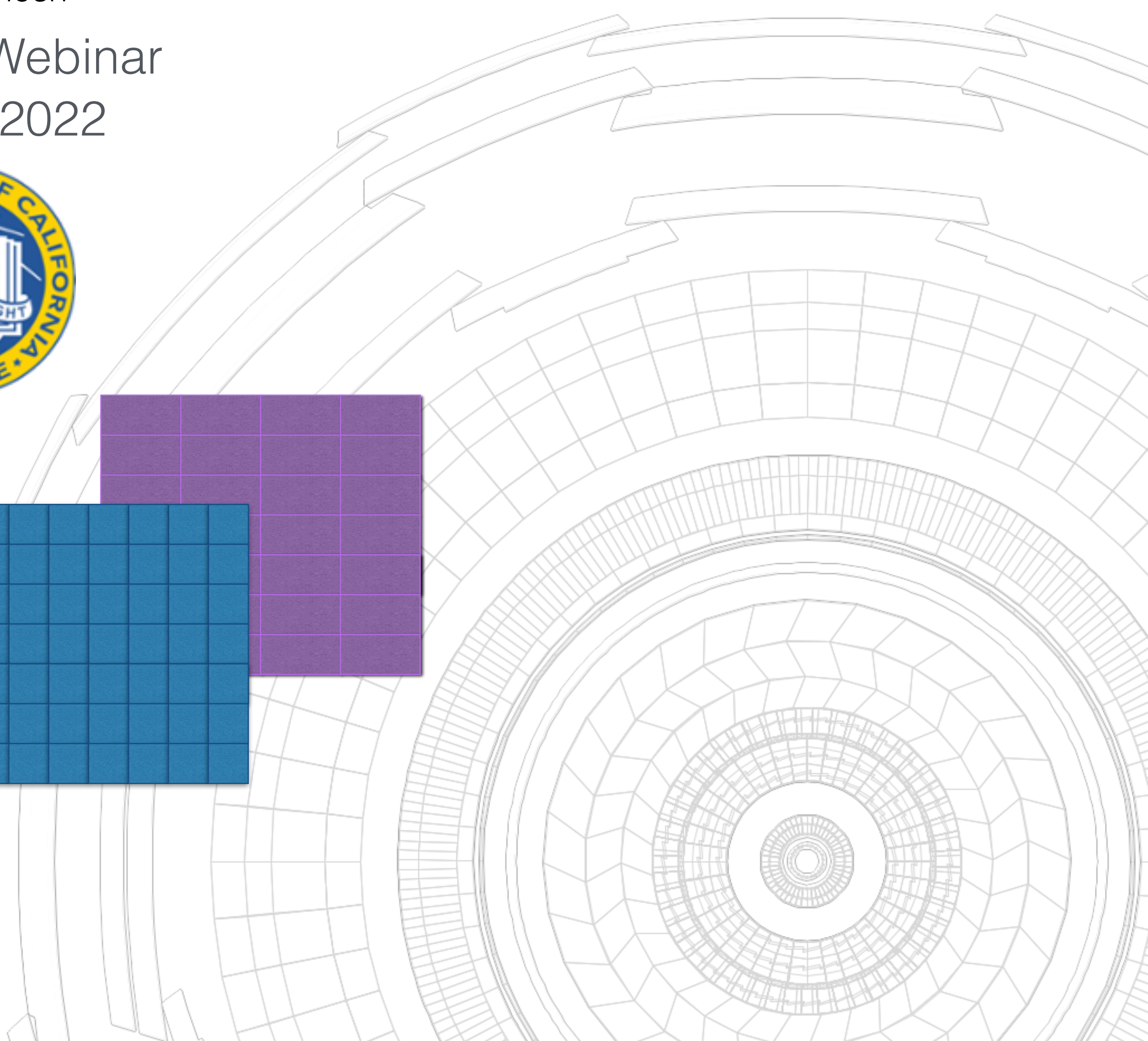
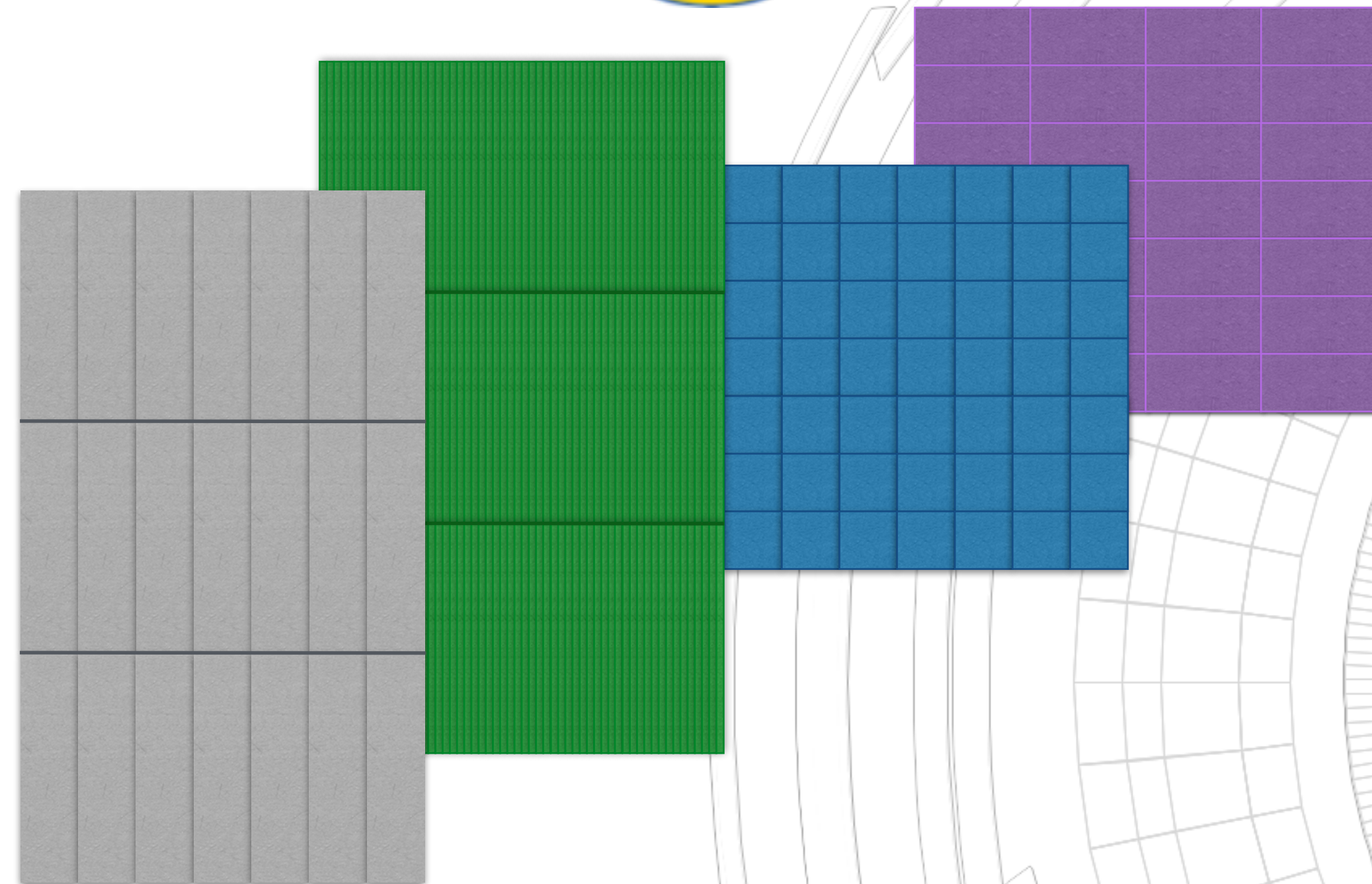


# Ethics in data science education for physicists

Aishik Ghosh

DESCOP Webinar

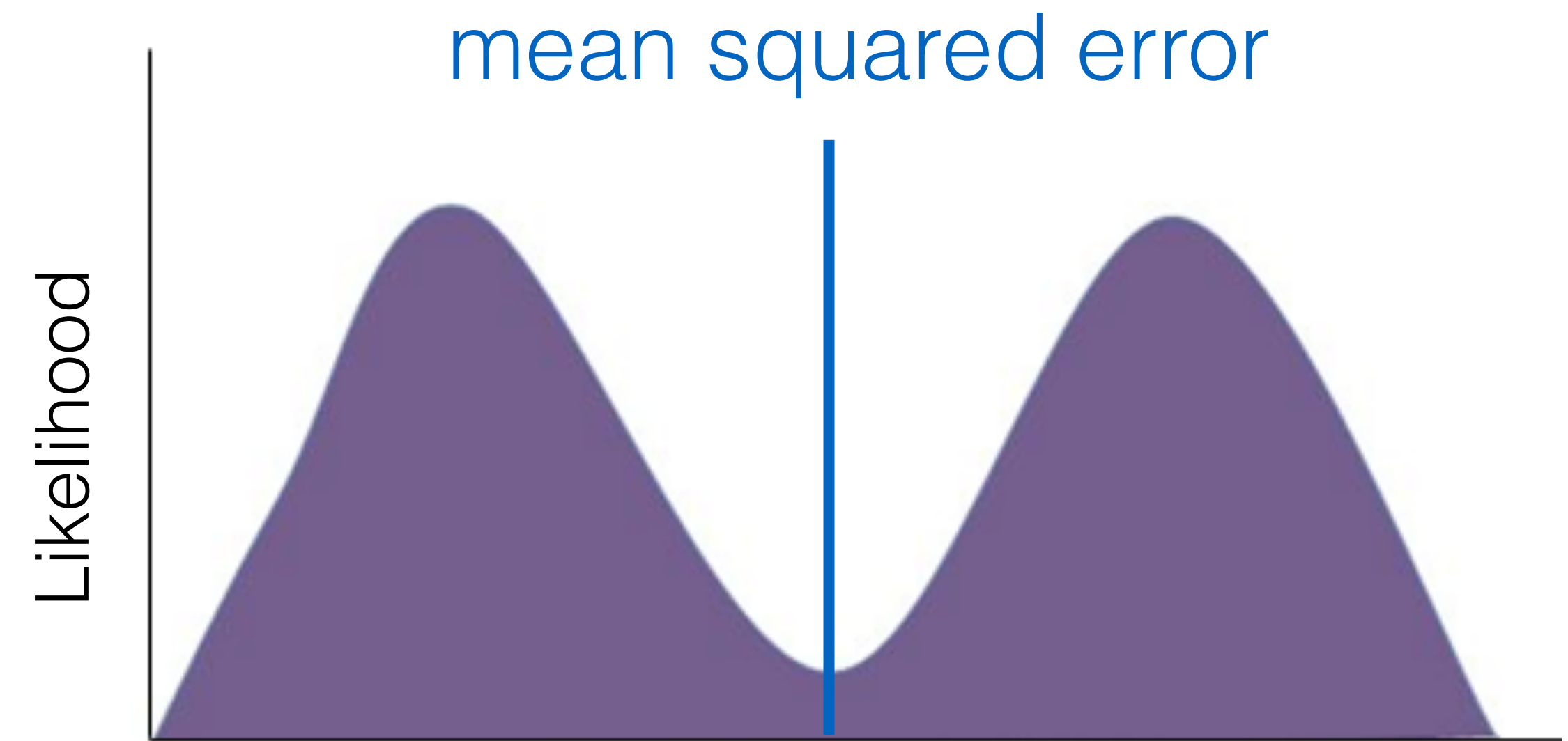
07 Dec 2022



# Data science for physics students

Transferable skills:

