

DATA SCIENCE FOR LARGE-SCALE SCIENTIFIC INFRASTRUCTURES

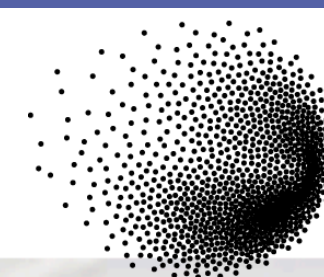
Luis Barba



Large-scale Scientific Infrastructures



SDSC



PSI

ETH zürich

EPFL



Swiss Neutron Sources SINQ and UCN



Swiss Light Source SLS

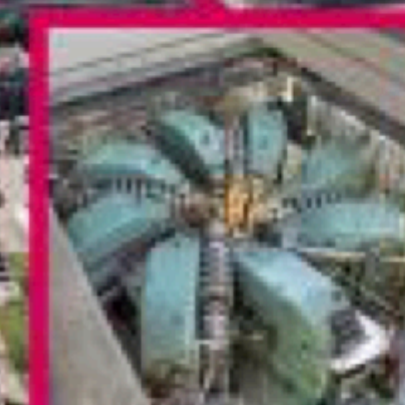


Swiss Free Electron Laser SwissFEL



Energy System Integration

Proton Therapy



Swiss Muon Source SμS



High-precision Particle Physics CHRISP

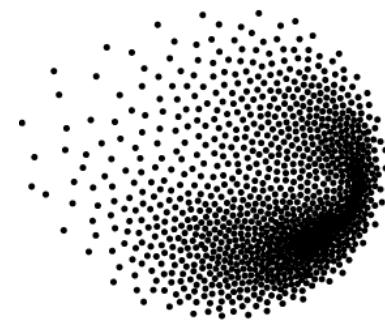
Hotlab

Radiopharmacy

PSI East

PSI West

Data Challenges in PSI and SDSC



PSI projects

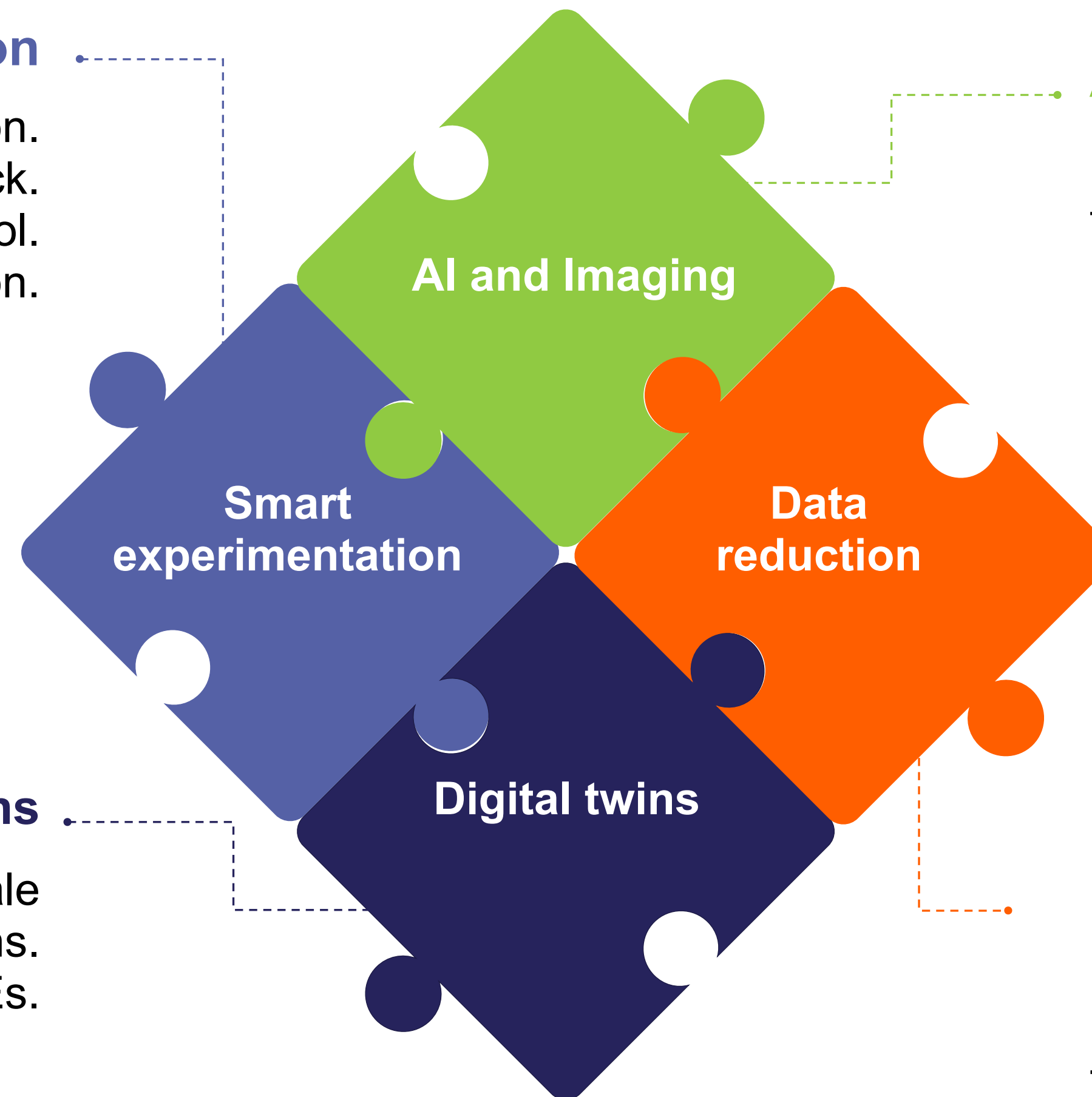


Smart experimentation

Smart/guided signal acquisition.
Real-time feedback.
Autonomous experimental control.
Uncertainty quantification.

AI and Imaging

Imaging reconstruction (e.g., X-ray tomography, ptychography).
Phase-retrieval, super-resolution.
New sensing modalities.
Data-driven inverse problem modeling.



Digital twins

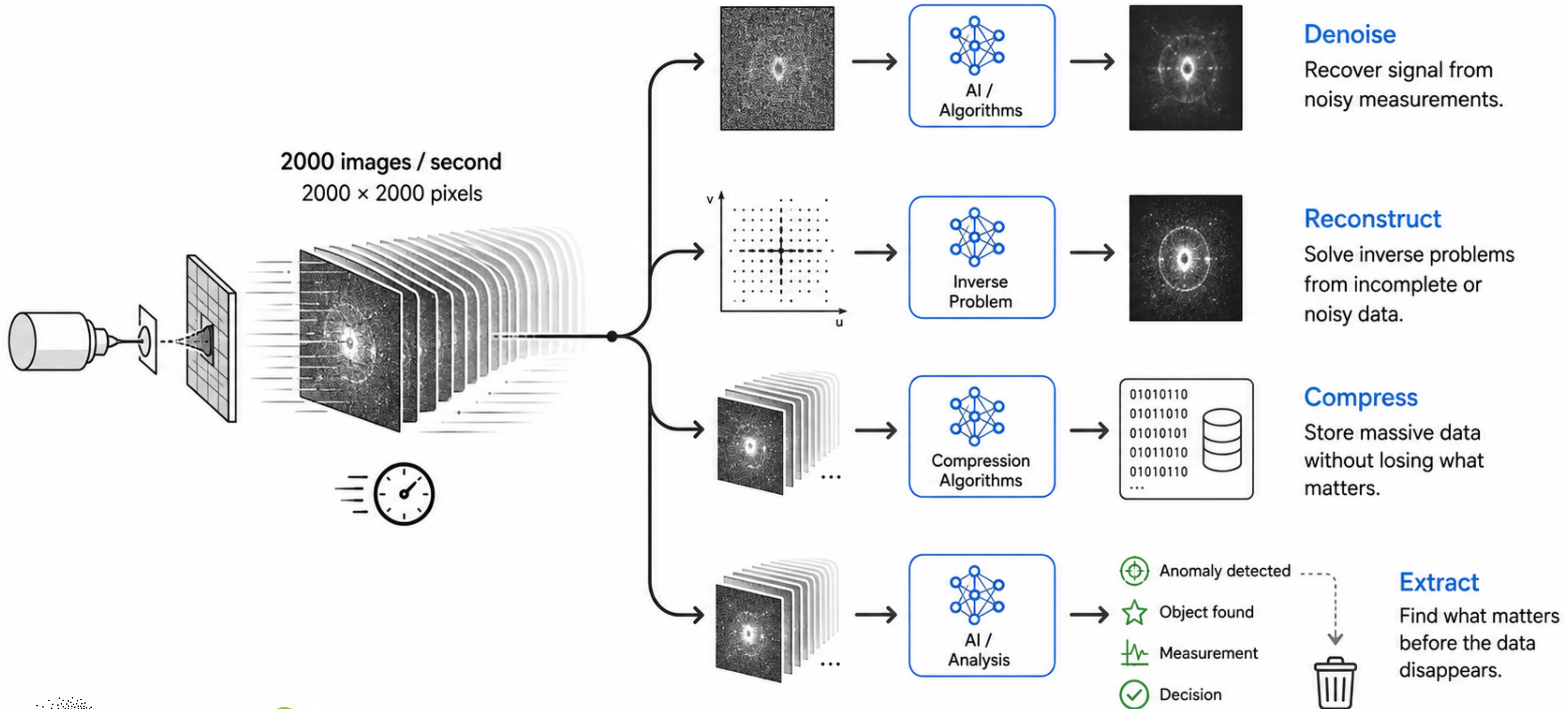
Approximation methods for large-scale simulations.
Physically-inspired NN, NeuralODEs.

Data reduction

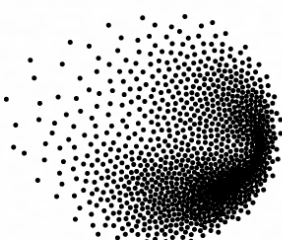
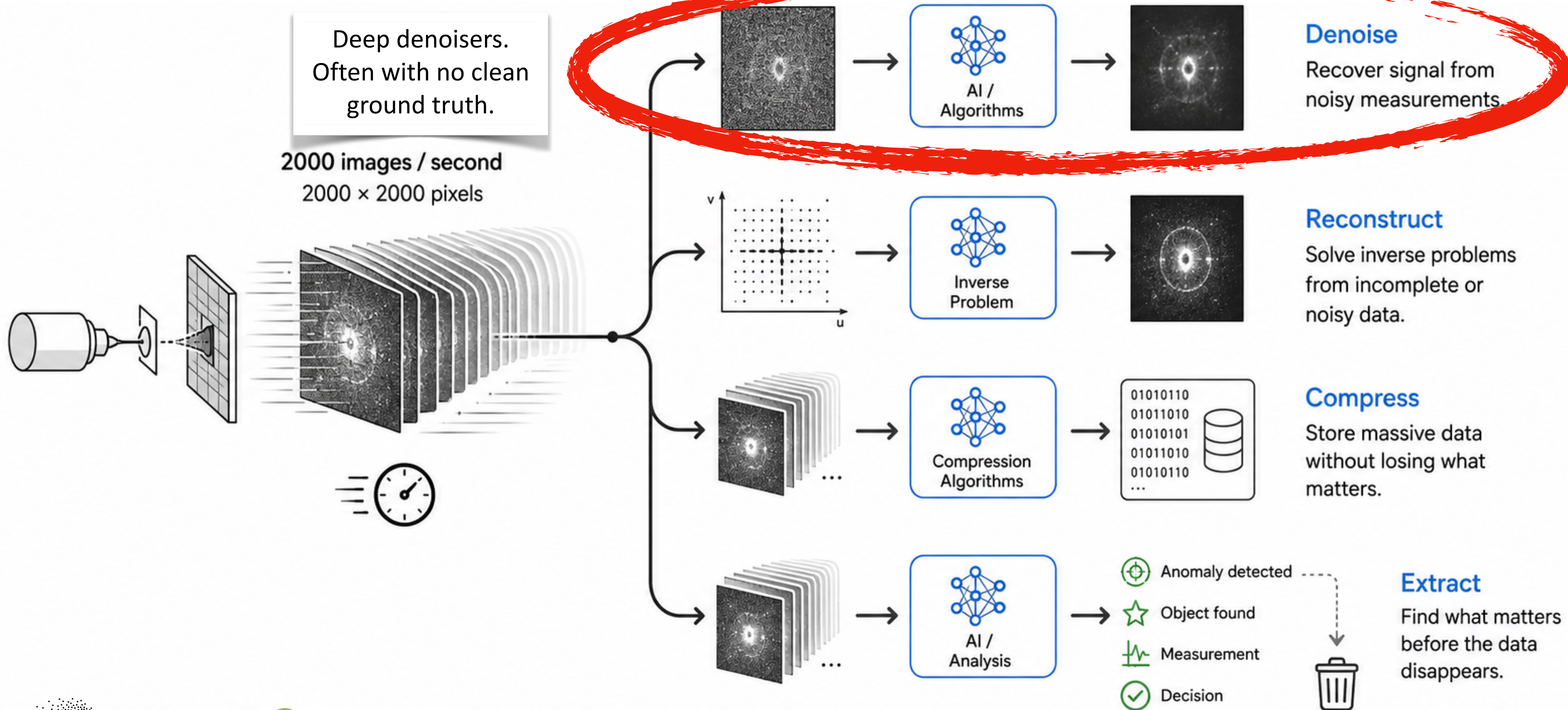
Data reduction and data compression (quantization, rate-distortion).
Time-series modeling.
Real-time data intake and (pre-)processing.



Data Challenges in PSI and SDSC



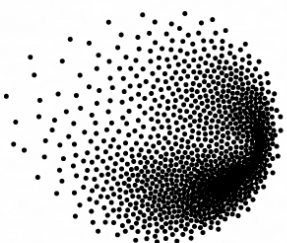
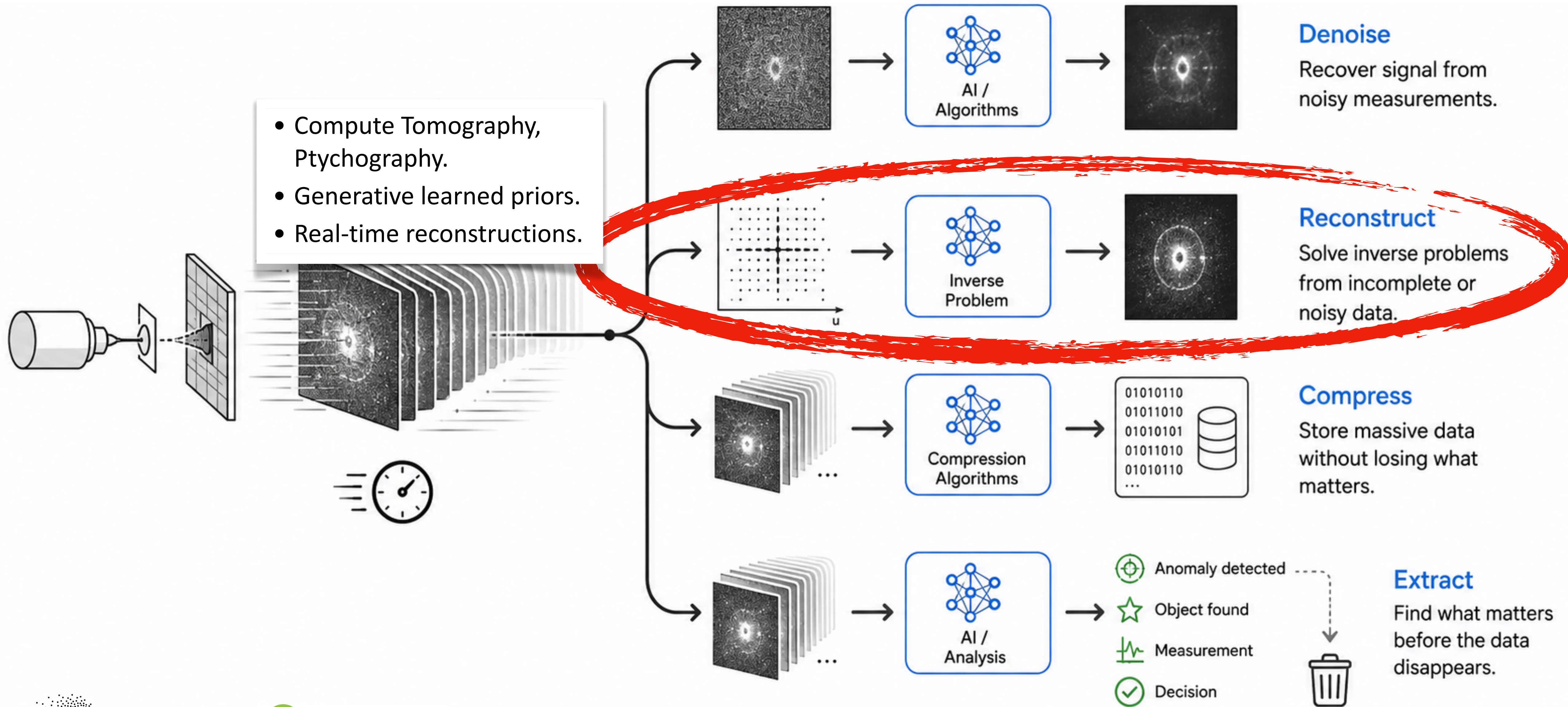
Data Challenges in PSI and SDSC



PSI



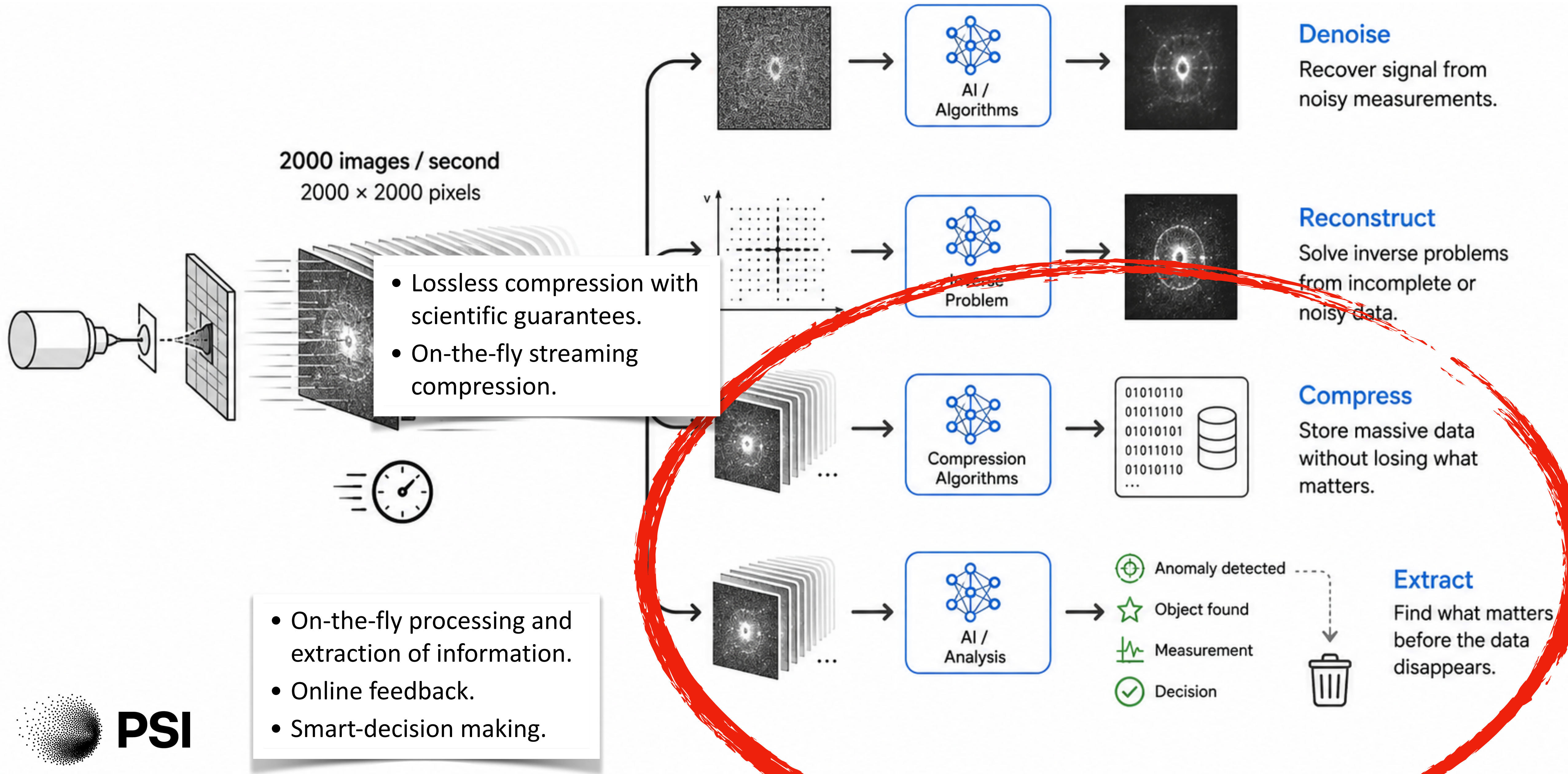
Data Challenges in PSI and SDSC



PSI



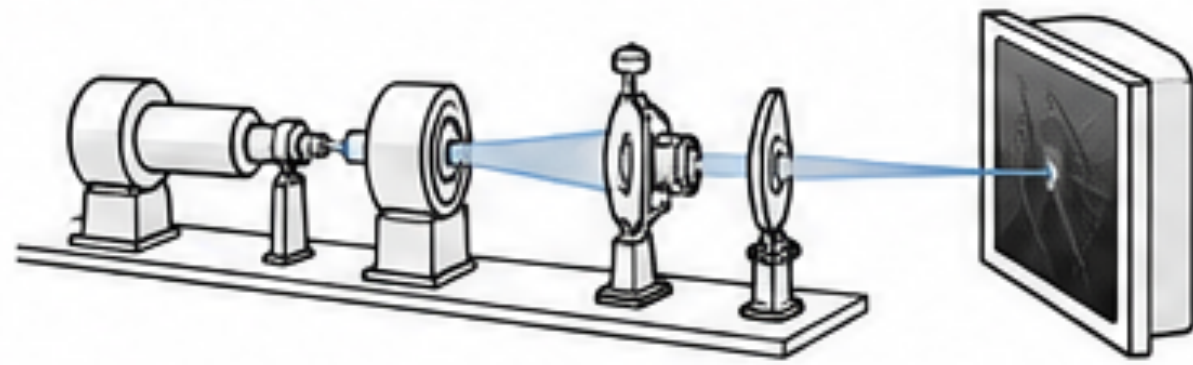
Data Challenges in PSI and SDSC



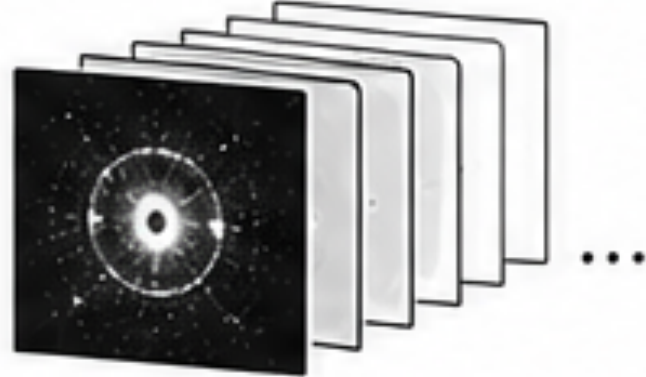
PSI and SKAO: Similar problems

PSI / Experimental Science

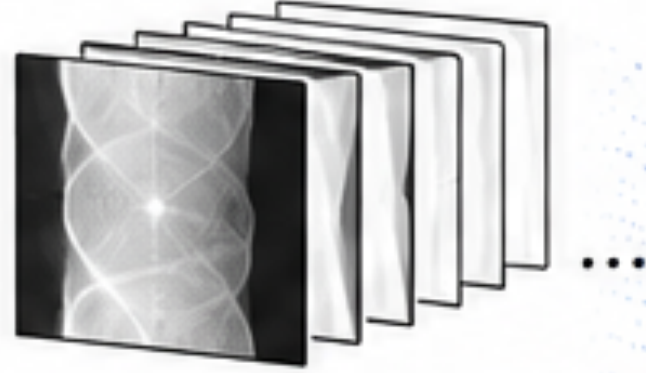
Examples of measurements



Diffraction patterns



Projections / sinograms



High-speed detector streams



High-rate experimental measurements

Different measurements. Similar questions.

