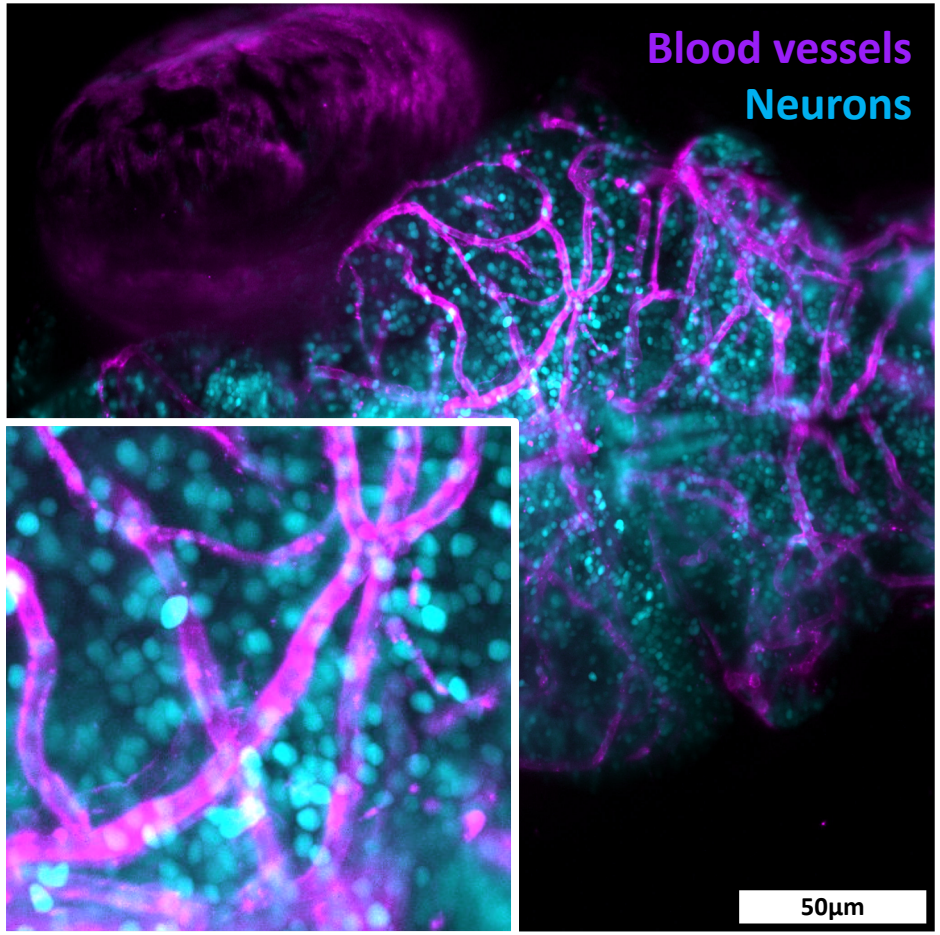
Raw data storage is a source a lot of problems



Zebrafish *in vivo* light sheet data from Thai Truong/Scott Fraser lab (U SoCal)

Hi, my name is Andrey. I work on:

Zebrafish brain light-sheet

- generating TBs of functional and structural data

Image analysis for friends

- analyzing 10s GB imaging data from other people/labs

Please @ me! aandreev@caltech.edu
Andrey Andreev (David Prober lab)
Slides are online: aandreev.net/LSFM

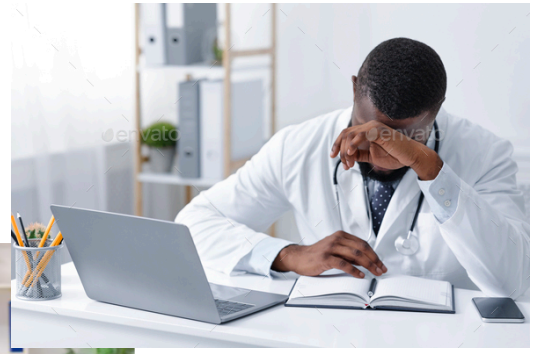
Caltech

Practice and theory of microscopy data management are… disconnected

Awesome expensive microscope

TBs of Data…

Smart scientists with USB sticks and a Dropbox



Great data management principles, concepts, and tools

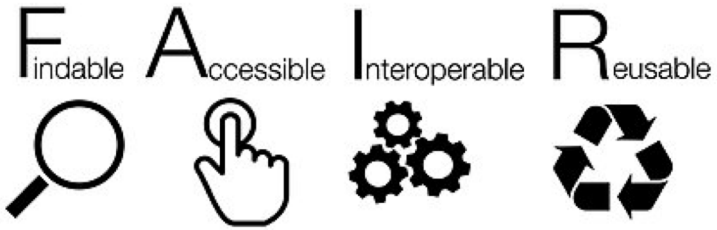
- Written Data Sharing plan
- FAIR data principles
- Standardized formats
- Carefully recorded metadata
- Databases with public API
- “Cloud” processing and storage
- Seamless sharing and collaboration
- Assigning DOIs to datasets
- Data archiving

