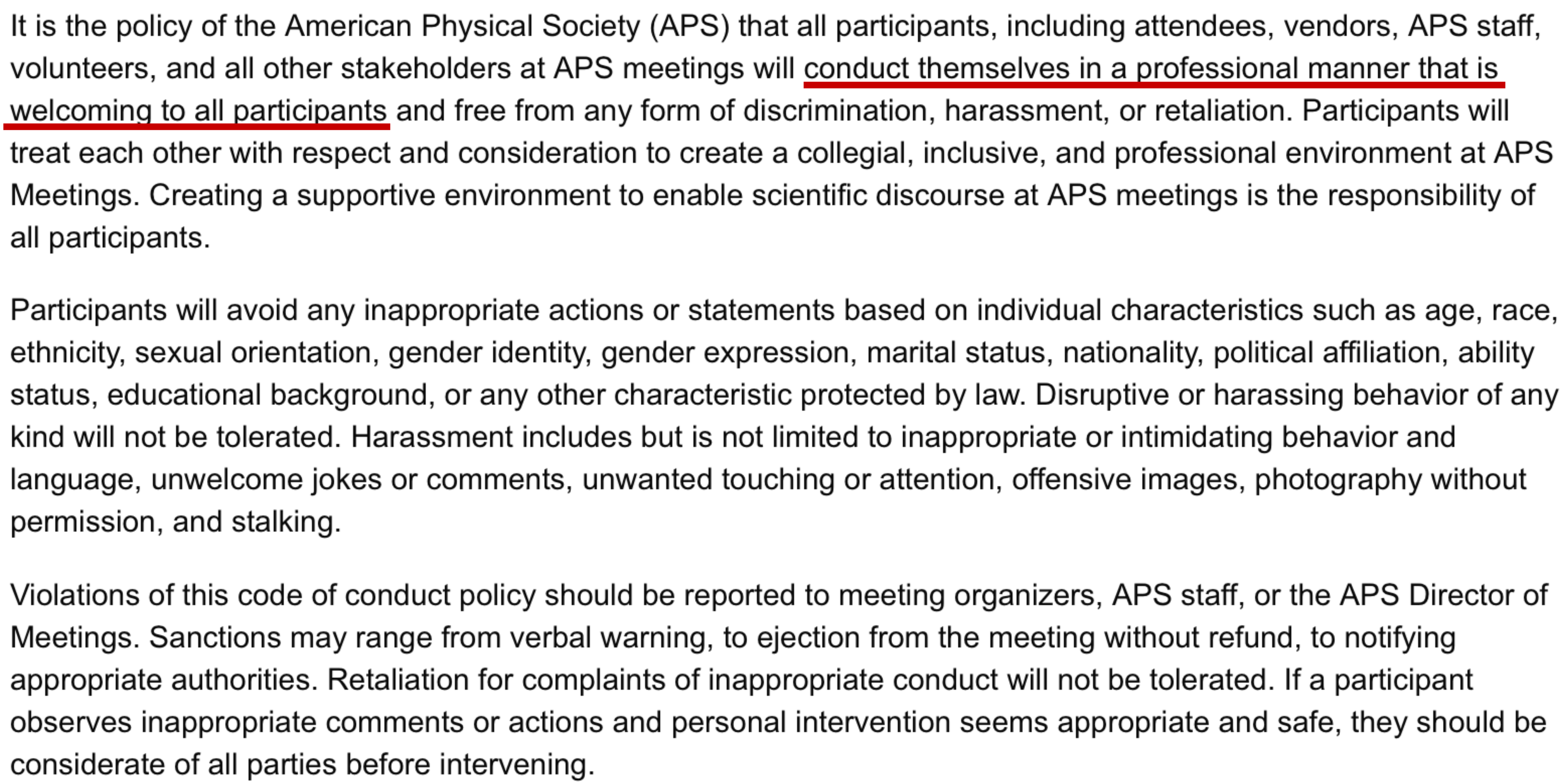
State of OSG

Frank Würthwein
OSG Executive Director
UCSD/SDSC

March 1st 2021

Welcome !!!

We follow the APS Code of Conduct :



It is the policy of the American Physical Society (APS) that all participants, including attendees, vendors, APS staff, volunteers, and all other stakeholders at APS meetings will conduct themselves in a professional manner that is welcoming to all participants and free from any form of discrimination, harassment, or retaliation. Participants will treat each other with respect and consideration to create a collegial, inclusive, and professional environment at APS Meetings. Creating a supportive environment to enable scientific discourse at APS meetings is the responsibility of all participants.

