Title: Transformers for scientific data - VIRTUAL - Helen Qu and Bhuvnesh Jain

Speakers: Bhuvnesh Jain

Series: Cosmology & Gravitation

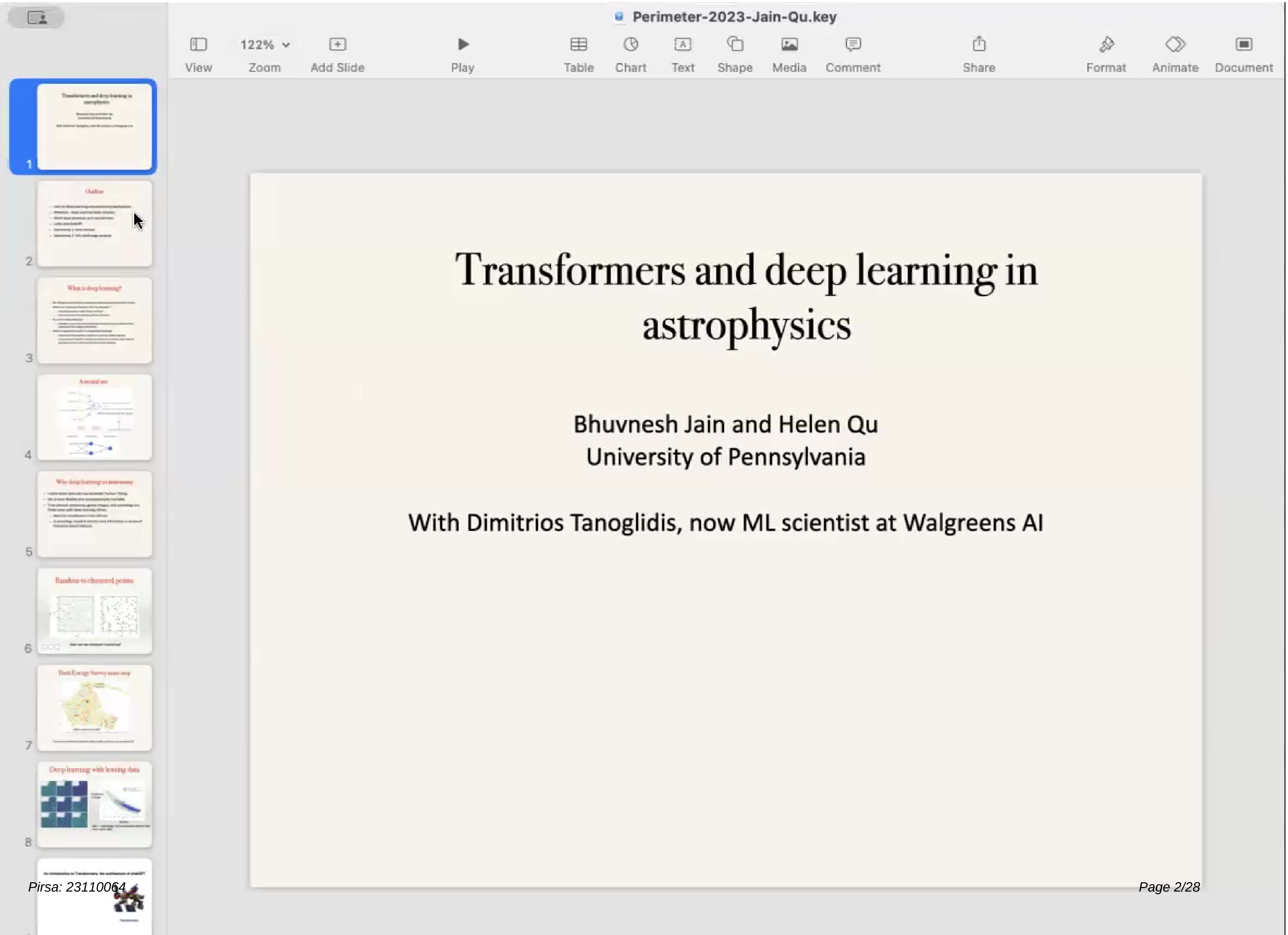
Date: November 14, 2023 - 11:00 AM

URL: https://pirsa.org/23110064

Abstract: The deep learning architecture associated with ChatGPT and related generative AI products is known as transformers. Initially applied to Natural Language Processing, transformers and the self-attention mechanism they exploit have gained widespread interest across the natural sciences. We will present the mathematics underlying the attention mechanism and describe the basic transformer architecture. We will then describe applications to time series and imaging data in astronomy and discuss possible foundation models.

---

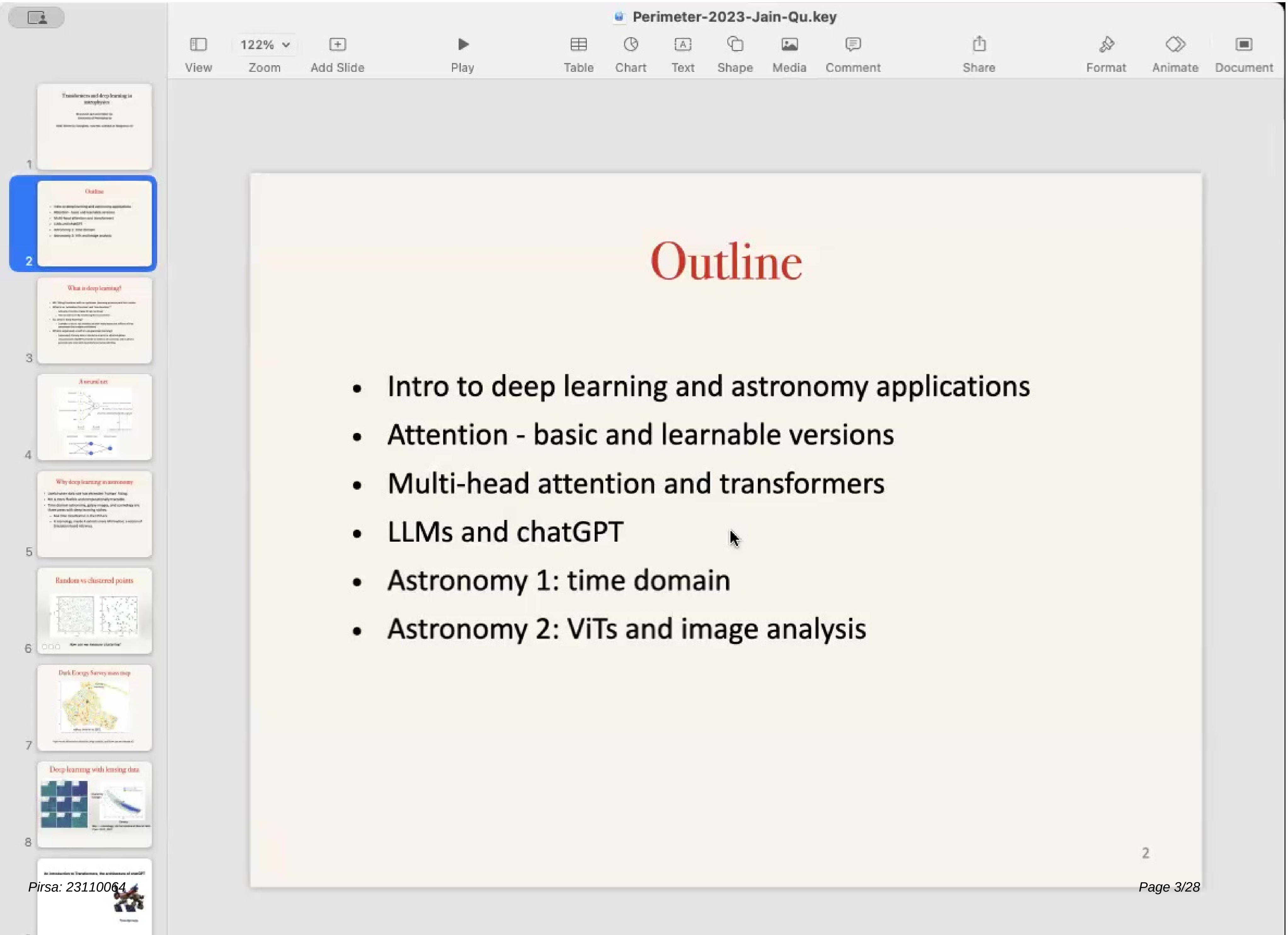
Zoom link https://pitp.zoom.us/j/91226066758?pwd=TWZ5RVliMjVKYXdLcHdya09lNWZhQT09



# Transformers and deep learning in astrophysics

Bhuvnesh Jain and Helen Qu
University of Pennsylvania

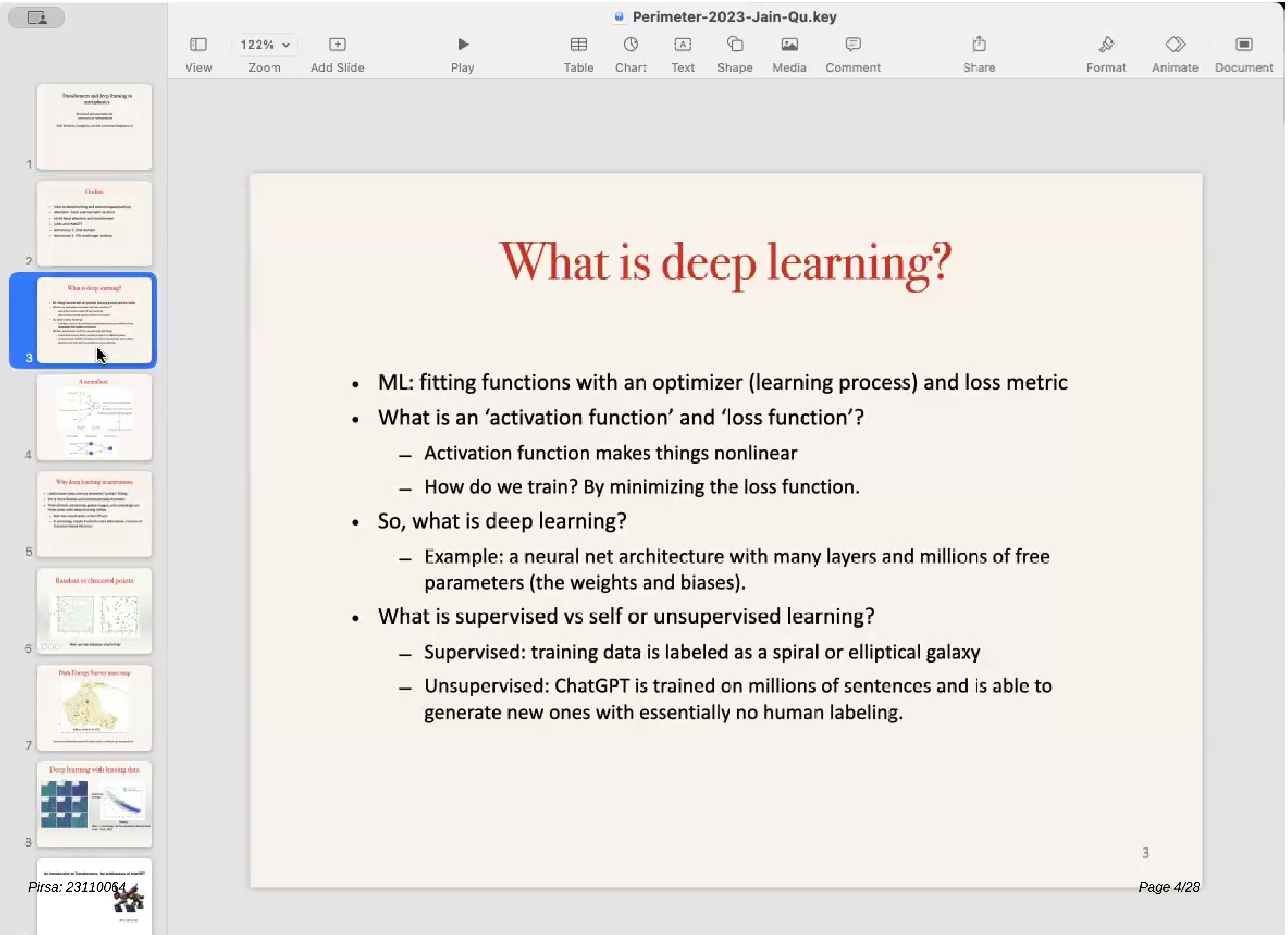
With Dimitrios Tanoglidis, now ML scientist at Walgreens AI



