

It's not who has the **best algorithm** who wins, it's who has the **most data**.

- Andrew Ng



Managing Machine Learning Projects with Google Cloud



Offered by [Coursera](#).



Previously this course was known as **Machine Learning for Business Professionals**.



Syllabus Week 1: **What is ML?** 1 - 11 | Week 2: **Employing ML** 12 - 26
| Week 3: **Discovering ML Use Cases** 27 - 37 | Week 4: **How to be successful at ML** 38 - 47

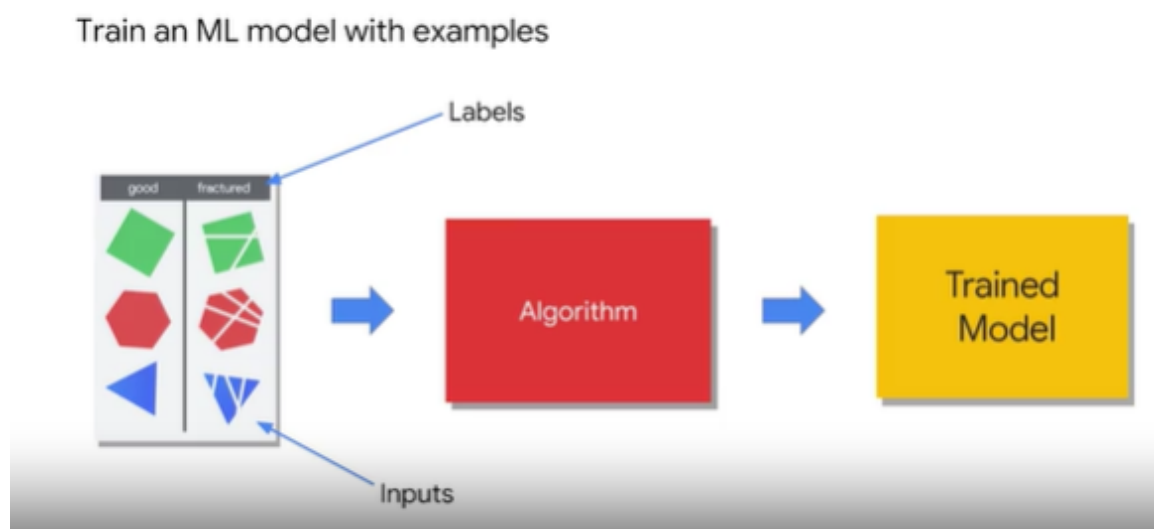


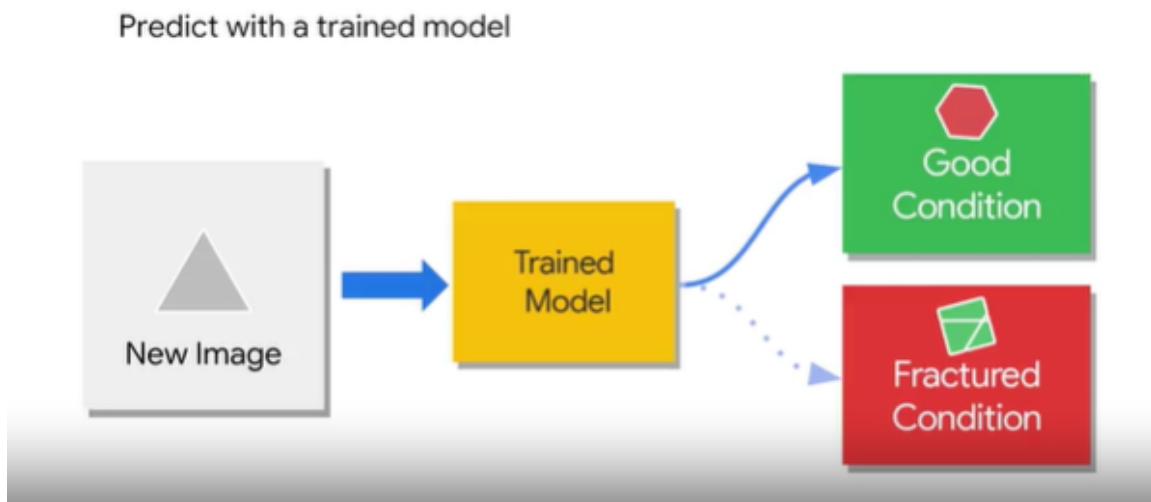
Teachers Carolyn Ujcic, Lak Lakshmanan, Valentine (Val) Fontana



The formatting guidelines of this note is given on the end page.

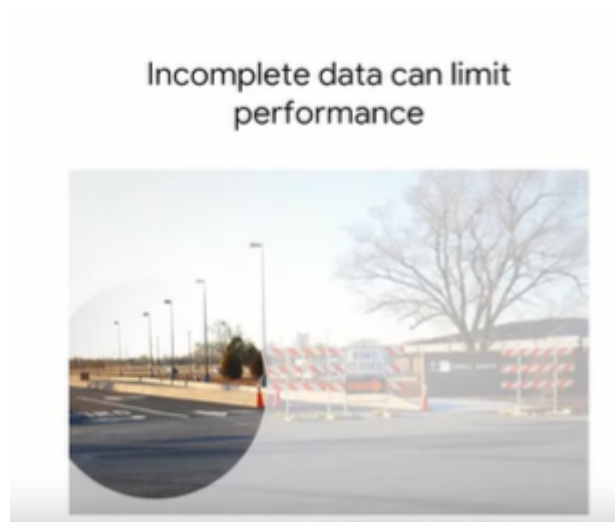
1. **Def: ML (Machine Learning) is a way to get predictive insights from data to make repeated decisions.**
2. **Backward looking vs forward looking data:**
 - Most analytics in your business is probably backward-looking, where you look at historical data to calculate metrics or identify trends.
 - Predictive analytics are forward looking.
 - A business analyst's reviews a report and sees that the demand is increasing for a specific product in a specific region. The analysts then suggests a new price for that product in that region to increase profit. Now, the business analyst is making a predictive insight, but is that scalable? Can that business analysts make such a decision for every product and every region? Can they dynamically adjust the price every second based on how many people want that item at that very instant?
In order to make decisions around predictive insights repeatable, you need machine learning.





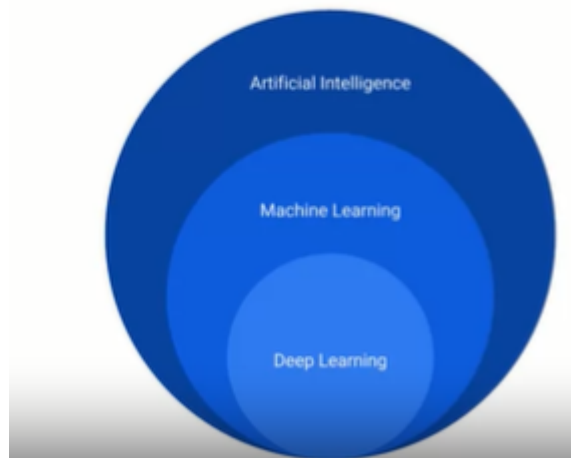
standard algorithms - these algorithms exist independently of the use case.

3. ML models aren't **programmed** with traditional logic, like if, this, then, that. Instead, you program them by giving them examples and letting them come up with the logic. Models will be as good as data.
4. **Qualities of good data**: broad coverage, clean & complete. The more wrong data (dirty data) you have, the more correct data you will need to provide to counterbalance.



5. **Difference b/w ML and AI**

ML is a type of AI



▼ What are expert programs?

AI programs designed to use if-then rules to emulate human inference and decision-making. They work well when you already know the logic. To create them, you first collect all of the domain of knowledge like a table of diseases and their symptoms. Then you convert this knowledge into if-then rules. The problem with these experts systems is that there aren't that many use cases where you can fully capture all the logic. In medicine, textbooks capture a lot of information for a doctor to make a diagnosis, but they don't capture everything. However, one area where it can be successful is in process control. In Japan, their trains have excellent on-time performance. Researchers in Japan have created expert systems that will adjust train schedules when there's a disruption in the same way the best human operators would. Another challenge with these systems is that it is expensive to code up that logic in the form of programs. Many times the logic is just a heuristic, and you spend years maintaining and fixing problems in the heuristic, as you discover the cases where it doesn't work. Why is ML a better tool than expert systems in many cases? Well, let's say you have a record of customer browsing activity on your shopping site, but you can't describe the logic for determining whether a customer will buy. For example, your logs might indicate that some small portion of people who looked at sunscreen ultimately bought sandals, and none of the people who looked at towels did. With ML, the model can learn based on customers' activity, how likely they are to buy sandals. Many business problems like this one are a better fit for ML because it is hard

to articulate the logic. But we do have examples of what actually happened historically.

