Artificial Intelligence
In Data Analysis

ESRF | The European Synchrotron

Vincent Favre-Nicolin
Algorithms & scientific Data Analysis

BIG DATA

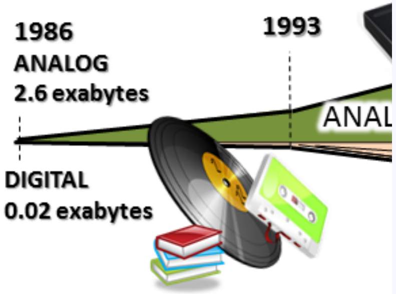
Global Information Storage Capacity

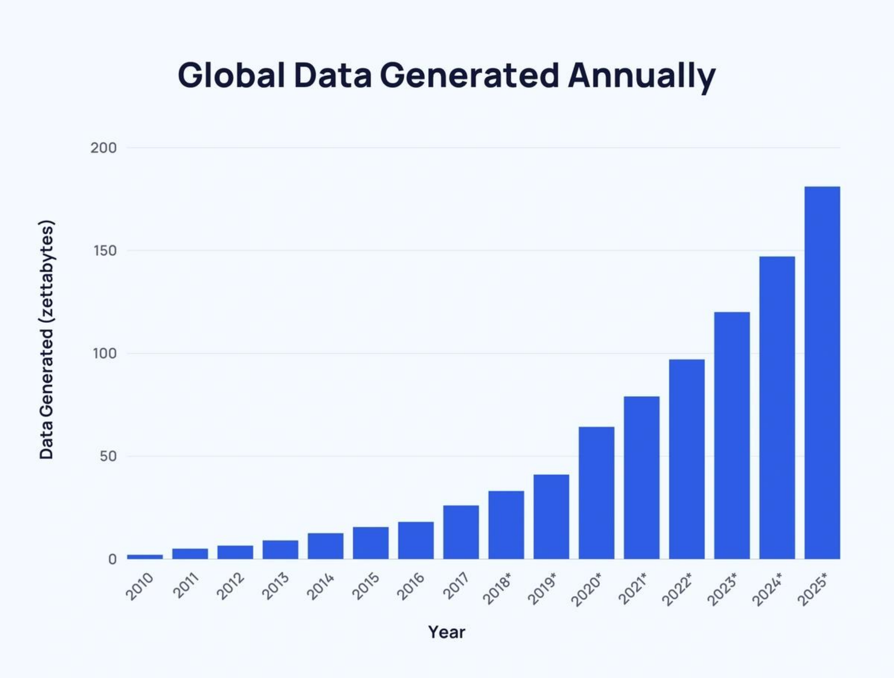
in optimally compressed bytes

2007 **ANALOG**

19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % **ANALOG**
- Portable media, flash drives: 2 %





- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

% digital:

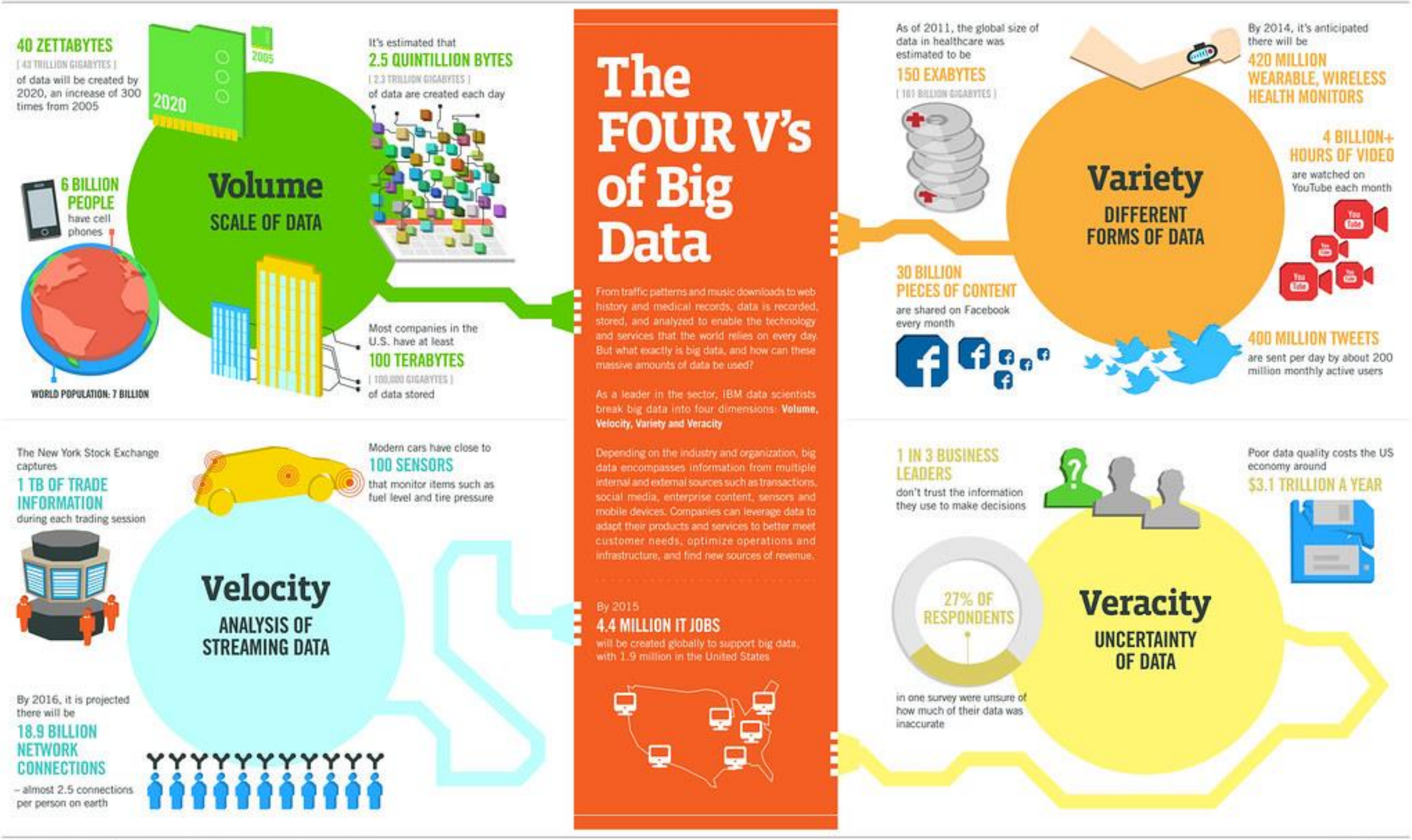
1 % 3 % 25 % 94 %

DIGITAL

280 exabytes

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

BIG DATA



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

BIG DATA ?

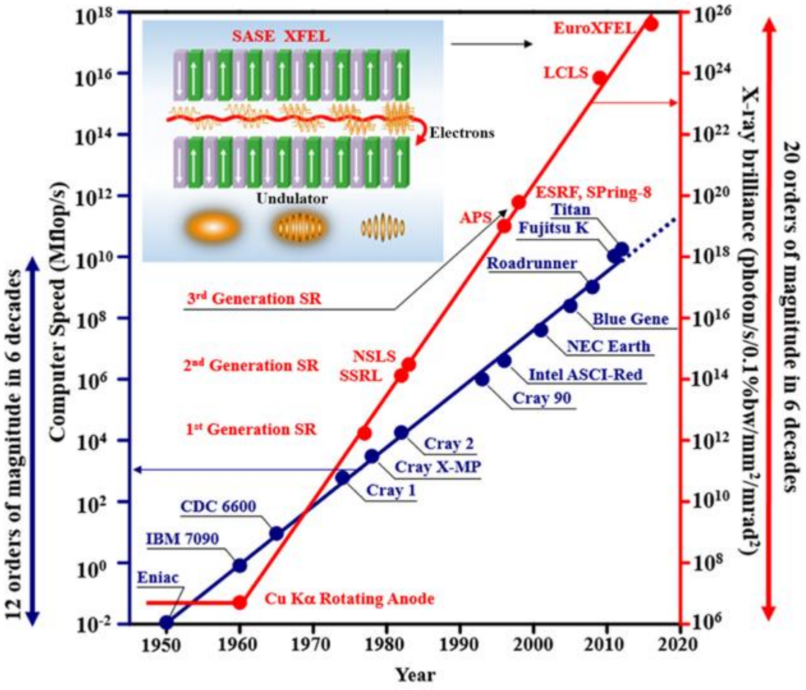
Your recent publications, h-index, R-factors, percentage of analysed datasets and number of proposals make you 23.5% welcome in our facility

Today, facilities (ESRF, ILL,…) gather bibliographic information regarding users, including:

- Relationship between a publication and an experimental session
- Publication delay after an experiment

This *could* (in principle) be used in the future to help during beamtime allocation panels

CHALLENGES FOR SYNCHROTRON / XFEL DATA ANALYSIS



- nb neurons in human brain: 10^{12}
- nb synapses10^{14}-10^{15}

More photons:

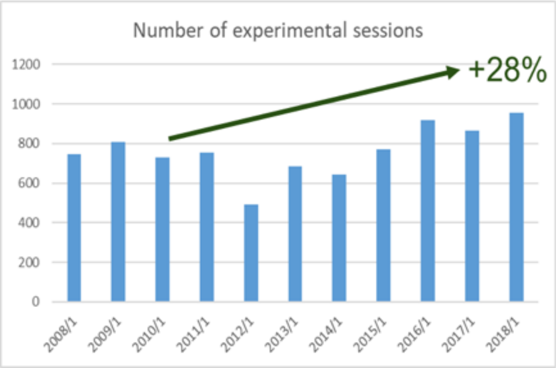
- 10^{12} photons/s (10^{17}-10^{18} photons/experiment)
- (more with pink beam)

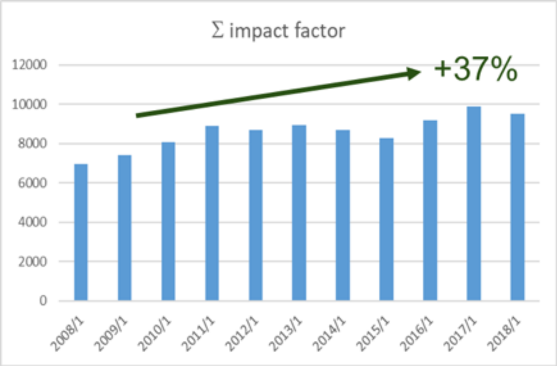
Need to:

- Process data
- Manage radiation damage
- Interpret individual datasets
- Explore encyclopaedia of data

Sample → Data collection → Data processing → Analysis (interpretation: segmentation, statistical analysis…)

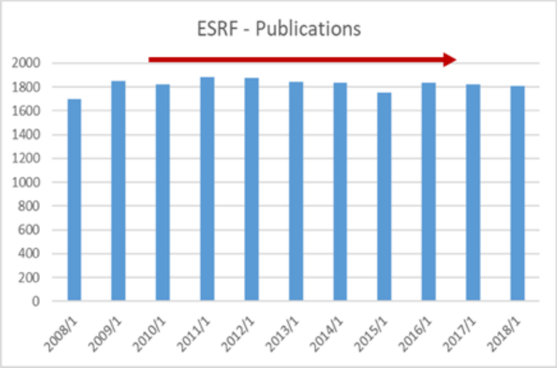
FROM DATA COLLECTION TO DATA EXPLOITATION





Beamtime usage and data exploitation over the 2008-2018 period

2/3 of the beamline portfolio are offering “ expert” emerging techniques: (ptychography, coherent diffraction imaging, serial crystallography, diffraction-contrast tomography, etc.)



Dramatic increase of raw data produced by most beamlines

#publications is stalling despite the increase in the #experiments !
Too much data ?
No time to analyse ?
Lack of good decisions during experiments ?

MACHINE LEARNING

There are several ways that machine learning can be applied to synchrotron data analysis, including:

Image processing: Machine learning algorithms can be used to process and analyze images obtained from synchrotron experiments, such as **segmentation**, **registration**, and **denoising**.

Data reduction: Machine learning can be used to reduce the large amount of data generated by synchrotron experiments, such as by **identifying patterns and features** in the data that are most relevant for a particular analysis.

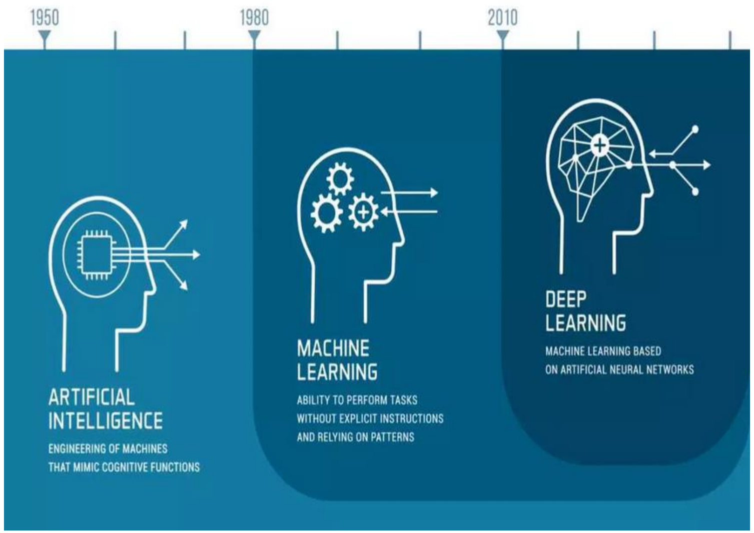
Modeling and simulation: Machine learning algorithms can be used to model and simulate the behavior of materials and systems under synchrotron conditions, such as by predicting the properties of a material based on its structure.

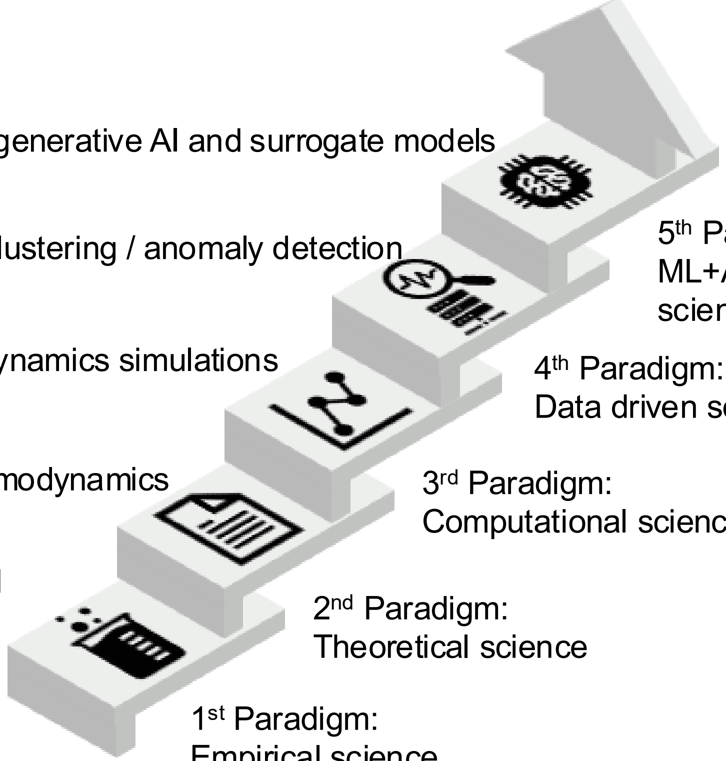
Data visualization: Machine learning can be used to **visualize and interpret large and complex data sets** generated by synchrotron experiments, such as by creating interactive and dynamic visualizations that allow scientists to **explore and analyze the data in new ways**.

Automation: Machine learning can be used to **automate the data analysis process**, such as by training a model to recognize patterns in the data and perform specific tasks without human intervention.

NB: above text is the answer from the *ChatGPT* AI to the question:
How can machine learning be applied to synchrotron data analysis ? ...

ARTIFICIAL INTELLIGENCE -> DEEP LEARNING





e.g. generative AI and surrogate models

5th Paradigm:
ML+AI driven science

“big data”, e.g. clustering / anomaly detection

4th Paradigm:
Data driven science

e.g. molecular dynamics simulations

3rd Paradigm:
Computational science

e.g. laws of thermodynamics

2nd Paradigm:
Theoretical science

experiments and observations

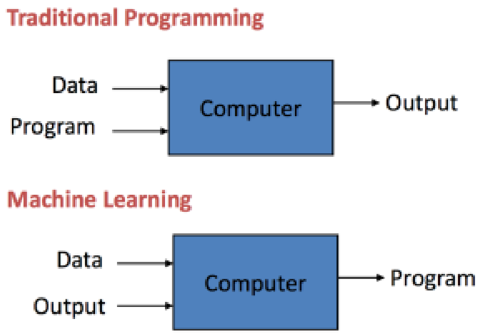
1st Paradigm:
Empirical science

https://www.microsoft.com/en-us/research/blog/ai4science-to-empower-the-fifth-paradigm-of-scientific-discovery/
https://plat.ai/blog/deep-learning-in-data-science/

Slide: L. Pithan

MACHINE LEARNING - CONCEPTS

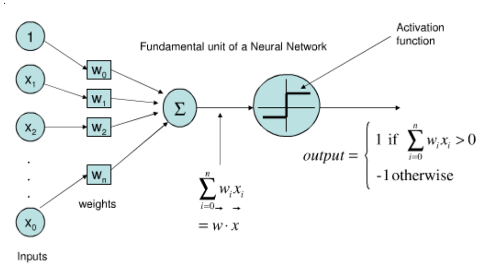
- **Artificial Intelligence**
  - Artificial Intelligence is when machines can do tasks that typically require human intelligence


Traditional Programming

Machine Learning

- **Machine Learning**
  - Machine learning is when machines can learn by experience and acquire skills without human involvement.

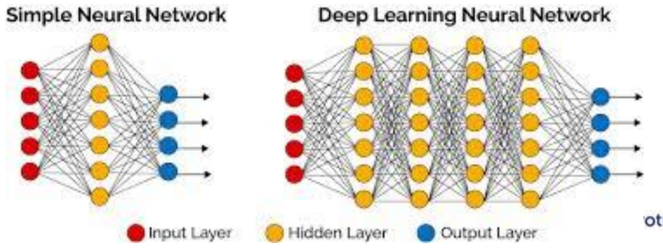
- **Deep Learning**
  - Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data.



- **What problems can be solved ?**

  "Just about any problem that requires "thought" to figure out is a problem deep learning can learn to solve."

