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

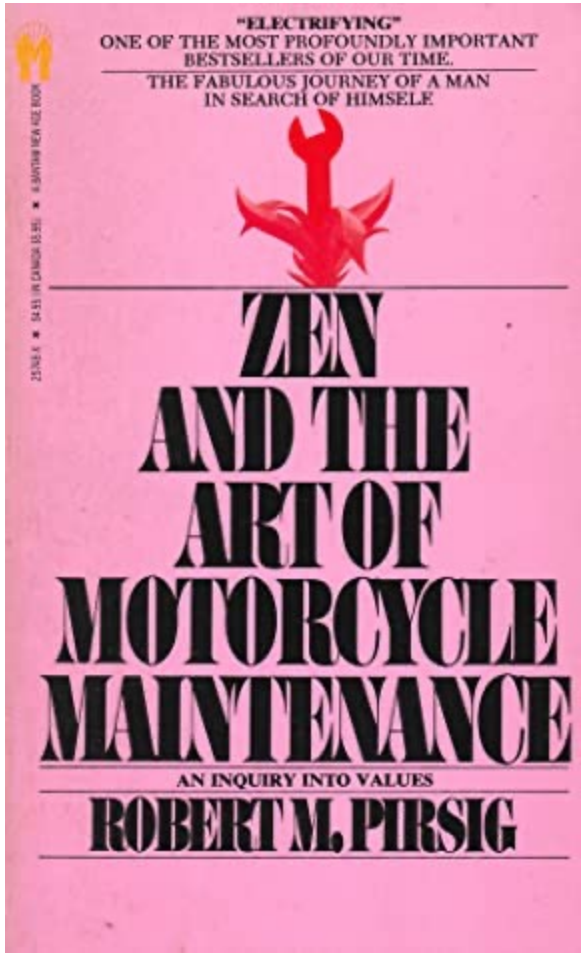
- Apache: “Community over code”
- OHDSI: “SCIENCE!”

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?

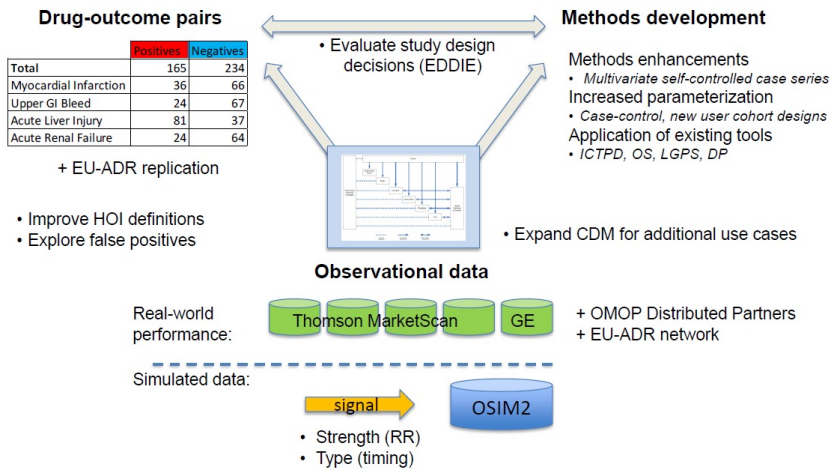


The start of OHDSI (OMOP)

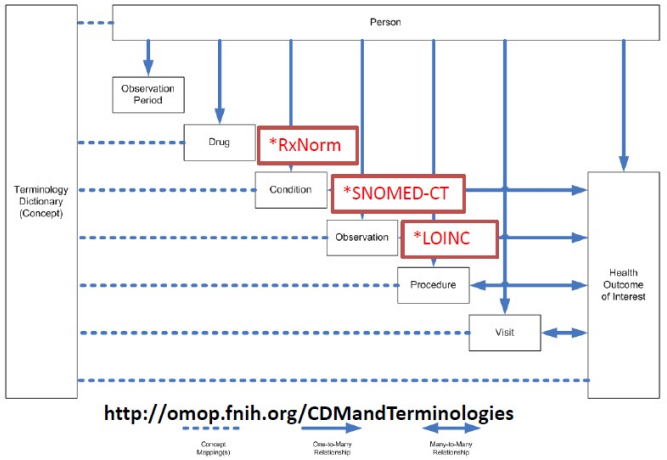
"Can we be misled if we apply current best practice research methods to observational data?"