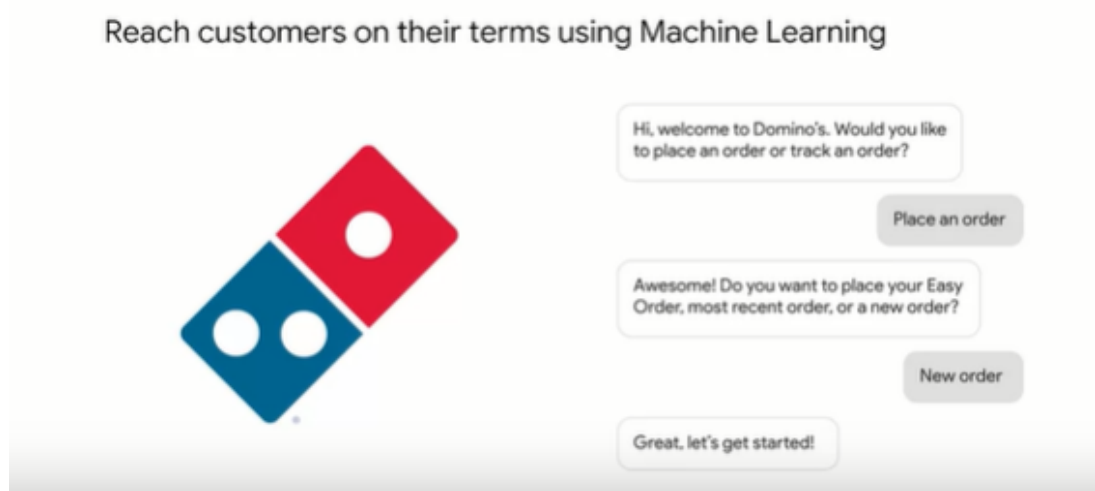
6. **Traditional** BI dashboards and data analytics - historical data, (vs) **predictive analytics** - unknown data.

ML works for making repeated decisions, not occasional ones. As you think about applying ML in your business, I want you to remember this part of the definition. It is often where you will get the most business value. That's because many of the benefits of ML, including scalability and efficiency, are only realized when you use it repeatedly.

Key ingredients for a recipe for ML: data, standard algorithms, predictive insights, and repeated decisions.

Where can we use ML? Quality control of a Space Shuttle? [less data] What about an annual sales forecast? [not frequent, also many statistical tools present] What about trying to model the likelihood that a truck might break down? [infrequent] But what if I told you that each day, mechanics decide whether they should inspect vehicles for issues to find if the truck is likely to break down? Then it becomes a frequent decision? We can use ML here.

7. Normal life **eg**: Do you get less spam in your inbox, or have you used a voice assistant to call your mom? Or search within Google Photos to find the best picture of your dog, or you might have watched that YouTube video that was recommended to you on the YouTube homepage. Think about your business processes. Where do you have major pain points? How might ML help?
- Aucnet built a real-time car image recognition system powered by Google Clouds ML technologies which reduced the time it took to list a car for auction from 20 minutes to only two to three minutes.
 - Using Google's Dialogflow chat bot technology, Domino's was able to improve the ordering experience even more for their customers. The bot can now handle more difficult or complex orders from customers.

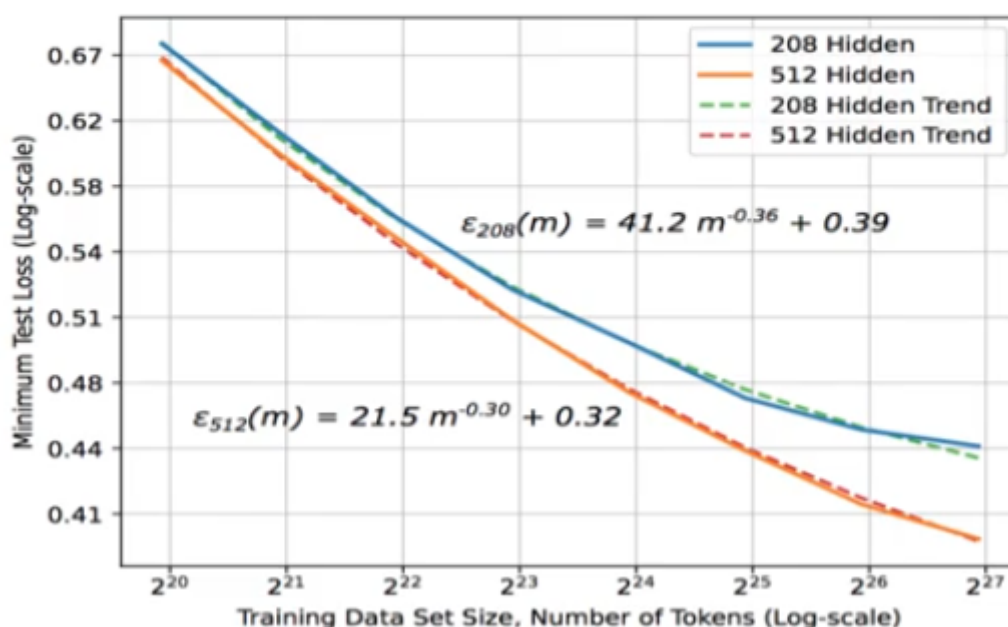


8. [Lab: Creating a Pizza bot with Diagflow and Cloud Image Classifier with Auto ML \(Vision API\).](#)

The **Vision API** is an API that uses machine learning and other Google services to extract information from images. It can do a variety of things, whether it's detecting the presence of certain classes within the image. Detecting and extracting text, determining if the image is safe to serve, or detecting the presence of corporate logos. Unlike the Vision API, **AutoML** Vision can be trained to use whatever classes you need for your business use case. You provide the labelled datasets.

9. Machine learning has been around since the 1970s, when it was used in fields such as astronomy. [So why the popularity now?](#) PART 1:

ML requires exponential increases in data size



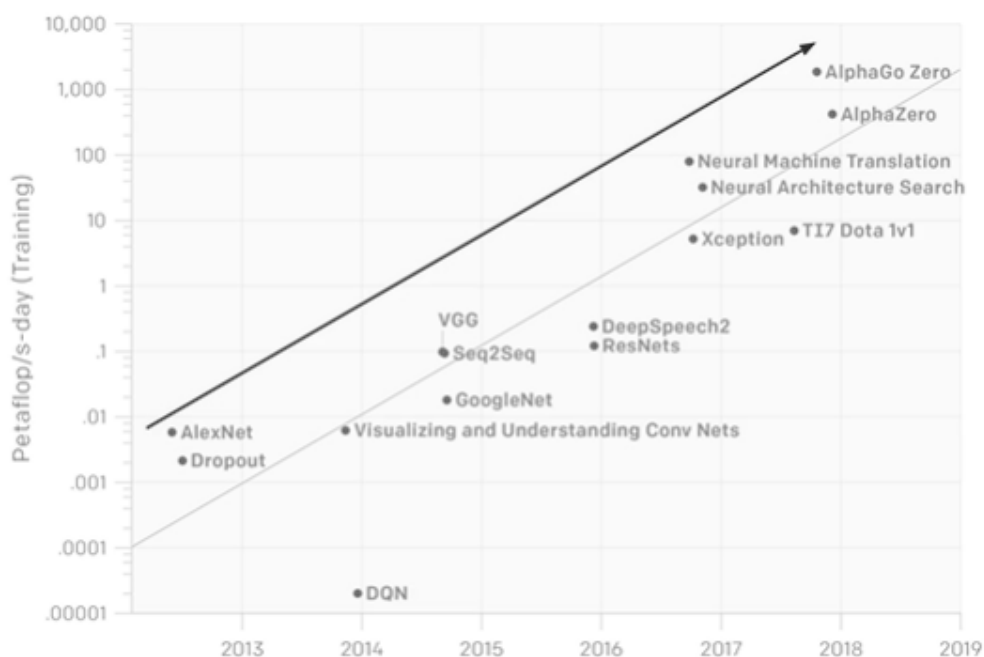
What made it ready was the convergence of 3 factors:

- Increasing availability of data
- Increasing maturity and sophistication of ML algorithms
- Increasing power and availability of computing hardware and software

The [cost of sending data](#) over the Internet in the US dropped from \$1,200 per megabit per second in 1998 to \$75 in 2005, that's a 94% decrease. During the same period, the [cost of storage](#) also fell significantly. Not just for disk storage, which is the sort of storage necessary for retrieving archived information, but also for RAM memory, which is the sort of storage necessary for big and fast applications. In 2005, the cost of 1 gigabyte of disk storage was 0.2% of what it was 10 years earlier, and the cost of RAM was less than 0.1% of what it was 10 years earlier. Then in 2007 the [iPhone](#) launched, bringing the desktop web to phones. Users started consuming and generating data constantly.

