Data Scientist: The Sexiest Job of the 21st Century

Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see possibilities. He began forming theories, testing hunches, and finding patterns that allowed him to predict whose networks a given profile would land in. He could imagine that new features capitalizing on the heuristics he was developing might provide value to users. But LinkedIn's engineering team, caught up in the challenges of scaling up the site, seemed uninterested. Some colleagues were openly dismissive of Goldman's ideas. Why would users need LinkedIn to figure out their networks for them? The site already had an address book importer that could pull in all a member's connections.

Luckily, Reid Hoffman, LinkedIn's cofounder and CEO at the time (now its executive chairman), had faith in the power of analytics because of his experiences at PayPal, and he had granted Goldman a high degree of autonomy.

He did this by ginning up a custom ad that displayed the three best new matches for each user based on the background entered in his or her LinkedIn profile. Within days it was obvious that something remarkable was taking place. The click-through rate on those ads was the highest ever seen. Goldman continued to refine how the suggestions were generated, incorporating networking ideas such as "triangle closing"—the notion that if you know Larry and Sue, there's a good chance that Larry and Sue know each other. Goldman and his team also got the action required to respond to a suggestion down to one click.

A New Breed

Goldman is a good example of a new key player in organizations: the "data scientist." It's a high-ranking professional with the training and curiosity to make discoveries in the world of big data. The title has been around for only a few years.

If your organization stores multiple petabytes of data, if the information most critical to your business resides in forms other than rows and columns of numbers, or if answering your biggest question would involve a "mashup" of several analytical efforts, you've got a big data opportunity.

Much of the current enthusiasm for big data focuses on technologies that make taming it possible, including Hadoop (the most widely used framework for distributed file system processing) and related open-source tools, cloud computing, and data visualization. While those are important breakthroughs, at least as important are the people with the skill set (and the mind-set) to put them to good use.

Who Are These People?

If capitalizing on big data depends on hiring scarce data scientists, then the challenge for managers is to learn how to identify that talent, attract it to an enterprise, and make it productive.

The first step in filling the need for data scientists, therefore, is to understand what they do in businesses. Then ask, What skills do they need? And what fields are those skills most readily found in?

More than anything, what data scientists do is make discoveries while swimming in data. It's their preferred method of navigating the world around them. In a competitive landscape where challenges keep changing and data never stop flowing, data scientists help decision makers shift from ad hoc analysis to an ongoing conversation with data.

Data scientists realize that they face technical limitations, but they don't allow that to bog down their search for novel solutions. Often they are creative in displaying information visually and making the patterns they find clear and compelling. They advise executives and product managers on the implications of the data for products, processes, and decisions.

Data scientists' most basic, universal skill is the ability to write code. This may be less true in five years' time, when many more people will have the title "data scientist" on their business cards. More enduring will be the need for data scientists to communicate in language that all their stakeholders understand and to demonstrate the special skills involved in storytelling with data, whether verbally, visually, or ideally both.

Perhaps it's becoming clear why the word "scientist" fits this emerging role. Experimental physicists, for example, also have to design equipment, gather data, conduct multiple experiments, and communicate their results. Some of the best and brightest data scientists are PhDs in esoteric fields like ecology and systems biology.

It's important to keep that image of the scientist in mind because the word "data" might easily send a search for talent down the wrong path. A data management expert might be great at generating and organizing data in structured form but not at turning unstructured data into structured data and also not at actually analyzing the data. And while people without strong social skills might thrive in traditional data professions, data scientists must have such skills to be effective.

Why Would a Data Scientist Want to Work Here?

Pay will of course be a factor. A good data scientist will have many doors open to him or her, and salaries will be bid upward. Several data scientists working at start-ups commented that they'd demanded and got large stock option packages. Our informal survey of the priorities of data scientists revealed something more fundamentally important. They want to be "on the bridge.". Considering the difficulty of finding and keeping data scientists, one would think that a good strategy would involve hiring them as consultants.

But the data scientists we've spoken with say they want to build things, not just give advice to a decision maker. One described being a consultant as "the dead zone all you get to do is tell someone else what the analyses say they should do." By creating solutions that work, they can have more impact and leave their marks as pioneers of their profession.

