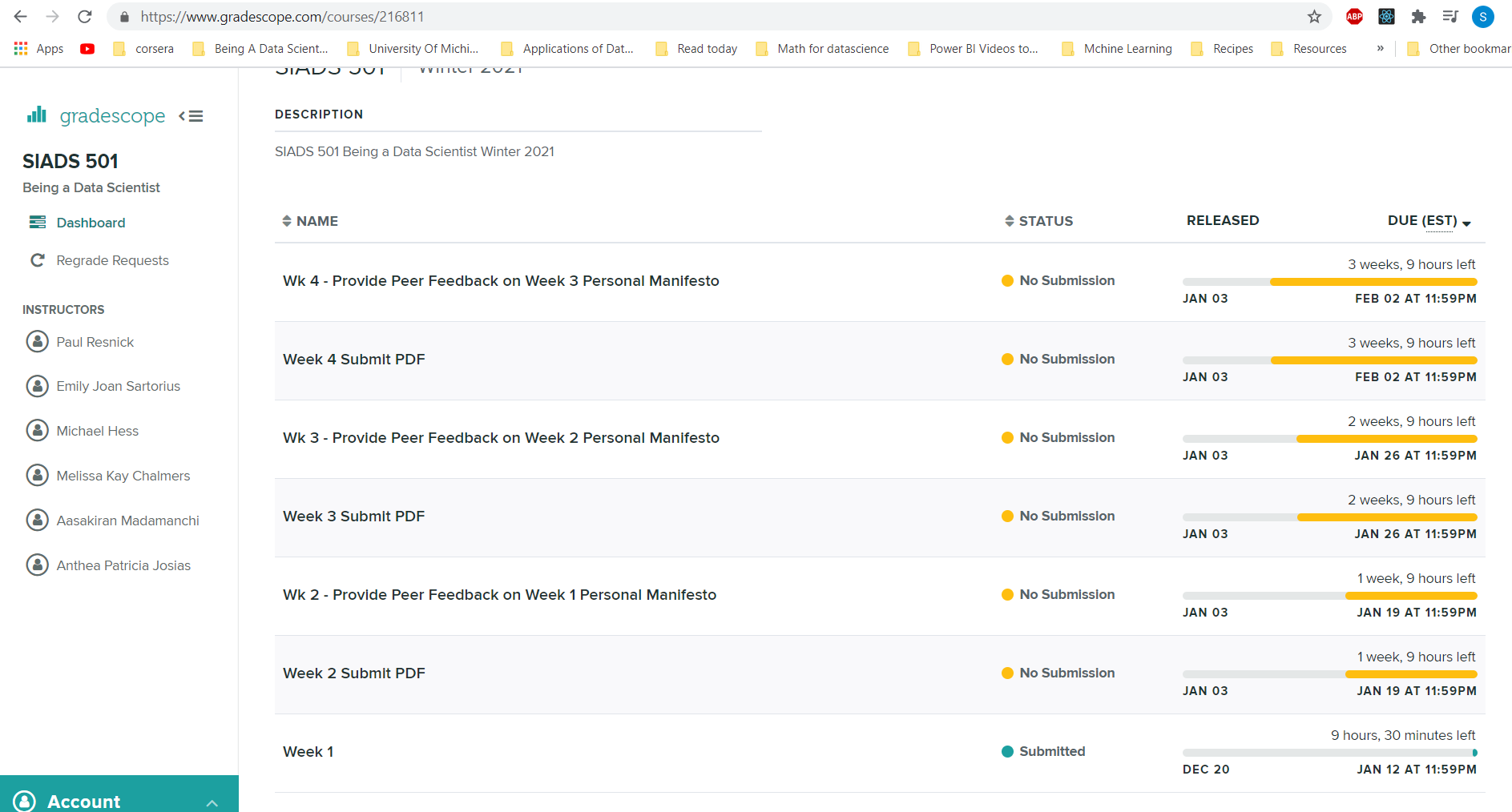
**NOTES FOR FILLING OUT THE PERSONAL MANIFESTO:**

**Refer to the table created which has been filled with the maxims and questions for the week**

**In the submission, delete the instructions.**



**SIADS 501 Class lectures notes**

At the end of the week,

you should be able to describe a project’s stages using vocabulary that we use,

of the four stages of a project,

recognize different types of problems and articulate

why each of those problems is the type that you've named it, and

you should have cultivated some curiosity about problem formulations. In particular,

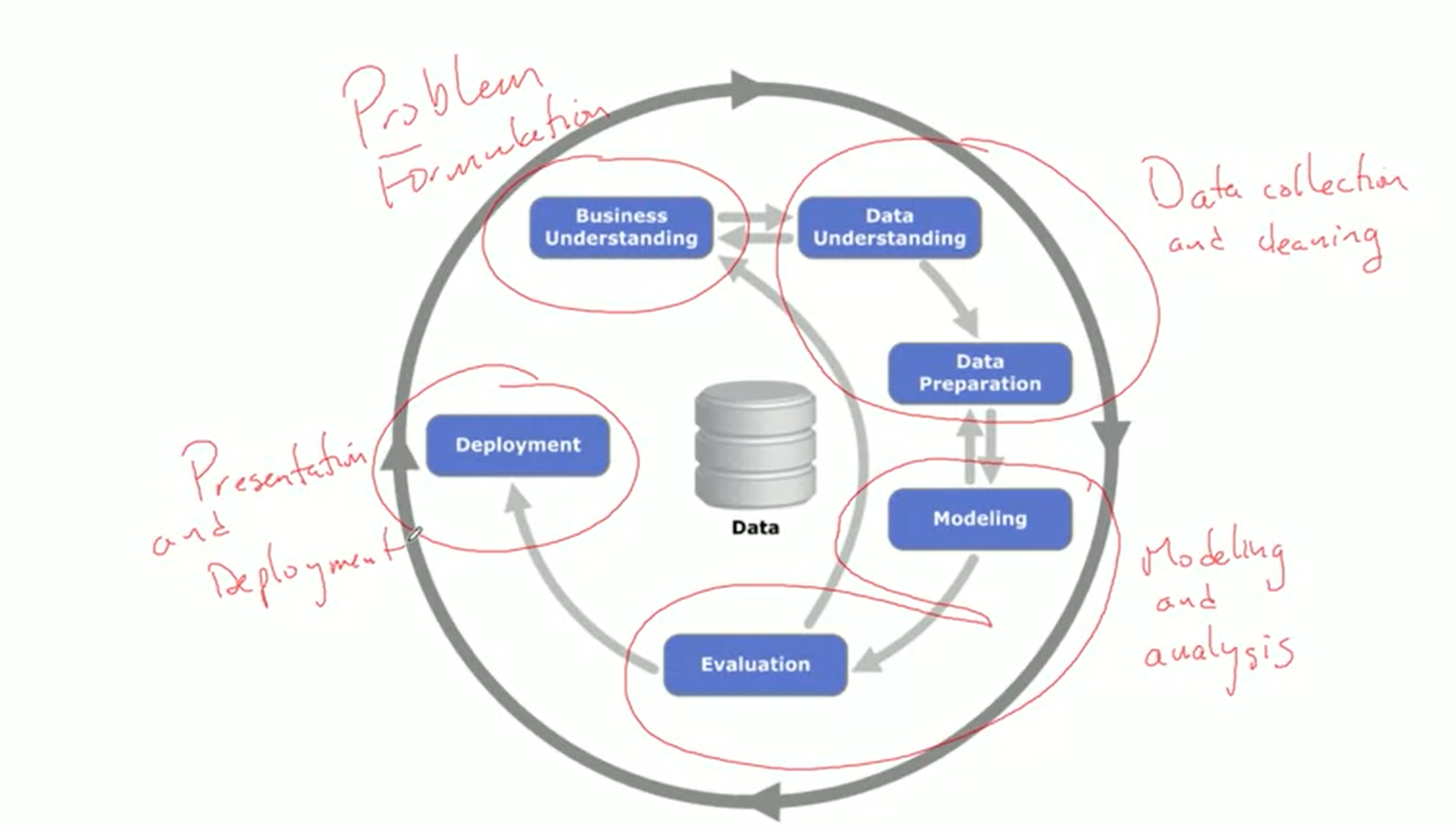
you should develop the habit of asking what's the real problem and

