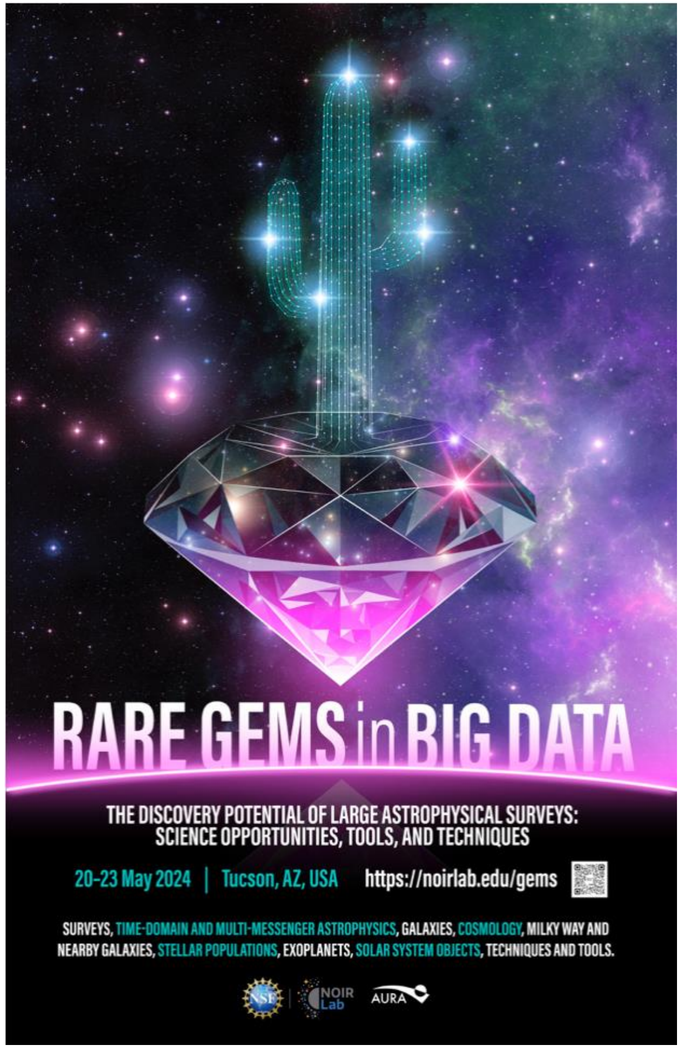
UNIVERSITY COLLEGE LONDON
CDT Data Intensive Science, Physics and Astronomy



Finding Pegasus:

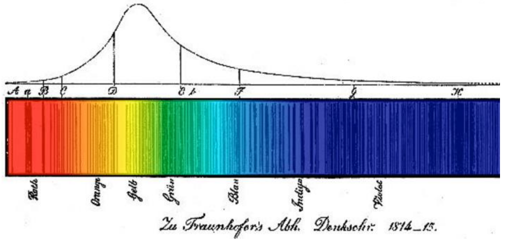
Leveraging the Manifold from Machine-Learning Dimensionality-Reduction to Enhance Unsupervised Anomaly Detection in DESI Spectra

Paul Nathan

Supervisors – Ofer Lahav, Nikos Nikolaou

Rare Gems in Big Data, 21 May 2024

Contact: Paul Nathan ucaprpn@ucl.ac.uk



DARK ENERGY SPECTROSCOPIC INSTRUMENT
U.S. Department of Energy Office of Science

The whole thing in 5 bullets

- Widescale use of unsupervised machine-learning techniques when performing anomaly detection in astronomical spectra.

- All these techniques struggle with high dimensional data hence we usually choose to work in lower dimension.

- Dimensionality reduction creates a manifold which will be model dependent and hence **the anomalies detected using it will also be model dependent.**

- We introduce the idea of thinking of anomaly detection models as working either **on manifold** and **off manifold** and note they can represent very different things.

- For a given manifold, **combining complementary on-manifold and off-manifold techniques should increase the range of anomalies we detect**

Introduction

- The volume of spectra available to modern astronomers means it’s a challenge to find anomalies
- Unsupervised ML anomaly-detection approaches are used extensively, however, they can struggle with high dimensional data.

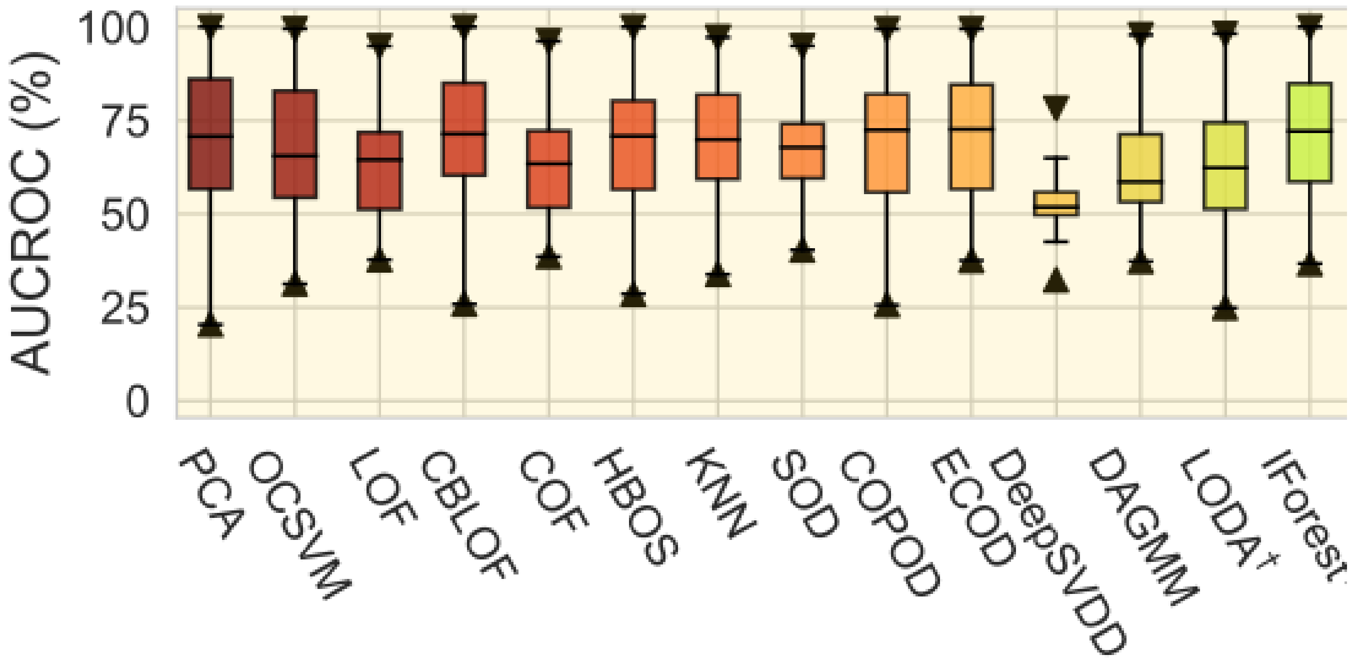
Introduction

- The volume of spectra available to modern astronomers means it's a challenge to find anomalies
- Unsupervised ML anomaly-detection approaches are used extensively, however, they can struggle with high dimensional data.

- A comprehensive benchmarking of anomaly detection methods available on **PyOD** and **scikit-learn** models was carried out by Han et al.(2022)(1)
- 14 different unsupervised algorithms were tested against 57 benchmark datasets

Credit: Han et al, 2022

- None was statistically the best and performance against a particular dataset was highly model dependent – "No Free Lunch"

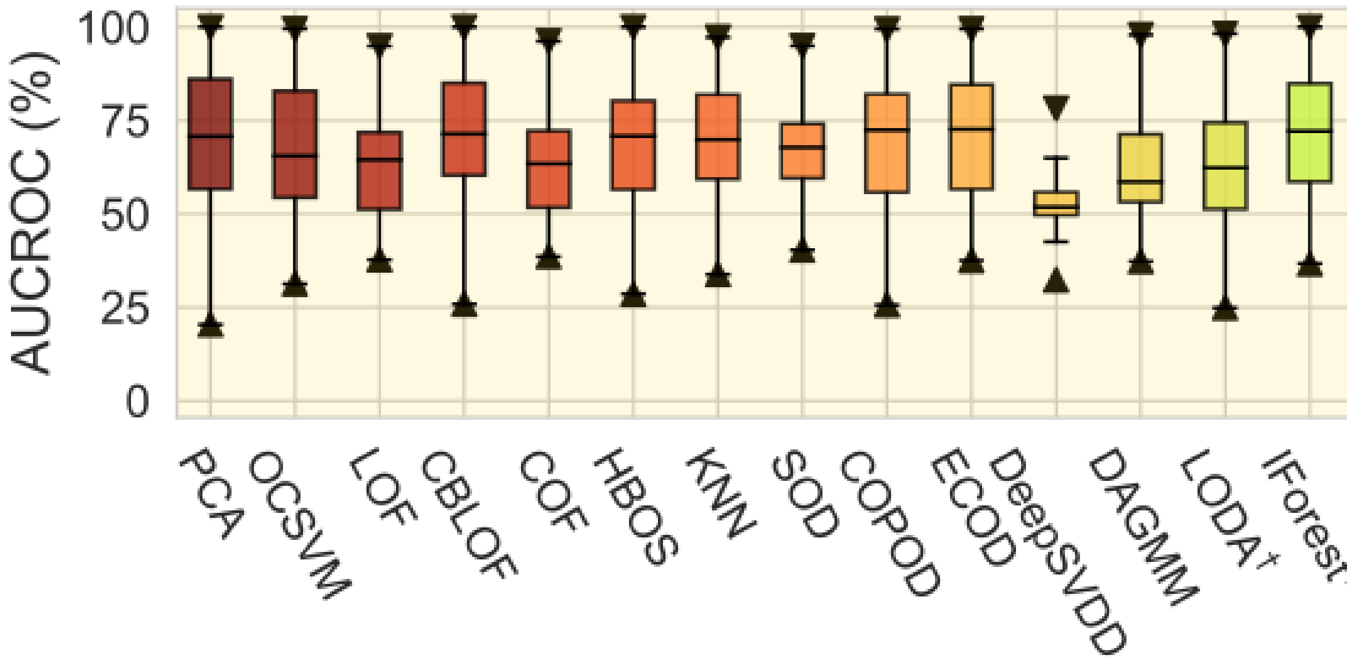


Introduction

- The volume of spectra available to modern astronomers means it's a challenge to find anomalies
- Unsupervised ML anomaly-detection approaches are used extensively, however, they can struggle with high dimensional data.

- A comprehensive benchmarking of anomaly detection methods available on **PyOD** and **scikit-learn** models was carried out by Han et al.(2022)[(1)]
- 14 different unsupervised algorithms were tested against 57 benchmark datasets

Credit: Han et al, 2022



- None was statistically the best and performance against a particular dataset was highly model dependent – "No Free Lunch"

- None perform particularly well with high-D datasets (apart from with the stylized MNIST datasets)

Why is High-D a problem?

- Raw data representation is often high-D because of how the data is collected rather than reflecting underlying physics
  - e.g. raw DESI spectra ~7800 bins
  - e.g. image of galaxy captured in 10,000s of pixels

Why is High-D a problem?

- Raw data representation is often high-D because of how the data is collected rather than reflecting underlying physics
  - e.g. raw DESI spectra ~7800 bins
  - e.g. image of galaxy captured in 10,000s of pixels

- This unnatural high-D representation lends us no physical intuition, is hard or impossible to work with computationally, and means we are afflicted by the **curse of dimensionality**[(2)]

Why is High-D a problem?

- Raw data representation is often high-D because of how the data is collected rather than reflecting underlying physics
    - e.g. raw DESI spectra ~7800 bins
    - e.g. image of galaxy captured in 10,000s of pixels

- This unnatural high-D representation lends us no physical intuition, is hard or impossible to work with computationally, and means we are afflicted by the **curse of dimensionality**[(2)]

- In high-D data becomes sparse as distances between points increases and points are found more towards the boundary of the space

- Conversely the distribution of these distances becomes tighter spread

- Concept of nearest neighbours – fundamental to many anomaly detection techniques – falls apart

Why is High-D a problem?

- Raw data representation is often high-D because of how the data is collected rather than reflecting underlying physics
  - e.g. raw DESI spectra ~7800 bins
  - e.g. image of galaxy captured in 10,000s of pixels

- This unnatural high-D representation lends us no physical intuition, is hard or impossible to work with computationally, and means we are afflicted by the **curse of dimensionality**[(2)]

- In high-D data becomes sparse as distances between points increases and points are found more towards the boundary of the space

- Conversely the distribution of these distances becomes tighter spread

- Concept of nearest neighbours – fundamental to many anomaly detection techniques – falls apart

- **Dimensionality reduction** should be beneficial but:
  - How do we do it and is it meaningful?

Dimensionality Reduction and the Manifold

- Dimensionality reduction relies on the Manifold Hypothesis, i.e. that most real-world high-D datasets reside close to a lower-D manifold.

    - An m-dimensional **manifold** is part of n-dimensional space (m<n) that locally resembles an m-dimensional hyperplane

Dimensionality Reduction and the Manifold

- Dimensionality reduction relies on the Manifold Hypothesis, i.e. that most real-world high-D datasets reside close to a lower-D manifold.
    - An m-dimensional **manifold** is part of n-dimensional space (m<n) that locally resembles an m-dimensional hyperplane
- This seems be valid for spectra, e.g. Yip et al. (2004)[(3)] and Portillo et al. (2020)[(4)]

Dimensionality Reduction and the Manifold

- Dimensionality reduction relies on the Manifold Hypothesis, i.e. that most real-world high-D datasets reside close to a lower-D manifold.

