Understanding our Universe with AI : aims and challenges

Rémy Koskas
Doctoral advisor: Jean Michel Alimi

LUTH Students Day

10 mars 2021

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

- Dark Matter only $\Rightarrow$ gravitation + cosmology

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

- Dark Matter only $\Rightarrow$ gravitation + cosmology
- AMR Code : the mesh is refined where "something happens" (where the number of particles per cell exceeds a certain threshold)

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

- Dark Matter only $\Rightarrow$ gravitation + cosmology
- AMR Code : the mesh is refined where "something happens" (where the number of particles per cell exceeds a certain threshold)
- For FUR [Full Universe Run] simulation, $L = 21$ Gpc, $N = 500 \cdot 10^9$ particles

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

- Dark Matter only $\Rightarrow$ gravitation + cosmology
- AMR Code : the mesh is refined where "something happens" (where the number of particles per cell exceeds a certain threshold)
- For FUR [Full Universe Run] simulation, $L = 21$ Gpc, $N = 500 \cdot 10^9$ particles
- From one simulation to another, the dark energy model is varied

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

- Dark Matter only $\Rightarrow$ gravitation + cosmology
- AMR Code : the mesh is refined where "something happens" (where the number of particles per cell exceeds a certain threshold)
- For FUR [Full Universe Run] simulation, $L = 21$ Gpc, $N = 500 \cdot 10^9$ particles
- From one simulation to another, the dark energy model is varied
  - An equation of state $w = P/\rho$ for the dark energy fluid is chosen :

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

- Dark Matter only $\Rightarrow$ gravitation + cosmology
- AMR Code : the mesh is refined where "something happens" (where the number of particles per cell exceeds a certain threshold)
- For FUR [Full Universe Run] simulation, $L = 21$ Gpc, $N = 500 \cdot 10^9$ particles
- From one simulation to another, the dark energy model is varied
  - An equation of state $w = P/\rho$ for the dark energy fluid is chosen : ΛCDM, quintessence ...
  - The cosmological parameters values ($\Omega_.$) are then imposed so that the n-uplet $(w, \Omega_m, \Omega_{DE}...)$ is compatible with supernovae and CMB data.

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

- Dark Matter only $\Rightarrow$ gravitation + cosmology
- AMR Code : the mesh is refined where "something happens" (where the number of particles per cell exceeds a certain threshold)
- For FUR [Full Universe Run] simulation, $L = 21$ Gpc, $N = 500 \cdot 10^9$ particles
- From one simulation to another, the dark energy model is varied
  - An equation of state $w = P/\rho$ for the dark energy fluid is chosen : ΛCDM, quintessence ...
  - The cosmological parameters values ($\Omega_.$) are then imposed so that the n-uplet $(w, \Omega_m, \Omega_{DE}...)$ is compatible with supernovae and CMB data.
- Those are REALISTIC models