- Statistical reasoning
- Tools for data visualisation, exploration
- An understanding of machine learning – what it can and cannot do
- Learning limitations / implications of statistical techniques – **ethics!**



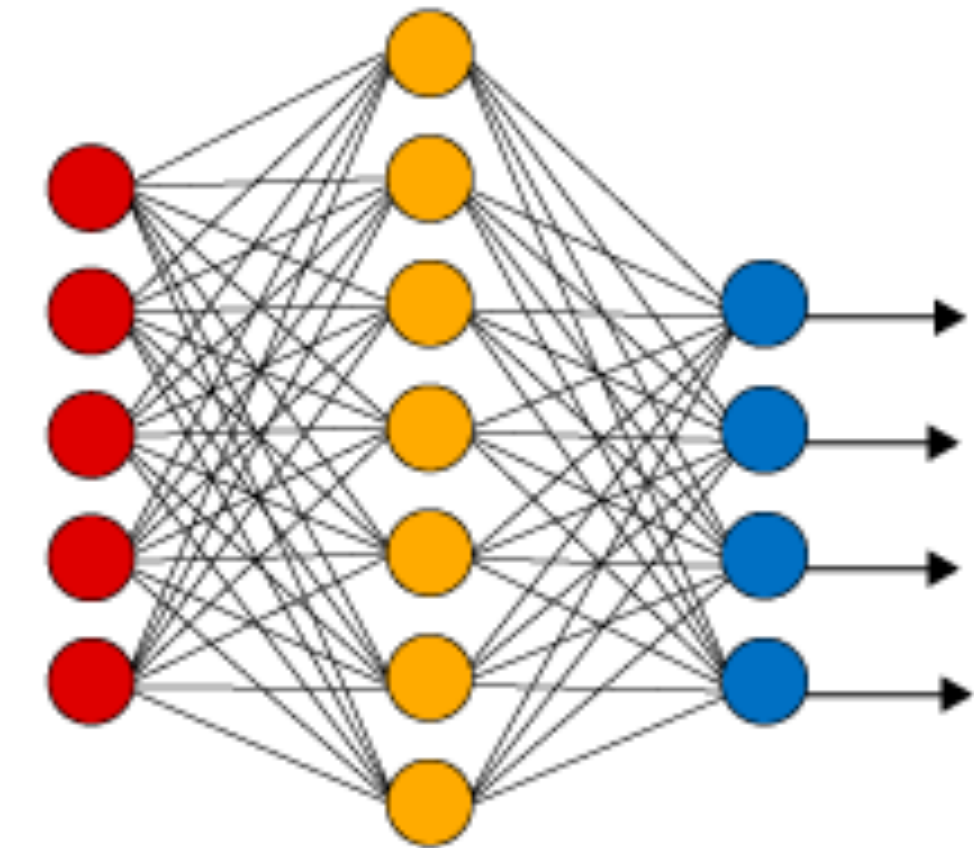
# Machine Learning

ML automatically learns patterns in training data and can make predictions on new data

Classification, regression, generation,

...

**Simple Neural Network**



Source: [https://www.reddit.com/r/Damnthatinteresting/comments/xhegp7/i\\_expanded\\_van\\_goghs\\_starry\\_night\\_with\\_openais/](https://www.reddit.com/r/Damnthatinteresting/comments/xhegp7/i_expanded_van_goghs_starry_night_with_openais/)

# Machine Learning

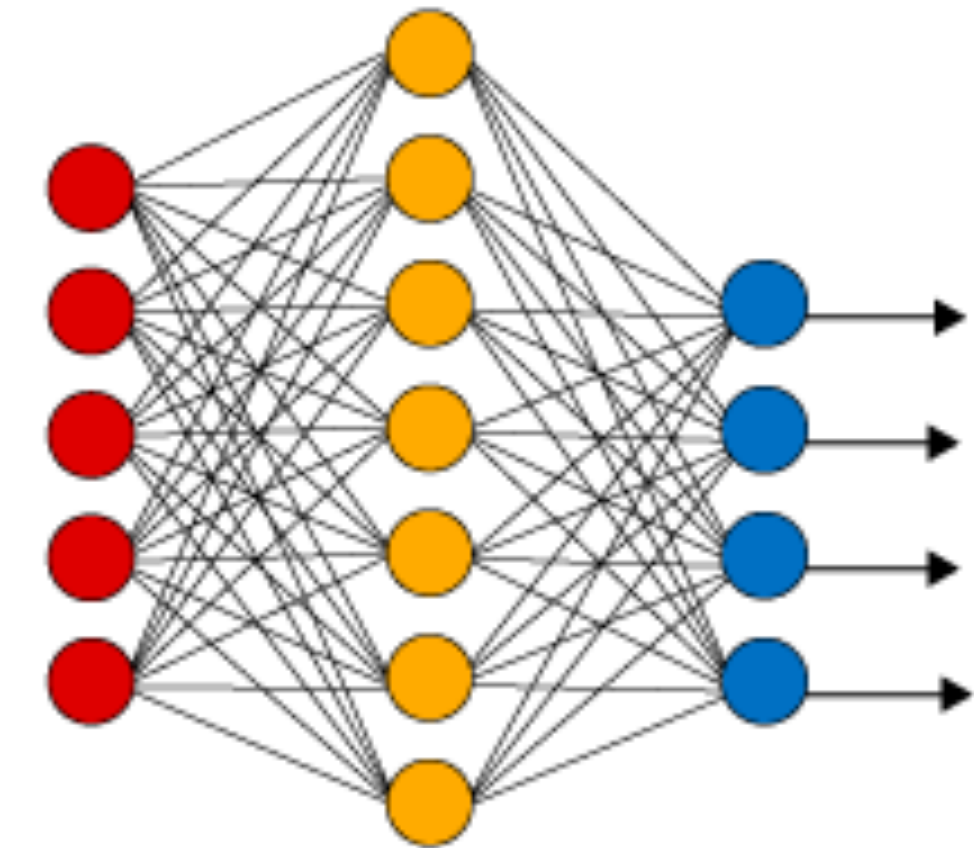
ML automatically learns patterns in training data and can make predictions on new data

Classification, regression, generation,

...



## Simple Neural Network



Source: [https://www.reddit.com/r/Damnthatinteresting/comments/xhegp7/i\\_expanded\\_van\\_goghs\\_starry\\_night\\_with\\_openais/](https://www.reddit.com/r/Damnthatinteresting/comments/xhegp7/i_expanded_van_goghs_starry_night_with_openais/)

# Machine Learning

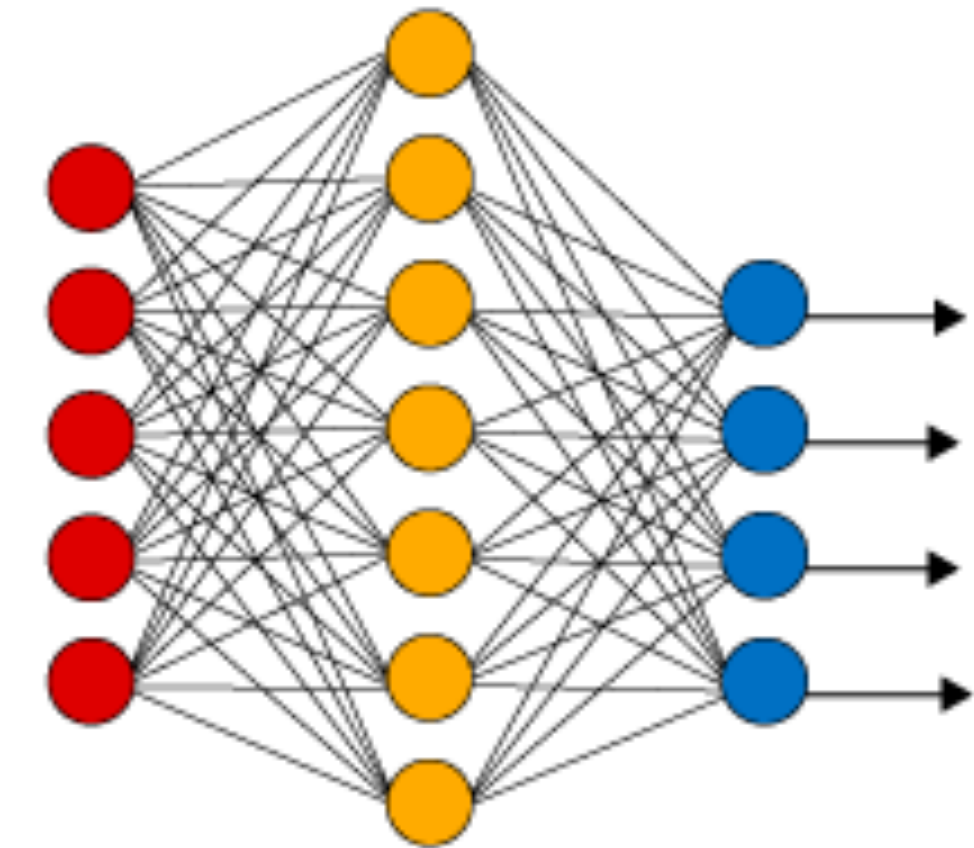
ML automatically learns patterns in training data and can make predictions on new data

Classification, regression, generation,

...



