

18:30 Intro

18:50 Synthetic Data - What is it, how to get it, and how can it help you

Paul Tiwald, mostly.ai

19:40 Announcements & Job Openings

Break & Buffet

20:10 AI in BioMedicine

Introduction and overview of the AI Institute at MedUni Vienna

Christoph Bock (Institute Director)

Scientific reasoning in large language models Matthias Samwald (Group Leader)

Applying diffusion models to protein design Moritz Schaefer (PostDoc)

How does deep learning help analyzing molecular data for precision medicine? Daria Romanovskaia (PreDoc)

Data Science @ Magenta

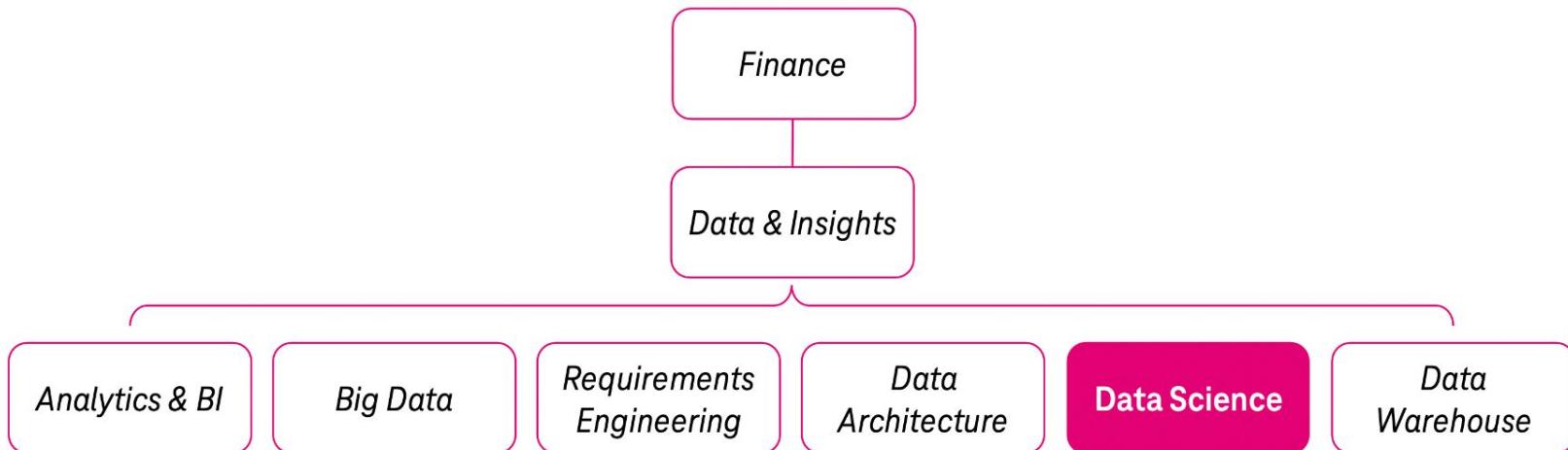
47th Deep Learning Meetup in Vienna | Jan 26th, 2023

T Magenta

Data Science @ Magenta

Magenta

Telecommunications provider offering mobile and fixed-line solutions for telephony and internet



Data Science Team

Responsible for delivering **advanced analytics & machine learning** solutions.

Data Science team currently consists of 6 people, Data & Insights in total is ~35 internals + ~15 externals.

Central team providing DS solutions to stakeholders across the company.

Topics include:

- Network quality & predictive maintenance
- ML-based tariff recommendation
- Customer segmentation
- Churn predictions
- Proactive notifications in case of service disruptions
- Customer fuzzy matching



Show case 1

Proactive Noise Detection

Scope

- Identify a defect's root cause in the cable network
- Direct technicians to the exact location of the defect
- Reduce downtime & improve customer satisfaction

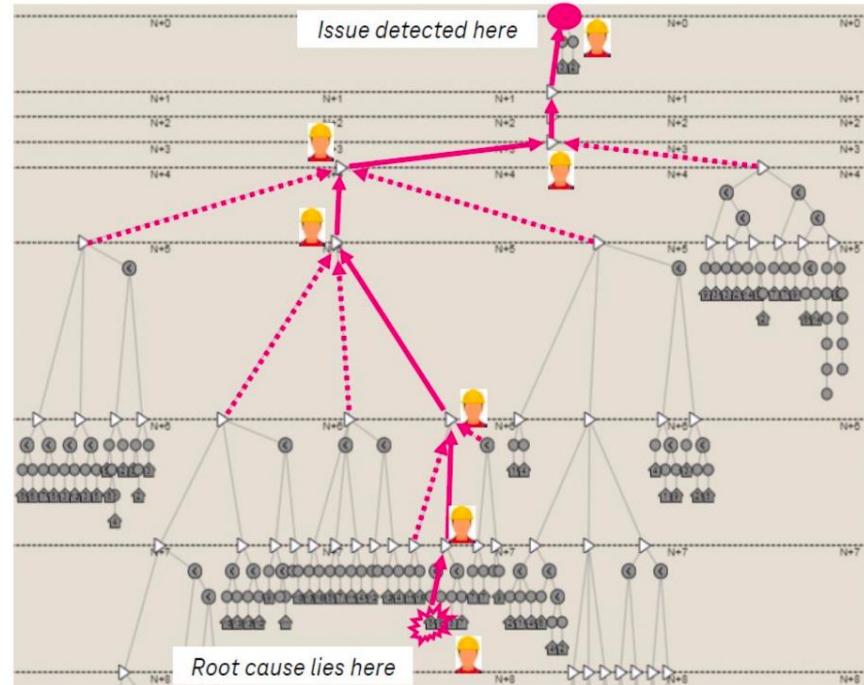
Result

Python-based automation of a simple business rule reduces time spent troubleshooting by **15-20 minutes per incident**

Results can be further optimized by using machine learning

Heiler et al. 2022, Identifying the root cause of cable network problems with machine learning.

<https://arxiv.org/abs/2203.06989>



Show case 2

Tariff recommendation

Predict interests



Tariff dimension	Interest scores
Incl. minutes	.6
Incl. SMS	.1
Incl. GB	.8
...	...

- lightGBM trained on historic tariff changes
- large volume of training data
- independent of current offer portfolio

Offer matching

Mobile Sim Only S	Mobile Sim Only M	Mobile Sim Only Unlimited Gold
+ Ohne Bindung JETZT NOCH MEHR GB! 30 GB mtl. 15 GB in AT + EU Details einblenden ↴	+ Ohne Bindung JETZT NOCH MEHR GB! 40 GB mtl. 20 GB in AT + EU Details einblenden ↴	+ Ohne Bindung 5G-fähig Unlimitiert davon 45 GB in der EU Details einblenden ↴
€ 15 MTL* Grundgebühr Entgeltbeitr. einschließen	€ 20 MTL* Grundgebühr Entgeltbeitr. einschließen	€ 35 MTL* Grundgebühr Entgeltbeitr. einschließen
Tarif wählen	Tarif wählen	Tarif wählen
*Zgl. Servicekosten ab € 33,- jahr		

- Best offer is selected based on multi-log-loss between interest scores and tariff capabilities relative to current tariff

Let's connect 😊

LinkedIn

<https://www.linkedin.com/in/heinz-eckert/>

GitHub

<https://github.com/heckert>

18:30 Intro

18:50 Synthetic Data - What is it, how to get it, and how can it help you

Paul Tiwald, mostly.ai

19:40 Announcements & Job Openings

Break & Buffet

20:10 AI in BioMedicine

Introduction and overview of the AI Institute at MedUni Vienna

Christoph Bock (Institute Director)

Scientific reasoning in large language models Matthias Samwald (Group Leader)

Applying diffusion models to protein design Moritz Schaefer (PostDoc)

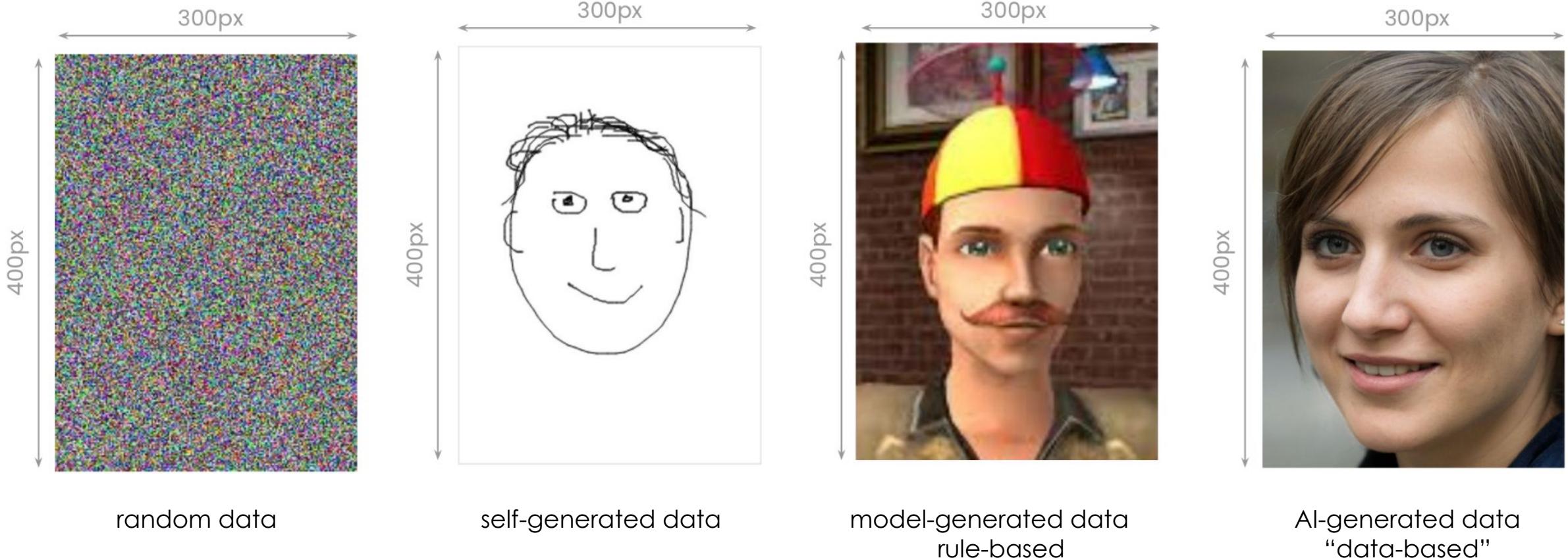
How does deep learning help analyzing molecular data for precision medicine? Daria Romanovskaia (PreDoc)



Synthetic Data

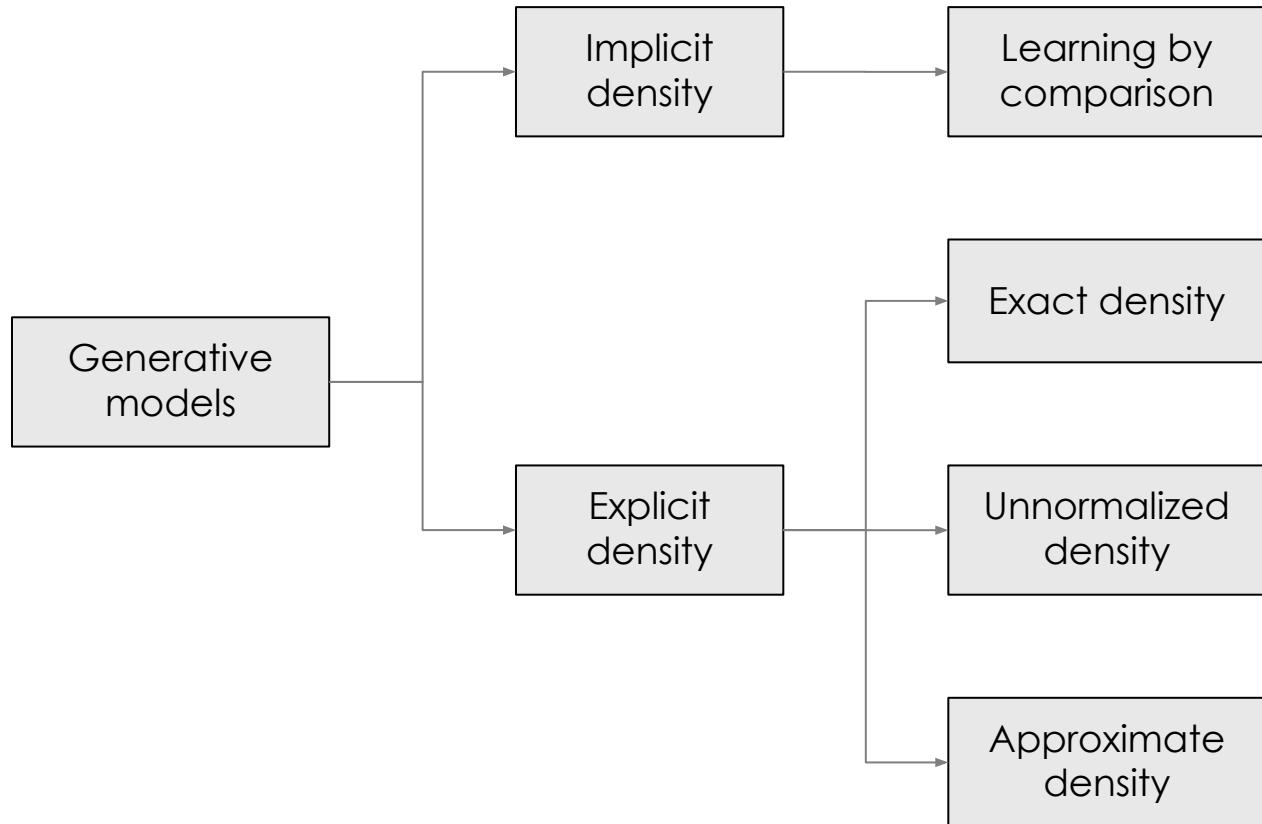
What is it, how to get it, and how can it help you

What is Synthetic Data



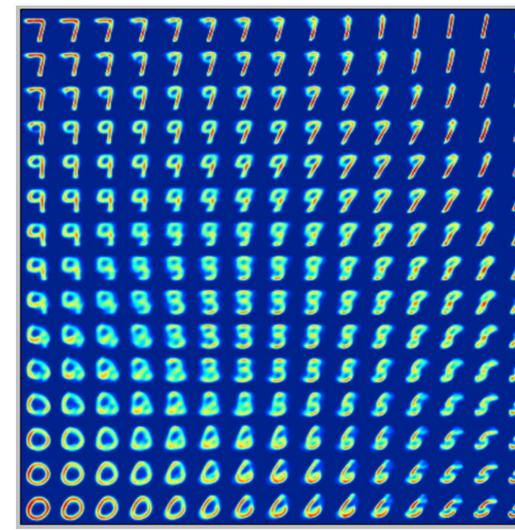
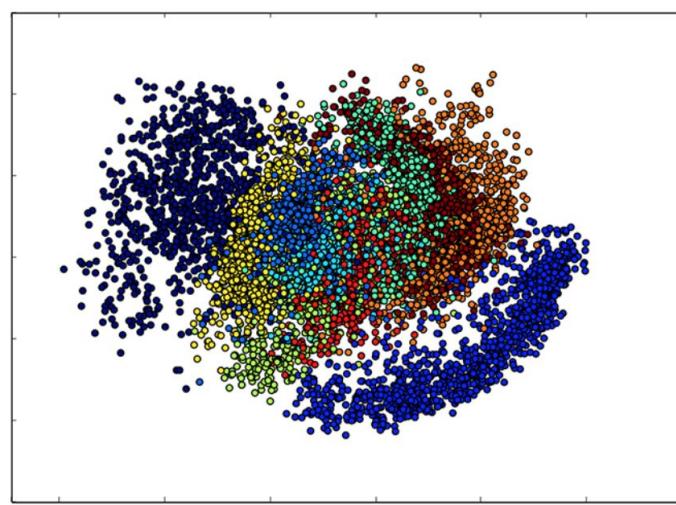
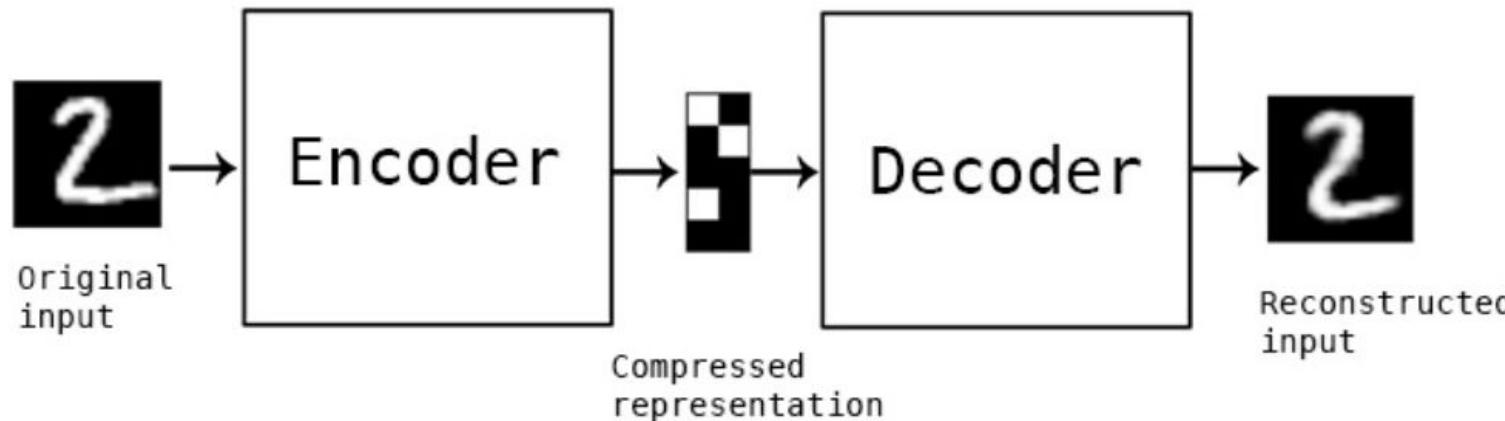
How to get it / create it?

