Applied Optimization for Wireless, Machine Learning, Big Data

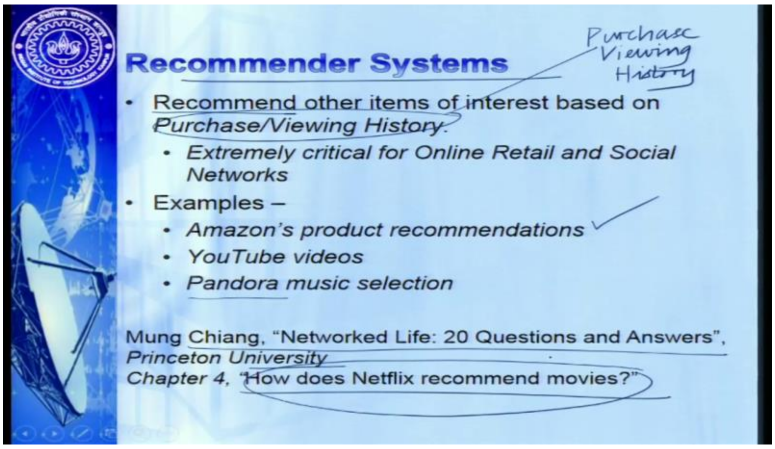
Prof. Aditya K. Jagannatham
Department of Electrical Engineering
Indian Institute of Technology, Kanpur

Lecture - 77
Introduction to Big Data: Online Recommender System (Netflix)

***Keywords**: Big Data, Online Recommender System*

Hello, welcome to another module in this massive open online course. So in this module let us start looking at yet another innovative application of optimization and this is in the field of Big Data and in fact, Big Data is something that has gathered significant amount of attention of late because of the tremendous rise in the amount of data that is being generated each day in various websites or various online services that are there. Big Data has several applications and we are going to look at one very specific and very relevant technique for Big Data known as a Recommender System.

(Refer Slide Time: 01:39)



Recommender Systems

- Recommend other items of interest based on *Purchase/Viewing History*
  - *Extremely critical for Online Retail and Social Networks*
- Examples –
  - *Amazon's product recommendations*
  - *YouTube videos*
  - *Pandora music selection*

Mung Chiang, "Networked Life: 20 Questions and Answers", *Princeton University*
Chapter 4, "How does Netflix recommend movies?"

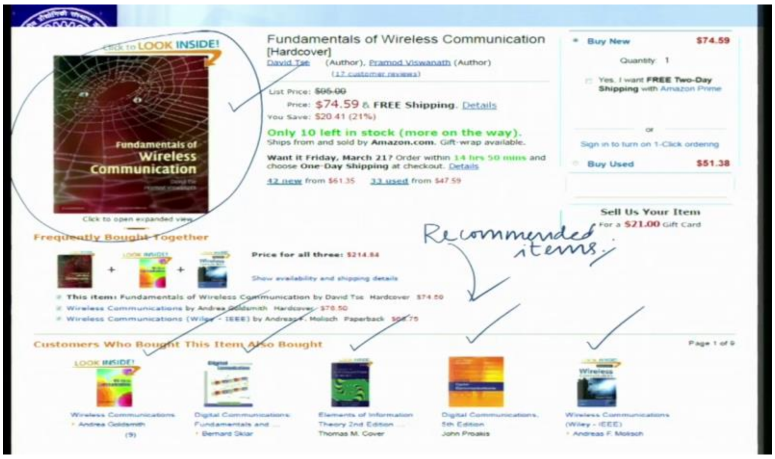
Purchase
Viewing
History

A Recommender System is something as the name implies, it recommends other items based on your purchase or viewing history. For instance, if you go to any E-commerce websites, you have several product recommendations, based on your viewing history of the items that you have browsed or based on even your past purchase history or even when you go to a video streaming site like YouTube, when you look at the different videos or when you are watching current video, the website automatically comes up with a recommendation of other videos that you might find a lot of similarity or that you will be highly interested in.