## Simple Neural Network



Source: [https://www.reddit.com/r/Damnthatinteresting/comments/xhegp7/i\\_expanded\\_van\\_goghs\\_starry\\_night\\_with\\_openai/](https://www.reddit.com/r/Damnthatinteresting/comments/xhegp7/i_expanded_van_goghs_starry_night_with_openai/)

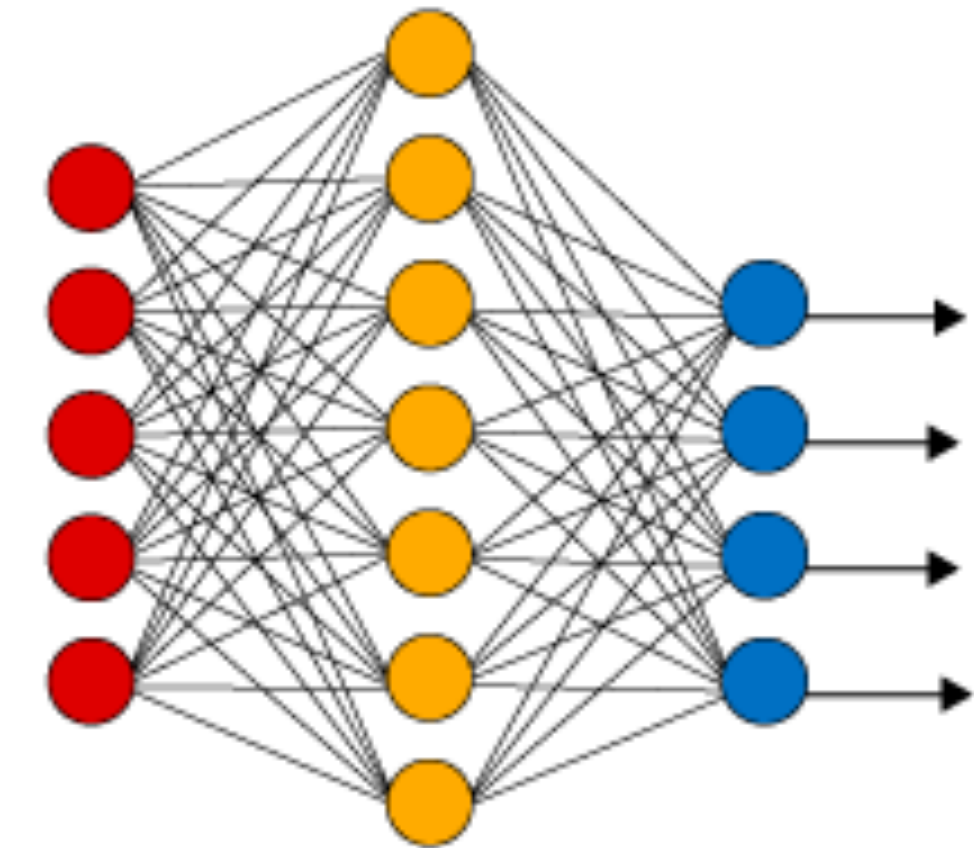
# Machine Learning

ML automatically learns patterns in training data and can make predictions on new data

Classification, regression, generation,

...

## Simple Neural Network



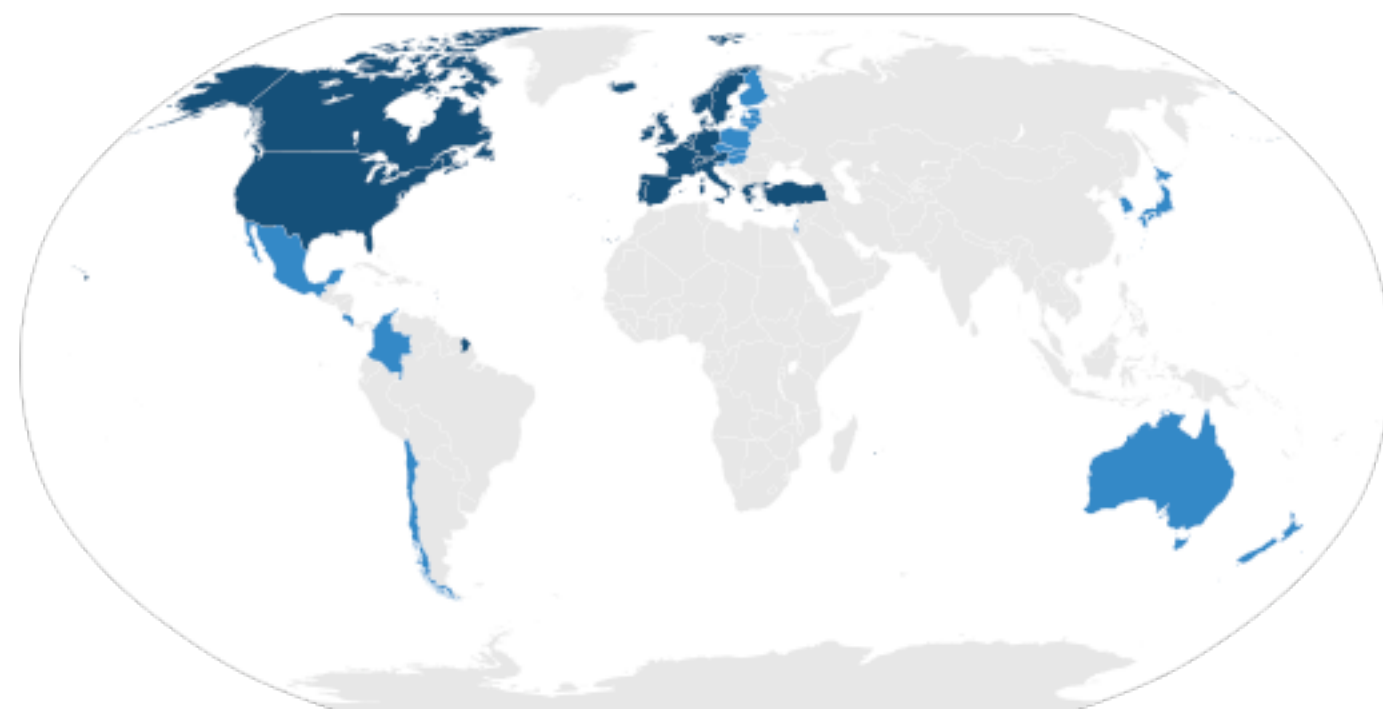
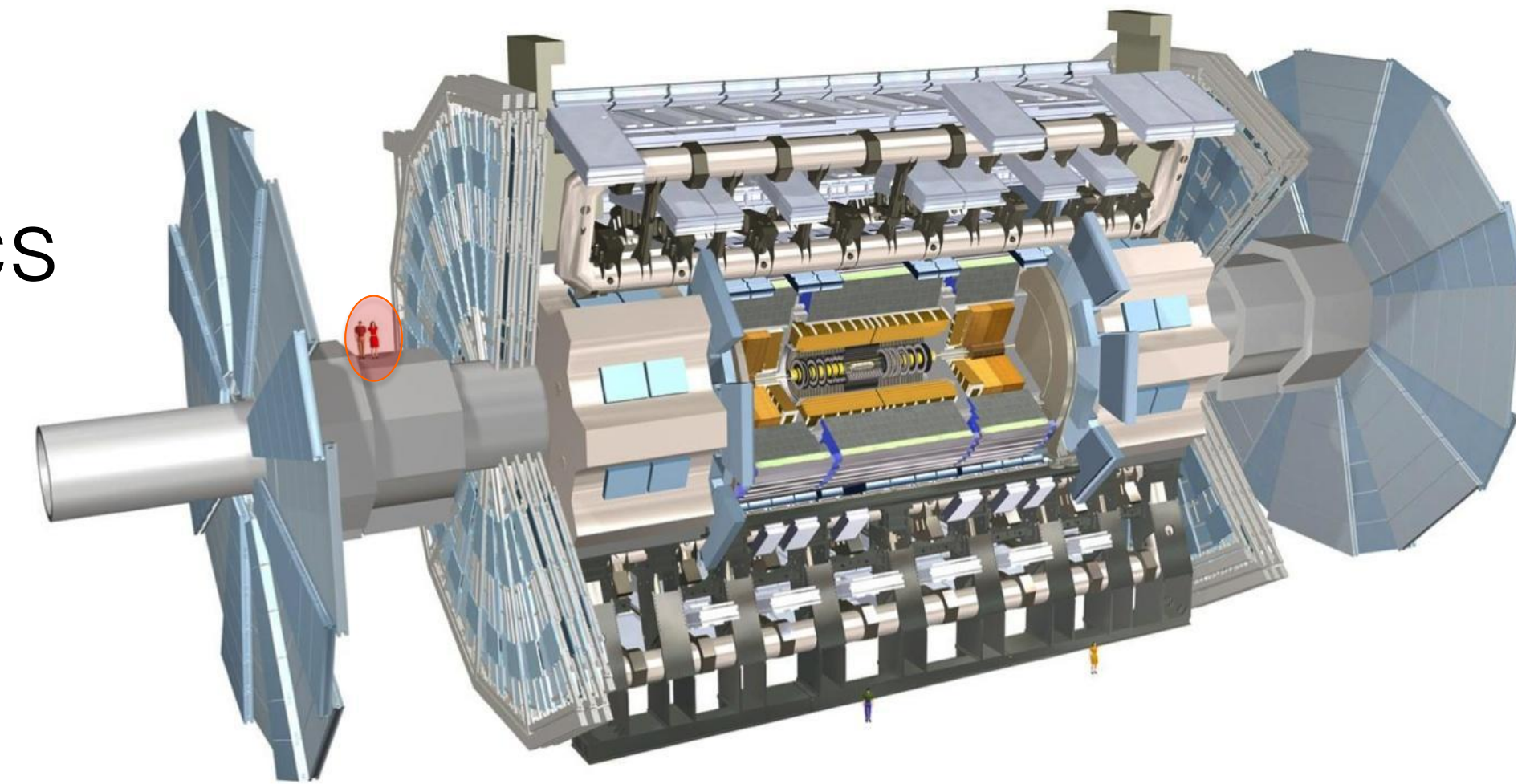
Source: [https://www.reddit.com/r/Damnthatinteresting/comments/xhegp7/i\\_expanded\\_van\\_goghs\\_starry\\_night\\_with\\_openais/](https://www.reddit.com/r/Damnthatinteresting/comments/xhegp7/i_expanded_van_goghs_starry_night_with_openais/)