Andrey Andreev, aandreev@caltech.edu
David Prober lab
Slides: aandreev.net/LSFM

Leaky data plumbing makes data unFAIR

Raw data from an expensive microscope

Published data

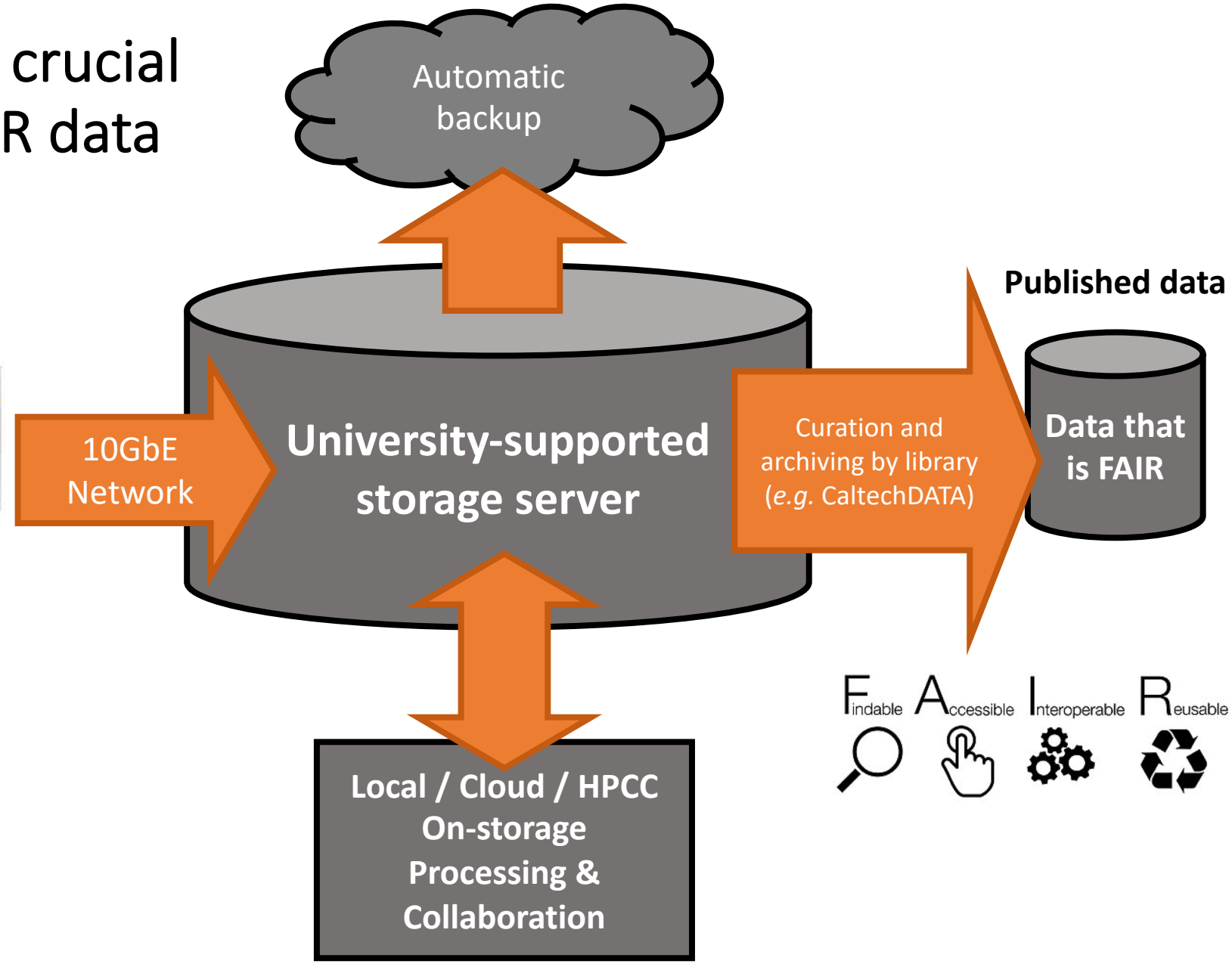


- ❑ **Data not backed up, lost**
- ❑ **Grant supporting storage is over**
- ❑ **Data not copied to USB / Dropbox**
- ❑ **Data never deposited in an archive**
- ❑ **Only copy is on researcher's laptop**
- ❑ **Data sharded across multiple drives**
- ❑ **Grad student who supported storage server left**

