Seven Sins of a Newbie Data Science

(and how not to commit them)

- Sarah Masud, Red Hat

About Me

Github: sara-02
Blog: themessier.wordpress.com

Me Learning To Give Back:

1. Open Source Contributions
2. Blogs
3. Meetups, Conferences
4. Mentorship
5. Program review committees

Let’s begin ;)

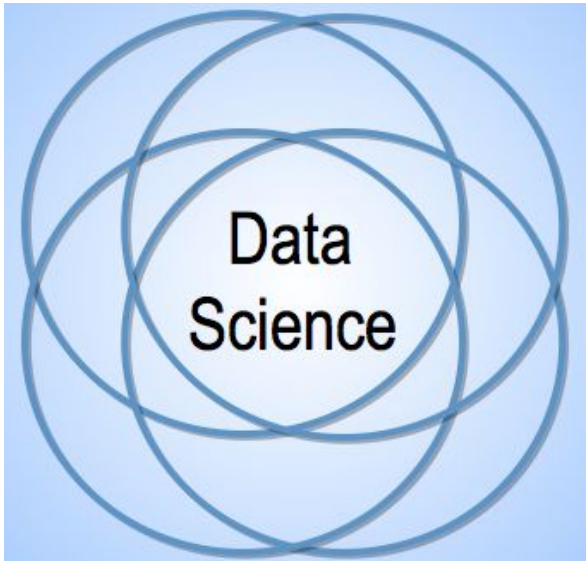


Image: https://commons.wikimedia.org/wiki/File:DataScienceLogo.png

Image: https://chroniclesofanassistant.wordpress.com/2010/11/14/first-day-of-work/