10. PART 2: There are also [changes related to algorithms](#). There have been lots of ML algorithms before. **Linear regression, decision trees, genetic**

algorithms, support vector machines, just to name a few. One in particular has revolutionized machine learning recently, neural networks. But for complicated mathematical reasons, those deep layers of neurons didn't work in computing. So people used **shallow neural networks**, which weren't very powerful. But then came advances in what we call **deep neural networks**. Many advances were made by Geoff Hinton at the University of Toronto. He and his team of students got deep learning to finally work, with the help of many other researchers. These deep neural networks are examples of **deep learning**, which propelled ML forward even further. One key advance was a special type of layer in a neural network called a **convolutional neural network**, or **CNN**. **Distributed computing** allowed large computing jobs to be split into manageable pieces. But distributed computing is subject to diminishing returns because there are network, speed, and reliability issues when computers have to talk to each other. **GPUs**, which are built as graphics cards for rendering and games, turn out to be really good at accelerating machine learning computations. Andrew Ng demonstrated in 2006 that GPUs function as ML accelerators. Between AlexNet, which started off the deep learning revolution and Neural Architecture Search, which powers Google's AutoML, the computational cost of training grew 100,000 times greater. That's because when you have more layers in a deep learning model, you need more data and your compute needs become more expensive.



Compute performance has hit a plateau (disobeying **Moore's Law** because in the past three years, growth has slowed dramatically as manufacturers run up against fundamental limits). One solution is to limit the power consumption of a chip. You can do that by building what are called application-specific chips, or **ASICs**. Google designed new types of hardware designed specifically for ML. The tensor processing unit, or **TPU**, is specifically optimized for ML. And it has more memory and a faster processor. Google has been working on the TPU for several years. And has made it available to other businesses.

Because of all of this technology, new hardware, distributed computing, new algorithms, it became feasible for more businesses to use ML, but it still wasn't easy. Part of the problem was that few ML software frameworks were designed for industry scale usage. If companies wanted to do ML, they would need the expertise to create their own frameworks. And that was a steep cost, especially because ML talent is still relatively rare. But over time, more and more ML frameworks are becoming available and are making it possible to use ML at high levels of abstraction. One such framework is TensorFlow. But even with frameworks like TensorFlow, implementing a custom model requires technical expertise. That's where **cloud technology** comes in for businesses that want to run ML models they already have in the cloud, but for whom the economics of owning the infrastructure doesn't make sense. Additionally, TPUs scale linearly, so you can train faster by running on more cores, which is another benefit of the economics of the cloud.

Reasons to do ML in the cloud

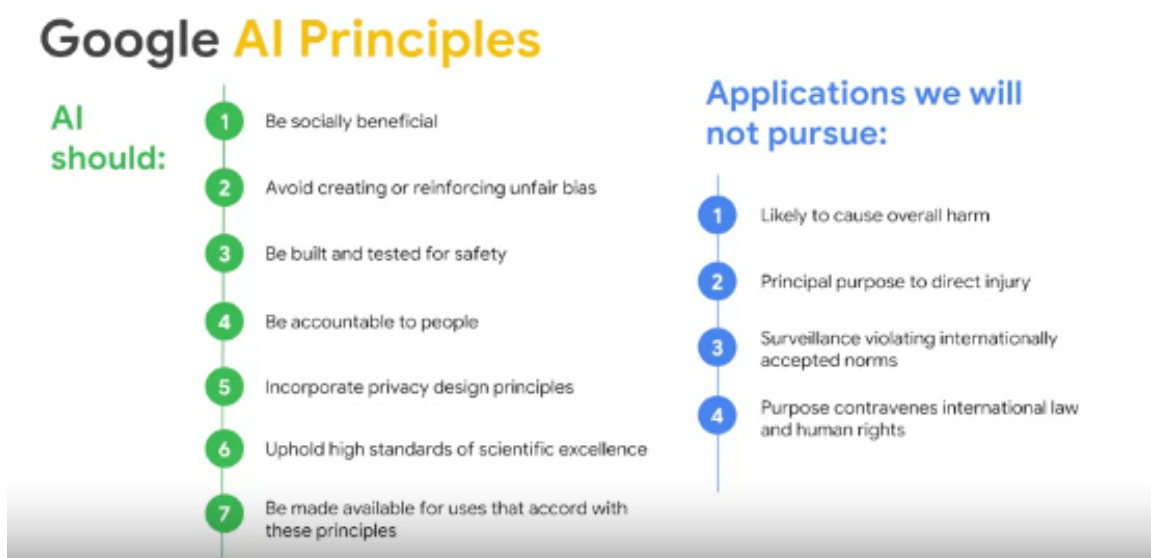
- IaaS allows businesses to rent instead of buy
- TPUs allow for linear scaling
- PaaS lets businesses think at a higher level of abstraction
- SaaS lets businesses concentrate on delivering value to their users

Reasons you may have to do ML without the cloud

- Data locality requirements
- Need to make predictions offline

A common paradigm is to run a fast but not great model locally. But use the cloud when a more accurate answer is needed. Thus, the cloud is recommended for most ML use cases.

11. Using AI Responsibly - Google Principles



Some ways to put these principles into action:



2 Avoid creating or reinforcing unfair bias

"AI algorithms and datasets can reflect, reinforce, or reduce unfair biases. We recognize that distinguishing fair from unfair biases is not always simple and differs across cultures and societies. We will seek to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief."

[2 from Week 2]

- Machine learning involves learning from examples. An example consists of **labels** and **features**. A label is just the true answer for an input. If you're trying to predict how much a customer in a bank will deposit in a given year, the amount that was actually deposited by that customer in that bank in a particular year is the label for the example. When we train a model, we compare the prediction for that example against the label for that example and use that information to update the model. [Phases of ML 13 - 19]



- Labels can be either numbers or they can be categories. The amount that somebody deposits in a year, for example, that's a number. Whether that amount is high or it's low, that's a category.

Regression vs. classification

Label Data Type	ML Problem Type	Example
Categorical	Classification	High vs. Low
Numerical	Regression	Dollars

The features are the inputs to the model or a distinctive attribute. There are other forms of machine learning that don't require a label, but they're less mature. In the case of images, the pixels of the image might be enough as features.

14. [Choosing the wrong label can cause problems.](#)

Choosing the wrong label can be problematic

- 1 Not everyone has the same amount of data, so it's hard to build a single table for everyone
- 2 Changing averages
- 3 Non-representative data

Aucknet

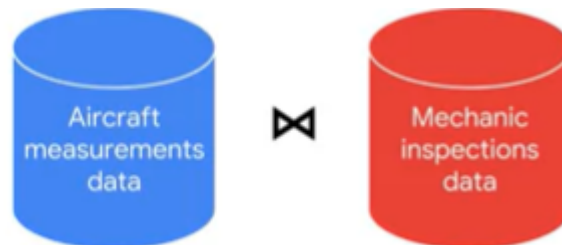
- feature - historical data (pictures) on condition of Tyre
- label - H.D data on price sold for
- regression problem

Youtube

- feature - country, how many of the people who watched this current video went to watch the next video, aggregate history of all of it to one billion plus users

- label - next video (H.D) because we identify videos by a video ID, the video ID is a label.
- classification problem [see: Week 2> ML Terminology in Context]

15. [Acquiring labelled data](#) - need to have all required data in one place.



Ways to label your data

- 1 Use a proxy label
- 2 Build a labeling system
- 3 Use a labeling service

AutoML offers labeling service.

16. Thus summarizing,

Formulating the ML problem

- 1 Choosing input features
- 2 Get labels
- 3 Choose an objective

17. Let's say you're running a retail website, where you're selling lots of different items. surprise: items that user might not have otherwise bought, clicks: novelty.

Quiz

Match up an objective that the model is optimizing for and the resulting outcome for retail.



