THE WILDERNESS

Jan Fiddler
THE RULES

- there are many different social media platforms
- there are many different terms of service
- there are many different laws and regulations
THE PATH

Samir Luther
WHY YOU SHOULD CARE

• you may be breaking the law
• you may be breaking codes of ethics/conduct
• advancing science
• more citations, better reputation, etc.
COPYRIGHTED DATA

- **situation:** you have permission to use this amazing dataset with which you can do great research, but it’s proprietary...
COPYRIGHTED DATA

• **situation:** you have permission to use this amazing dataset with which you can do great research, but it’s proprietary...

• **solution:** there’s no problem. just use it and do the cool research - that you can’t share the data is unfortunate, but that’s the way it is.
COPYRIGHTED DATA

- **situation:** you don’t have permission to use this amazing dataset with which you can do great research, but you already collected/used the data...
COPYRIGHTED DATA

- **situation**: you don’t have permission to use this amazing dataset with which you can do great research, but you already collected/used the data…

- **solution**: ask for permission, and stop using the data until you have received approval.
EXISTING DATA

• **situation**: you found an existing dataset that is almost, but not exactly, what you need.
EXISTING DATA

- **situation:** you found an existing dataset that is almost, but not exactly, what you need.

- **solution:** check if you can expand or refine the dataset, before considering collecting your own data.
PERMITTED DATA

• situation: you have found this amazing source of suitably licensed and freely sharable data with which you can do great research.
PERMITTED DATA

- **situation**: you have found this amazing source of suitably licensed and freely sharable data with which you can do great research.

- **solution**: fantastic, it looks like you’re on the right track - let’s figure out how to collect and share this data.
CASE STUDIES

- MIRFLICKR dataset
- ImageCLEF photo annotation task
- MediaEval placing task
- YFCC100M dataset
by Silke Gerstenkorn
by Dave Wild
by Hugo A.B. Olivas
by Martin P. Szymczak
by Mani Babbar
by Lee Otis
MIRFLICKR

• the **good**
  
  • relatively large and well-annotated dataset
  
  • freely usable due to Creative Commons licenses
  
  • dataset includes images, tags, features, exif, code, tools

• the **bad**

  • hosting and downloading was challenging
IMAGECLEF: PHOTO ANNOTATION

Figure 1. Images annotated with the concept 'reflection'.

Figure 2. Images retrieved for the query 'traffic light trails'.
Which of the following concepts are clearly present in the picture below? Tick all that apply

- [ ] Cat
- [x] Dog
- [ ] Horse
- [ ] Fish
- [ ] Bird
- [ ] Insect
- [ ] Animal (other)
- [ ] None of the above

[click to view image in larger size]
IMAGECLEF: PHOTO ANNOTATION

• the good
  • diverse and challenging concepts compared to other tasks
  • revealed trends in how participants approached the task

• the bad
  • concepts and evaluation metrics evolved over time, making year-over-year comparisons difficult
  • code of participants not shared
  • data not shared with non-participants
  • annotation funding
MEDIAEVAL: PLACING TASK

Sean Davis
MEDIAEVAL: PLACING TASK

George Megas
MEDIAEVAL: PLACING TASK
MEDIAEVAL: PLACING TASK

- the **good**
  - accuracy increased and then started plateauing
  - baseline methods provided some bar of entry
- the **bad**
  - training set grew over time, so even with the same test set year-over-year comparisons were difficult
  - participants didn’t learn as much from each other as we hoped
YFCC100M
YFCC100M

Original Metadata
- title
- tags
- description
- geo-tag
- uploader info
- capture device
- date
- URL to the original item

Expansion Packs
- autotags: presence of visual concepts
- Exif
- place labels

Multimedia Commons

AWS S3 repository
- images
- videos

Pre-computed features
- deep features: CNN codes, VLAD, ...
- conventional features: SIFT, FCTH, ...

Annotated subsets
- YLI-MED: multimedia event detection
- YLI-GEO: geo-location

Tools
- tutorials
- workshops
YFCC100M

• the **good**
  • large and richly annotated dataset
  • overlaps with other well-known datasets
  • images, videos, metadata, features all in the cloud

• the **bad**
  • photos and videos disappeared before a copy could be made of them
  • hosted across two platforms, and gaining access is not easy
  • stored in an organized yet impractical way
  • random selection biased towards prolific photographers
REFLECTING
REFLECTING

- annotating, storing, hosting, serving data is not necessarily cheap
- handling large amounts of non-text data is a pain
- the format in which to store data is not obvious
- the easier you make it for the data to be used, the more it will be used and the fewer questions you get
REFLECTING

- user data requires legal and privacy considerations
- registration walls and additional license agreements make the data less free and less accessible
- no control over repository = no control over its future
PLAN AHEAD

- consider what data needs to be collected
- obtain written permission from all stakeholders
- adhere to licensing, privacy & deletion requirements
PLAN AHEAD

• consider who will use your data and how
• consider how to process, format & annotate the data
• secure enough space to store all data
• be aware of platform limitations
DOs

• make it easy for people to see and use the data
• consider offering an API as single point of access
• consider offering code, features, etc.
• consider offering separate download options
• consider allowing anyone to contribute
DO

- link-only social media platforms
  - collect more data than you really need
  - devise an approach that can deal with missing data
- full-content social media platforms
  - also collect more data than you really need
  - keep some as backup in case of mistakes or data corruption
- random sampling, but not too random
DO

• consider applying for research/academic grants/programs
• use a spot instance instead of on-demand compute
FINAL WORDS

• collecting and sharing data is hard
• a well-planned approach is key to success
• make sure the data can live on even if you move on