OBSERVATIONAL MEDICAL OUTCOMES PARTNERSHIP

OMOP experiment



OMOP Common Data Model



"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

HADES

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!

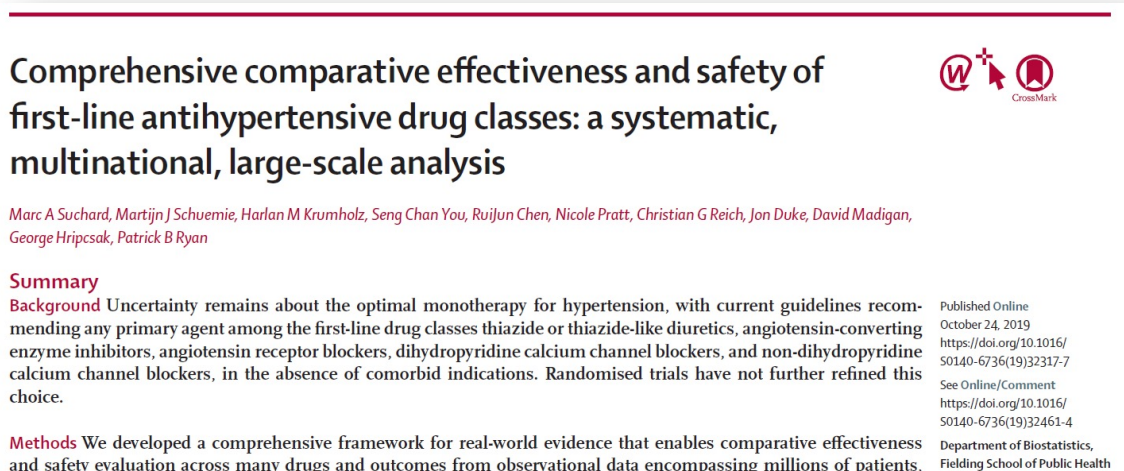
- 21 methods research papers used HADES
- >>25 clinical research papers used HADES
- Impacted decisions by EMA (and probably FDA)

Harvard Data Science Review • 2.1

How Confident Are We About Observational Findings in Health Care: A Benchmark Study

Martijn J. Schuemie, M. Soledad Cepede, Marc A. Suchard, Jianxiao Yang, Yuxi Tian Alejandro Schuler, Patrick B. Ryan, David Madigan[1], George Hripcsak

[1]Professor of Statistics, Columbia University



Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis

Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, RuiJun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, Patrick B Ryan

Summary

Background Uncertainty remains about the optimal monotherapy for hypertension, with current guidelines recommending any primary agent among the first-line drug classes thiazide or thiazide-like diuretics, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, dihydropyridine calcium channel blockers, and non-dihydropyridine calcium channel blockers, in the absence of comorbid indications. Randomised trials have not further refined this choice.

Methods We developed a comprehensive framework for real-world evidence that enables comparative effectiveness and safety evaluation across many drugs and outcomes from observational data encompassing millions of patients.

Published Online October 24, 2019 https://doi.org/10.1016/S0140-6736(19)32317-7

See Online/Comment https://doi.org/10.1016/S0140-6736(19)32461-4

Department of Biostatistics, Fielding School of Public Health

The best is ahead!

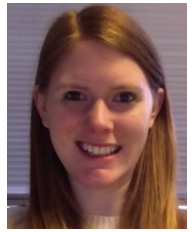
- EMA has identified ErasmusMC as the coordinating center for the Data Analysis and Real World Interrogation Network (DARWIN).
- DARWIN will run on OHDSI's principles and software, generating evidence for the EU.