  Bernard Marr

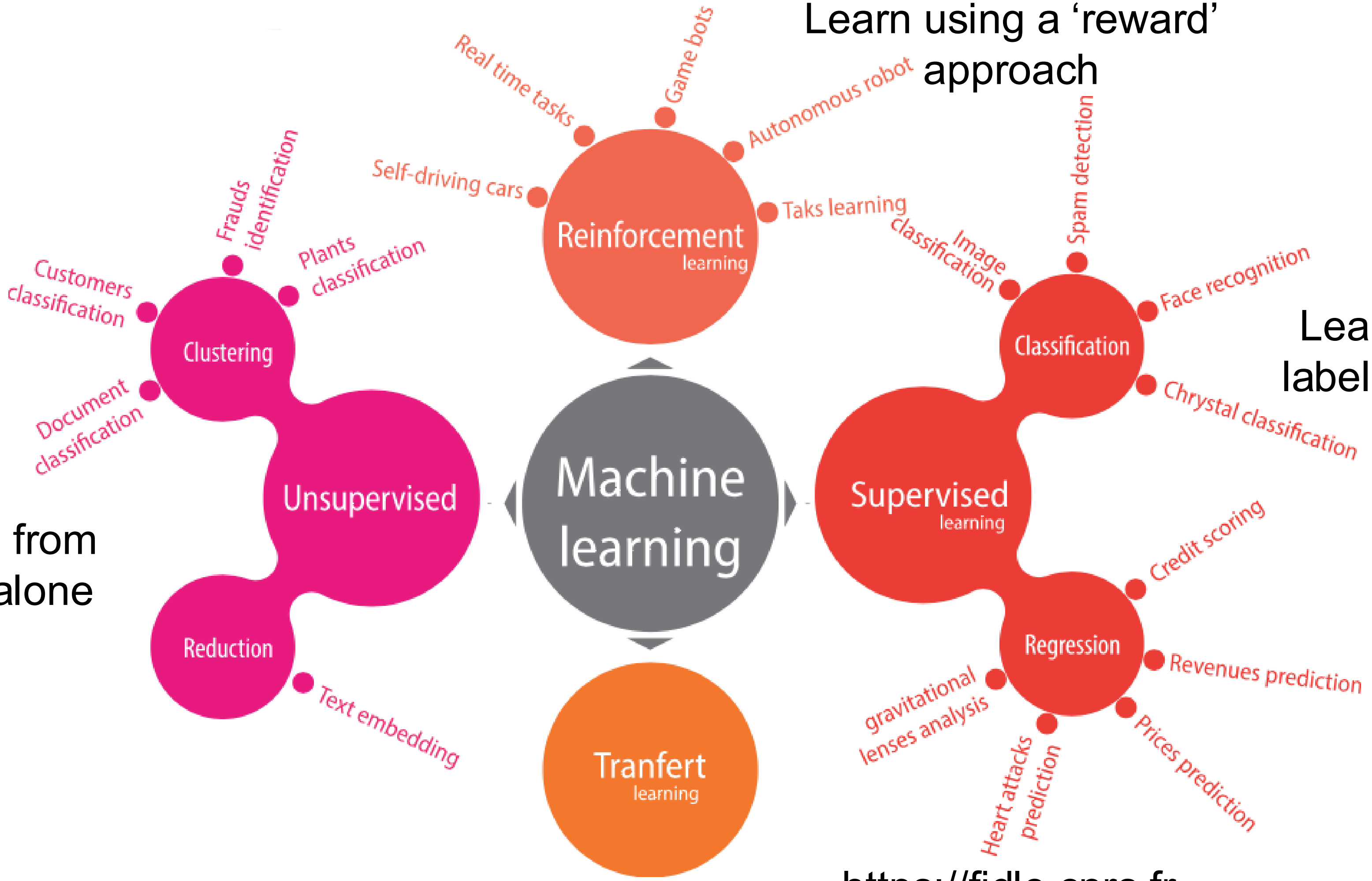


MACHINE LEARNING CATEGORIES

Learn using a ‘reward’ approach

Learn from labelled data

Learn from data alone

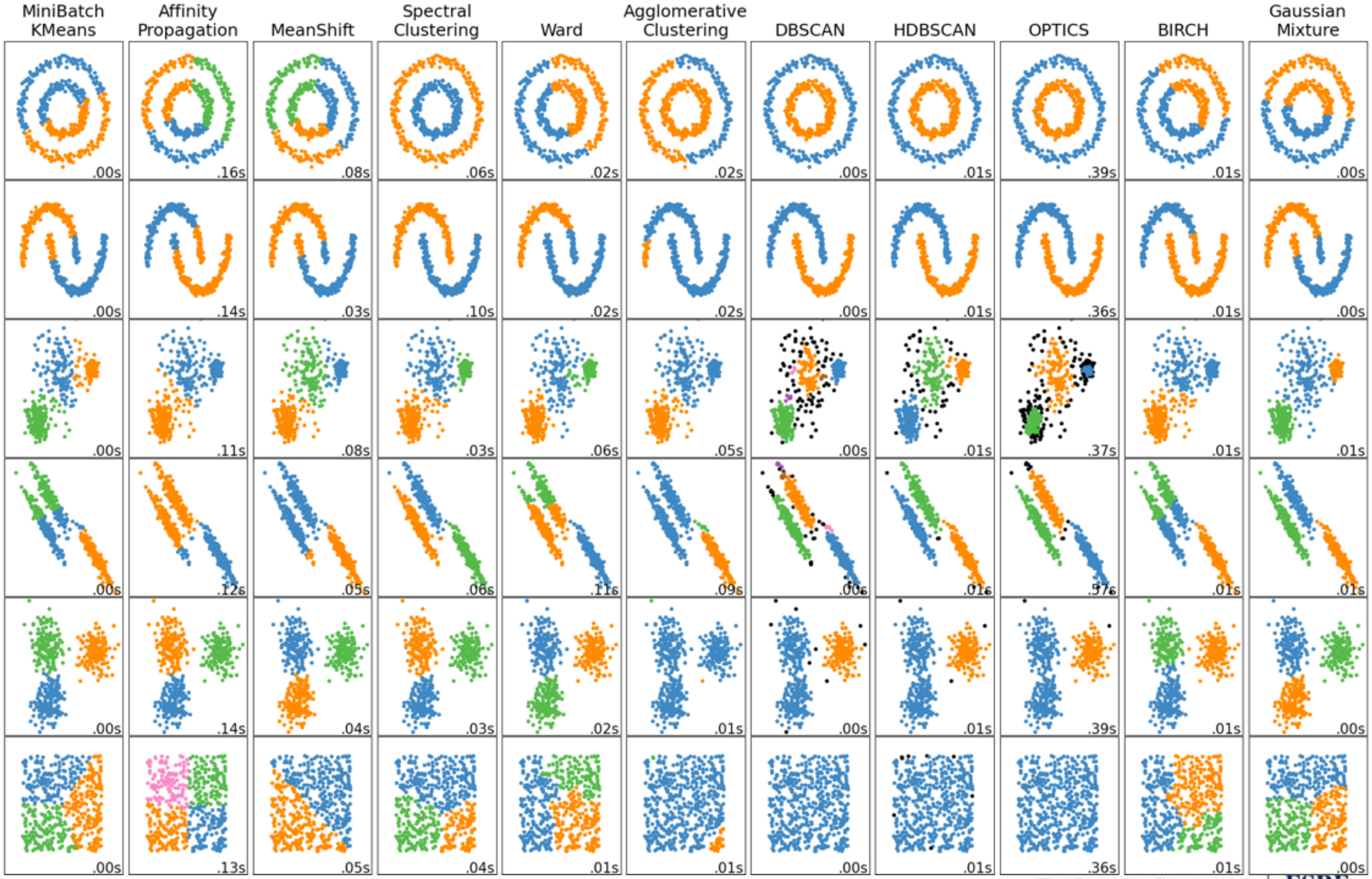


https://fidle.cnrs.fr

https://cloud.univ-grenoble-alpes.fr/index.php/s/wxCztjYBbQ6zwd6

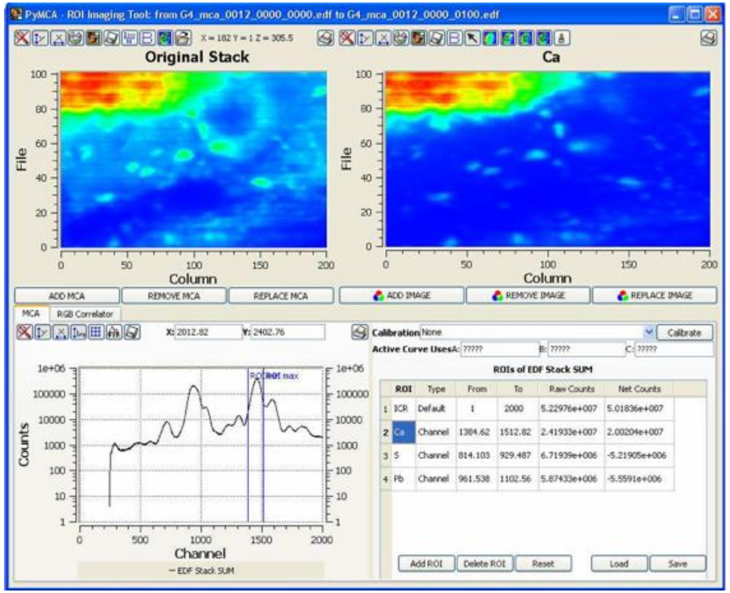
EXAMPLE: CLUSTERING ANALYSIS USING SCIKIT-LEARN

https://scikit-learn.org/stable/user_guide.html



You don't always need *deep learning* or a GPU for machine learning !

'CLASSICAL' MACHINE LEARNING: PCA, ICA,



X: 2012.82 Y: 2402.76

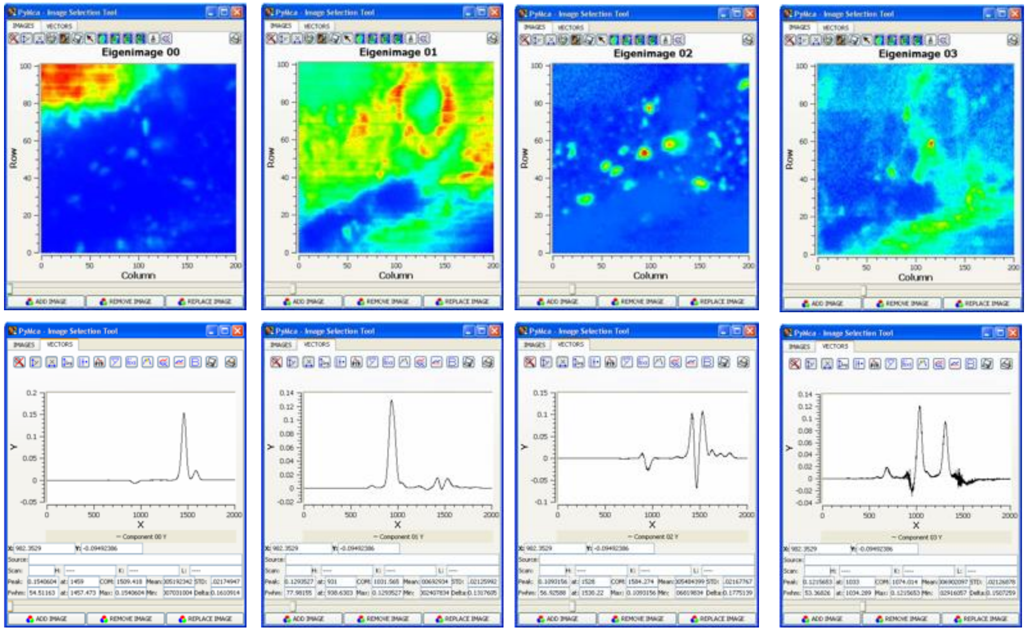
Calibration None

Active Curve Uses A: ????? B: ????? C: ?????

ROIs of EDF Stack SUM

	ROI	Type	From	To	Raw Counts	Net Counts
1	ICR	Default	1	2000	5.22976e+007	5.01836e+007
2	Ca	Channel	1384.62	1512.82	2.41933e+007	2.00204e+007
3	S	Channel	814.103	929.487	6.71939e+006	-5.21905e+006
4	Pb	Channel	961.538	1102.56	5.87433e+006	-5.5591e+006

Data: 101x200 map of fluorescence spectra (2000 points each)
=> Use unsupervised decomposition (PCA)



Machine learning has been possible in PyMca for a *long* time:

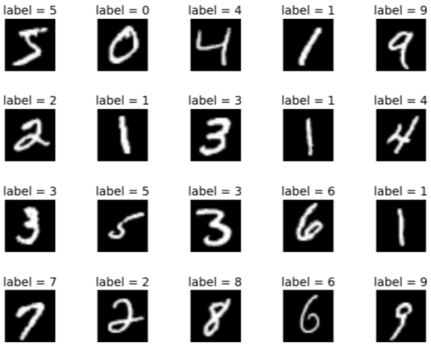
- **Principal component analysis**
- **Independent component analysis**
- **Non-negative matrix factorisation**

- ✓ Allows easy (no training) analysis of datasets
- ✓ Totally generic
- ✓ ..but does not tell you what the different compents are

Slide figures: V. A Solé

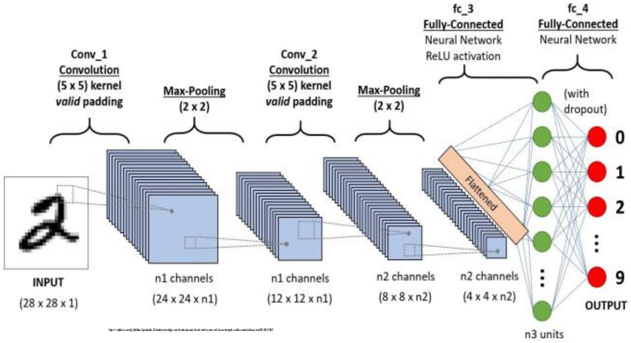
SUPERVISED CONVOLUTIONAL NEURAL NETWORKS: KEY CONCEPTS

Training Dataset

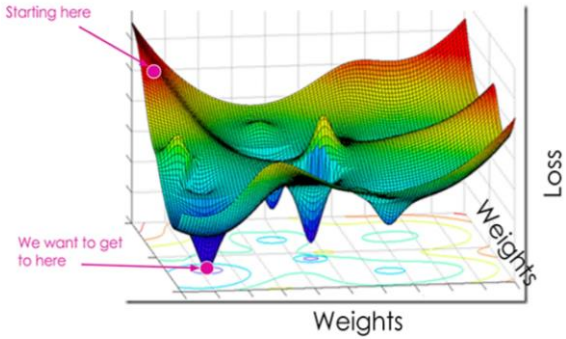


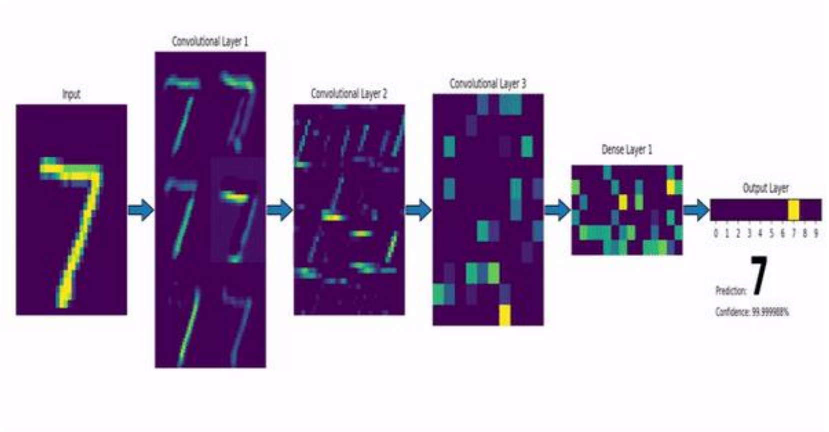
https//corochann.com/mnist-dataset-introduction-532/

Model



Loss Function





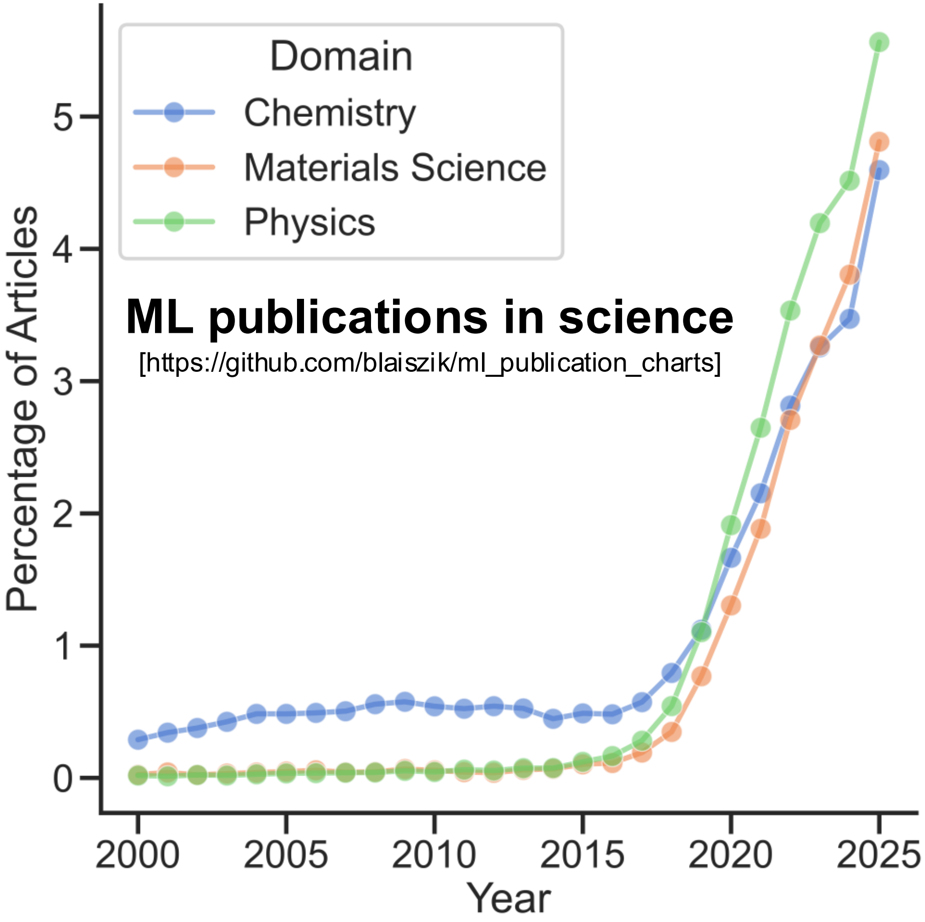
- Convolutional layers extract features from images
- The minimization of the loss function adjusts the weights to make the model fit the statistical distribution of the training set
- When the training is completed the model can generalize the prediction to new data

Slide: M. Masto

MACHINE LEARNING FOR X-RAYS & NEUTRONS

Design be
experimer

Better
processing



ML publications in science

[https://github.com/blaiszik/ml_publication_charts]

ir
(si

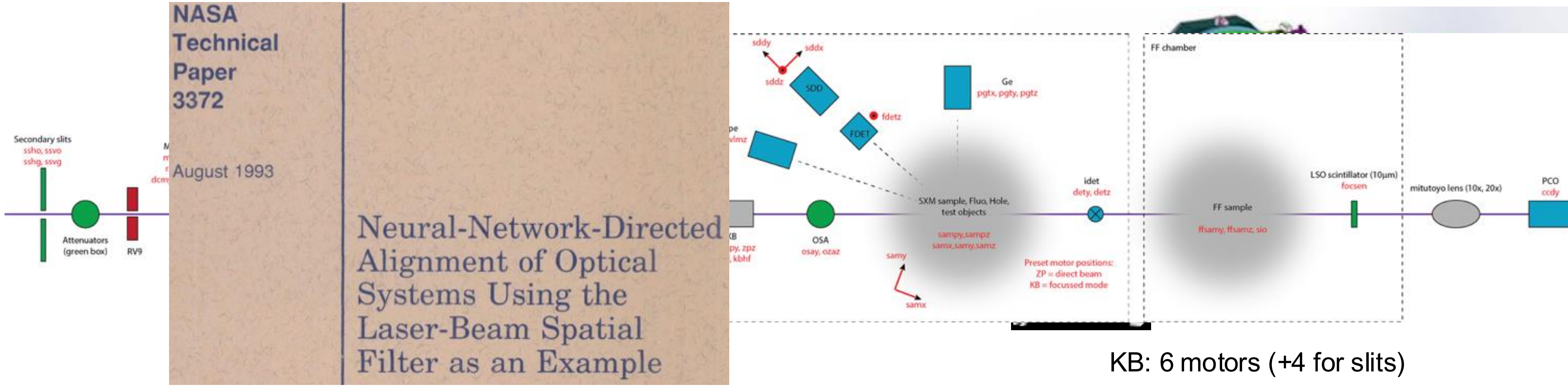
ter
etation
datasets)

2019 works

zPPTZE

1. **Design better experiments**
2. Faster data processing
3. Better processing
4. Interpretation (single dataset)
5. Interpretation (multiple datasets)

BEAMLINE OPTICS ALIGNMENT



NASA
Technical
Paper
3372

August 1993

Neural-Network-Directed
Alignment of Optical
Systems Using the
Laser-Beam Spatial
Filter as an Example

KB: 6 motors (+4 for slits)

From source to detector, there can be >10 optical elements, each with 2-10 motors:

- **Undulators**
- **Mirrors**
- **Slits**
- **Monochromator**
- **Nano-focusing optics**
- **Sample alignment**
- **Detectors** (distance, orientation, energy thresholds, analyser)

All alignment usually done progressively from the source, but some elements with >=4 motors (actuators), and/or non-perfect optical properties are difficult to parametrise.