# Who am I ?

Postdoctoral Scholar at UC Irvine & Berkeley Lab

Undergrad in India, PhD from Paris-Saclay University

Research: Machine Learning for particle / astro physics

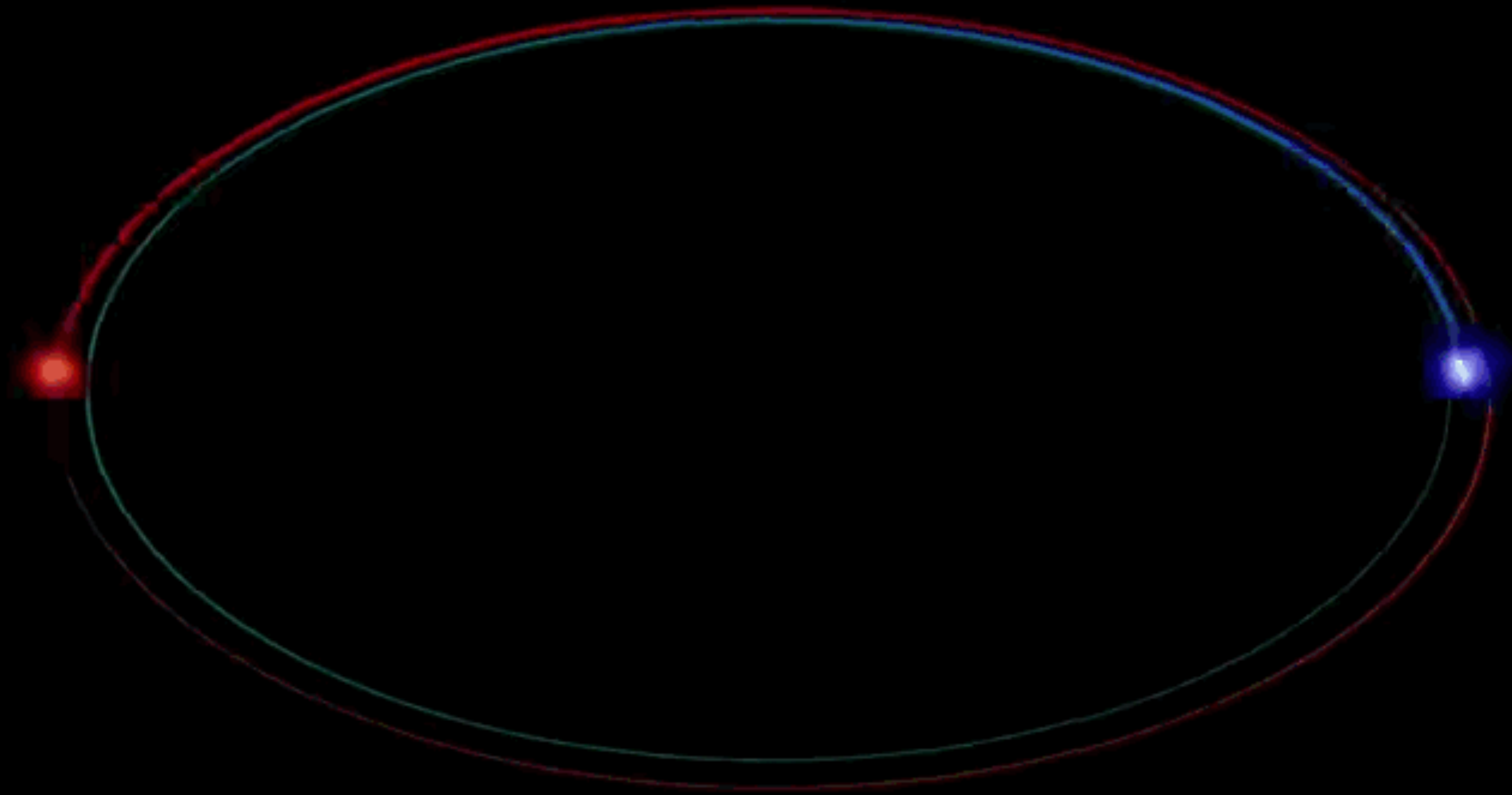


Guidelines for Trustworthy AI: Helped formulate EU's proposed AI regulation policy