And similarly music websites like Pandora for instance, which is a music website, it comes up with a set of music videos or music albums or songs that would be of very high interest to you and some of these you might not be aware of. So by coming up with this highly specific set of recommendations, it is a win-win situation because you cannot browse the infinite number of products on an E-commerce website and similarly, it is also beneficial for the website because enticing the customers to this possible set of objects that the customer is interested in, are increasing their business. So it is a win-win situation for everyone, it saves your time, increases the business of the website and so on and for this part this module will be referring in particular to this very interesting book by Professor Mung Chiang titled “Networked Life: 20 Questions and Answers” by Princeton, Princeton University and the chapter that we are talking about is Chapter 4, which is “How does Netflix recommend movies?”.

So all such systems which basically recommend various options for you to buy or browse are known as recommender systems. The more closely your recommendations match the interest of the consumer, the better your recommender system is then the better is going to be the efficiency of your website. So the idea is to design the best recommender system which comes up with a very specific and highly interesting set of recommendations.

(Refer Slide Time: 05:04)



Click to LOOK INSIDE!

Fundamentals of Wireless Communication
[Hardcover]
David Tse (Author), Pramod Viswanath (Author)
(17 customer reviews)

List Price: ~~$95.00~~
Price: $74.59 & FREE Shipping. Details
You Save: $20.41 (21%)

Only 10 left in stock (more on the way).
Ships from and sold by Amazon.com. Gift-wrap available.

Want it Friday, March 21? Order within 14 hrs 50 mins and choose One-Day Shipping at checkout. Details

42 new from $61.35 33 used from $47.59

Click to open expanded view

Buy New $74.59
Quantity: 1
Yes, I want FREE Two-Day Shipping with Amazon Prime
or
Sign in to turn on 1-Click ordering
Buy Used $51.38

Sell Us Your Item
For a $21.00 Gift Card

Frequently Bought Together

Price for all three: $214.84
Show availability and shipping details

- This item: Fundamentals of Wireless Communication by David Tse Hardcover $74.59
- Wireless Communications by Andrea Goldsmith Hardcover $78.50
- Wireless Communications (Wiley - IEEE) by Andreas F. Molisch Paperback $60.75

Customers Who Bought This Item Also Bought

Page 1 of 9

Wireless Communications
Andrea Goldsmith
(9)

Digital Communications: Fundamentals and ...
Bernard Sklar

Elements of Information Theory 2nd Edition ...
Thomas M. Cover

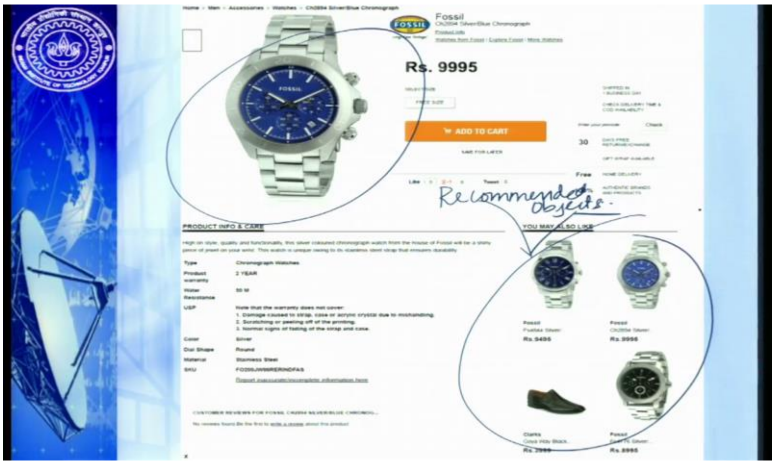
Digital Communications, 5th Edition
John Proakis

Wireless Communications (Wiley - IEEE)
Andreas F. Molisch

Recommended items.

For instance, is a simple snapshot from one of the E-commerce websites, you have a book that you are interested in buying. This is the book you are interested in and the website comes up with an alternative set of recommendations. So you look at these alternatives or set of recommendations. So the recommender systems look at patterns of different users, their purchase histories and come up with recommendations based on what other people who had similar interests have purchased or have viewed and so on.

(Refer Slide Time: 06:02)



Home > Men > Accessories > Watches > CH2894 Silver/Blue Chronograph

Fossil

CH2894 Silver/Blue Chronograph

Rs. 9995

ADD TO CART

30

Free

Recommended objects.

PRODUCT INFO & CARE

Type Chronograph Watches

Product warranty 2 YEAR

Water Resistance 50 M

USP Note that the warranty does not cover:
1. Damage caused to strap, case or acrylic crystal due to mishandling.
2. Scratching or peeling off of the printing.
3. Normal signs of fading of the strap and case.