# Outline

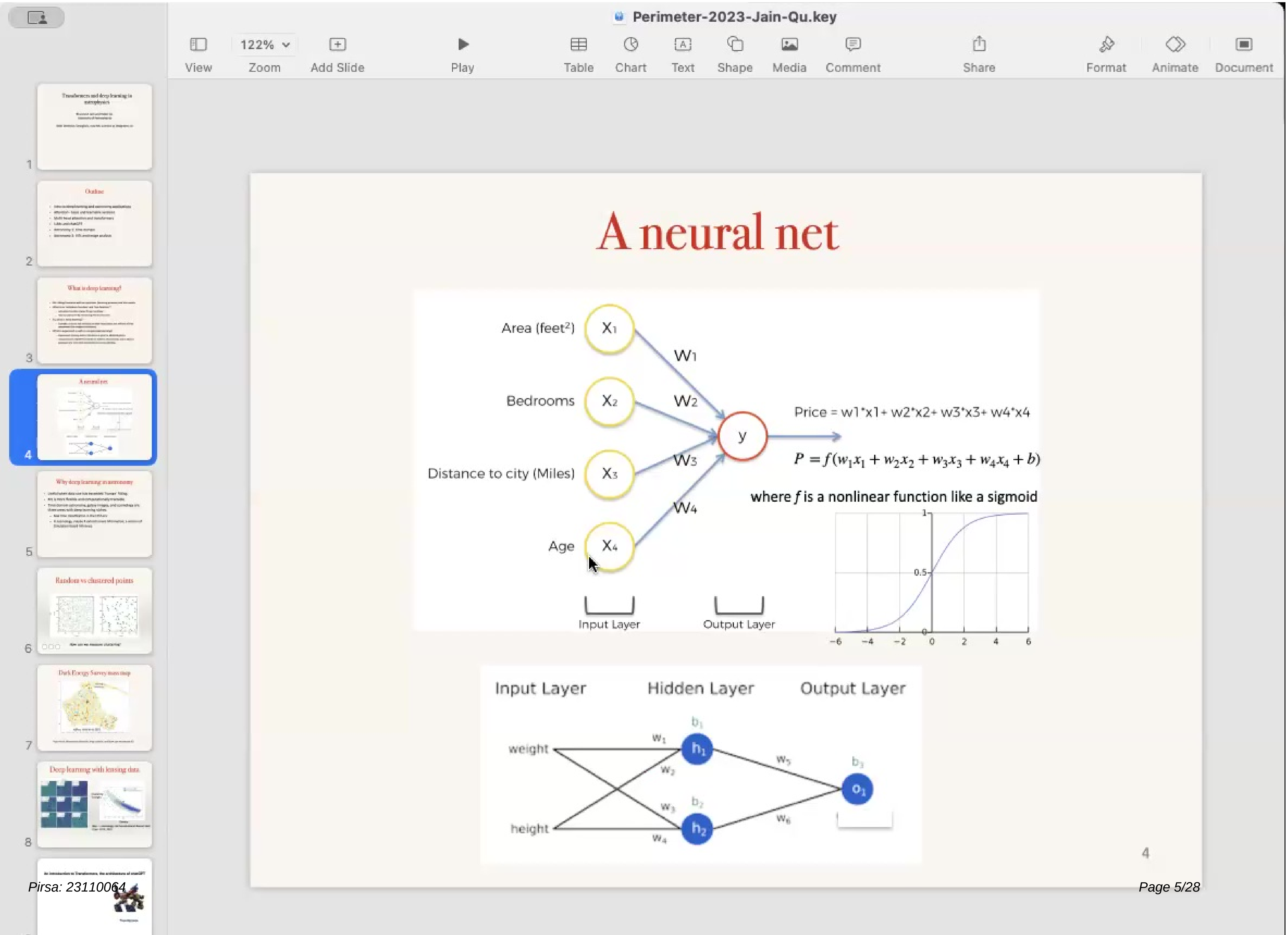
- Intro to deep learning and astronomy applications
- Attention - basic and learnable versions
- Multi-head attention and transformers
- LLMs and chatGPT
- Astronomy 1: time domain
- Astronomy 2: ViTs and image analysis

2


Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
1
2
Outline
3
What is deep learning?
4
A neural net
5
Why deep learning in astronomy
6
Random vs clustered points
7
Dark Energy Survey mass map
8
Deep learning with lensing data

# What is deep learning?

- ML: fitting functions with an optimizer (learning process) and loss metric
- What is an 'activation function' and 'loss function'?
  - Activation function makes things nonlinear
  - How do we train? By minimizing the loss function.
- So, what is deep learning?
  - Example: a neural net architecture with many layers and millions of free parameters (the weights and biases).
- What is supervised vs self or unsupervised learning?
  - Supervised: training data is labeled as a spiral or elliptical galaxy
  - Unsupervised: ChatGPT is trained on millions of sentences and is able to generate new ones with essentially no human labeling.



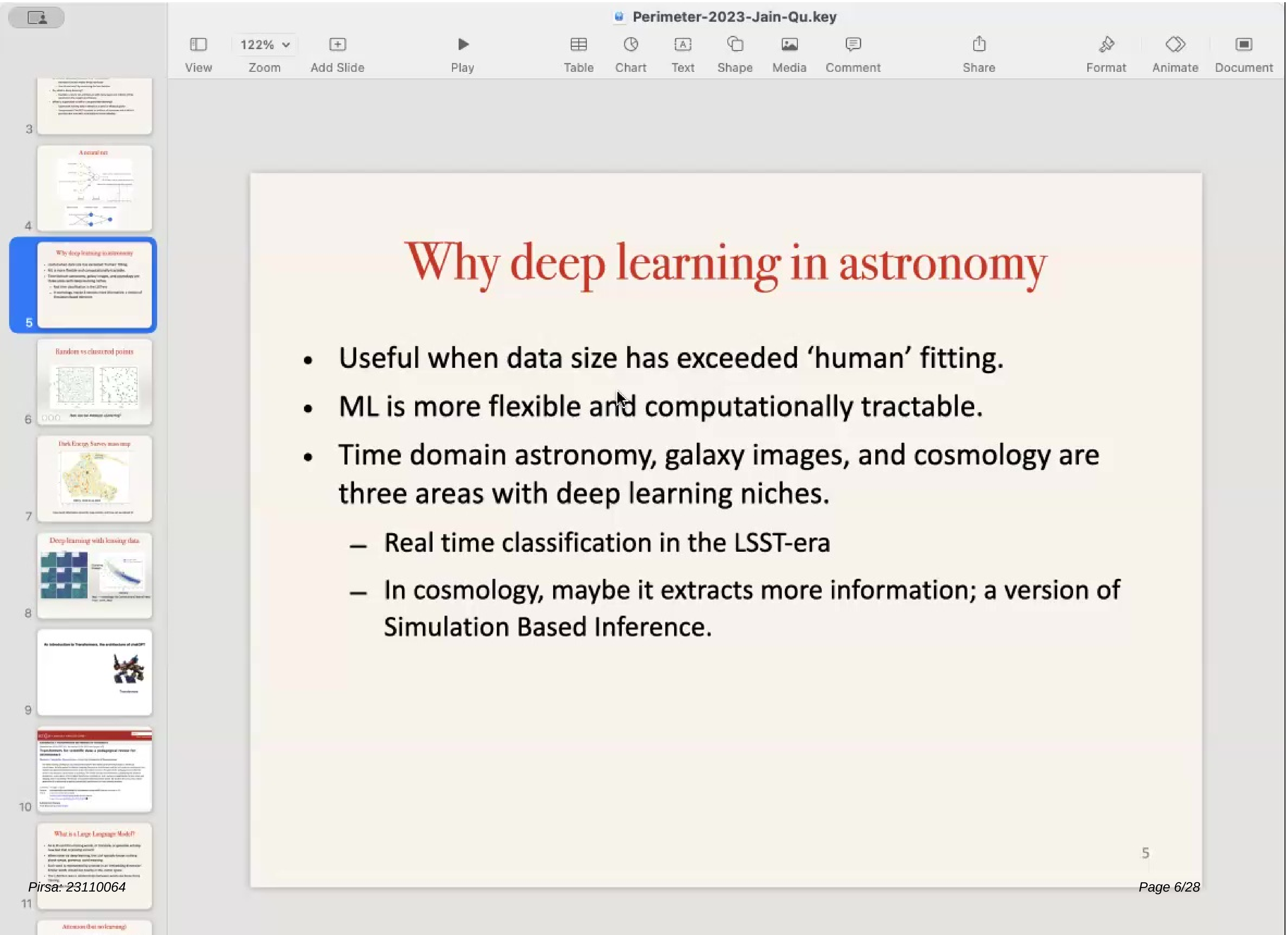
# A neural net

Price = w1*x1+ w2*x2+ w3*x3+ w4*x4

$$P = f(w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b)$$

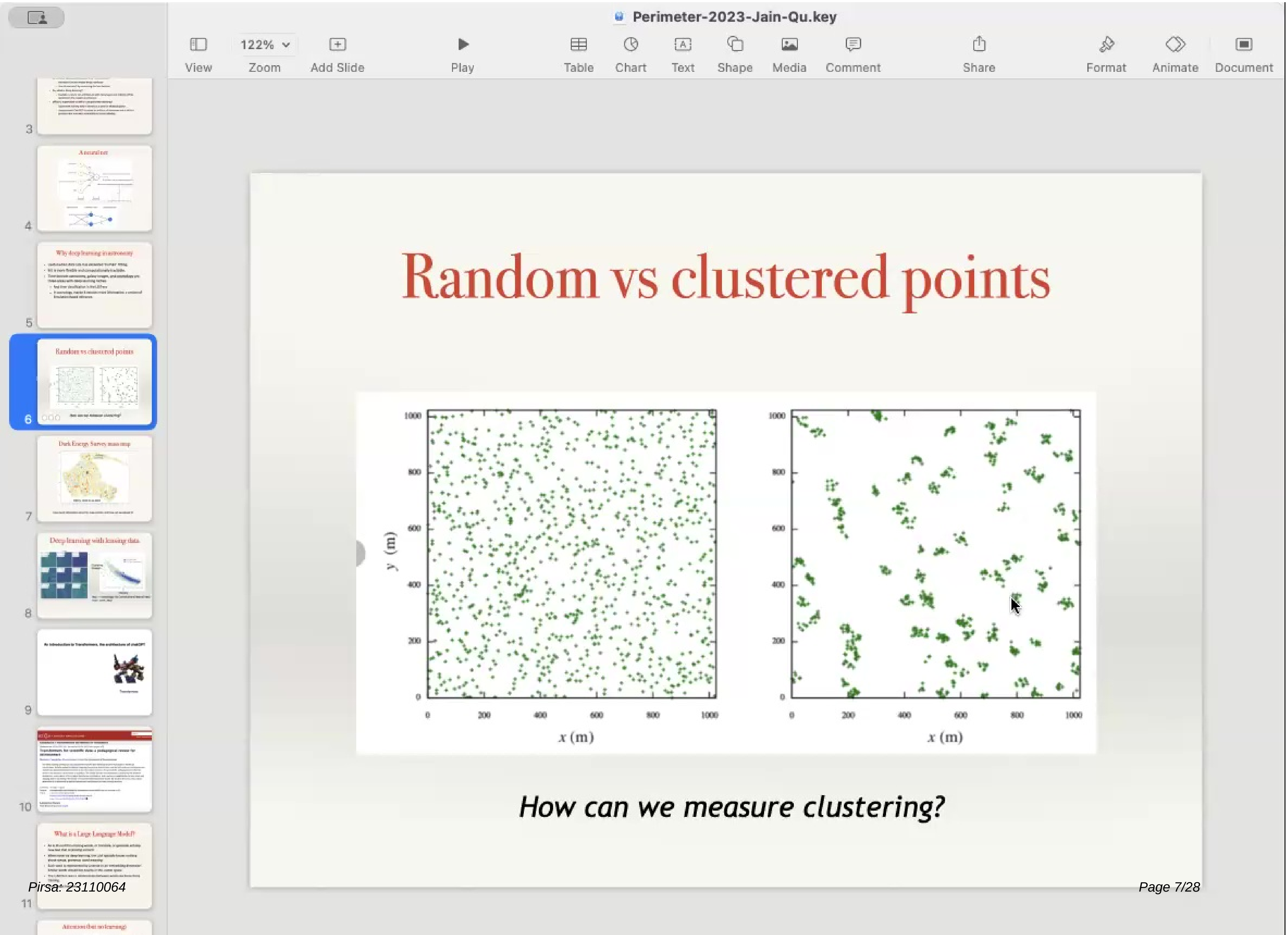
where $f$ is a nonlinear function like a sigmoid

4



# Why deep learning in astronomy

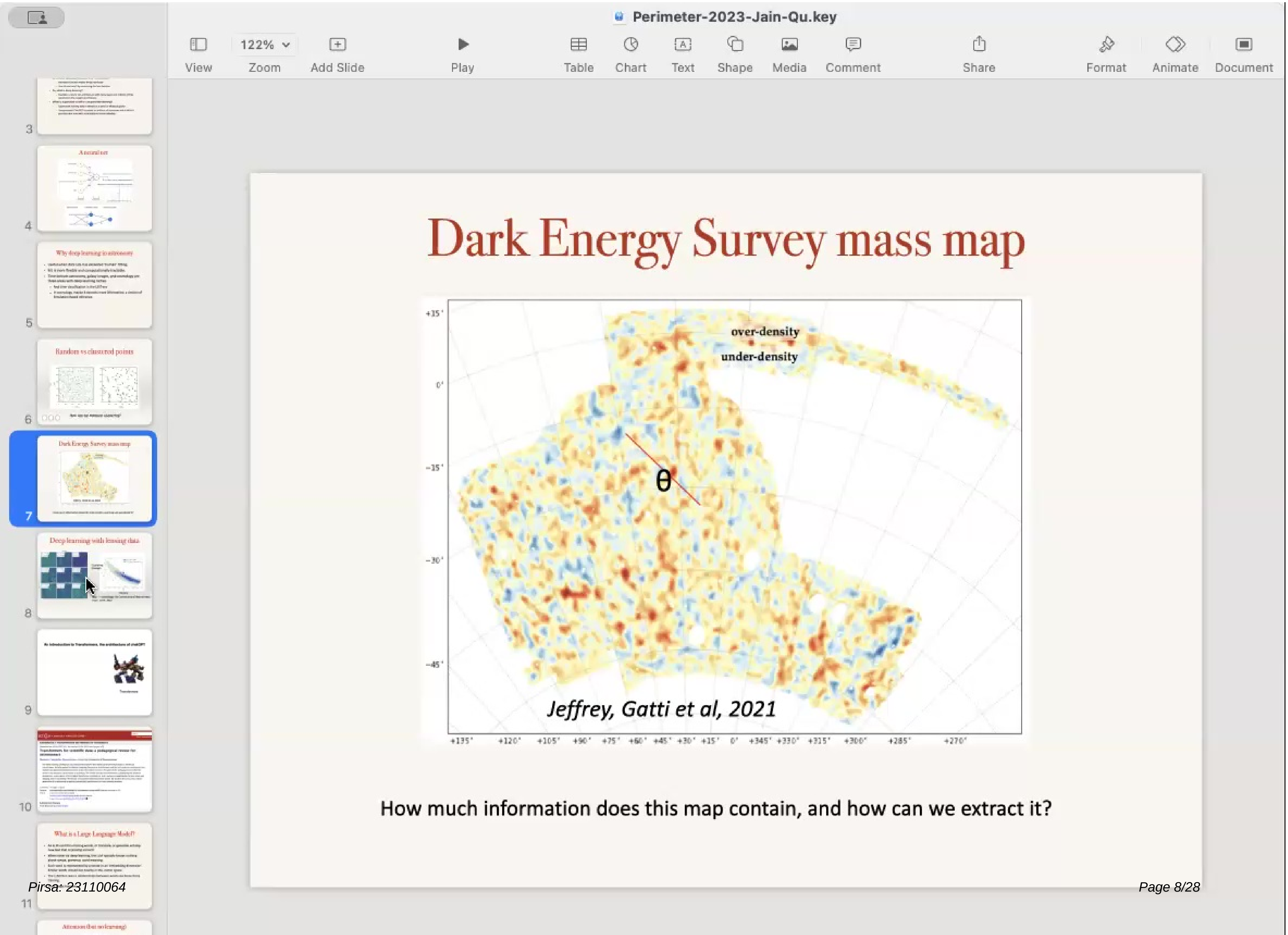
- Useful when data size has exceeded ‘human’ fitting.
- ML is more flexible and computationally tractable.
- Time domain astronomy, galaxy images, and cosmology are three areas with deep learning niches.
  - Real time classification in the LSST-era
  - In cosmology, maybe it extracts more information; a version of Simulation Based Inference.


Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
3
4
A neural net
5
Why deep learning in astronomy
6
Random vs clustered points
How can we measure clustering?
7
Dark Energy Survey mass map
8
Deep learning with lensing data
9
10
11
What is a Large Language Model?
Attention (but no learning)

# Random vs clustered points

1000
800
600
400
200
0
y (m)
0
200
400
600
800
1000
x (m)
1000
800
600
400
200
0
0
200
400
600
800
1000
x (m)

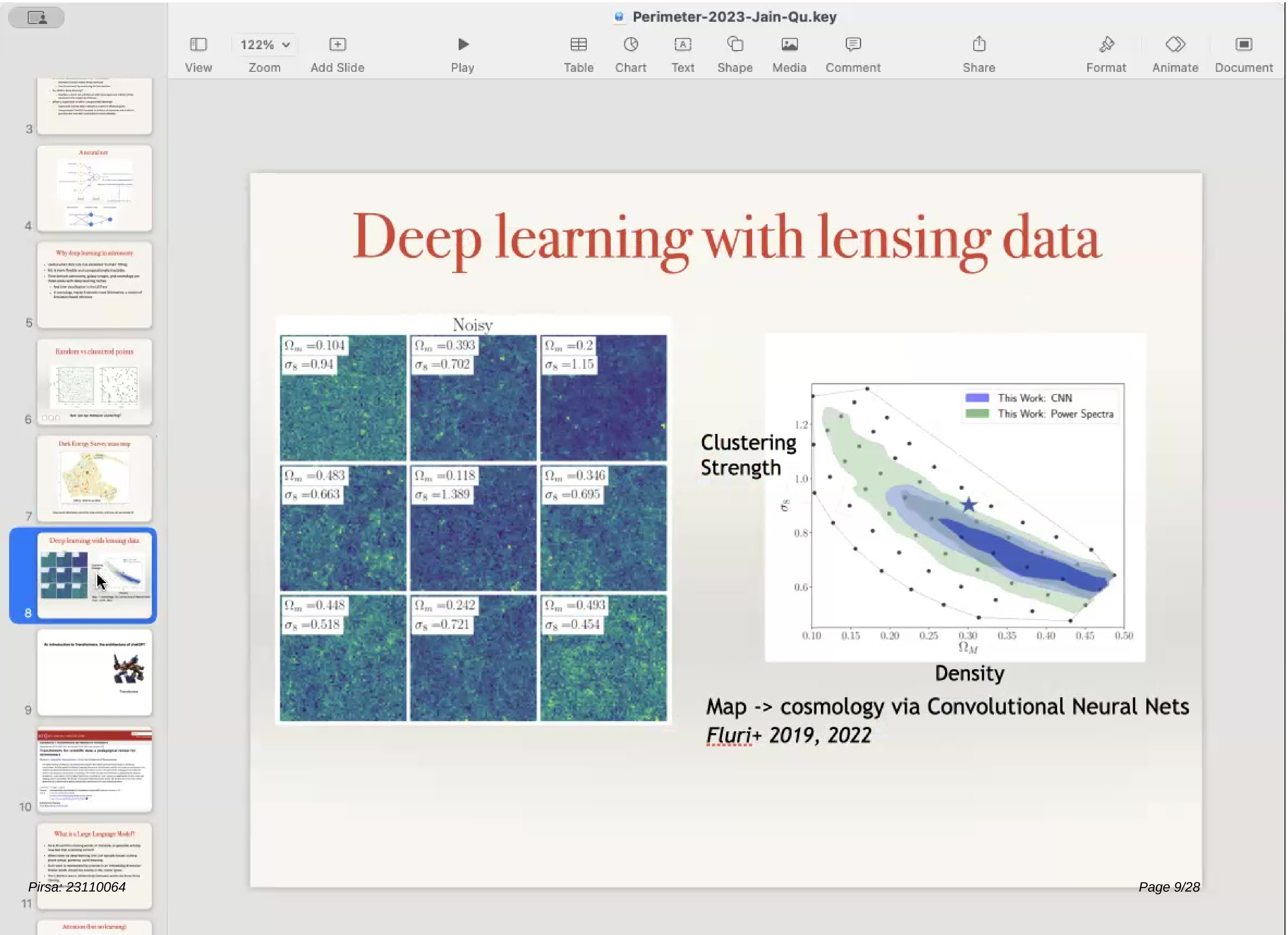
*How can we measure clustering?*


Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
3
A neural net
4
Why deep learning in astronomy
5
Random vs clustered points
6
Dark Energy Survey mass map
7
Deep learning with lensing data
8
An introduction to Transformers, the architecture of chatGPT
9
10
What is a Large Language Model?
11
Attention (but no learning)
+15°
0°
-15°
-30°
-45°
+135°
+120°
+105°
+90°
+75°
+60°
+45°
+30°
+15°
0°
+345°
+330°
+315°
+300°
+285°
+270°
over-density
under-density
θ

# Dark Energy Survey mass map

*Jeffrey, Gatti et al, 2021*

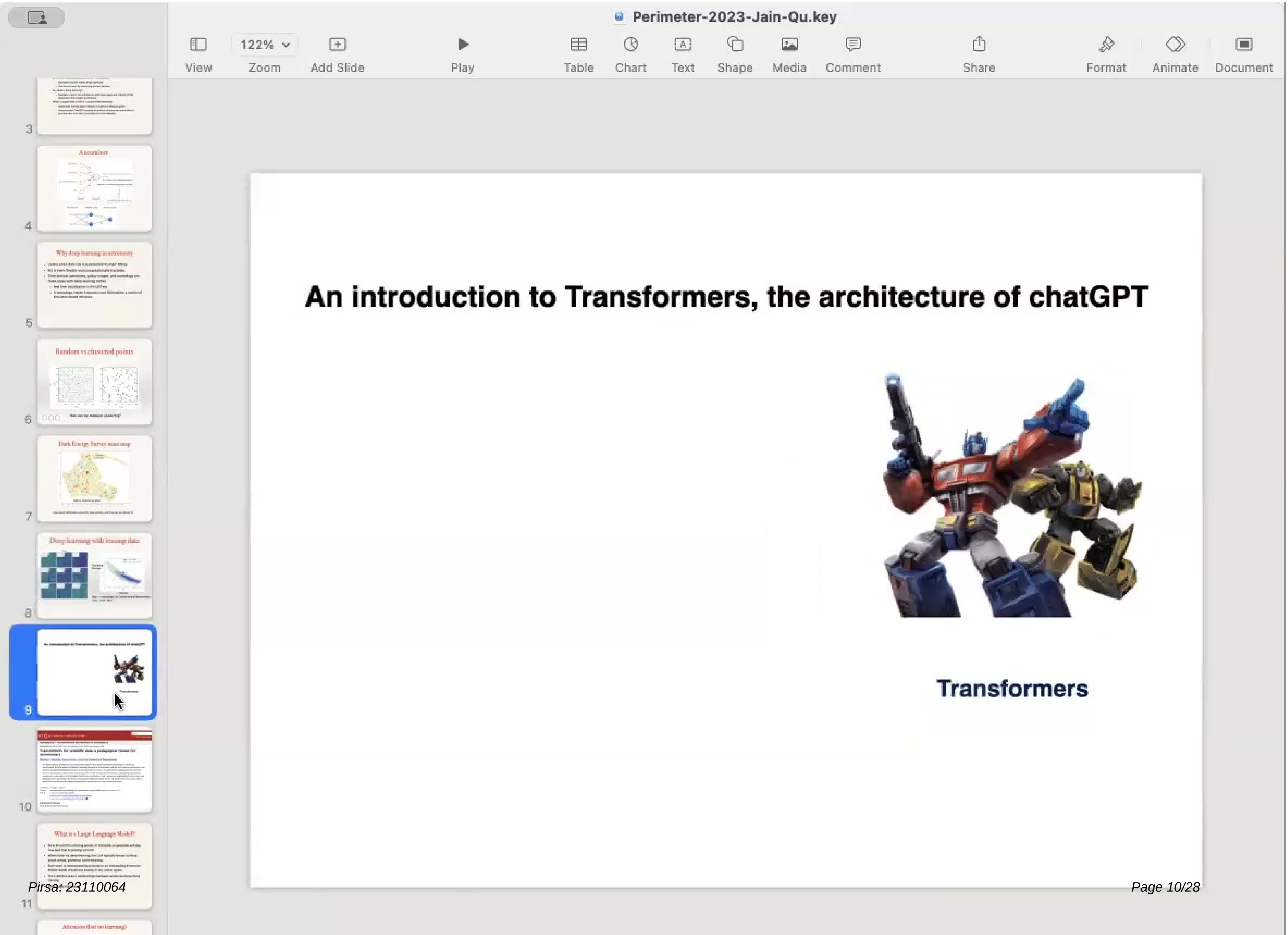
How much information does this map contain, and how can we extract it?


Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
3
A neural net
4
Why deep learning in astronomy
5
Random vs clustered points
6
Dark Energy Survey mass map
7
Deep learning with lensing data
8
9
10
What is a Large Language Model?
11

# Deep learning with lensing data

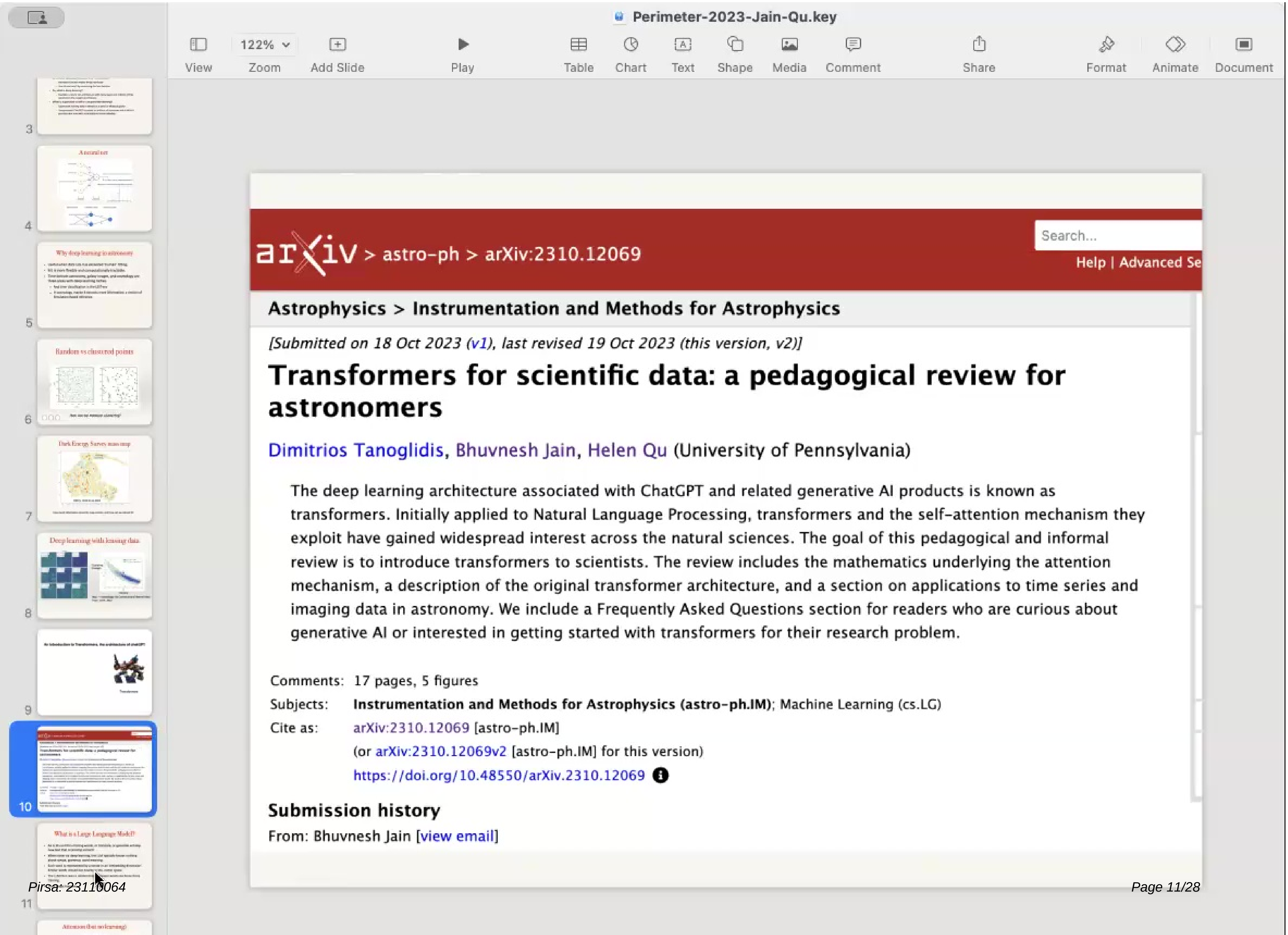
Map -> cosmology via Convolutional Neural Nets

*Fluri+ 2019, 2022*


Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
3
A neural net
4
Why deep learning in astronomy
5
Random vs clustered points
6
Dark Energy Survey mass map
7
Deep learning with lensing data
8
9
10
What is a Large Language Model?
11
Attention (but no learning)

# An introduction to Transformers, the architecture of chatGPT

Transformers


Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
3
4
A neural net
5
Why deep learning in astronomy
6
Random vs clustered points
7
Dark Energy Survey mass map
8
Deep learning with lensing data
9
An introduction to Transformers, the architecture of chatGPT
10
11
What is a Large Language Model?
Attention (but no learning)
arXiv > astro-ph > arXiv:2310.12069
Search...
Help | Advanced Se

