Welcome to MBAN 5520
Statistics and Predictive Modeling

Today

1. Who am I? Who are you?
2. What is MBAN 5520
3. How is it Delivered? How is evaluation done?
4. Introduce a common **process** for data mining – how to start and what are the steps?

Who am I?

- B.A. (Statistics) UNB, M.Math (Statistics), Ph.D. (Management Sciences) Waterloo
- Joined Saint Mary's in 1979
- Dean of Sobey School (1991-1993, 1996-2002)
- Associate Vice President (Enrolment) and Registrar (2005-2017)
- Other hats I've worn (department chairperson, associate dean, MBA director, director of recruitment, director of the language centre)

I am passionate about numbers and data.

I like decision-making that is grounded in data.

I like exploring data for patterns.

I am interested in how people understand and interpret numeric information and the ways different people are challenged in processing numeric information.

We can’t be effective in giving advice if others do not understand us.

Who are you?

- Where are you from?
  - Send me a picture from you neighbourhood/city/favorite place

- I sent you an email with a variety of questions to understand who you are. Many responded, but I would like to hear from all of you.

- Pictures and unique comments help me remember and understand you.

- We will be looking at the diversity in data. More diversity makes it harder to predict some things (outcomes) but also aids us in differentiating cases. Diversity in the class creates challenges in teaching, but allows me to learn more.

Numeracy

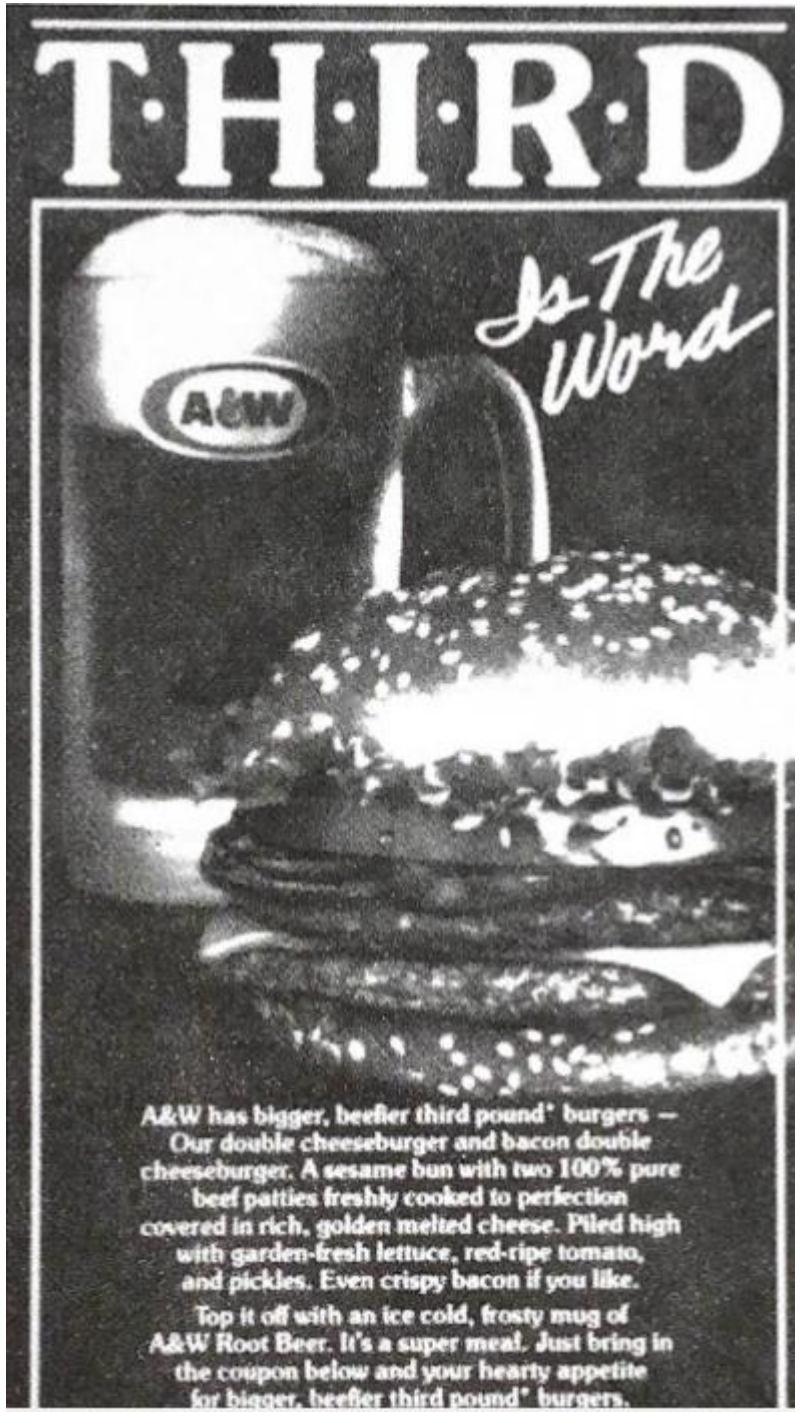
- Do we really understand numbers?
- On 6 August 2021, it was reported that 76.6% of NS population was vaccinated, but the same source also reported that only 66.2% were vaccinated.
- Other sources reported that 87% were vaccinated and another said 75.2% vaccinated.
- All figures used the same Nova Scotia Department of Health data released on that day.
- So, what was the vaccination rate? Why are the numbers different?

