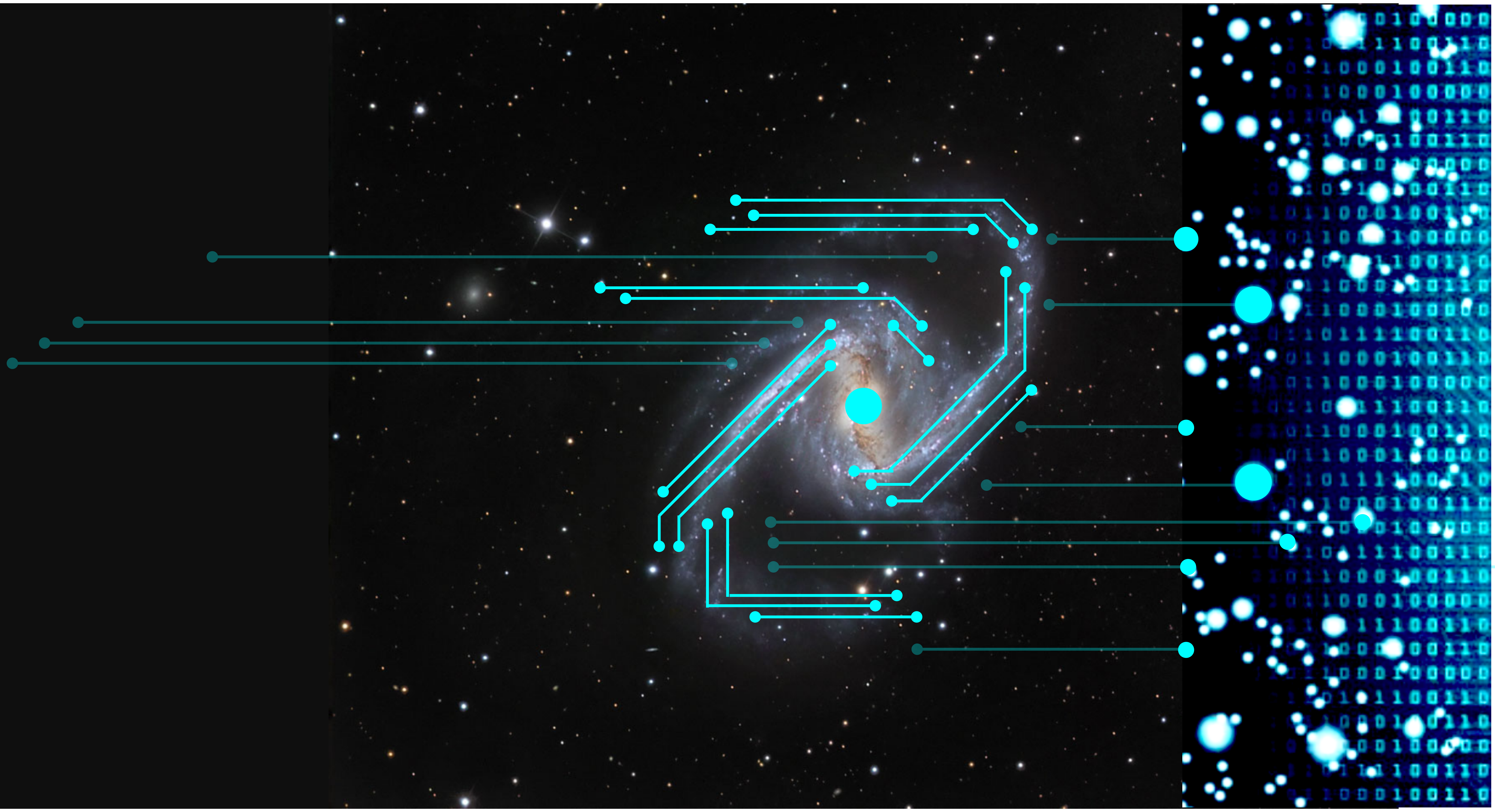
AI for Astronomy in the SKA Era: learning semantically meaningful classification targets for radio astronomy

