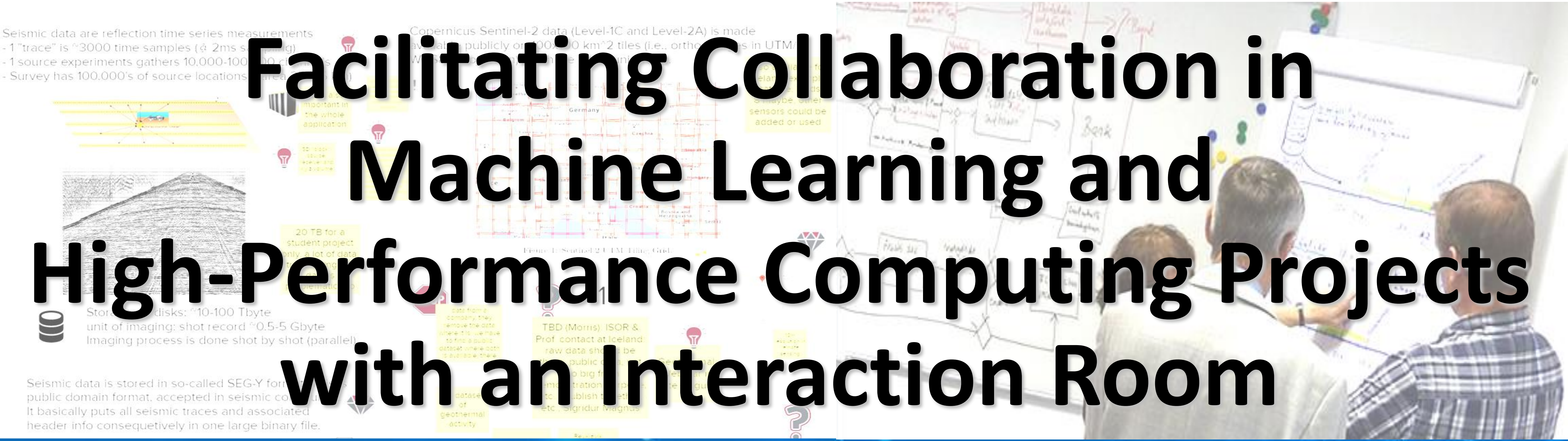
UNIVERSITY OF ICELAND
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

RAISE
Center of Excellence

The CoE RAISE project receives funding from the European Union's Horizon 2020 – Research and Innovation Framework Programme H2020-INFRAEDI-2019-1 under grant agreement no. 951733


Seismic data are reflection time series measurements
- 1 "trace" is ~3000 time samples (@ 2ms s
- 1 source experiments gathers 10.000-100
- Survey has 100.000's of source locations
Copernicus Sentinel-2 data (Level-1C and Level-2A) is made
sensors could be added or used
20 TB for a student project
Stor disks: ~10-100 Tbyte
unit of imaging: shot record ~0.5-5 Gbyte
Imaging process is done shot by shot (parallel)
Seismic data is stored in so-called SEG-Y for
public domain format, accepted in seismic co
It basically puts all seismic traces and associated
header info consequetively in one large binary file.
TBD (Morris): ISOR &
Prof. contact at Iceland
Germany

Facilitating Collaboration in Machine Learning and High-Performance Computing Projects with an Interaction Room

Matthias Book, Morris Riedel*, Helmut Neukirchen, Ernir Erlingsson
University of Iceland
*Double affiliation with Jülich Supercomputing Centre, Germany

https://uni.hi.is/helmut/

The Nature of Software Development: Social learning

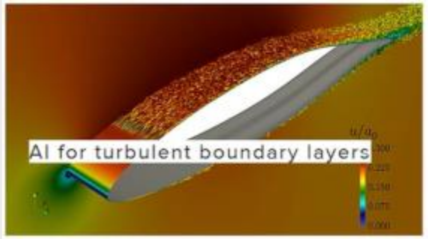
“Because software is **embodied knowledge**, and that knowledge is initially **dispersed**, **tacit**, **latent**, and **incomplete**, software development is a **social learning** process.”

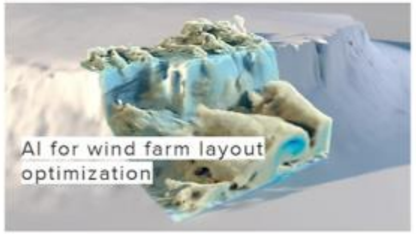
Howard Baetjer, Jr.: Software as Capital. IEEE Computer Society Press, 1998

Social Learning: even more complicated when collaborative and interdisciplinary

- Project "Research on AI- and Simulation-Based Engineering at Exascale":
  - Engineering problems to be solved by simulation & AI on (pre-)exascale HPC systems.
  - Experts from different fields & partners: Engineering, HPC, AI/ML & SE to work together.

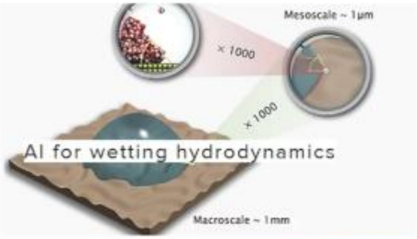
    ⇒Different stakeholders with different knowledge & different implicit assumption.


AI for turbulent boundary layers


AI for wind farm layout optimization

AI for data-driven models in reacting flows

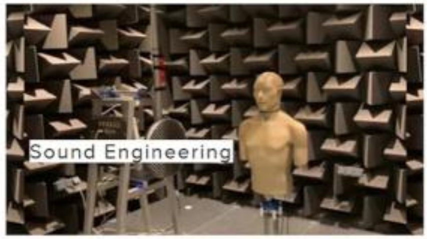
Smart models for next-generation aircraft engine design


AI for wetting hydrodynamics

Event reconstruction and classification at the CERN HL-LHC

Seismic imaging with remote sensing for energy applications

Defect-free metal additive manufacturing


Sound Engineering

The Interaction Room (IR)

