Making the Case for the ESGF and Apache: Long-Term Software Stewardship

Chris Mattmann

Chief Architect, Instrument and Science Data Systems,
Jet Propulsion Laboratory, California Institute of Technology

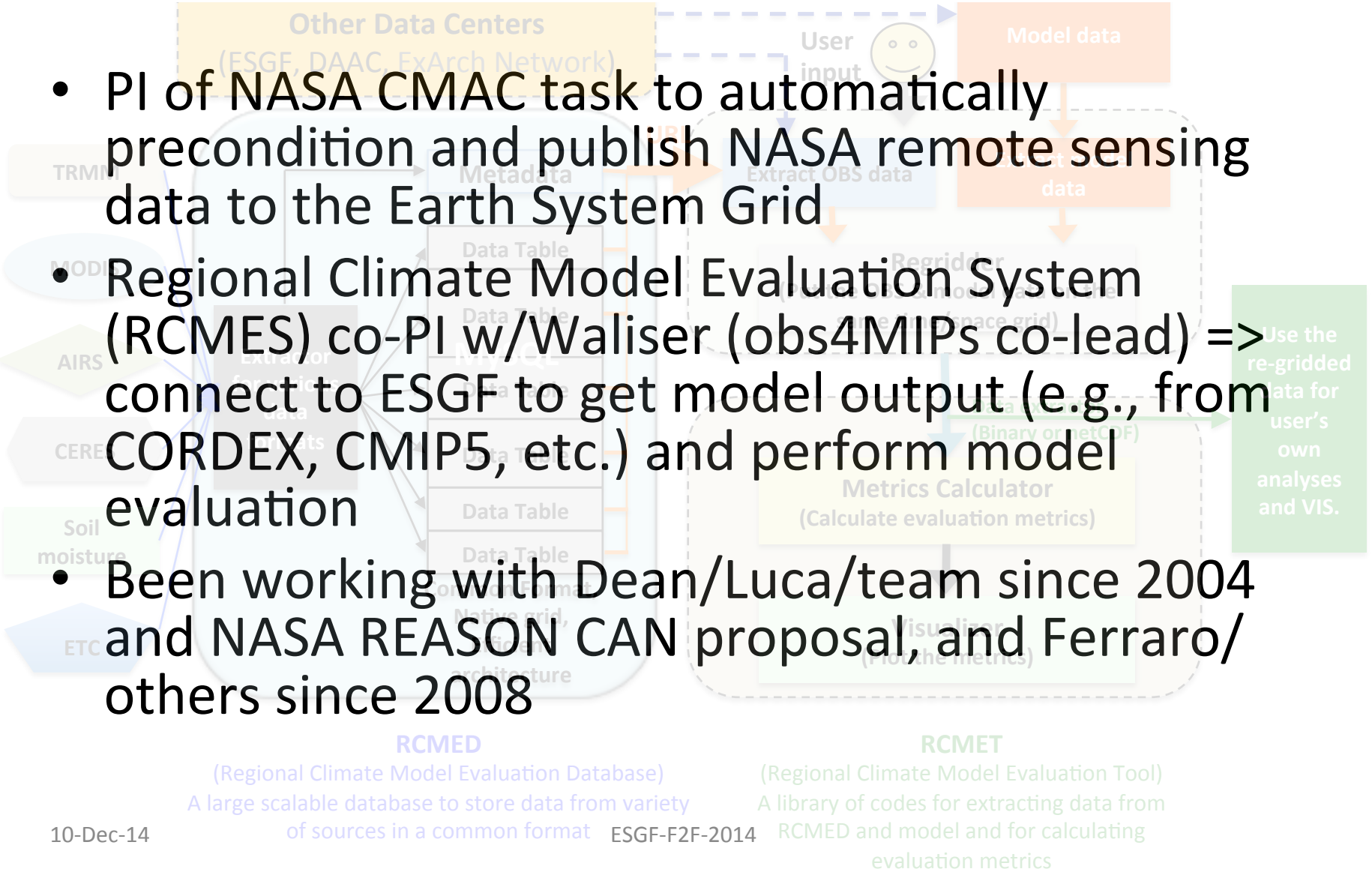
Adjunct Associate Professor, USC
Director, Apache Software Foundation

And you are?

