rainforest

www.rainforestqa.com

MACIEJ GRYKA

GIVING YOUR MACHINE LEARNING MODEL AN API

MACHINE LEARNING

MACHINE LEARNING

- black box (the model)
- feature vector goes in
- prediction comes out
- classification / regression, supervised / unsupervised

TRAINING A MODEL

- get some data
- clean, extract features?
- train the model
- evaluate (explain train/test split)
- iterate?
- profit

MACHINE LEARNING: PITFALLS OF UPDATING MODELS

- When you iterate and change the feature vector, all your old data has to be re-processed.
- If you add new features that you don't have for old data, you better be ready to start data collection from scratch.

SWEET, I HAVE GOOD PREDICTION ACCURACY. MY JOB HERE IS DONE.

me, a while ago

K, LET'S USE THE MODEL IN PROD.

people at work

...

me

WHAT WE NEED

- An HTTP endpoint that takes a bunch of data (a feature vector) and returns a prediction (a vector of numbers).
- Should be fast-ish (for us ~1s is fine, 10s probably not) and scalable
- Retraining and updating should be easy.
- Obvious which model is currently being used.
- Able to prevent common errors and figure out what's up when stuff goes wrong.

LET'S BUILD AN API

- use flask (obviously)
- load a model (e.g. from S3) into memory
- start accepting requests:
  - validate input
  - process data, if needed
  - get prediction from the model
  - return the response

