Generating Synthetic Data with Python

Alberto Labarga

Barcelona PyDay November 2023

Noé Casas

Vyron Vasileiadis

Adrià Blanco

Guillermo Abad

Javier Sabariego

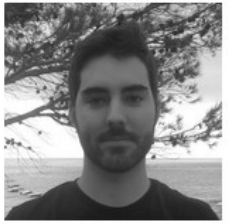
Christian Adell

Pavel Sulimov

Ferran Jovell

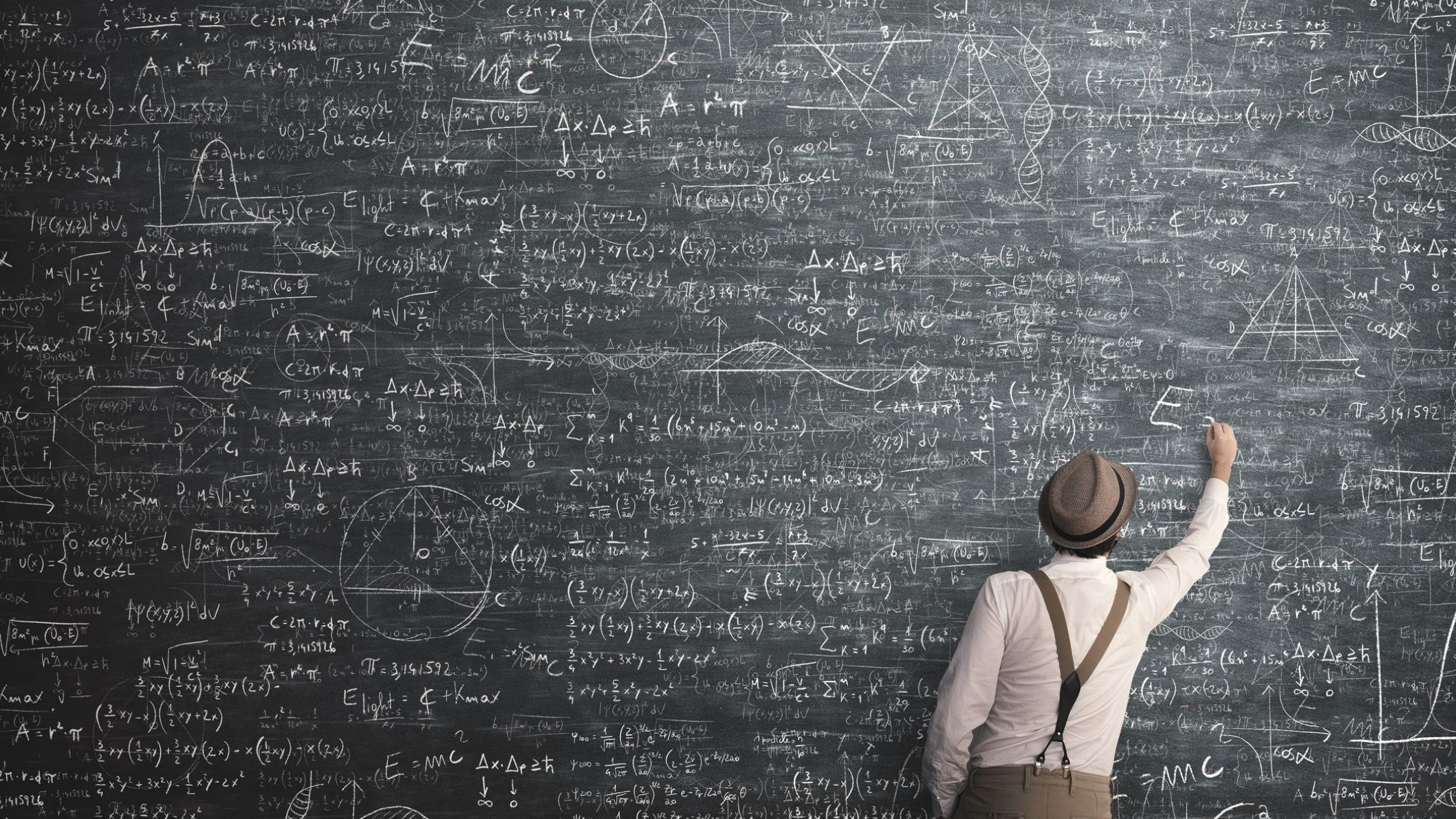
Alberto Labarga

Diego Giaquinta



Oriol Abril Pla

Gabriel de Maeztu



alabarga / **synthetic-data-pybcn**

Type / t

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

synthetic-data-pybcn Public

Pin | Unwatch 1

main | 1 branch | 0 tags | Go to file | Add file | Code

alabarga Add files via upload — eceee2f now — 3 commits

data	Add files via upload	now
models	Add files via upload	now
.gitignore	Initial commit	4 days ago
LICENSE	Initial commit	4 days ago
README.md	Update README.md	4 days ago
requirements.txt	Add files via upload	now
synthetic_data_generation.ipynb	Add files via upload	now

https://github.com/alabarga/synthetic-data-pybcn

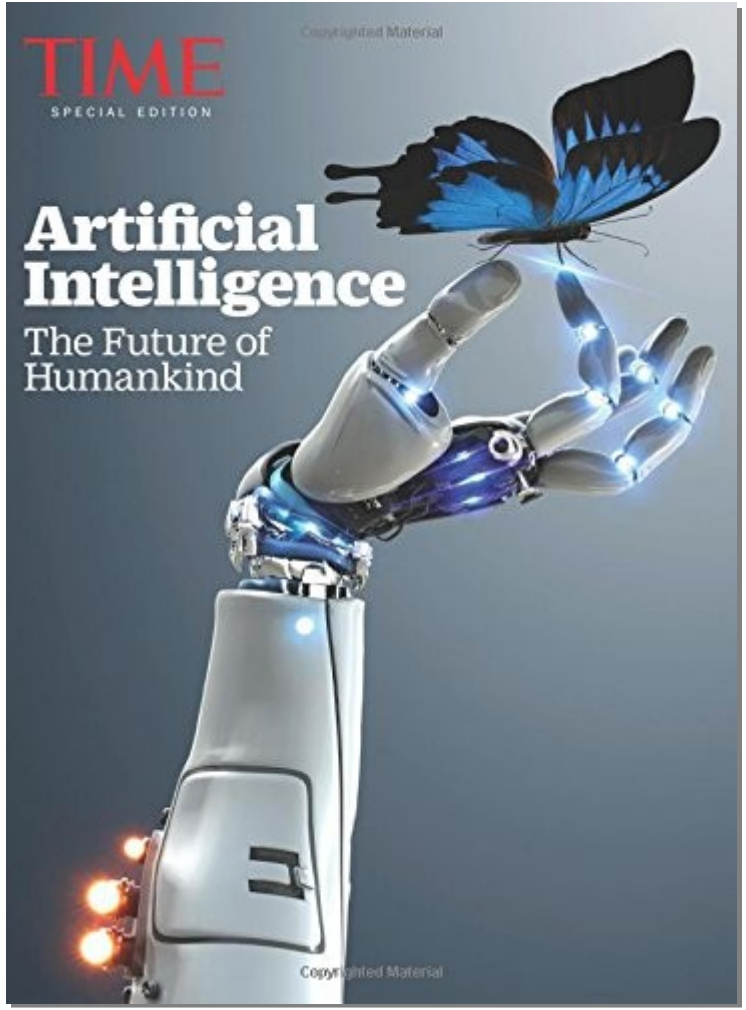
Telecommunications Engineer
Head of Biomedical Data Hub @BSC
More than 20 years crunching data

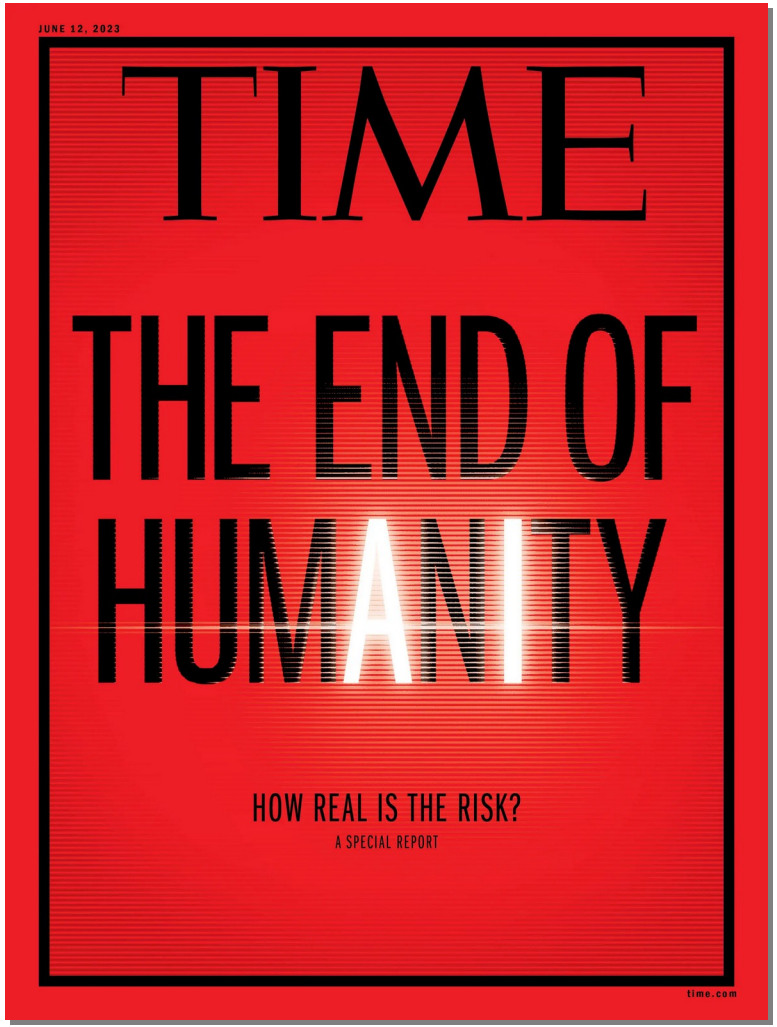
open data - open source – open science

I am Alberto

alabarga
alabarga
/in/albertolabarga

Barcelona Supercomputing Center
Centro Nacional de Supercomputación


TIME
SPECIAL EDITION
Artificial Intelligence
The Future of Humankind


JUNE 12, 2023
TIME
THE END OF HUMANITY
HOW REAL IS THE RISK?
A SPECIAL REPORT
time.com

FL

Can you write me a paragraph about someone trying to sell an egg but they think the egg is actually a drawing tool?