Astrophysics > Instrumentation and Methods for Astrophysics

*[Submitted on 18 Oct 2023 (v1), last revised 19 Oct 2023 (this version, v2)]*

# Transformers for scientific data: a pedagogical review for astronomers

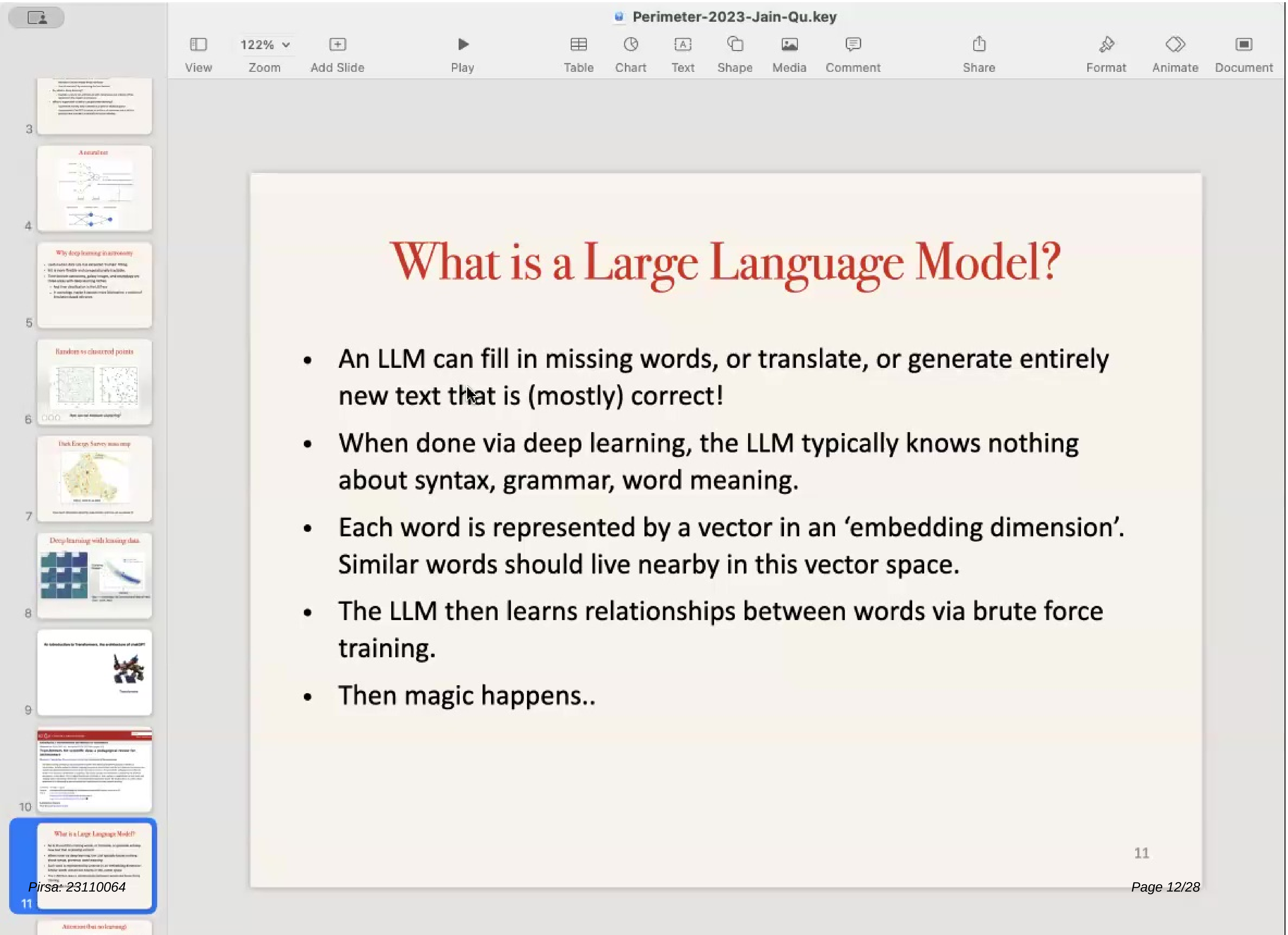
Dimitrios Tanoglidis, Bhuvnesh Jain, Helen Qu (University of Pennsylvania)

The deep learning architecture associated with ChatGPT and related generative AI products is known as transformers. Initially applied to Natural Language Processing, transformers and the self-attention mechanism they exploit have gained widespread interest across the natural sciences. The goal of this pedagogical and informal review is to introduce transformers to scientists. The review includes the mathematics underlying the attention mechanism, a description of the original transformer architecture, and a section on applications to time series and imaging data in astronomy. We include a Frequently Asked Questions section for readers who are curious about generative AI or interested in getting started with transformers for their research problem.

Comments: 17 pages, 5 figures
Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**; Machine Learning (cs.LG)
Cite as: arXiv:2310.12069 [astro-ph.IM]
(or arXiv:2310.12069v2 [astro-ph.IM] for this version)
https://doi.org/10.48550/arXiv.2310.12069

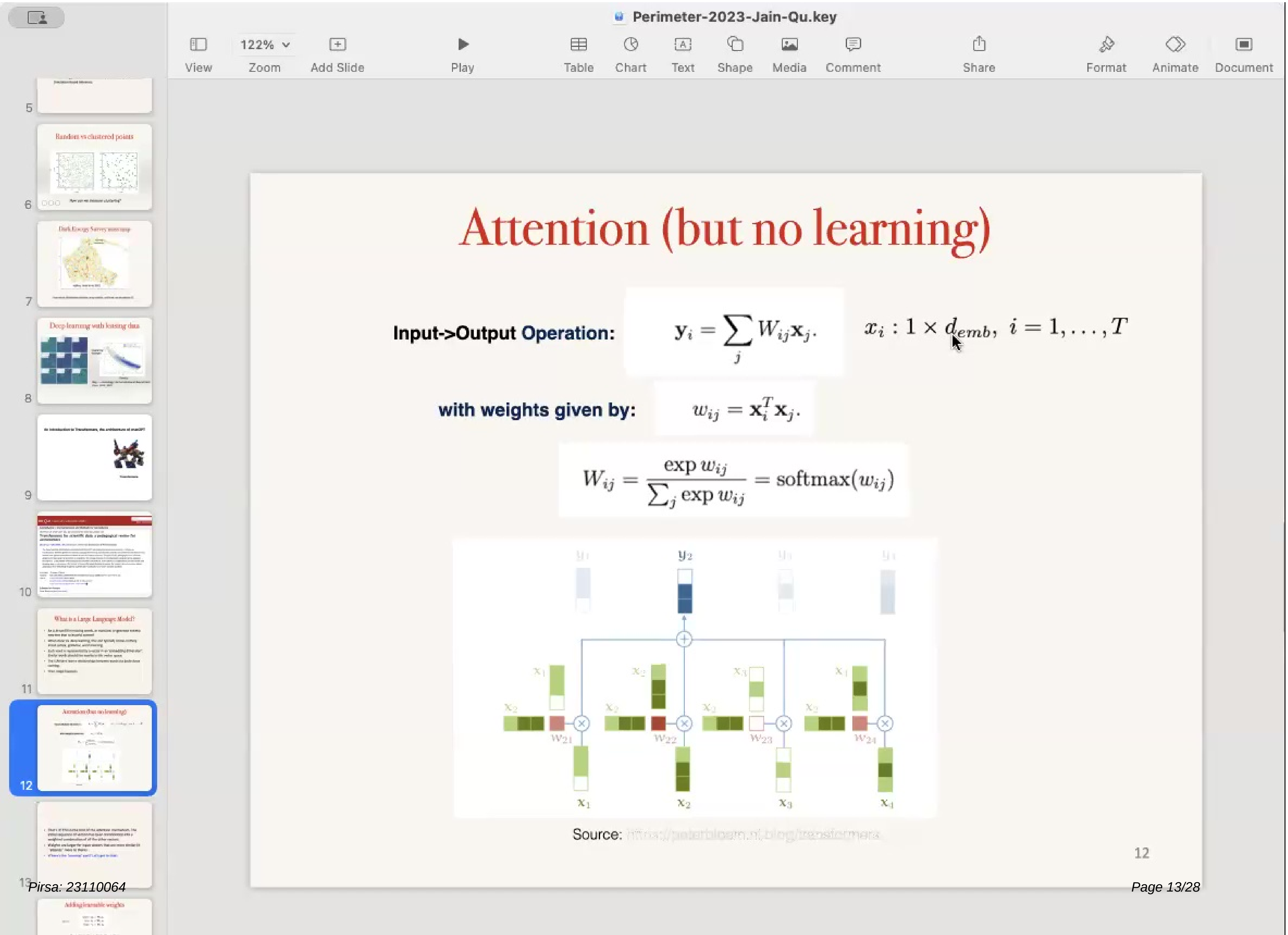
## Submission history

From: Bhuvnesh Jain [view email]



# What is a Large Language Model?

- An LLM can fill in missing words, or translate, or generate entirely new text that is (mostly) correct!
- When done via deep learning, the LLM typically knows nothing about syntax, grammar, word meaning.
- Each word is represented by a vector in an ‘embedding dimension’. Similar words should live nearby in this vector space.
- The LLM then learns relationships between words via brute force training.
- Then magic happens..


Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
5
Random vs clustered points
6
Dark Energy Survey mass map
7
Deep learning with missing data
8
9
10
What is a Large Language Model?
11
Attention (but no learning)
12
13
Adding learnable weights
y1
y2
y3
y4
x1
x2
x3
x4
w21
w22
w23
w24

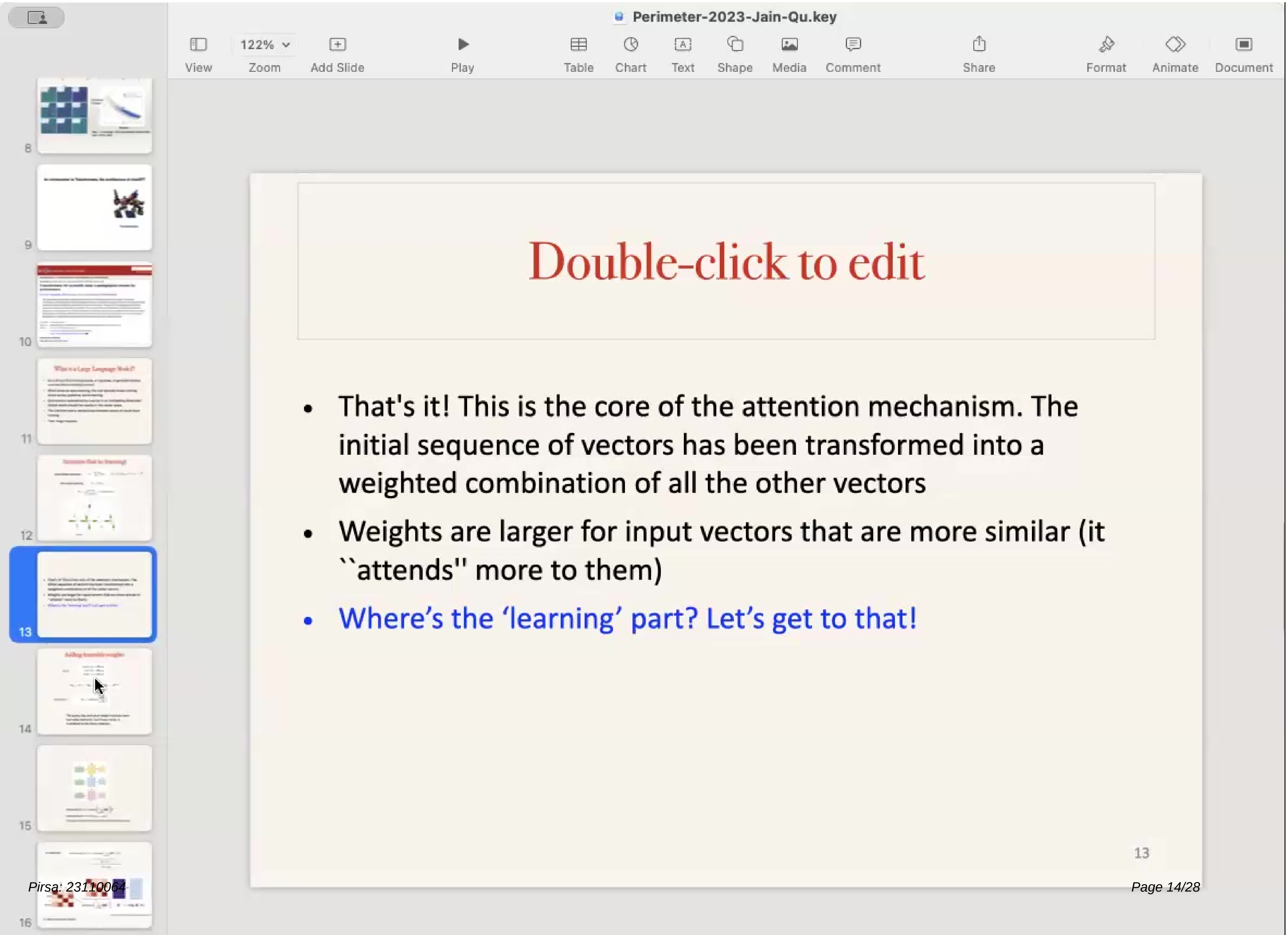
# Attention (but no learning)

**Input->Output Operation:** $\mathbf{y}_i = \sum_j W_{ij}\mathbf{x}_j.$ $\quad x_i : 1 \times d_{emb},\ i = 1, \ldots, T$

