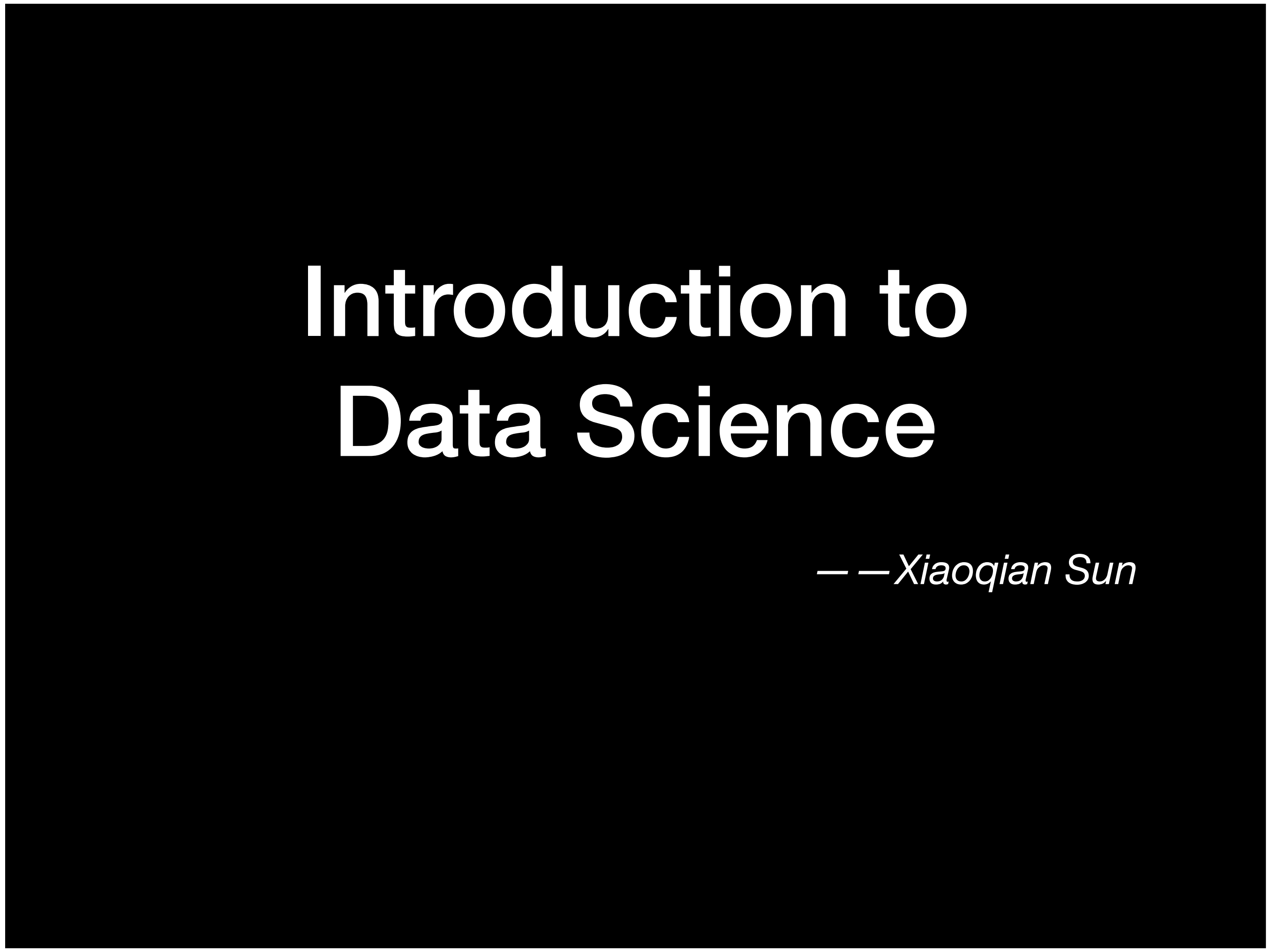
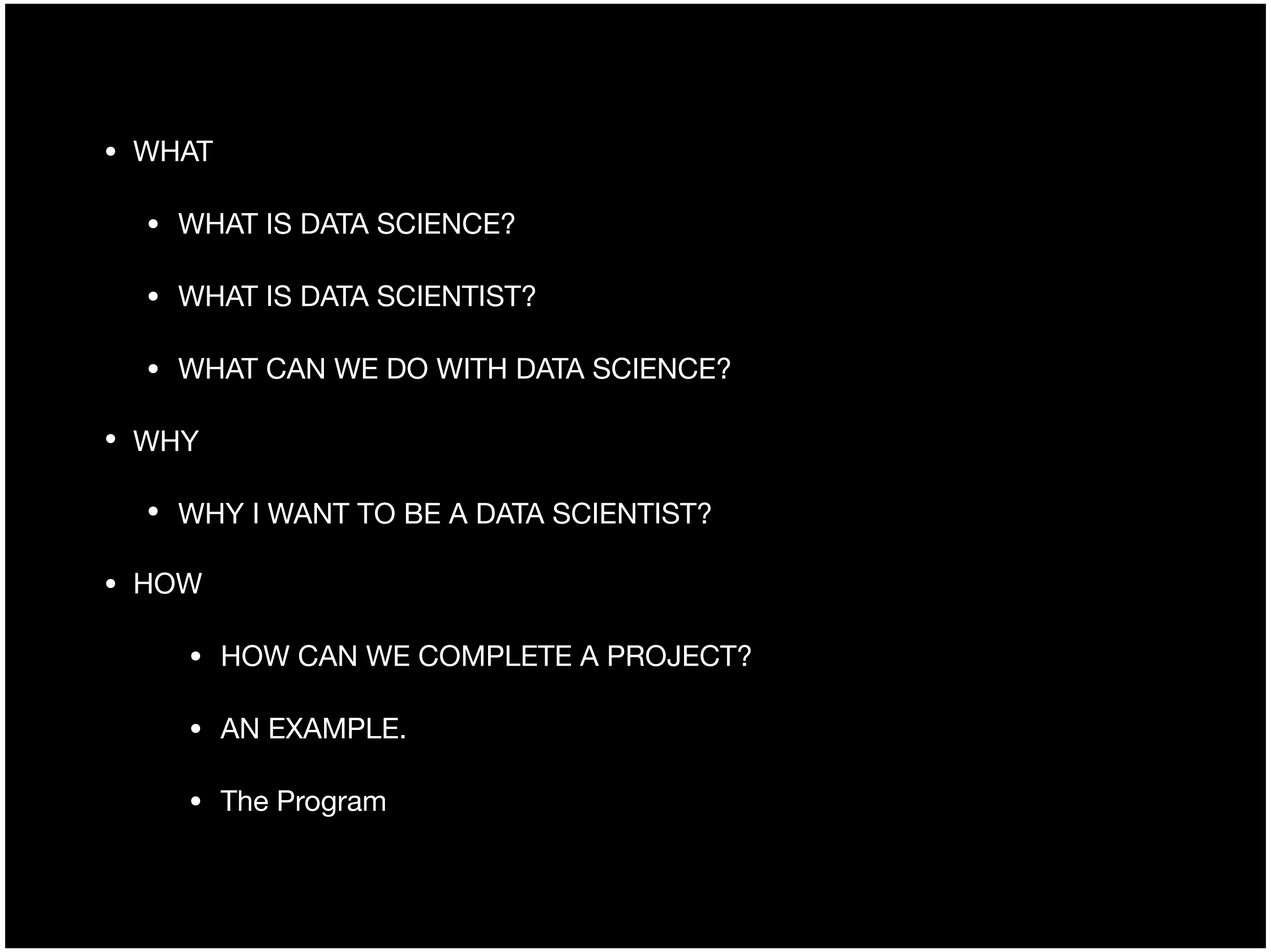


Introduction to Data Science

——Xiaoqian Sun

Hello guys, welcome to my class.
I'm Xiaoqian Sun and you can also call me Lexi.
Today, I'll give you an intro to Data Science. Hope you enjoy it. Let's begin.



- WHAT
  - WHAT IS DATA SCIENCE?
  - WHAT IS DATA SCIENTIST?
  - WHAT CAN WE DO WITH DATA SCIENCE?
- WHY
  - WHY I WANT TO BE A DATA SCIENTIST?
- HOW
  - HOW CAN WE COMPLETE A PROJECT?
  - AN EXAMPLE.
  - The Program

This is today's content table, including 3 big parts, what, why and how?

For the what part, I will ask you three questions. I'll give you my answer but I also want you guys to think about it and tell me your understanding.

For the why part, I'll share my own story with you.

For the how part:

Firstly, I'll talk about the process to complete a data science project. After that, I'll give you an example to illustrate that. At last, let me talk a little bit about what I'm doing and what can I provide for you guys to help you start learning data science.

WHAT IS
DATA SCIENCE?

Let me ask you a question: what is data science?

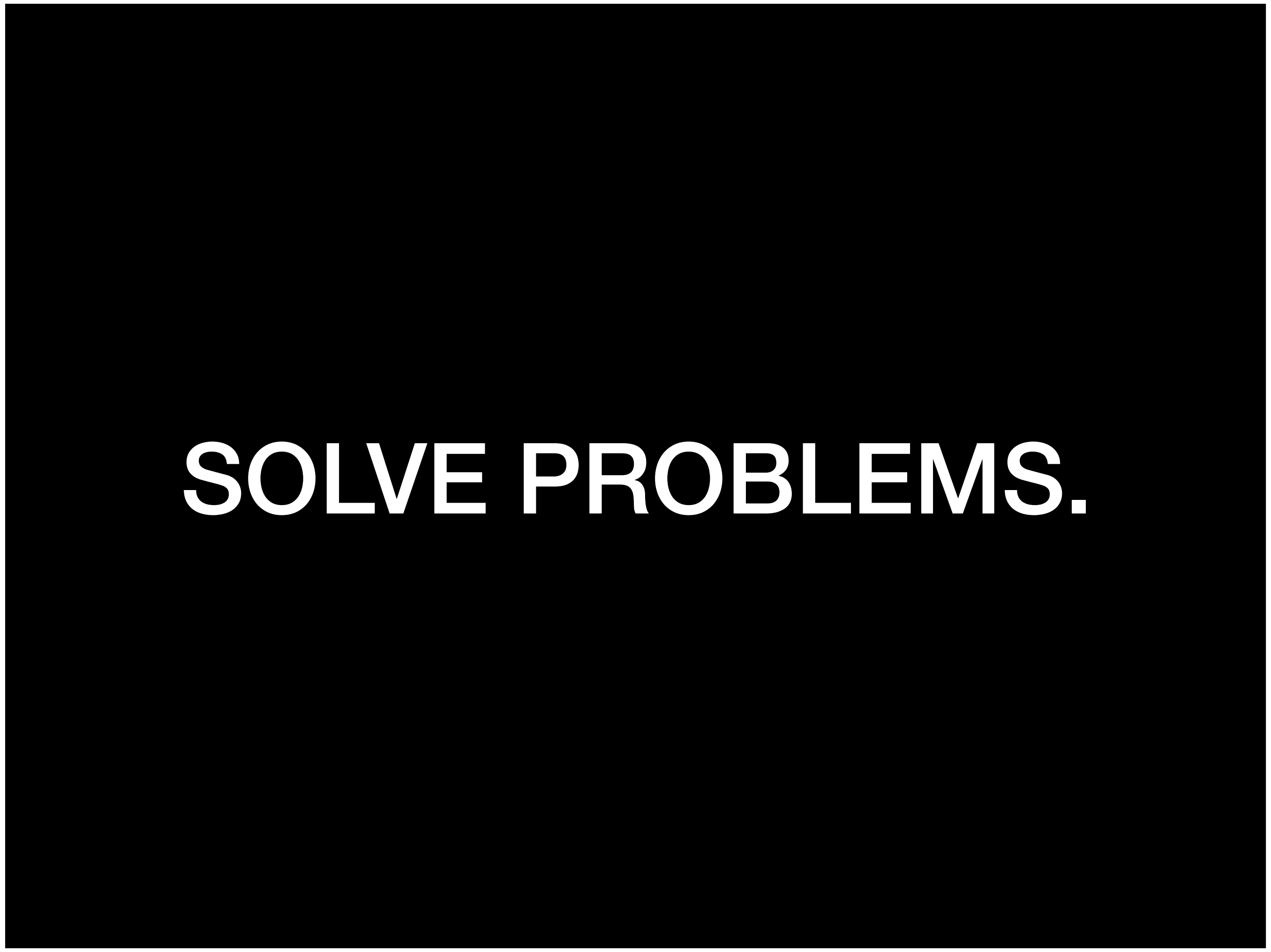
Actually, this is the first question that my professor asked me in my introduction to data science in my data science program. I was too shy to answer this question in front of this class and it was a pity because I had the same answer with my professor and obviously, I missed an opportunity to leave a good first impression on my professor in my first class. I would imagine that if I had answered that question, whether my master's study would be different.