(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

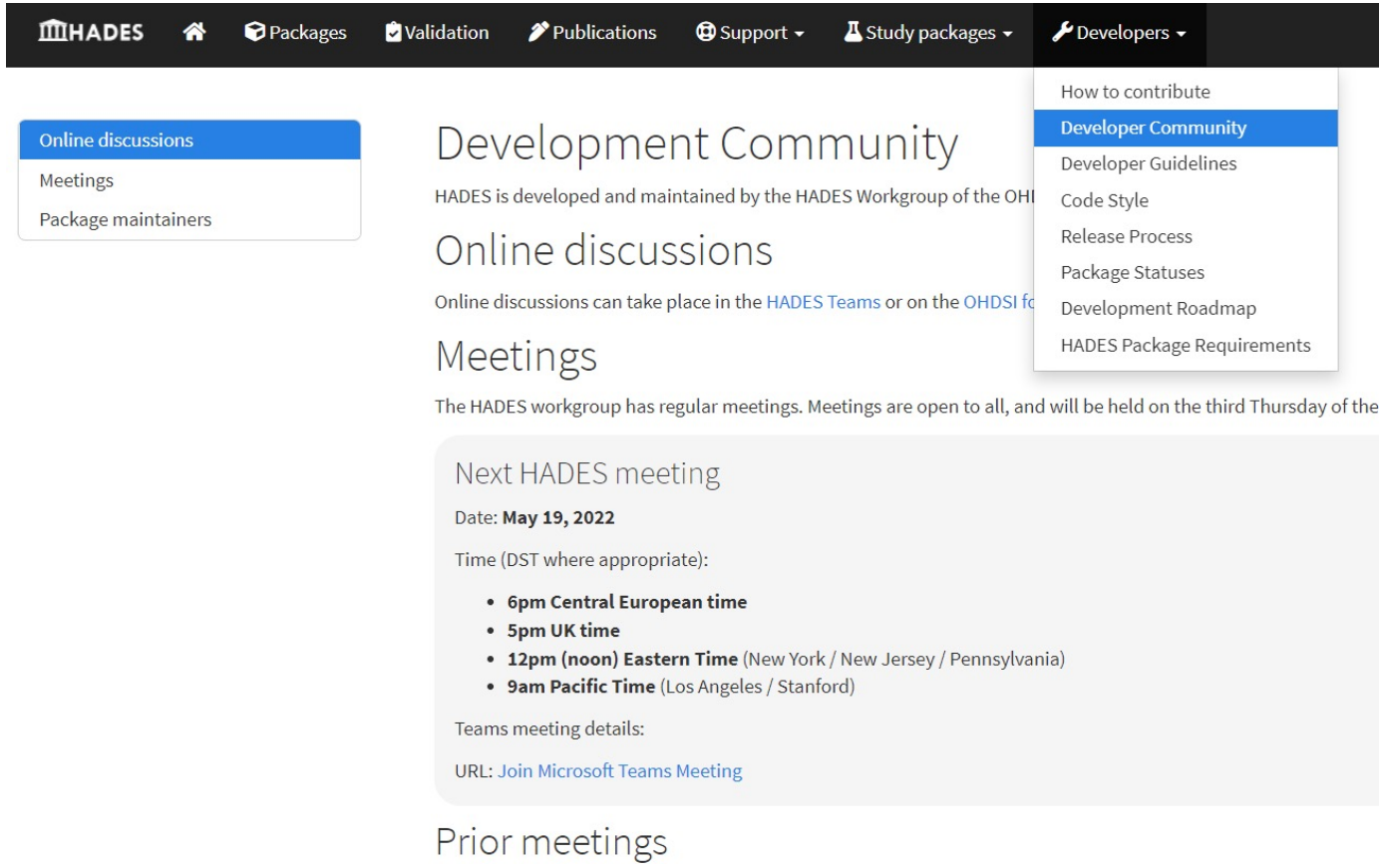




- 2014-2018: Multiple analytics R packages
  - Cyclops, CohortMethod, PatientLevelPrediction…
- 2018: Establishing Methods Library
  - Packages verified through empirical evaluation
  - Focus on stability and validity
- 2020: Rebranding as HADES
- 2020-: Growing the community