Anna Scaife - Jodrell Bank Centre for Astrophysics

The Alan Turing Institute

MANCHESTER 1824
The University of Manchester

SKAO



21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research
26 October 2022

as595
radastrat
anna.scaife@manchester.ac.uk

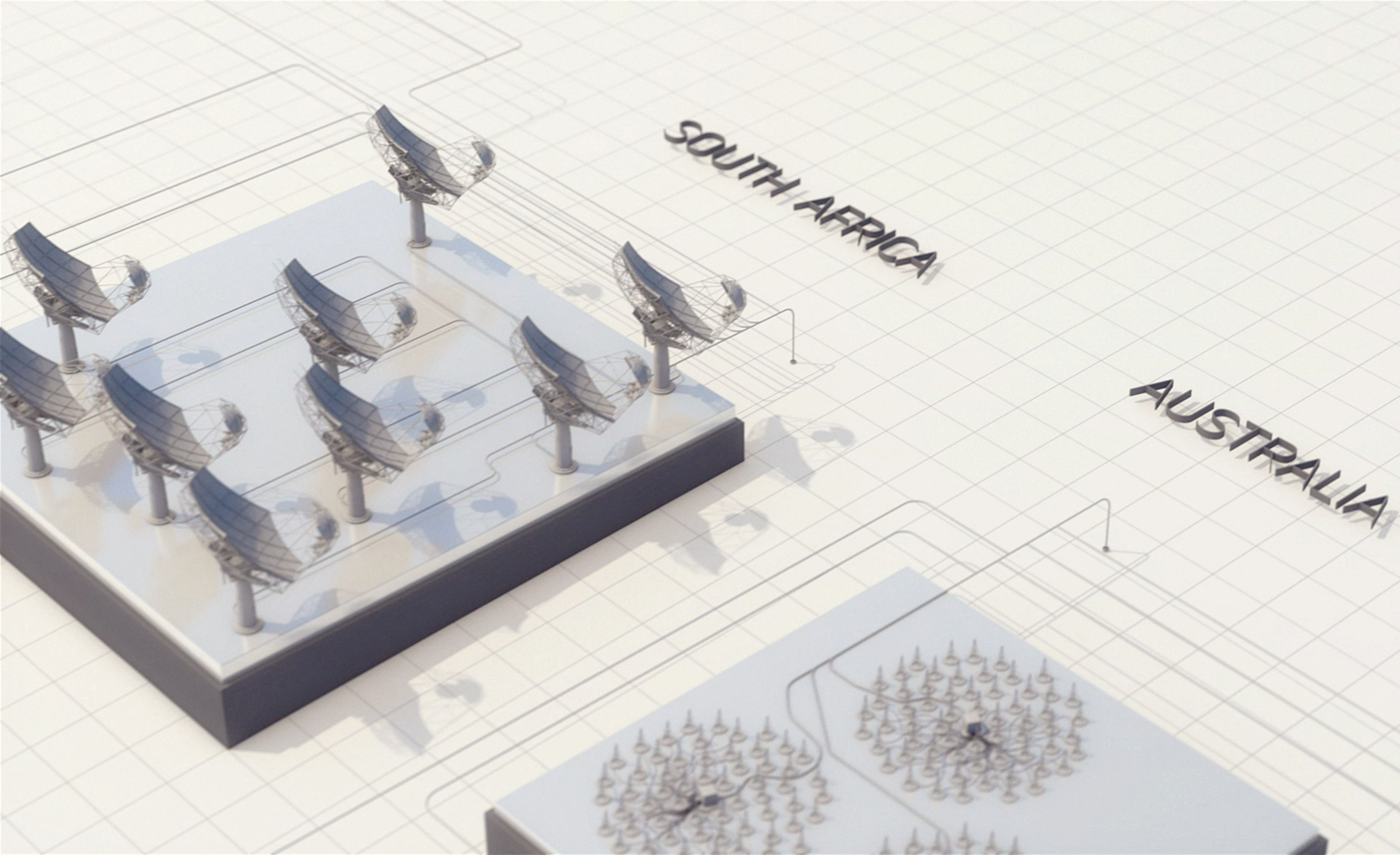

ACAT 2022
BARI

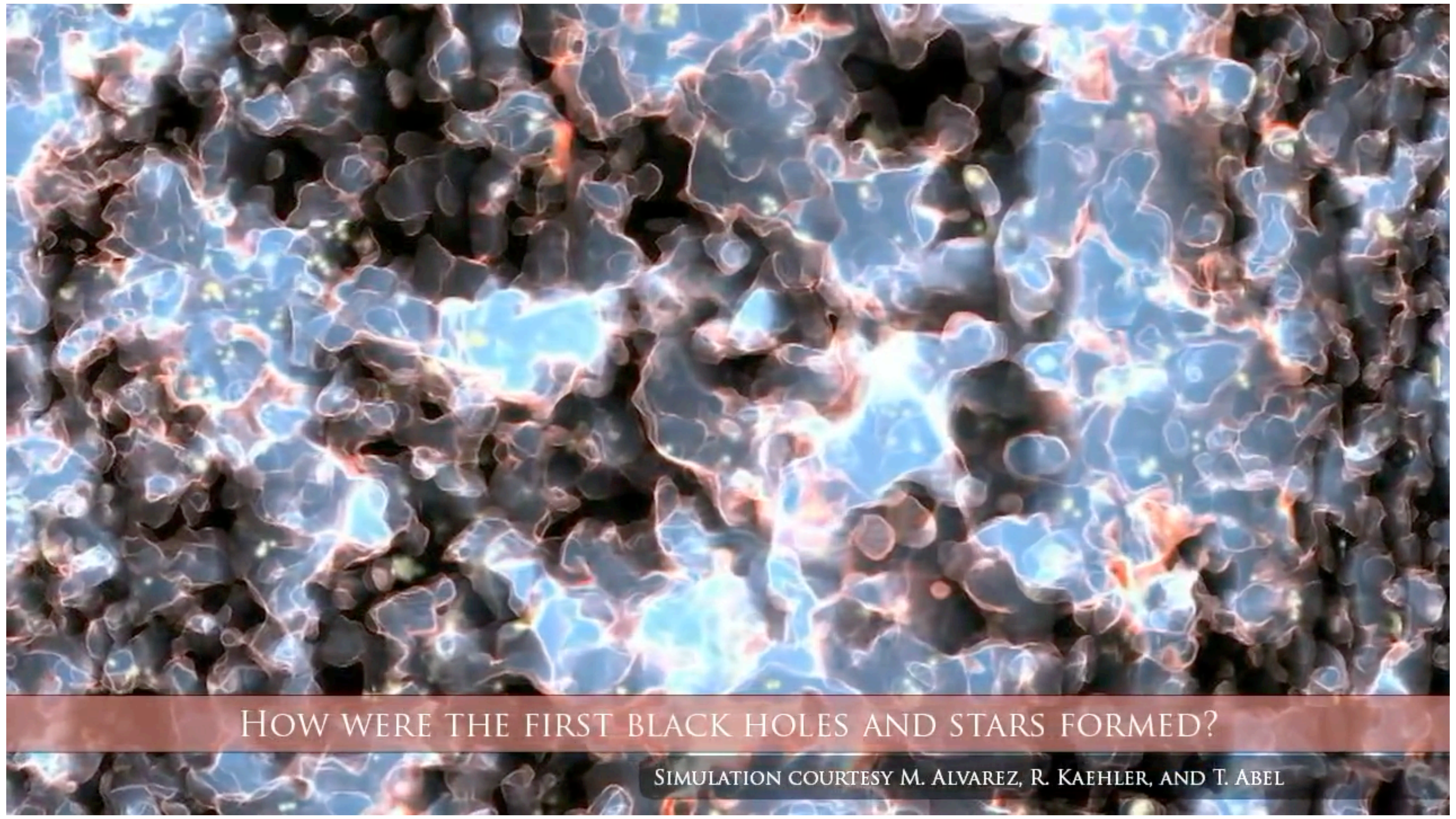

ACAT 2022
BARI

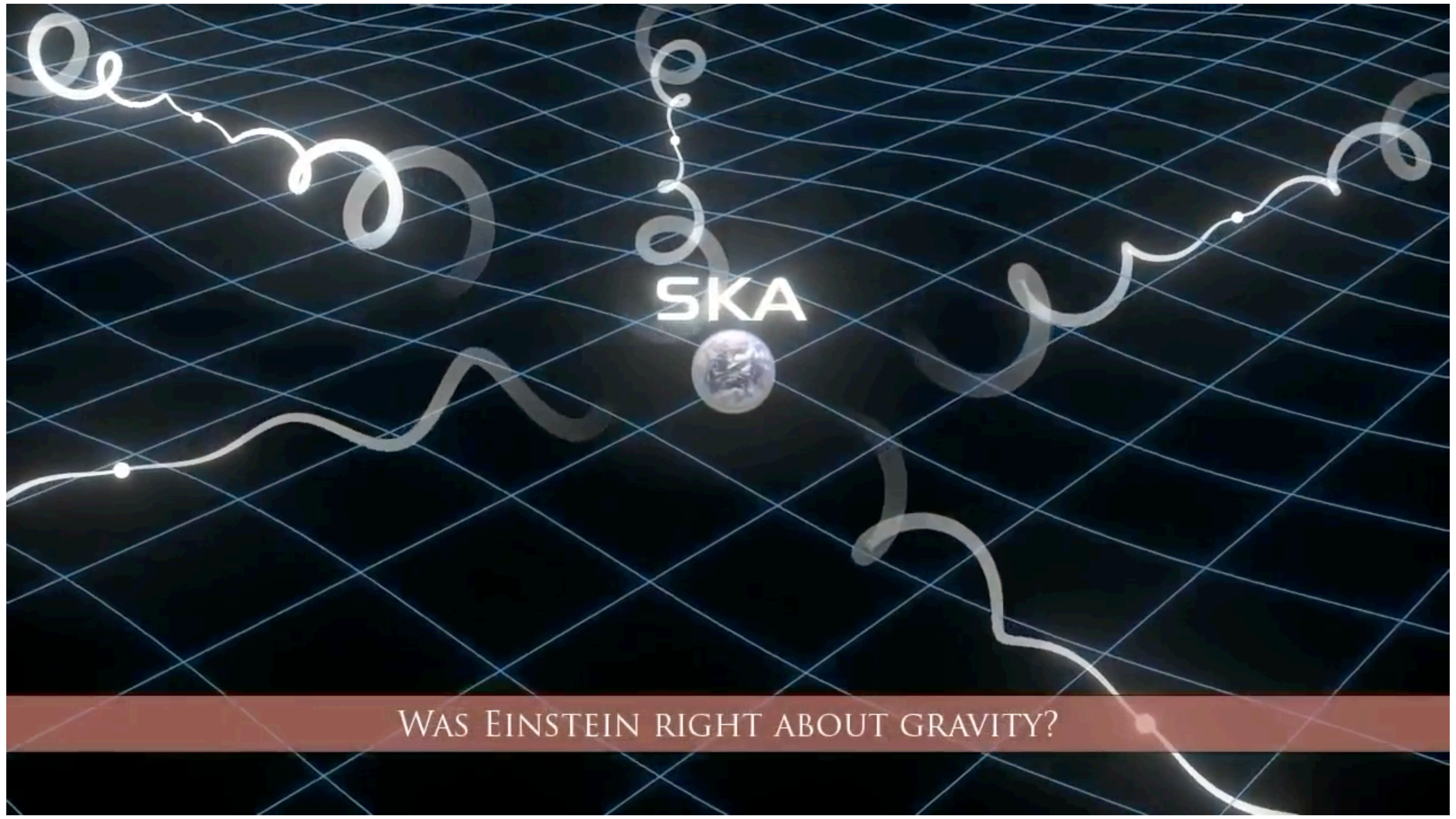

ACAT 2022
BARI

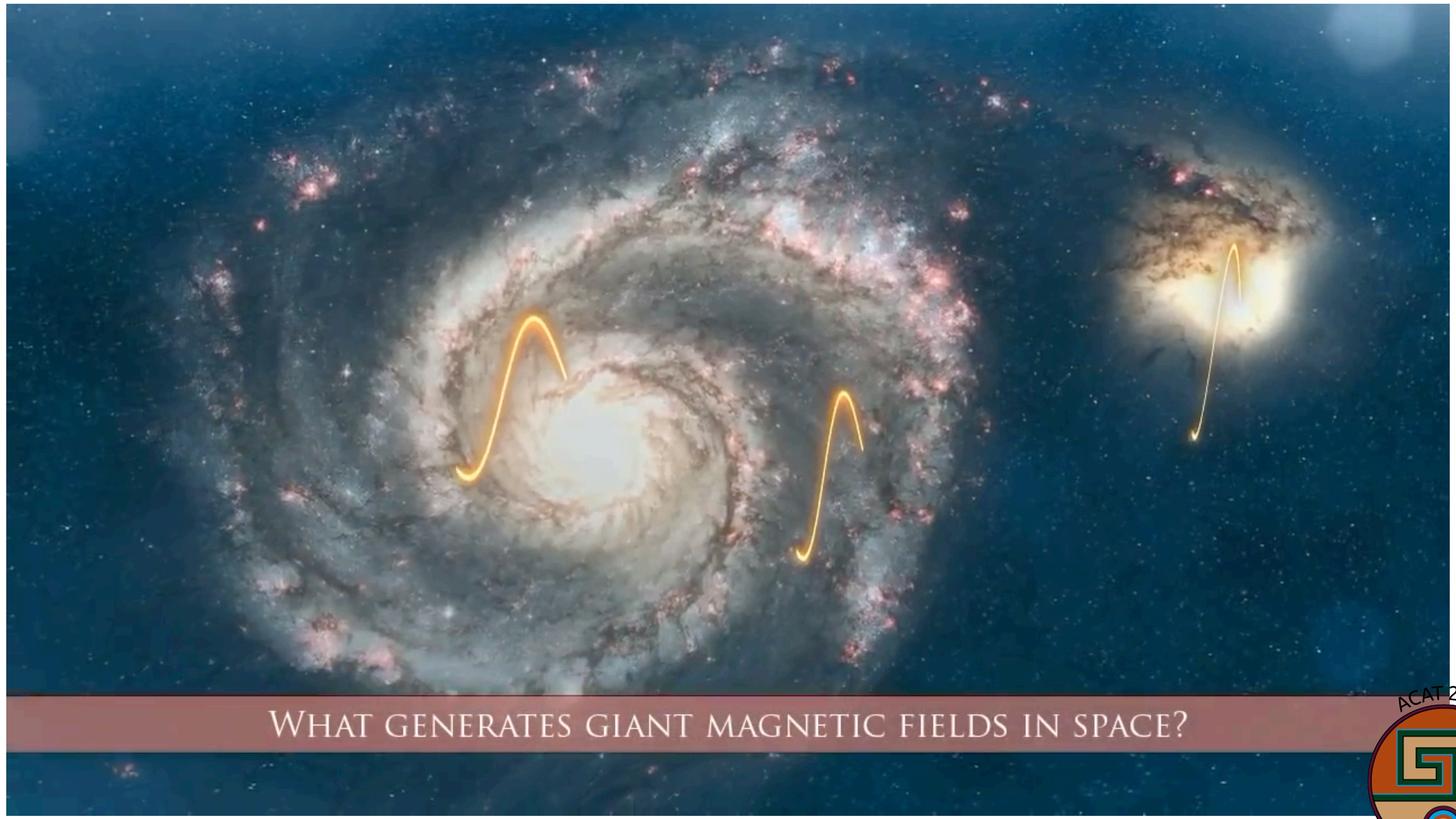

ACAT 2022
BARI

ACAT 2022
BARI