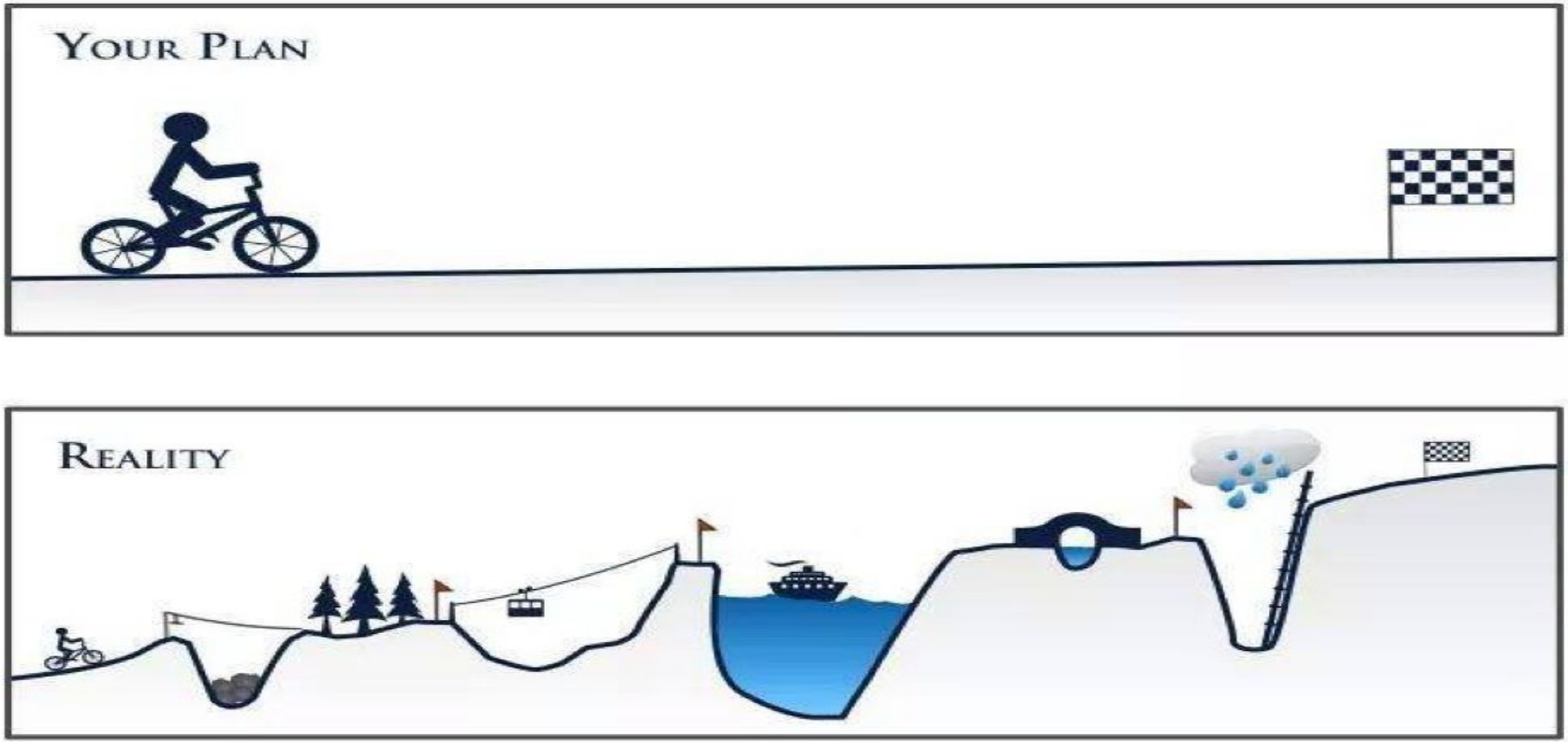


Image: https://www.kdnuggets.com/2016/10/big-data-science-expectation-reality.html

1: The Problem Statement

At College:

"On a ***loan data-set***, using ***logistic regression*** determine if person will ***default or not.***"

1: The Problem Statement

At Work:

“We have been collecting **these data points** since past 3 years. See **what can be done** to **monetize it**.”

1: The Problem Statement

Solution

1. Understand the business needs!
2. Then understand the data collected.
3. Finally translate the vague problem into a known one.

2: Show Me the data

At College:

“Use the data from **Kaggle**, **UCLA registry**, **Image-Net, Wikipedia**...”

Image: https://me.me/i/show-me-the-data-9747283

2: Show Me the data

At Work:

“Use **whatever data is legally available**, but get this problem solved!”

2: Show Me the data

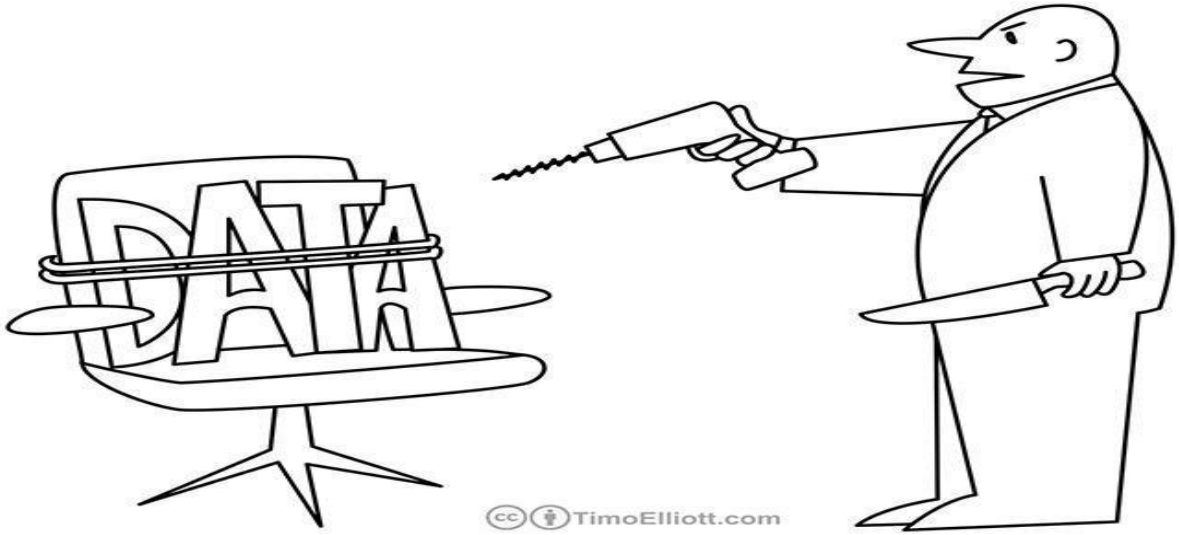
Solution:

1. Don’t expect someone to give you the data willingly!
2. Learn to deal with lack of labelled data.
3. Learn Web Scraping/Data ingestion pipelines.

3. Using A Missile Gun To Kill The Chicken

At College:

“Sounds **cool**! Let me use this **SOTA** algorithm.”



"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Image: https://pbs.twimg.com/media/B83v847CUAAQHKg.jpg:large

3. Using A Missile Gun To Kill The Chicken

At Work:

“Provide us with a **cheap**, **accurate**, **stable** solution.”