Definition	Vaccination Rate
All Nova Scotians – 2 doses	66.2%
Eligible Nova Scotians – 2 doses	75.2%
All Nova Scotians – 1 or 2 doses	76.6%
Eligible Nova Scotians – 1 or 2 doses	87%

Numeric communication is frequently done poorly. Data definitions are important.

People don't have the patience to listen/read definitions.

You hear what you want to hear.



- A&W was the largest drive-in fast food restaurant before McDonalds.
- In 1982, McDonalds introduced the very successful Quarter Pounder.
- Soon after, A&W introduced **Third is the Word** – a 1/3 burger.
- It was not successful. They tried in 2007 with the 1/3 Angus and later the 1/3 Sirloin. No luck.
- Why?

When potential customers were asked why they bought ¼ pounder rather than the 1/3 burger, the most common reply was

“Why would I buy 1/3 of a pound when I can get ¼ pound for the same price?”

- How numbers are communicated matters a lot!
- What if McD called it a 4oz burger and A&W called theirs a 5oz burger?

- The way numbers are presented affects the way they are interpreted.
- These differences in interpretation vary markedly across the broader population.
- There is no “best” way to communicate, but some ways are better/worse than others.
- Data analytics is all about communication – making the data talk to you AND how to communicate your findings to others.

What is data analytics (data mining)?

The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases.

At the end of my 6 weeks with you, you will

- Have a basic understanding of the process of data mining.
- Be aware of some of the most common applications of data mining.
- Be able to effectively use Excel for simple data exploration and analysis.
- Be able to use Python for similar tasks and some advanced models.
- Be aware of challenges associated with data quality.
- Have a basic understanding of probability and variability and their role in decision-making using data.
- Understand some of the complexity in measuring the “performance” of predictive models and errors in decision-making.
- Understand classical statistical hypothesis testing methodology and the common errors and misunderstandings in using it.

Zoom

- All lectures in this course will be broadcast using **Zoom**, even though classes are face-to-face.

- What this means, is that if you cannot come to campus, you can still attend class via the Zoom link. But Normally I expect you to be here in person.

- All Zoom sessions will be **recorded** and a link to the recording will be posted by the end of each day.

Brightspace

- All course materials and activities will be on Brightspace.
- There will be a folder for each Week, including readings, slides and recordings
- Caution: I have limited experience with Brightspace for anything other than posting files. Be patient with me.

Assessment

- Grades can be useful to foster learning and to evaluate subject mastery.
- But they are not great at doing both at once!
- I am more interested in learning and have limited confidence that we are very good at assessing mastery.
- BUT, I have a problem I need help with. We have stated 3 learning objectives for this course. For AACSB accreditation, we have to demonstrate assessment of our success in achieving these “learning objectives”.
- I have no plans on incorporating these into my grades, but I welcome your ideas on how to do the evaluation for AACSB.

Textbook and Quizzes

- There is no required “textbook”, but I will be posting my own text material.
- I edit the “text” each year and hope to stay ahead of the class!

Quizzes

- Associated with each “week”, there will be 10 multiple choice questions on the assigned “chapter”.
- The quiz will occur near the start of class and be done in-person.
- The only purpose of the quiz is to get you to read before class.

- **Assignments** – there will be 5 assignments worth 8 points each.
- Assignments will be posted each Tuesday and due Saturday. I will try to keep them short.
- Several assignments will use Excel. If Python is required for an assignment, you can use AI to do the coding. "How to" issues with Excel will be covered in class as well as short "How to" videos.