Taxonomy of deep generative models



- only care about generation, not $p(x)$
 - instead of maximizing the density, compare real vs generated sample (classification problem)
 - examples: GAN, GMMN
-
- directly learn density $p(x)$
 - examples: autoregressive models (Transformer, RNNs), flow-based models
-
- learn unnormalized density $E(x) \propto p(x)$
 - examples: EBM
-
- learn approximation (e.g. lower bound) of density $L(x) \leq p(x)$
 - examples: VAE, diffusion models

Variational Autoencoder

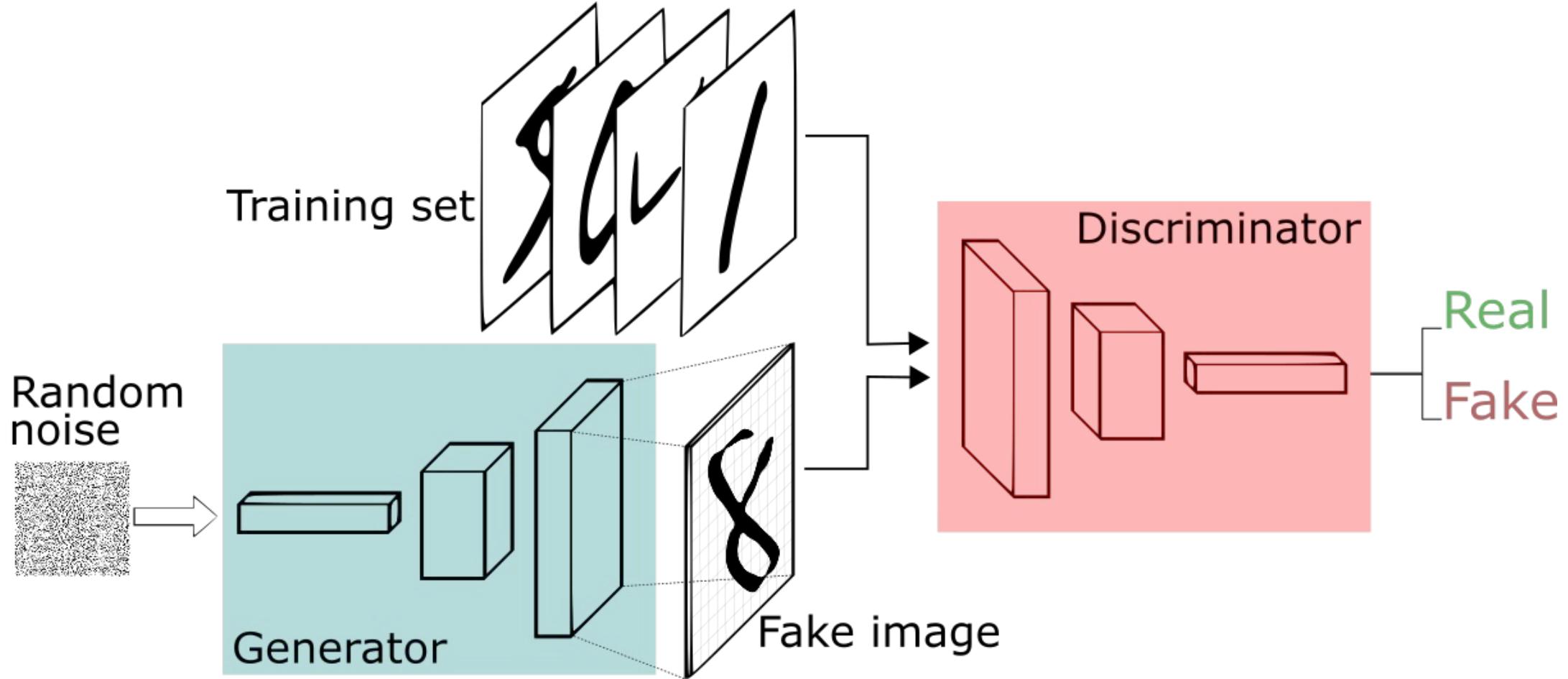


Applications:

- Dimensionality reduction
- Feature extraction
- Image denoising
- Image compression/search
- Anomaly detection
- Missing value imputation

[Keras tutorial on Auto-Encoders](#)

Generative Adversarial Networks



Generative Adversarial Networks



2014



2015



2016



2017

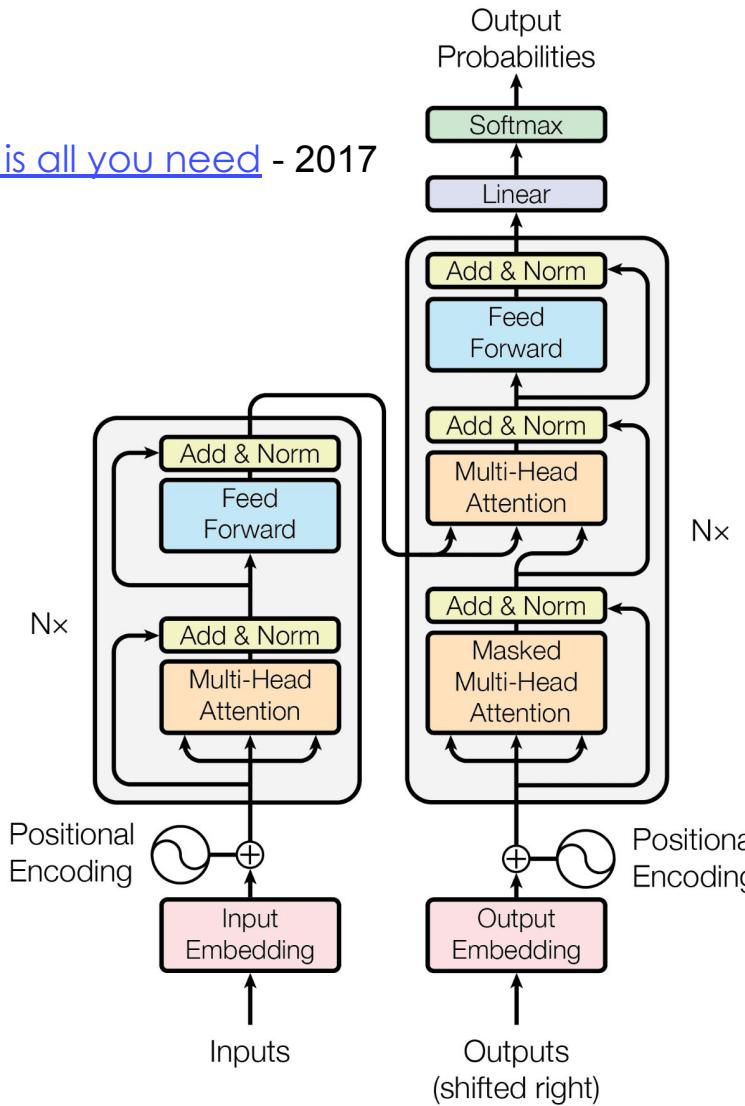
[The Malicious Use of Artificial Intelligence:
Forecasting, Prevention, and Mitigation](#) - 2018



[NVIDIA's StyleGAN2](#) - 2020

Transformers

[Attention is all you need](#) - 2017



- two important features:
auto-regressive nature
attention mechanism
- basis of Large Language Models:
Chat GPT (GPT3), Bert
- not limited to language models:
Whisper - automatic speech recognition system
Vision transformer

Transformers - ChatGPT

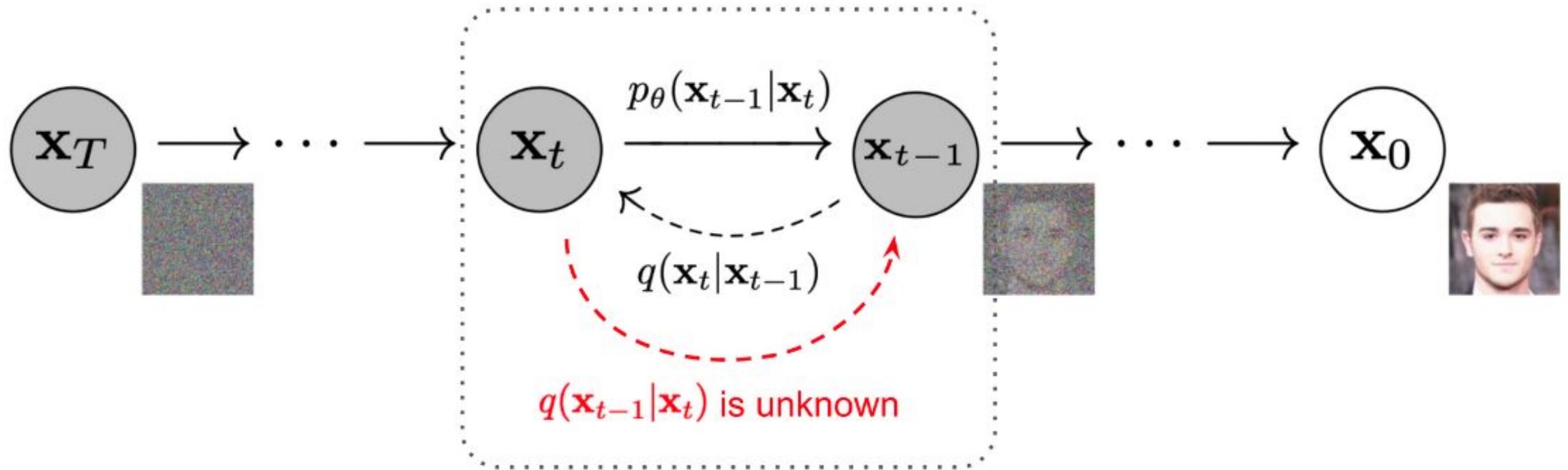
me:

I'm giving a talk titled "Synthetic Data - What is it, how to get it, and how can it help you" - can you write a short abstract for this presentation?

Chat GPT:

In this presentation, we will explore the concept of synthetic data and its potential uses in various industries. We will discuss what synthetic data is, how it is generated, and the different methods for creating it. Additionally, we will delve into the benefits and limitations of using synthetic data and provide examples of how it can be applied in real-world scenarios. The presentation will also cover best practices for acquiring and utilizing synthetic data to improve the performance of machine learning models and other applications.

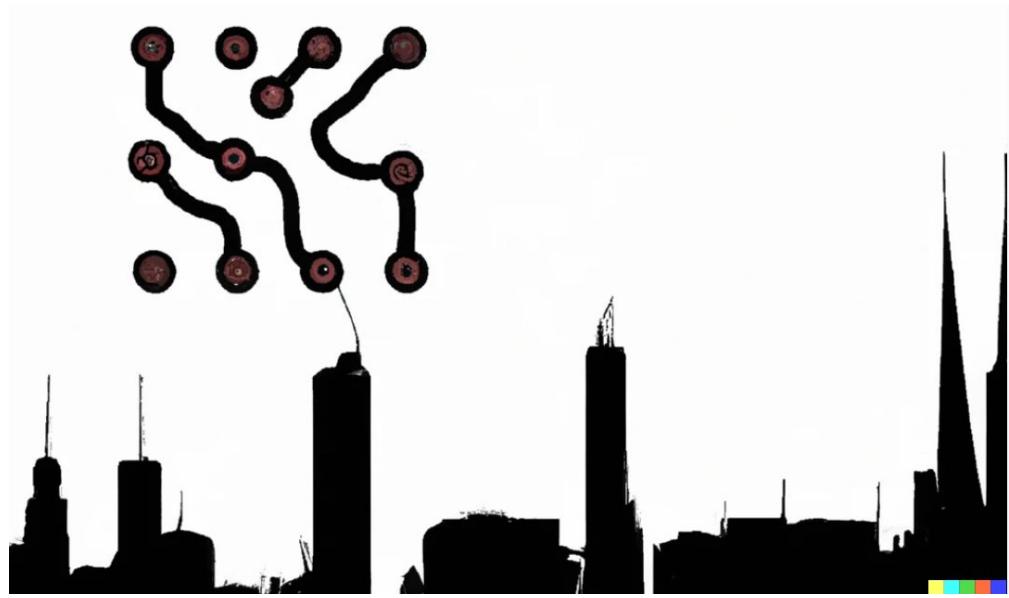
Diffusion models



[Ho et al. 2020](#), [Lilian Weng 2021](#)

Diffusion models - Dall-E 2

Create a Logo for the Vienna Deep Learning meetup.
It should integrate the city's skyline and components of neural networks.



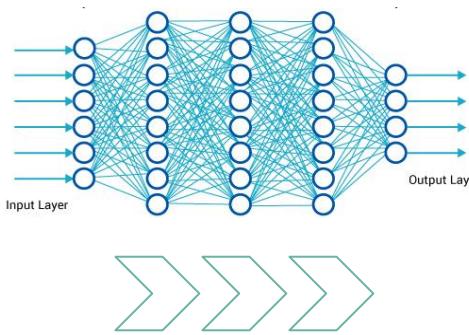
Diffusion models - Dall-E 2



Tabular Synthetic Data

NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Mary	25y	female	Book	12€	4/2/19	8:12
John	72y	male	Pizza	34€	4/2/19	18:12
...						
Bill	18y	male	Swim	6€	4/4/19	10:02
Bill	18y	male	Shoes	123€	4/4/19	12:32

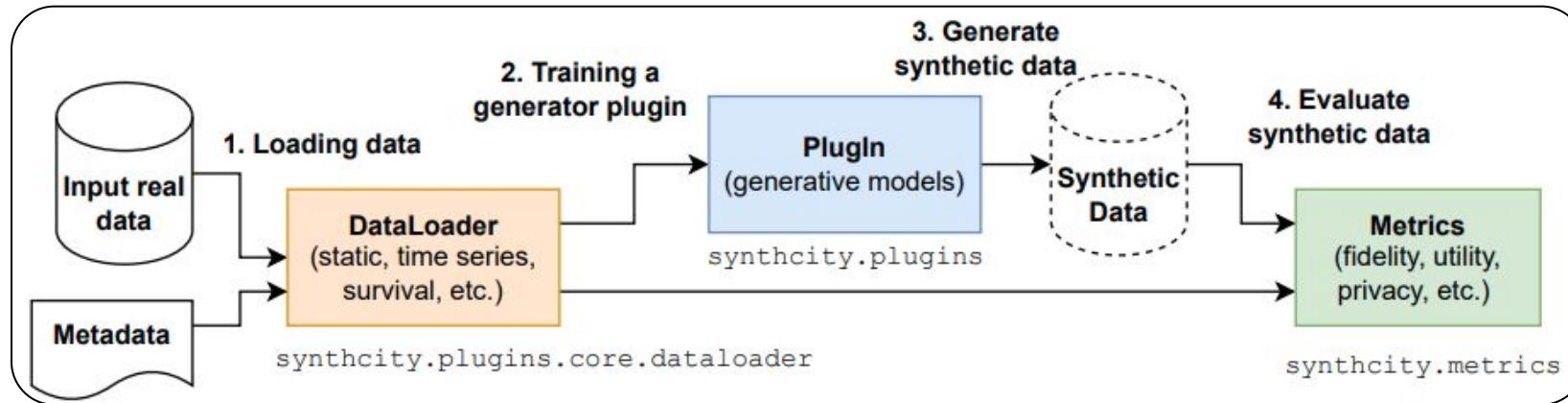
Real Data



NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Kim	29y	female	Amazon	236€	4/4/19	12:32
Kim	29y	female	Zalando	36€	4/4/19	18:58
...						
Brian	82y	male	Beer	6€	4/2/19	21:32
Sue	24y	female	Sushi	12€	4/2/19	21:32

Synthetic Data

Where to get tabular synthetic data



SynthCity:
<https://github.com/vanderschaarla/b/synthcity>

Smarter Synthetic Data

MOSTLY.AI

Sign in to start generating synthetic data

* E-mail address

* Password

Forgot password?