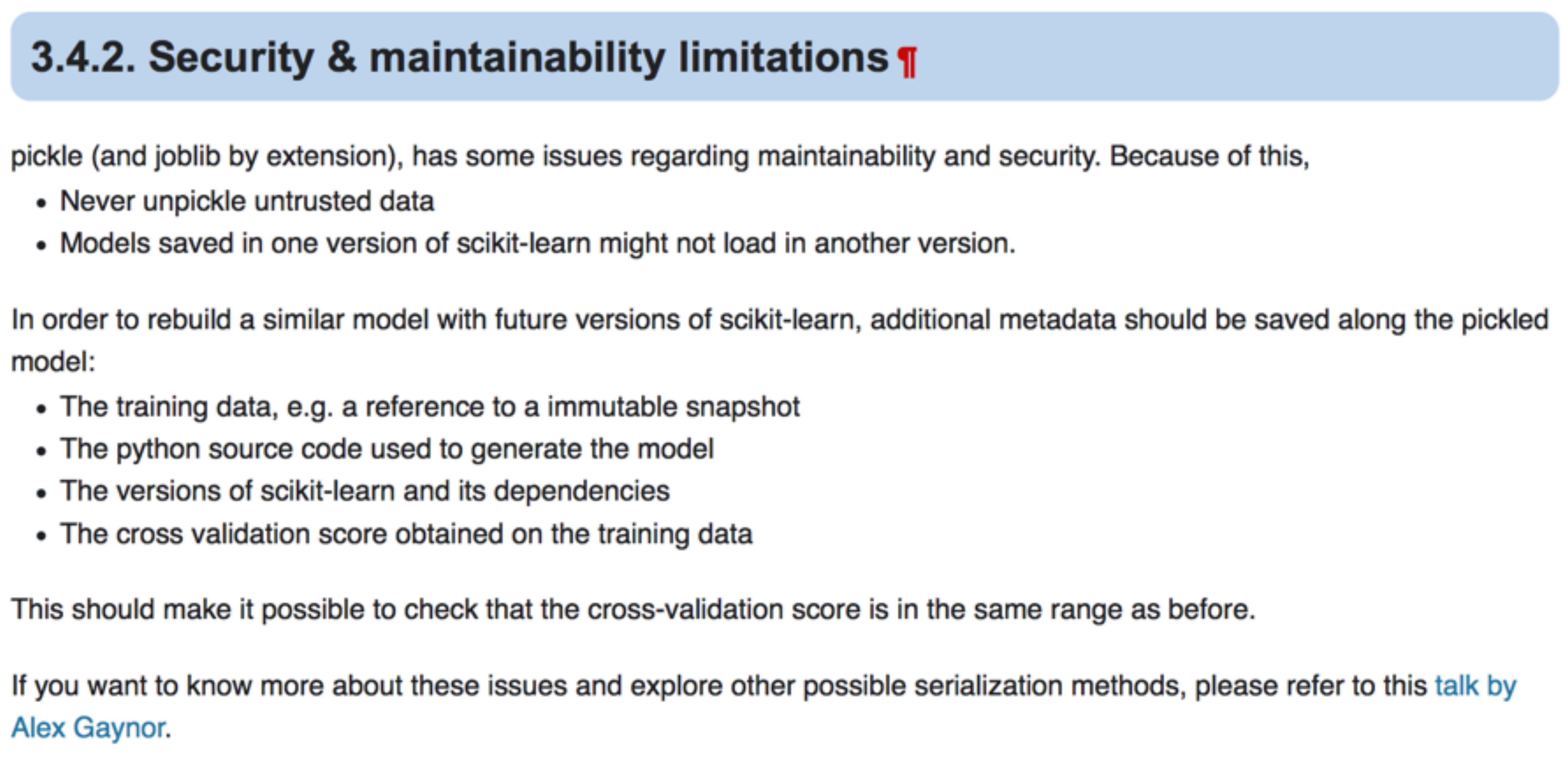
LET'S BUILD AN API

- use flask (obviously)
- load a model (e.g. from S3) into memory
- start accepting requests:
  - validate input
  - process data, if needed
  - get prediction from the model
  - return the response

SERIALIZING ML MODELS

- `pickle` all the things.
- OK, I guess `joblib` is better here.
- Except Pickles are for Delis (PyCon 2014 talk by Alex Gaynor).
  - you're at the mercy of someone else's code for non-transparent things
  - you implicitly rely on package versions

SERIALIZING ML MODELS



3.4.2. Security & maintainability limitations ¶

pickle (and joblib by extension), has some issues regarding maintainability and security. Because of this,

- Never unpickle untrusted data
- Models saved in one version of scikit-learn might not load in another version.

In order to rebuild a similar model with future versions of scikit-learn, additional metadata should be saved along the pickled model:

- The training data, e.g. a reference to a immutable snapshot
- The python source code used to generate the model
- The versions of scikit-learn and its dependencies
- The cross validation score obtained on the training data

This should make it possible to check that the cross-validation score is in the same range as before.

If you want to know more about these issues and explore other possible serialization methods, please refer to this talk by Alex Gaynor.

http://scikit-learn.org/stable/modules/model_persistence.html#security-maintainability-limitations

SERIALIZING ML MODELS

- What’s important to keep track of:
  - training + test data,
  - meaning of your feature vector,
  - how it was computed,
  - performance scores,
  - be able to recreate the object, not only reinstantiate it.