Improve the measurement?

Denoising · Calibration

Reconstruct what we cannot directly observe?

Inverse problems · Imaging

Reduce the amount of data while preserving what matters?

Compression · Data reduction

Extract useful information efficiently?

Detection · Classification · Discovery

SKAO / Radio Astronomy

Examples of measurements



SKA-LOW

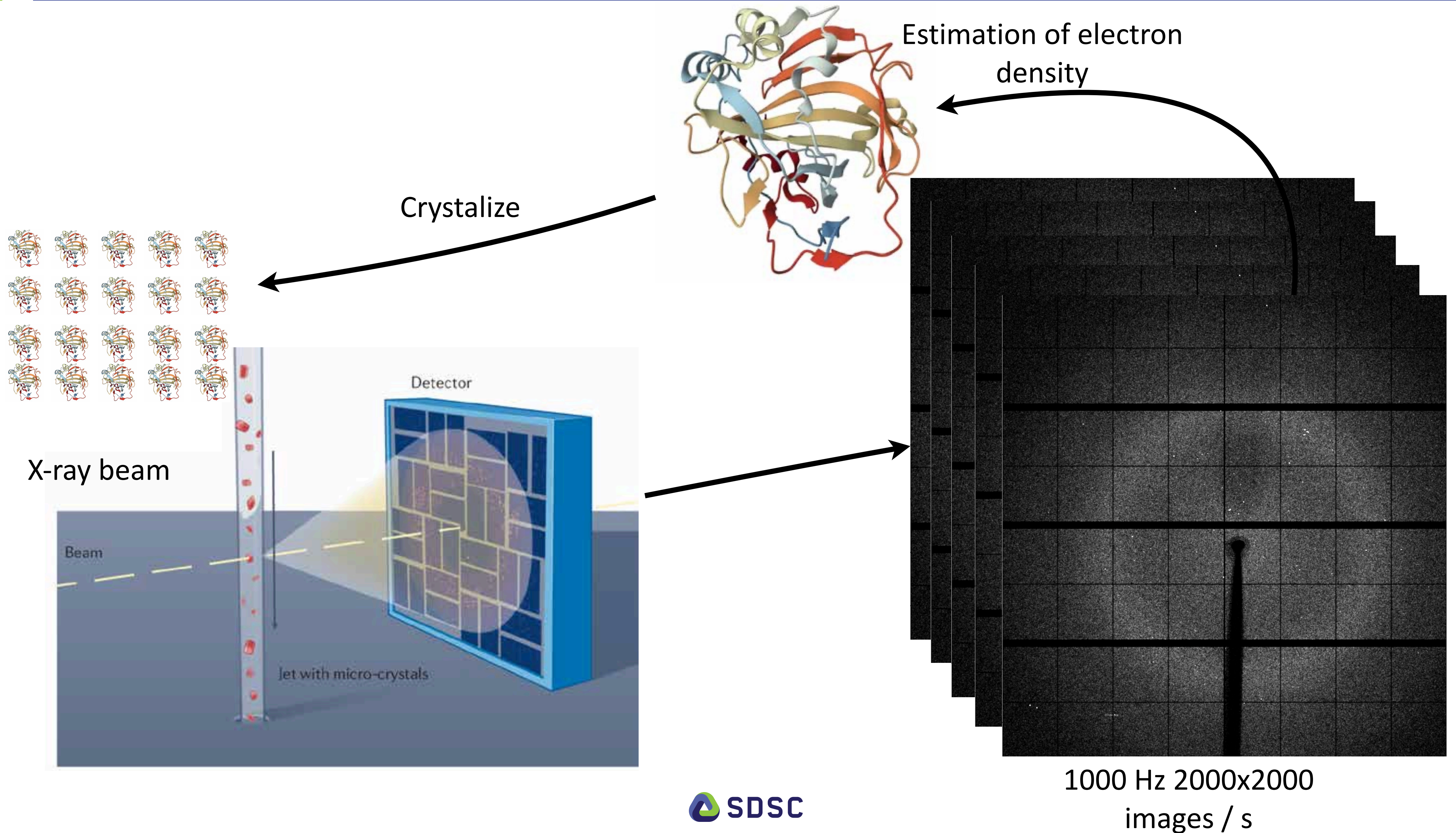
Antenna array

SKA-MID

Dish array

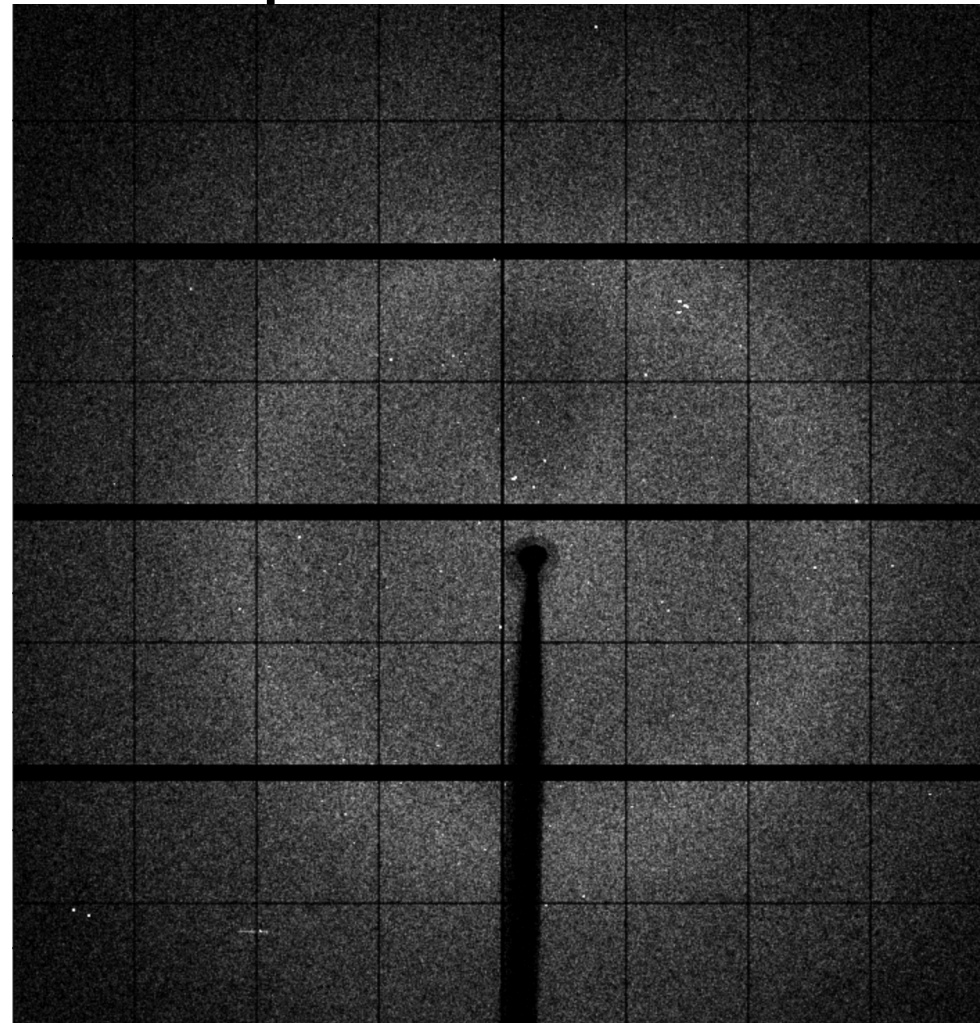
High-rate astronomical measurements
(visibilities / correlations)

X-ray Crystallography



X-ray Crystallography

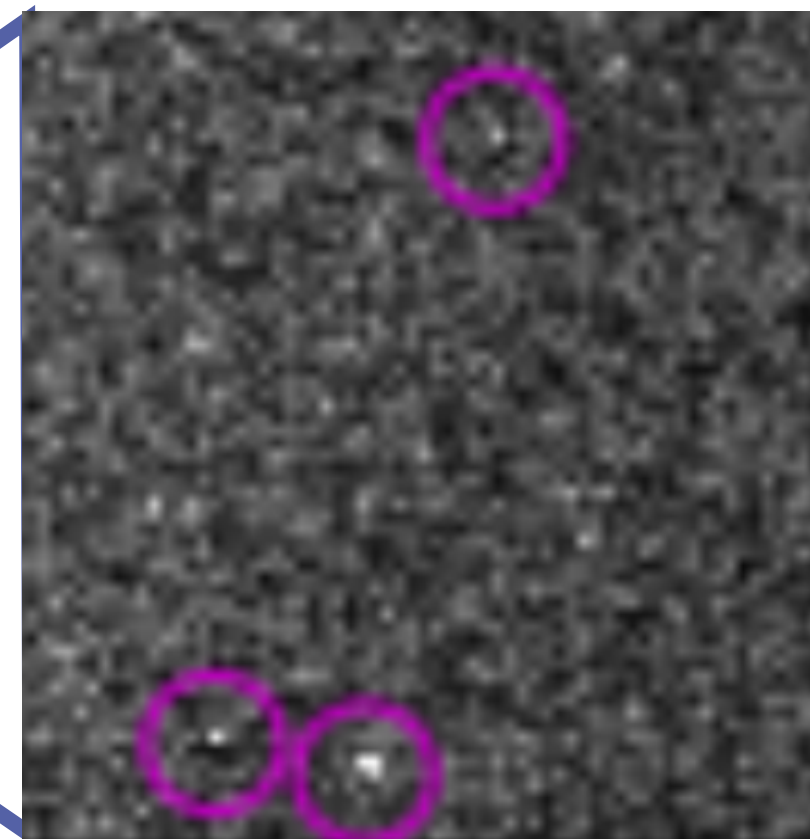
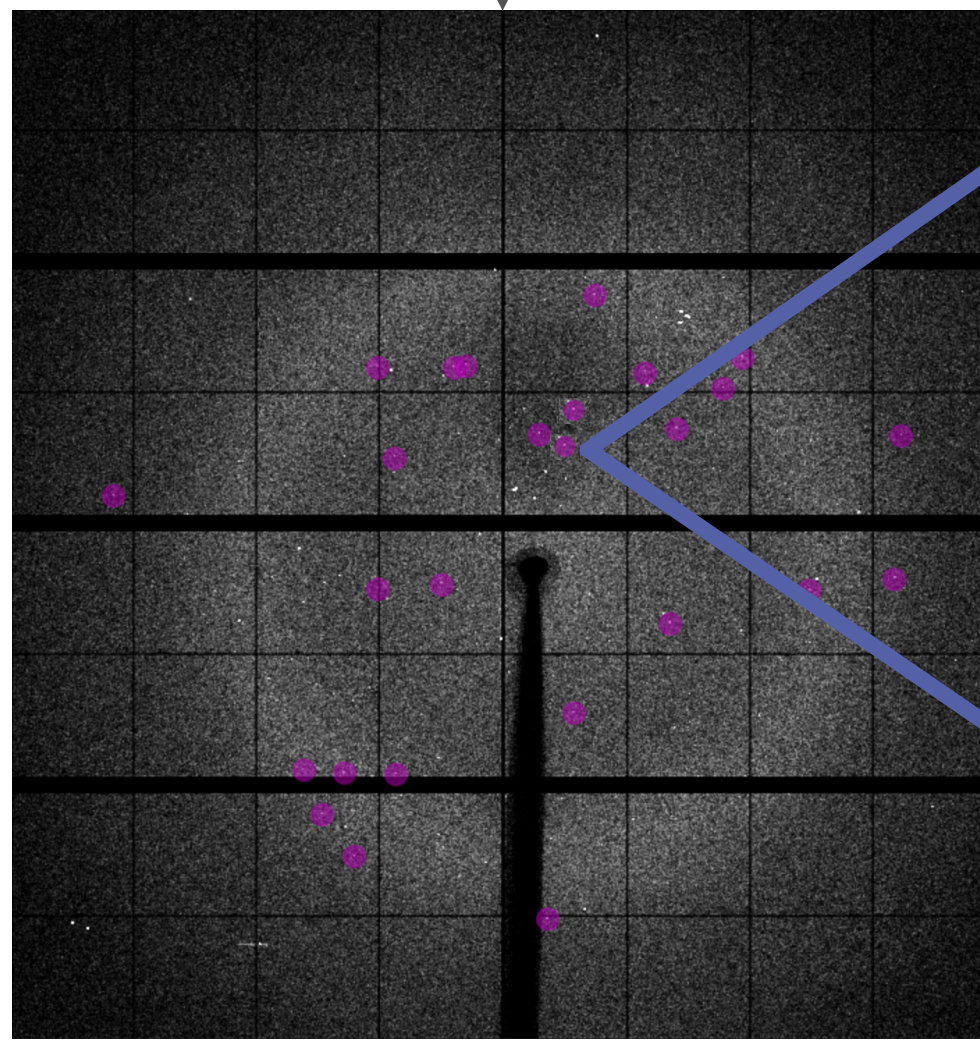
Spot detection



Indexing: identifies each spot and predicts locations that contain useful information (red circles)