    - An m-dimensional **manifold** is part of n-dimensional space (m<n) that locally resembles an m-dimensional hyperplane

- This seems be valid for spectra, e.g. Yip et al. (2004)[(3)] and Portillo et al. (2020)[(4)]

- We can find low-D manifold using:
    - linear projection e.g. PCA or
    - non-linear manifold learning: e.g. t-SNE, Local Linear Embedding, AE, VAE

Dimensionality Reduction and the Manifold

- Dimensionality reduction relies on the Manifold Hypothesis, i.e. that most real-world high-D datasets reside close to a lower-D manifold.
    - An m-dimensional **manifold** is part of n-dimensional space (m<n) that locally resembles an m-dimensional hyperplane
- This seems be valid for spectra, e.g. Yip et al. (2004)[(3)] and Portillo et al. (2020)[(4)]
- We can find low-D manifold using:
    - linear projection e.g. PCA or
    - non-linear manifold learning: e.g. t-SNE, Local Linear Embedding, AE, VAE
- <u>**But the manifold we find will be different depending on the model we use**</u>
- Linear methods will find a hyperplane; non-linear methods can find more complex shaped manifolds

How does the Manifold affect Anomaly Detection?

- Two main ways of identifying outliers :

How does the Manifold affect Anomaly Detection?

- Two main ways of identifying outliers :
    - 1) Look for points which are isolated or extreme in the distribution because of distance from neighbouring points, or because of the relative under-density of surrounding points
        - in high-D this is problematic so we usually transform to low-D and look for extreme and isolated points on the low-D manifold instead: **ON-MANIFOLD ANOMALIES**

How does the Manifold affect Anomaly Detection?

- Two main ways of identifying outliers :
  - 1) Look for points which are isolated or extreme in the distribution because of distance from neighbouring points, or because of the relative under-density of surrounding points
    - in high-D this is problematic so we usually transform to low-D and look for extreme and isolated points on the low-D manifold instead: **ON-MANIFOLD ANOMALIES**
  - 2) Construct a low-D manifold which well represents the bulk of the dataset. Points far from this manifold (high reconstruction error) are assumed to be outliers: **OFF-MANIFOLD ANOMALIES**

How does the Manifold affect Anomaly Detection?

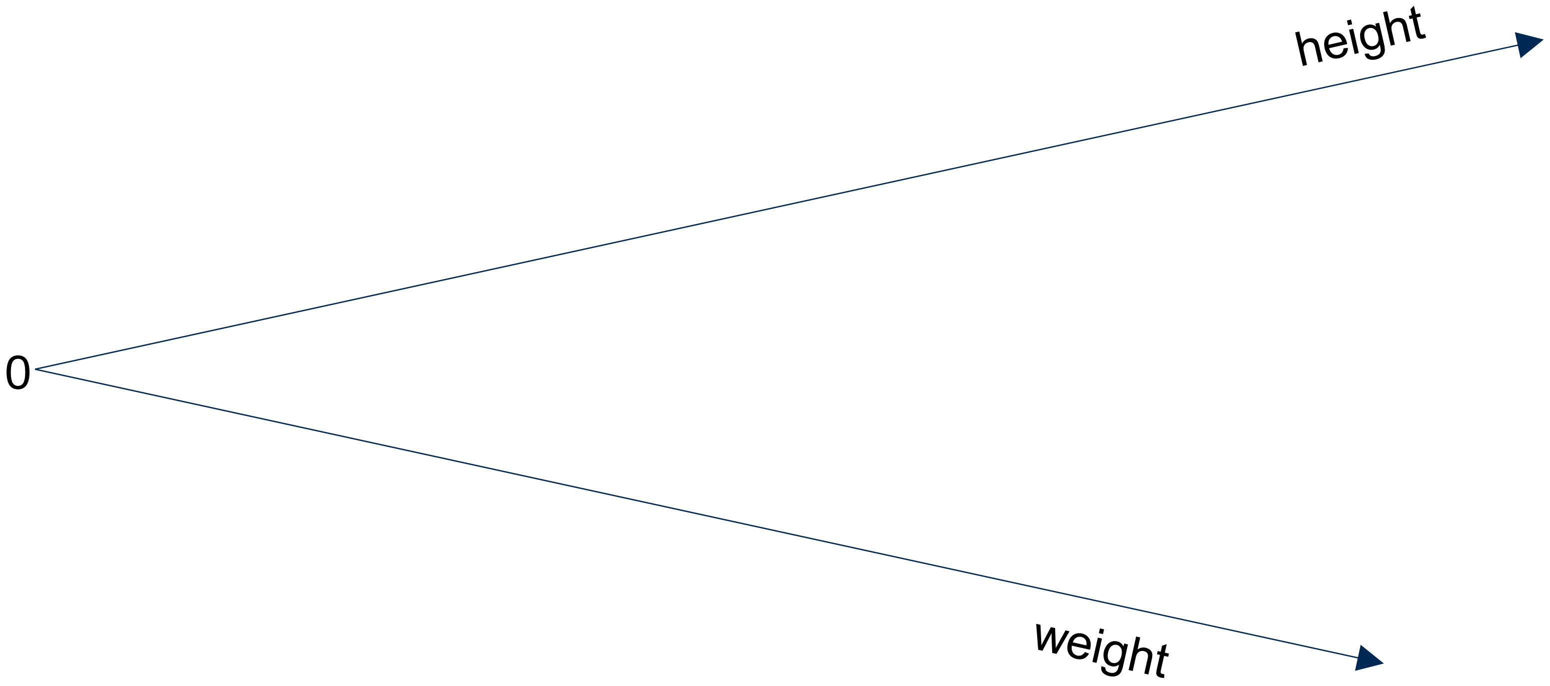
- Two main ways of identifying outliers :
  - 1) Look for points which are isolated or extreme in the distribution because of distance from neighbouring points, or because of the relative under-density of surrounding points
    - in high-D this is problematic so we usually transform to low-D and look for extreme and isolated points on the low-D manifold instead: **ON-MANIFOLD ANOMALIES**
  - 2) Construct a low-D manifold which well represents the bulk of the dataset. Points far from this manifold (high reconstruction error) are assumed to be outliers: **OFF-MANIFOLD ANOMALIES**
- The manifold is model dependent – **therefore the anomalies detected will also be model dependent**
- Furthermore **<u>on-manifold and off-manifold anomalies are likely to be quite different</u>**

How does the Manifold affect Anomaly Detection?

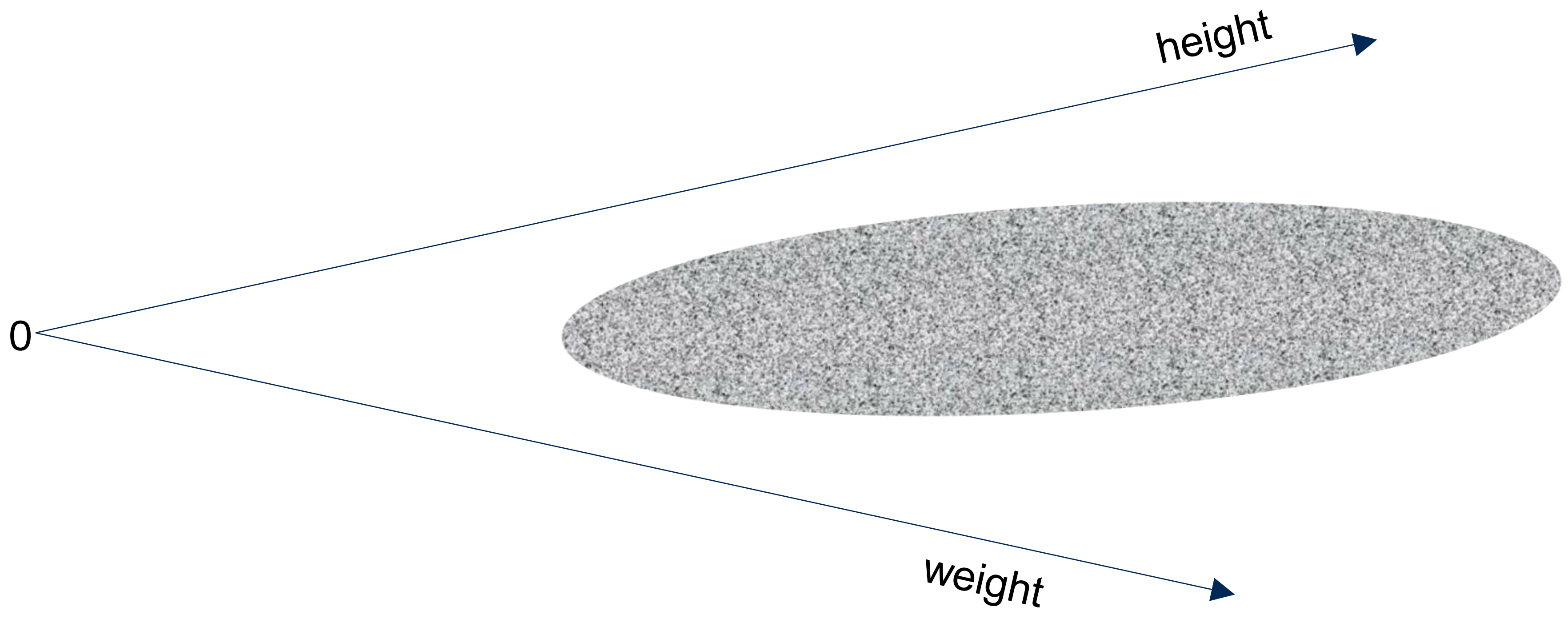
- Two main ways of identifying outliers :
  - 1) Look for points which are isolated or extreme in the distribution because of distance from neighbouring points, or because of the relative under-density of surrounding points
    - in high-D this is problematic so we usually transform to low-D and look for extreme and isolated points on the low-D manifold instead: **ON-MANIFOLD ANOMALIES**
  - 2) Construct a low-D manifold which well represents the bulk of the dataset. Points far from this manifold (high reconstruction error) are assumed to be outliers: **OFF-MANIFOLD ANOMALIES**
- The manifold is model dependent – **therefore the anomalies detected will also be model dependent**
- Furthermore **<u>on-manifold and off-manifold anomalies are likely to be quite different</u>**
- If we've constructed the manifold well then
  - on-manifold outliers ≈ extremes in current thinking
  - off-manifold outliers ≈ new or rare physics

