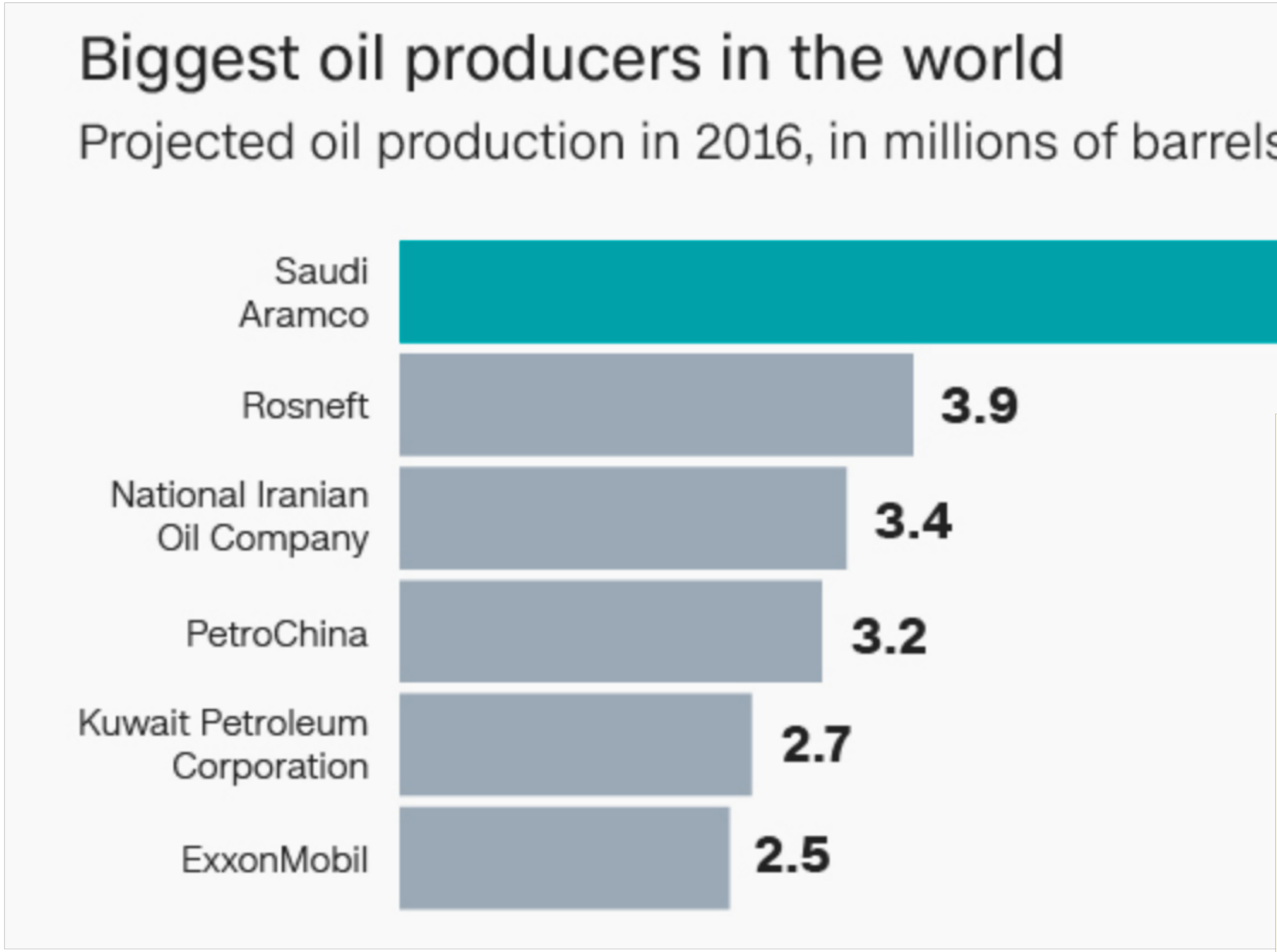
Big Data and Extreme Computing Series 2: Edge Computing