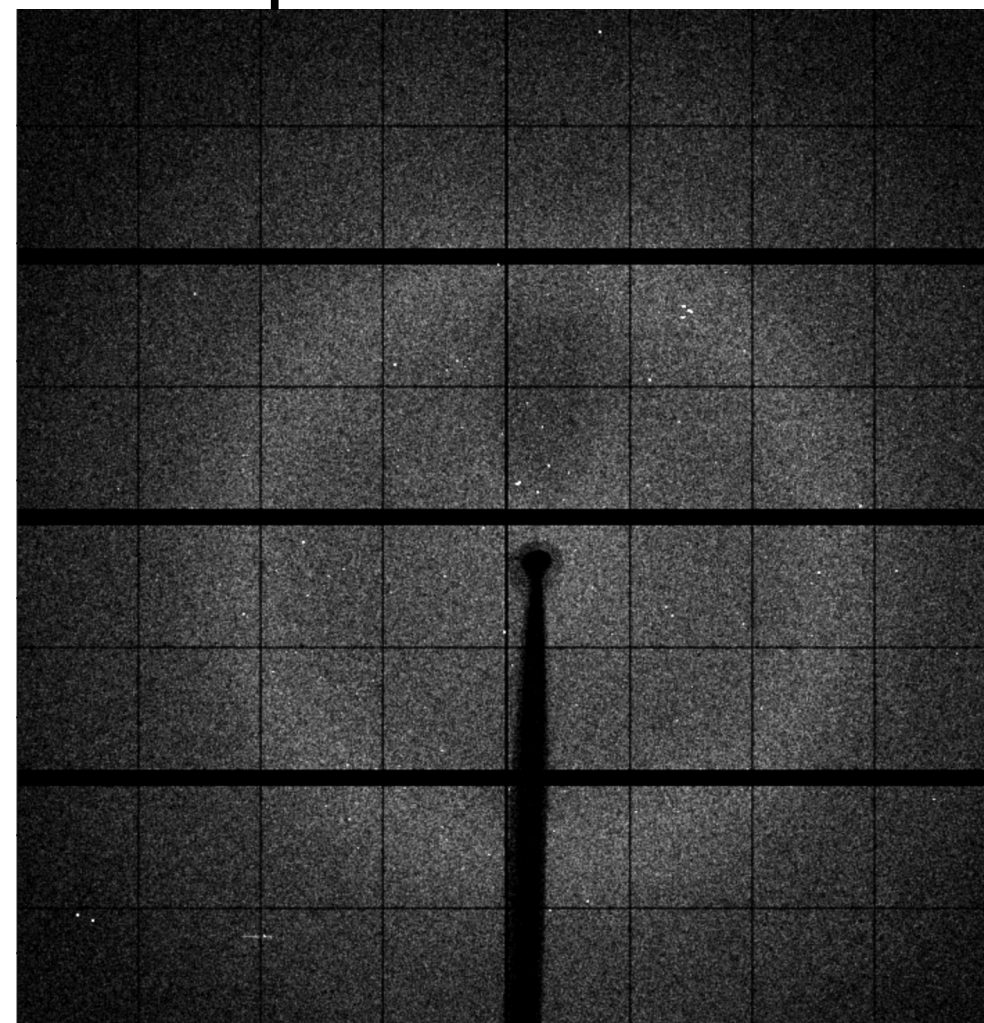
ML

Intensities are Fourier transform of electron density

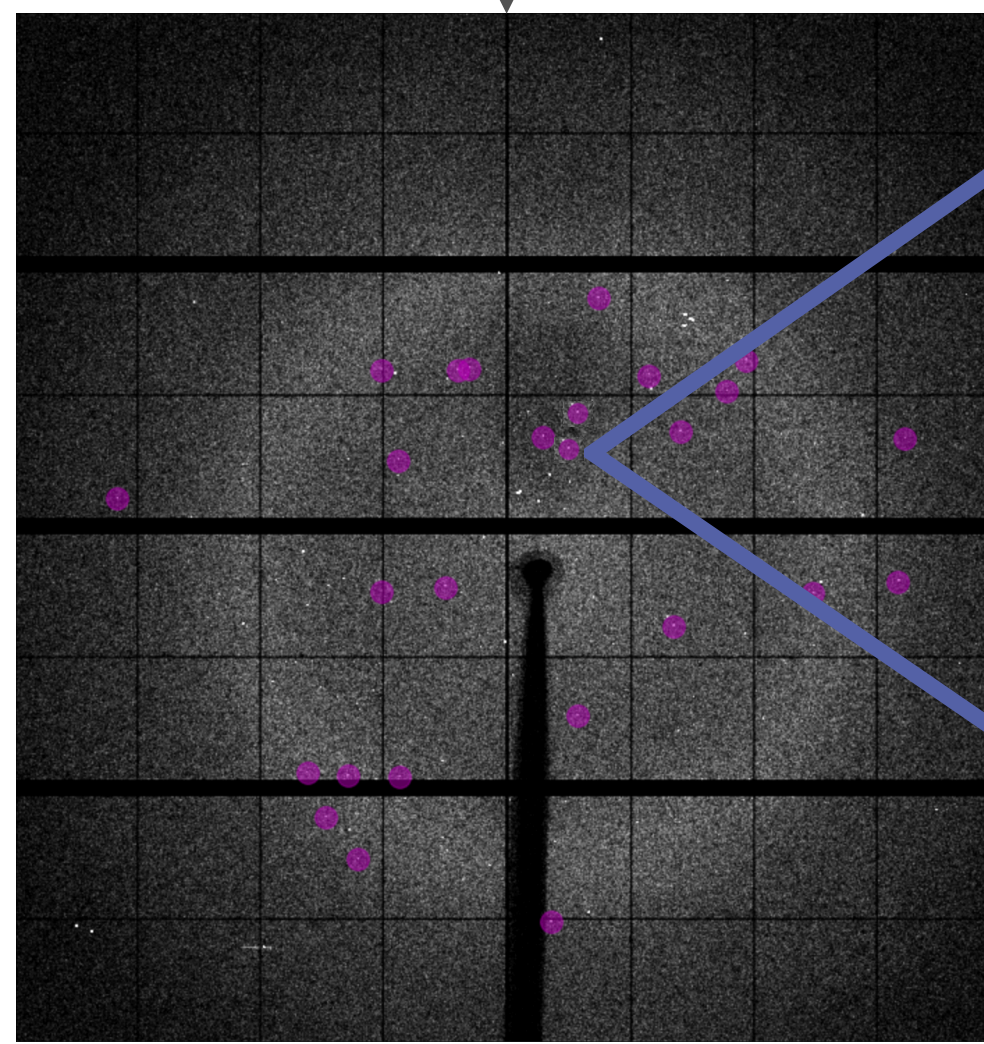


X-ray Crystallography

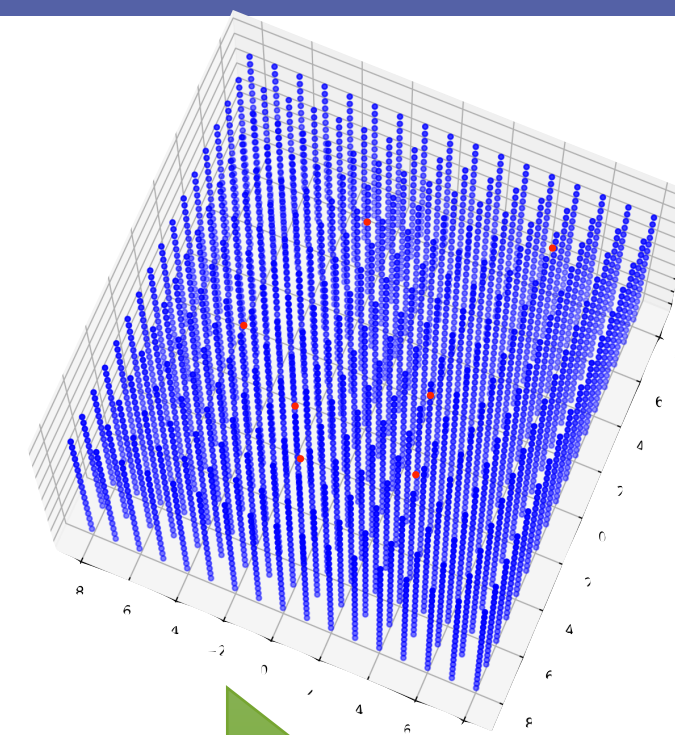
Spot detection



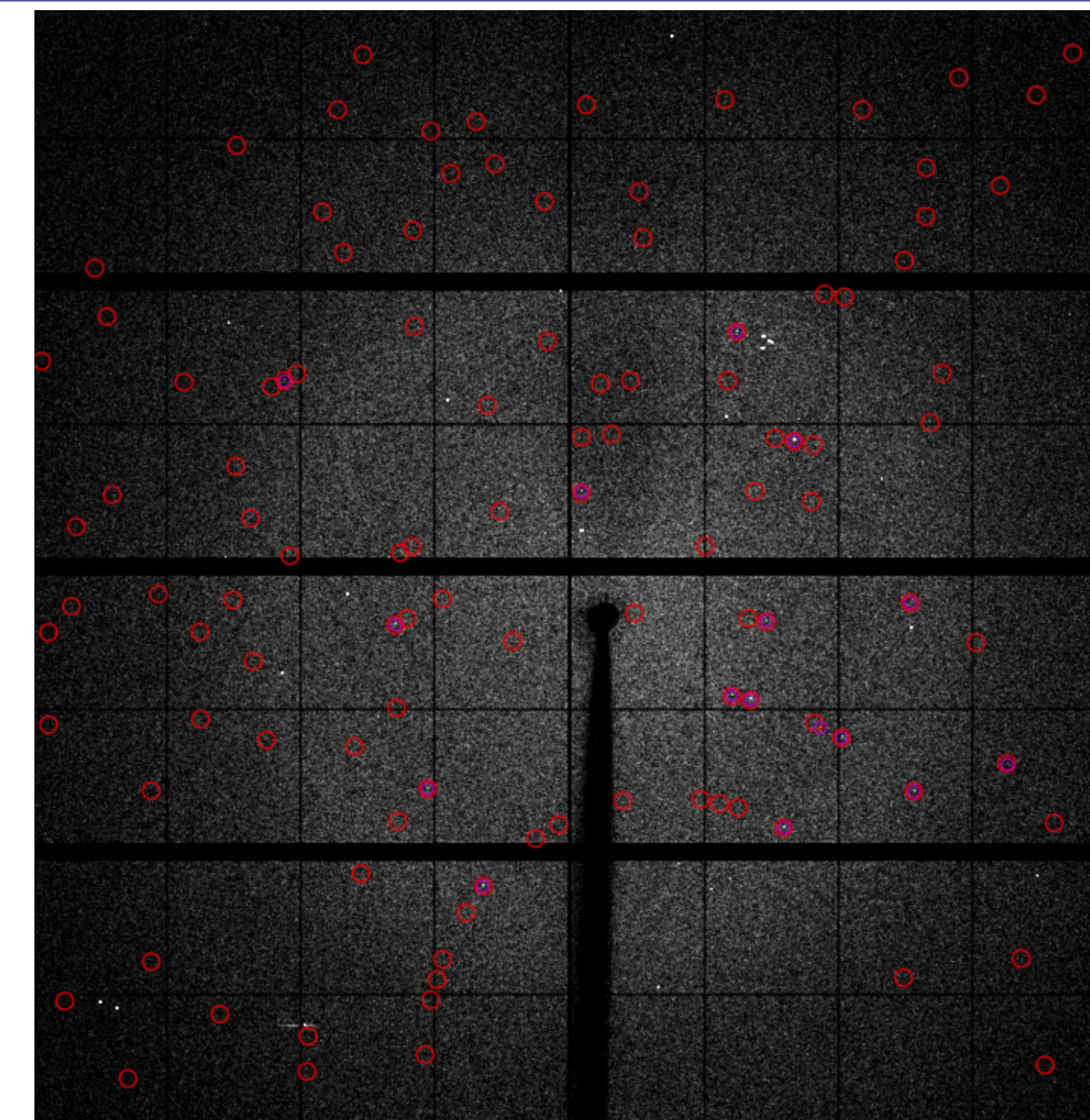
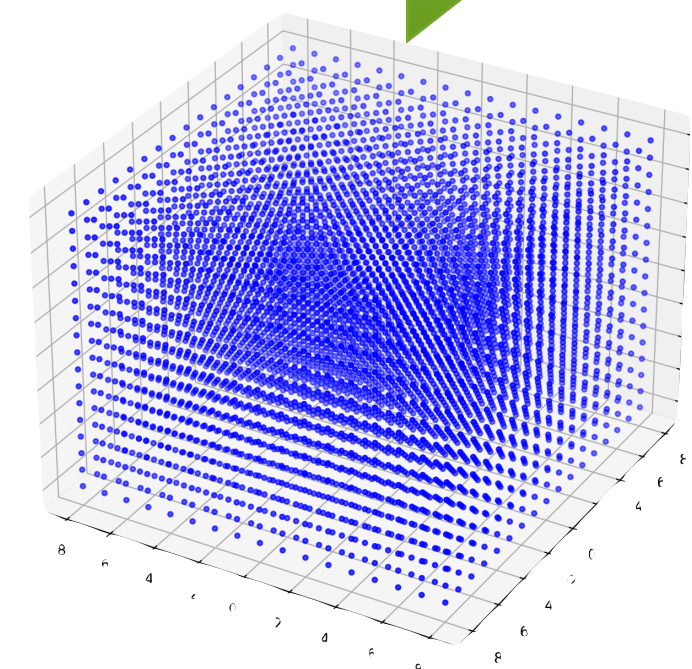
ML



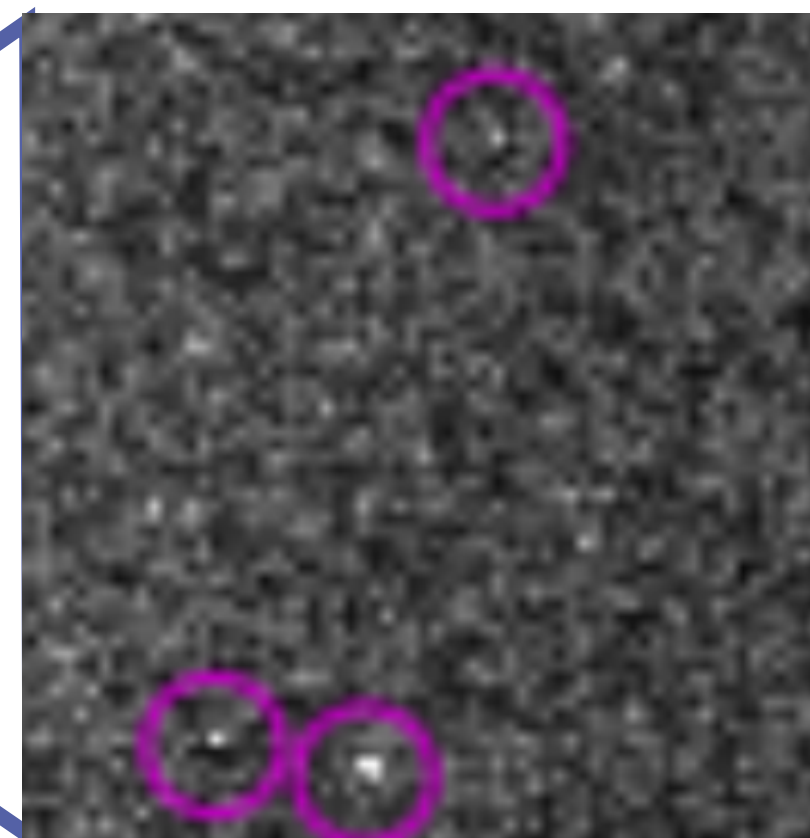
Indexing: identifies each spot and predicts locations that contain useful information (red circles)



Indexing

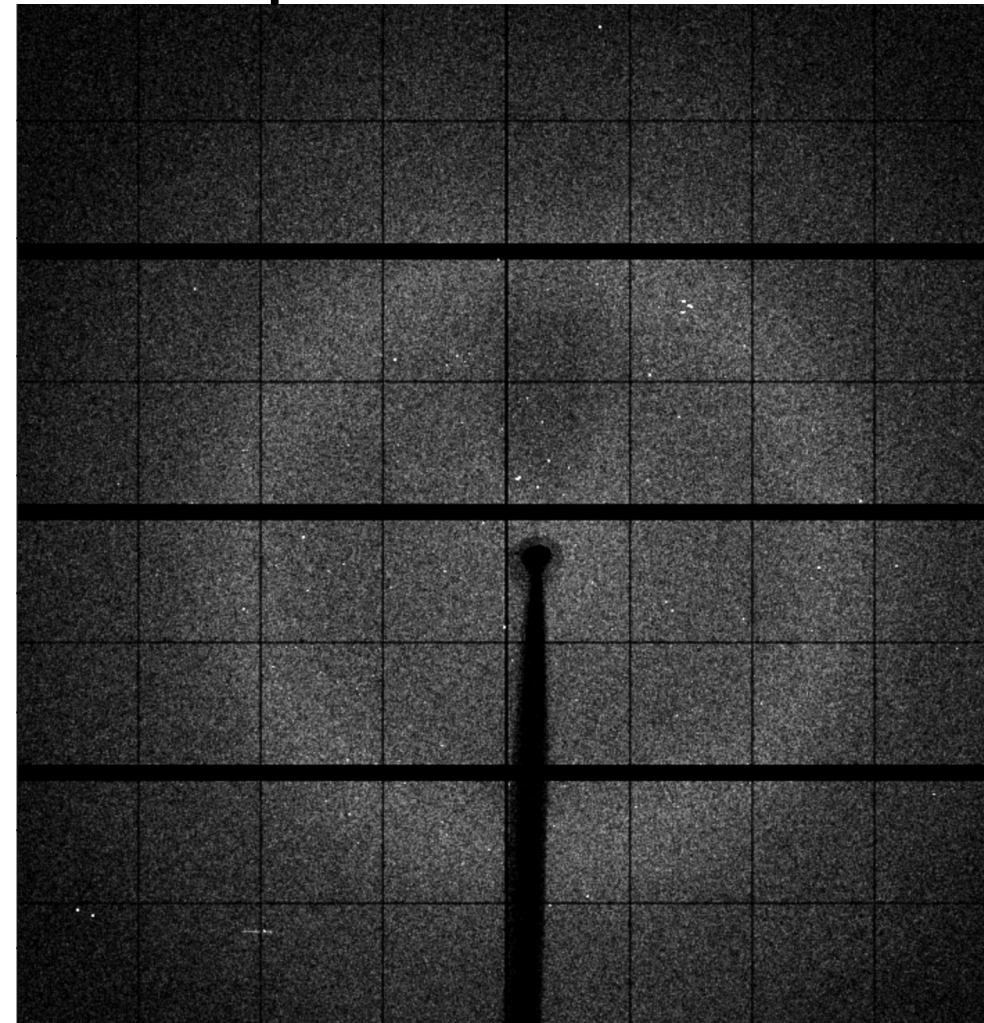


Intensities are Fourier transform of electron density

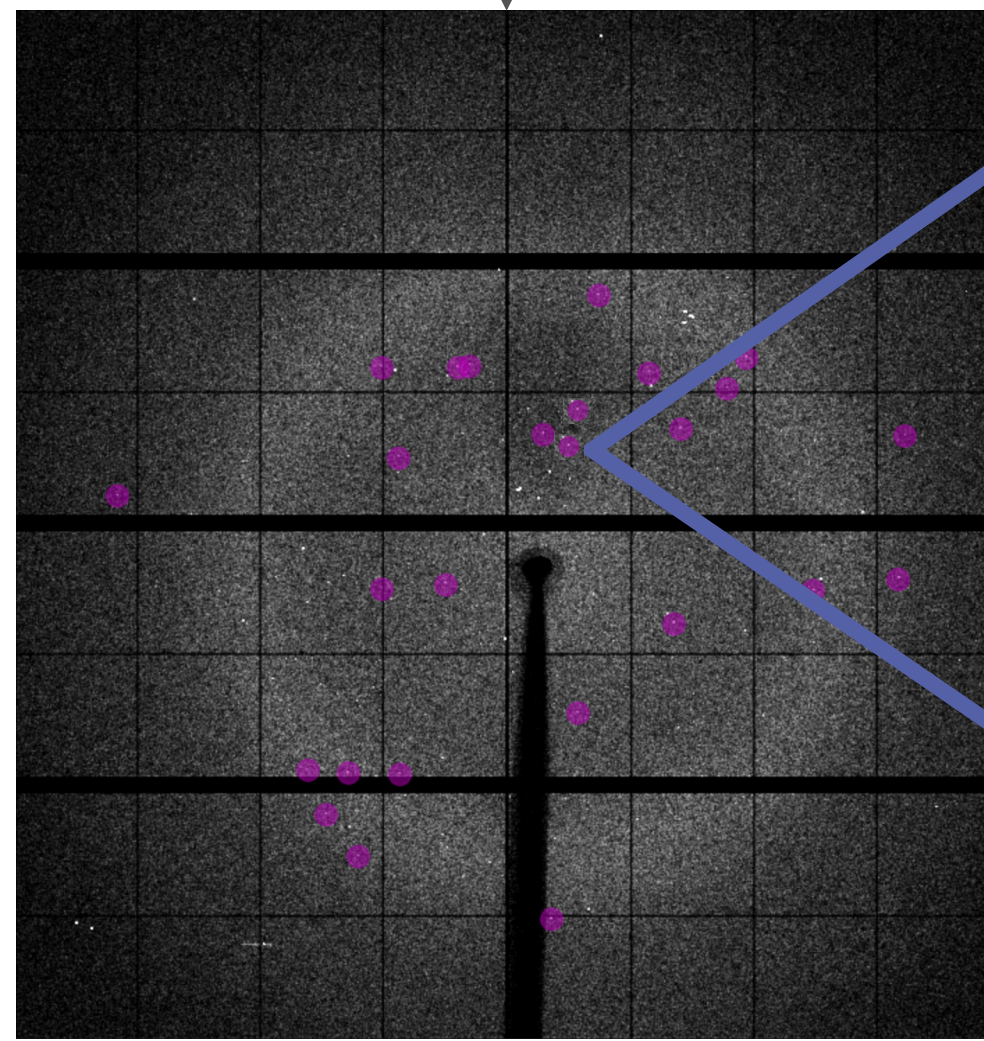


X-ray Crystallography

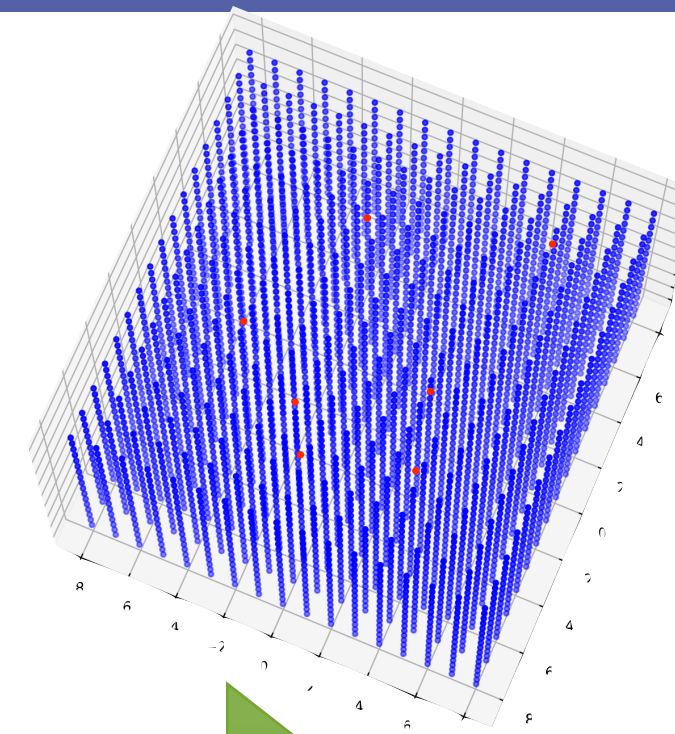
Spot detection



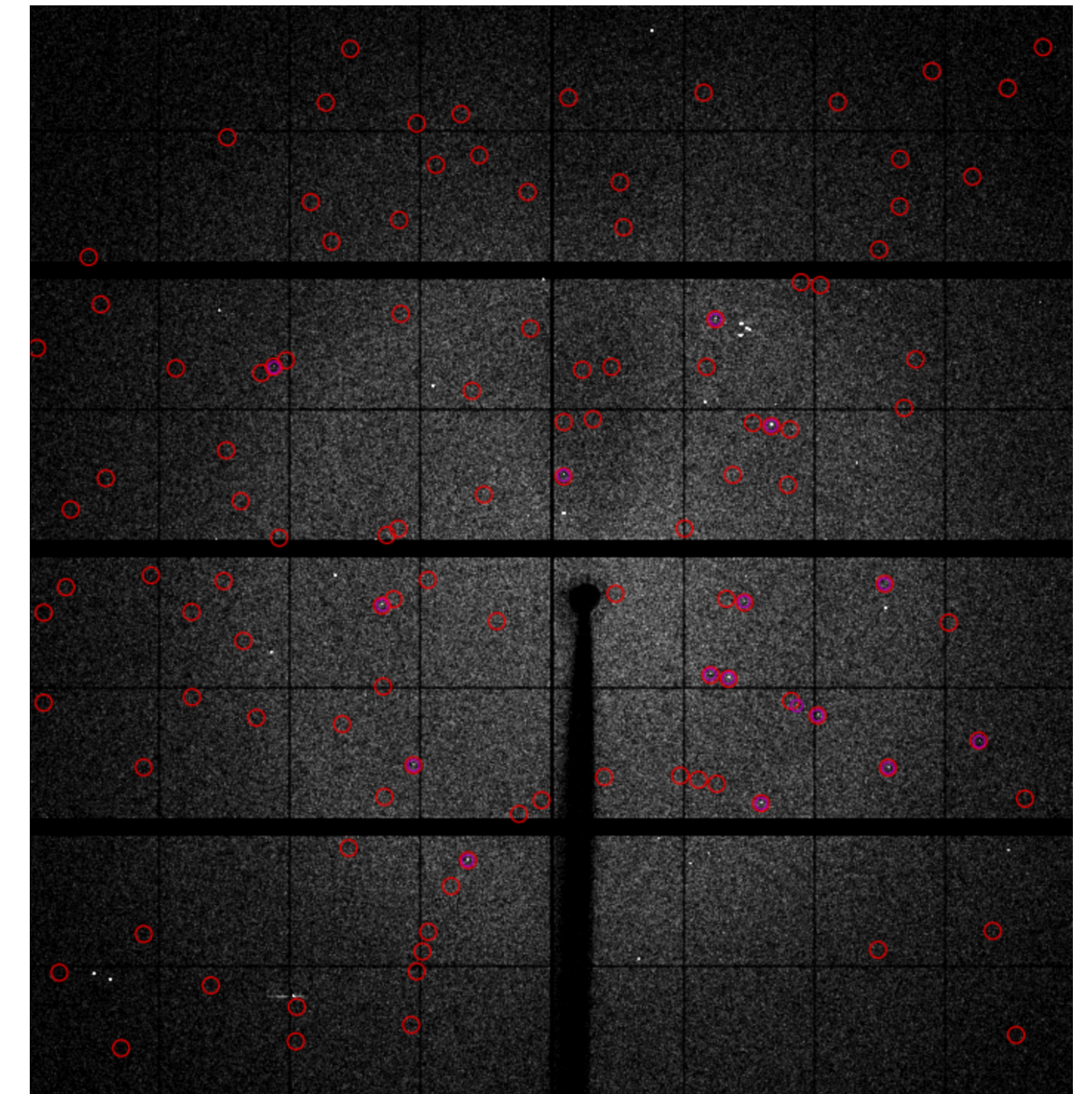
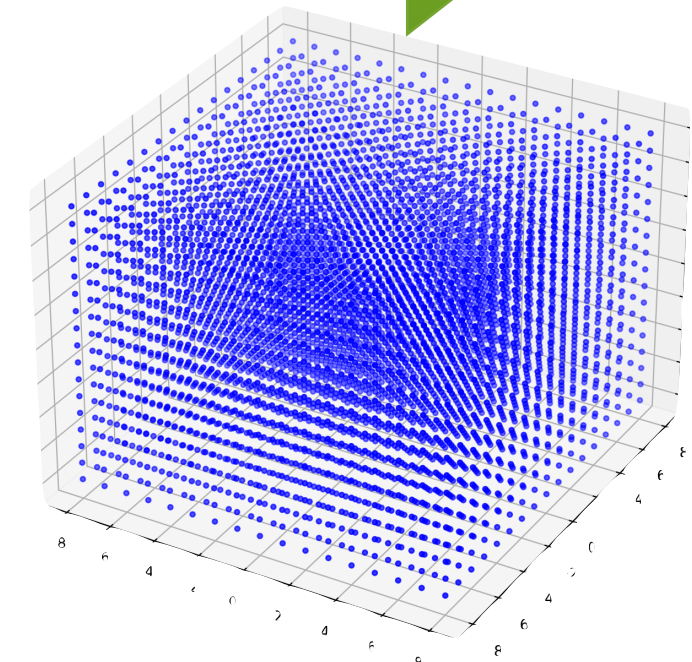
ML



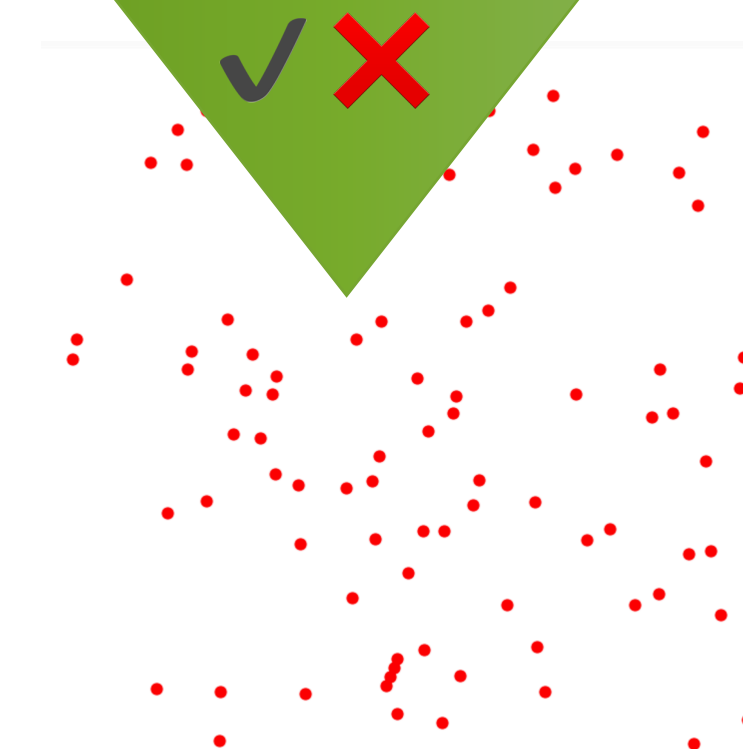
Indexing: identifies each spot and predicts locations that contain useful information (red circles)