I know being shy would let me miss something. This is also my true story. I went to dog park yesterday and a girl in the dog park asked my dog's name. I told her and she found that my dog's name is Japanese and she told me she comes from japan. I was very happy. I'm learning Japaness now, and like their culture, their architectures and TV shows. I wanted to make friends with her but I was too shy to open my mouth. After exchanging our dog's name, our talk ended. Bu lucky me, we met at the parking garage, by saying hi to her dog, we talked a little bit. Next time, if I would meet her at the dog park, I will start the talk.

Now, you are right infront of the screen, you don't have to feel embrassing if you have different answer from me. Go ahead and think about it.

WHAT IS DATA SCIENTIST?

And next question is what is data scientist?



SOLVE PROBLEMS.

All right.
My answer is that: data science is a methodology to solve problem.
A data scientist is a problem solver based on solid data.
In another word, Data scientists examine which questions need answering and where to find the related data, and answer that question based on the data they collected.

WHAT CAN WE DO WITH DATA SCIENCE?

The third question I think might be helpful to answer is: what can we do with data science?

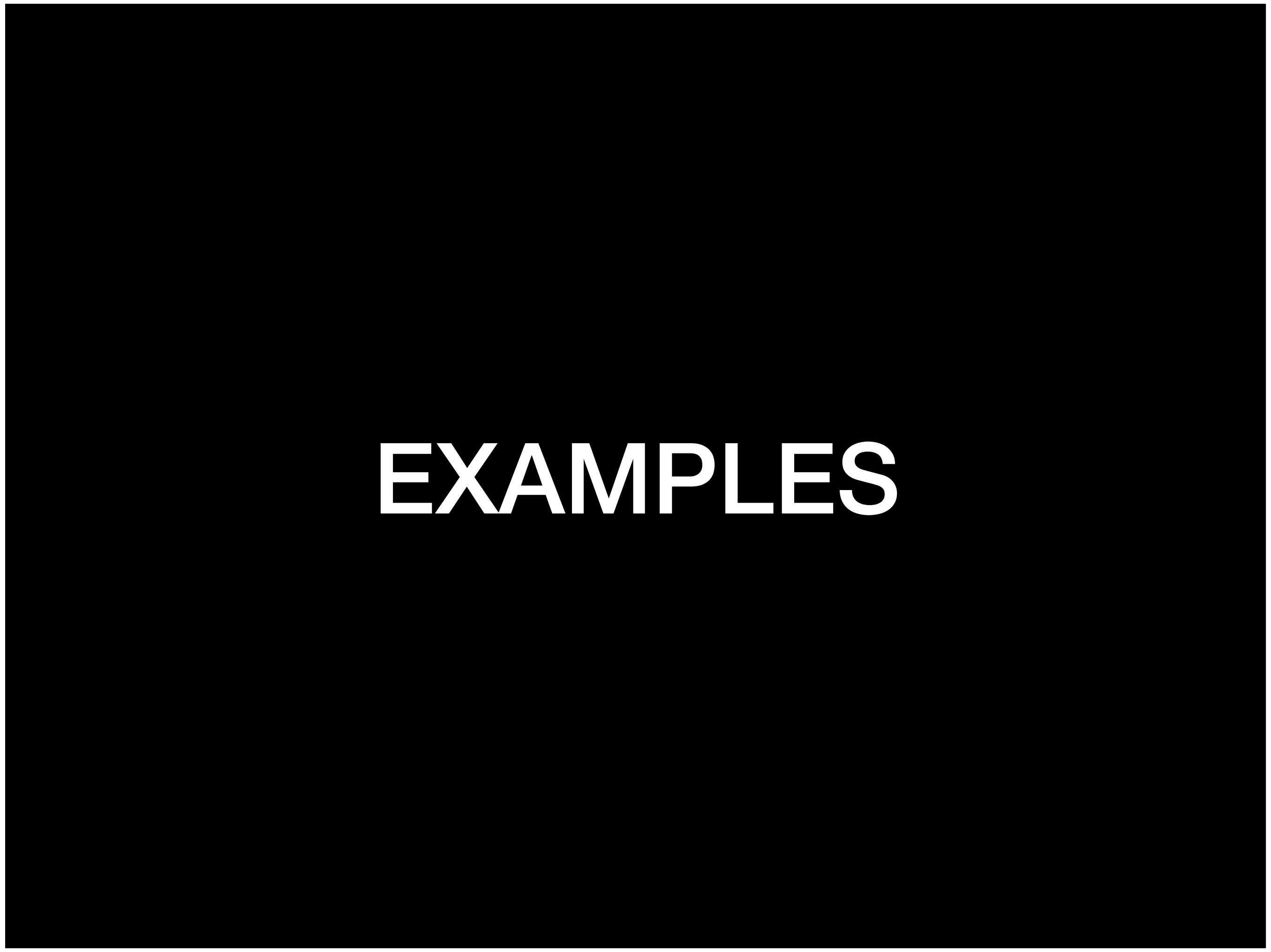
The implication behind this question is:

Is data science interesting enough that I should consider having a career as a data scientist?

If I learn data science, what can I do in the future?

What if I want to be a doctor, how can I benefit if I learn data science.

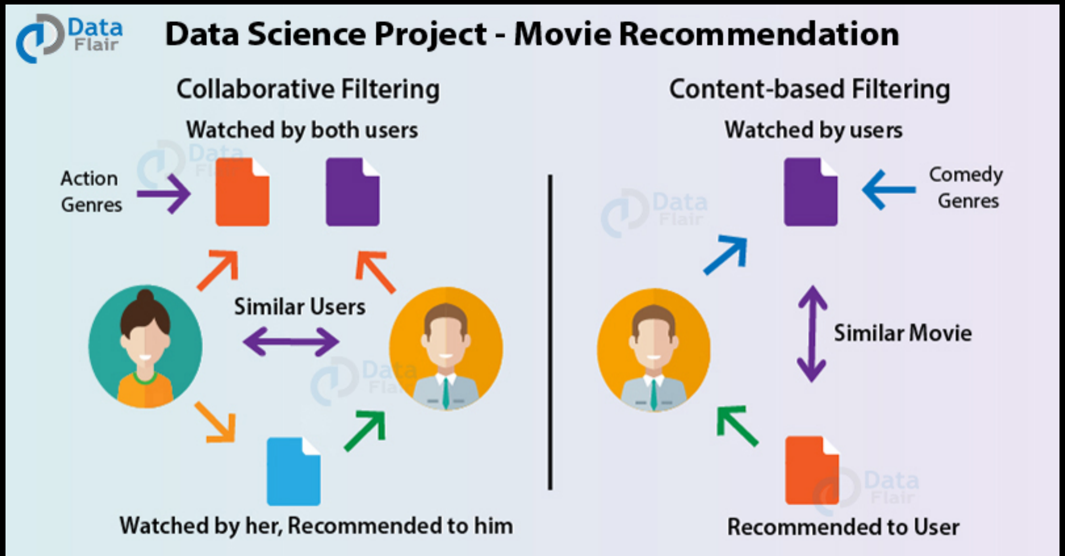
What if I want to be a journalist, is data science helpful for me to write really good stories or documentary?



EXAMPLES

Let me give you some examples.

Movie Recommendation system



Have you ever been on an online streaming platform like Netflix, Amazon Prime? I'm sure you have.

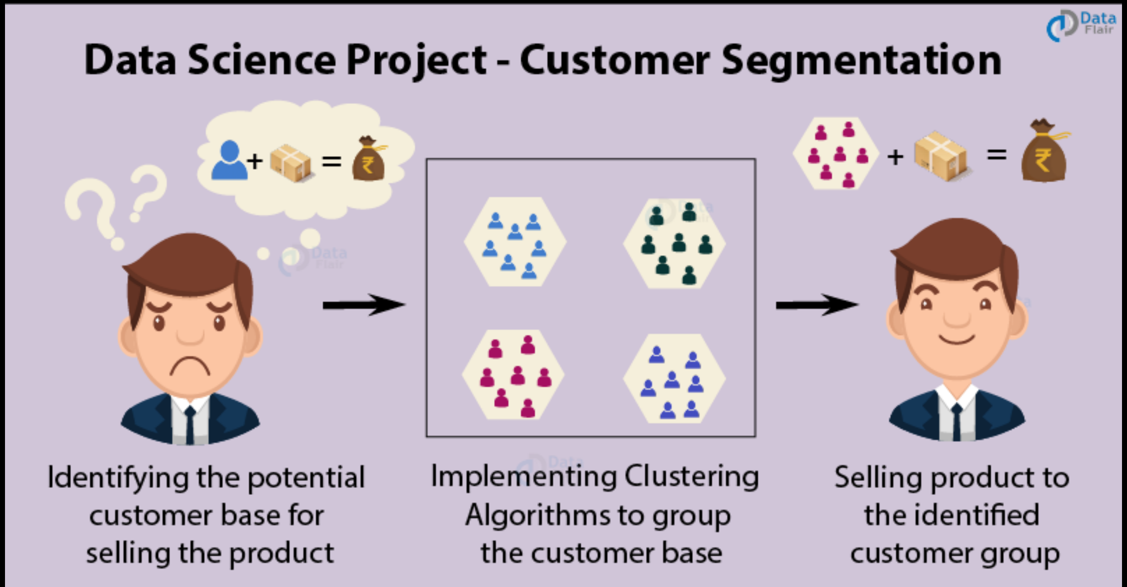
You might watch a movie and after some time, that platform started recommending you different or similar movies and TV shows. A system which is capable of learning your watchlist pattern and providing you with relevant suggestions is called Recommendation system.