**with weights given by:** $w_{ij} = \mathbf{x}_i^T \mathbf{x}_j.$

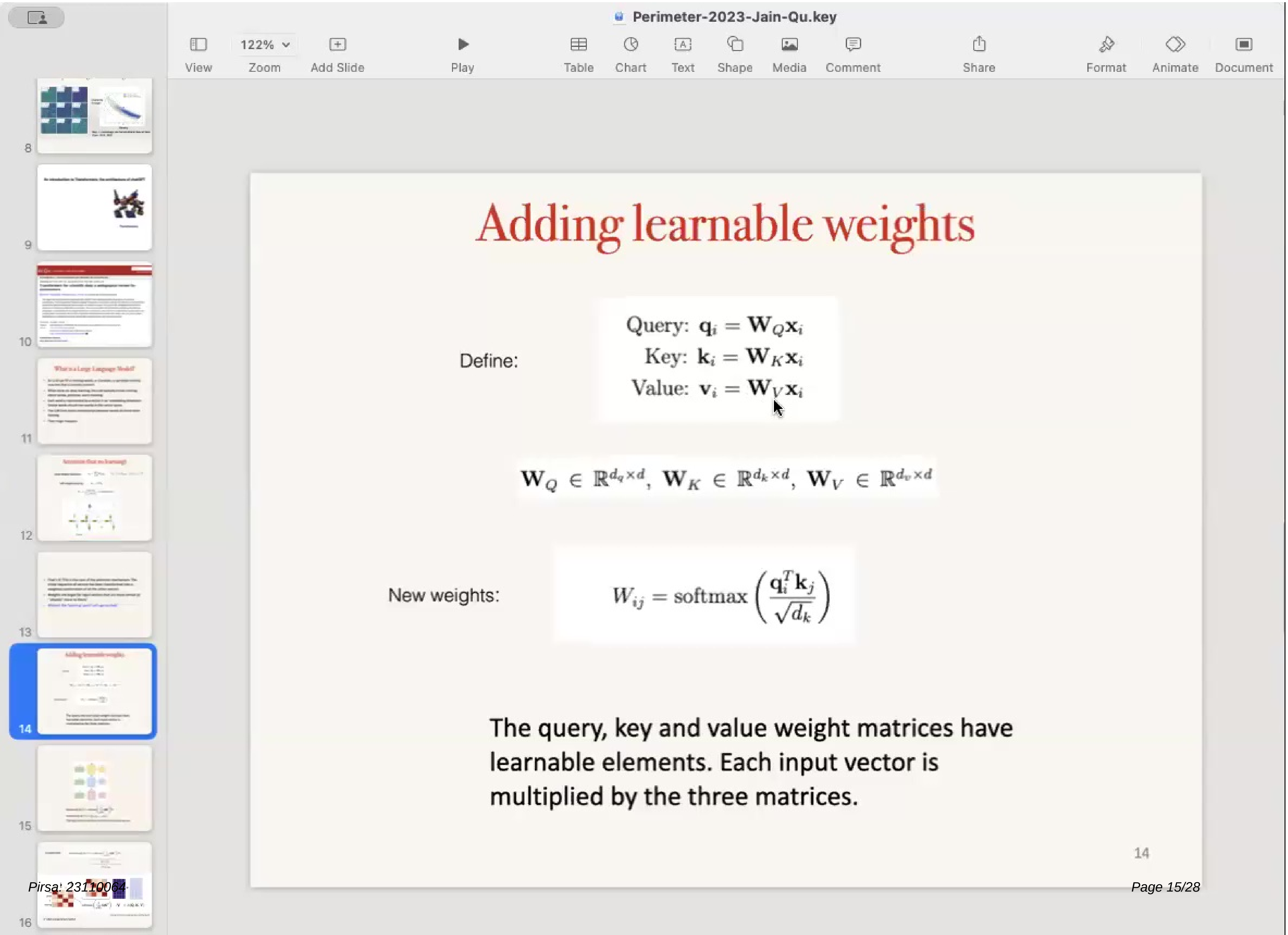
$$W_{ij} = \frac{\exp w_{ij}}{\sum_j \exp w_{ij}} = \text{softmax}(w_{ij})$$

Source: https://peterbloem.nl/blog/transformers



# Double-click to edit

- That's it! This is the core of the attention mechanism. The initial sequence of vectors has been transformed into a weighted combination of all the other vectors
- Weights are larger for input vectors that are more similar (it ``attends'' more to them)
- Where's the 'learning' part? Let's get to that!



# Adding learnable weights

Define:

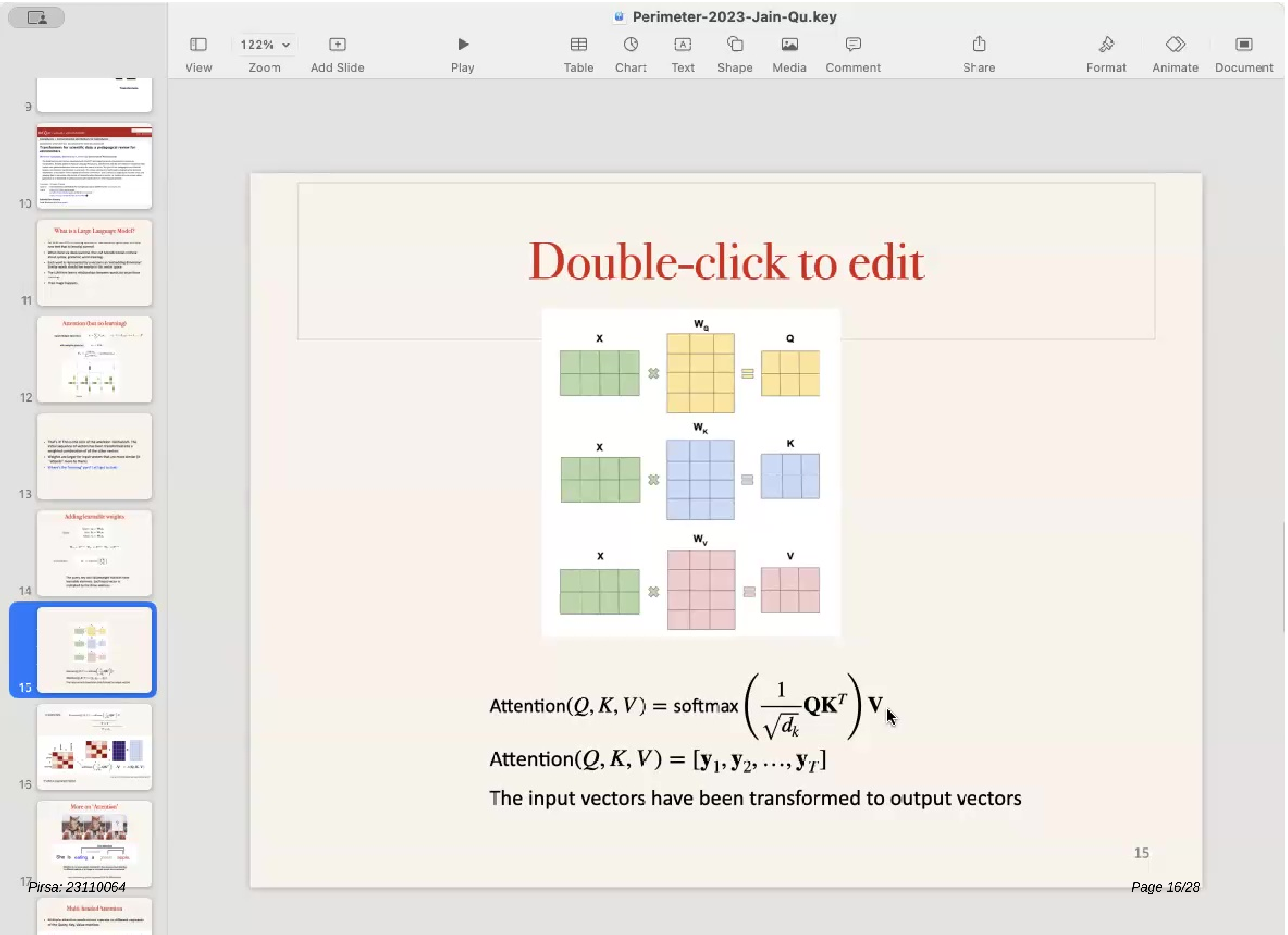
$$\begin{aligned} \text{Query: } \mathbf{q}_i &= \mathbf{W}_Q \mathbf{x}_i \\ \text{Key: } \mathbf{k}_i &= \mathbf{W}_K \mathbf{x}_i \\ \text{Value: } \mathbf{v}_i &= \mathbf{W}_V \mathbf{x}_i \end{aligned}$$

$$\mathbf{W}_Q \in \mathbb{R}^{d_q \times d},\ \mathbf{W}_K \in \mathbb{R}^{d_k \times d},\ \mathbf{W}_V \in \mathbb{R}^{d_v \times d}$$

New weights:

$$W_{ij} = \text{softmax}\left(\frac{\mathbf{q}_i^T \mathbf{k}_j}{\sqrt{d_k}}\right)$$

The query, key and value weight matrices have learnable elements. Each input vector is multiplied by the three matrices.

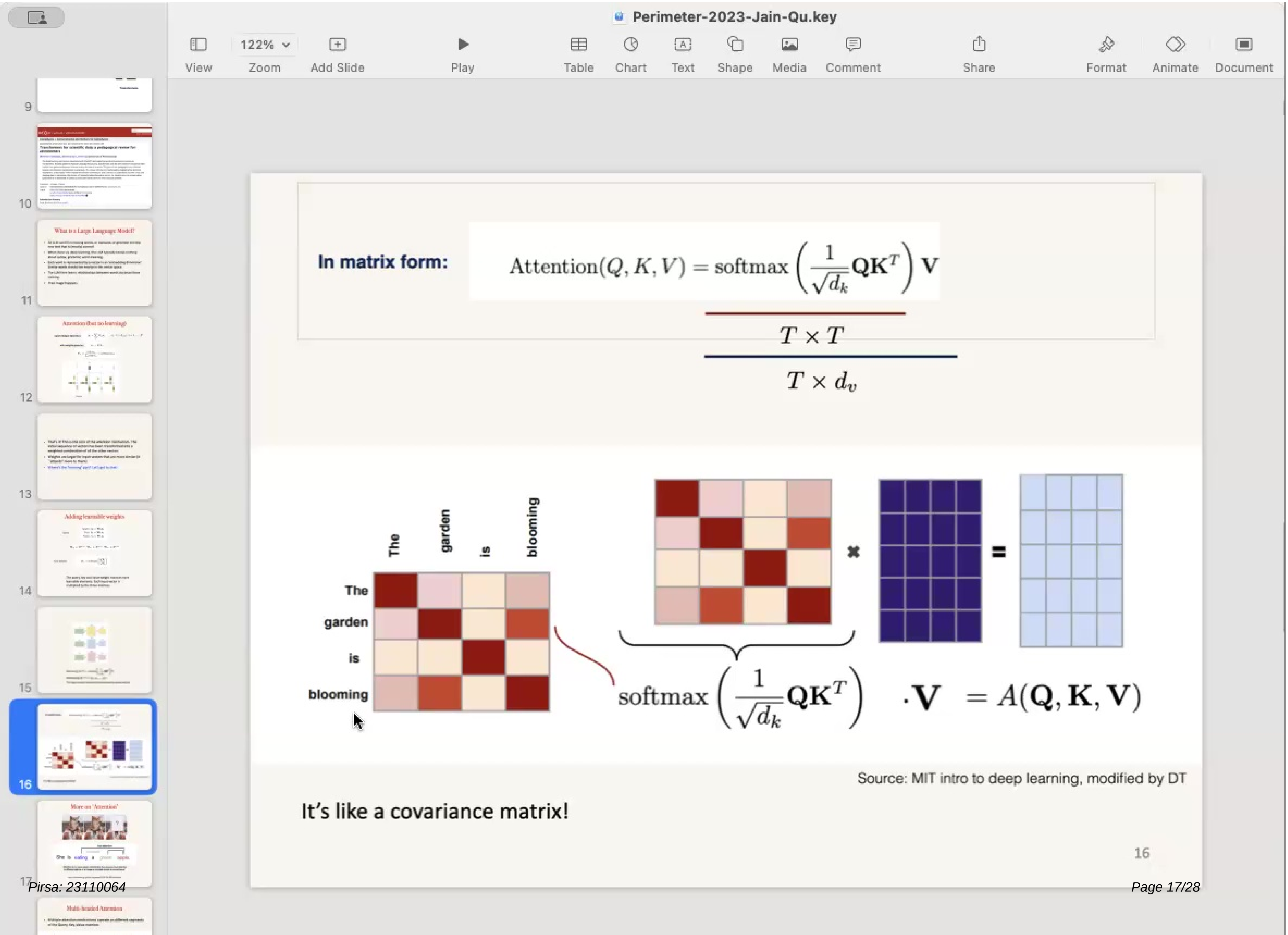


# Double-click to edit

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{1}{\sqrt{d_k}}\mathbf{QK}^T\right)\mathbf{V}$$

$$\text{Attention}(Q, K, V) = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T]$$

The input vectors have been transformed to output vectors



**In matrix form:**

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{1}{\sqrt{d_k}}\mathbf{QK}^T\right)\mathbf{V}$$

$T \times T$

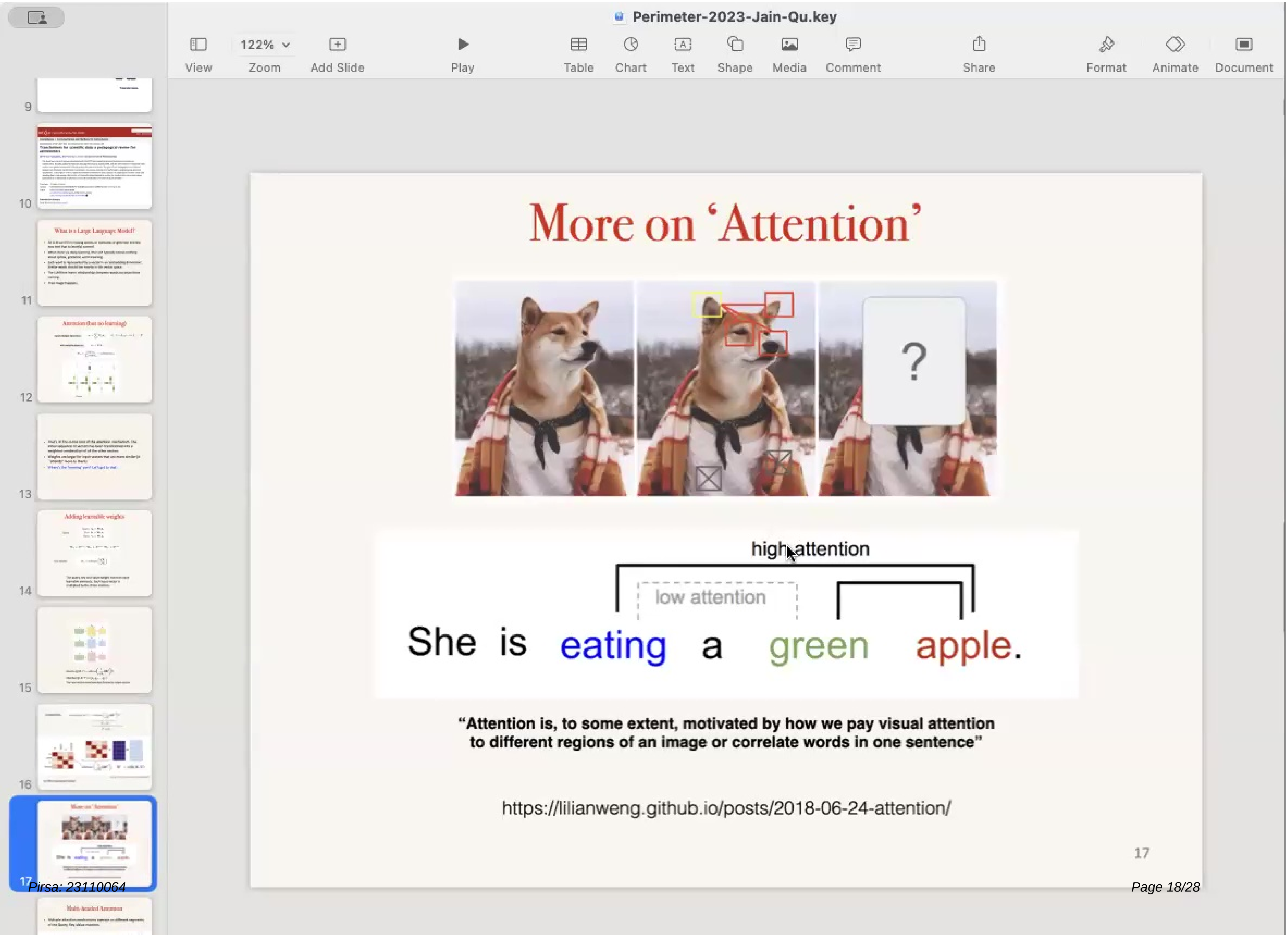
$T \times d_v$

$$\text{softmax}\left(\frac{1}{\sqrt{d_k}}\mathbf{QK}^T\right) \cdot \mathbf{V} = A(\mathbf{Q}, \mathbf{K}, \mathbf{V})$$

Source: MIT intro to deep learning, modified by DT

It’s like a covariance matrix!