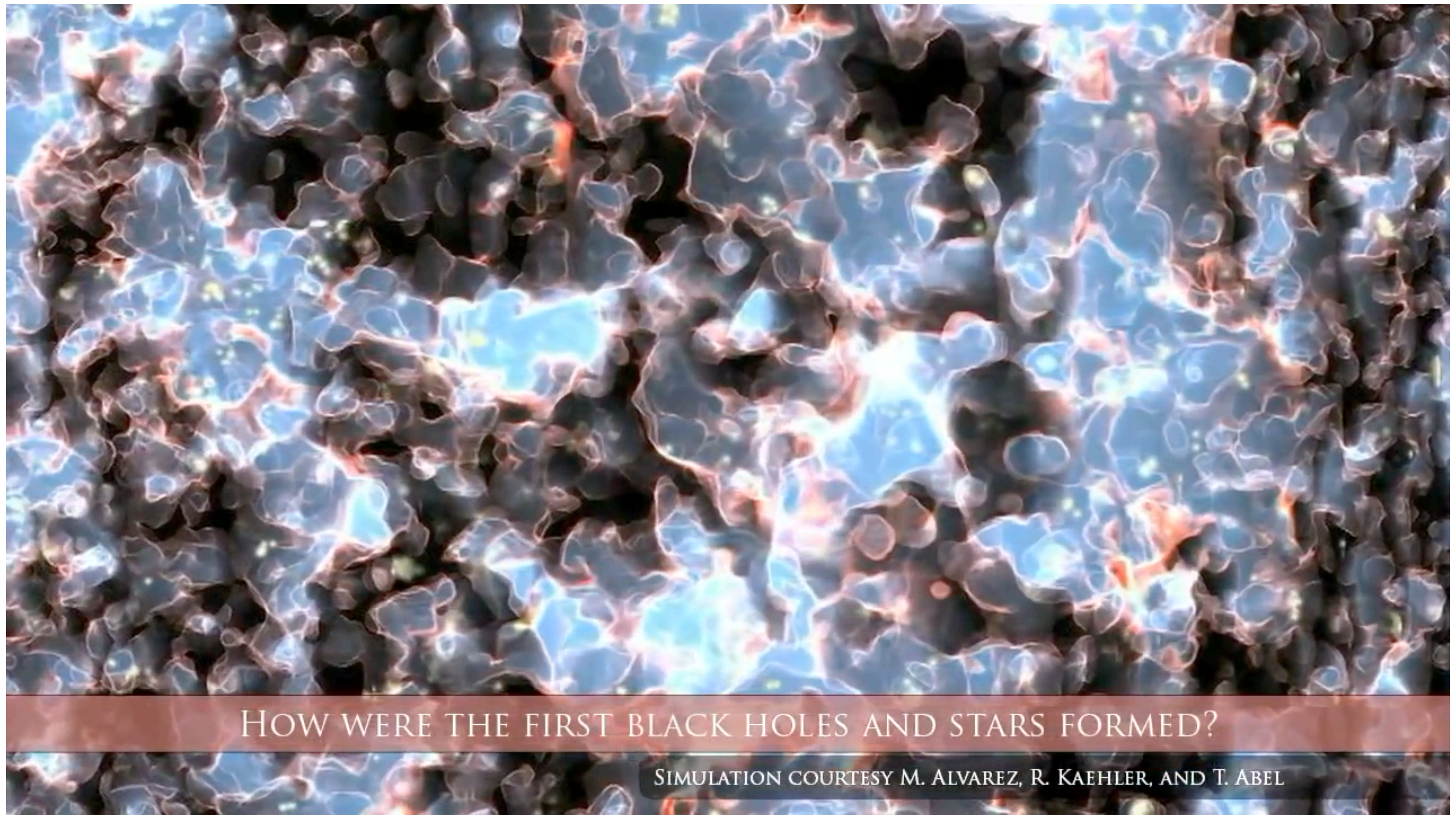

SOUTH AFRICA
AUSTRALIA

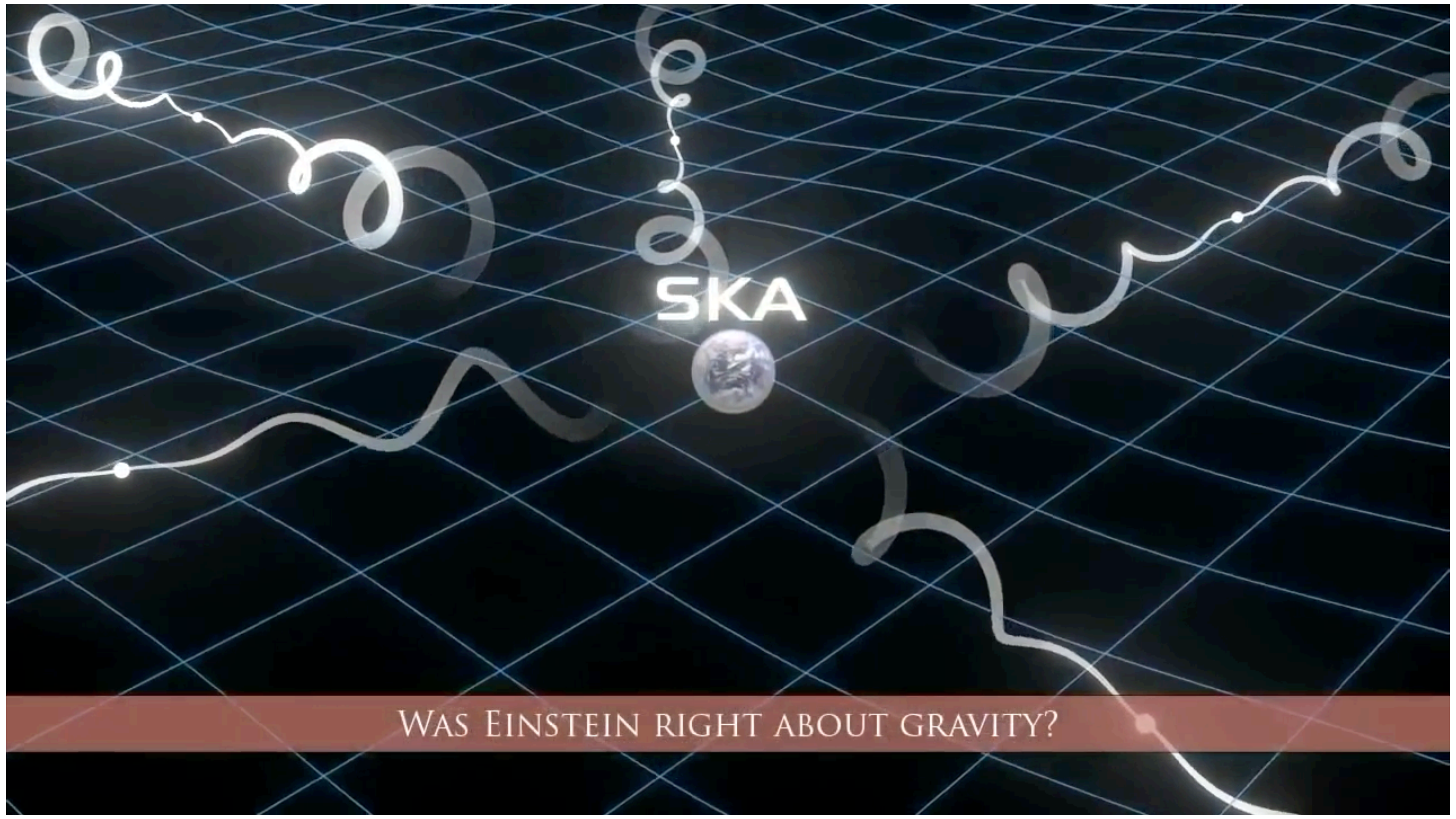




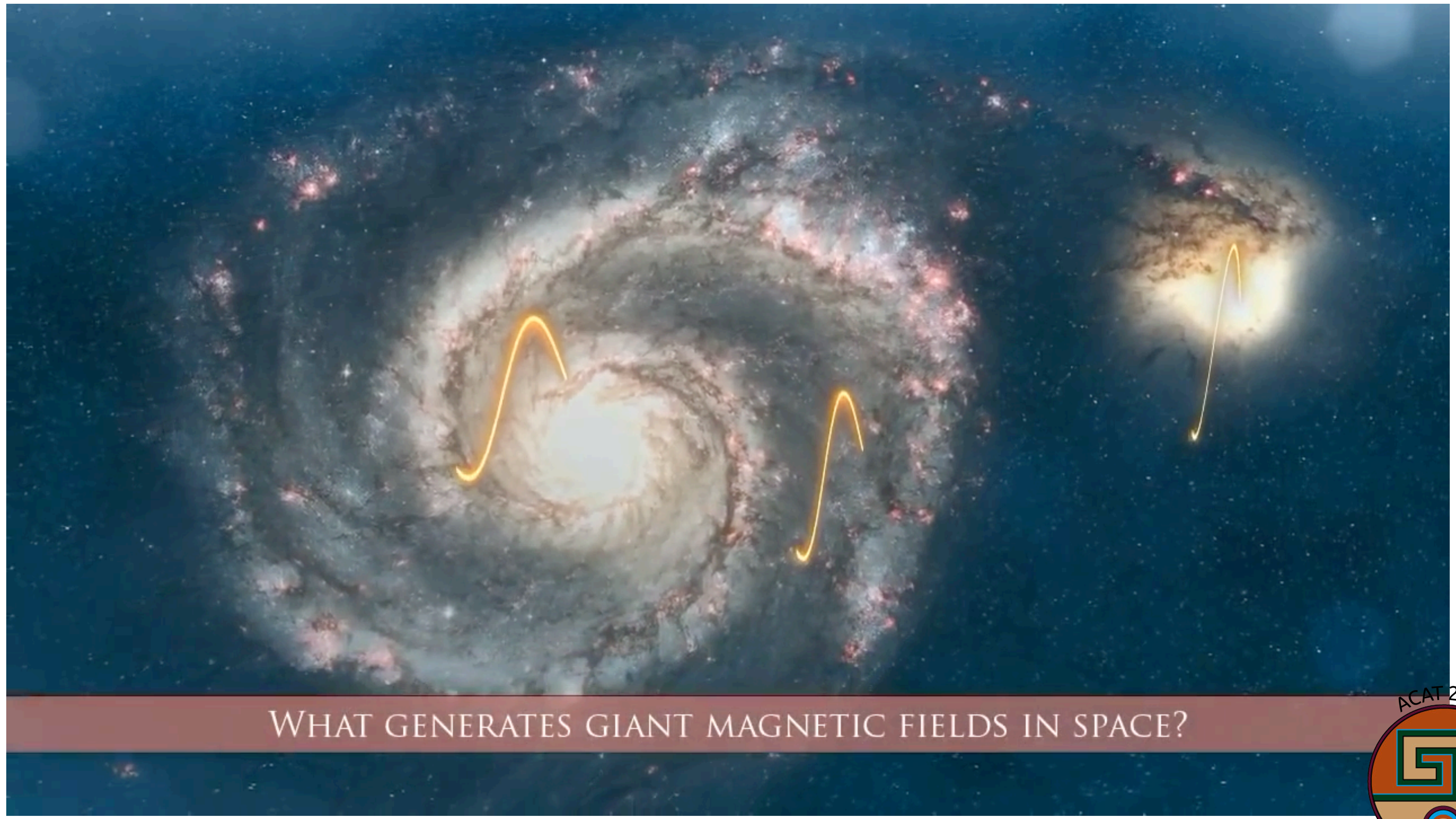


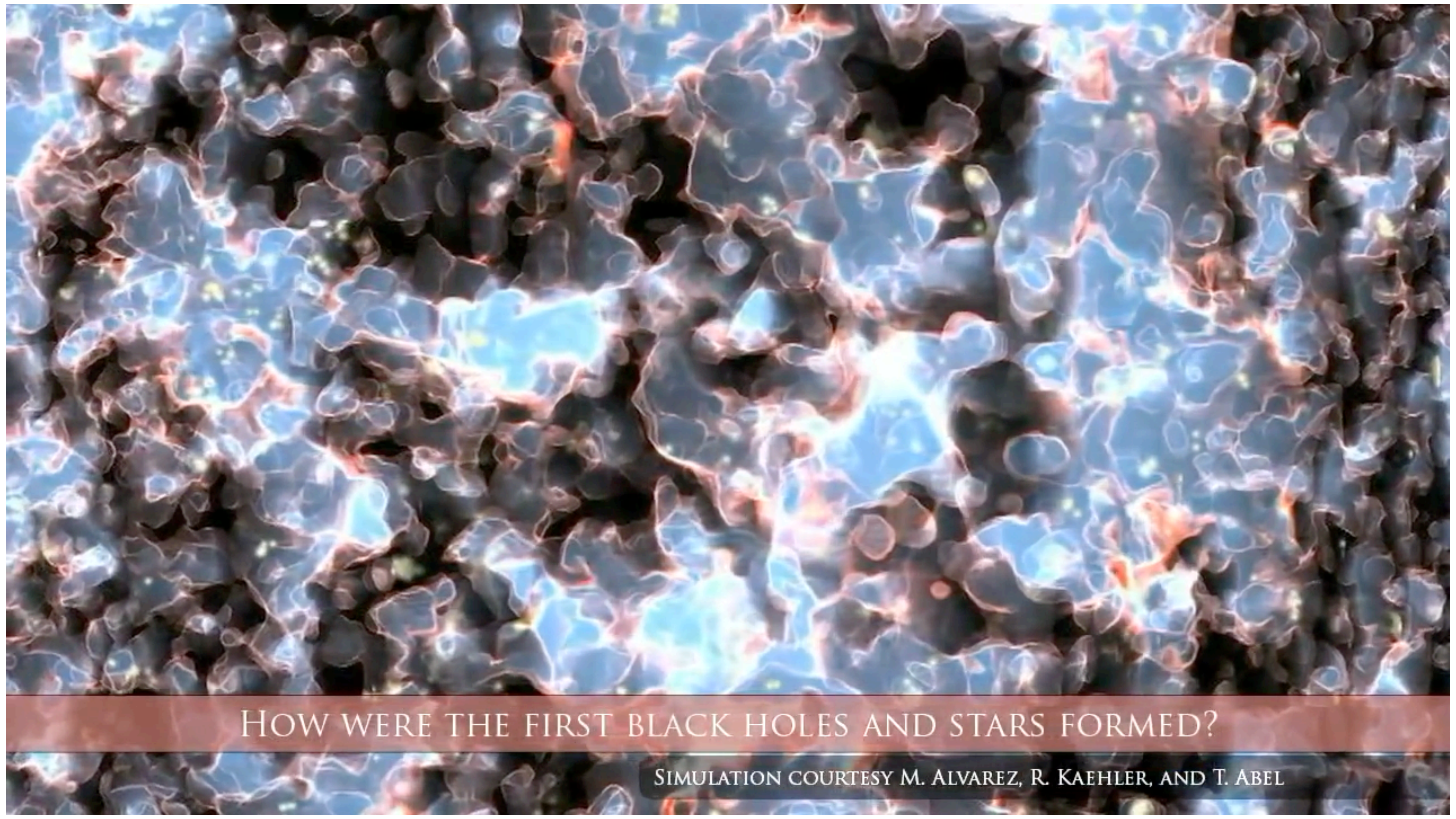
ACAT 2022 BARI

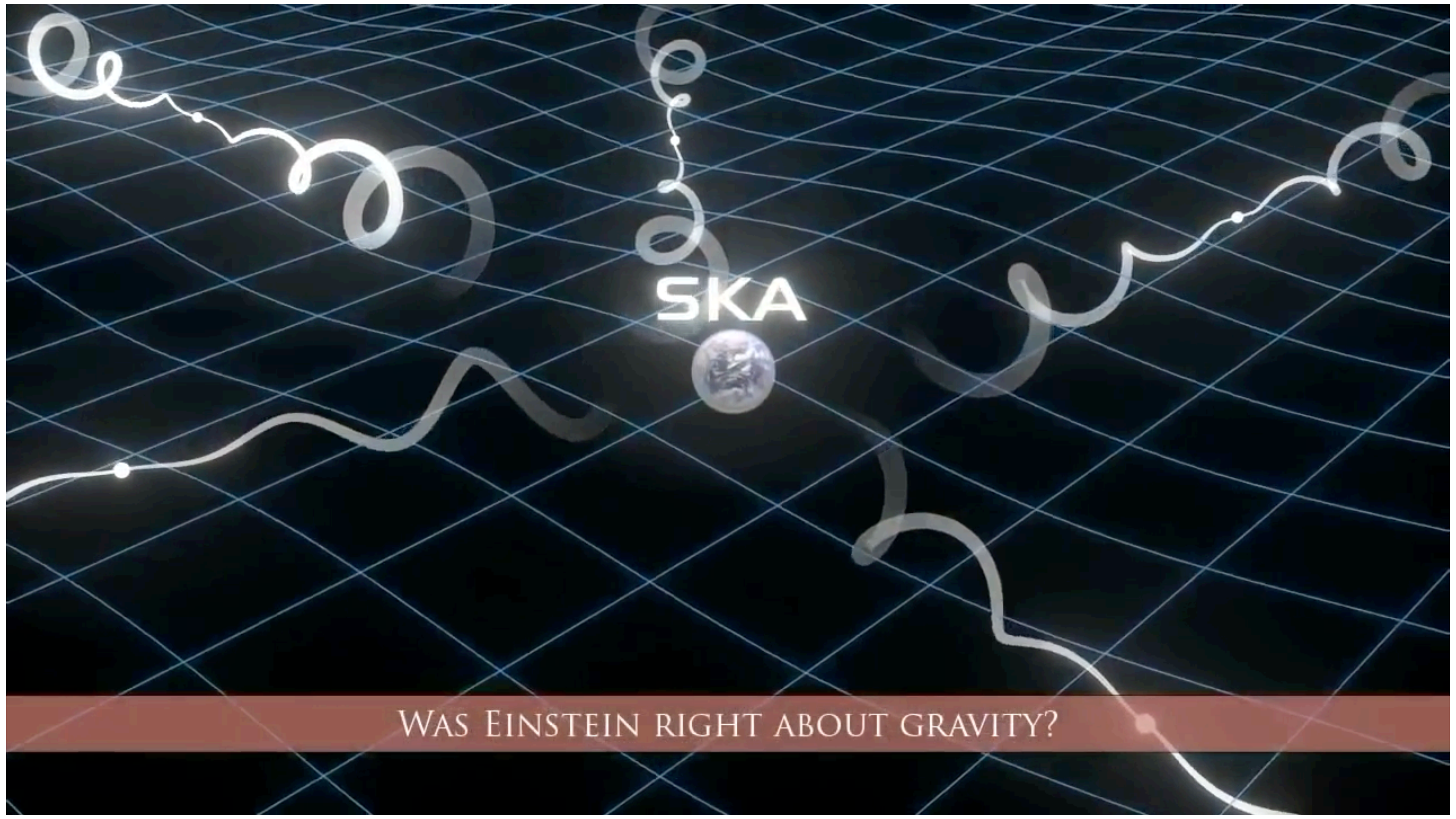

How were the first black holes and stars formed?
Simulation courtesy M. Alvarez, R. Kaehler, and T. Abel

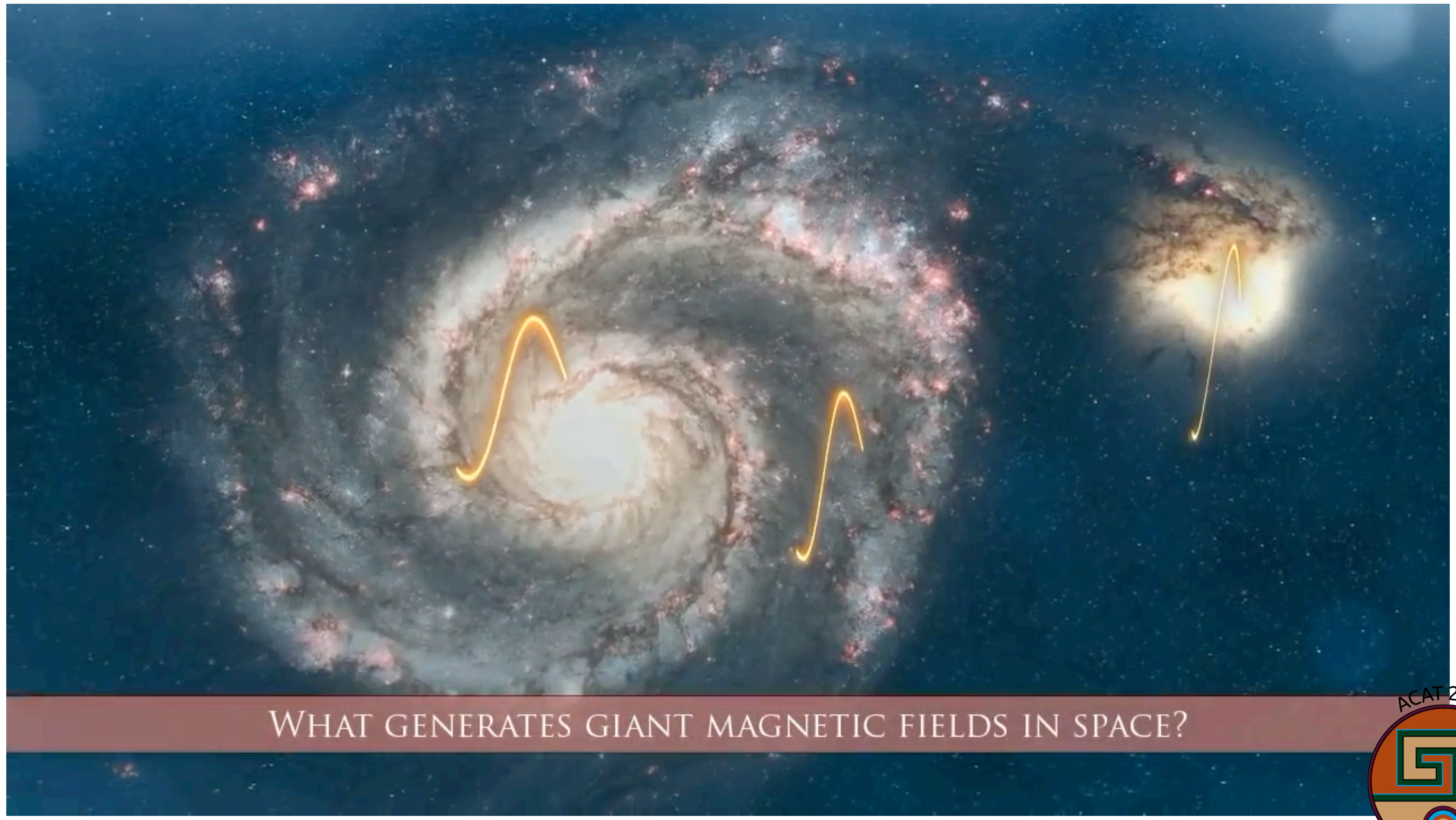

SKA
Was Einstein right about gravity?


Are we alone?

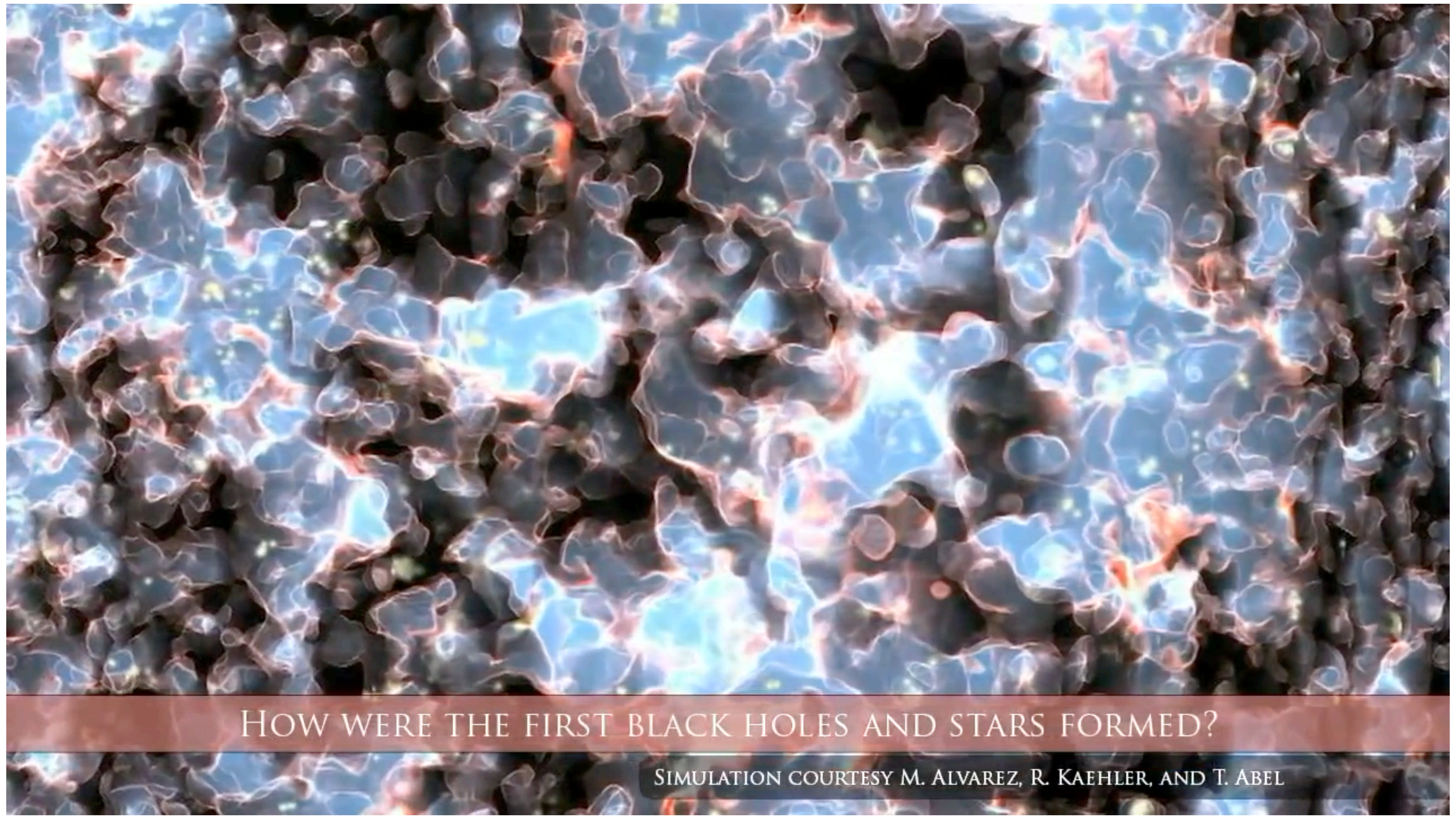

What generates giant magnetic fields in space?