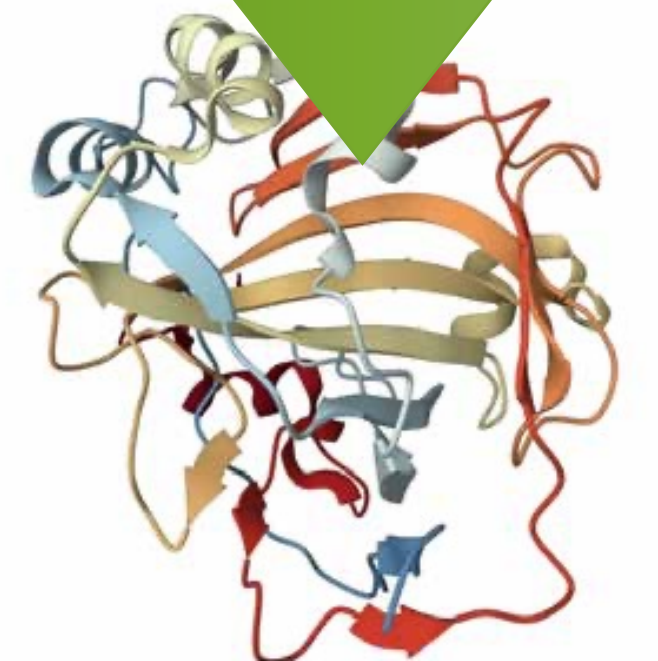
Indexing



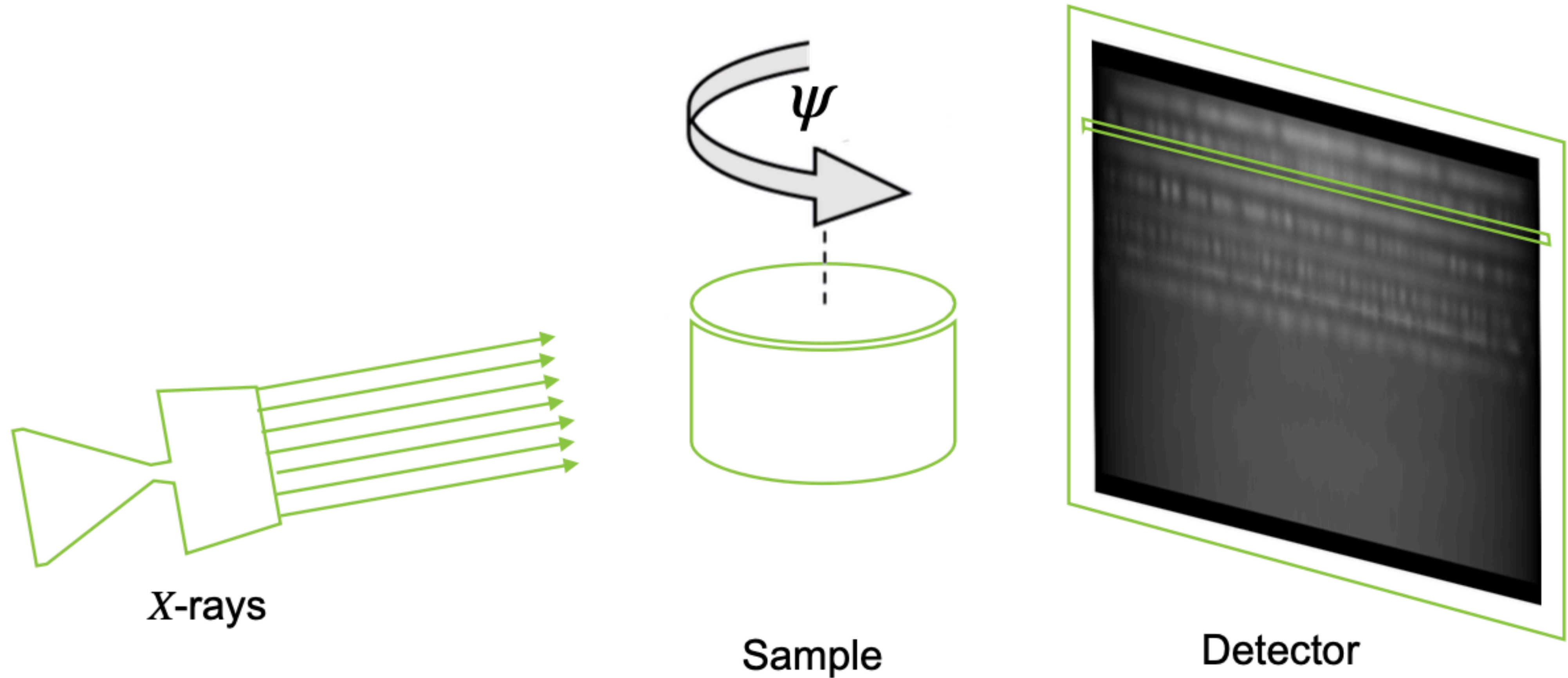
Compression



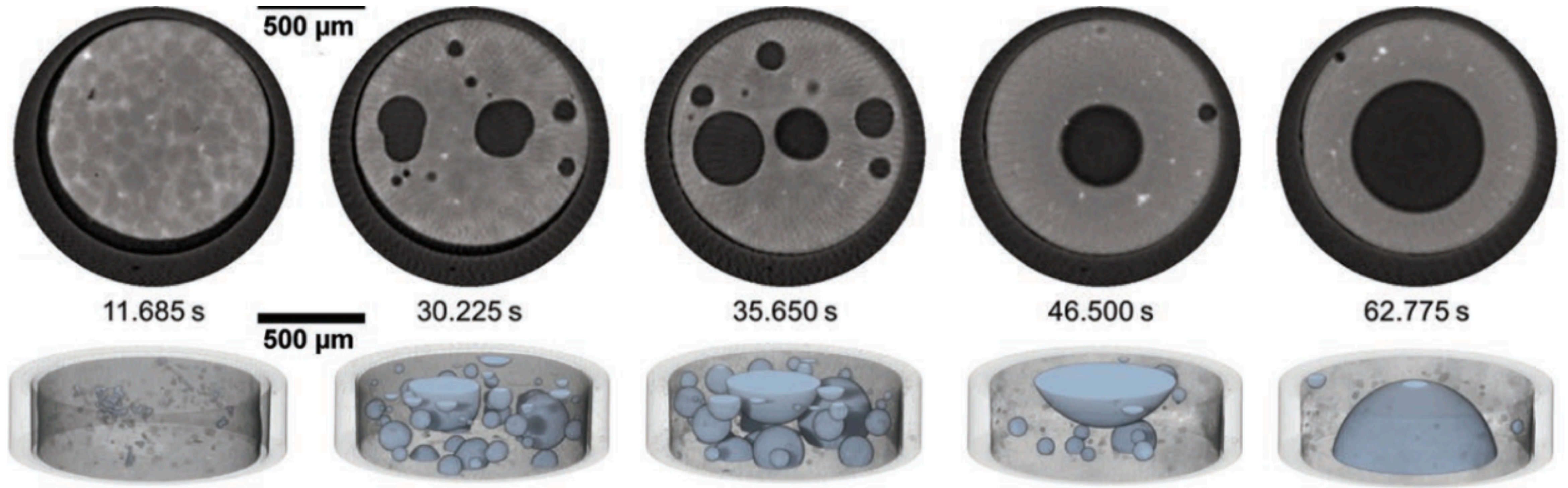
Reconstruction



Time Resolved Computed Tomography

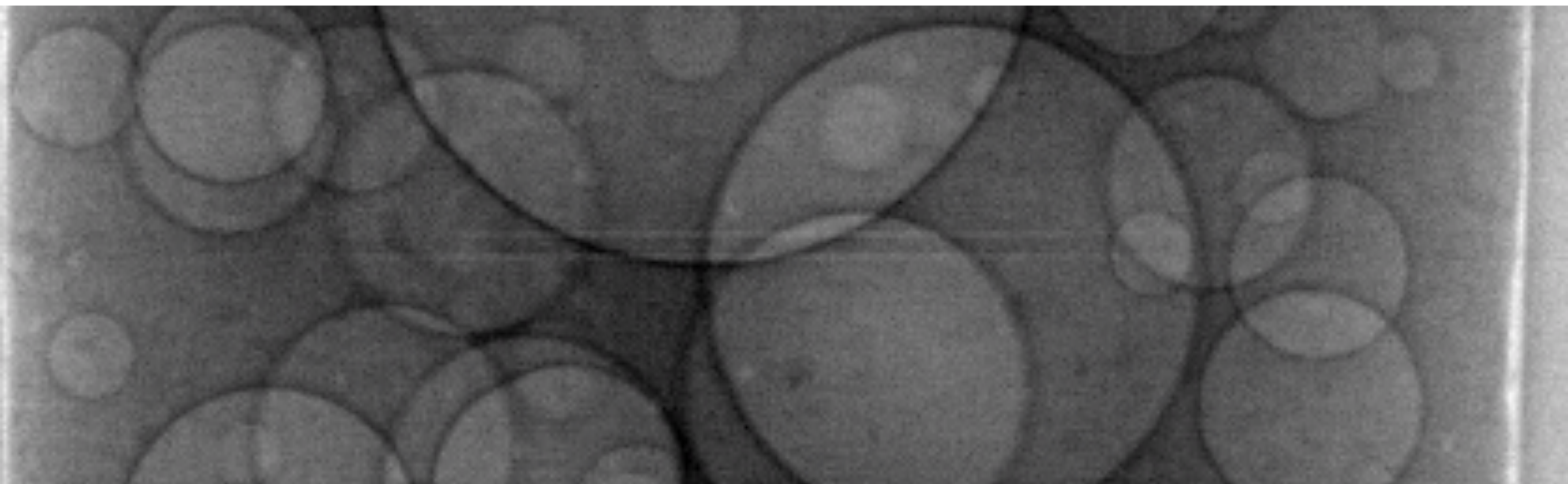


Time Resolved Computed Tomography



3D object is changing over time.
Some changes can be extremely
fast.

Time Resolved Computed Tomography



To capture 4D dynamical scenes,
we need thousands of projections
per second.



Tera-bytes of raw data in a few
seconds.

Time Resolved Computed Tomography

We need to apply compression to be able to store this data.

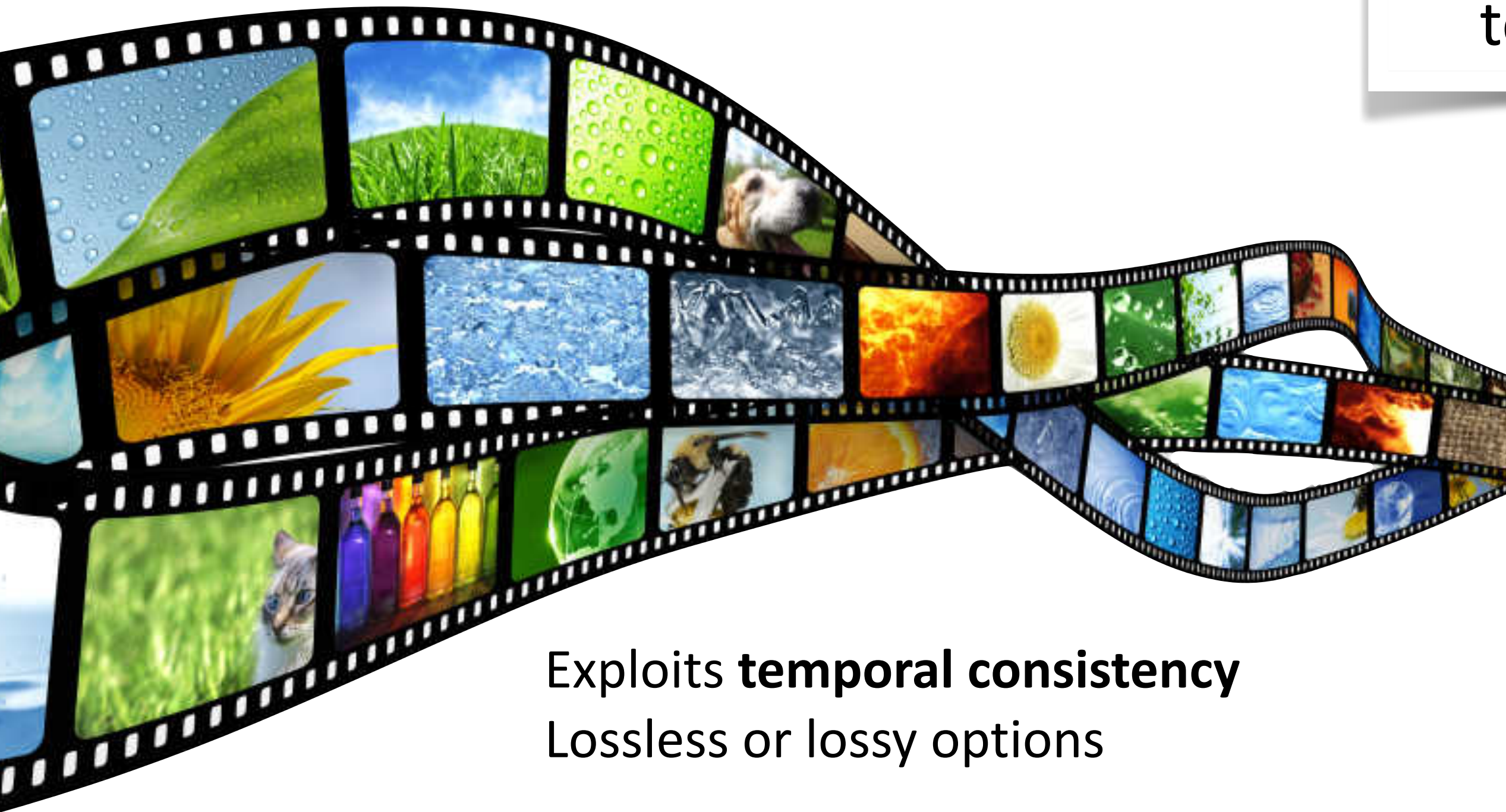
To capture 4D dynamical scenes, we need thousands of projections per second.



Tera-bytes of raw data in a few seconds.

Time Resolved Computed Tomography

HEVC video compression

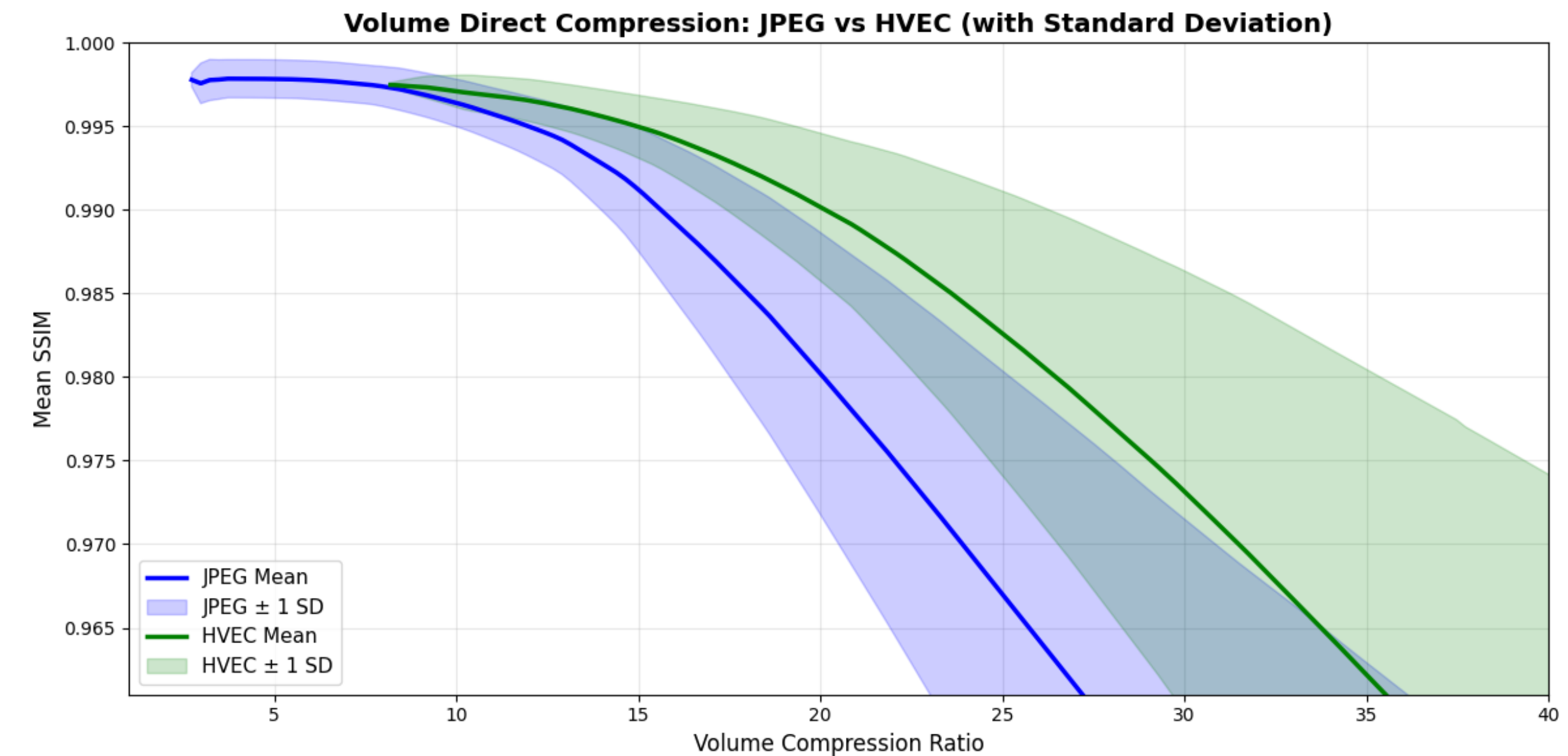


Exploits **temporal consistency**
Lossless or lossy options

Scientific work requires close
to lossless performance.

Lossless compression
cannot compress noise.

Video vs 2D compression

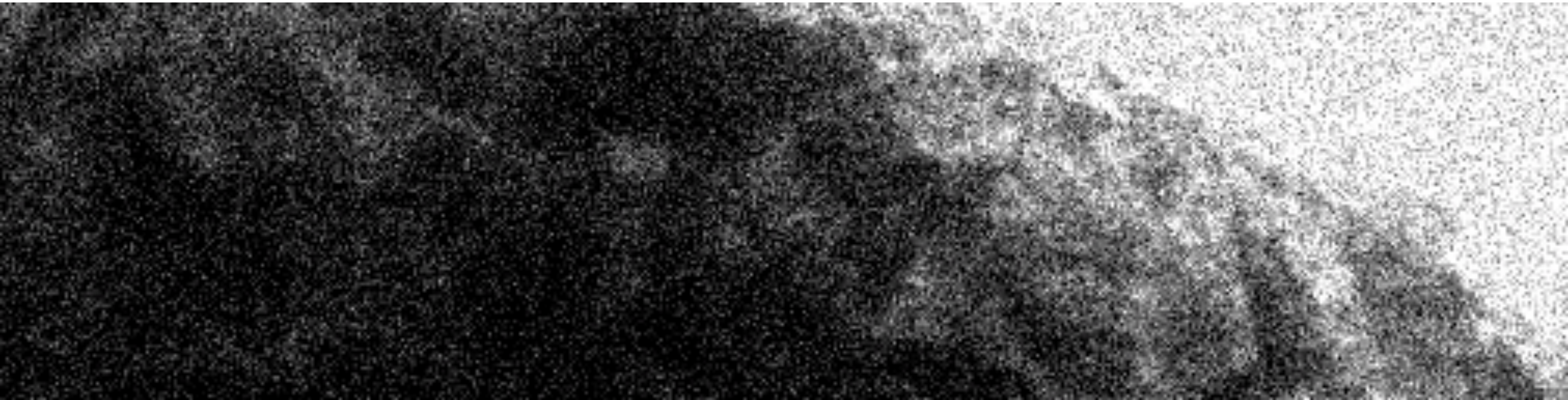


Lossless can provide up to
4x more compression.

Noisy acquisitions in time-resolved CT

When dealing with biological data, we have to **reduce the dose** dramatically to avoid frying the subject.

Produces data with lots of Poisson noise.

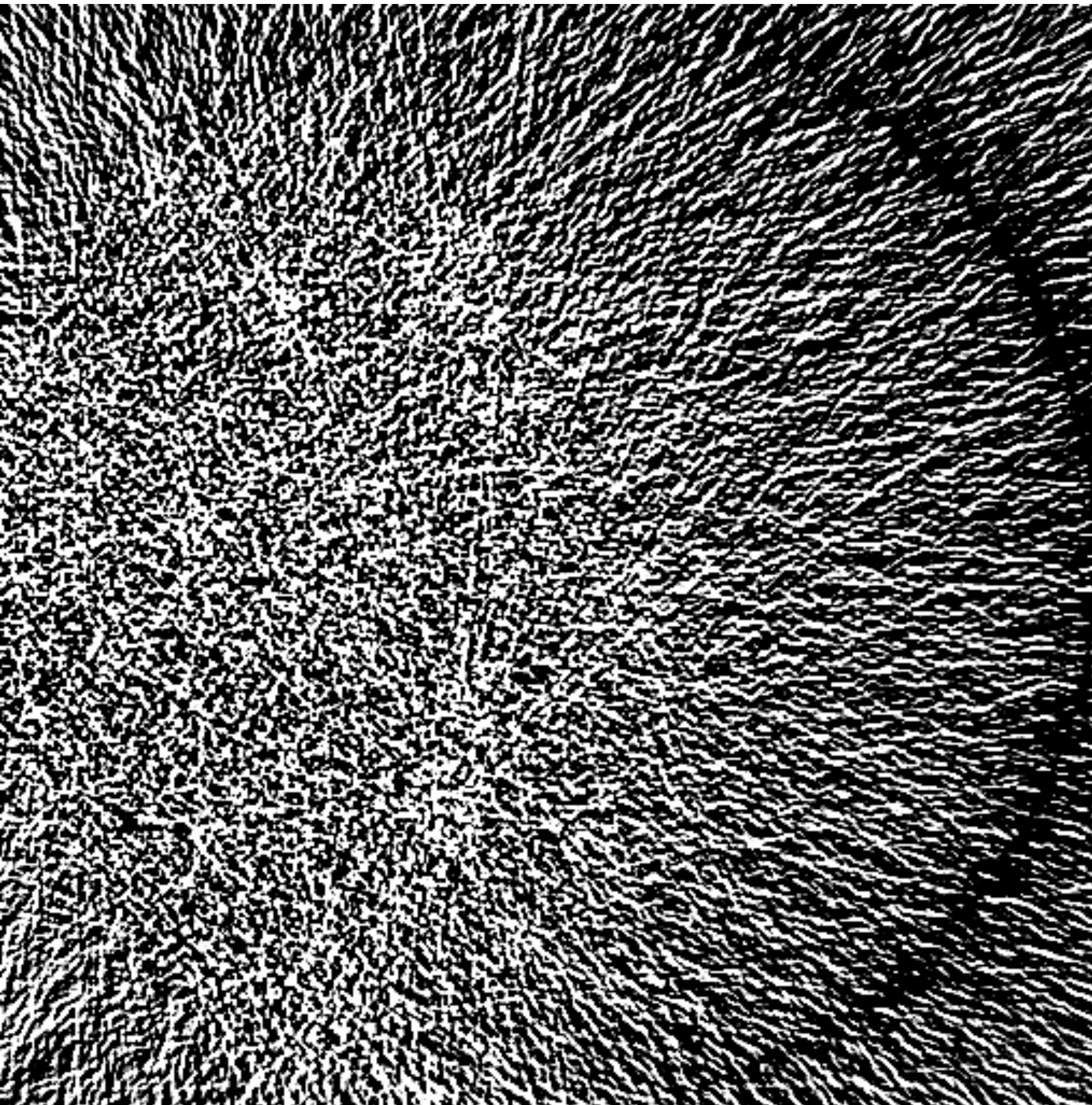


20x less dose than the ideal setting.