Application/Industry Perspective

David Keyes
KAUST
Saudi Arabia

Upstream petroleum industry



Aramco cost to produce: approx. USD 10 per barrel

Source: CNN Business

A typical seismic imaging campaign

10 crews of about 900 people each on 24-hour operations in Saudi Arabia, with an estimated 270 billion barrels of oil, eager to discover more

A typical seismic imaging campaign

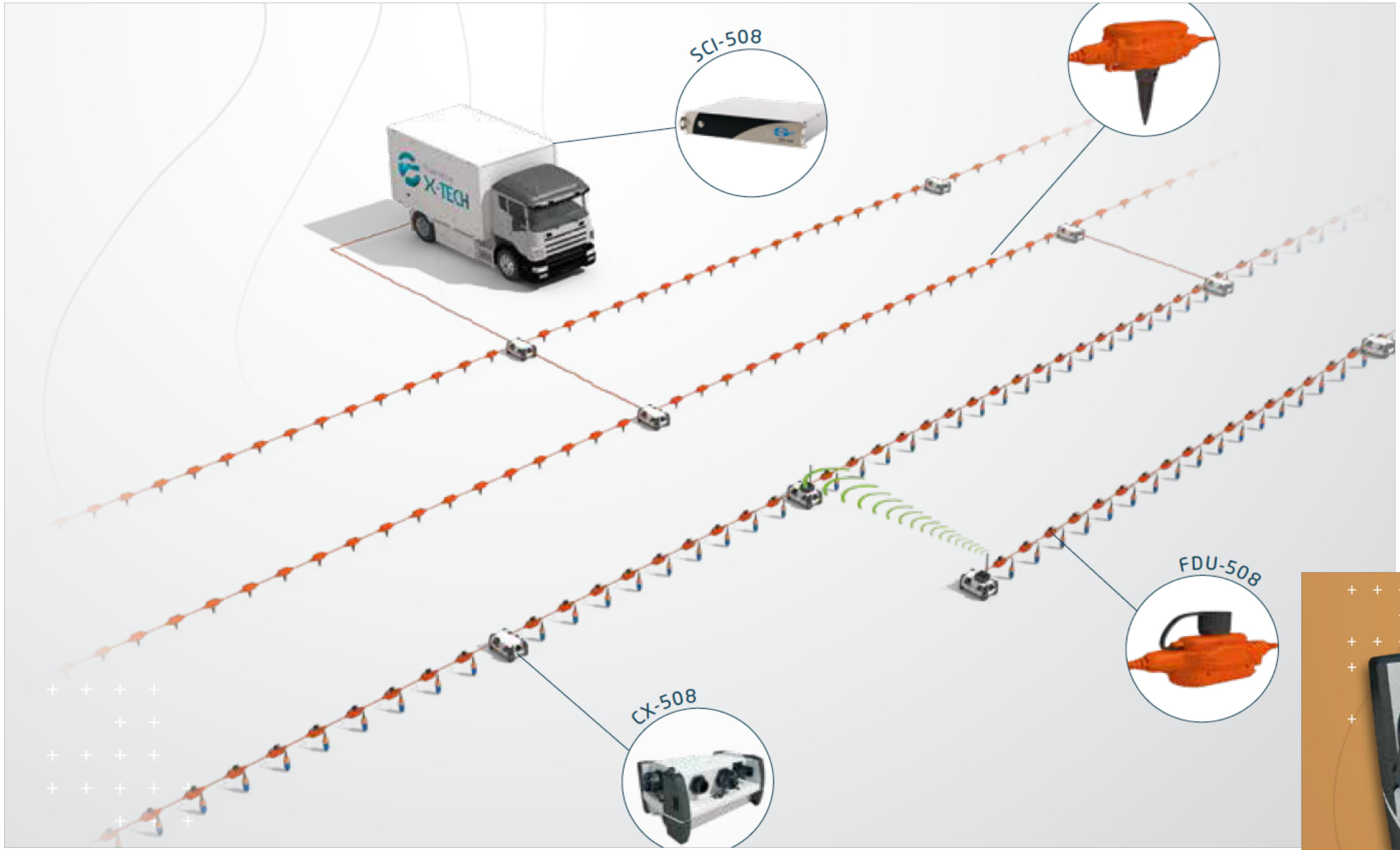
Approximately 15,000 “shots” per day on average (86,400 sec/day)