- **An example**

[Target Knows You Better Than You Do!]

How Target Knows You Better Than You Do

How companies know your secrets - a true story:

- Andrew Pole had just started working for Target department stores in 2002, when 2 colleagues from Marketing stopped by his desk and asked

“If we wanted to figure out if a customer is pregnant, even if she didn’t want us to know, can you do that?”

“Can you give us a list?” the marketers asked.

“We knew that if we could identify them in their second trimester, there’s a good chance we could capture them for years. As soon as we get them buying diapers from us, they’re going to start buying everything else too.

- Most of us have shopping habits that we change infrequently.
- Target shoppers only buy certain things from Target, even though Target sells almost everything they might need.
- Making them change their shopping habits is extremely hard to do.
- Starting a family is one of the life moments where everything changes – a marketing opportunity.

- We disclose a lot of personal information every day without thinking about it – social media, loyalty cards, any web activity, … we leave digital finger prints.

Whenever possible, Target assigns each shopper a unique code — known internally as the Guest ID number — that keeps tabs on everything they buy.

“If you use a credit card or a coupon, or fill out a survey, or mail in a refund, or call the customer help line, or open an e-mail we’ve sent you or visit our Web site, we’ll record it and link it to your Guest ID,”

Also linked to your Guest ID is demographic information like your age, whether you are married and have kids, which part of town you live in, how long it takes you to drive to the store, your estimated salary, whether you’ve moved recently, what credit cards you carry in your wallet and what Web sites you visit.

Target can buy data about your ethnicity, job history, the magazines you read, if you’ve ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of coffee, paper towels, cereal or applesauce, your political leanings, reading habits, charitable giving and the number of cars you own.

- At Target, some expectant mothers sign up for the store's baby registry.
- Target has every sales transaction of that customer for years before the pregnancy and all the days of the pregnancy until delivery.
- How does shopping behavior change over the 9 months?

- After 3 months, pregnant women start buying larger quantities of unscented body lotion.
- During the first 20 weeks, pregnant women load up on supplements such as calcium, magnesium and zinc.
- Closer to the delivery date, they start buying more scent-free soap, hand sanitizer, washcloths, cotton balls.

- Pole found 25 products that when analyzed could be used to create a pregnancy prediction score.
- They could also estimate the due date quite accurately.
- They were able to identify tens of thousands of customers who were likely pregnant.

- How would you use this data?
- They also know all your other shopping habits.
- They know what type of marketing promotion you respond to best – email, coupons, discounts, events,…

CREEPY!!!

- Take a fictional Target shopper named Jenny Ward, who is 23, lives in Atlanta and in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug.
- There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August.
- What's more, because of the data attached to her Guest ID number, Target knows how to trigger Jenny's habits.
- They know that if she receives a coupon via e-mail, it will most likely cue her to buy online.
- They know that if she receives an ad in the mail on Friday, she frequently uses it on a weekend trip to the store.
- And they know that if they reward her with a printed receipt that entitles her to a free cup of Starbucks coffee, she'll use it when she comes back again.

- About a year after Pole created his pregnancy-prediction model, a man walked into a Target outside Minneapolis and demanded to see the manager.
- He was clutching coupons that had been sent to his daughter, and he was angry, according to an employee who participated in the conversation.
- “My daughter got this in the mail!” he said.
- “She’s still in high school, and you’re sending her coupons for baby clothes and cribs?
- Are you trying to encourage her to get pregnant?”

- The manager didn’t have any idea what the man was talking about.
- He looked at the mailer. Sure enough, it was addressed to the man’s daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants.

- The manager apologized and then called a few days later to apologize again.
- On the phone, though, the father was somewhat abashed.
- “I had a talk with my daughter,” he said.
- “It turns out there’s been some activities in my house I haven’t been completely aware of. She’s due in August. I owe you an apology.”

What is “Problem Solving”?

Everyone says we need to develop better problem solving skills, but what does this mean?

When we are solving a problem or performing a new task for the first time, where do you start?

First Steps in Problem Solving

What questions should we ask in a data mining problem?

- Define the task – make sure you fully understand what you are supposed to do/solve.
- Define a “process” for solving.