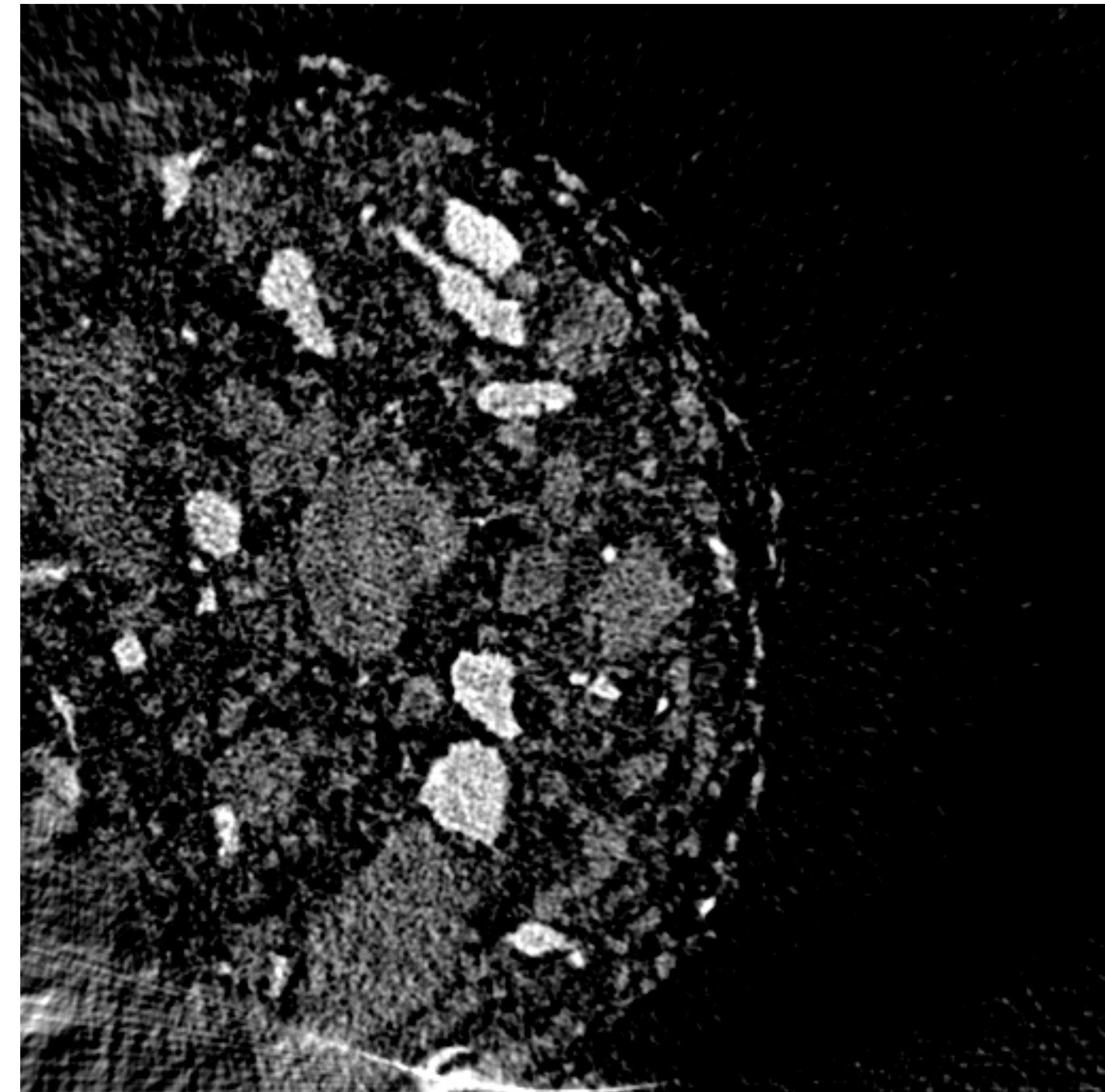
Reconstruction from projections

Directly into the pipeline, noisy acquisitions produce terrible reconstructions.

Noisy reconstruction



Ground Truth



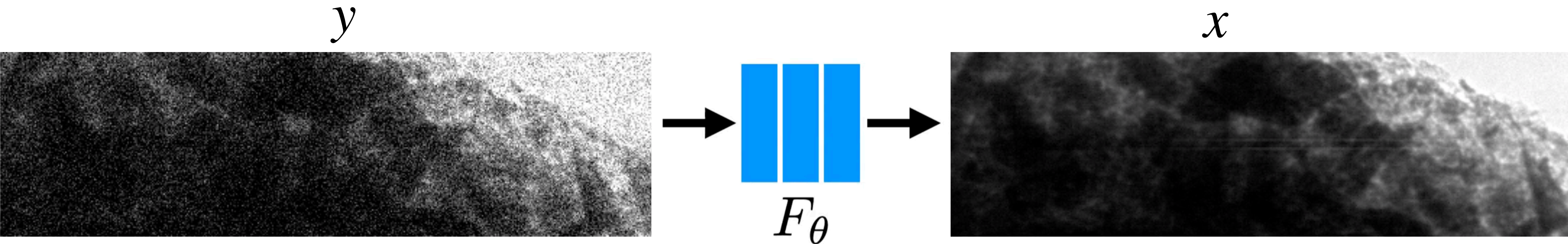
We need denoising that works on unseen new distributions.

Noise2Clean

Assuming $y = x + \varepsilon$
and Gaussian noise

Classical Supervised
learning regression task,
usually solved with a UNet

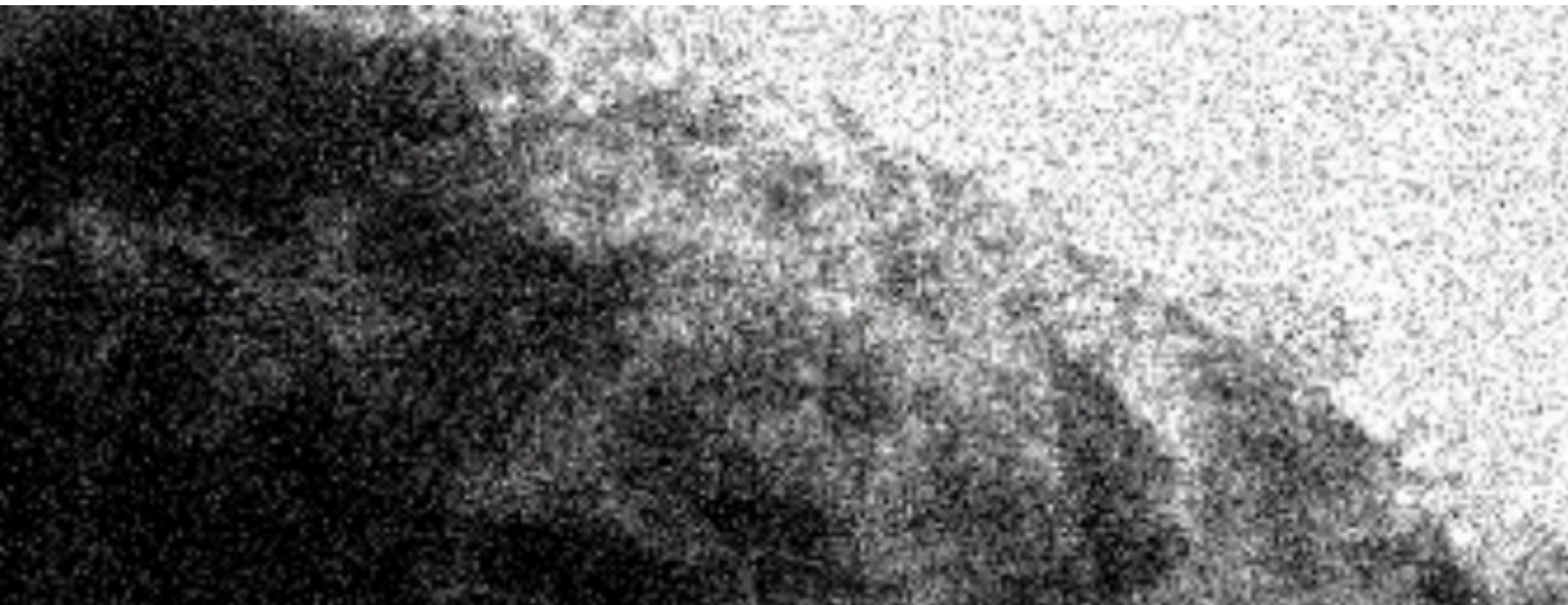
Minimize the risk $L(\theta) = E_{x,\varepsilon} \|F_\theta(y) - x\|^2$



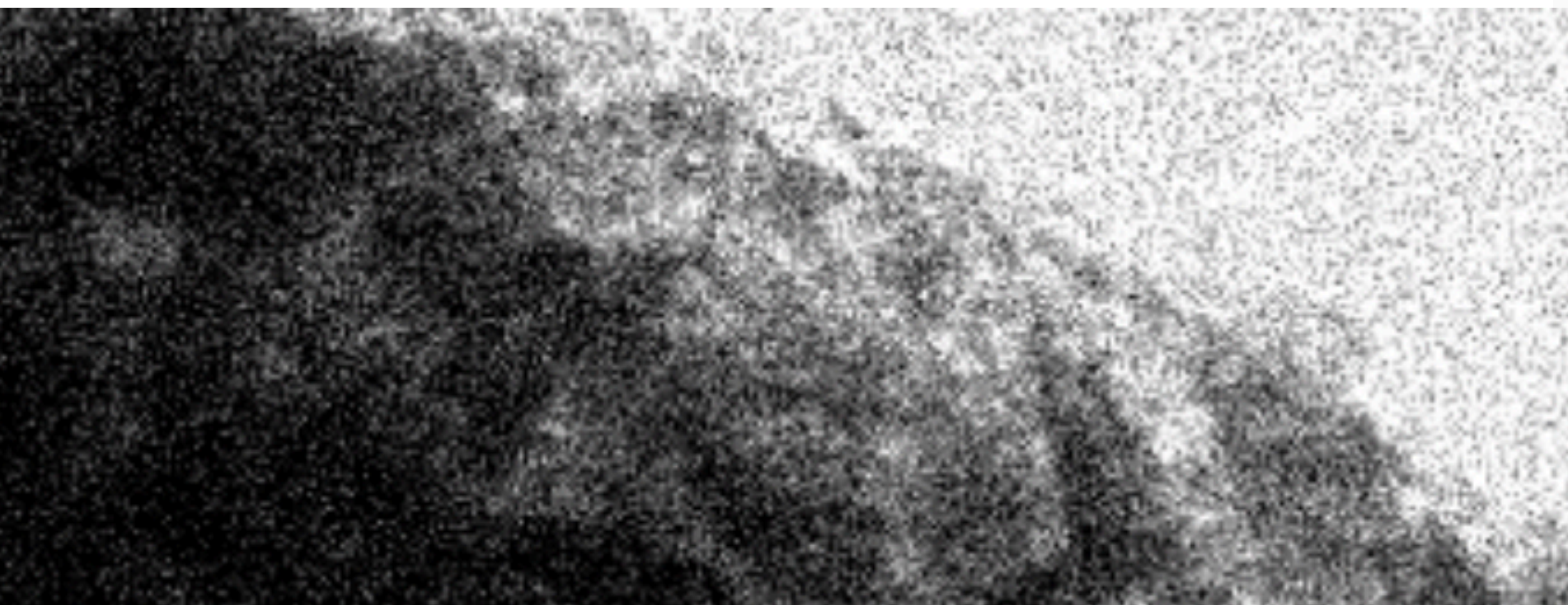
Noise2Noise

Noise2Noise: Learning Image Restoration without Clean Data, Lehtinen et al. (2018)

y_1



y_2



Noise2Clean

Noisy input

Neural network f

Clean target y
requires ground truth

Noise2Noise

Noisy input

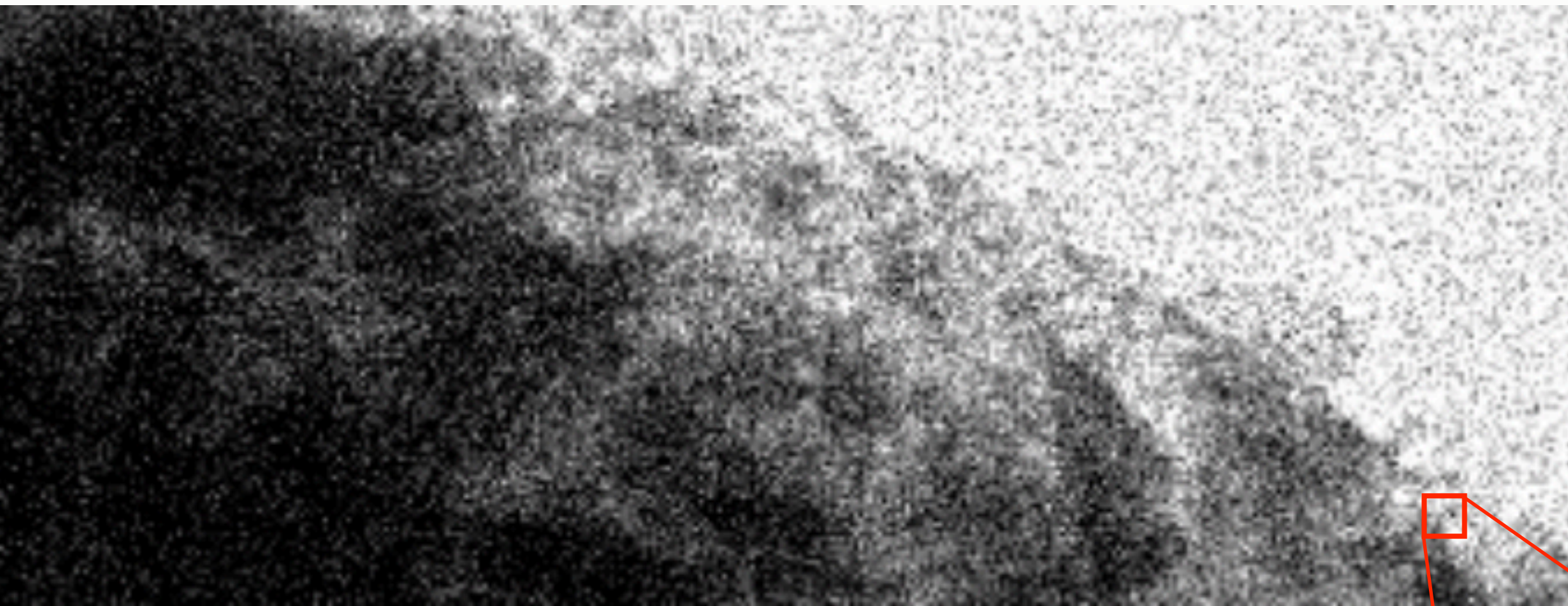
Neural network f

Noisy target y'
independent of input

Same optimal denoiser: $f^*(x) = E[y | x]$
Condition: $E[\text{noisy target} | y] = y$ and target is independent of input

Noise2Void

$$y = x + \varepsilon$$



Noise2Void (N2V): train from a single noisy image, no pairs at all.

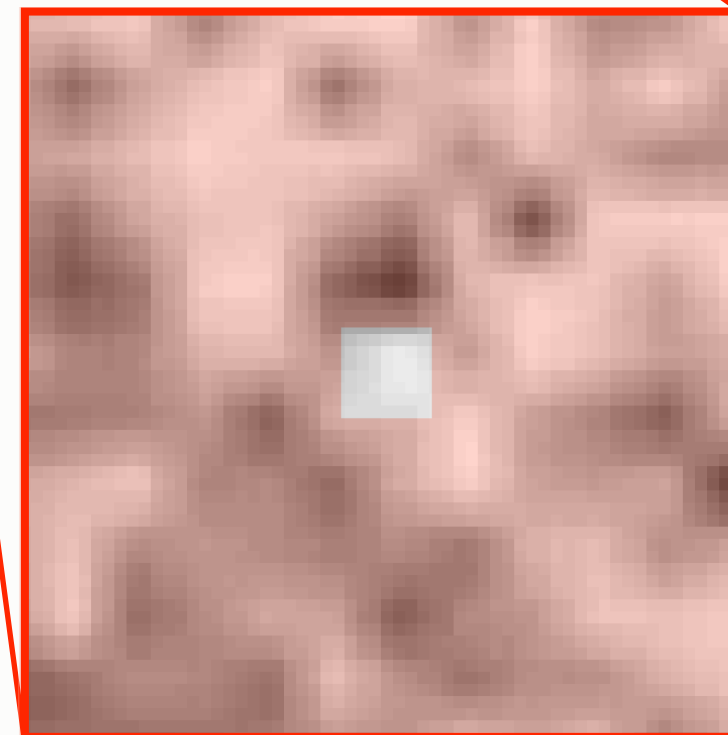
We can try to minimize the risk using the context

$$L(\theta) = E_{x,\varepsilon} \|F_{\theta}(y_{\setminus i})_i - y_i\|^2$$

Core assumptions:

Noise is independent across pixels and zero-mean.

Predict the pixel value from the context.



The minimizer recovers

$$F_{\theta}^*(y) = E[x_i | c = y_{\setminus i}]$$

Putting denoisers together

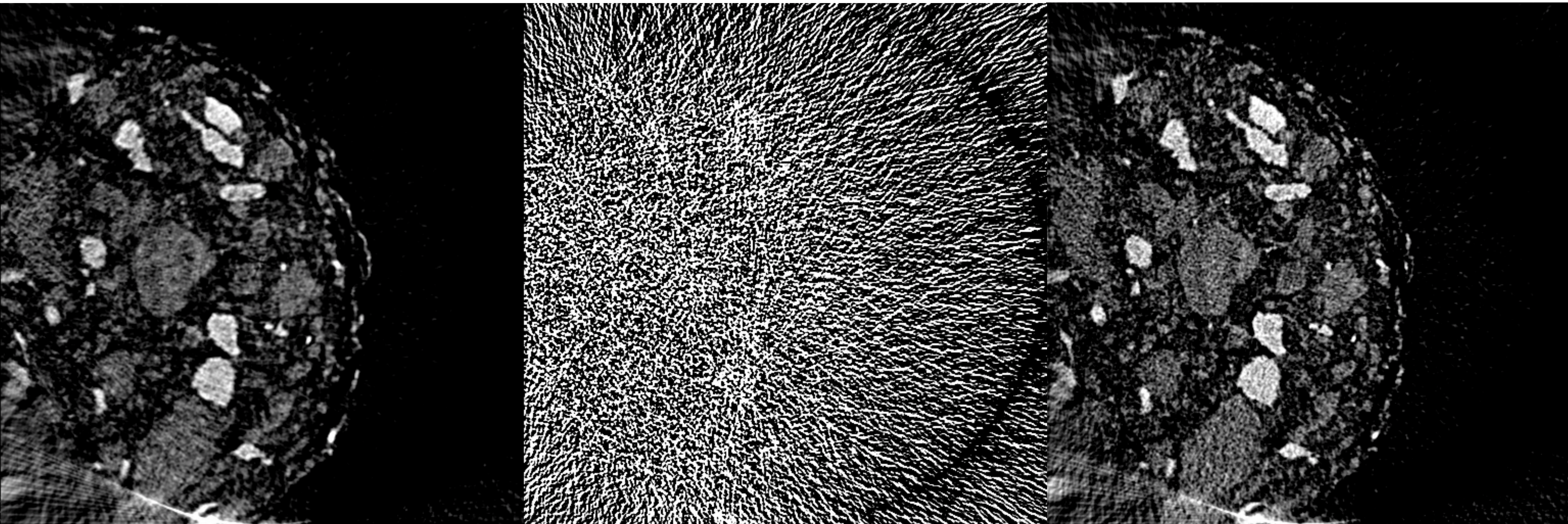
Temporal consistency +
Noise2Void

No clean data, pure self-
supervised training.

