



Kristianstad  
University  
Sweden



International Conference  
**NEW PERSPECTIVES  
in SCIENCE EDUCATION**



# The Chronological Ascent to a Spatially Grounded World Model: Merging Geometric Architectures with Large Language Models to Meet Future Education Applied to Industry 6.0

*Charlotte Sennersten & Kamilla Klonowska*

Computer Science Department,  
Kristianstad University, Sweden

## Content Outline

- **Ground truth** in reality (UNESCO perspective)
- A decade-long timeline (**2016-2026**)
- Geometry: **3D spatial relationships** alongside **natural and programming languages**
- From Leonardo da Vinci's 1514 geometry to modern computational methods
- World model(s) and graph database(s)
- **Points, voxels, and bounding boxes**
- Voxels, collision boxes, and **text-based queries within 3D space**
- **Systematisation of language and spatial concepts**
- Education 6.0 and Industry 6.0 frameworks
- **Challenges** in joint curriculum and syllabus development

*“This highlights an educational paradigm which indirectly points towards a disconnect between general generative AI models ‘appearing’ to understand the ‘reality’ that they do not understand expressed in UNESCO’s ‘Guidance for Generative AI in Education and Research’. This calls for a slight urgency since generative AI are not yet informed by observations of the real world while in science real-world observations constitute scientific ground truth.”*

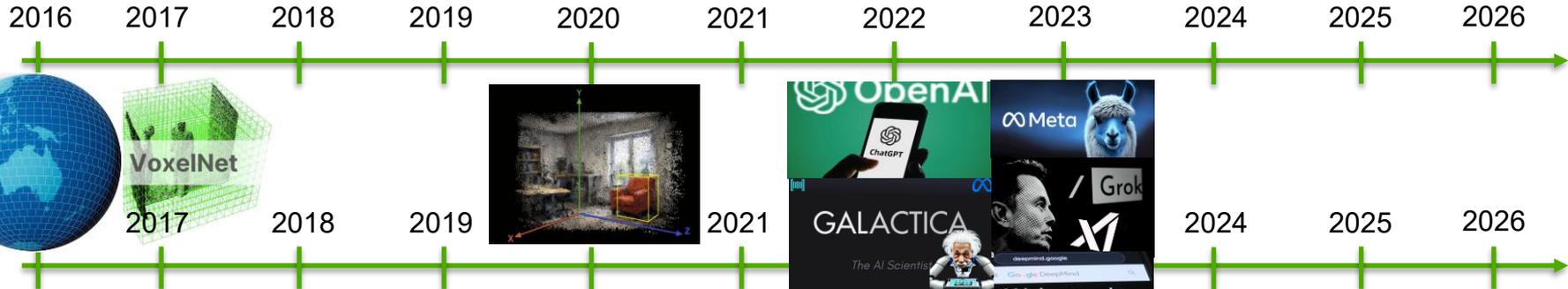
The disconnect between GenAI models ‘appearing’ to understand the text that they use and generate, and the ‘reality’ that they do not understand the language and the real world can lead teachers and students to place a level of trust in the output that it does not warrant. This poses serious risks for future education. Indeed, GenAI is not informed by observations of the real world or other key aspects of the scientific method, nor is it aligned with human or social values. For these reasons, it cannot generate genuinely novel content about the real world, objects and their relations, people and social relations, human-object relations, or human-tech relations. Whether the apparently novel content generated by GenAI models can be recognized as scientific knowledge is contested.

