# Open Source Science for *Large-Scale Data Mining in Earth Observation*

Earth observation data AI processing system architecture,
and international public-private partnerships concept

NASA workshop #2: State-of-the-Art in Mission Data Processing Systems
Mar 1-4, 2022

*terrabyte* initiative @ DLR & LRZ, Germany
presenter: Conrad M Albrecht

# Background & Objectives

1. *Large-Scale Data Mining in Earth Observation**
   - a **key challenge** for deep/machine learning in Earth observation is **limited availability of labels** for model training due to the human labor-intensive collection process
   - (among others) **weakly-supervised learning concerns** artificial intelligence methods to
     - **compress data for multi-purpose** model training (self-supervised learning), and to
     - **handle auto-generated, noisy labels*****

2. the *DLR terrabyte*** Mission Data Processing System (MDPS)
   - What are your key and driving requirements?
     - **efficient data assembly** of co-registered spatio-temporal geo-information for machine learning
     - **co-location of data and compute** for Big Geo-Data processing
     - **geo-data discovery** and cross-layer operations
     - platform interoperability for **federated system designs**
   - Who are the customers and stakeholders/primary user?
     - Earth Observation Center of the German Aerospace Center (national research institute)
     - other research centers of the Helmholtz Association (national research institutes)
     - Technical University of Munich (academia)
     - collaboration with corporate research and industry
   - What external factors constrain your solution?
     - hosting a mix of public and private datasets (data access management)
     - implementation of the GDPR
     - etc. (not in focus for this presentation)