recognizing the dangers that can come from proxies for the outcomes that you really care about because they'll get gamed

**Stages**

**Problem formulation  
Data collection and cleaning  
Analysis and Modeling  
Presentation and Integration into action**

Cross Industry Standard for Data Mining – CRISP-DM



**The "six phases of a project" have been jocularly described as:**

1. Enthusiasm
2. Disillusionment
3. Panic
4. Search for the guilty
5. Punishment of the innocent
6. Praise and honors for the non-participants

**Problem Formulation – Taxonomy**

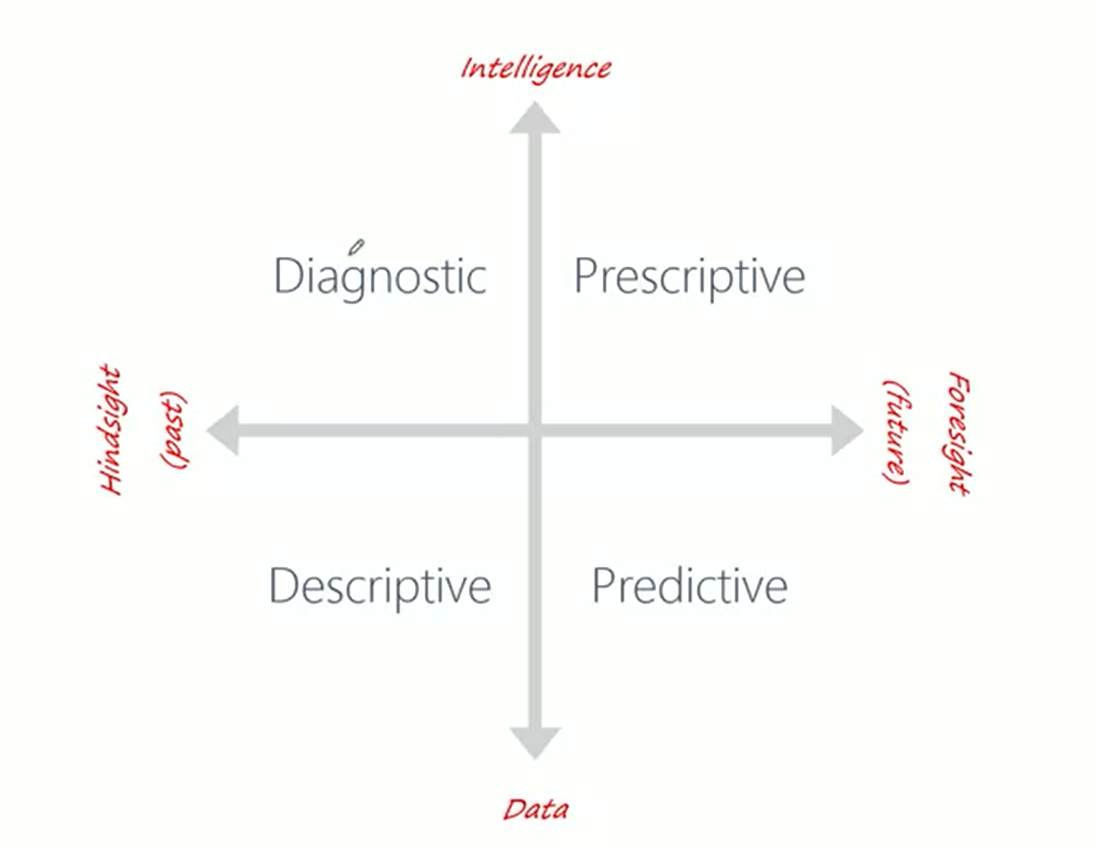
1. **A) Classification / Probability Estimation -- DISCRETE  
    Label – Ok or Spam  
    Multilabel**

**Known label groups  
Predict what kind of thing it is**

1. **Regression -------CONTINUOUS**
2. **Similarity matching**
3. **Clustering Groups are not predefined and no labels for the groups  
   Could be built on top of Similarity**
4. **Co-occurrence Grouping (Market basket Analysis)**
5. **Profiling and Anomaly detection   
   For creating personas – typical behavior  
   Can help in detecting Anomalies**
6. **Link prediction  
   Is person C likely to become friends with Person D  
   What will they read next**
7. **Data Reduction   
   Can we roll up these things to a single component feature   
   (PCA)**
8. **Causal Modeling**

**Does smoking lead to death**

1. **Thresholds and alerts. Summary statistics, trends – mainly inference  
     
     
   11) Optimization – allocate people to cashier station to minimize wait time.   
     