65-ton vibrator trucks drop steel plates on the ground to create acoustic pulses

A typical seismic imaging campaign

50,000 “strings” deployed to recover the acoustic signals, each containing about 10 geophones, on fiber optic cables; about 500,000 receivers per shot

A typical seismic imaging campaign



About 3 TB of acoustic data collected per day from multiple shots detected by multiple devices

Million-channel recording device with built-in GPS and long-life batteries

Data collection

- So far, all data is gathered and preserved in raw form
- The vibrator signal sweeps a frequency range
- The more longer the time spent on a frequency range the better the signal to noise ratio
- Typically sweep duration is 6s to 12s
- Listening time is also typically 6s to 12s
- Total shot recording time is the sum of the sweep & listening time, so 12s to 24s per shot
- With 24s per shot one single source can generate up to 3600 shot/day
- To achieve higher productivity many sources (10 to 20) acquire shots at the same time

Meta data on the data

- Typical sampling interval is 2ms
- Maximum frequency is approximately 100Hz, so that 2ms is usually sufficient; in 1ms sampling interval is also possible
- The precision is usually 24 bits and data are stored using 32bits
- The acquisition cost is much higher than the storage so all the data is kept
- At recording time, limited processing can happen such as receiver stacking, basic filtering, and resampling

Meta data on the data

- Acquisition is typically in remote areas
- Field data are stored at acquisition time on disks
- These data are periodically (daily) copied on external storage (disks) and theses disks are physically transferred to centralized storage
- Backup tapes are also created
- 3 TB per day is about 1 PB/year (*not* the SKA ☺)

SKA will generate 11 EB per day when operational

- Further processing produces far more data than acquisition
- *Processing* a full campaign (e.g., 100 km x 100 km x 8 km) takes several months
- Currently, little artificial intelligence is used, compared to standard physics-based inversion algorithms

Final product

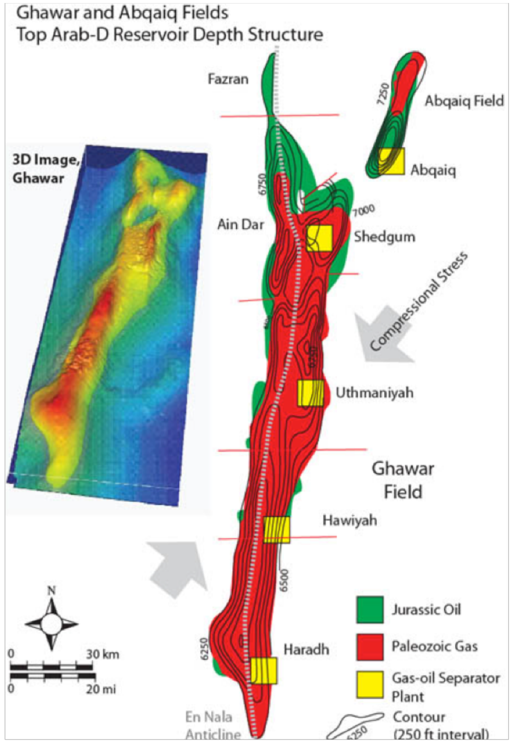
- Size of the final product is usually similar to the the size of the input data
- Final product can include 3d volume of reflection coefficients (impedance contrasts) and/or physical parameters such as acoustic and shear velocities
  - This is fed into the reservoir production simulation codes
  - Pumping scenarios used to optimize production
- Volume of data increases year on year
  - Higher frequencies, denser shots, and denser receivers
- Processing the data into a subsurface image is a far larger application of “big data”

Comment

- More important to pay attention to the data generated by the various imaging techniques on supercomputers than on the acquisition
- Even so, the world's largest oil field is already resolvable at imaging scales within a Top 20 supercomputer