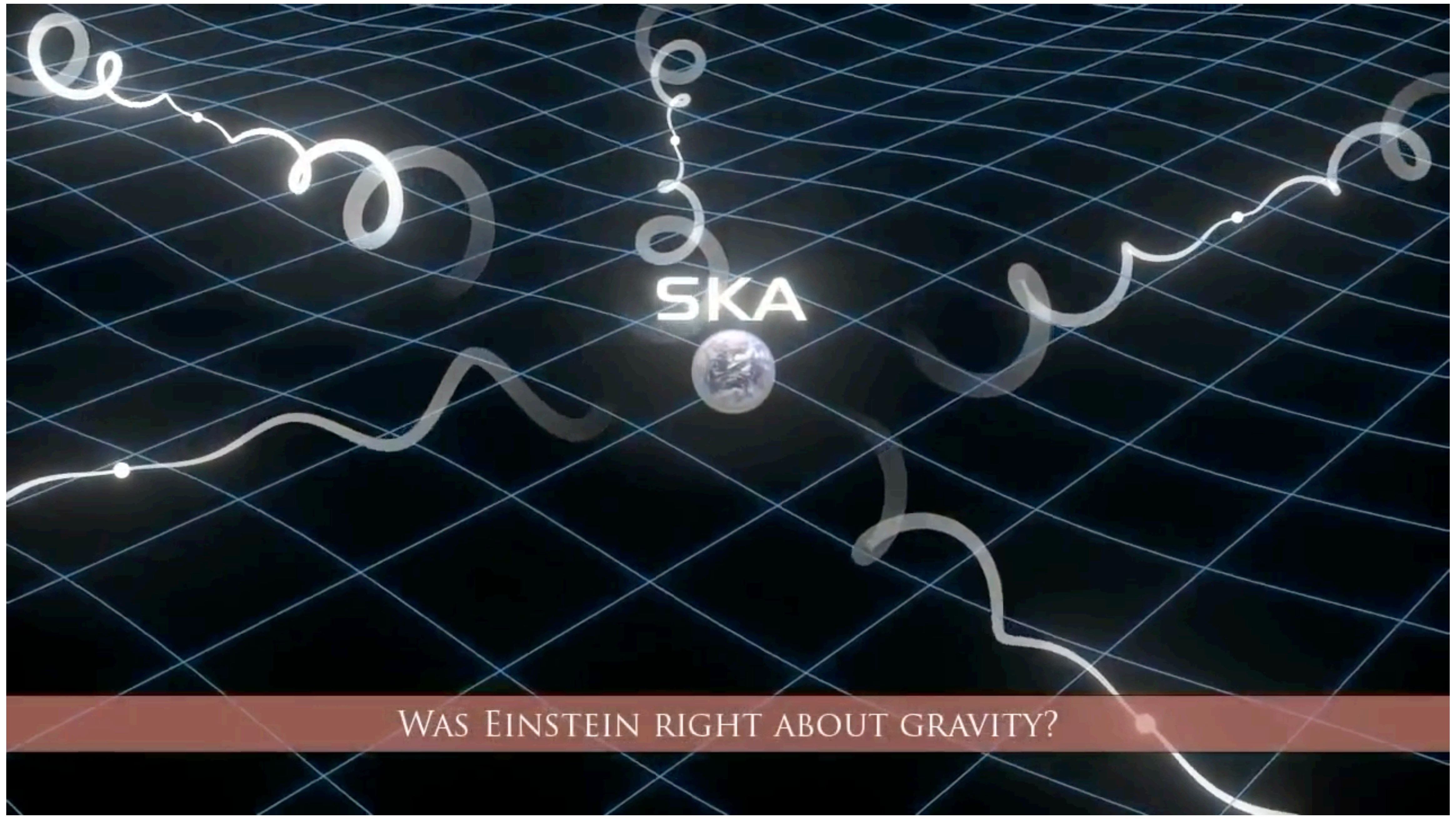




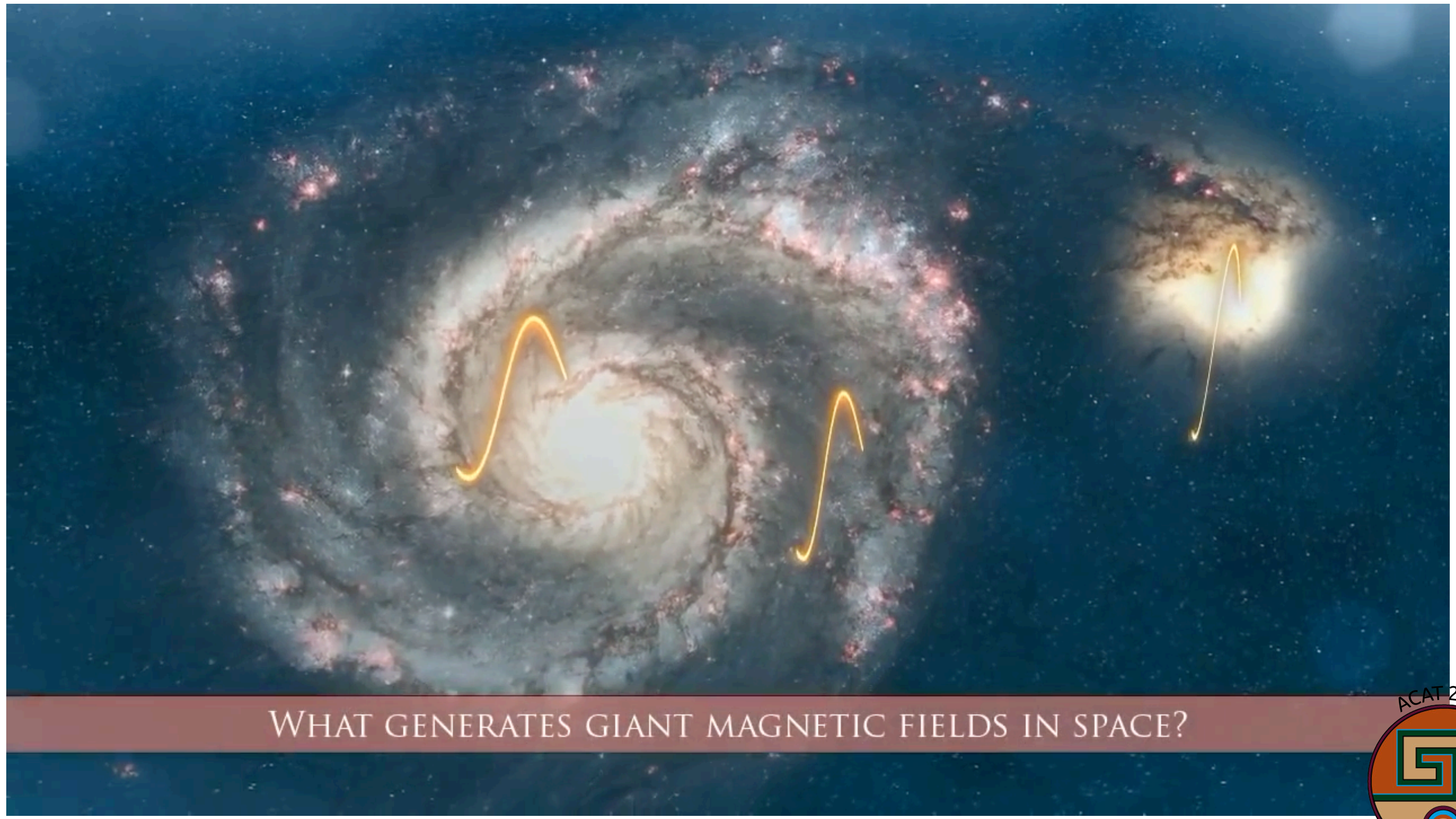


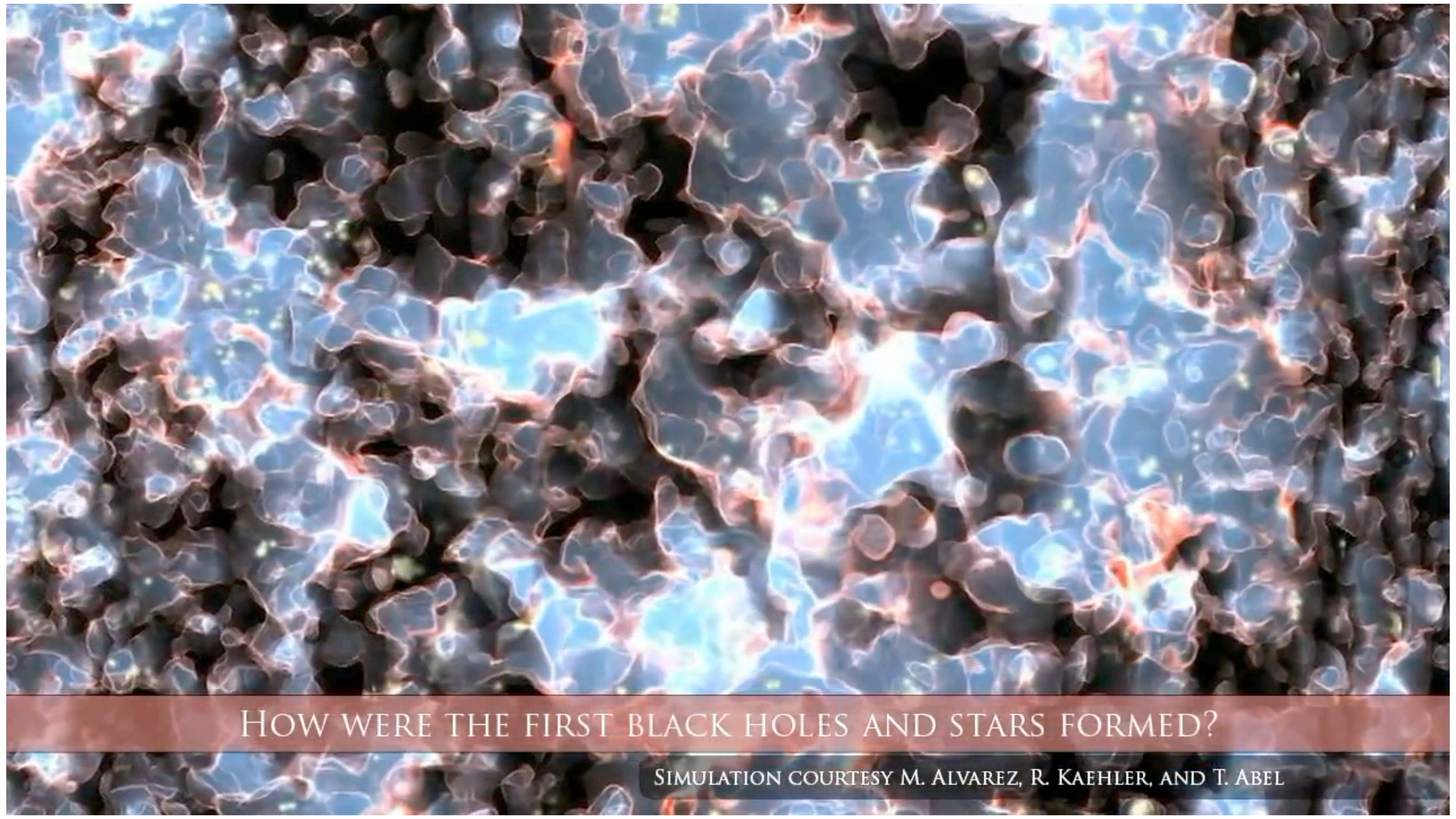
ACAT 2022
BARI

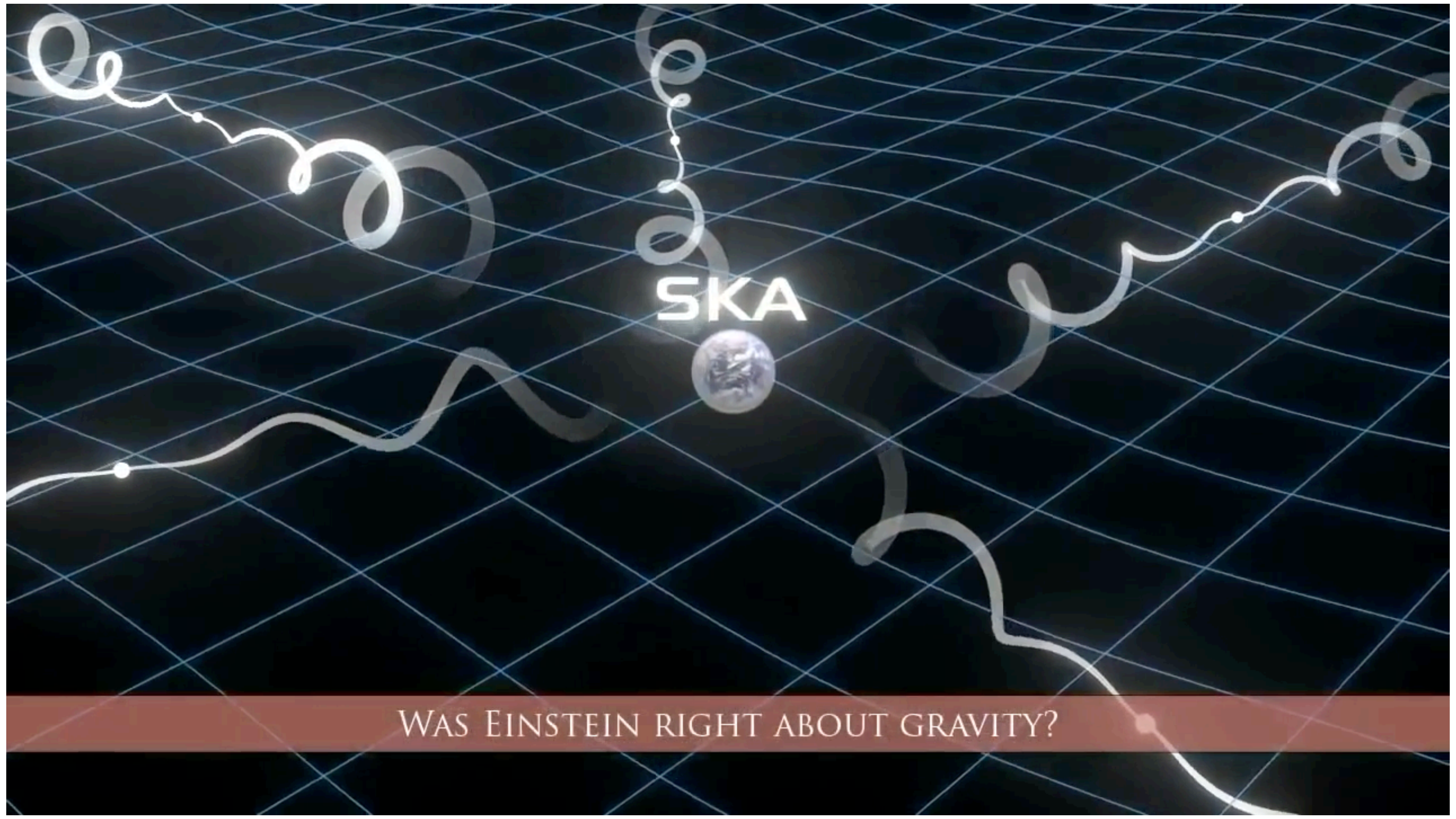

How were the first black holes and stars formed?
Simulation courtesy M. Alvarez, R. Kaehler, and T. Abel