I'm not a robot

reCAPTCHA
Privacy - Terms

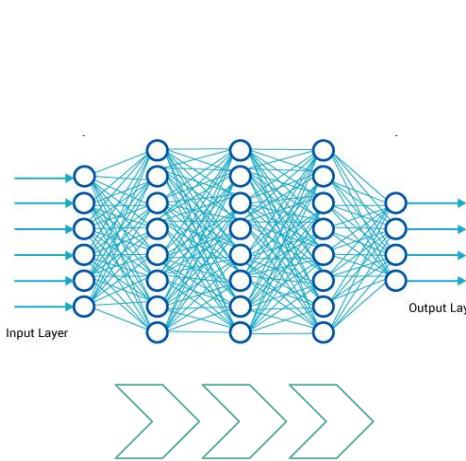
MOSTLY AI's SD Platform:
<https://synthetic.mostly.ai>

How can it help you?

Tabular Synthetic Data

NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Mary	25y	female	Book	12€	4/2/19	8:12
John	72y	male	Pizza	34€	4/2/19	18:12
...						
Bill	18y	male	Swim	6€	4/4/19	10:02
Bill	18y	male	Shoes	123€	4/4/19	12:32

Real Data



Private

NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Kim	29y	female	Amazon	236€	4/4/19	12:32
Kim	29y	female	Zalando	36€	4/4/19	18:58
...						
Brian	82y	male	Beer	6€	4/2/19	21:32
Sue	24y	female	Sushi	12€	4/2/19	21:32

Synthetic Data

Open

Legacy Anonymization is not safe

WIRED

NetFlix Cancels Recommendation
Contest After Privacy Lawsuit

The
New York
Times

Researchers spotlight the lie of
'anonymous' data

AOL: "This was a screw up"

The
Guardian

Sticky data: Why even
'anonymized' information can
still identify you

Big data privacy is a bigger issue than you think

Harvard
Business
Review

There's No Such Thing as Anonymous Data

Sorry, your data can still be
identified even if it's anonymized

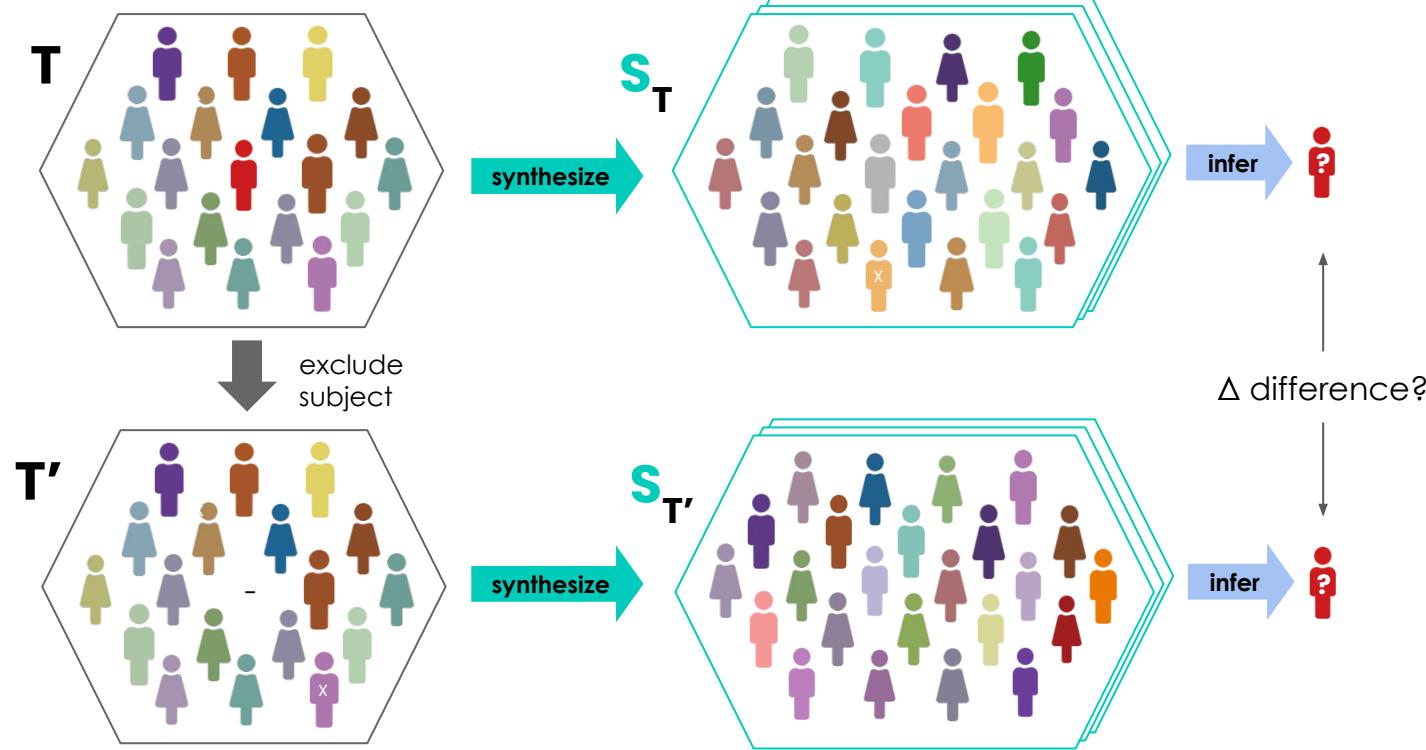
'Anonymised' data can never be totally
anonymous, says study

SCIENCE & TECHNOLOGY

You're not so anonymous

Saying it's Anonymous Doesn't Make It So: Re-
identifications of "anonymized" law school data

How to test Synthetic-Data Privacy



We must not be able to infer more about an individual, when that person is included in the database used for synthesis.

How to test Synthetic-Data Privacy

Accuracy scores for 50 randomly chosen subjects, that were part of training

	NB	SVM	KNN	RF	LR	FRNN	ENS	DUM	RMEAN
Target T	42.8±5	43±6.6	41.6±9.3	49.8±9.4	38.3±3.7	49.1±9.8	45±6.8	32	44.2±3.8
Synthetic S_T	42.1±4.2	39.5±7.3	36.2±6.8	37.9±5.9	36.5±6	37±6.3	39.7±6.1	32	38.4±2

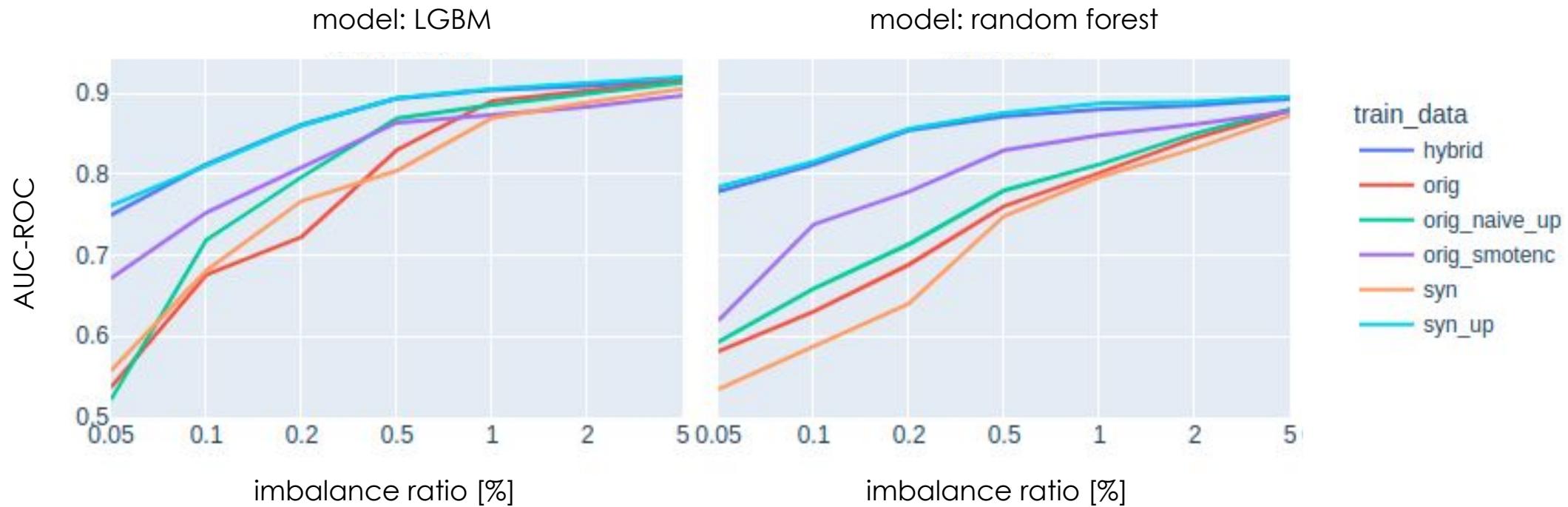
Accuracy scores for 50 randomly chosen subjects, that were NOT part of training

	NB	SVM	KNN	RF	LR	FRNN	ENS	DUM	RMEAN
Target T'	42.1±5	39.9±7.1	36.9±5.7	39.7±5.1	37.6±4.1	39.1±5	39.7±5.9	32	39.3±1.6
Synthetic $S_{T'}$	43.7±4.2	40.4±6.4	35.9±6.4	39.1±6.2	36.9±4.6	38±5.8	40.5±7	32	39.2±2.4

