# LLO 8200: Introduction

## Youtube generated Subtitles

So we’re gonna get started and talk about

um what is data science we this term

gets bandied around a lot we talk about

things like big data and so on

it's just thinking about

um what is this thing and and how is it

used and also as you know kind of we'll

do a couple of different examples uh

when has it been used well to you know

for great success within an organization

and what are some of the failures that

we can see

and the point of this class is really to

point us towards ways in which data can

be used effectively for decision making

so I want to think a little bit about

times when data was used effectively for

decision making and uh when it wasn't

so many many companies over the last

particularly you know a couple of

decades have made data analytics the the

central focus of what they do and I've

gotten really good at using a data

analysis

to inform their decision making probably

one of the absolute most famous is

Netflix so we've got this is Netflix's

market capitalization over time

um about 10 years before the the date of

this recording Netflix was about a 22

billion dollar company

by the time we got to 2022 it was about

a 300 billion dollar company

and now you know it's it has certainly

fallen off but it's still a 150 billion

dollar company more or less

so how did they go from being a 20

million dollar company 10 years ago to a

150 billion dollar company now

and what people within the company will

say is it had everything to do with

their use of analytics now Netflix got

started

um actually sending out DVDs to people

they were a mail delivery DVD service

indirect competition competition with

video stores

what they offered along with mail

delivery and no overdue dates was a

recommendation system if you liked this

movie then you might like that movie

it didn't work all that well uh it

helped a little bit and it was one of

the the first kind of major deployments

of those kinds of recommendation systems

ever but it really wasn't the key to the

business at the time and here we're

talking when it was much lower market

cap

um in the early 2000s in fact what

customers said that they liked best

about Netflix was that they didn't have

overdue fines that the the primary

competitor which is video stores and

particularly Blockbuster would made

almost all of their money from charging

Hefty overdue fines for overdue DVDs

but

what happened within Netflix was they

developed a core of individuals with

great expertise in using data analytics

to make predictions the prediction that

they started with was if somebody likes

one movie which other movies is that

person gonna like they took that and

they did a bunch of other stuff with it

so

as they moved into streaming Netflix had

this capability to begin with of doing

some recommendations

but they used that in a huge number of

other ways so this is from Netflix

themselves and they're talking about how

they use data and they think of

themselves as a data first company and a

company that

uses data and data analysis in every

aspect of what they do and so they've

got these three different roles within

the organization analysts engineers and

visualizers

and here they're talking about the ways

that each of them are using data to

inform the work that they do and most

importantly the decisions that they make

so for the analyst they they highlight

deep dive analysis and Metric

development what they're really talking

about is predicting outcomes and

predicting outcomes in two really

specific ways one is what's going to

predict a user's level of Engagement

with the platform how can we figure out

which users are going to use the

platform more and what characteristics

of the platform are going to encourage

people to use it more

they try all different kinds of things

to get this done uh one of the the early

Innovations from Netflix was an auto

play feature at the end of a given

episode they would immediately play the

next episode they implemented that they

reviewed the analysis and they found

that it increased engagement time

substantially

the other thing that they want to

predict at Netflix is membership who

will join and how long they'll join for

and so similarly they're looking at the

characteristics of individuals they're

looking at the characteristics of the

content that they're offering and

they're constantly trying to figure out

all right how can we predict

who will join and how long they'll join

for and go to the flip side of joining

as a member of Netflix is who's going to

leave

and they want they'll change as many

things as they can in order to affect

that outcome

now in the engineering side here we're

really talking about the delivery of the

content and they're very clever in terms

of thinking about

predicting things like when will the

system be under load uh you know and as

we think about the U.S most people most

of the time are going to be watching

Netflix at the end of a long day so you

know something like between 8 pm and

midnight they'll have this massive load

on their servers they're very very good

at predicting when the load will happen

and being ready so that when people are

asking for whatever video that they want

they're going to get it delivered

