**Statistics classes don’t teach you about money**

**The practical first step in a data science project**

[Statistics](http://bit.ly/quaesita_statistics) and [data science](http://bit.ly/quaesita_datascim) classes don’t do much to prepare students for hashing out project specifications on the job. Here are a few problems:

* They teach you to take the problem statement for granted, but what if your boss doesn’t have the skills to make sensible requests?
* They don’t teach you about money and data budgeting, so you learn how to calculate the ideal data requirements but not the cost-benefit thinking…
* They don’t teach you negotiation skills, especially the flexibility that helps you navigate the gray areas of negotiating a workable budget to execute data collection.
* They are more likely to be oriented on [inherited data than on primary data](http://bit.ly/quaesita_provenance) settings, since that’s easier for your prof to grade.
* They leave out the real-world details.



Modified from photo by [Ante Hamersmit](https://unsplash.com/@ante_kante?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

**A data scientist’s first day on the job**

Imagine that you’re a [data scientist](http://bit.ly/quaesita_datascim) who has been hired to [estimate](http://bit.ly/quaesita_vocab) the average height of pine trees in the forest pictured below.



How tall are the trees in this forest? Photo by [Marita Kavelashvili](https://unsplash.com/@maritafox?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

*(Note: the links in this article take you to my lighthearted explanations of any jargon terms that crop up.)*

**Facts versus statistics**

If we were to perfectly measure every single tree, we would get something far better than an [estimate](http://bit.ly/quaesita_vocab); we would get a **fact**. The actual [truth](http://bit.ly/quaesita_saddest) about the heights of the trees in this forest. When you have facts, you don’t need [statistics](http://bit.ly/quaesita_statistics).

When you have facts, you don’t need statistics.

Should you then go out and measure every tree’s Planck Length (the smallest unit of length in physics, with one unit equal to 0.00000000000000000000000000000000001616255 meters)? Which instrument would you use to get such a precise measurement? I bet you don’t have it lying around in your garage, especially since it hasn’t been invented yet.

Even if we settled for humanity’s most precise measuring device (orders of magnitude too imprecise if you have your heart set on Planck Length), one tree measured with it would likely be much too expensive for whatever purpose motivated your boss to hire you.

Furthermore, even if you settled for plank length instead of Planck Length and allowed yourself round off to the nearest meter, measuring *every* tree would be overkill… your forest is much too big. Would your boss approve of your completionist desire to collect-’em-all?

Statistical sampling is all about getting a perspective on your problem that’s less perfect than a fact but good enough.

If you’re thinking like a good [statistician](http://bit.ly/quaesita_statistics), you’re immune to the perfectionist impulse — why measure the whole [population](http://bit.ly/quaesita_vocab) when you can get a *good enough* [estimate](http://bit.ly/quaesita_vocab) by taking a [sample](http://bit.ly/quaesita_vocab)? Sure, this introduces [uncertainty](http://bit.ly/quaesita_uncertainty) (we’re no longer dealing with facts) but perhaps we can live with that. Let’s measure a good enough sample of trees so we don’t have to measure them all!

**But wait… what is “good enough”?**

We haven’t even gone anywhere near a tree and we’re already stumbling into two hurdles with our seemingly-simple tree measuring task:

* **If we’re not measuring in Planck Length, how precise should the measurement be?**
* **If we’re not measuring all the trees, how many trees should we measure?**

The way through both of these questions is to understand ***why*** your project exists in the first place: what is the purpose of the task and what does *“good enough”* actually mean? This is a cost-benefit kind of question which you can’t answer without understanding the real world aspects of the project.

Begin with why: why are you collecting data? What is the purpose of your project? What does *“good enough”* actually mean?

Unfortunately, if you’re the new hire on your team, setting the bar for “good enough” is, strictly speaking, someone else’s job. This someone is usually The Boss. Unless \*you’re\* the boss, it’s not your call to make. If you treat real-world data problems like homework questions, this will be a struggle for you.



Modified from photo by [micheile dot com](https://unsplash.com/@micheile?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

**Statistics classes don’t teach you about money**

The first problem is that classroom courses for [data science professionals](http://bit.ly/quaesita_universe) rarely rub your nose in [data](http://bit.ly/quaesita_hist) budgeting. Most homework problems ask you to take [sample size](http://bit.ly/quaesita_vocab) for granted, wiring your brain to work with [inherited data](http://bit.ly/quaesita_notyours) but doing nothing to help you handle data collection negotiations in the real world.