Care and Feeding

Data scientists don't do well on a short leash. They should have the freedom to experiment and explore possibilities. That said, they need close relationships with the rest of the business. As the story of Jonathan Goldman illustrates, their greatest opportunity to add value is not in creating reports or presentations for senior executives but in innovating with customer-facing products and processes.

LinkedIn isn't the only company to use data scientists to generate ideas for products, features, and value-adding services. E is already using data science to optimize the service contracts and maintenance intervals for industrial products. Google, of course, uses data scientists to refine its core search and ad-serving algorithms. Zynga uses data scientists to optimize the game experience for both long-term engagement and revenue. Netflix created the well-known Netflix Prize, given to the data science team that developed the best way to improve the company's movie recommendation system. The test-preparation firm Kaplan uses its data scientists to uncover effective learning strategies.

Data scientists tend to be more motivated, too, when more is expected of them. The challenges of accessing and structuring big data sometimes leave little time or energy for sophisticated analytics involving prediction or optimization.

The Hot Job of the Decade

Hal Varian, the chief economist at Google, is known to have said, "The sexy job in the next 10 years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?"

If "sexy" means having rare qualities that are much in demand, data scientists are already there. They are difficult and expensive to hire and, given the very competitive market for their services, difficult to retain. There simply aren't a lot of people with their combination of scientific background and computational and analytical skills.

The problem with that reasoning is that the advance of big data shows no signs of slowing. If companies sit out this trend's early days for lack of talent, they risk falling behind as competitors and channel partners gain nearly unassailable advantages. Think of big data as an epic wave gathering now, starting to crest. If you want to catch it, you need people who can surf.

IBM: Data Science is a Teamsport

Introduction

The greatest challenge of the big data revolution is making sense of all the information generated by today's vast digital economy. The more data you have, the better the quality of your reports and strategic recommendations, right? Sure…if you can analyze that data intelligently and quickly, and make it actionable with valuable insights. Otherwise, more data can mean more problems: messy data, storage woes, security risks, frustrated business teams and overloaded IT staff.

"*Successful organizations build data science teams that incorporate different skill sets and responsibilities, instead of relying on a few elite individuals*"

In practice, several people work on a team to build data products. Your analyses will only be as good as the team that is responsible for collecting, building and analyzing the underlying data.

Which talents and abilities define the members of a data science team, and how do they complement each other? Read on and find out.

1. The Data Scientist

Data scientists are often referred to as "unicorns" because they have a rare combination of talents: they handle a variety of responsibilities and skill sets covering mathematics, statistics, domain expertise, communications and more. Basically, the job of the data scientist is to look for hidden patterns. They accomplish this by applying advanced analytics techniques including (but not limited to) machine learning, modeling, statistics and visualization.

Often, data scientists will construct models to predict outcomes or discover underlying patterns; their game plan is to produce actionable insights that can be used to improve future outcomes.

A skilled data scientist explores and examines data from multiple disparate sources. They don't just collect and report on data; they look at it from many angles, determine what it means and then recommend ways to apply the findings. They need to make sure their queries are correct and must be able to back up their conclusions with sound models and trusted data as the Data Scientist is often expected to present recommendations to management and executive teams.

Top skills for Data Scientists

Data scientists are distinguished by their strong business acumen, plus the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. The data scientist often becomes the liaison between the IT department and C-level executives.

Data scientists are inquisitive and curious: exploring, asking questions, doing what-if analysis, and questioning existing assumptions and processes. A data scientist's technical skills often include multiple programming languages, familiarity with big data management and analysis tools like Apache Hadoop and Spark, and experience with tools that help them visualize data and insights.

2. The Data Engineer

At a high level, data engineers help gather, organize and clean the data that data scientists will ultimately use to build their analysis. If data scientists extract value from data, data engineers make sure data flows smoothly from source to destination so it can be processed.

Data engineers are often tasked with laying the groundwork for a data analyst or data scientist to easily retrieve the data needed for their evaluations and experiments.

