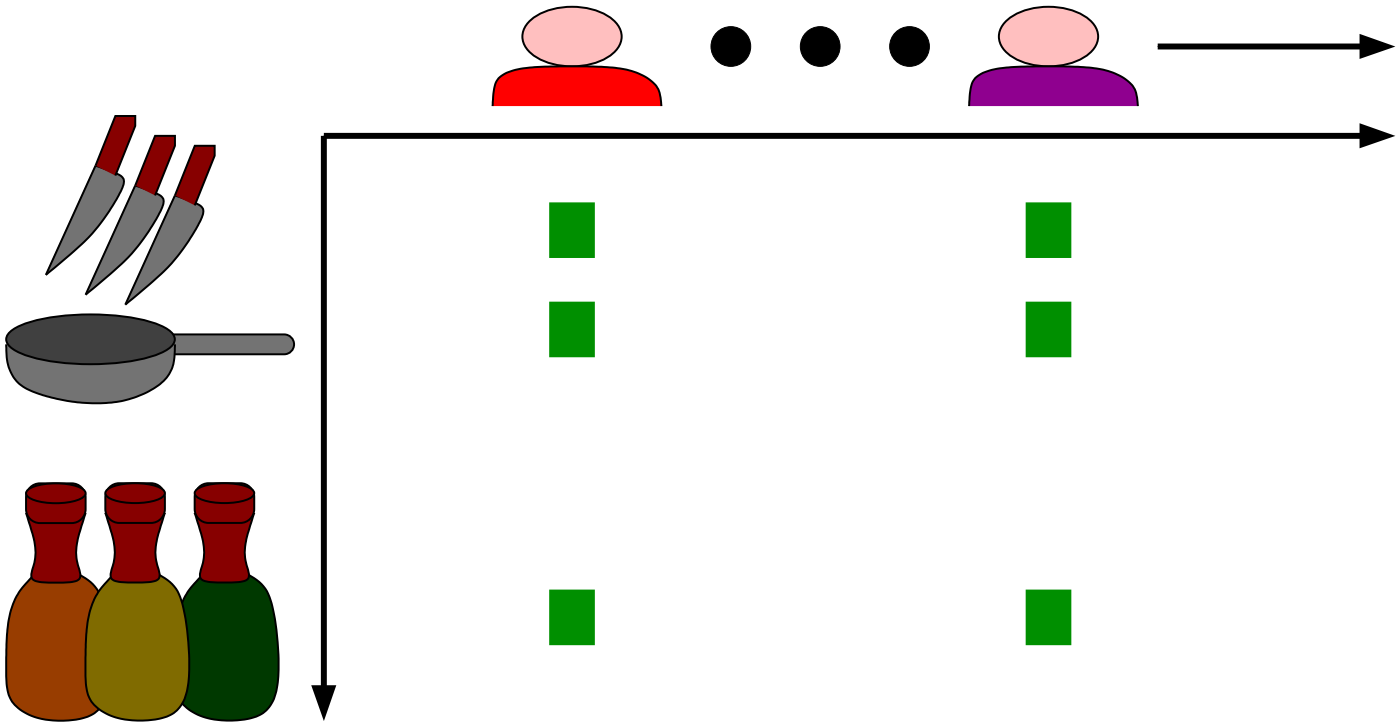
What is the structure in this data?

Table of which customers purchased which items:



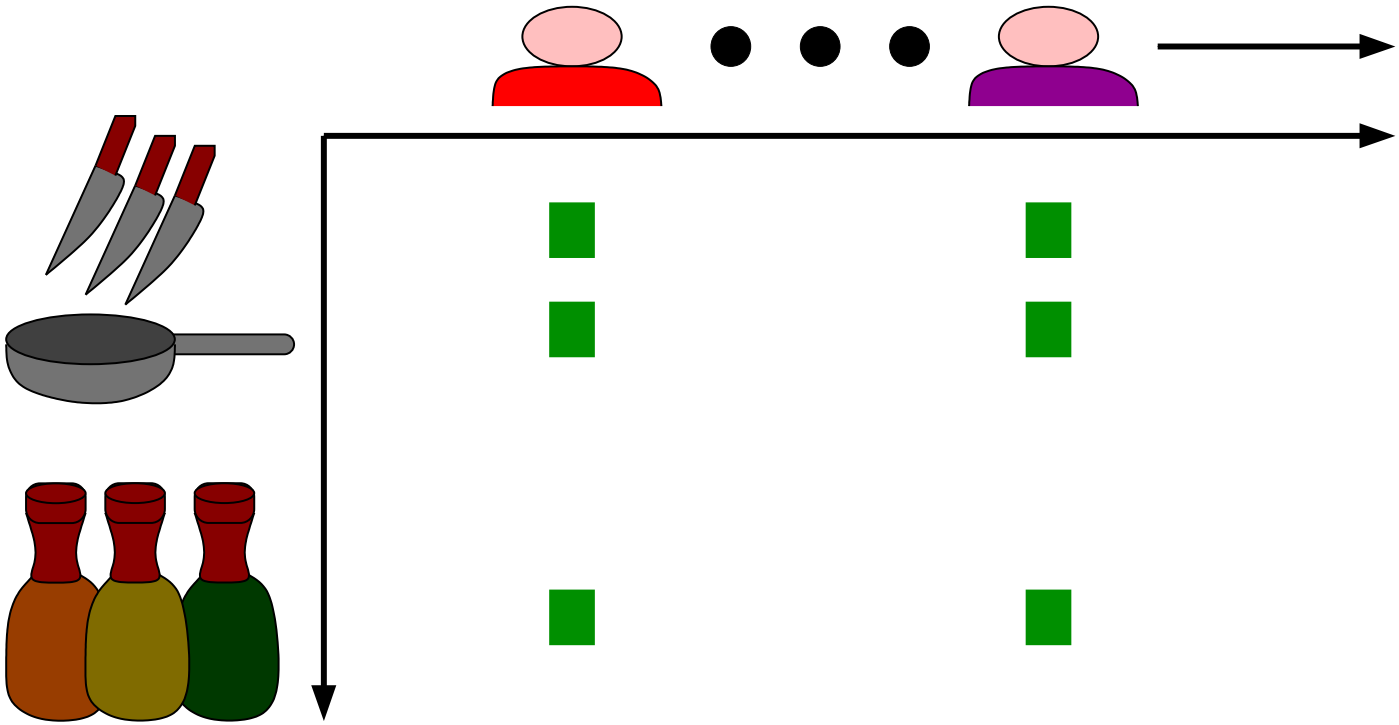
What is the structure in this data?

Table of which customers purchased which items:



What is the structure in this data?