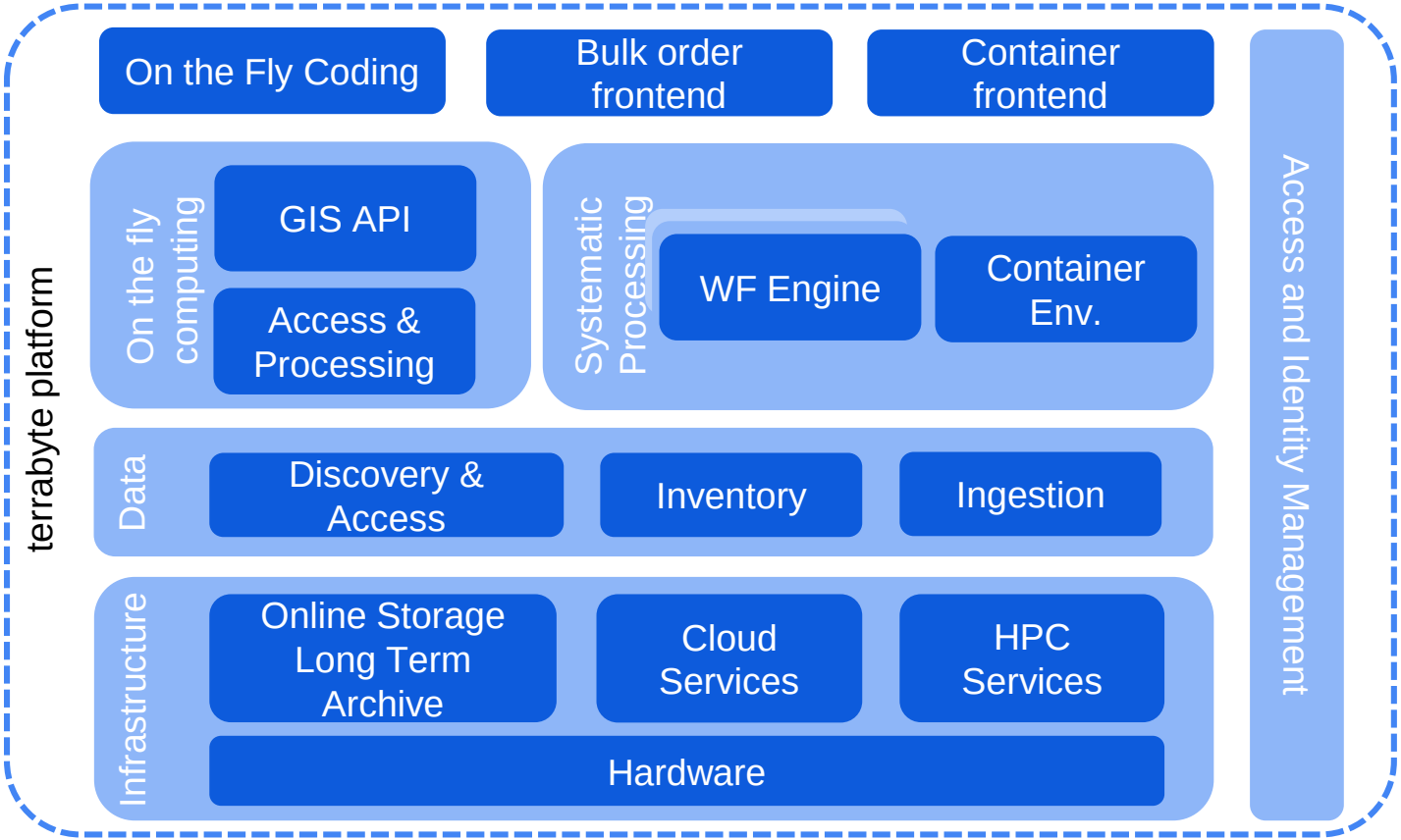


* https://www.asg.ed.tum.de/sipeo/team/dr-rer-nat-conrad-albrecht

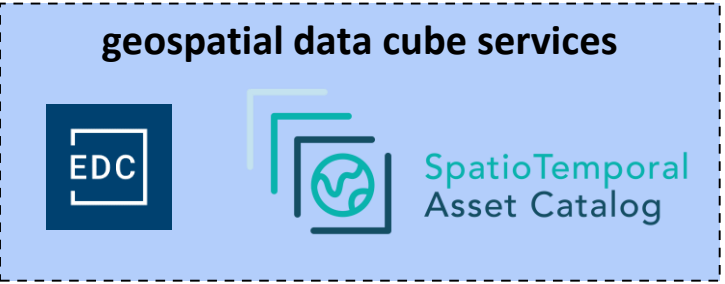
** https://www.lrz.de/presse/ereignisse/2021-07-22-terrabyte_ENG_

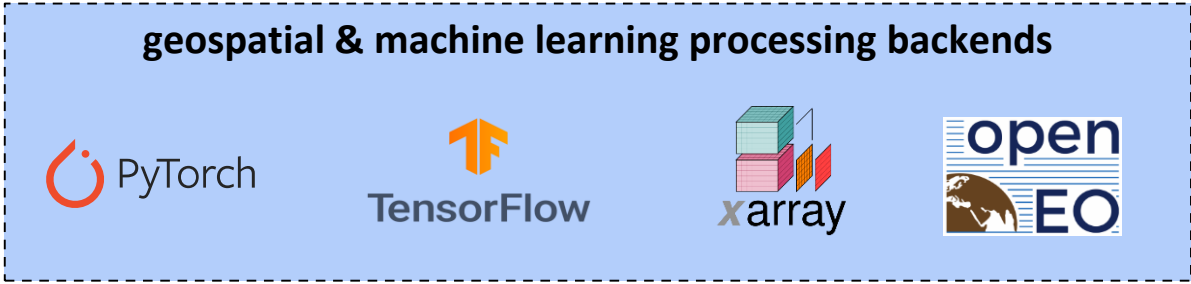
*** https://doi.org/10.1109/BigData52589.2021.9672060

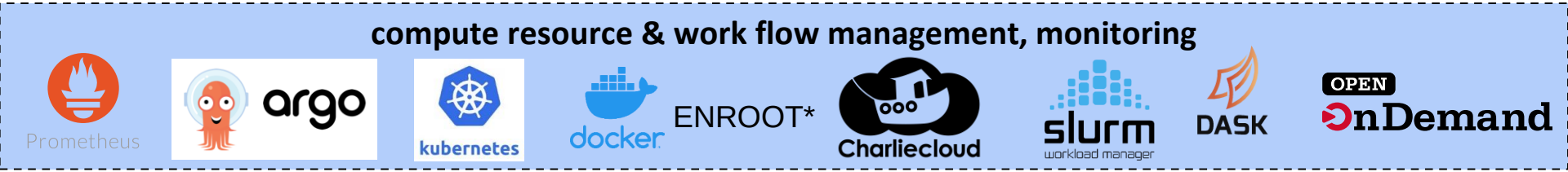
# High-Level Data Processing System Architecture: base system