SBA Research, 2020

Machine-Learning on Synthetic Data

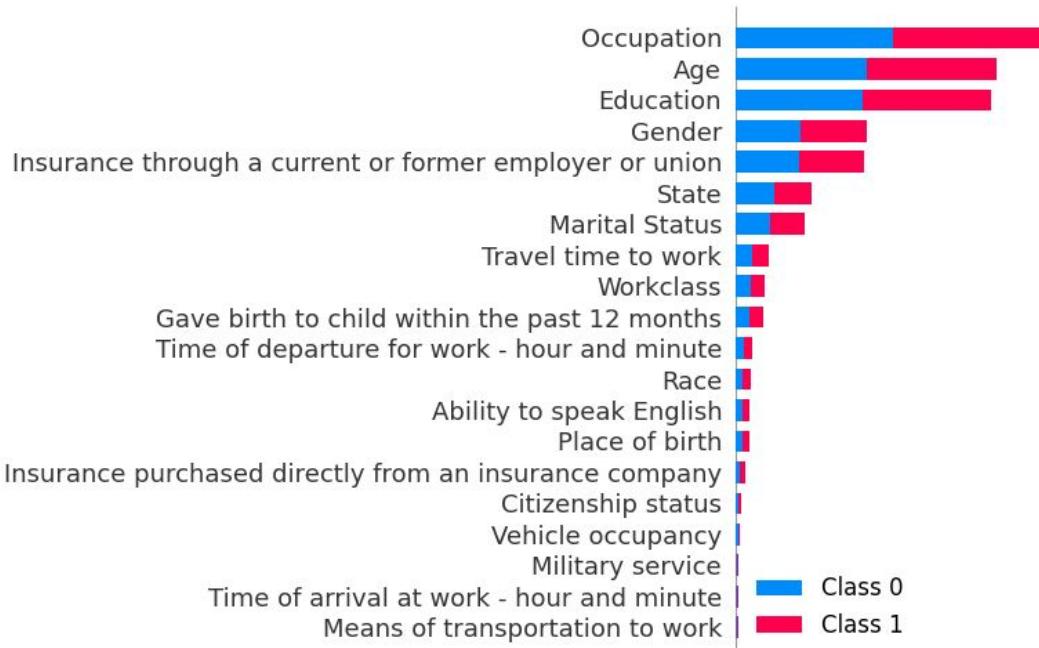
Upsampling of minority classes



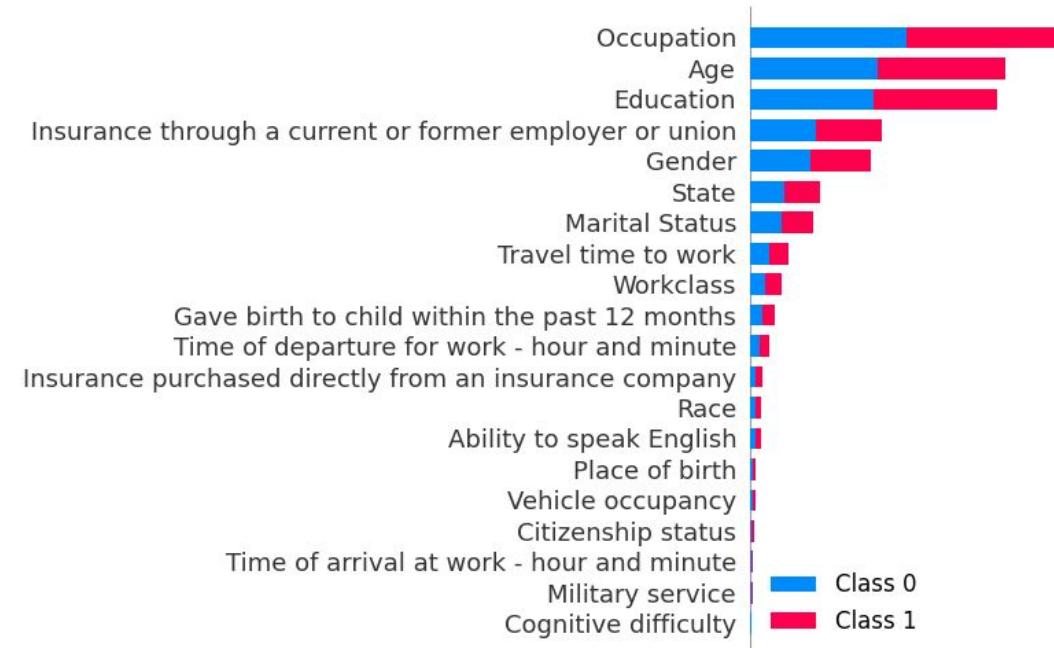
Explainability of Synthetic Data

SHAP values of real and synthetic data

Train-Real-Explain-Real



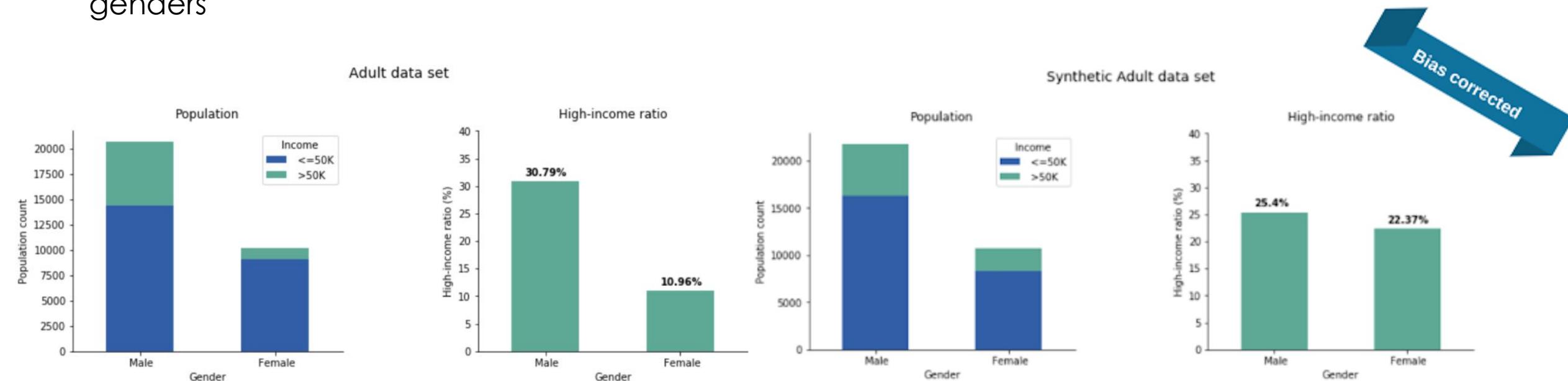
Train-Synthetic-Explain-Synthetic



Fair Synthetic Data

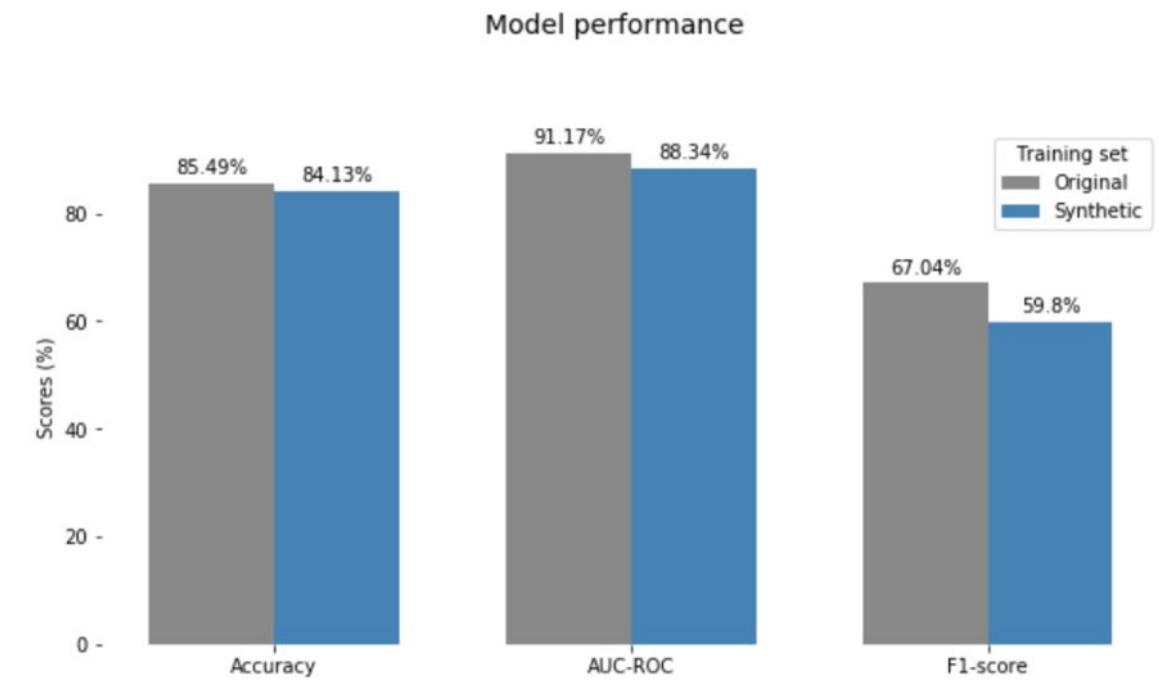
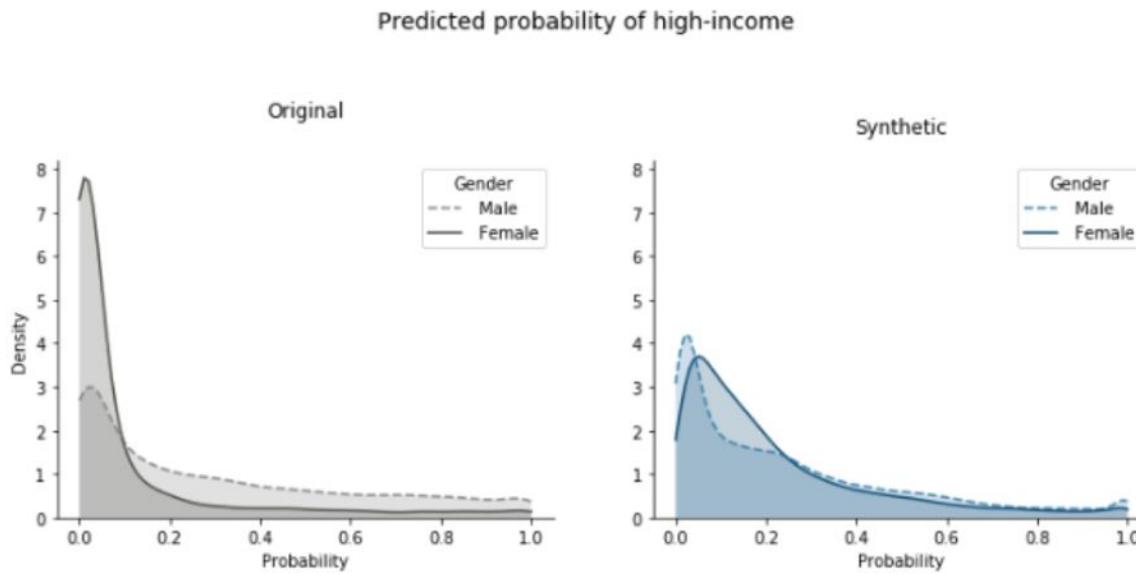
Example

- Income inequality in the Adult census data set: violation of **demographic parity**, as there are over 30% of all men but only ~11% of women earning more than \$50,000 per year.
- The bias-mitigated synthetic Adult data set: the ratio of high-income individuals are equal across genders



[Representative & Fair Synthetic Data](#)

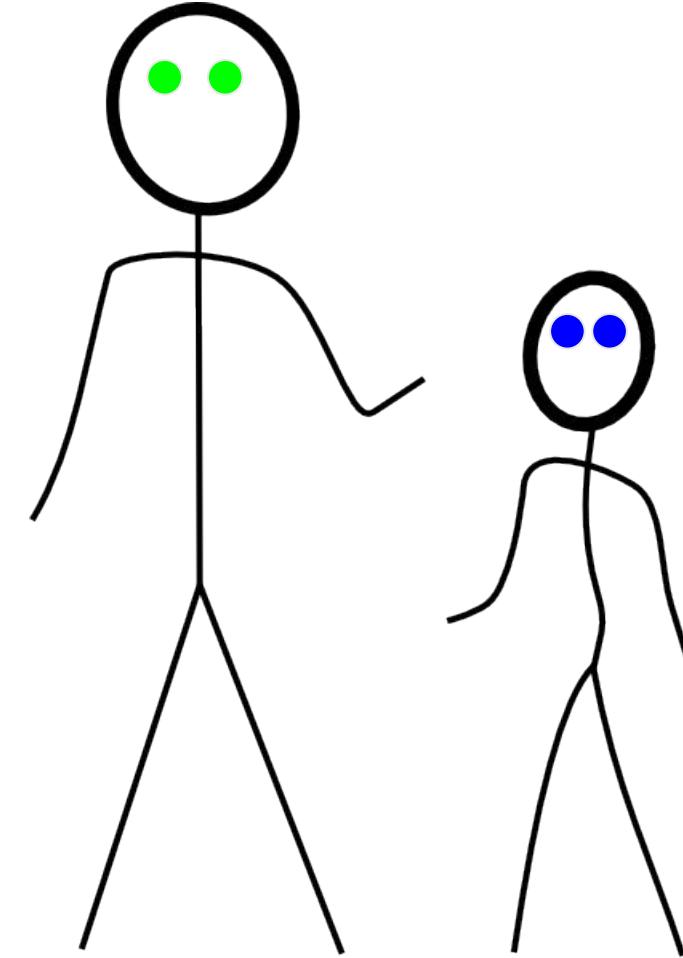
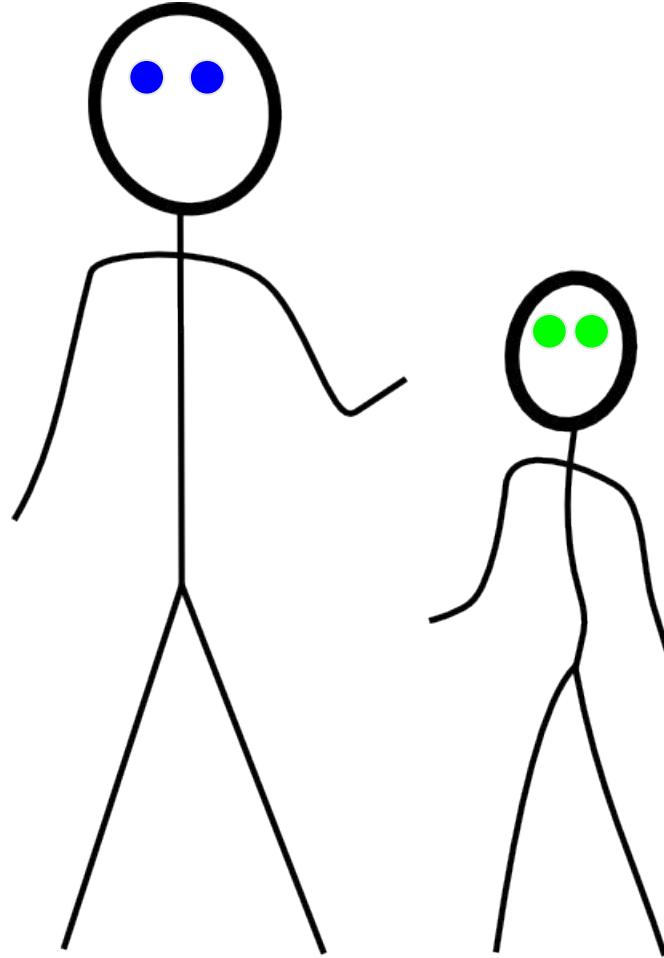
Fair Synthetic Data



predictions of ML model trained on synthetic data are fairer

training an ML model on fair synthetic data hardly reduces performance

Simulation with Synthetic Data



Simulation with Synthetic Data



International Journal of Research in
Marketing

Volume 39, Issue 4, December 2022, Pages 988-1018

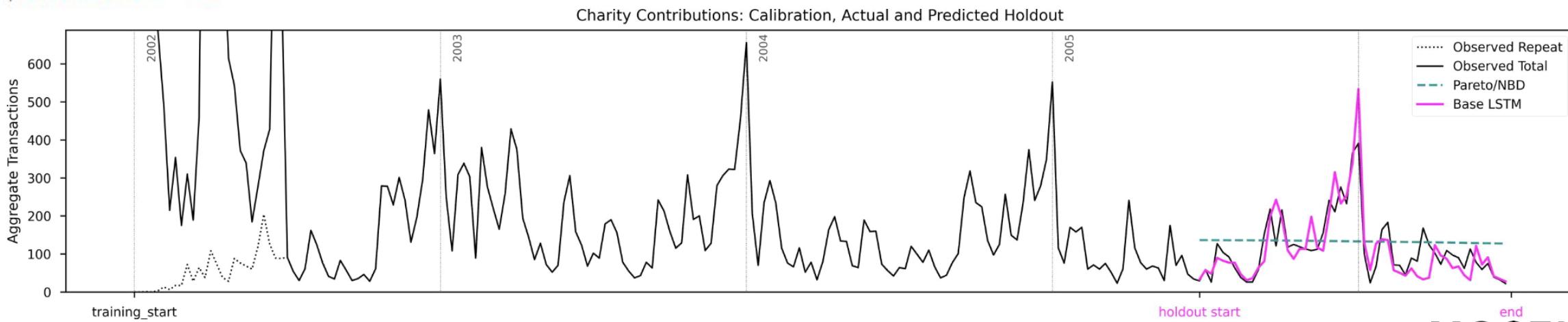


Full length article

Customer base analysis with recurrent neural networks

Jan Valentin^a , Thomas Reutterer^a , Michael Platzer^b
, Klaudius Kalcher^b

[Customer base analysis with recurrent neural networks](#)



What can Synthetic Data do for you?

- Privacy-preserving Analytics
- Privacy-preserving Machine Learning
- Data Collaboration
- Data Augmentation
- Data Diversity

References

- check out our blog
<https://blog.mostly.ai/>

[Synthetic Behavioral Data](#)

[Synthetic Geo Data](#)

[Synthetic Text Data](#)

[Fair Synthetic Data](#)

[Synthetic Data Benchmarks](#)

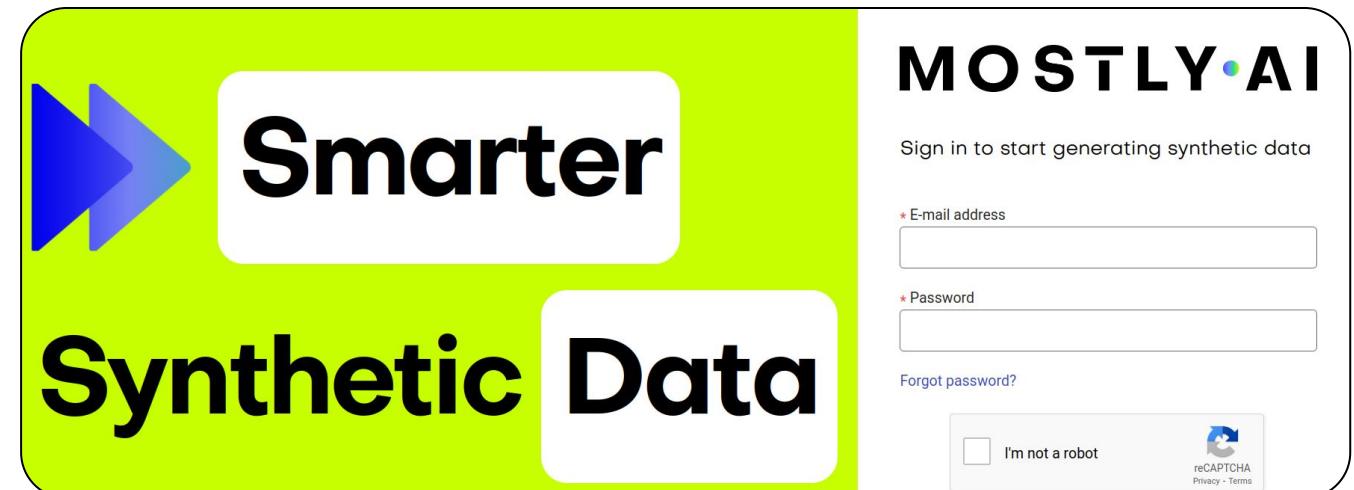
[JRC Report on Synthetic Data](#)

[AI-based Re-Identification Attacks](#)

[Privacy Assessment of Synthetic Data](#)

and many more

- try our free version
<https://synthetic.mostly.ai>





Self-supervised Learning: The dark matter of intelligence

Yann LeCun

Events



Social Artificial
Intelligence Night

Ifh III
st. pölten

saint.fhstp.ac.at

Julian Schrittwieser ist Keynotespeaker 2023

Nicht verpassen! Julian Schrittwieser – Experte von Google DeepMind und Mastermind von AlphaGo und AlphaZero, hält die Keynote auf der Social Artificial Intelligence Night am 24. März 2023.

[MEHR INFOS HIER!](#)



Wann: 24. März 2023 | 16:30 – 23:00 Uhr

Wo: FH St. Pölten



saint.fhstp.ac.at

Fach-Vorträge

Call for Contribution

Poster



Wann: 24. März 2023 | 16:30 – 23:00 Uhr

Wo: FH St. Pölten

Job Openings

Send us announcements & job openings!
contact@vdlm.at

Why Magenta Telekom?

Those who strive to simplify Magenta customers lives, deserve exceptional benefits and a great work-life-balance.

Welcome
& Work

- Onboarding**
- Flexible working hours**
- Mobile working**
- 40 working hours per week, no All-In contract!**

Social
Support
&
Insurance

- Pension plan**
- Employee assistance services**

Developmen
t & Health

- T-Academy:** Important lessons are one click away – on our online learning platform.
- Fitness offers**

Daily Life

- Meals:** Enjoy the offerings of our Canteen
- Food Allowance**

Mobility: We fund 2/3 of your annual transit ticket in large cities

Products
& Family

Smart phones, tariffs & entertainment: There are discounts available to you.

Family & friends: Attractive internet and mobile packages for your family and friends.

Childcare

Magenta®

<https://www.magenta.at/karriere/jobs>

<https://www.magenta.at/karriere/info-corner>





Artificial Intelligence & Data Science Consultant
(Junior) Consultant Digital Transformation

(Senior) Pentester
Consultant IT-Security
(Junior) Consultant Application Security
(Junior) Consultant Cyber Wargaming and Simulation
(Senior) Consultant Data Protection

Deloitte.