A recommendation system is a platform that provides its users with various contents based on their preferences and likings.

A recommendation system will find similarities between users based one their basic information, like age, gender, location and watching list. For example, if this film was watched by user A, user A and B are similar user, like they have similar age, same gender, similar watching list, and there is a bigger chance that this film will be recommended to user B.

A recommendation system will also find a similarity between the different products. For example, Netflix Recommendation System provides you with the recommendations of the movies that are similar to the ones that you have watched in the past. They assume that if you watched this before, you might also like the similar products since this is your film gene.

Customer Segment



Data Science Project - Customer Segmentation

Identifying the potential customer base for selling the product

Implementing Clustering Algorithms to group the customer base

Selling product to the identified customer group

And the second example is customer segment, which is a real essential part in commercials. It's easy to understand right? Bank won't recommend credit card to people with much low credit, or high-risk financial product to people who cannot stand high risk.

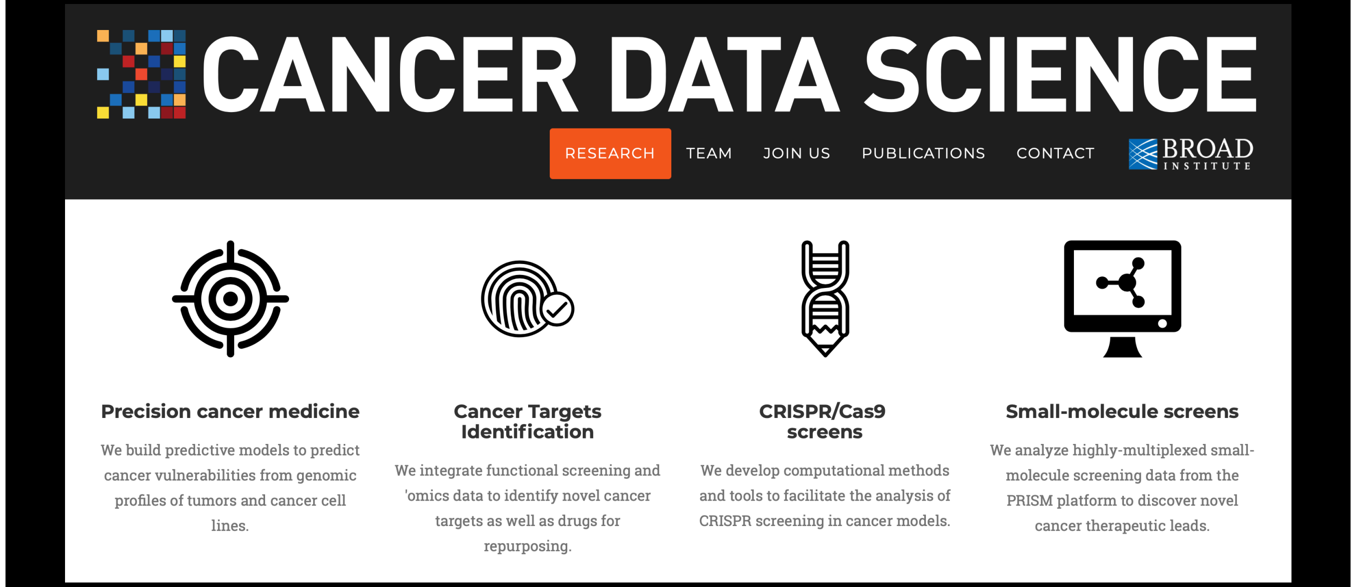
So, What is customer segments? Customer Segmentation is the process of dividing customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and spending habits and so on.

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. In my mind, understanding your customer is the very first step of starting business, only in that way can you figure out what kind of stratgies you want to use, what kind of problem you want to solve for them.

Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

For example, mobile company. If you are a new user, they will give you a big welcome offer to let you explore as much as possible. And then they will give you 15% off of monthly fee to keep you stay with them if you open auto recharge. And once you become loyal customer, they will give you little surprise from time to time to re-engage you.

Fighting Cancer with Machine Learning



CANCER DATA SCIENCE

Precision cancer medicine

We build predictive models to predict cancer vulnerabilities from genomic profiles of tumors and cancer cell lines.

Cancer Targets Identification

We integrate functional screening and 'omics data to identify novel cancer targets as well as drugs for repurposing.

CRISPR/Cas9 screens

We develop computational methods and tools to facilitate the analysis of CRISPR screening in cancer models.

Small-molecule screens

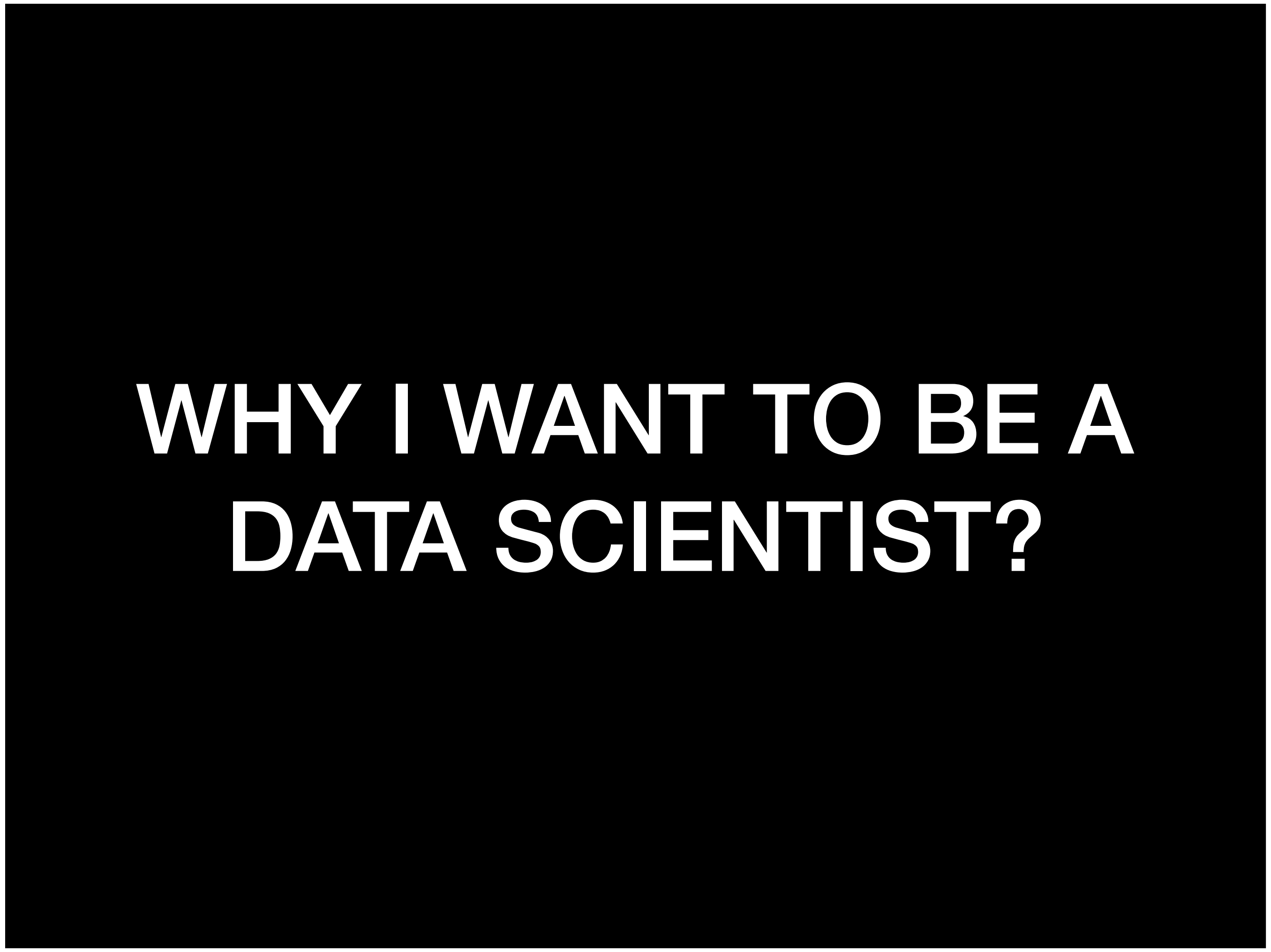
We analyze highly-multiplexed small-molecule screening data from the PRISM platform to discover novel cancer therapeutic leads.

Next example is about medical. We all know cancer is terrifying and people are fighting with cancer really hard to find ways to cure cancer.

A major obstacle for treating cancer is a lack of precision medicines. Many potential targeted therapies fail to transition from preclinical models to patients due to incomplete knowledge of the drug's mechanism of action and/or absence of robust biomarkers to identify relevant patient populations.
To find biomarkers that predict sensitivity to genetic or chemical perturbations, you will work with some of the largest experimental cancer biology datasets in the world.

If you want to learn more about this pic, simply google: fight cancer with machine learning. Data scientists in this organization are using big data and machine learning techniques to find biomarkers to help improve the accuracy of classifying and recognizing patient groups.

Well, I want to say, there is a whole world outside waiting for you to explore and play with. Data science is a tool to help you get in the filed that you are interested in, that you would like to devote yourself into, that you would like to choose to start a career.



WHY I WANT TO BE A DATA SCIENTIST?

Then how do I know data science? Why I want to be a data scientist? Let me introduce myself in detail, and I hope my experience can help you to figure out something even a little bit.

My bachelor's degree is Economics. I didn't pick this major by myself. Actually, my parents picked it for me and lots of my friends also let their parents to pick the major for them. In my age, students were all focused one their course, Chinese, Math, Chemistry, Physics. We lack the opportunity to explore the world and find out our real interest.

Well, My father is a cable engineer and my mother is a project manager. In their thoughts: mechanical engineering is so exhausting for girls , while economics is always important as long as it's not the end of the world. So I went to university to study economics. Econ is interesting and I developed a brand new perspective to look at the world, but somehow, I don't want to work in this industry for my whole life.

After graduation, I went to Meituan, one of the world largest online to offline internet company, to work as a product manager. Basically, I was in charge the life circle of the product, including designing system, testing, launching, iterating promotions and tracking operating data. In that role, I contacted data everyday and I wanted to be more professional in data, so I went to gwu for further study.