There is a difference between popular in the sense that things are rated highly and popular (4) in the sense that people buy them a lot (2). The hardest part of a recommendation system is not the algorithm because the algorithm is standard. It's not the data because you probably have purchase transactions data already. **In this case of a recommendation system, one of the hardest parts is getting the objective right.** In reality, the objective that most retailers pick will be a weighted combination of all these factors. The retailers will typically tweak the relative weights as the company strategy changes. There is no right answer, it's all a trade-off. So discuss what the right objective is with your business stakeholders.

18. When we [train a model](#), we compare the prediction of that model against the label and use that information to update the model, and that part is basically just applying the standard machine learning algorithm to your data. Although it's possible to train a machine learning model only one time, **the best-performing machine learning models are trained continuously** by taking a trained model from earlier and retraining it with new data. That's because the world is always changing and only recent data will have a record of these newer changes. If you've built a workflow where humans label the data, then you've done half the work already. The next step is connecting their output to the model. Not only that, you set yourself up for success by keeping humans in the loop. Many people assume that this isn't

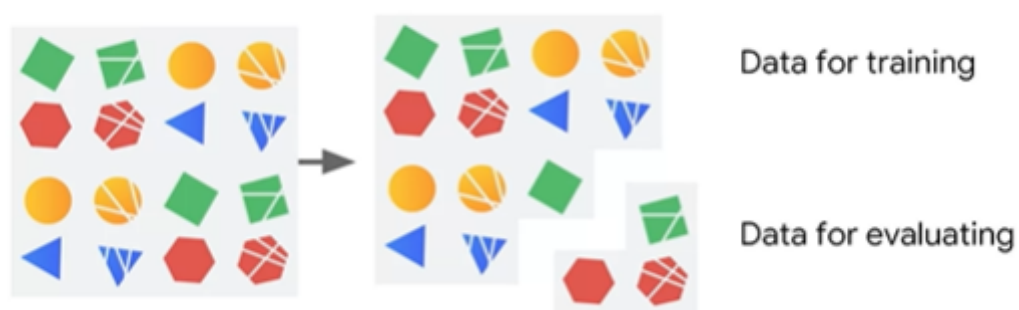
unnecessary, but **for great machine learning, it absolutely is necessary to have humans in the loop.**

Here are some standard algorithms to choose from for ML

ML task	Standard algorithm
Image classification	ResNet (originally by Microsoft Research, an implementation open-sourced by Google)
Text classification	FastText (open-sourced by Facebook Research)
Text summarization	Transformer (open-sourced by Google)

For business problems, mostly you will probably use standard algorithms. People create new algorithms all the time and they compare them with previously known algorithms, but the improvements are very, very minor.

19. [How to get an idea on how good the model is?](#)



We don't show the model all of the data that we have (why? original training data can't evaluate - model memorize), instead we split that data into two parts. Maybe we show the model 80 percent of the data (**training data**), and have it learn from that, and then we try out the model on the remaining 20 percent (**test data**). Create a **confusion matrix**. Accuracy can be computed simply by finding the fraction of times that the model is correct.

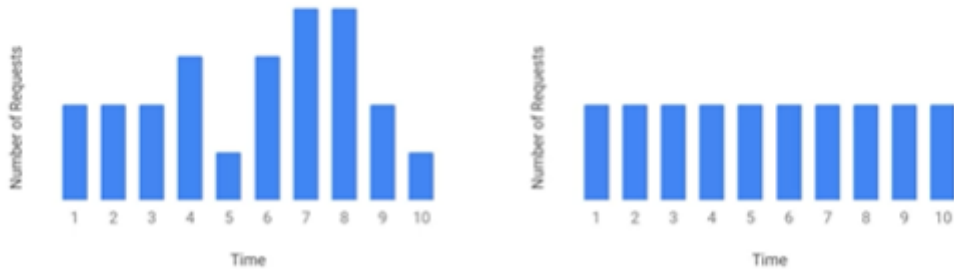
Confusion Matrices measure performance relative to expectations for Classification



Usually accuracy is enough, but in some cases, you will be creating a machine learning model for rare occurrence. For example, you may be building a model to identify credit card fraud. The vast majority of transactions are not fraudulent. So, even if the model says, nothing is ever fraudulent, it'll only get a very small fraction wrong. The accuracy will be something like 99.999 percent. In such cases, we use other metrics computed from the confusion matrix. Matrix that emphasize the performance on the rare stuff.



Interactive and periodic prediction

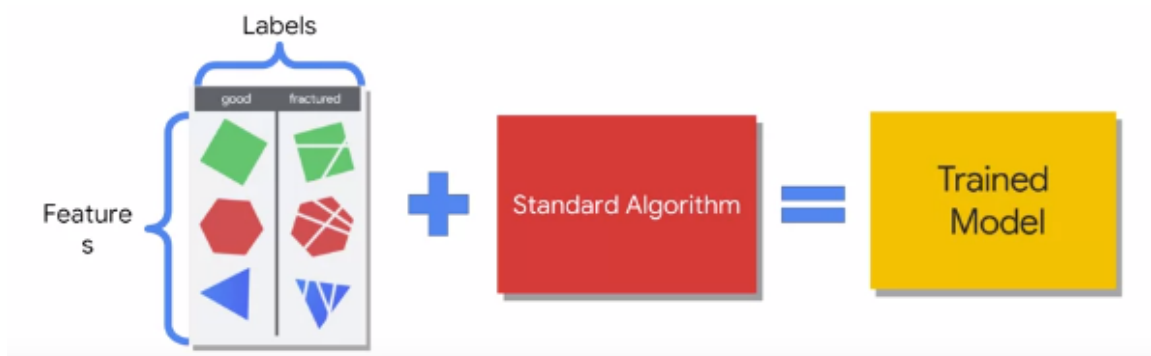


20. Make different models [combination - diff features, complexity (neural n/w layers)] and select what works best.

ML Best Practices

- 1 ML involves experimentation
- 2 Start Simple
- 3 Don't use your test data during experimentation
- 4 Do pilot projects with end-users

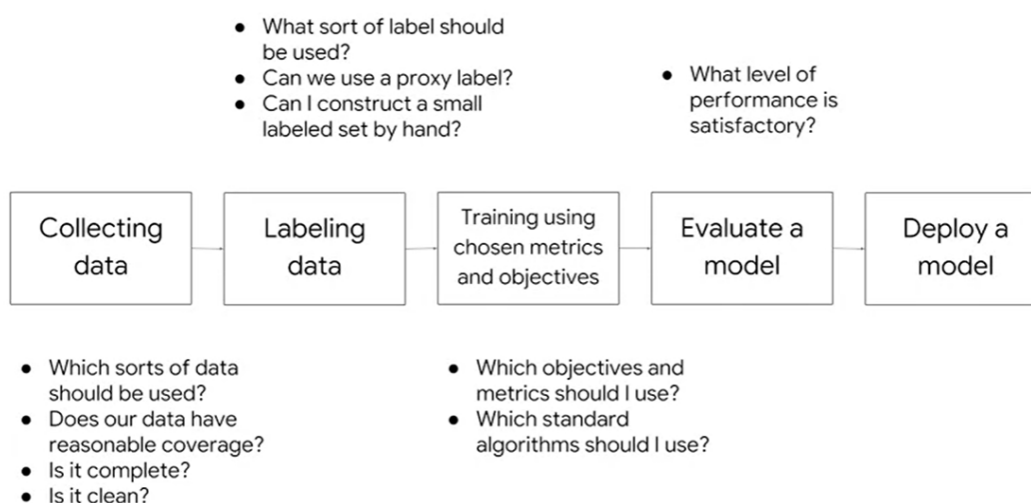
21. The first stage of training looks like this.



The process of evaluation & deployment looks like this.



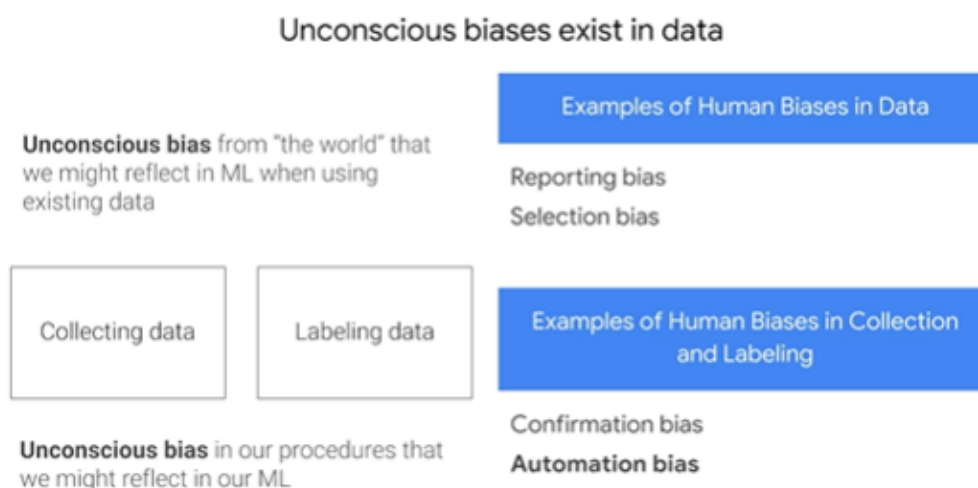
The more complete process looks like this.