Impact of AI on science

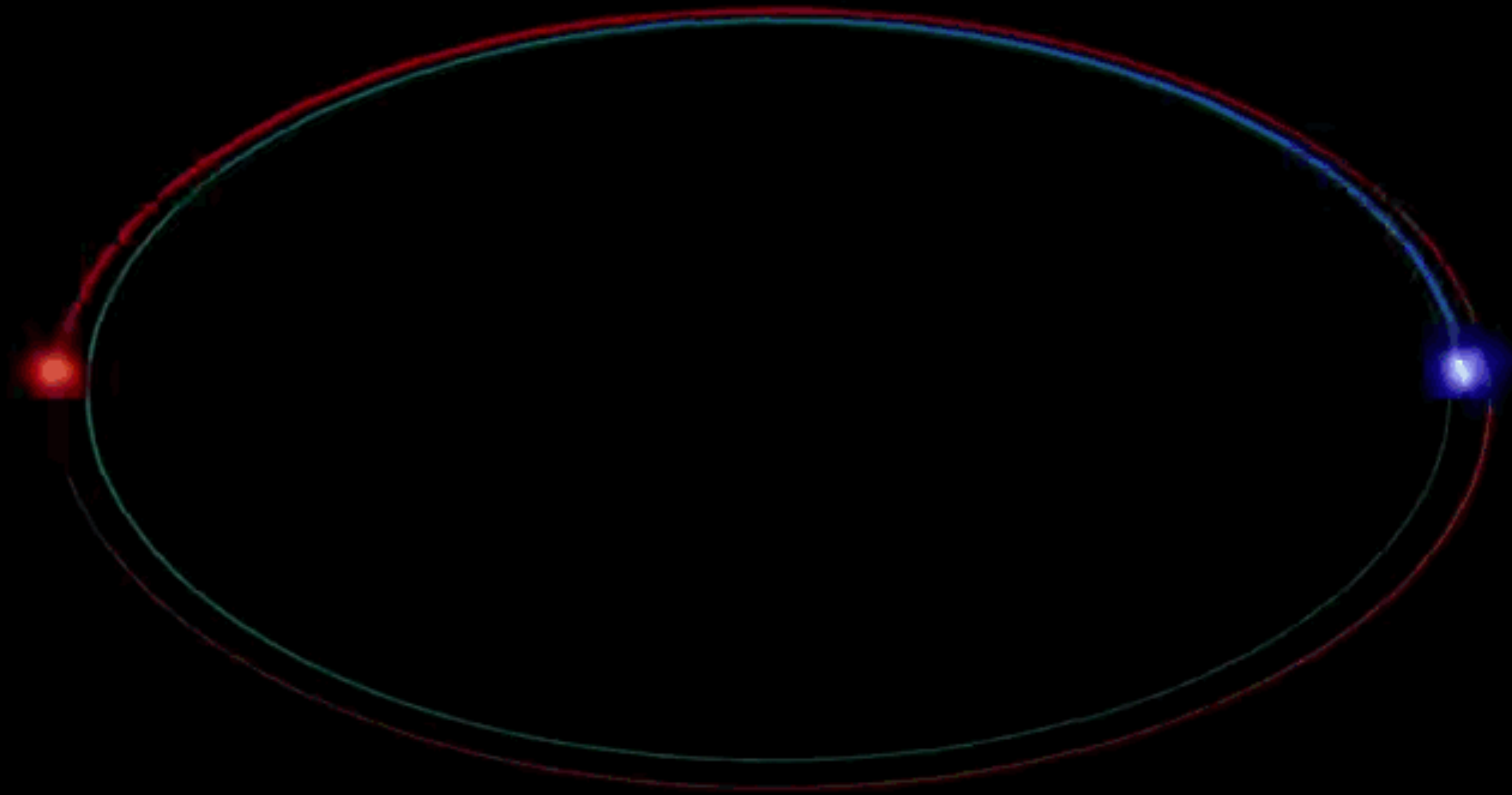
Deeply concerned about AI ethics

# Smash particles at Large Hadron Collider

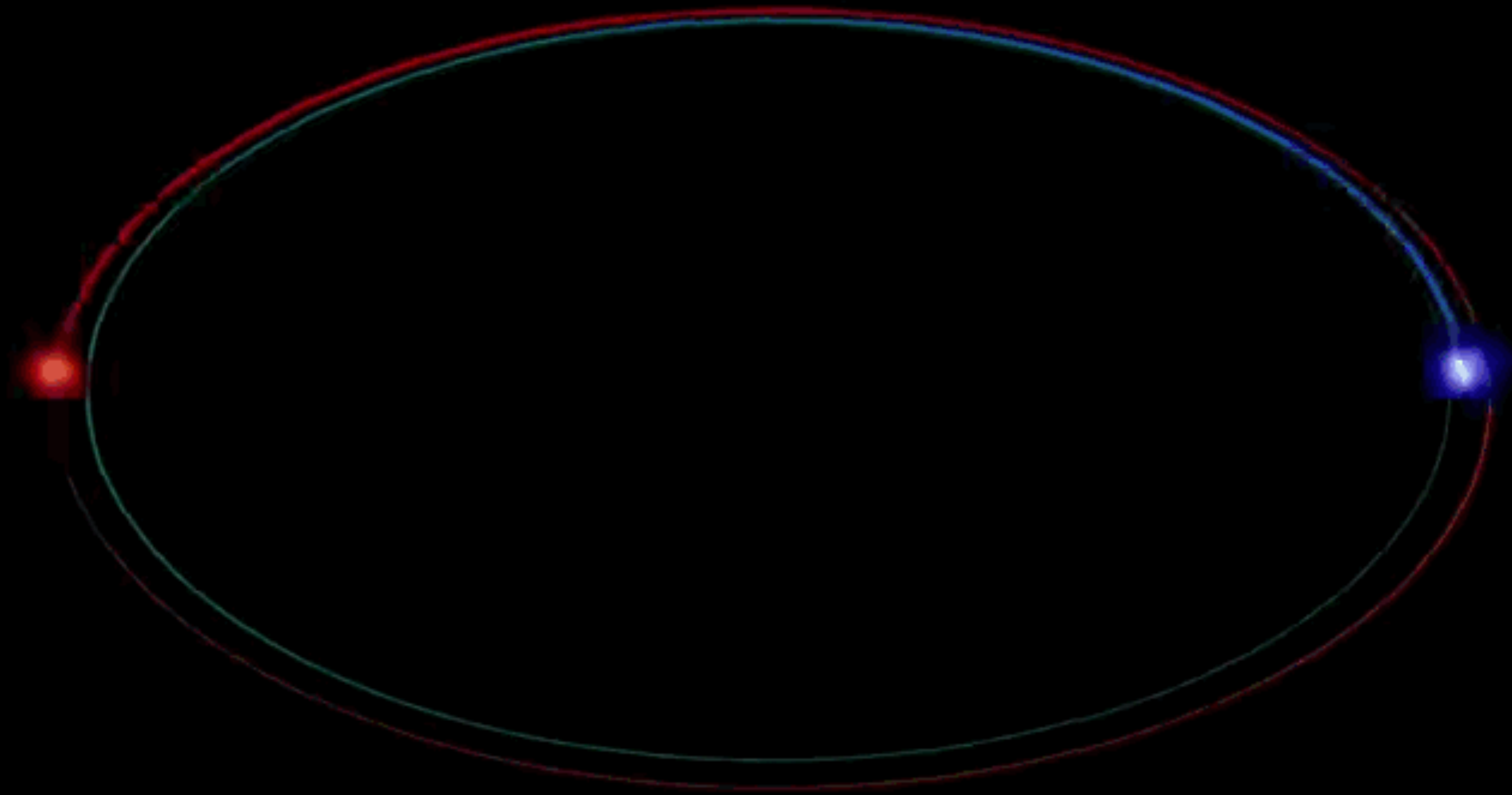




# Smash particles at Large Hadron Collider



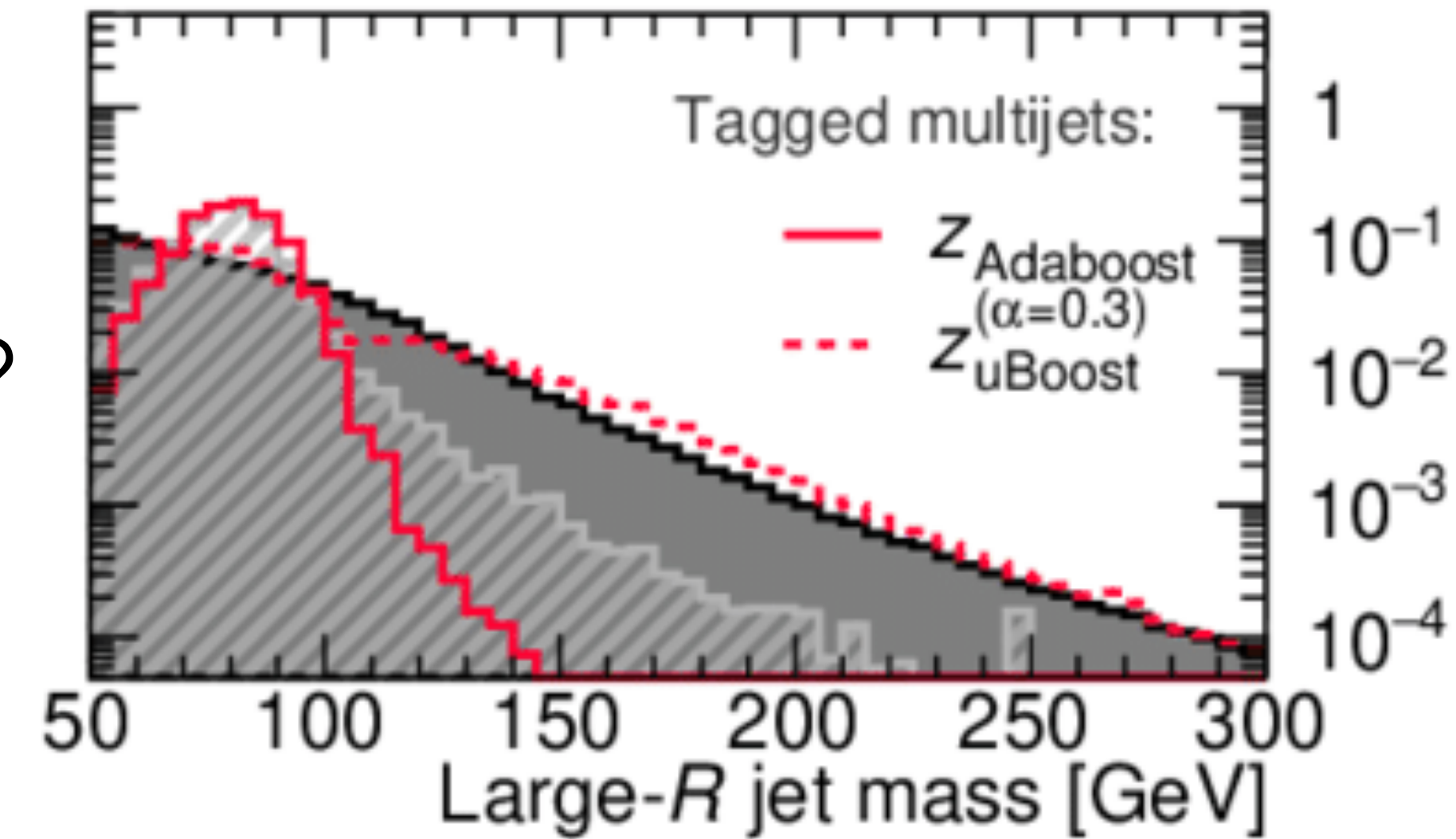
# Smash particles at Large Hadron Collider



## At LHC, we care deeply about biases & risks

ML training data comes from simulations

- Data prep process transparent and auditable ?
- Are there any systematic biases ?
- Will ML generalise to new data ? Different category of collision events?
- **Is the model overconfident ? (Answer: Almost always, yes)**
- What is the model really learning ?



Surely, scrutiny for social applications even higher ...?

At a multi-stake holder conference on AI policy:

ML models tested robustly for 'critical' applications like particle physics, transport control systems, ...

Deployed with far fewer requirements for applications to society!

[Software](#) to predict criminals – models not made public, ML to screen job applications [learns historical biases against women](#), ...



# Why ethics

---

“A physics education provides strong foundation for jobs in diverse fields”

⇒ Data science tools we teach for physics today will be used for diverse applications tomorrow

- Statistical techniques & metrics from science find their way into social applications, are the underlying assumptions still valid ? (Eg. [Predpol](#) violates assumptions on data collection method)
- Physicists bear some responsibility for dual-use technology we build (Fast inference for triggers —> weapons tomorrow ?)
- Science sounding jargon used to justify biased algorithms (Eg. Crime determination algorithms that claim high % of confidence)
- Over-investment in prediction technology rather than prevention of underlying cause, statistical / ML methods could distract us from meaningful solutions (in justice system, education)
- AI ethics an active area of research - consider a career in it!

## Why ethics

“A physics education provides strong foundation for jobs in diverse fields”

⇒ Data science tools we teach for physics today will be used for diverse applications tomorrow

- Statistical techniques & metrics from science find their way into social applications, are the underlying assumptions still valid ? (Eg. [Predpol](#) violates assumptions on data collection method)
- Physicists bear some responsibility for dual-use technology we build (Fast inference for triggers → weapons tomorrow ?)
- Science sounding jargon used to justify biased algorithms (Eg. Crime determination algorithms that claim high % of confidence)
- Over-investment in prediction technology rather than prevention of underlying cause, statistical / ML methods could distract us from meaningful solutions (in justice system, education)
- AI ethics an active area of research - consider a career in it!

## Why ethics

“A physics education provides strong foundation for jobs in diverse fields”

⇒ Data science tools we teach for physics today will be used for diverse applications tomorrow

- Statistical techniques & metrics from science find their way into social applications, are the underlying assumptions still valid ? (Eg. [Predpol](#) violates assumptions on data collection method)

- Physicists bear some responsibility for dual-use technology we build (Fast inference for triggers)

Sciencification of social issues: Optimise quantitative metrics, ignore what cannot be quantified, treat proxy metrics as the full story

- Science sounding jargon used to justify biased algorithms (Eg. Online determination algorithms that claim high % of confidence)
- Over-investment in prediction technology rather than prevention of underlying cause, statistical / ML methods could distract us from meaningful solutions (in justice system, education)
- AI ethics an active area of research - consider a career in it!

## Why ethics

“A physics education provides strong foundation for jobs in diverse fields”

⇒ Data science tools we teach for physics today will be used for diverse applications tomorrow

- Statistical techniques & metrics from science find their way into social applications, are the underlying assumptions still valid ? (Eg. [Predpol](#) violates assumptions on data collection method)

- Physicists bear some responsibility for dual-use technology we build (Fast inference for triggers)

Sciencification of social issues: Optimise quantitative metrics, ignore what cannot be quantified, treat proxy metrics as the full story

- Science sounding jargon used to justify biased algorithms (Eg. Online determination algorithms that claim high % of confidence)

Those of us with a physics training are prone to this mistake – An example at the LHC!

- Over-investment in prediction technology rather than prevention of underlying cause, statistical / ML methods could distract us from meaningful solutions (in justice system, education)
- AI ethics an active area of research - consider a career in it!