SERIALIZING ML MODELS

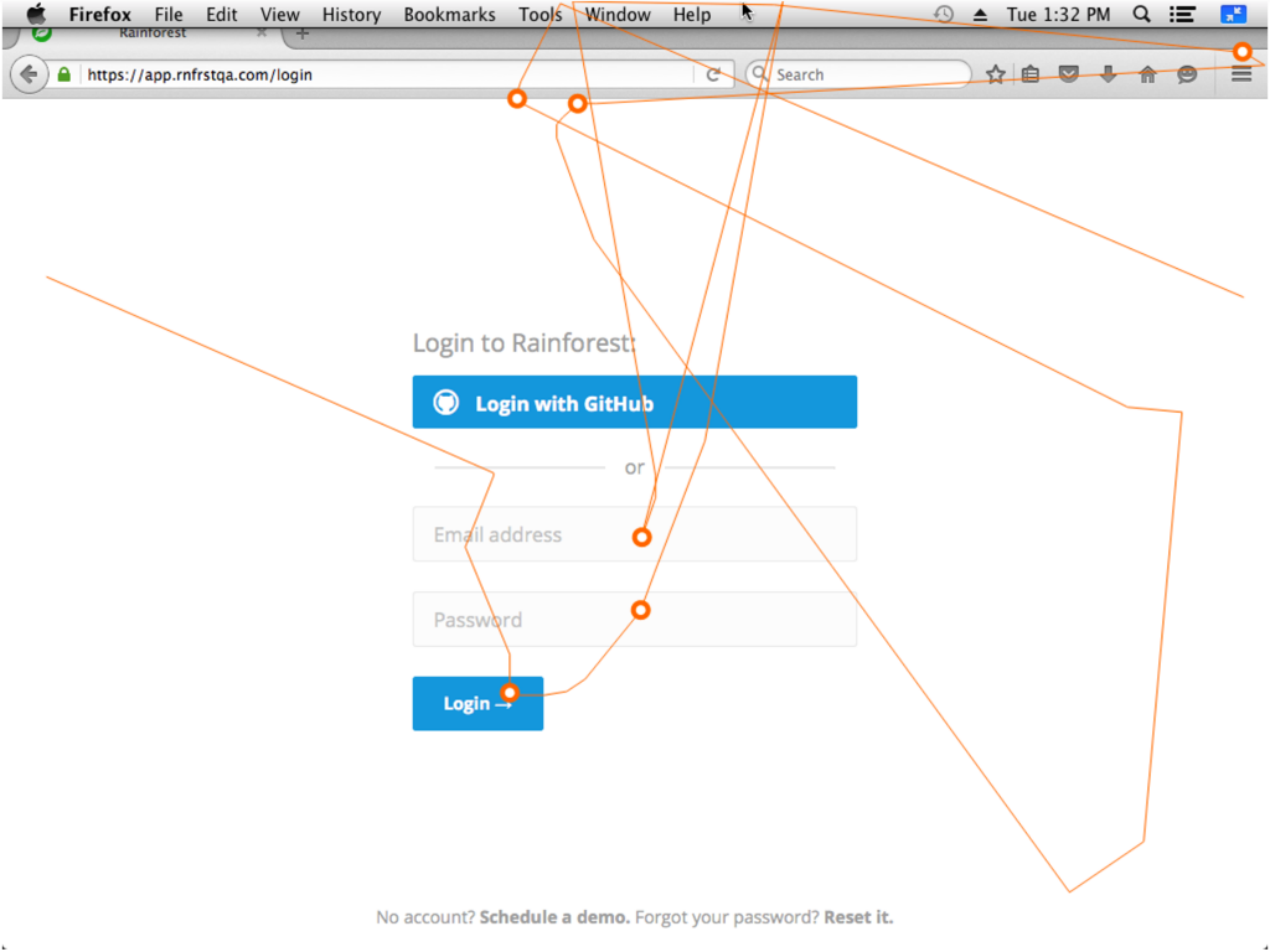
- Our solution: `destimator`.
- Saves the model together with a bunch of metadata. Always shipped around together.
- Still at the mercy of `joblib` doing the bulk of work - but now:
  - we know exactly which model we're dealing with,
  - we can re-train it if we want to (we have the data).

DESTIMATOR

- We keep track of:
  - data (training + test)
  - feature names
  - git hash
  - distribution info (Python + packages versions)
  - performance numbers
  - creation timestamp

DESTIMATOR: WHAT WE USE IT FOR AT rainforest

- Crowdsourced web app testing.
- If you use `selenium` and don’t love writing and rewriting your tests on every change, have a look!
- We deal with humans and need to verify their work.
- Most of them are great (hello testers!), but a small fraction are trying to cheat.
- We use machine learning (and elbow grease) for fraud detection.



Login to Rainforest:

Email address

Password

No account? **Schedule a demo.** Forgot your password? **Reset it.**

DESTIMATOR: LIMITATIONS

- Rely on `joblib` for heavy lifting. Not bad per-se, but you have to understand the trade-offs. E.g. it does nothing to improve security, so only load trusted models.
- Increases the size of the model you have to ship around (not a problem for us, but keep in mind). Probably better to store reference to the data, instead of the data itself?
- MVP - only serves our narrow use case for now. E.g. we'd need to change how we store performance numbers for regression.

DANKE!

also, we're hiring :D

maciej@rainforestqa.com

@maciejgryka