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

Package	Version	Maintainer(s)	Availability	Open issues	Open pull-requests	Build status	Coverage
Andromeda	v0.6.0	Adam Black	CRAN	6	0	R-CMD-check passing	codecov 89%
BigKnn	v1.0.1	Martijn Schuemie	GitHub	1	0	R-CMD-check passing	codecov 96%
Capr	v1.0.3	Martin Lavallee	GitHub	3	0	R-CMD-check passing	codecov 59%
CirceR	v1.1.1	Chris Knoll	GitHub	1	1	R-CMD-check failing	codecov 87%
CohortDiagnostics	v2.2.4	Gowtham Rao	GitHub	25	0	R-CMD-check failing	codecov 83%
CohortGenerator	v0.4.0	Anthony Sena	GitHub	7	0	R-CMD-check passing	codecov 99%
CohortMethod	v4.2.2	Martijn Schuemie	GitHub	15	0	R-CMD-check passing	codecov 86%
Cyclops	v3.1.2	Marc Suchard	CRAN	10	0	R-CMD-check passing	codecov 67%
DatabaseConnector	v5.0.2	Martijn Schuemie	CRAN	11	2	R-CMD-check passing	codecov 49%
EmpiricalCalibration	v3.0.0	Martijn Schuemie	CRAN	1	0	R-CMD-check passing	codecov 89%
EnsemblePatientLevelPrediction	v0.0.2	Jenna Reps	GitHub	3	0	R-CMD-check passing	codecov 91%
Eunomia	v1.0.2	Frank DeFalco	GitHub	0	0	R-CMD-check failing	codecov unknown
EvidenceSynthesis	v0.2.3	Martijn Schuemie	CRAN	1	0	R-CMD-check passing	codecov 41%
FeatureExtraction	v3.2.0	Anthony Sena	GitHub	37	3	R-CMD-check failing	codecov unknown
Hydra	v0.3.0	Martijn Schuemie	GitHub	2	1	R-CMD-check passing	codecov 87%
MethodEvaluation	v2.2.0	Martijn Schuemie	GitHub	3	0	R-CMD-check passing	codecov 59%
OhdsiSharing	v0.2.2	Lee Evans	GitHub	1	1	R-CMD-check failing	codecov 0%
ParallelLogger	v3.0.0	Martijn Schuemie	CRAN	3	0	R-CMD-check passing	codecov 82%
PatientLevelPrediction	v5.0.5	Jenna Reps & Peter Rijnbeek	GitHub	23	2	R-CMD-check passing	codecov 89%
ROhdsiWebApi	v1.3.1	Gowtham Rao	GitHub	17	3	R-CMD-check passing	codecov 62%
SelfControlledCaseSeries	v3.2.1	Martijn Schuemie	GitHub	10	0	R-CMD-check passing	codecov 82%
SelfControlledCohort	v1.6.0	Jamie Gilbert	GitHub	1	0	R-CMD-check passing	codecov 96%
SqlRender	v1.9.0	Martijn Schuemie	CRAN	14	1	R-CMD-check passing	codecov 78%

Special callout: Maxim and the Rabbits

- Maxim Moinat became the maintainer of WhiteRabbit, RabbitInAHat, and Usagi in 2020
- 2021 Titan Award winner