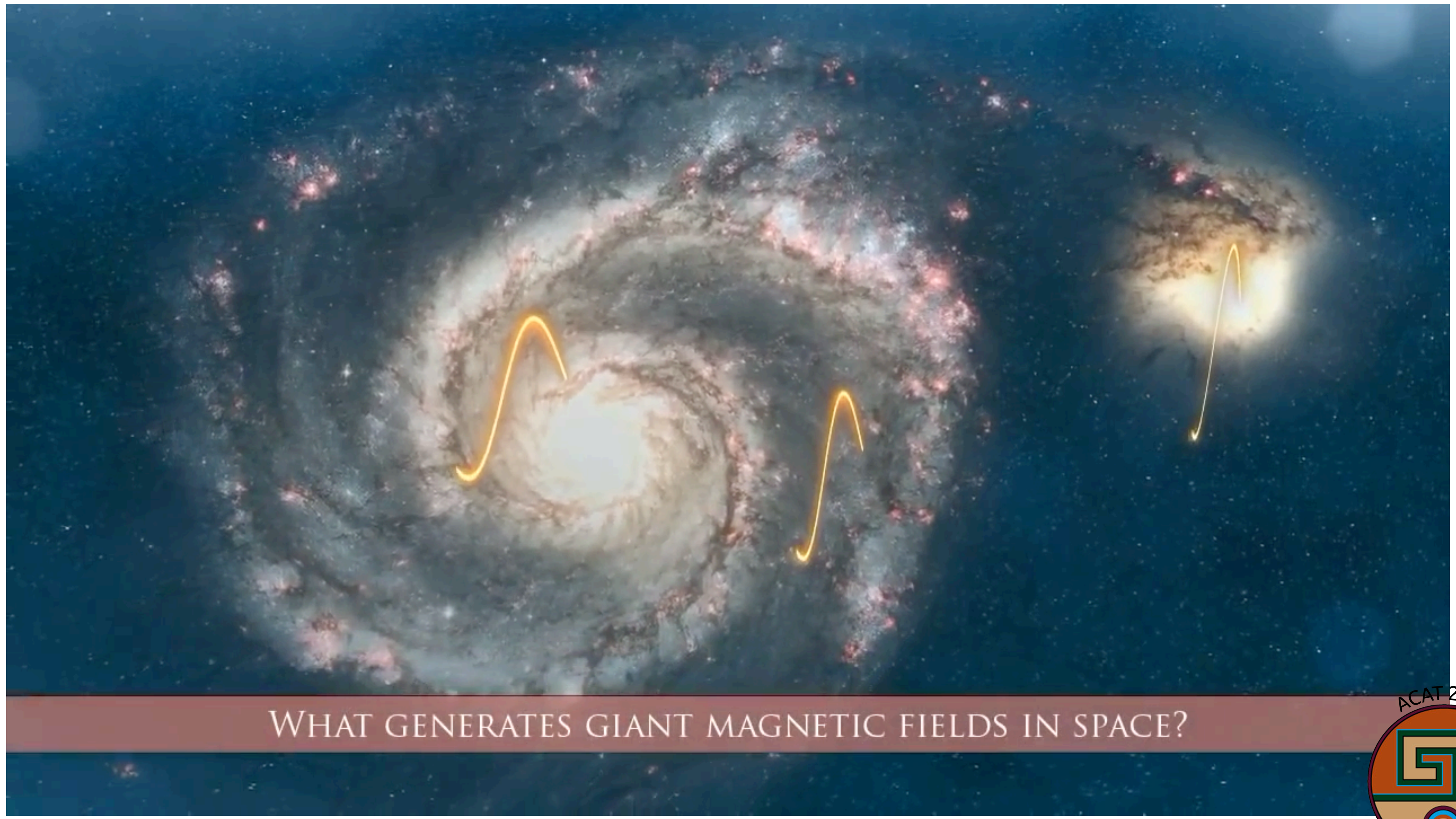

SKA
Was Einstein right about gravity?


Are we alone?

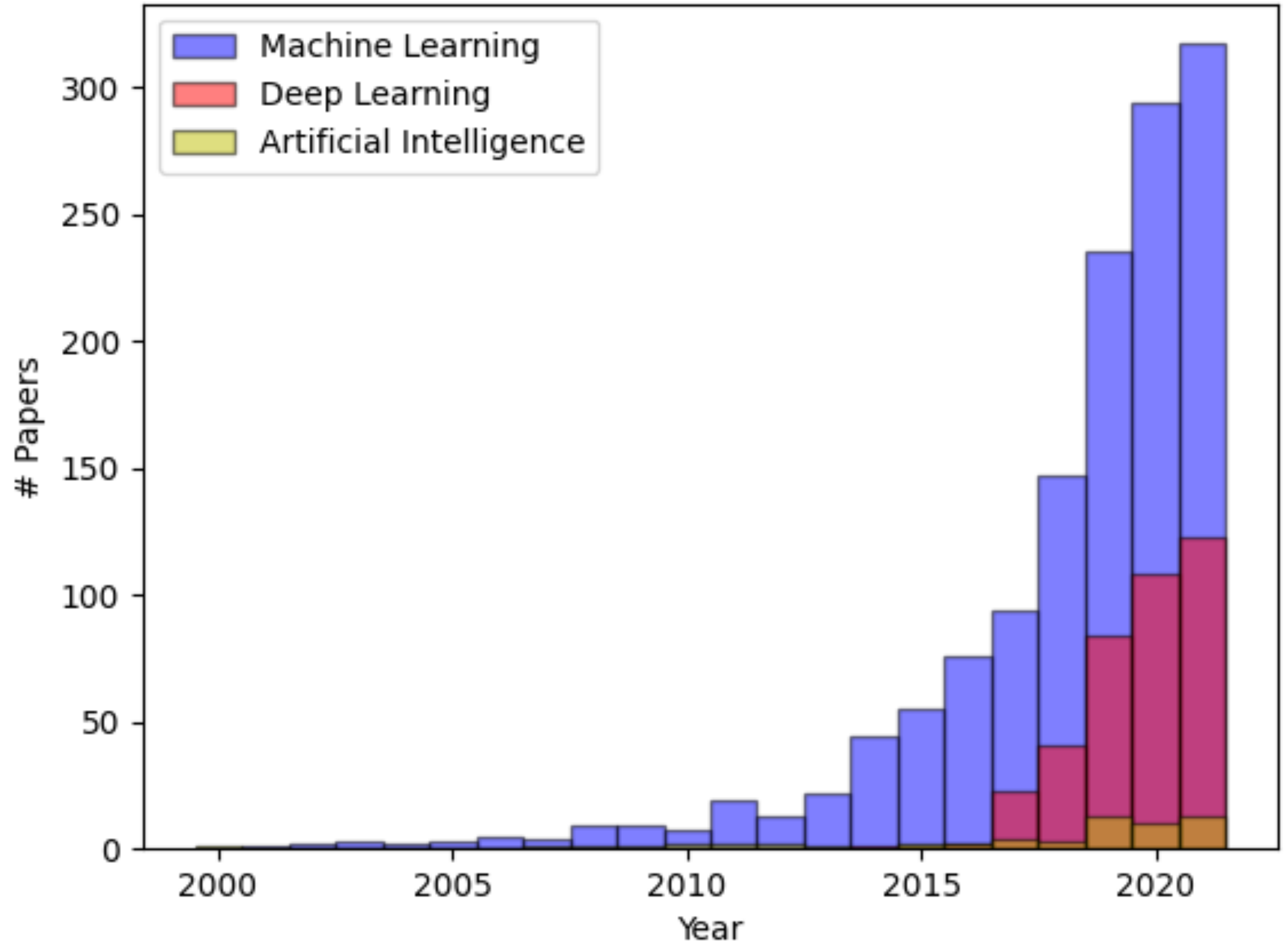

What generates giant magnetic fields in space?







ACAT 2022
BARI



Astronomy papers on the arxiv that include the keywords “machine learning”, “deep learning”, or “artificial intelligence” in the abstract or title.

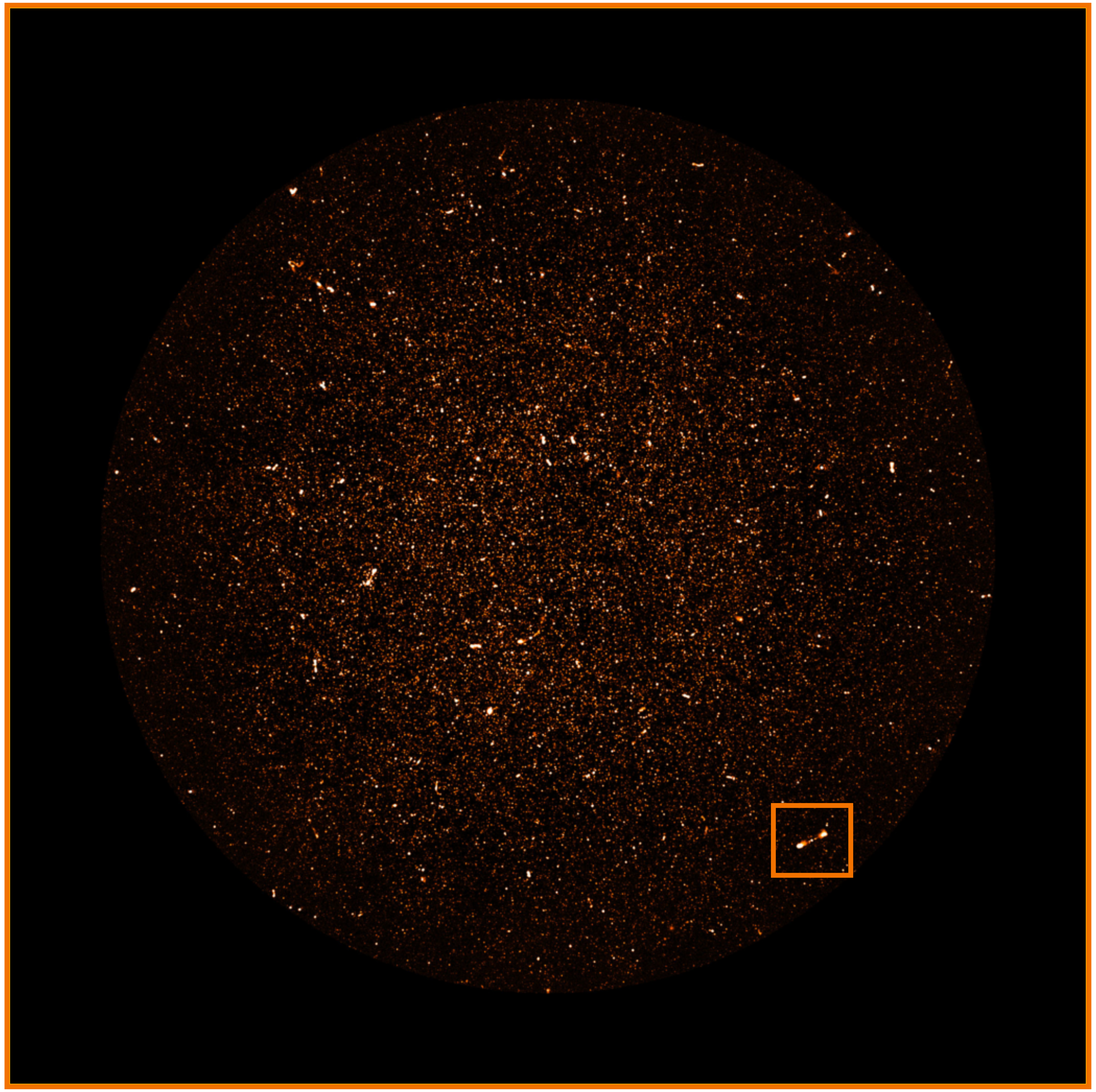
Scaife & Walmsley, in prep.

The SKA will be the world’s largest radio observatory

It is designed to answer some of the most important questions in modern astrophysics

It is a big data machine

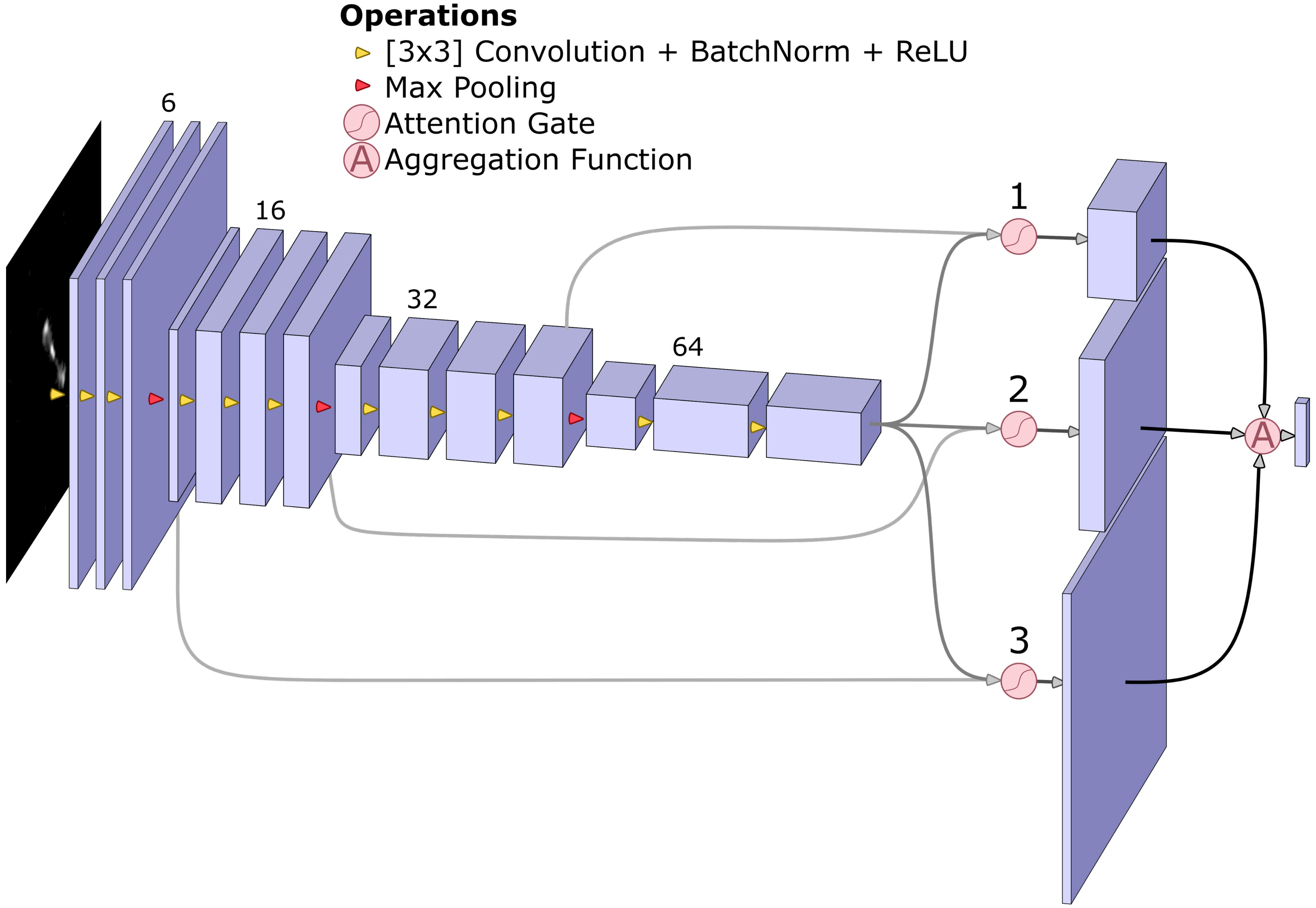
Mauch et al +, 2019, ApJ, arXiv:1912.06212