In data science program, I learned program language, such as R, python. I learned how to used data in a totally different way, not just seeing the trend going up or down, but to digging into the data and extracting the hidden pattern from the data.

In my second year of maters' study, I started to do a project with my professor. The project is about to find lower-dimension metric that contribute to stratify lupus patients. Our goal is also trying to find biomarker to help to classify and recognize lupus patient as early as possible. We had some really good results and we are now working with more samples to testify and validate the results we got before.

Thought this project, I developed my strong interest in studying genetic data. So, if you ask me why I want to be a data scientist? The answer shot be:

MAKE A DIFFERENCE.

I want to make a difference. I want to become a professional data scientist, focusing on disease data, to improve life quality of patients based on solid data. I would feel that my life worth it if I could do something meaningful, if I could bring hope to patients.

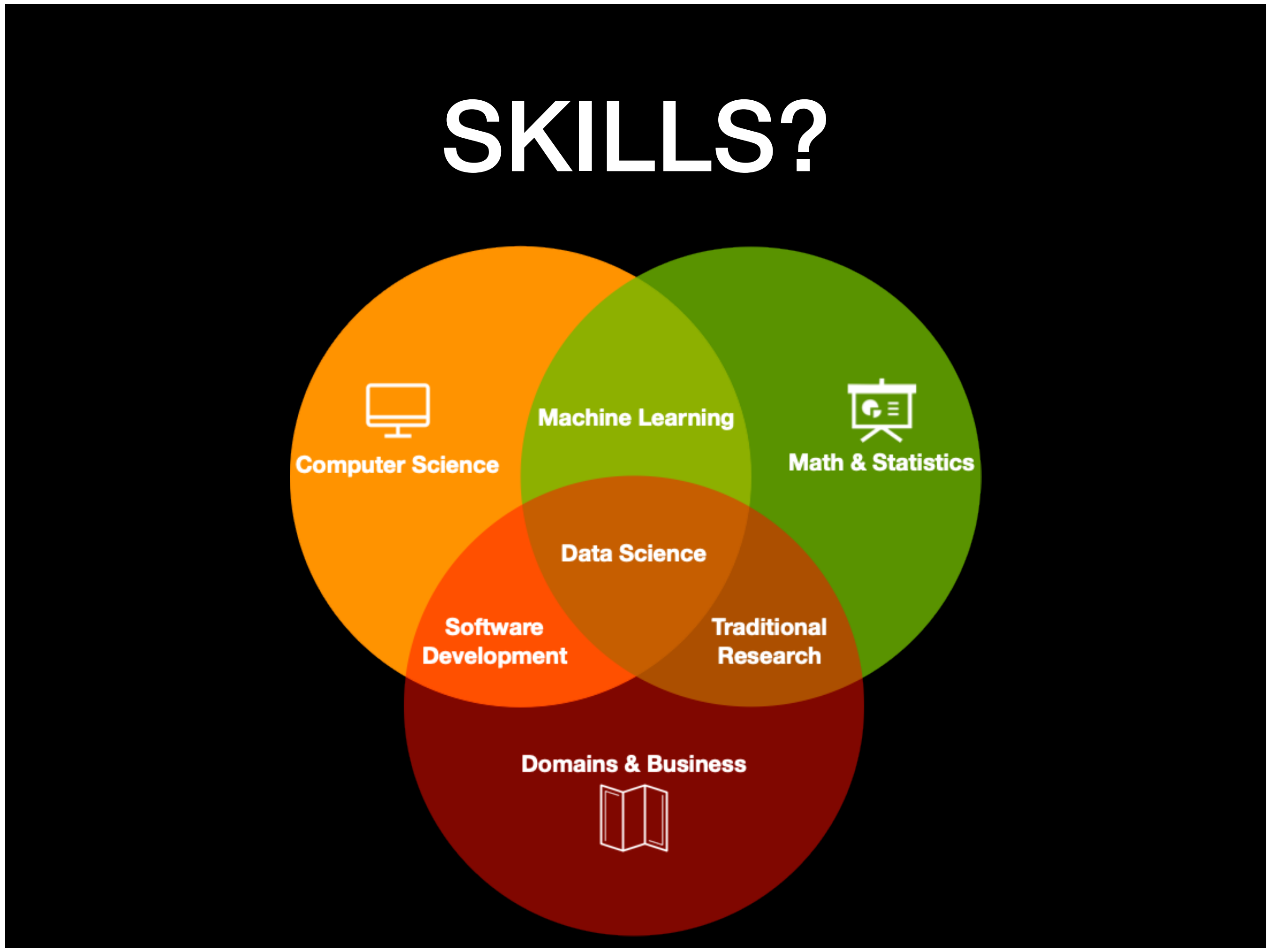
You guys are so lucky that you are exposed to information from various source today. You have bigger chance to find where your passion lies in, to figure things out, to set a life-long goal to guide you to move forward much earlier than me. You guys have many powerful accesses to outstanding people in that industry who would like to give you some instructions and share experience with you.

But no worries, life is long. So take your time and explore the world like nothing can stop you. If you make some choices that you might find wrong later, no big deal, cut loss in time and go for what really motivates you.

Good luck everybody.

HOW CAN I BECOME A DATA SCIENTIST?

Then Let’s talk about how can I become a data scientist?
We should answer the 4th question: what skills do we need to become a data scientist.



SKILLS?

My answer lies in this picture.

Data scientists today are people who blend with many different skills. This diagram shows the set of skills that a data scientist need to have.

Computer science is a methodology to write code for obtaining, cleaning and building models to analyze. You can choose program language you prefer. For me, I'm a heavy Python user.

Math & Statistics is to use mathematic algorithms and equations to properly do calculations on our dataset.
By using computer science to automate the execution of mathematic algorithms on the dataset is called machine learning.

When we add domain and business knowledge about the dataset into the composition, we are generating useful knowledge and conclusion about the dataset that will help guide our decision, this is what data science is trying to achieve. For example, when we study experimental cancer biology datasets, say we want to find biomarkers to classify cancer patients, we won't use algorithms to identify relationship among biomarkers, but we will focus on clustering and grouping.

If we only use traditional math & stats techniques combined with domain knowledge, that's the traditional research. Because without the help of computer, our computational power is limited to deal with large and complex dataset.

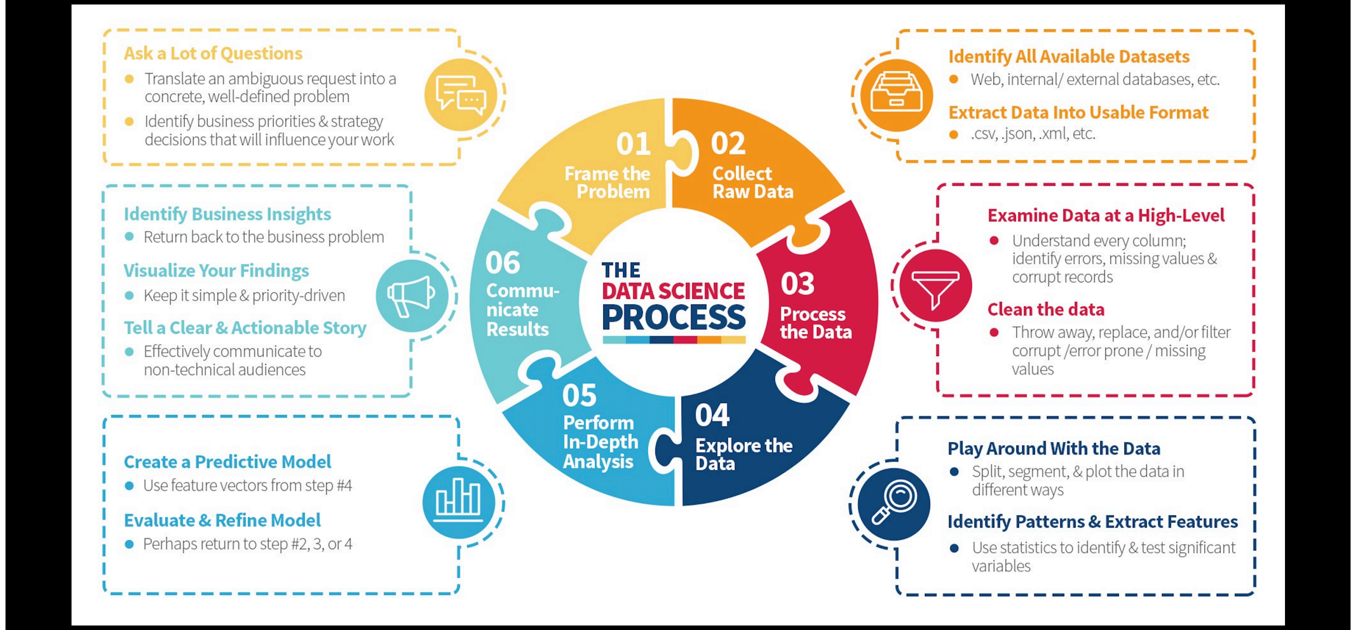
If we only use computer science techniques to display domain knowledge, that's the field of software development, not solving potential or existing problems with data.

In another word, data scientists are able to perform complex analysis on huge datasets. Once they have done this, they also have the ability to write and make informative visulizatons to communicate their technical findings to technical people.

HOW CAN WE COMPLETE A PROJECT?

For next part, I will talk about the procedures to do a data science project.

DATA SCIENCE DECONSTRUCTED



THE DATA SCIENCE PROCESS

01 Frame the Problem

Ask a Lot of Questions
- Translate an ambiguous request into a concrete, well-defined problem
- Identify business priorities & strategy decisions that will influence your work

02 Collect Raw Data

Identify All Available Datasets
- Web, internal/ external databases, etc.

Extract Data Into Usable Format
- .csv, .json, .xml, etc.