- Chief Architect at NASA JPL in Pasadena, CA USA
- Software Architecture/ Engineering Prof at Univ. of Southern California
- One of original PMC members for Apache Nutch
  - predecessor to Hadoop
- Apache Board of Directors involved in
  - OODT (VP, PMC), Tika (PMC), Nutch (PMC), Incubator (PMC), SIS (PMC), Gora (PMC), Airavata (PMC)

And what do you know about ESGF?

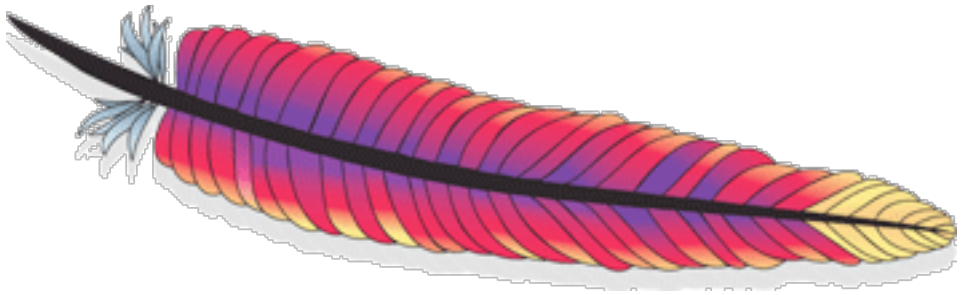
- PI of NASA CMAC task to automatically precondition and publish NASA remote sensing data to the Earth System Grid
- Regional Climate Model Evaluation System (RCMES) co-PI w/Waliser (obs4MIPs co-lead) => connect to ESGF to get model output (e.g., from CORDEX, CMIP5, etc.) and perform model evaluation
- Been working with Dean/Luca/team since 2004 and NASA REASON CAN proposal, and Ferraro/ others since 2008



Jet Propulsion Laboratory
California Institute of Technology

The Apache Software Foundation

- Largest open source software development entity in the world
  - Over 3000+ committers
  - Over 4200+ contributors
  - Over 400+ members
- 150+ Top Level Projects
  - 34 Incubating
  - 32 Lab Projects
- 28 retired projects in the "Attic"
- Over 1.6 *million* revisions
- OpenOffice downloaded *10 million* times per day
- 501(c)3 non-profit organization incorporated in Delaware



-Over 10M successful requests served a day across the world

-HTTPD web server used on 100+ million web sites (53+% of the market)

Apache is a well recognized brand

GOVERNOR ARNOLD SCHWARZENEGGER

November 5, 2009

Apache Software Foundation

It is a great pleasure to extend my greetings to all those attending ApacheCon and congratulations on your tenth anniversary.

I applaud your incredible work over the past decade and appreciate you choosing California as the place to celebrate this fantastic milestone. Our state is a land of innovation, and you have likewise fostered great technological advancements that have touched the lives of millions of people around the world.

Whether managing financial systems, positioning satellites or powering websites through the Apache HTTP Server, your open source projects play key roles in making our information age possible. Thank you for your extraordinary accomplishments and commitment to discovery.

On behalf of all Californians, I send my gratitude to everyone in attendance for your participation, and I offer my best wishes for a rewarding conference and continued success.