Much of radio astronomy is driven by population analyses

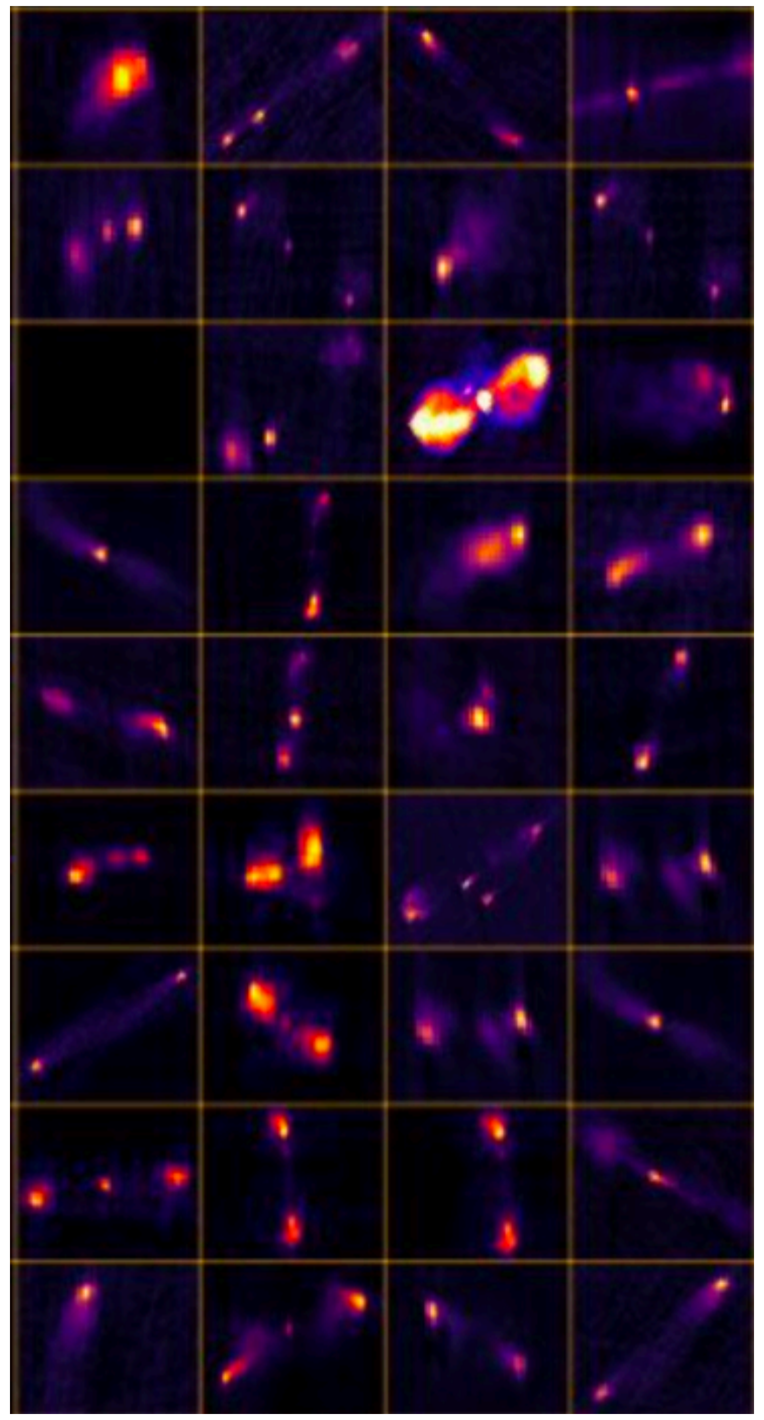
Populations need to be extracted from observational data

New discoveries need to be separated from known populations



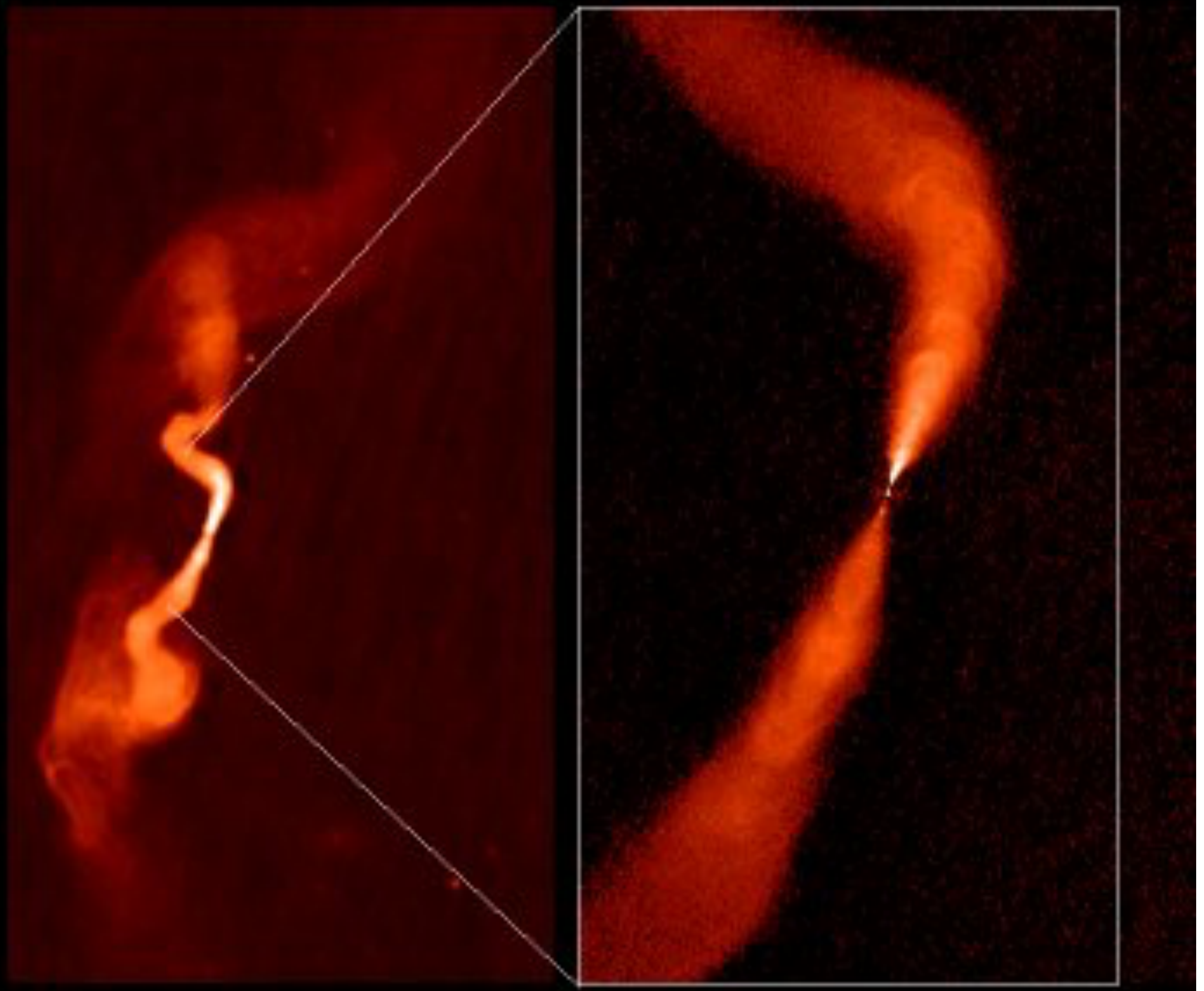
Bowles, AMS +, 2021, MNRAS, arXiv:2012.01248

Survey	Sources per Square Degree
NVSS (1998)	~50
FIRST (1995)	~90
LoTSS (2017)	~750
ASKAP (Australian SKA Pathfinder)	~2900*

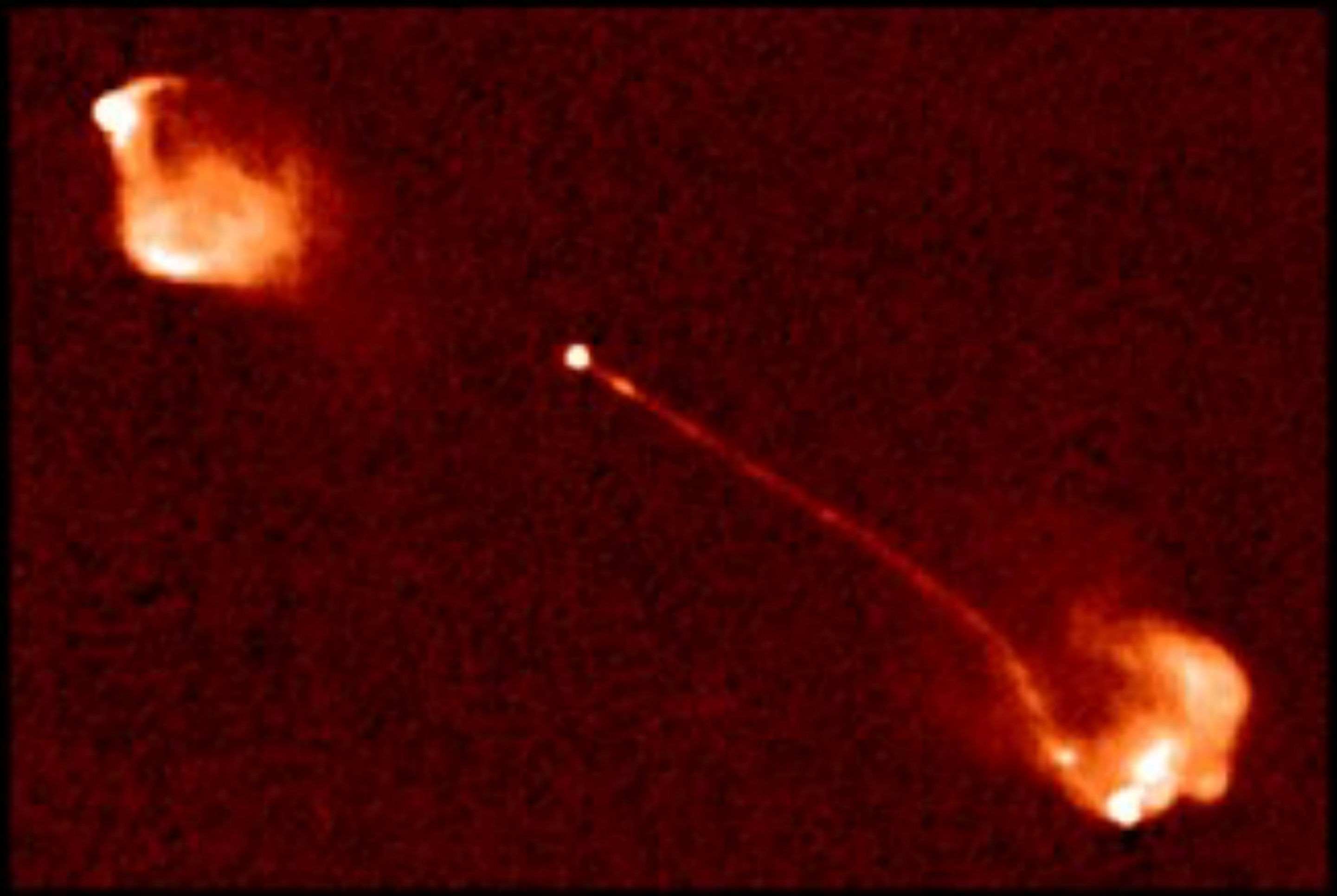


Experts: ~1 min per source (125,000 sources / yr of full time work)
Radio Galaxy Zoo: 300,000 sources 12,000 users over 5.5 years
Machine Learning: 100 million sources in ~15 min

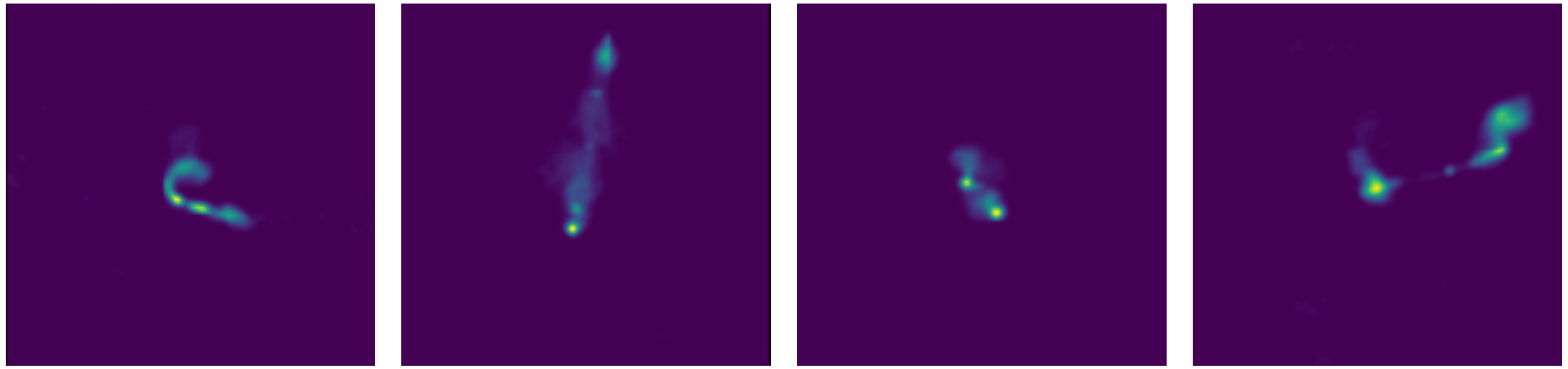
*estimated using their goal of 60 million extragalactic synchrotron sources.
Mol, J David 2011 (LOFAR Beam former); https://ned.ipac.caltech.edu/level5/March01/Andernach/Ander3.html; W. Williams Oct. 2019 Colloquium Slide 28; Johnston, S., Taylor, R., Bailes, M., et al. (2008); Image credit: NRAO/AUI/NSF