Miao., F., and Holmes., W., “Guidance for Generative AI in Education and Research”, UNESCO, 2023, <https://doi.org/10.54675/EWZM9535> p.16

# Ten Years (Snapshot) of Building Digitized 'Reality'

Academia

3DLLM



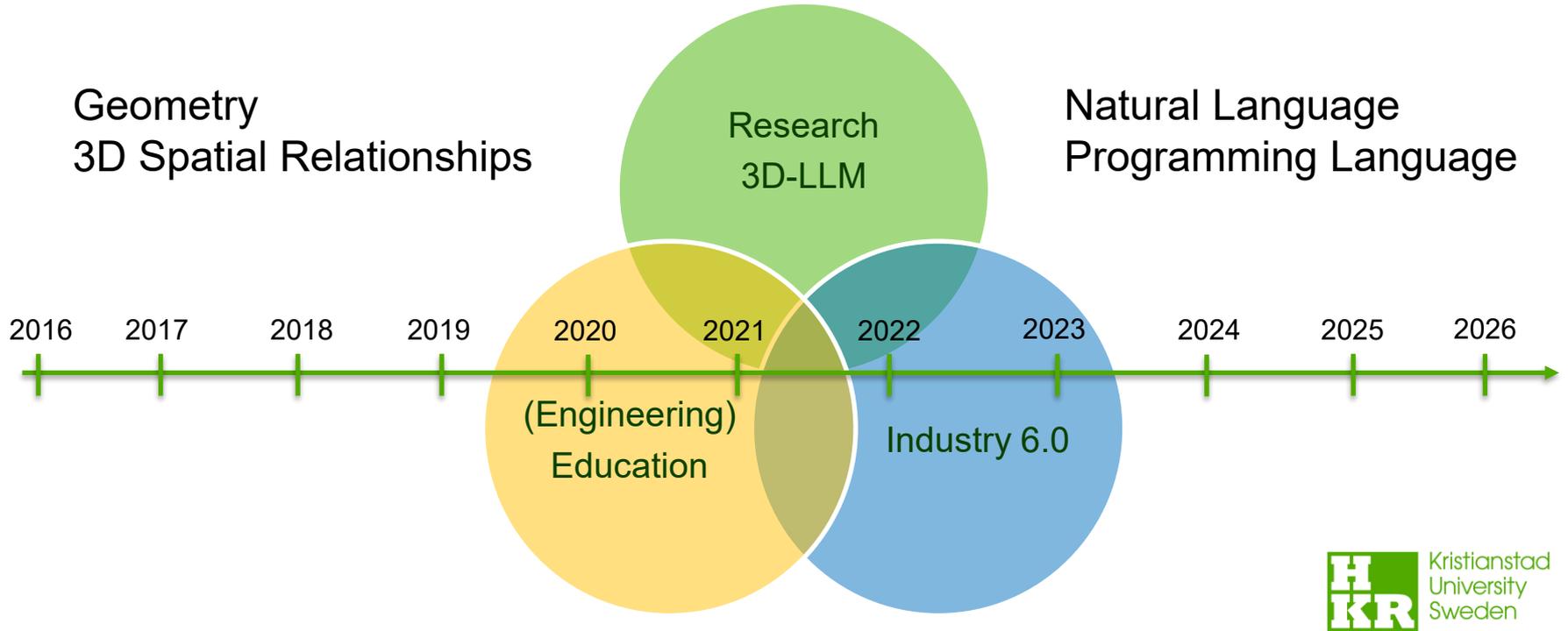
Industry

Computation  
Data Management

ScanQA  
ScanRefer

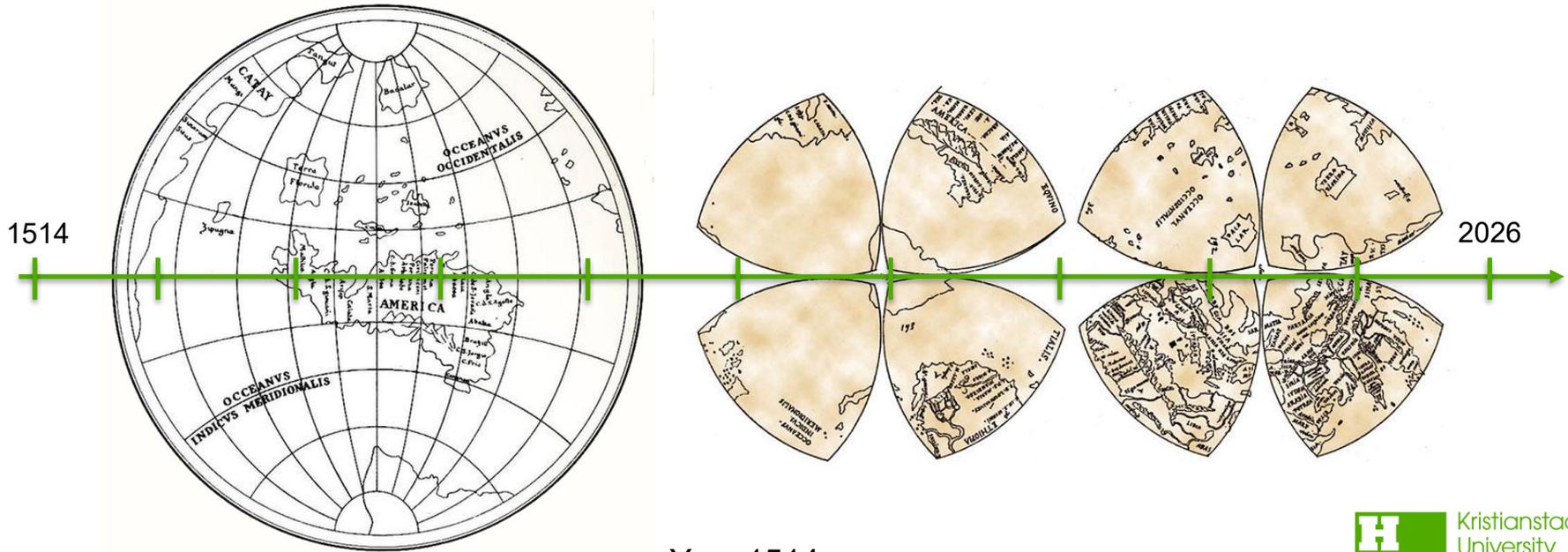
MATH-DT  
workshop  
in the USA

*This highlights an educational paradigm which indirectly points towards a disconnect between general generative AI models 'appearing' to understand the 'reality' that they do not understand expressed in UNESCO's 'Guidance for Generative AI in Education and Research'. This calls for a slight urgency since generative AI are not yet informed by observations of the real world. While in science real-world observations constitute scientific