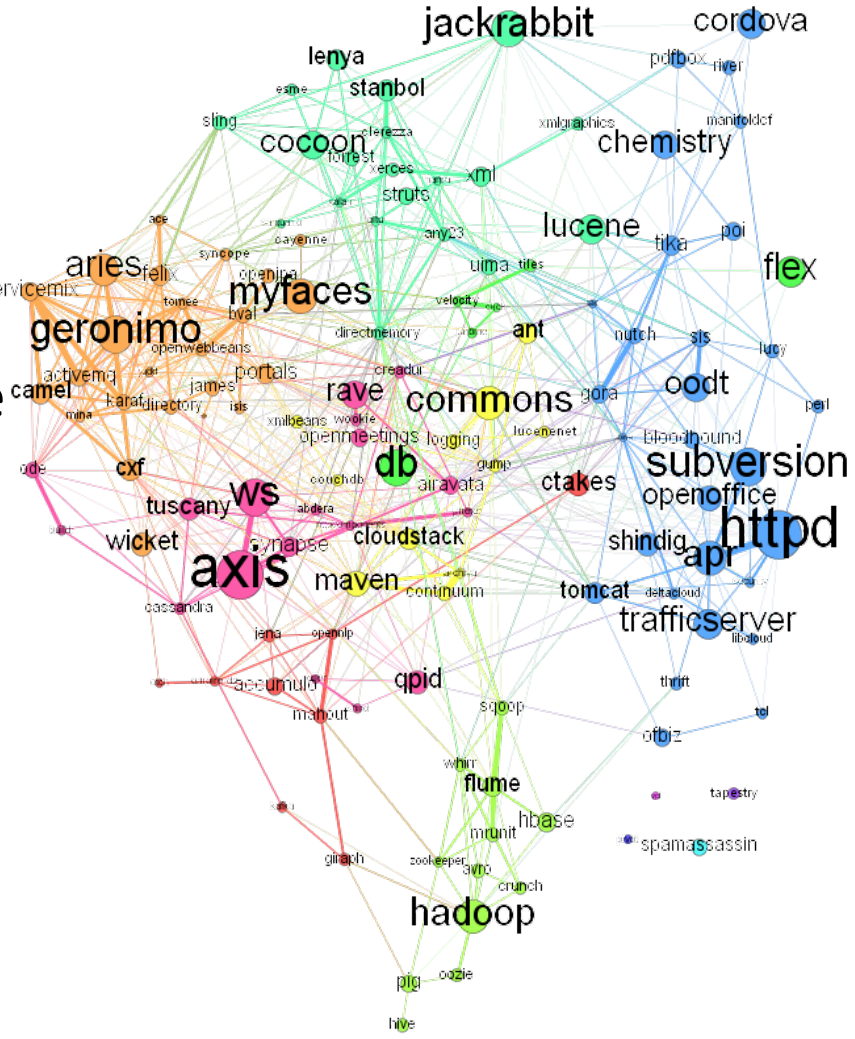
Sincerely,

Arnold Schwarzenegger

STATE CAPITOL · SACRAMENTO, CALIFORNIA 95814 · (916) 445-2841

Grow “communities”

- …software is a side effect
- Apache is a foundation that embodies a set of tried and true “loose” principles that govern the health, measurement and evolution of communities