Instrumentation artefacts

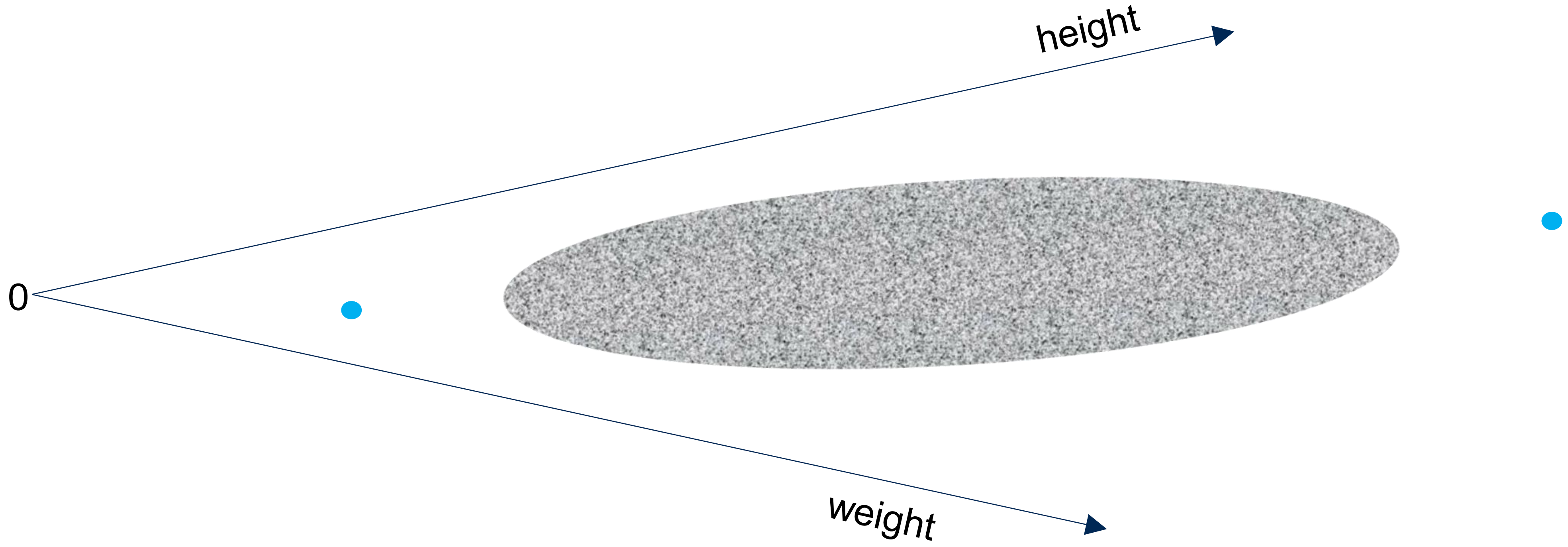
3D → 2D example: Horses



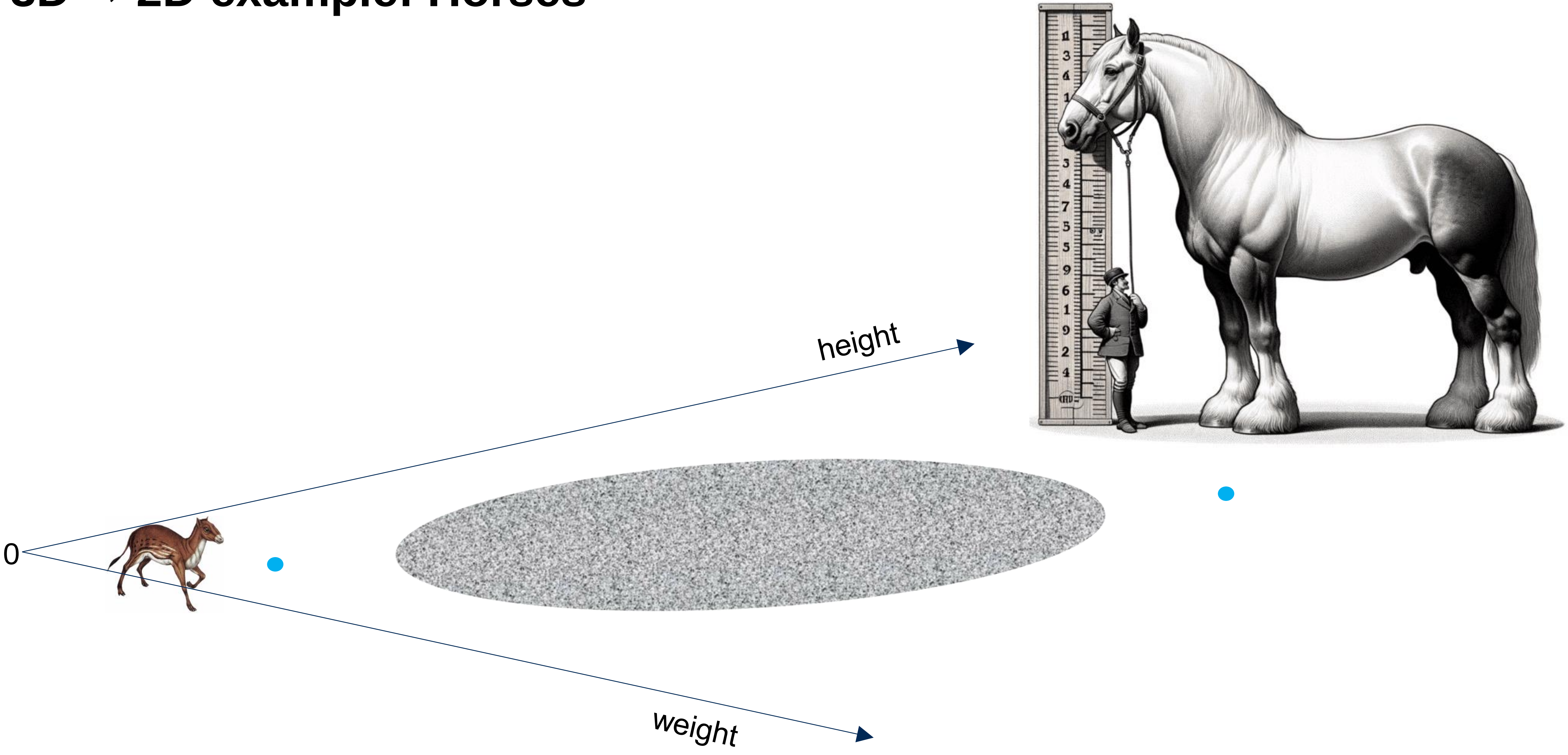
3D → 2D example: Horses



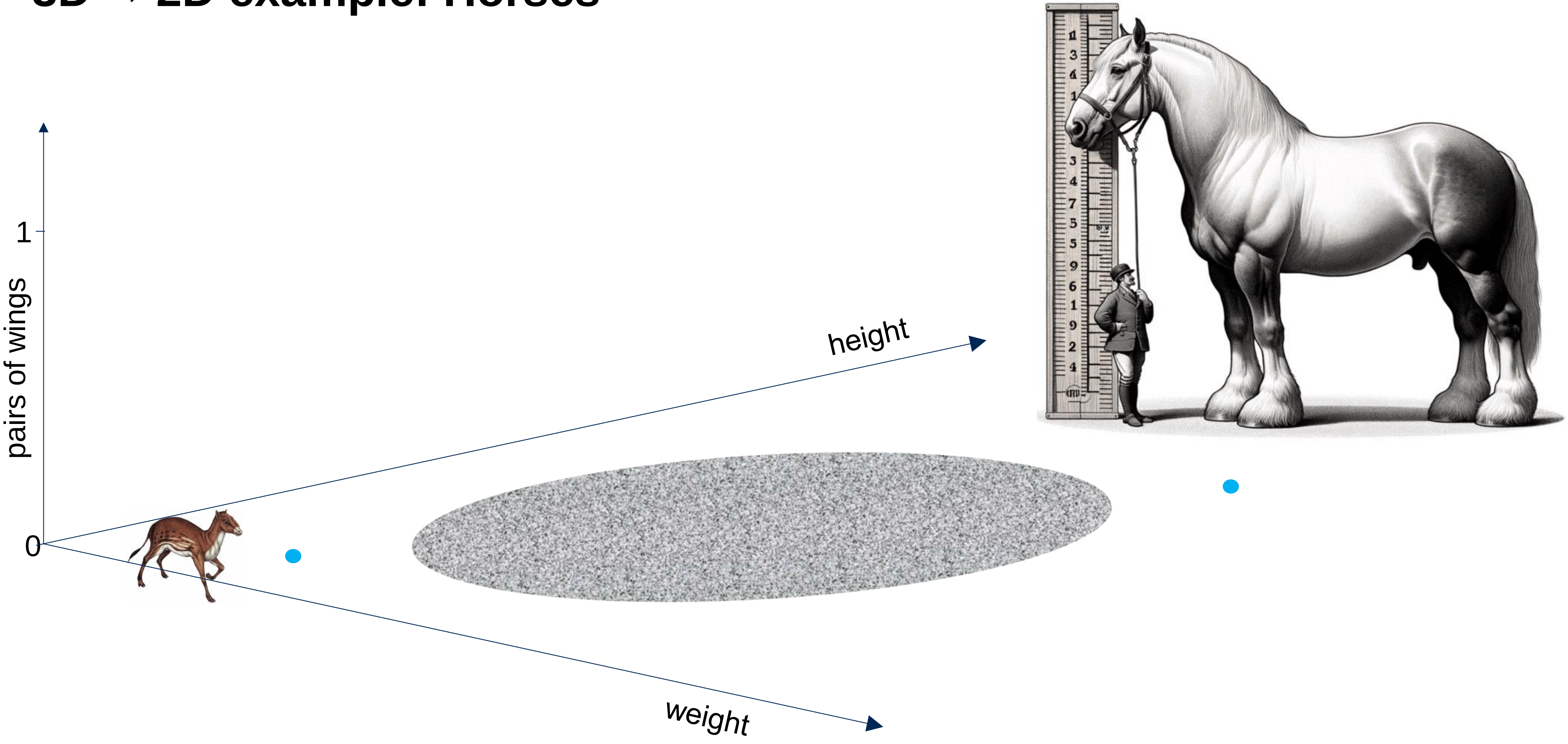
3D → 2D example: Horses



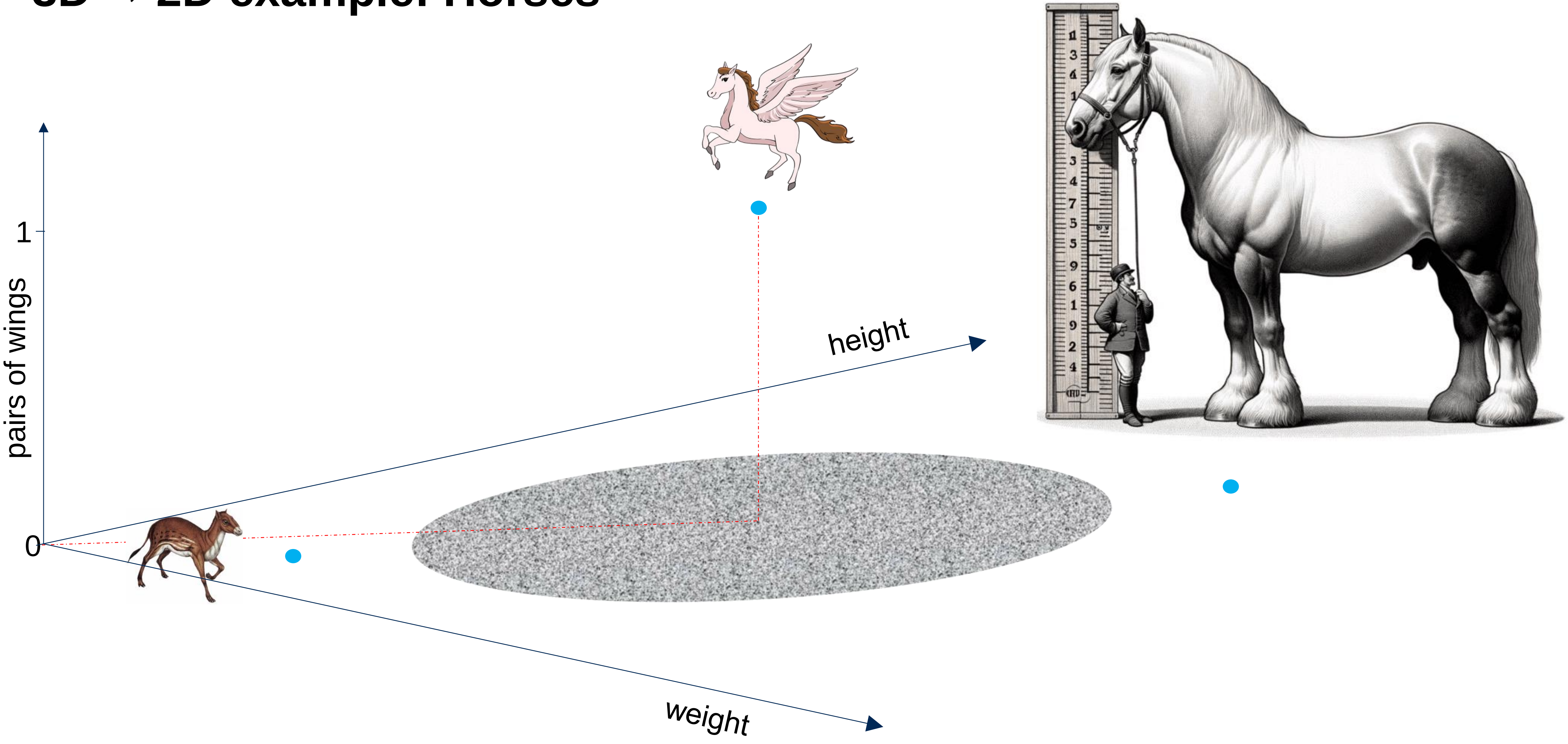
3D → 2D example: Horses



3D → 2D example: Horses

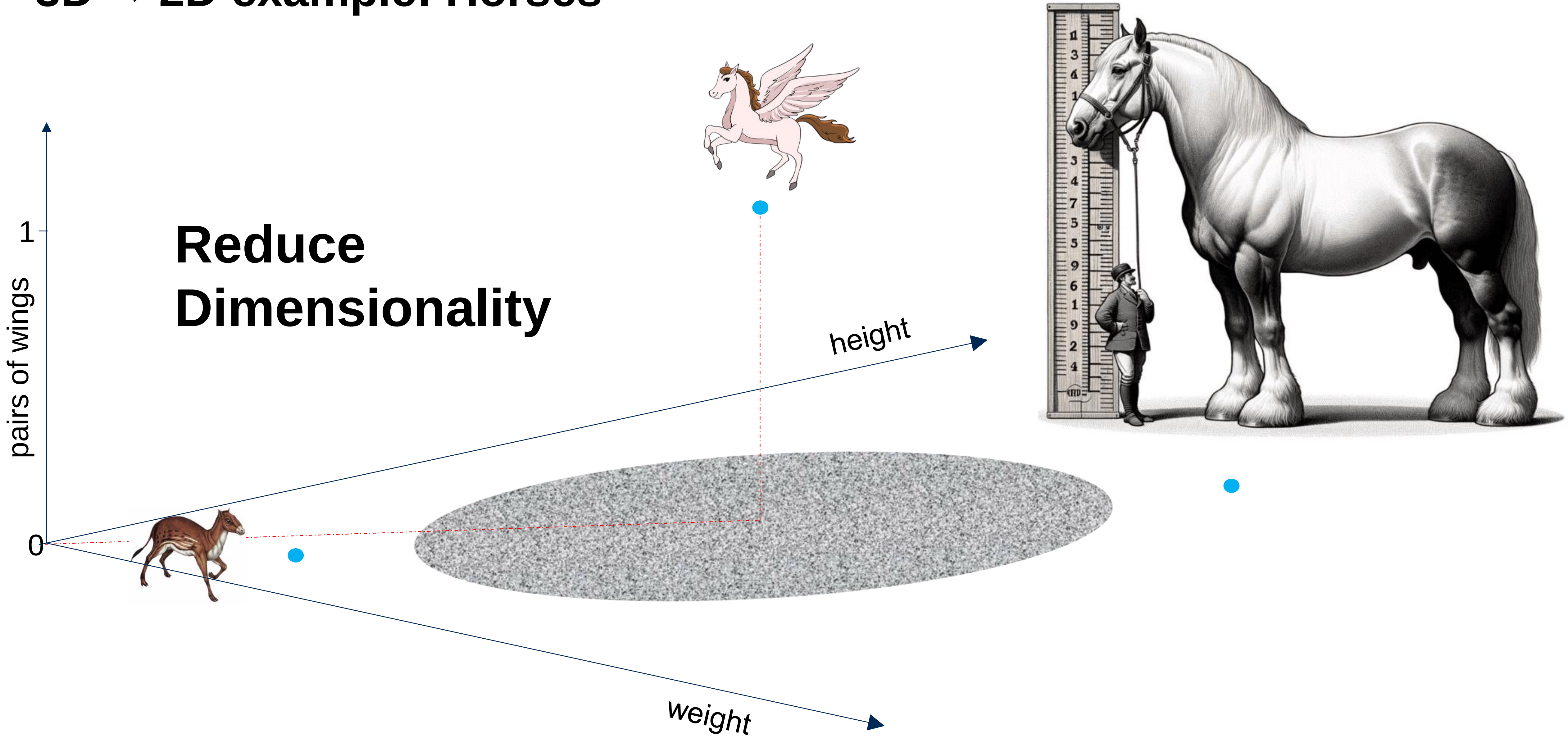


3D → 2D example: Horses

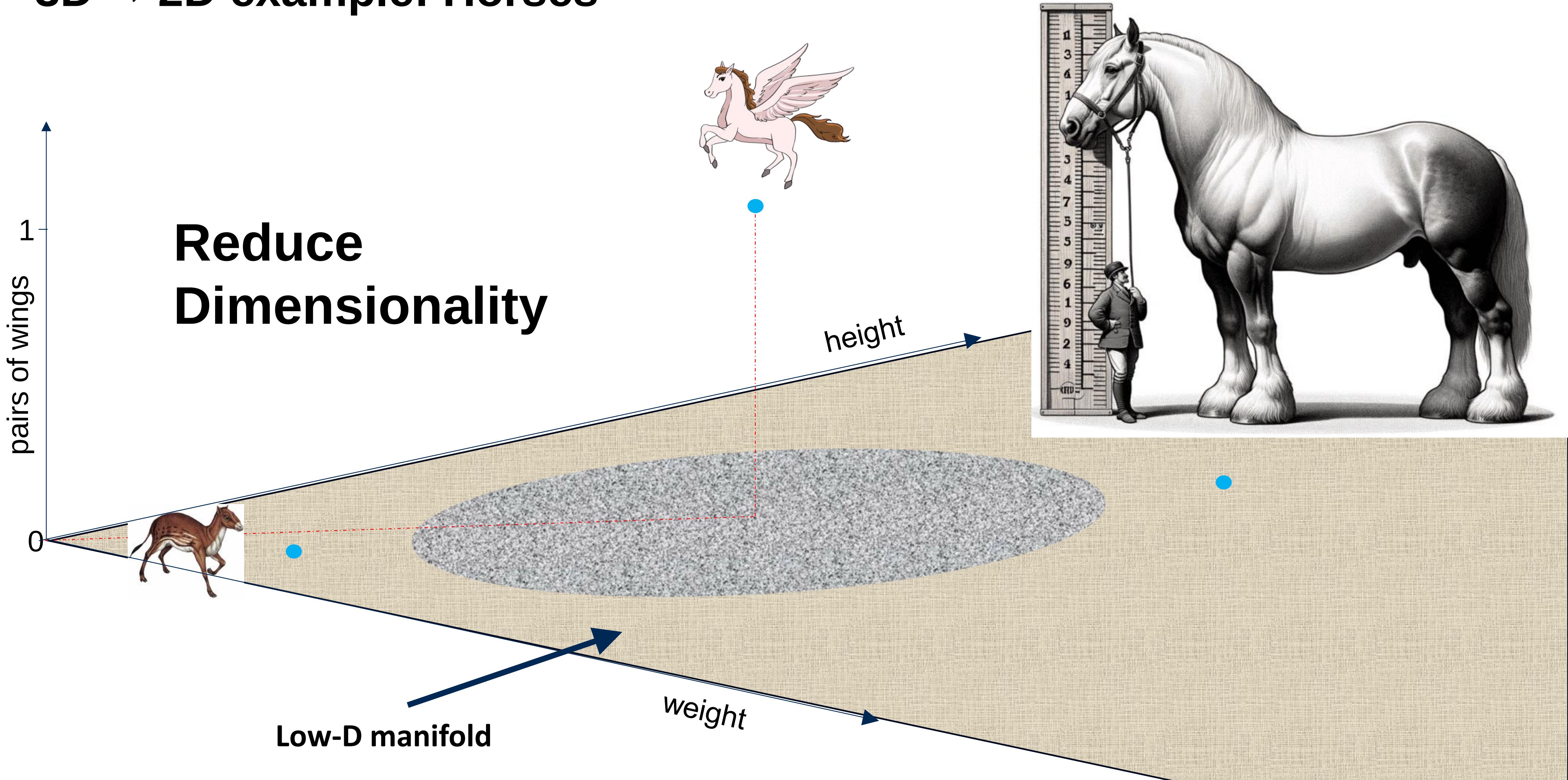


3D → 2D example: Horses

Reduce Dimensionality

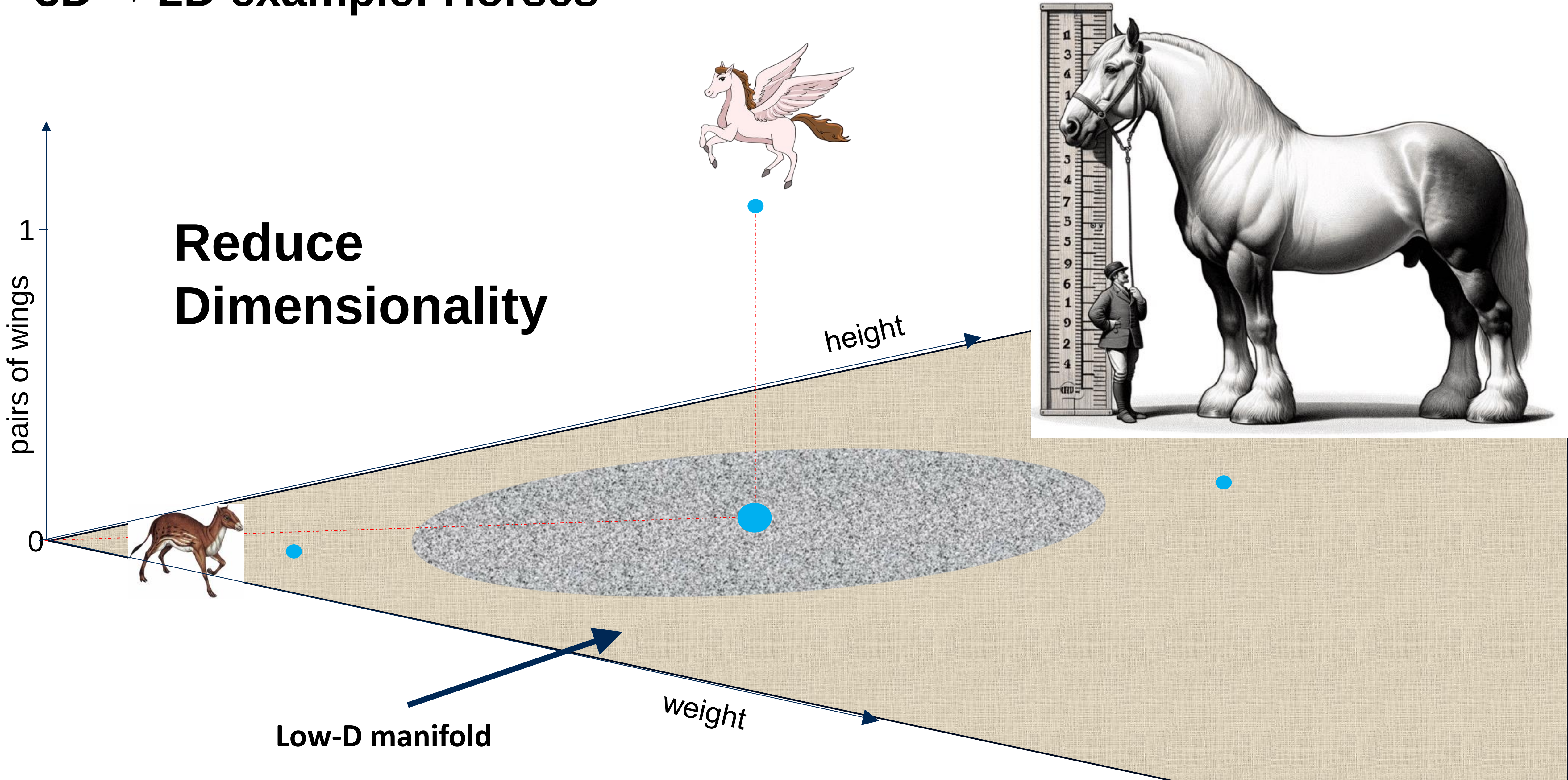


3D → 2D example: Horses



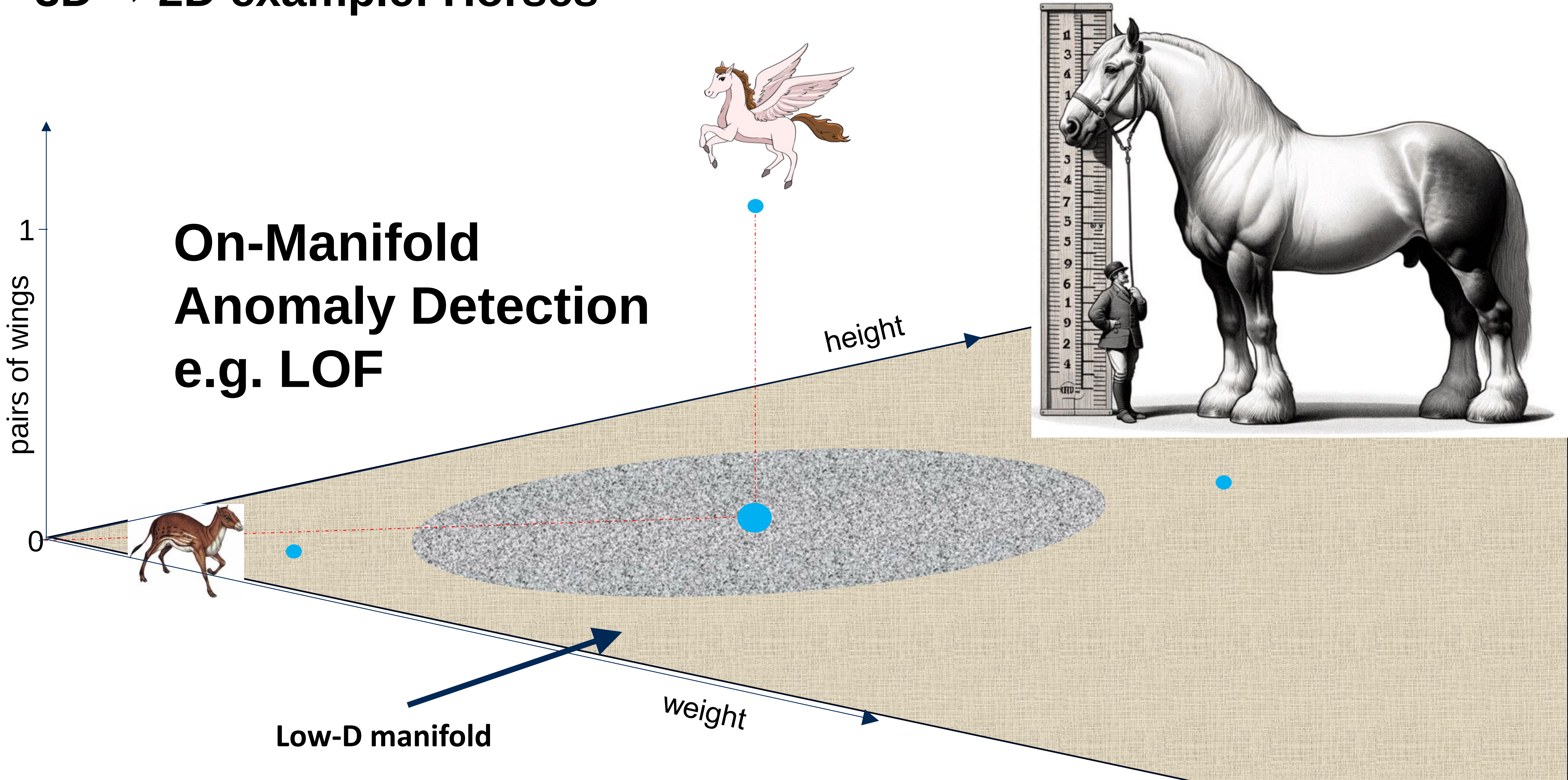
3D → 2D example: Horses

Reduce Dimensionality



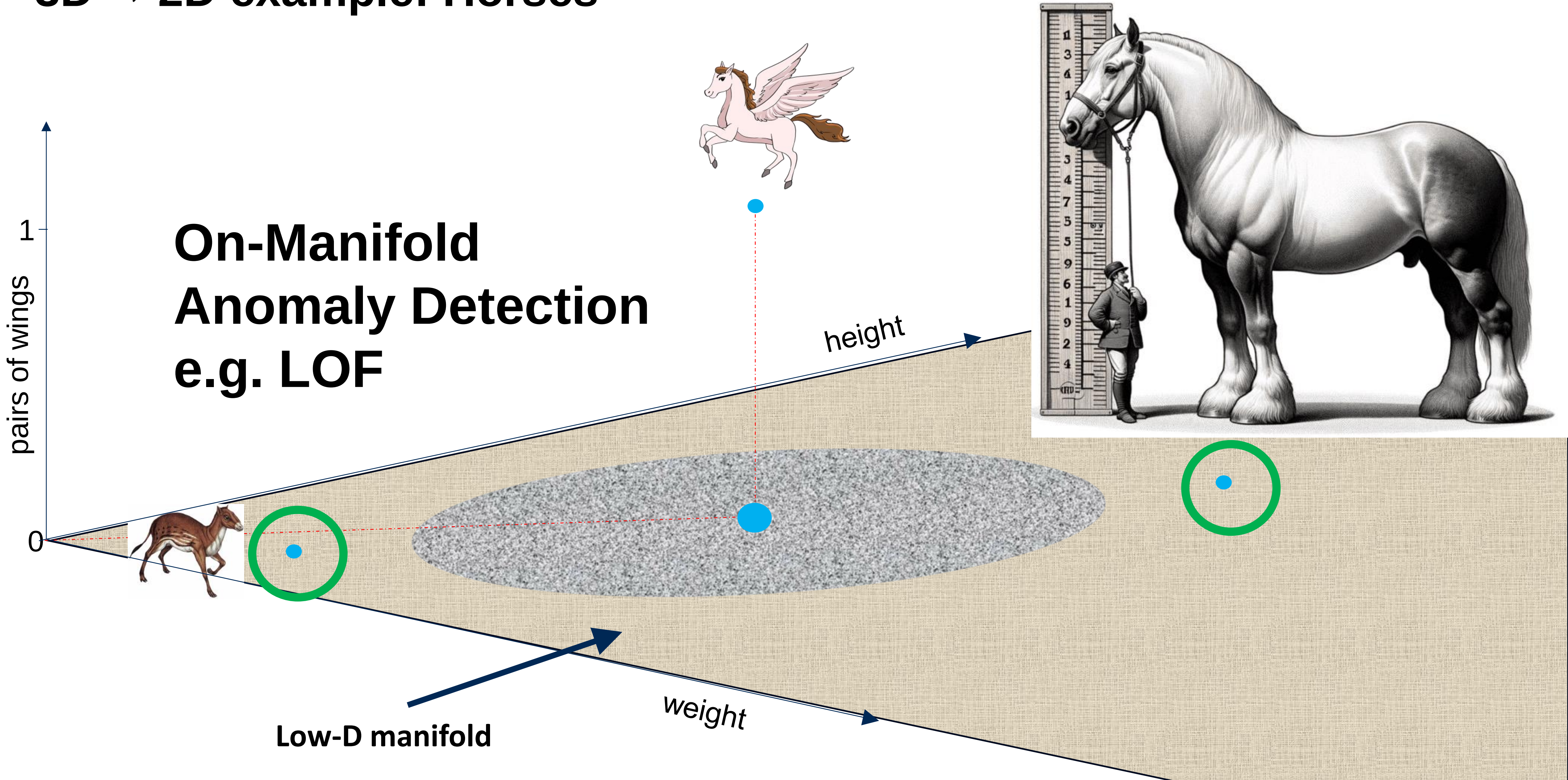
