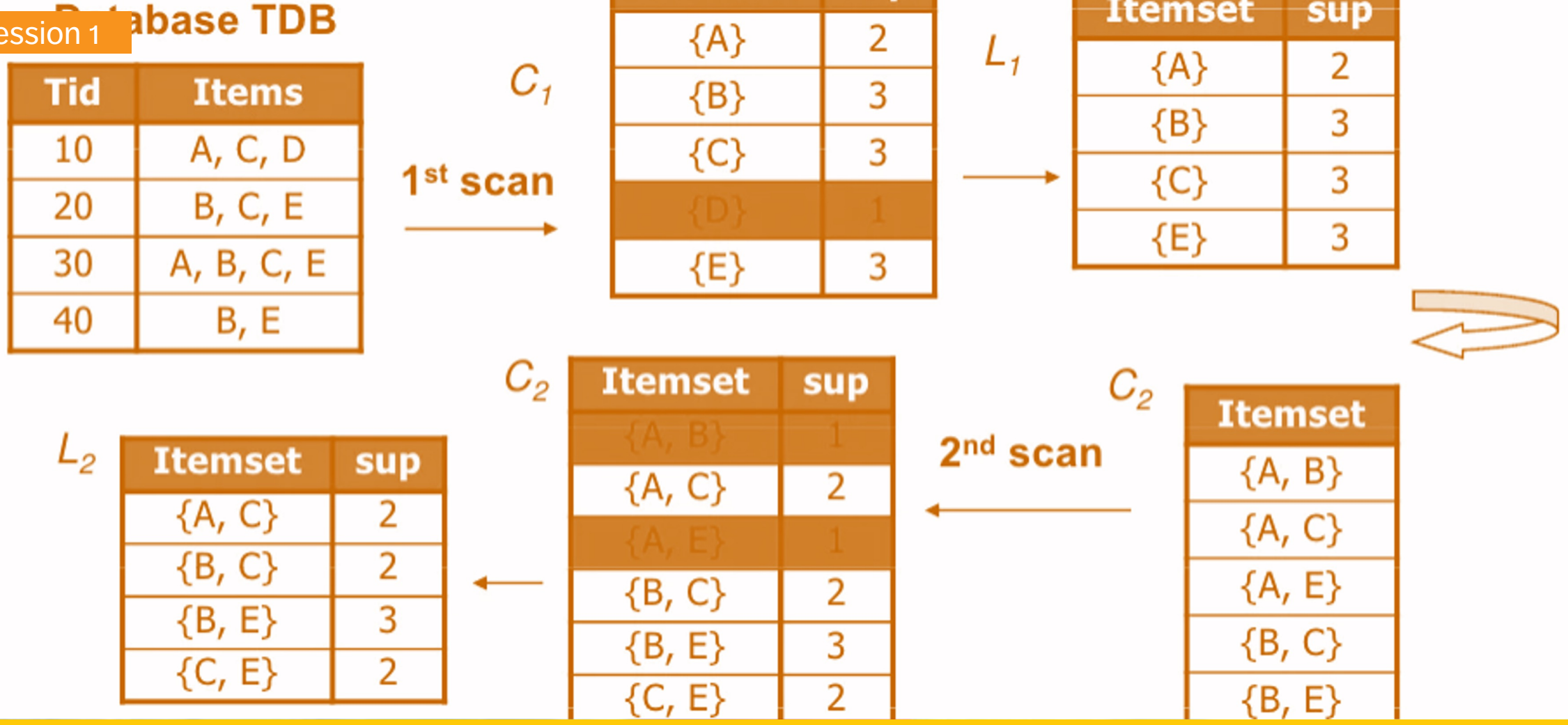


2. Association Rules Overview

Overview

Association rules discovery (ARD) is a type of unsupervised learning that finds **connections** among the attributes and levels of a dataset's observations.

We might analyze a dataset on the physical activities and purchasing habits of North Americans and discover that

- runners who are also triathletes (the **premise**) tend to drive Subarus, drink microbrews, and use smart phones (the **conclusion**)
- individuals who have purchased home gym equipment are unlikely to be using it 1 year later

Overview

The presence of a **correlation** between the premise and the conclusion does not imply the existence of a **causal relationship** between them.

It is difficult to prove causation *via* data analysis; in practice, decision-makers pragmatically (often erroneously) focus on "**there's no smoke without fire**."

Example: being a triathlete does not cause one to drive a Subaru, but Subaru Canada thought that the connection was strong enough to offer to reimburse the registration fee at an IRONMAN 70.3 competition (at least in 2018)!