AI:

<https://1brd.ly/4l4ql?st=iaquo92mhtb1> // Artificial Intelligence & Data Science Consultant

<https://1brd.ly/VDa3q?st=od5b1adfs9jo> // (Junior) Consultant Digital Transformation

Pentesting (auch andere Städte in Österreich möglich):

<https://1brd.ly/jD3B5?st=nsftoeuhbies> // (Senior) Pentester / RedTeaming Consultant

<https://1brd.ly/bDPdb?st=18flbrhto4n7> // (Senior) Pentester / RedTeaming Consultant Wien

<https://1brd.ly/3A4eb?st=c9d8o3v21q47> // (Senior) Pentester / RedTeaming Consultant Oberösterreich

<https://1brd.ly/MDadM?st=pojpn809h852> // (Senior) Pentester / RedTeaming Consultant Tirol

<https://1brd.ly/pY56Y?st=agofgrjfcnk> // (Senior) Pentester / RedTeaming Salzburg

<https://1brd.ly/LDWn7?st=h2tcjp0jvsr5> // (Senior) Pentester / RedTeaming Steiermark

<https://1brd.ly/rDbYA?st=7djon8f2iaa3> // Consultant IT-Security

<https://1brd.ly/aD48y?st=7c68939ffne4> // (Junior) Consultant Application Security

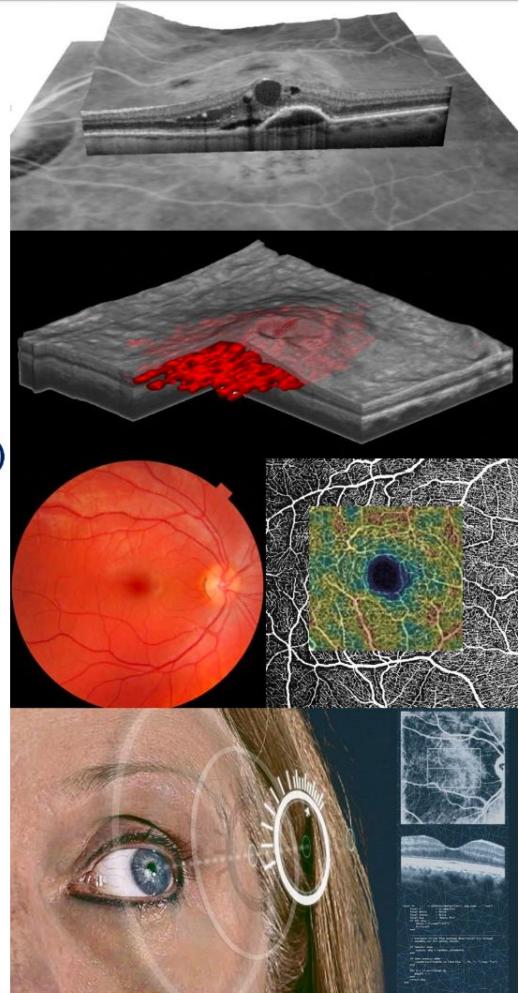
<https://1brd.ly/qD0Bd?st=pjal8995d3jb> // (Junior) Consultant Cyber Wargaming and Simulation

<https://1brd.ly/nDkj?st=oj60n7hs64d9> // (Senior) Consultant Data Protection



Multiple PhD Positions at OPTIMA Lab in Deep Learning for Retinal Imaging

- Interdisciplinary and International Research Lab at MedUni Wien
 - Computer scientists, software engineers and medical doctors
- Availability of large curated datasets and high-performance compute (HPC)
 - Well-structured and annotated imaging data lakes
 - Multimodal imaging (1Mio OCT scans) and electronic health records (EHR)
- Multiple collaborations with industry
 - Pharma and imaging device companies
- Part of FWF Austrian Science Fund project
 - In collaboration with Institute for Machine Learning at JKU Linz
 - JKU: Günter Klambauer and Sepp Hochreiter
- Contact: hrvoje.bogunovic@meduniwien.ac.at



Speech and Hearing Science Lab

Medical University of Vienna

Open Postdoc position on pathological speech

- Images / videos of vocal folds
 - Image analysis & visualization
- Audio
 - Speech analysis and synthesis
- Text
 - Clinical text mining and meta data retrieval

*data science - artifical intelligence -
deep learning - clinical science -
engineering science*

! Apply via shs-lab.at !



General Hospital of Vienna



Der Wissenschaftsfonds.



MEDICAL IMAGING CLUSTER
MEDICAL UNIVERSITY OF VIENNA



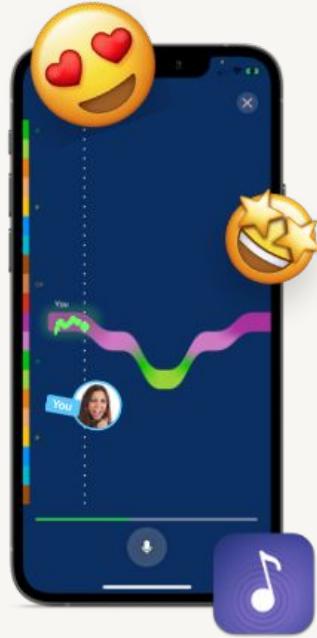
MEDIZINISCHE
UNIVERSITÄT WIEN

Philipp Aichinger
Department of Otorhinolaryngology
Division of Phoniatrics-Logopedics

CP phonicscore The MusicTech Company from Vienna



PracticeBird.com



Sawti.app



OpenSheetMusicDisplay.org

Open positions @PhonicScore for a 6 months project on

Onset and pitch detection in polyphonic and drum recordings

Tasks

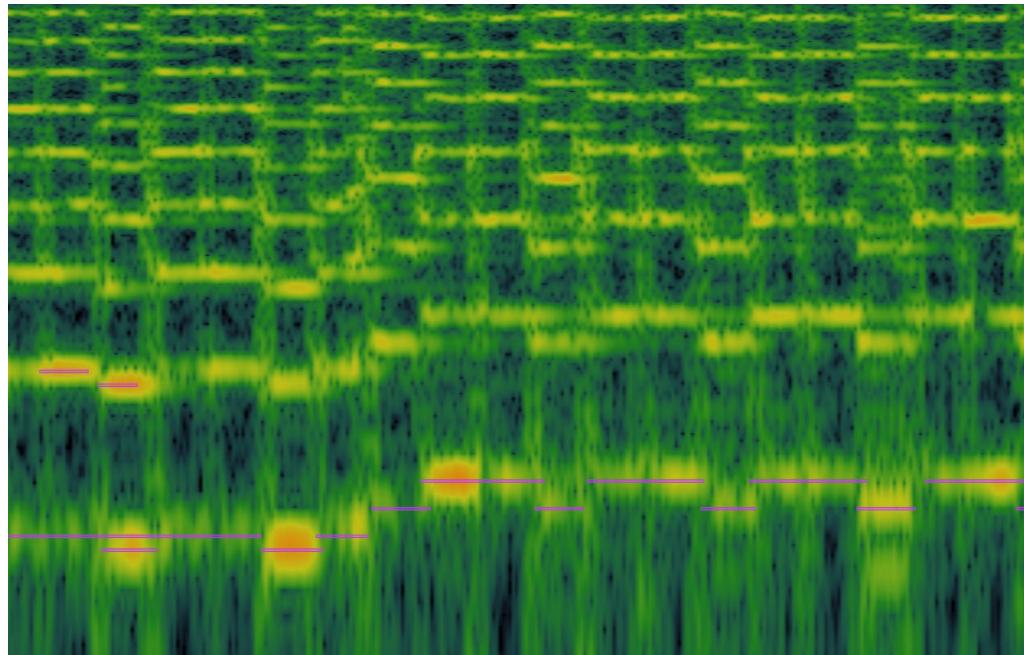
- creating and improving DL-models using TensorFlow
- training with Google Cloud Services
- optimizing models for Android/iOS using TFLite

Details

- Time frame: Feb-Aug 2023
- 20-40h/week
- Main setup ready from first project
- High chances for follow-up project(s)
- Client delivers annotated audio data
- Coding in Python and C++
- Junior and Senior positions

Contact:

PhonicScore , Matthias Uiberacker
m.uiberacker@phonicscore.com



AI in BioMedicine

Prof. Christoph Bock, Prof. Matthias Samwald,

Dr. Moritz Schaefer, Daria Romanovskaia

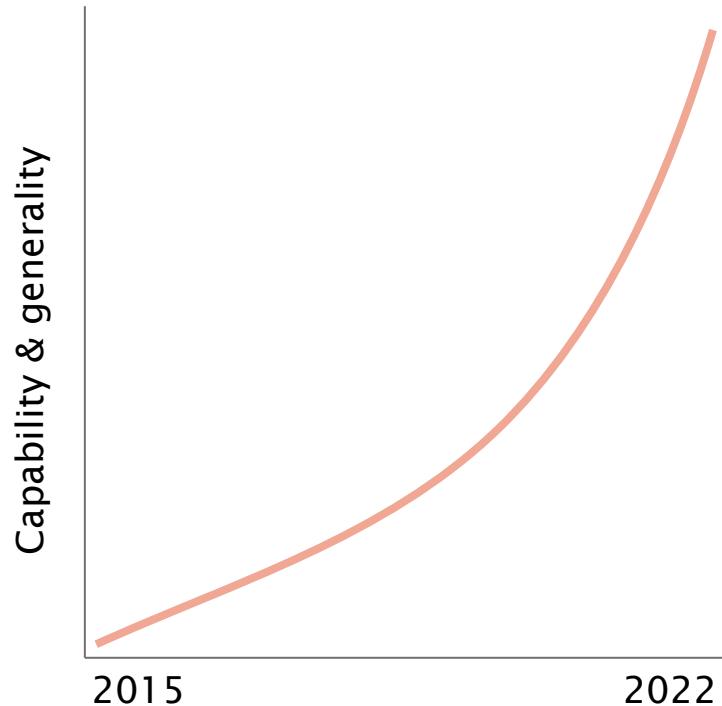
AI Institute at the Medical University of Vienna



Reasoning in large language models

Matthias Samwald

From language models to general AI



2015	"The Unreasonable Effectiveness of Recurrent Neural Networks"
2017	Transformer architecture
2018	"Pre-train, then transfer"
2020	Scaling hypothesis (GPT-2 -> GPT-3)
2022 - Q1	Instruction tuning (InstructGPT, T0)
2022 - Q2	Chain-of-thought reasoning
2022 - Q4	Self-improvement

From language models to general AI

Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact  
that it plainly and indubitably proved the fallacy of all the plans for  
cutting off the enemy's retreat and the soundness of the only possible  
line of action--the one Kutuzov and the general mass of the army  
demanded--namely, simply to follow the enemy up. The French crowd fled  
at a continually increasing speed and all its energy was directed to  
reaching its goal. It fled like a wounded animal and it was impossible  
to block its path. This was shown not so much by the arrangements it  
made for crossing as by what took place at the bridges. When the bridges  
broke down, unarmed soldiers, people from Moscow and women with children  
who were with the French transport, all--carried on by vis inertiae--  
pressed forward into boats and into the ice-covered water and did not,  
surrender.
```

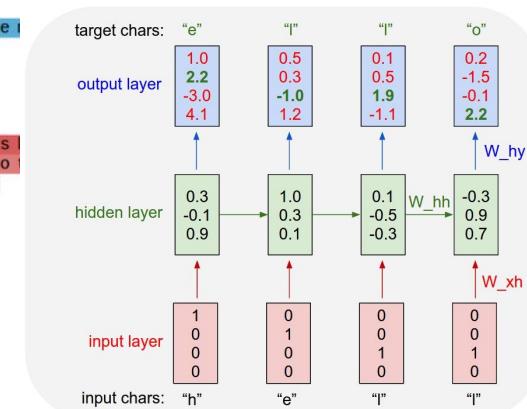
Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the  
contrary, I can supply you with everything even if you want to give  
dinner parties," warmly replied Chichagov, who tried by every word he  
spoke to prove his own rectitude and therefore imagined Kutuzov to be  
animated by the same desire.
```

```
Kutuzov, shrugging his shoulders, replied with his subtle penetrating  
smile: "I meant merely to say what I said."
```

Cell that robustly activates inside if statements:

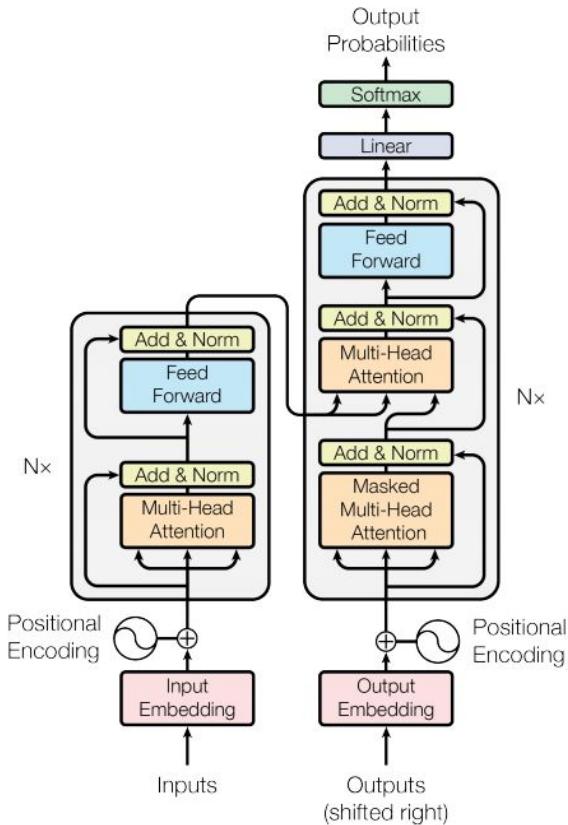
```
static int __dequeue_signal(struct sigpe  
    siginfo_t *info)  
{  
    int sig = next_signal(pending, mask);  
    if (sig) {  
        if (current->notifier) {  
            if (sigismember(current->notifier_ma  
                if (!(current->notifier)(current->no  
                    clear_thread_flag(TIF_SIGPENDING);  
                    return 0;  
                }  
            }  
            collect_signal(sig, pending, info);  
        }  
        return sig;  
    }
```



"The Unreasonable Effectiveness of Recurrent Neural Networks",
Andrej Karpathy

2015	"The Unreasonable Effectiveness of Recurrent Neural Networks"
2017	Transformer architecture
2018	"Pre-train, then transfer"
2020	Scaling hypothesis (GPT-2 -> GPT-3)
2022 - Q1	Instruction tuning (InstructGPT, T0)
2022 - Q2	Chain-of-thought reasoning
2022 - Q4	Self-improvement

From language models to general AI



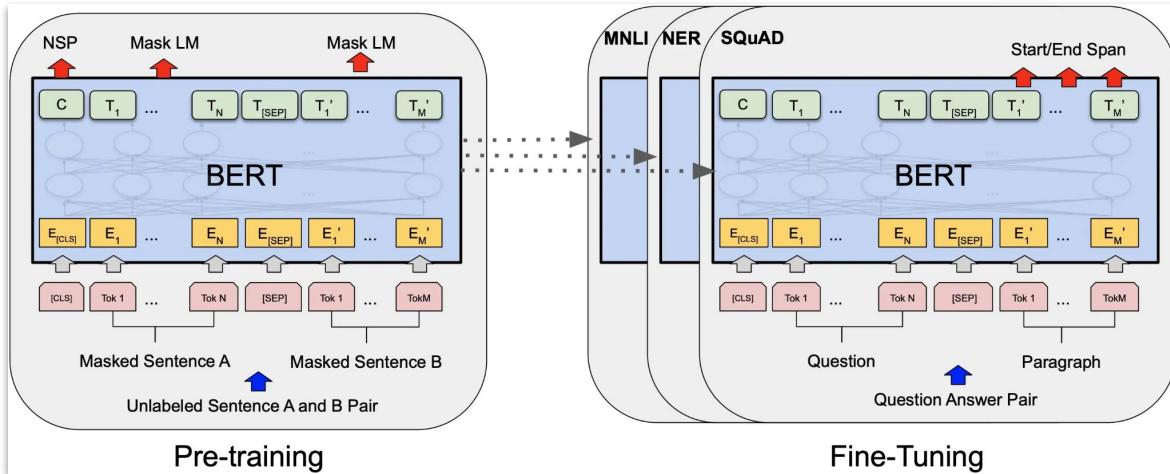
"Attention Is All You Need", Vaswani *et al.*

2015	"The Unreasonable Effectiveness of Recurrent Neural Networks"
2017	Transformer architecture
2018	"Pre-train, then transfer"
2020	Scaling hypothesis (GPT-2 -> GPT-3)
2022 - Q1	Instruction tuning (InstructGPT, T0)
2022 - Q2	Chain-of-thought reasoning
2022 - Q4	Self-improvement

From language models to general AI

NLP's ImageNet moment has arrived

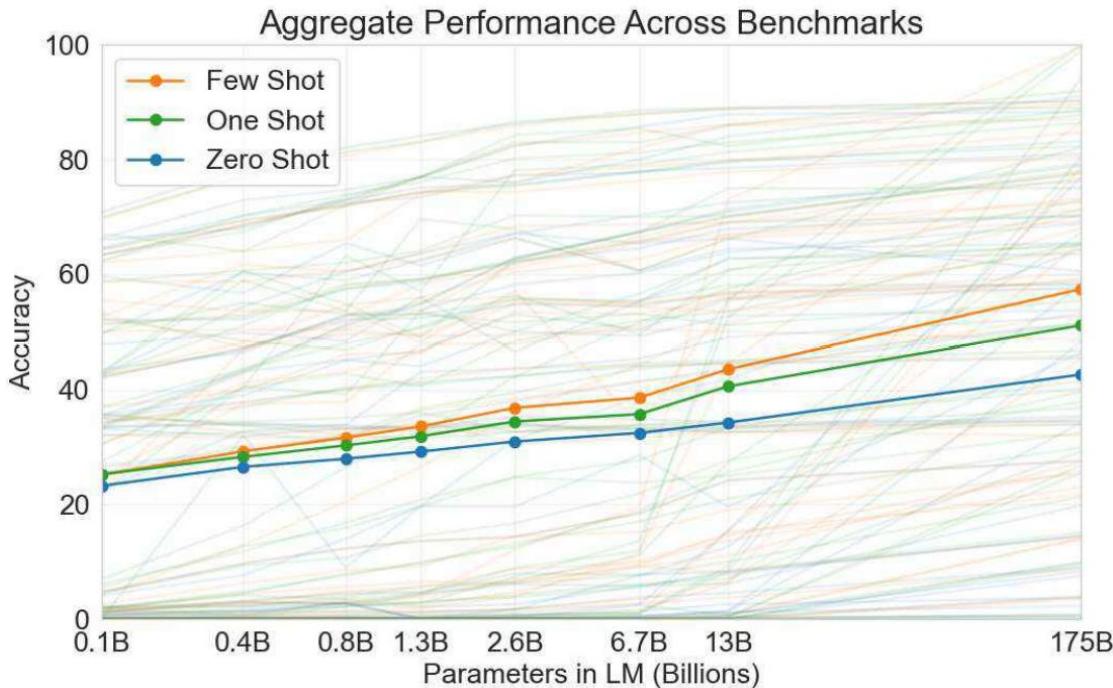
Big changes are underway in the world of NLP. The long reign of word vectors as NLP's core representation technique has seen an exciting new line of challengers emerge. These approaches demonstrated that pretrained language models can achieve state-of-the-art results and herald a watershed moment.