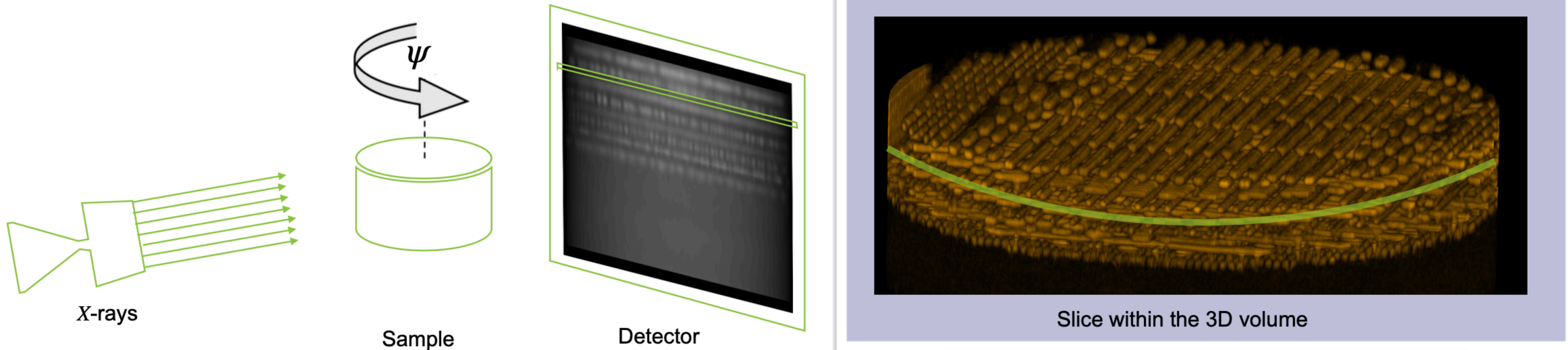
Reconstruction 27.8 PSNR / 0.73 SSIM

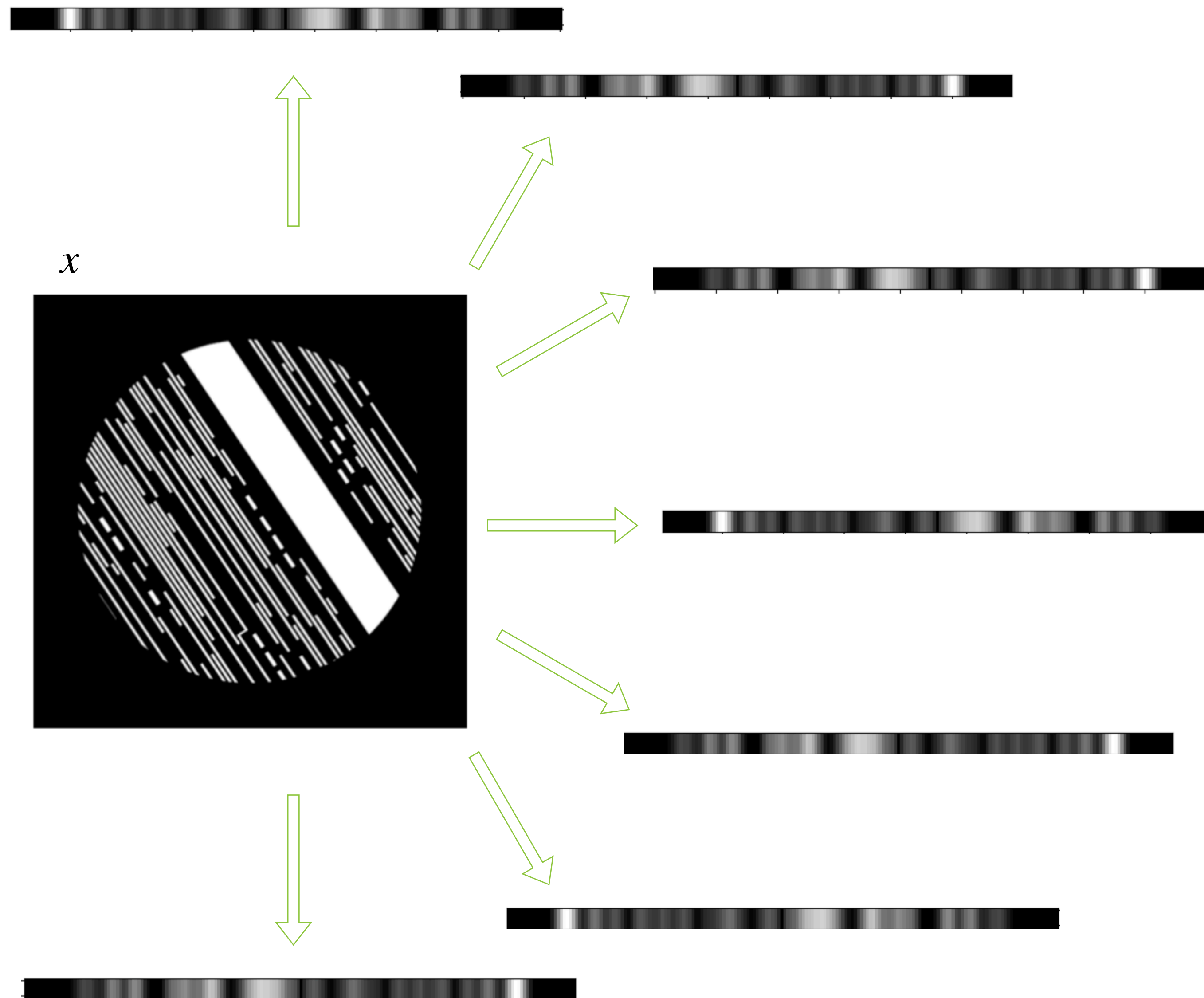
Noisy input 4.6 PSNR / 0.017 SSIM

Ground Truth



DIFFUSION PRIORS FOR TOMOGRAPHY





For each layer, we measure projections onto different directions

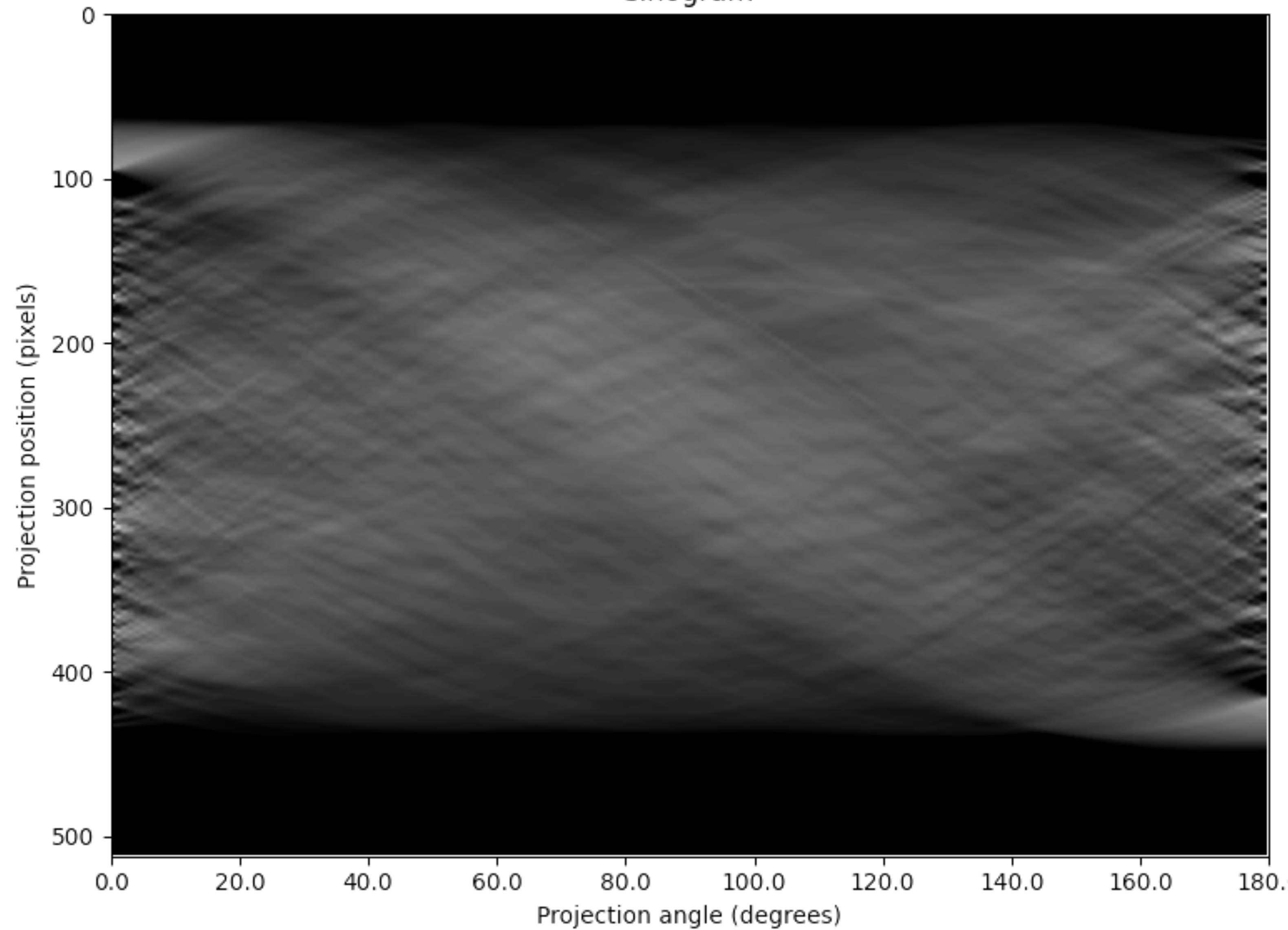
Measurements come from a linear model $A_i x + \varepsilon$, where A_i is a linear projection

TOMOGRAPHIC SETTING

Ground Truth object



Sinogram

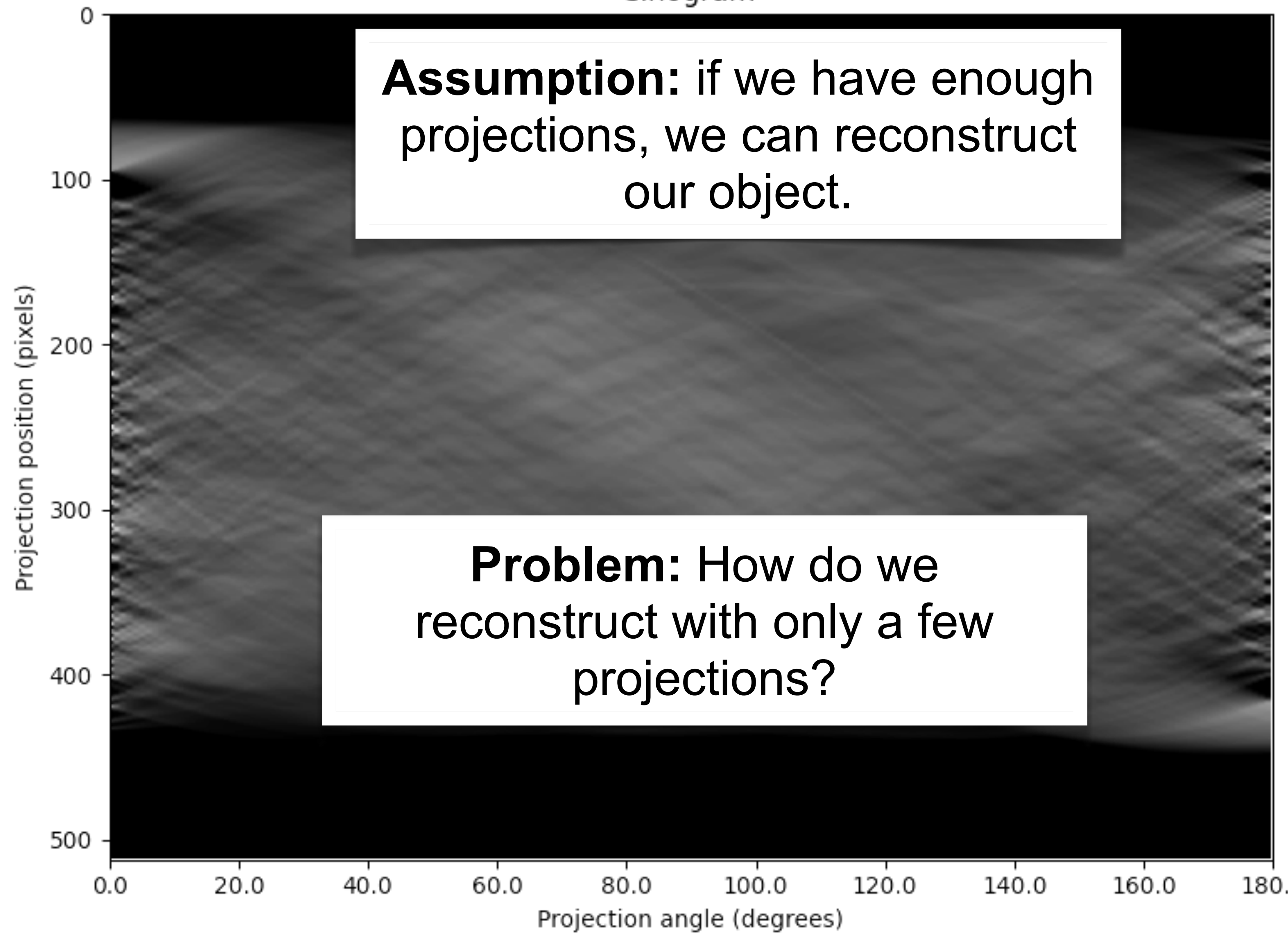


TOMOGRAPHIC SETTING

Ground Truth object

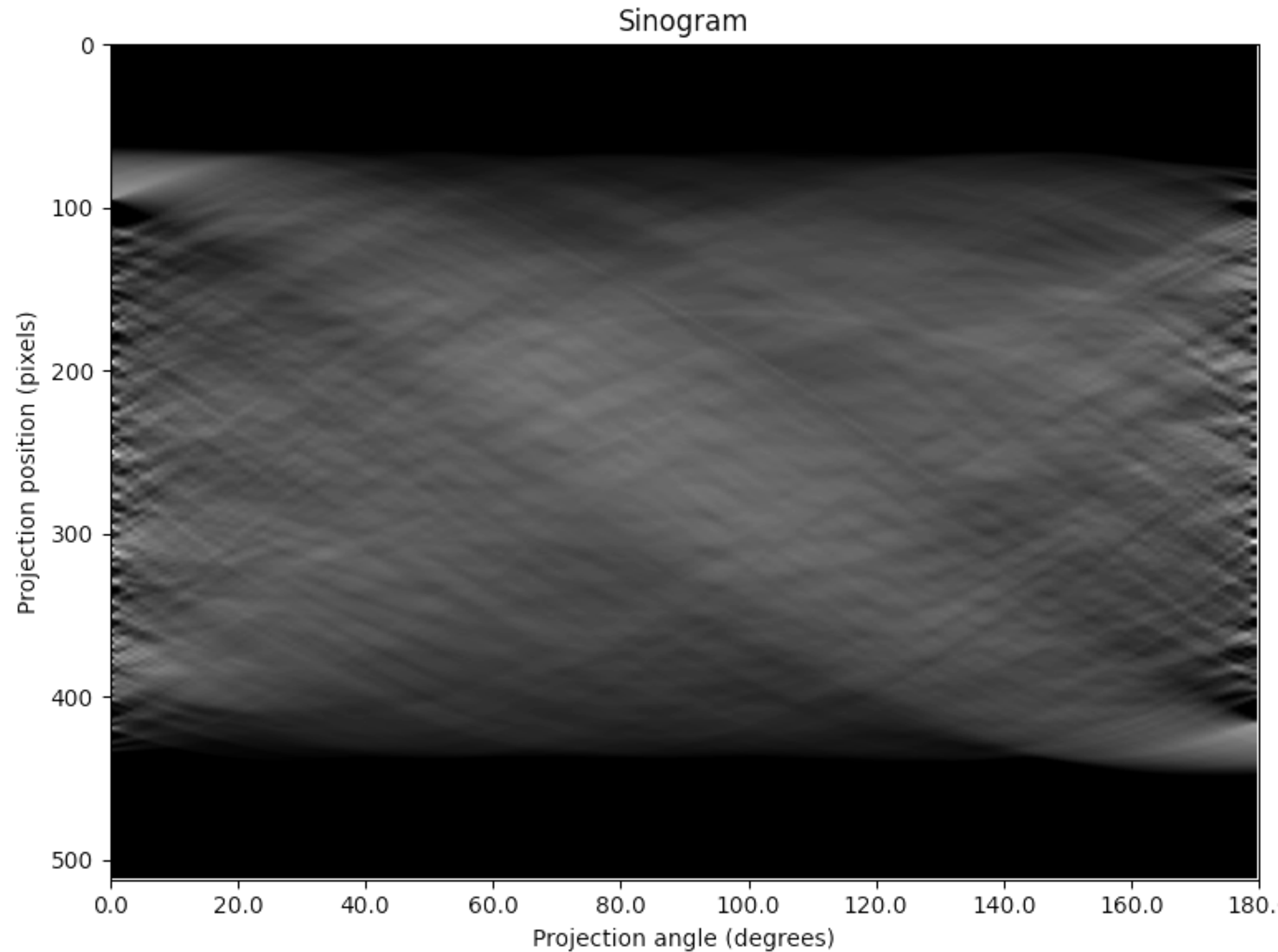


Sinogram

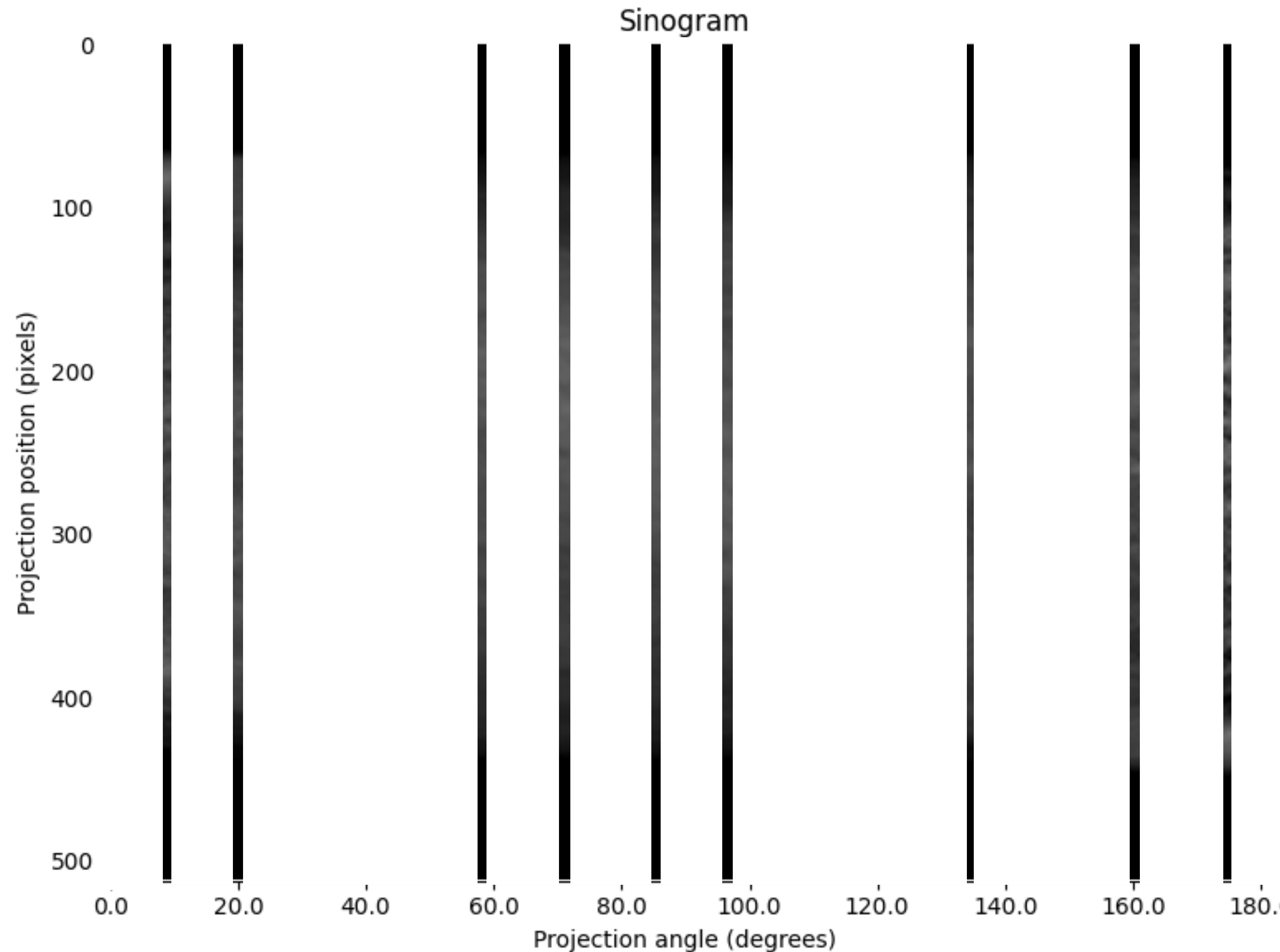
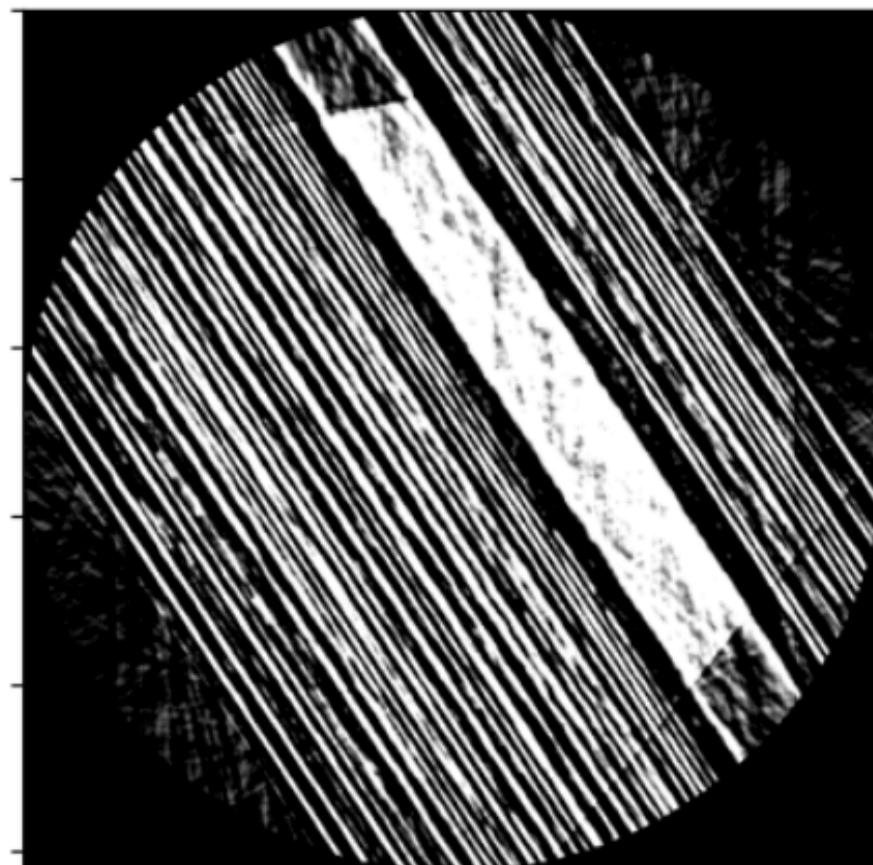


SPARSE PROJECTIONS

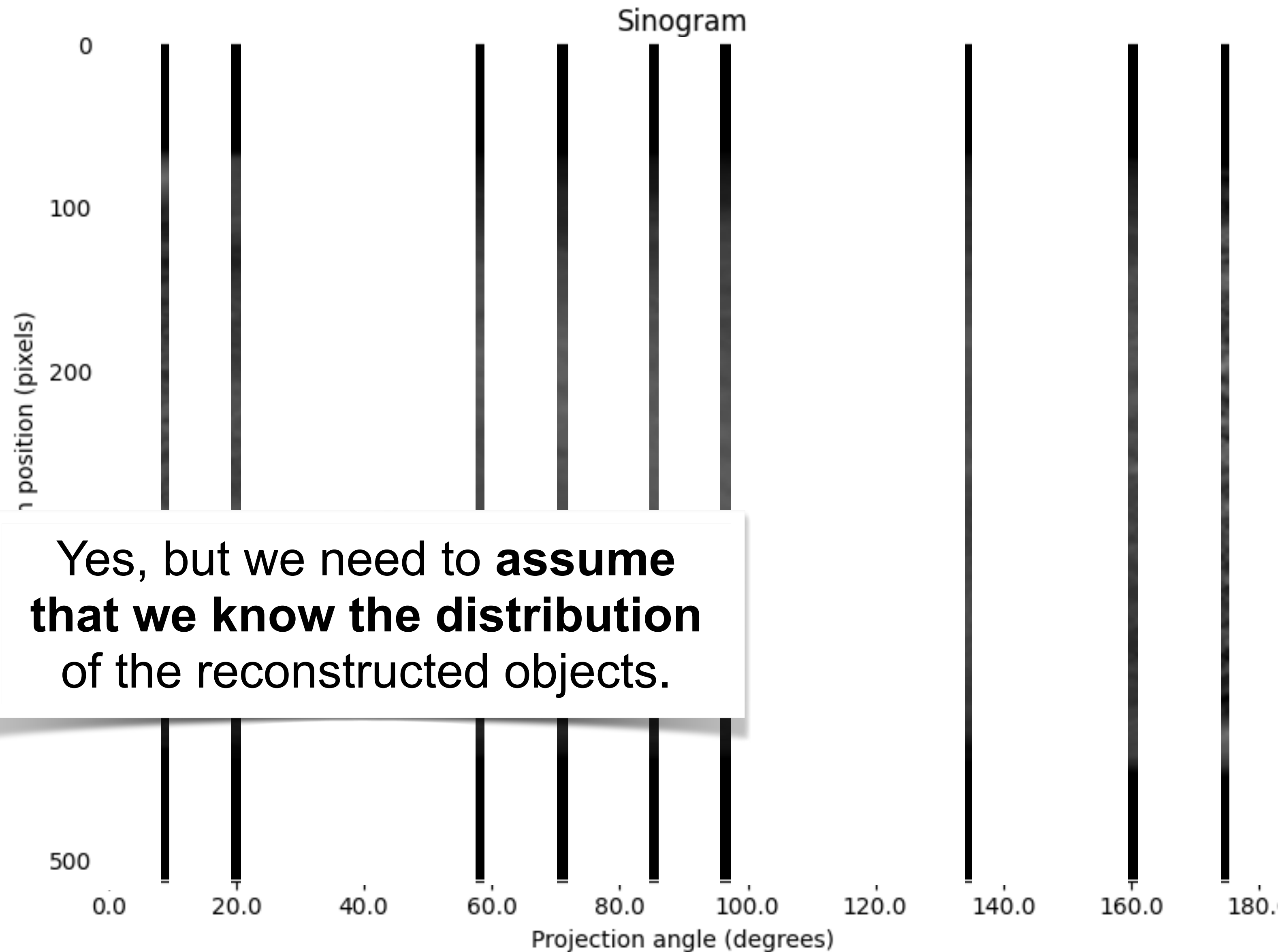
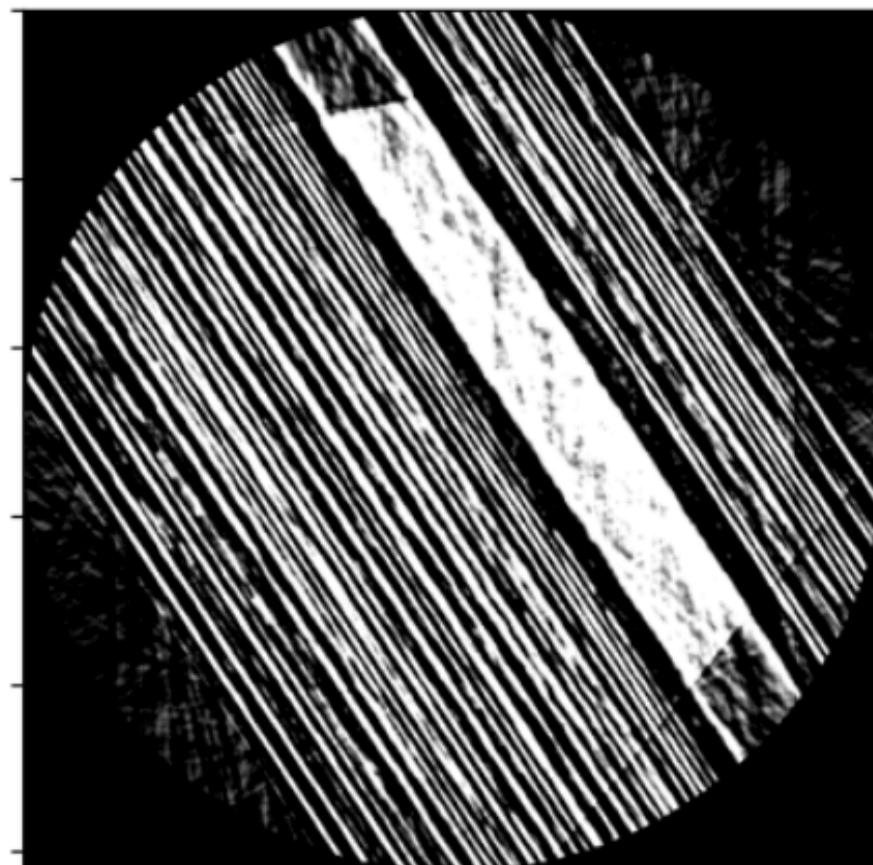
- Instead of measuring in every direction
- We measure projections in only a **few directions** (budget)