03 Process the Data

Examine Data at a High-Level
- Understand every column; identify errors, missing values & corrupt records

Clean the data
- Throw away, replace, and/or filter corrupt /error prone / missing values

04 Explore the Data

Play Around With the Data
- Split, segment, & plot the data in different ways

Identify Patterns & Extract Features
- Use statistics to identify & test significant variables

05 Perform In-Depth Analysis

Create a Predictive Model
- Use feature vectors from step #4

Evaluate & Refine Model
- Perhaps return to step #2, 3, or 4

06 Communicate Results

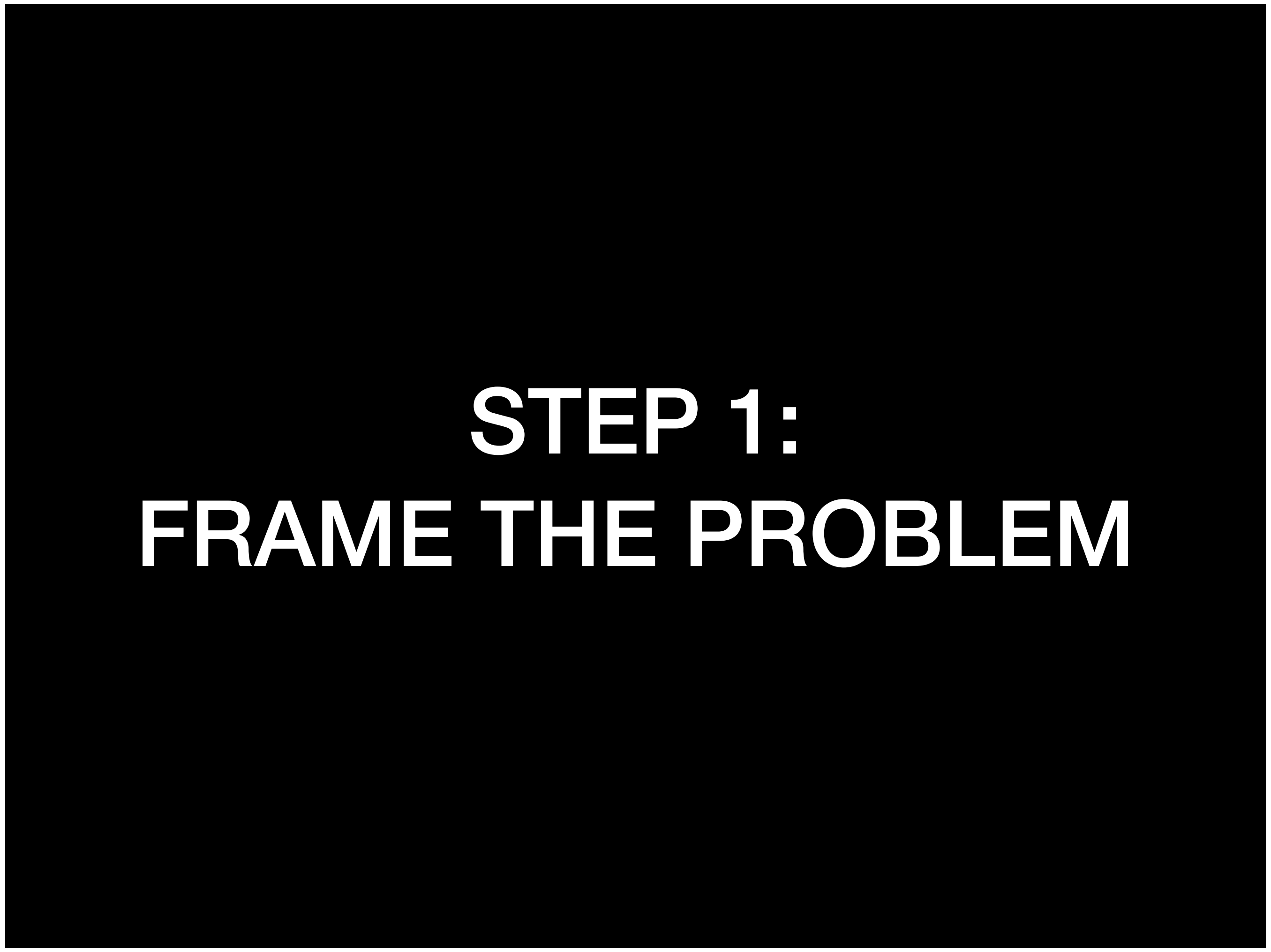
Identify Business Insights
- Return back to the business problem

Visualize Your Findings
- Keep it simple & priority-driven

Tell a Clear & Actionable Story
- Effectively communicate to non-technical audiences

A simple Google search of “how to learn Data Science” returns thousands of learning plans, degree programs, tutorials, and bootcamps. It’s never been more difficult for a beginner to find signal in the noise. Rather than jump around, I would like to empower you to begin. We use this process to break down complex project into small parts and guide us to progress.

First of all, let’s deconstruct the data science process. One thing you should notice is that: The whole is greater than the sum of it’s parts. Data science is not a single discipline, but a craft at intersection of many, so let’s look at the story of data science process.



STEP 1: FRAME THE PROBLEM

In the first step we ask a lot of questions to frame the problem.

We try to figure out the situation: what’s the problem right now, what do you need, what’s your goal? We try to get the big picture of the problem.
For example, a company comes to you. they want to improve the efficiency of employees. At this time, you should have a big question mark in your heart: this is what you want, but is this the real question that the company is facing?

Then, you asked a lot of questions, like what was the efficiency before? when it started to drop? what happened that time? which employees working efficiency dropped, all of them or just part of them? What happened to that group of employees?
At last, we make a hypothesis based on the information you collected: that the drop of efficiency might because of the drop of happiness of employees. In another way, they feel unhappy in this company, they are not willing to contribute anymore, but they might not have any other better choice, they choose to stay.

So how to improve employee efficiency becomes how to let their employees feel happy again.

That's the beginning step, you don't only take the information provided by your client, but also ask by yourself, to figure out the real issue.

STEP 2:
COLLECT RAW DATA

Next step, when we figure out what kind of problem we are trying to solve, we collect new data beside the part provided by your client.

For example, before the company provided you with lots of data including profit, cost, sales, but the data dose little to describe how happy these employee are.

So, we collect new data about factors that really influence happiness score, such as salary, family condition, trust or working environment.

STEP 3:
PROCESS DATA

When we get the data, we try get it ready to use.

First of all, we should understand this dataset thoroughly.
What does each variables mean? Like: What is trust? How do you define working environment?

Secondly, we also need to build a pip line to clean the data. Like: how do we deal with missing values or outliers? Such as what do we do if one salary is 1 million per month?

When we get the data prepared, we move to next step.

STEP 4:
EXPLORE THE DATA

In step 4: We play around the data to get familiar with it.

What pattern exist in each factor? What's the relationship between each two factors.

For example, if we found salary has strong relation with family condition, we might just choose one in next step to build our model.

In this step, we use data mining and data visualization techniques. Results from this step is an important part in your final communication with your client, these clear and easy-understanding graphs will also let them understand their problem clearly.

STEP 5:
PERFORM
IN-DEPTH ANALYSIS

In next step we perform in-depth analysis on the dataset.

Actually, you usually face a pool of models to choose from when you come to this stage

Sometimes, we have to do a lot of research, read lots of publications of related topics, like how other people analyzed this problems, what kind of techniques they used, what results they got.
Well, sometimes, you may also face a new question and no precursors in this area. you'll get so excited and explore by yourself to extract the hidden patters in the data set.

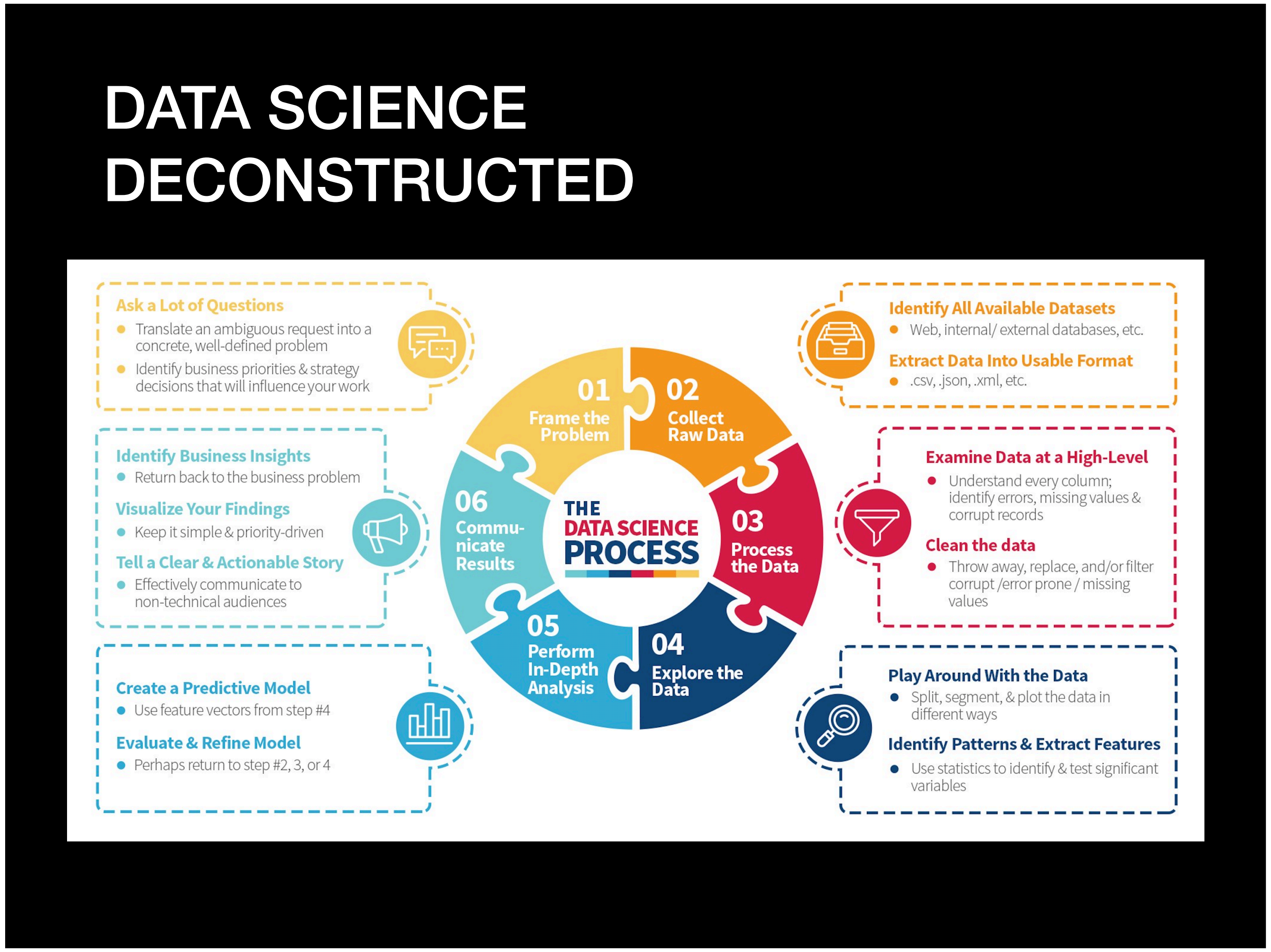
Continue with happiness score question. Eventually, we try to figure out how to improve happiness of employees right? We cannot improve all, like increasing salary, giving much more freedom to their own work, renovating the whole building to improve working environment. That's too much investment. We try to find which factor is most powerful. So we build a model to see which factor has most strong relationship with happiness score so that we choose the best fit.

STEP 6:
COMMUNICATE RESULT

Assume that we have the first version result: We figure out that salary and trust contribute most to improve happiness score.

At this time, we come to communicate with our client to tell them what we found, how was the process and why it is trustable, to convince them our advice is actionable.

In this communication, if our clients have more thoughts and concerns, we might repeat the whole process until we find the truth.