16



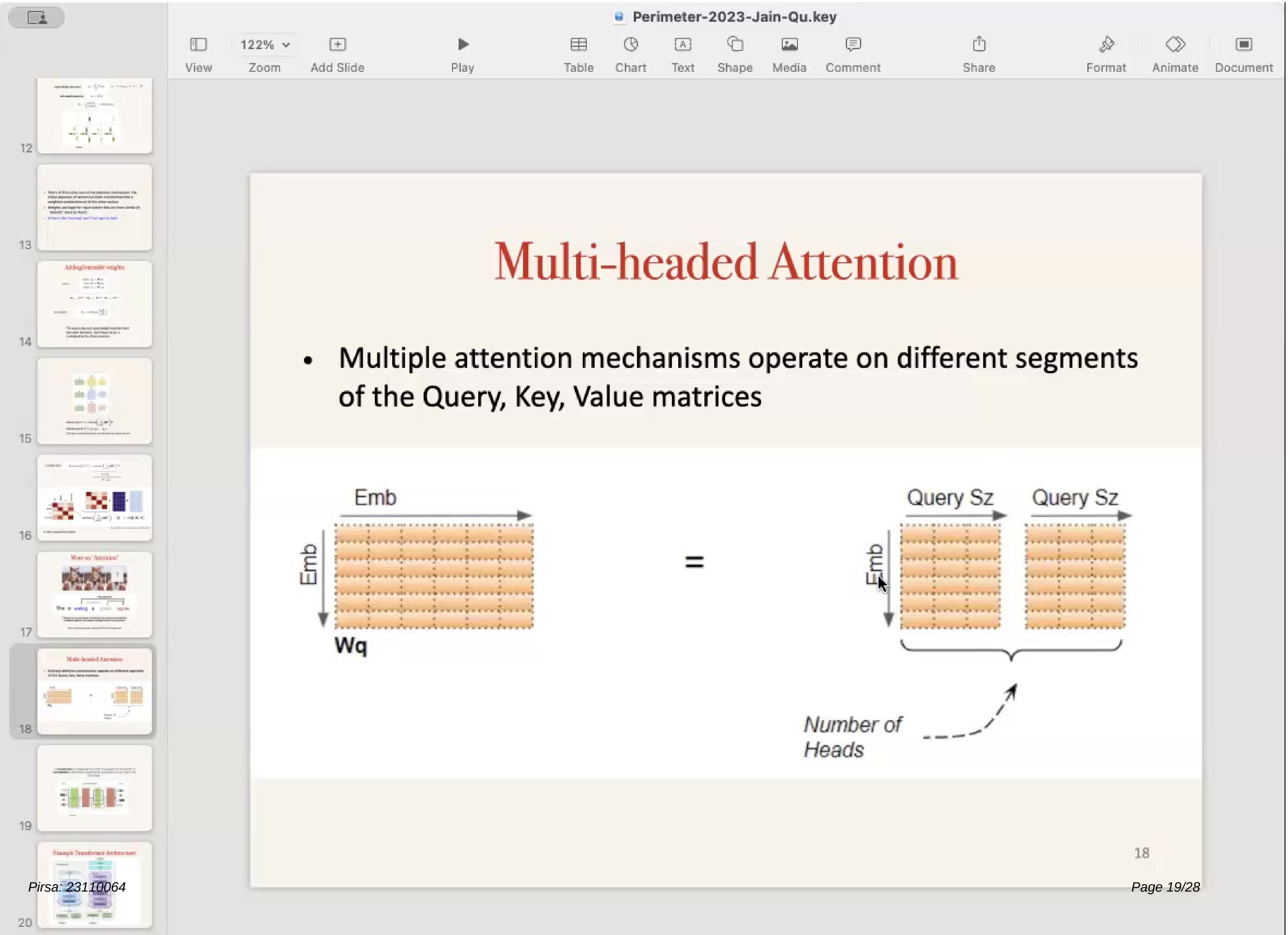
# More on ‘Attention’

She is eating a green apple.

**“Attention is, to some extent, motivated by how we pay visual attention to different regions of an image or correlate words in one sentence”**

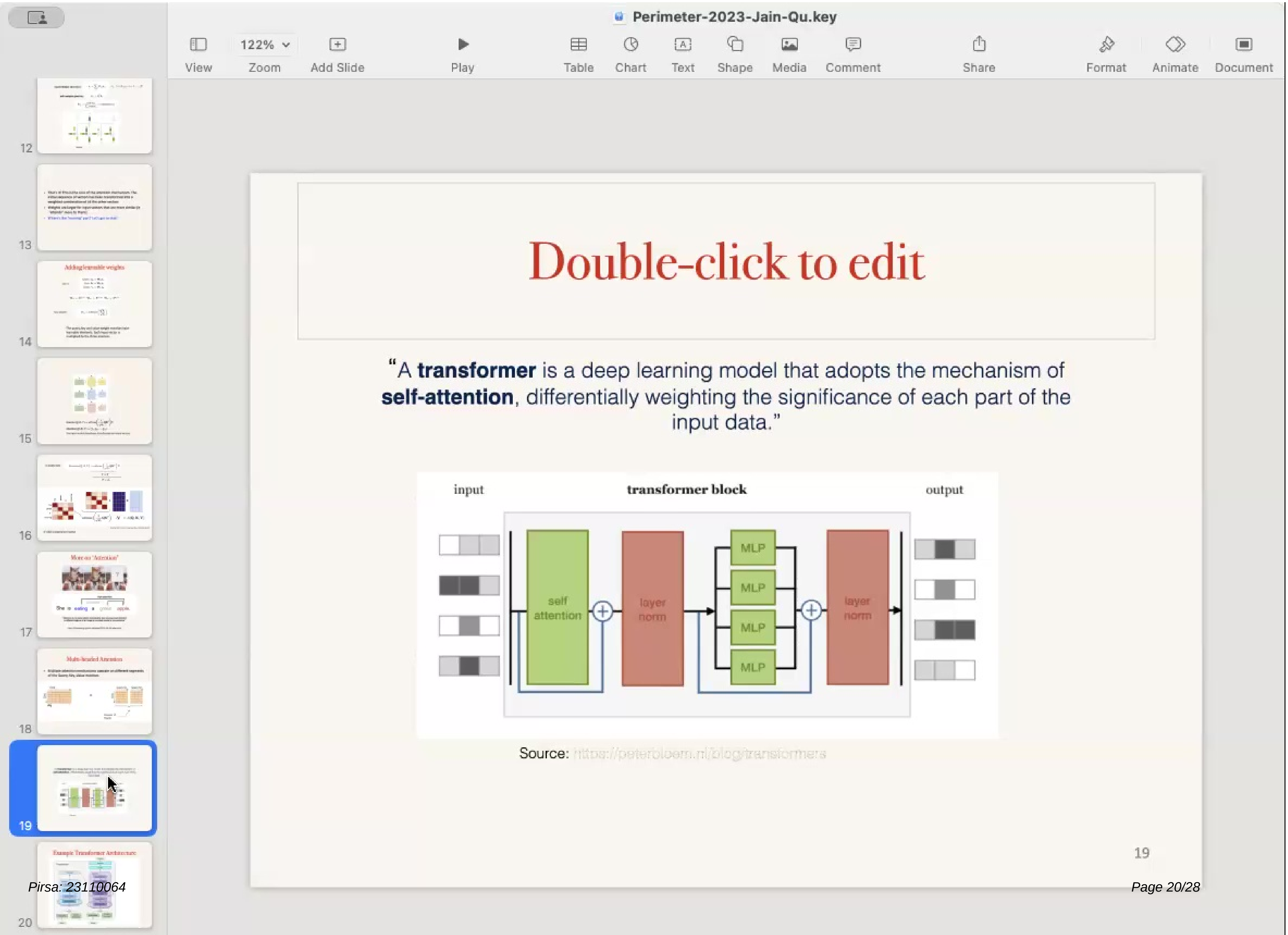
https://lilianweng.github.io/posts/2018-06-24-attention/

17



# Multi-headed Attention

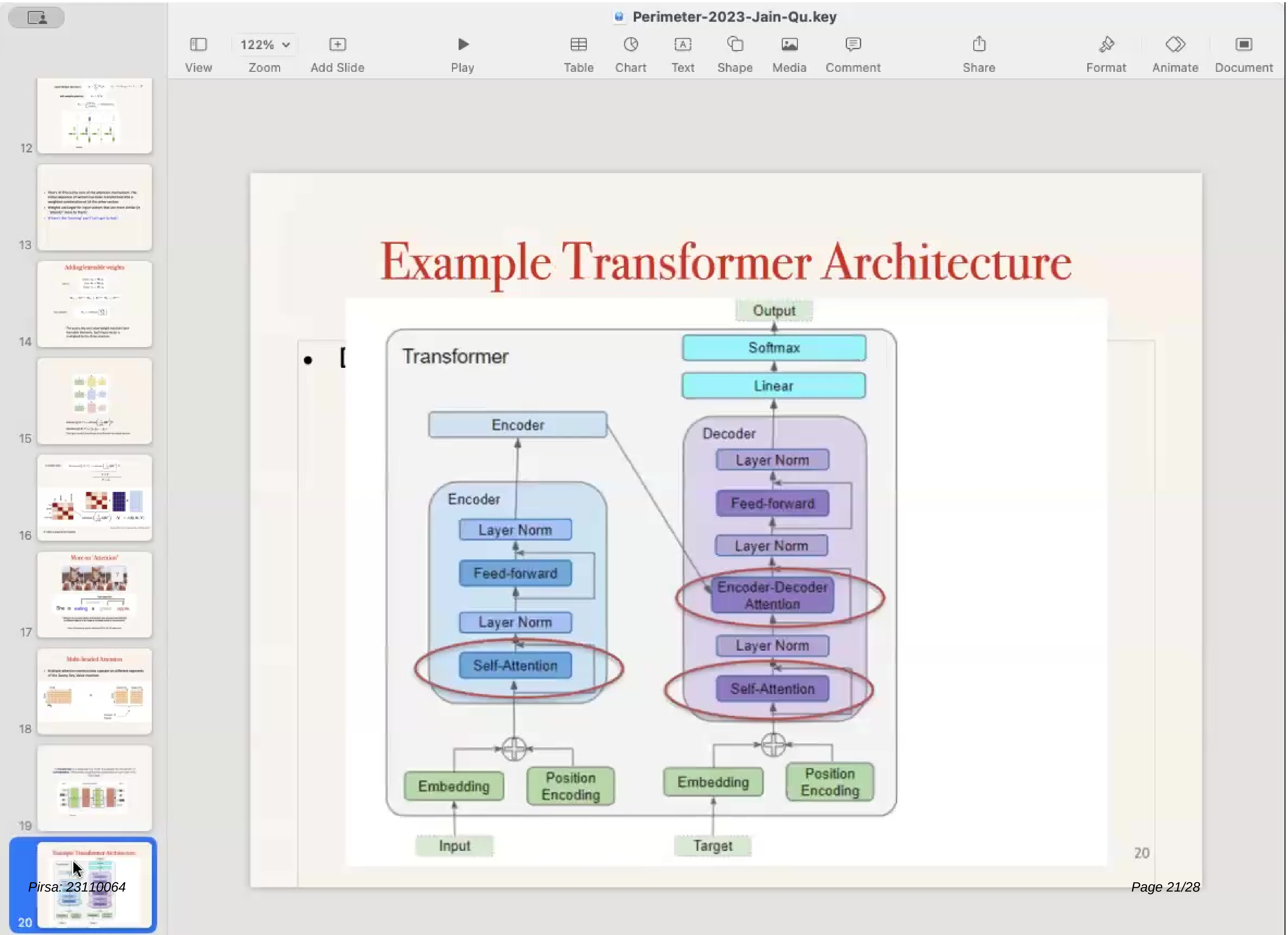
- Multiple attention mechanisms operate on different segments of the Query, Key, Value matrices



# Double-click to edit

"A **transformer** is a deep learning model that adopts the mechanism of **self-attention**, differentially weighting the significance of each part of the input data."