WHITE RABBIT

RABBIT IN A HAT

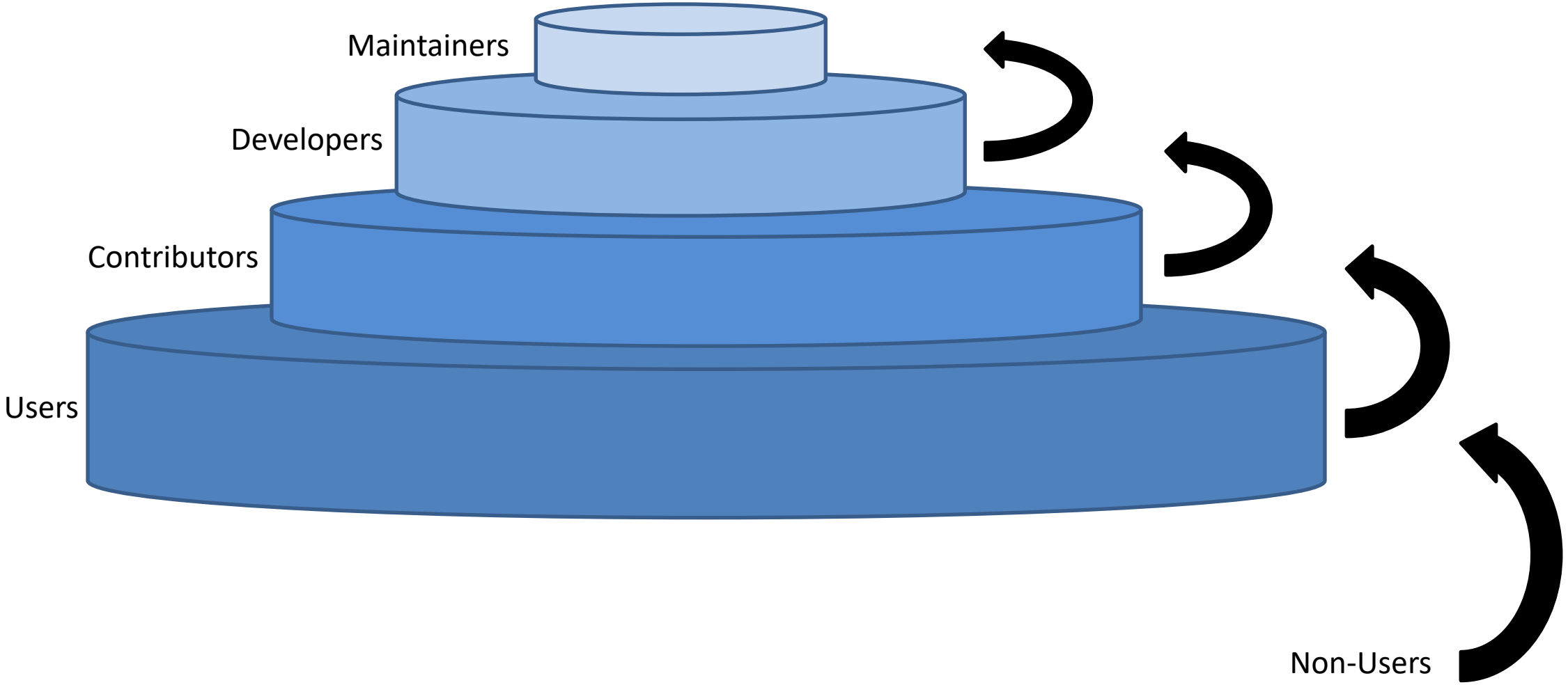
USAGI

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)

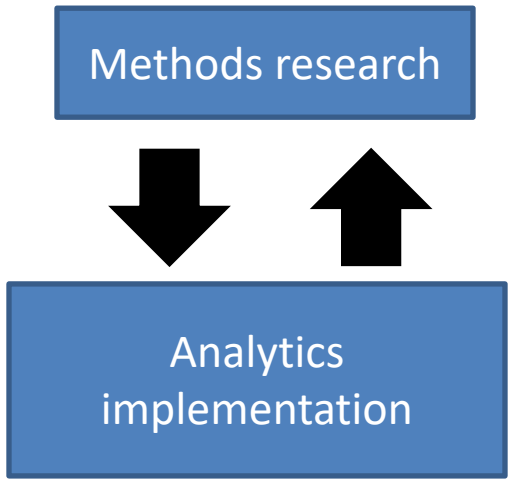
- Allow more institutions to lead / provide maintainers
- Encourage commercial partners to contribute
- Give credit where credit is due
  - Advertising contributing organizations
  - Acknowledging contributions by individuals
- …?