Lining up geological & computational scales

- Volume of the Earth: approximately 10^{12} km^3 ($4\pi R^3/3$, $R = 6371$ km)
- Volume of Ghawar reservoir: approximately 10^{12} m^3… fits in box 300 km × 33 km x 100 m (with 10% padding)
- Ghawar's volume is 1 part in 10^9 of the Earth … and a very important one-billionth it *is* ☺
- To resolve the earth to 1 kilometer, *or equivalently* Ghawar to 1 meter, requires 1 Tera-word of data per gridded field
  - at double precision (8 Bytes/per word) requires 8 TeraBytes
  - this is *not* a daunting problem today



Ghawar: world's largest continuous-pressure reservoir

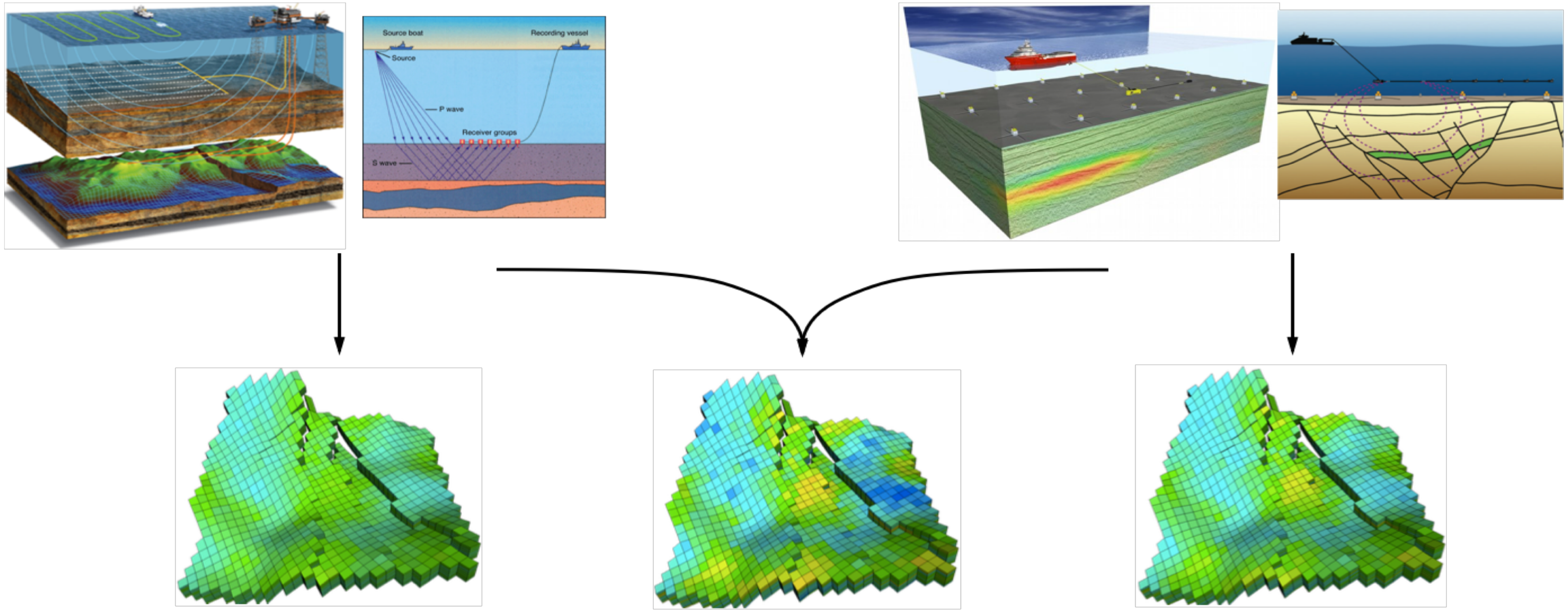
Lining up geological & computational scales

- KAUST's Shaheen XC40 has approximately 792 TeraBytes of DRAM
  - about *100 copies* of the Earth at 1 km or Ghawar at 1 m resolution fit within Shaheen
- Enough for
  - several components per grid cell
  - auxiliary workspaces for constitutive properties
  - sparsely stored adjoint, Jacobian, preconditioner …
  - visualization, etc.
- We cannot resolve all relevant scales (pores, etc.)