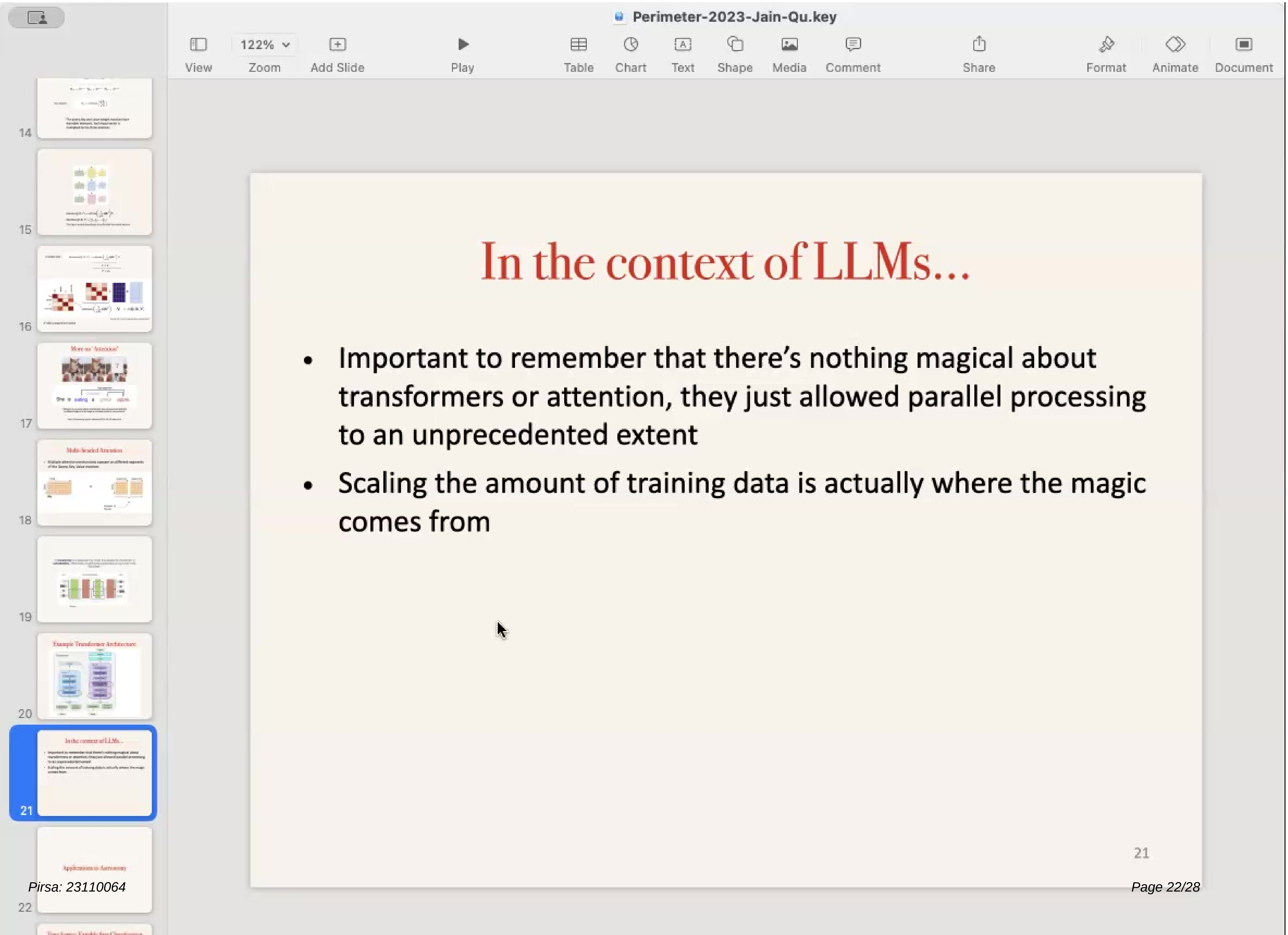
Source: https://peterbloem.nl/blog/transformers



# Example Transformer Architecture

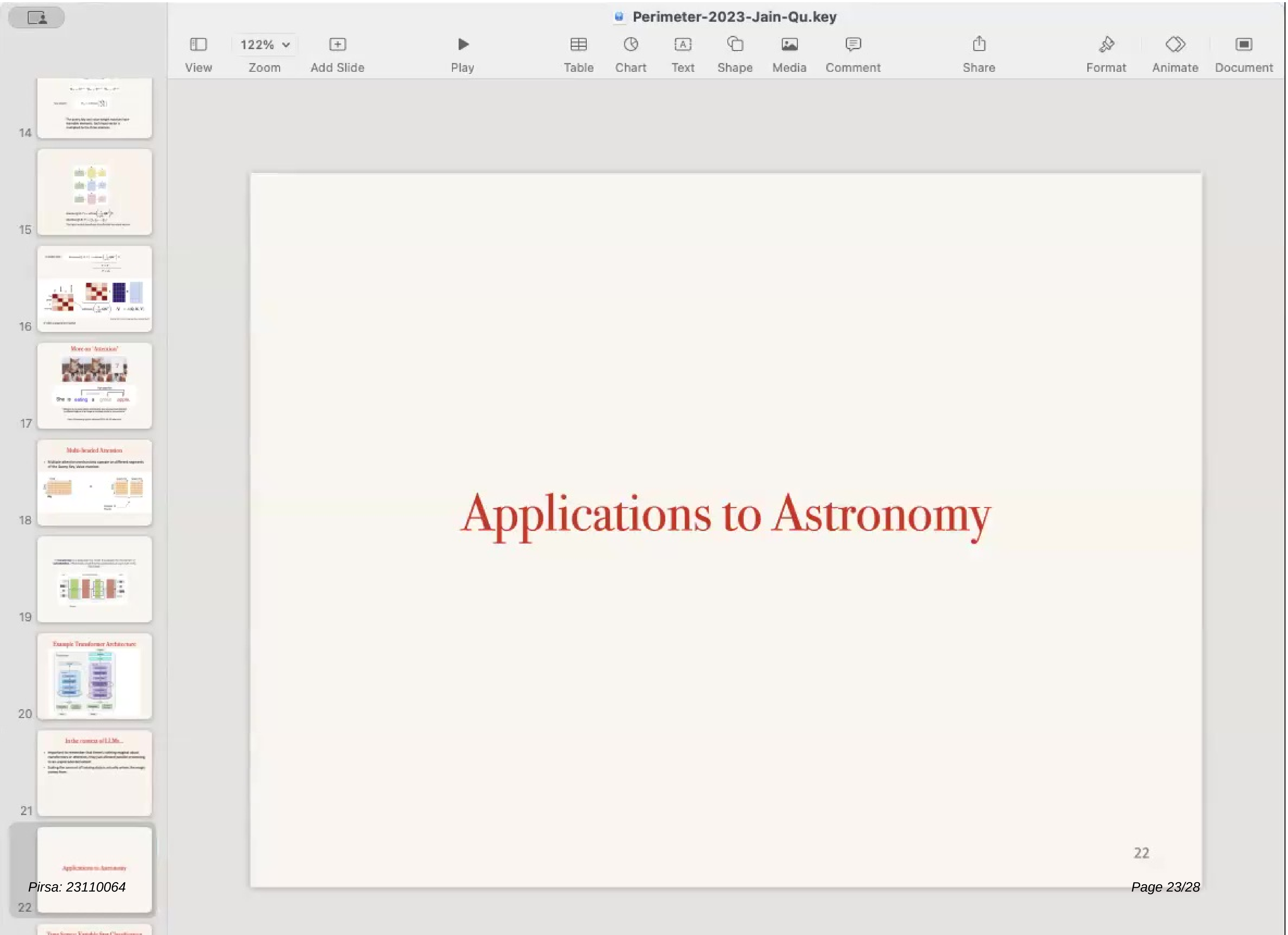
- [

20



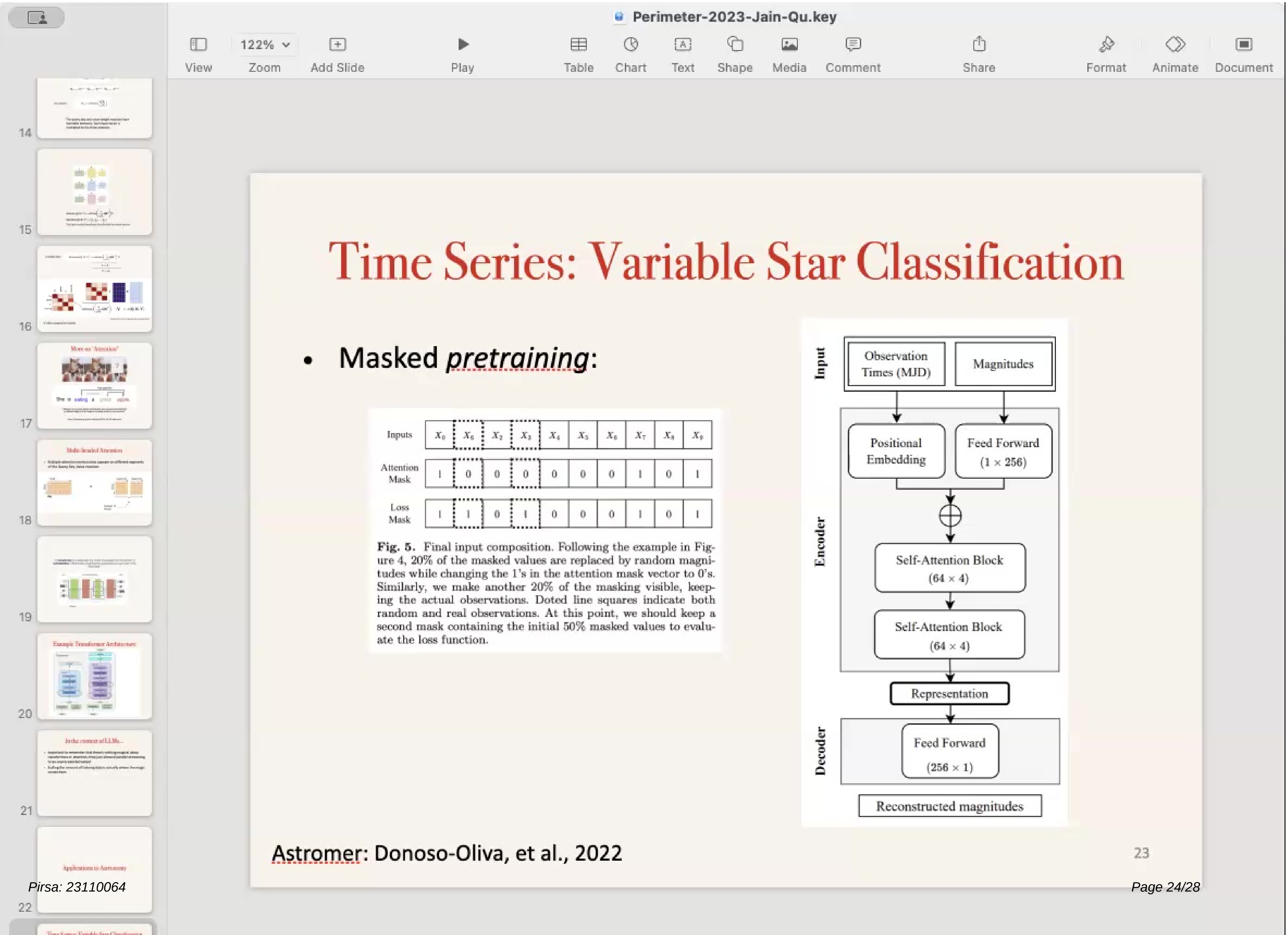
# In the context of LLMs…

- Important to remember that there's nothing magical about transformers or attention, they just allowed parallel processing to an unprecedented extent
- Scaling the amount of training data is actually where the magic comes from



# Applications to Astronomy

22



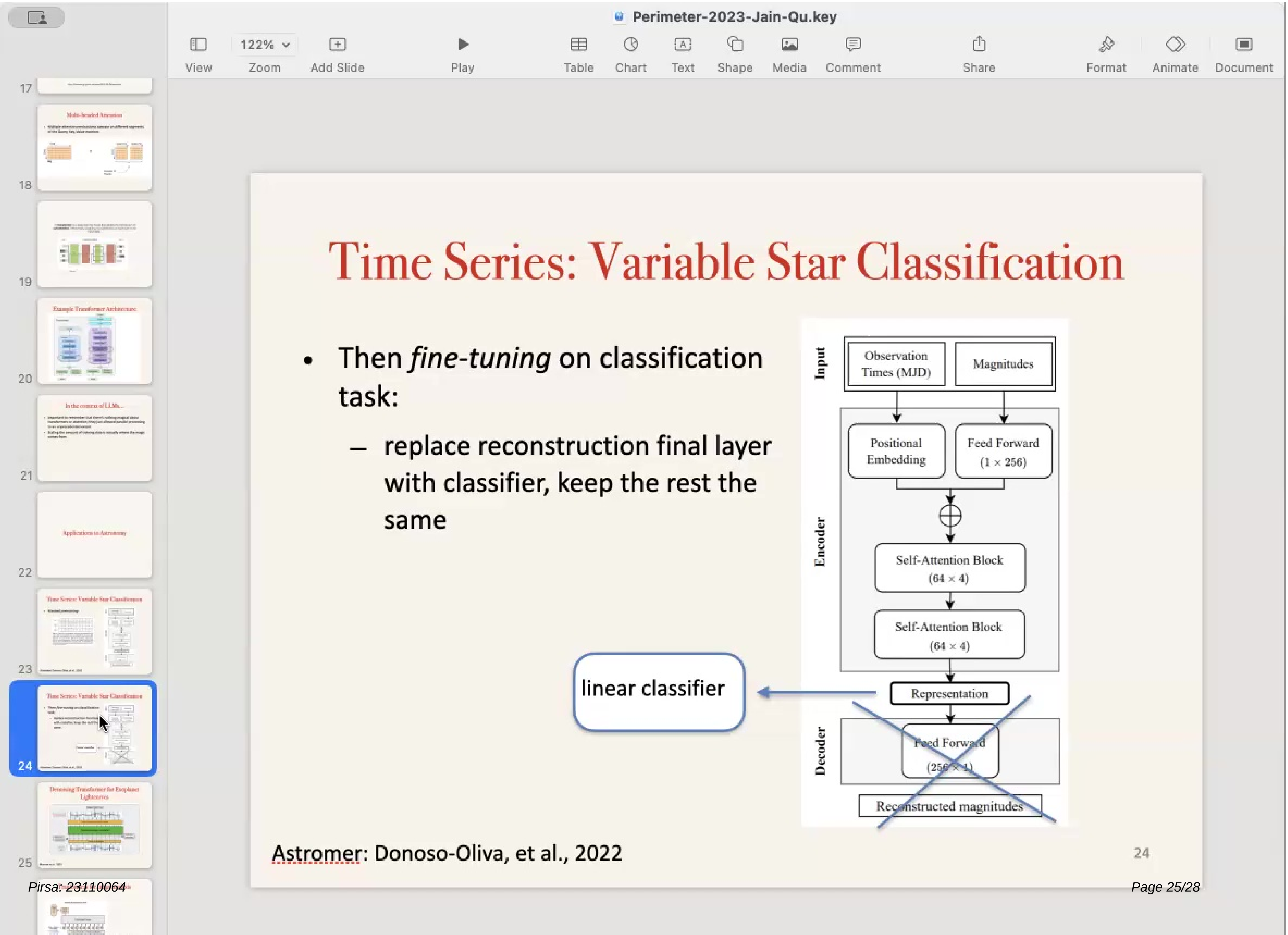
# Time Series: Variable Star Classification

- Masked *pretraining*:

| Inputs | $X_0$ | $X_6$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Attention Mask | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Loss Mask | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |

**Fig. 5.** Final input composition. Following the example in Figure 4, 20% of the masked values are replaced by random magnitudes while changing the 1's in the attention mask vector to 0's. Similarly, we make another 20% of the masking visible, keeping the actual observations. Doted line squares indicate both random and real observations. At this point, we should keep a second mask containing the initial 50% masked values to evaluate the loss function.