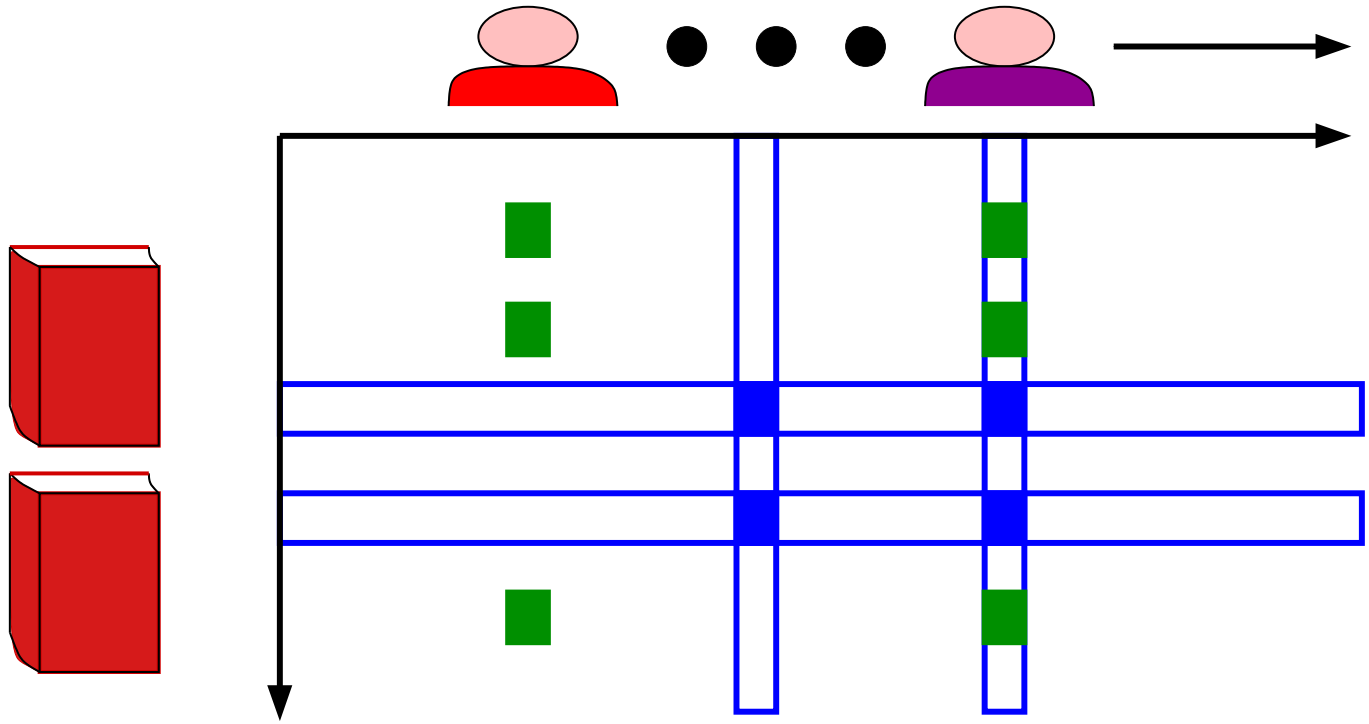
Table of which customers purchased which items:



What is the structure in this data?

Rectangles represent a shared interest of many customers

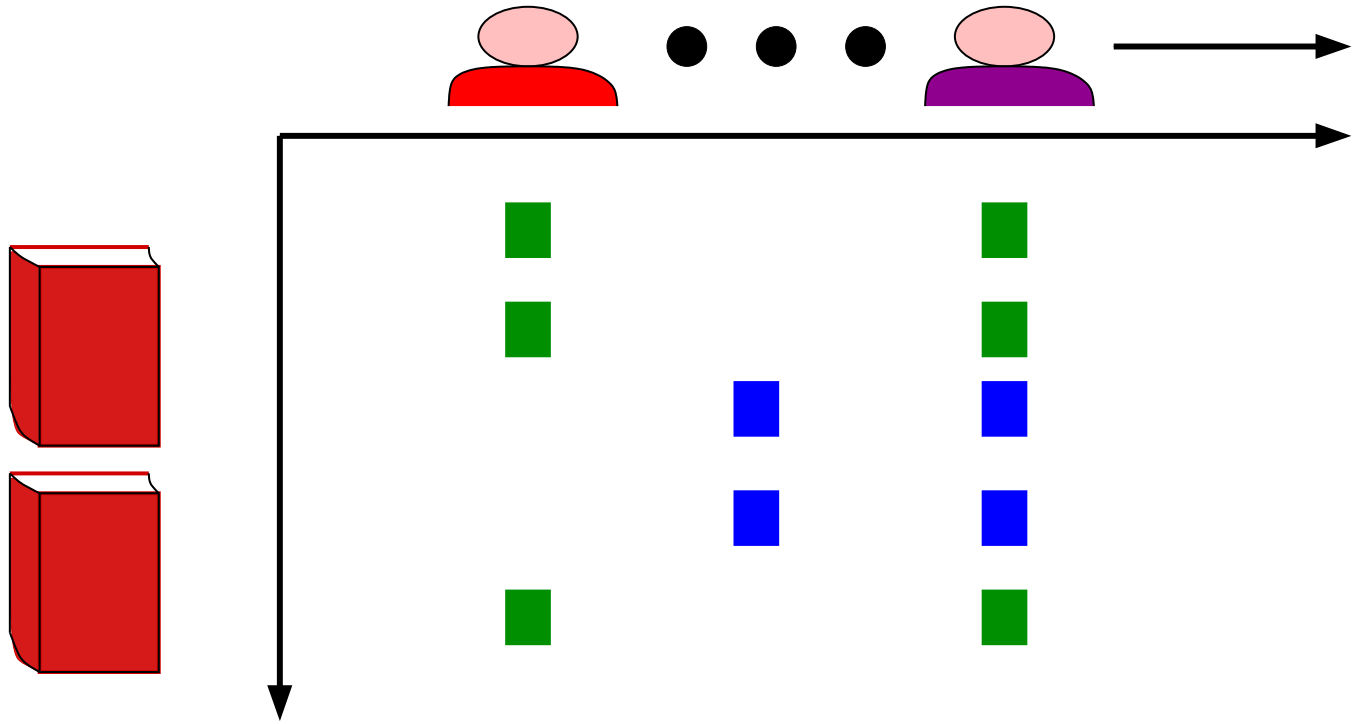
Table of which customers purchased which items:



What is the structure in this data?

Rectangles represent a shared interest of many customers

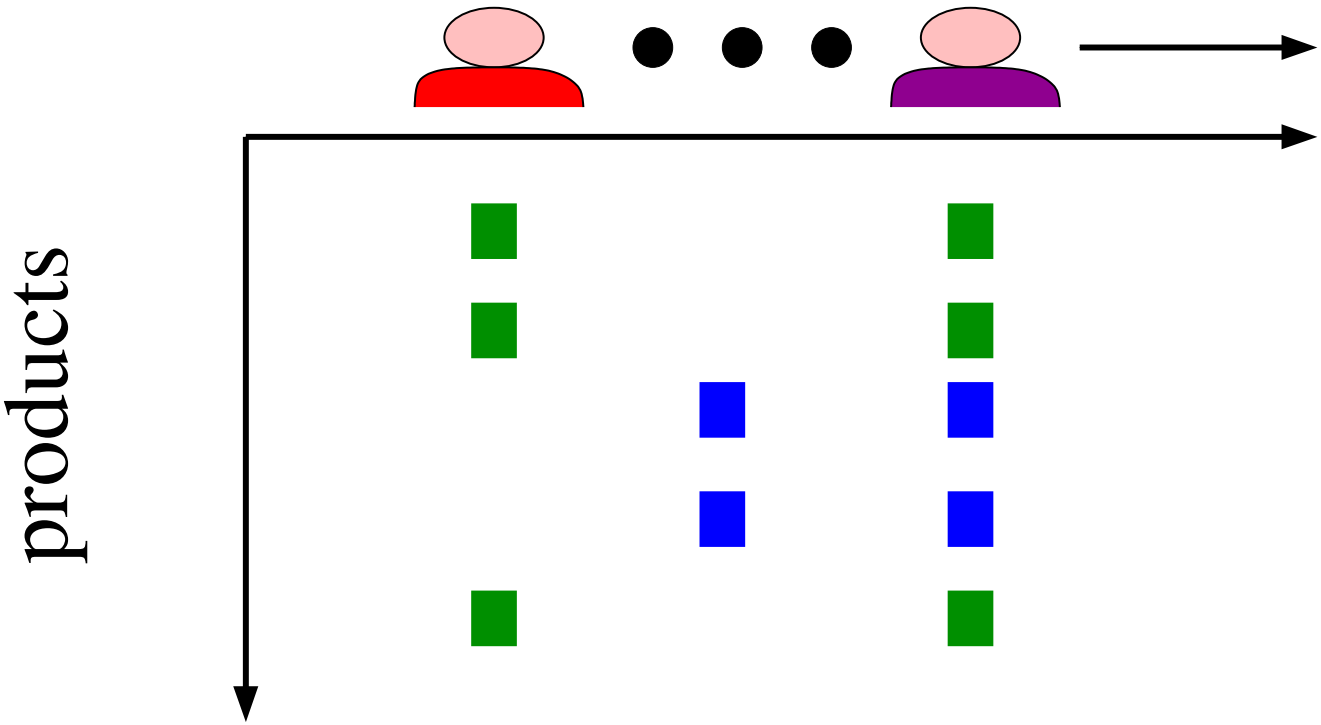
Table of which customers purchased which items:



What is the structure in this data?