Sure! Here's a paragraph:

The person was trying to sell their egg at a street market, but they thought it was actually a drawing tool. They were confidently telling potential buyers about the egg's ability to create beautiful, detailed sketches with its delicate, curved shape. Some curious passersby stopped to listen and even gave the egg a try, but they were quickly disappointed to find that it was just an ordinary egg and not a drawing tool at all. The person was embarrassed and quickly moved on to the next item in their sales pitch.

FL

Can you write a version of their sales pitch?

Sure! Here's a possible sales pitch for the egg-turned-drawing-tool:

Hello, friends! I have a very special item to show you today. This is no ordinary egg - it's a one-of-a-kind drawing tool. With its delicate, curved shape, this egg is perfect for creating beautiful, detailed sketches. Just hold it in your hand and let your creativity flow. You'll be amazed at the incredible art you can create with this amazing egg. And the best part? It's only $5! Don't miss out on this incredible deal. Come and see for yourself how this egg can unleash your inner artist.

FL

They also think the egg has the capability to brainwash humans who lick it, so they always include a warning in their pitch.

Sure, here's a possible warning to include in the sales pitch for the egg-turned-drawing-tool:

Warning: This egg has the unique ability to brainwash humans who lick it. Please be careful when handling this egg and do not lick it under any circumstances. We want you to enjoy the incredible benefits of this egg as a drawing tool, but we also want to ensure your safety and well-being. Thank you for understanding.

Reality-centric AI

- Transform **healthcare** systems?
- Alleviate **traffic** control?
- Solve the **energy** crisis?
- Prevent **climate** catastrophes?
- Avoid the next **financial** crisis?
- Transform **education**?

Reality-centric AI

- Not just a few breakthroughs away…..
- Because of the complexity of the real-world – a different AI/ML paradigm
- Solving these complex human-centric problems is our biggest task as AI researchers!

Reality-centric AI

- AI which aims to solve **real-world problems**
- AI which operates **effectively** and **accountably** given the inherent and unavoidable complexities of the real world
- AI which **empowers**, and does not marginalize humans