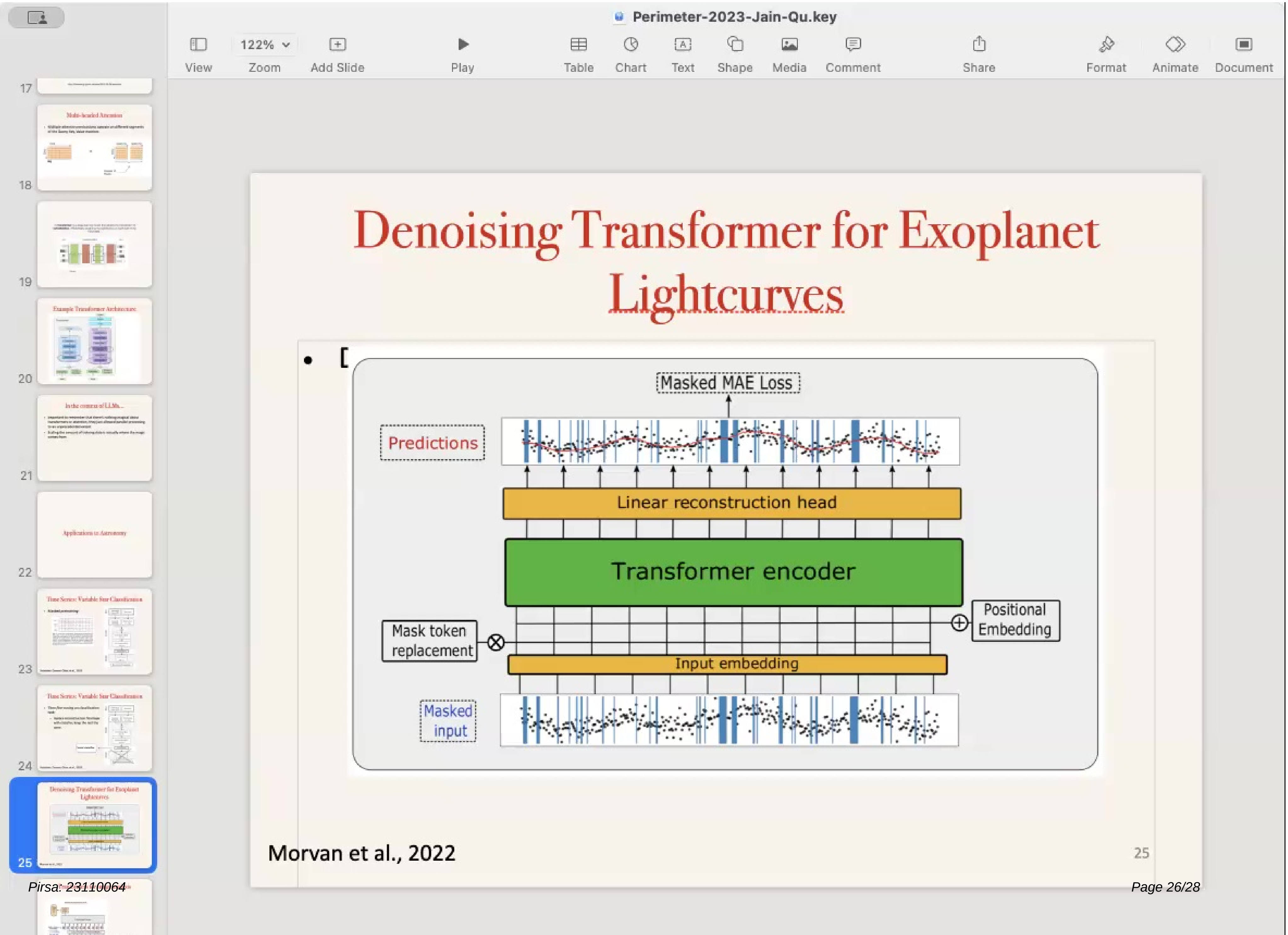
Astromer: Donoso-Oliva, et al., 2022



# Time Series: Variable Star Classification

- Then *fine-tuning* on classification task:
  - replace reconstruction final layer with classifier, keep the rest the same

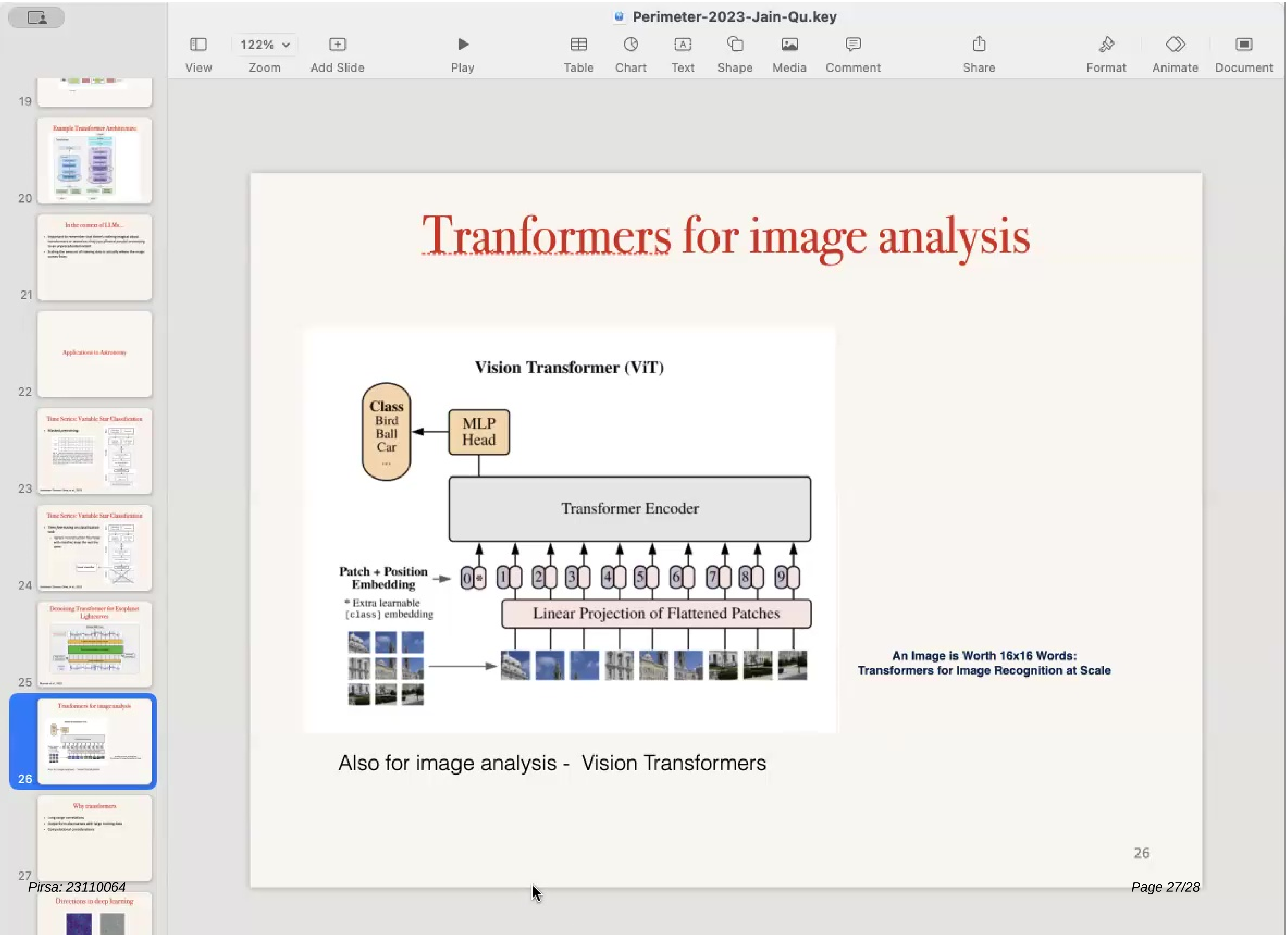
Astromer: Donoso-Oliva, et al., 2022



# Denoising Transformer for Exoplanet Lightcurves

- [

Morvan et al., 2022

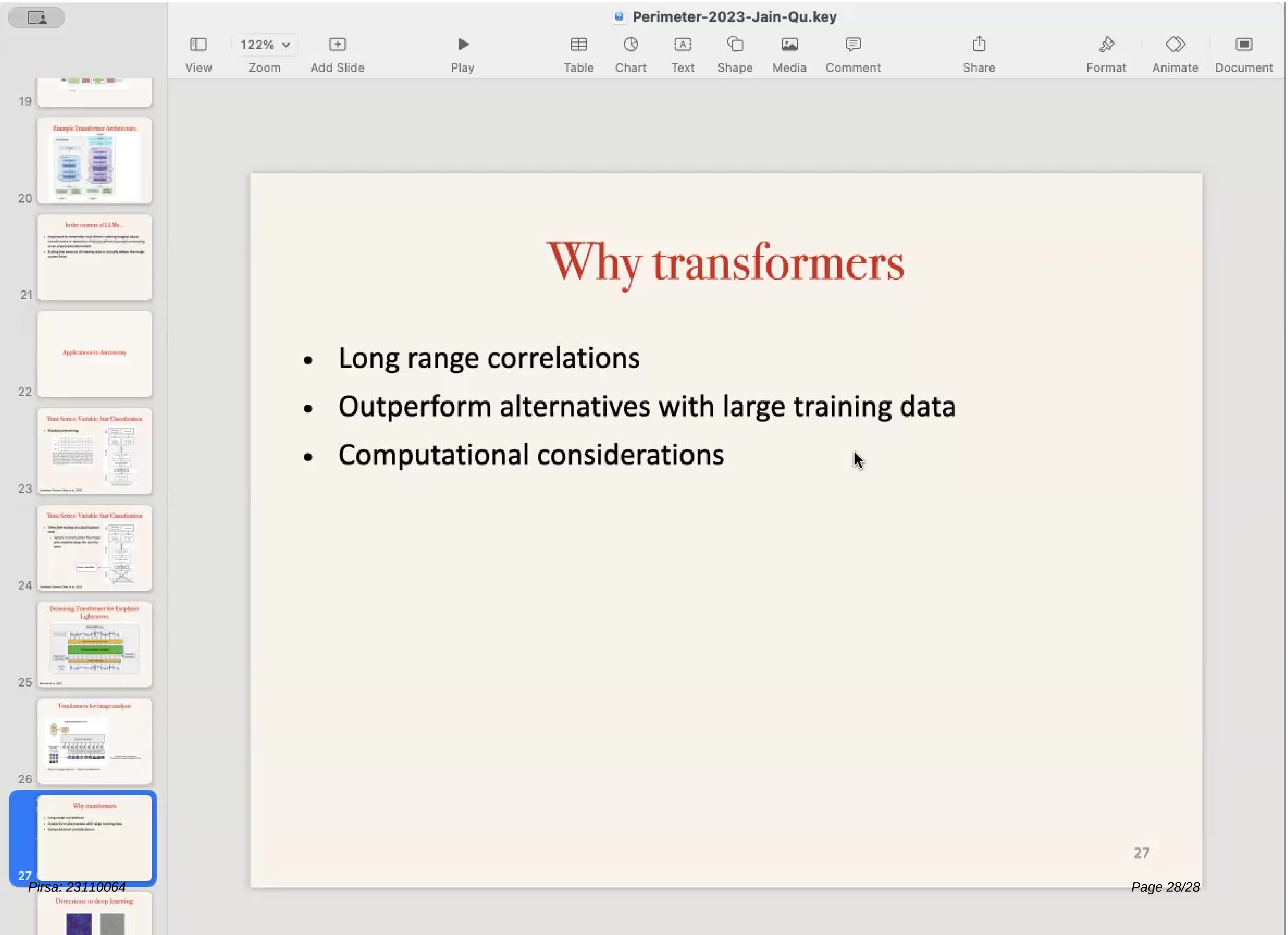

Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
19
Example Transformer Architecture
20
In the context of LLMs...
21
Applications in Astronomy
22
Time Series: Variable Star Classification
23
Time Series: Variable Star Classification
24
Denoising Transformer for Exoplanet Lightcurves
25
Transformers for image analysis
26
Why transformers
27
Directions in deep learning
Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

# Tranformers for image analysis

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Also for image analysis - Vision Transformers

26


Perimeter-2023-Jain-Qu.key
View
122%
Zoom
Add Slide
Play
Table
Chart
Text
Shape
Media
Comment
Share
Format
Animate
Document
19
Example Transformer Architecture
20
In the context of LLMs...
21
Applications to Astronomy
22
Time Series: Variable Star Classification
23
Time Series: Variable Star Classification
24
Denoising Transformer for Exoplanet Lightcurves
25
Transformers for image analysis
26
Why transformers
27
Directions in deep learning

# Why transformers

- Long range correlations
- Outperform alternatives with large training data
- Computational considerations

27