

# Setting Up Your Public Data for Success

Rachael Tatman  
March 25, 2019



@rctatman

# Why share data?

1. “Broader impact”
2. 9% more citations  
(Piwowar & Vision, 2013)
3. A philosophical  
commitment to the ideals  
of open science
4. **You want people to use it**



**Expectation!**



**Reality :'(**



# How can you make your public datasets successful?

- Success = someone using your dataset
  - Downloads are probably the easiest things to track
- What needs to happen for someone to download your data?
  - They know that it exists
  - They know how to use it
  - They need to find the content interesting enough to work with





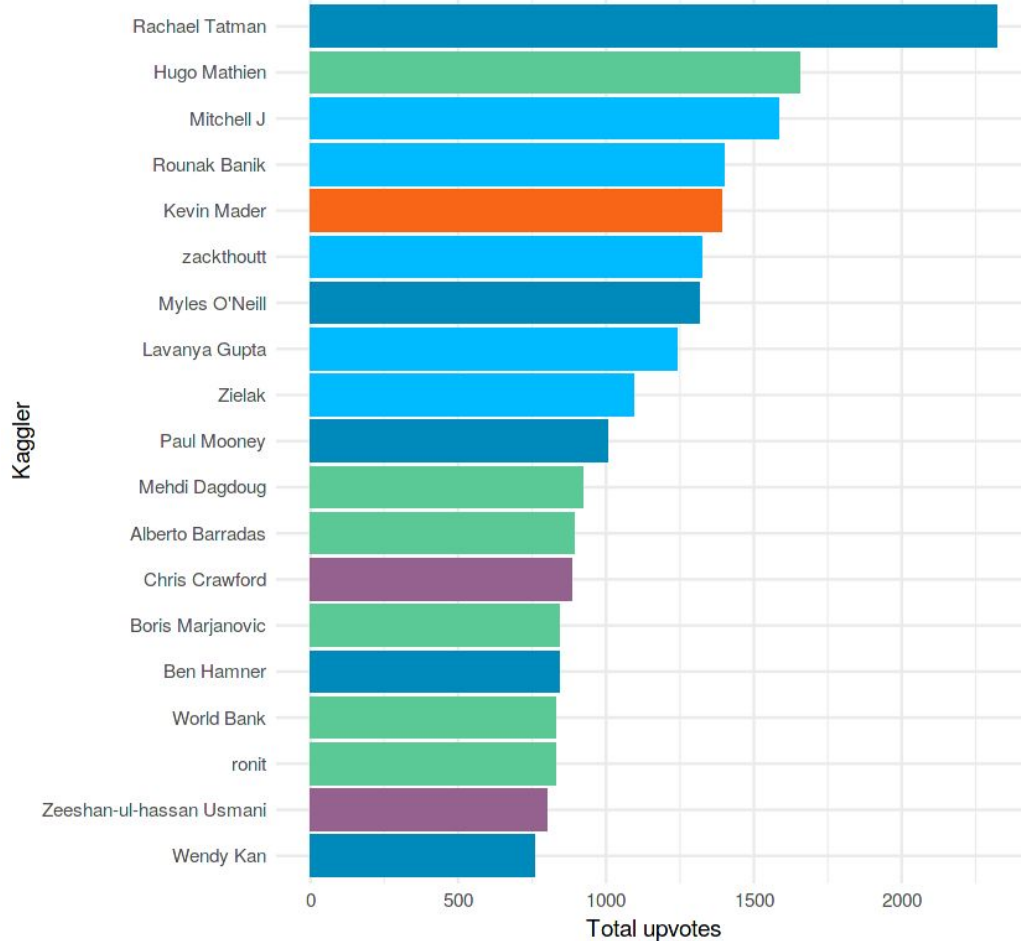
# How come I can help?

- I've seen Kaggle's datasets platform grow from 500 to more than 15k datasets (and uploaded around 100 of my own)
- I spend a lot of time interacting with the most voracious consumers of public data: **aspiring data scientists**
- I ran the numbers! [We make our own data public](#), through, so you're free to poke around on your own.