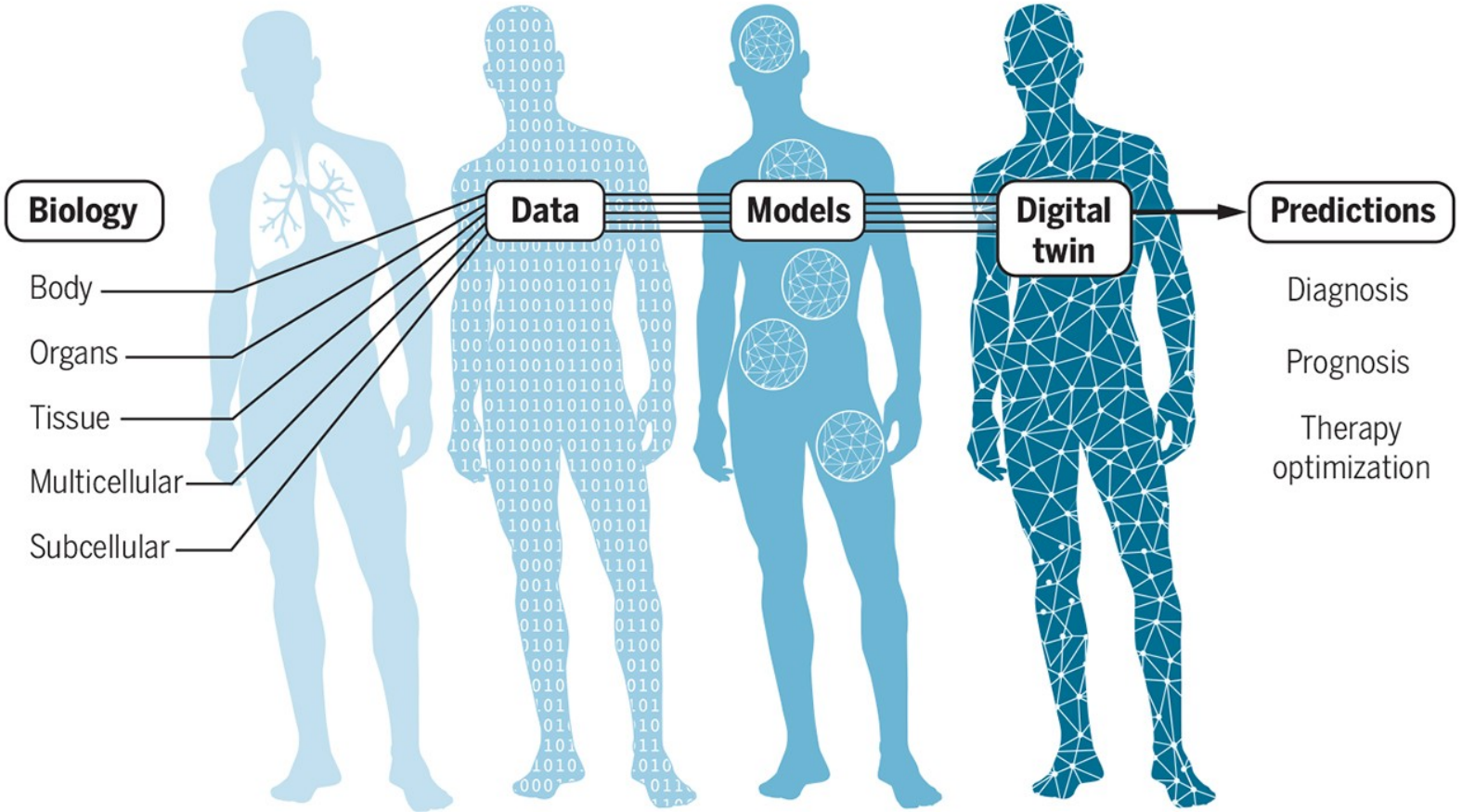
Destination **EARTH**

Barcelona
Supercomputing
Center

vCity: a human-centric platform for urban digital twins

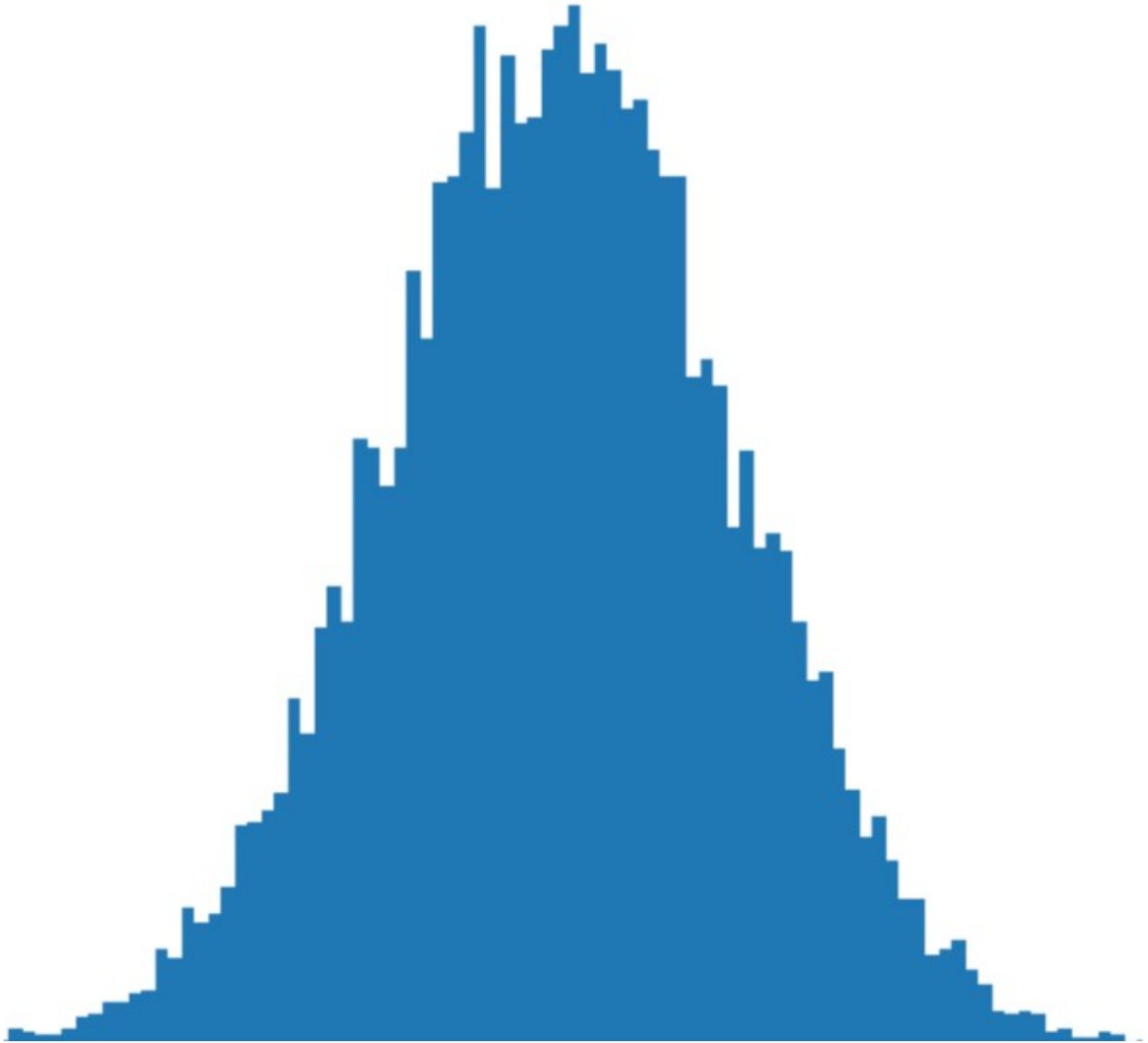
Building a personalized digital twin

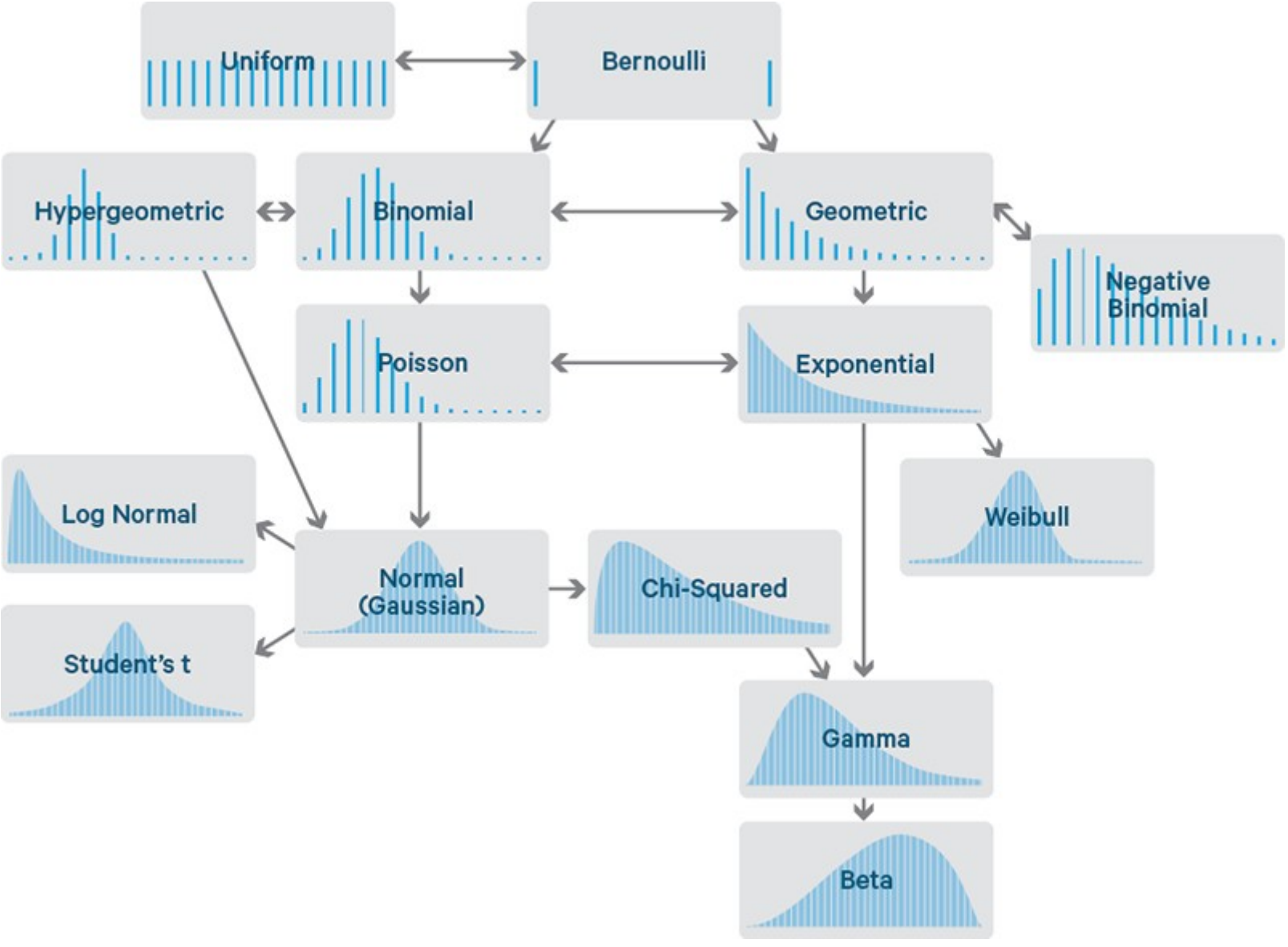
Data from multiple scales are needed to build computational representations of biological processes and body systems that are affected by viral infection. These submodels are integrated and personalized with clinical data from individual patients. The digital twin can then be used to derive predictions about diagnosis, prognosis, and efficacy and optimization of therapeutic interventions.