AUTOMATED ACQUISITION: SCANNING X-RAY MICROSCOPY

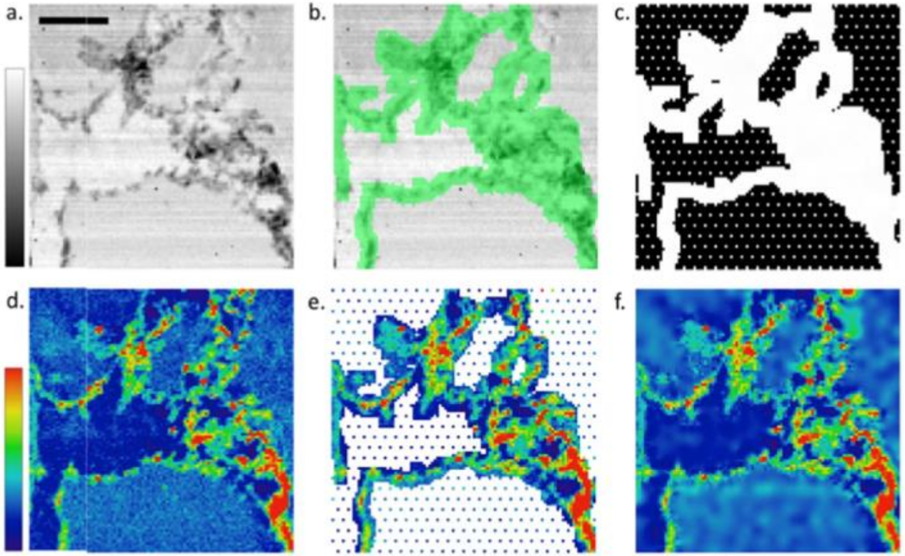


Figure 1. A rapidly acquired STXM map (**a**) (80 µm × 80 µm, 50 × 50 pixels, 20 ms dwell time/pixel, scale bar = 20 µm) is used to create a mask (**c**) which is dense only in the regions of interest (green areas in **b**). A sub-sampled sparse XRF acquisition (**e**) is approximately 3 times faster than a full one (**d**). When the sparse acquisition is reconstructed through biharmonic in-painting (**f**) it can be processed with the usual XRF workflows producing similar results. In this specific case Panels d, e and f show Si XRF signal collected at 1.95 keV on a foraminifera section over an area of 80 µm × 80 µm, with 1.6 µm step size and 3 s acquisition time/pixel.

Fluorescence maps

Using a 2-pass data acquisition:

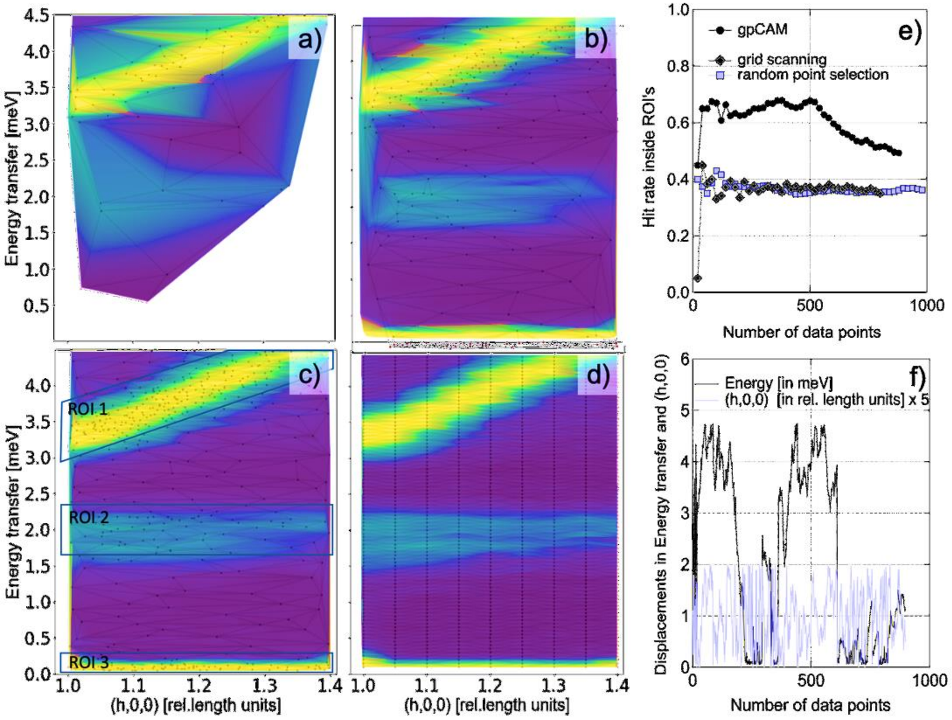
- Low resolution
- Identify relevant areas
- High resolution in selected areas

Allows:

- Faster scanning
- Less radiation damage

Sci. Rep. 10(2020), 9990

AUTOMATED DATA ACQUISITION



a) 100 points
b) 250 points
c) 500 points
d) 800 [grid]

Automated measurement on the triple-axis spectrometer at ILL (Thales):

- First measure a few points
- Then automatically determine where the most relevant points should be measured using a ***gaussian process***

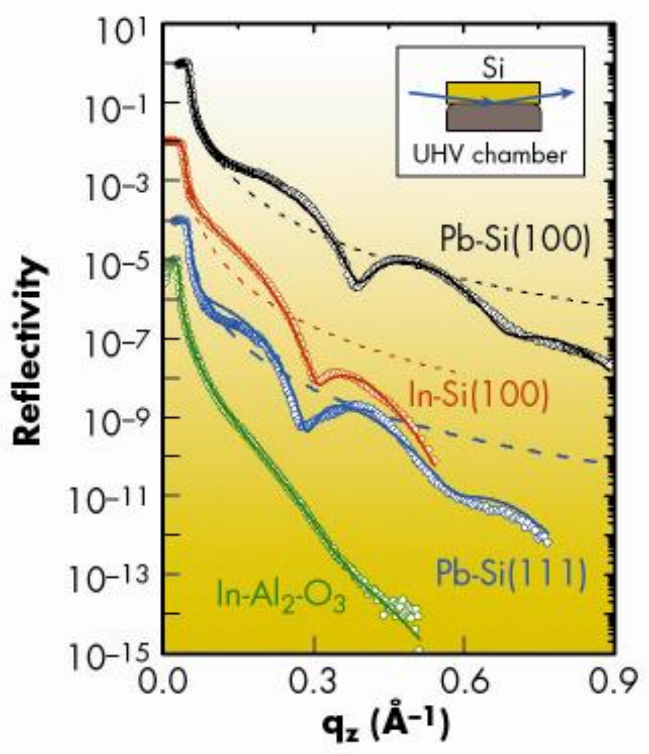
- Generic approach which can be used for any type of experiment
- Models can be physics-aware by optimising specific areas (strong signal gradient, …)

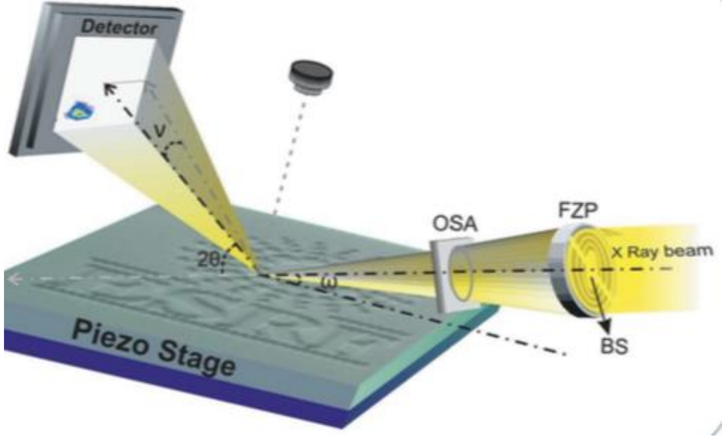
https://gpcam.lbl.gov/

Nature Reviews Physics 3 (2021), 685

1. Design better experiments
2. **Faster data processing**
3. Better processing
4. Interpretation (single dataset)
5. Interpretation (multiple datasets)

X-RAY REFLECTIVITY





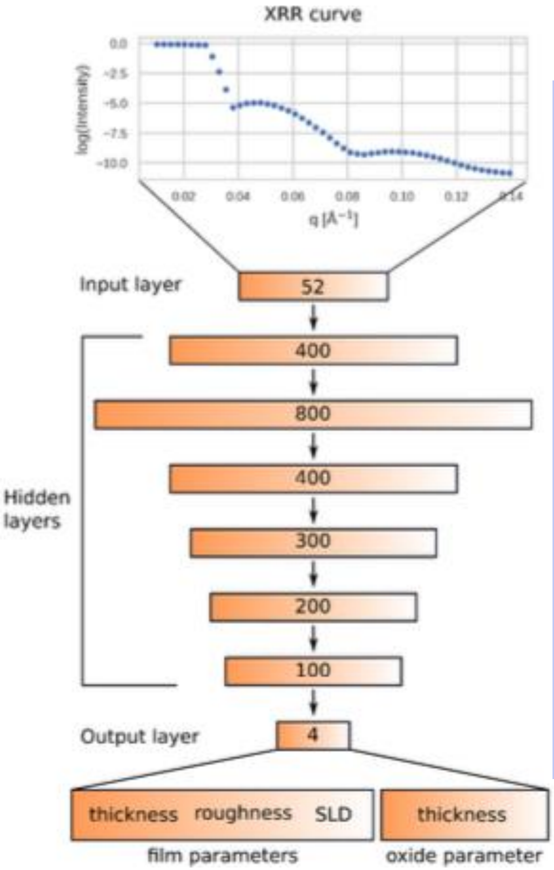
X-ray reflectivity: measure the (specular) reflected intensity allows retrieving the electronic density ρ(z) as a function of depth.

This can be **fitted** (Monte-Carlo, least squares, genetic algorithms) to retrieve the thickness and roughness of layers.

Fast (continuous) scanning allows with a bright beam allows to scan surfaces up to 10^3-10^4 positions/s, allowing to acquire reflectivity curves for thousands-millions of points.

Phys. Rev. Lett. 98, 116101 (2007)
J. Appl. Cryst. **47**, 762–769 (2014)

X-RAY REFLECTIVITY FITTING WITH NEURAL NETWORK

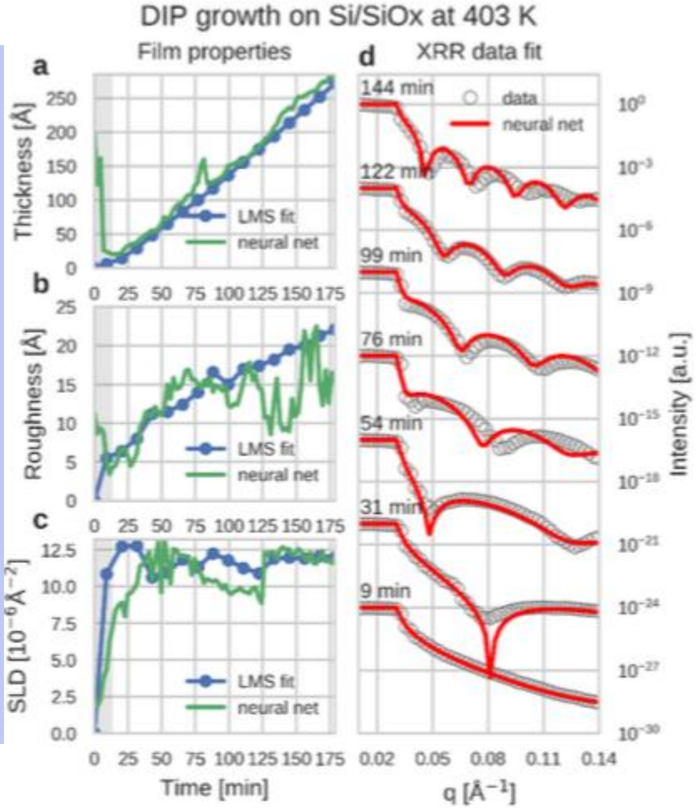


Training a neural network with a range of film parameters:

- Thickness
- Roughness
- Scattering length density
- Substrate thickness

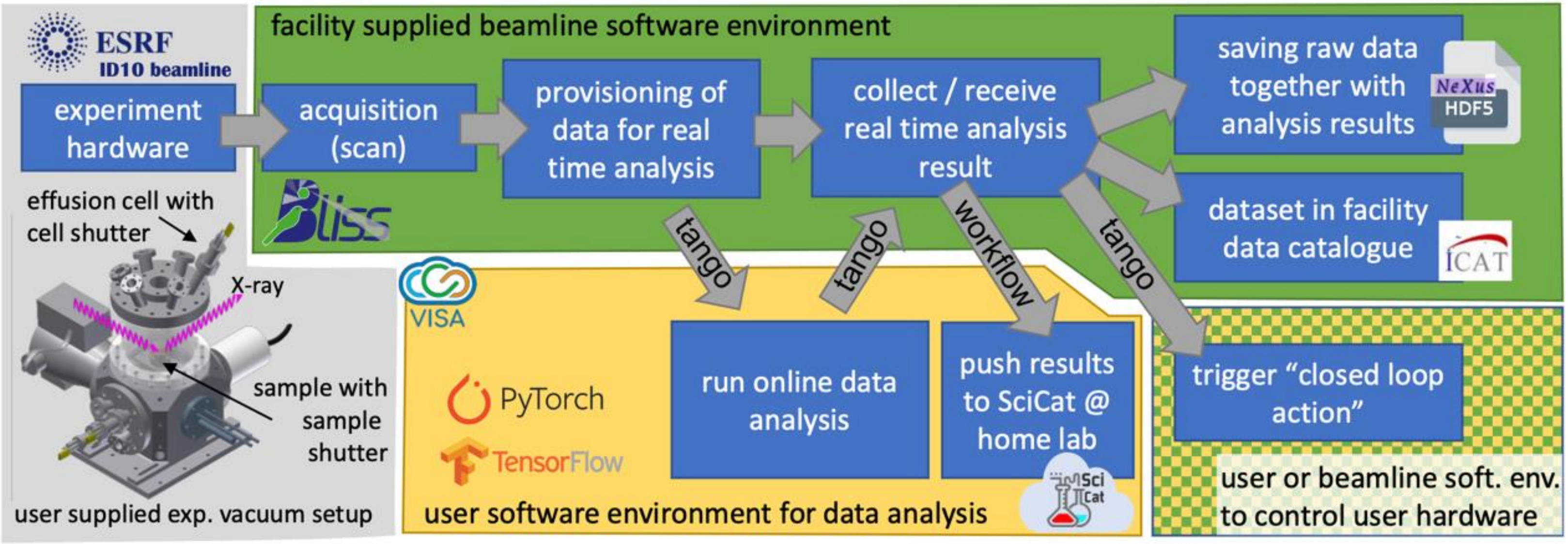
The resulting NN allows to predict the layers parameters with a **speed of 0.03-77 milliseconds per curve** – fast enough for live analysis.

Important: predictions are limited to the range of the simulated training set.



https://arxiv.org/abs/1910.02898

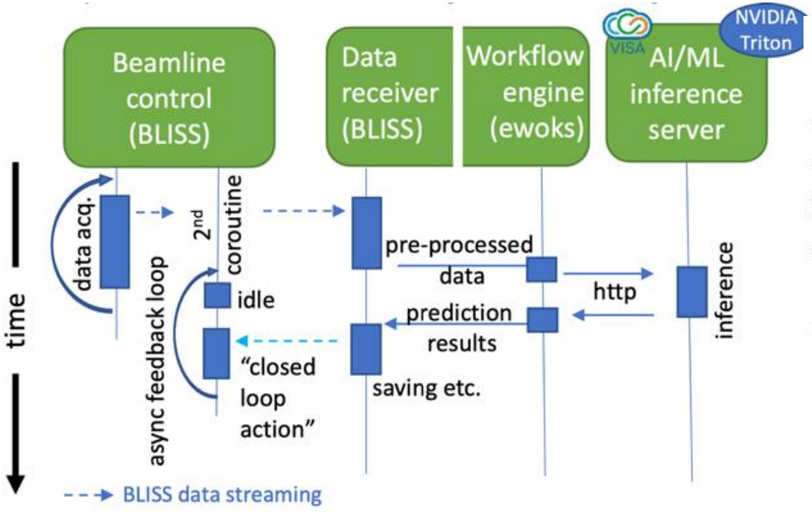
ML AUTOMATED PROCESSING / ACQUISITION WITH BLISS

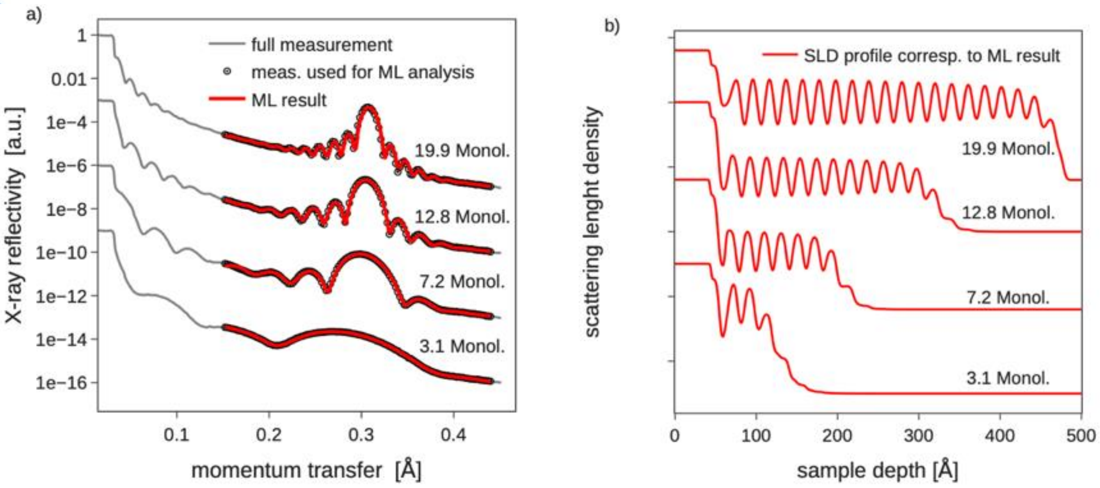


Reflectivity @ID10 general workflow

Pithan et al, arXiv:2306.11899

ML AUTOMATED PROCESSING / ACQUISITION WITH BLISS





Using ML to tune / trigger actions during an experiment: live processing of data on a dedicated server for ML inference.