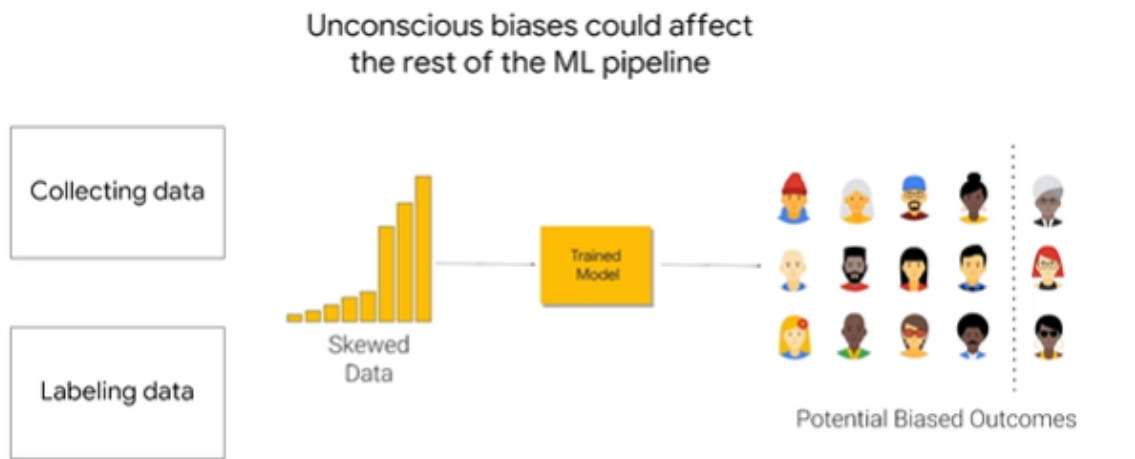
What's common to all of these questions? **They're decided by you**, the creator of the model. Machine learning is not automatic, it is highly driven by the decisions that we make in terms of how we train the machine learning model. The human impact doesn't stop with you. Maybe the model is selecting which customers are higher priority. So who's calls you're going to answer first? Maybe the model is selecting which products you're going to recommend to a specific customer. Maybe the model is selecting which transactions are likely to be fraudulent, and that you are going to examine further. Maybe the model is selecting which trees are being cut illegally. The end customers of your business therefore see some kind of an effect because of the predictions of the model. They're provided with some specific quality of service, or they get some kind of a product, or the wait in the long run queue. They see one product or they see another, they can go through their credit card purchase, or the transaction gets denied. Your end

customers see the impact of the decisions that you make. So the decisions that you make as you do machine learning end up having an impact on the real world because of the way that the predictions of the machine learning model are going to get used.

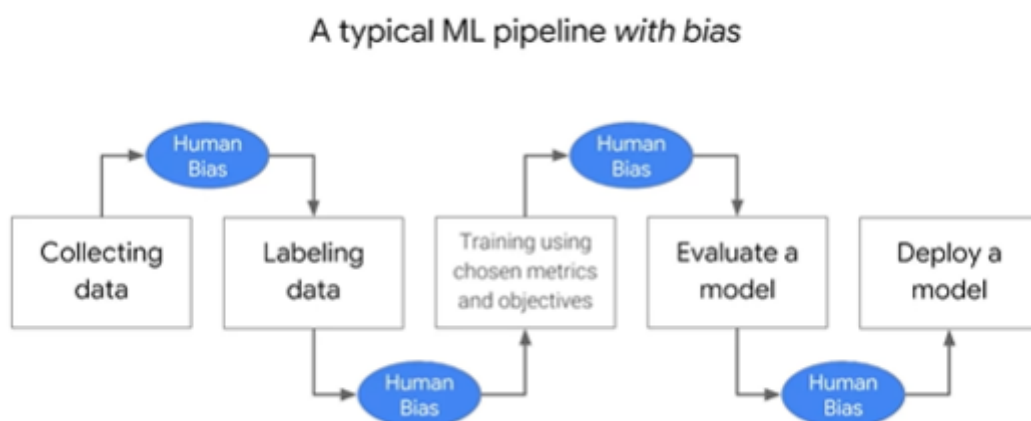
22. More about **bias** in data:



Human biases that exist in data because data found in the world has existing biases with regard to properties like gender, race, and sexual orientation. **Selection bias** happens because the subjects that get into our samples represent a privileged type of user. **Confirmation bias** refers to only looking for data which confirms our hypothesis. *eg :someone who buys a tie would look for men's shirts. So we might restrict our data collection to only men's apparel items. But women do wear ties.* An **automation bias** refers to the biases which crop up when the data we use is just the data that is easily automatable, the data that we can quickly get. *eg: If men's apparel and women's apparel are run by different parts of our organization, it is quite possible that this is why we might have collected only men's shirt sales or a fancy digital data collection system at some locations, but other locations are still sending you faxes.*



So what's the impact of biases in collecting data and labeling data? it affects the entire pipeline. In machine learning, we begin with our data. So the biases in the original data are going to get reflected downstream in our models and consequently, they are going to result in biased outcomes. The result is that biases can appear at every point of the data pipeline. [eg: *model to find fraudulent transactions.*]

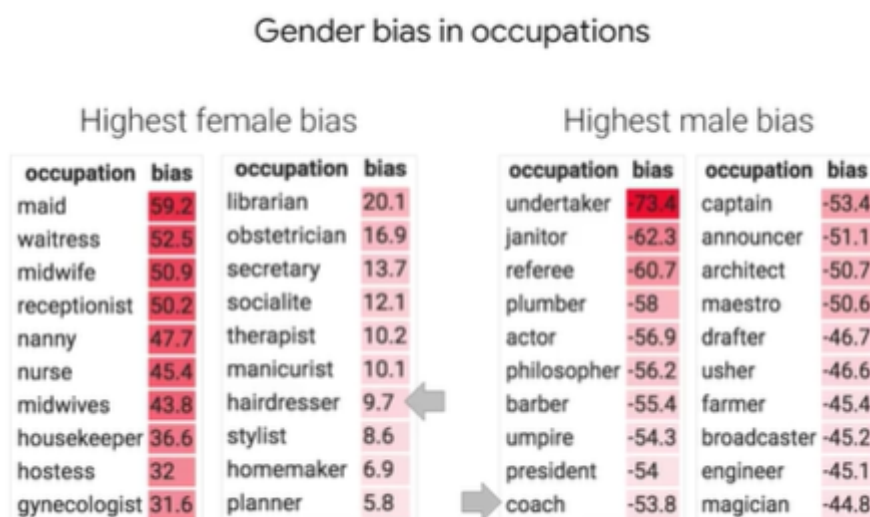


2 Avoid creating or reinforcing unfair bias

ML models learn from existing data collected from the real world, and so an accurate model may learn or even amplify problematic pre-existing biases in the data based on race, gender, religion, or other characteristics.

eg: a model to decide whether or not to make a loan was given

Machine learning models are trained on real world data, you may have removed the gender from your loan application process. But the use of occupation as an input will mean that gender leaks through. Occupations tend to be associated with specific genders. So by using occupation as an input to our machine learning model, that decides whether or not to extend a loan, we may be providing a backdoor entry to the use of gender. This property of machine learning models leads to reinforcing existing biases in society. It's one of the things that we have to be on our guard against. It comes down to understanding our data, really understanding it, and understanding some of the implicit biases that ride along with that data.



23. [Activity: Understanding Learned Relationships using the Embedding Projector Review](#)
24. We've built a model, [how do we know whether the model is fair?](#) One challenge is that there's no one standard definition of fairness. Whether decisions are made by humans or by machines. Both in law and philosophy,

this is a question that engenders lots of debate. This because defining fairness is difficult. eg: *blue & orange people loan case scenario* [see: *Week 2 > Fairness in ML Part 1*]. As in eg. the group-unaware constraint holds that the thresholds are the same. The demographic parity constraint, requires that the same percentage within each subgroup will receive the loan, and the equal opportunity constraint requires that the same fraction within each subgroup, who can pay them off, will receive the loan. Now, none of these definitions is necessarily better than the other. Each of them has different mathematical properties, and you cannot satisfy them all at the same time, and which constraint you choose depends a lot on the context of the machine learning use case. But you've got to choose.

