User Data Collection in Open Source: Sempervierens

github.com/njsmith/sempervirens

Matthias Bussonnier – UC Berkeley, Berkeley Institute for data science

Problem in open source

- How many people are using my library ?
- With which frequency ?
- Which version ?
- Which API/submodules are more used ?
- On which OS ?

current solution.

- Run survey on mailing list, Twitter, download count, pop con + extrapolate …

```
import requests
requests.post('https://myserver.com/analytics',
                    data:{version:mylib.__version__})
```

- Hard to “ask” user (eg numpy)
- Reinvent the wheel for each project.
- Legal issues / IRB.
- More maintenance, need domain/server… etc.
- More bugs
- Ethical questions

An idea: let's collaborate

Sempervierens (evergreen)

(github: njsmith/sempervirens)

- Inspire by Mozilla Telemetry.
- 1 Library to Bridge Client <-> Backend

IDE Side APIs

- should_get_user_consent() -> Bool
- set_user_content(x:Bool)
- get_license_text() -> Unicode
- Disable system-wide by IT
- Collect data on custom URL

Library Side APIs

- Opt-in for library authors
- First iteration:
  - increment(project, os.name/version)
  - increment(project, use_feature/call_func)
  - No "Performance" analytics
  - No-OP of no-consent.

Aggregate, upload

- Every now and then post results.
- Neutral third party
  - Publish only aggregated statistics
  - Possibly whitelisted
- Get quantitative results
  - Funding Agencies.
  - Deprecation cycles.
  - Function usage documentation
  - ...

Thanks

- github.com/njsmith/sempervirens