Top skills for Data Engineer

Data engineers are hard-core engineers who understand the internal workings of database software. They compile and install database systems, write complex queries, scale them to multiple machines, manage backups and deploy disaster recovery systems. They develop, construct, test and maintain architectures such as databases and large scale data processing systems.

3. The Developer

A successful project usually productizes the data science work so it can serve an internal stakeholder or external customers. Developers often come in at the end of the data science workflow.

They are responsible for building the applications where the models will sit applications that leverage the insights and data gathered from the rest of the data science team. This requires a fair amount of programming time, and it can be a challenging job. Lack of integrated technologies can stifle developers' progress, making it difficult to embed the data science team's analysis.

Another requirement: developers must have a full component of programming skills at the ready. This includes expertise in building web services, front-end development skills and strong knowledge of user interface functionality and features. In addition, they should be familiar with application programming interfaces (APIs) and using them to integrate various data products and sources into applications.

To build game-changing mobile, web and enterprise applications that will disrupt markets, developers need the data and the tools to turn their vision into reality. The

developer must collaborate with the data scientist, data engineer and business analyst to ensure alignment between the business objectives and the analytics back end.

4. The Business Analyst

The business analyst, sometimes referred to as the citizen analyst, provides businesslevel expertise and guidance to the data science teams. Their responsibility is to apply domain knowledge and make datainformed decisions.

The business analyst understands what the business needs, but doesn't have the technical background to develop a detailed analysis. Flexible and user-friendly technologies allow them to develop some business-level analysis without coding and without having to ask the data scientist.

5. Data science teams: The new agents of change

Skilled data scientists, data engineers, developers and business analysts are transformative figures in modern business. They are the beating heart of the big data economy. It's not just that they are designing new systems; they are going to bat for new sources of data and new ways to use that data. Of course, IT still has to build the system, but the data science professionals are the ones who help departments collaborate to solve problems and speed innovation.

The best data products are the ones that the end user doesn't even notice. The technology to collect and analyze massive volumes of business data is available now, and you can exploit it to your company's benefit.

The Future of Cognitive Computing

Let me begin with a word of caution: It's very easy when we get into areas of Cognitive Computing and Artificial intelligence to rapidly drop down to Deep machine learning and algorithms. These are really exciting and fascinating areas of technology, but I think the thing we must keep in mind is "Towards what end?"

What is it that we are actually trying to do?

It's all about the outcomes:

- Changing the world
- Changing entire industries
- Seeing things and getting insights that we have never been able to grasp before.

I encourage you to keep thinking about what we can do with this technology. How can we impact society and the human state in ways that we've never been able to before? Cognitive computing is not trying to replicate what the human brain does. Cognitive computing is a system that can handle massive amounts of unstructured data.

Unstructured data – "dark data" – accounts for 80% of all data generated today. Most of that data is "dark" – we cannot make sense of that data. It is noisy or formats that cannot be read by traditional systems. Furthermore, the amount of dark data is expected to grow to over 93% by 2020. Astrophysicists already know about "dark matter".

Dark matter cannot be seen, but we know it exists by the impact on gravity. The same thing is happening with our data. Think of the number of solutions that data holds. This is not a journey to reproduce what the human mind can do.

The objective is to analyze and garner insight from this massive amount of data.

Oil & Gas

Modern facilities have more than 80,000 sensors in place, and a single reservoir will produce more than 15 petabytes of data in its lifetime.

Cognitive computing can:

- help companies prevent drilling in the wrong place.
- help with flow optimization (pumping too much or too little).

Facilities generate more data than current technology can deal with. This is a huge opportunity.

Retail

Consumers post 500 million tweets and 55 million Facebook updates each day. From our partnership with Twitter, we now have direct access to tap into this "big hose" of data. This

data can help identify buying patterns, preferences, insights, and where society is moving – Insights that can be moved across every form of commerce.

Internet of Things

IoT is one of the next great frontiers. Watson is unmatched in natural language processing. Watson also can handle images and vision. Think about signal processing. Machine-to-machine data will dominate the market in just a few years. That is noisy, unstructured data, perfect for cognitive computing.

Internet of Things
Smart, connected appliances will grow from less than **1%** of the market today to **more than half** in 2020. This could be appliances, or could be smart connected cities. Think of the implications of city security or traffic management.