DATA SCIENCE DECONSTRUCTED

THE DATA SCIENCE PROCESS

01 Frame the Problem

02 Collect Raw Data

03 Process the Data

04 Explore the Data

05 Perform In-Depth Analysis

06 Commu-nicate Results

Ask a Lot of Questions
- Translate an ambiguous request into a concrete, well-defined problem
- Identify business priorities & strategy decisions that will influence your work

Identify All Available Datasets
- Web, internal/ external databases, etc.

Extract Data Into Usable Format
- .csv, .json, .xml, etc.

Examine Data at a High-Level
- Understand every column; identify errors, missing values & corrupt records

Clean the data
- Throw away, replace, and/or filter corrupt /error prone / missing values

Play Around With the Data
- Split, segment, & plot the data in different ways

Identify Patterns & Extract Features
- Use statistics to identify & test significant variables

Create a Predictive Model
- Use feature vectors from step #4

Evaluate & Refine Model
- Perhaps return to step #2, 3, or 4

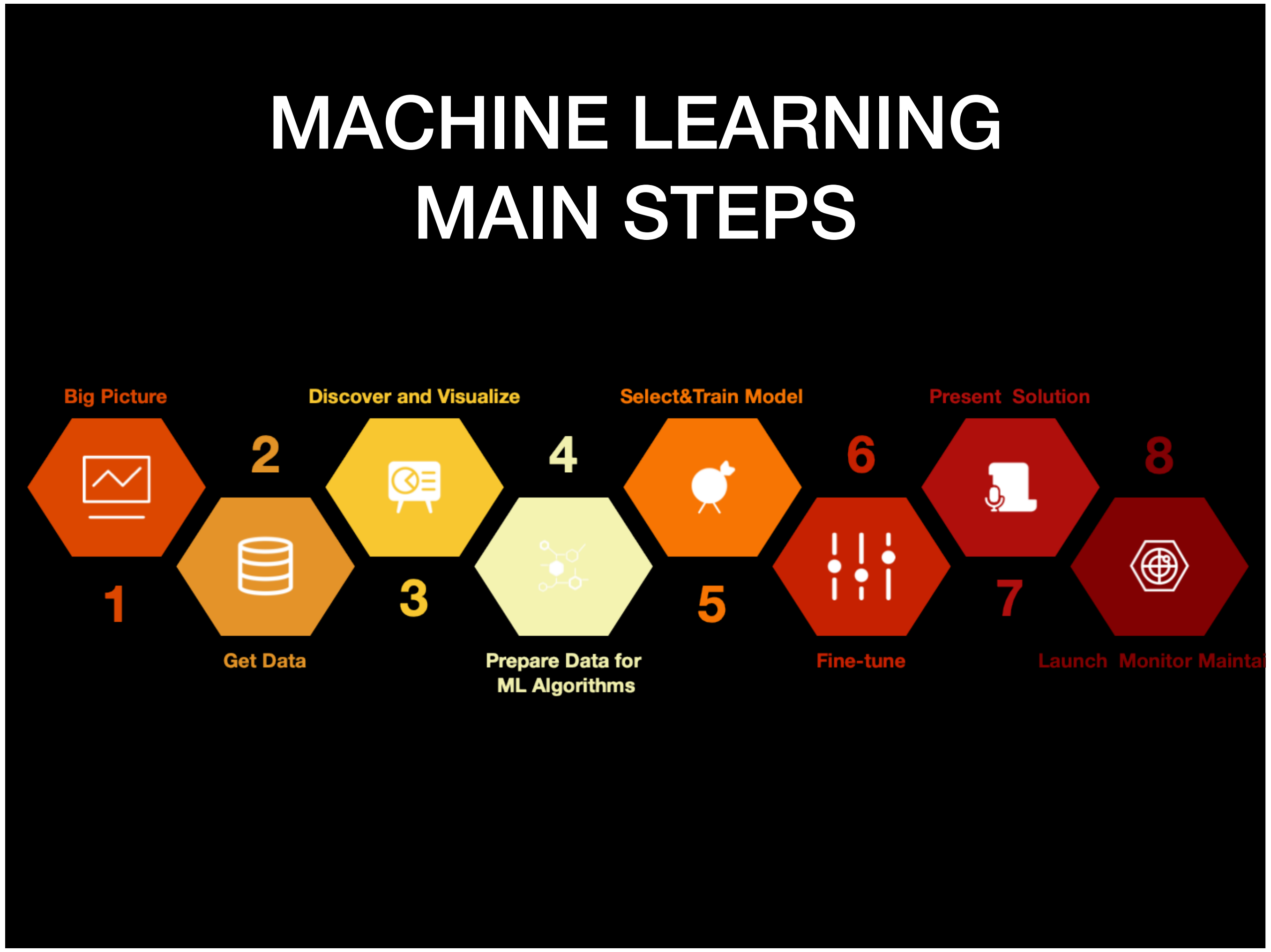
Identify Business Insights
- Return back to the business problem

Visualize Your Findings
- Keep it simple & priority-driven

Tell a Clear & Actionable Story
- Effectively communicate to non-technical audiences

This is the whole framework of completing a data science project.

Data science work-flow is a non-linear, iterative process, and that there are many skills and tools required to cover the full data science process.



MACHINE LEARNING MAIN STEPS

1 Big Picture

2 Get Data

3 Discover and Visualize

4 Prepare Data for ML Algorithms

5 Select&Train Model

6 Fine-tune

7 Present Solution

8 Launch Monitor Mainta

This slides is about machine learning main steps which is a more detailed process.

First you should have a big picture of the whole project, and then we get the data, explore it and get it prepared for ML algorithms.
When we select the right model and train the model, next step we fine-tune the parameters in the model to get a more accuracy result and more efficient calculating process. If you don't know any althgorithms and don't know what does it mean, that's totally fine, we will show you more in next several courses.

After we present solution, we collect opinions and thoughts, and repeat the whole process. Then we launch it, monitor operating data, and iterate promotions.

So the whole process is a never-end process. We collect information from different sources to support us to improve our system.

Actually this is the process in which how product manager manage the life circle. Remember what I talked before? After we get an idea, we find data to support our idea is actionable, is benefitable, and then we design the system to perform this idea, discuss this plan with co-workers from different departments, including risk control, technology, user design and so one. After successfully launch the system, we collect operating data and user feedback to make promotions.

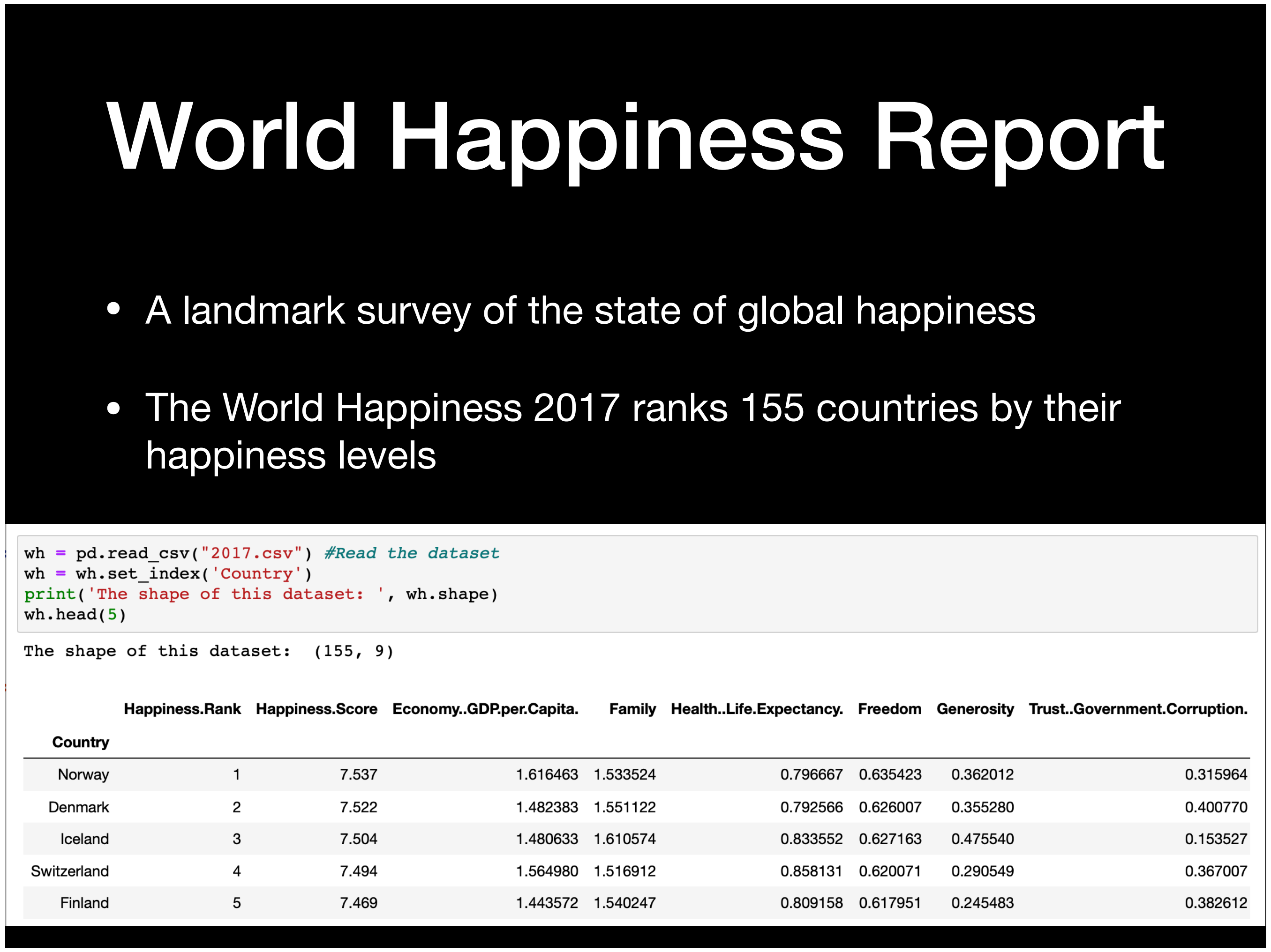
As a product manager, we try to mine the deeper demand of our users and satisfy them. As a data scientist, we try achieve this goal based on solid data.

AN EXAMPLE

So, to help you better understand the process, I'll illustrate a project by doing some basic visualization. Let's continue with happiness score data.

Let's assume that we try to understand the relationship between happiness score and other factors, and try to have a conclusion how to improve happiness score of people.

We've already have the whole picture, and we know to problem we try to solve. Let's move to step 2, collect raw data.