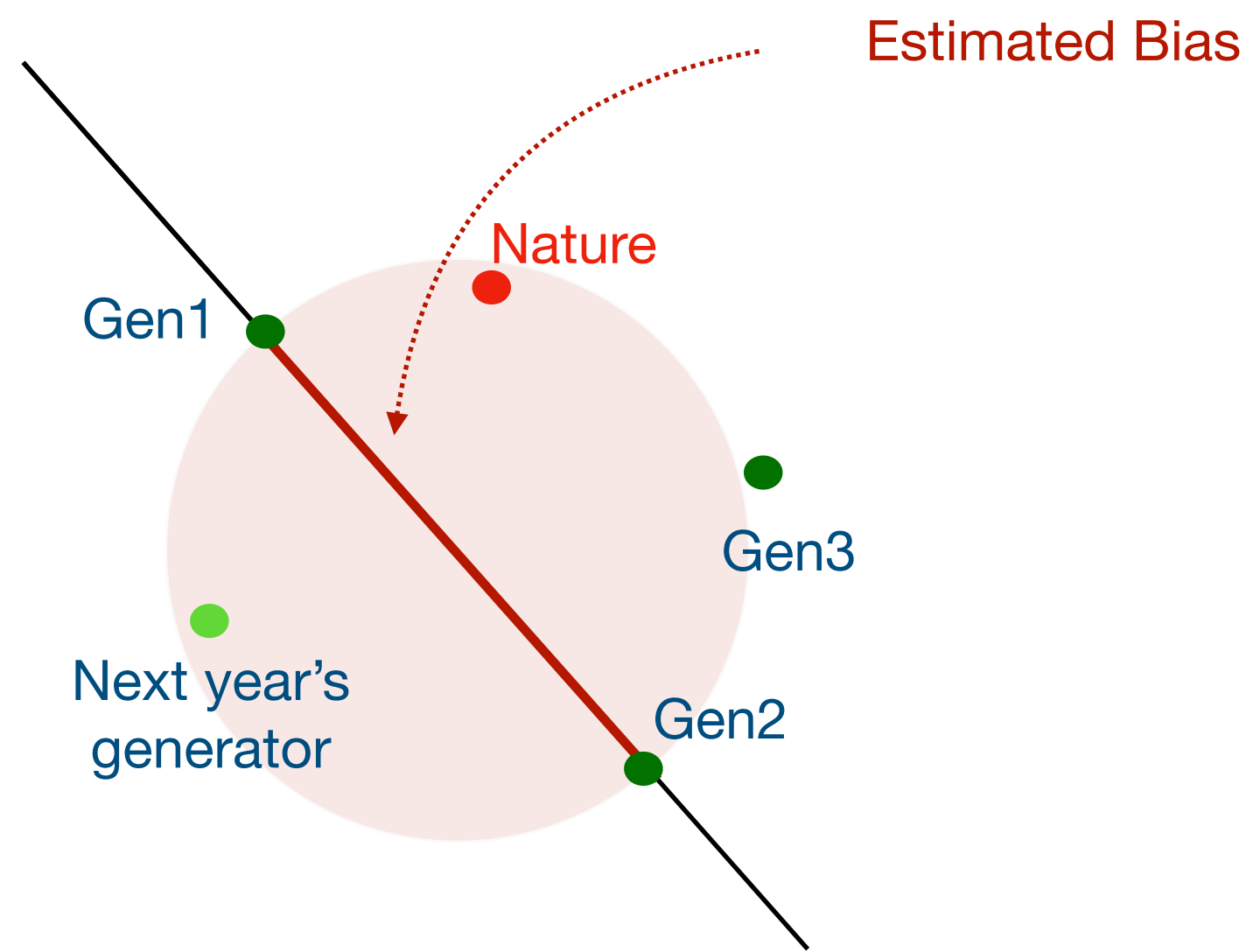


# “Just use de-biasing methods”

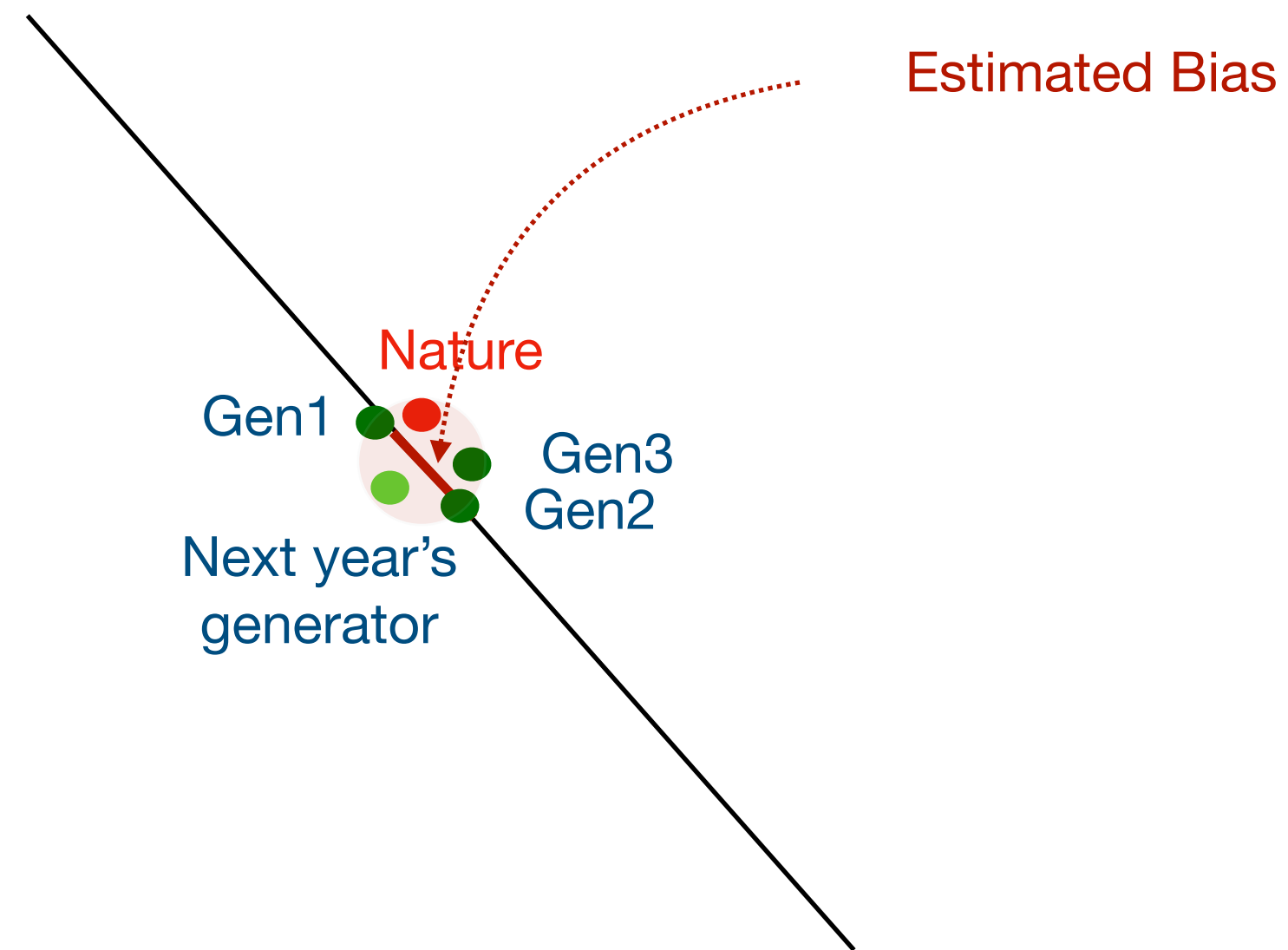
If our technique is unbiased, all points should lie together

[arXiv:2109.08159](https://arxiv.org/abs/2109.08159)