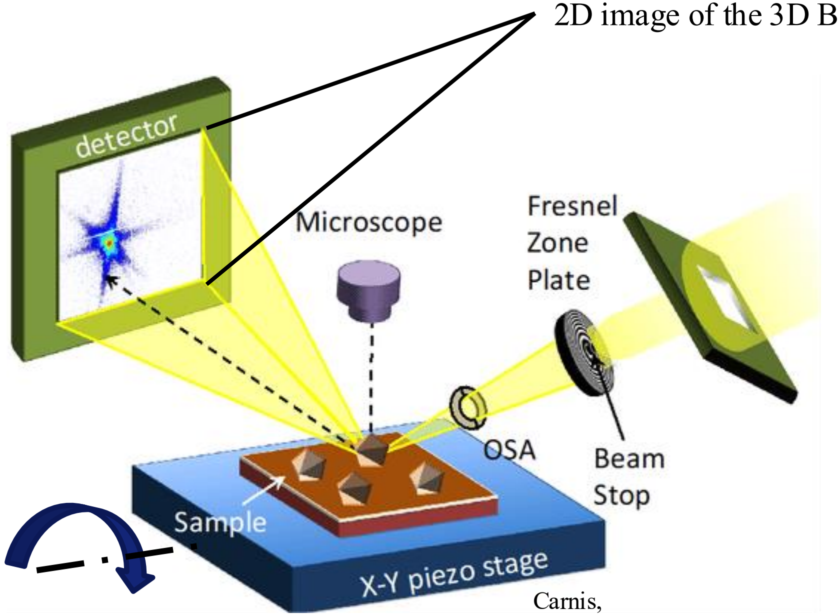
Automatically decide when the measurement should stop (enough data, target reached for operando,..)

ML can be used to automatically fit reflectivity data.
Good example of ML application as the non-linearity of the reflectivity vs the model can make the fit complicated.

ML reflectivity fit:J. Appl. Cryst. 55, 362

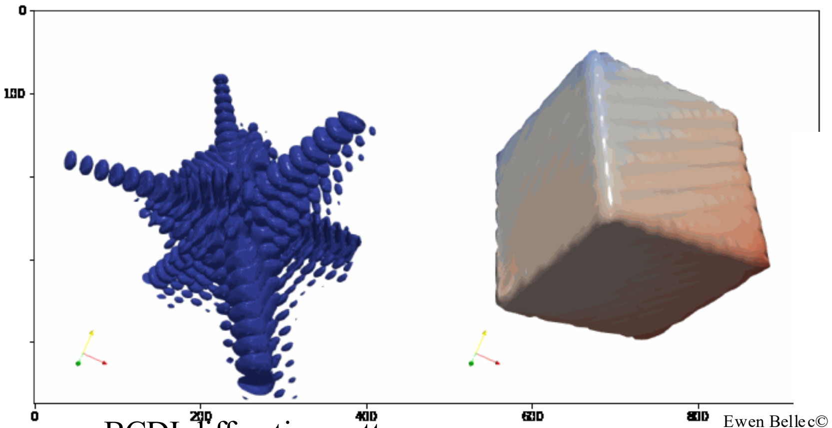
Pithan et al, arXiv:2306.11899

COHERENT IMAGING



Rocking axis

Carnis, 2017



BCDI diffraction pattern

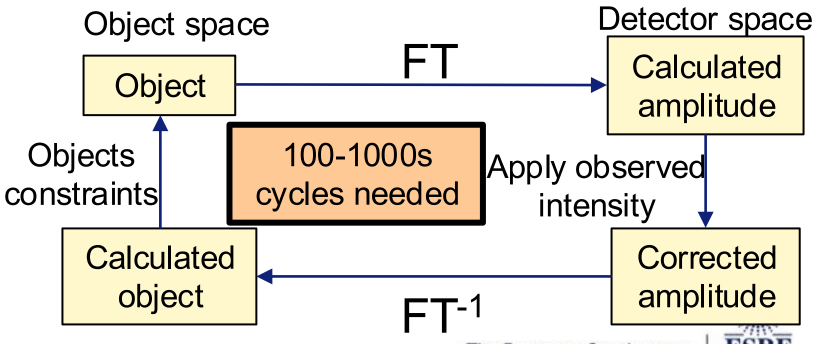
Crystalline sample

Ewen Bellec©

coherent imaging techniques: the **far-field projection** is recorded, being related by Fourier transform(s) to the object density (refraction index, strain).

=> **high-resolution two and three-dimensional imaging** (resolution down to less than 10 nm).

The **phase of the scattered intensity is lost**, preventing a direct reconstruction → need **phase retrieval algorithms**.



Object space

Detector space

Object

FT

Calculated amplitude

Objects constraints

100-1000s cycles needed

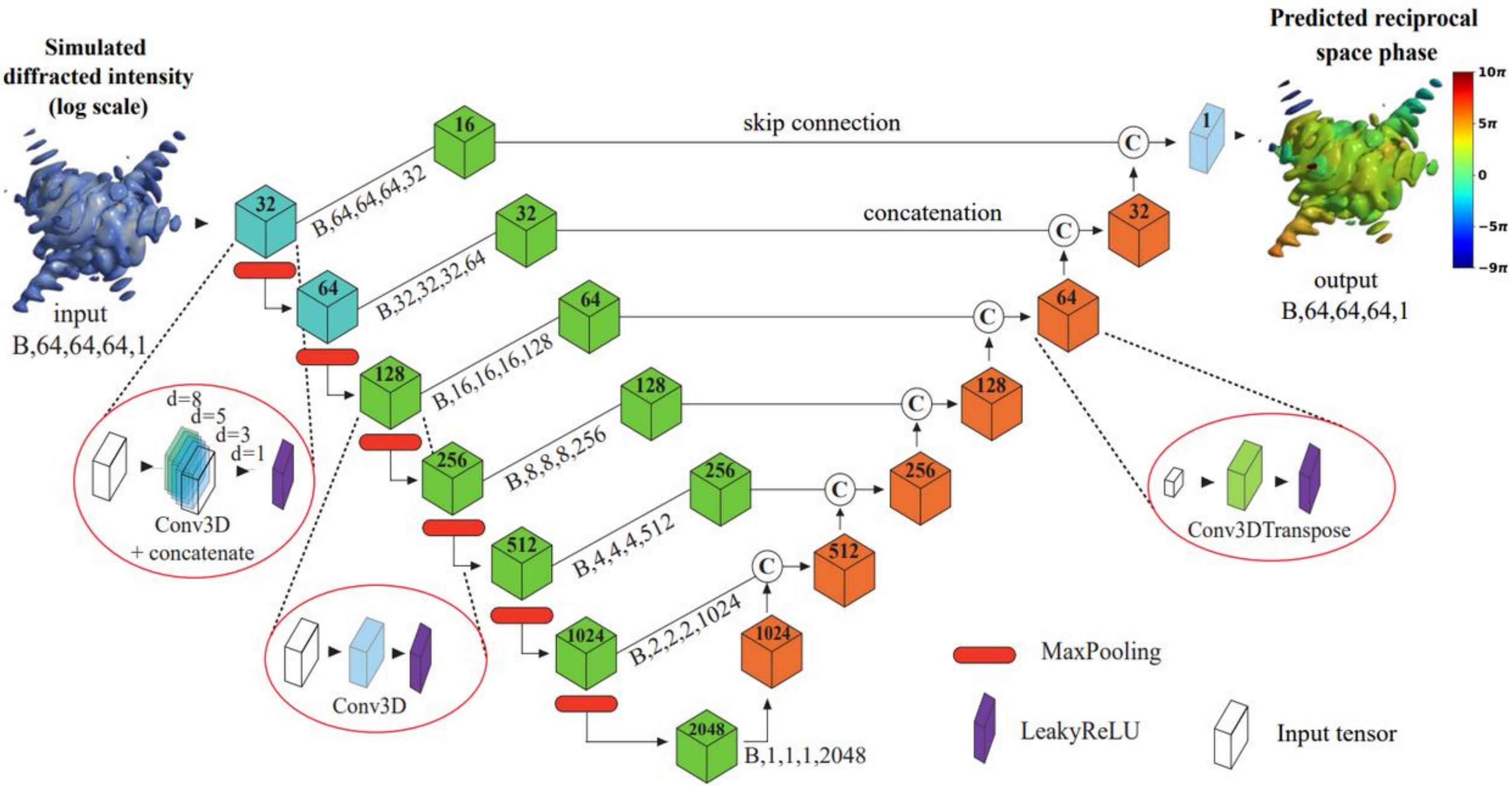
Apply observed intensity

Calculated object

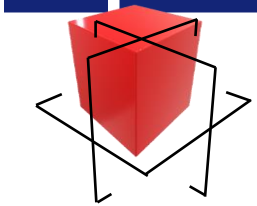
Corrected amplitude

FT^{-1}

BRAGG COHERENT DIFFRACTION IMAGING

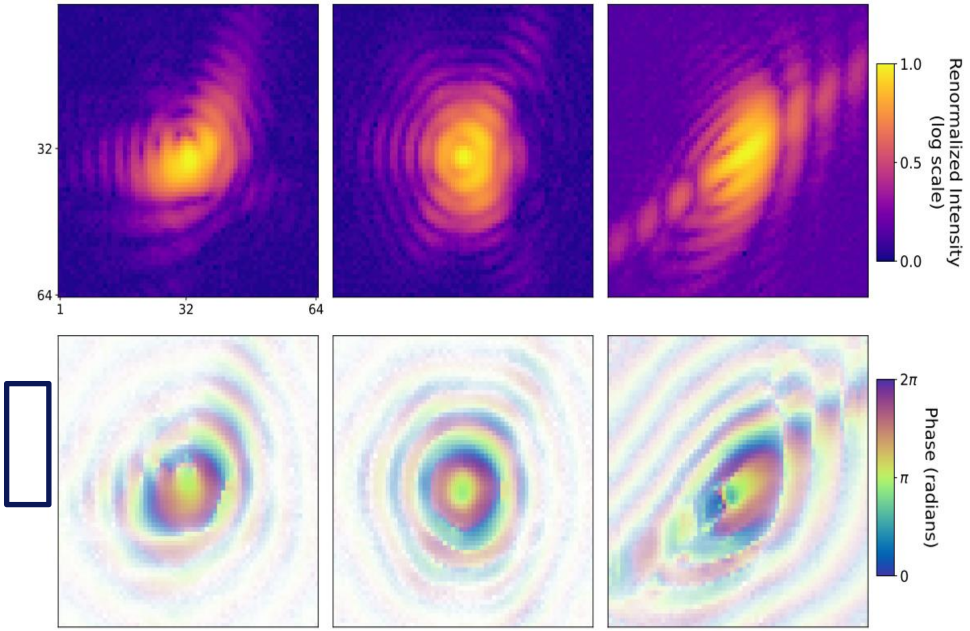


BRAGG COHERENT DIFFRACTION IMAGING



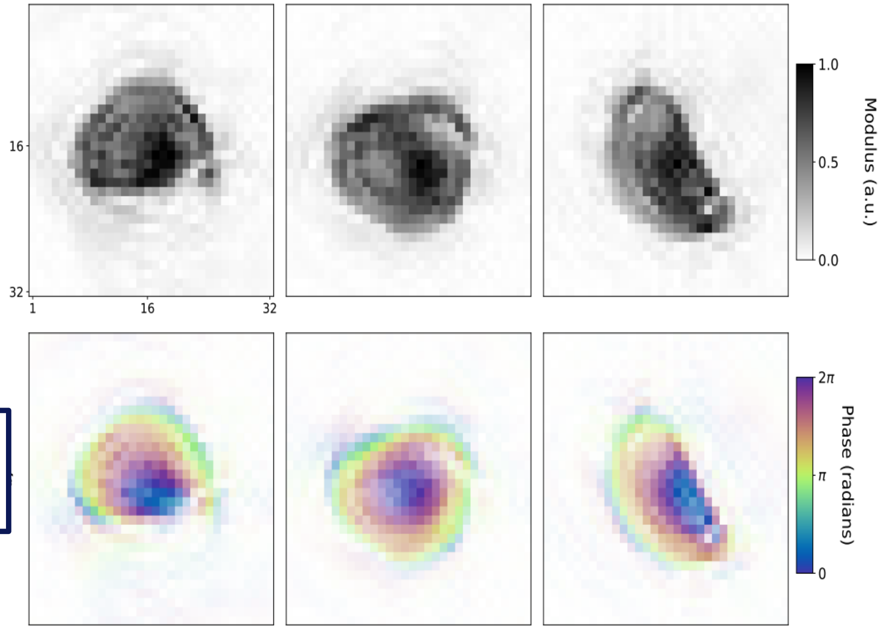
Platinum nanoparticle on YSZ substrate

Experimental measurement



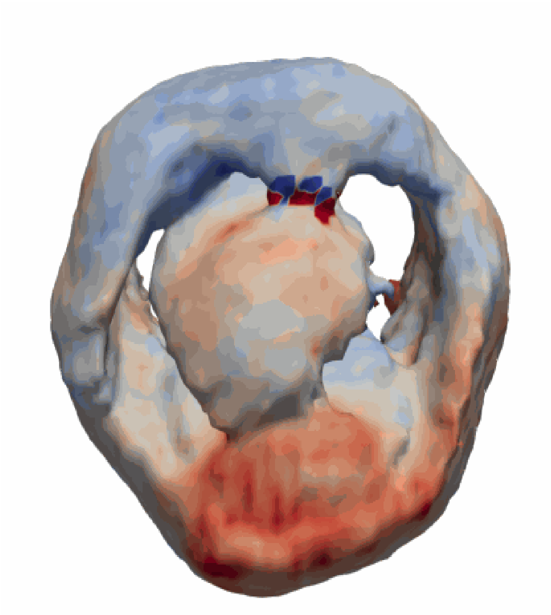
DL phase prediction

Reconstructed object



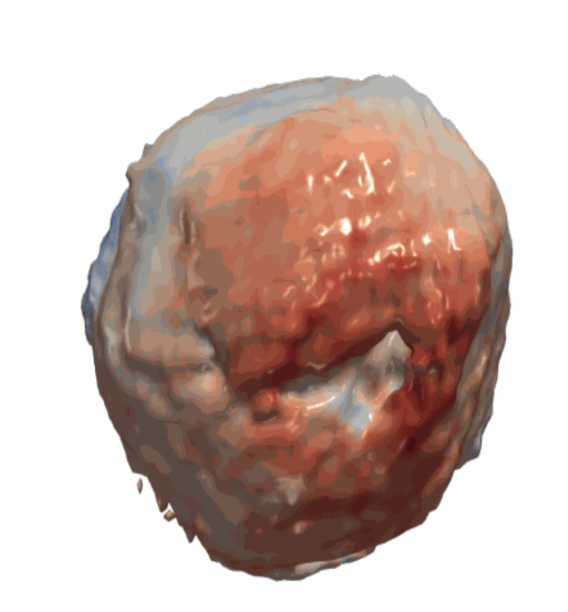
BRAGG COHERENT DIFFRACTION IMAGING

Iterative Phase Retrieval:
Best of **60** runs of
400 HIO + 1000
RAAR + 300 ER



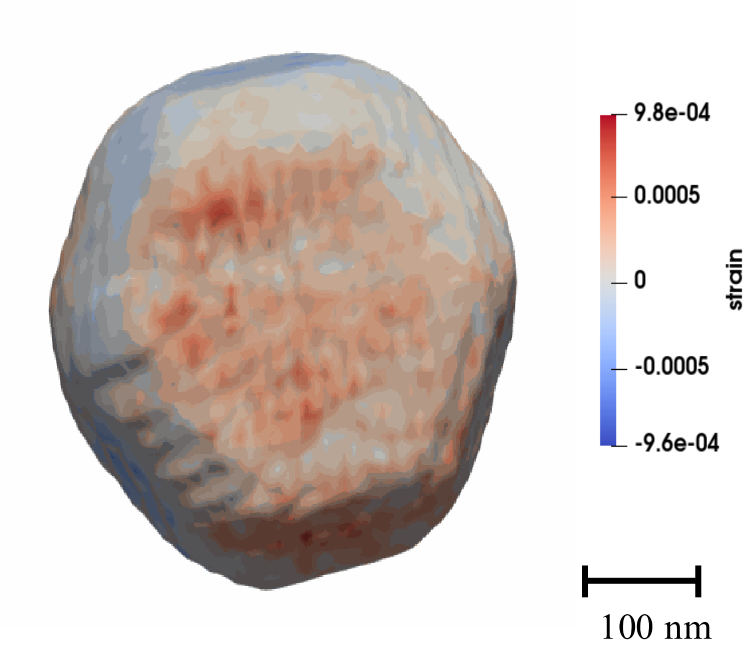
t ~ 4 mins

Deep Learning prediction



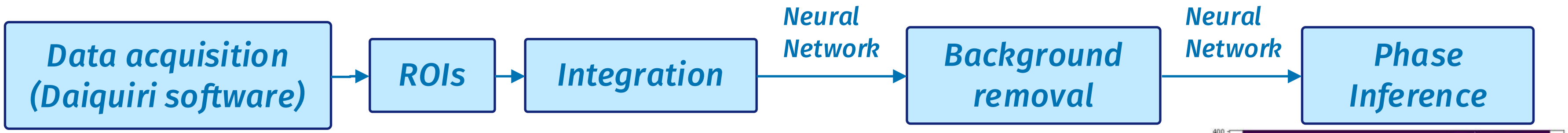
t ~ 4 s

Deep Learning initial guess
+
1 run of 300ER

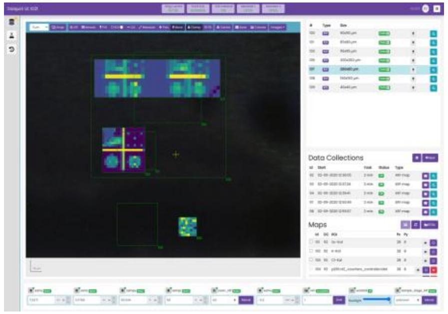


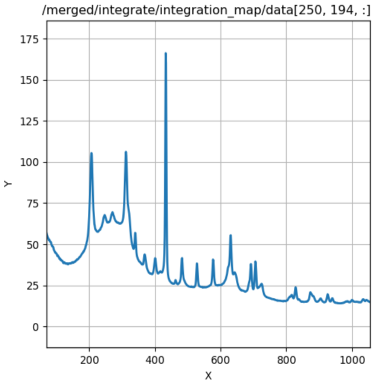
t ~ 5 s

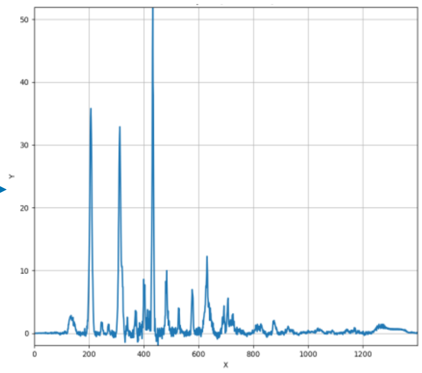
POWDER DIFFRACTION MAPS (ID13)

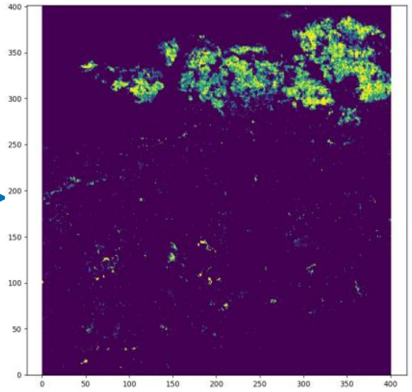


Data acquisition (Daiquiri software) → ROIs → Integration → Neural Network → Background removal → Neural Network → Phase Inference









Azurite probabilistic map ($Cu_3(CO_3)_2(OH)_2$)

- **Training of the Neural Network:**
  - Set of .cif files (possible crystallographic phases)
  - Generation of synthetic data
  - 15 .cif files -> +2 hours training (on dedicated machine)
  - Training can be done offline

- **Assisted phase inference:**
  - 160k frames scan -> ~3-4 min to predict (on dedicated machine)