## Kagglers whose datasets most upvoted

Top 20 Kagglers aggregated by the upvotes their Datasets received



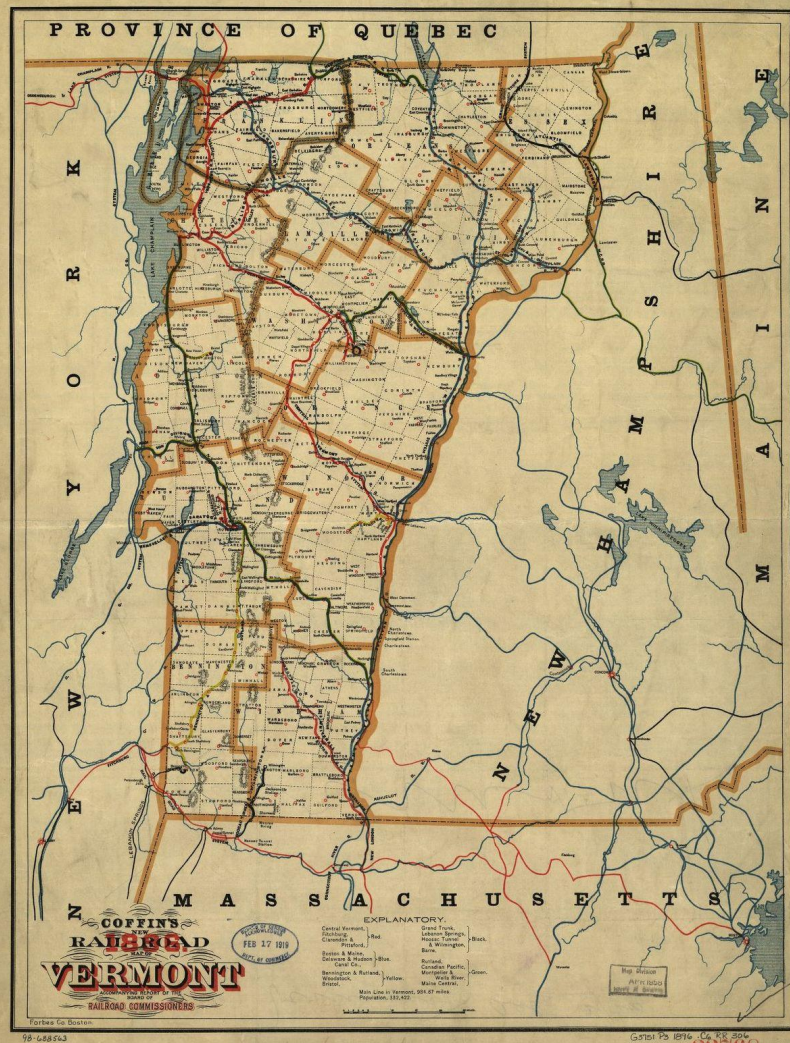
<https://www.kaggle.com/nulldata/popular-kagglers-on-the-datasets-analysis>





# Data Sharing Roadmap

1. During data collection
2. Before data release
3. After data release



# During Data Collection

- Consider the potential audience for your data
  - Researchers
  - Journalists
  - Learners and educators
  - Hobbyists
- Collect data that's as rich as possible while maintaining anonymity (if relevant)
  - Think about what potential audiences might be interested in investigating
  - Consider reporting k-anonymity, t-closeness, l-diversity
  - If you're sharing raw data files you can't use differential privacy





# Before Data Release:

- Prepare clear documentation and metadata
- Provide sample code
- Release data as drop-in replacement for a popular dataset
- Pose questions that could be answered using your data