World Happiness Report

- A landmark survey of the state of global happiness
- The World Happiness 2017 ranks 155 countries by their happiness levels

```
wh = pd.read_csv("2017.csv") #Read the dataset
wh = wh.set_index('Country')
print('The shape of this dataset: ', wh.shape)
wh.head(5)
```

```
The shape of this dataset:  (155, 9)
```

	Happiness.Rank	Happiness.Score	Economy..GDP.per.Capita.	Family	Health..Life.Expectancy.	Freedom	Generosity	Trust..Government.Corruption.
Country								
Norway	1	7.537	1.616463	1.533524	0.796667	0.635423	0.362012	0.315964
Denmark	2	7.522	1.482383	1.551122	0.792566	0.626007	0.355280	0.400770
Iceland	3	7.504	1.480633	1.610574	0.833552	0.627163	0.475540	0.153527
Switzerland	4	7.494	1.564980	1.516912	0.858131	0.620071	0.290549	0.367007
Finland	5	7.469	1.443572	1.540247	0.809158	0.617951	0.245483	0.382612

This is the dataset we collect for solving our problems, called world happiness report.

First, let's have a basic idea what is world happiness report.

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th.

Next, I used python code to read csv file (it's just a form of file, like excel, word). And show top 5 rows of this data frame to take a look.
We can see that this dataset contains 155 countries happiness score, and 9 columns to describe one countries, including Happiness rand, happiness score, GDP, family extra.

The screen short is not complete, let's go to the code workspace to take a look.

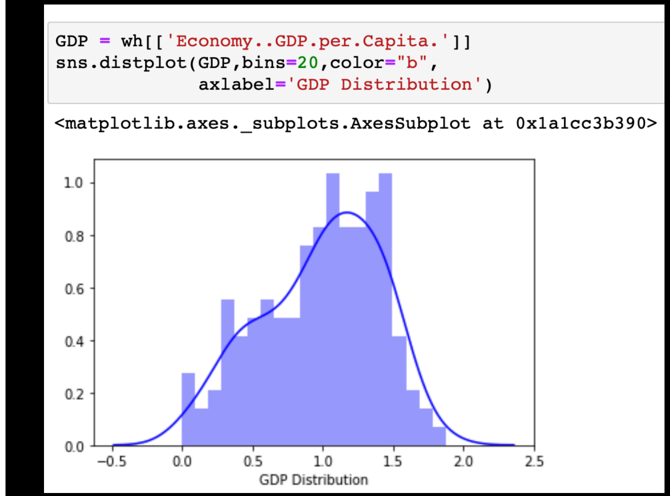
What I am writing is python.
First, I import some packages. A package is a collection of Python modules. The distinction between module and package is just at the file system level. Different packages contain different moduls for various purpose. Import time for time processing, import numpy to deal with data manipulation.

Let me explain about these variables for you.
Happiness Rank: Rank of the country based on the Happiness Score.

Happiness Score: A metric measured in 2015 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest.”
life expectancy: in terms of healthy years
Trust: trust in government and business
Freedom: perceived freedom to make life decisions

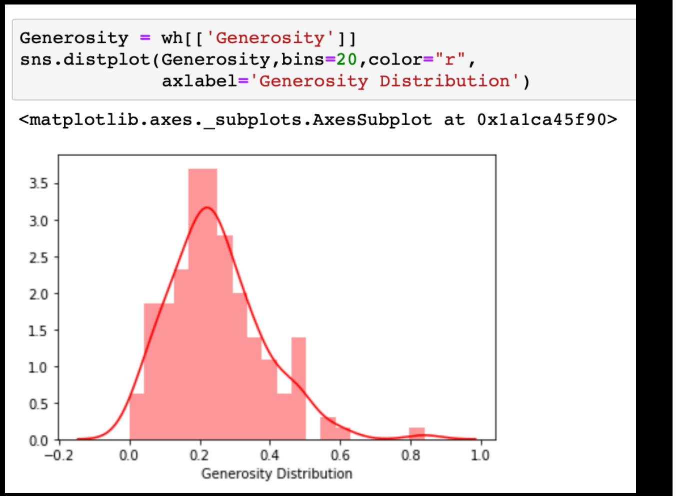
For step 3, We checked missing values, outliers and our data is clean now. Let’s move to step 4. Explore the data.

Distribution



```
GDP = wh[['Economy..GDP.per.Capita.']]
sns.distplot(GDP,bins=20,color="b",
             axlabel='GDP Distribution')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1cc3b390>
```



```
Generosity = wh[['Generosity']]
sns.distplot(Generosity,bins=20,color="r",
             axlabel='Generosity Distribution')
```

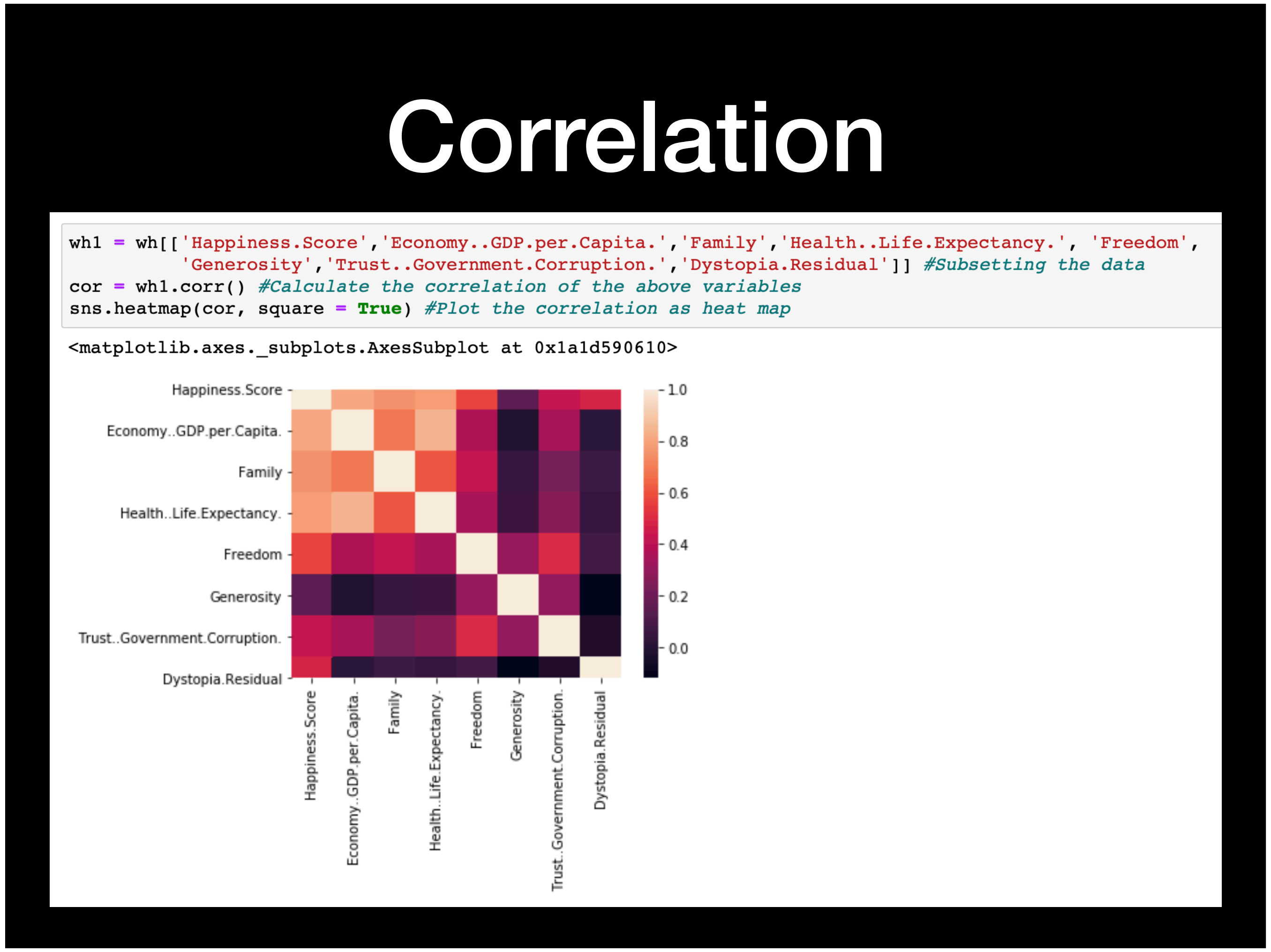
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1ca45f90>
```

- GDP distribution is left-skewed
- Generosity distribution is right-skewed

Let’s look at the distribution of two factors, GDP and Generosity.
GDP is left skewed which means mean is less than the mode, while generosity is right skewed, which means mean of generosity is more than mode.

Skewed data arises quite naturally in various situations. Incomes are skewed to the right because even just a few individuals who earn millions of dollars can greatly affect the mean, and there are no negative incomes. In skewed data, the tail region may act as an outlier for the statistical model and we know that outliers adversely affect the model’s performance especially regression-based models. So there is a necessity to transform the skewed data to close enough to a Gaussian distribution or Normal distribution. This will allow us to try more number of statistical model.

I will talk about how we transfer from this kind of data later.



Correlation

```
wh1 = wh[['Happiness.Score','Economy..GDP.per.Capita.','Family','Health..Life.Expectancy.', 'Freedom',
          'Generosity','Trust..Government.Corruption.','Dystopia.Residual']] #Subsetting the data
cor = wh1.corr() #Calculate the correlation of the above variables
sns.heatmap(cor, square = True) #Plot the correlation as heat map
```

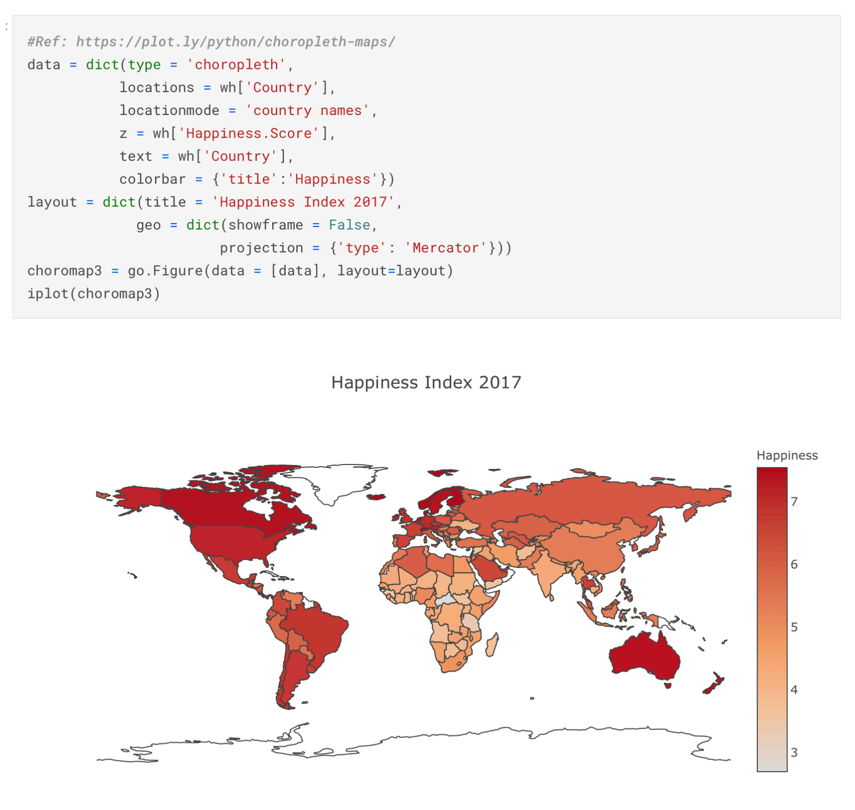
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1d590610>
```

And we will try to understand the correlation between few variables. For this, we compute the correlation matrix among the variables and plotted as heat map.