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



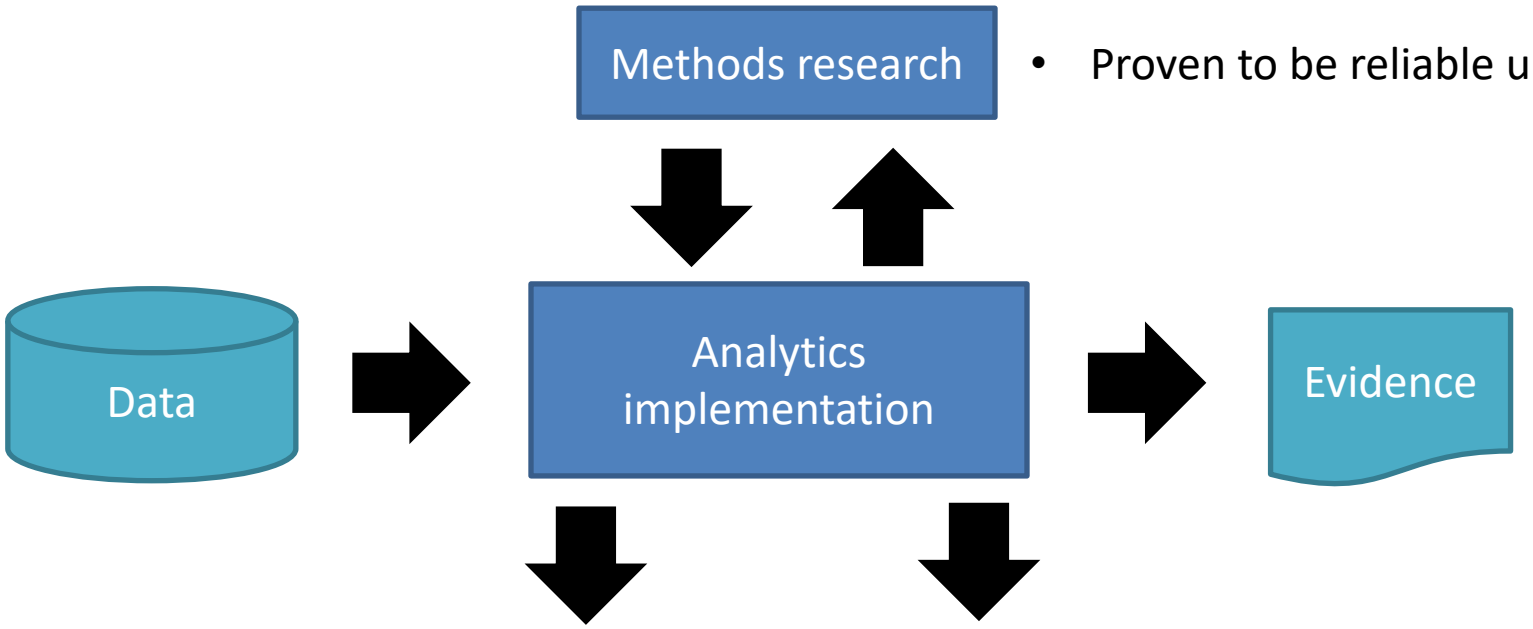
Software requirements will continue to evolve (sometimes rapidly)