Participants will avoid any inappropriate actions or statements based on individual characteristics such as age, race, ethnicity, sexual orientation, gender identity, gender expression, marital status, nationality, political affiliation, ability status, educational background, or any other characteristic protected by law. Disruptive or harassing behavior of any kind will not be tolerated. Harassment includes but is not limited to inappropriate or intimidating behavior and language, unwelcome jokes or comments, unwanted touching or attention, offensive images, photography without permission, and stalking.

Violations of this code of conduct policy should be reported to meeting organizers, APS staff, or the APS Director of Meetings. Sanctions may range from verbal warning, to ejection from the meeting without refund, to notifying appropriate authorities. Retaliation for complaints of inappropriate conduct will not be tolerated. If a participant observes inappropriate comments or actions and personal intervention seems appropriate and safe, they should be considerate of all parties before intervening.

Private chats are enabled throughout in zoom.
Please follow code of conduct, and especially in private chats.

Simple Rules for Virtual Meetings

- During sessions, all attendees are muted by default.
  - Only the Host & co-hosts can unmute you.
  - During the breaks, you will be able to unmute yourself, and are welcome to talk with others as you see fit.
- Raise your hand if you want to speak.
  - Co-hosts will call on raised hands during Q&A after each talk.
- Feel free to add questions and/or comments to the chat at any time.
  - Co-hosts will answer Q's and/or ask speaker during Q&A after talk.
  - If time runs out, speakers may answer Q's to their talks during the following talks in the same session.
- We will keep zoom session alive during breaks, and you are welcome to continue Q&A then.

The success of the virtual AHM depends on all of us working together within the limitations of being virtual.

OSG "Statement of Purpose"

OSG is a consortium dedicated to the advancement of all of open science via the practice of distributed High Throughput Computing (dHTC), and the advancement of its state of the art.

OSG "Strategy"

- OSG is a collaboration between CI, software, and science professionals.
- **OSG appears differently to different people** …
    - **Trainer, service provider, compute and data infrastructure, scientific user community, …**
- It is governed by the OSG Council, maintaining its by-laws, and electing an executive director for 2 year renewable terms to coordinate a program of work.

Four categories of participants in the OSG Consortium

- The individual researchers and small groups through the Open Science Pool.
- The campus Research Support Organizations
  - Teach IT/CI organizations & support services so they can integrate with OSG
  - Train the Trainers (to support their researchers)
- Multi-institutional Science Teams
  - XENON, GlueX, SPT, Simons, and many many more
  - Collaborations between multiple campuses
- The 4 “big science” projects:
  - US-ATLAS, US-CMS, LIGO, IceCube

The Open Science Pool Community

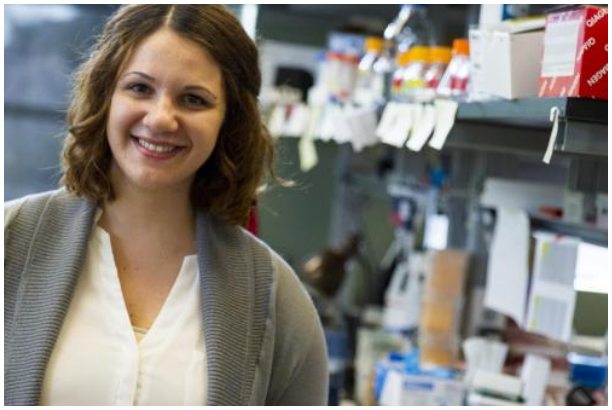
A community for all researchers in the US, from undergraduates to post-graduates.

Following some pointers to science examples.

Natasha Pavlovikj

Graduate Student in Computer Science working with post-doc in Food Science and Technology at UNL.

Developed "ProkEvo", an automated, reproducible, and scalable framework for high-throughput bacterial population genomics analyses.



Natasha analyzed 110 TB of genomics data on OSG in 2020.