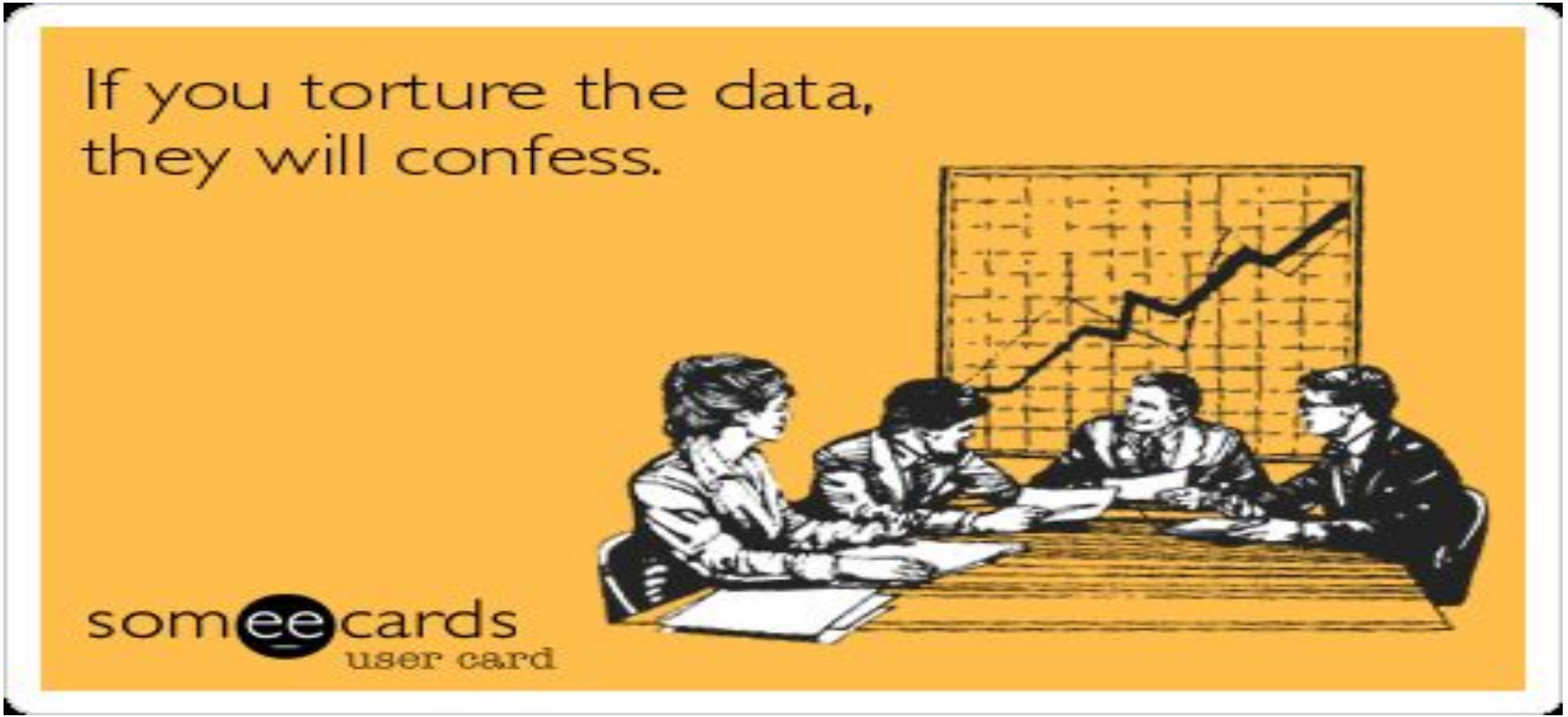


Image: https://www.someecards.com/usercards/viewcard/if-you-torture-the-data-they-will-confess-94dd7

3. Using A Missile Gun To Kill The Chicken

Solutions:

1. Not every problem needs to be a DS problem!
2. Use switch cases if that is enough.
3. Understand the **business constraints**.

4: The Value of Your Work

At College:

1. Accuracy of model.
2. Number of research papers.
3. Subject grade!

4: The Value of Your Work

At work

1. RoI.
2. RoI.
3. RoI.

Image: https://me.me/i/show-me-the-money-memes-11885126

4: The Value of Your Work

Solution:

1. Understand the **business**.
2. Optimise for **Accuracy vs Cost**.
3. Keep the **end user** in mind.

5: Serving the model

At College

“It about building **most accurate system**, running it from the **terminal**. And that is it!”

5: Serving the model

At Work:

1. How many concurrent users can we serve?
2. What time delay can we afford, before we lose the customer?

5: Serving the model

Industry:

1. How is the model exposed to UI?
2. Can the model be distributed?
3. Can the model scale with increase in data?

6. Know Thy Audience

At College:

“Technical mentors, peers.”

6. Know Thy Audience

At Work:

“Audience is always a mixed Baggage.”