**spi package by Quentin Bruel*

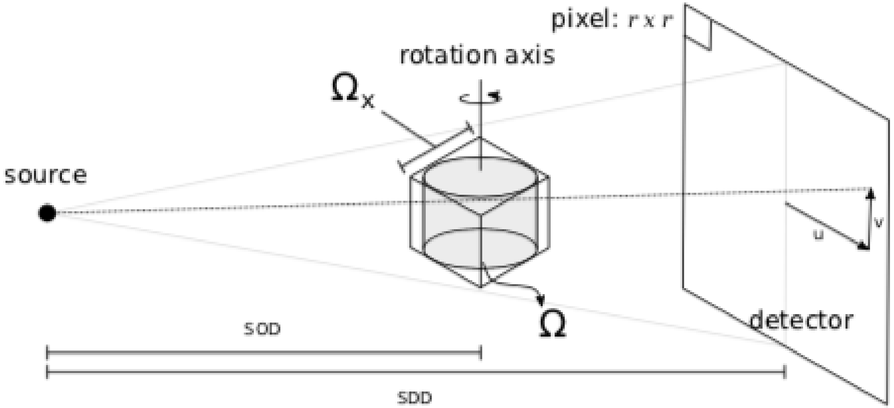
https://gitlab.esrf.fr/spi/spi

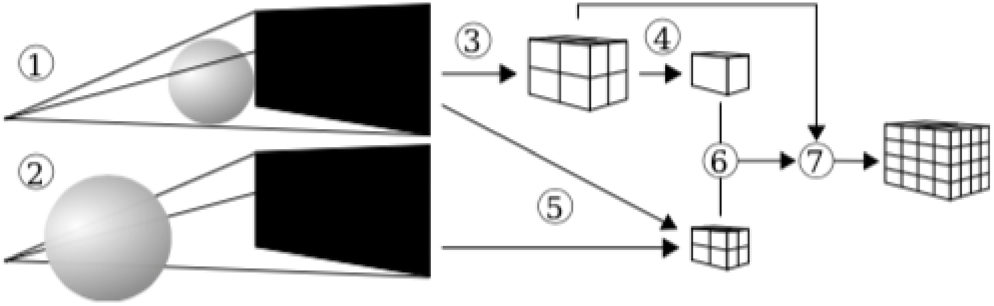
Slide: E. Gutierrez Fernandez

1. Design better experiments
2. Faster data processing
3. **Better processing**
4. Interpretation (single dataset)
5. Interpretation (multiple datasets)

Key idea: try to get **more information** from the same datasets (**higher signal/noise**) compared to traditional algorithms

PHASE CONTRAST IMAGING: IMPROVE RESOLUTION





In phase contrast imaging, valuable improvements include:

- **High resolution object from low resolution image**
- **Exploiting low signal/noise data**

This would allow:

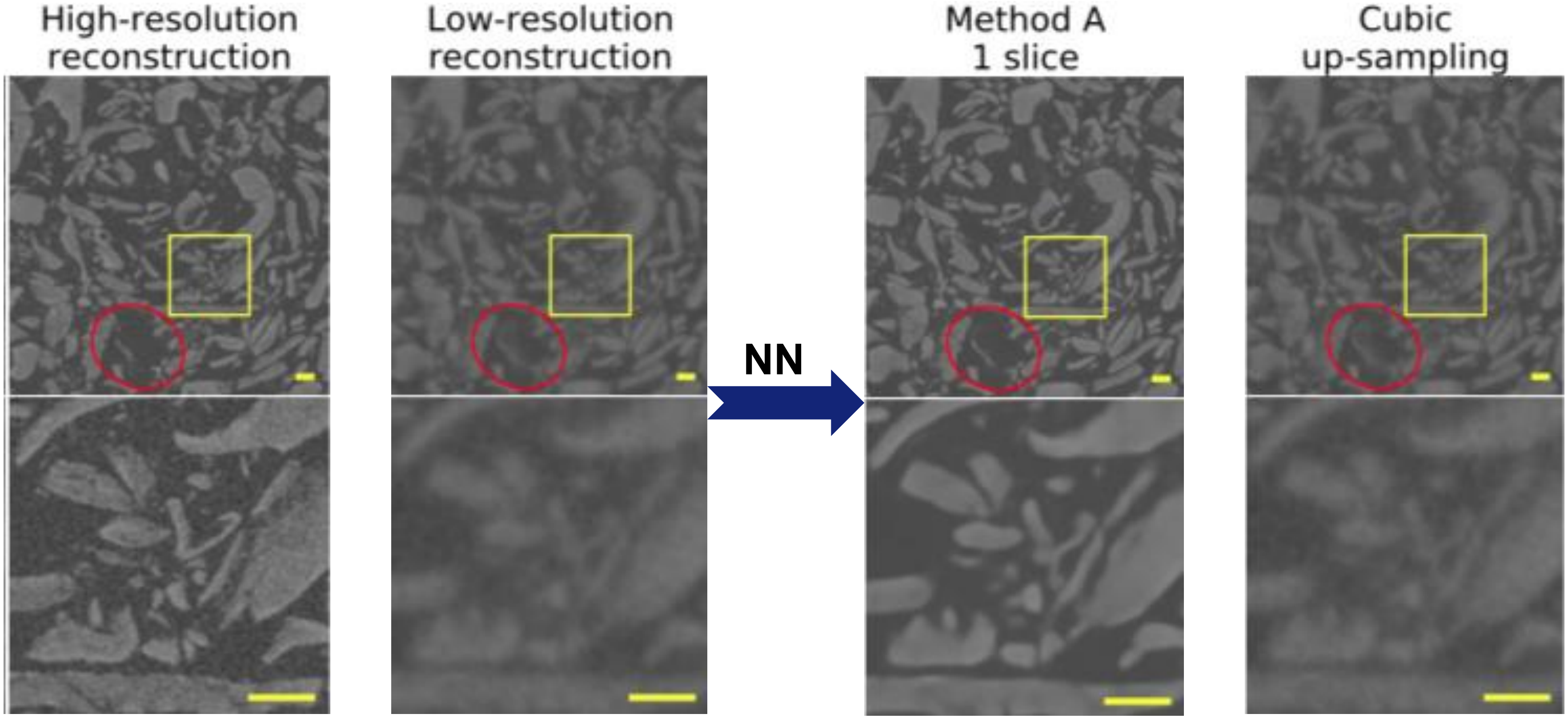
- **Faster data acquisition** (fewer projections)
- **Less radiation damage**

Method: create a neural network by acquiring a dataset at high resolution, and train it against the same data with a reduced resolution.

The NN is then tested against a different part of the same object.

Appl. Sci. 2019, 9, 2445

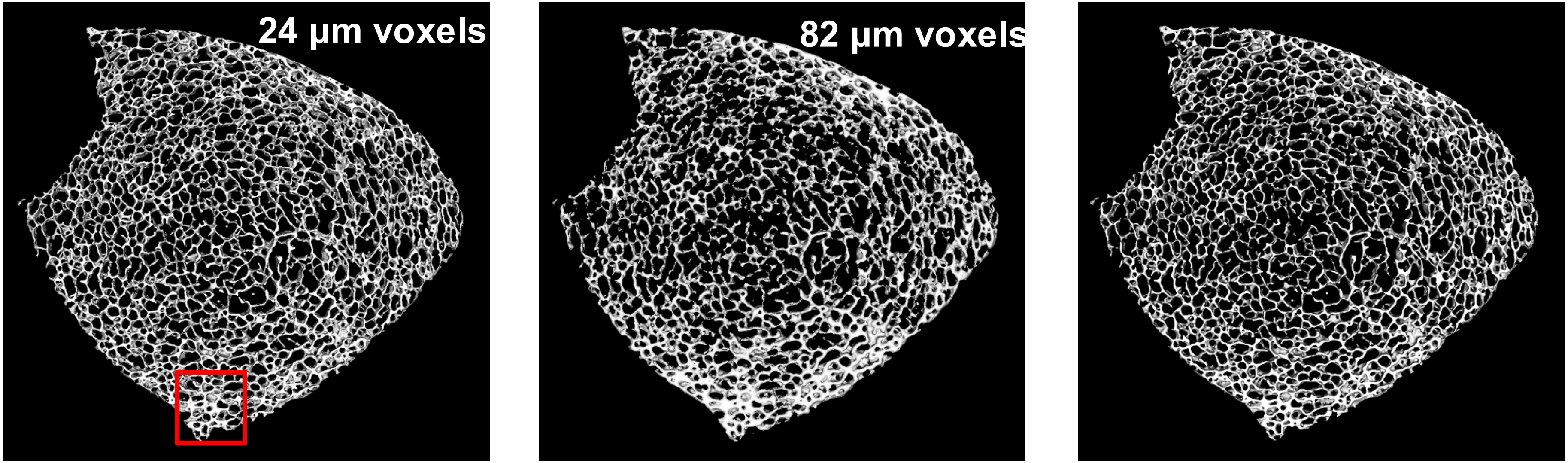
PHASE CONTRAST IMAGING: IMPROVE RESOLUTION



Example NN reconstruction on an oatmeal sample, training from a high resolution sub-dataset (17µm voxels) to reconstruct low-resolution ones (68µm voxels)

Appl. Sci. 2019, 9, 2445

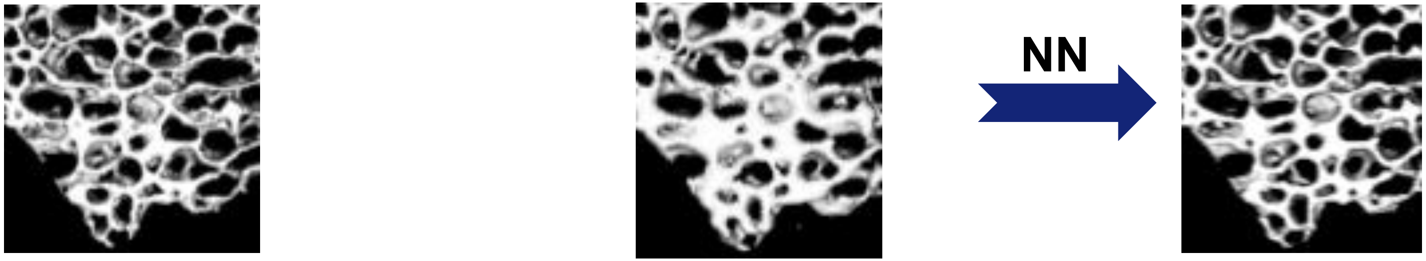
PHASE CONTRAST: IMPROVE RESOLUTION *IN VIVO*



High Resolution (from µCT)

Low Resolution

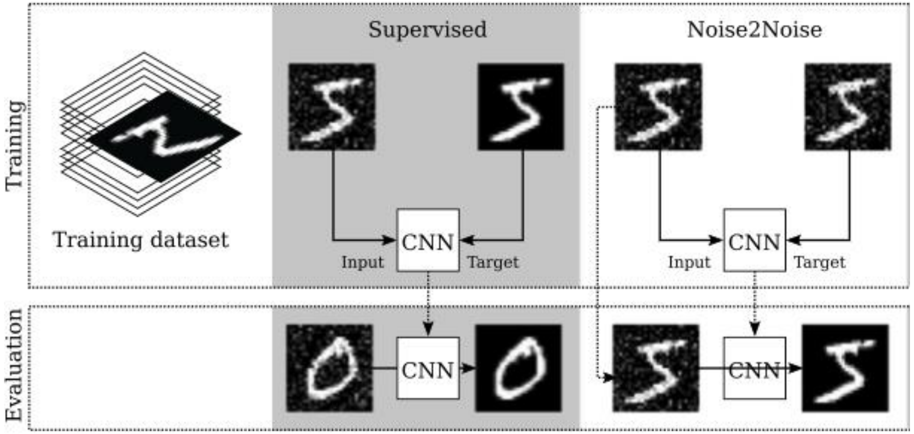
Super Resolved image



Application to the study of ostheoporosis

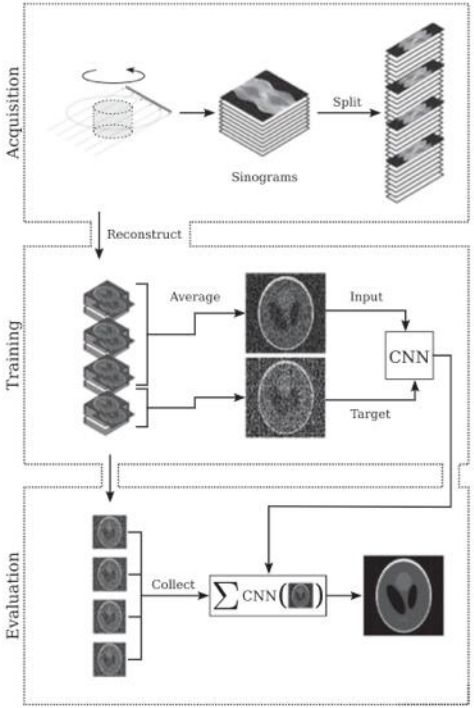
Slide courtesy of F. Peyrin
PhD Y Li, 2019 ; B. Sixou, F. Peyrin (unpublished)

MACHINE LEARNING – DENOISING IMAGES & VOLUMES



Unsupervised denoising (noise2noise):

- Use two images => only difference is noise
- Have a NN learn the noise model
- Use the NN for denoising

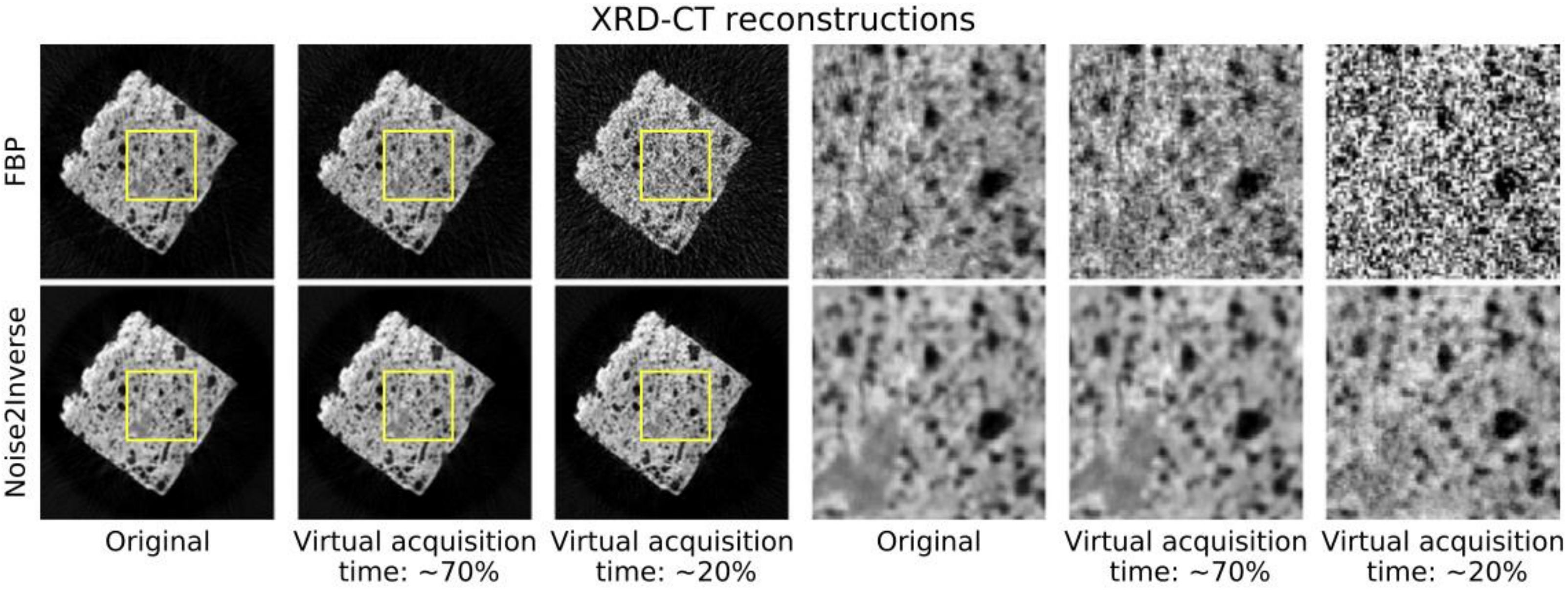


Noise2inverse: similar approach, but combined with the tomography reconstruction (filtered back-projection)

Slide courtesy N. Vigano

Hendriksen et al, *IEEE Transactions on Computational Imaging* 6 (2020), 1320
Hendriksen et al, *Scientific Reports* 11 (2021), 11895

MACHINE LEARNING – DENOISING IMAGES & VOLUMES



Example analysis on diffraction tomography (XRD-CT)

- FBP: classical reconstruction without noise analysis, with datasets at different acquisition time (more noise)
- Noise2inverse: with the 'noise-aware' NN

=> Increase resolution and/or enable low-dose acquisition

Slide courtesy N. Vigano

Hendriksen et al, *IEEE Transactions on Computational Imaging* 6 (2020), 1320
Hendriksen et al, *Scientific Reports* 11 (2021), 11895

1. Design better experiments
2. Faster data processing
3. Better processing
4. **Interpretation (single dataset)**
5. Interpretation (multiple datasets)

SEGMENTATION: MAPPING THE BRAIN CONNECTIONS

- Understand how neurons work together to interpret sensory information and to generate behavior
- Decipher the logic of neural circuits underlying learning and cognition
- Find ways to cope with neurological diseases
- Get inspired to design next generation computing architectures and improve artificial intelligence

How we go about it today

- MRI – probabilistic connections between brain regions (1 voxels contains tens of thousands of neurons)
- Visible light microscopy – sparse information (~ 1 in 500 000 neurons)
- EM – comprehensive mapping but we only have one *Drosophila* brain so far which took a few years of data collection

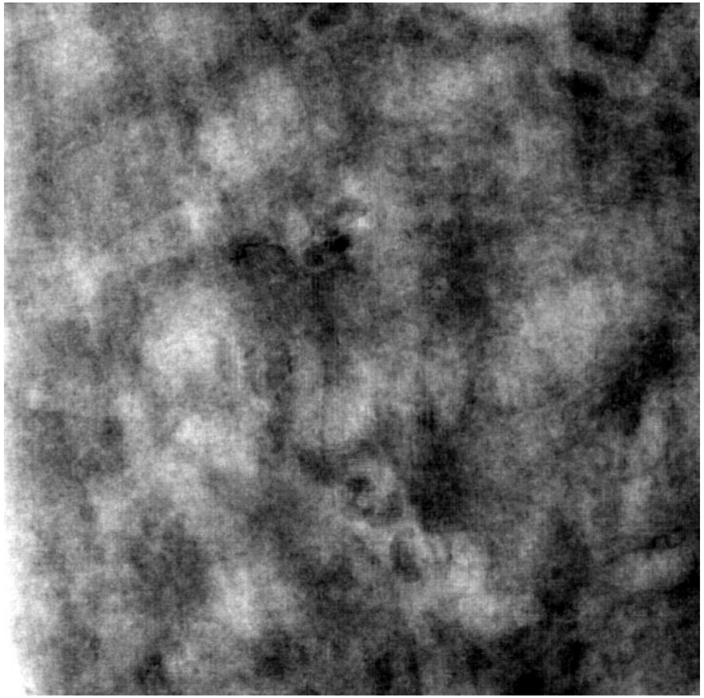
Slide courtesy of A. Pacureanu

HOLO-TOMOGRAPHY: RESOLVING NEURAL NETWORKS

Connectomics in mouse cortex – complementarity with FIB-SEM & TEM
Data collection ~4h (id16A)