Andrey Andreev, aandreev@caltech.edu
David Prober lab
Slides: aandreev.net/LSFM

Centralized storage is a crucial step towards more FAIR data

Raw data from an expensive microscope



Andrey Andreev, aandreev@caltech.edu
David Prober lab
Slides: aandreev.net/LSFM

But wait, I think that…

Andrey Andreev, aandreev@caltech.edu
David Prober lab
Slides: aandreev.net/LSFM

Only ~~light-sheet~~ fancy microscopes need a lot of data storage	Everyone will benefit from fast, accessible, reliable storage with automatic backups. It allows less pain of data sharing, scaling up of experiments, or starting new projects.
Centralized storage is too expensive	It can costs as little as $10-30/TB to store and backup data, similar to what labs pay already independently. You don't need to go "enterprise" to store 1000 TB
We have network storage already	It is too expensive or very limited, we checked
You can use cloud, we have free Dropbox license	"Cloud" as primary storage is often more expensive, slow, unreliable, and limited. Hard to share and collaborate on large dataset. Hard to isolate sensitive equipment from the Internet
We'll need to hire & train IT staff	IT professionals estimate <1 hr/month support time after initial install, and storage runs on widely-used hardware and software (e.g. Dell hardware & Windows Server; no special training required)
You should just buy your own server if you need it so much	We do that right now, but it is more expensive, in money and time, and even labs with few people or resources deserve good data plumbing
Everyone is used to the *status quo*	It is outdated. Newly-hired professor in neuroscience will need to store TBs of data, quickly. New postdoc will start a data-heavy project. Plumbing lubricates research
You can just use USB drives / sticks	USB drives are unreliable, can be lost, can transfer viruses, data is hard to track and organize
People will deposit tons of their trash data	It's okay, storage is cheap. Library professionals are great at curation and archiving, and can help with retiring data. It is more expensive to lose or create inaccessible data
We'll need fast network, and we barely got 1Gbps	Yes, some faculty will want you to upgrade. It is overdue anyways. Meanwhile they are still using these slow networks

We are drafting a position statement with options-solutions

Physical side: Modern requirements for centralized storage, backup, network, sharing: 500TB per lab, 10GbE minimum, 1-2-3, options like fast flash storage, “share as link” worldwide *etc* Infrastructure tiered options with cost estimates	**People side:** Science is interaction between people. People $/hr >> infrastructure costs How to make change? Lobby & pooling resources between labs / departments Existing services (*e.g.* IT & library) will be able to be integrated

We are drafting a position statement with options-solutions

Physical side: Modern requirements for centralized storage, backup, network, sharing: 500TB per lab, 10GbE minimum, 1-2-3, options like fast flash storage, “share as link” worldwide *etc* Infrastructure tiered options with cost estimates	**People side:** Science is interaction between people. People $/hr >> infrastructure costs How to make change? Lobby & pooling resources between labs / departments Existing services (*e.g.* IT & library) will be able to be integrated

Product 2: Current Status survey

University	Cloud storage vendors	Cloud storage limit	NAS volume, TB	NAS cost, $/TB/yr	HPC initial storage quota, TB	HPC storage cost
1	Box	NA		$137	2TB for year	$355/TB for 5yr
2	Google, Box, Microsoft	NA	5TB per PI	$200	??	??
3	Box, Google, Microsoft	NA	NA	$460	N/A	$1200-$3600/TB/yr
4	Dropbox, Microsoft, Google	NA	2GB	NA	NA	NA
5	Google, Microsoft	NA	10TB/departmen	??	??	??

Product 3: Case Studies & Horror Stories

- Keeping data for 2 years to analyze with new tools
- USC centralized storage for 12+ microscopy people
- CaltechData project by the library

- Mailing USB hard-drives from UK to Germany
- Leaving imaging facility with stack of 20 USB drives
- USB drives are used with acquisition computer with outdated Windows
- Postdoc uses unlimited Google Drive “from the previous university”
- …And more…

Please @ me! I would love to talk to you and your CIO aandreev@caltech.edu
Andrey Andreev (David Prober lab)
Slides: aandreev.net/LSFM

Caltech

Thanks to: Tom Morell, Kristin Briney (Caltech); Valerie Thomas, Dan Koo, Francesco Cutrale, Jeremy Wiemer (USC); Sandra Gesing (U Notre Dame); Jacqueline D Campbell, Sarah Nusser (Iowa State); Damian Dalle Nogare (NIH); Eric Wait, Blair Rossetti (Janelia); Uri Manor (Salk); Ben Steventon (Cambridge), Daniel Waiger (Hebrew U of Jerusalem), Vikas Trivedi (EMBL)