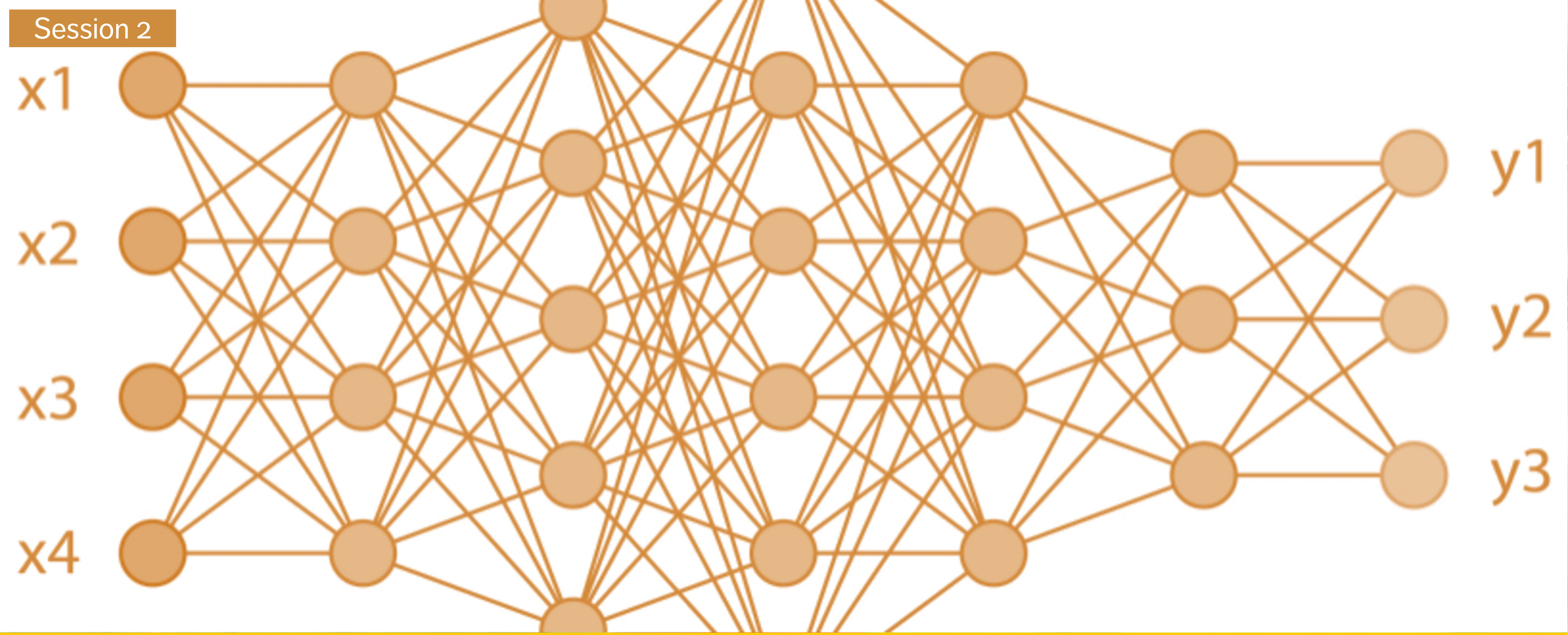
Session 2



4. Classification Overview

Overview

In **classification**, a sample set of data (the **training** set) is used to determine rules and patterns that divide the data into pre-determined groups, or classes (supervised learning; predictive analytics).

The training data usually consists of a **randomly** selected subset of the **labeled** (target) data.

Value estimation (regression) is similar to classification when the target variable is **numerical**.