Interaction Room
communicate • understand • decide

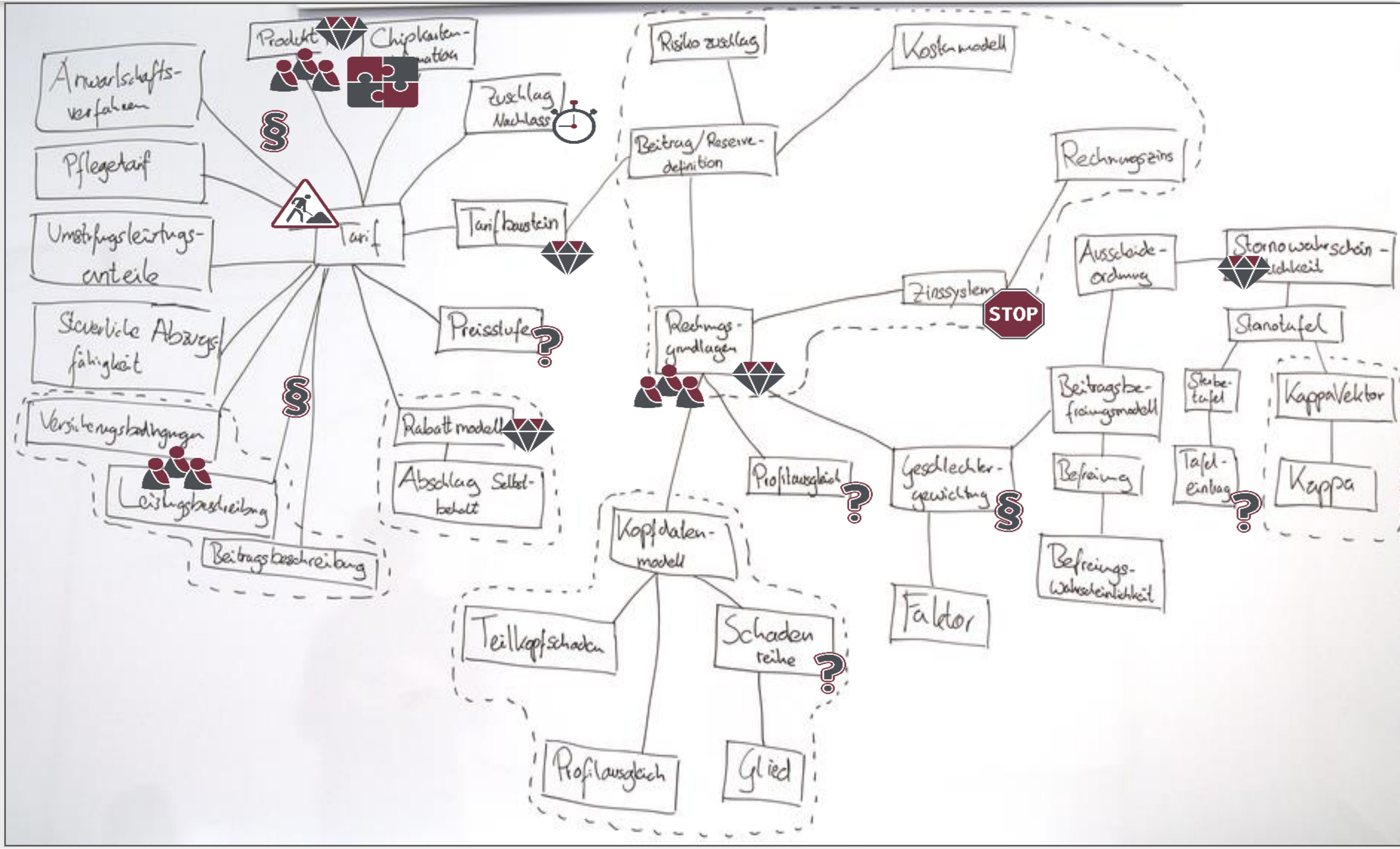
- Successfully used in business information system development:
  - Domain experts and software experts need to **collaborate** and **understand each other**.

    Book, Grapenthin, Gruhn: “Seeing the forest and the trees: focusing team interaction on value and effort drivers”, Proc. ACM SIGSOFT 20th Intl. Symp. on Foundations of Software Engineering, 2012.

- The Interaction Room is
  - a **dedicated room** (physical, but virtual worked well) for the project team
    - where experts from different domains feel at home
  - with **large whiteboards** (analog or digital) on the walls with canvases focusing on different aspects
    - but without a classic conference table
  - to **visualize and discuss** key project aspects informally
    - instead of going over tedious documents / without the goal of creating a spec (might be done subsequently)
  - to **identify implicit assumptions**, **uncertainties**, **risks**, etc.
    - by adding annotations to the canvases.

Example: Object canvas for an insurance system with annotations to capture implicit assumptions/knowledge

- **Annotations added:**
  - Value,
  - Complexity,
  - Usability,
  - Uncertainty,
  - Legal issues,
  - etc.



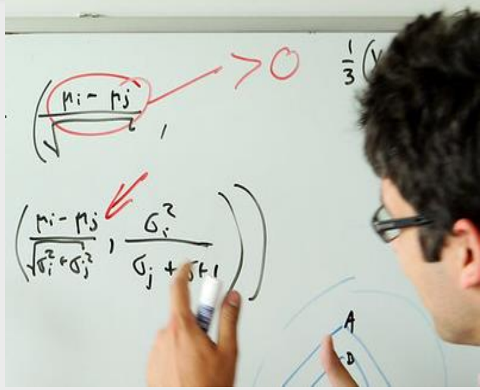
Interaction Room: A Pragmatic Approach to Conceptualizing Software

- **Informal**, **high-level sketches** of software models.
  - sacrifice formality (no modeling language), consistency, completeness,
  - in favor of pragmatism and interdisciplinary understanding.
- **Informal sketches** + **formalized annotations** serve as **catalysts for the identification**, **understanding, and discussion** of the most **critical aspects**.
  - Interdisciplinary communication,
  - Identification of goals, complexity drivers, risks & uncertainties, trade-offs.
  - **Shift** attention
    - from what is obviously visible
    - **to what is invisible**, what is implied, what is unknown (=what makes or breaks a project).
- Proven for business information systems.

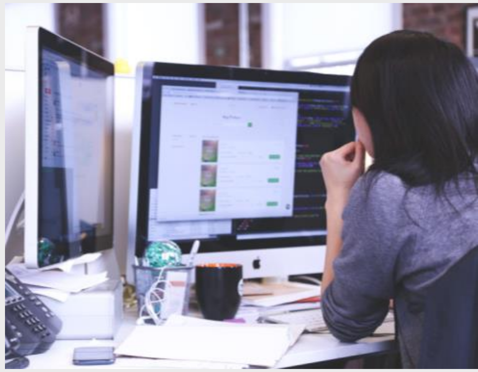
Interaction Room for ML & HPC?

- Goal: **Facilitate collaboration** of experts from
  - the **natural science/engineering domain,**
  - the **HPC domain,**
  - the **AI/ML domain,**
  - the **computer science/software engineering domain**.

- **Adapt proven Interaction Room concepts**:
  - But: **canvases** and **annotations** needed that are **specific to HPC/AI/ML needs**.



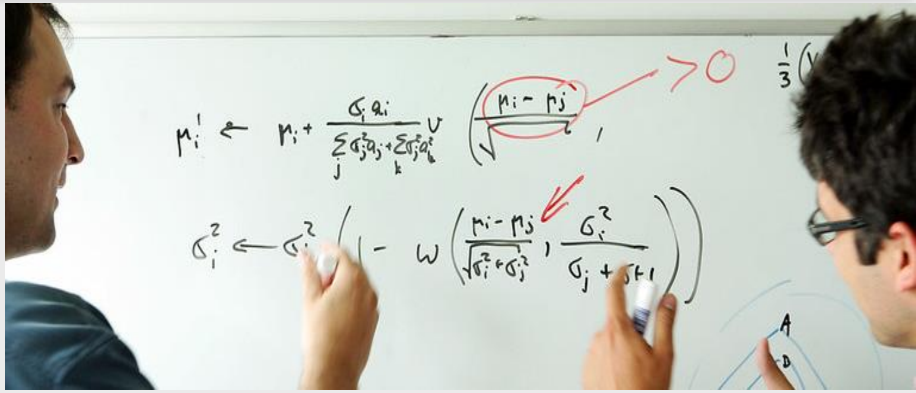
CC BY 2.0 Steve Wilson https://www.flickr.com/photos/125303894@N06/14387366382



Public domain https://www.piqsels.com

Interaction Room for ML & HPC!