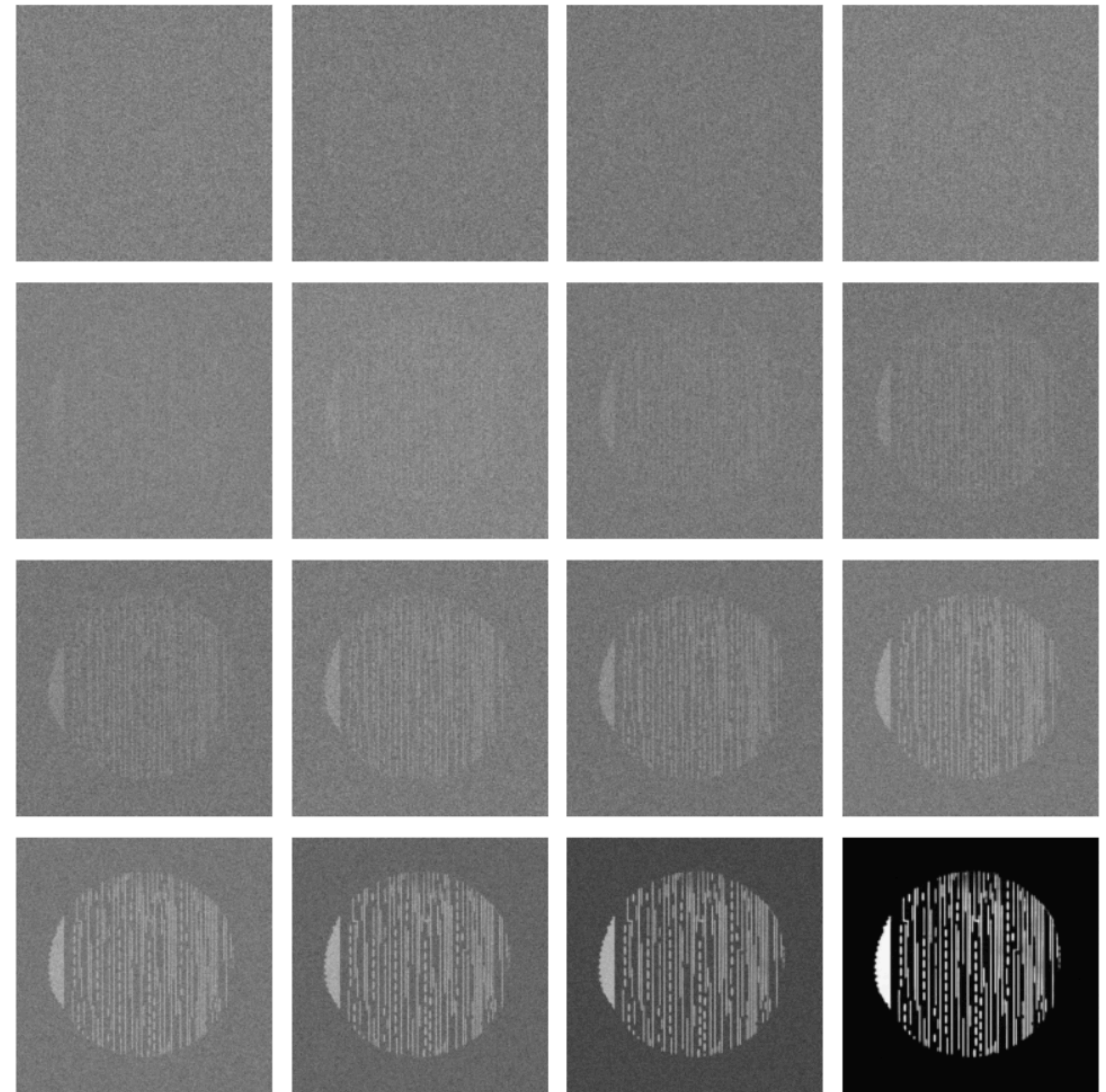
- Instead of measuring in every direction
- We measure projections in only a **few directions** (budget)
- Can we come up with a better reconstruction algorithm?



- Instead of measuring in every direction
- We measure projections in only a **few directions** (budget)
- Can we come up with a better reconstruction algorithm?

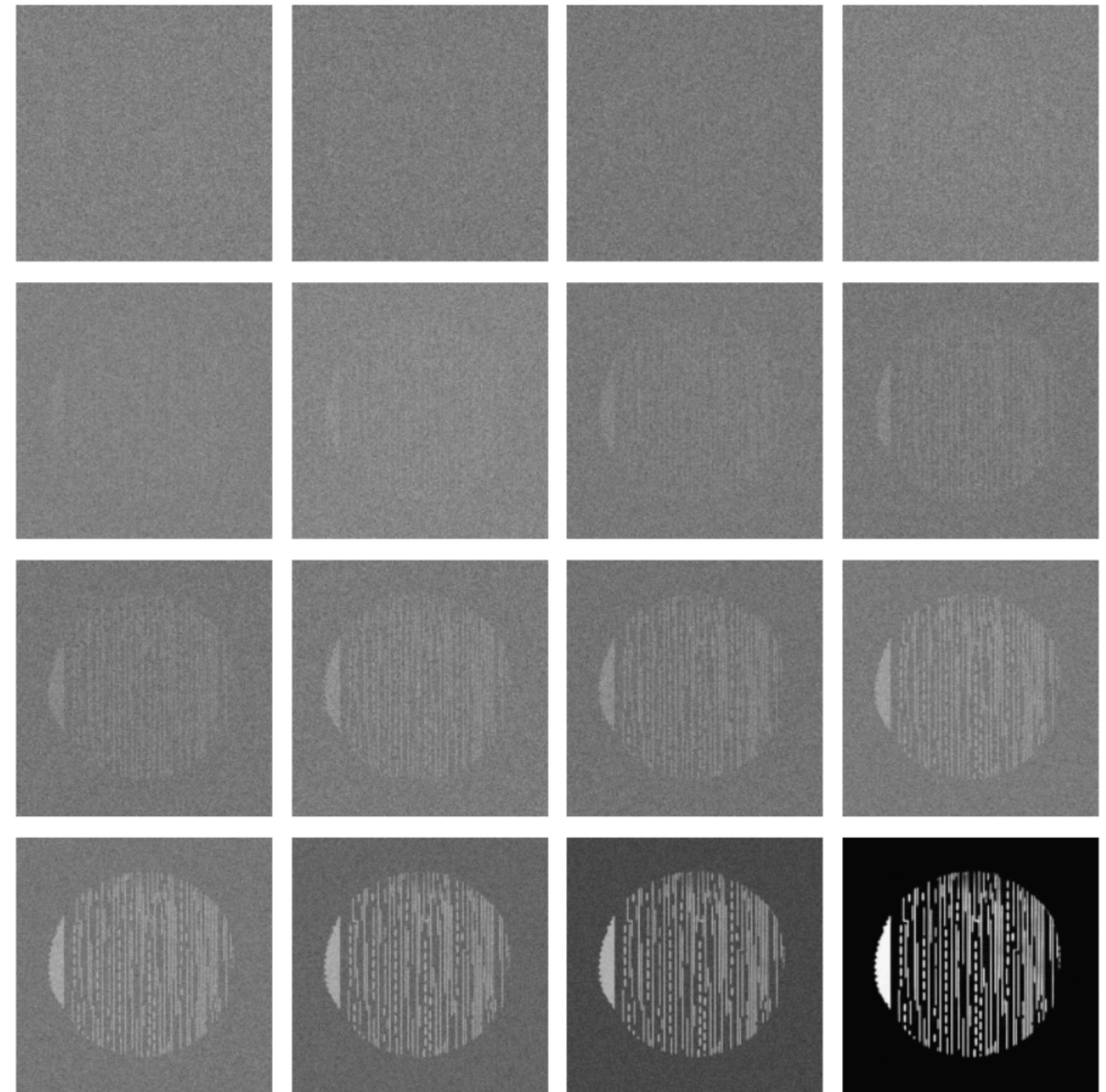


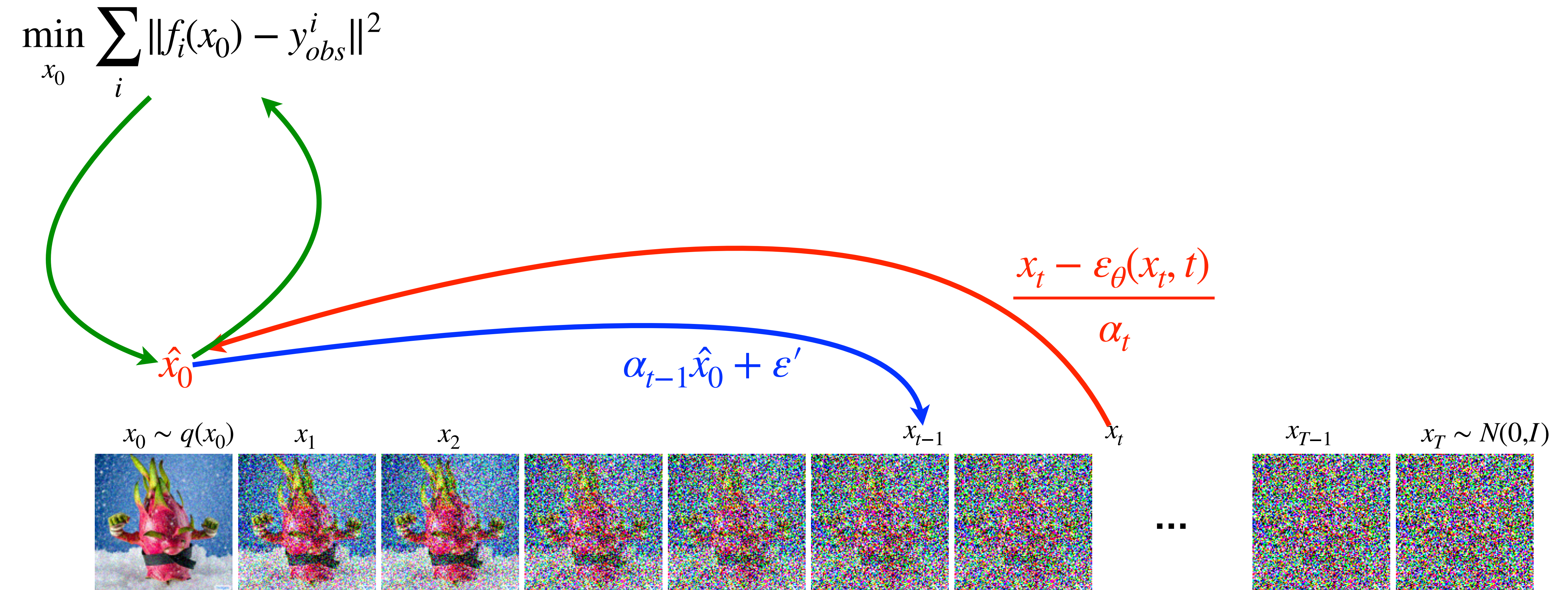
- Diffusion models are trained to produce an image from gaussian noise.
- Over say 1000 steps, starting from gaussian noise, the model moves towards the image distribution it was trained on.
- We work for this project with tomograms of the blueprints of chips.



- Diffusion models are trained to produce an image from gaussian noise.
- Over say 1000 steps, starting from gaussian noise, the model moves towards the image distribution it was trained on.
- We work for this project with tomograms of the blueprints of chips.

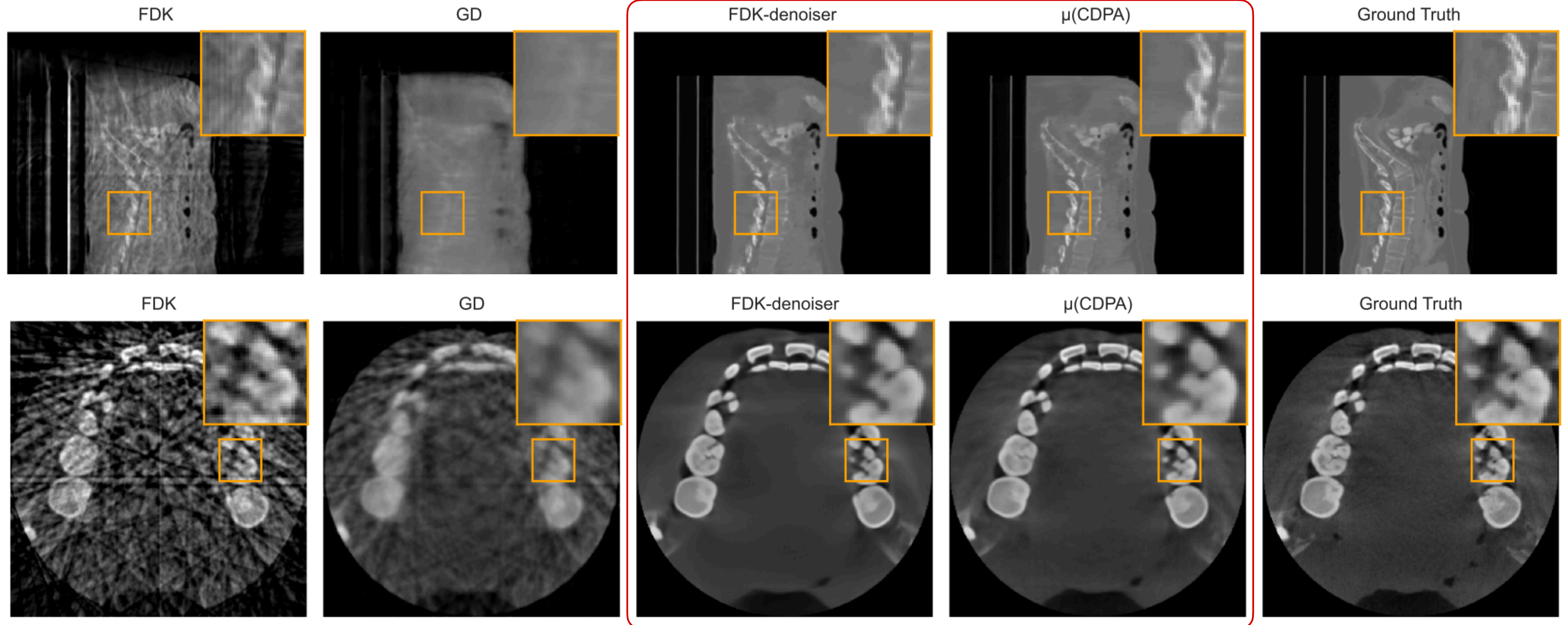
Challenge: guide the generative process to match the partial observations we have