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

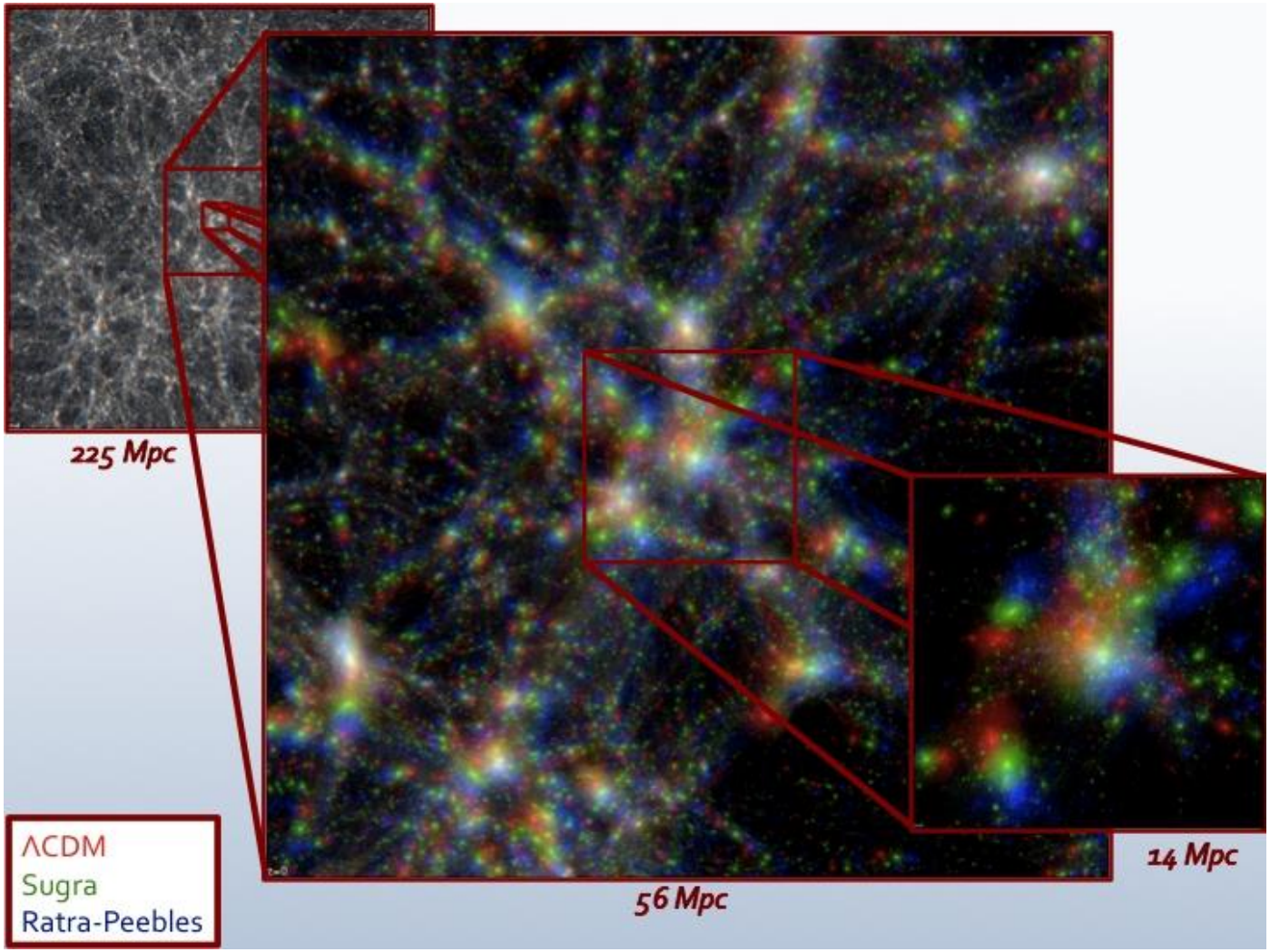
- Dark Matter only $\Rightarrow$ gravitation + cosmology
- AMR Code : the mesh is refined where "something happens" (where the number of particles per cell exceeds a certain threshold)
- For FUR [Full Universe Run] simulation, $L = 21$ Gpc, $N = 500 \cdot 10^9$ particles
- From one simulation to another, the dark energy model is varied
  - An equation of state $w = P/\rho$ for the dark energy fluid is chosen : ΛCDM, quintessence ...
  - The cosmological parameters values ($\Omega_.$) are then imposed so that the n-uplet $(w, \Omega_m, \Omega_{DE}...)$ is compatible with supernovae and CMB data.
- Those are REALISTIC models $\Rightarrow$

Numerics

Dark Energy Universe Simulation `deus-consortium.com`, LUTH, 2012
High performance cosmological simulation

- Dark Matter only $\Rightarrow$ gravitation + cosmology
- AMR Code : the mesh is refined where "something happens" (where the number of particles per cell exceeds a certain threshold)
- For FUR [Full Universe Run] simulation, $L = 21$ Gpc, $N = 500 \cdot 10^9$ particles
- From one simulation to another, the dark energy model is varied
  - An equation of state $w = P/\rho$ for the dark energy fluid is chosen : ΛCDM, quintessence ...
  - The cosmological parameters values ($\Omega_.$) are then imposed so that the n-uplet $(w, \Omega_m, \Omega_{DE}...)$ is compatible with supernovae and CMB data.
- Those are REALISTIC models $\Rightarrow$ the final Universes are very close one to each other [Alimi&al 2010]

Making the difference ...



Questions

Almost perfect superposition of the halos

Questions

Almost perfect superposition of the halos
Natural question : is there any detectable cosmological imprint in the difference ?

Questions

Almost perfect superposition of the halos
Natural question : is there any detectable cosmological imprint in the difference ?

More formally,

Questions

Almost perfect superposition of the halos
Natural question : is there any detectable cosmological imprint in the difference ?

More formally,

- Can we predict/discriminate the cosmology by analyzing the halos ?