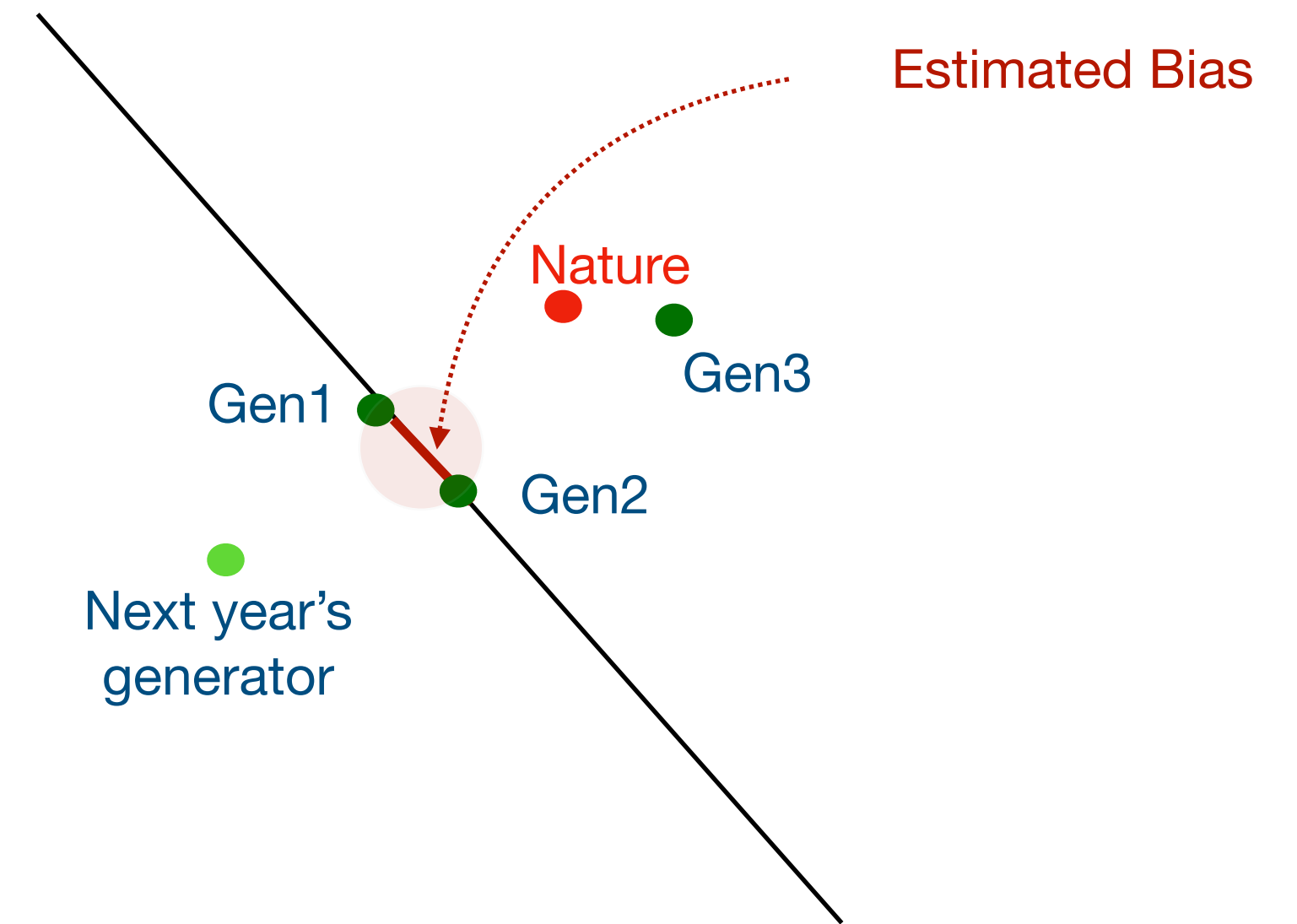
Default



What you want with de-biasing



What you get with de-biasing

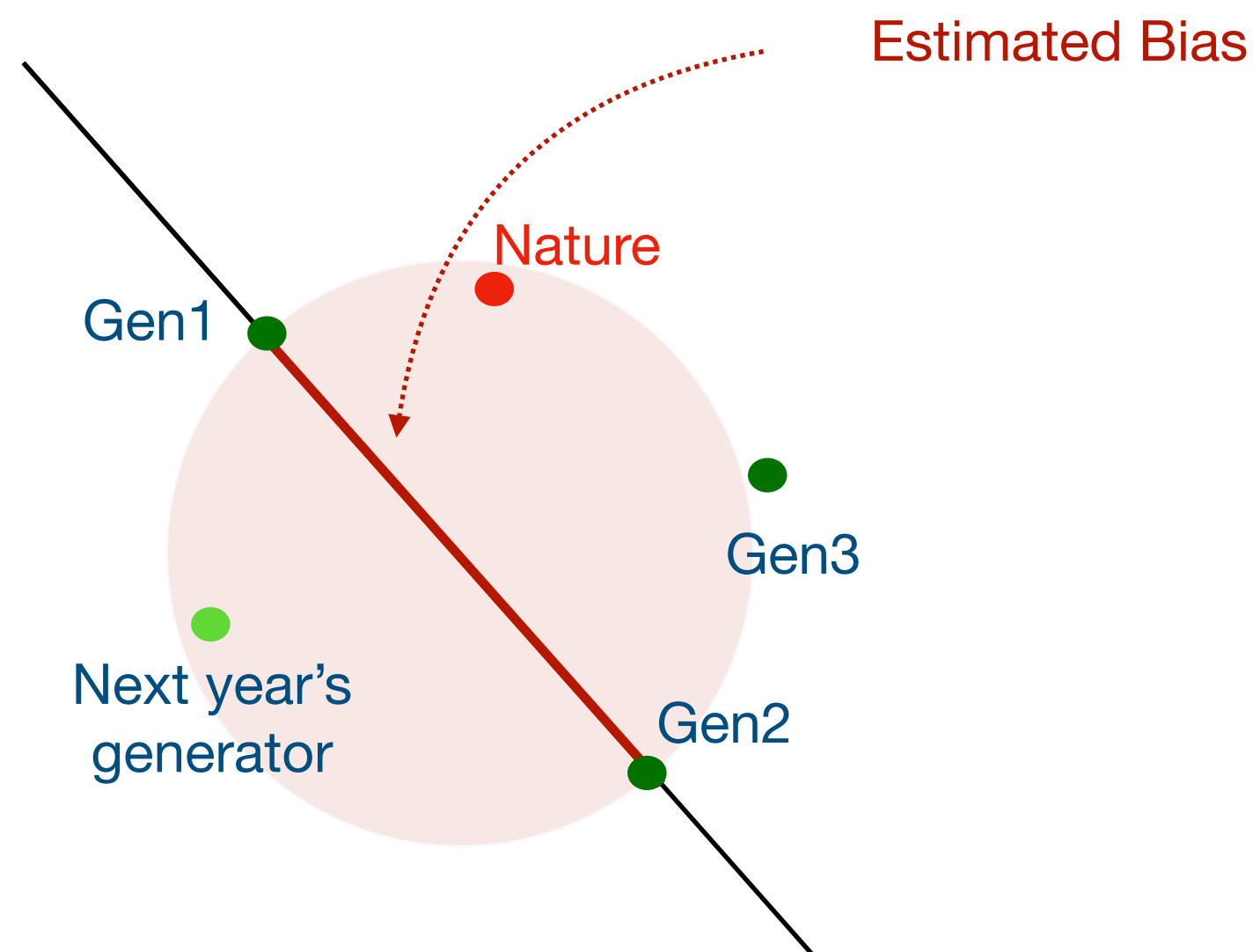


# “Just use de-biasing methods”

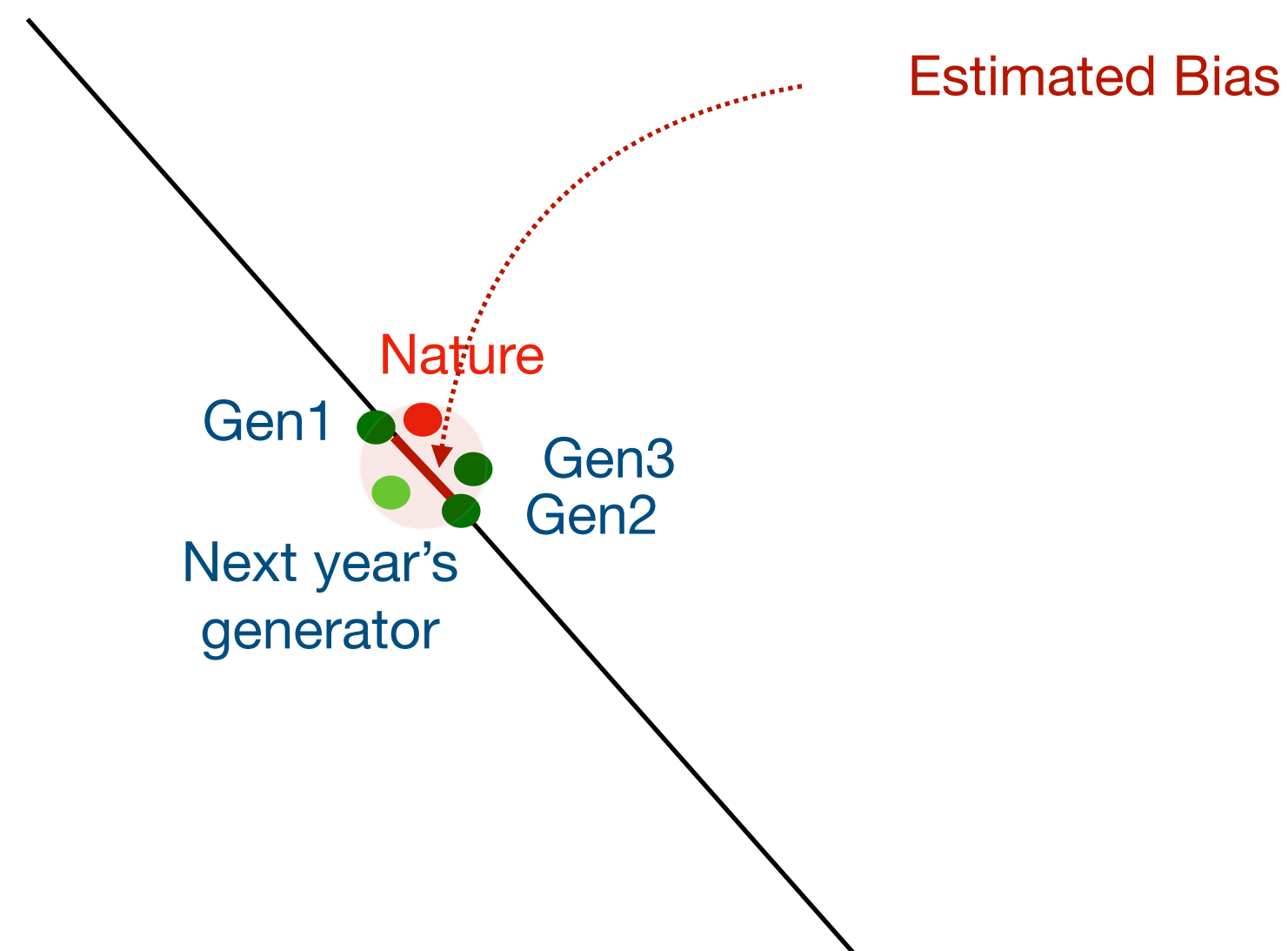
If our technique is unbiased, all points should lie together

[arXiv:2109.08159](https://arxiv.org/abs/2109.08159)

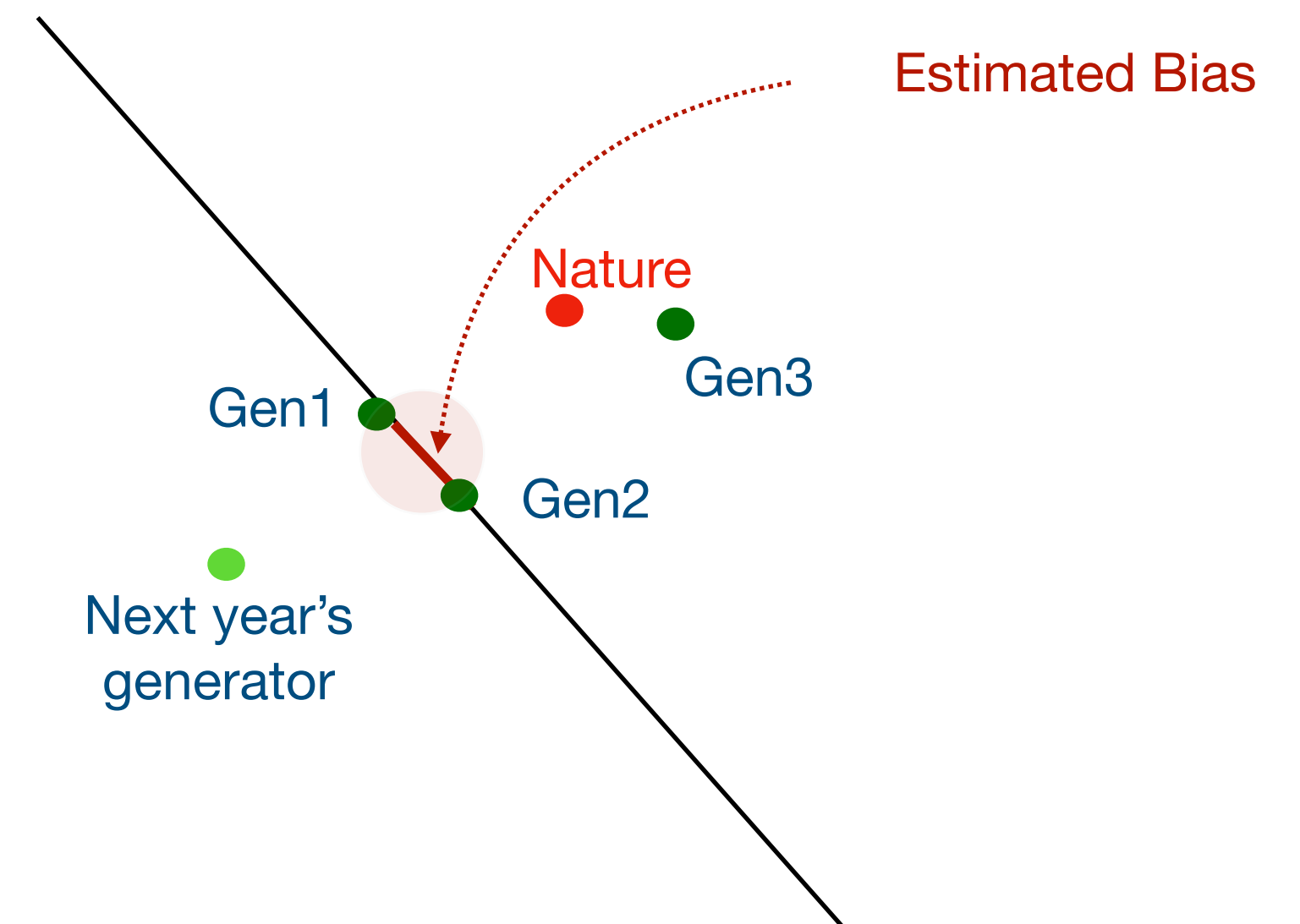
Default



What you want with de-biasing



What you get with de-biasing



True bias: Volume covered by the points

Proxy measure of bias: Distance between Gen1 & Gen2

Physicists often suggested to use 'de-biasing' method here, but it would only reduce the proxy-measure, not the true bias

We showed this with additional studies, **much harder to prove in social applications!**

## Cases in society: Harm the vulnerable

---

In 2013, the Netherlands government deployed a model to detect welfare fraud by people receiving childcare benefits.

[Wrongly accused 30,000 parents](#). It used incorrect data about people, used statistical correlations to cause individuals of fraud (eg. Having a dual nationality with Turkey/Morocco/ Eastern Europe)

**Before, each case was reviewed by humans, now they had no recourse. Took 6 years to disband the system.**

## Cases in society: Harm the vulnerable

---

In 2013, the Netherlands government deployed a model to detect welfare fraud by people receiving childcare benefits.

[Wrongly accused 30,000 parents](#). It used incorrect data about people, used statistical correlations to cause individuals of fraud (eg. Having a dual nationality with Turkey/Morocco/ Eastern Europe)

**Before, each case was reviewed by humans, now they had no recourse. Took 6 years to disband the system.**

But even when models are accurate (eg. predicting risk of default), do people not have a [right to explanation](#) and feedback?