- ML/AI/HPC-specific canvases that address:
  - Crucial **interdisciplinary discussion points**,
  - **Typical HPC/ML/AI project phases**.



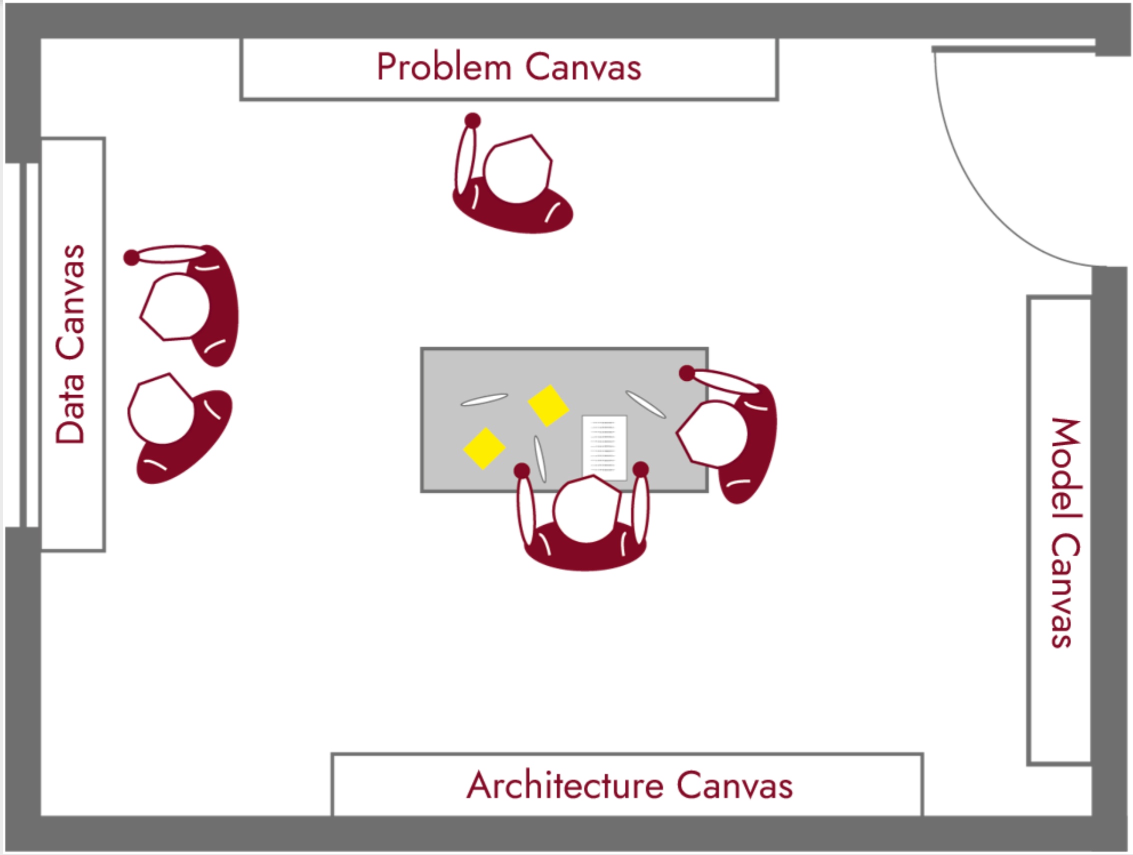
CC BY 2.0 Steve Wilson https://www.flickr.com/photos/125303894@N06/14387366382

- **Different project types** needing **different canvas types**, e.g.:
  - **Simulation sciences** (“classic” HPC),
    - Canvases proposed in earlier position paper, but never tried in practice.
  - **ML & HPC** (“High-Performance Data Analytics”)
    - Covered in the remainder as applied in the CoE RAISE project.

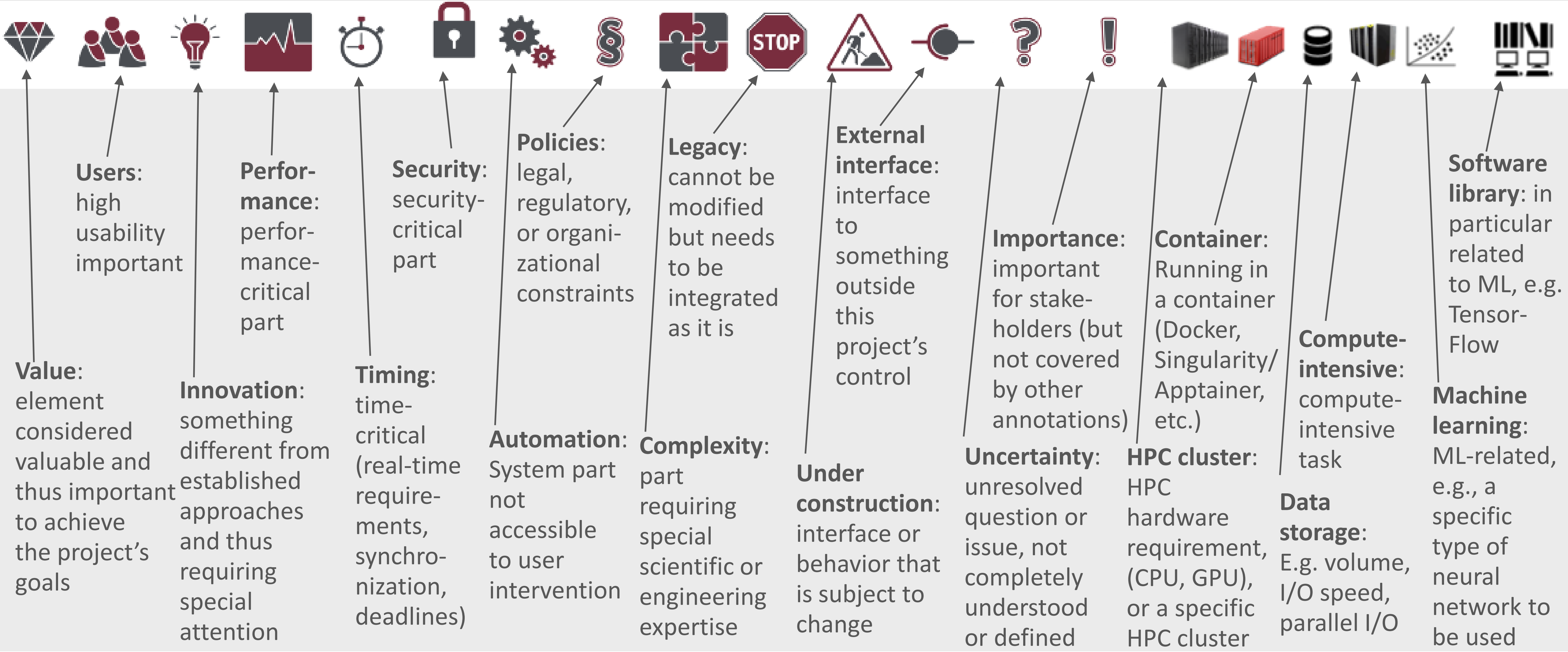
Book, Riedel, Neukirchen, Goetz: *Facilitating Collaboration in High-Performance Computing Projects with an Interaction Room.* 4th ACM SIGPLAN International Workshop on Software Engineering for Parallel Systems (SEPS 2017)