   12) Ranking – Google search results**



**Proxy measures in school rankings**

**SAT scores  
Funding facilities, acceptance rate**

|  |
| --- |
| **MAXIM AND QUESTIONS TABLE** |
| **Most prediction problems are really causal modeling problems in disguise  This means are you trying to truly predict based on current inference or are you trying to predict what will happen if you change something** |
| **First, outcome proxies will be gained.** |
| **The original formulation is rarely the right formulation.** |
| “Your model is not what the data scientists design, it’s what the engineers build. |
| **Consulting: The five whys**  **What is the question to be answered?**  **Why do you care about that?**  **Why is that?**  **Why is that?**  **Why is that?** |
| **Who will be using the results? For what decisions? Landlord, city planners** |
| **Who is the stakeholder** |
| • Unclean data is the canary in the coal mine. |
| • Don't treat the symptom, figure out what the root cause is. |
| • Celebrate your cleaning efforts. |
| • the primary beneficiary of your documentation efforts is your future self |
| • data diagnosis can be a manual task, but data cleaning should always be automated in a rerunnable program |
| • Always take extra steps to confirm the meaning of data |
| Sustaining doubt is harder work than sliding into certainty |
| We are pattern seekers, believers in a coherent world, in which regularities (such as a sequence of six girls) appear not byaccident but as a result of mechanical causality or of someone’s intention. We do not expect to see regularity produced by a random process, and when we detect what appears to be a rule, we quickly reject the idea that  the process is truly random. |
| misclassifying a random event as systematic. We are far too willing to  reject the belief that much of what we see in life is random. |
| Week 3: • Multiple comparisons means more opportunities for false positives • If you look at enough variables, you'll find something statistically significant • I'll believe that result when I see it replicated on a new data set. • While it is easy to lie with statistics, it is even easier to lie without them. ~ Frederick Mosteller • Correlation doesn't imply causation • what are some potential confounders? • is this a case of Simpson's paradox? • Don't assume is a confounder, it might be a mediator or a collider  Week 4: • the truth is not self-evident • the data doesn't speak for itself • the models don't speak for themselves • the results don't speak for themselves • A picture is worth 1,000 words • Tell the story. You are not the hero of your story • Maybe stories are just data with a soul. — Brené Brown • Data makes people think, emotions make them act. — Antonio Damasio • is there an intuitive explanation for this model output? • Which features does the model focus on? • Is this a situation where we need an explainable model? • we should present this uncertainty in terms of hypothetical concrete outcomes rather than statistical abstractions |
|  |
|  |
|  |

|  |
| --- |
| **Ethical Committment** |
| **Don’t overhype** |
| **Speak truth to power** |
| **WEEK 3** |
|  |
|  |
|  |
| **Ethical committment** |
| I will always hold out some data for cross validation to avoid overfitting |

**CHAPTER 2**

**Problem Formulation will identify tasks**

* 1. **Classification and class probablilty estimation (Categorical)**
  2. **Regression (Numeric)**
  3. **Similarity**
  4. **Clustering**
  5. **Co-occurrence(association- market basket analysis)**
  6. **Profiling (behavir description)**
  7. **Link prediction**
  8. **Data reduction**
  9. **Causal modeling**

**Reading Chapter 2**

**Insights :**

Recognizing familiar problems and their solutions avoids wasting time and resources reinventing the wheel. It also allows people to focus attention on more interesting parts of the process that require human involvement—parts that have not been automated, so human creativity and intelligence must come into play.

**Causal modeling** In all cases, a careful data scientist should always include with a causal conclusion the exact assumptions that must be made in order for the causal conclusion to hold (there always are such assumptions—always ask).  
Example: The discovery of the “placebo effect” in medicine illustrates a notorious situation where an assumption was overlooked in carefully designed randomized experimentation.

Stage :Problem formulation  
What is it: Maxim

**Data mining and its results**

There is another important distinction pertaining to mining data: the difference between (1) mining the data to find patterns and build models, and (2) using the results of data mining.

Stage : Data Analysis and Modeling

What is it : Maxim

Insight

Often the key to a great success is a creative problem formulation by some analyst regarding how to cast the business problem as one or more data science problems. High-level knowledge of the fundamentals helps creative business analysts see novel formulations.

Stage : Data Analysis and Modeling

What is it : Maxim

Insight

If solving the business problem is the goal, the data comprise the available raw material from which the solution will be built. It is important to understand the strengths and limitations of the data because rarely is there an exact match with the problem. It is also common for the costs of data to vary.

Stage : Data collection and cleaning

What is it : Maxim

Insight

It is not unusual for a business problem to contain several data mining tasks, often of different types, and combining their solutions will be necessary. (My notes: a problem could be a combination: Such a problem usually requires unsupervised approaches such as profiling, clustering, anomaly detection, and co-occurrence grouping.)

Stage :Problem formulation  
What is it: Maxim

Insight

 To facilitate such qualitative assessment, the data scientist must think about the comprehensibility of the model to stakeholders (not just to the data scientists). And if the model itself is not comprehensible (e.g., maybe the model is a very complex mathematical formula), how can the data scientists work to make the behavior of the model be comprehensible.

Stage :Data Analysis and modeling  
What is it: Question

Insights

“Your model is not what the data scientists design, it’s what the engineers build.

Stage : Deployment (Since the ultimate model is the one built and deployed)

What is it: Maxim

Insights

 The CRISP cycle is based around exploration; it iterates on approaches and strategy rather than on software designs.

Stage : Deployment (Since the ultimate model is the one built and deployed)

What is it: Maxim

Insight

 In analytics, it’s more important for individuals to be able to formulate problems well, to prototype solutions quickly, to make reasonable assumptions in the face of ill-structured problems, to design experiments that represent good investments, and to analyze results.

Insights

The same reasoning applies to any computation of summary statistics: have you thought about the problem you would like to solve or the question you would like to answer? Have you considered the **distribution** of the data, and whether the chosen statistic is appropriate? (mean/median)

Stage : Problem Formulation

What is it: Question

**Chris Wiggins interview**

**Insight**

So there’s an intellectual challenge there that is not exactly the intellectual challenge of machine learning. It’s more the intellectual challenge of trying to use machine learning to answer questions from a real-world domain

Stage : Problem Formulation  
What is it: Question

Insight

**So from web logs to every event when somebody interacts with the mobile app, there are copious, copious data available to this company to figure out: What is it that the readers want? What is it that they value?**

Stage : Problem Formulation  
What is it: Question

**Insight**

One of the driving themes of my work has been taking domain questions and asking: How can I reframe this as a prediction task?