Questions

Almost perfect superposition of the halos
Natural question : is there any detectable cosmological imprint in the difference ?

More formally,

- Can we predict/discriminate the cosmology by analyzing the halos ?
- If so, what are the discriminatory halo properties ?

Machine Learning

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics
  - N-body simulations - learn to associate to nitial conditions the final matter distribution [Lucie-Smith *et al* 2018]

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics
  - N-body simulations - learn to associate to nitial conditions the final matter distribution [Lucie-Smith *et al* 2018]
  - Cosmological estimation From the TOTAL matter distribution, estimate by deep learning the Ω's of the background cosmology. [Ravanbakhsh *et al* 2016]

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics
  - N-body simulations - learn to associate to nitial conditions the final matter distribution [Lucie-Smith *et al* 2018]
  - Cosmological estimation From the TOTAL matter distribution, estimate by deep learning the Ω's of the background cosmology. [Ravanbakhsh *et al* 2016]
- We took a different approach

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics
  - N-body simulations - learn to associate to nitial conditions the final matter distribution [Lucie-Smith *et al* 2018]
  - Cosmological estimation From the TOTAL matter distribution, estimate by deep learning the Ω's of the background cosmology. [Ravanbakhsh *et al* 2016]
- We took a different approach

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics
  - N-body simulations - learn to associate to nitial conditions the final matter distribution [Lucie-Smith *et al* 2018]
  - Cosmological estimation From the TOTAL matter distribution, estimate by deep learning the Ω's of the background cosmology. [Ravanbakhsh *et al* 2016]
- We took a different approach
  1. We have a lot of halos simulated in different cosmologies.

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics
  - N-body simulations - learn to associate to nitial conditions the final matter distribution [Lucie-Smith *et al* 2018]
  - Cosmological estimation From the TOTAL matter distribution, estimate by deep learning the Ω's of the background cosmology. [Ravanbakhsh *et al* 2016]
- We took a different approach
  1. We have a lot of halos simulated in different cosmologies.
  2. Each halo is described through quantitative properties (profile parameters and so on)

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics
  - N-body simulations - learn to associate to nitial conditions the final matter distribution [Lucie-Smith *et al* 2018]
  - Cosmological estimation From the TOTAL matter distribution, estimate by deep learning the Ω's of the background cosmology. [Ravanbakhsh *et al* 2016]
- We took a different approach
  1. We have a lot of halos simulated in different cosmologies.
  2. Each halo is described through quantitative properties (profile parameters and so on)
  3. We conceive a ML engine to associate to any processed halo (= the set of the attributes' values) the background cosmology among a list of possibilities (classification task) or the values of cosmological parameters (regression task)

Machine Learning

Where does the cosmology lie among the morphological or velocity-related attributes of the halos ? $\Rightarrow$ weak signals to be detected
Machine learning (aka AI) could do so

- Already used in cosmology/astrophysics
  - N-body simulations - learn to associate to nitial conditions the final matter distribution [Lucie-Smith *et al* 2018]
  - Cosmological estimation From the TOTAL matter distribution, estimate by deep learning the Ω's of the background cosmology. [Ravanbakhsh *et al* 2016]
- We took a different approach
  1. We have a lot of halos simulated in different cosmologies.
  2. Each halo is described through quantitative properties (profile parameters and so on)
  3. We conceive a ML engine to associate to any processed halo (= the set of the attributes' values) the background cosmology among a list of possibilities (classification task) or the values of cosmological parameters (regression task)
  4. We try to determine which properties are important to achieve the recognition - those are the "cosmologically impregnated" attributes. this is a physical output

Looking for the hidden horse

It is crucial to understand how the resulting engine works.

Looking for the hidden horse

It is crucial to understand how the resulting engine works. In particular, we have to check that the classification is achieved only by physical means, ignoring any cosmological clue coming from the numerical nature of the simulation. In other words, the engine should work on a real Universe.

Looking for the hidden horse

It is crucial to understand how the resulting engine works. In particular, we have to check that the classification is achieved only by physical means, ignoring any cosmological clue coming from the numerical nature of the simulation. In other words, the engine should work on a real Universe.

Clever Hans : Encyclopedia Britannica

Looking for the hidden horse

It is crucial to understand how the resulting engine works. In particular, we have to check that the classification is achieved only by physical means, ignoring any cosmological clue coming from the numerical nature of the simulation. In other words, the engine should work on a real Universe.