# Open-Source Software Stack (incomplete) sample list we evaluate

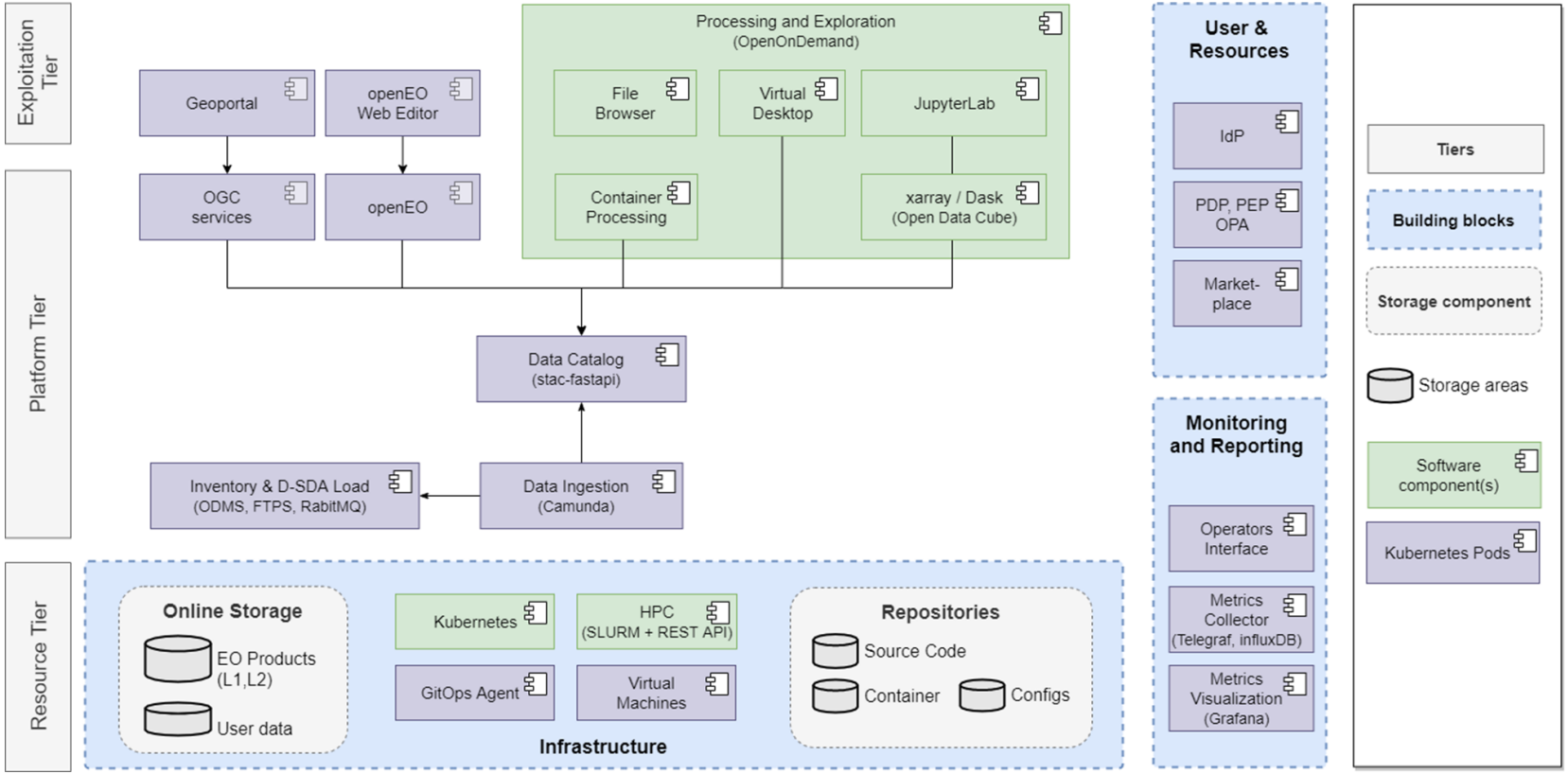



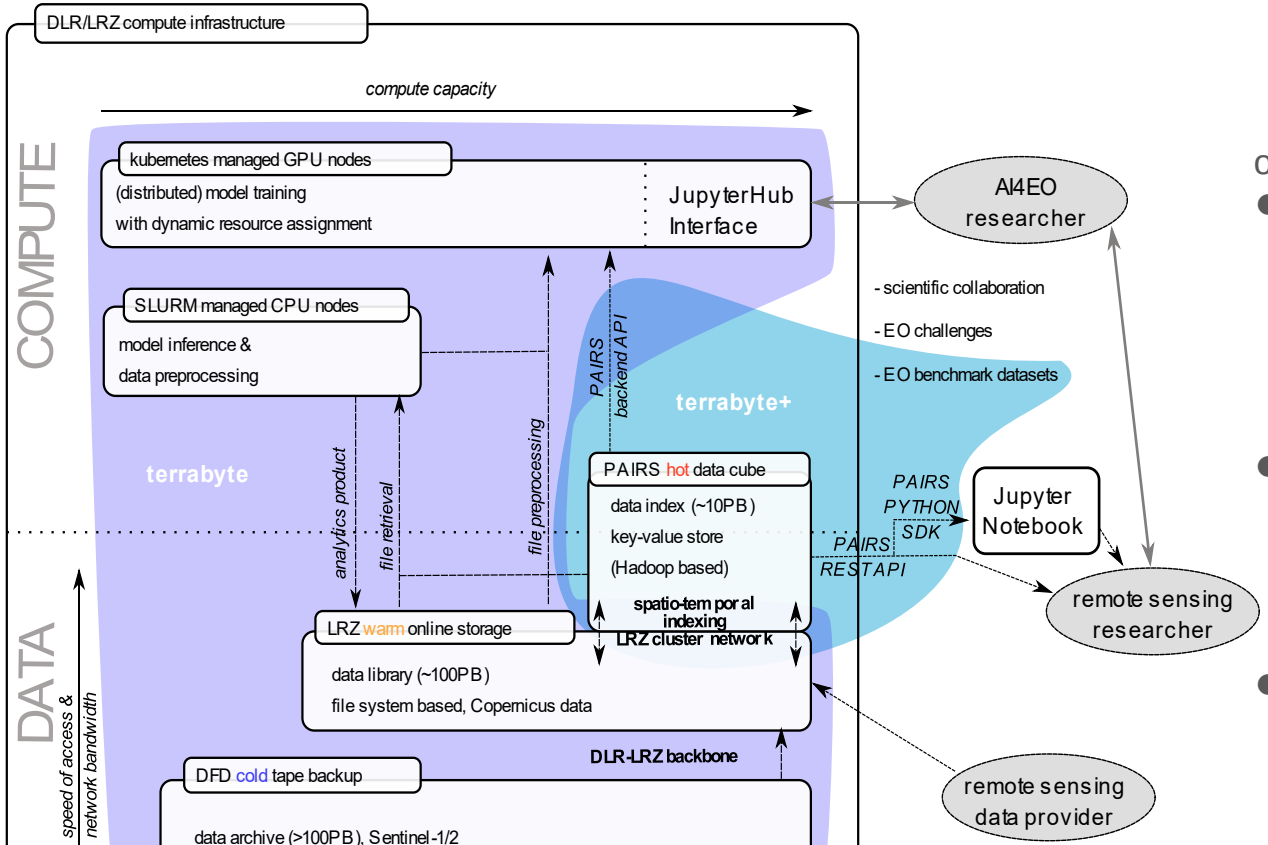



* https://github.com/NVIDIA/enroot

# High-Level Data Processing System Architecture: details



# High-Level Data Processing