Open source software and Science Obviously!

Martijn Schuemie
What happened?

• 2007 Martijn was a scientist, and did not believe in open source: “Scientists should do science, not maintain software”

• 2022 Martijn is considered by some a leader in open source software. Spends about 25% of his time maintaining open-source software
Unique about OHDSI

• Apache: “Community over code”

• OHDSI: “SCIENCE!”
What is science?

“The real purpose of the scientific method is to make sure Nature hasn't misled you into thinking you know something you don't actually know.”

What assumptions am I making?

Can I test those assumptions? Or has someone else already tested them?
The start of OHDSI (OMOP)

“Can we be misled if we apply current best practice research methods to observational data?”

“YES!!!” “But through science we can do better.”
Open-source software at the core of OHDSI

Methods research
Improving observational research methods through (empirical) science

ATLAS
Implementing best practices for observational research

Clinical research
Improving health care by generating evidence

Open Source allows for transparency, reproducibility, and therefore critical scientific evaluation
Our software has been extremely successful!

• 21 methods research papers used HADES
• >>25 clinical research papers used HADES
• Impacted decisions by EMA (and probably FDA)
The best is ahead!

- EMA has identified ErasmusMC as the coordinating center for the Data Analysis and Real World Interrogation Network (DARWIN).
- DARWIN will run on OHDSI’s principles and software, generating evidence for the EU.

(But... we must continue to ensure our tools meet DARWIN’s need)
Evolution of the HADES community

• 2014: 1st version of ACHILLES R package
  – Spawned SqlRender and DatabaseConnector

• 2014-2018: Multiple analytics R packages
  – Cyclops, CohortMethod, PatientLevelPrediction...

• 2018: Establishing Methods Library
  – Packages verified through empirical evaluation
  – Focus on stability and validity

• 2020: Rebranding as HADES

• 2020-: Growing the community
HADES community growth in action

- Welcomed developers in addition to scientists
- Defined standards and interfaces
  - Common Data Model
  - Standard release process and versioning
  - Code style guidelines
- Open meetings and open discussions
  - 3rd Thursday of the month
HADES community growth in action

• Welcomed developers in addition to scientists
• Defined standards and interfaces
  – Common Data Model
  – Standard release process and versioning
  – Code style guidelines
• Open meetings and open discussions
  – 3rd Thursday of the month
• 2021 Unit-test-a-thon
HADES community growth in action

- Welcomed developers in addition to scientists
- Defined standards and interfaces
  - Common Data Model
  - Standard release process and versioning
  - Code style guidelines
- Open meetings and open discussions
  - 3rd Thursday of the month
- 2021 Unit-test-a-thon
- 12 HADES maintainers
  Adam Black, Anthony Sena, Chris Knoll, Jamie Gilbert, Jenna Reps, Frank DeFalco, Gowtham Rao, Lee Evans, Marc Suchard, Martin Lavallee, Peter Rijnbeek