  … but we can already out-resolve typical detail from seismic inputs
  - with adaptive discretization we can scale down to wells

Geophysical Joint Inversion: motivation

Survey images from Schlumberger, Rock Solid Images, and EMGS



seismic alone

joint

electromagnetic alone

- The combination of two imaging modalities takes advantage of their complementary strengths allowing improved contrast and resolution

Joint inversion: formulation

- In joint inversion of data acquired from two different physics modalities, we seek to minimize a regularized nonlinear least squares functional of the form:

$$J(\mathbf{m}^{\mathrm{s}}, \mathbf{m}^{\mathrm{e}}) = \underbrace{\boxed{\frac{1}{2}\|\mathbf{f}^{\mathrm{s}}(\mathbf{m}^{\mathrm{s}}) - \mathbf{d}^{\mathrm{s}}_{\mathrm{obs}}\|^2} + \boxed{\frac{1}{2}\|\mathbf{f}^{\mathrm{e}}(\mathbf{m}^{\mathrm{e}}) - \mathbf{d}^{\mathrm{e}}_{\mathrm{obs}}\|^2}}_{\text{data misfit}} + \underbrace{\boxed{\alpha^{\mathrm{s}}\|\mathbf{m}^{\mathrm{s}}\|_{\mathrm{TV}}} + \boxed{\alpha^{\mathrm{e}}\|\mathbf{m}^{\mathrm{e}}\|_{\mathrm{TV}}}}_{\substack{\text{single-field}\\ \text{regularization}}} + \underbrace{\boxed{\beta R(\mathbf{m}^{\mathrm{s}}, \mathbf{m}^{\mathrm{e}})}}_{\substack{\text{incoherence}\\ \text{between fields}}}$$

- The first two terms represent data misfits between the outputs of the individual physics models (seismic & electromagnetic) and their associated observations.
- The next two terms represent regularization of each parameter field (here, edge-preserving ("total variation") regularization).
- The last term is a "structural similarity" term that penalizes incoherence between the two parameter fields.
  - A popular choice is the cross product of the gradients of the two fields, which favors parameter fields that have contours of similar shape.
  - Alternative forms of imposing structural similarity between $\mathbf{m}^{\mathrm{s}}$ and $\mathbf{m}^{\mathrm{e}}$ are being investigated, including vector total variation and nuclear norm.

Joint inversion: formulation

- In joint inversion of data acquired from two different physics modalities, we seek to minimize a regularized nonlinear least squares functional of the form:

$$J(\mathbf{m}^{\mathrm{s}}, \mathbf{m}^{\mathrm{e}}) = \underbrace{\boxed{\frac{1}{2}\|\mathbf{f}^{\mathrm{s}}(\mathbf{m}^{\mathrm{s}}) - \mathbf{d}^{\mathrm{s}}_{\mathrm{obs}}\|^2} + \boxed{\frac{1}{2}\|\mathbf{f}^{\mathrm{e}}(\mathbf{m}^{\mathrm{e}}) - \mathbf{d}^{\mathrm{e}}_{\mathrm{obs}}\|^2}}_{\text{data misfit}} + \underbrace{\boxed{\alpha^{\mathrm{s}}\|\mathbf{m}^{\mathrm{s}}\|_{\mathrm{TV}}} + \boxed{\alpha^{\mathrm{e}}\|\mathbf{m}^{\mathrm{e}}\|_{\mathrm{TV}}}}_{\substack{\text{single-field}\\ \text{regularization}}} + \underbrace{\boxed{\beta R(\mathbf{m}^{\mathrm{s}}, \mathbf{m}^{\mathrm{e}})}}_{\substack{\text{incoherence}\\ \text{between fields}}}$$

- The $\mathbf{s}$ terms represent seismic parameters and observations
- The $\mathbf{e}$ terms represent electromagnetic parameters and observations
- The cross term represents the coupling (zero for independent)

Coupling mechanisms

cross-gradient

$$\hat{\mathcal{R}}_{\mathrm{cg}}(m_1, m_2) = \frac{1}{2} \int_\Omega |\nabla m_1 \times \nabla m_2|^2 \, dx$$

normalized cross-gradient

$$\mathcal{R}_{\mathrm{ncg}}(m_1, m_2) = \int_\Omega \left| \frac{\nabla m_1}{|\nabla m_1|} \times \frac{\nabla m_2}{|\nabla m_2|} \right|^2 \, dx$$

vectorial total variation

$$\mathcal{R}_{\mathrm{VTV}}(m_1, m_2) := \int_\Omega \sqrt{|\nabla m_1|^2 + |\nabla m_2|^2 + \varepsilon} \, dx$$