Public Safety
New York City surveillance cameras and sensors generate 520 TB of data per day, largely unstructured, and untapped.

Security

In 2014, more than 1 billion personal data records were compromised by cyber attacks. Security is no longer about firewalls. Security is now about behavioral analysis of people and systems. Systems can predict abnormalities and react in real time.

Energy

More than **680 million** smart meters will be installed globally by 2017 – producing more than **280 PB** of new data to be analyzed and acted upon. Digital meters are in many countries around the world, but the data is dark. It is difficult to integrate renewables without understanding demands.

Healthcare

One of the biggest opportunities, of course, is healthcare. Healthcare is an enormous industry, prime for disruption and new forms of insight. This is one of the industries where we (IBM) have doubled down, not only through electronic medical records, patient population healthcare, but also medical imaging.

Each person will generate **1 million GB** of health-related data in their lifetime equivalent to about **300 million** books. In that data is really the secret to our own health and well being.

Transportation

By 2020, **75%** of the world's cars will be connected… and they will produce **350 MB** of data per second to be assessed and acted upon. These devices will need to be cognitive. They will need to make realtime decisions about the environment based on learning about the environment and learning about driver behavior.

Every single industry is being swamped with data. Every industry is trying to access that 80-90% of dark data and get insights to differentiate. We are at an industry inflection point.

We as humans have a number of abilities that machines will have a hard time replicating; maybe never.

Human :

- Compassion
- Intuition
- Design
- Value judgement
- Common sense

Machines:

- Total/Instant Recall
- Deep learning
- Discovery
- Large-scale math
- Fact checking

The opportunity is man and machine. We are seeing this in every discipline and every industry that we go into with Watson. We as humans have a normal distribution (statistics) of skills. What we're finding is that we can move that distribution. We can take the best experts and make them better by introducing man and machine.

Since 2011, this field has exploded.

- Many image processing tools
- Many buying optimization tools
- Many voice recognition tools

Each of these is really a point solution to improve a one-dimensional aspect of a business model. It really is like a tool – hammer, screwdriver. Very few, besides IBM are trying to build an entire toolkit/platform of capability for all industries with this cognitive computing capability.

We took Watson, which was one system, and brought it to IBM Cloud. We decomposed one system (Q&A) to individual services that can be composed to create meaningful solutions.

This system effectively had five parts at the time:

- Machine Learning
- Question Analysis
- Natural Language Processing
- Feature Engineering
- Ontology Analysis

We (IBM) have built out a suite of services that enable you to make your own mini-Watson for a solution to your problem. We decided to make it cloud based and composable for all industries.

Ecosystem

This has become the platform for our ecosystem. Our plan is to develop not just dozens, but hundreds of these services on the Watson cloud as fast as we can. We have a pipeline of these services in a rich environment.

What is the essence of cognitive?

Cognitive systems must learn at scale.

- Learning at scale in the data
- Learning at scale for your solution/business
- Reasoning with a goal to take an action

Cognitive systems interact with humans it's the interaction of man and machine that will produce capabilities going forward; Not just about automating systems.

Rethinking What's Possible

The possibilities are immense.

Healthcare

We at IBM believe that image analytics and machine learning can change the course of health care. Two-thirds of medical information is contained in images (X-ray, MRI, CAT scan). We know the diagnosis is not what it needs to be. Radiologists look at thousands of images in a single day human issues such as fatigue set in. We are going to train Watson to read those images with the patient data. Watson will use Unstructured medical record data across the entire patient's history to generate recommendations in minutes. This will change the course of healthcare.

Seismology

They are building cognitive environments for decision making.

- Where do I drill the next well?
- Do I bid on that land to get the oil under it?
- What happens to oil reserves over time?

We are going to help transform a very intensive data industry.

Education & Accelerated Learning

There is a direct analogy between health care and education. Medical is treating to the average. Education is very much the same. Teachers teach to the average. Think about having Watson engage with the individual student to observe learning patterns and intervene. The engagement has a direct correlation to accelerate learning and change education systems.