Overview

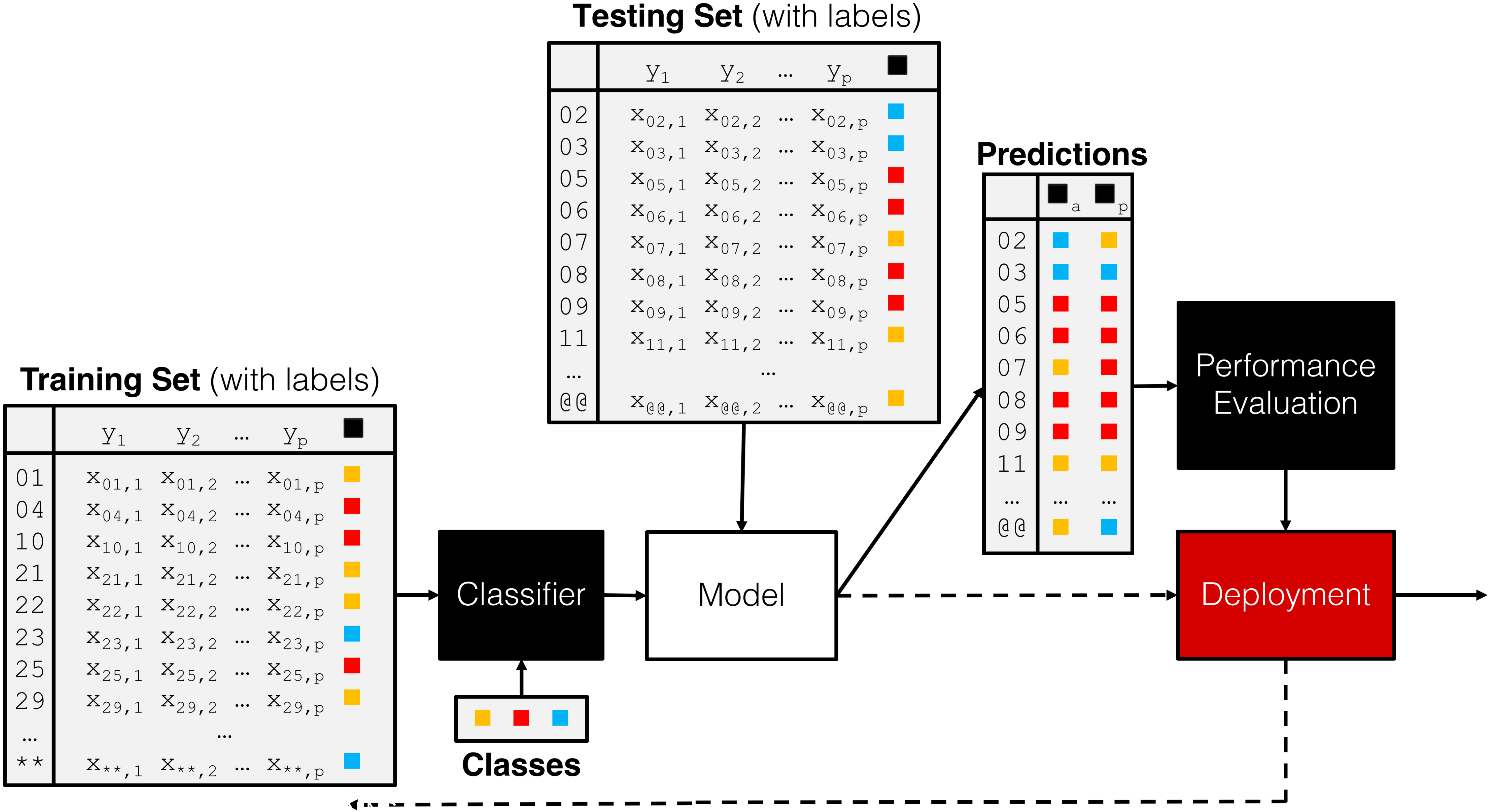
In the **testing** phase, the model is used to assign a class to observations for which the label is hidden, but ultimately known (the **testing** set).

The performance of a classification model is evaluated on the testing set, **never** on the training set. In the **absence** of testing data, classification may be **descriptive** but not predictive.

Technical issues include:

- selecting the features to include in the model
- selecting the algorithm
- etc.