Stage : Problem Formulation  
What is it: Question

**Insight**

**The key is usually to just keep asking, “So what?” You’ve predicted something to this accuracy? So what? Okay, well, these features turned out to be important. So what? Well, this feature may be related to something that you could make a change to in your product decisions or your marketing decisions. So what?**

Stage : Analysis/Deployment (cycle)  
What is it: Maxim

**Insight**

The great thing about predictions is that you can be wrong, which I think is hugely important. I can’t sleep at night if I’m involved in a scientific field where you can’t be wrong.

Stage : Analysis   
What is it: Data Ethics

Insight

It takes a long time to become an expert in something. It takes years of mistakes.

Stage : Analysis   
What is it: Expertise

**Insight**

The world’s like that. The world doesn’t hand you models. It doesn’t come to you with a model and say, “Diagonalize this Hamiltonian.”[10](https://learning-oreilly-com.proxy.lib.umich.edu/library/view/data-scientists-at/9781430265993/9781430265986_Ch01.xhtml#Fn10) It comes to you with observations and a question usually being asked by the person who gathered those data. How do you explore a data set that you’ve been handed?

Stage : Problem definition   
What is it: Question

* **Chapter 2 - Business Problems and Data Science Solutions**

**Insight 1:**

To facilitate such qualitative assessment, the data scientist must think about the *comprehensibility* of the model to stakeholders (not just to the data scientists). And if the model itself is not comprehensible (e.g., maybe the model is a very complex mathematical formula), how can the data scientists work to make the behavior of the model be comprehensible.   
The biggest challenge in presenting your analysis is the communication of the findings so others find them actionable and at the very least relevant.  
Stage :Data Analysis and modeling  
What is it: Question

            I**nsight 2:**  
 Your model is not what the data scientists design, it’s what the engineers build.

            Transferring the model design to the engineers is the other end of  of this

            communication channel. These are the developers and deployers who are working on

            constructing the model for deployment  
 **Stage** : Deployment (Since the ultimate model is the one built and deployed)  
 **What is i**t:  Maxim **Insight 3:** The CRISP cycle is based around exploration; it iterates on *approaches* and *strategy*

            rather than on software designs.

            There is a different cadence to Data Science projects that typical software development

projects . The exploratory aspect being paramount.  
 **Stage** : Deployment (Since the ultimate model is the one built and deployed)  
            **What is it**: Maxim

**Insight 4:** The same reasoning applies to any computation of summary statistics: have you

thought about the problem you would like to solve or the question you would like

to answer? Have you considered the distribution of the data, and whether the

chosen statistic is appropriate? (mean/median)

Understanding the distribution of data and not just the statistics (as Chris Wiggins also

states) helps us choose the approach to analysing the data.  
           **Stage** : Problem Formulation  
           **What is it**: Question

* **Chris Wiggins interview**

**Insight 1:**

The key is usually to just keep asking, “So what?” You’ve predicted something to this accuracy? So what? Okay, well, these features turned out to be important. So what? Well, this feature may be related to something that you could make a change to in your product decisions or your marketing decisions. So what?

The questions when you analyze your data typically lead to insights but often to more questions.

**Stage** : Modeling and Analysis/Deployment (cycle)

**What is it**: Maxim

**Insight 2:**

The great thing about predictions is that you can be wrong, which I think is hugely important. I can’t sleep at night if I’m involved in a scientific field where you can’t be wrong.  
**And**   
It takes a long time to become an expert in something. It takes years of mistakes.

This is a field where iterative cycles will offer corrections. And so don’t consider the first iteration to be perfection. This is also what I will find most challenging with my “fix-it” engineering mentality.  
**Stage** : Modeling and Analysis  
**What is it**: Expertise

Erin Shelman

Insights

I’m a co-organizer of the Seattle chapter of PyLadies, which is an international mentorship group whose goal is to help women become active participants and leaders in the Python open-source community.

Stage : Presentation

What is it : Ethics

Insights

Erin talks about two systems her team has built:

Over the last year and a half I’ve mostly worked on Recommendo,

building new algorithms and the real-time scorer. Segmento offers customer segmentation as a service.  
These are two approaches to segmenting and slicing the data and is the questions being asked of the data and extracting information from it.   
Stage : Problem Definition (since they defined the solution from the raw data)  
What is it : Goal

Insight

Erin talks about the close relationships the datascience team keeps with the web team.

We also work hard to have good relationships with the people who are ultimately responsible for getting our work in front of customers—primarily the web team. So we keep in touch with them regularly and let them know what is going on from our side. . We also make sure that when they find bugs that we respond right away.

Stage : Presentation and Deployment  
What is it : Goal

Insight

Erin’s team decided not to build a tool to alert customer when a product was to be replenished wen they learnt that the stylist would lose her commission if this was built. Feeling the pulse of the business and responding to it shows humanity

“However, through the course of our chat we learned that she was not interested in our tool because if a customer she chatted with replenished their product online, our stylist wouldn’t be credited with the sale. “

Stage : problem formulation  
What is it : ethical commitment

Insight

How to find insights in data is presented here:

The most interesting types are data are those collected for one purpose and used for another. For example, one of our developers, Jason Wilson, had a really cool idea to look at what was purchased when people asked for gift receipts. Then you could make recommendations for the most gifted products for an upcoming holiday.

Stage : problem formulation  
What is it : goal

**Insight**

Erin demonstrates how analysis and modeling used in data science is applied to her specific data and scenario here:

I used record linkage to solve the problem. Record linkage is a technique used to find duplicates in things like census data and medical records. In survey data it’s typical to have typos and variations in name spellings and you want to link those separate records into a single entry.

I forced matching on things like product type and brand, and then used fuzzy string matching to measure the similarity between product descriptions. My output was a probability that two items were the same “record” for each candidate product.

Stage : Data Analysis and Modeling

What is it : Expertise ( I think it demonstrates expertise in her field)

**Insight**

What Erin looks for in other people’s work- Presentation. What can I say? I’m shallow. I don’t just mean visual presentation (though it’s important), but the ability to convey results both technically and non-technically. The work needs to communicate the point clearly and coherently.