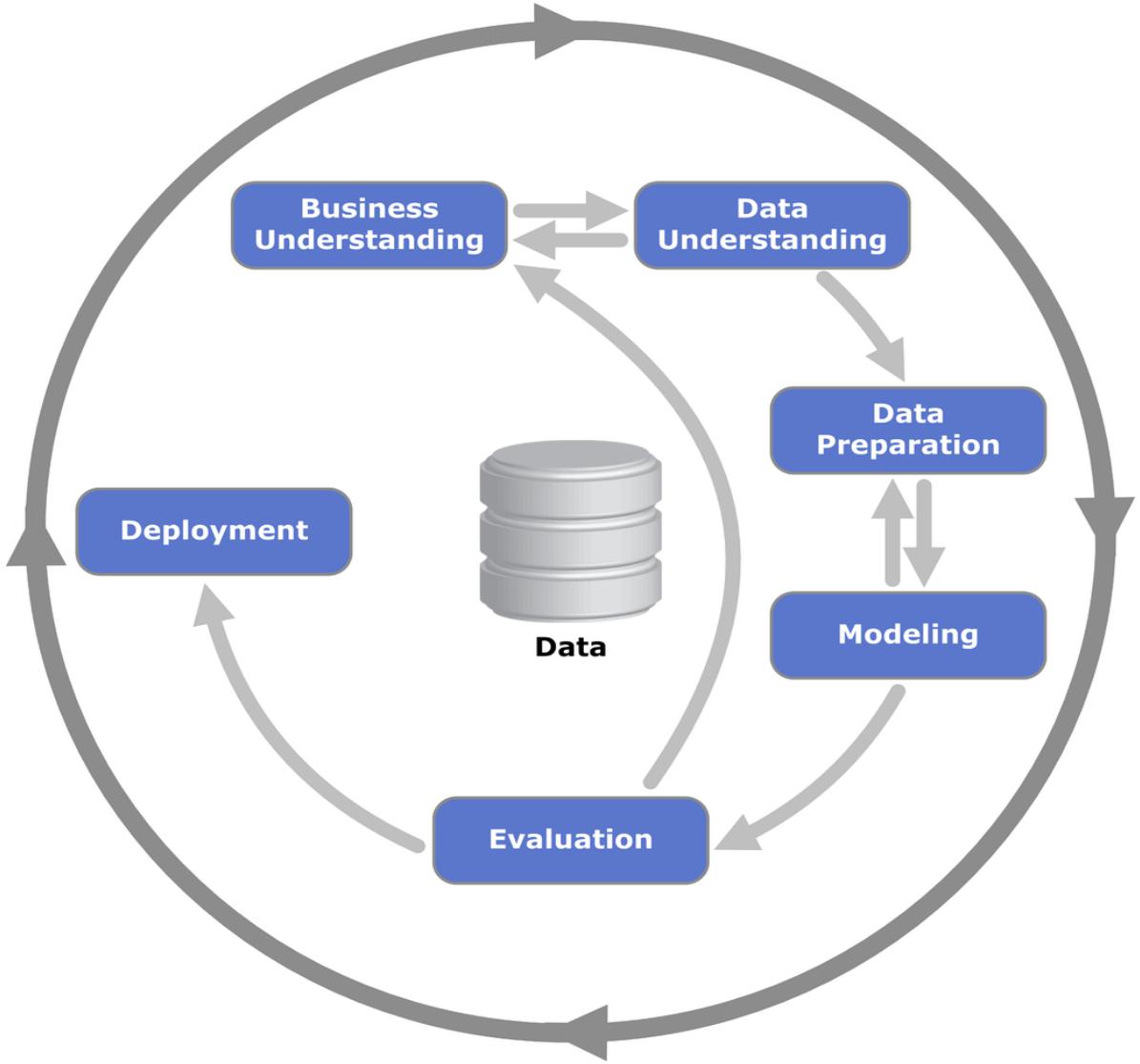
- Have you ever seen anything like this before?
- Experience is valuable. Ask an expert?
- Where should a novice begin?

- These are all important and necessary steps, but they leave out many aspects of problem solving (in general) and problem solving using data, in particular.

- There are several established frameworks for doing data analytics projects.

- We will look at one, CRISP-DM, and see how it fits with the Target story and how it connects to the strategy in the AI reading.

CRISP-DM – Cross Industry Standard Process for Data Mining



CRISP-DM – Cross Industry Standard Process for Data Mining

- It is a circular problem solving process (most frameworks are circular).
- It recognizes that you may have to backtrack.
- Sometimes there is very frequent backtracking between early stages.
- It breaks the process into 6 stages (phases).
- Stages are not independent – what you do at one stage is influenced by others – look ahead and look backwards.
- Some analysts will combine activities between adjacent stages (overlapping pairs).