"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Devlin *et al.*

2015	"The Unreasonable Effectiveness of Recurrent Neural Networks"
2017	Transformer architecture
2018	"Pre-train, then transfer"
2020	Scaling hypothesis (GPT-2 -> GPT-3)
2022 - Q1	Instruction tuning (InstructGPT, T0)
2022 - Q2	Chain-of-thought reasoning
2022 - Q4	Self-improvement

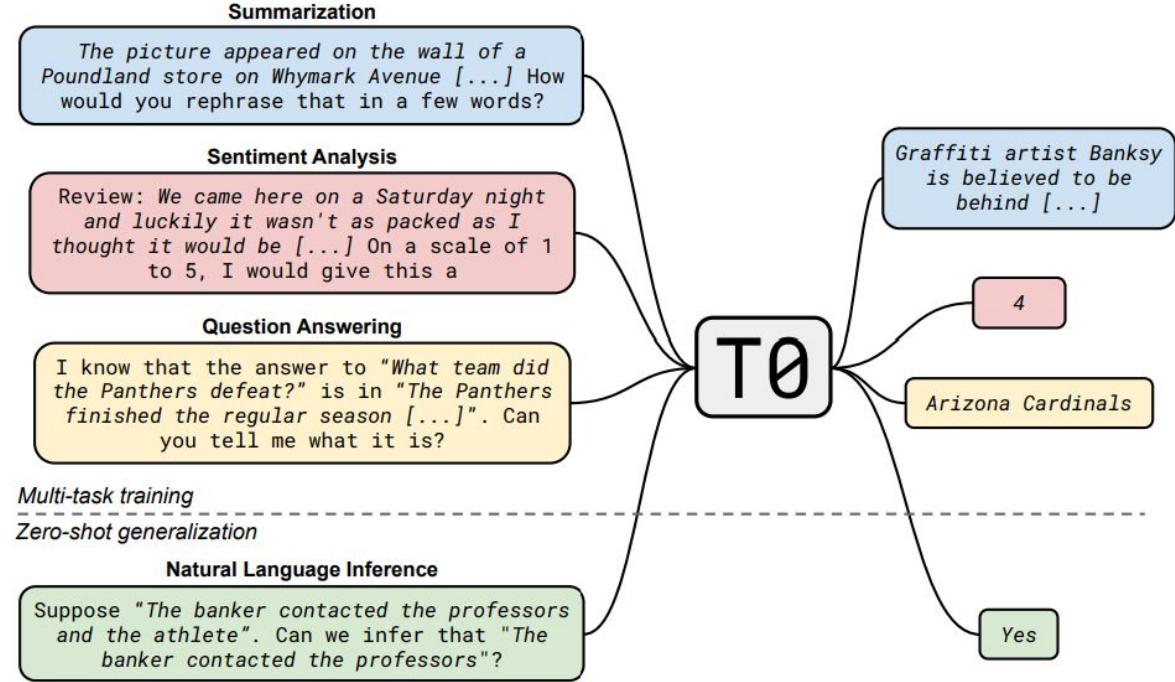
From language models to general AI



“Language Models are Few-Shot Learners”, Brown *et al.*

2015	“The Unreasonable Effectiveness of Recurrent Neural Networks”
2017	Transformer architecture
2018	“Pre-train, then transfer”
2020	Scaling hypothesis (GPT-2 → GPT-3)
2022 - Q1	Instruction tuning (InstructGPT, T0)
2022 - Q2	Chain-of-thought reasoning
2022 - Q4	Self-improvement

From language models to general AI



Multitask Prompted Training Enables Zero-Shot Task Generalization”, Sanh *et al.*

“Training language models to follow instructions with human feedback”, Ouyang *et al.*

2015	“The Unreasonable Effectiveness of Recurrent Neural Networks”
2017	Transformer architecture
2018	“Pre-train, then transfer”
2020	Scaling hypothesis (GPT-2 -> GPT-3)
2022 - Q1	Instruction tuning (InstructGPT, T0)
2022 - Q2	Chain-of-thought reasoning
2022 - Q4	Self-improvement

From language models to general AI

PROMPT:

Explain the moon landing to a 6 year old.

GPT-3 COMPLETION: 🤦

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

[...]

InstructGPT COMPLETION: 👍

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Multitask Prompted Training Enables Zero-Shot Task Generalization”, Sanh *et al.*

“Training language models to follow instructions with human feedback”, Ouyang *et al.*

2015	“The Unreasonable Effectiveness of Recurrent Neural Networks”
2017	Transformer architecture
2018	“Pre-train, then transfer”
2020	Scaling hypothesis (GPT-2 -> GPT-3)
2022 - Q1	Instruction tuning (InstructGPT, T0)
2022 - Q2	Chain-of-thought reasoning
2022 - Q4	Self-improvement

From language models to general AI

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 **X**

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

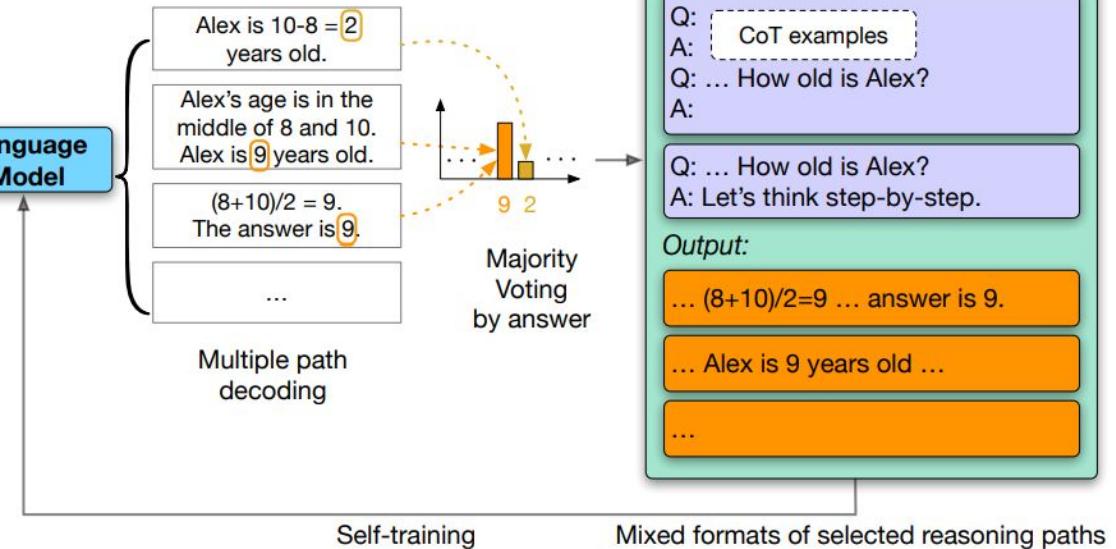
“Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, Wei et al.

“Large Language Models are Zero-Shot Reasoners”, Kojima et al.

2015	“The Unreasonable Effectiveness of Recurrent Neural Networks”
2017	Transformer architecture
2018	“Pre-train, then transfer”
2020	Scaling hypothesis (GPT-2 -> GPT-3)
2022 - Q1	Instruction tuning (InstructGPT, T0)
2022 - Q2	Chain-of-thought reasoning
2022 - Q4	Self-improvement

From language models to general AI

Q: John buys 20 cards and 1/4 are uncommon. How many uncommon cards did he get?
A: John gets $20 * 1/4 = 5$ uncommon cards. The answer is 5.
...
CoT examples
Q: Amy is 10. Jake is 8. Alex's age is right in the middle. How old is Alex?
A:
Training-set questions or self-generated questions



“Which of these assistant responses is less harmful? Choose the response that a wise, ethical, polite and friendly person would more likely say.”

“STaR: Bootstrapping Reasoning With Reasoning”, Zelikman et al.
“Large Language Models Can Self-Improve”, Huang et al.
“Constitutional AI: Harmlessness from AI Feedback”, Bai et al.

“The Unreasonable Effectiveness of Recurrent Neural Networks”

Transformer architecture

“Pre-train, then transfer”

ng hypothesis (GPT-2 -> GPT-3)

ction tuning (InstructGPT, T0)

Chain-of-thought reasoning

2022 - Q2

Self-improvement

2022 - Q4

Our mission:

Make advanced AI systems
robust, transparent and
fit to radically transform
science and medicine

USMLE [test:731] Three days after induction chemotherapy, a 9-year-old boy is admitted to the hospital for fever and neutropenia. He was initially diagnosed with t(9;22)-positive acute T cell lymphoblastic leukemia following workup for fatigue and pancytopenia. He is fully conscious. His blood pressure is 110/65 mm Hg, temperature is 39.5°C (103.1°F), pulse is 98/min, and respiratory rate is 14/min. Physical examination including examination of the central venous catheter site shows no source of infection. Broad-spectrum antibiotics are initiated. Which of the following factors about this patient most strongly indicates a poor prognosis?

- A) Age B) Cytogenetics C) Fever and neutropenia D) Initial presentation

✓ Let's use step by step inductive reasoning, given the medical nature of the question.

The first thing that stands out is that the patient is 9 years old. In general, children have a better prognosis than adults when it comes to leukemia.

The next thing to consider is the patient's cytogenetics. The fact that the patient has t(9;22)-positive acute T cell lymphoblastic leukemia indicates that the patient has a specific chromosomal abnormality that is associated with a poor prognosis.

The next thing to consider is the patient's fever and neutropenia. Fever and neutropenia are both signs of infection, and infection

is a serious complication in leukemia patients, particularly in those who are immunocompromised. Finally, we need to consider the patient's initial presentation. The fact that the patient was initially diagnosed with leukemia following workup for fatigue and pancytopenia indicates that the patient's leukemia was advanced at the time of diagnosis.

All of these factors indicate a poor prognosis for this patient.

Therefore, among A through D, the answer is B, cytogenetics.

✗ Let's think step by step like a medical expert. The first step is to gather information about the patient. In this case, we are given the following information:

-The patient is 9 years old
-The patient has t(9;22)-positive acute T cell lymphoblastic leukemia
-The patient is currently experiencing fever and neutropenia

The next step is to analyze the information and look for any patterns or clues that could help us answer the question.

In this case, the most important factor that indicates a poor prognosis is the patient's age. Children under the age of 10 have a much higher chance of relapsing after treatment and a lower chance of achieving remission.

Therefore, among A through D, the answer is A.

✗ Let's derive the differential diagnosis step by step:

- 1) The patient has fever and neutropenia.
- 2) The patient is 9 years old.
- 3) The patient has t(9;22)-positive acute T cell lymphoblastic leukemia.
- 4) The patient's initial presentation was fatigue and pancytopenia.

The most likely cause of the patient's fever and neutropenia is infection. The patient's age, cytogenetics, and initial presentation are all risk factors for infection.

Therefore, the factor that most strongly indicates a poor prognosis is the patient's age.

Therefore, among A through D, the answer is A.

Pattern A Correct reasoning step

Pattern B Correct recall of knowledge

Pattern C Correct reading comprehension

Pattern D Incorrect reasoning step

Pattern E Incorrect or insufficient knowledge

Pattern F Incorrect reading comprehension



ThoughtSource

Question:

In fruit flies, red eyes are dominant over sepia eyes. A male with red eyes is crossed with a female with sepia eyes producing first generation offspring that all had sepia eyes. What does this indicate about the flies that were crossed?

- A) Both parents are heterozygous for the trait.
- B) The male is heterozygous and the female is homozygous recessive for the trait.
- C) The male is homozygous dominant and the female is heterozygous for the trait.

Chain-of-thought:

Red eyes are dominant over sepia eyes. This means that if a fly has one allele for red eyes and one allele for sepia eyes, their eyes will be red.

Each fruit fly inherits one of the alleles from the father and one of the alleles from the mother.

A male with red eyes was crossed with a female with sepia eyes. This means that the male fly has at least one allele for red eyes, and the female fly has two alleles for sepia eyes. So we know the exact alleles of the mother, but we are still unsure about the alleles of the father.

All of the offspring had sepia eyes. This means that they all had two alleles for sepia eyes. Since one allele is inherited from the father, this means that the father had one allele for sepia eyes, i.e. the male is heterozygous.

Answer:

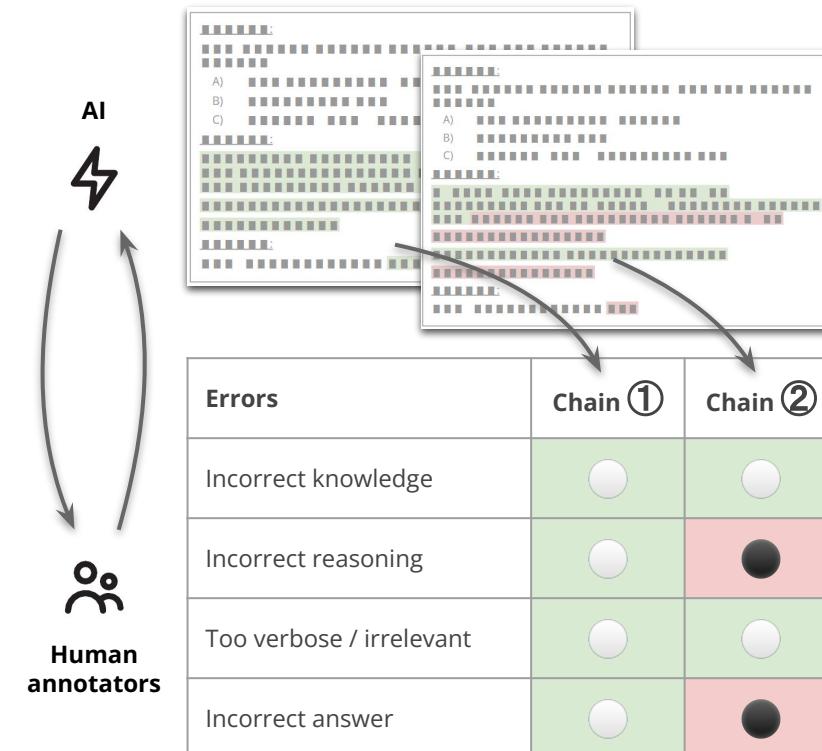
Therefore, among A through C, the answer is **B**. (correct answer)

AI models generate complex reasoning chains and answers



(example generated with text-davinci-002 model)

Human annotators evaluate reasoning chains.
Annotations can be used to further train AI models.



<https://github.com/OpenBioLink/ThoughtSource/>

Question:

In fruit flies, red eyes are dominant over sepia eyes. A male with red eyes is crossed with a female with sepia eyes producing first generation offspring that all had sepia eyes. What does this indicate about the flies that were crossed?

- A) Both parents are heterozygous for the trait.
- B) The male is heterozygous and the female is homozygous recessive for the trait.
- C) The male is homozygous dominant and the female is heterozygous for the trait.

Chain-of-thought:

Red eyes are dominant over sepia eyes. This means that if a fly has one allele for red eyes and one allele for sepia eyes, their eyes will be red.

Each fruit fly inherits one of the alleles from the father and one of the alleles from the mother.

A male with red eyes was crossed with a female with sepia eyes. This means that the male fly has at least one allele for red eyes, and the female fly has two alleles for sepia eyes. So we know the exact alleles of the mother, but we are still unsure about the alleles of the father.

All of the offspring had sepia eyes. This means that they all had two alleles for sepia eyes. Since one allele is inherited from the father, this means that the father had one allele for sepia eyes, i.e. the male is heterozygous.

Answer:

Therefore, among A through C, the answer is **B.** (correct answer)



AI models generate complex reasoning chains and answers

(example generated with text-davinci-002 model)

 ThoughtSource Annotator

tfidf ▾ Visualisation threshold  Download current Matthias Logout

Question

Stars are organized into patterns called constellations. One constellation is named Leo. Which statement best explains why Leo appears in different areas of the sky throughout the year?

- A. Earth revolves around the sun.**
- B. The sun revolves around Earth.
- C. The constellations revolve around Earth.
- D. Earth revolves around the constellations.

The first thing we need to know is that the Earth revolves around the sun. This means that the Earth is constantly moving and changing its position in relation to the sun. The second thing we need to know is that the constellations are fixed patterns in the sky. This means that they do not change their position in relation to the Earth. So, when we look at the sky, we see different constellations at different times of the year because the Earth is in a different position in relation to the constellations. The constellations appear to move across the sky because the Earth is moving.

