CIS 422 Projects

- Group projects: up to 4 students in each group
- Your project must exemplify new skills that you learned from this class:
  - Data modeling, SQL, Indexing, ER diagrams, algorithms
- Two deadlines:
  - proposal + final submission

CIS 422 Projects

- Data-driven project:
  - Choose a real-world data set
  - Design a relational schema for it
  - Convert and load the into a SQL-like system
    - MonetDB, Cassandra, CouchDB, Neo4j, MongoDB, ...
  - Design several queries to extract interesting facts from your data
  - Write a project report, report your findings
  - Submit both your report and your code

CIS 422 Projects

- Data-Driven project deliverables:
  - A copy of the data set (or a link to it)
  - Scripts for converting the data into a format suitable for the target system + converted files
  - ER model of the database design
  - DDL statements to define the schema
  - DML statements to populate the database + Indexes
  - queries, with explanation and observed result
  - A project report (pdf)

CIS 422 Projects

- The Project report must:
  - Explain the project goals
  - Explain the chosen domain / data set
  - Explain the decisions made in the design stage
  - Document the database design with ER diagrams
  - In case indexes are created, report the observed performance improvement
  - Explain the purpose of each SQL query and report its observed results
- Everything that you write in the report must be reproducible by the instructor

CIS 422 Projects

- Caveats:
  - **Only original work is acceptable**
  - You cannot reuse examples from books, slides, websites and/or other classes
  - You must put the instructor in the position of reproducing every step of your project
    - data transformation
    - schema creation and data loading
    - indexing
    - queries

Plausible data sets

- UCI ML Data Set Repository
  - http://archive.ics.uci.edu/ml/datasets
- Market Basket Analysis
  - http://archive.ics.uci.edu/ml/datasets/online+retail
- Census Data
  - https://ipums.org
- Survey data
  - https://www.kaggle.com/kaggle/kaggle-survey-2018

Plausible data sets

- DB performance data set
  - http://www.tpc.org/tpch/
- House pricing data
  - https://www.zillow.com/research/data/
- NBA-related data
  - Too many data sets to be listed!

Example: Market Basket Analysis

- Your database tracks products, customers and transactions for a grocery store
- Products are identified by a unique ID, and have properties like price, availability, ...
- Customers have a unique ID, a name, an address..
- “Transactions” represent the act of one customer buying several items

Example: Market Basket Analysis

- "Transactions" represent the act of one customer buying several items

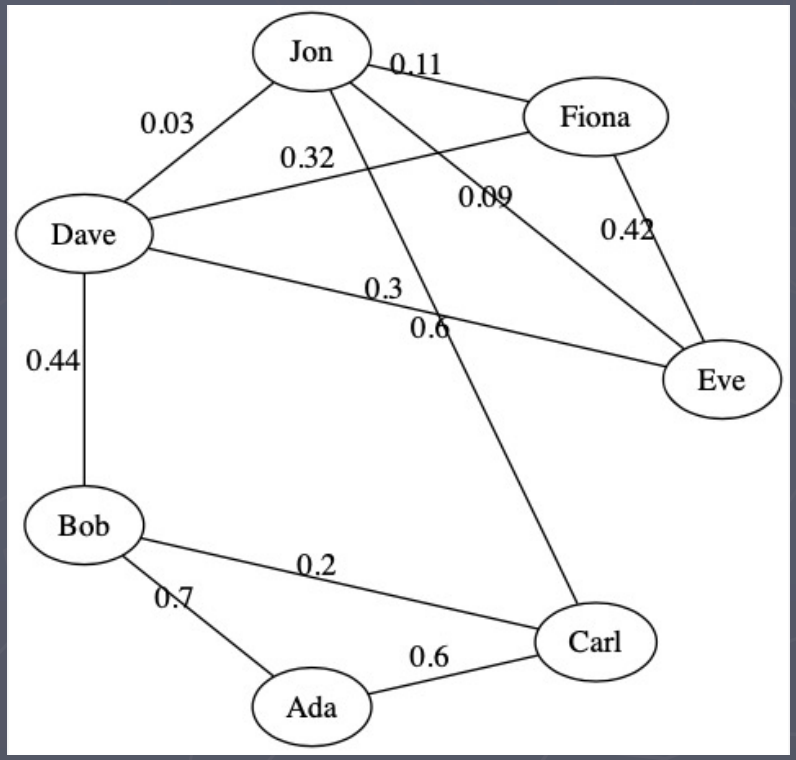
GroceryStoreTransactions				
TxnID	CustID	ProdID	Qty	UnitPrice
1	Ada	Beer	6	11 USD
1	Ada	Olives	2	4 USD
2	Bob	ToiletP	100	250 USD
2	Bob	Water	8	12 USD
3	Ada	Beer	6	11 USD