The color palette in the side represents the amount of correlation among the variables. The lighter shade represents high correlation. We can see the line here has the lightest color, that’s because the factor is completely correlated with itself

We can see that happiness score is highly correlated with GDP per capita, family and life expectancy. It is least correlated with generosity.

World Map

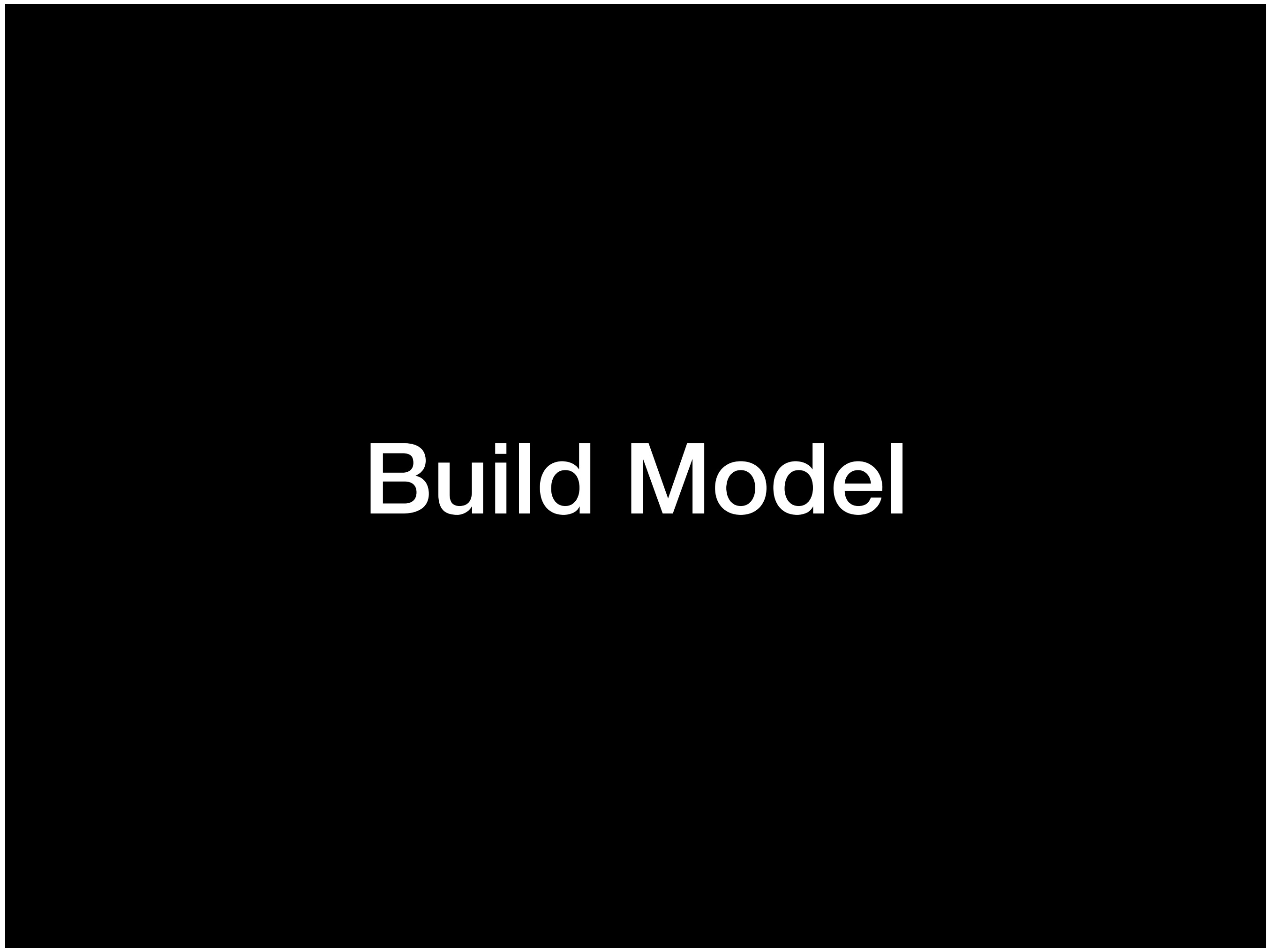


```
#Ref: https://plot.ly/python/choropleth-maps/
data = dict(type = 'choropleth',
           locations = wh['Country'],
           locationmode = 'country names',
           z = wh['Happiness.Score'],
           text = wh['Country'],
           colorbar = {'title':'Happiness'})
layout = dict(title = 'Happiness Index 2017',
             geo = dict(showframe = False,
                       projection = {'type': 'Mercator'}))
choromap3 = go.Figure(data = [data], layout=layout)
iplot(choromap3)
```

For the happiness score world map, we can have a big picture of how this score distributed among the world.

The color scale in the side represents the happiness score scope. The darker means the score is higher.

We can notice that happiness score of African countries are obviously lower than that of other countries. The happiest countries are concentred in European countries, Canada, united states and Australia.

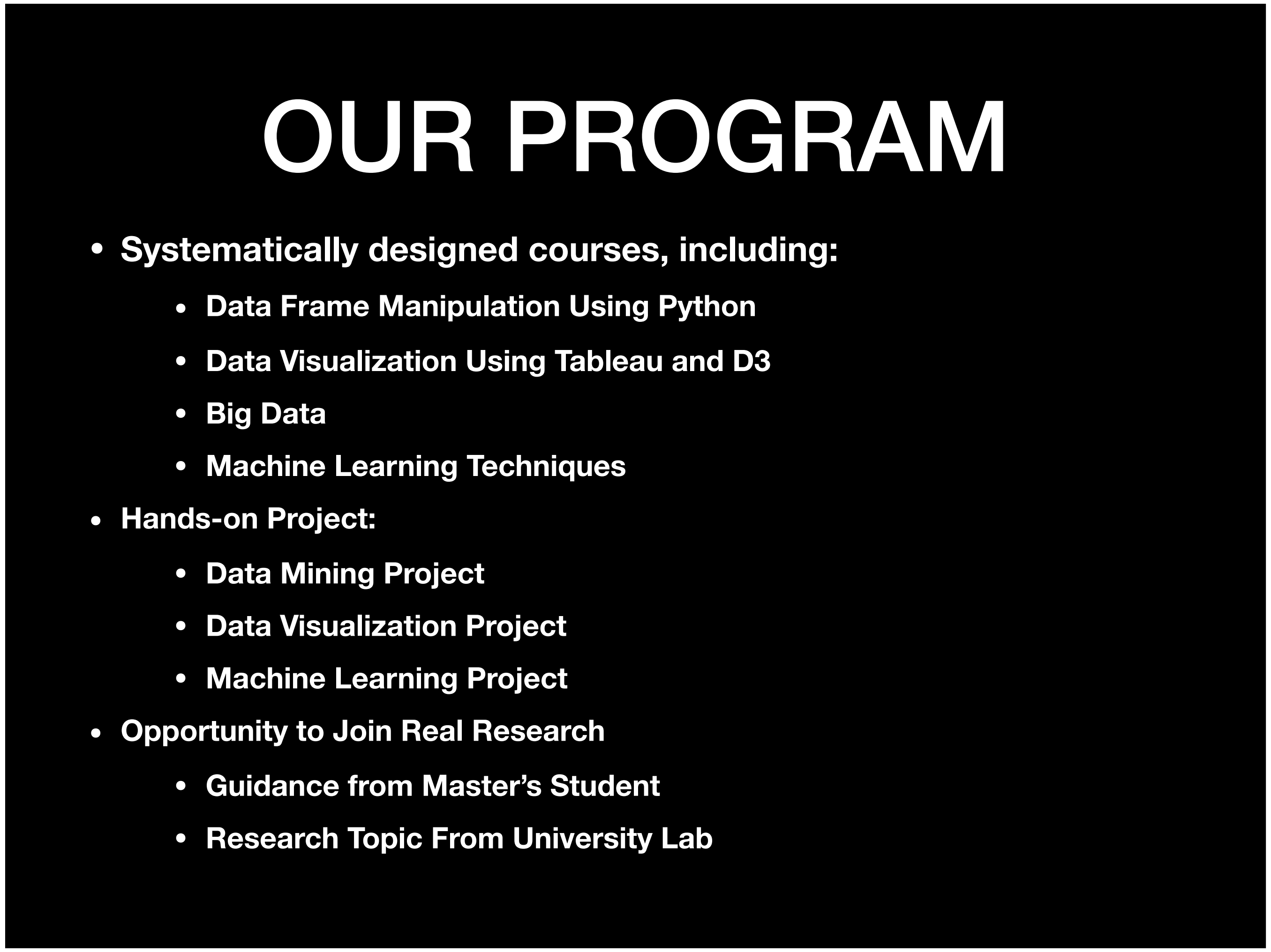


Build Model

After we have learned the scope of the dataset, we move to step 5, we try to choose the most suitable model to explore which factor or factors contribute most to happiness score.

Since this is our first class, I don’t want to flush you with too much information. ’ll leave this part in later section.

In the last section, let me explain a little bit what can I provide for you guys.



OUR PROGRAM

- **Systematically designed courses, including:**
  - **Data Frame Manipulation Using Python**
  - **Data Visualization Using Tableau and D3**
  - **Big Data**
  - **Machine Learning Techniques**
- **Hands-on Project:**
  - **Data Mining Project**
  - **Data Visualization Project**
  - **Machine Learning Project**
- **Opportunity to Join Real Research**
  - **Guidance from Master's Student**
  - **Research Topic From University Lab**

So let me introduce our program. All in all, this program will provide with three part: Systematically designed courses, Hands-on Project and the Opportunity to Join Real Research.

For the courses, All the courses mentioned above get you ready to join real research

The courses Include Data Frame Manipulation Using Python, Data Visualization Using Tableau and D3, Big Data and Machine Learning Techniques.
For hands on project, we design 4 projects for you to implement all the techniques you learned in the courses to let you gain some practical experience, not just knowing some concepts.

If you pass the final exam, we only have one exam, you have the opportunity to join ral research with guidance from maters students and all researches are from university lab. The topics are of different field such as health care, medical science or archeology. We will present you several choices at the end of whole courses to let you choose from, as long as you were qualified.

CONTACT ME

xiaoqiansun1992@gmail.com

If you have any question, you are welcome to contact me.

I would like to try my best to answer your questions. Hope you will also fall in love with data science. Bye guys.