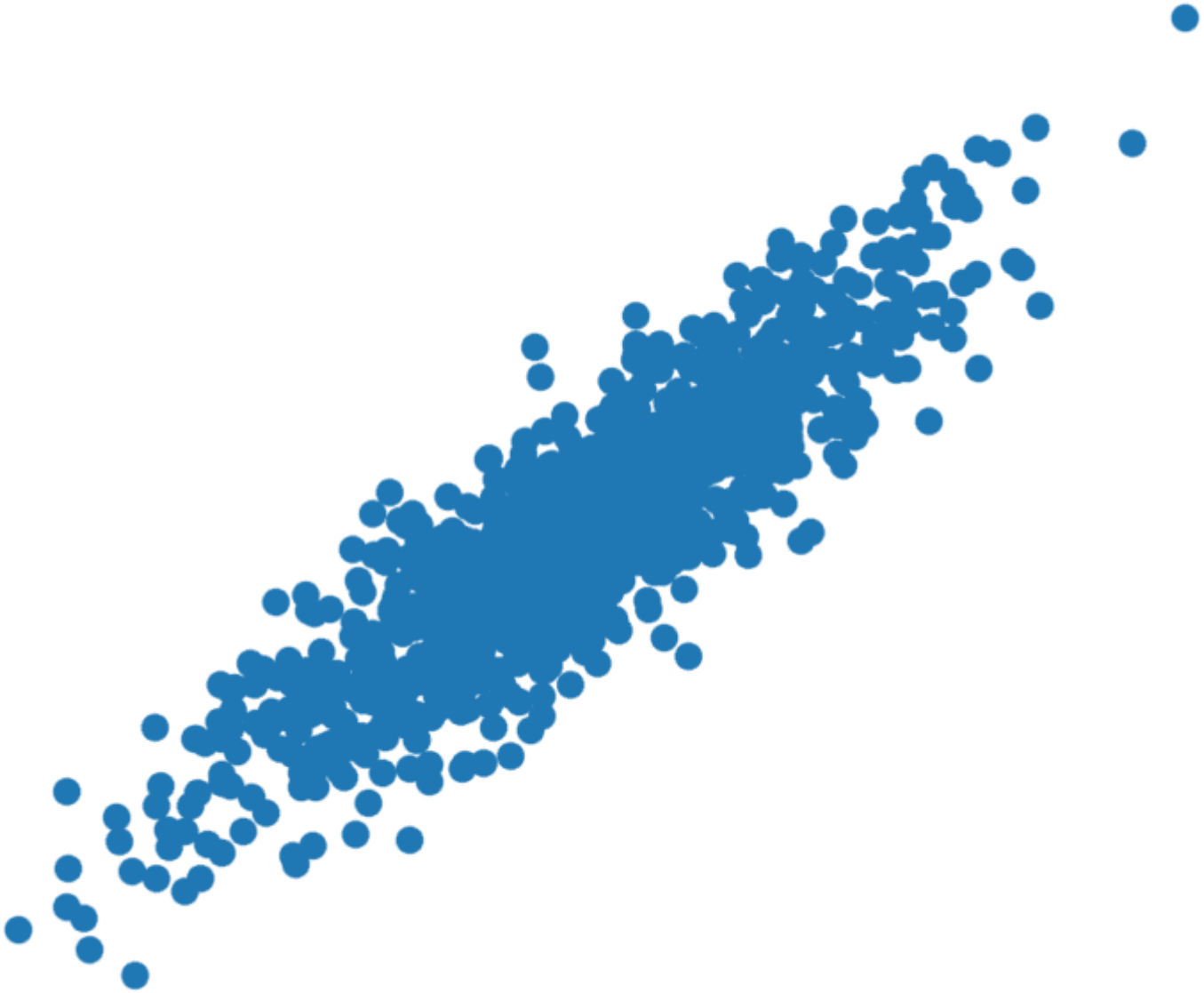
42

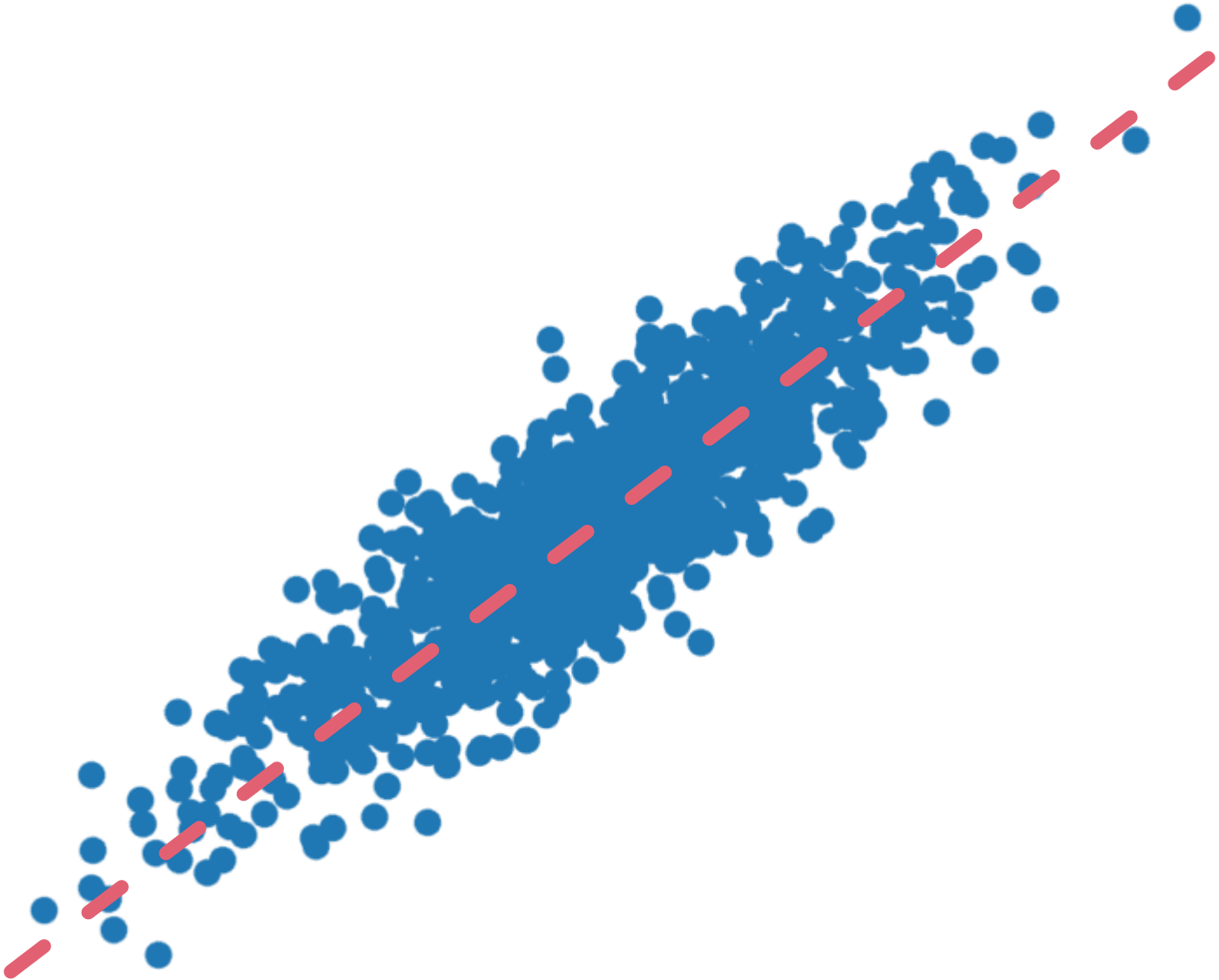
[55, 45, 48, 33, **42**, 83, 21, 75, 30, 24]

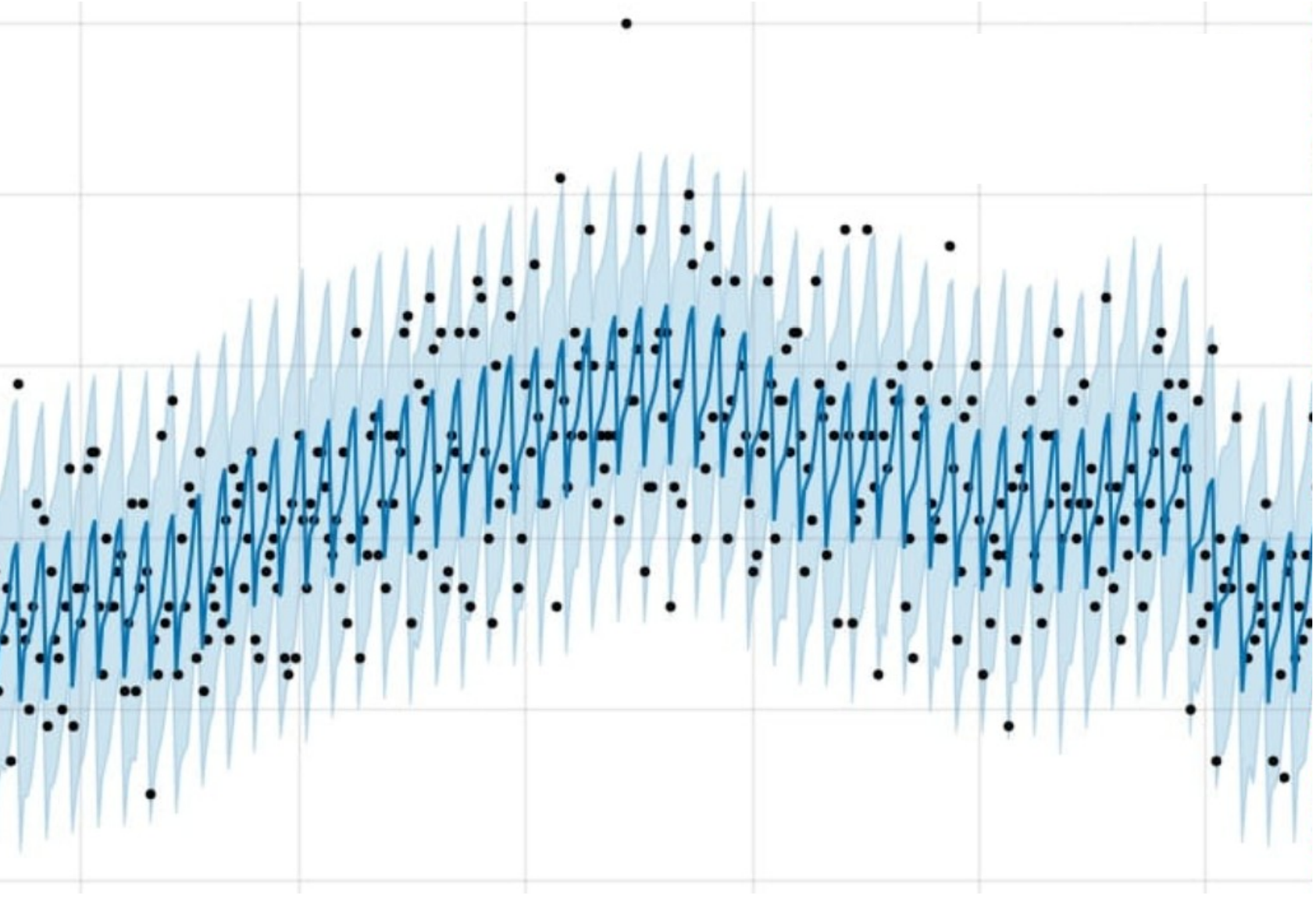


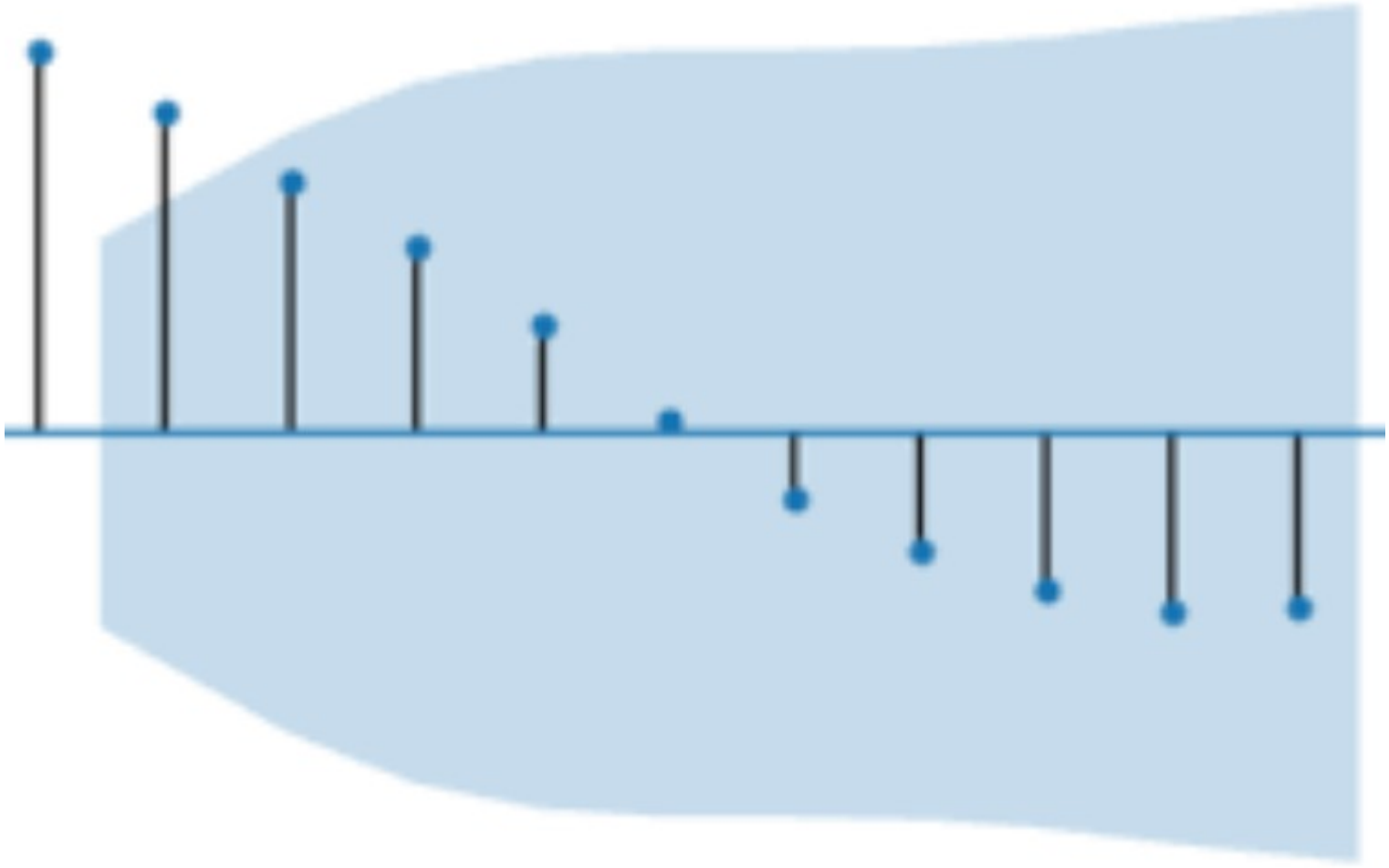

Uniform
Bernoulli
Hypergeometric
Binomial
Geometric
Negative Binomial
Poisson
Exponential
Log Normal
Weibull
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta

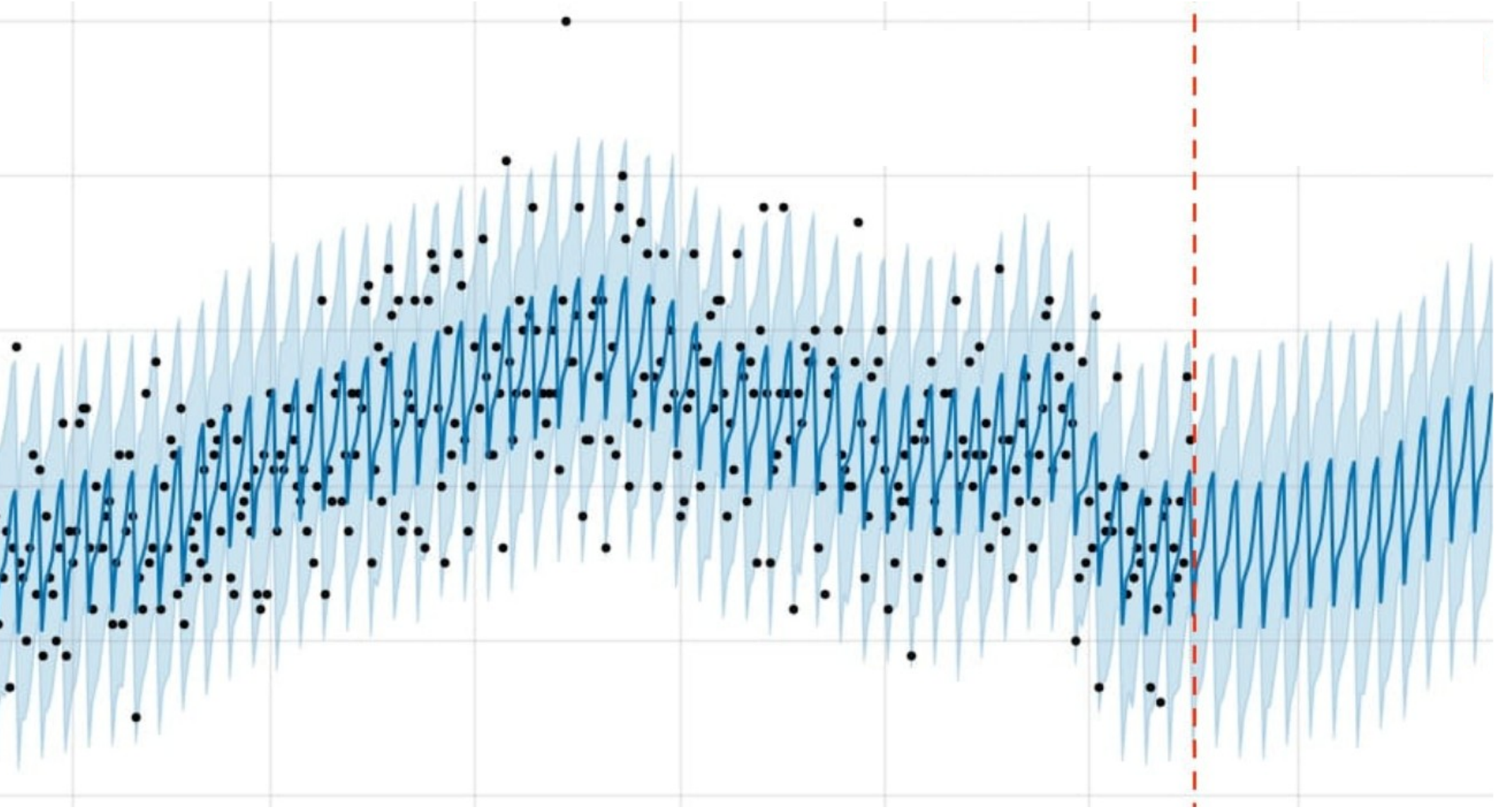
```
[55, 45, 48, 33, 42, 83, 21, 75, 30, 24]

[93, 44, 72, 83, 71, 77, 56, 43, 84, 82]
```