Interaction Room: Canvases for ML & HPC Projects

- **Problem canvas:**
  - Goal and scope of research question (=the scientific domain) to be solved.
- **Data canvas:**
  - Understand the data to be processed (what data is available/needed, formats, size, access, etc.).
- **Model canvas:**
  - ML models to be used (clustering, classification, deep learning, parameters, integrate with simulations, etc.).
- **Architecture canvas:**
  - ML libraries, numerical solvers, HPC hardware (CPU/GPU)/specific clusters.



Interaction Room: Annotations for ML & HPC Projects



Value: element considered valuable and thus important to achieve the project’s goals

Users: high usability important

Innovation: something different from established approaches and thus requiring special attention

Performance: performance-critical part

Timing: time-critical (real-time requirements, synchronization, deadlines)

Security: security-critical part

Automation: System part not accessible to user intervention

Policies: legal, regulatory, or organizational constraints

Complexity: part requiring special scientific or engineering expertise

Legacy: cannot be modified but needs to be integrated as it is

Under construction: interface or behavior that is subject to change

External interface: interface to something outside this project’s control

Uncertainty: unresolved question or issue, not completely understood or defined yet

Importance: important for stakeholders (but not covered by other annotations)

HPC cluster: HPC hardware requirement, (CPU, GPU), or a specific HPC cluster

Container: Running in a container (Docker, Singularity/ Apptainer, etc.)

Data storage: E.g. volume, I/O speed, parallel I/O

Compute-intensive: compute-intensive task

Machine learning: ML-related, e.g., a specific type of neural network to be used

Software library: in particular related to ML, e.g. TensorFlow

Interaction Room: Process

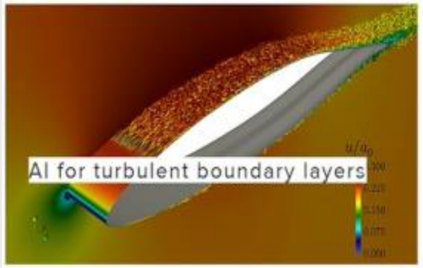
- **Moderator** facilitates discussion between stakeholders from different domains.
- Start with **Problem canvas**
  - Moderator and/or stakeholders fill canvas with informal drawings.
  - **Annotations added** by stakeholders:
    - **ad-hoc when they come up**,
    - in an **extra annotation round** without discussion (to encourage shy people to raise issues: “add at least one”).
      - But: Afterwards discussed and more info added or removed (if not agreed).
- Typically followed by **Data canvas**,
- Then: **Model canvas**,
- Finally, **Architecture canvas**.
- Not necessary one-time sequential flow, but **iterative refinement of canvas contents**.
  - Current “primary” canvas if in focus, but may add to any “secondary” as well.
- Refine in further IR sessions, e.g. in later project stages.

Case Studies

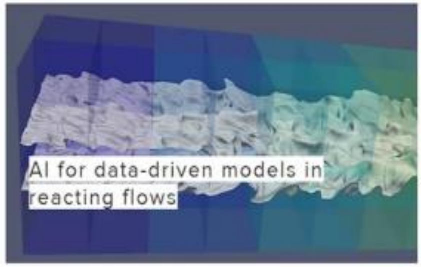
- Virtual Interaction Room:
  - **Collaborative remote digital whiteboards** (MURAL boards) for the canvases,
    - Allowed to fill canvas and add annotations in parallel in a highly collaborative style (better than physical IR).
  - **Video conferencing** (for video and audio).
- **9 projects** involving HPC, ML & simulation.
- Ca. 4-8 participants per project:
  - **Moderator,**
  - **Engineers,**
  - **ML experts,**
  - **HPC experts,**
  - **Software Engineers.**

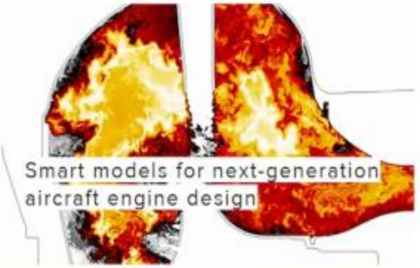
Some participants had multiple roles (e.g. moderator and software engineer at same time).

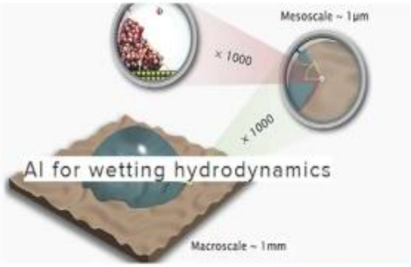
- **Initial session + later refinement sessions.**
  - Later sessions by project without moderator: varying extent of refinement in each project.