Rectangles represent a shared interest of many customers

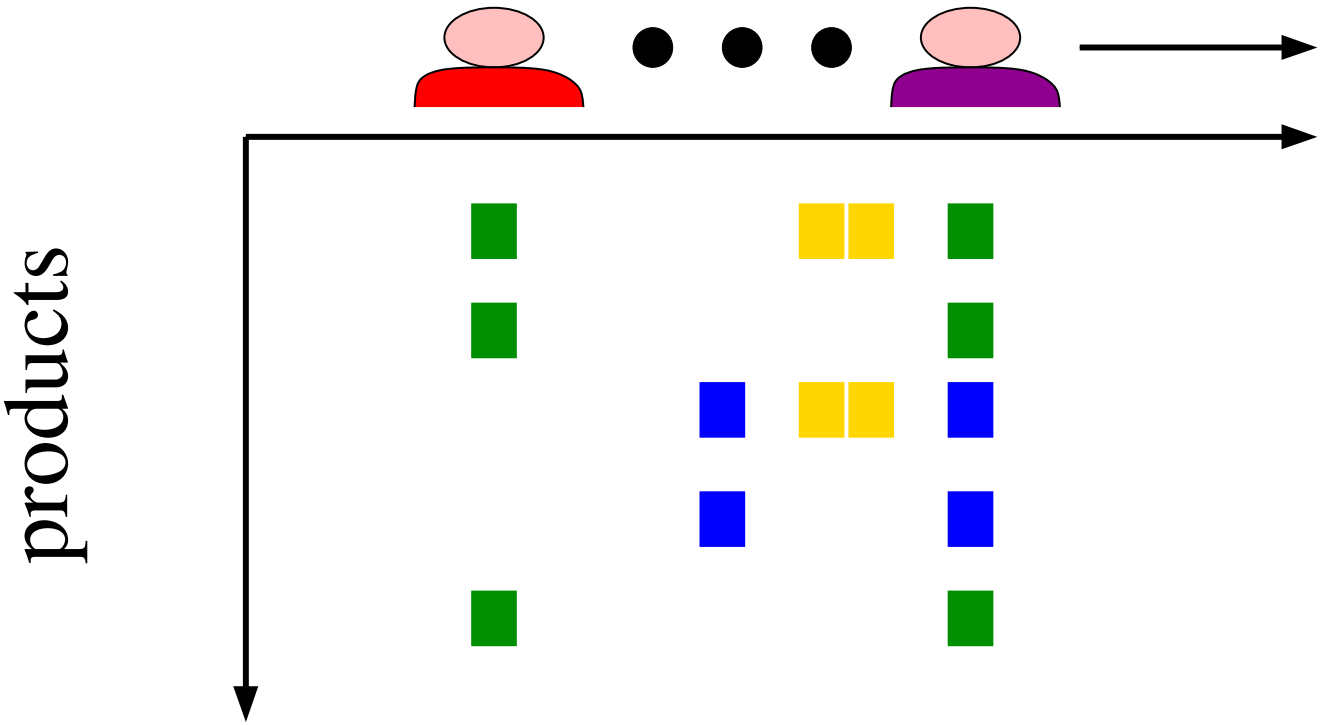
Table of which customers purchased which items:



What is the structure in this data?