Business Understanding

- What is our objective?
- What does success look like? Can we measure it?
- Focus on outcomes, not activity
- Who are the actors? Who makes the decisions? Who implements them?
- Do they share values/objectives or is there conflict?
- Do they understand/appreciate data? Or only when it supports their beliefs?

In business understanding you are trying to define the problem/opportunity from the perspective of the business. The most frequent cause of project failure is solving the wrong problem.

You are also trying to rephrase the **business problem** as a

- **I**f I know XXX then I can do YYY.
- Can the data tell me XXX?

We must translate the business objective(s) into a set of questions that data mining can answer.

Target DM questions

With the story about Target, what would be appropriate data mining questions?

A	How do I increase sales?
B	How do I change customer behaviour?
C	Which of my customers are pregnant?
D	All of the above.

Target:

- If I know which customers are pregnant and when the baby is due, I can send them appropriate promotions to better engage them as customers.
- Can you determine which customers are pregnant and when the baby is due?

Data Understanding

- What data do I have?
- What data do I need?
- How do I get it?
- What does the data look like?
- What is the quality of the data?

Target:

- We have a wealth of data that we have captured about customers and we can buy more data.
- But with respect to pregnancy, we have data about women that have signed up for our baby shower registry.
- We know when the baby is due and we have detailed purchase records (date of purchase, items bought, which Target store, cash or credit,….).
- We can construct a purchasing profile before pregnancy and see how the profile changes through the pregnancy.

Data Preparation

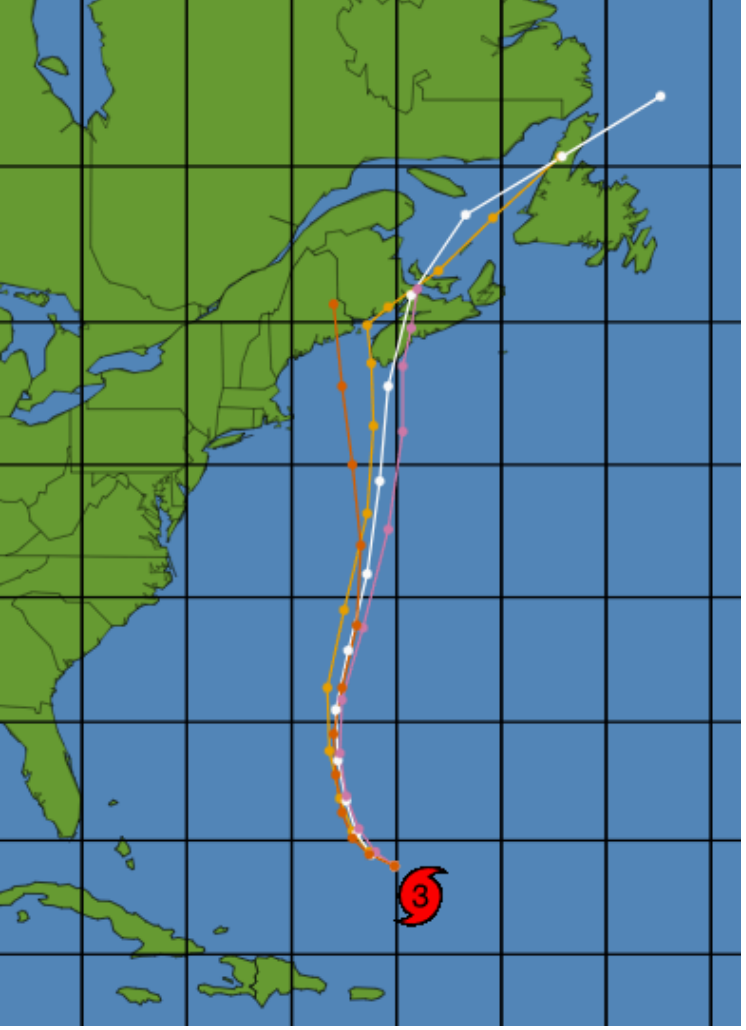
- Select and extract the data from the database(s)
- Clean the data for outliers
- Transform, recode, relabel, construct data
- Filter and sort data