Session 2



Training Set (with labels)

	y_1	y_2	…	y_p	■
01	$x_{01,1}$	$x_{01,2}$	…	$x_{01,p}$	🟨
04	$x_{04,1}$	$x_{04,2}$	…	$x_{04,p}$	🟥
10	$x_{10,1}$	$x_{10,2}$	…	$x_{10,p}$	🟥
21	$x_{21,1}$	$x_{21,2}$	…	$x_{21,p}$	🟨
22	$x_{22,1}$	$x_{22,2}$	…	$x_{22,p}$	🟨
23	$x_{23,1}$	$x_{23,2}$	…	$x_{23,p}$	🟦
25	$x_{25,1}$	$x_{25,2}$	…	$x_{25,p}$	🟥
29	$x_{29,1}$	$x_{29,2}$	…	$x_{29,p}$	🟨
…		…			
**	$x_{**,1}$	$x_{**,2}$	…	$x_{**,p}$	🟦

🟨 🟥 🟦

Classes

Testing Set (with labels)

	y_1	y_2	…	y_p	■
02	$x_{02,1}$	$x_{02,2}$	…	$x_{02,p}$	🟦
03	$x_{03,1}$	$x_{03,2}$	…	$x_{03,p}$	🟦
05	$x_{05,1}$	$x_{05,2}$	…	$x_{05,p}$	🟥
06	$x_{06,1}$	$x_{06,2}$	…	$x_{06,p}$	🟥
07	$x_{07,1}$	$x_{07,2}$	…	$x_{07,p}$	🟨
08	$x_{08,1}$	$x_{08,2}$	…	$x_{08,p}$	🟥
09	$x_{09,1}$	$x_{09,2}$	…	$x_{09,p}$	🟥
11	$x_{11,1}$	$x_{11,2}$	…	$x_{11,p}$	🟨
…			…		
@@	$x_{@@,1}$	$x_{@@,2}$	…	$x_{@@,p}$	🟨

Predictions

	■$_a$	■$_p$
02	🟦	🟨
03	🟦	🟦
05	🟥	🟥
06	🟥	🟥
07	🟨	🟥
08	🟥	🟥
09	🟥	🟥
11	🟨	🟨
…	…	…
@@	🟨	🟦

Applications

Medicine and Health Science

- predicting which patient is at risk of suffering a second, fatal heart attack within 30 days based on health factors (blood pressure, age, sinus problems, etc.)

Social Policies

- predicting the likelihood of requiring assisting housing in old age based on demographic information/survey answers

Marketing and Business

- predicting which customers are likely to switch to another cell phone company based on demographics and usage

Applications

Other uses include:

- Predicting that an object belongs to a particular class.
- Organizing and grouping instances into categories.
- Enhancing the detection of relevant objects

  avoidance: “this object is an incoming vehicle”

  pursuit: “this borrower is unlikely to default on her mortgage”

  degree: “this dog is 90% likely to live until it’s 7 years old”
- Predicting the inflation rate for the coming two years based on a number of economic indicators.

Examples

Scenario:

A motor insurance company has a fraud investigation dept. that studies up to 30% of all claims made, yet money is still getting lost on fraudulent claims.

Questions: can we predict

- whether a claim is likely to be fraudulent?
- whether a customer is likely to commit fraud in the near future?
- whether an application for a policy is likely to result in a fraudulent claim?
- the amount by which a claim will be reduced if it is fraudulent?

Examples

Scenario:

Customers who make a large number of calls to a mobile phone company's customer service number have been identified as churn risks. The company is interested in reducing said churn.

Questions: can we predict

- the overall lifetime value of a customer?
- which customers are more likely to churn in the near future?
- what retention offer a particular customer will best respond to?

Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue]
Real World Data Mining Applications, 2015

Objective

The U.S. Internal Revenue Service (IRS) estimated that there were large gaps between **revenue owed** and **revenue collected** for 2001 and for 2006.

Using DoR data, the authors sought to increase **efficiency** in the audit selection process and to **reduce the gap** between revenue owed and revenue collected.

Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Methodology

1. **data selection and separation:** experts selected several hundred cases to audit and divided them into training, testing and validating sets
2. **classification modeling** using MultiBoosting, Naïve Bayes, C4.5 decision trees, multilayer perceptrons, support vector machines, etc.
3. **evaluation of all models** on the testing set – models performed poorly until the size of the business being audited was recognized to have an effect, leading to two separate tasks (large/small businesses).
4. **model selection/validation** compared the estimated accuracy between different classification model predictions and the actual field audits (MultiBoosting with Naïve Bayes was selected as the final model; suggesting improvements to increase audit efficiency).

Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

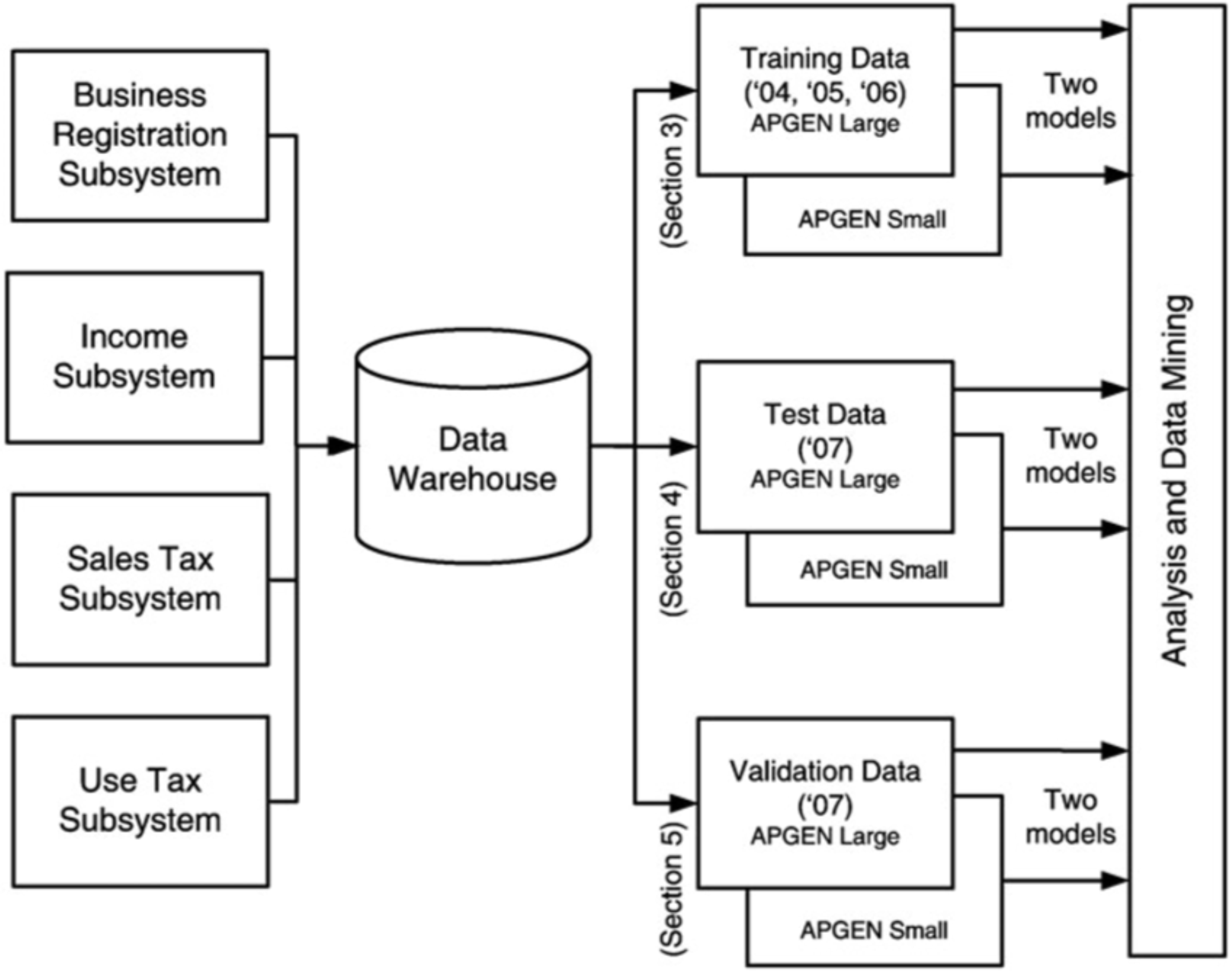
Data

Selected tax audit cases from 2004 to 2007, collected by the audit experts, which were split into training, testing and validation sets:

- the **training data** set consisted of *Audit Plan General* (APGEN) *Use Tax* audits and their results for the years 2004-2006
- the **testing data** consisted of APGEN Use Tax audits conducted in 2007 and was used to test or evaluate models (for Large and Smaller businesses) built on the training dataset
- while **validation** was assessed by actually conducting field audits on predictions made by models built on 2007 Use Tax return data processed in 2008.

Case Study

Minnesota Tax Audits



Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue](#)
Real World Data Mining Applications, 2015

Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue](#)
Real World Data Mining Applications, 2015

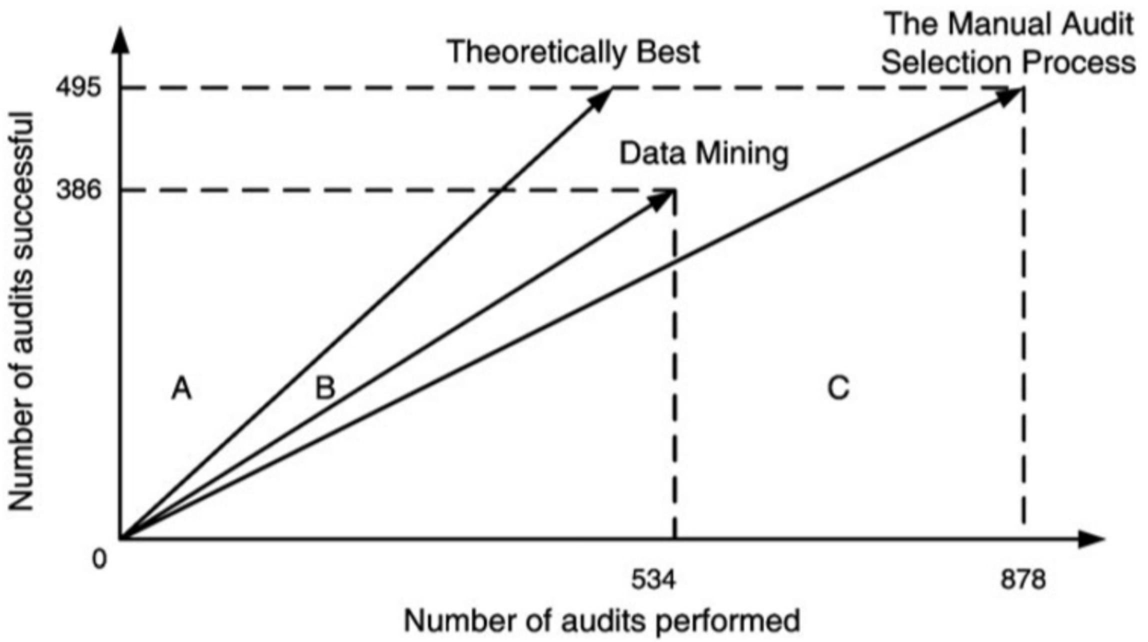
Strengths and Limitations of the Algorithms

- Naïve Bayes classification assumes independence of the features, which rarely occurs in real-world situations. This approach is also known to potentially introduce bias to classification schemes. In spite of this, classification models built using it have a successfully track record.
- MultiBoosting is an **ensemble technique** that uses committee (i.e. groups of classification models) and "group wisdom" to make predictions; unlike other ensemble techniques, it is different from other ensemble techniques in the sense that it forms a committee of sub-committees, which has a tendency to reduce both bias and variance of predictions.