6. Know Thy Audience

Solution:

1. Know you concepts well.
2. Teaching DS to your grandma style of conversations.

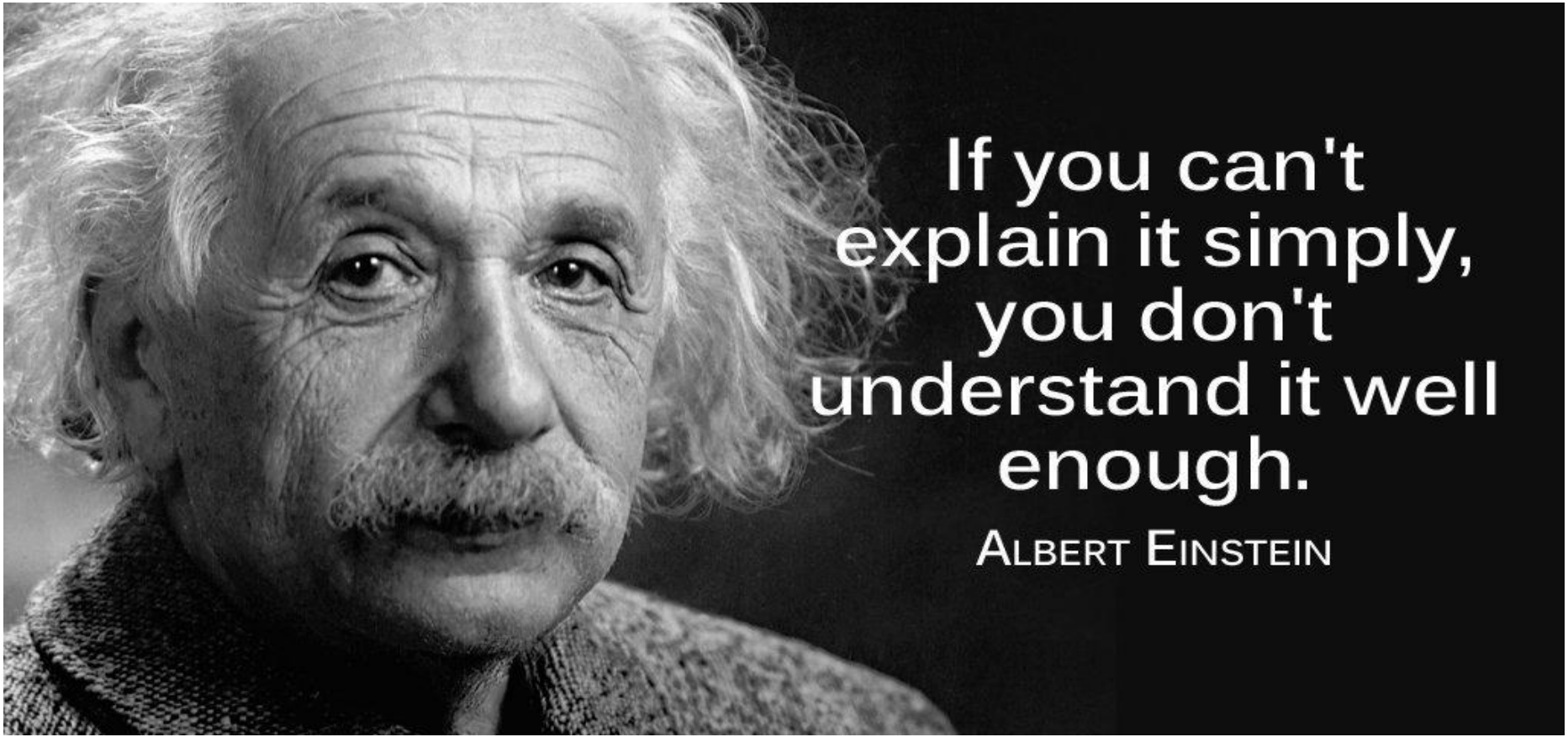


Image: http://www.combine-lab.com/if-you-cant-explain-it-simply-you-dont-understand-it-well-enough/

7. Entropy sets in

At College:

“Build once, use once, and then forget it!”

7. Entropy sets in

At Work:

“The same model and code can be used in production for years without replacement.”

7. Entropy sets in

Solution:

1. Build scalable robust models.
2. Perform regular model evaluation.
3. Re-train the model from time to time.

Love the problem, not your solution.
Learn to **Unlearn → Relearn → Remodel.**

BECAUSE ...

Image: https://www.cafepress.com/+entropy_always_wins_3_shot_glass,1289685014

Thank You
Q & A