Pre-K (ages between 2-3) vocabulary is a direct correlation to long-term potential. Watson has the capability to double or triple the vocabulary of that 2-3 year old. There is a huge opportunity to change education.

Genomics

Eventually genomic data will likely swamp image data. Genomics will generate more data than can be understood by doctors (humans alone) – Hundreds of mutations, thousands of pathways that cause a tumor can manifest itself. Think about using a Watson to help humans understand what those mutations mean, and what is the right treatment. IBM is working with a number of leading edge genomic institutions in the US and Canada to explore what Watson can do in this area. Frankly, it is the only way we're going to deal with genomic data.

What IBM looks for in a Data Scientist

Job seekers sometimes ask how IBM defines "data scientist." It's an important question since more and more would-be data scientists are fighting for attention in an increasingly lucrative labor market.

The first step is to distinguish between what we see as true data scientists and other professionals working in adjacent roles (for instance, data engineers, business analysts, and AI application developers).

The definition demonstrates that to achieve the true potential of data science, we need data scientists with very particular experiences and skills specifically, we need people with the experiences and skills required to run and complete data science projects:

1. Training as a scientist, with an MS or PhD
2. Expertise in machine learning and statistics, with an emphasis on decision optimization
3. Expertise in R, Python, or Scala
4. Ability to transform and manage large data sets
5. Proven ability to apply the skills above to real-world business problems
6. Ability to evaluate model performance and tune it accordingly

Let's look at those qualifications in the context of our definition of data science.

1. Training as a scientist, with an MS or PhD

This is less about the degree itself and more about what you learn when you get an advanced degree. In short, you learn the scientific method, which starts with the ability to take a complex yet abstract problem and break it down into a set of testable hypotheses. A determined person can learn these skills outside of academia or via the right mix of online training and practice so there's some flexibility around having the actual degree but direct experience applying the scientific method is a must.

To get published, candidates have to present their work in a way that allows others to review and reproduce it. You must also provide evidence that the results are valid and the methods are sound. It's possible to get an abstract sense of those values, but there's no substitute for the negative and positive reinforcement from mentors or the rejection or acceptance of journals and reviews.

2. Expertise in machine learning and statistics, with an emphasis on decision optimization

Applying the scientific method to business problems lets us make better decisions by predicting what will happen next. In addition, decision optimization (aka operations research) is a fast-growing aspect of data science. Indeed, the goal of data science is to

help make better decisions by probabilistically estimating what's likely to occur in the future.

3. Expertise in R, Python, or Scala

Being a data scientist doesn't require you to be as good at programming as professional developers, but the ability to create and run code that supports the data science process is mandatory and that includes the ability to use statistical and machine learning packages in one of the popular data science languages.

But it's important to note that the reason to use these languages isn't that they're free, but for the innovation and the freedom to take them where you want to go.

4. Ability to transform and manage large data sets

The true data scientist will know how to pull data sets together from multiple sources and multiple data types with the help of his or her data science team. The data itself might be a combination of structured, semi-structured, and unstructured data living on multiple clouds.

5. Proven ability to apply the skills above to real-world business problems

It's the ability to communicate with non-data scientists in order to make sure that data science teams have the data resources they need and that they're applying data science to the right business problems. This requires good storytelling skills, and in particular, the ability to map mathematical concepts to common sense.

6. Ability to evaluate model performance and tune it accordingly

Expertise in machine learning in general. Data scientists who lack this skill can easily believe that they've created and deployed effective models when in fact their models are badly over-fit to the available training data.

A data scientist is fundamentally different from a business analyst or data analyst, who often serve as product owners on data science teams, with the important role of providing subject matter expertise to the data scientists themselves.

Pokemon Go & Ingress

It’s the 90s all over again. Pokémon fever is sweeping the globe thanks to a new Augmented Reality version of the monster collecting game, which sees players hunt through real-world locations for digital critters.

To play, you walk around the real-world, following a Google Maps-like interface, until signs of a nearby Pokémon appear on screen. When you hold up your smartphone camera, the Pokémon is overlaid on the screen using augmented-reality, allowing you to capture it.