3D → 2D example: Horses

On-Manifold
Anomaly Detection
e.g. LOF



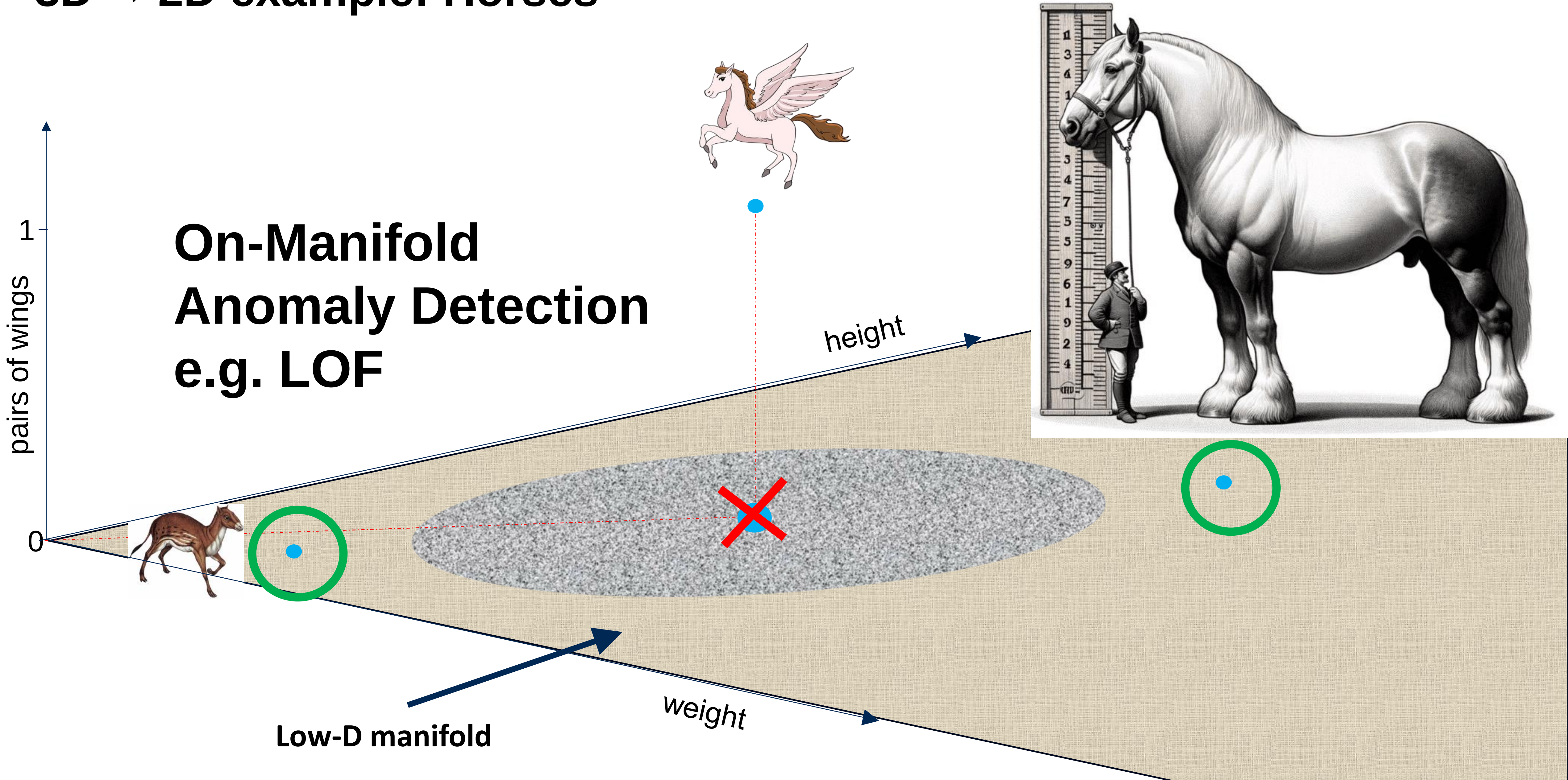
3D → 2D example: Horses

On-Manifold Anomaly Detection e.g. LOF



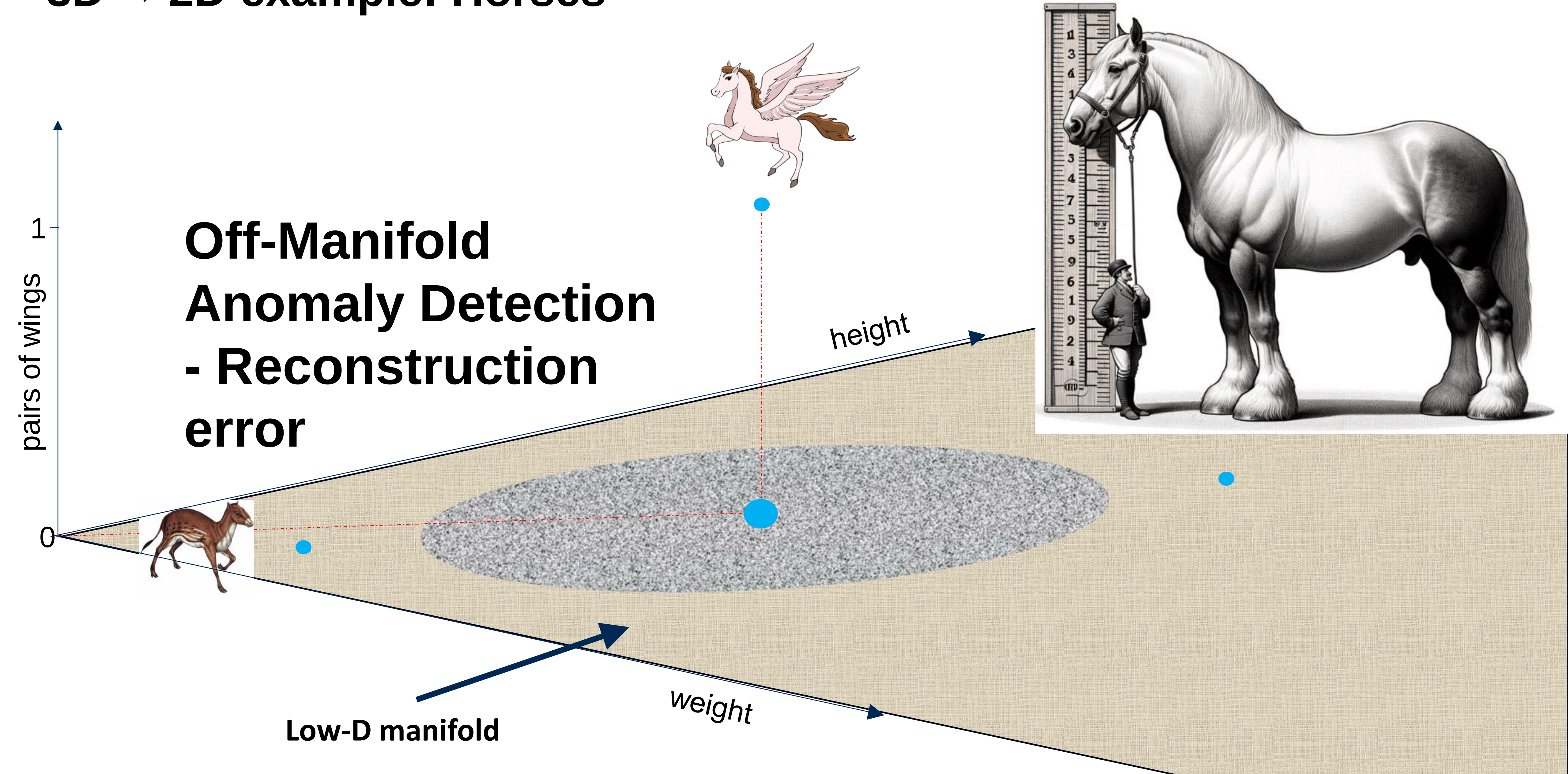
3D → 2D example: Horses

On-Manifold
Anomaly Detection
e.g. LOF



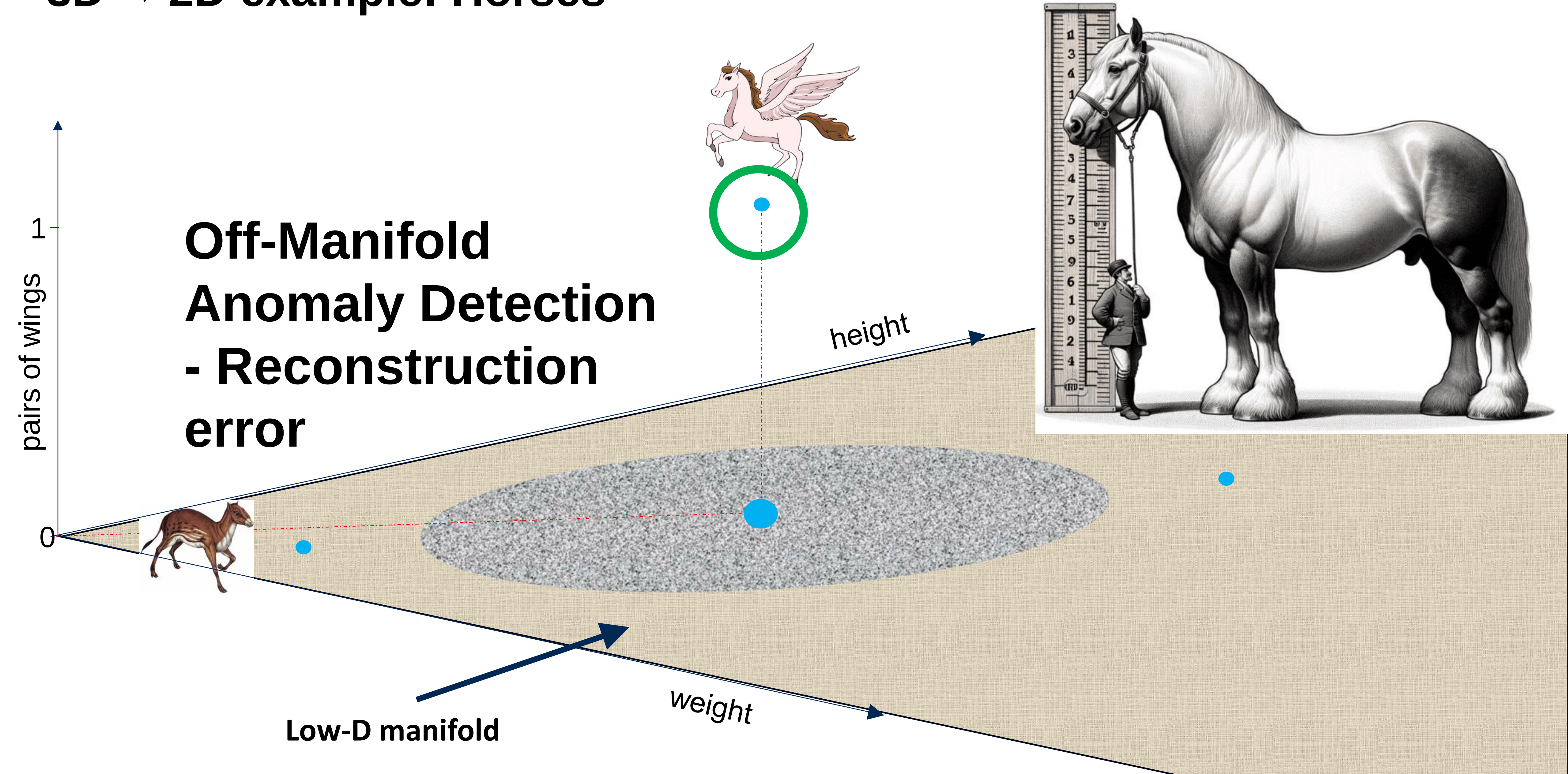
3D → 2D example: Horses

Off-Manifold Anomaly Detection - Reconstruction error

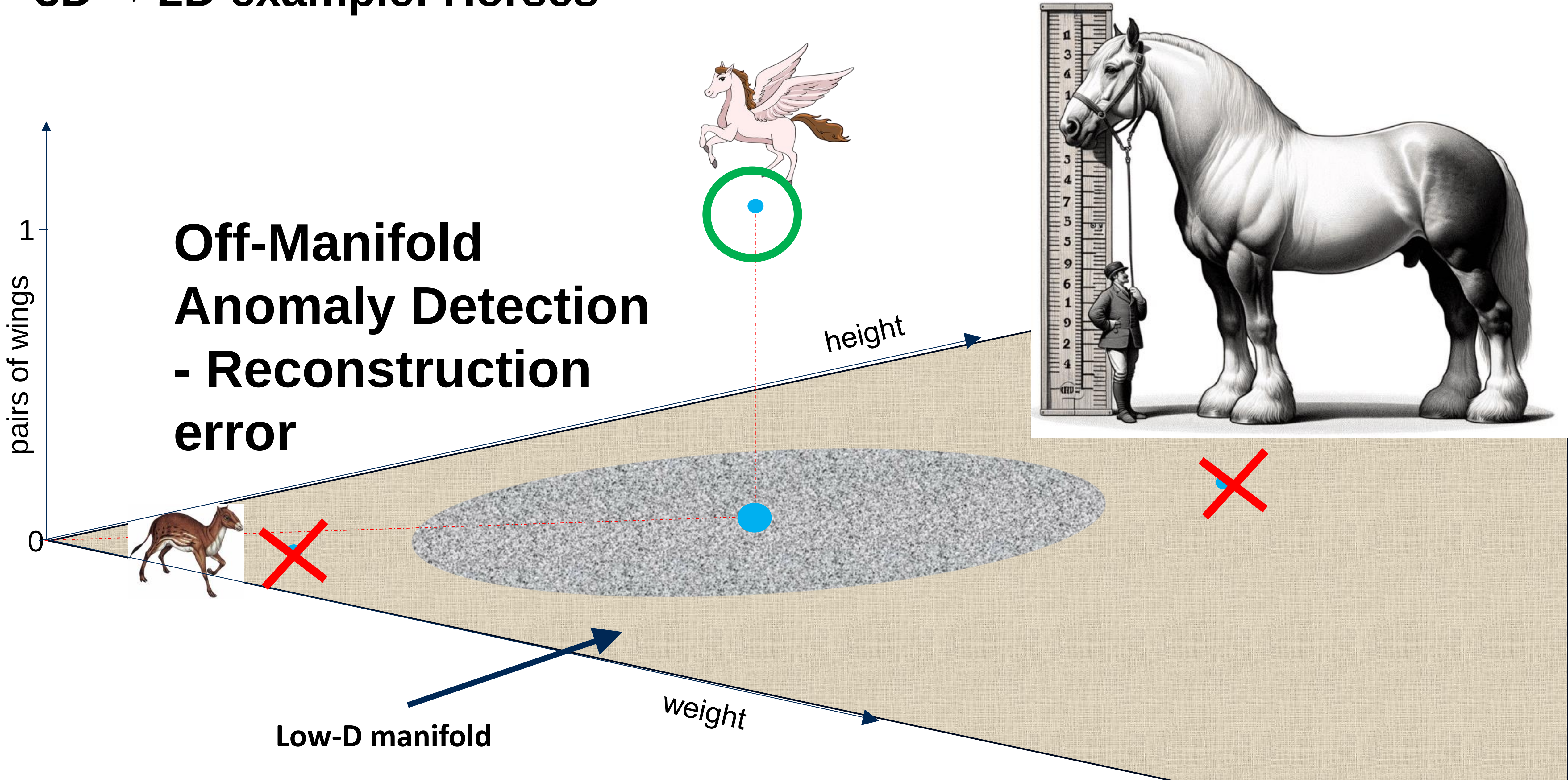


3D → 2D example: Horses

Off-Manifold Anomaly Detection - Reconstruction error

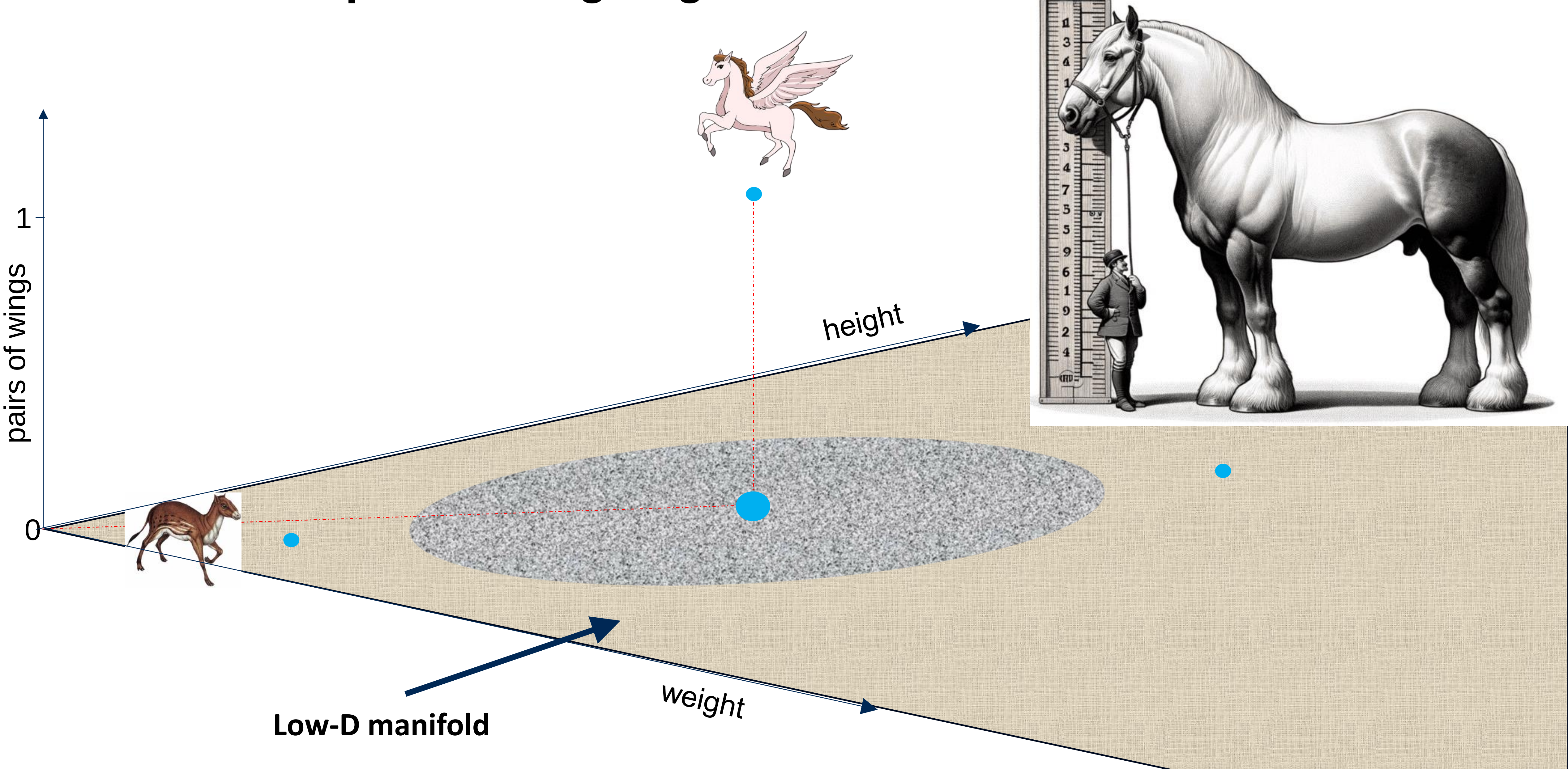


3D → 2D example: Horses

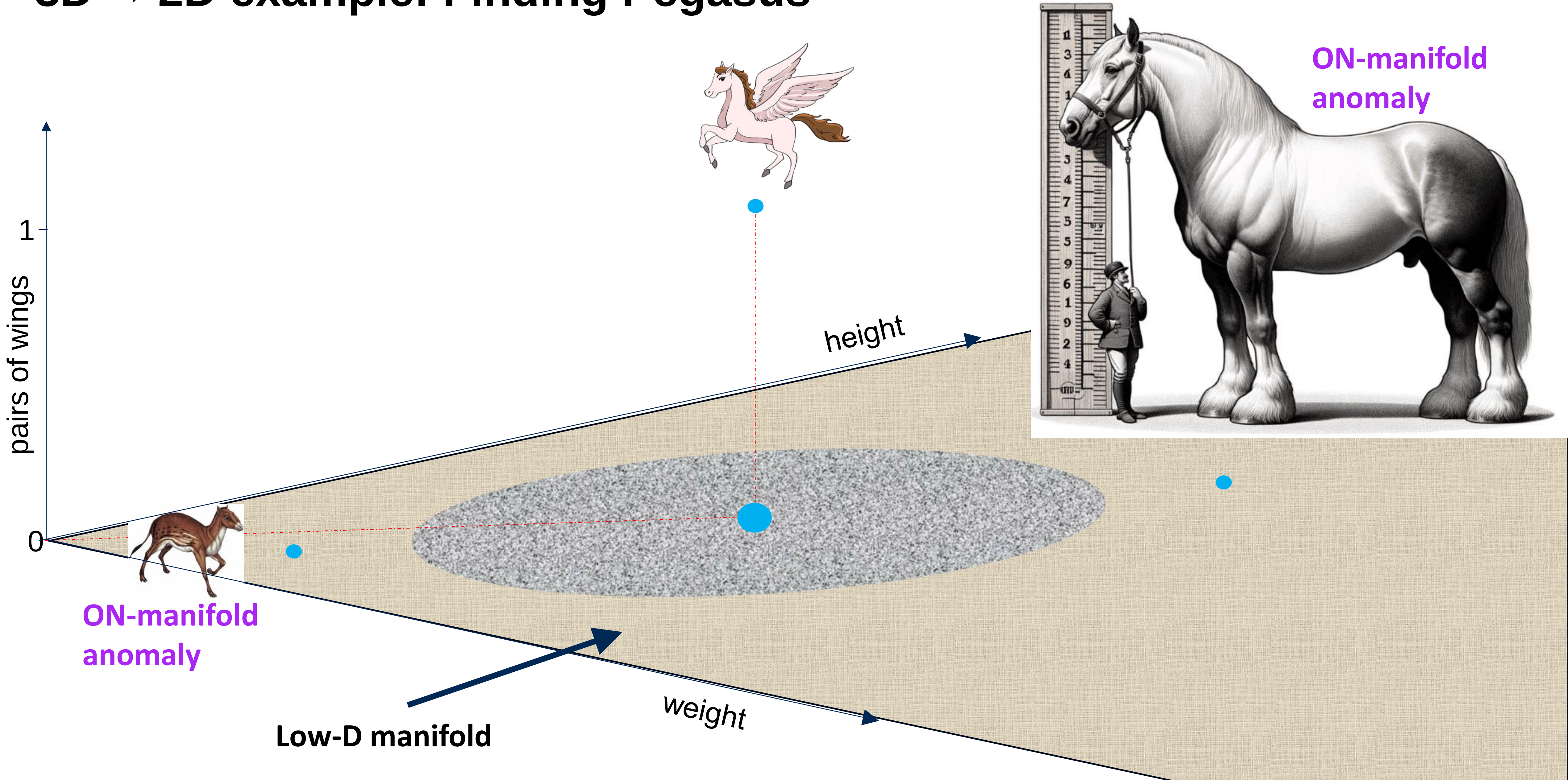


Off-Manifold Anomaly Detection - Reconstruction error