25. Allison Woodruff & her team asked different groups what is the [appropriate definition of unfairness](#)? The result varies greatly depending on who you're speaking to and their ethical, legal and economic priorities. Unfairness, is not immediately obvious, and it requires asking nuanced, social, political and ethical questions. It takes a holistic analysis to consider which definition of fairness is appropriate for your project. They asked different groups, what is the appropriate definition of unfairness? The result varies greatly depending on who you're speaking to and their ethical, legal and economic priorities. Unfairness, is not immediately obvious, and it requires asking nuanced, social, political and ethical questions. It takes a holistic analysis to consider which definition of fairness is appropriate for your project.

Determining the appropriate unfairness definition

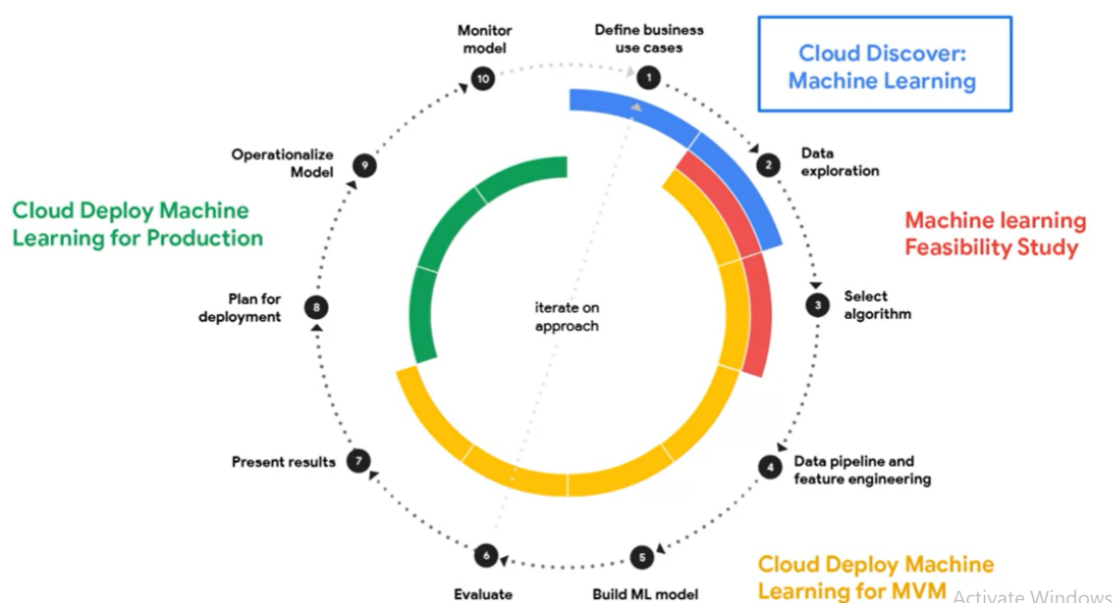
But what is the appropriate definition of unfairness?

- Ethical and moral imperatives
- Legal requirements
- Regulatory risk
- Public relations and brand risk
- User trust
- Agreements with local communities and markets
- ...and many other considerations!

Checklist of data prone to bias related issues.: biometrics, race, skin color, religion, sexual orientation, socioeconomic status, income, country, location, health, language or dialect? For example, zip code or other geospatial data, is often correlated with socioeconomic status, and/or income. Image or video data, can reveal information about race, gender, and age.

26. [Activity: Applying Fairness Concerns with the What-If tool Review](#)

27. [What are the use cases of ML? \[27-32\]](#)



Above figure - processes in Google Cloud Discover

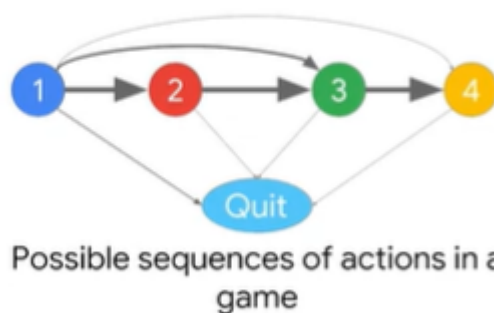
- 1 ML can be used to simplify many rule-based systems
- 2 Automate business processes with ML
- 3 Predict business metrics
- 4 Understand unstructured data with ML
- 5 Personalize applications with ML
- 6 Improve user experiences with ML

RankBrain improved performance significantly



28. [1] The first way to think about machine learning use cases is, as a way to simplify rule-based systems. That was the essential idea behind RankBrain, our deep neural network for search ranking. It outperformed many human-built signals and we could replace many of the hand-coded rules with machine learning. The neural network ended up improving our search quality dramatically. Plus, the system could continually improve itself based on what users actually preferred. *eg: Acme widgets smoke alarms systems.*
29. *eg: Netmarble. Here unusual action is skipping a step to complete a level.*

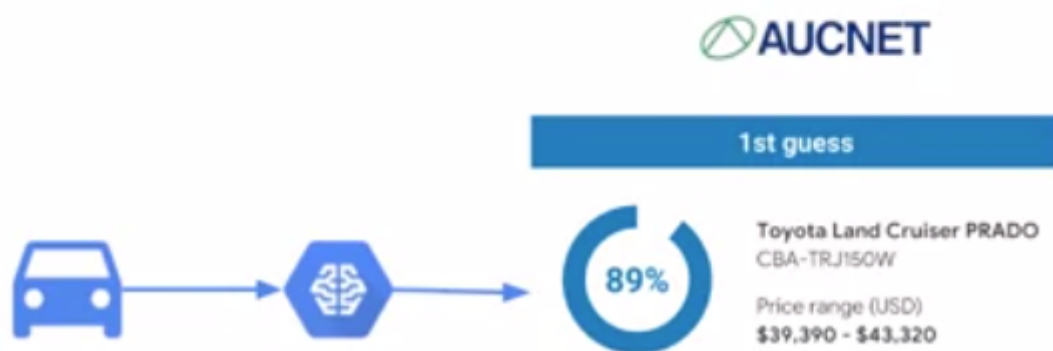
Detecting unusual player actions



So, the key idea here, is to do a pilot project, and that success will embolden other departments to carry out their own machine learning projects. Be the pioneer.

30. [2] eg: Aucknet, Ocado, Ananda Developers

Aucnet reinvents car valuation



The purpose of streamlining a business process is not to reduce costs, the purpose is to reduce drudgery, increase accuracy, gain more visibility into the business. In other words, it's about doing more, not about spending less.

31. [4] Unstructured data is data that can't be easily compared. Language is just one example of unstructured data. Other examples include pictures, videos, sounds, etc. All of these are examples of data that you cannot easily hold in a relational database. Bloomberg aggregates financial news, data, and analytics worldwide, but disseminating that data across all languages proved to be a challenge. Google Translate allowed Bloomberg to bring all that data to their customers regardless of language. This table shows a subset of the language pairs available in the Translation API. There are 97 pairs in all. Do you have customers all over the world? Do you have dynamic content like product reviews? Then use automatic translation to serve them. Are you a marketplace with buyers and sellers who don't speak a common language? Use a translation service to connect them seamlessly.



more eg: Kewpie (baby food), Arcules

32. [5] , [6] Personalization allows a business to scale, to reach more customers and to provide services in context, in the moment. The primary way that personalization is done using machine learning is with what we call recommendation systems or recommenders. You see recommenders all the time. They're the most common form of machine learning that end-users are faced with. Recommenders help you find content that goes together. Content that you're looking for or that you didn't know that you wanted. eg: WeightWachters.

There are many advantages of Personalization.