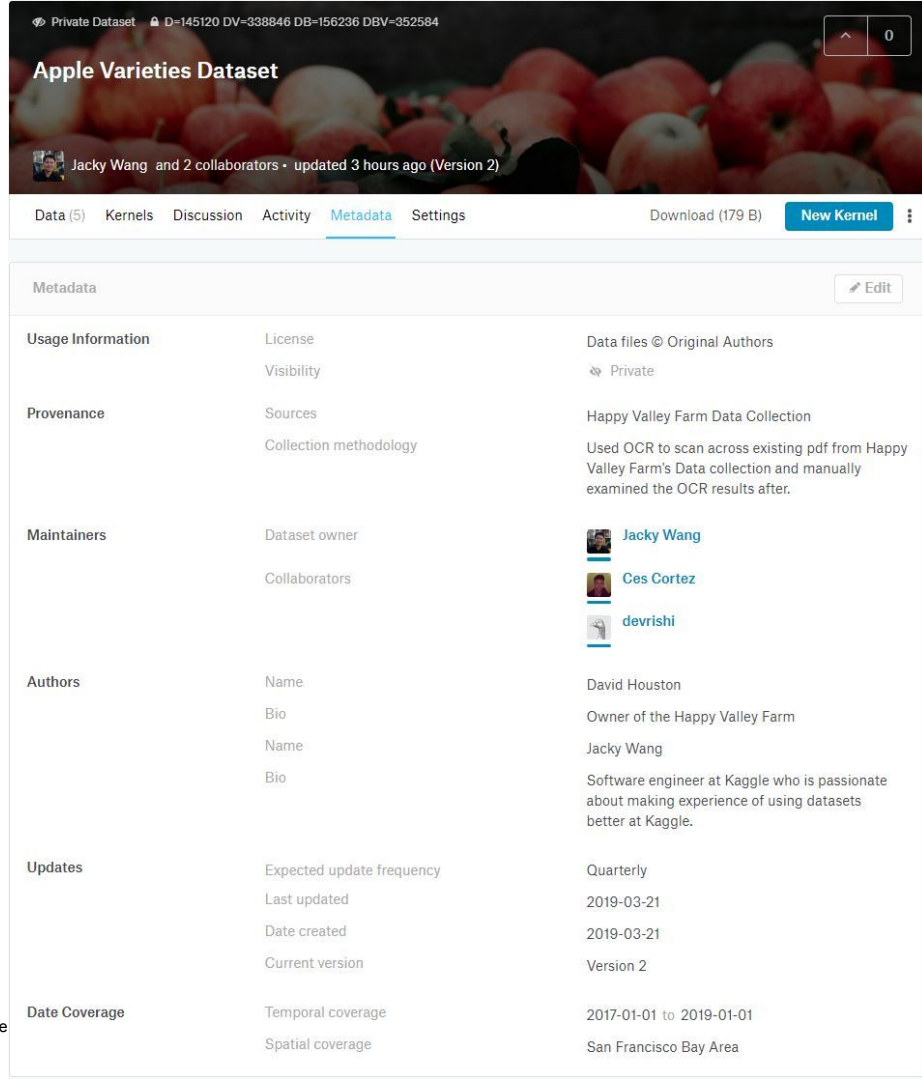


# Before Data Release:




## Prepare clear documentation and metadata

- Consider including:
  - License
  - Provenance
  - Maintainers
  - Authors
  - Update schedule
  - Dataset coverage
- Gebru et al 2018 has a nice discussion of metadata for machine learning datasets

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumeé III, and K. Crawford. Datasheets for machine learning datasets. arXiv preprint arXiv:1803.09010, 2018.



The screenshot shows the Kaggle interface for the 'Apple Varieties Dataset'. At the top, it indicates the dataset is 'Private' and provides identifiers: D=145120, DV=338846, DB=156236, and DBV=352584. The dataset is owned by Jacky Wang and 2 collaborators, and it was updated 3 hours ago (Version 2). The 'Metadata' tab is selected, showing a table with the following information:

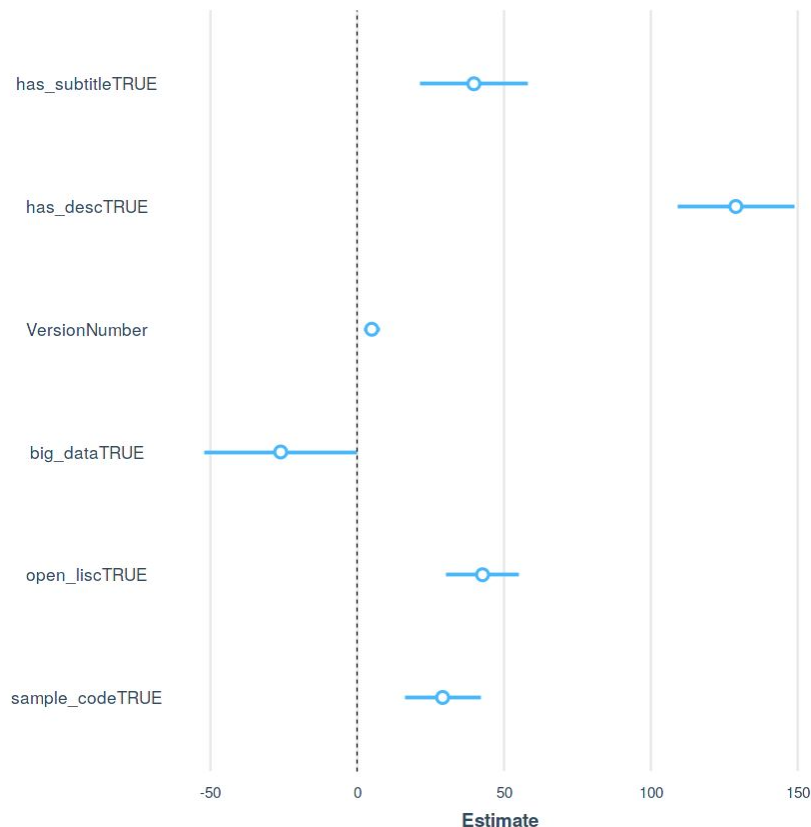
Metadata		
<b>Usage Information</b>	License	Data files © Original Authors
	Visibility	Private
<b>Provenance</b>	Sources	Happy Valley Farm Data Collection
	Collection methodology	Used OCR to scan across existing pdf from Happy Valley Farm's Data collection and manually examined the OCR results after.
<b>Maintainers</b>	Dataset owner	 <a href="#">Jacky Wang</a>
	Collaborators	 <a href="#">Ces Cortez</a>  <a href="#">devrishi</a>
<b>Authors</b>	Name	David Houston
	Bio	Owner of the Happy Valley Farm
	Name	Jacky Wang
	Bio	Software engineer at Kaggle who is passionate about making experience of using datasets better at Kaggle.
<b>Updates</b>	Expected update frequency	Quarterly
	Last updated	2019-03-21
	Date created	2019-03-21
	Current version	Version 2
<b>Date Coverage</b>	Temporal coverage	2017-01-01 to 2019-01-01
	Spatial coverage	San Francisco Bay Area

# Effects of Documentation & Metadata

If you hold everything else steady, then you can increase the # of dataset downloads by:

- Updating your dataset (+5 per update)
- Adding sample code (+29)
- Adding a subtitle (+39)
- Using an open license like CC-BY-NA or CC (+40)
- Adding a description of what's in your dataset (+129)

Large (>1 Gigabyte) datasets aren't downloaded more; **people are happy to use small data!**





# Before Data Release:

Release data as drop-in replacement for a popular dataset

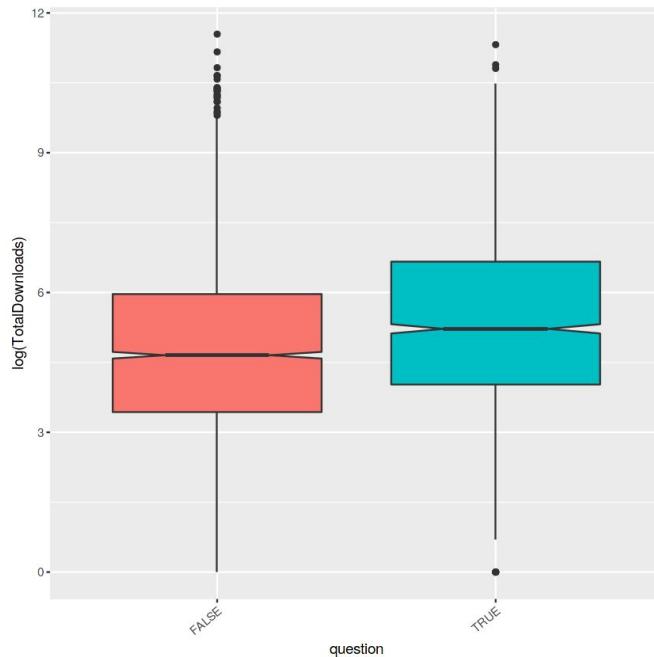
- **Example:** MNIST (LeCun et al 2010) is a digit recognition dataset and one of the most popular machine learning datasets ever released
- Datasets released as drop-in replacements:
  - Fashion-MNIST (Xiao et al 2017)
  - KMNIST (Clanuwat et al 2018, NeurIPS)
- Extremely appealing to learners