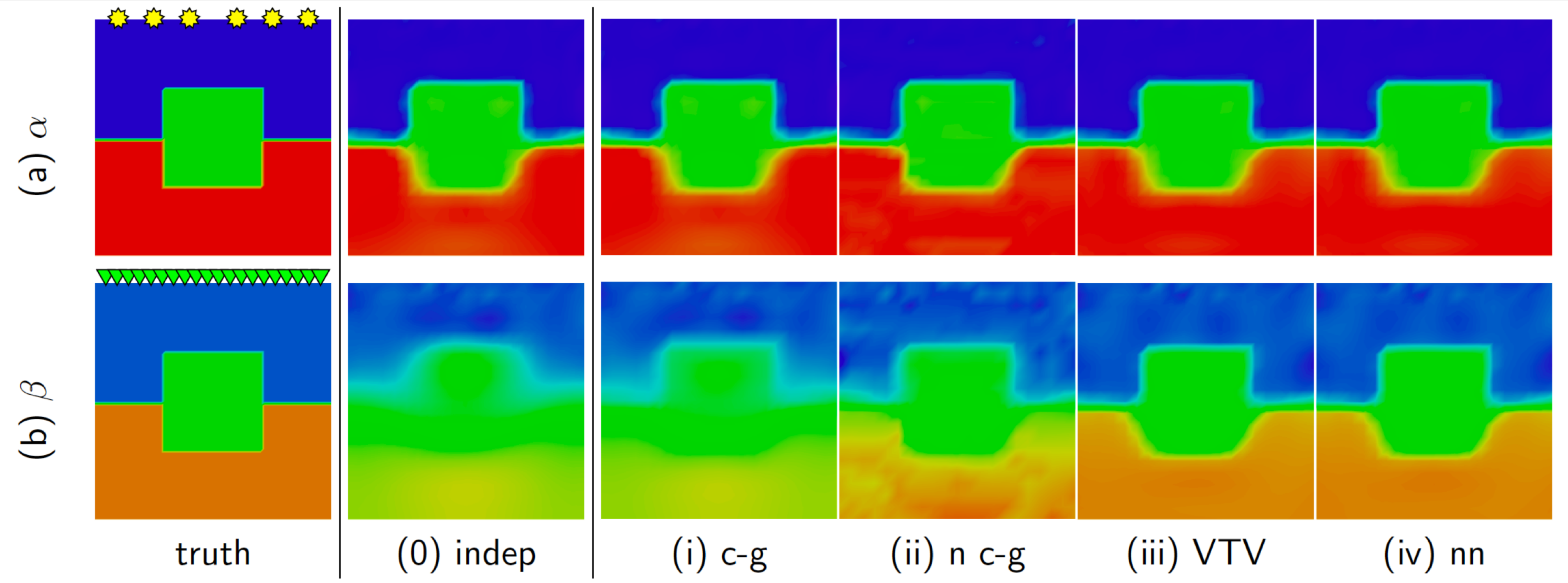
nuclear norm

$$\mathcal{R}_*(m_1, m_2) := \int_\Omega \|\mathbf{G}(x)\|_* \, dx$$

where

$$\mathbf{G}(x) := [\nabla m_1 | \nabla m_2] = \begin{bmatrix} \partial_{x_1} m_1 & \partial_{x_1} m_2 \\ & \vdots & \\ \partial_{x_d} m_1 & \partial_{x_d} m_2 \end{bmatrix}$$

Joint inversion: 2D model



Inversion of acoustic data for two fields—density (top) and bulk modulus (bottom). Left shows the "truth" fields. Next are independent ('indep') inversions. The cross-gradient penalty ('c-g')and normalized cross-gradient ('n c-g') are shown next. The Vectorial Total Variation function ('VTV') is next, followed by the nuclear norm ('nn').