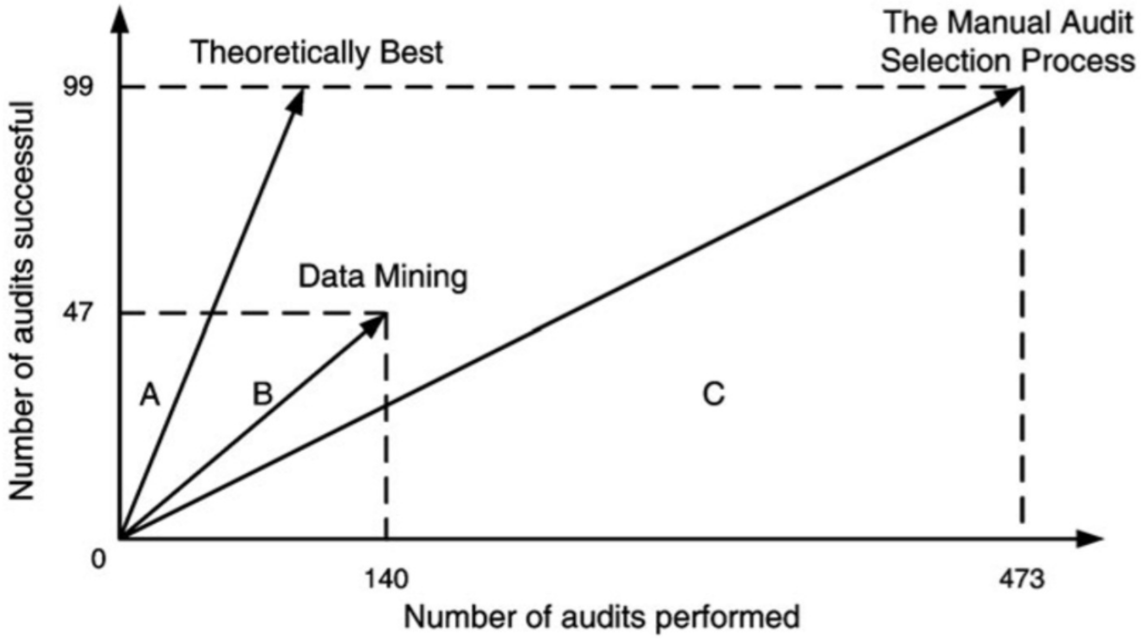
Case Study

Minnesota Tax Audits

APGEN
Large



APGEN
Small



Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue](#)
Real World Data Mining Applications, 2015

Take-Aways

- Many models were churned out before the team made a final selection.
- Past performance of a model family in a previous project can guide the selection, but remember the *No Free Lunch (NFL) Theorem*: nothing works best all the time!
- The feature selection process could very well require a number of visits to domain experts before the feature set yields promising results.
- Data analysis teams should seek out individuals with a good understand of both data and context.
- Domain-specific knowledge has to be integrated in the model in order to beat random classifiers, on average.
- Even slight improvements over the current approach can find a useful place in an organization – data science is not solely about Big Data and disruption!

General Classification Comments

Classification is linked to **probability estimation**

- approaches based on regression models could prove fruitful

Rare occurrences (often more interesting or important):

- historical data at Fukushima's nuclear reactor prior to the meltdown could not have been used to learn about meltdowns, for instance
- predicting no meltdown will yield correct predictions roughly 99.99% of the time, but will miss the point of the exercise

No Free-Lunch Theorem: no classifier works best for all data.

With big datasets, algorithms must also consider efficiency.

Suggested Reading

Classification Overview

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Classification and Value Estimation

- Overview
- Case Study: Minnesota Tax Audit

Spotlight on Classification

*Overview (advanced)

- Formalism

Exercises

Classification Overview

1. How would you use standard statistical modeling techniques to answer the questions presented in the two scenarios in the slides?
2. Identify scenarios and questions that could use classification and/or value estimation in your every day work activities.