Rectangles represent a shared interest of many customers

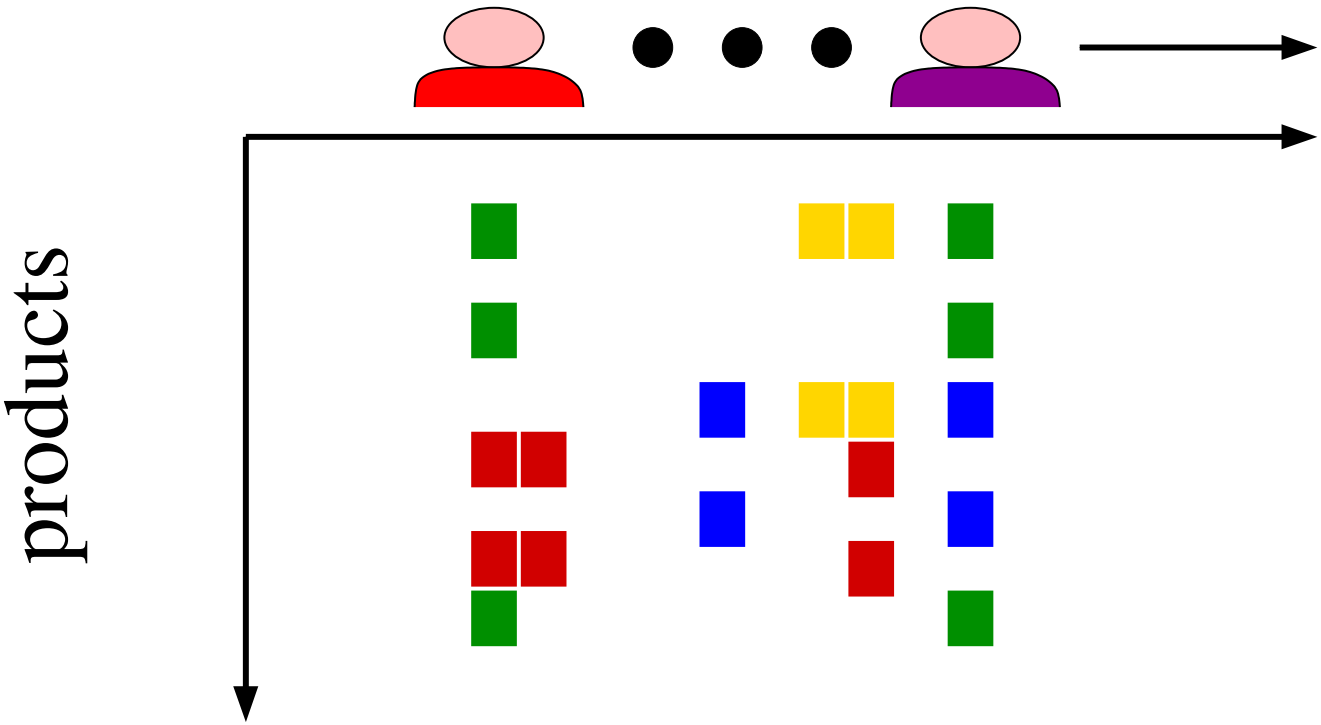
Table of which customers purchased which items:



What is the structure in this data?