 Answer: A) Earth revolves around the sun.



- Incorrect reasoning
- Insufficient knowledge
- Incorrect reading comprehension
- Too verbose

The first thing to note is that the constellations don't actually move. They just appear to move because of the way Earth moves. The second thing to note is that Earth revolves around the sun. So, as Earth moves around the sun, the constellations appear to move in the sky. This is why Leo appears in different areas of the sky throughout the year.

 Answer: A) Earth revolves around the sun.



- Incorrect reasoning
- Insufficient knowledge
- Incorrect reading comprehension
- Too verbose



<https://github.com/OpenBioLink/ThoughtSource/>

ThoughtSource

Scientific / medical datasets

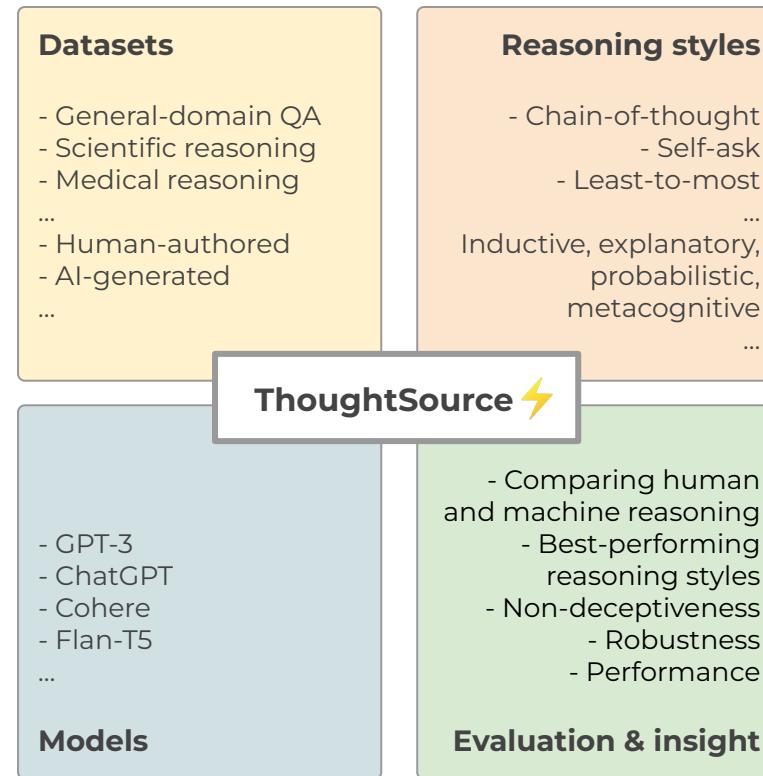
- WorldTree V2
- EntailmentBank
- OpenBookQA
- MedQA
- MedMCQA
- PubmedQA
- [...]

General-domain QA datasets

- CommonsenseQA
- StrategyQA
- QED

Math word problem datasets

- AQUA
- ASDiv
- GSM8K
- MAWPS
- SVAMP



**Contributors
welcome :)**



<https://github.com/OpenBioLink/ThoughtSource/>

Thanks!

🌐 <https://samwald.info/>

🐦 <https://twitter.com/matthiassamwald>

🐙 <https://github.com/OpenBioLink/ThoughtSource/>

@ matthias.samwald [at] meduniwien.ac.at



Applying diffusion models to protein design

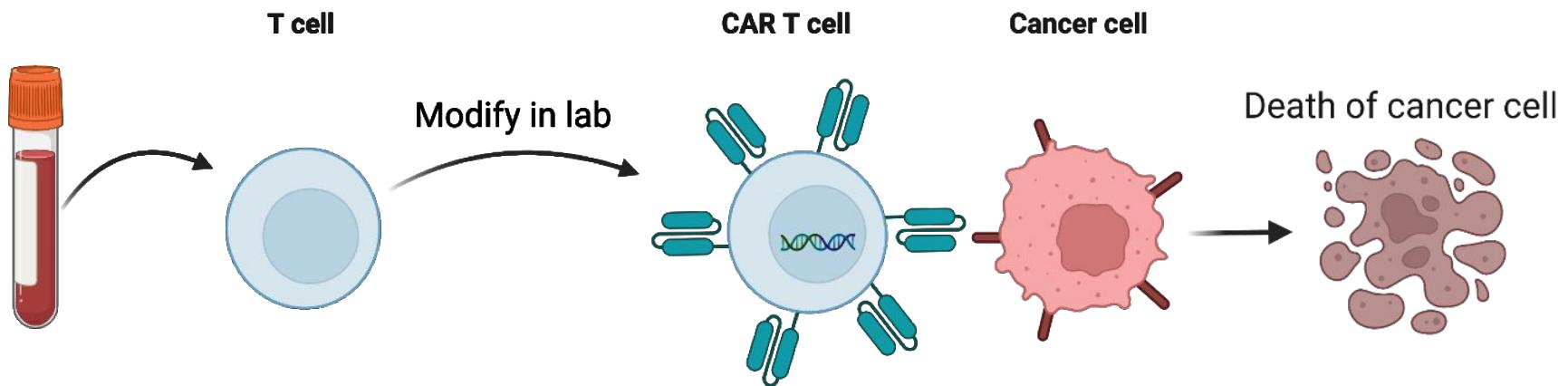
Moritz Schaefer

Postdoc @ AI Institute/Bock lab

MEDICAL UNIVERSITY
OF VIENNA

Applying diffusion models to protein design
Institute of Artificial Intelligence

“CAR T cell therapy” - Supercharging immune cells to fight cancer



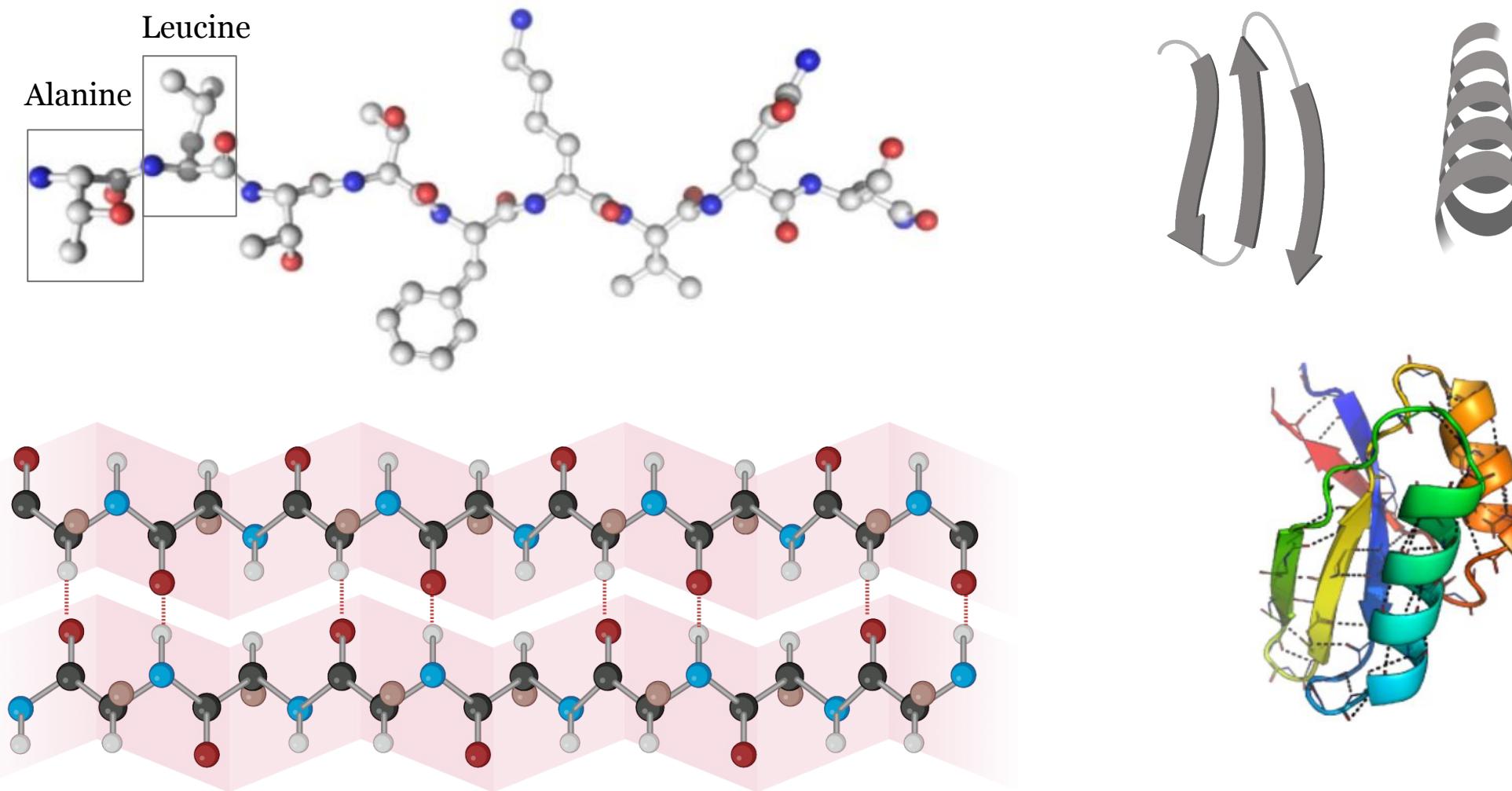
T cells: Killer machines of our immune system

Receptor: Protein that binds specific molecules

CAR: Receptor that helps T cells to bind to cancer

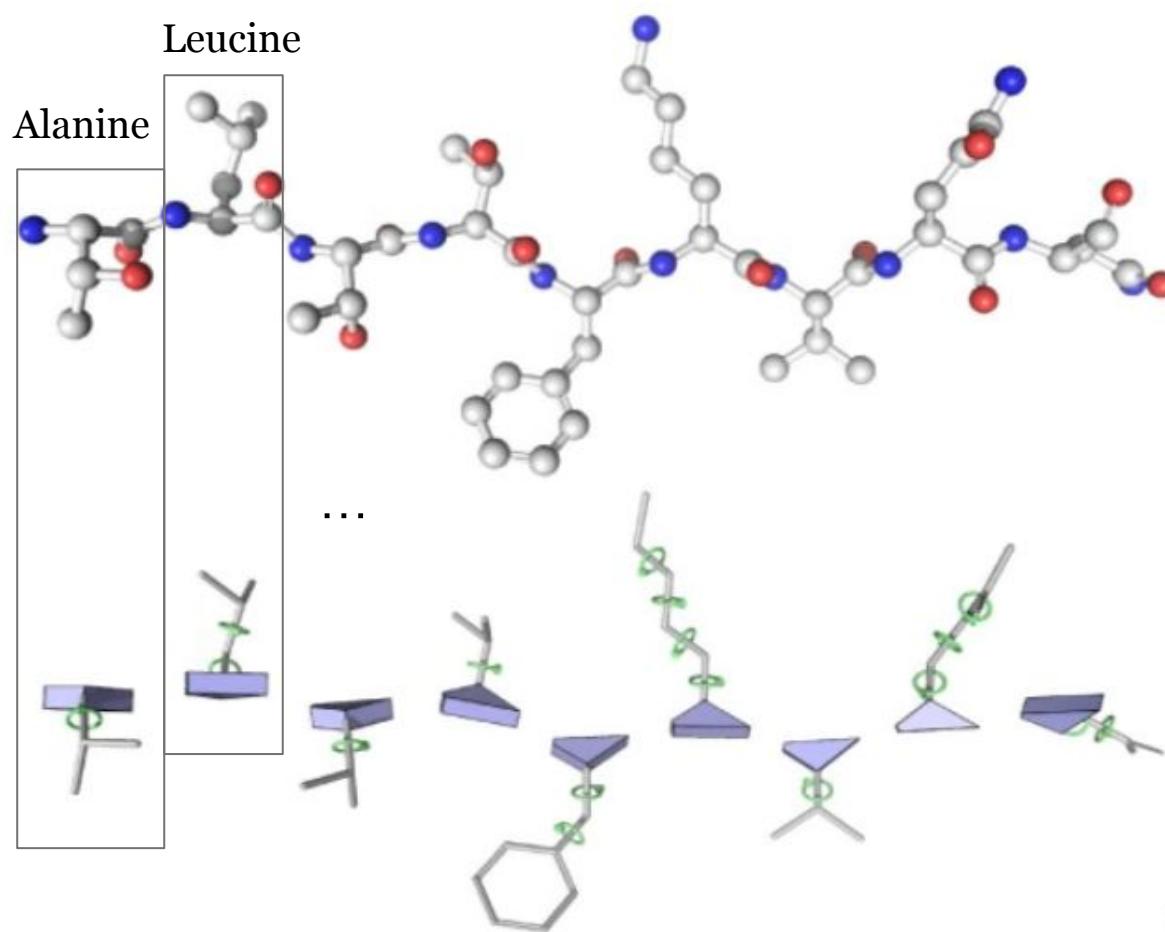
(Biorender)

Proteins - the functional building blocks of life

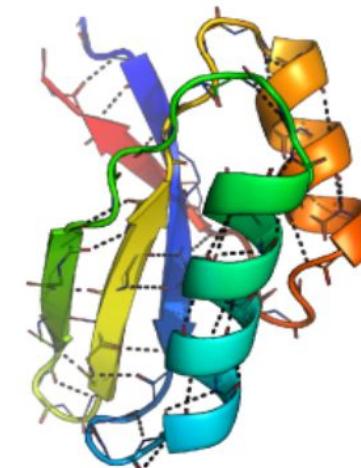


(Jumper et al., 2021; Anand et al., 2022; Biorender)

Proteins - the functional building blocks of life

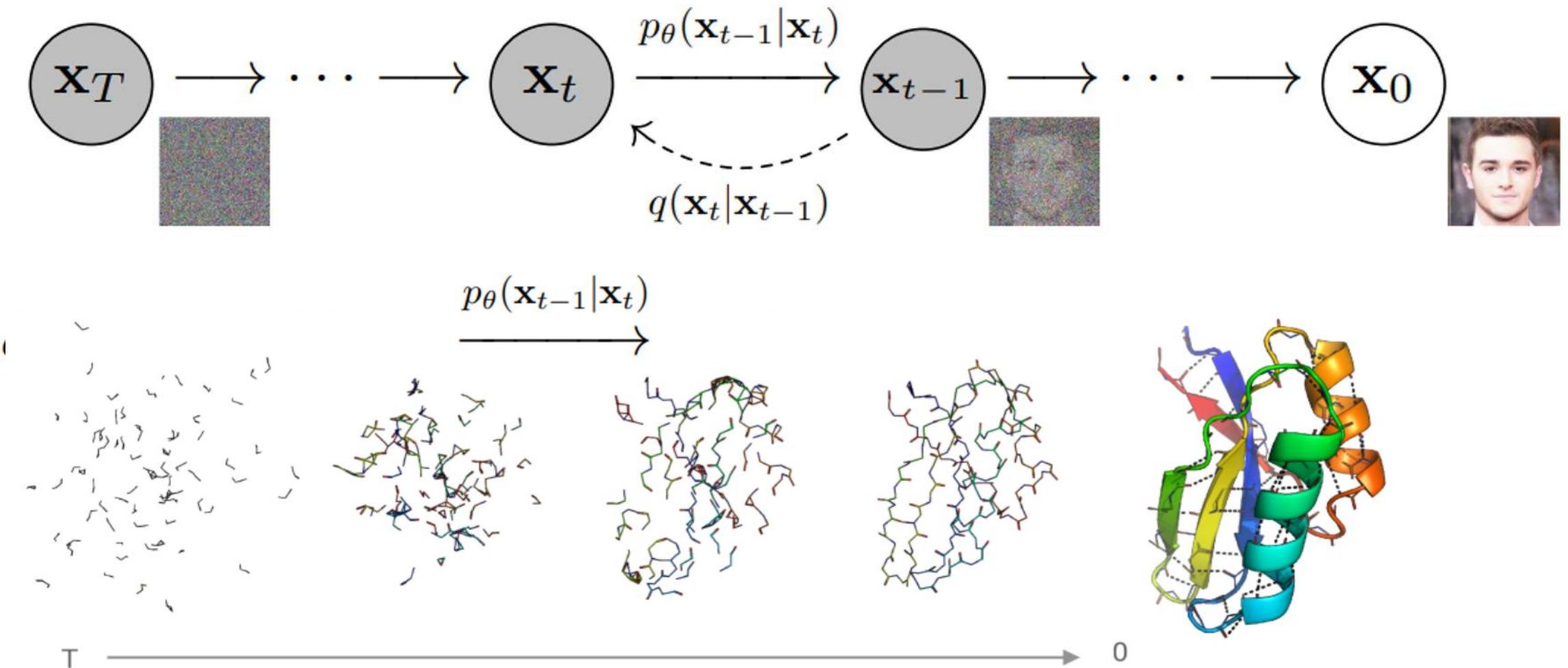


- Amino-acid positions $\mathbf{x} \in \mathbb{R}^{3 \times n}$
- Amino-acid orientation $\mathbf{q} \in SO(3)^n$
- Amino-acid type $\mathbf{s} \in \mathbb{A}^n$
- Side-chain angles $\chi \in [-\pi, \pi]^{4 \times n}$