AI for turbulent boundary layers

AI for wind farm layout optimization


AI for data-driven models in reacting flows


Smart models for next-generation aircraft engine design

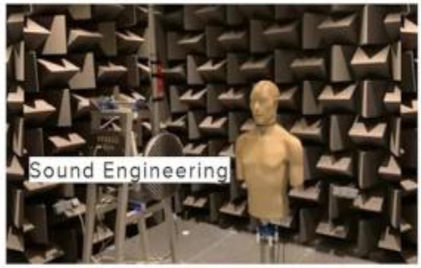


AI for wetting hydrodynamics

Event reconstruction and classification at the CERN HL-LHC

Seismic imaging with remote sensing for energy applications

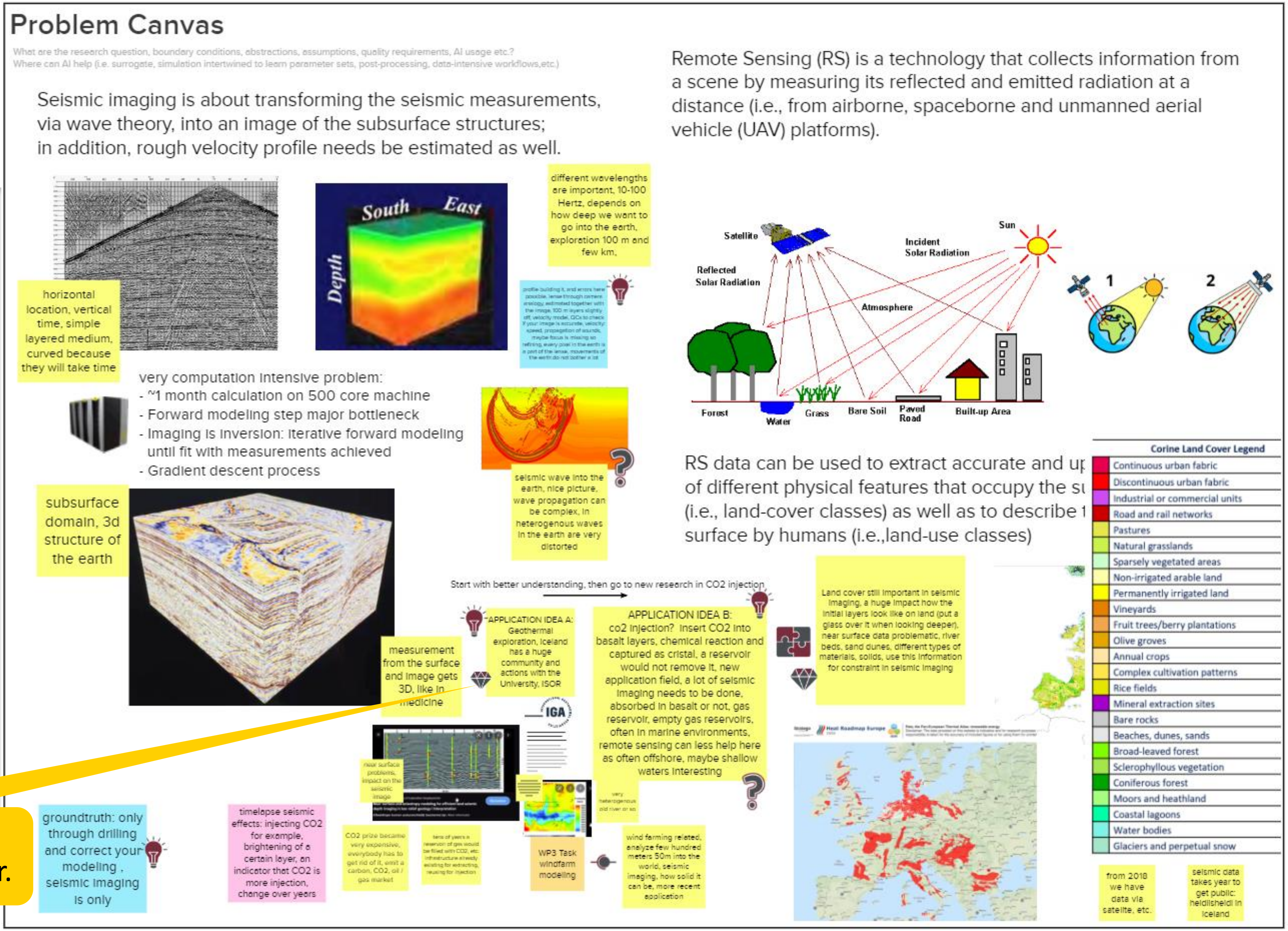
Defect-free metal additive manufacturing


Sound Engineering

Case study Seismic Imaging: Problem Canvas

- "Geophysical tomography":
  - To find basalt layers to inject CO_2.
  - Combined with remote sensing.
- Annotations often used together with sticky notes containing further information.

"Value" and "Innovation" annotation often used together.



Problem Canvas

What are the research question, boundary conditions, abstractions, assumptions, quality requirements, AI usage etc.?
Where can AI help (i.e. surrogate, simulation intertwined to learn parameter sets, post-processing, data-intensive workflows,etc.)

Seismic imaging is about transforming the seismic measurements, via wave theory, into an image of the subsurface structures; in addition, rough velocity profile needs be estimated as well.

horizontal location, vertical time, simple layered medium, curved because they will take time

different wavelengths are important, 10-100 Hertz, depends on how deep we want to go into the earth, exploration 100 m and few km,

profile building it, and errors here possible, lense through camera analogy, estimated together with the image, 100 m layers slightly off, velocity model, QCs to check if your image is accurate, velocity: speed, propagation of sounds, maybe focus is missing so refining, every pixel in the earth is a part of the lense, movements of the earth do not bother a lot

very computation intensive problem:
- ~1 month calculation on 500 core machine
- Forward modeling step major bottleneck
- Imaging is inversion: iterative forward modeling until fit with measurements achieved
- Gradient descent process

seismic wave into the earth, nice picture, wave propagation can be complex, in heterogenous waves in the earth are very distorted

subsurface domain, 3d structure of the earth

measurement from the surface and image gets 3D, like in medicine

Start with better understanding, then go to new research in CO2 injection

APPLICATION IDEA A: Geothermal exploration, Iceland has a huge community and actions with the University, ISOR

APPLICATION IDEA B: co2 injection? insert CO2 into basalt layers, chemical reaction and captured as cristal, a reservoir would not remove it, new application field, a lot of seismic imaging needs to be done, absorbed in basalt or not, gas reservoir, empty gas reservoirs, often in marine environments, remote sensing can less help here as often offshore, maybe shallow waters interesting

near surface problems, impact on the seismic image

very heterogenous old river or so

groundtruth: only through drilling and correct your modeling , seismic imaging is only

timelapse seismic effects: injecting CO2 for example, brightening of a certain layer, an indicator that CO2 is more injection, change over years

CO2 prize became very expensive, everybody has to get rid of it, emit a carbon, CO2, oil / gas market

tens of years a reservoir of gas would be filled with CO2, etc. infrastructure already existing for extracting, reusing for injection

WP3 Task windfarm modelling

wind farming related, analyze few hundred meters 50m into the world, seismic imaging, how solid it can be, more recent application

Remote Sensing (RS) is a technology that collects information from a scene by measuring its reflected and emitted radiation at a distance (i.e., from airborne, spaceborne and unmanned aerial vehicle (UAV) platforms).

RS data can be used to extract accurate and up
of different physical features that occupy the su
(i.e., land-cover classes) as well as to describe t
surface by humans (i.e.,land-use classes)

Corine Land Cover Legend
Continuous urban fabric
Discontinuous urban fabric
Industrial or commercial units
Road and rail networks
Pastures
Natural grasslands
Sparsely vegetated areas
Non-irrigated arable land
Permanently irrigated land
Vineyards
Fruit trees/berry plantations
Olive groves
Annual crops
Complex cultivation patterns
Rice fields
Mineral extraction sites
Bare rocks
Beaches, dunes, sands
Broad-leaved forest
Sclerophyllous vegetation
Coniferous forest
Moors and heathland
Coastal lagoons
Water bodies
Glaciers and perpetual snow

Land cover still important in seismic imaging, a huge impact how the initial layers look like on land (put a glass over it when looking deeper), near surface data problematic, river beds, sand dunes, different types of materials, solids, use this information for constraint in seismic imaging