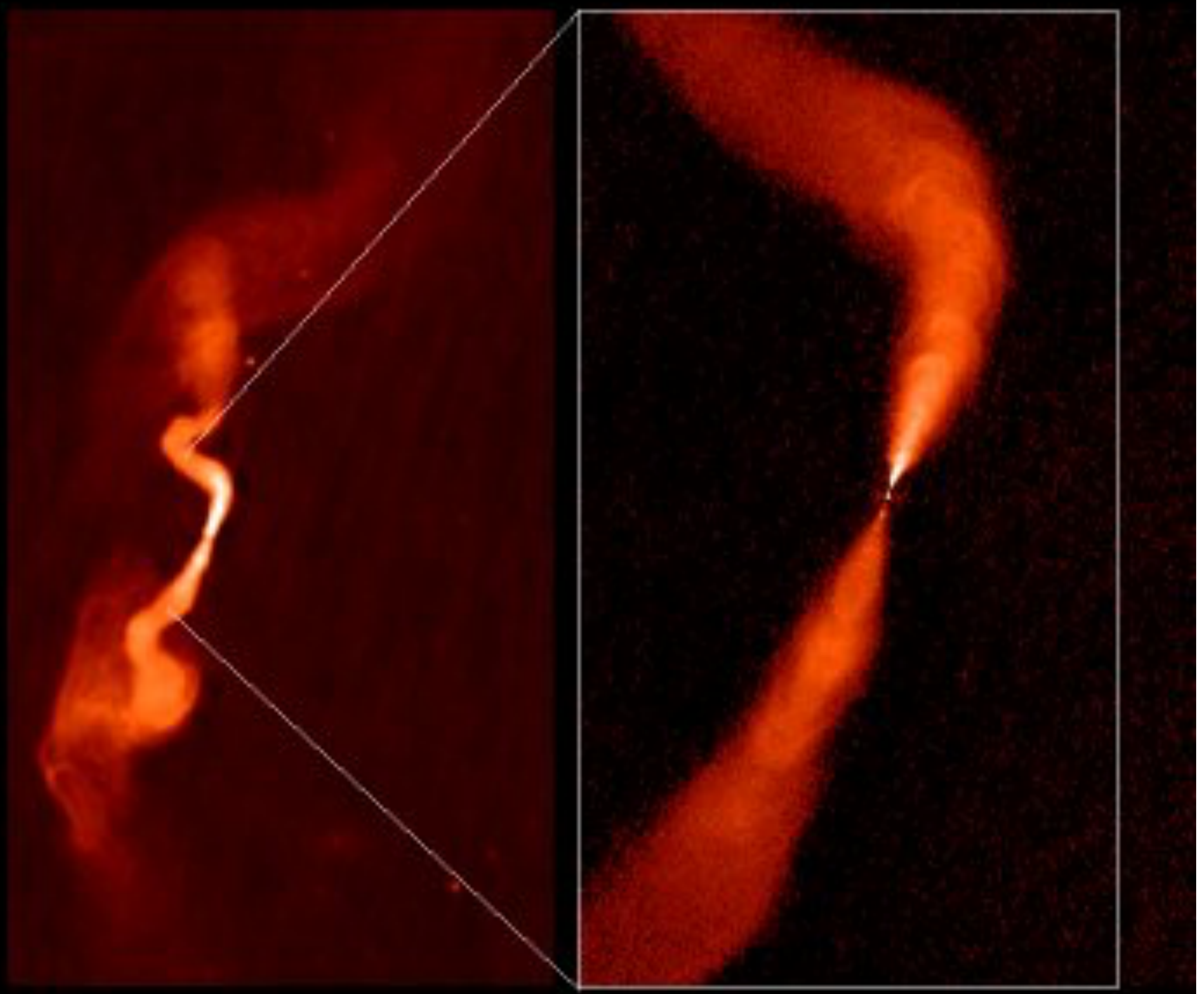
FR Class I source: radio galaxy 3C31



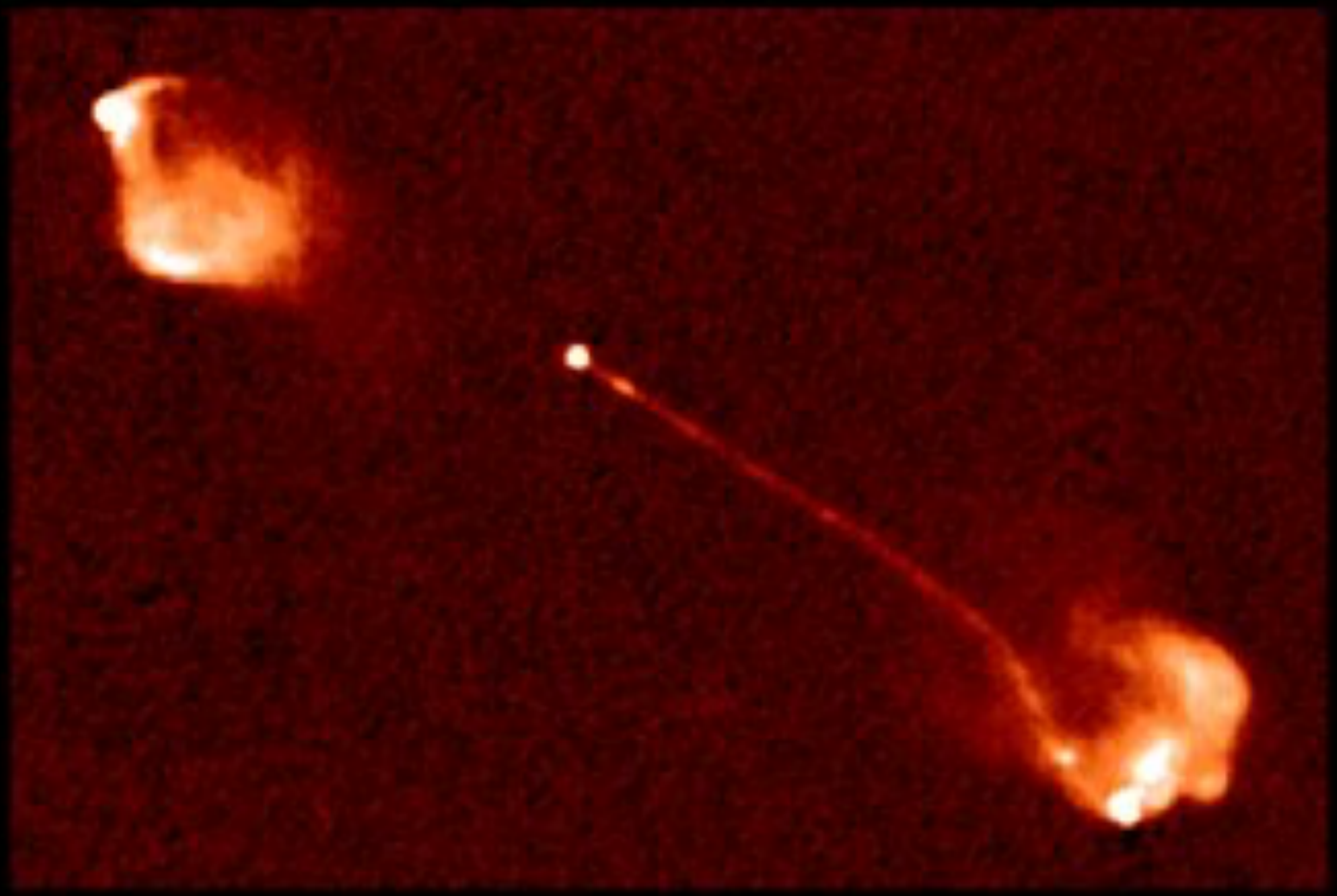
FR Class II source: quasar 3C175



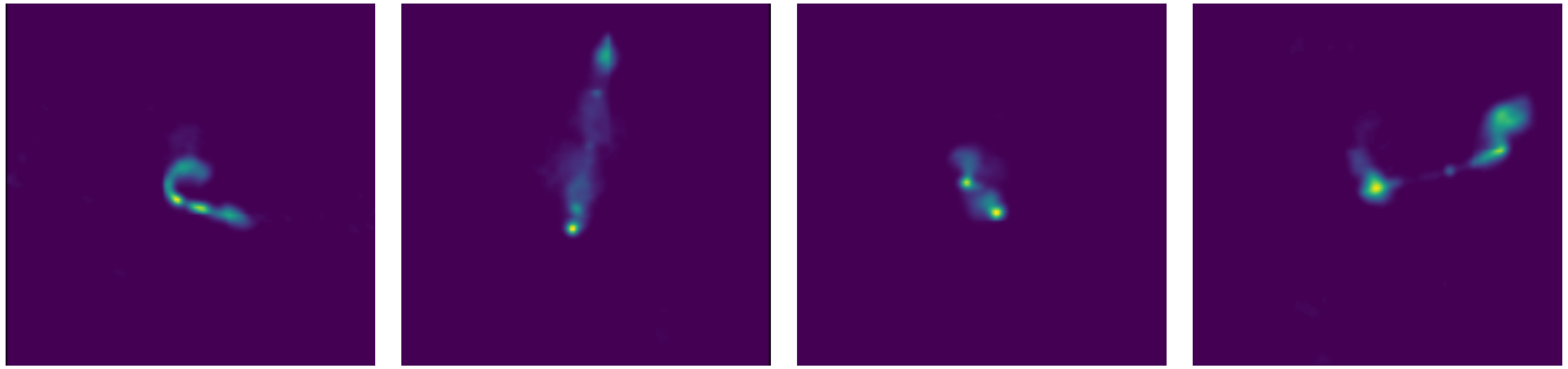
- Large archival databases, but only small *labelled* datasets
- Significant and variable *class imbalances*
- Need for carefully *calibrated uncertainties* on model outputs
- Need for *biases* in model outputs to be quantitatively estimated



FR Class I source: radio galaxy 3C31



FR Class II source: quasar 3C175



- **Large archival databases, but only small *labelled* datasets**
- Significant and variable *class imbalances*
- Need for carefully *calibrated uncertainties* on model outputs
- Need for *biases* in model outputs to be quantitatively estimated

- **Large archival databases, but only small *labelled* datasets …**

- **Large archival databases, but only small *labelled* datasets …**

Approach 1: Get more labels.

- Label more data using experts –> *expensive: probably why you're in this situation in the first place…*
- Ask for help from citizen scientists –> *provides non-expert labels; requires higher consensus*

- **Large archival databases, but only small *labelled* datasets …**

Approach 1: Get more labels.

- Label more data using experts –> *expensive: probably why you're in this situation in the first place…*
- Ask for help from citizen scientists –> *provides non-expert labels; requires higher consensus*

Approach 2: Make better use of unlabelled data.

- Generative Adversarial Networks –> *stability issues; biases*
- Semi-supervised learning
- Self-supervised learning

} *domain / dataset shift issues*

- …

ACAT 2022
BARI

- **Large archival databases, but only small *labelled* datasets ...**

Approach 1: Get more labels.

- Label more data using experts –> *expensive: probably why you're in this situation in the first place…*
- Ask for help from citizen scientists –> *provides non-expert labels; requires higher consensus*

Approach 2: Make better use of unlabelled data.

- Generative Adversarial Networks –> *stability issues; biases*
- Semi-supervised learning
- Self-supervised learning

domain / dataset shift issues

- …

Approach 3: Change the labels.

ACAT 2022

BARI

- **Large archival databases, but only small *labelled* datasets …**

Approach 1: Get more labels.

- Label more data using experts –> *expensive: probably why you're in this situation in the first place…*
- Ask for help from citizen scientists –> *provides non-expert labels; requires higher consensus*

Approach 2: Make better use of unlabelled data.

- Generative Adversarial Networks –> *stability issues; biases*
- Semi-supervised learning
- Self-supervised learning

domain / dataset shift issues

- …

Approach 3: Change the labels.

The challenge: find 10 *plain English* semantic tags that can be used to label radio galaxies in a way that allows us to separate scientific classes.

- **Large archival databases, but only small *labelled* datasets …**

Approach 1: Get more labels.

- Label more data using experts —> *expensive; probably why you're in this situation in the first place…*
- Ask for help f … *igher consensus*

Approach 2: Ma

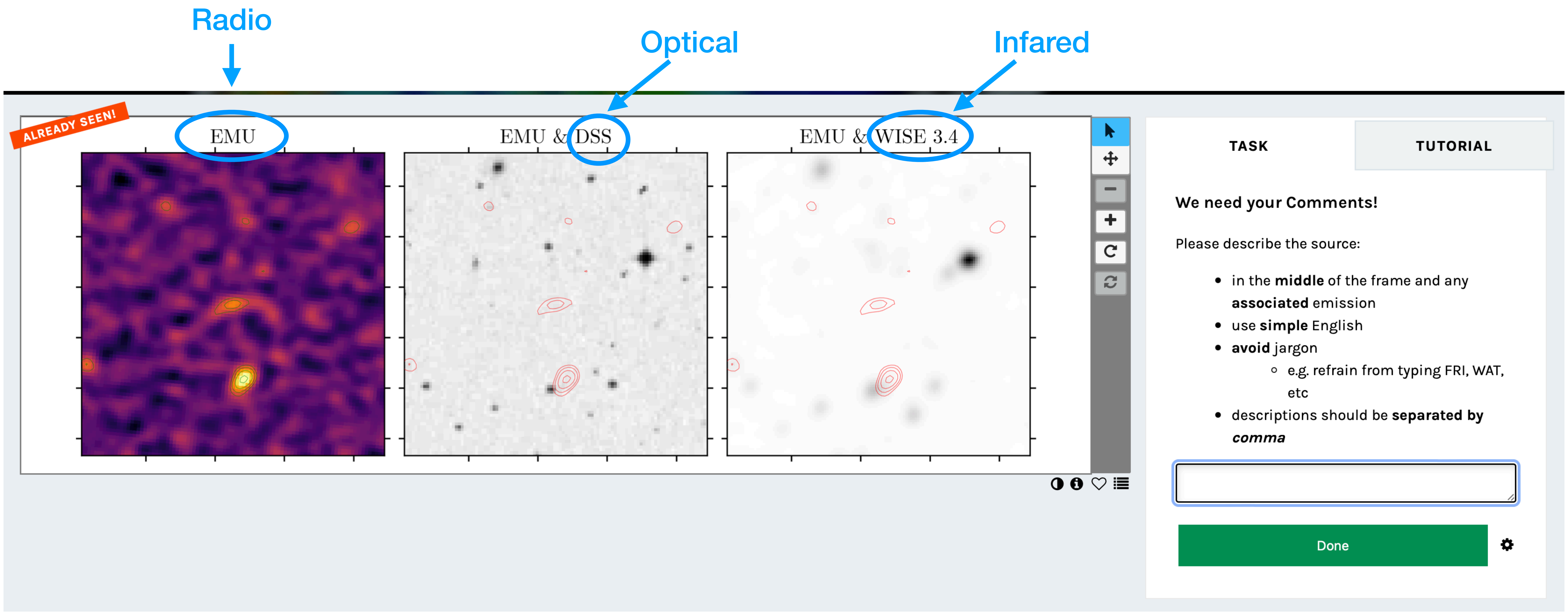
- Generative Ac
- Semi-supervis
- Self-supervise
- …