ground truth. A key aspect is to reach 3D structure comprehension including shape, orientation and location into practice into our current educations to meet this reality of ours. The worldwide shift from static, textbook-based instruction to spatially grounded World Models and Industry 6.0 innovations is creating an educational environment that nurtures curiosity, deep understanding, and scientific literacy across the globe, while preparing learners for an interconnected, intelligent future.*





The World Map utilizes a so-called "octant projection," where the Earth is divided into eight spherical triangles.



Year 1514

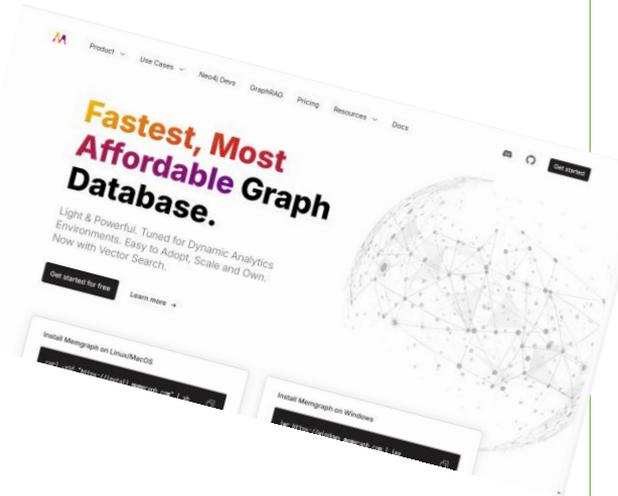
Article: Leonardo da Vinci's World Map by Christopher Tyler | Christopher Tyler



How do we map our world?  
For computational benefit and spatial indexation?