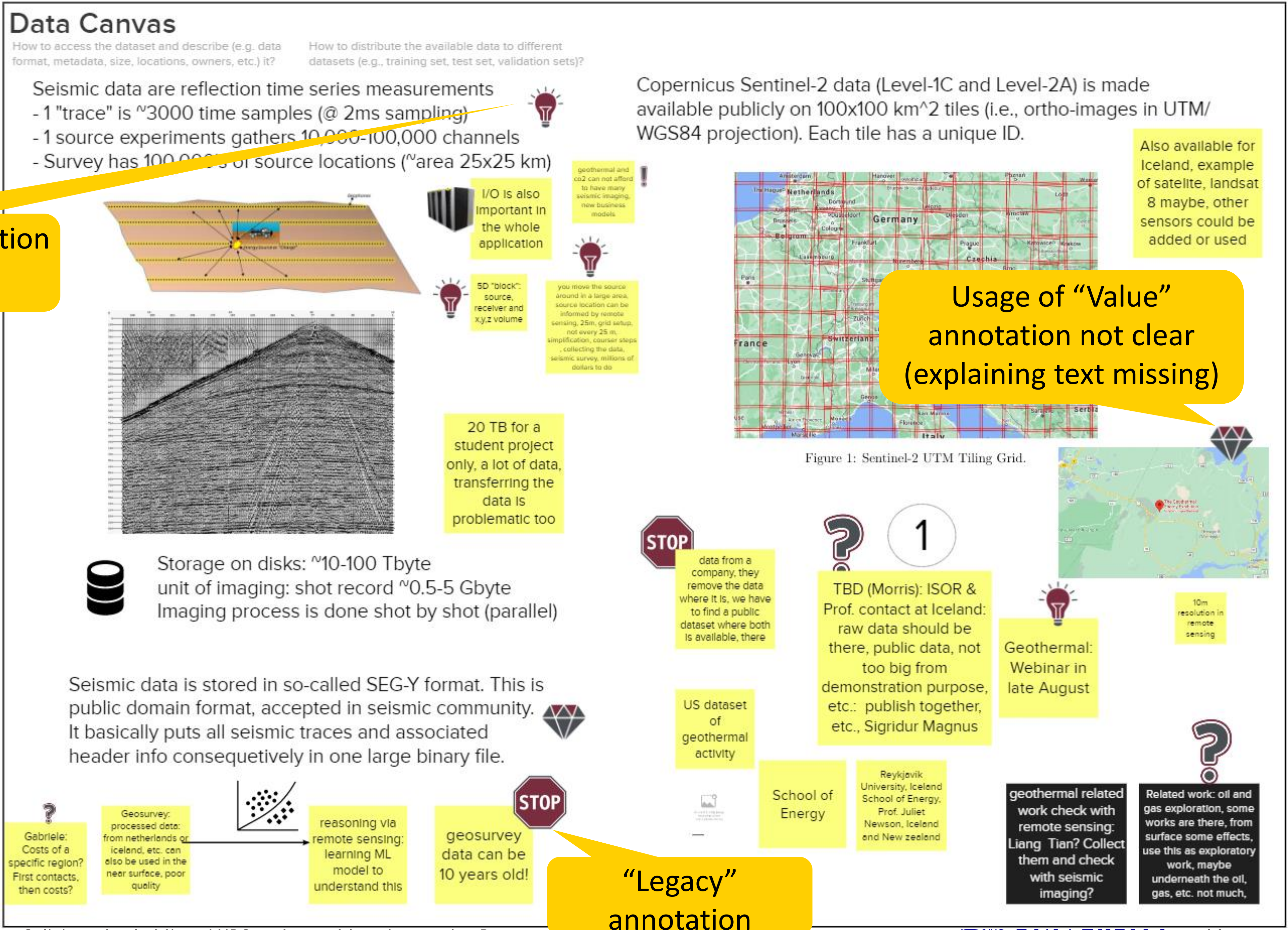
from 2018 we have data via satelite, etc.

seismic data takes year to get public: heildilsheildl in Iceland

Case study Seismic Imaging: Data Canvas

Usage of "Innovation" annotation not always clear

- Better guidance wrt. annotation meaning needed.
- Maybe also lack of moderation.



Data Canvas

How to access the dataset and describe (e.g. data format, metadata, size, locations, owners, etc.) it?

How to distribute the available data to different datasets (e.g., training set, test set, validation sets)?

Seismic data are reflection time series measurements
- 1 "trace" is ~3000 time samples (@ 2ms sampling)
- 1 source experiments gathers 10,000-100,000 channels
- Survey has 100,000's of source locations (~area 25x25 km)

I/O is also important in the whole application

geothermal and co2 can not afford to have many seismic imaging, new business models

5D "block": source, receiver and x,y,z volume

you move the source around in a large area, source location can be informed by remote sensing, 25m, grid setup, not every 25 m, simplification, courser steps , collecting the data, seismic survey, millions of dollars to do

20 TB for a student project only, a lot of data, transferring the data is problematic too

Storage on disks: ~10-100 Tbyte
unit of imaging: shot record ~0.5-5 Gbyte
Imaging process is done shot by shot (parallel)

Seismic data is stored in so-called SEG-Y format. This is public domain format, accepted in seismic community. It basically puts all seismic traces and associated header info consequetively in one large binary file.

Gabriele: Costs of a specific region? First contacts, then costs?

Geosurvey: processed data: from netherlands or iceland, etc. can also be used in the near surface, poor quality

reasoning via remote sensing: learning ML model to understand this

geosurvey data can be 10 years old!

Copernicus Sentinel-2 data (Level-1C and Level-2A) is made available publicly on 100x100 km^2 tiles (i.e., ortho-images in UTM/WGS84 projection). Each tile has a unique ID.

Also available for Iceland, example of satelite, landsat 8 maybe, other sensors could be added or used

Figure 1: Sentinel-2 UTM Tiling Grid.

data from a company, they remove the data where it is, we have to find a public dataset where both is available, there

1

TBD (Morris): ISOR & Prof. contact at Iceland: raw data should be there, public data, not too big from demonstration purpose, etc.: publish together, etc., Sigridur Magnus

Geothermal: Webinar in late August

10m resolution in remote sensing

US dataset of geothermal activity

School of Energy

Reykjavik University, Iceland School of Energy, Prof. Juliet Newson, Iceland and New zealand

geothermal related work check with remote sensing: Liang Tian? Collect them and check with seismic imaging?