Recommendation systems are good for users

- 1 Help users find related content
- 2 Help users explore new items
- 3 Improve user decision making

Recommendation systems are good for businesses

- 1 Increase user engagement
- 2 Present more information about users
- 3 Change user behavior



Any businesses stock what customers within 10 miles want from them. Radio stations choose programming that nearby listeners like etc. This pressure to serve a small local consumer base meant that niche products often weren't worth stocking. But e-commerce and machine-learning changed all of that. According to an article in Wired by Chris Anderson, in 1988, a book called 'Touching the Void' was published. The book told the story of a near death experience in the Peruvian Andes, but somehow went out of print. A decade later in the era of the Internet, 'Into Thin Air' was published and became hugely popular. Publishers were surprised to see there was suddenly demand for 'Touching the Void'. A few years earlier, no retailer would have been able to profitably stock a product that was in the long tail as 'Touching the Void' was. But with the advantage of a larger pool

of consumers and the data and using machine-learning, Amazon was able to recognize the appeal of this old product and begin marketing it. **Thus Personalization allows a business to serve the long tail.**

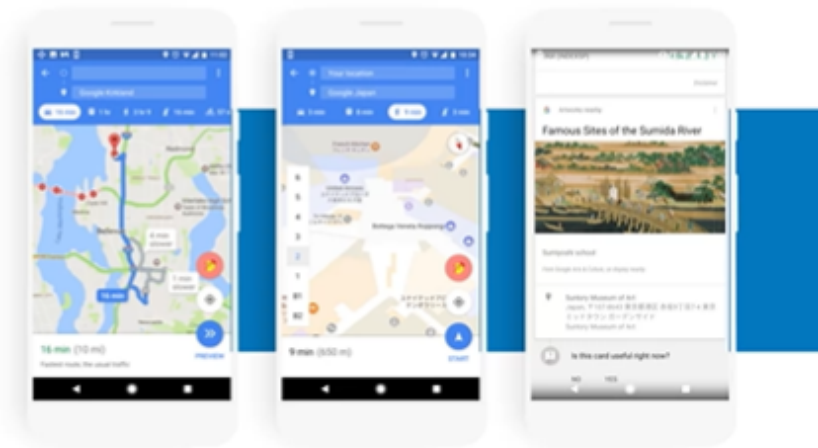
33. I was in a subway station called Roppongi, and Maps told me that I was on floor number two. How did maps know? Whatever the data sources it uses, wi-fi points, barometric pressure, typical walking speed it's pretty obvious this cannot be a simple set of rules. Plus, of course, you need the relevant data to train the model, and the relevant data to keep the model remaining fresh. And once you have the data, you can now use machine learning to avoid having to write the logic. Maps here is anticipating information that you might want to know if you are in a multi-story building. What else can Maps anticipate? Looking at the map, still in Japan, I glance at my phone between meetings and notice that I was getting a recommendation. Maps is now connecting my history, that I like art, that I like museums, that I'm in Japan, to now recommend things to me. This is even more of a data problem. **The machine learning is what allows the original limited how-to-get-from-point-a-to-point-b map to now become a virtual assistant.**

The journey to ML leads to **personalization**

Solve problems in a data-driven way, building analytics expertise

Data insights that you couldn't obtain before

Build just-in-time, personalized experiences



So machine learning is about **scaling beyond hand written rules**. But then you start being able to do things that you could never achieve if you are hand writing rules. So think about your business. Your business analysts today are essentially looking only at the bulk of your business, the 80% of things that 20% do. You may be thinking of machine learning as a way to do the things in the middle, of being able to take the data you happen to have and training a machine learning model. But think of machine learning as a way to get to the things on the right (tail end), of being able to personalize your services for each of your customers. And notice the question at the bottom of the card on the right, asking for user feedback to keep improving the model.

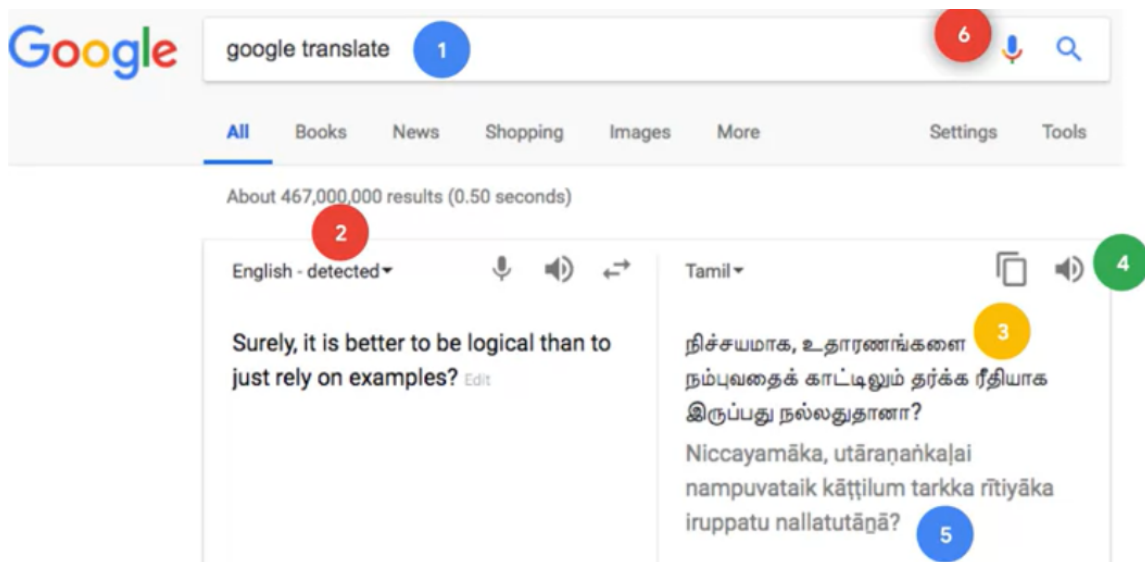
34. **creative uses of ML [34-35]** PART 1: super-resolution, create new content (NSynth, Magenta - music, style transfer - paintings, etc)
35. PART 2: *eg: bicycle rental*. Even though rides have many characteristics, there's a limit to how fine-grained your analytics can go, because with every new dimension you aggregate on, you reduce the number of rentals per bin. Thinking back, there were already bins that were nearly empty after splitting by day of week and station. Fundamentally, this is where machine learning can help. Machine learning can help you predict the value for things that haven't happened before, such as rentals on a particular day of the week, at a particular station based on similar stations.



PART 3: In 2018, AirAsia established a groundwork to use machine learning to optimize pricing for a range of services. They used Google Cloud Machine Learning Engine to sort and predict demand for ancillary services such as baggage, seats, and meals. For a low-cost carrier, a good stream of revenue comes from these ancillary services, so it's very important to accurately predict demand for these services and be prepared. For example, to make sure that a meal is in stock when a customer wants to purchase it. There are lots of factors that going to being able to successfully predict the number of vegetarian meals that AirAsia needs to stock. It's not as simple as knowing the size of the plane. Planes to and from India, for example, will tend to have a lot more vegetarians. But not just that, the time of day probably matters, as does which flight is connected to this one. After all, you can't just stop the plane and pick up a vegetarian meal. The more factors there are that you have to balance, the more useful machine learning becomes.

36. Up until this point, we have been talking about machine learning models as if they stand alone. Yet, when we talked about real companies that use machine learning, whether it is Netmarble or AirAsia, you found that we were talking about a set of machine learning models that address a set of business needs. That's because machine learning **models don't exist in isolation**. You almost never built and put into production a single ML model, just like putting into production a software library. What you put into production are applications that use a software library, and that is true of machine learning also. You will put into production applications that use machine learning, and these applications won't use just one ML model, they

will use many. So an important design consideration is that you should think about how to create a seamless experience out of many models. eg: *Google Translate*.



37. [Activity: Generating Novel Rhythms with ML](#)

38. Final module of this course includes:

Learn how to ...

Formulate a data strategy

Develop governance around uses of ML

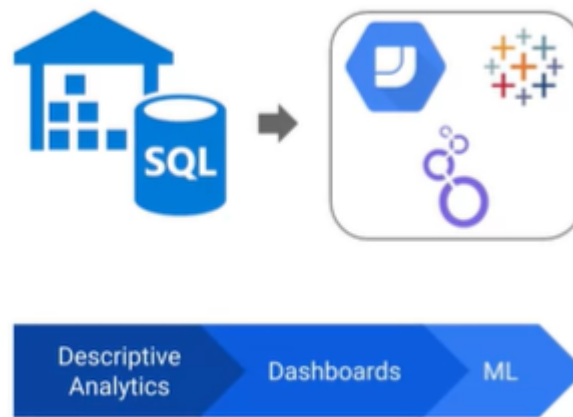
Build successful ML teams

Create a culture of innovation

39. [Seven actions as part of your data strategy \[39-41\]](#). [1] Respect the opportunity- eg: movie recommendation system In traditional data analytics, you start with the logic and then collect data to apply that logic. With machine learning, you start with data and then apply a standard algorithm to your data to obtain predictions. Once you identify the need to capture what movies users are watching, you can get a bit more fine-grained.