Stage : Presentation

What is it : Maxim (to sell your model you need a good presentation!)

Jake Porway

Insights

Similar to Erin’s application of statistical algorithms to retail industry, Jake applies these algorithms for mitigating social issues

“I’m really excited because we at DataKind focus on applications of data science to make the world a better place. So using the same technologies that help Netflix recommend movies you want to watch, we apply similar techniques to problems, like sourcing clean water, combating human rights violations, or addressing other pressing social issues. It really feels like a brave new world, and the chance to use data science skills to do something good at this time is just incredibly rewarding”

Also

What has really shocked and surprised me in a good way is that there’s almost no limit to where data and data science can be applied.

Stage : Analysis and Modeling

What is it : Ethical commitment

Insights

Being a liaison between the people who have been working for social improvement and data scientists who can help with technology is Datainds primary goal:” So at the very least, we want to scale DataKind so we can create many similar environments where the power of data science for good can flourish and be usable by the general population.”

Stage : Presentation and Deployment

What is it : Goal

Insights:

Jake explains how data science helped evaluate the effectiveness of a bike donation program in Africa

“They wanted to know if and how well that investment was working. Our volunteer team discovered by looking at the GPS data from their services, that there wasn’t a statistically significant increase in the range for the groups that had the bikes. “

Stage : Presentation and Deployment

What is it : Goal

I**nsights**

Sometime diligence and good observation is all the analysis you need to perform on the data. When statisticians took a good hard look at the data collected in an experiment they realized there were some who were gaming the system.

“Another lesson from this project was that we shouldn’t underestimate how much little things like that can transform an organization. And so finding out about the data was just a simple analysis that found a problem in data quality.”

Insights:

Jake precautions us against finding ourselves doing this:

 As a data scientist, you may find yourself running a version of what are essentially psychological experiments on users. That’s something that people really need to think deeply about.

Stage : Presentation and Deployment

What is it : Ethical commitment

Insights

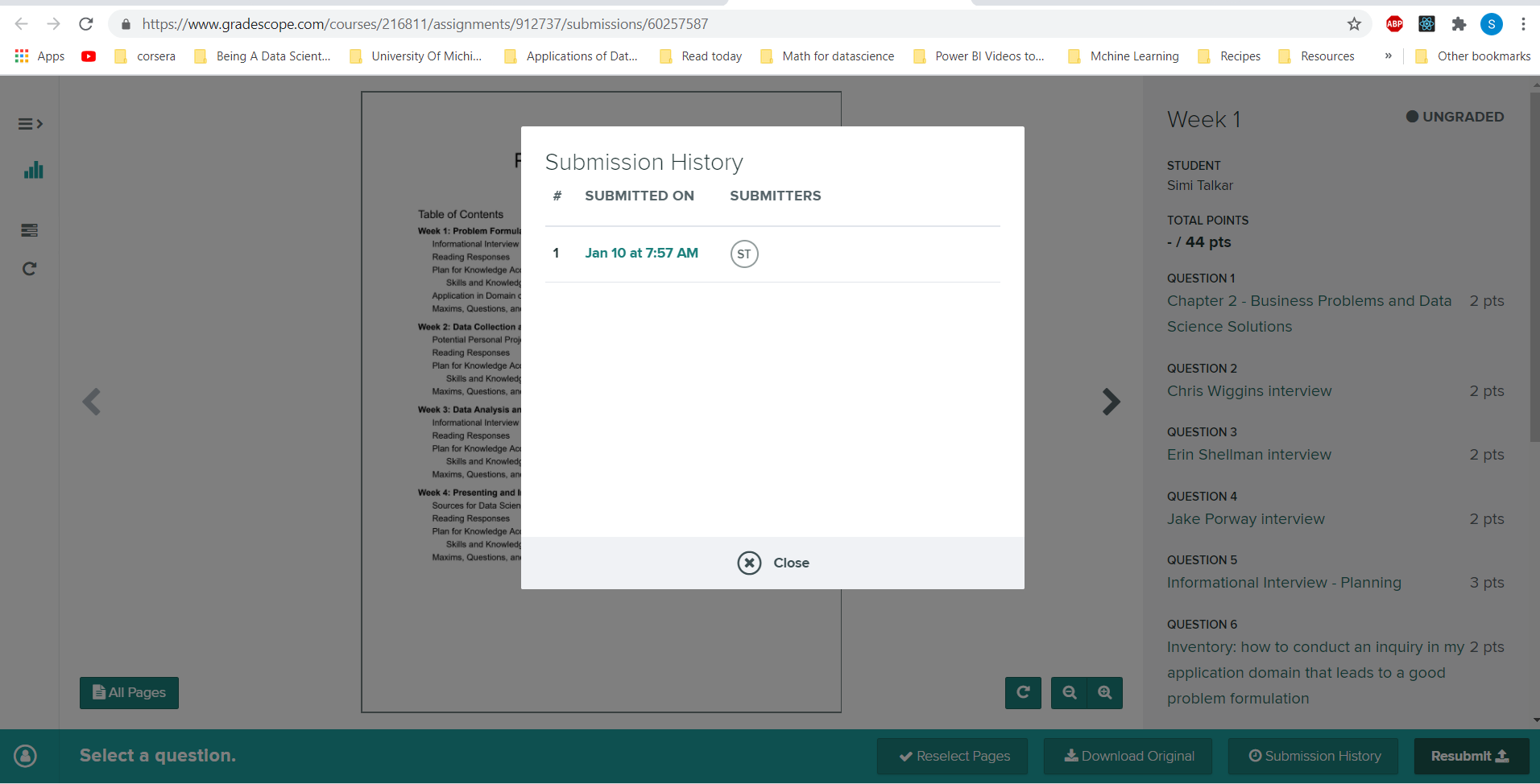
To quote Jake, who questions everything but is an optimist

George Box has been saying for a century that all models are wrong, but some are useful.

Stage : Analysis and modeling

What is it : Maxim

**Problems**



Week 2

Small samples yield extreme results more than large samples do

W.E.I.R.D

Western Educated Industrialized Rich Democratic

Data Cleaning questions:

* Does what I’m looking at making sense?
* Does the data match the column label?
* Does the data abide by the appropriate rules for its field?

\*\* Are the characters in a name only alphabetical (Brendan) or are there numbers in it (B4rendan)?

\*\* Is the numerical portion of a phone number 10 digits (5558675309) or not (675309)?

* Compute summary statistics for the numerical data. Do they make sense?