<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
<th>Maintainer(s)</th>
<th>Availability</th>
<th>Open issues</th>
<th>Open pull-requests</th>
<th>Build status</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andromeda</td>
<td>v0.6.6</td>
<td>Adam Black</td>
<td>CRAN</td>
<td>6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BigKan</td>
<td>v1.6.1</td>
<td>Martijn Schuemie</td>
<td>GitHub</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capr</td>
<td>v1.6.3</td>
<td>Martin Lavallee</td>
<td>GitHub</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CircR</td>
<td>v1.1.1</td>
<td>Chris Knoll</td>
<td>GitHub</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CohortDiagnostics</td>
<td>v3.3.4</td>
<td>Gowtham Rao</td>
<td>GitHub</td>
<td>7</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CohortGenerator</td>
<td>v0.4.8</td>
<td>Anthony Sena</td>
<td>GitHub</td>
<td>7</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CohortMethod</td>
<td>v4.2.2</td>
<td>Martijn Schuemie</td>
<td>GitHub</td>
<td>15</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyclops</td>
<td>v3.1.2</td>
<td>Marc Suchard</td>
<td>CRAN</td>
<td>10</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DatabaseConnector</td>
<td>v5.6.0</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
<td>11</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EmpiricalCalibration</td>
<td>v3.6.6</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EnsemblePatientLevelPrediction</td>
<td>v0.6.2</td>
<td>Jenna Reps</td>
<td>GitHub</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eunomia</td>
<td>v1.9.2</td>
<td>Frank DeFalco</td>
<td>GitHub</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EvidenceSynthesis</td>
<td>v0.2.2</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FeatureExtraction</td>
<td>v2.1.6</td>
<td>Anthony Sena</td>
<td>GitHub</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydra</td>
<td>v0.3.0</td>
<td>Martijn Schuemie</td>
<td>GitHub</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MethodEvaluation</td>
<td>v2.2.0</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OhsdSharing</td>
<td>v2.2.2</td>
<td>Lee Evans</td>
<td>GitHub</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ParallelLogger</td>
<td>v3.0.0</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PatientLevelPrediction</td>
<td>v5.6.5</td>
<td>Jenna Reps &amp; Peter Rijnbeek</td>
<td>GitHub</td>
<td>21</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R0HdsdWebApi</td>
<td>v1.2.1</td>
<td>Gowtham Rao</td>
<td>GitHub</td>
<td>17</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SelfControlledCaseSeries</td>
<td>v3.2.1</td>
<td>Martijn Schuemie</td>
<td>GitHub</td>
<td>10</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SelfControlledCohort</td>
<td>v1.6.6</td>
<td>Jamie Gilbert</td>
<td>GitHub</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SqlRender</td>
<td>v1.9.8</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
<td>14</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Special callout: Maxim and the Rabbits

- Maxim Moinat became the maintainer of WhiteRabbit, RabbitInAHat, and Usagi in 2020
- 2021 Titan Award winner
Where do we go from here?

- Grow the community
  - More users, contributors, developers, maintainers, leaders

- While keeping science at the core of our software
Growing the community
How do we grow the community?

Provide on-ramps

• Make it easier for users to get started with our tools
  – Easier installation
  – Allow users to run analyses without writing R code

• Provide development opportunities
  – Allow developers to run studies
  – Allow scientists to learn how to code

• Hack-a-thons

• Traineeships
  – Khieron

• …?
How do we grow the community?

Make the economics work

(2007 Martijn’s main problem was lack of financial backing)

• Allow more institutions to lead / provide maintainers
• Encourage commercial partners to contribute
• Give credit where credit is due
  – Advertising contributing organizations
  – Acknowledging contributions by individuals
• ...?
How do we grow the community?

Provide community structure
- Code of conduct
- Bylaws
- Clearly defined interfaces between software components
  - Allows decoupling
Keep science at the core

“If we knew what it is we were doing, it would not be called research. Would it?” - Albert Einstein

Software requirements will continue to evolve (sometimes rapidly)

Continue to demand science behind our software
Keeping science at the core

Core OHDSI software principle: *make sure Nature hasn’t misled you*

Methods research

• Proven to be reliable under certain circumstances

Data ➔ Analytics implementation ➔ Evidence

Unit test results ➔ Diagnostics

>= 80% coverage for most HADES packages!

For example
• Negative control distribution
• Cohort diagnostics
• Covariate balance
Keeping science at the core

Data → Analytics implementation → Methods research → Evidence → Best practices - Decision support?

Unit test results → Diagnostics → User
What can science learn from open source?

• Scientists still tend to work in silos or cliques
• A study is often kept secret until the paper is published
  – OHDSI is at the forefront of study transparency
• In science, what would be the equivalent of
  – An issue?
  – A pull request?
  – A fork?
• Could we re-invent the scientific discussion?
So what happened?

• The economics changed
  – My new employer allows for spending time on software maintenance
  – Publishing papers and writing grants no longer the end goal

• Open science happened
  – Sharing code and functional software now recognized as necessary for science to progress

• Observational research must do better
  – Adoption of new methods requires they’re readily available.
Concluding thoughts

- Open-source software is at the core of OHDSI’s science
  - For obvious reasons (in hindsight)
- We have already been incredibly successful.
  - Now we have to deliver on our promise (think DARWIN)
- Our open-source community has grown, but now needs to mature
- I’m excited about this future, but it will require a community
  - Shoutout to the Open-Source Community Workgroup!
- ...
Concluding thought

SCIENCE!