European countries are increasingly engaging in this discussion

# Misinformation

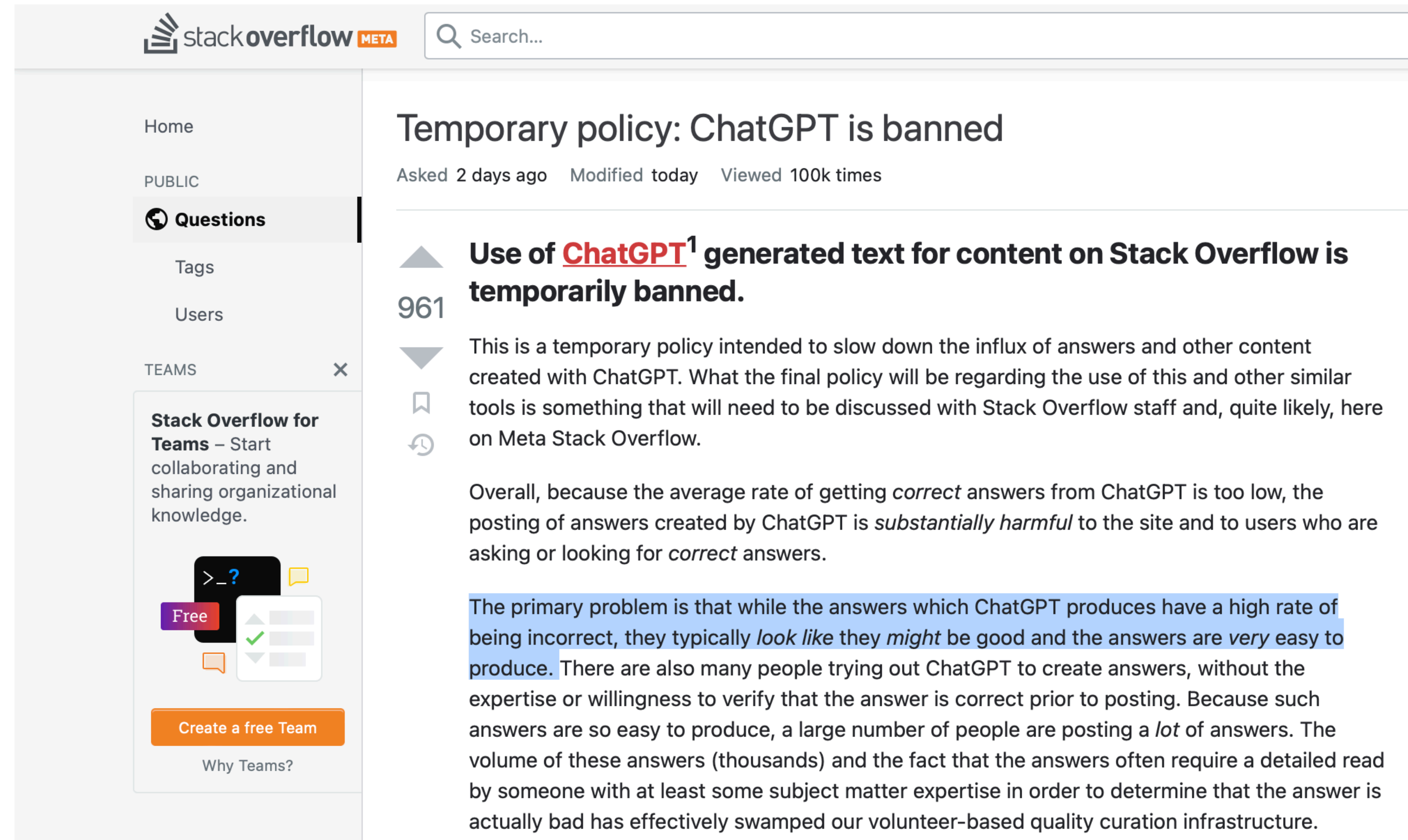
Large Language Models trained on content from the internet can today chat with you like a human

Could ChatGPT be your AI therapist?  
Could it teach you programming?

## Problems:

- Data: It learnt from the internet, what could go wrong?
- Model: It learns to sound plausible, not factual

And mistakes in code are the easiest to find, other dangers of such deployment harder to quantify



The screenshot shows a Stack Overflow page with the following elements:

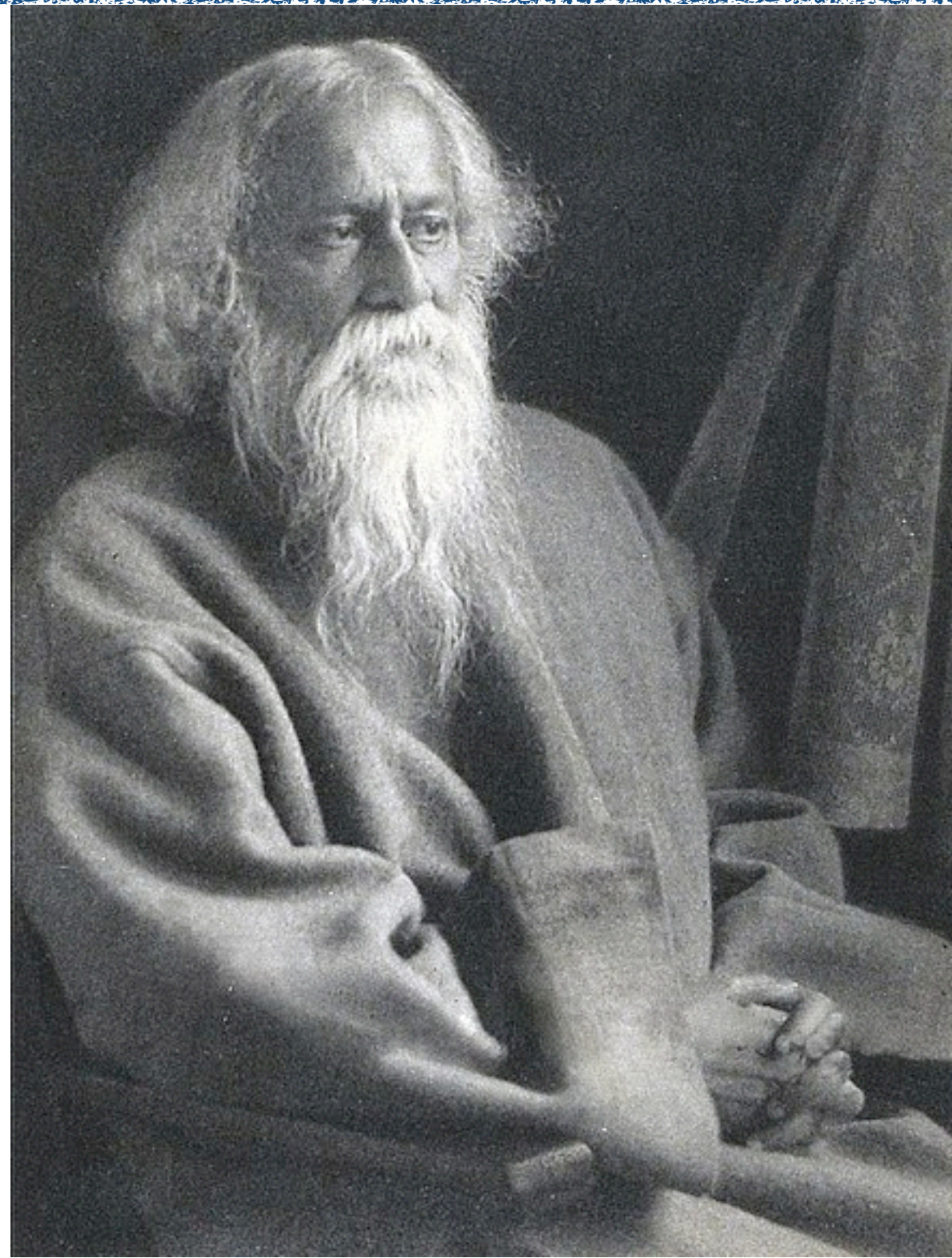
- Header:** Stack Overflow logo with a 'META' tag and a search bar.
- Navigation:** Home, PUBLIC, Questions (selected), Tags, Users, TEAMS (with a close icon).
- Left Sidebar:** A promotional card for 'Stack Overflow for Teams' with the text 'Start collaborating and sharing organizational knowledge.' and a 'Create a free Team' button.
- Main Content:**
  - Title:** 'Temporary policy: ChatGPT is banned' (961 votes).
  - Metadata:** 'Asked 2 days ago', 'Modified today', 'Viewed 100k times'.
  - Body:**
    - A paragraph explaining the temporary policy: 'This is a temporary policy intended to slow down the influx of answers and other content created with ChatGPT. What the final policy will be regarding the use of this and other similar tools is something that will need to be discussed with Stack Overflow staff and, quite likely, here on Meta Stack Overflow.'
    - A paragraph stating: 'Overall, because the average rate of getting *correct* answers from ChatGPT is too low, the posting of answers created by ChatGPT is *substantially harmful* to the site and to users who are asking or looking for *correct* answers.'
    - A highlighted paragraph: 'The primary problem is that while the answers which ChatGPT produces have a high rate of being incorrect, they typically *look like they might* be good and the answers are *very easy to produce*. There are also many people trying out ChatGPT to create answers, without the expertise or willingness to verify that the answer is correct prior to posting. Because such answers are so easy to produce, a large number of people are posting a *lot* of answers. The volume of these answers (thousands) and the fact that the answers often require a detailed read by someone with at least some subject matter expertise in order to determine that the answer is actually bad has effectively swamped our volunteer-based quality curation infrastructure.'

## Conclusion (1/2)

---

- Discussed the importance of ethics in data science education for physics students. It's a two-way street of idea exchange
- Easier to demonstrate dangers of bias / mistakes in physics data
- We cannot ignore problems just because they are difficult to quantify
- Science inspired / technological solutions can entrench systemic problems
- Next speakers will develop the discussion about the much more challenging and important questions, impact on society

# Conclusion



Rabindranath Tagore talked about being ruled by “the machine”, that offers no flexibility to individuals, that dehumanises, and forms rigid systems people must follow, justified by “science”, [in the year 1917](#).

Today, a rush for automation with technological solutions may further entrench systemic problems into unquestionable “machines” with no explainability, flexibility / human-in-the-loop, justified by bad science.

Unless we are cognisant of ethics

# Goodhart's Law

When a measure becomes a target, it ceases to be a good measure



# Goodhart's Law

When a measure becomes a target, it ceases to be a good measure

=> Dangerous to optimise proxy metrics of uncertainty