Trustworthy AI Systems: Lessons Learned from an Arms Control Application

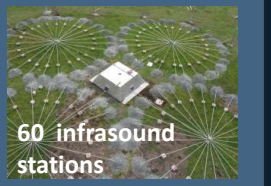
Elena Tomuta,
Chief, Software Applications Section,
International Data Centre, CTBTO

CTBT: Comprehensive Nuclear Test-Ban-Treaty

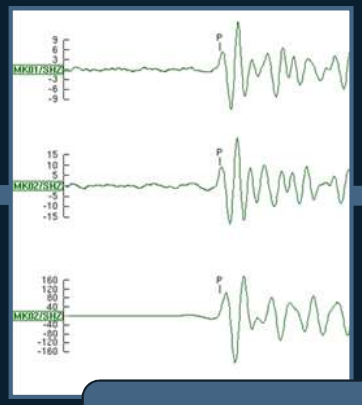
- Establishes a ban on all nuclear explosions by everyone, everywhere: on the Earth's surface, in the atmosphere, underwater and underground.

- Establishes a verification regime, including an international monitoring system.

CTBTO Verification Regime: Processing Sesimic, Hydroacoustic and Infrasound data



Acquire data



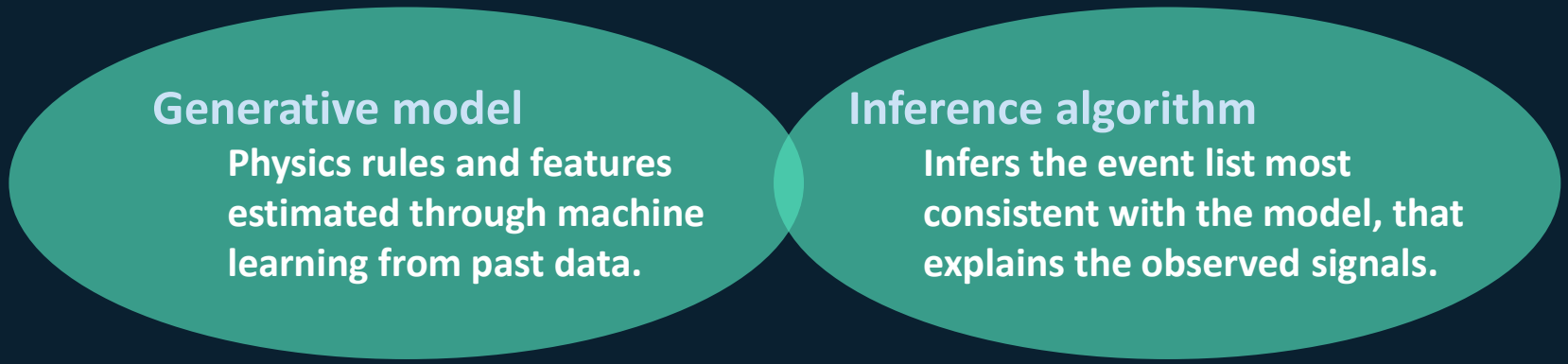
Detect signals

Build Events

Review and correct automatic event bulletins

Building Events: from a rule-based system to machine learning

NET-VISA (**N**ETwork processing **V**ertically **I**ntegrated **S**eismic **A**nalysis)
University of California at Berkeley, by Prof. Stuart Russell and Dr. Nimar Arora



Trust:

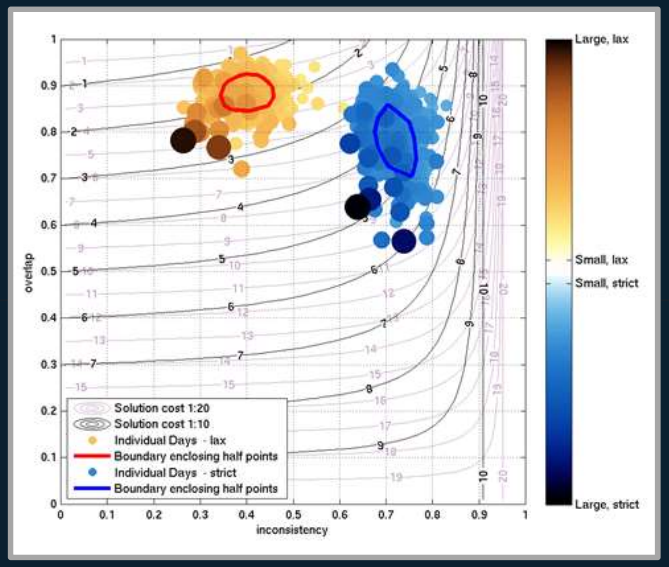
Firm belief in the reliability, truth or ability of someone or something.

What does better mean for us?

What is our ground truth?

Characterizing the event set:

- Overlap: Percentage of ground truth events found
- Inconsistency: Percentage of false events built



ability

Event Quality

- Location accuracy
- Quality of associations

Risk

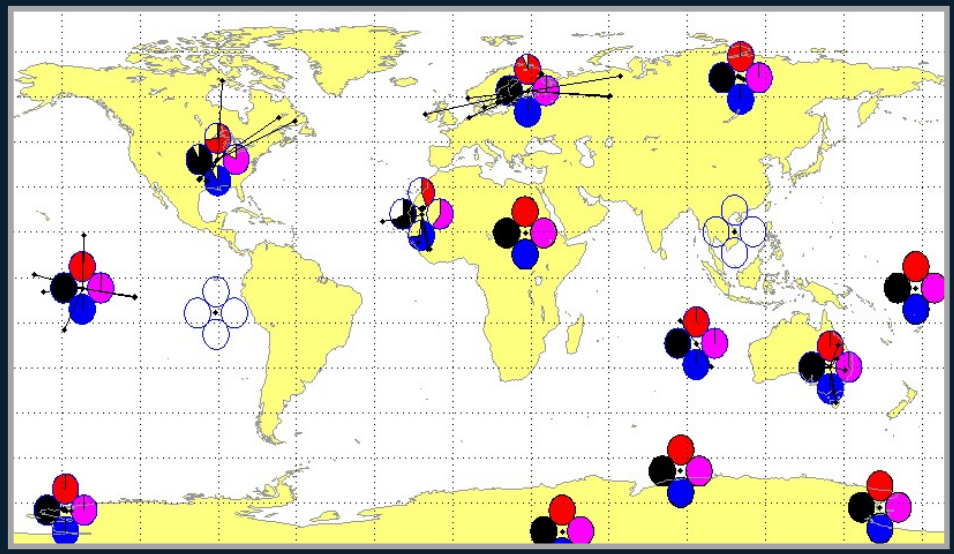
Bias:

- Training data dominated by natural events.

reliability

How do we ensure that we do not miss man-made events?

- Model design
- Targeted tests



Stakeholder "culture" influences perception

- "The model is not physical"
- "It's a black-box"

belief

Changing perception:

- Transparency: code can be examined
- Explain the model and the algorithm
- Explain individual events
- Involve stakeholders in testing and use

What does better mean for us?

Stakeholder perception

ground truth

Bias

Risk

belief reliability ability

Transparency

Explain

Stakeholder involvement