Rectangles represent a shared interest of many customers

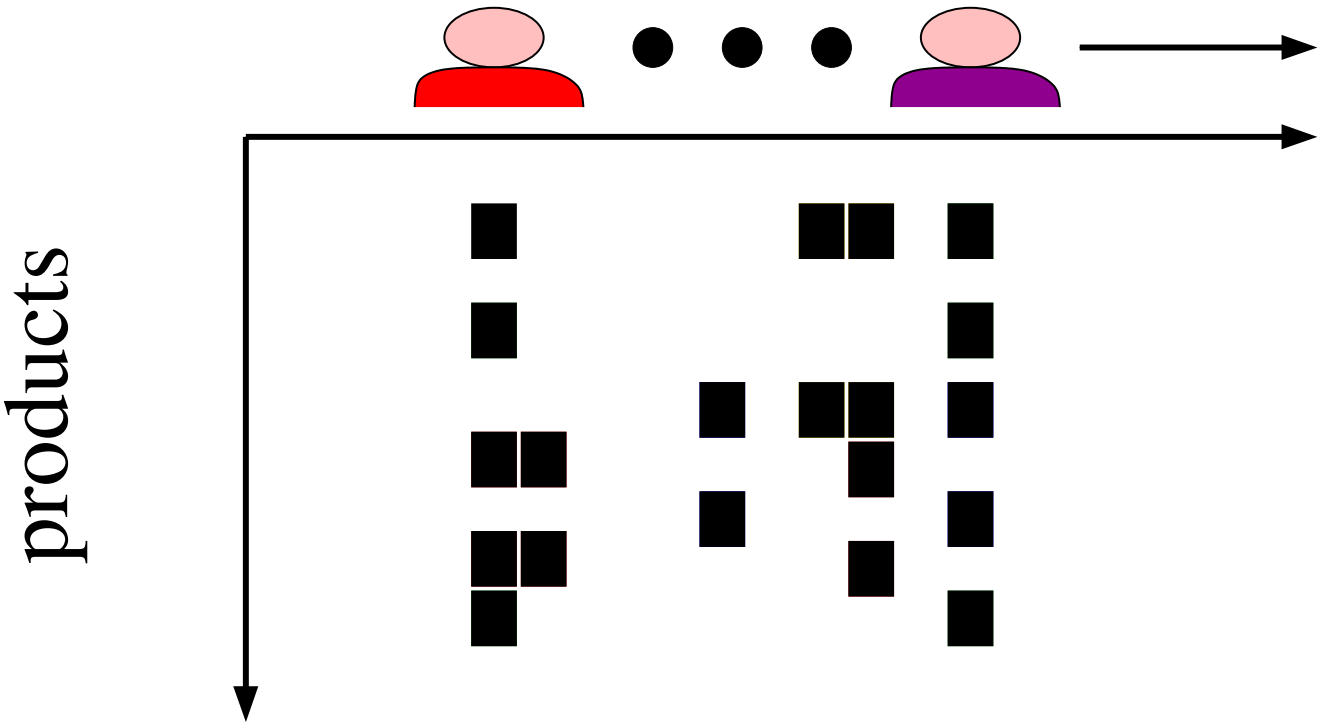
Table of which customers purchased which items:



What is the structure in this data?

Rectangles represent a shared interest of many customers

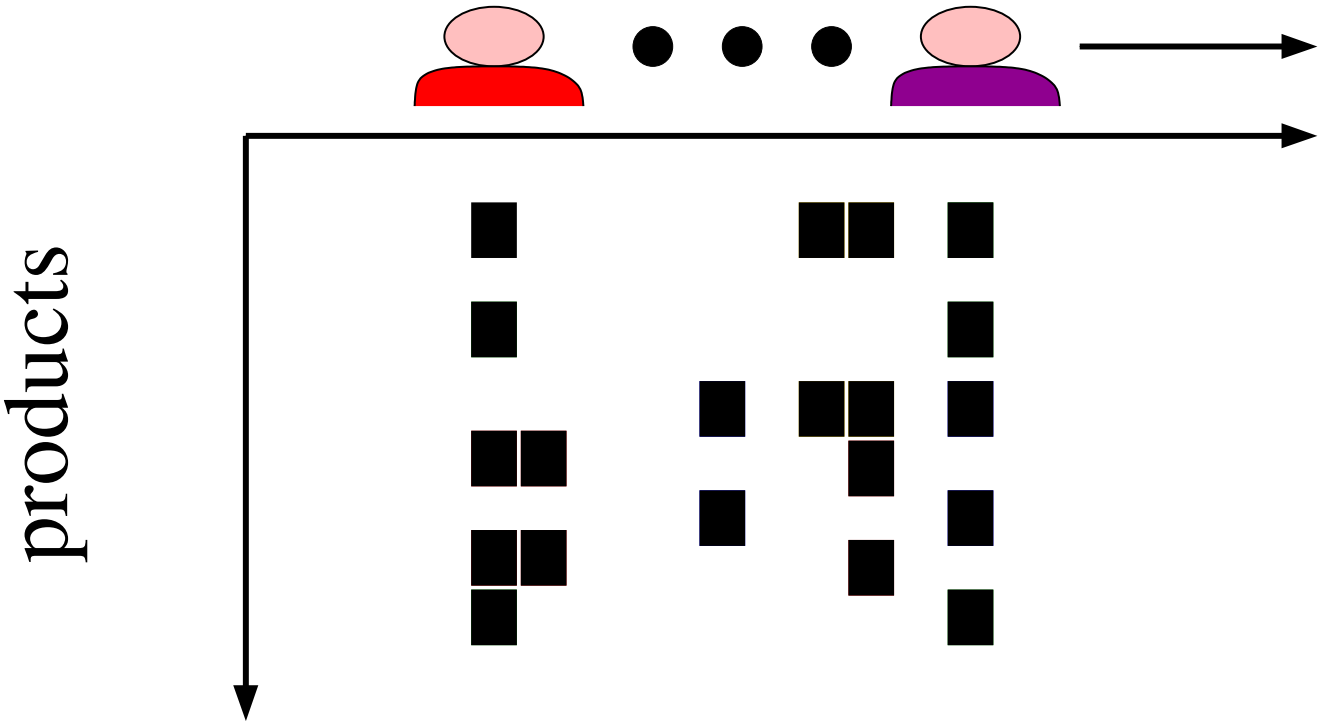
Table of which customers purchased which items:



What is the structure in this data?

Rectangles represent a shared interest of many customers

Table of which customers purchased which items:



What is the structure in this data?
Rectangles represent a shared interest of many customers

Common belief: a small number of "interests" explain the data

Computational Issues

Question

Can we find these patterns in the data quickly?

Computational Issues

Question

Can we find these patterns in the data quickly?

(This is called **Nonnegative Matrix Factorization**)

Computational Issues

Question

Can we find these patterns in the data quickly?

(This is called **Nonnegative Matrix Factorization**)

Many algorithms work well in practice, but no known theoretical explanation!

Computational Issues

Question

Can we find these patterns in the data quickly?

(This is called **Nonnegative Matrix Factorization**)

Many algorithms work well in practice, but no known theoretical explanation! Is there an efficient algorithm that works on every input?

Computational Issues

Question

Can we find these patterns in the data quickly?

(This is called **Nonnegative Matrix Factorization**)

Many algorithms work well in practice, but no known theoretical explanation! Is there an efficient algorithm that works on every input?

In joint work with Sanjeev Arora, Rong Ge and Ravi Kannan:
YES:

Computational Issues

Question

Can we find these patterns in the data quickly?

(This is called **Nonnegative Matrix Factorization**)

Many algorithms work well in practice, but no known theoretical explanation! Is there an efficient algorithm that works on every input?

In joint work with Sanjeev Arora, Rong Ge and Ravi Kannan:
YES: solve a system of polynomial inequalities with few variables

Beyond Worst-Case

Question

Can we give a theoretical explanation for why heuristics work so well in practice?

Beyond Worst-Case

Question

Can we give a theoretical explanation for why heuristics work so well in practice?

There must be some property of the problems we actually want to solve that makes them easier

Beyond Worst-Case

Question

Can we give a theoretical explanation for why heuristics work so well in practice?

There must be some property of the problems we actually want to solve that makes them easier

---

In joint work with Sanjeev Arora and Rong Ge:
We found such a property (that empirically holds),

Beyond Worst-Case

Question

Can we give a theoretical explanation for why heuristics work so well in practice?

There must be some property of the problems we actually want to solve that makes them easier

---

In joint work with Sanjeev Arora and Rong Ge:
We found such a property (that empirically holds), and makes these problems **much easier** (than the worst-case)!

Many More Applications...

Topic Modeling

Information Retrieval

Clustering

Automated Diagnosis

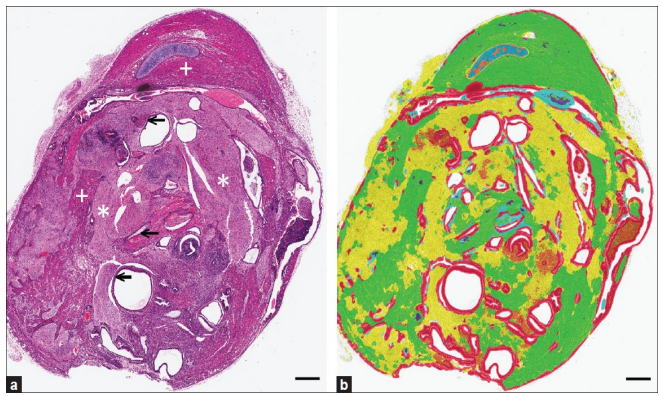


Image Segmentation

Face Recognition

Thanks!