Example: Market Basket Analysis

- Plausible queries:
- What are the top 10 sets of three items that are bought together most frequently?
- Which customer is the most similar to Bob in his/her buying behavior?
- For more ideas: https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

Example: Social Networks

- Your database tracks people and their social connections
- The data looks like a graph

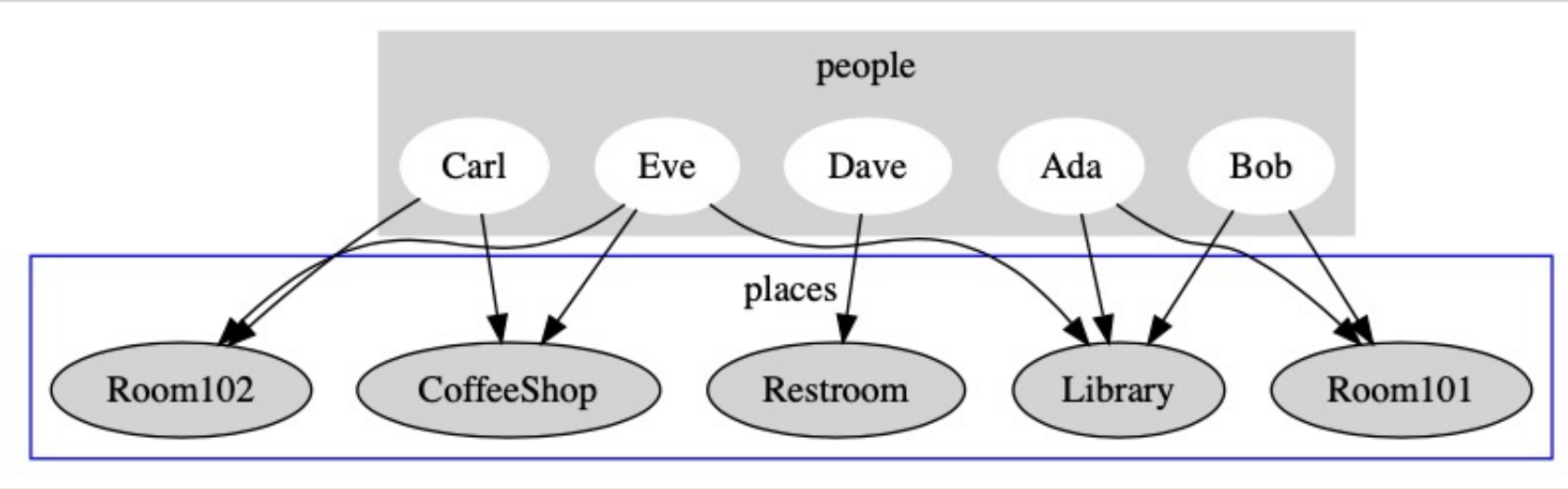


Example: Social Network

- Plausible queries:
- Who are the most connected users?
- Who are the most isolated users?
- How can we suggest new friends to Bob?
- Is it true that every two users are connected by a chain of at most 3 connections?

Example: Contact Tracing

- Your database tracks people and the places they visit
- For each visit we record a timestamp and a duration



Example: Contact Tracing

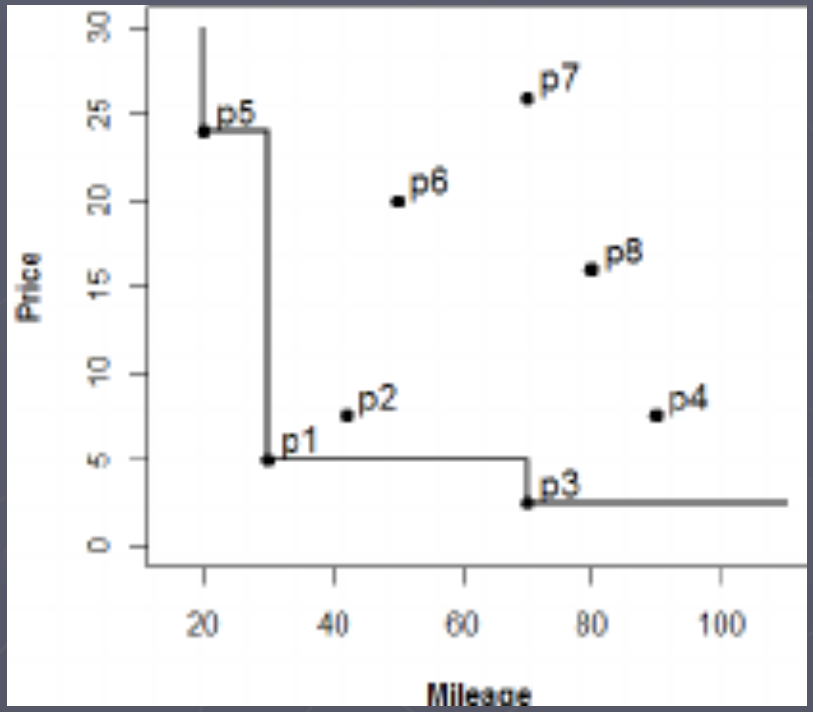
- Your database tracks people and the places they visit
- For each visit we record a timestamp and a duration
- Plausible queries:
  - If Bob exhibits symptoms, who are the top-10 individuals that should be tested first?