\*\* If you’re dealing with time elapsed data is the minimum value negative (- 10 seconds)?

\*\* If you’re dealing with annual wage data for blue collared workers is the maximum value something outlandish ($1,000,000)?

* Look at how many values are nulls? Is the number of nulls acceptable? Is there a pattern as to where there are null values?
* Are there duplicates and is that okay?

**WEEK 3**

|  |  |  |
| --- | --- | --- |
| **Problem Formulation** | | **Analysis Technique** |
| **Regression** | * 1. Simple Linear Regression | |
|  | * 1. Neural network | |
| **Classification** | 1. Logistic regression | |
|  | 1. Decision Tree | |
|  | 1. Random Forest | |
|  | 1. Naïve Bayes | |
|  | 1. Neural Network | |
| **Similarity matching** | 1. Jaccard similarity | |
|  | 1. Cosine similarity | |
|  | 1. Neural Network | |
| **Clustering** | 1. K Means | |
|  | 1. PCA | |
| **Co-occurance grouping** | * + 1. KNN | |
|  | * 1. Association rule mining | |
| **Profiling** | Use the other techniques above and fine tune | |
|  | * + 1. Cluster centroid after clustering | |
| **Outlier** | 1. Linear regression with point with large residual error | |
| **Link Prediction** | 1. Classification and regression + Network features | |
| **Data Reduction** | 1. PCA/SVD Matrix factorization | |
|  |  | |
| **Causal Modeling** | 1. Causal diagrams | |
|  | 1. Structural equation modeling | |

In a previous article, we discussed the alpha error rate (or false-positive error rate), which is the probability of falsely rejecting the null hypothesis.[[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4840791/#ref1)] In any study, when two or more groups are compared, there is always a chance of finding a difference between them just by chance. This is known as a Type 1 error, in contrast to a Type 2 error, which consists of failing to detect a difference that truly exists. Conventionally, the alpha error is set at 5% or less which ensures that when we do find a difference between the groups, we can be at least 95% confident that this is a true difference and not a chance finding.

**Over-fitting and cross validation**

Choosing a model that does not for the training data too closely.

Signal is the pattern that you see in the sample data that you will also see in the test data

Noise is the things you should have ignored in the training data.

Bias Variance trade-off

**Overcoming Over-fitting**

**Cross-validation K-fold**

**The normal process – leave off some data for testing (don’t use it for cross validation either to avoid leakage)**

**Week 3 Reading**

**Overfitting in Machine learning**

* ***Overfitting in Machine Learning: What is it and how to prevent it***

***Insight 1***

*“If the algorithm is too complex or flexible (e.g. it has too many input features or it’s not properly regularized), it can end up “memorizing the noise” instead of finding the signal.”*

A simpler regularized model will ignore the noise in the unseen test data better.

Stage: Data Analysis and modeling

What it is: Maxim

Insight 2

**“Both bias and variance are forms of prediction error in machine learning.”**

Stage: Data Analysis and modeling

What it is: Maxim

**Insight 3:**

“**If our model does much better on the training set than on the test set, then we’re likely overfitting.”**

The bias in such a situation is low and the variance is high.

Stage: Data Analysis and modeling

What it is: Maxim

**Insight 4:**

**“Another tip is to start with a very simple model to serve as a benchmark.”**

Since playing this tug-of-war between bias and variance is an art form (asper Prof Resnick), this tip can serve well as you try more complex algorithms to check and see if te additional complexity is worth it.

Stage: Data Analysis and modeling

What it is: Maxim

**Insight 5:**

Solutions to over-fitting include

1. Cross-validation
2. Train with more data
3. Remove features
4. Early stopping
5. Regularization

* ***Common pitfalls in statistical analysis: The perils of multiple testing***

***Insight 1***

***“***IS ADJUSTMENT OR COMMON SENSE NEEDED FOR MULTIPLE TESTING?”

Stage: Data Analysis and modeling

What it is: Question

The below are the common sense tactics to adopt when evaluating significance

***Insight 3***

“*Readers should evaluate the quality of the study and the actual effect size instead of focusing only on statistical significance”*

Stage: Data Analysis and modeling

What it is: Maxim

Insight 3:

“*Results from single studies should not be used to make treatment decisions; instead, one should look for scientific plausibility and supporting data from other studies which can validate the results of the original* study”

Stage: Data Analysis and modeling

What it is: Maxim

“Authors should try to limit comparisons between groups and identify a single primary endpoint; using a composite endpoint or global assessment tool is also an acceptable alternative to using multiple endpoints.”

Stage: Data Analysis and modeling

What it is: Maxim

* ***P-Hacking and the problem with Multiple Comparisons***

***Insights 1***

**“**P-hacking is a close cousin of the multiple comparison problem, but here the motivation is a bit more sinister.”

**With multiple comparison, the chance of p-value falling below significance level is made a possibility at least once within several trials** The big problem with p-hacking is that we simply do not know if the strength of the relationship found is purely an artifact of the sample, the analytical method used, or legitimate judgment calls made by the researcher.  We just don’t know.

Stage: Data Analysis and modeling

What it is: Ethical consideration

**Insights 2:**

“With HARKing, the analyst presents a hypothesis as if he/she set out to test that relationship right from the very beginning, but had already completed all of the analysis and knows how the data turns out.  [It is, effectively, the very opposite of the scientific method](http://jom.sagepub.com/content/early/2014/03/18/0149206314527133).”

Stage: Data Analysis and modeling

What it is: Ethical consideration

* ***Correlation vs. Causation: An Example***

***Insight 1***

“The common problem in these articles is that they take two correlated trends and present it as one phenomenon causing the other.”

Ask yourself if the two variables being related are correlated or if there is truly a cause-effect relationship

**Stage**: Data Analysis and modeling/Presentation and Deployment

**What it is:** Question

**Insight 2**

*“Humans*[*naturally see patterns*](http://bigthink.com/endless-innovation/humans-are-the-worlds-best-pattern-recognition-machines-but-for-how-long)*where they don’t exist, and we like to tell a cohesive story about what we think is going on (the narrative fallacy). However, the world usually does not have defined causes and effects, and we*[*must settle for correlations*](https://www.forbes.com/sites/gilpress/2013/04/19/big-data-news-roundup-correlation-vs-causation/)*.”*

Unless **other tests** such as randomized controlled trial is conducted, it will be premature to come to a conclusion of causation over correlation.

