

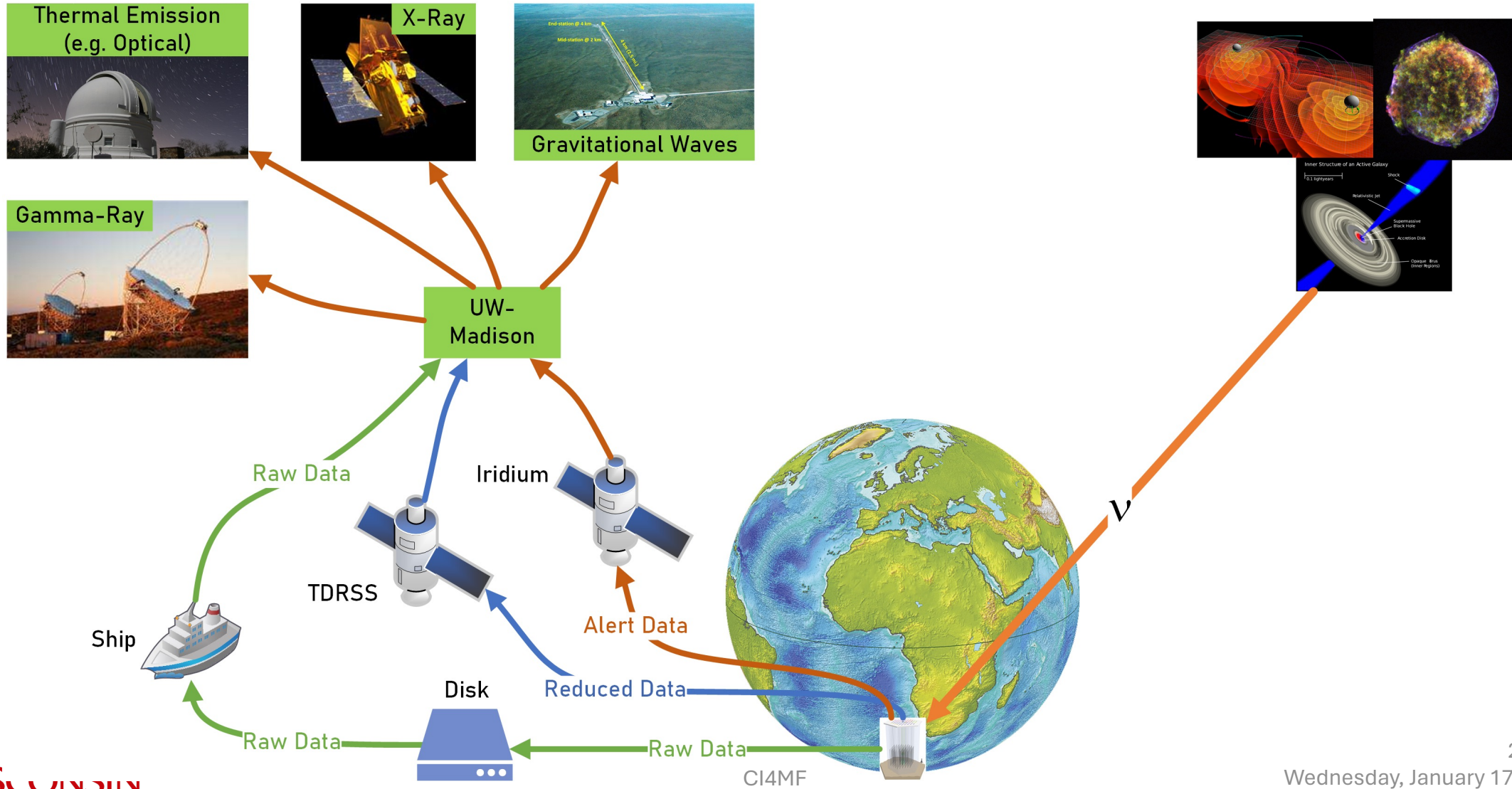
IceCube Data Management

Benedikt Riedel
UW-Madison

CI4MF Panel
17 January 2024



Multi-Messenger Astrophysics



IceCube Data Movement

- Global heterogeneous resources pool
- Mostly shared and opportunistic resources
- Not a lot of data re-use
- Data is centrally hosted at UW-Madison
 - In- and Output data needs to be moved back and forth from/to UW-Madison
 - Good way to become friends with your campus' network team
- Archival
 - Raw Data: At NERSC (GridFTP, transitioning to HTTP)
 - Processed Data: DESY-Zeuthen (GridFTP, transitioning to HTTP)



Enabling technologies?

- HTTP-based transfers – GridFTP is deprecated
- Token-based Authentication – Can manage ourselves compared to X.509
- Column-based Data Formats

Paradigm-shifting Changes?

- AI Training: Demand large VRAM GPUs or sets of GPUs
 - Limited local resources – Cost/benefit is not there yet to buy a “ML Training” box (~\$250-500k/unit)
 - Need to be outsourced to HPC centers – Move data to HPC centers
- AI Inference:
 - Streaming data vs. data pipelines
 - Offloading to a third party? – Inference-as-a-Service
- Distributed Filesystems
 - Object store vs. Filesystem
 - File Cataloging/metadata – 12+ years of data
- Event-based processing
- Better calibration from IceCube Upgrade – Multiple re-processing ahead



Challenges?

- Funding
 - Do more with less
 - Scope creep – FAIR, Open Data, ever increasing datasets
 - How much AI?
- Shifting more to external resource providers
- Shifting technologies
- Archival

Thank you!

Questions?