Applications of ProkEvo:

- Outbreak detection
- Source tracking
- Understanding epidemics
- Public Health Surveillance
- Pathogen transmission
- Discover ecological properties

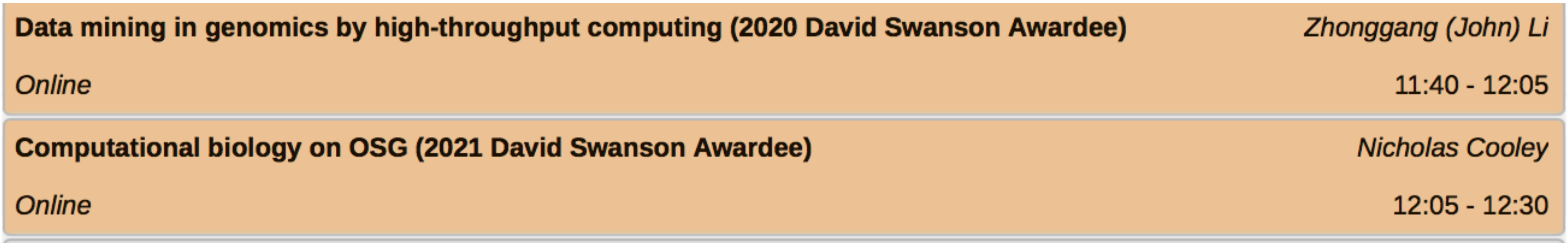
Example Analysis for publication:
23k genomes analyzed in 26 days
Pegasus workflow with 200k++ jobs
producing 1.2TB output data.

Natasha's talk at OSG AHM 2020

Resulting publication:
DOI: 10.1101/2020.10.13.336479

More Examples Today

David Swanson Award Recipients:



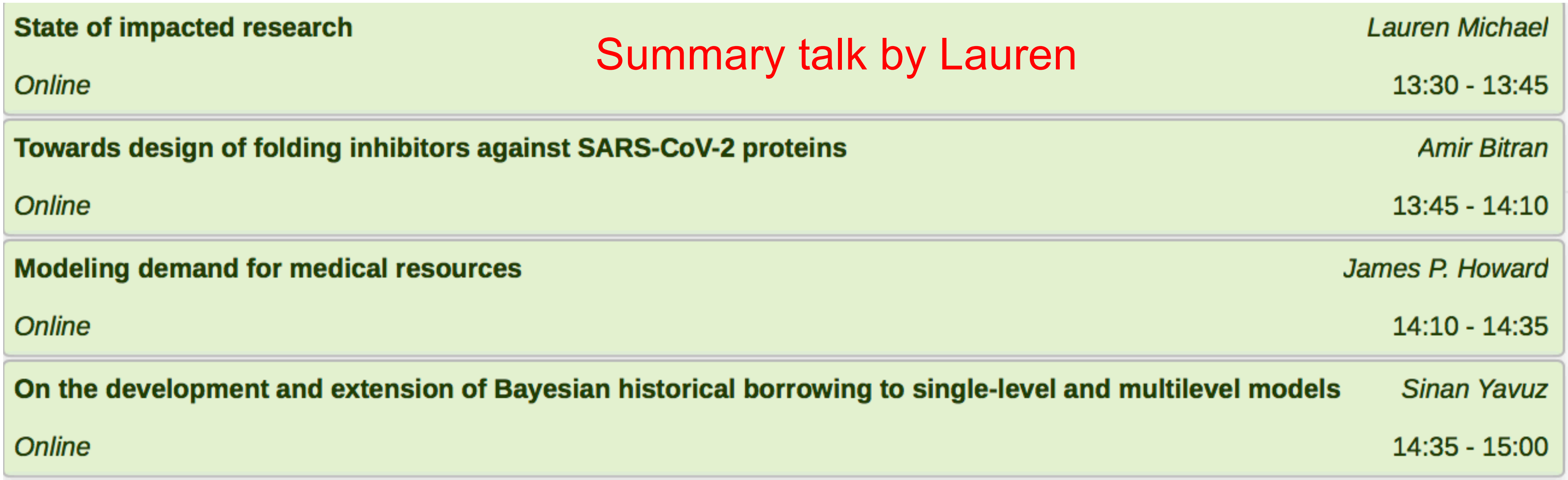
Data mining in genomics by high-throughput computing (2020 David Swanson Awardee) *Zhonggang (John) Li*

Online 11:40 - 12:05

Computational biology on OSG (2021 David Swanson Awardee) *Nicholas Cooley*

Online 12:05 - 12:30

Monday Afternoon Session:



State of impacted research *Lauren Michael*

Online 13:30 - 13:45

Summary talk by Lauren

Towards design of folding inhibitors against SARS-CoV-2 proteins *Amir Bitran*

Online 13:45 - 14:10

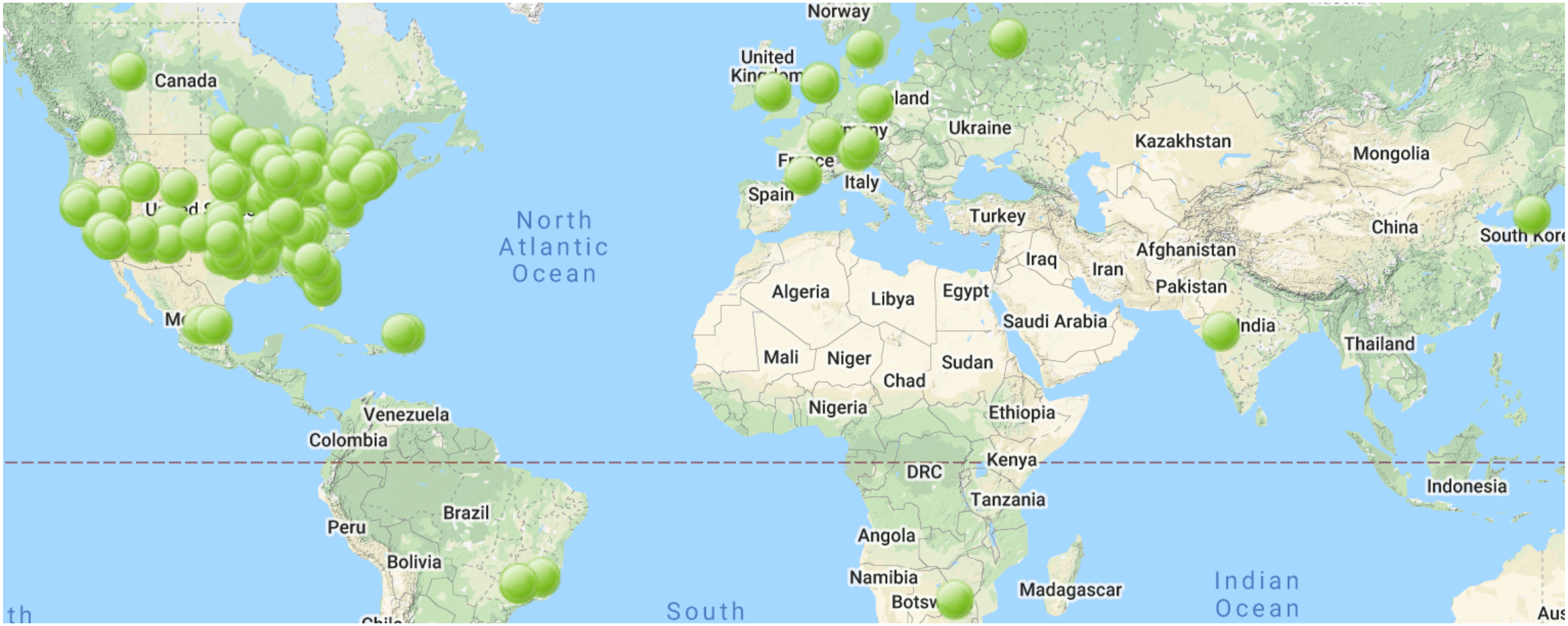
Modeling demand for medical resources *James P. Howard*

Online 14:10 - 14:35

On the development and extension of Bayesian historical borrowing to single-level and multilevel models *Sinan Yavuz*