Clever Hans : Encyclopedia Britannica

"The 'Clever Hans' effect occurs when the learned model produces correct predictions based on the 'wrong' features. This effect [...] goes undetected by standard validation techniques has been frequently observed [...] where the training algorithm leverages spurious correlations in the data." [Kauffman & al 2020]

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ?

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)
- Since $\bar{\rho}$ is different in each cosmology (because of different w 's and the constraint of realistic models ...) so is m_p.

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)
- Since $\bar{\rho}$ is different in each cosmology (because of different w 's and the constraint of realistic models ...) so is m_p.
- $\Rightarrow$ So, if the machine is able to detect that

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)
- Since $\bar{\rho}$ is different in each cosmology (because of different w 's and the constraint of realistic models ...) so is m_p.
- $\Rightarrow$ So, if the machine is able to detect that
  1. all first-cosmology halo masses in the train set are multiple of the same base mass (that the machine should determine)

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)
- Since $\bar{\rho}$ is different in each cosmology (because of different w 's and the constraint of realistic models ...) so is m_p.
- $\Rightarrow$ So, if the machine is able to detect that
  1. all first-cosmology halo masses in the train set are multiple of the same base mass (that the machine should determine)

And

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)
- Since $\bar{\rho}$ is different in each cosmology (because of different w 's and the constraint of realistic models ...) so is m_p.
- $\Rightarrow$ So, if the machine is able to detect that
  1. all first-cosmology halo masses in the train set are multiple of the same base mass (that the machine should determine)

     And
  2. all second-cosmology halo masses are multiple of another base mass

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)
- Since $\bar{\rho}$ is different in each cosmology (because of different w 's and the constraint of realistic models ...) so is m_p.
- $\Rightarrow$ So, if the machine is able to detect that
  1. all first-cosmology halo masses in the train set are multiple of the same base mass (that the machine should determine)

     And

  2. all second-cosmology halo masses are multiple of another base mass

  then the machine will also be able to classify the halos of the test set (simply by looking if their masses are multiple of m_p^{Λ} or m_p^{RP}).

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)
- Since $\bar{\rho}$ is different in each cosmology (because of different w 's and the constraint of realistic models ...) so is m_p.
- $\Rightarrow$ So, if the machine is able to detect that
  1. all first-cosmology halo masses in the train set are multiple of the same base mass (that the machine should determine)

     And

  2. all second-cosmology halo masses are multiple of another base mass

  then the machine will also be able to classify the halos of the test set (simply by looking if their masses are multiple of m_p^{Λ} or m_p^{RP}).

This is a typical Clever Hans : the data embody clues w.r.t. the target variables of purely arithmetical nature, which are thus not reproducible out of this set of simulations - on real observations, for example.

Looking for the hidden horse

Do we observe Clever Hans effects if we use brute data in our work ? A lot ...
Example :

- Consider only two cosmological models, say, ΛCDM and Ratra-Peebeles.
- All the masses are integer multiple of the particle mass m_p. (Each halo contains an integer number of particles)
- Since $\bar{\rho}$ is different in each cosmology (because of different w 's and the constraint of realistic models ...) so is m_p.
- $\Rightarrow$ So, if the machine is able to detect that
  1. all first-cosmology halo masses in the train set are multiple of the same base mass (that the machine should determine)

     And

  2. all second-cosmology halo masses are multiple of another base mass

  then the machine will also be able to classify the halos of the test set (simply by looking if their masses are multiple of m_p^{Λ} or m_p^{RP}).

This is a typical Clever Hans : the data embody clues w.r.t. the target variables of purely arithmetical nature, which are thus not reproducible out of this set of simulations - on real observations, for example. This kind of effects should be carefully hunted if we want to obtain physically reliable results.

Conclusion and perspectives

- Our objective is to detect and localize cosmological imprints in halos properties.
- Signals are weak, AI could be a solution.
- But we do not only want an AI that works on the simulation, we want an AI that would work on real/physical examples.
- This is why one should understand how the AI works and track any Clever Hans.
- This means that, crucially, one has to pave the way with much physical knowledge.

Perspectives :

- Extension to projected halos
- Application to observed clusters (weak lensing ...)
- Extension of cosmological detection to gravity-theory detection

Understanding our Universe with AI : aims and challenges

Rémy Koskas
Doctoral advisor: Jean Michel Alimi

LUTH Students Day

10 mars 2021