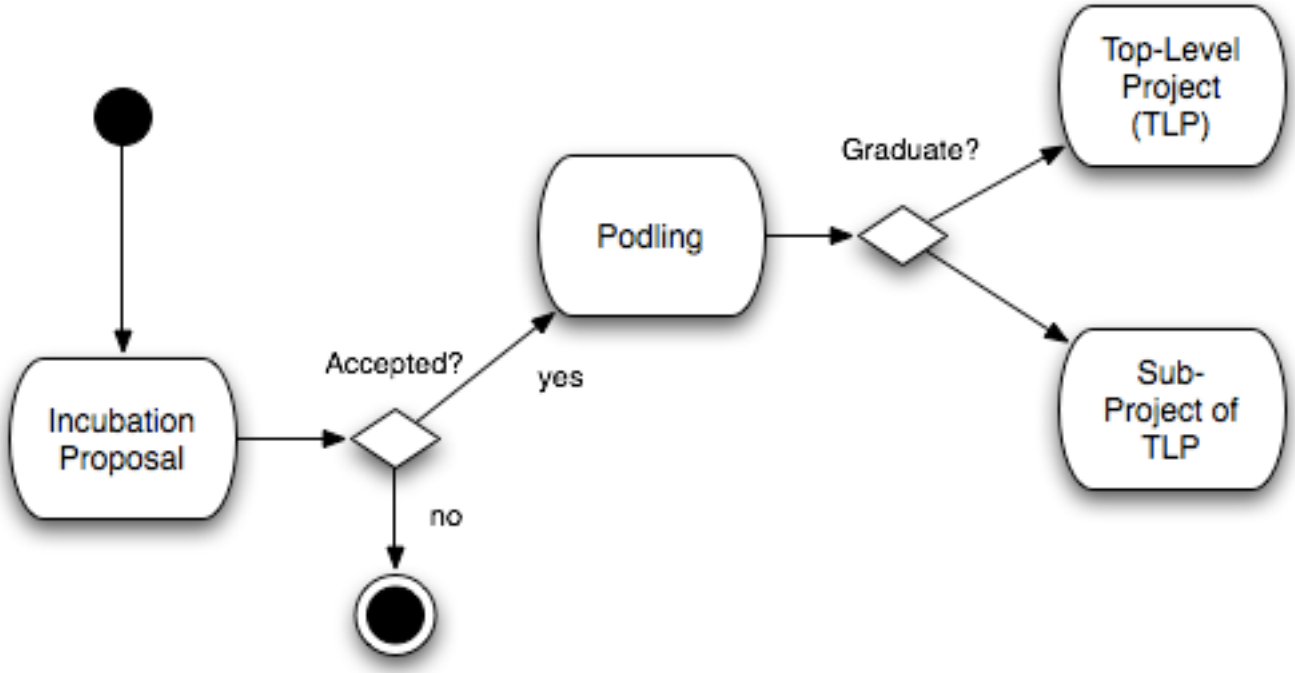
Fielding, Roy T. "Shared leadership in the Apache project." Communications of the ACM 42.4 (1999): 42-43.

Some principles of Apache: "The Apache Way"

- Release software (early, often, whatever)
  - *Communities that don't make releases are dead*
- Add contributors from diverse organizations
  - *Be resilient in the face of any one organization pulling their funding/resources/people from the project - <don't shame them when they do; life happens> - projects that don't add contributors are dead*
- Release software under permissive license, ALv2
  - *Related to BSD, MIT, adds patent termination/indemnification downstream*
- Perform license/component vetting according to well established collective works guidelines
- It takes at a minimum 3 people to get things done
  - *PMCs need at least 3 active members*

Apache Maturity Model

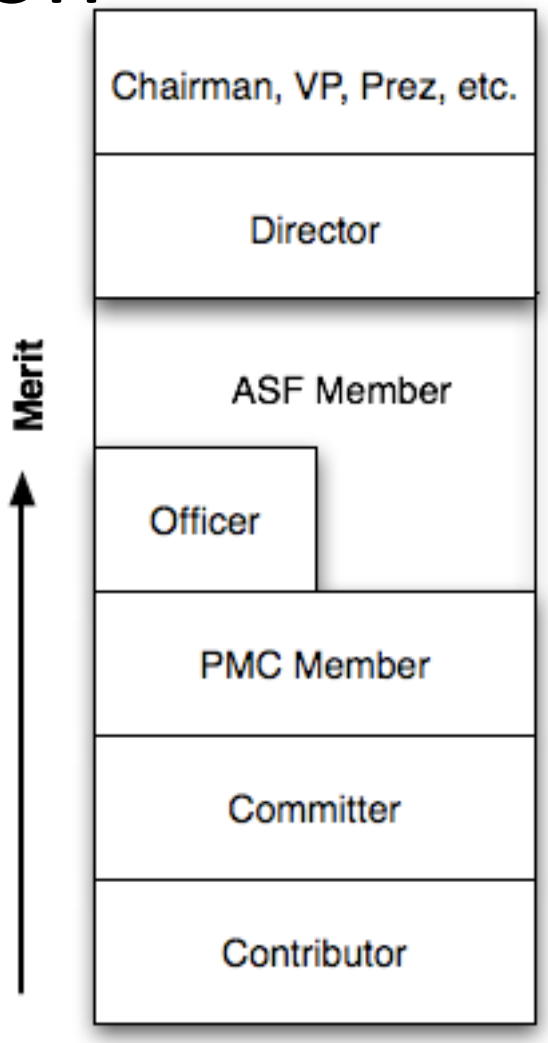
- Start out with Incubation
- Grow community
- Make releases
- Gain interest
- Diversify