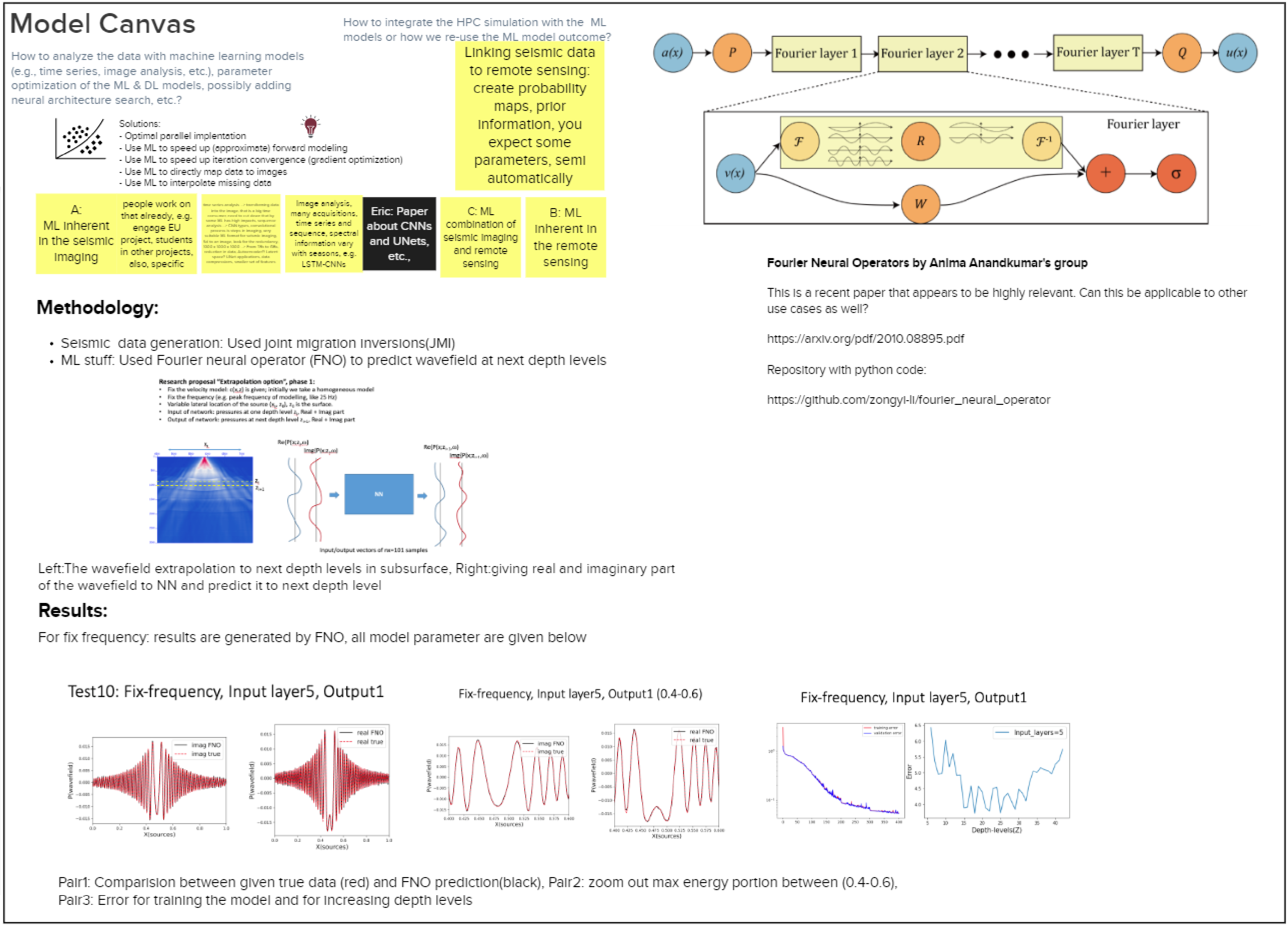
Related work: oil and gas exploration, some works are there, from surface some effects, use this as exploratory work, maybe underneath the oil, gas, etc. not much,

Usage of "Value" annotation not clear (explaining text missing)

"Legacy" annotation

Case study Seismic Imaging: Model Canvas

- Only one annotation used.
  - **Maybe need more annotations specific to ML models?**
- Nevertheless, valuable information provided by informal text.
  - ⇒Canvas is useful.



Model Canvas

How to analyze the data with machine learning models (e.g., time series, image analysis, etc.), parameter optimization of the ML & DL models, possibly adding neural architecture search, etc.?

How to integrate the HPC simulation with the ML models or how we re-use the ML model outcome?

Solutions:
- Optimal parallel implentation
- Use ML to speed up (approximate) forward modeling
- Use ML to speed up iteration convergence (gradient optimization)
- Use ML to directly map data to images
- Use ML to interpolate missing data

Linking seismic data to remote sensing: create probability maps, prior information, you expect some parameters, semi automatically

A: ML inherent in the seismic imaging

people work on that already, e.g. engage EU project, students in other projects, also, specific

time series analysis --> transforming data into the image, that is a big time consumer, need to cut down that by some ML has high impacts, sequence analysis --> CNN types, convolutional process is steps in imaging, very suitable ML format for seismic imaging, ful to an image, look for the redundancy 1000 x 1000 x 1000 --> From TBs to GBs, reduction in data. Autoencoder? Latent space? Uber applications, data compression, smaller set of features.

Image analysis, many acquisitions, time series and sequence, spectral information vary with seasons, e.g. LSTM-CNNs

Eric: Paper about CNNs and UNets, etc.,

C: ML combination of seismic imaging and remote sensing

B: ML inherent in the remote sensing

Methodology:

- Seismic data generation: Used joint migration inversions(JMI)
- ML stuff: Used Fourier neural operator (FNO) to predict wavefield at next depth levels

Research proposal "Extrapolation option", phase 1:
- Fix the velocity model: $c(x,z)$ is given; initially we take a homogeneous model
- Fix the frequency (e.g. peak frequency of modelling, like 25 Hz)
- Variable lateral location of the source (x_s, z_0), z_0 is the surface.
- Input of network: pressures at one depth level z_i, Real + Imag part
- Output of network: pressures at next depth level z_{i+1}, Real + Imag part

Left:The wavefield extrapolation to next depth levels in subsurface, Right:giving real and imaginary part of the wavefield to NN and predict it to next depth level

Results:

For fix frequency: results are generated by FNO, all model parameter are given below

Test10: Fix-frequency, Input layer5, Output1
Fix-frequency, Input layer5, Output1 (0.4-0.6)
Fix-frequency, Input layer5, Output1
imag FNO
imag true
real FNO
real true
P(wavefield)
X(sources)
training error
validation error
Input_layers=5
Error
Depth-levels(Z)

Pair1: Comparision between given true data (red) and FNO prediction(black), Pair2: zoom out max energy portion between (0.4-0.6), Pair3: Error for training the model and for increasing depth levels

Fourier Neural Operators by Anima Anandkumar's group

This is a recent paper that appears to be highly relevant. Can this be applicable to other use cases as well?