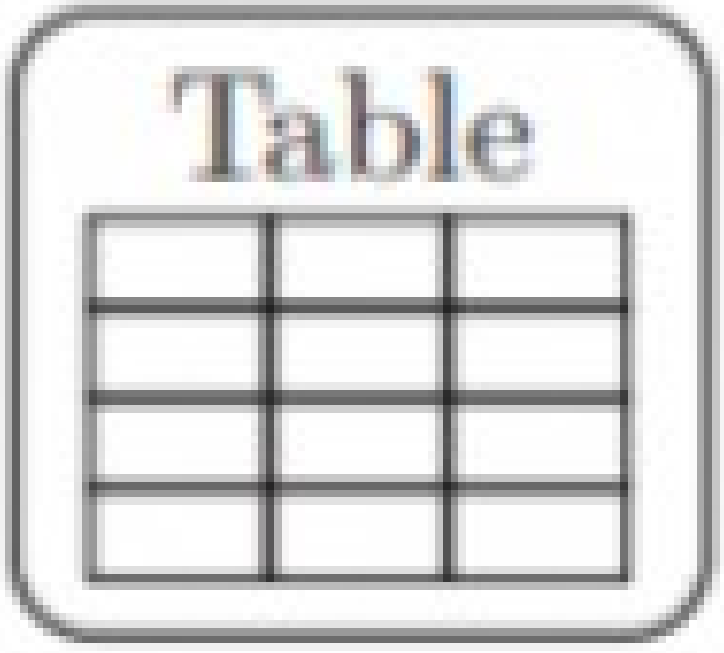




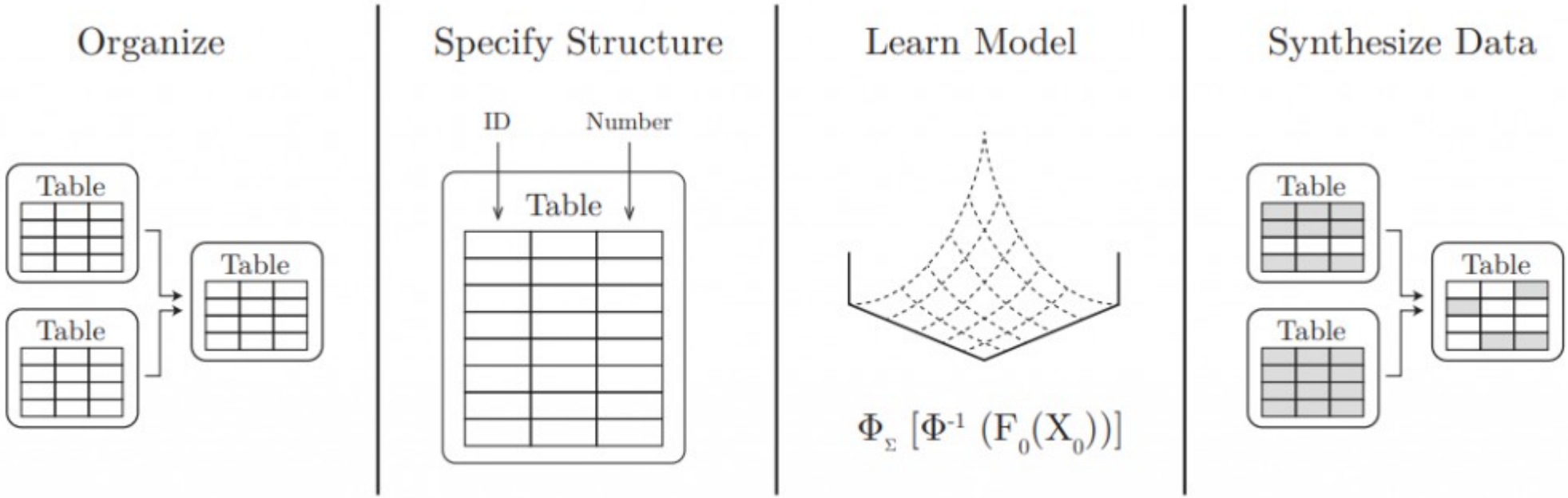




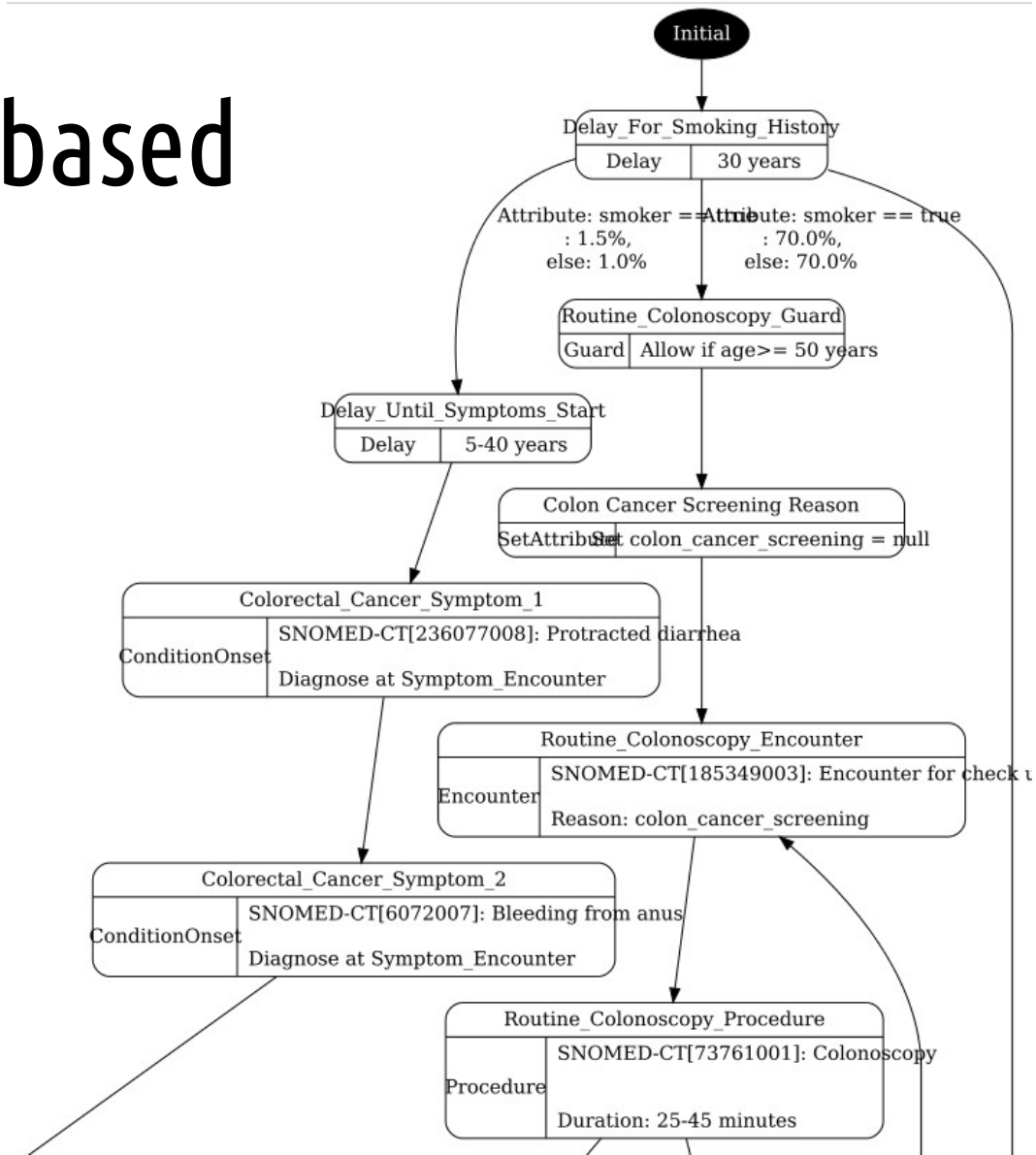




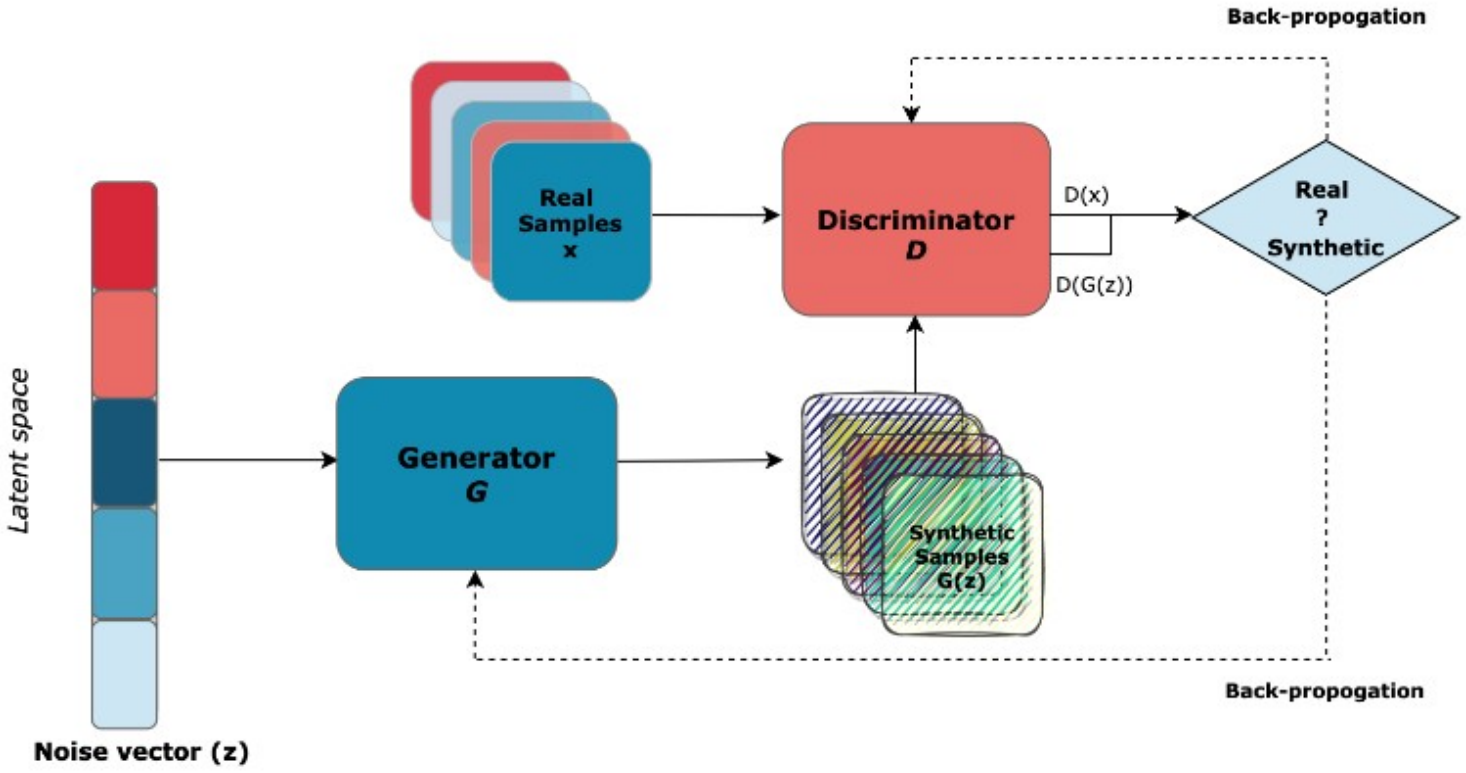
Table

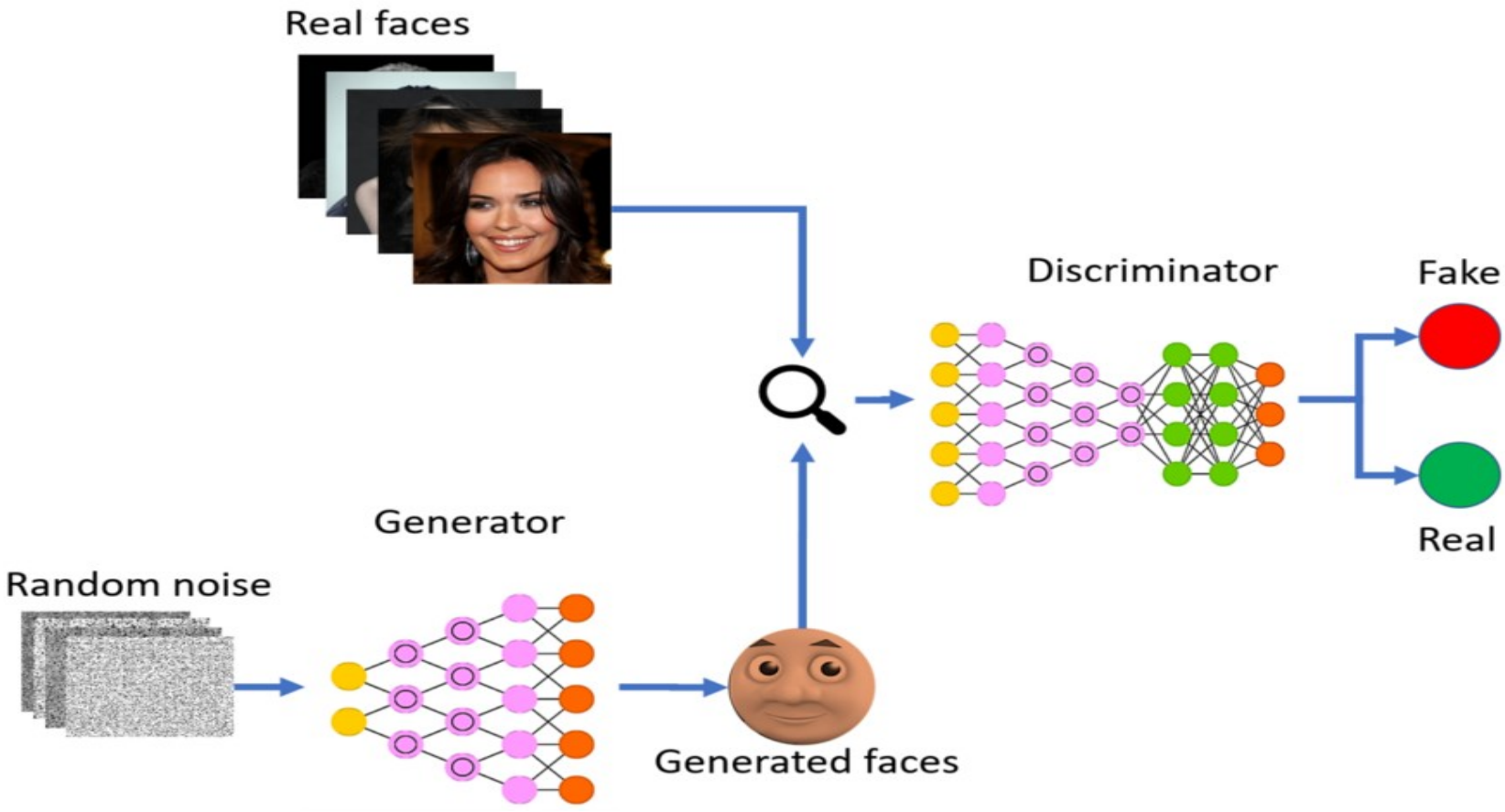

Organize
Table
Table
Table
Specify Structure
ID
Number
Table
Learn Model
$\Phi_{\Sigma}\,[\Phi^{-1}\,(F_0(X_0))]$
Synthesize Data
Table
Table
Table