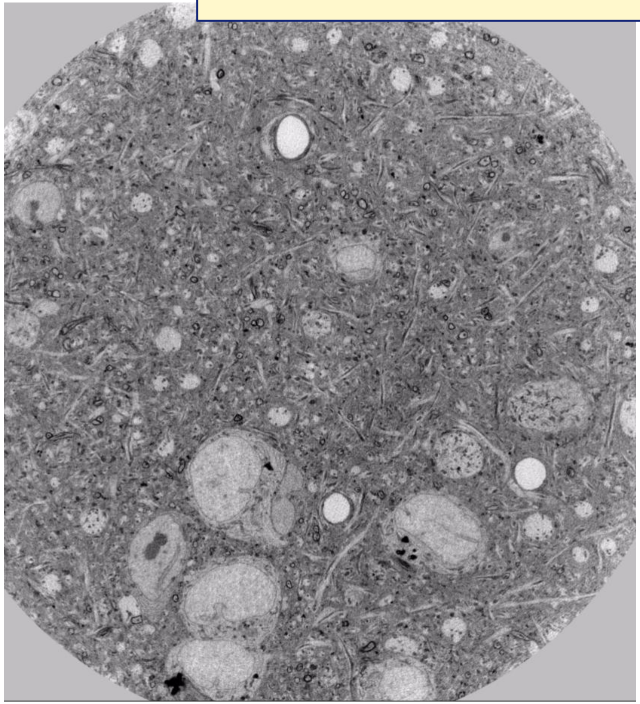
See Alexandra Joita-Pacureanu
ESRF webinar

2K or 4K pixels



Phase maps
(object rotation)

Holo-tomography
reconstruction

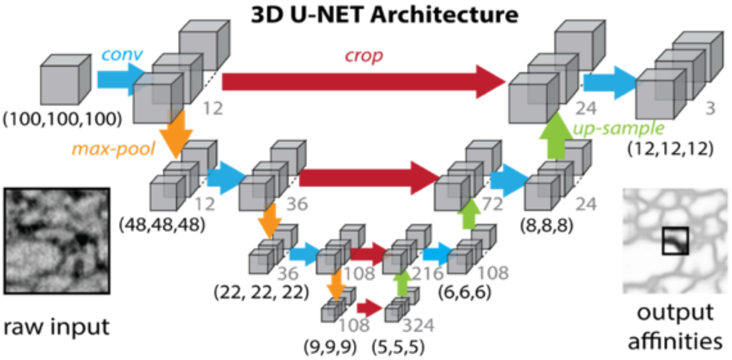


Pixel size 40 nm

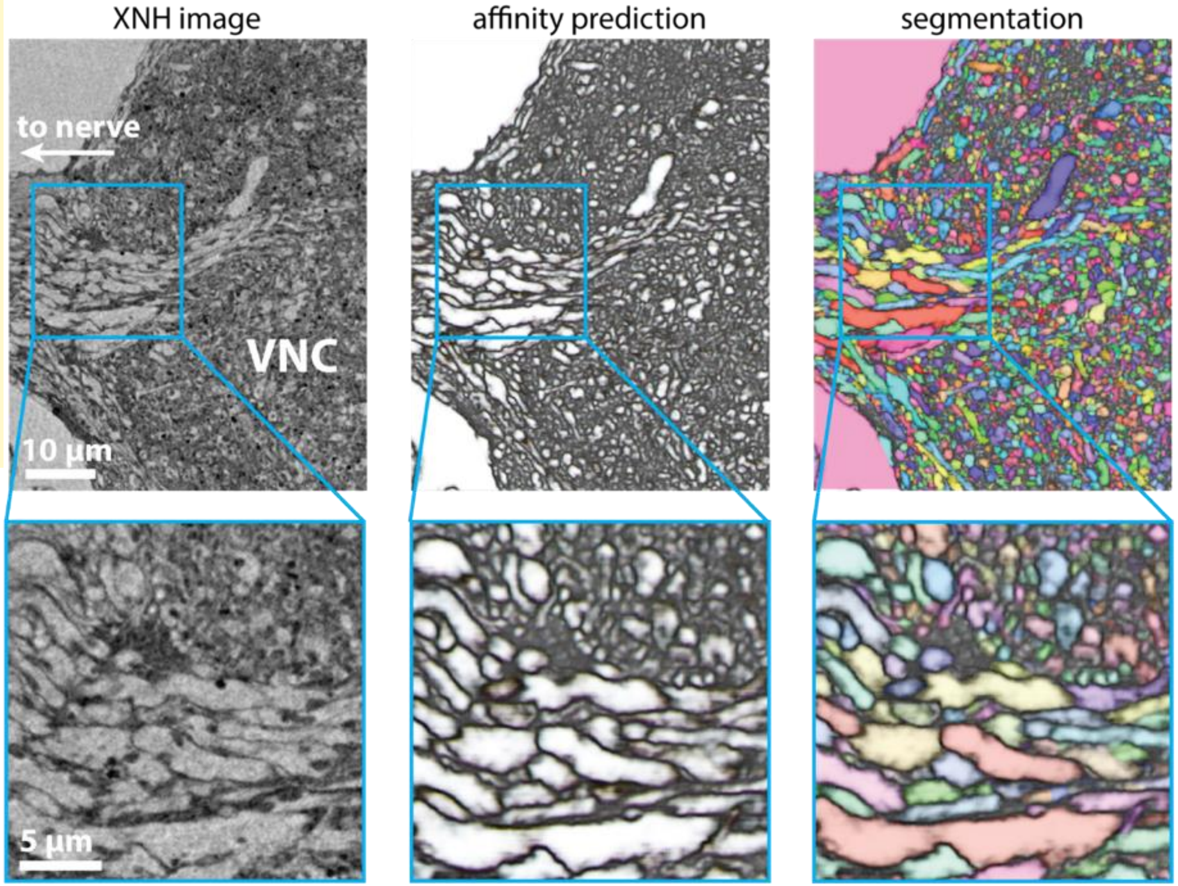
Nat. Neuroscience (2020) & bioRxiv 653188, A Pacureanu
W Lee, A Kuan, J Maniates-Selvin, Harvard Medical School

SEGMENTATION: MAPPING THE BRAIN CONNECTIONS

- Neuronal circuits are very complex and densely packed
- TB sized images
- Manual annotation takes tremendous resources – it took 60 human years to annotate 30% of a fruit fly brain
- Recent developments in deep learning give hope that automatic analysis is feasible



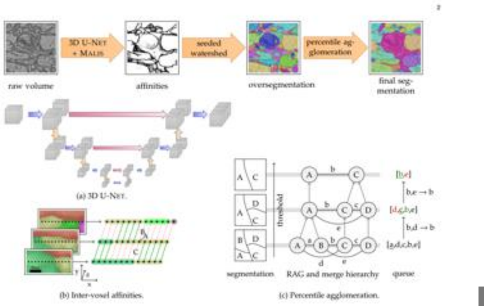
3D U-NET architecture adapted from
Funke et al., TPAMI, 2018



Pacureanu et al. bioRxiv 2019

Slide courtesy of A. Pacureanu

AUTOMATIC SEGMENTATION USING DEEP LEARNING

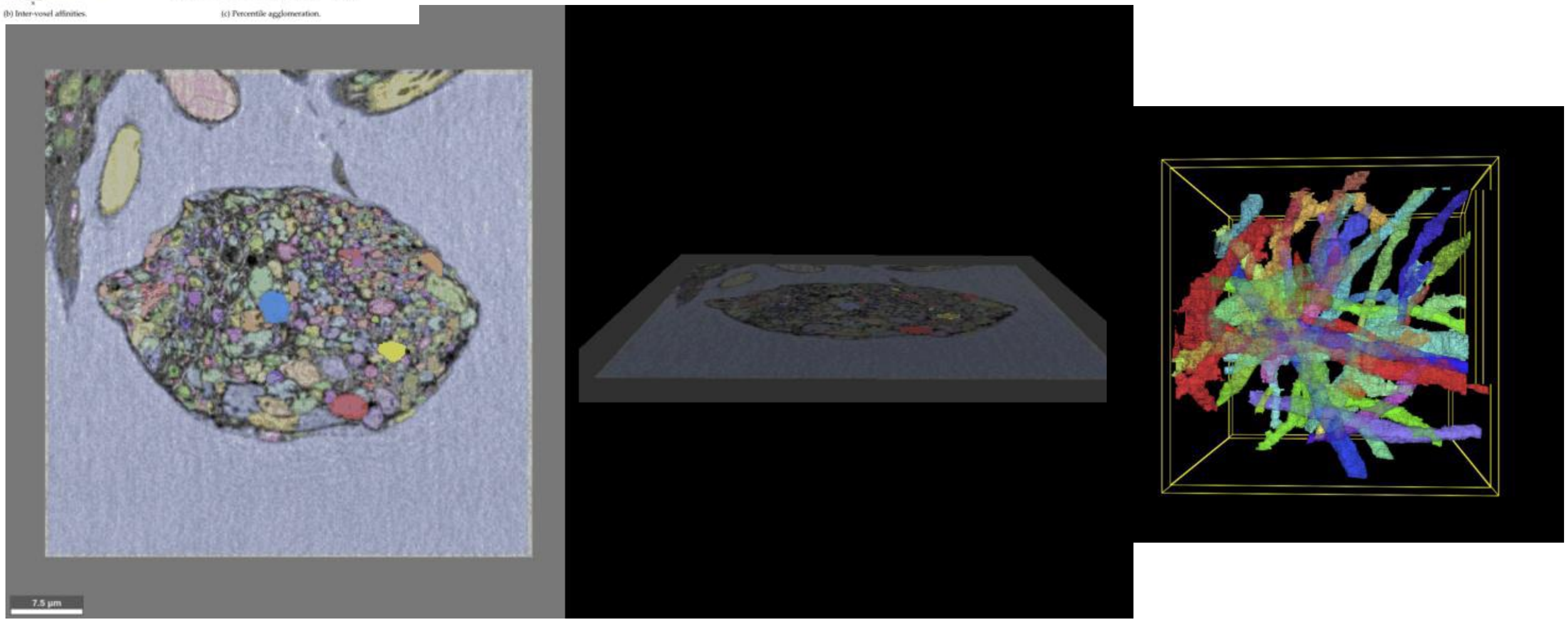


Funke et al. 2018

Deep Learning
Automatic segmentation through convolutional neural networks

VNC: ventral nerve cord

Automated segmentation of a portion of the adult *Drosophila* VNC. Right, volumetric view of selected large-diameter neurons



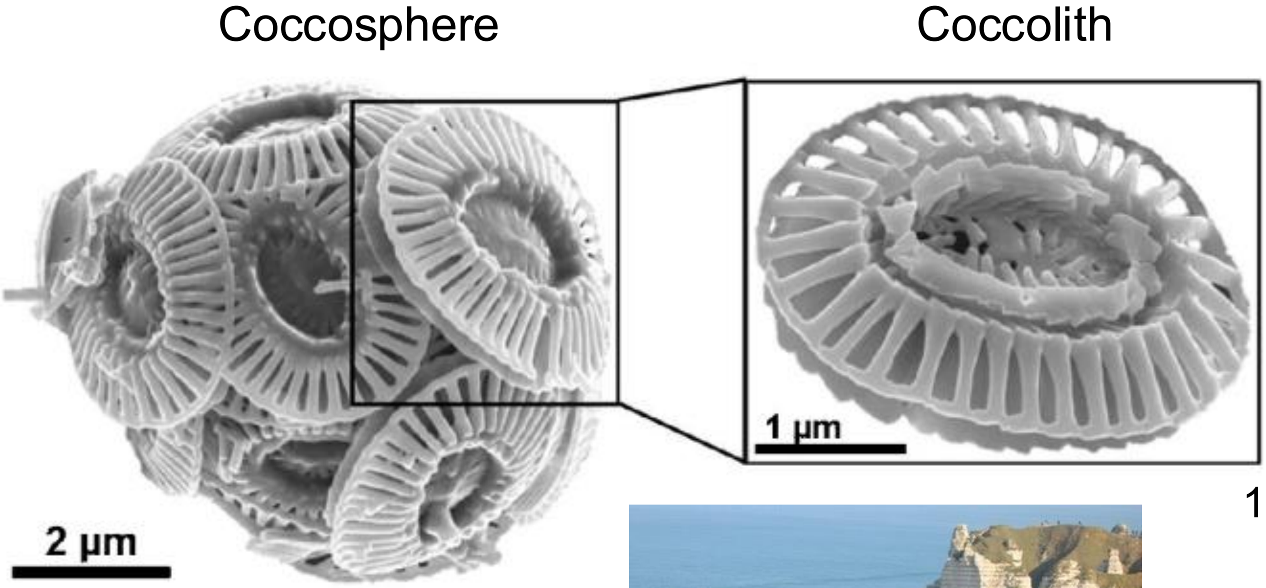
Automatic segmentation in drosophila VNC and mouse cortex

Nat. Neuroscience (2020) & bioRxiv 653188, A Pacureanu
W Lee, A Kuan, J Maniates-Selvin, Harvard Medical School

1. Design better experiments
2. Faster data processing
3. Better processing
4. Interpretation (single dataset)
5. **Multiple datasets, databases…**

MARINE ALGAE - COCCOLITHOPHORES

Emiliania huxleyi



Hoffmann et al (2014)

Production is light dependent

1 per 2 hours

Calcite crystals $CaCO_3$

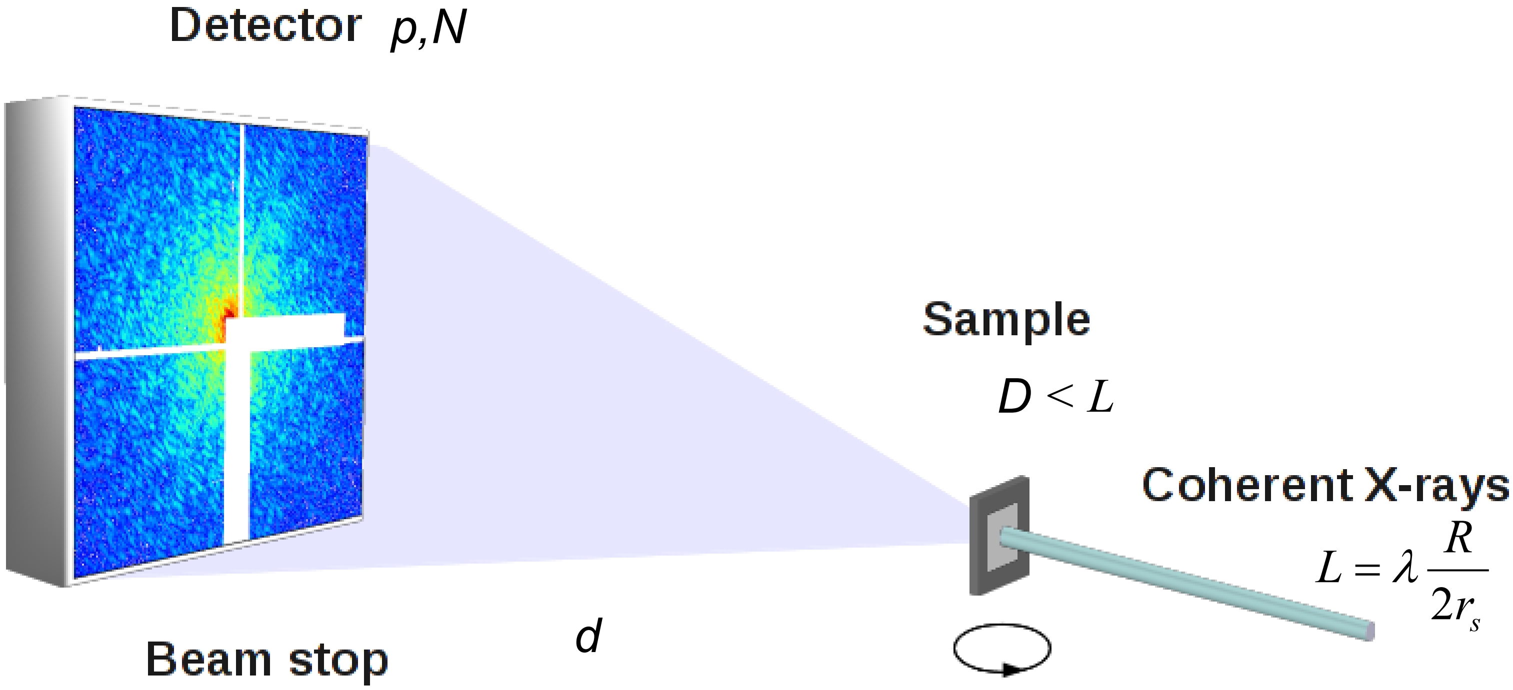
10-20 coccoliths per coccosphere

Most important calcifying organisms on Earth

Beuvier *et al.*, Nature Comm. 10, 751 (2019)
Slide: Y. Chushkin

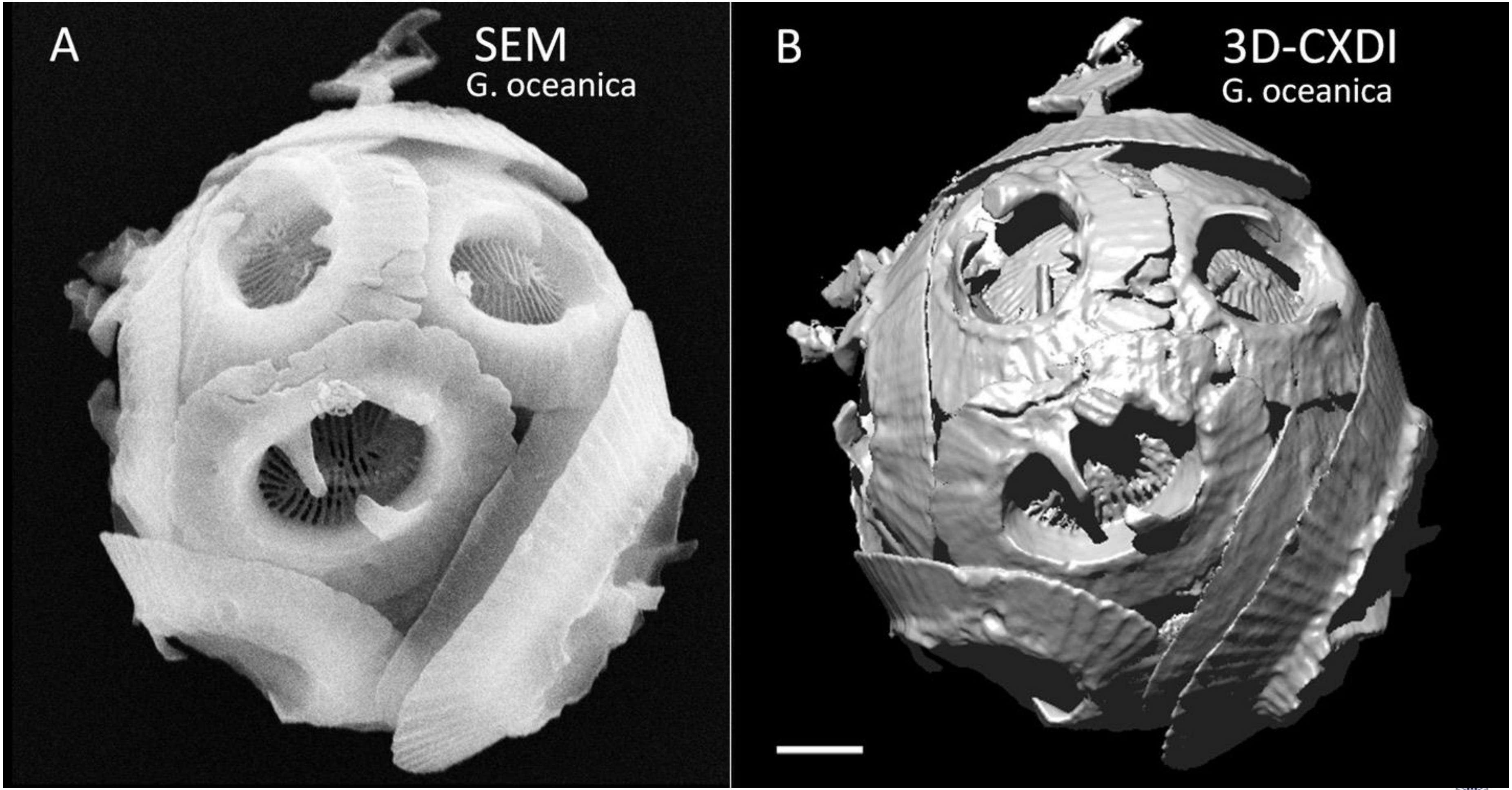
CDI – ID1O BEAMLINE

Isolated sample <7 µm is illuminated with coherent plane wave
Voxel size 5 to 32nm