Color Silver

Dial Shape Round

Material Stainless Steel

YOU MAY ALSO LIKE

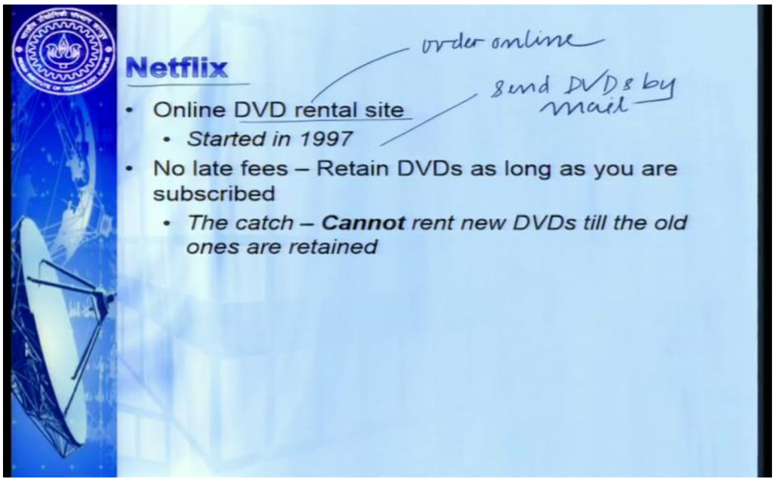
Fossil Rs.9495

Fossil Rs.9995

Clarks

Fossil Rs.8995

(Refer Slide Time: 06:22)



Netflix

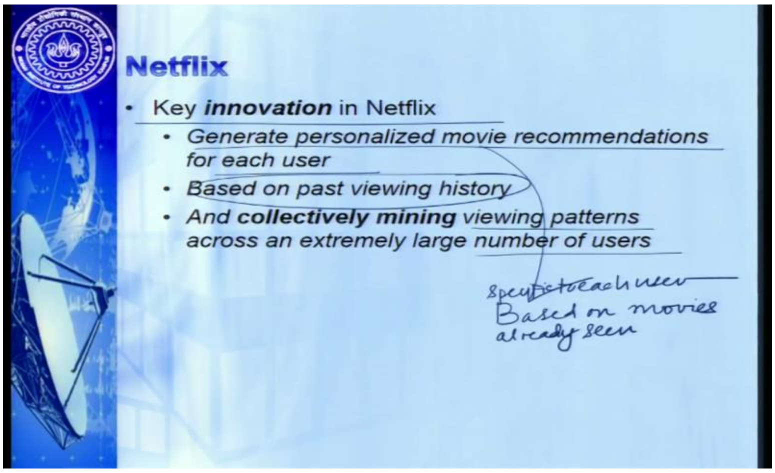
- Online DVD rental site
  - *Started in 1997*
- No late fees – Retain DVDs as long as you are subscribed
  - *The catch – **Cannot** rent new DVDs till the old ones are retained*

order online

send DVDs by mail

One particular interesting application that we are going to talk about is that of Netflix and it is an Online DVD rental site started in 1997. The model of Netflix is that you send DVDs by mail or regular post which you can order online and they will be sent by mail to you for a fee of cost. Now once you watch the DVDs, you send them back you get a new set of DVDs, specific to that particular website.

(Refer Slide Time: 07:18)



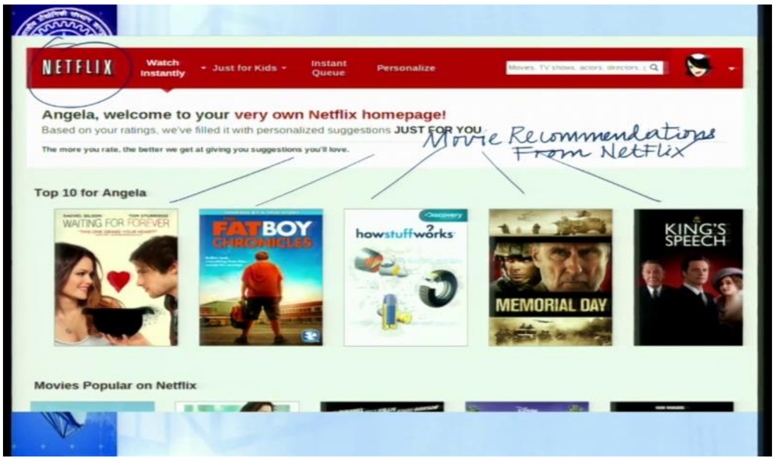
Netflix

- Key ***innovation*** in Netflix
  - *Generate personalized movie recommendations for each user*
  - *Based on past viewing history*
  - *And **collectively mining** viewing patterns across an extremely large number of users*