Continue to demand science behind our software

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

Keeping science at the core

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



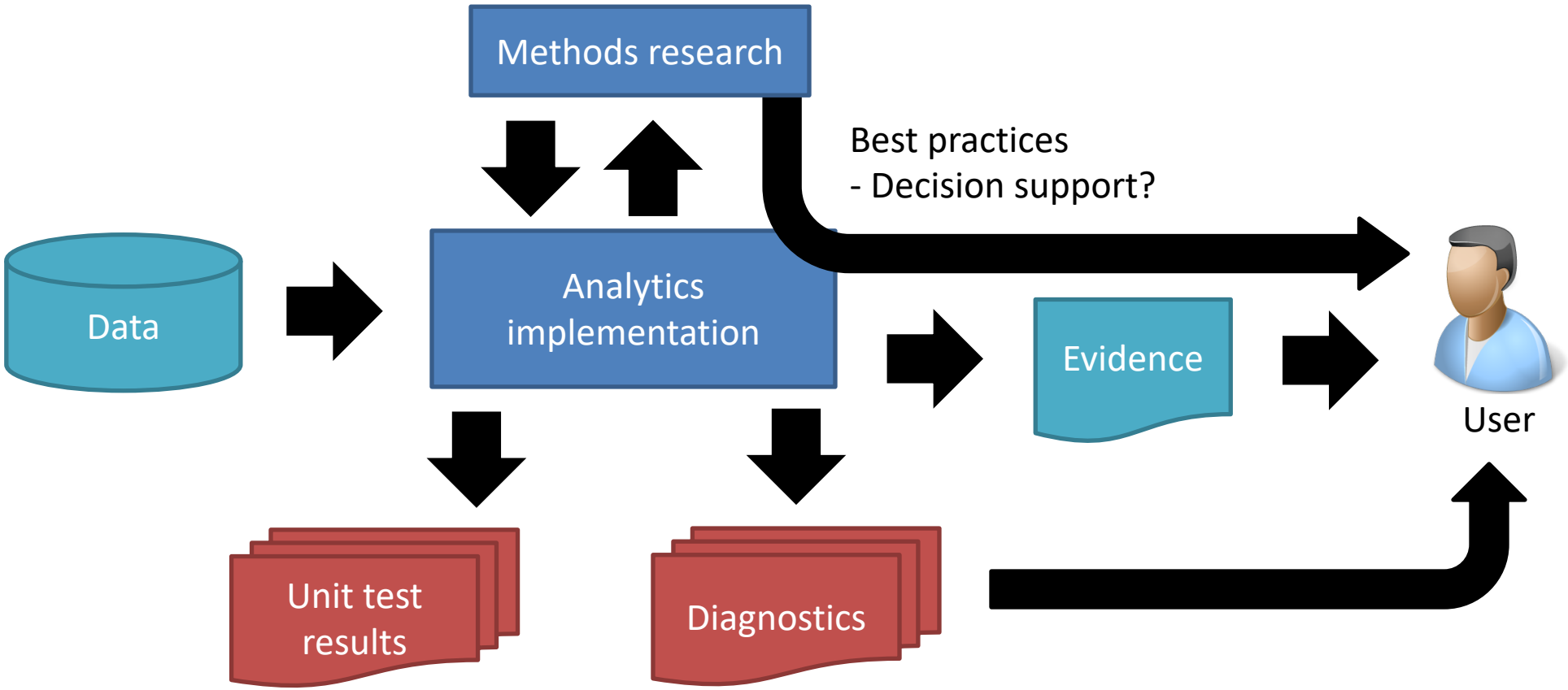
- Proven to be reliable under certain circumstances

>= 80% coverage for most HADES packages!

For example

- Negative control distribution
- Cohort diagnostics
- Covariate balance

Keeping science at the core

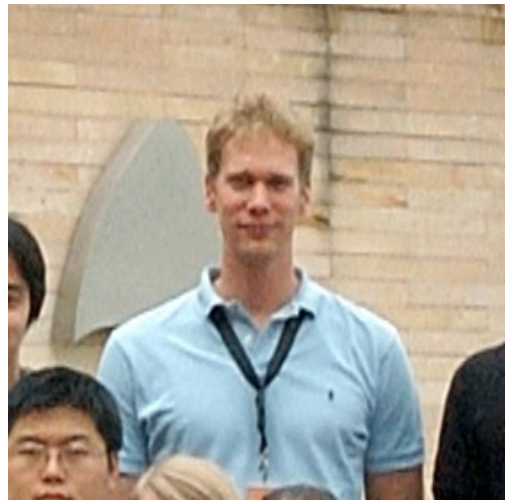


What can science learn from open source?

- 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!