Online 14:35 - 15:00

OSG Compute Federation



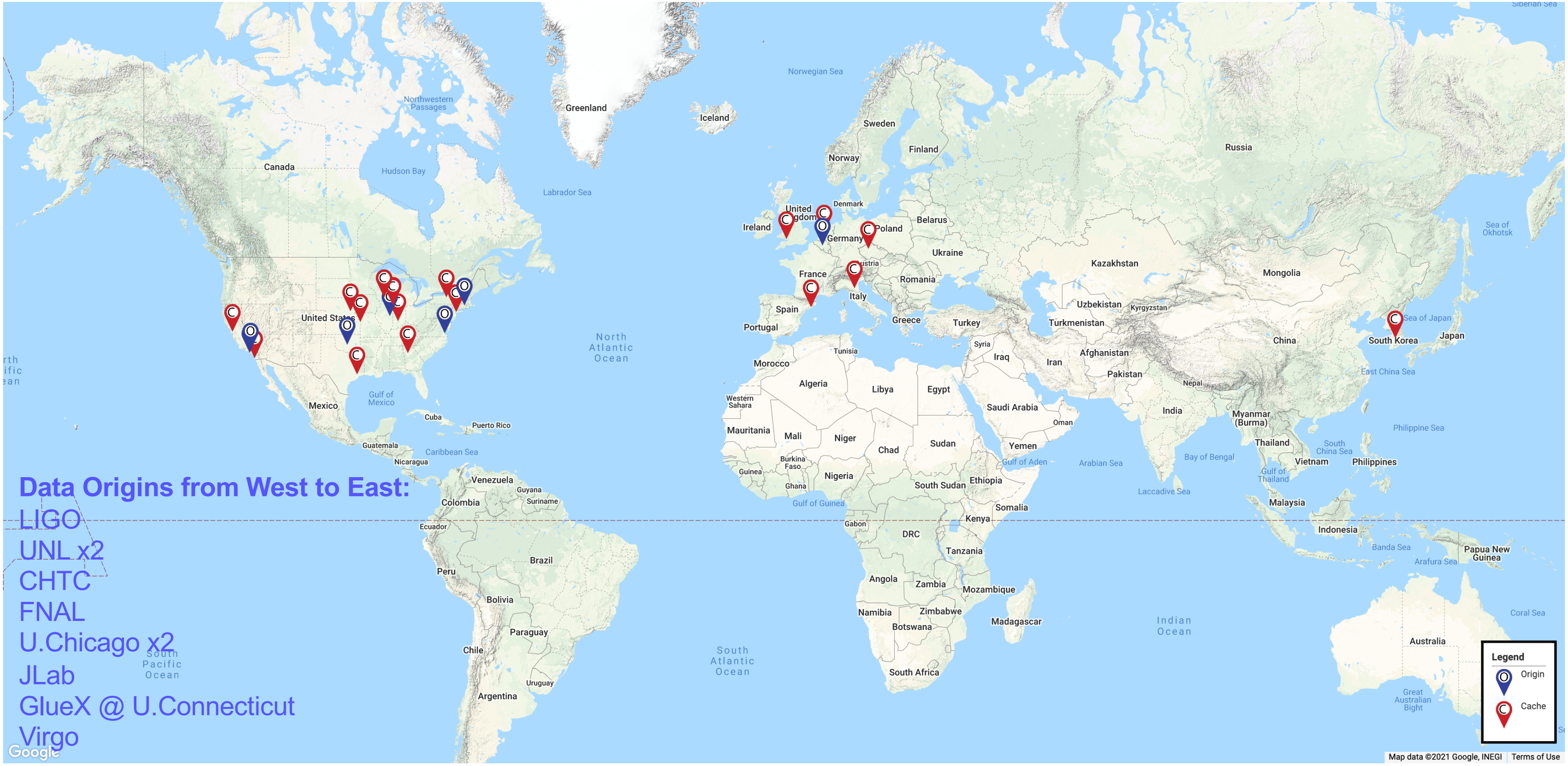
136 "green dots" listed on this map

… but the real story is so much more complicated ….

Open Science Grid

OSG Data Federation

UCSD



Data Origins from West to East:

LIGO
UNL x2
CHTC
FNAL
U.Chicago x2
JLab
GlueX @ U.Connecticut
Virgo

17 Caches … 6 of which are in R&E network backbone

10 Data Origins … one of which is for all of open science

The “real story”

- OSG supports a **modular software stack** and a “**Fabric of Services**” that allows organizations to create their own dHTC environment.
  - **dHTC environments from multiple organizations typically have non-trivial overlaps.**
    - They may share resources
    - They may share services
    - They share data in a global namespace for public and private data
- **OSG operates one feature complete instance of such an environment for the common good of all of open science**
  - We support the common good by democratizing access
  - We eat our own dog food by operating software & sevices we support
  - We teach others to follow our lead to support their researchers and collaborations.

A Feature-Complete dHTC Environment

- **Open Science Pool**
  - Submission infrastructure that functions as compute “Access Point”
  - Workload management system
    - Complex workflows across heterogenous resources possible.
      - Easy to run workflows comprised of 100,000 jobs or more with complex dependencies between sets of jobs (full support of arbitrary DAGs).
  - Homogeneous runtime environment across heterogeneous resources
    - Includes a dozen or more types of GPUs from NVIDIA and AMD
    - Includes some FPGAs
    - 100’s of application “modules” and 100’s of curated containers
- **Open Science Data**
  - Storage that functions as “Data Entry Point”
  - Transparent “Data Access” via Global Content Delivery Network
  - Supporting Public and Private Data

Any Researcher in the US can request access to this dHTC environment

Open Science Grid

UCSD

The OSG Data Story

One of the best hidden secrets of OSG

In 2020, we saw for the first time many individual Researchers with larger than TB data sets, even one with a 100TB data set.

Large Data no longer just a feature of large collaborations in physics and astronomy.