Sennersten et al., 2016





## How do we map/represent our world? Points (nodes), voxels, ...

Dividing 3D space into cubes (voxels) – while optimizing performance and speed.

End-to-End Learning  
data management to perceptual learning. Apple researchers in USA  
on of VoxelNet, the first end-to-end trainable network to unify feature  
diction from raw LIDAR point clouds [8]. The key innovation was the  
layer. This allowed the model to transform disordered points within a 3D  
volumetric representations. By removing the need for manual feature  
-view projections, VoxelNet proved that a neural network could learn  
f objects with varying geometries (e.g., pedestrians and cyclists) directly  
the "Voxel" as the standard unit for 3D deep learning, much like the pixel

of voxels within the VFE layer (see figure 2), several key components  
is defined as a volumetric, three-dimensional pixel; utilizing this concept

space is partitioned into a uniform grid.  
grid is designated as a voxel.  
d within the same voxel are grouped together.  
layer, it is necessary to:  
to their respective voxels.  
the point level within each voxel.  
atures into a fixed-length feature vector for each voxel.



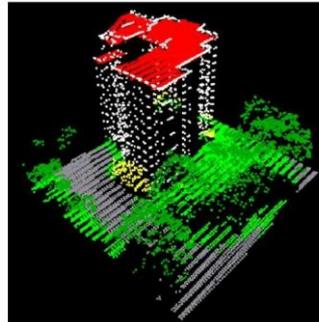
VoxelNet



operates on the raw point cloud and produces the 3D detection results using a single  
end-to-end learning network", source [8].

patial Understanding and Natural Language: Insights from ScanQA and

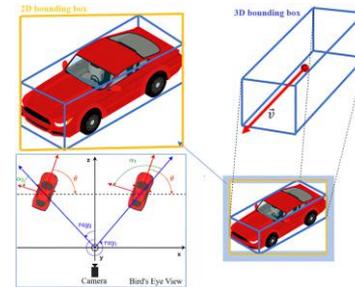
ation of 3D spatial knowledge with natural language processing, two significant  
ScanRefer, serve as prominent examples. Both projects conducted



Point Cloud



Voxel Cloud





NEW PERSPECTIVES  
in SCIENCE EDUCATION

2020, advanced the field of 3D vision-and-language understanding. **SceneQA** focused on 3D question answering within spatial contexts, while **SceneReloc** addressed 3D object localization in RGB-D scans using natural language descriptions. These endeavors pursued distinct objectives in examining the relationship between language and 3D environments and demonstrated that 3D large language models (3D-LLMs) significantly outperformed existing baseline approaches in both 3D question answering and spatial grounding tasks. Furthermore, these models exhibited the capability for complex task decomposition, including providing comprehensive instructions for identifying ingredients within specific 3D spaces and facilitating navigation to target objects through conversation-based waypoints.

**3.1 SceneQA**  
The objective of this project carried out in Japan was to accurately identify and localize the referenced object by predicting its 3D bounding box within a given scan [9]. For instance, when presented with a point cloud of a room and a description such as "the red chair next to the window", the model generated the corresponding 3D bounding box for that specific chair (see figure 3). To effectively integrate the language with 3D object data, both "language embedding" and "3D geometry fusion" techniques were employed, enabling the model to interpret the provided description as well as the spatial context of the scene. **SceneQA** data is available at GitHub [10].