Rule-based


Initial
Delay_For_Smoking_History
Delay
30 years
Attribute: smoker == true
: 70.0%,
else: 70.0%
Routine_Colonoscopy_Guard
Guard
Allow if age>= 50 years
Delay_Until_Symptoms_Start
Delay
5-40 years
Colon Cancer Screening Reason
SetAttribute
Set colon_cancer_screening = null
Colorectal_Cancer_Symptom_1
ConditionOnset
SNOMED-CT[236077008]: Protracted diarrhea
Diagnose at Symptom_Encounter
Routine_Colonoscopy_Encounter
Encounter
Reason: colon_cancer_screening
Colorectal_Cancer_Symptom_2
ConditionOnset
SNOMED-CT[6072007]: Bleeding from anus
Diagnose at Symptom_Encounter
Routine_Colonoscopy_Procedure
Procedure
SNOMED-CT[73761001]: Colonoscopy
Duration: 25-45 minutes

Generative Adversarial Networks




Real faces
Discriminator
Fake
Real
Generator
Random noise
Generated faces

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

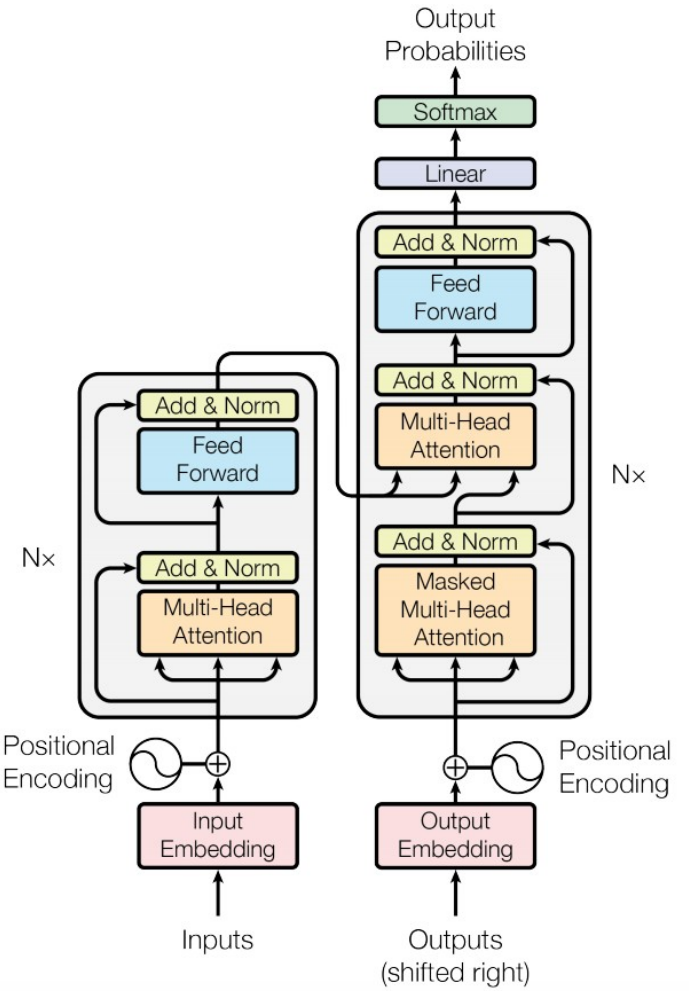
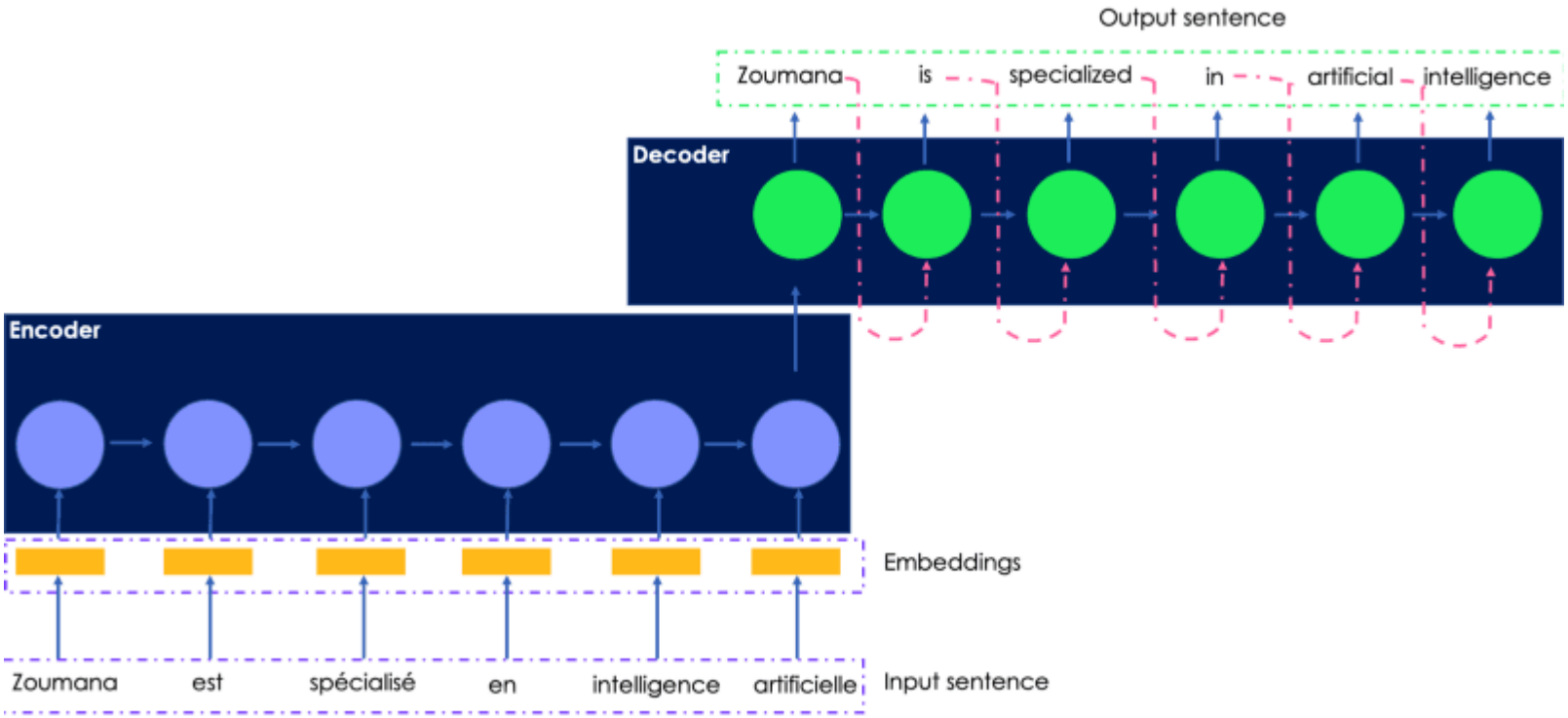
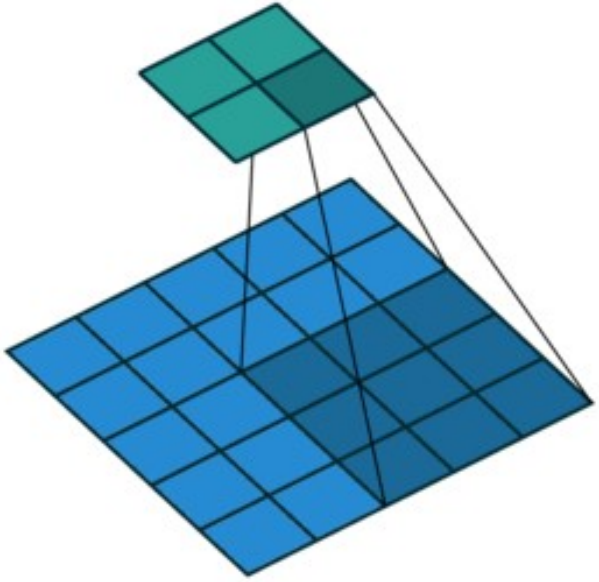
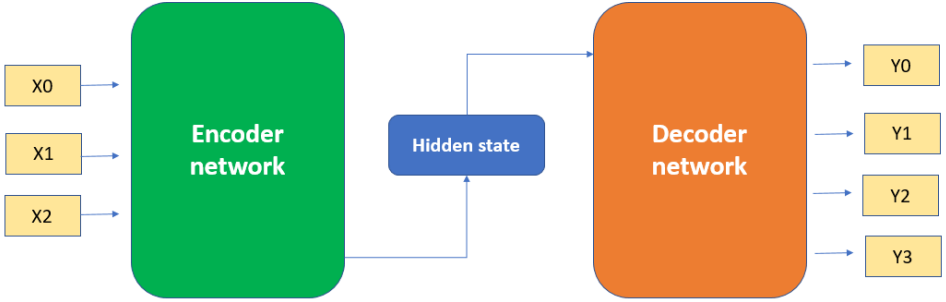