similarly all of the aspects of

maintaining the the catalog and

delivering it to the relevant users at

the right time they're constantly

predicting when that's going to happen

and what's what uh Innovations can be

made to make sure that that happens for

users when they want and they've got

visualizers which is really interesting

these individuals as it talks about are

interested in analytic tools and

Analysis and particularly what they're

constantly trying to do is take the data

that's being generated by the operations

of Netflix and turn that into actionable

intelligence for people across the

company so Netflix insists that it

doesn't exclusively make content

decisions based on analytics but

analytics are very important for Content

decisions now these content decisions

aren't necessarily going to be made by

experts in like data analysis right so

um they're the the visualizers are there

to help translate what can sometimes be

very complicated data into something

that can help people make decisions

okay uh so this is a company that you

know like absolutely at its heart

um uh has uh data analysis

what happened to companies that didn't

use data analysis during that same time

period uh here's the uh this is the

stock price of Blockbuster um

Blockbuster was a quite a valuable stock

um 29 trading at 29

um in 2002 a very very profitable

company

um

but Blockbuster pretty much assumed that

the way that they were doing business

was working just fine and they didn't do

a ton to look at how consumer Behavior

could be modeled or predicted based on

different changes that they were making

they just assumed people would you know

rent movies fail to bring them back on

time and pay overdue fines that worked

until it didn't and you can see the

company you know very very quickly lost

value in tandem with the rise of Netflix

and in fact went bankrupt and was

later purchased by another company

ceasing to exist at all

so here we have an example of a company

that really didn't start as a data

analytics company but used its existing

abilities to turn itself into a data

science first company and derived huge

benefits from doing so

it's not all success stories right I

mean there are times when things go

wrong so uh in the about two years ago

from before the recording of this

um uh lecture the company Zillow

announced that one of its major

divisions

um it's

a home buying division was going to

cease to exist and they lost something

like 420 million dollars uh in the

course of about three months which led

to massive reorganizations and layoffs

so what happened at Zillow so let's talk

a little bit about what Zillow is first

and Zillow is primarily a data science

company as well

uh their

um their primary product for a long time

was just ads now the way that they got

ads was that they provided people with

what was called they called a zestimate

right you go on Zillow and it tells you

what homes are worth what the predicted

value of a home is going to be and so

what they were doing was looking at

various things that might predict what a

house would sell for and putting them

together into a single number that says

okay this is what we think your house

might be worth

uh and of course we can't really know

what a house is worth until it sells

right we're not all selling our houses

all the time and so that information

people found very interesting they

wanted to know what their own house was

worth what their neighbor's house or in

a neighborhood they might move to what

are the houses were there when people

visited the website they got ads and

that was a very profitable business what

Zillow thought they could do though was

take those predictions and act on them

so Zillow didn't just start to keep

making predictions Zillow

would go and buy houses because they

thought they knew what houses were going

to be worth in the future they had

predictions about what houses were going

to be worth so they buy a house at a

value that the customer is willing to

sell at and they had an automated way of

doing this and they saw that the house

to the the customer sells the house to

Zillow Zillow makes some some repairs

what's called house flipping just doing

some basic mostly cosmetic repairs turns

it around puts it back on the market

sometimes delaying by a bit because they

thought the market value would rise over

some period of time and they thought

they could make a tiny profit from doing

that

so and they had every reason to think so

here's overall what house prices look

like from 1990 through 2022 and in

particular at the time when Zillow was

thinking about going into this business

they were making predictions from the

past between the end of the Great

Recession about you know 2009 or so

through

um 2013 home values Rose very steadily

and they were predicted to continue to

do so and so you can see the red arrow

is where our models might suggest that

home values end up they got there

eventually but not until much later

one other thing happened to Zillow

during this time period and it was all

about how much it costs to build things

so these are construction materials

there's a very long-term Trend in

construction materials starting way back

in 1950 and going through 2023.