Modalities of Data Use

1. Science that produces data.

   Small input large output

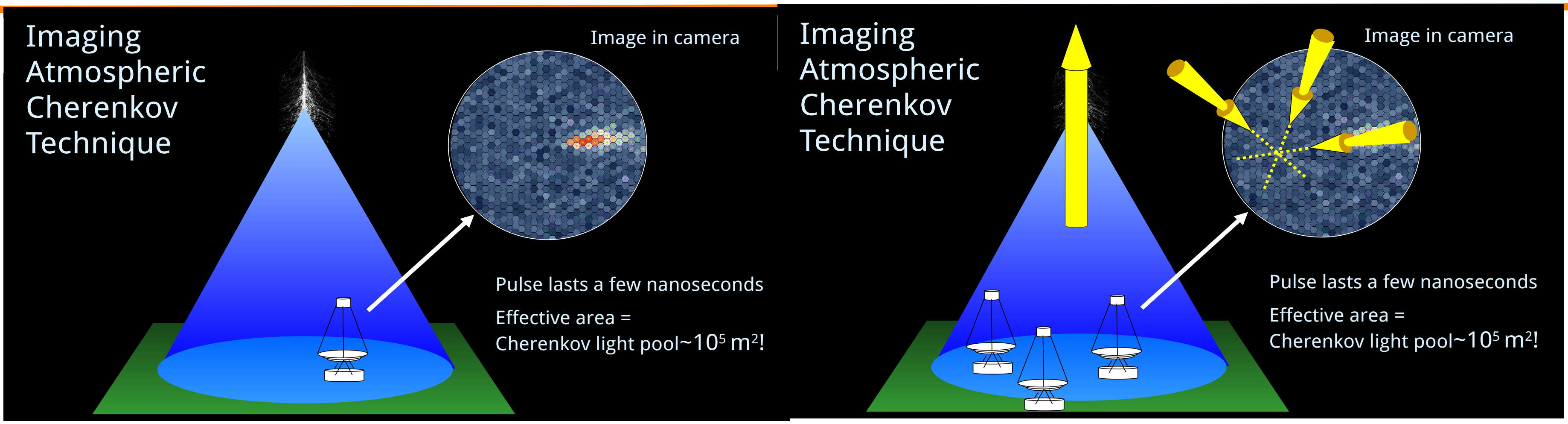
2. Science that processes data once.

   Large input small output

3. Science that reuses the same data over and over.

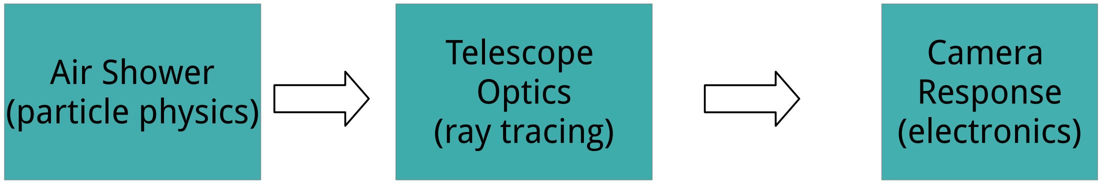
We will show some examples for each, with pointers to more information.

1. Producing 400TB on OSG


Imaging Atmospheric Cherenkov Technique
Image in camera
Pulse lasts a few nanoseconds
Effective area =
Cherenkov light pool~10^5 m^2!
Imaging Atmospheric Cherenkov Technique
Image in camera
Pulse lasts a few nanoseconds
Effective area =
Cherenkov light pool~10^5 m^2!

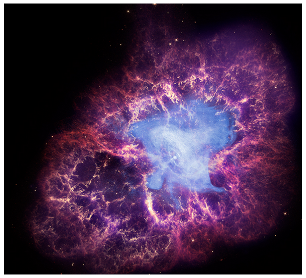
Multiple telescopes allow measurement of direction via triangulation

Simulation Chain:



Air Shower (particle physics) → Telescope Optics (ray tracing) → Camera Response (electronics)