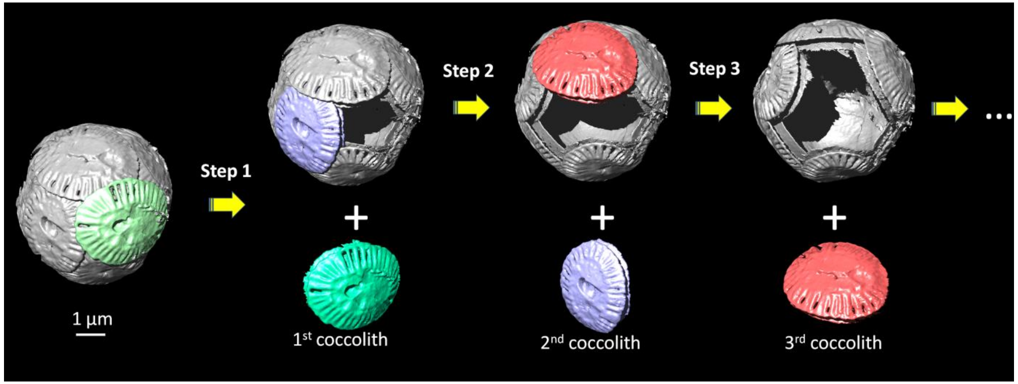
Slide: Y. Chushkin

3D-CXDI



Beuvier *et al.*, Nature Comm. 10, 751 (2019)
Slide: Y. Chushkin

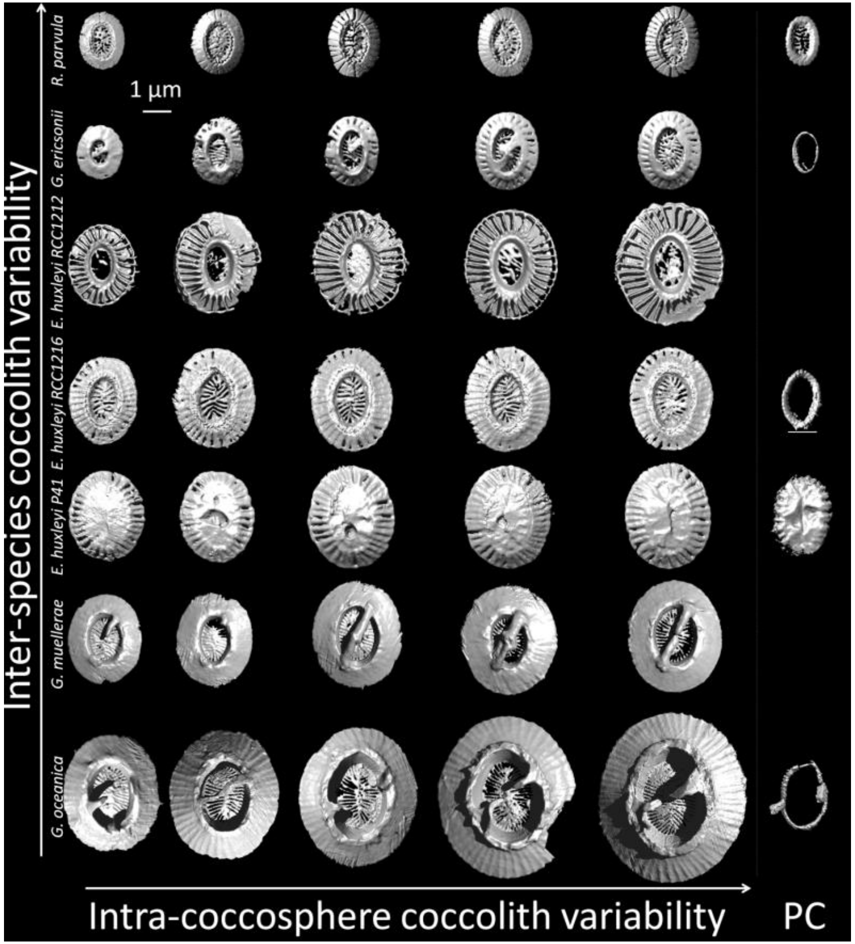
COMPLEX SEGMENTATION



3D coherent imaging of $CaCO_3$ coccospheres (produced by phytoplankton and a large contributor to CO_2 storage) => extraction of a large collection of coccoliths

How do we make sense of large collections of reconstructed objects ?

Similarities, multivariate analysis, principal component analysis… Can we go **beyond model-based statistics** ?



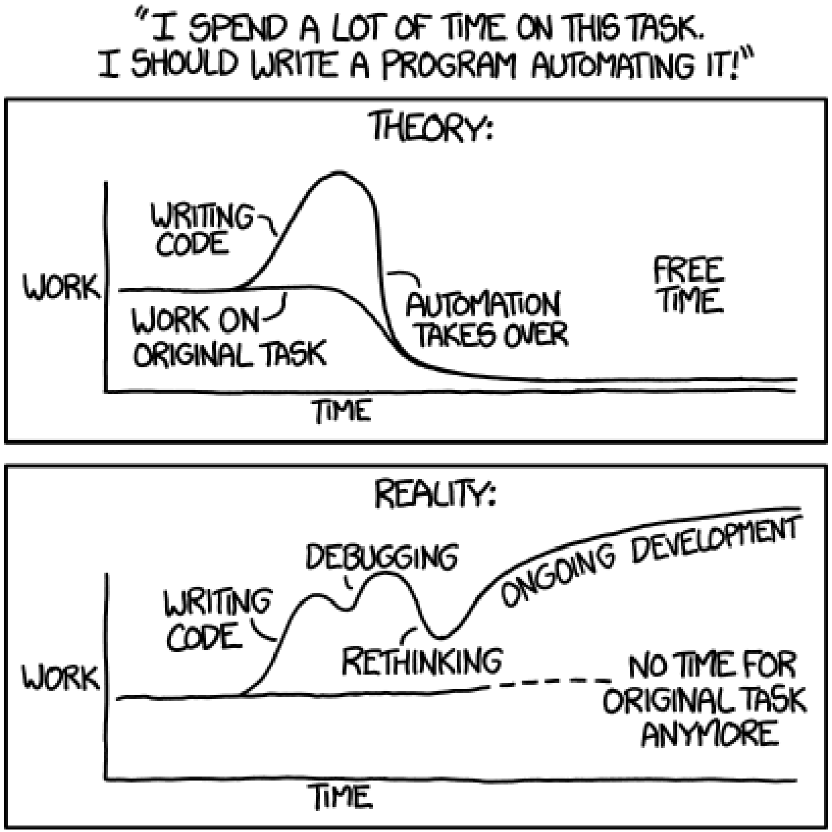
Beuvier et al, Nat. Comm. **10**, 751 (2019).

LIMITS OF MACHINE LEARNING

Machine learning:

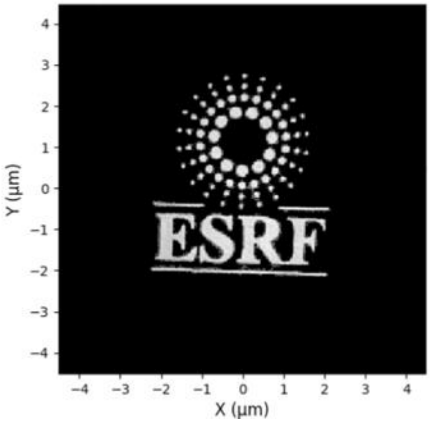
- Better experiments
- Faster data processing
- Better processing (problem-tuned)
- Interpretation

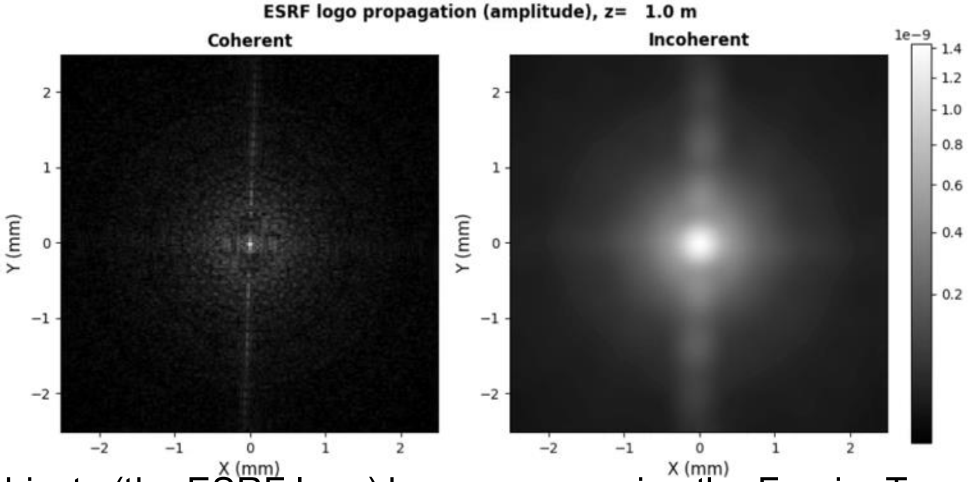
- What should not be done with machine learning ?
- What limits for NN ? Blind spots ?
- What good practices to avoid common mistakes ?



https://xkcd.com/1319/

BIAS IN MACHINE LEARNING





Coherent diffraction allows to reconstruct objects (the ESRF logo) by e.g. measuring the Fourier Transform of the object (which is only blurry with incoherent illumination).

Ask the difference between coherent diffraction and standard diffraction to ML chatbots:

BIAS IN MACHINE LEARNING: CHATGPT VS CHATCGT

Question: what is the difference between classical and coherent diffraction ?

ChatGPT: "(…) The main difference between the two techniques is the **coherence of the beam used**. In standard diffraction, the beam is typically not fully coherent, meaning that **the waves in the beam do not have a well-defined phase relationship with one another**. In contrast, **in coherent diffraction, the beam is fully coherent**, meaning that the waves in the beam have a well-defined phase relationship with one another. This difference in coherence has several important implications for the diffraction patterns obtained and the information that can be obtained from them.

For example, because the phase of the waves in a coherent beam is well-defined, **the diffraction patterns obtained using coherent diffraction are typically much more detailed and informative** than those obtained using standard diffraction. This allows for **higher resolution imaging** and better determination of the atomic structure of the material. Additionally, coherent diffraction allows for the use of phase retrieval algorithms, which can be used to obtain the phase of the scattered waves, and thus the amplitude and phase of the electron density in the material."

ChatCGT (self-professed Marxist AI from the French CGT union): "Coherent diffraction is a concept which **opposes classical diffraction**. Classical diffraction relies on the hypothesis that an object is **composed of small points** which produce waves which propagate and superpose with each other. On the other hand, coherent diffraction is based on the fact that the object is made of **multiple small elements** which have distinct sizes and natures and which interfere with each other. Coherent diffraction is thus a richer and more complex form of diffraction of the object properties. **It is considered as a more "egalitarian" form of diffraction, as it treats every constituent of the object in a more equal manner, contrary to classical diffraction which does not take into account differences**."

OPEN SCIENCE, OPEN DATA

UNESCO Recommendation on Open Science

SCHRIFTEN ZUR
WISSENSCHAFTLICHEN WELTAUFFASSUNG

LOGIK DER
FORSCHUNG

ZUR ERKENNTNISTHEORIE DER
MODERNEN NATURWISSENSCHAFT

VON

KARL POPPER

VERLAG VON JULIUS SPRINGER · 1935

« non-reproducible single occurrences are of no significance to science »

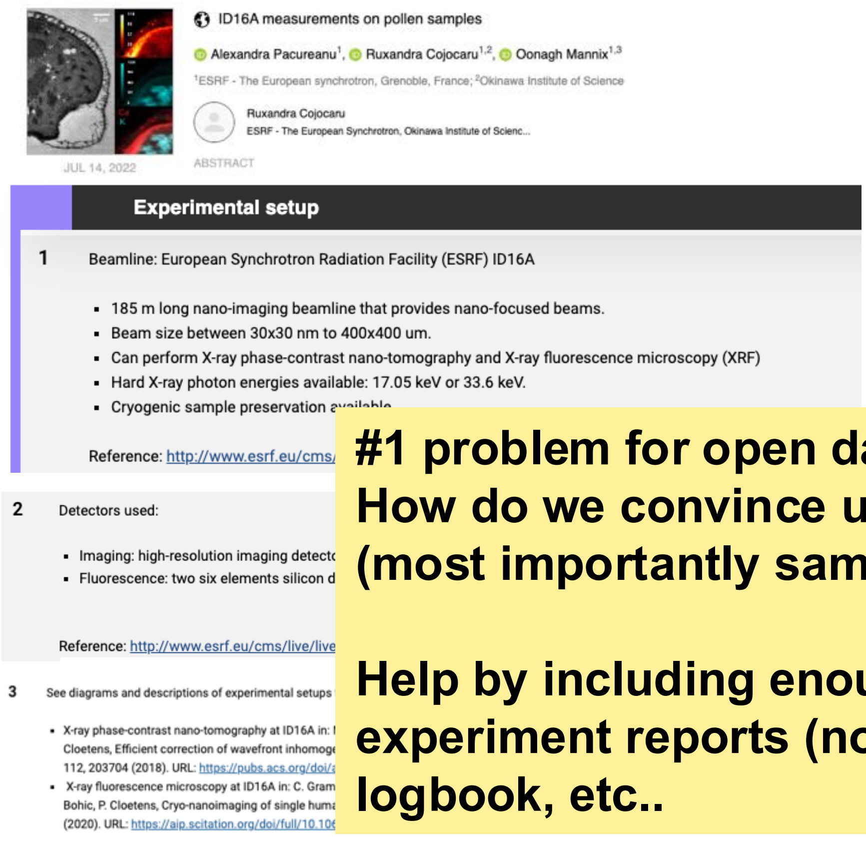
Karl Popper

- The current period is less enlightened that most of us hoped from the *open information* era of the internet: "alternate facts", "filter bubbles", etc…
- We can't solve all those problems…
- But we have a duty to increase the **trust in scientific outputs**
- And a cornerstone for that is open - reliable & FAIR- data

(**Findable, Accessible, Interoperable, Reusable**)

OPEN DATA REQUIRES RICH METADATA – WE NEED FAIR DATA

We can learn from FAIR data (**Findable, Accessible, Interoperable, Reusable**)



ID16A measurements on pollen samples

Alexandra Pacureanu[1], Ruxandra Cojocaru[1,2], Oonagh Mannix[1,3]

[1]ESRF - The European synchrotron, Grenoble, France; [2]Okinawa Institute of Science

Ruxandra Cojocaru
ESRF - The European Synchrotron, Okinawa Institute of Scienc...

JUL 14, 2022

ABSTRACT

Experimental setup

1 Beamline: European Synchrotron Radiation Facility (ESRF) ID16A

- 185 m long nano-imaging beamline that provides nano-focused beams.
- Beam size between 30x30 nm to 400x400 um.
- Can perform X-ray phase-contrast nano-tomography and X-ray fluorescence microscopy (XRF)
- Hard X-ray photon energies available: 17.05 keV or 33.6 keV.
- Cryogenic sample preservation available.

Reference: http://www.esrf.eu/cms/

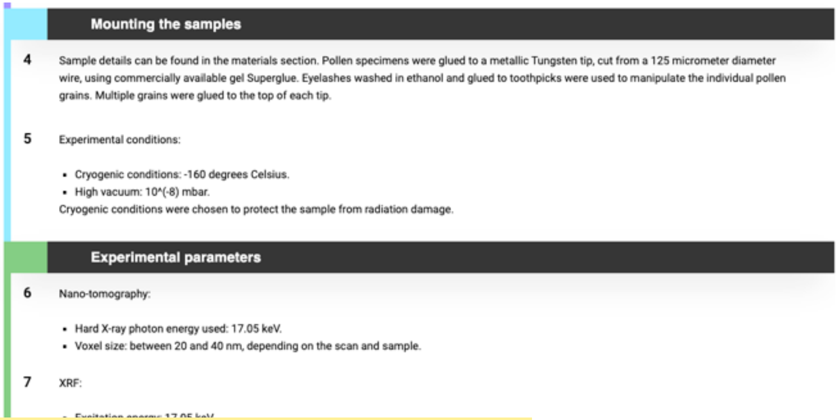
2 Detectors used:

- Imaging: high-resolution imaging detecto
- Fluorescence: two six elements silicon d

Reference: http://www.esrf.eu/cms/live/live

3 See diagrams and descriptions of experimental setups

- X-ray phase-contrast nano-tomography at ID16A in:
Cloetens, Efficient correction of wavefront inhomoge
112, 203704 (2018). URL: https://pubs.acs.org/doi/a
- X-ray fluorescence microscopy at ID16A in: C. Gram
Bohic, P. Cloetens, Cryo-nanoimaging of single huma
(2020). URL: https://aip.scitation.org/doi/full/10.106



Mounting the samples

4 Sample details can be found in the materials section. Pollen specimens were glued to a metallic Tungsten tip, cut from a 125 micrometer diameter wire, using commercially available gel Superglue. Eyelashes washed in ethanol and glued to toothpicks were used to manipulate the individual pollen grains. Multiple grains were glued to the top of each tip.

5 Experimental conditions:

- Cryogenic conditions: -160 degrees Celsius.
- High vacuum: 10^(-8) mbar.

Cryogenic conditions were chosen to protect the sample from radiation damage.

Experimental parameters

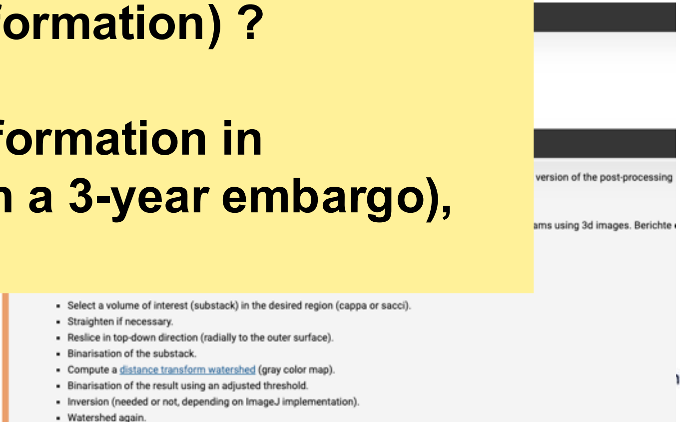
6 Nano-tomography:

- Hard X-ray photon energy used: 17.05 keV.
- Voxel size: between 20 and 40 nm, depending on the scan and sample.

7 XRF:

#1 problem for open data:
How do we convince users to fill in rich metadata (most importantly sample information) ?

Help by including enough information in experiment reports (now with a 3-year embargo), logbook, etc..



version of the post-processing

ams using 3d images. Berichte

- Select a volume of interest (substack) in the desired region (cappa or sacci).
- Straighten if necessary.
- Reslice in top-down direction (radially to the outer surface).
- Binarisation of the substack.
- Compute a distance transform watershed (gray color map).
- Binarisation of the result using an adjusted threshold.
- Inversion (needed or not, depending on ImageJ implementation).
- Watershed again.