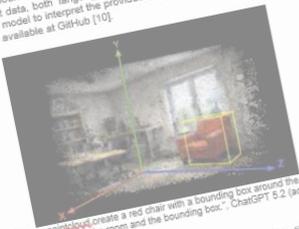


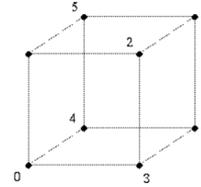
Fig. 3. Prompt "From a pointcloud create a red chair with a bounding box around the chair standing by the window. Show x, y, and z for the room and the bounding box." ChatGPT 5.2 (accessed 19/12/2024)

**3.2 SceneRefer**  
The objective of the German Canadian project was to move beyond localization and address inquiries related to 3D scenes. Each input consisted of a complete 3D scan, obtained from an RGB-D setup such as **ScanNet**, accompanied by a question formulated in natural language [11]. To facilitate a clearer understanding, the following table (table 1) illustrates the functioning of both approaches. **SceneReloc** data is available at GitHub [12].

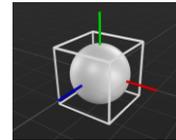
Feature	SceneQA	SceneRefer
Primary Task	Answer questions about a 3D scene	Localize a referenced object in 3D
Input	3D scan + natural language question	3D scan + natural language reference
Output	Answer text + (optional) object bounding boxes	3D bounding box of referenced object
Focus	Scene reasoning and understanding	Object grounding via language
Dataset Size	~41 k QA pairs	~81 k object descriptions
Complexity	Reasoning about relations + answering questions	Finding one referenced object

In summary, the two methodologies are outlined as follows:  
**SceneQA**: "Respond to a question pertaining to the entirety of this 3D scene."  
**SceneReloc**: "Identify the object referenced within this 3D scan."

**Definition:**  
A voxel is a volume element, typically a cubic cell.



**Collision Box**



WHERE is the object spatially when asking questions via language?





## Systemize up language via tokenization:

1. Citations [START\_REF] and [END\_REF]
2. Step-by-step reasoning: <work> (internal working memory context)
3. Mathematics: ASCII operations into individual characters: !"#%&\*+,-./:;<=> ?~^\_` and parantheses are ( ) [ ] { }.
4. Numbers: splitting digits into individual tokens, 737612.62 -> 7, 3, 7, 6, 1, 2, ., 6, 2.
5. SMILES formula
6. Amino acid sequences: [START\_AMINO] and [END\_AMINO], MIRLGAPQTL -> M, I, R, L, G, A, P, Q, T, L.
7. DNA sequences: [START\_DNA] and [END\_DNA], CGGTACCCTC -> C, G, G, T, A, C, C, C, T, C.



with spatial points/nodes



**New Perspectives  
in Science  
Education**

# International Conference NEW PERSPECTIVES in SCIENCE EDUCATION



5. **VoxelNet**: Enhancing 3D Object Detection through Sparse Computational Efficiency

Observing that dense prediction heads are often computationally inefficient, given that typically less than 1% of the 3D space contains relevant objects, researchers from Hong Kong introduced **VoxelNet** (15) in 2023. Instead of using dense grids of data or manually designed reference points known as "anchors", this network used a sparse approach meaning it only processed the parts of the 3D space where actual data points (voxels) contained points exist. By doing so, it could efficiently identify and predict the presence of 3D objects directly from the information contained within these voxels, making the process both faster and more resource efficient. **VoxelNet** showed that a simpler approach such as adding extra down-sampling steps to increase the area each voxel could "see" could then perform better than complicated, multi-stage detection systems.

6. **Digital Twins and Mathematical Rigor**

Advancements in 3D modelling and LLMs have increased focus on DTs, which virtually replicate physical assets in real time. In digital twins, geolocation data, longitude and latitude, obtained from example using GPS for location. Mapping and surveying often use grid coordinates along the x and y-axis, northings indicate distance along the y-axis while eastings show distance along the x-axis. The three-dimensional nature of Earth is characterized by its curvature, distinguishing it from a flat, two-dimensional surface.