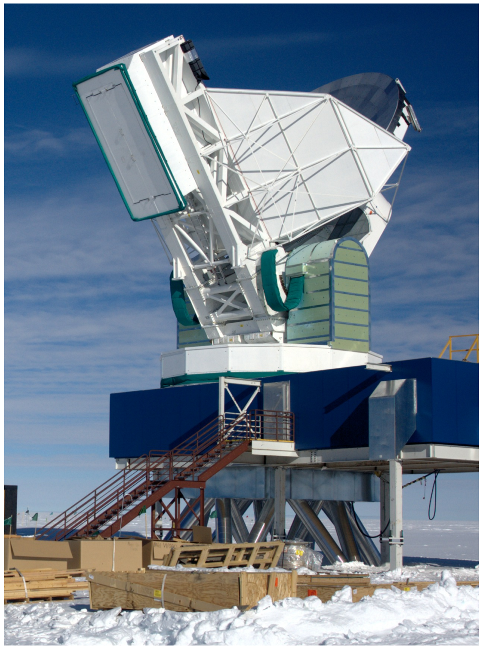
10^9 particle showers, $2 \cdot 10^7$ CPU hours, 400 TB



Challenge: measure sources extensions that are 1/10th of the size of intrinsic resolution.
=> excellent simulations required.

Nepomuk Otte presentation @ OSG AHM 2020

2. Analyzing O(PB) on OSG



SPT-3G: 12,000 detectors sampled at 1502Hz
Collecting >TB/day of data.

Challenge: turn RAW data into CMB map of sky.

Slice data in GB chunks per job.
Run many many jobs.

Roughly 5-10 Million CPU hours per science paper

Next gen CMB instrument in late 2020 will collect ~70TB/day of data.

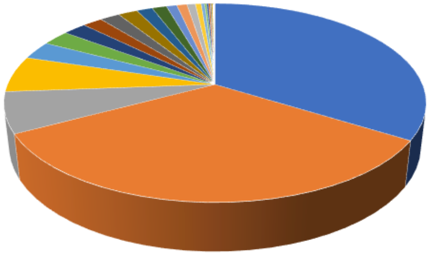
Nathan Whitehorn @ OSG AHM 2020

3. About 100 Projects used the caching infrastructure in 2020

Average data reuse ~ 100 times

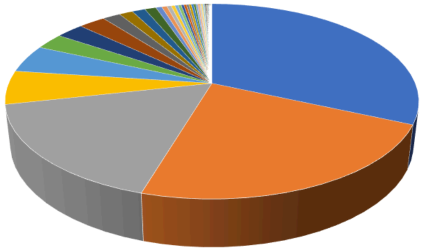
Projects with largest Working Sets are not the most read, or reused data

New in 2020:
More individual researchers !!!



Read Volume

22 PB



Working sets

212 TB

Projects with largest working sets

Owner	Working Set	Data Read	Reuse
Npavlovikj (UNL)	110 TB	346 TB	3.1
LIGO (Private)	46 TB	6.9 PB	152
CHTC	11 TB	380 TB	34
elagin	5.4 TB	8.1 TB	1.5
tuxino	5.2 TB	6.3 TB	1.2
LIGO (Public)	7.8 TB	1.9 PB	243
runyu	3.4 TB	5.3 TB	1.6
gladstein	2.1 TB	3.7 TB	1.7

Projects with largest reuse

Owner	Working Set	Data Read	Reuse
Nova	370 GB	7.0 PB	20,000
gziegler	24 GB	280 TB	12,000
odgerk	113 GB	630 TB	5,700
Dune	33 GB	161 TB	5,000
cgomes02	100 GB	385 TB	4,000
SBND	100 GB	203 TB	1,300
MicroBoone	412 GB	385 TB	950
Minerva	600 GB	450 TB	780

(ignored projects with working sets smaller than 10 GB)

Wide Range of Sciences use the caching infrastructure

Bioinformatics (npavlonikj)
Evolutionary Biology (gladstein)
Plant Genetics (gziegler)
Computer Engineering (cgomes02)
Instrumentation R&D (elagin, runyu, odgerk, tuxino)
Experiments in (Astro)Physics (LIGO, Nova, Dune, SBND, MicroBoone, Minerva)

Projects with largest working sets

Owner	Working Set	Data Read	Reuse
Npavlovikj (UNL)	110 TB	346 TB	3.1
LIGO (Private)	46 TB	6.9 PB	152
CHTC	11 TB	380 TB	34
elagin	5.4 TB	8.1 TB	1.5
tuxino	5.2 TB	6.3 TB	1.2
LIGO (Public)	7.8 TB	1.9 PB	243
runyu	3.4 TB	5.3 TB	1.6
gladstein	2.1 TB	3.7 TB	1.7

Projects with largest reuse

Owner	Working Set	Data Read	Reuse
Nova	370 GB	7.0 PB	20,000
gziegler	24 GB	280 TB	12,000
odgerk	113 GB	630 TB	5,700
Dune	33 GB	161 TB	5,000
cgomes02	100 GB	385 TB	4,000
SBND	100 GB	203 TB	1,300
MicroBoone	412 GB	385 TB	950
Minerva	600 GB	450 TB	780