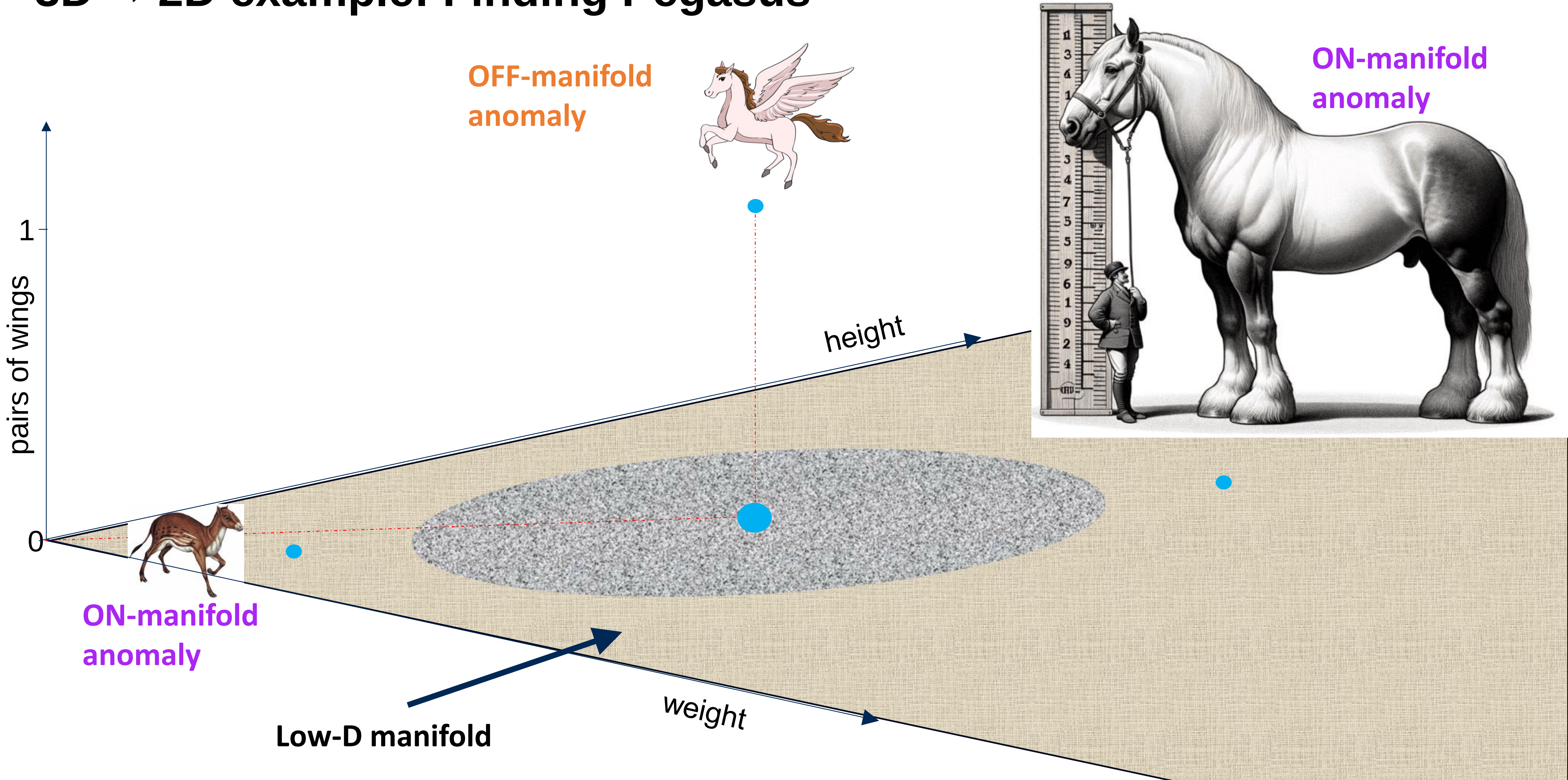
3D → 2D example: Finding Pegasus



3D → 2D example: Finding Pegasus

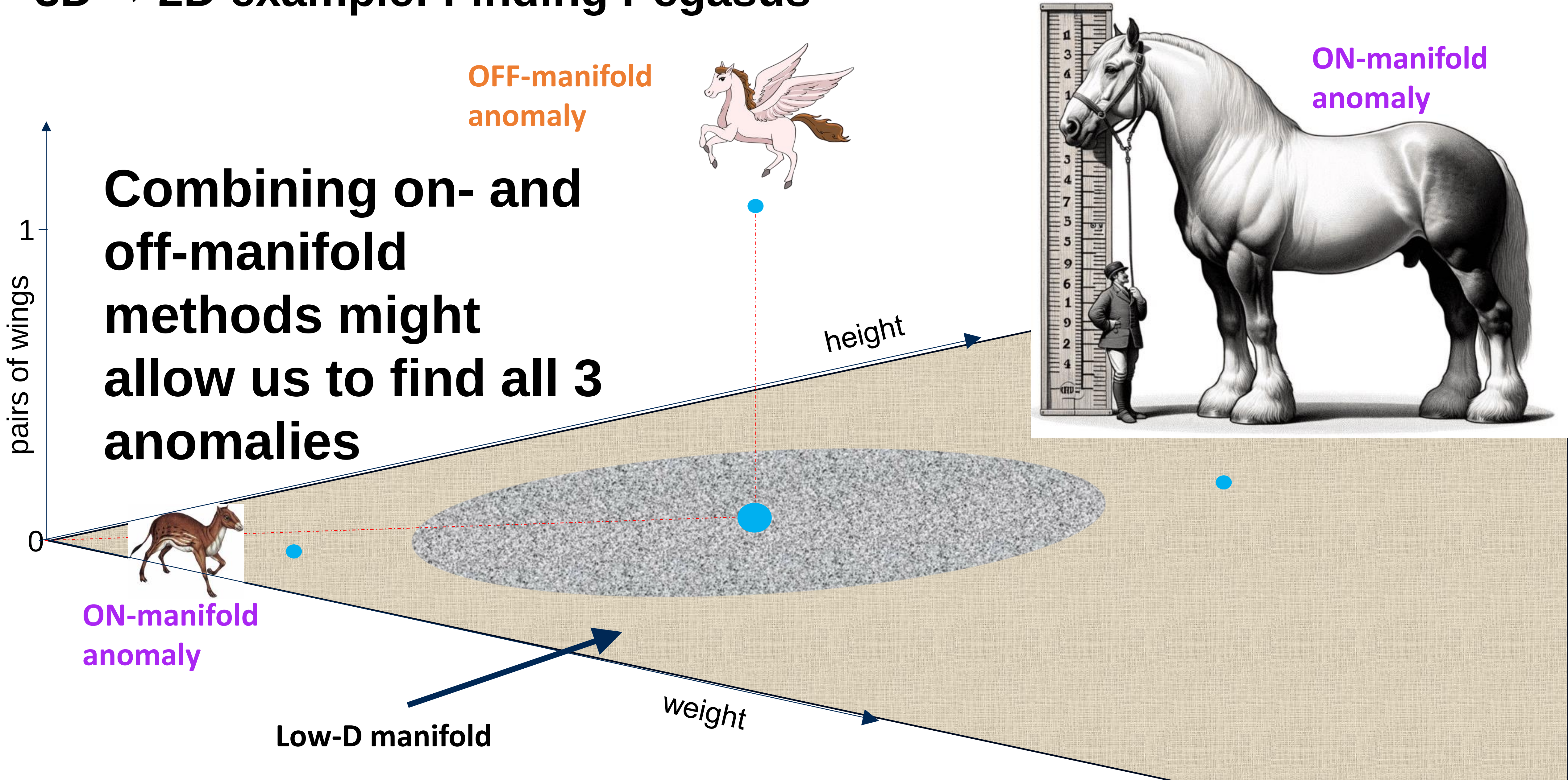


3D → 2D example: Finding Pegasus



3D → 2D example: Finding Pegasus

Combining on- and off-manifold methods might allow us to find all 3 anomalies



~7800D → 6D example: DESI Spectra from BGS

- Data is from the DESI Bright Galaxy Sample – Iron Datataset DR1 – “good spectra”
- Normalised, downsampled ~x5 and deredshifted. Minimal other preprocessing so far.
- ~55,000 spectra split ~26,000 Stars and ~19,000 Galaxies based on DESI target type

~7800D → 6D example: DESI Spectra from BGS