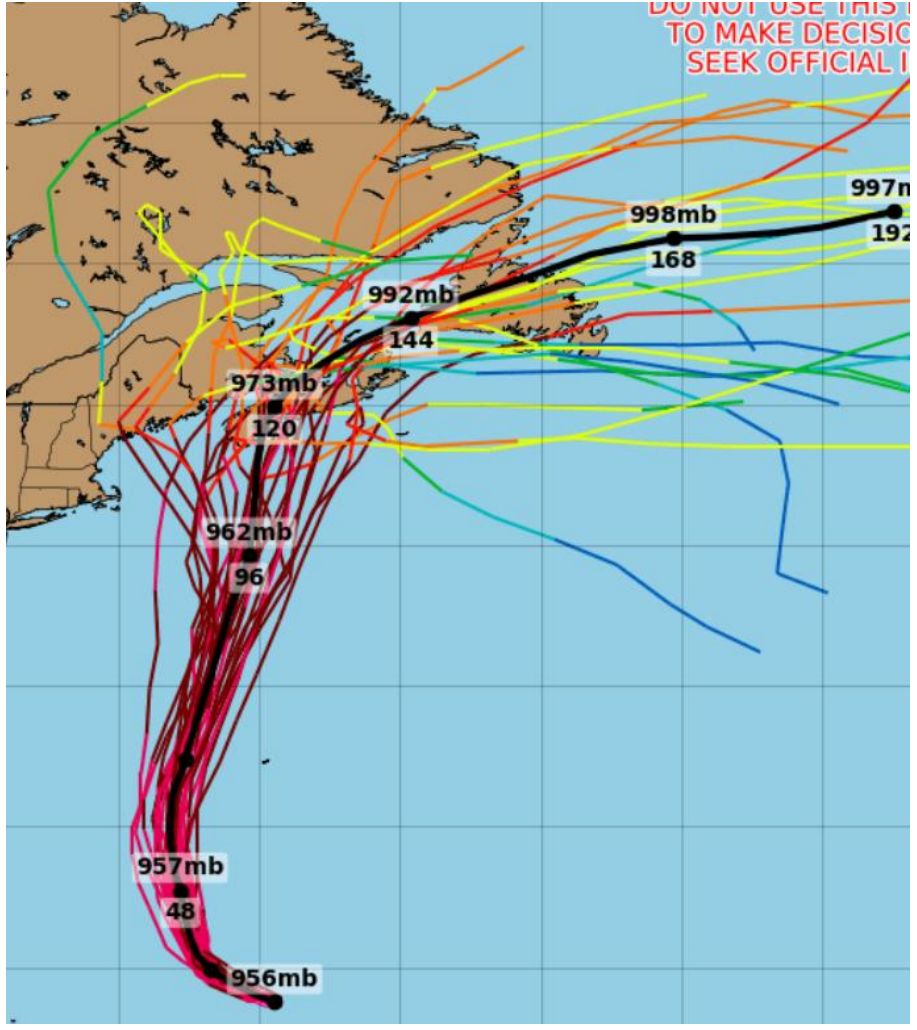
Target – data preparation:

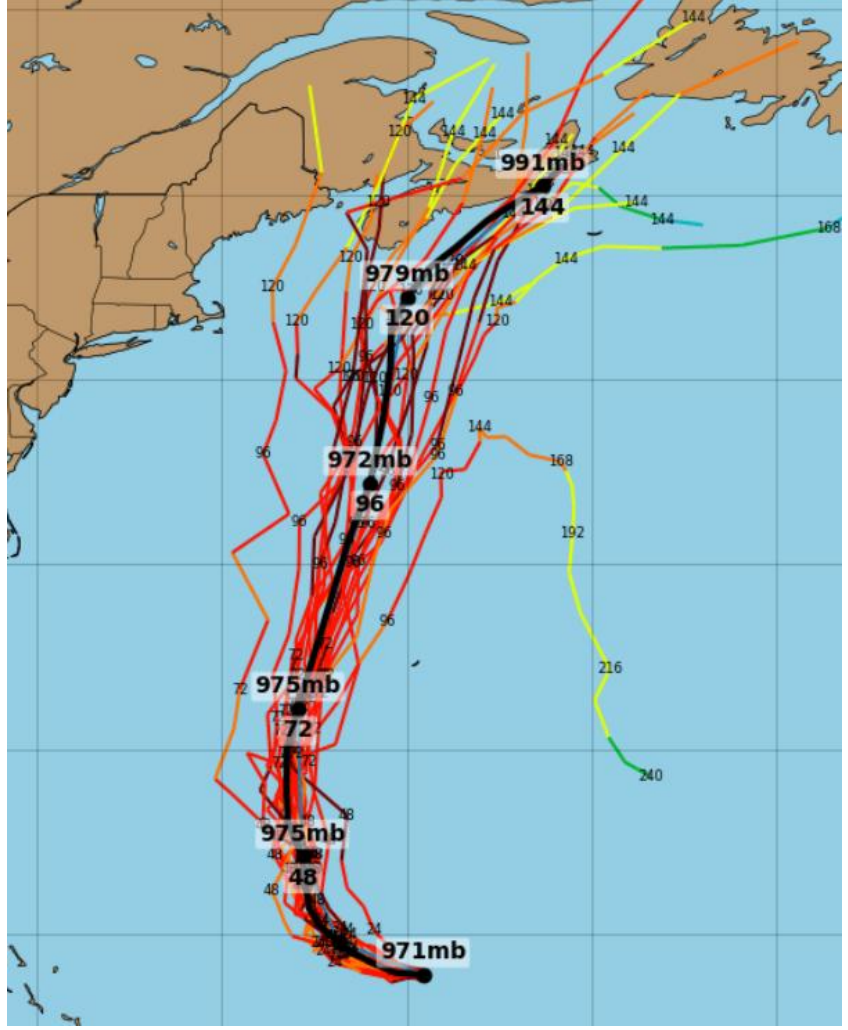
- Need to organize the purchase transactions into a useable form.
- May need to group transactions by product category (e.g., body lotions),
- and summarize transactions (e.g., average number of purchases of lotions per month, or average volume of body lotion purchased per month).