System Architecture: *hot* data cube extension



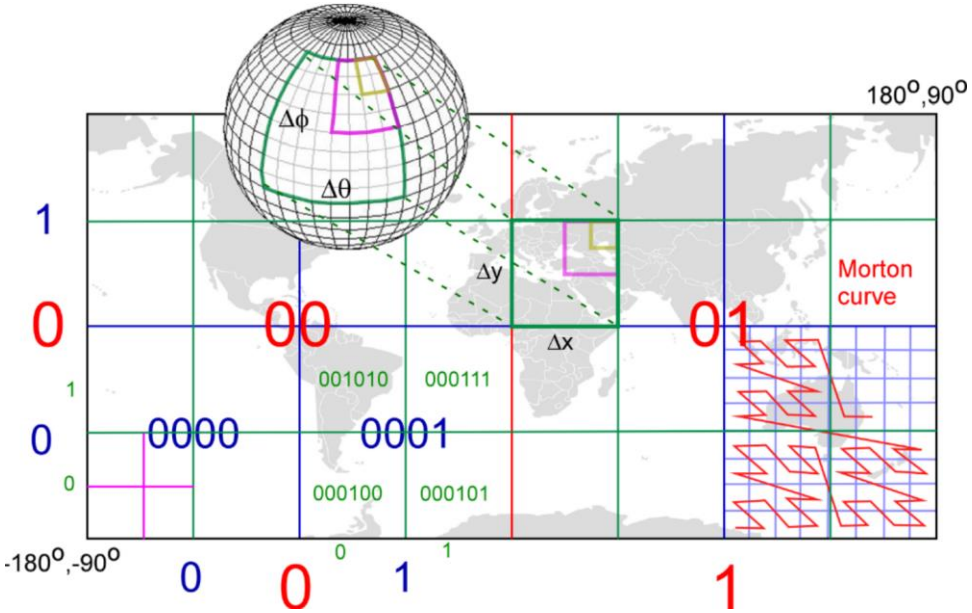

objectives:

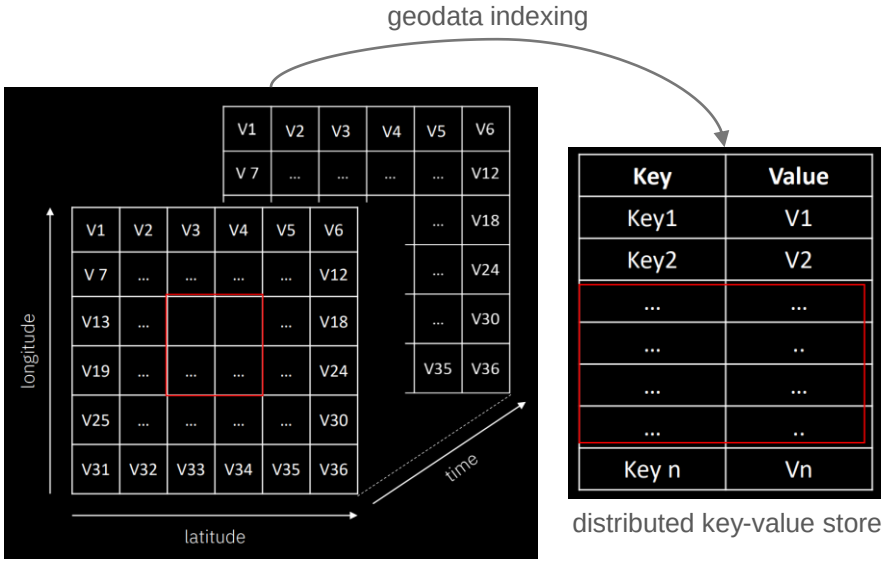
- accelerate academic research **serving up-to-date, retrievable open geospatial benchmark datasets** for machine learning
- seamless collaboration with international researchers through **browser-based code-development interfaces**
- host **geospatial data cubes for AI community challenges** in Earth observation

IBM PAIRS platform: https://doi.org/10.1007/978-3-030-55462-0_1 https://pairs.res.ibm.com/tutorial