**Stage**: Data Analysis and modeling/Presentation and Deployment

**What it is:** Maxim/Ethical considerations?

* ***Simpson’s Paradox in Real Life*** *or* ***Ignoring a Covariate: An Example of Simpson’s Paradox***

**Insights 1:**

*“If the populations re separated in parallel into a set of descrptive categories, the population with higher overall incience may ye exhibit a lower incidence within each category”*

This occurs because you have to take into account the **weight** of each contrbuting category to the overall when looking at the overall increase or decrease.

**Stage**: Data Analysis and modeling

**What it is:** Maxim

**Insight 2**

“The actual effect of smoking will be underestimated from these data, as the population seen in the original survey was already subject to selection: the small proportion of the older women smoking is likely to be due not only to a low proportion in that cohort being smokers, but also to those who had smoked being less likely to survive to be seen in the original study”

Beware of survival bias

**Stage**: Data collection and cleaning

**What it is:** Maxim

* ***Conditioning on a collider***

***Insights 1:***

*“Causal inference from observational data boils down to assumptions you have to make[ and third variables you have to take into account.”*

*The third variable is the confounding variable that cause spurious correlation to be derive between the other two variables.\*

**Stage**: Data analysis and modeling

**What it is:** Maxim

**Insights 2:**

*“Assuming that women (X1) have worse chances to get into the limelight than men, but overstating the implications of your evidence (X2) helps with getting into the limelight; we could find that women in the limelight (conditioning on Y) are more likely to have overstated their evidence because the more tempered women simply didn’t make it. That’s obviously just wild speculation, but in everyday life, people are very willing to speculate about confounding variables, so why not speculate a collider for a change?”*

If the third variable is a mediator ot a collider then it is a bad idea to condition on it.

I will never knowingly report on an analysis whicj ocnditions on a mediator or collider.

WEEK 4 Readingg

Presentations

SUCCESS

S -- Simple

U – Unexpected/some surprose

C—Concrete

C – Credible

E - Emotional

S - Stories

**Black box model: For certain inputs you get outputs**.

The model gives outputs for certain inputs,

but people don't see or can't

understand how that output is derived from the inputs

**Not Black Box – interpretable – if a human can**

By contrast, a model is interpretable not black

box if a human can guess the output for a given input,

if a human can describe what features in

the input are affecting the output,

and if a human can describe

input changes that would affect the output

**Aim is to make a model interpretable**

**There are many model families that are not interpretable such as some neural networks . But for the sake of interpretability you might be sacrificing accuracy. This depends on the domain (stocks)**

In machine learning, **accuracy** is measured by comparing the output of a machine learning model to the known actual values from the input data set.

**WEEK 4 READING**

* ***A History Lesson On the Dangers Of Letting Data Speak For Itself***
* ***Storytelling for Data Scientists***
* ***Interpretability is crucial for trusting AI and machine learning***
* ***The Signal and the Noise, Chapter 2***
* ***The Signal and the Noise, Chapter 6***
* ***How Not to Be Misled by the Jobs Report***
* ***But what is this "machine learning engineer" actually doing?***
* ***How we scaled data science to all sides of Airbnb over 5 years of hypergrowth***

***{Stage the insight is relevant for: problem formulation;  
data collection & cleaning;  
 modeling & analysis;   
presentation & deployment} -   
{which of the following it is:  
 expertise;  
 goal;  
 maxim;   
question;  
 ethical commitment}***

* ***A History Lesson On the Dangers Of Letting Data Speak For Itself***

**Insight 1:**

***“S****emmelweis’s data met three key criteria—it was*truthful*,*valuable*and*actionable*—but he ultimately failed to see his ideas adopted in his lifetime. The Hungarian physician stumbled in one essential area—the communication of his data.”*

Stage: ***presentation & deployment***What it is: Goal

Insight 2:

The four oversights that can contribute to miscommunication of data analysis:

. Timeliness and Clarity

Communicate your results early and often clarifying margin of error. This is an iterative process after all and keeping people in the loop will refresh their memories and help them keep abreast on the next milestone of your discovery.  
Note: Idea echoed by my interviewee Sandeep Pawar

**.** Audience and the Curse of Knowledge (cognitive bias)

Definition : [Cognitive bias](https://en.wikipedia.org/wiki/Cognitive_bias)  occurs when an individual, communicating with other individuals, unknowingly assumes that the others have the background to understand

Respect your audiences’ capabilities as well as limitations. Maintain an amicable relation with them. To foster change, you have to take people along with you. Antagonizing them serves neither them nor you any purpose.

Note: Idea echoed by my interviewee Erin Shellman interview of week 1

**.** Narrative Evokes Emotion

**“T***he data-driven doctor missed an opportunity to weave his facts into a compelling data story that connected with his audience on an emotional level.”*

Data speaks to human when presented at an emotional level. As sentient beings, we need to feel empathy, to react favorably to the change being suggested.

**.** The Power of Data Visualization

Data tables have their usages, but visualization when done right has an immediate impact of conveying results.

Stage**: *Presentation & Deployment***What it is: Maxims

* ***Storytelling for Data Scientists***

**Insight 1**

*“Artistotle’s prescription for effective communication : Ethos establishes the credibility of the speaker, and Logos appeals to the reasoning used. Yet, both are useless without Pathos, the way to the heart of the audience.”*

Appealing to the “human” aspect of you audience will transmit you message to them more effectively than only reasoning and logic.

Stage: Presentation and deployment  
What it is: Maxim

**Insight 2**

“*On a daily basis, Data Scientists need to explain results of complex models simply and understandably to stakeholders. They also need to pitch projects. When doing either, storytelling should be the tool of choice for Data Scientists.”*

Weave a story around your data. Just like early maxims are instilled in children with fairy tales and mythology, adults respond to stories as well and will absorb the takeaways more readily from a captivatingly narrated tale.

Stage: Presentation and deployment  
What it is: Maxim

**Insight 3**

**SUCCESs**

**“Concrete:** Use concrete examples. Help your listeners or readers picture the situation that you’re trying to solve.”

Vague theories may lead to inaction or ineffectual response. I find this during the MADS course as well. I learn much more from a lecture where concept is linked to an working example.