The game is made by Niantic, a former Google subsidiary, and is based on a similar game called Ingress that sees players competing to hold territory by visiting it in the real world players capture “exotic matter” rather than Pokémon.

When still part of Google, Ingress was seen as a data gold mine, helping the firm improve its location services by gathering data as players visited landmarks. It’s not clear if Niantic has plans to use Pokémon Go data for similar purposes, but popular locations in Ingress also appear as key zones in Pokémon Go.

Bukalapak AI Recommendation

The Origin Story

The origin story for our AI starts on the day I moved back to Indonesia. Having just dropped out of my Ph.D. in computer vision a couple of years before, I was eager to start a culture of AI research within Bukalapak and see what kind of AI we can build from our mountains of data.

It turns out, one of the core principle at Bukalapak since day one is “store everything, we’ll be able to build something useful with those data someday.”

Designing and Building the AI

We started with one of the most common types of recommendation AIs: collaborative filtering. The idea is deceptively simple: if a user viewed product X and then viewed product Y, then another user viewing product X might also be interested with product Y. We are effectively using the traffic data to form associations between our millions of products.

One crucial context: prior to the AI, we use Elastic’s “more like this” feature to give our visitors a product-to-product recommendation. We took a couple of samples and visually compare the output from Elastic and our AI. The samples show good quality recommendations. Instead of showing similar products, we present our visitors with various product alternatives for their shopping journey.

Validating the AI

We use an approach called A/B testing to validate whether the AI-based recommendation that we put on that slot gives more impact than the Elastic-based similar products. We then track the user’s journey within our site, taking careful note of the relevant metrics.

In our case, when a user visit a product page, the recommendations at the bottom of the page are being populated from either the reco AI or from Elastic. We then use our internally built A/B testing framework to track multiple metrics related with purchases to gather real-world objective data regarding the performance of the AI.

The result was astounding. We observed a significant increase in visitor purchase when they were being presented with recommendations coming from the new AI.

Convinced with the positive impact of the AI, we decided to spend another month building automated data pipelines and process to incrementally train the AI and build fresh recommendations every day for our visitors. At the end of the first month alone, we saw that the AI had generated an additional US$1 million of revenue for our sellers.

Expanding to More AI Products

What we have established at Bukalapak so far consist of two streams of research efforts. The first stream is for building regression or neural-net based AI models from millions of rows of structured and labeled data.

We use an AutoML platform to rapidly build basic models to test out the feasibility of using AI to approach the problem. This platform enables our data scientists to quickly iterate and validate their models, typically within a couple of weeks instead of months. The second stream is for building more complex AI models. We establish several AI R&D teams to build and improve various AI services ranging from using computer vision to recognize objects/watermarks within product images, to using natural language processing to understand search queries better and assist our users in finding the right product for their shopping journey.

Recommendations for Building AIs for Your Industry

Here are three recommendations from us if you want to build AIs for your particular industry/business:

1. **Collect, store, and verify that you have the data**

   Starting the project with an already organized petabyte-scale data warehouse helps saves months within the project.

2. **Initiate, validate, iterate**

   Sometimes starting with a simple model will be enough as long as you can quickly validate the model against real-world traffic or data. Then you can iterate and build progressively more complex and better quality models.

3. **A/B test, A/B test, A/B test**

   Running manual analysis using a small number of samples are fine for quick and early validations, but don’t forget to validate your model using real-world data and A/B tests, doing so will help you avoid possible selection bias and gather objective data to guide your AI development.

Netflix: House of Card

In any business, the ability to see into the future is the killer app, and Netflix may be getting close with “House of Cards.” The series, directed by David Fincher, starring Kevin Spacey and based on a popular British series, is already the most streamed piece of content in the United States and 40 other countries, according to Netflix.

Netflix, which has 27 million subscribers in the nation and 33 million worldwide, ran the numbers. It already knew that a healthy share had streamed the work of Mr. Fincher, the director of “The Social Network,” from beginning to end. And films featuring Mr. Spacey had always done well, as had the British version of “House of Cards.”

Film and television producers have always used data, holding previews for focus groups and logging the results, but as a technology company that distributes and now produces content, Netflix has mind-boggling access to consumer sentiment in real time.