# Open-Source Software Stack *hot* data cube extension

# PAIRS*,**: *Hot* Data Cube design



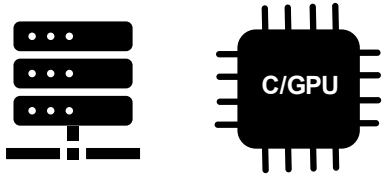

+ **data cube**: unified, nested spatial grid
+ **key-value store**: scalable even for global and sparse datasets





* architecture: https://doi.org/10.1109/BigData.2015.7363884
** sample applications: https://doi.org/10.1109/BigData.2016.7840910

# Compute & Storage Infrastructure




- terrabyte system hosted at Leibniz Supercomputing Centre (LRZ) co-located with SuperMUC NG*
- cluster nodes partitioned into CPU and GPU machines attached to Infiniband (200Gb/s) storage system (~40PB HDD & ~1PB SSD) including S3 RESTful API access
- core datasets: global Sentinel and Landsat product data since mission start, available at terrabyte and near-line at *EOC National Satellite Data Archive (D-SDA)*
- additional cloud storage machines (storage: NL-SAS HDD & NVMe SSD)
- *hot* data cube with dedicated nodes enabling fine-grained, custom spatio-temporal indexing and cross-dataset queries of geo-information





* https://doku.lrz.de/display/PUBLIC/SuperMUC-NG

# System Interoperability: *Call to the Open Big-Geo Data Community*

OGC®
Making location count.
https://ogcapi.ogc.org

1. existing OpenGeospatial Consortium standards
   - OpenGeospatialConsortium WebProcessingService: https://www.ogc.org/standards/wps
   - OpenGeospatialConsortium DataAccess(and)ProcessingAPI: https://www.ogc.org/blog/4665
   - SpatialTemporalAssetCatalogue API: https://github.com/radiantearth/stac-api-spec
   - SentinelHub processing API: https://docs.sentinel-hub.com/api/latest/reference/#tag/process

2. Cross-platform, cross-datalayer spatio-temporal data cube join RESTful API
   - follow-up on OGC initiative „Towards Data Cube Interoperability“: https://www.ogc.org/projects/initiatives/gdc
   - **objective**: pursue a user-centric, pragmatic approach to define an (extendable) API for data cube filtering and joining

# terrabyte‘s Implementation Strategy


- **focus of terrabyte** initiative is collaborative science projects and experiments with scalable **Big Geo-Data processing** solutions **in hybrid HPC-cloud** environment
- **software stack** is designed around the **open-source paradigm**
- opportunity to **collaborate with international partners** to establish cross-platform **WebAPI interface standards through OGC**
- LRZ implements information security management system according to standards ISO/IEC 20000-1 and ISO/IEC 27001
- **cost efficiency through** scalable GPFS file system coupled to an extendable cluster of C/GPU nodes added in multiple stages of project; terrabyte targets **consolidation of DLR‘s in-house compute resources for AI workloads**

# Supporting Open-Source Science

*NASA's Science Mission Directorate defines open-source science as a **collaborative culture** enabled by technology that empowers the **open sharing** of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding.*

- ***collaborative culture***: close interaction with academia wrt. the development of state-of-the-art AI methodologies in Earth observation on international level: https://ai4eo.de/partners
- ***open sharing***: integration of scientific projects on terrabyte through national and German funding following the FAIR principles*
- *technological challenges*:
  - rapidly evolving plurality of open-source technologies with limited software life cycles
  - balance system security and open-source, open-access
- *accellerator to „open"*: lively international consortia embracing all relevant players of the geospatial domain to define technically solid, vendor-neutral industry standards



* https://www.go-fair.org/fair-principles

# THX 2 …

- Marc Jochemich (DLR Washington, DC)
- Jonas Eberle & Maximilian Schwinger (German Remote Sensing Data Center @ DLR)
- Nicolay Hammer, Johannes Albert-von der Gönna, and Juan Durillo Barrionuevo (LRZ Big Data & AI team)
- Xiaoxiang Zhu („AI for Earth Observation“ lab @ TUM & DLR)
- Hendrik Hamann & Johannes Schmude (Physical Analytics @ IBM Research)

… for support, collaboration & discussions