The 2023 MATH-DT workshop in the USA (16) underscored the need for foundational mathematical advances to transition from generic physical laws to personalized applications. Researchers emphasized that DTs rely on solving inverse problems, such as using sensor data to determine world properties. This has brought attention to what is called **Uncertainty Quantification** (UQ), a foundational mathematical discipline required for the development and operation of DTs. UQ involves virtual models accurately reflect reality and support reliable decision-making, including initial modeling, identifying and managing unknowns at every stage of the industrial pipeline.

## Academia 6.0

## Education 6.0



## Industry 6.0

*Education 6.0 is an emerging paradigm that integrates AI, VR/AR, and advanced technology to create a personalized, adaptive, and ethical learning experience. It shifts the focus from mere knowledge acquisition to creation ("Maker Knowledge") and sustainability, aiming to prepare students for the flexible and automated job market (Industry 6.0).*

*Industry 6.0 is an emerging, futuristic industrial paradigm focusing on symbiotic, self-evolving systems that integrate advanced AI, robotics, biotechnology, and quantum computing with human intelligence for total sustainability.*



# International Conference NEW PERSPECTIVES in SCIENCE EDUCATION



## Our Joint International and National Curricula Development Challenges to support Scientific and Engineering Reasoning - A System Approach



**Academia 6.0** Education 6.0 is an emerging paradigm that integrates AI, VR/AR, and advanced technology to create a personalized, adaptive, and ethical learning experience. It shifts the focus from mere knowledge acquisition to creation ("Maker Knowledge") and sustainability, aiming to prepare students for the flexible and automated job market (Industry 6.0).



REFERENCES

[1] Miao, F., and Holmes, W., "Guidance for Generative AI in Education and Research", UNESCO, 2023, <https://doi.org/10.54675/EWZM9530>

[2] Semnersten, C., Davie, A., and Lindley, C., "VoiceNet: An Agent Based System for Spatial Data Analytics", COGNITIVE 2016, The Eight International Conference on Advanced Cognitive Technologies and Applications, Rome, Italy, 2016, download\_fulltext.pdf

[3] Semnersten, C., Lindley, C., Evans, B., Grace, A., and Wise, M., "VoiceNet: An Agent Based System for Spatial Data Analytics", COGNITIVE 2016, The Eight International Conference on Advanced Cognitive Technologies and Applications, Rome, Italy, 2016, download\_fulltext.pdf

[4] VoiceNet: 4D data integration platform, <https://www.voice-net.com/>, accessed 13/02/2024

[5] Semnersten, C., Leccobbe, M., Panetto, H., and De Santis, M., "Digital Twin Paradigm: A Systematic Literature Review", Computers in Industry, Volume 130, 2021, <https://doi.org/10.1016/j.compind.2021.103469>

[6] Semnersten, C., Evans, B., and Lindley, C., "VoiceNet: A Geo-Located Spatial Temporal Software Application", The Eleventh International Conference on Advanced Technologies and Applications, Venice, Italy, 2019, [https://www.researchgate.net/publication/344129193\\_VoiceNet\\_A\\_Geo-Localized\\_Spatial\\_Temporal\\_Software\\_Application](https://www.researchgate.net/publication/344129193_VoiceNet_A_Geo-Localized_Spatial_Temporal_Software_Application)

[7] Semnersten, C., Milford, M., and Bayazit, T., "PointCloud3D: Crack Detection in Applications", Personalized and Adaptive Learning, Point-Cloud-Based Deep Neural Network, <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[8] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[9] Azuma, Y., Miyazawa, M., and Takahashi, Y., "Understanding of 3D Object Localization in Robot Navigation", <https://doi.org/10.1109/ICRA48129.2021.9551115>, 2021, <https://arxiv.org/abs/2011.04472>

[10] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[11] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[12] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[13] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[14] Chen, Y., "World Models in Digital Twins (MATH-DT)", <https://doi.org/10.1109/ICDT48129.2021.9551115>, 2021, <https://arxiv.org/abs/2021.02078>