Stage: Presentation and deployment  
What it is: Maxim

* ***Interpretability is crucial for trusting AI and machine learning***

*“The algorithms inside the black box models do not expose their secrets. They don't, in general, provide a clear explanation of why they made a certain prediction. They just give us a probability, and they are opaque and hard to interpret. Sometimes there are thousands (even millions) of model parameters, there’s no one-to-one relationship between input features and parameters, and often combinations of multiple models using many parameters affect the prediction. Some of them are also data-hungry. They need enormous amounts of data to achieve high accuracy. It’s hard to figure out what they learned from those data sets and which of those data points have more influence on the outcome than the others.”*

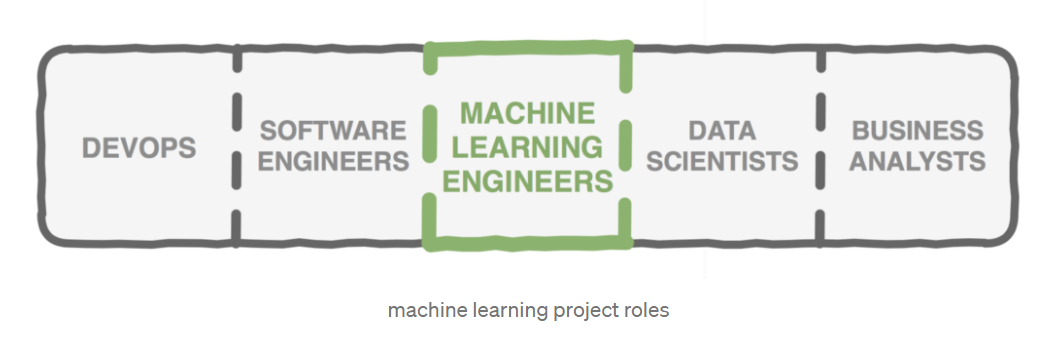
What secrets does the black box algorithmic model hold. Is it too complex to be interpretable? If they learn the wrong thing will you be able to detect that? Can we trust the model to make fair decisions upon deployment?

**Stage** : Presentation and Deployment  
**What it is** : Question

* ***But what is this "machine learning engineer" actually doing?***

**Insight 1**

**How do a machine learning engineer and a data scientist fit into a software development model**



**Status** : Deployment

**What it is** : Question

**Insight 2**

*“As good and customizable as the current ML frameworks are, sooner or later your teammates will have an intriguing use case that is not achievable with any of them. Well, not with standard APIs. But when you dig into their internals, tweak them a little and mix in another library or two, you make it possible.****You abuse the frameworks and use them to their full potential.****That requires both extensive programming and machine learning knowledge, something that is quite unique to your role in the team.”*

Not all your requirements will be available out of the box, you will have to develop expertise in clubbing together libraries and APIs for your project specific goal.

**Status** : Deployment

**What it is** : Expertise

**Insight 3:**

*“You must stay up to date with the state of art technologies and constantly look for the places in which the overall product performance could be improved.”*

**An expertise you (ML engineer) will have to work towards – will be to scout around for the best product for scheduling, monitoring..**

**Status** : Deployment

**What it is** : Expertise

**Insight 4:**

“***What is interpretability?***

*Interpretability means giving explanations to the end users for a particular decision or process. More specifically, it entails:*

* *Understanding the main tasks that affect the outcomes.*
* *Explaining the decisions that are made by an algorithm.*
* *Finding out the patterns/rules/features that are learned by an algorithm.*
* *Being critical about the results.*
* *Exploring the unknown unknowns for your algorithm.”*

Stage:  
What it is:

* ***The Signal and the Noise, Chapter 2***

***Insight 1***

*“This does not mean that you must feel totally dispassionate about a political event in order to make a good prediction about it. But it does mean that a fox’s aloof attitude can pay dividends.”*

*Going in with pre-set ideas and a unilateral way of thinking, can cloud our critical abilities to analyze the data. In the assigned reading, people were picking factors they “assumed” to be important like gender, race and an off chance incident to predict an event. A more rational approach that will lead to a prediction with admittedly known error, is that a quantifiable result can be derived from quantifiable inputs.*

*(Example: The inputs Nate Silver considered were poll numbers – weighted by the distance of the poll from the election, funds, reliability of the polling firms ….)*

*Stage: Data Analysis  
What it is: Maxim*

***Insight 2***

*“This probabilistic principle also holds when I am forecasting the outcome in an individual race. How likely is a candidate to win, for instance, if he’s ahead by five points in the polls?”*

*Stage: Data Analysis  
What it is: Maxim*

* ***How Not to Be Misled by the Jobs Report***

**Insight 1  
“**But what if all the worries were based on nothing more than random statistical noise?”

Uncertainty has to be assumed and conveyed when predictions are made and the question has to asked about how much the range of uncertainty is. You also have to question what the sampling error is and take that into account.

NOTE: Definition of sampling error: In statistics, **sampling error** is the **error** caused by observing a sample instead of the whole population. The **sampling error** is the difference between a sample statistic used to estimate a population parameter and the actual but unknown value of the parameter.

Stage: Data Analysis  
What it is: Maxim

*‘*

* ***How we scaled data science to all sides of Airbnb over 5 years of hypergrowth***

Insight 1

“Just as it became impossible to meet every new member of the community, it was now also impossible to meet and work with every employee. We needed to find a way to democratize our work, broadening from individual interactions, to empowering teams, the company, and even our community.”

And   
“Empowering teams is about removing the burden of reporting and basic data exploration from the shoulders of data scientists so they can focus on more impactful work”

Democratization is one of the keys to scaling data science in a company. When all employees of a company have at least the basic knowledge about data science, they can be on the look out for data, lend a hand in the building of tools to use models (ML engineers for instance), will be in step with the scientists when data is presented to them.

Stage: All  
What it is: Goal

Insight 2

While answering questions and measuring things is certainly part of the job, at Airbnb we characterize data in a more human light: **it’s the voice of our customers.**

**And**

**“**Thus, data science is an act of interpretation — we translate the customer’s “voice” into a language more suitable for decision-making.”

Voices, experiences, feelings, seemingly intangible human emotions are converted to data derivable formats for decision making.

Stage: Presentation and deployment  
What it is: Ethical committment (to look beyond the cold hard numbers)

Insight 3

*“But when decision-makers don’t understand the ramifications of an insight, they don’t act on it. When they don’t act on it, the value of the insight is lost.”*

The entire effort of problem formulation, data collection , analysis goes to waste if the result is not clearly communicated with a purpose and/or actionable goal emerging out of it.

Stage: All  
What it is: Maxim