Specific to each user
Based on movies already seen

Now, the key innovation that we are talking about is basically the website generates personalized movie recommendations for each user, based on your past viewing history. So this collectively mines the viewing patterns of an extremely large number of users, that is all the movies you have seen and the movies that a large number of users have seen, to extract information and then based on this mining of this collective data of users and movies, you come up with the specific set of recommendations for a particular reason for that matter for each user. So it comes up with a set of recommendations or comes up with a set of ratings for you for the movies which you have never seen. That is in essence what the problem is. So it recommends movies to you based on what Netflix thinks are movies that you are going to rate highly which means it has to generate a predicted rating for you, for a set of movies that you have not seen and then, choose a set of movies based on what Netflix thinks you would rate very highly.

(Refer Slide Time: 09:15)



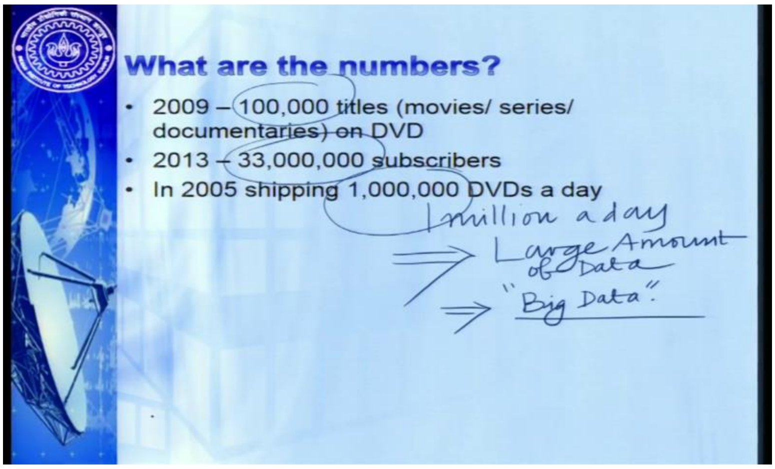
Angela, welcome to your very own Netflix homepage!
Based on your ratings, we've filled it with personalized suggestions JUST FOR YOU
The more you rate, the better we get at giving you suggestions you'll love.

Movie Recommendations From NetFlix

Top 10 for Angela

Movies Popular on Netflix

(Refer Slide Time: 09:52)



What are the numbers?

- 2009 – 100,000 titles (movies/ series/ documentaries) on DVD
- 2013 – 33,000,000 subscribers
- In 2005 shipping 1,000,000 DVDs a day

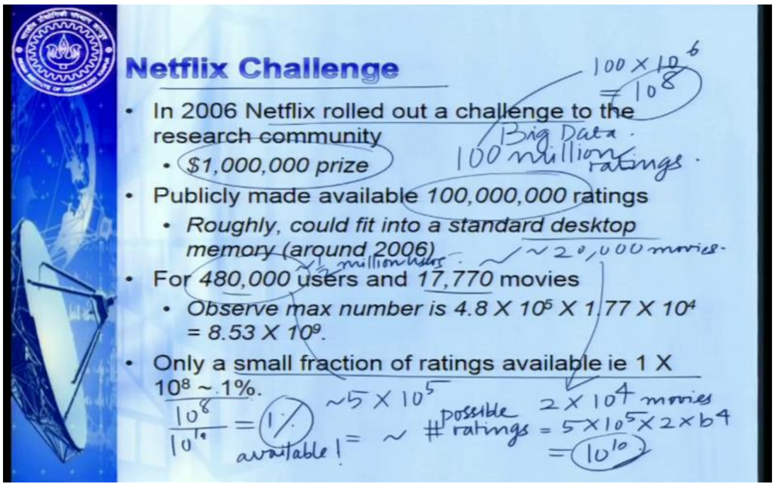
1 million a day

$\Rightarrow$ Large Amount of Data

$\Rightarrow$ "Big Data".