- When the project is ready, graduate into
  - Top-Level Project (TLP)
  - Sub-project of TLP
- Increasingly, Sub-projects are <u>discouraged</u> compared to TLPs

Apache Organization

- Apache is a meritocracy
  - You earn your keep and your credentials
- Start out as *Contributor*
  - Patches, mailing list comments, etc.
  - No commit access
- Move onto *Committer*
  - Commit access, evolve the code
- *PMC Members*
  - Have binding VOTEs on releases/personnel
- *Officer (VP, Project)*
  - PMC Chair
- *ASF Member*
  - Have binding VOTE in the state of the foundation
  - Elect Board of Directors
- *Director*
  - Oversight of projects, foundation activities



Many topical projects at Apache

Some Metrics:

1. Open Office downloaded **10 million times a day**
2. HTTPD still powers **53% of the Internet According to recent Netcraft Survey**
3. Apache Hadoop (ecosystem) **1.5Billion dollar industry**

Many topical **Science Projects**

- Biomedical, Earth science, radio astronomy, physics, remote sensing, eScience,

SD Times SOFTWARE DEVELOPMENT

http://www.sdtimes.com/default.aspx

SEARCH

As of February 16, 2013 02:29 PM

HOME >> OPINIONS

Apache does science

By Chris Mattmann

February 12, 2013 — Apache is more than just building Web servers: We do science too!

The Apache Software Foundation is one of the leading open-source clearinghouses responsible for the technology that empowers the Internet, like the HTTPD Web server; the technology that powers Big Data, like Hadoop; and more recently the technology that powers consumer office productivity applications, like Open Office. This is common knowledge in the tech sector. What isn't common knowledge to the tech sector is that the ASF has grown in recent years from being solely focused on technology communities to being also focused on communities that support science. Yes, science people. Apache does science too.

Take the Apache OODT project, originating from within the walls of NASA over the last decade, and including huge staff time from NASA, other government agencies and university partners. OODT allows the general software enthusiast to manage data the same way that NASA's next generation of Earth science remote sensing missions do, and the same way that NASA's Planetary Data System does. (PDS is the archive for all planetary missions over the last 40 years.)

Besides NASA, OODT includes a number of contributors: from next-generation astronomical ground-based instruments like the Square Kilometre Array (which will generate over 700TB of data per second when it sees first light in 2020); from Big Data efforts in climate science; and from biomedical informatics systems at the U.S. National Cancer Institute, helping in the management of data related to the early detection of cancer in the Early Detection Research Network project.

Many projects **funded by Govt Entities**

- Accumulo, NiFi, OODT, Airavata, cTAKES
  - NSAx2, NASA, NSF, HHS respectively

NSA partners with Apache to release open-source data traffic program

Summary: *The National Security Agency has released a new open-source program for data network interoperability.*

By Steven J. Vaughan-Nichols for Linux and Open Source | November 25, 2014 -- 17:26 GMT (09:26 PST)

Get the ZDNet Must Read News Alerts - US newsletter now

Many of you probably think that the National Security Agency (NSA) and open-source software get along like a house on fire. That's to say, flaming destruction. You would be wrong.

In partnership with the Apache Software Foundation, the NSA announced on Tuesday (https://www.nsa.gov/public_info/press_room/2014/nifi_announcement.shtml) that it is releasing the source code for Niagarafiles (Nifi). The spy agency said that Nifi "automates data flows among multiple computer networks, even when data formats and protocols differ".

Details on how Nifi does this are scant at this point, while the ASF continues to set up the site where Nifi's code will reside.

In a statement, Nifi's lead developer Joseph L Witt said the software "provides a way to prioritize data flows more effectively and get rid of artificial delays in identifying and transmitting critical information".