Modelling

- What is a model?
- The output of the model may be
  - a formula for scoring likelihood of being in a category (class);
  - a decision rule for assigning the customer to a category (if XXX then YYY);
  - a sample of customers that “look like” the customer and you use the behavior of the sample to predict what this customer will do.







All models are wrong, but some are useful – George Box

- Does problem fit the characteristics of any of the common data mining applications?
- May need to use multiple applications – e.g, use feature selection to find most important variables and then build a classification model.

- There are many different approaches and new ones are being created all the time.
- How do you calibrate your model (determine coefficients for a formula)?
- If there are multiple possible models, which should you choose?

Target:

- "… able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score.
- More importantly, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy." -
- Classification model identifies whether pregnant and value estimation model predicts the due date.

Evaluation

- All models make mistakes.
- Predictions are hard to do “accurately”.
- Predictions are not exact.
- Some cases will be misclassified.

- What are appropriate performance measures?
- What are the consequences of errors?
- What is “success”?

Performance measures

Which of the following are measures of performance for Target's model to predict whether a customer is pregnant?

A	Sales increased 8%
B	27,866 customers were predicted to be pregnant.
C	Model correctly identified 85% of those in the baby registry as being pregnant.
D	Among those predicted to be pregnant, only 38% were in the baby registry.

Target:

- "... Take a fictional Target shopper named Jenny Ward,
  - who is 23,
  - lives in Atlanta and
  - in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug.
- There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August."

Deployment

- Does the model answer the question?
- Does the user understand the model? Will the user accept the output?
- How do you maintain the model?
- Are there ethical issues with respect to implementation?
- Were there unintended consequences?
- Should we rephrase the business problem and data problem and repeat the process?

Target:

- What's more, because of the data attached to her Guest ID number, Target knows how to trigger Jenny's habits.
- They know that if she receives a coupon via e-mail, it will most likely cue her to buy online.
- They know that if she receives an ad in the mail on Friday, she frequently uses it on a weekend trip to the store.
- And they know that if they reward her with a printed receipt that entitles her to a free cup of Starbucks coffee, she'll use it when she comes back again.

- Target used additional models to determine the best ways to get the customers to buy more.
- The model identified “tens of thousands” of quality prospects that Target would not have been able to identify otherwise. So yes, it addressed the business problem/opportunity.
- But during deployment they uncovered ethical issues (privacy) that had not been recognized at the business understanding stage.
- Success gave credibility to the data analysis group and appears to have stimulated the search for more opportunities.

Summary

- Data mining is a **PROCESS** not a tool or formula.
- It starts with identifying what the business issue is and how data may be able to inform the decision.
- Data can be messy.
- It is useful to have mental pictures of different types of approaches to modelling to guide you (common DM applications).
- You are not finished until the business says you are successful.

Perks Café - CRISP Exercise

Background Story

Perks Café is a popular local coffee shop that has launched a loyalty card program. Every time customers make a purchase, they scan their card. The café's owners want to increase revenue, but they aren't sure what their next step should be. They believe data analytics could help.