So in 2009, Netflix had about a 100000 titles and in 2013 about 33 million subscribers. In 2005, it is close to shipping about 1 million DVDs a day which is a large amount of data, implies this is a Big Data problem. So from this large amount of data, how do you mine the patterns and that is to be found out.

(Refer Slide Time: 11:00)



Netflix Challenge

- In 2006 Netflix rolled out a challenge to the research community
  - *$1,000,000 prize*
- Publicly made available *100,000,000* ratings
  - *Roughly, could fit into a standard desktop memory (around 2006)*
- For *480,000* users and *17,770* movies
  - *Observe max number is 4.8 X 10⁵ X 1.77 X 10⁴ = 8.53 X 10⁹.*
- Only a small fraction of ratings available ie 1 X 10⁸ ~ .1%.

$100 \times 10^6 = 10^8$

Big Data.

100 million ratings.

~½ million users.

~ 20,000 movies.

$\sim 5 \times 10^5$

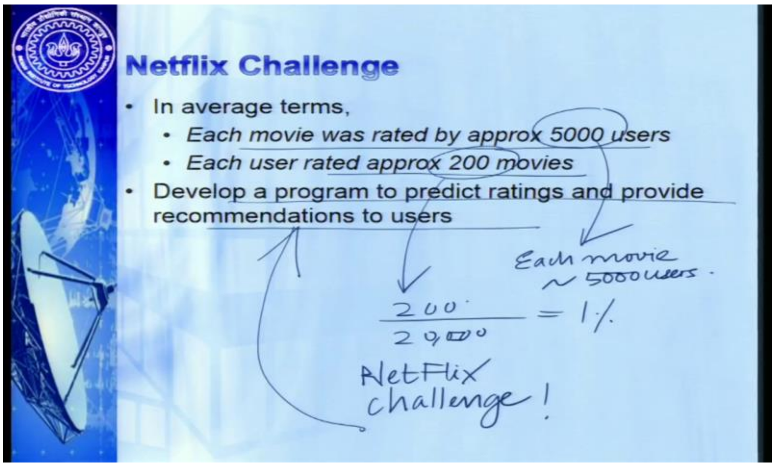
2×10^4 movies

$\frac{10^8}{10^{10}} = 1\%$ available!

$= \sim$ # possible ratings $= 5 \times 10^5 \times 2 \times 10^4 = 10^{10}$

Now, in 2006, Netflix rolled out an interesting challenge, this is termed as the Netflix challenge, to the research community and what happened in that challenge is it made available 100 million ratings. 100 million is $100 \times 10^6 = 10^8$ ratings which could fit into a standard memory of a standard desktop. Now, at that point it had about 480000 users and of course, these ratings were about 480000, that is roughly half a million users, approximately 20000 movies. So the approximate number of possible ratings is 10^{10}. Now the actual number of ratings is only 10^8. Now the reason for that is very simple because you have half a million users, you have twenty thousand movies but each user has not seen every movie. So each user has probably seen a fraction of the movies.

(Refer Slide Time: 13:37)



Netflix Challenge

- In average terms,
  - *Each movie was rated by approx 5000 users*
  - *Each user rated approx 200 movies*
- Develop a program to predict ratings and provide recommendations to users

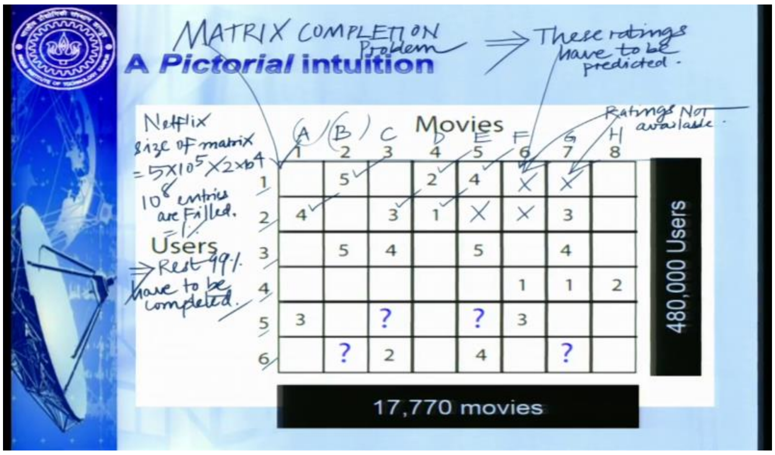
Each movie
~ 5000 users.