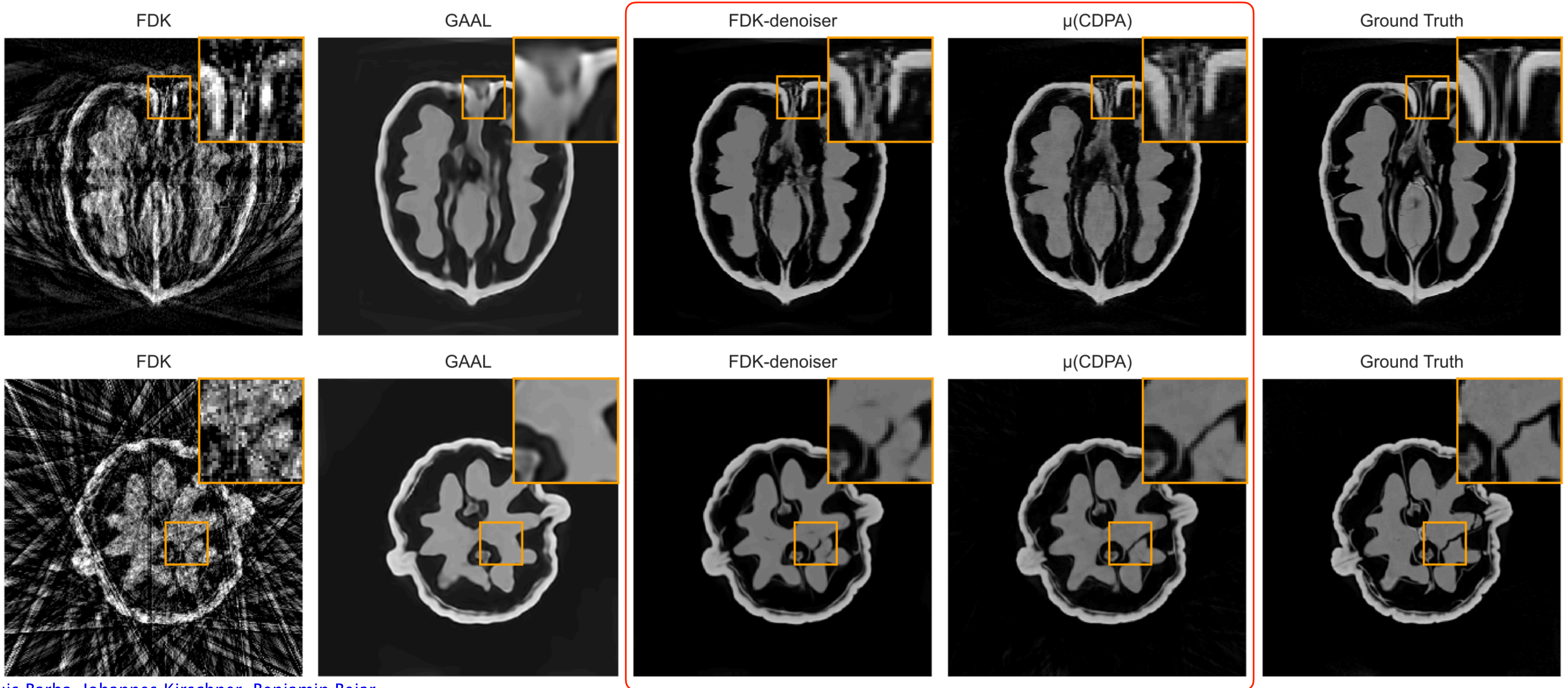
CDPA FOR CBCT RECONSTRUCTIONS

Reconstructions on 20 equally-spaced
measurements of 256^3 volumes



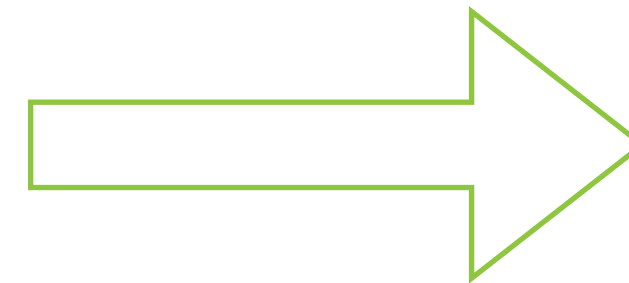
CDPA FOR CBCT RECONSTRUCTIONS

Reconstructions on 20 equally-space measurements of 256^3 volumes

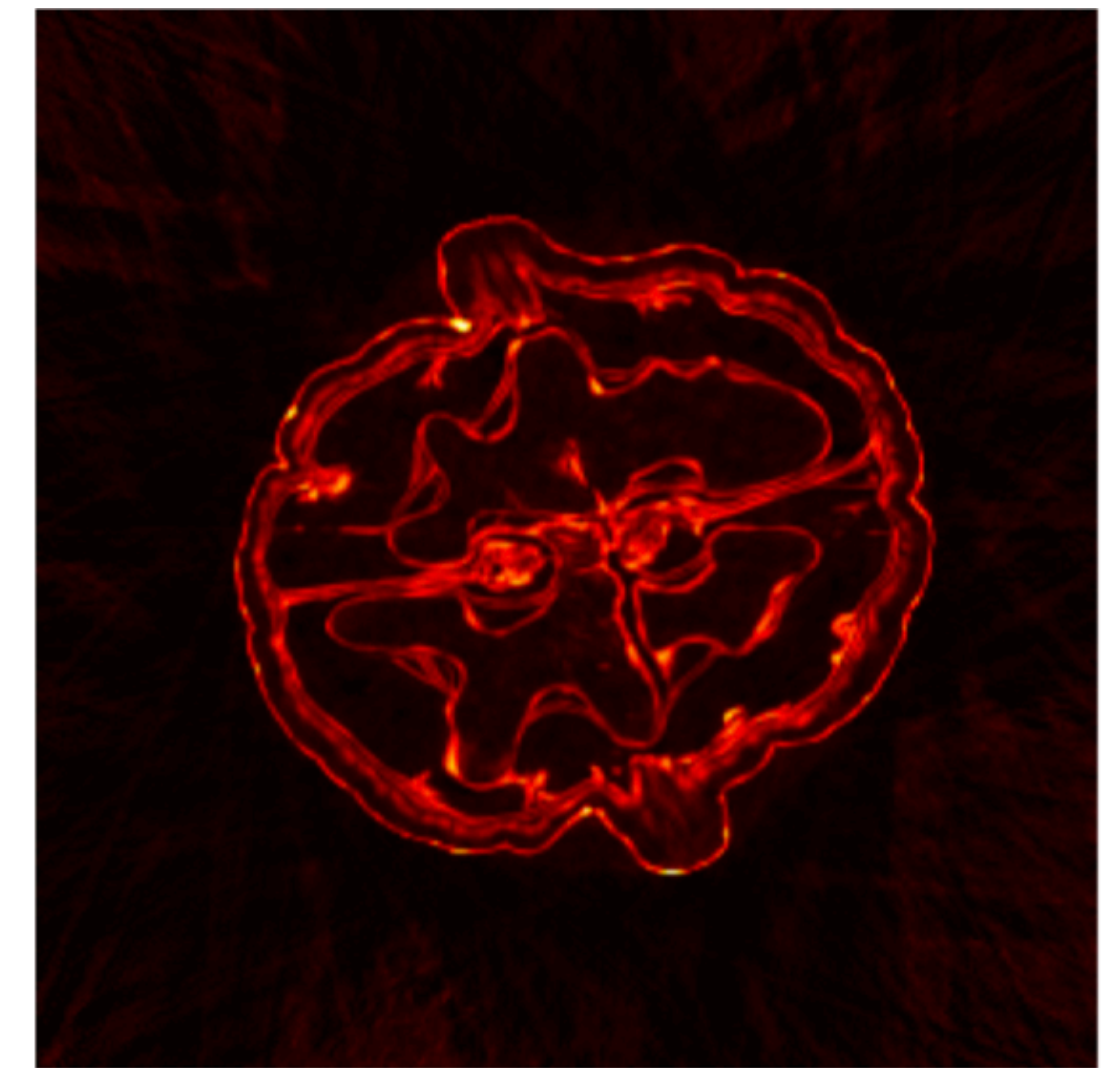


CDPA: UNCERTAINTY QUANTIFICATION

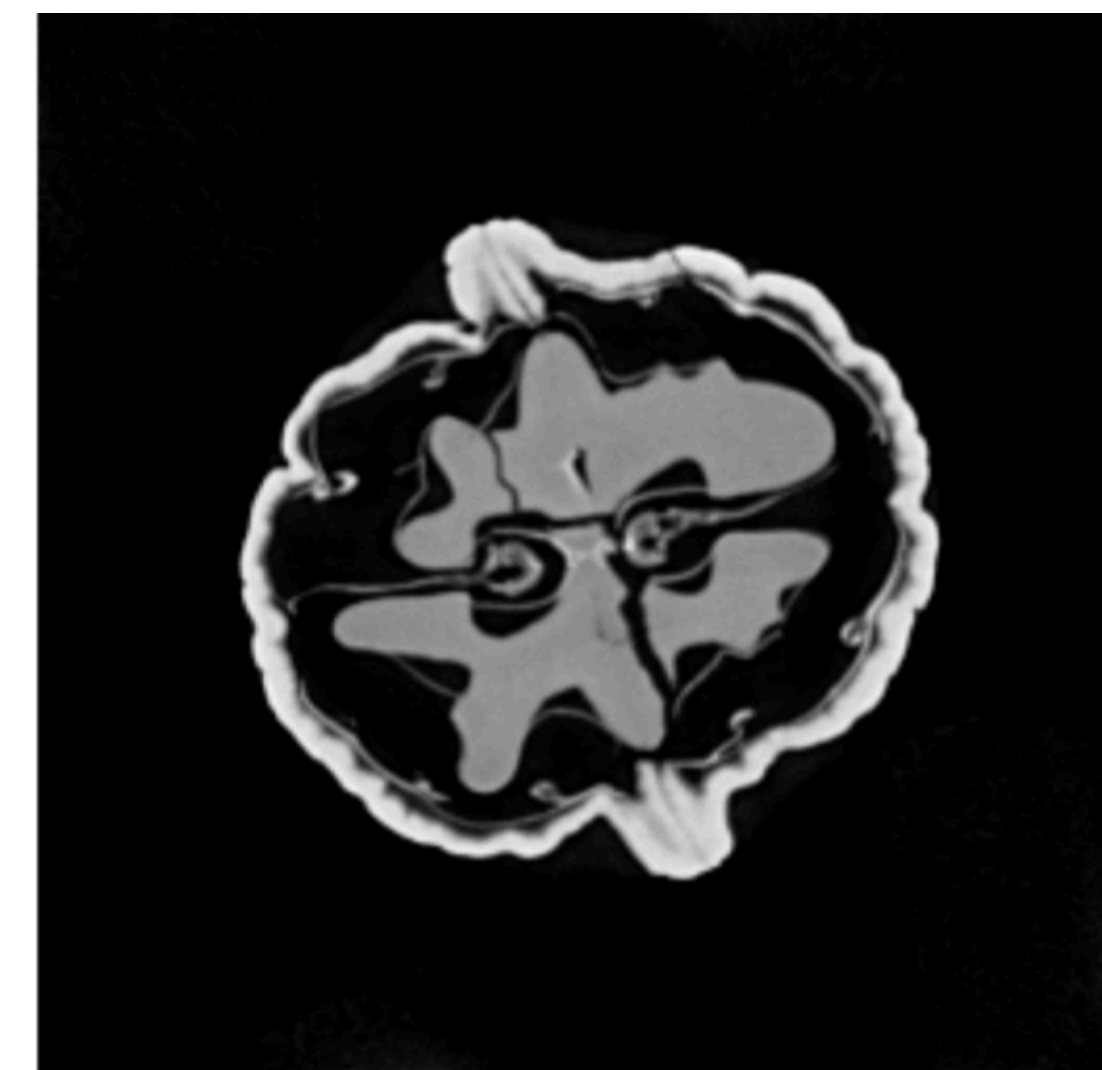
independent diffusion samples



std

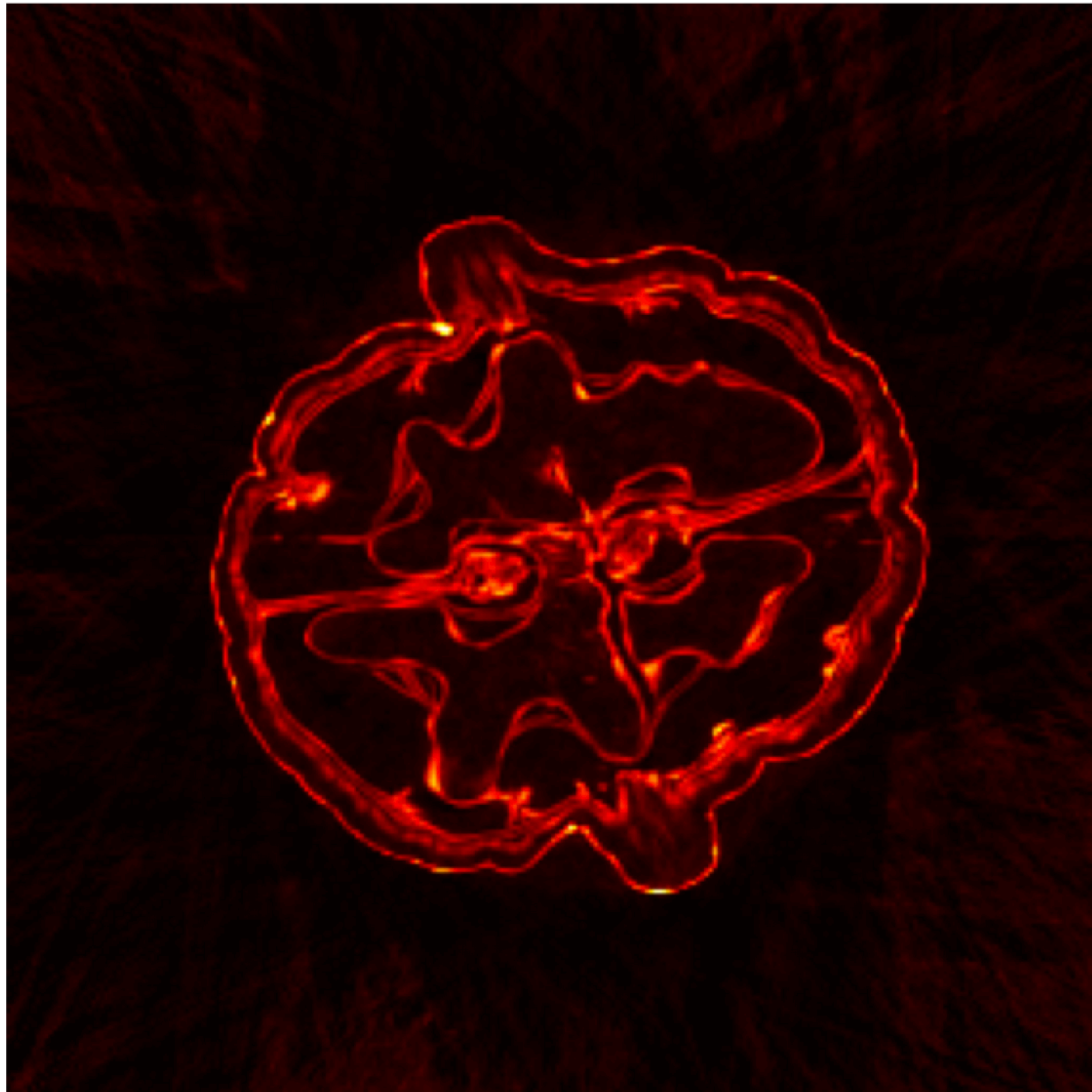


mean

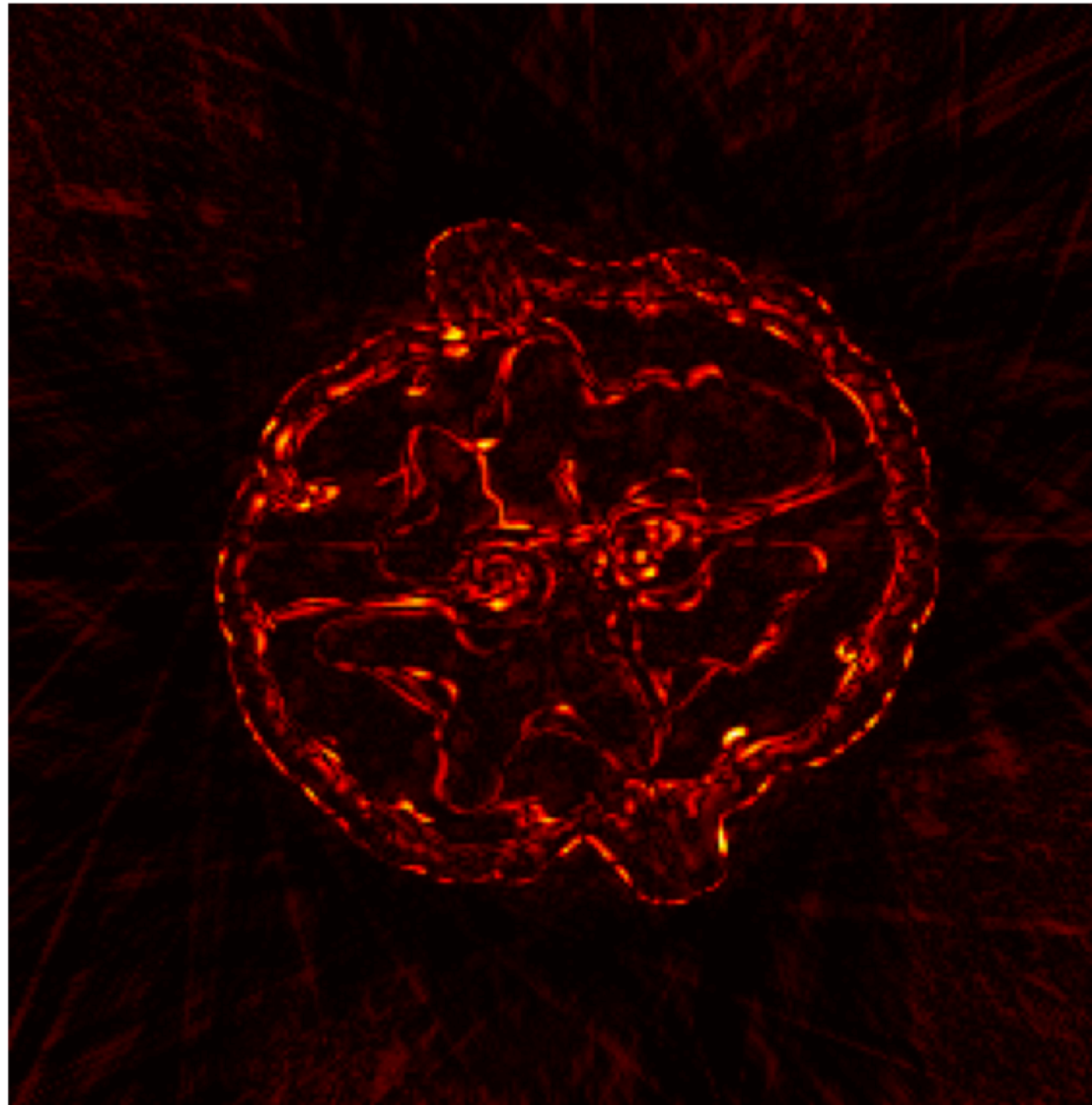


CDPA: UNCERTAINTY QUANTIFICATION

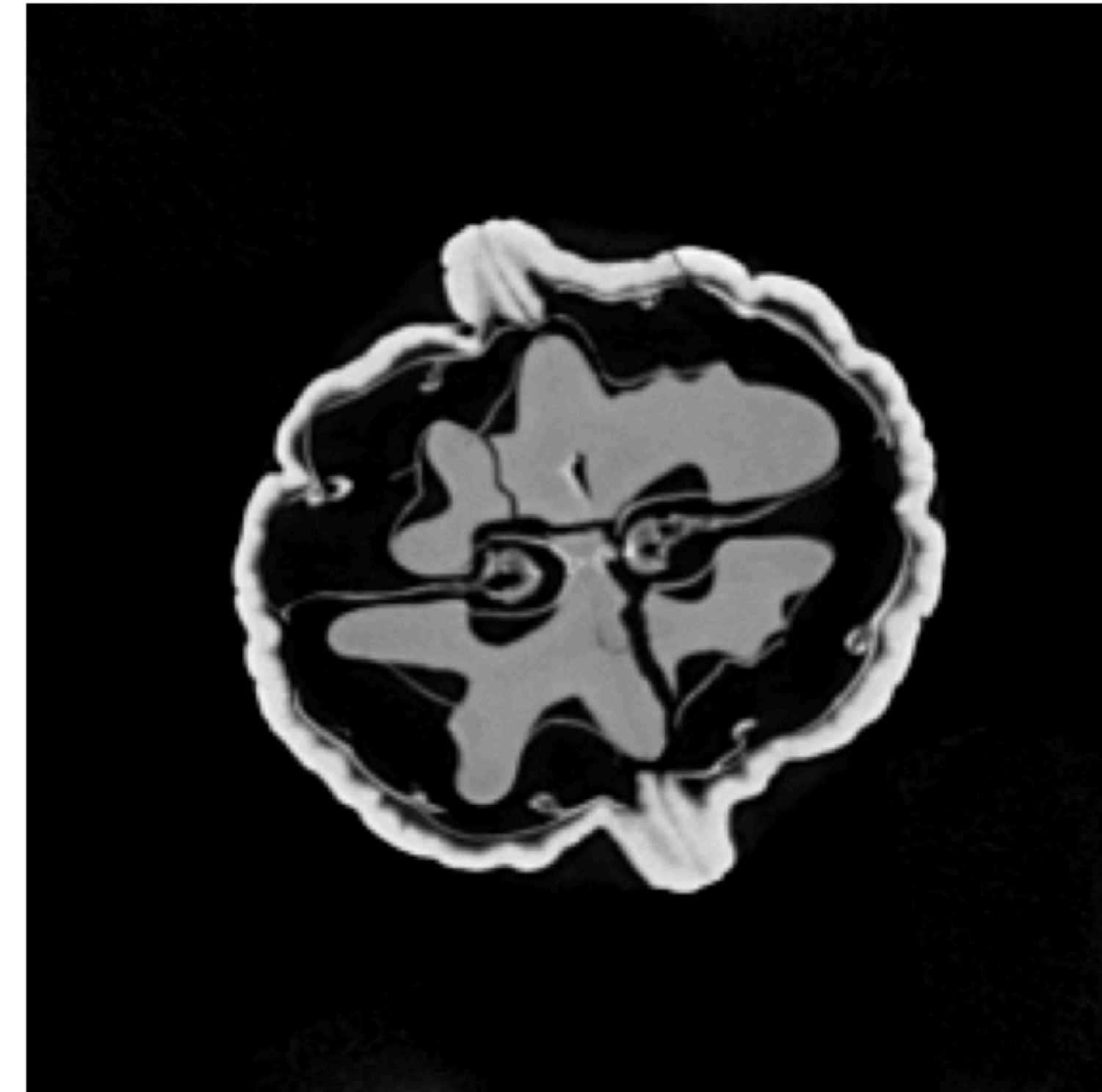
STD of 10 CDPA samples



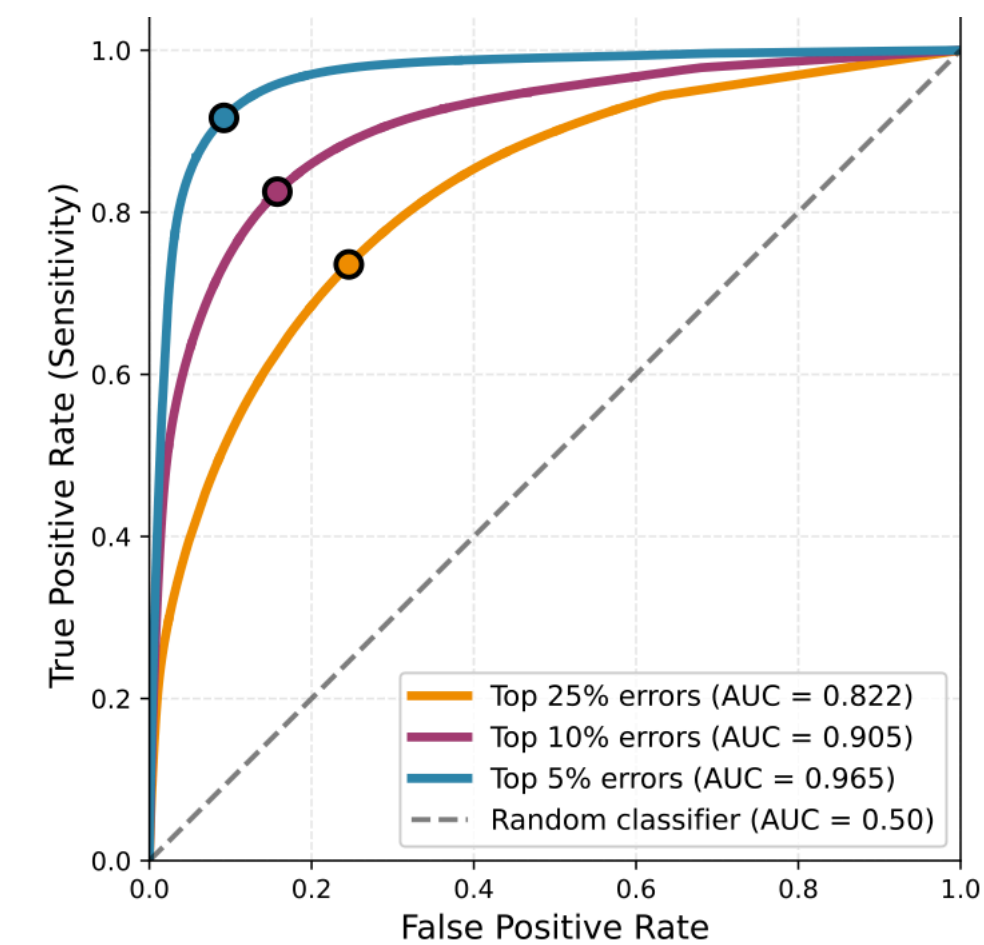
Abs Difference ($\mu(\text{CDPA}) - \text{GT}$)



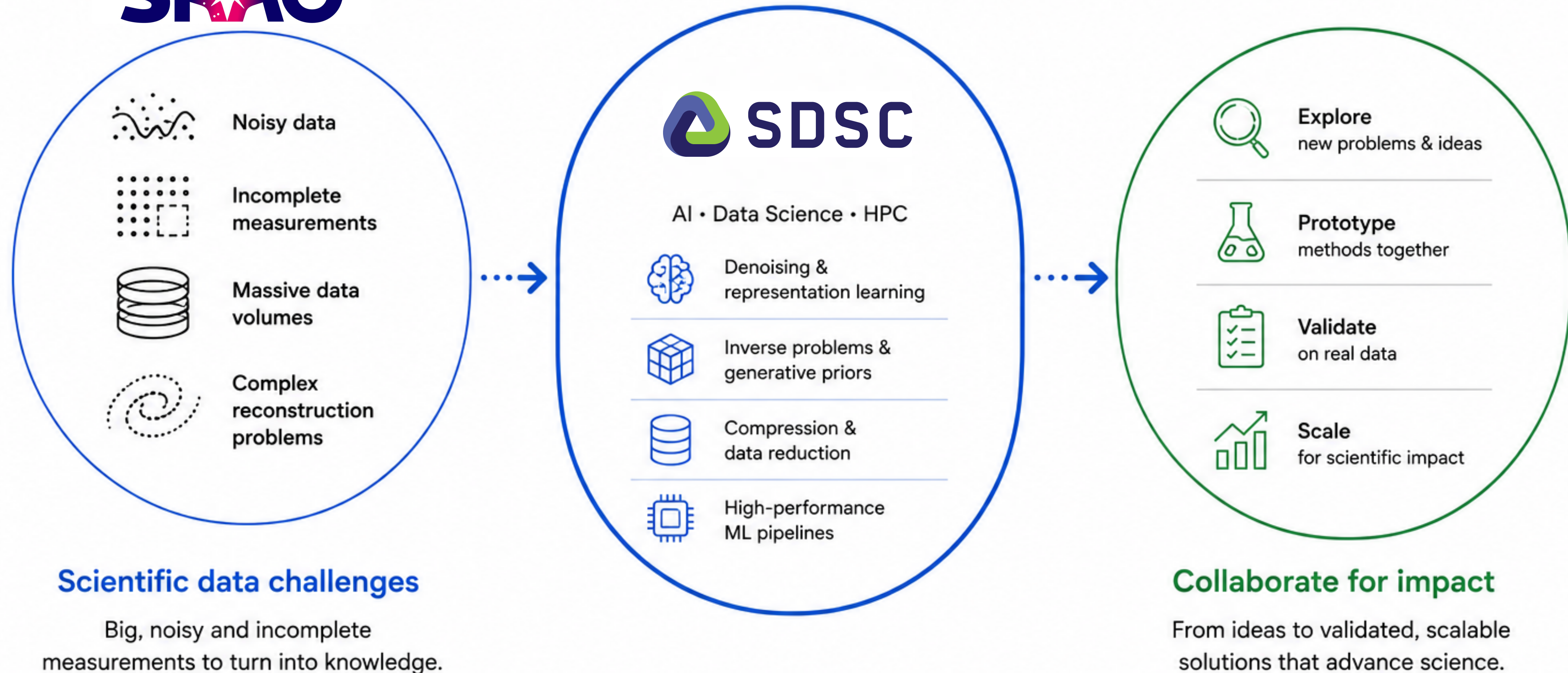
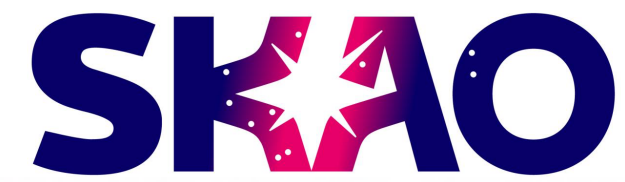
Ground Truth (GT)



High uncertainty is a great predictor for high error.



CONCLUSIONS



Let's explore these challenges together.

Different expertise. Shared data challenges. Better solutions together.