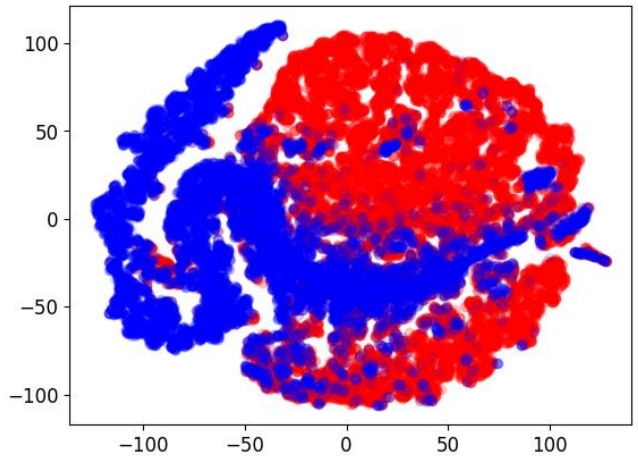
- Data is from the DESI Bright Galaxy Sample – Iron Datataset DR1 – “good spectra”
- Normalised, downsampled ~x5 and deredshifted. Minimal other preprocessing so far.
- ~55,000 spectra split ~26,000 Stars and ~19,000 Galaxies based on DESI target type
- Used a variety of DR techniques to build a 6-D manifold from the original 7800 D data – PCA, a linear AE, and AE
- Then visualized the 6D → 2D using a manifold-learning technique tSNE.

~7800D → 6D example: DESI Spectra from BGS

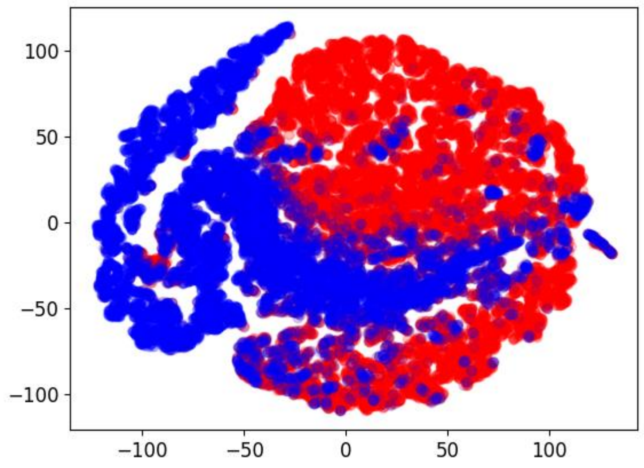
- Data is from the DESI Bright Galaxy Sample – Iron Datataset DR1 – “good spectra”