Market Basket Analysis

ARD is aka as **market basket analysis**.

Example: purchase of bread and milk, but that is unlikely to be of interest given the frequency of market baskets containing milk (**or** bread).

If the presence of milk is **independent** of the presence of bread (and *vice-versa*), and if 70% of baskets contain milk and 90% contain bread, say, we would expect **at least** 90% × 70% = 63% of all baskets to contain **both**.

If we observe both in 72% of baskets, say (a 1.15-fold increase), we conclude that there is a **weak correlation** between the milk and bread purchases.

Market Basket Analysis

Sausages and buns are not purchased as frequently as milk and bread, but they might still be purchased as a pair more often than one would expect.

If the presence of sausage is **independent** of the presence of buns (and *vice-versa*), and if 10% of baskets contain sausages and 5% contain buns, say, we would expect **at least** 10% × 5% = 0.5% of all baskets to contain **both**.

If we observe both in 4% of baskets, say (an 8-fold increase), we conclude that there is a **strong correlation** between the sausage and buns purchases.

Market Basket Analysis

How can we **act** on this insight? Supermarkets could advertise a sale on sausages while **simultaneously** (and quietly) raising the price of buns. This could have the effect of bringing in a higher number of customers into the store, hoping to increase the **sale volumes** for both items while keeping the **combined price of the two items constant**.

Little Story: a supermarket found an association rule linking the purchase of beer and diapers and consequently moved its beer display closer to its diapers display, having confused correlation and causation.

What do you think might actually be happening here?

Applications

Typical uses include:

- finding **related concepts** in text documents – looking for pairs (triplets, etc) of words that represent a joint concept: {San Jose, Sharks}, {Michelle, Obama}, etc.;
- detecting **plagiarism** – looking for specific sentences that appear in multiple documents, or for documents that share specific sentences;
- identifying **biomarkers** – finding diseases frequently associated with a set of biomarkers;

Applications

Typical uses include:

- making predictions and decisions based on association rules (there are pitfalls)
- altering circumstances to take advantage of correlations (suspected causal effect)
- using connections to modify the likelihood of certain outcomes (see above)
- imputing missing data
- text autofill and autocorrect
- etc.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Objective

Using data from the *Danish National Patient Registry*, the authors sought connections between different **diagnoses:** how does a diagnosis at some point in time allow for the prediction of another diagnosis at a later time?

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Methodology

1. compute the **strength of correlation** for pairs of diagnoses over a 5 year interval (on a representative subset of the data)
2. test diagnoses pairs for **directionality** (one diagnosis repeatedly occurring before the other)
3. determine reasonable **diagnosis trajectories** (thoroughfares) by combining smaller (but frequent) trajectories with overlapping diagnoses
4. **validate** the trajectories by comparison with non-Danish data
5. **cluster** the thoroughfares to identify a small number of **central medical conditions** (key diagnoses) around which disease progression is organized

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Data

The *Danish National Patient Registry* is an electronic health registry containing administrative information and diagnoses, covering the whole population of Denmark, including private and public hospital visits of all types:

- inpatient (overnight stay)
- outpatient (no overnight stay)
- emergency visits.

The data set covers 15 years of such visits, from January ’96 to November ’10, and consists of 68 million records for 6.2 million patients.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Challenges and Pitfalls

- Access to the **patient registry** was protected and could only be granted after approval by *the National Board of Health*.
- There are gender-specific differences in diagnostic trends, but many diagnoses were made predominantly in different sites, suggesting the stratifying by **site** as well as by **gender**.
- In the process of forming small diagnoses chains, they had to compute the correlations using **large groups** for each pair of diagnoses (1 million diagnosis pairs = 80+ million samples) to compensate for **multiple testing** (1000s years' worth of CPU run time) – pre-filtering steps were used to avoid this pitfall.

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Project Summary and Results

The dataset was reduced to **1,171 significant trajectories**.

These thoroughfares were clustered into patterns centred on 5 key diagnoses for disease progression:

- **diabetes**
- **chronic obstructive pulmonary disease** (COPD)
- **cancer**
- **arthritis**
- **cerebrovascular disease**

Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Project Summary and Results

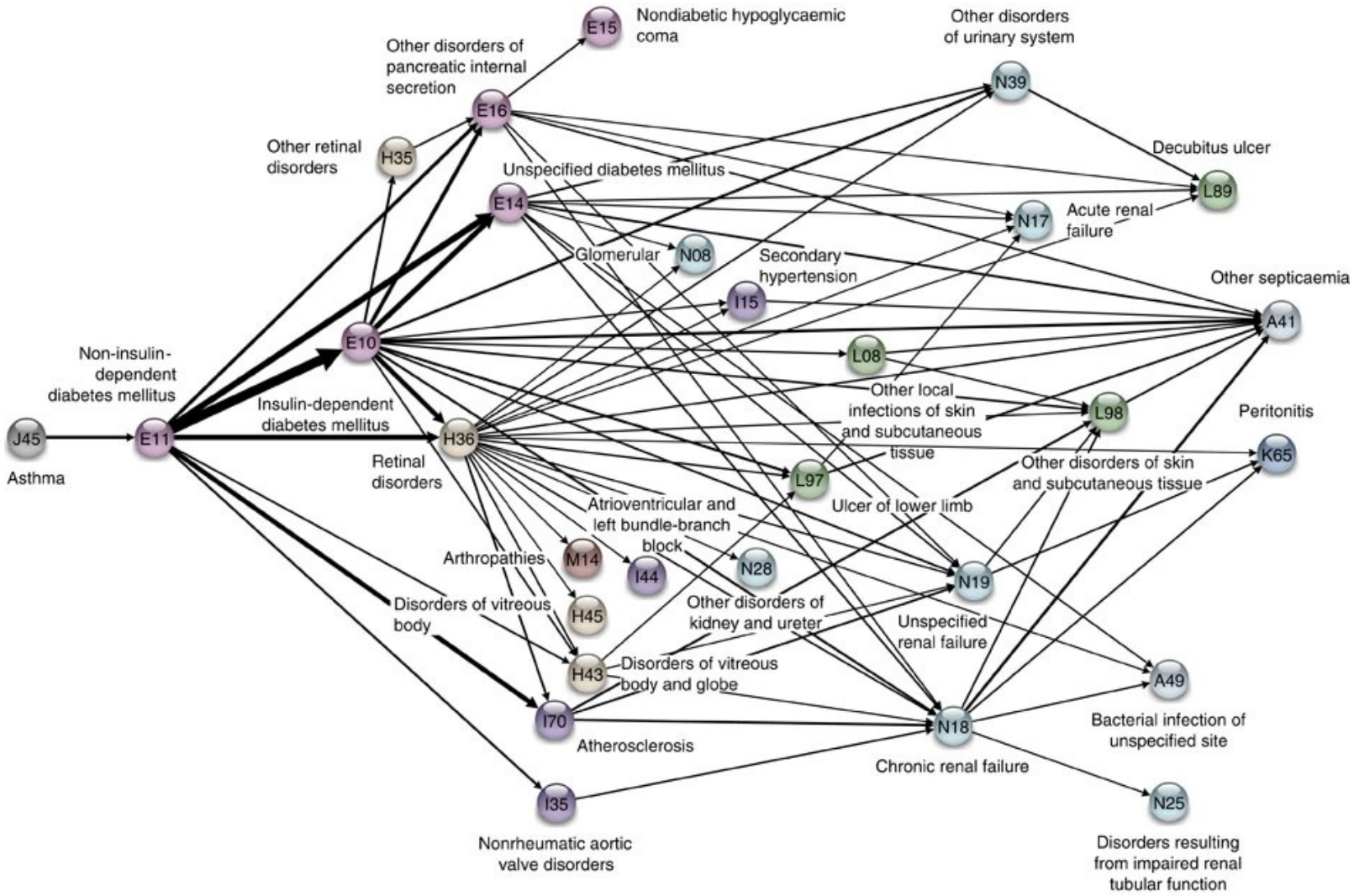
Early diagnoses for these central factors can help reduce the risk of adverse outcome linked to future diagnoses of other conditions.

Among the specific results, the following “surprising” insights were found:

- a diagnosis of anemia is typically followed months later by the **discovery of colon cancer**
- a diagnosis of gout was identified as **a step on the path** toward eventual diagnosis of cardiovascular diseases
- COPD is **under-diagnosed** and **under-treated**

Case Study

Danish Medical Data



Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Suggested Reading

Association Rules Overview

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Association Rules Mining

- Overview
- Case Study: Danish Medical Data

Exercises

Association Rules Overview

1. Of what types of machine learning tasks are the following problems representative?
   - Identifying risk factors associated to breast/prostate cancer.
   - Predicting whether a patient will have a second, fatal heart attack within 30 days of the first on the basis of demographics, diet, clinical measurements, etc.
   - Establishing the relationship between salary and demographic information in population survey data.
   - Predicting the yearly inflation rate using various indicators.

2. What are some examples of supervised, unsupervised, semi-supervised, reinforcement machine learning tasks in the business world? In a public policy/government setting? In a scientific setting?