Your task today is to think like consultants and apply the CRISP-DM framework. Since we only have 15 minutes, we will focus on the first two phases:

- Business Understanding
- Data Understanding

Part A: Business Understanding

Discuss in your group:

1. What is the real business problem that Perks Café faces?
2. How should the café define success?
   - Broad goal and then more specific and measurable

Part B: Data Understanding

Now think about the data.

1. What types of customer or transaction data should the café collect to answer their questions?
2. What problems or limitations might they face with this data?

Deliverable

Be prepared to share **one business objective** and **one data issue** from your group with the class.

Part B: Data Understanding

Now think about the data.

1. What types of customer or transaction data should the café collect to answer their questions?
   - Purchase frequency
   - Basket size (how many items per transaction)
   - Types of products bought (coffee, muffins, sandwiches, etc.)
   - Time/day of purchases
   - Loyalty program usage (who uses it, how often)
2. What problems or limitations might they face with this data?
   - Missing or incorrect loyalty scans
   - Customers sharing loyalty cards
   - Data that is too old to be useful
   - Inconsistent or incomplete records

Deliverable

Be prepared to share **one business objective** and **one data issue** from your group with the class.

Modeling

- Why is it valuable to know common data mining models?
    - Have I seen a problem like this before?

- What is your role?
    - What are things that a data analyst can do for you?

- As a data analyst, where do you start?
    - Do I recognize this problem?
    - Is there a best practice for analyzing this problem?
    - What challenges should we look out for?

Descriptive and Predictive Applications

- Data mining applications are usually grouped into descriptive and predictive applications.

Descriptive Applications are ones in which you may be summarizing or finding patterns, but without these descriptions being *supervised*. We are not directing our approach to analyzing the data. We call this **Unsupervised Learning**.

I find the "unsupervised learning" language confusing.

- In **Predictive Applications** we are trying to predict an outcome based upon information we have at hand.
- We are building a formula, where several variables are used to predict another.
- So the search for the “formula” is being “supervised”, in the sense that we are looking for patterns that predict a specific outcome.
- We call this **Supervised Learning**.

Why do we call it “learning”?

- Once we have completed our analysis, we have created a methodology that a computer can then use to repeat this work on a new data set. And with each new set of data, the methodology (machine) should get better at doing it.
- We are teaching a machine (computer) to learn – **Machine Learning**.

- And this learning will lead to **Artificial Intelligence**.
- Video “Machine Learning in 5 Levels of Difficulty” posted to Top Hat. Great intro.

Descriptive – Unsupervised Learning

- **Clustering** - grouping cases (people, transactions, …) such that within the group, cases are **similar**, but the groups are quite different from one another.
- **Co-occurrence grouping** – grouping cases where two (or more) cases usually occur together (e.g., products that are often purchased together)
- **Similarity matching** – finding cases that are "**similar**" to one that you are examining.
- **Profiling** – in contrast to Clustering where we let the data form the cluster, in Profiling, we define a group and then try to describe common characteristics (what do our best customers look like?)

- **Data reduction** – can we compress our data set into something more manageable without losing valuable knowledge?
- **Feature selection** – with so many variables, can we combine some to form new variables that contain the most important information? Is there one subset of variables that contains the most important information and other variables are redundant?

- E.g., looking at a transcript to understand student ability is difficult. Different courses, different grades, different profs, too much information.
- Can I get better insight by reducing the data to overall GPA, or GPA each year?

- **Causal analysis** – If we find a relationship among the variables, can we explain why A affects B?

- Some predictive models are very accurate (e.g., predicting who will cancel their subscription), but we can't explain why they work. If we want to change an outcome (get a customer to stay with us rather than switch), then we also need a causal model to tell us what to do.
- If we use a model to decide who gets approved for a loan or who doesn't, we may need to defend our decision and explain "why" the model rejected a client.

When facing a new problem, getting started is often easier if you can fit your problem into a common category

Target: "Which of our customers is pregnant"

This sounds like a classification problem. So how do you start a classification problem?

Summary

- Data mining is a **PROCESS** not a tool or formula.
- It starts with identifying what the business issue is and how data may be able to inform the decision.
- Data can be messy.
- It is useful to have mental pictures of different types of approaches to modelling to guide you (common DM applications).
- You are not finished until the business says you are successful.

What is next?

- Business Understanding is context specific so we will jump to Data Understanding next.

- What does data look like?
- Where do we get it?
- What problems should we be aware of?