# Before Data Release:

Pose questions that could be answered using your data

Kaggle datasets with a “?” in their descriptions are downloaded an extra 450 times (on average).



# After Data Release

- Update your dataset periodically
  - On average a datasets is downloaded an additional 5 times for every update
- Reach out to relevant communities
- Consider hosting your data on multiple platforms





# After Data Release:

Reach out to relevant communities

- Send an announcement to relevant professional organization's mailing lists
- Release a data paper
- Announce your data release on social media
  - Relevant Facebook groups
  - Tweet about it!
  - Don't forget platforms that aren't US-Centric (WeChat, WhatsApp)
- Talk to your university's PR department



# After Data Release:

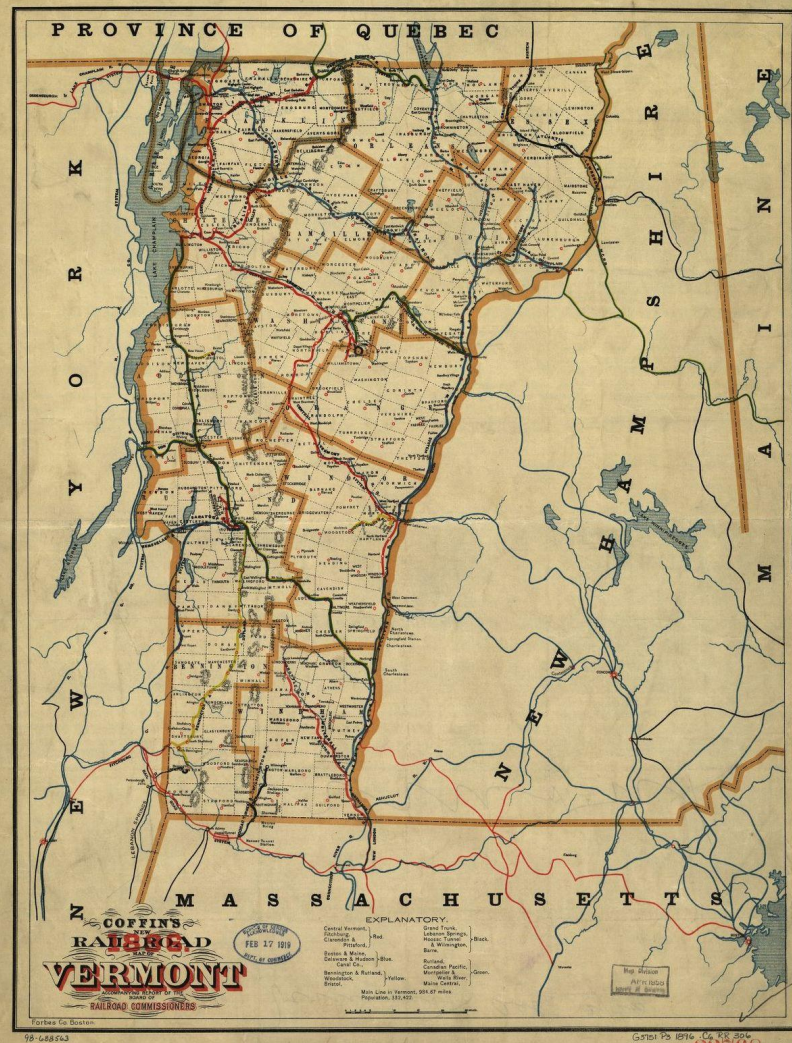
Host your data on multiple platforms

- Redundancy is good and will save your butt (RIP SQLShare)
- Sharing on platforms with existing communities can help surface your data to them
- Make sure at least one dataset hosting uses schema.org standards so your data will be included in Google's [Dataset Search](#)