Example: Vacation Planning

- Your database proposes deals to someone who is planning a vacation
- Each deal can be evaluated according to several properties
  - Price (the lower the better)
  - Hotel rating (the higher the better)
  - Hotel reviews (the higher the better)
  - .. and many others

Example: Vacation Planning

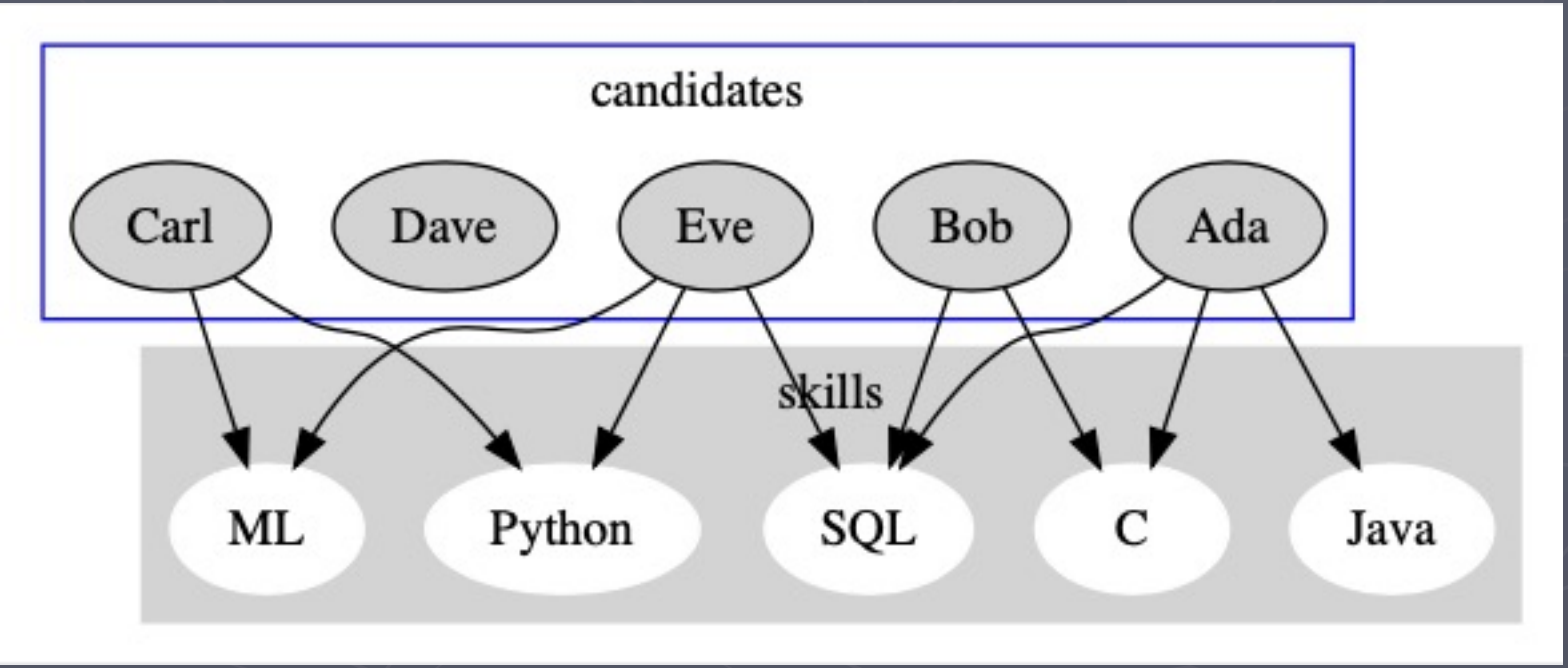
- An offer o_1 is "dominated" by another offer o_2 if o_2 is as good as o_1 in all the dimensions, and strictly better in at least one dimension
- Skyline queries: Identify all the offers that are not dominated by any other offer



- More ideas: http://delab.csd.auth.gr/papers/IISA2015tpm.pdf

Example: Hiring Decisions

- Your database tracks candidates and their skills
- Each open position requires a set of skills



Example: Hiring Decisions

- Plausible queries:
- Find the best matching candidates for each open position
- Find the best matching positions for each candidate
- If the hiring decisions are known: identify discriminative hiring policies
  - Measure Fairness Metrics:

    Group Fairness vs. Individual Fairness
  - https://dataresponsibly.github.io/courses/documents/spring20/Lecture1.pdf