- Normalised, downsampled ~x5 and deredshifted. Minimal other preprocessing so far.
- ~55,000 spectra split ~26,000 Stars and ~19,000 Galaxies based on DESI target type
- Used a variety of DR techniques to build a 6-D manifold from the original 7800 D data – PCA, a linear AE, and AE
- Then visualized the 6D → 2D using a manifold-learning technique tSNE.

STARS
GALAXIES

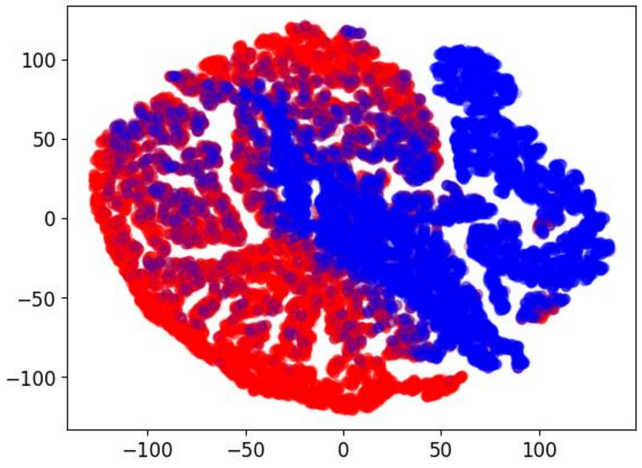
PCA



Linear AE ≅ PCA



AE



- (Note the unsupervised classification – separation between stars and galaxies)
- We can see how the (2D visualisations of the) low-D manifolds are **model-dependent**

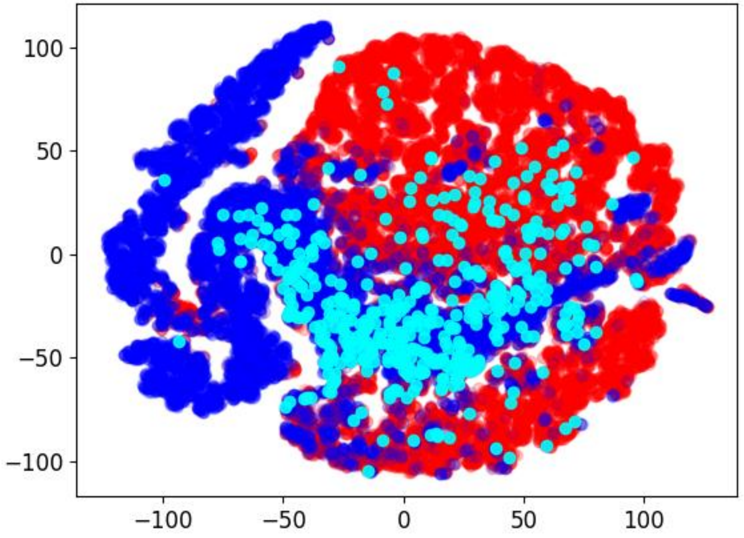
~7800D → 6D example: DESI Spectra from BGS – PCA manifold