Jonathan Friedland, the company’s chief communications officer, said, “Because we have a direct relationship with consumers, we know what people like to watch and that helps us understand how big the interest is going to be for a given show. It gave us some confidence that we could find an audience for a show like ‘House of Cards’ ”.

It is impossible to say that “House of Cards” is a hit because Netflix, to the consternation of some of its more traditional competitors, is not participating in ratings. While careers and entire networks have been made and lost based on the mysterious alchemy of finding a hit, Netflix seems to be making it look easy, or at least making it a product of logic and algorithms as opposed to tradition and instinct.

A cable executive who has talked to Amazon says that its Prime service, a nascent effort to get into original content, will also lean hard on data-driven approaches to determine its programming.

“I think it is a little hysterical to say that Big Data will win the day now and forever, but it is clear that having a very molecular understanding of user data is going to have a big impact on how things happen in television,” he said.

“Netflix and Amazon know when you stop and start a program, whether you wanted the whole thing, all of that,” said Rick Smolan, whose most recent book was “The Human Face of Big Data.” “Programmers have been wandering out and shooting a shotgun into the night sky and hoping they hit something, and I end up paying $150 for channels full of nothing I want to watch. These guys know what they are aiming at.”

Netflix's command of data, including mine, isn't foolproof. It thinks I like "The West Wing," which I don't, and it thinks I am a sucker for every quirky little indie movie that floats in, which I am not. But when it came to guessing if "House of Cards" might appeal to me — politics, media and Mr. Fincher are all hot buttons — the deck was stacked in its favor.

Gojek Finding Fantastic Driver

Problem Statement

Place of Interest, in our context, is defined as a place where a high number of pickups happen. This can be at any popular location, such as a mall, train station, or even a large residential complex. In the earlier versions of our app, when a user booked a GO-RIDE or a GO-CAR at a POI.

We wanted to decrease this friction by letting the user choose the exact '***gate'*** or entrance they would want to be picked up from. The allocated driver would then directly arrive at that point without the need to coordinate over calls or chat messages.

Our Solution

In order to scale, we had to automate this process, i.e., build an algorithm that can correctly identify the clusters of popular pickup points the same way a human brain intuitively does, by looking at the plots manually. To do this, we took advantage of historical data we had about where exactly customers were being picked up from around such POIs.

We had to build an algorithm that can not only correctly identify separate clusters from each other, but also allocate a central point for each cluster that best represented where pickups frequently happen.

Clustering Algorithms

We experimented with various clustering algorithms, of which we'll discuss DBSCAN and KMeans. While both are unsupervised algorithms, the former is a density-based clustering algorithm that effectively removes noise, and the latter is a distance-based that does not account for noise in the dataset.

DBSCAN

For DBSCAN to work effectively, clusters need to have similar densities and cannot be scattered; any data point will be considered an outlier if it is further than the distance metric, eps, from the other data points.

The geospatial pickup data we were dealing with contained both of the following extremes:

- If a particular pickup spot around a POI can be more widely preferred and consequently have a higher density. This means that a pickup cluster can have a density that is higher or lower than other clusters for that POI.
- Scattered pickup locations can always occur if GPS accuracy of the device is low. This would mean that any pickup cluster in a building's basement, or one with slightly scattered pickups, would not even be considered as a cluster.

The density of pickups will almost always be different across different POIs, and is likely to be different across pickup clusters within the same POI as well. Keeping this in mind, generalising the DBSCAN algorithm across POIs would be difficult.

KMeans

KMeans is a distance-based clustering algorithm, where each point is assigned to the nearest cluster centre. Although the algorithm doesn't account for outliers, it was much more effective in correctly identifying the clusters that DBSCAN was unable to.

However, KMeans has one major drawback: k, or the number of clusters, needs to be an input to the model. Next, we needed to automate the process of accurately determining the value of k for each POI.

Conclusion

Once we found the number of clusters that best represented the dataset for that POI, we then assign the cluster centres (as determined by KMeans) as pickup points, or gates. But we weren't done yet. The next step was to find the names of these gates — we had to automate this and with that came another set of challenges.