$\frac{200}{20,000} = 1\%$

NetFlix
challenge!

So only 1 percent of the ratings are available which means we have to predict the rest of the ratings. Based on these predictions, you come up with the movie recommendations for each user. In an E-commerce website, if every person has bought every item then there is no set to cover. The challenge is because few people have bought few items and it is not even that few people have bought the same items, different people have bought a different item. So from this checker kind of matrix, we have to come up with the ratings and recommendations for each user. So each movie was rated by approximately 5000 users. So each user has seen about a percent of the movies. Now, we have to develop a program to predict the ratings and provide recommendations for you. So this was the Netflix challenge.

(Refer Slide Time: 16:05)



A Pictorial intuition

MATRIX COMPLETION Problem ⇒ These ratings have to be predicted.

Netflix
size of matrix
$= 5 \times 10^5 \times 2 \times 10^4$
10^8 entries are Filled.
$= 1\%$
⇒ Rest 99% have to be completed.

Ratings NOT available.

Movies

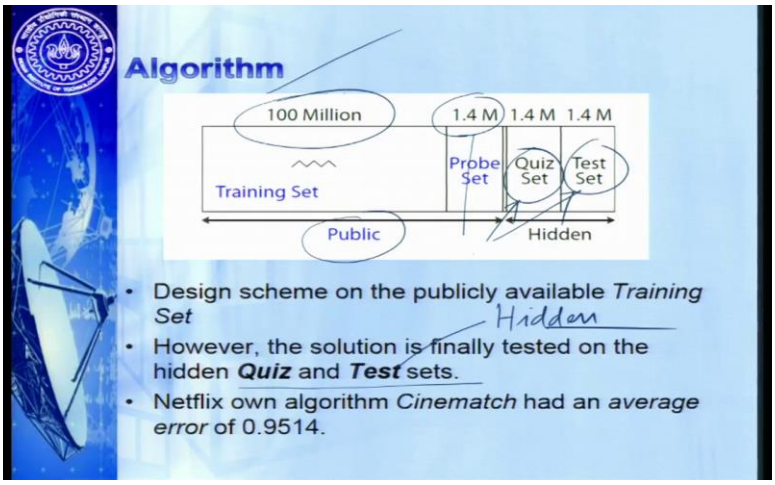
Users	A 1	B 2	C 3	D 4	E 5	F 6	G 7	H 8
1		5		2	4	X	X	
2	4		3	1	X	X	3	
3		5	4		5		4	
4						1	1	2
5	3		?		?	3		
6		?	2		4		?	

480,000 Users

17,770 movies

Now you have a set of movies A B C D E F G H and you have users 1 2 3 4 5 6. For instance, user 1 has not seen movie A, but he has seen movie B and rated movie B. User 1 has seen movie D, rated movie. User 2 has seen movie A, movie C, movie D rated. So these are the ratings that are available. The empty blocks are ratings. So for instance user 1 has not seen movie 6 which implies these ratings have to be predicted. You treat this as a matrix each row corresponds to user each column corresponds to movie. So now some users have rated some movies, therefore, some entries of this matrix are filled, the rest of the entries of this matrix are vacant. And therefore, we have to complete this matrix. This problem is known as a Matrix Completion Problem. So you have half million users, 20000 movies. So the size of the matrix is humongous.

(Refer Slide Time: 19:20)



Algorithm

100 Million
1.4 M
1.4 M
1.4 M
Training Set
Probe Set
Quiz Set
Test Set
Public
Hidden

- Design scheme on the publicly available *Training Set*
- However, the solution is finally tested on the hidden ***Quiz*** and ***Test*** sets.
- Netflix own algorithm *Cinematch* had an *average error* of 0.9514.

Hidden

The way the contest was organized was the training set was made available to the public and there is a probe set and one has to eventually test what is the performance of the algorithm that is being proposed. Now the quiz set and the test set are hidden and finally, the algorithm is tested on the quiz and test set. So whichever algorithm performs the best that is gives the recommendations on the test set which are closest to the ratings that are given by the users, that is selected.

So we will stop here and start looking at exploring this problem in the subsequent modules. Thank you very much.