https://arxiv.org/pdf/2010.08895.pdf

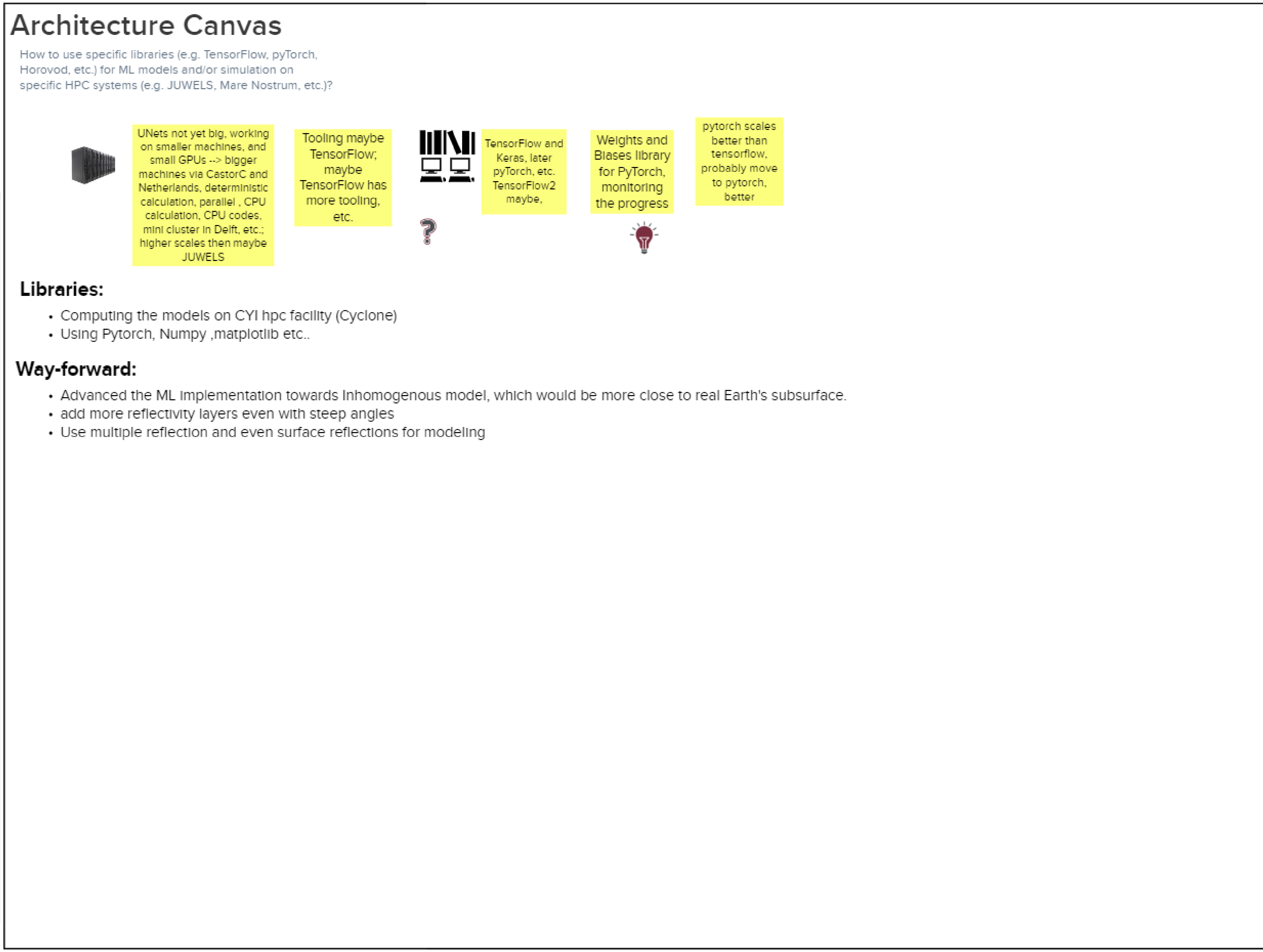
Repository with python code:

https://github.com/zongyi-li/fourier_neural_operator

Case study Seismic Imaging: Architecture Canvas

- Architecture canvases typically not very crowded:
  - **HPC-related Architecture might be straightforward**?
  - **More guidance might be needed from moderators?**
  - Nevertheless, valuable information provided by informal text.

    ⇒Canvas is useful.



Architecture Canvas

How to use specific libraries (e.g. TensorFlow, pyTorch, Horovod, etc.) for ML models and/or simulation on specific HPC systems (e.g. JUWELS, Mare Nostrum, etc.)?

UNets not yet big, working on smaller machines, and small GPUs --> bigger machines via CastorC and Netherlands, deterministic calculation, parallel , CPU calculation, CPU codes, mini cluster in Delft, etc.; higher scales then maybe JUWELS

Tooling maybe TensorFlow; maybe TensorFlow has more tooling, etc.

TensorFlow and Keras, later pyTorch, etc. TensorFlow2 maybe,

Weights and Biases library for PyTorch, monitoring the progress

pytorch scales better than tensorflow, probably move to pytorch, better

Libraries:

- Computing the models on CYI hpc facility (Cyclone)
- Using Pytorch, Numpy ,matplotlib etc..

Way-forward:

- Advanced the ML implementation towards Inhomogenous model, which would be more close to real Earth's subsurface.
- add more reflectivity layers even with steep angles
- Use multiple reflection and even surface reflections for modeling

Discussion & Outlook

- No empiric evaluation – but in our experience: IR for HPC & ML leads to
  - fostering **better communication of stakeholders**,
  - more **explicit externalization of discussion outcomes**.
  - Other projects where domain experts and ML experts had to work together without using an Interaction Room, just saw some unstructured discussions – one even failed.
- **Extent of using annotations varied** in the different use cases:
  - **Moderator should always attend** and needs to **encourage using annotations**.
    - (We had no explicit annotation round as virtual boards allowed adding annotations any time.)
  - Important to **provide** as drag-n-drop not just icons, but also **text explaining meaning**.
  - **Better icons needed** for annotations beyond those copied from business systems:
    - e.g. HPC cluster vs. Compute-intensive: vs.
- **More annotations specific to Model canvas** needed, less from business IR.
  - Evaluation needed what ML model issues are is relevant.

Helmut Neukirchen et al.: HPC Systems Engineering in ML and HPC projects with an Interaction Room

Summary

- Interaction Room **facilitates collaboration of the involved stakeholders.**
  - More and more broad collaborative and multi-institutional projects.

Bernholdt et al. “A survey on sustainable software ecosystems to support experimental and observational science at Oak Ridge National Laboratory”, Int. Conf. Computational Science (ICCS) 2022,

- **Informal sketches** + **formalized annotations:**
  - **catalyst for the identification**, **understanding and discussion** of the **critical aspects**.
- **Canvases** and **annotations** need to be **specific to the domain**, so far:
  - Business information systems,
  - HPC simulation sciences,
  - HPC ML projects.
- Also an aid to improve **software sustainability**:
  - Archiving the canvases that evolved throughout the project:
    - **Capture knowledge and assumptions that are typically not documented anywhere else**.
    - Important in scientific projects with high staff turnover (temporary contracts, PhD students).

drive. enable. innovate.

European Center of Excellence in
Research on AI- and Simulation-Based Engineering at Exascale
(CoE RAISE)
https://www.coe-raise.eu

RAISE
Center of Excellence

The CoE RAISE project receives funding from
the European Union’s Horizon 2020 –
Research and Innovation Framework Programme
H2020-INFRAEDI-2019-1 under grant agreement no. 951733