# Data Sharing Roadmap

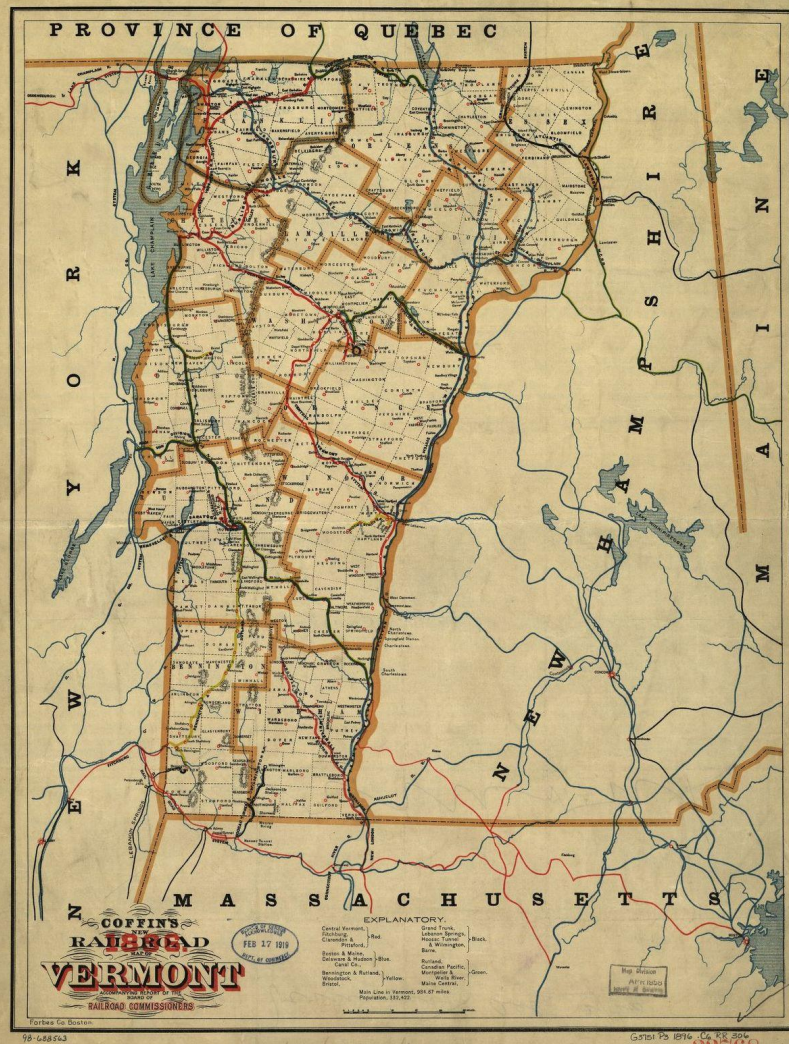
1. During data collection
2. Before data release
3. After data release





# Data Sharing Roadmap

1. During data collection
  - a. Consider your audience
  - b. Collect rich data
2. Before data release
3. After data release



# Data Sharing Roadmap

1. During data collection
2. Before data release
  - a. Documentation & metadata
  - b. Sample code
  - c. Drop-in replacement
  - d. Pose questions
3. After data release





# Data Sharing Roadmap

1. During data collection
2. Before data release
3. After data release
  - a. Update your data
  - b. Reach out to relevant communities
  - c. Redundant hosting





# Thanks! Questions?

@rctatman

rachael@kaggle.com

Slides:

Data: <https://www.kaggle.com/kaggle/meta-kaggle>

Code: <https://www.kaggle.com/rctatman/what-makes-a-dataset-successful>

The image features the Kaggle logo, which consists of the word "kaggle" in a lowercase, sans-serif font. The letters are a light blue color. A small trademark symbol (TM) is located at the top right of the letter "e". The logo is centered horizontally and vertically on a dark gray background. The background is covered with a repeating pattern of light gray isometric cubes, each composed of smaller cubes, creating a 3D effect.

kaggle™