so you'll notice something about this a

long steady rise in the price of

construction materials from 1975

to about 2020 so we're going from about

here to about here

um during that time period it was very

predictable how much house prices would

go up uh and excuse me housing

construction prices like how much it

costs to do stuff with a house this was

a key input for Zillow because they

would spend some time fixing up the

houses that they had bought

from 2020 through

2022 we saw the the fastest increase in

that particular area of the economy that

we've ever seen

so construction prices went up faster

during those two years than they ever

had in the since we've measured this

particular thing

so Zillow had this idea about how things

work they thought they knew what house

prices were going to be worth and they

thought that they would go up and they

thought they knew how much it would cost

to fix things up and it wouldn't cost

you know it cost a little bit more than

it did in the past but not much and

between those two they thought they

could make a huge profit

in fact what happened was their models

were not correct right that they

um

what they thought was going to happen

based on the past was not at all what

happened during the course of these two

years so they were left with a large

number of houses that were worth less

than they thought they were and they

also weren't able to fix them up because

it was far too expensive and so they

lost a huge amount of money very very

rapidly

so here we are you know we've got this

this idea of data analysis using data to

predict things

um and we want to know the the strengths

and weaknesses of these approaches and I

think that there's a really good summary

of this from this article on what

happened at Zillow

data and models aren't good at things

that have happened not have not happened

before it reminds me of the famous Yogi

Berra statement that prediction is hard

particularly about the future

a computer will do whatever you ask it

to do but the outcome depends on what

you ask

this class is all about what you ask

how can we ask the right questions of a

data analysis so we avoid using it

poorly and instead are able to implement

it so that any organization not just a

for-profit company but a school a

college a non-profit you can have an

understanding of how their operations

are working and what might change what

might happen in the future particularly

as things that are happening within the

company change

so let's talk a little bit about some of

the terms and particularly I want to

introduce this concept of a model

so in this class just like your previous

class we'll talk about dependent

variables so this is what we're

interested in the measurable phenomenon

so when I was talking about Netflix it

would be things like whether or not

somebody's going to subscribe how long

somebody's going to spend engaged on the

platform for Zillow it would be

something like what's a house going to

be worth

and then we've got independent variables

these are predictors this is anything

that might help to predict the outcome

so in the case of something like Netflix

this might be what somebody has watched

in the past

um it might be what time of day it is

people have different lengths of

Engagement depending on what time of day

that they tune in

um it might be the um

uh the they might and Netflix does this

it might be the location that the person

is they they vary the content that's

delivered depending on the the location

of the person across the country

anything that might help to predict that

outcome

for home prizes it would definitely be

the neighborhood and the value of homes

that had sold previously how many square

foot the house has all that kind of

stuff and then a model is how we

establish the relationship between the

predictors and the outcome

we say okay I see these characteristics

of the situation what does that tell me

about the outcome and what we do is we

resolve that to a set of measurable

things we actually break that down into

this the precise relationship between

those inputs and that outcome

what that does in the end is it says

okay when this input changes I'm going

to expect these changes in the outcome

once we've got a model then we can move

forward and say okay what does that

model predict would happen if we saw

some changes in the predictors

so that's absolutely crucial and you can

see how for Netflix that worked out

great they had some really good models

that predicted how individuals were

going to engage with a platform they

continue to refine and expand those and

they did a really good job of

understanding the situation led to a

very successful company for Zillow they

also had a model they thought about the

relationship between

current home prizes and future home

prices and all these other

characteristics but things changed their

model was not working anymore and

resulted in a lot of losses

okay

so

what does data science do and basically

we're going to take this in two

different areas the first thing it does

is it establishes the nature of the

dependent variable through what we would

call a descriptive analysis

here we're going to be thinking about

things like how do we measure the the

central tendency of the dependent

variable is it should we look at the

mean are there other measures that we

should look at

where's the outcome highest or lowest

you know where are people most engaged

with the platform like Netflix where are

they least engage where are home prices

lowest and highest when were they

highest or lowest can we look back in

time and see when these values were

highest or lowest

and then how does this outcome vary as a

function of other stuff what what can we

say just kind of thinking it through do

we see that users of a certain age are

more likely to engage with the platform

do we see that users where it looks like

their income is higher or lower or more

likely to engage with with different

types of content all the different

things that we might possibly measure or

infer how does the outcome change as a

function of those

so that's descriptives and that's just

kind of like you know kind of what it

sounds like it's describing but it's

really thinking about the dependent

variable and like just looking at any

patterns that we can establish

now when we talk about modeling we get

much more specific and we're really

interested in prediction what is that

dependent variable going to be how sure

can that it can we be that it's going to

be a certain value particularly if we're

thinking about in the future

um

so the first thing that we would need to

establish then is how much does that

dependent variable increase or decrease

as a function of the independent

variables what's the rate of change of

that dependent variable when we're

taking into account other independent

variables

um so we could say you know for

um every

um uh time that uh somebody has engaged

with Netflix in the past what does that

predict about their future engagement is

it one to one do people increase or

decrease their engagement over time what

does that look like

how well does that hold up now this is

really key and we'll talk about this a

lot in this class okay we've got a model

we've established some relationships

um are there times when it's stronger or

weaker does it how much confidence

should we really have in the results of

this model how much error is really what

we're talking about how wrong can the

model be

and then the last and again this is

really important we may be able to do

all this modeling and make some

predictions

can we do something can we make

different decisions and that's this

question of actionable insights

so that gives us just a brief

introduction to thinking about this

we're thinking about how organizations

might use data

and they can use it for descriptive

purposes they can use it for modeling

purposes and really what they want to do

is make different decisions the

decisions that they make can be

extremely consequential and can lead to

success or failure as we've seen