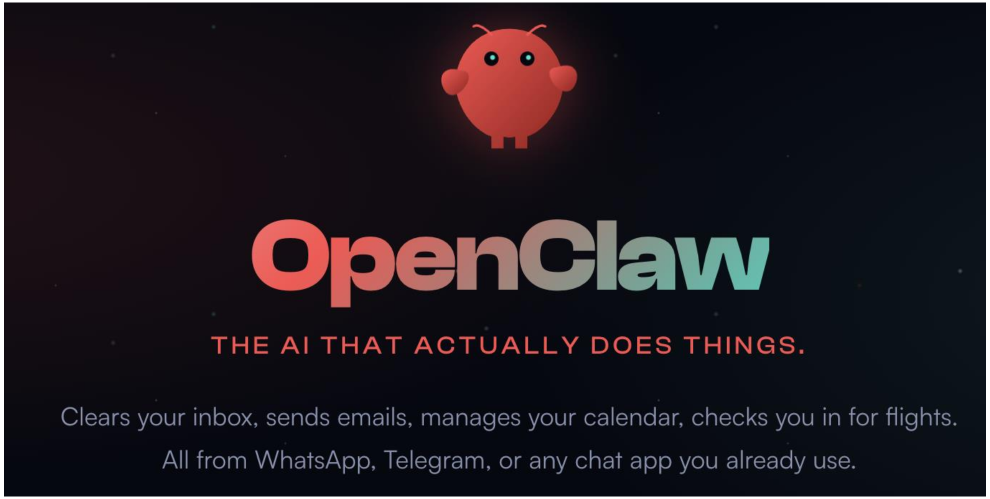
https://www.protocols.io/view/id16a-measurements-on-pollen-samples-bpfwmjpe.html

AI AGENTS – THE (NEAR) FUTURE

*In artificial intelligence, an **intelligent agent** is an entity that **perceives its environment**, takes **actions autonomously** to achieve goals, and may improve its performance through machine learning or by acquiring knowledge. [wikipedia]*

Basically- *your* AI assistant

(*) May include a number of privacy/security issues

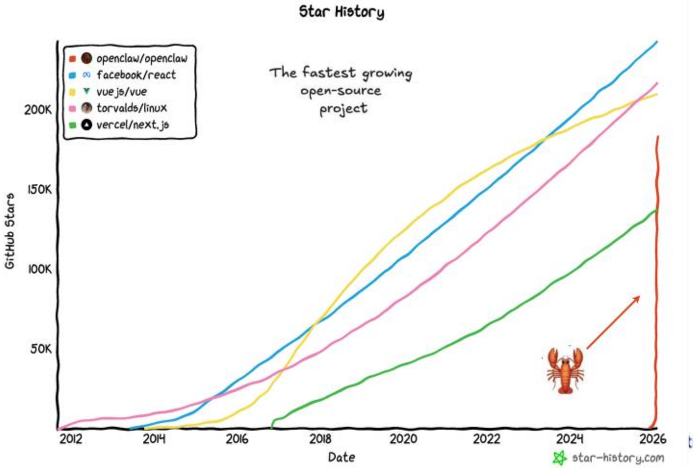


OpenClaw

THE AI THAT ACTUALLY DOES THINGS.

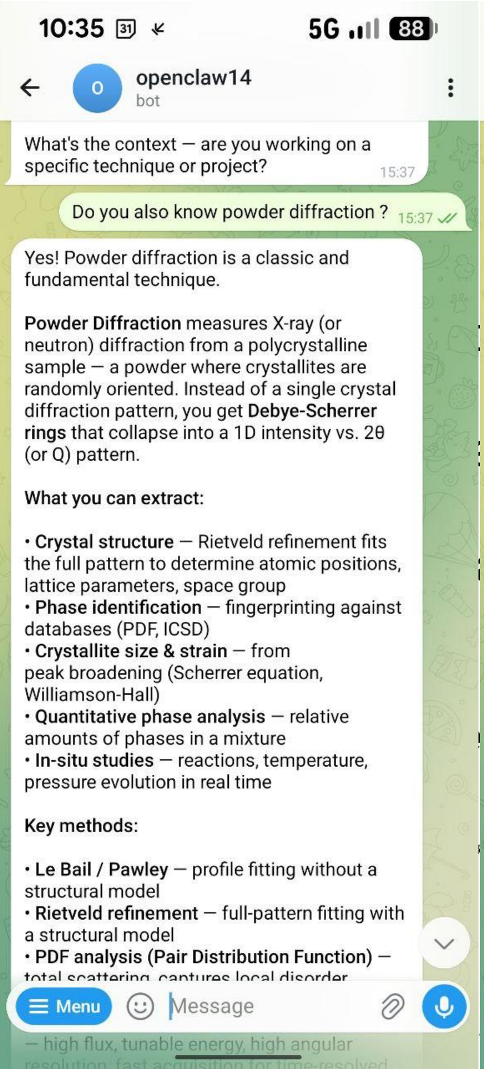
Clears your inbox, sends emails, manages your calendar, checks you in for flights.
All from WhatsApp, Telegram, or any chat app you already use.

https://openclaw.ai/



https://ppaolo.substack.com/p/openclaw-system-architecture-overview

AI AGENTS – THE (NEAR) FUTURE



What's the context — are you working on a specific technique or project? 15:37

Do you also know powder diffraction ? 15:37

Yes! Powder diffraction is a classic and fundamental technique.

Powder Diffraction measures X-ray (or neutron) diffraction from a polycrystalline sample — a powder where crystallites are randomly oriented. Instead of a single crystal diffraction pattern, you get **Debye-Scherrer rings** that collapse into a 1D intensity vs. 2θ (or Q) pattern.

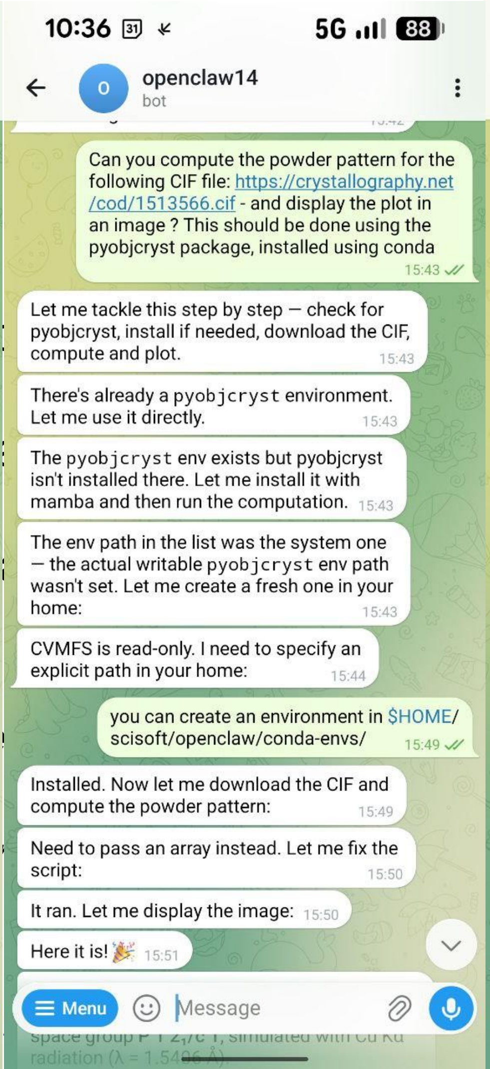
What you can extract:

- **Crystal structure** — Rietveld refinement fits the full pattern to determine atomic positions, lattice parameters, space group
- **Phase identification** — fingerprinting against databases (PDF, ICSD)
- **Crystallite size & strain** — from peak broadening (Scherrer equation, Williamson-Hall)
- **Quantitative phase analysis** — relative amounts of phases in a mixture
- **In-situ studies** — reactions, temperature, pressure evolution in real time

Key methods:

- **Le Bail / Pawley** — profile fitting without a structural model
- **Rietveld refinement** — full-pattern fitting with a structural model
- **PDF analysis (Pair Distribution Function)** — total scattering, captures local disorder

— high flux, tunable energy, high angular resolution, fast acquisition for time-resolved



Can you compute the powder pattern for the following CIF file: https://crystallography.net/cod/1513566.cif - and display the plot in an image ? This should be done using the pyobjcryst package, installed using conda 15:43

Let me tackle this step by step — check for pyobjcryst, install if needed, download the CIF, compute and plot. 15:43

There's already a `pyobjcryst` environment. Let me use it directly. 15:43

The `pyobjcryst` env exists but pyobjcryst isn't installed there. Let me install it with mamba and then run the computation. 15:43

The env path in the list was the system one — the actual writable `pyobjcryst` env path wasn't set. Let me create a fresh one in your home: 15:43

CVMFS is read-only. I need to specify an explicit path in your home: 15:44

you can create an environment in $HOME/scisoft/openclaw/conda-envs/ 15:49

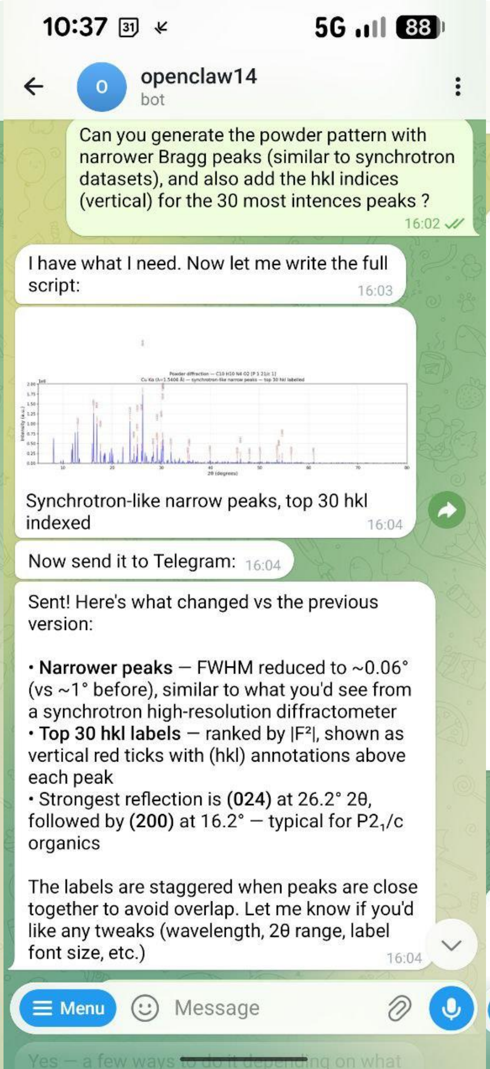
Installed. Now let me download the CIF and compute the powder pattern: 15:49

Need to pass an array instead. Let me fix the script: 15:50

It ran. Let me display the image: 15:50

Here it is! 🎉 15:51

space group P 1 2_1/c 1, simulated with Cu Kα radiation (λ = 1.5406 Å).



Can you generate the powder pattern with narrower Bragg peaks (similar to synchrotron datasets), and also add the hkl indices (vertical) for the 30 most intences peaks ? 16:02

I have what I need. Now let me write the full script: 16:03

Synchrotron-like narrow peaks, top 30 hkl indexed 16:04

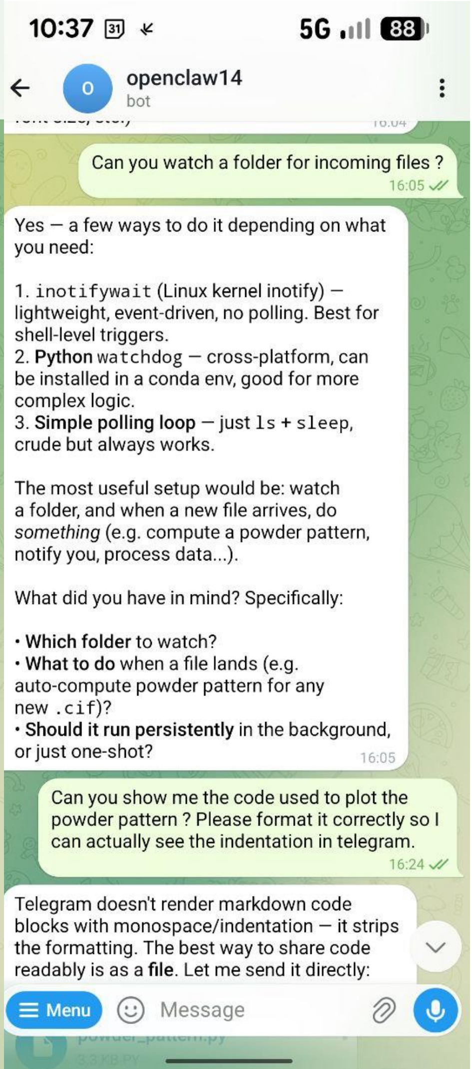
Now send it to Telegram: 16:04

Sent! Here's what changed vs the previous version:

- **Narrower peaks** — FWHM reduced to ~0.06° (vs ~1° before), similar to what you'd see from a synchrotron high-resolution diffractometer
- **Top 30 hkl labels** — ranked by |F²|, shown as vertical red ticks with (hkl) annotations above each peak
- Strongest reflection is **(024)** at 26.2° 2θ, followed by **(200)** at 16.2° — typical for P2_1/c organics

The labels are staggered when peaks are close together to avoid overlap. Let me know if you'd like any tweaks (wavelength, 2θ range, label font size, etc.) 16:04

Yes — a few ways to do it depending on what



Can you watch a folder for incoming files ? 16:05

Yes — a few ways to do it depending on what you need:

1. `inotifywait` (Linux kernel inotify) — lightweight, event-driven, no polling. Best for shell-level triggers.
2. **Python** `watchdog` — cross-platform, can be installed in a conda env, good for more complex logic.
3. **Simple polling loop** — just `ls` + `sleep`, crude but always works.

The most useful setup would be: watch a folder, and when a new file arrives, do *something* (e.g. compute a powder pattern, notify you, process data...).

What did you have in mind? Specifically:

- **Which folder** to watch?
- **What to do** when a file lands (e.g. auto-compute powder pattern for any new `.cif`)?
- **Should it run persistently** in the background, or just one-shot? 16:05

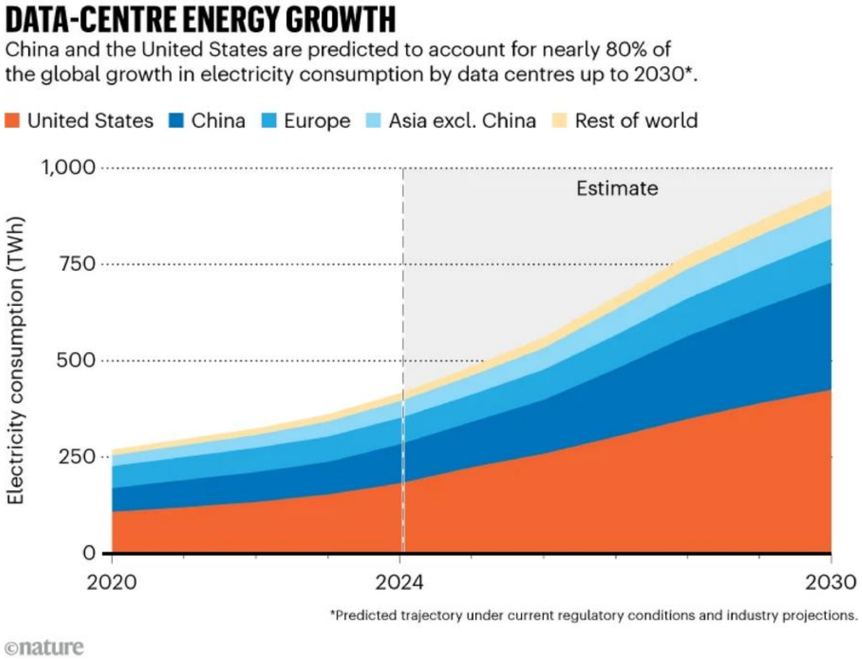
Can you show me the code used to plot the powder pattern ? Please format it correctly so I can actually see the indentation in telegram. 16:24

Telegram doesn't render markdown code blocks with monospace/indentation — it strips the formatting. The best way to share code readably is as a **file**. Let me send it directly:

powder_pattern.py

AI/DEEP LEARNING & ETHICS IN SCIENCE

- If you **use AI to do your research work**, are you cheating ?
  - To **write reports ? Articles ?**
    - Today = not acceptable
  - To **analyse data** ?
    - If you are in full control of the analysis steps, OK !
    - What if you get AI Agents to do most of the work for you ?
    - Are you still a scientist ?
  - Essential: be transparent about what AI tools you used
- Whatever the answer – remember **Darwin's natural selection**:
  - Practice which will dominate in a few years will be based on *results* (scientific discoveries)
- What about the **environment** ? Huge energy consumption from AI..



DATA-CENTRE ENERGY GROWTH

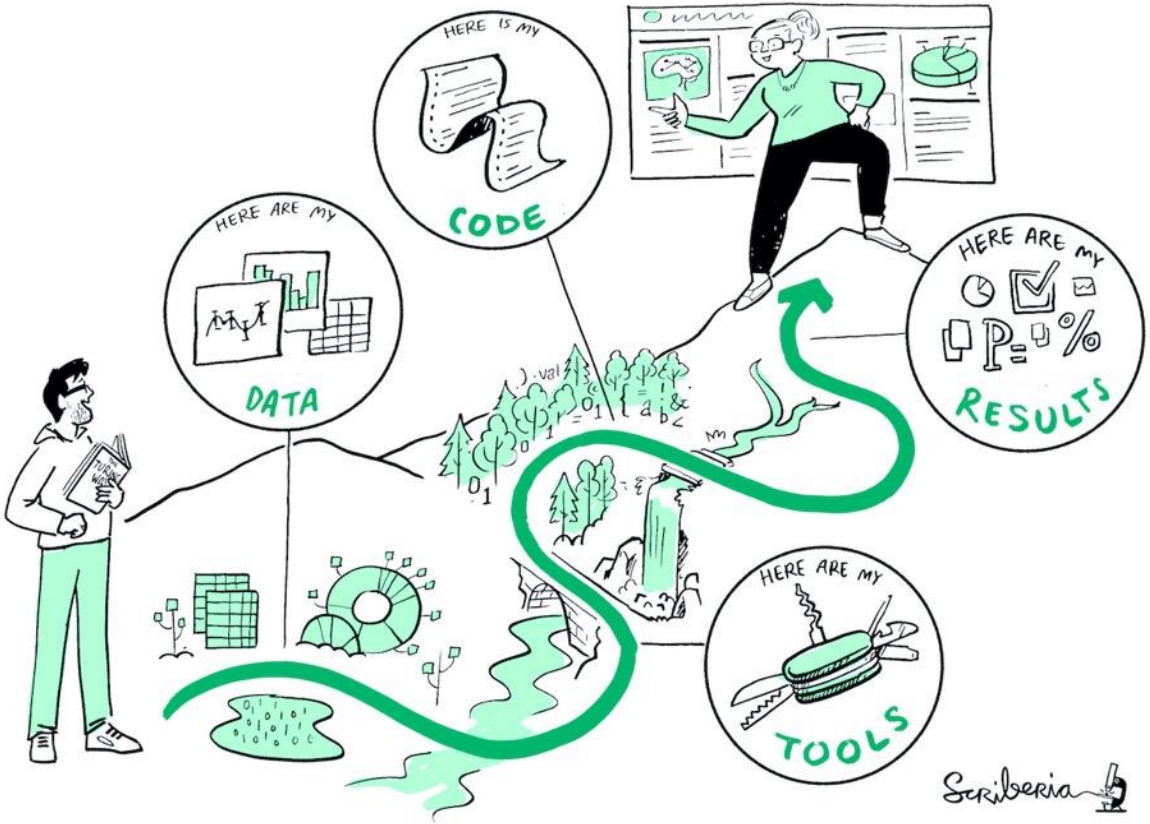
China and the United States are predicted to account for nearly 80% of the global growth in electricity consumption by data centres up to 2030*.

*Predicted trajectory under current regulatory conditions and industry projections.

©nature

Projected energy consumption from data center => 1 PWh in 2030 (=annual electricity consumption for Japan)

THANK YOU FOR LISTENING



Questions ?

Illustration: https://zenodo.org/record/7587336