- Take the PCA-generated manifold and look both off- and on-manifold for outlying points
- Identify 1% of total population as outliers under both methods

~7800D → 6D example: DESI Spectra from BGS – PCA manifold

- Take the PCA-generated manifold and look both off- and on-manifold for outlying points
- Identify 1% of total population as outliers under both methods

PCA



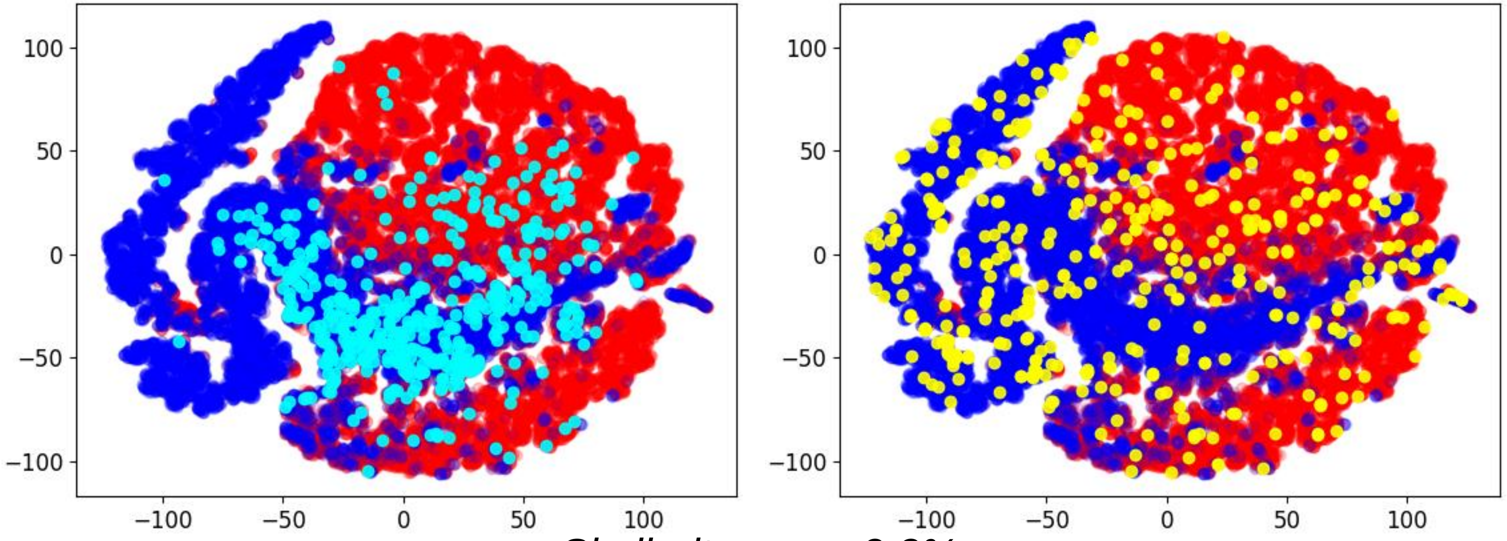
OFF-manifold

STARS
GALAXIES

~7800D → 6D example: DESI Spectra from BGS – PCA manifold

- Take the PCA-generated manifold and look both off- and on-manifold for outlying points
- Identify 1% of total population as outliers under both methods

PCA **LOF**



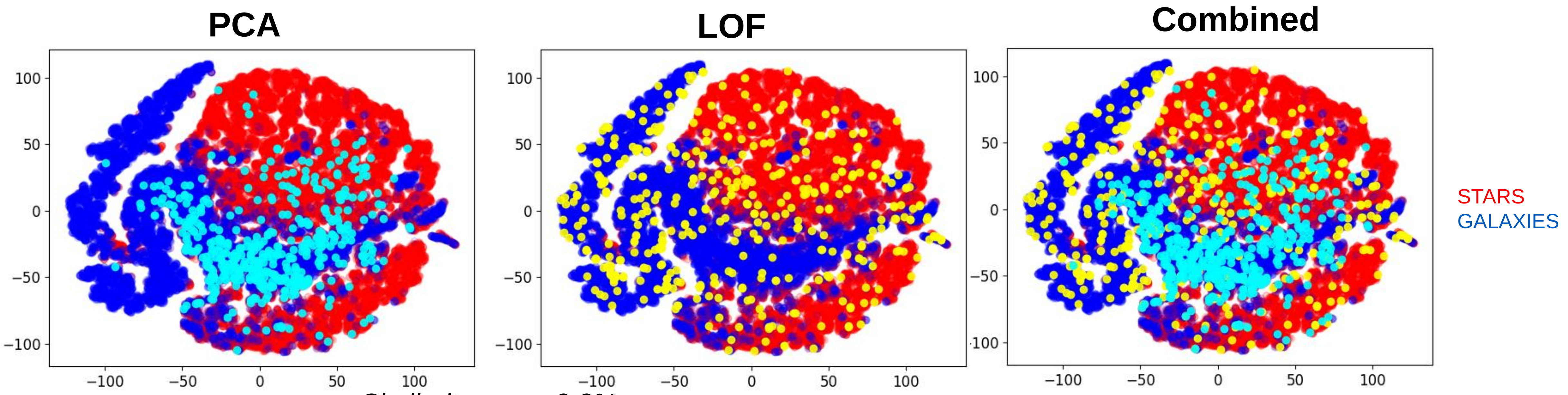
Similarity score:2.3%

OFF-manifold **ON-manifold**

STARS
GALAXIES

~7800D → 6D example: DESI Spectra from BGS – PCA manifold

- Take the PCA-generated manifold and look both off- and on-manifold for outlying points
- Identify 1% of total population as outliers under both methods



Similarity score:2.3%

OFF-manifold

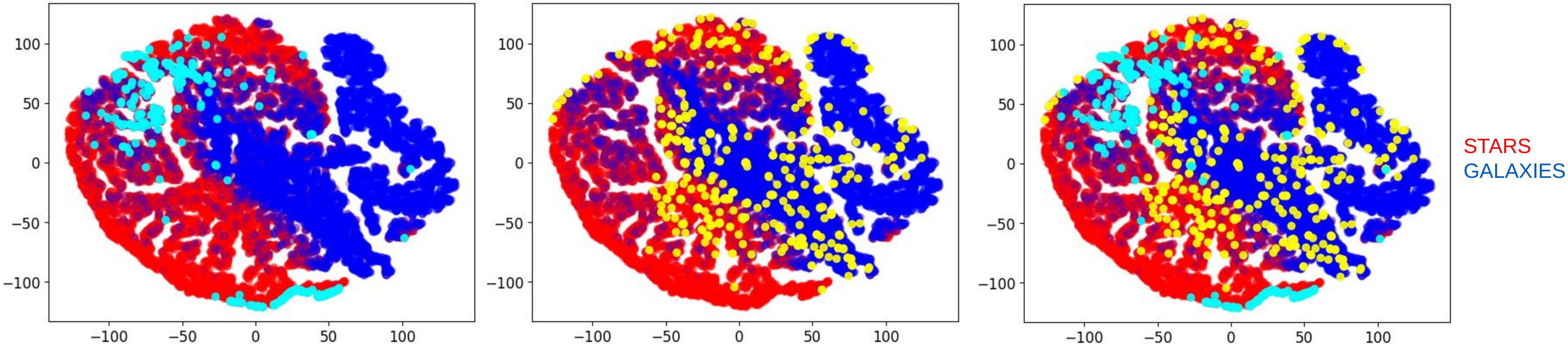
ON-manifold

- By combining an on- and off-manifold method we should be able to detect more anomalies

~7800D → 6D example: DESI Spectra from BGS – AE manifold

- Take the AE-generated manifold and look off- and on-manifold for outlying points
- Identify 1% of total population as outliers under both methods

AE **LOF** **Combined**



Similarity score: 2.2%

OFF-manifold **ON-manifold**

- By combining an on- and off-manifold method we should be able to detect more anomalies

Takeaways

- Unsupervised anomaly detection is model dependent
- **It is helpful to split techniques/anomalies between on- and off-manifold**. In general these will not produce the same result.
- For a given manifold, **combining complementary on- and off-manifold techniques should widen the number of anomalies we detect in high-D data**. Many of you are intuitively combining AD techniques already but we hope viewing the problem from the perspective of the manifold will inform these choices.
- This is very much a work in progress and we will also be looking to apply these ideas to bigger DESI datasets, more AD techniques, test the impact of different levels of preprocessing and also to test standard benchmark datasets in other domains
- **Please get in contact** ucaprpn@ucl.ac.uk if you want to discuss further any of the topics raised here.

Papers in preparation

DESI: Identifying Anomalous Spectra with Variational Autoencoders

Constantina Nicolaou,[1★] R.P. Nathan,[1] Ofer Lahav,[1] Antonella Palmese,[2] The DESI Collaboration
[1] *University College London, Gower St, London WC1E 6BT, UK*
[2] *Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, United States*

Accepted XXX. Received YYY; in original form ZZZ

PAPER IN PREPARaTION

ABSTRACT

The tens of millions of spectra being captured by the Dark Energy Spectroscopic Instrument (DESI) provide tremendous discovery potential. In this work we show how Machine Learning, in particular a Variational Autoencoder (VAE), can detect anomalies in a sample of approximately 200,000 DESI spectra comprising galaxies, quasars and stars. We demonstrate that the VAE can compress the dimensionality of a spectrum by x100, while still retaining enough information to accurately reconstruct spectral features. We then detect anomalous spectra in two ways: those with high reconstruction error and those which are isolated in the VAE latent representation. The anomalies identified fall into two broad categories: spectra with artefacts and spectra with unique physical features. Awareness of the former can help to improve the DESI spectroscopic pipeline; whilst the latter can lead to the identification of new and unusual objects. To further curate the list of outliers, we use Astronomaly which employs Active Learning to provide personalised outlier recommendations for visual inspection. In this work we also explore the VAE latent space and find that different object classes and sub-classes are separated despite being unlabelled. We demonstrate the interpretability of this latent space by identifying tracks within it that correspond to various spectral characteristics. For example, we find tracks that correspond to increasing star formation and increase in broad emission lines along the Balmer series. In upcoming work we will be applying the methods presented here to search for both systematics and astrophysically interesting objects in much larger datasets of DESI spectra.

Key words: techniques: spectroscopic – methods: statistical – methods: data analysis – galaxies: peculiar – [methods: machine

Finding Pegasus: Leveraging the Manifold from Machine-Learning Dimensionality-Reduction to Enhance Unsupervised Anomaly Detection in DESI Spectra

R. P. Nathan,[1★] O. Lahav,[1] and N. Nikolaou[1]
[1] *University College London, Gower St, London WC1E 6BT, UK*

Accepted XXX. Received YYY; in original form ZZZ

PAPER IN PREPARaTION

ABSTRACT

Large-scale surveys like DESI mean we live in a Golden Age when it comes to astronomical spectra. The sheer volume of spectra available, however, combined with their high-dimensional representation means it can be a challenge to find anomalous instances – be they instrumentation artefacts, rare objects or "unknown unknowns." Machine-Learning techniques have been used for a number of years to identify anomalies and are mostly well suited to the task of looking for anomalies at scale. Unsupervised anomaly-detection approaches have been used extensively, however, they can struggle with high-dimensional data. The purpose of this work is to highlight some of the issues key to the high-D data problem – usually thought of collectively as the Curse of Dimensionality – with a particular focus throughout on anomaly detection. In particular, we look at this problem from the perspective of the manifold that is created when dimensionality-reduction techniques are employed – either explicitly or implicitly – to get round the high-D problem. We will give illustrations – both simple and then using real DESI data – of what difference this manifold can make in practice and how it can bring significant model dependence to the set of anomalies detected. We discuss different unsupervised anomaly-detection techniques and introduce the terms on-manifold or off-manifold as a helpful way of categorizing them. We illustrate that by combining on- and off-manifold techniques, we might increase the number of anomalies detected – which will be of especial importance in recall-sensitive tasks. And we suggest that this might

- ”Identifying Anomalous Spectra with Variational Autoencoders”, **Constantina Nicolaou et al. [2024]**

- ”Finding Pegasus: Leveraging the Manifold from Machine-Learning Dimensionality-Reduction to Enhance Unsupervised Anomaly Detection in DESI Spectra”, **R.P. Nathan et al. [2024]**

Selective Bibliography

References

1. Han et al., 2022, “ADBench: Anomaly Fetection Benchmark”, 36th Conference on Neural Information Processing Systems
2. R. Bellman, 1957, “Dynamic programming”
3. Yip C. W., et al., 2004, “Distribution of Galaxy Spectral types in the Sloan Digital Survey”, The Astronomical Journal, 128, 2603
4. Portillo S. K. N., Parejko J. K., Vergara J. R., Connolly A. J., 2020, “Dimensionality Reduction of SDSS Spectra with Variational Autoencoders”,The Astronomical Journal, 160, 45

Further Reading

- Julie Delon, “The Curse of Dimensionality”
- Aurélien Geron, “Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow”, 2023
- Christopher M. Bishop, ”Pattern Recognition and Machine Learning”, 2006
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville, “Deep Learning”, 2016
- Kevin P. Murphy, “Probabilistic Machine Learning: An Introduction”, 2022

Questions?