- 1 Design systems so that you will have more data next year
- 2 Break down data silos
- 3 Transition from data lakes to data warehouses
- 4 Learn about your data
- 5 Integrate pilots into your tools
- 6 Run ML models on real-time data to extract the most value

40. [2] , [3] The second thing you need to do as part of your data strategy, is to break down data silos. It's important to have all the data organized in such a way that you can query together, and create an ML training dataset. Why do we say that? Even if there's a culture of data sharing, the problem is that your data is spread throughout the organization, and stored in different databases. No one will give an ML engineer access to their production databases. Even if they have a cluster on which they have a copy of their data, they're concerned that the ML person will take all the their compute. So the correct technical solution is to separate compute from storage. In other words, built a data warehouse. Then all that the other organization needs to do is to provide access to the data. Then, whichever department queries data, pays for their own compute. This is the principle behind Google **BigQuery**.
41. [4] - [6] Building a dashboard is one of the best ways to understand your data. Dashboards are visualizations of your data that can be seen widely across the company. Use tools such as Data Studio, Looker, Tableau, etc, connected to your data warehouse.



Use technologies like vision API, Natural Language API, Auto ML texts, and Auto ML vision, to enrich your unstructured data.

Choose the modeling software based on the value of the use case



Just as a global infrastructure free warehouse such as BigQuery was useful for collecting historical data, a global infrastructure free messaging systems such as Cloud Pub/Sub is also necessary for collecting real time data. As with BigQuery, the department that owns the data in Pub/Sub can provide access to others without assuming the cost of their compute.

Cloud Data Catalog can help maintain metadata across your data

42. Lab : Predicting visitor purchases with BQML model.

43. When you're implementing ML use cases, it's important to **balance** data **access** within your company, against the **security** [43 - 45] implications of that access.



To create a model that returns users specific results, you typically need to access to user specific data. Fortunately there are techniques you can use to remove some sensitive data from the data sets, while still training effective ML models.

Three goals for ML and Privacy

- 1 Identify sensitive data
- 2 Protect sensitive data by removing, masking or coarsening
- 3 Create public governance documentation

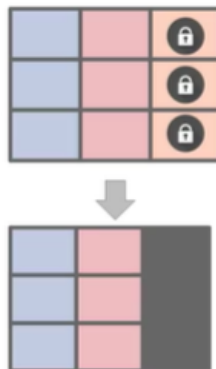
Types of sensitive data

- 1 Specific columns in structured datasets
- 2 Patterned text, e.g., credit card numbers
- 3 Unstructured data, like audio, video and images
- 4 Combinations of fields

Protecting sensitive data

- 1 Remove it
- 2 Mask it
- 3 Coarsen it

Ways to remove sensitive data



When data is restricted to specific columns and structured datasets, you can create a view that doesn't provide access to these columns.

Ways to remove sensitive data



Use the DLP API to remove sensitive data

Data engineers are capable of deciding which columns to retain on the basis of relevance

Masking sensitive data

```
Friends, Romans, countrymen,  
lend me your ears;  
I come to bury Caesar, not to  
praise him.
```



```
Friends, Romans, countrymen,  
lend me your ears;  
I come to bury 6d086db, not to  
praise him.
```

If some or all of the sensitive data fields can be masked without affecting ML training. Then several techniques can be used to mask that data. The most common approach, is to use substitution cipher, which involves

replacing all occurrences of a plain text identifier by its hashed and all encrypted value.

14. Demo: data masking

15. You can decrease the precision or granularity of data in order to make it more difficult to identify sensitive data within the data set. You trade accuracy for sensitivity. GPS locations - decimal based latitudes and longitudes rounded up to single digit precision. Zip codes - US zip codes in a five plus four form can identify a household but can be coarsened to include just the first three digits. This is a zip three format. This limits the ability to identify a specific user by putting many users in the same bucket. IP address - a good coarsening technique is to zero out the last octet of the IPv4 addresses. That is the last eighty bits if you are using IPv6. Numeric quantities - numbers can be binned to make them less likely to identify an individual. For instance, an exact birthday is often not required, just the decade or year in which a user was born. And similar numeric fields can be coarsened by substituting ranges.

Ways of coarsening data

- 1 GPS locations
- 2 Zip codes
- 3 Numeric quantities
- 4 IP addresses

Common principles for establishing a policy framework

- 1 Establish a secure location for documentation
- 2 Exclude sensitive information from documentation

Common principles for establishing a policy framework

- 3 Document all sources of incoming sensitive data
- 4 Document all locations of stored sensitive data
- 5 Document sensitive data locations where remediation is difficult

Common principles for establishing a policy framework

- 6 Establish a process to scan for new sensitive data
- 7 Document who has access and for what reason
- 8 Document the process by which employees request access

Common principles for establishing a policy framework

- 9 Establish a process to review access policies
- 10 Establish a process to communicate and enforce policies

46. As long as the use cases fall along the most popular, most common sets of use cases, you don't really need expert researchers. Your existing information technology and business analysts teams can be up skilled. The IT folks can learn how to use Auto ML. The business analyst teams can learn how to use BigQuery ML.

Most common use cases don't require ML experts

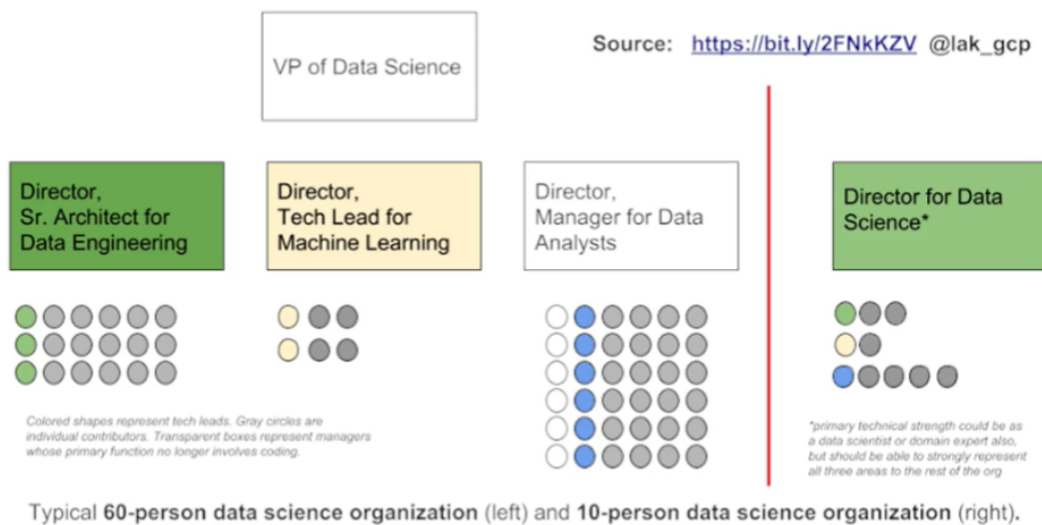
- ✓ IT personnel can use AutoML
- ✓ Analysts can use BigQuery ML

Who does what on a data science team?



Your team as a whole, should know how to build end-to-end machine learning solutions, to real world problems. Data engineers build pipelines to routinely ingest and transform data. Data analysts build dashboards and create reports to support decision-making throughout the organization. You don't want your ML engineers building an ML Model to automate any decision-making, until your data analytics team has done the analysis by hand a few times, and you learn that this is a decision your business tends to make routinely. Finally, ML engineers build predictive models using curated data.

What a team should look like (size vary, approximate ratio remains same:



47. Creating a culture of innovation:

Use three simple principles:



Focus on the user



Think 10X



Launch and iterate

Focus on the user



Who is my **customer**?

What do they **expect**?

10X Thinking	How might we get cars to run forever between fueling stops?	Charge battery with renewable energy	Research into smart sensors, revolutionary batteries, extract brake power, self-driving cars
Improvement Thinking	How can we get cars to be more fuel-efficient?	Replace engine with a diesel one	Build new partnerships with manufacturing companies
	Challenge	Solution	Means to solve

Launch and iterate requires a shift in mindset. It says that you don't need to perfect something on paper first to be successful. In fact you don't even need to have every detail figured out. Launch and iterate encourages experimentation and learning from each output with help from your peers, leadership and feedback from your customers. eg: *Google Glass*.



Incorporating ML for your business requires a change in mindset and general attitude about failure



THE END

▼ Formatting guidelines:

Bold: Important

Italic: Examples

Bold Italic: Key terms

Every numbered point is emphasizing a key idea, and that is highlighted in **blue text**. If those ideas are questions, then their answers have been represented in **purple text**. **Green text** denotes lab or activity.

Using this note in conjunction with the course will help you grasp the note better.