[15] Chen, Y., "World Models in Digital Twins (MATH-DT)", <https://doi.org/10.1109/ICDT48129.2021.9551115>, 2021, <https://arxiv.org/abs/2021.02078>

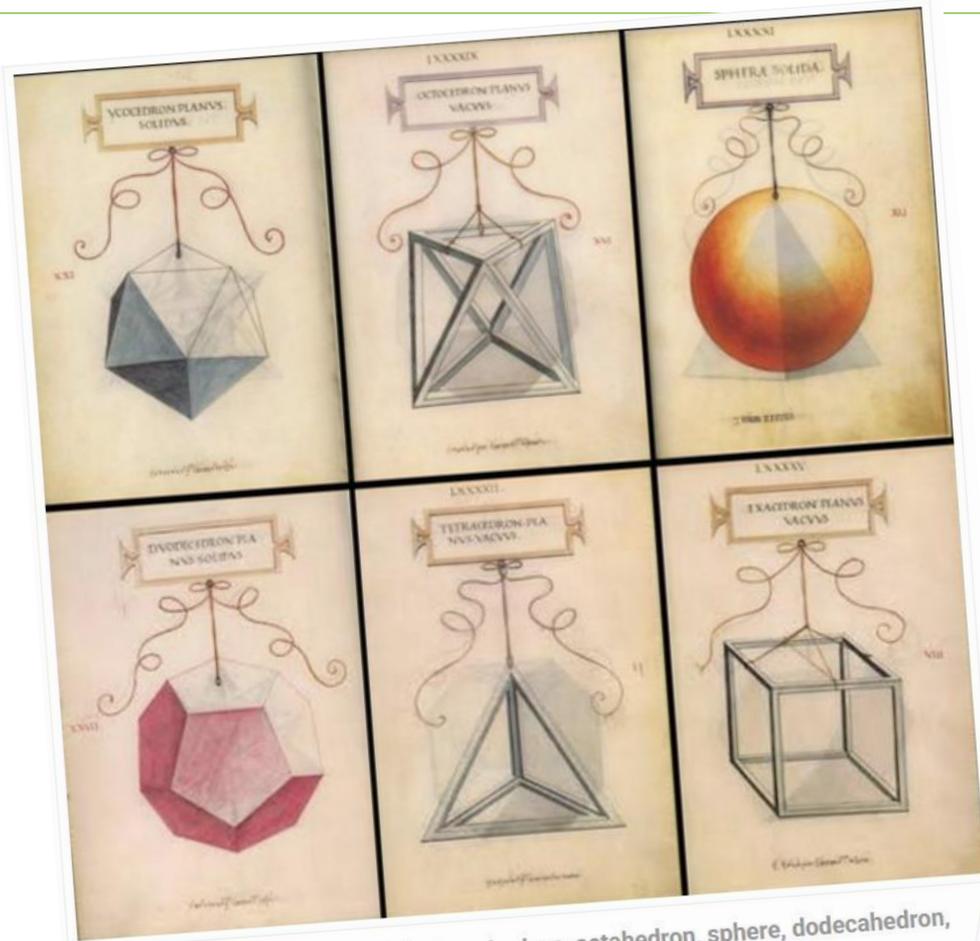
[16] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[17] Anil, H., "Mathematical Opportunities in Digital Twins (MATH-DT)", <https://doi.org/10.1109/ICDT48129.2021.9551115>, 2021, <https://arxiv.org/abs/2021.02078>

[18] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[19] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>

[20] Semnersten, C., Milford, M., and Bayazit, T., "Point-Cloud-Based Deep Neural Network", <https://doi.org/10.48550/arXiv.2015.04472>, 2021, <https://arxiv.org/abs/2015.04472>



Shapes' names from top left: icosahedron, octahedron, sphere, dodecahedron, tetrahedron, cube