Stop treating data like it’s priceless. Data isn’t sacred; it’s a resource like any other.

Other homework problems teach you to calculate the [sample size](http://bit.ly/quaesita_vocab) you need without ever preparing you for the next bit: how to scare up the money you’d need to actually get your hands on this ideal sample size. (Not to mention the etiquette of explaining a power analysis budget curve to a boss with a numbers allergy.) One of this educational oversight’s most pugnacious manifestations is a habit of treating data as priceless, resulting in odd behaviors that look damned-near infantile to every other adult on your team. In the real world, there’s scarcity and nice things cost money. This applies to data too. Data isn’t sacred; it’s a resource like any other.

**Do bosses understand what they’re asking for?**

The second problem is your boss’s skill level. If you take charge of the situation (you leader you!) and do the work without taking the time to fully understand your boss’s vision, you’re in danger of crafting a solution that doesn’t fit the problem.

On the other hand, if you approach your boss with a request for measurement and sample size specifications, well, here be dragons too.

Suppose your boss answers, *“Twenty trees measured in feet, please.”*

It takes skill to convert a vision for the project into [sample size](http://bit.ly/quaesita_gistlist) requirements and until you know [your boss’s decision-making skill level](http://bit.ly/quaesita_ytjenny), it’s hard to judge whether their response is well-considered or lazy. It could be exactly what you need in order to move forward, but unless your boss has experience with data and measurement, their off-the-cuff answer might shoot the project in the foot. There’s a solid chance they’re sending you on a wild goose chase.

Until you’ve worked closely with your boss, you won’t know.

Let’s talk about decision skills! Watch on YouTube at [bit.ly/quaesita\_ytjenny](http://bit.ly/quaesita_ytjenny)

**Assumptions, assumptions, assumptions**

As soon as you’re dealing with uncertainty, you’re going to need a bridge from the facts you have (your sample of a few trees) and the facts you wish you had (your population of all the trees in the forest). That bridge is [assumptions](http://bit.ly/quaesita_saddest). Assumptions are what make a [statistics](http://bit.ly/quaesita_fisher) project tick.

DATA + ASSUMPTIONS = INFERENCE

The tricky part is that your boss — not you! — is the one who’s responsible for setting the project’s [assumptions](http://bit.ly/quaesita_jupimoon). If you’re not the decision-maker, then your job is to serve as an interpreter between mathematics and whatever’s in your boss’s head. That’s another skill they rarely cover in class.

**[Tough Love for Naïve Data Newbies](https://towardsdatascience.com/tough-love-for-naïve-data-newbies-5dd376693eea" \t "_blank)**

**[3 mistakes that statistically-ignorant data nerds make](https://towardsdatascience.com/tough-love-for-naïve-data-newbies-5dd376693eea" \t "_blank)**

[towardsdatascience.com](https://towardsdatascience.com/tough-love-for-naïve-data-newbies-5dd376693eea" \t "_blank)

**Leaving out the real world**

I’ll cover this one in the next article, but the short version is that school leaves most of the real-world details out. Just like this paragraph.

**The real first step in a data project**

[Decision scientists](http://bit.ly/quaesita_di) and more seasoned [data scientists](http://bit.ly/quaesita_datascim) start every project by interviewing the boss carefully to make sure that the specs of the data collection request are clear and that they match the boss’s vision for the project, while balancing the cost-benefit aspects of the data collection process. Alas, this is a skillset you’re unlikely to pick up in class. Without it, there’s a good chance that you’ll either usurp the boss’s role or panic and do exactly what the boss says. Both are bad!

If you’re an inexperienced data worker, there’s a good chance that you’ll either usurp the boss’s role or panic and do exactly what the boss says. Both are bad!

It’s only safe to move from the realm of facts to the realm of uncertainty when the person in charge has a clear vision of what “good enough” means for the project and has the ability (via their own skill or a colleague’s help) to convert this into language that data professionals can work with. Everything should start with purpose — the ***why*** of the project — and carefully consider the cost-benefit realities of information.

And that means that your first real task on any data project has relatively little to do with numbers and much more to do with psychology and communication.

Every data project begins with one essential step: understanding your boss and your business.

Every data project begins with one essential step: understanding your boss and your business. Skip this step at your peril!