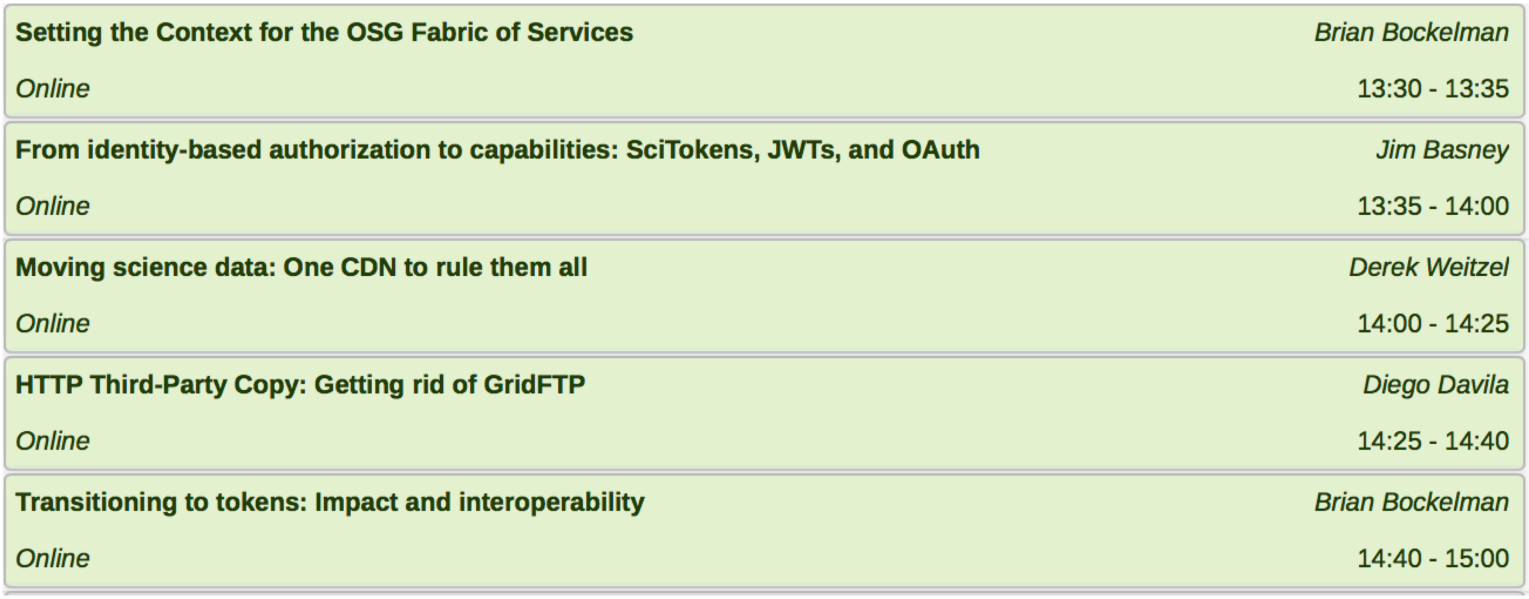
(ignored projects with working sets smaller than 10 GB)

Advancing the State of the Art

This year's biggest stories of change:

- Replacing gridFTP with HTTP(S)
- Replacing authorization based on person with authorization based on capability.
- OSG 3.6 release came out last week => first release without any Globus Toolkit legacy software.

See Wednesday Afternoon Session:



Setting the Context for the OSG Fabric of Services	*Brian Bockelman*
Online	13:30 - 13:35
From identity-based authorization to capabilities: SciTokens, JWTs, and OAuth	*Jim Basney*
Online	13:35 - 14:00
Moving science data: One CDN to rule them all	*Derek Weitzel*
Online	14:00 - 14:25
HTTP Third-Party Copy: Getting rid of GridFTP	*Diego Davila*
Online	14:25 - 14:40
Transitioning to tokens: Impact and interoperability	*Brian Bockelman*
Online	14:40 - 15:00

Summary & Conclusion

- OSG continues to **advance all of open science via the practice of dHTC, and the advancement of its state of the art**.
  - Lot's of "Big Data" across many science domains
- Open Science Pool as strategy to **democratize access to dHTC**
- The end of an era is coming: the **first OSG release without Globus** has arrived

Acknowledgements

- This work was partially supported by the NSF grants OAC-2030508, OAC-1841530, OAC-1836650, and MPS-1148698