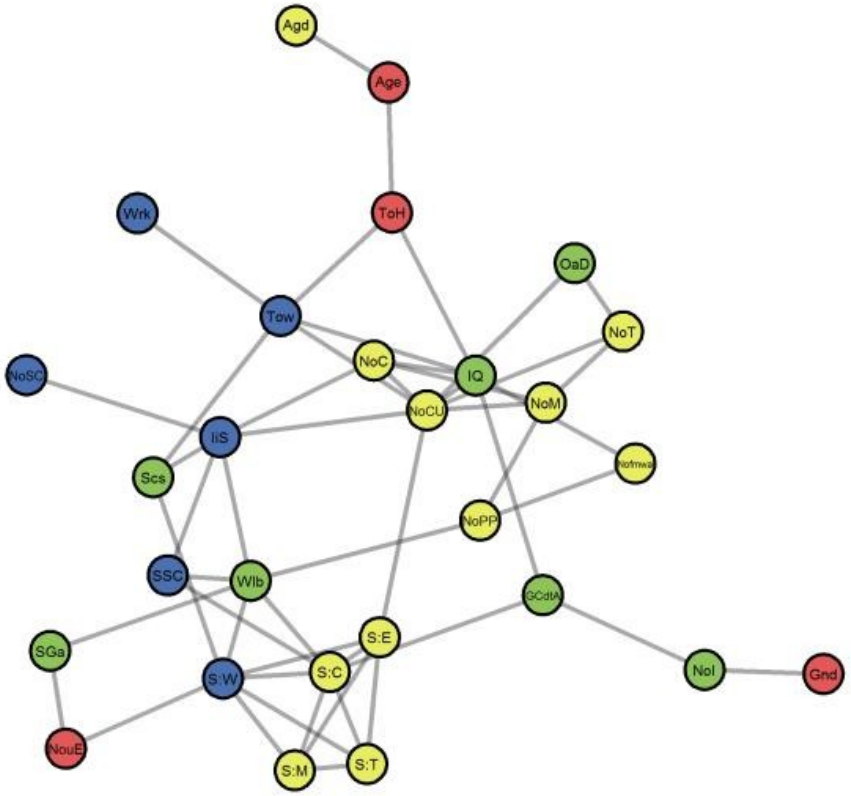
Figure 1: The Transformer - model architecture.


Output sentence
Zoumana
is
specialized
in
artificial
intelligence
Decoder
Encoder
Embeddings
Zoumana
est
spécialisé
en
intelligence
artificielle
Input sentence




X0
X1
X2
Encoder network
Hidden state
Decoder network
Y0
Y1
Y2
Y3

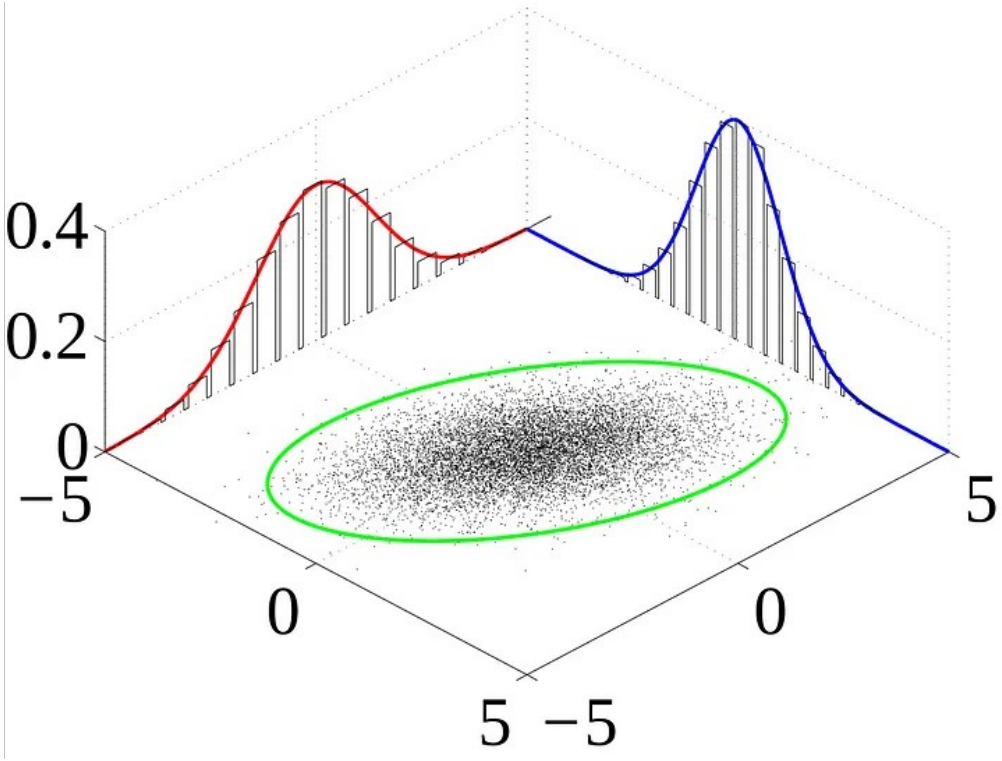
Complex Relational Structure e.g. sparse dependencies (variables only dependent on small subset of other variables) and heterogeneous functional relationships (dependencies) between variables



Modality	Component	Invariances
Images	Convolutional	Spatial translation
Text	Recurrent	Temporal translation
Tabular	Fully connected	-

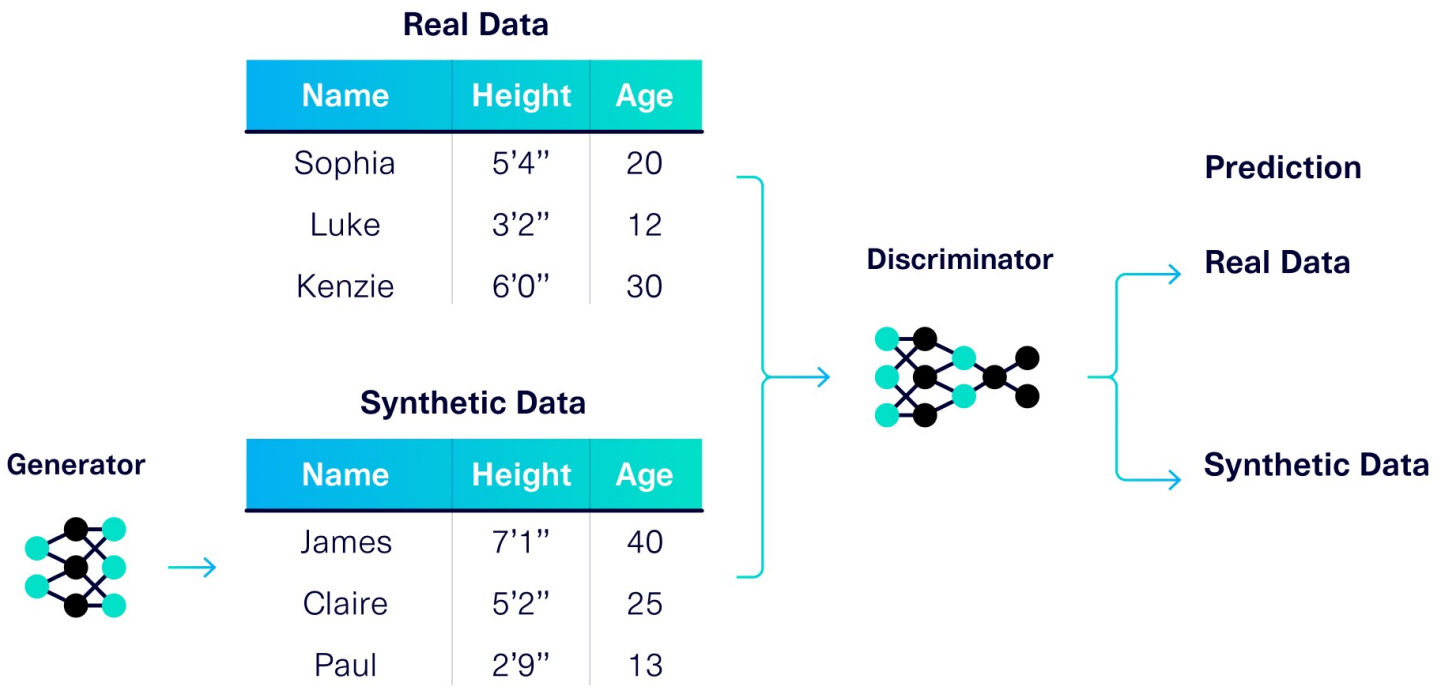
Gaussian copulas

Sklar's theorem: Any multivariate joint distribution can be written in terms of univariate marginal distributions functions and a copula describing the variables' dependence structure



CTGAN

CTGAN is a type of GAN that's used for generating synthetic tabular data.



Evaluation

FIDELITY

How similar is this synthetic data as compared to the original training sets

Kullback-Leibler (KL) divergence, pairwise correlation difference

UTILITY

How useful is this synthetic data for our downstream machine learning applications

Accuracy, F1-score, ROC, and AUC-ROC

PRIVACY

Has any sensitive data been inadvertently synthesized by our model

Membership inference, re-identification and attribute inference attacks