Work led by **Micah Bowles**

Machine Learning and the Physical Sciences @ 36th Conference on Neural Information Processing Systems (NeurIPS 2022)

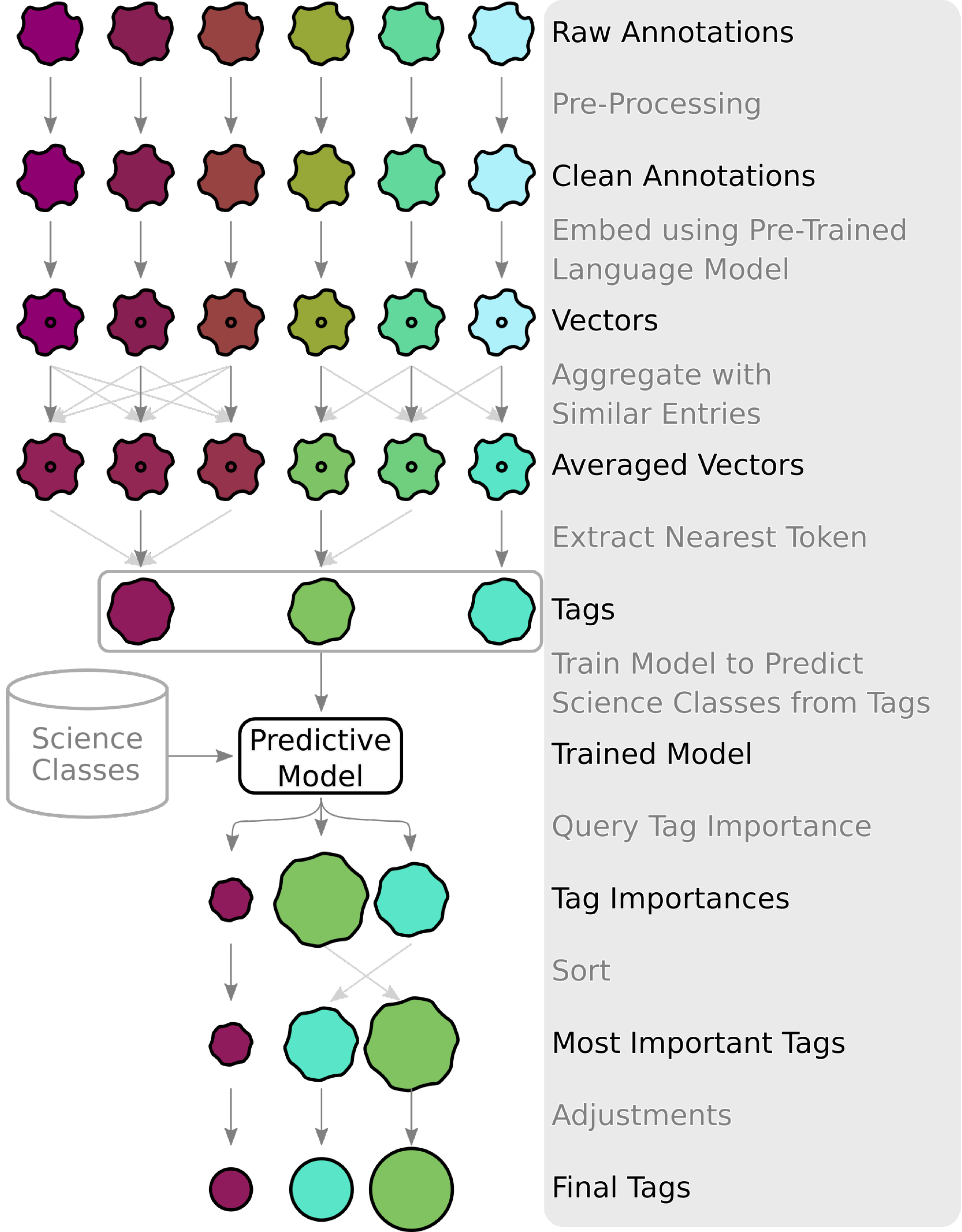
Approach 3: Change the labels.

The challenge: find 10 *plain English* semantic tags that can be used to label radio galaxies in a way that allows us to separate scientific classes.



- Users were asked to provide plain English annotations for a set of ~300 radio galaxies;
- Experts were asked to label the same galaxies using a set of 22 astrophysical classifications.

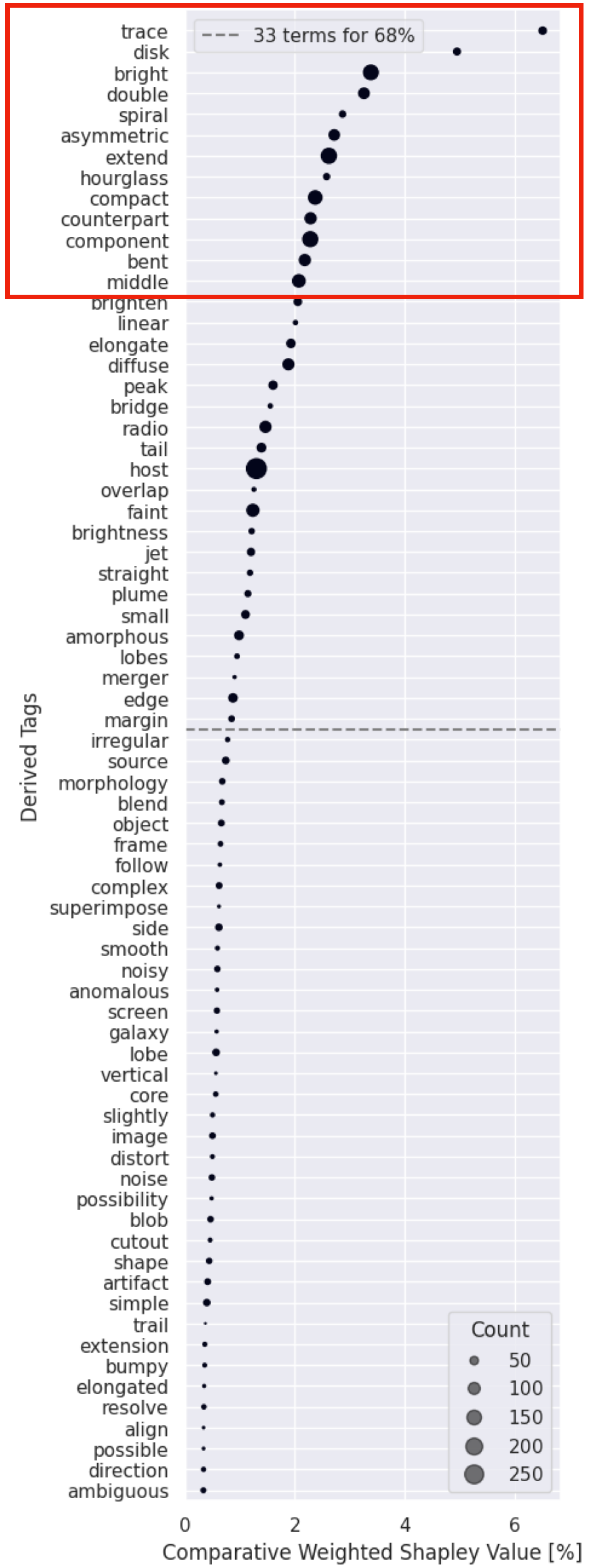
Bowles et al. 2022, accepted NIPS 2022; submitted MNRAS

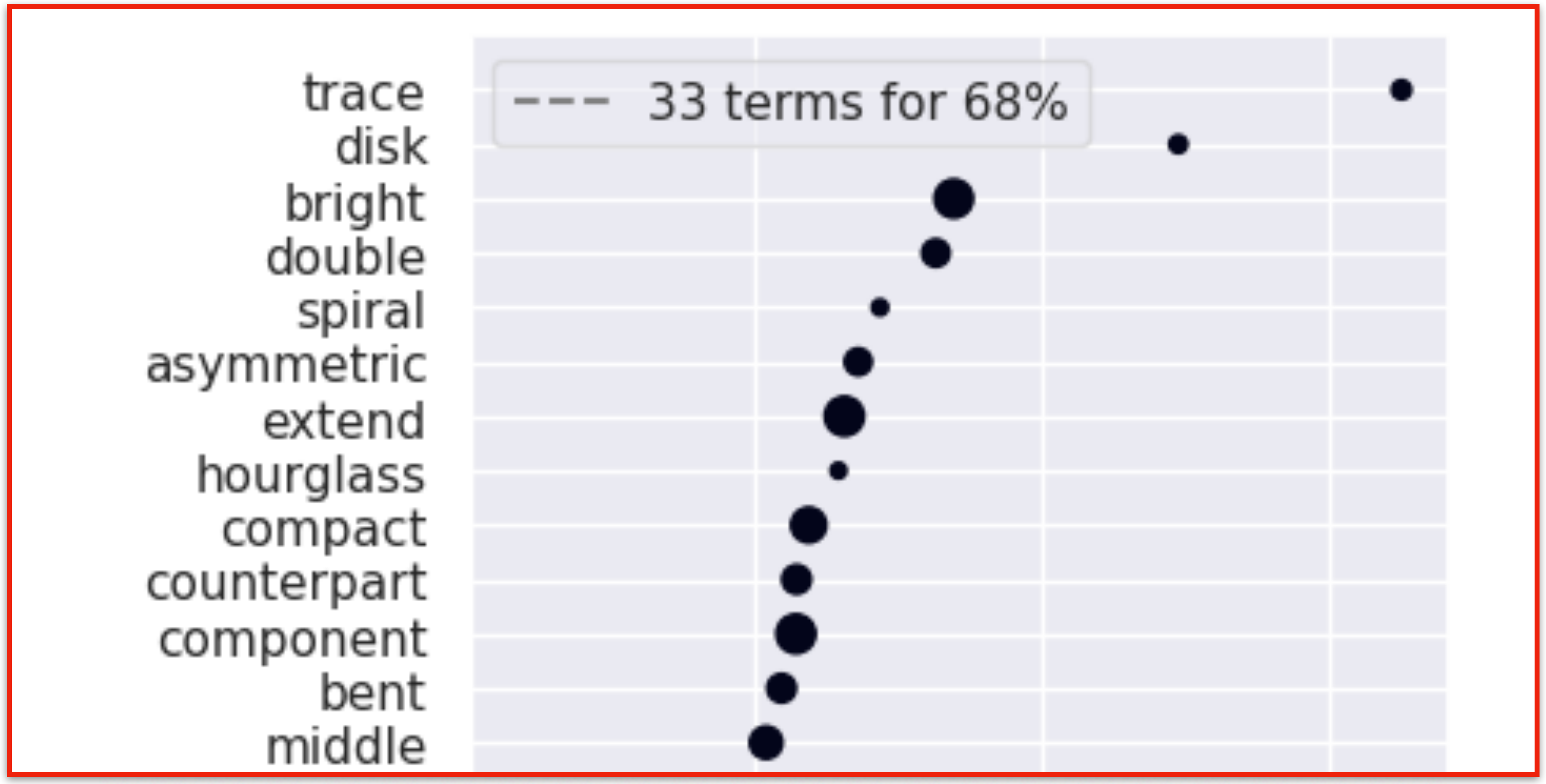


Aggregate similar annotations to create “tags”

Identify most important tags to form a taxonomy

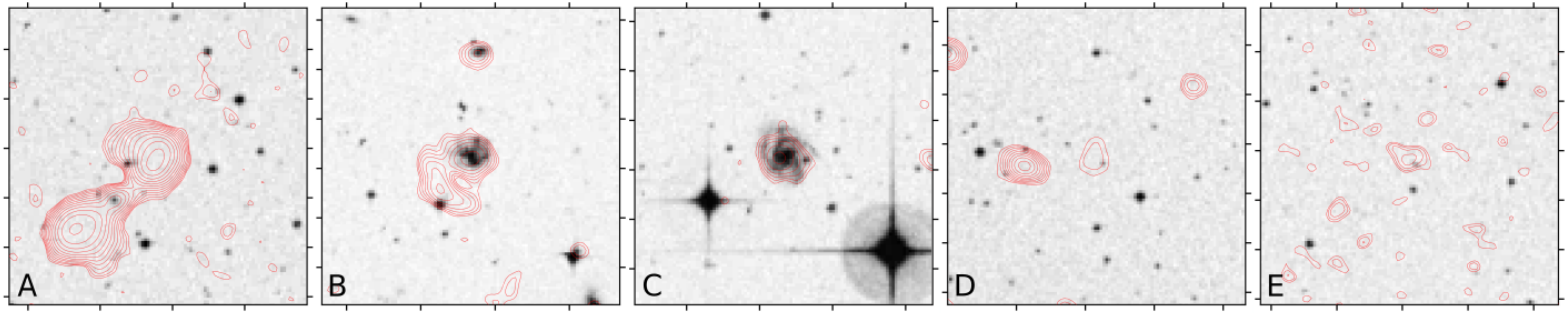
Bowles et al. 2022, accepted NIPS 2022; submitted MNRAS





Proposed for Algorithmic Assignment	
asymmetric brightness	Integrated flux ratio between source sections.
asymmetric structure	Symmetric components around host.
compact	Angular extent of the components.
diffuse	Proportion of assembly mask with emission.
double	A 'component' number of two.
edge brightened	Relative radial brightness distribution.
extended	Angular extent of the source.
faint	Integrated relative flux.
host	Whether or not a host is identifiable.
peak	Peak within the assembly mask.
small	Angular extent of assembly mask.
traces host galaxy	Assembly mask and host emission correlation.

Proposed for Tagging
amorphous, bent, bridge, core, hourglass, jet, lobe, merger, plume, tail



	Coordinates (J2000)	Query	Tags
A	21h 02m 16s -54° 23′ 36″	hourglass \ (amorphous ∪ traces host galaxy ∪ bent)	diffuse, double, edge brightened, extended, faint, host, hourglass, jet, lobe, peak
B	20h 40m 36s -53° 15′ 53″	(merger ∩ bridge) \ faint	bridge, extended, host, merger, traces host galaxy
C	20h 59m 43s -53° 58′ 52″	amorphous	amorphous, compact, extended, host, traces host galaxy
D	20h 23m 29s -56° 17′ 08″	amorphous	amorphous, compact, core, faint, host, small
E	21h 02m 34s -58° 04′ 04″	amorphous	amorphous, asymmetric structure, core, extended, faint, host

Final thoughts …

as595
radastrat
anna.scaife@manchester.ac.uk

Final thoughts …

- There are wider advantages to plain language descriptors of complex physical phenomena: collaboration, inclusivity, language barriers, barriers to participation, interdisciplinarity;
- Moving away from historical labelling schemes mitigates against learned biases and allows for new relationships (and potentially new physics) to be identified;
- The methodology we use is domain agnostic and can be repurposed for other branches of astronomy and physics more widely;
- Must be mindful of the *anglocentric* nature of our current experiment and the potential biases that may introduce.

as595
radastrat
anna.scaife@manchester.ac.uk