ESGF Software and Community Questions

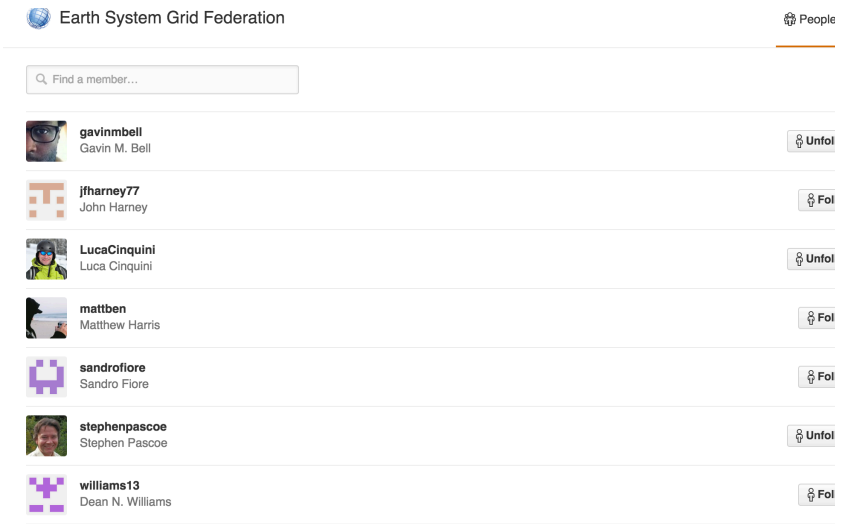
Jet Propulsion Laboratory
California Institute of Technology

- Some questions and comments I hear about ESGF software from time to time
  - *Who owns the ESGF software?*
  - *Who funds ESGF software?*
  - *Who owns ESGF component X?*
  - *When will component X be released?*
  - *Person X just made this great continuous integration and test framework at their institution – how to get others to use it?*
  - *Who should we ask about software component X?*
- Answers to these questions are complex, multi-organizational, across funding and political boundaries
- More on the next slide..


BENEVOLENT
DICTATOR

OK it's not just Luca, but..

- According to Github, there are 7 members of the ESGF org, 5 contributors to esgf-installer, 13 to esgf-web-fe, 3 to esgf-publisher, etc.
- Sincere question(s):
  - How many people are actually developing ESGF "core" software (and not downstream software or ops software that uses it, etc.)
  - What is core; what is downstream? What are the implications of either? Why the disparity in members of the org and contributors?
  - What happens if <set of people developing ESGF software> <insert [get hit by bus; lightning, etc.]>?
  - Why do some of the software related "governance" and "funding" questions keep coming up?
- Myth Busting: *software* governance != *operational* <or insert [other] here governance> - recall paper from Roy



Disclaimer: *Prior slide was meant to be provocative*

Punch line

- ***ESGF should consider software community governance as part of its overall governance structure***
- ***ESGF should consider becoming an Apache Incubator project and I am volunteering to help***
- Github is great and a place to collaborate and share software
  - Pull/Request model is fantastic and collaborative
- Github != software community building
  - It's a tool and enabling capability
- Apache and other foundations have great Git and Github tooling/support – Apache projects can be mirrored at Github and Pull requests are integrated into Apache communities

Potential Benefits

- *Who's responsible for software/component/release/etc. X?*
  - The Project Management Committee
- *What happens if funding agency X removes its funding or if agency Y takes away person Z?*
  - ASF projects resilient to the loss of contributor/funding source/org
- *What are the rules for getting my patch to ESGF core, downstream app, etc., part of ESGF?*
  - The answer isn't talk to <Luca/etc. – kidding> but it is "follow the Apache meritocracy" and process
- Cross pollination with other Apache big time projects
  - Research efforts considering things like Spark, Shark, Mesos, ESGF at Apache encourages cross pollination and interest
- Good news: ESGF already embodies *many of the principles of Apache*!
- **More good news: Apache has separate, independent funding from diverse tech companies and one of its goal is to be around in 50 years**

Thanks!

- Chris Mattmann
  @chrismattmann/Twitter
  chris.a.mattmann@nasa.gov