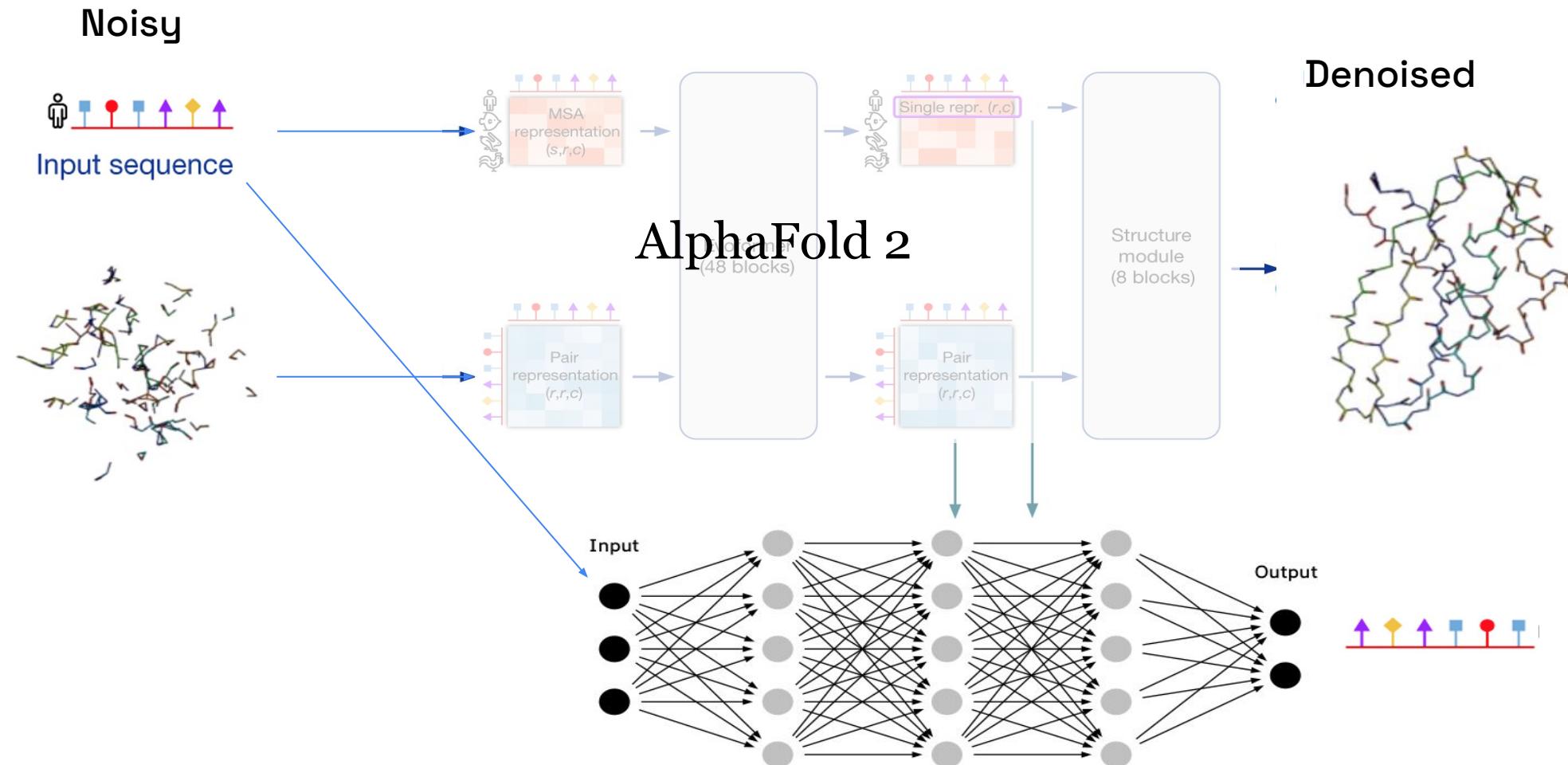
(Jumper et al., 2021, Anand et al., 2022)

Diffusion models - From image generation to proteins



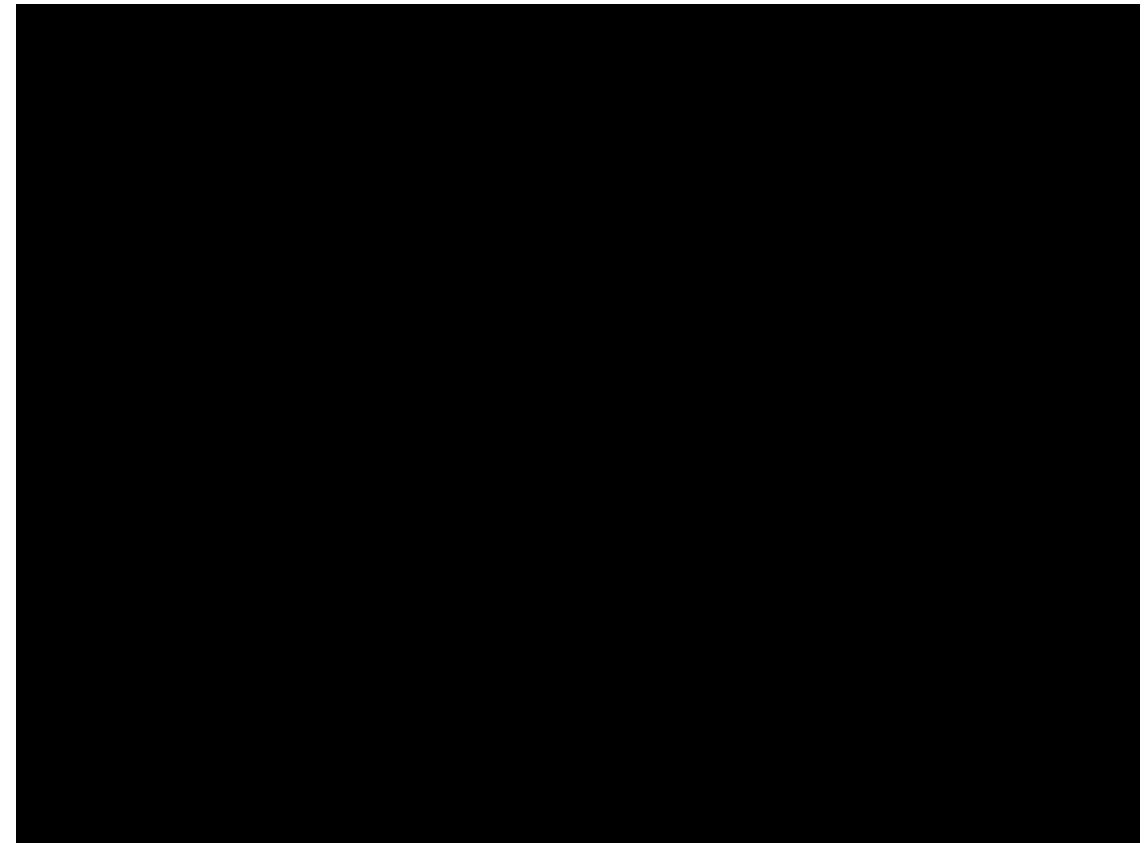
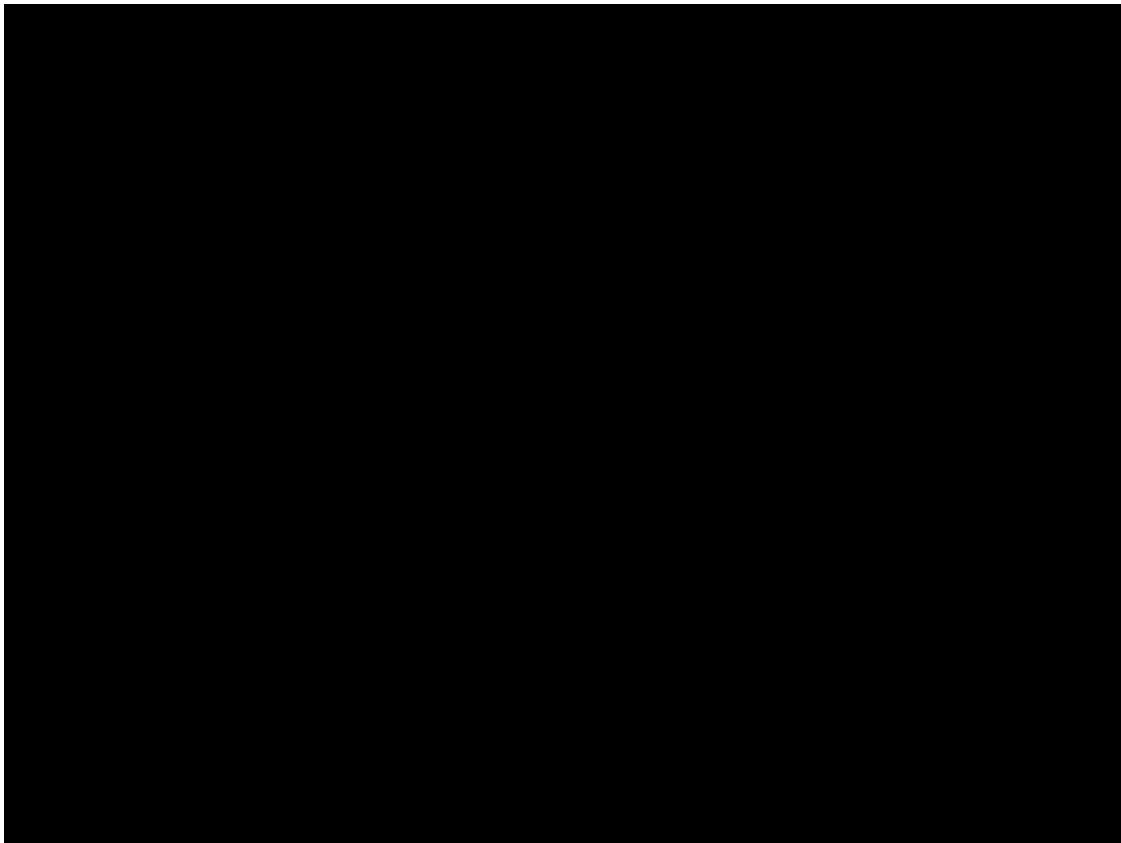
(Ho et al., 2021)

AlphaFold 2 is an excellent denoising model

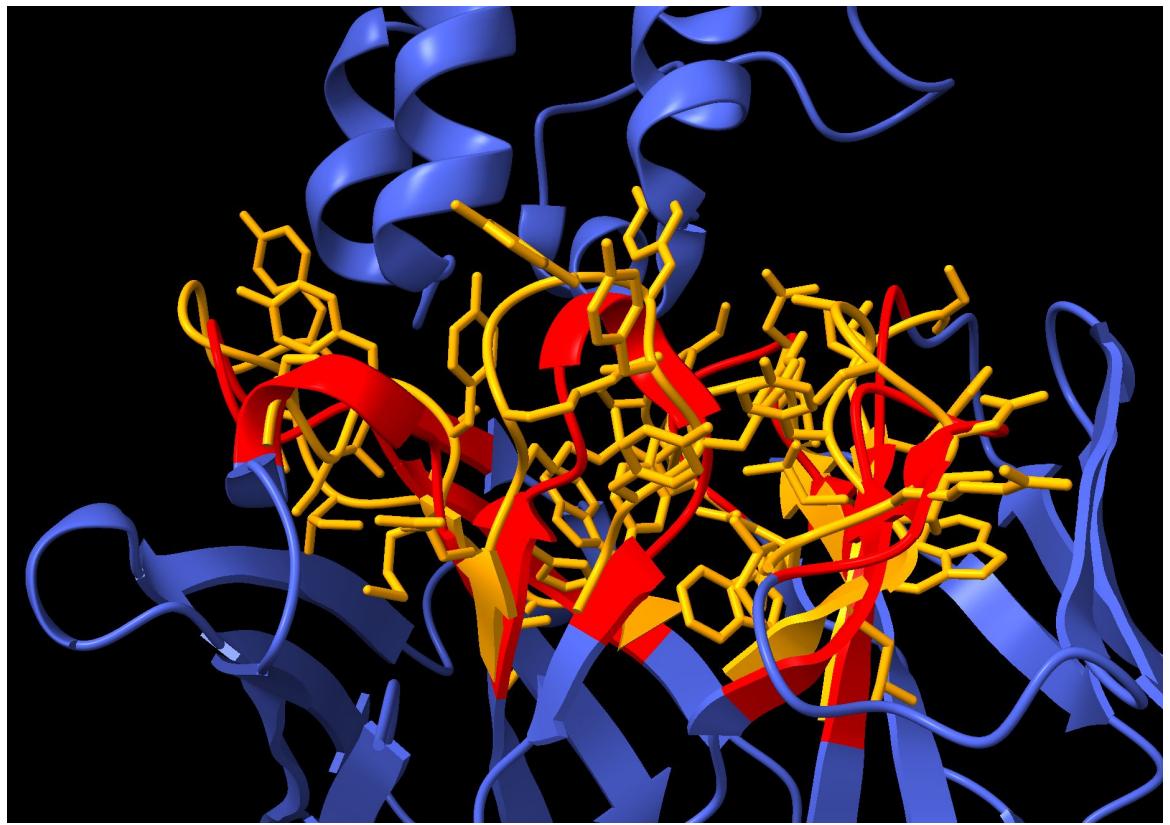
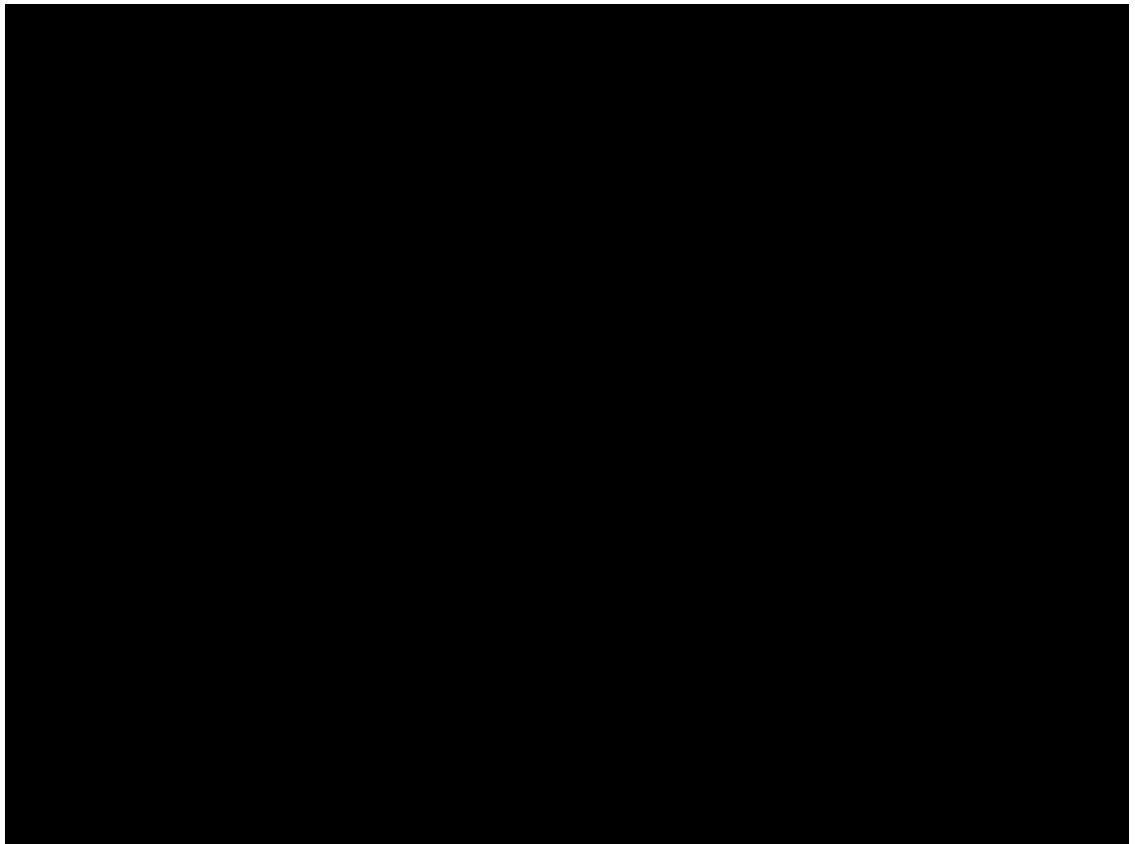


(Jumper et al., 2021; Anand et al., 2022; Watson et al., 2022)

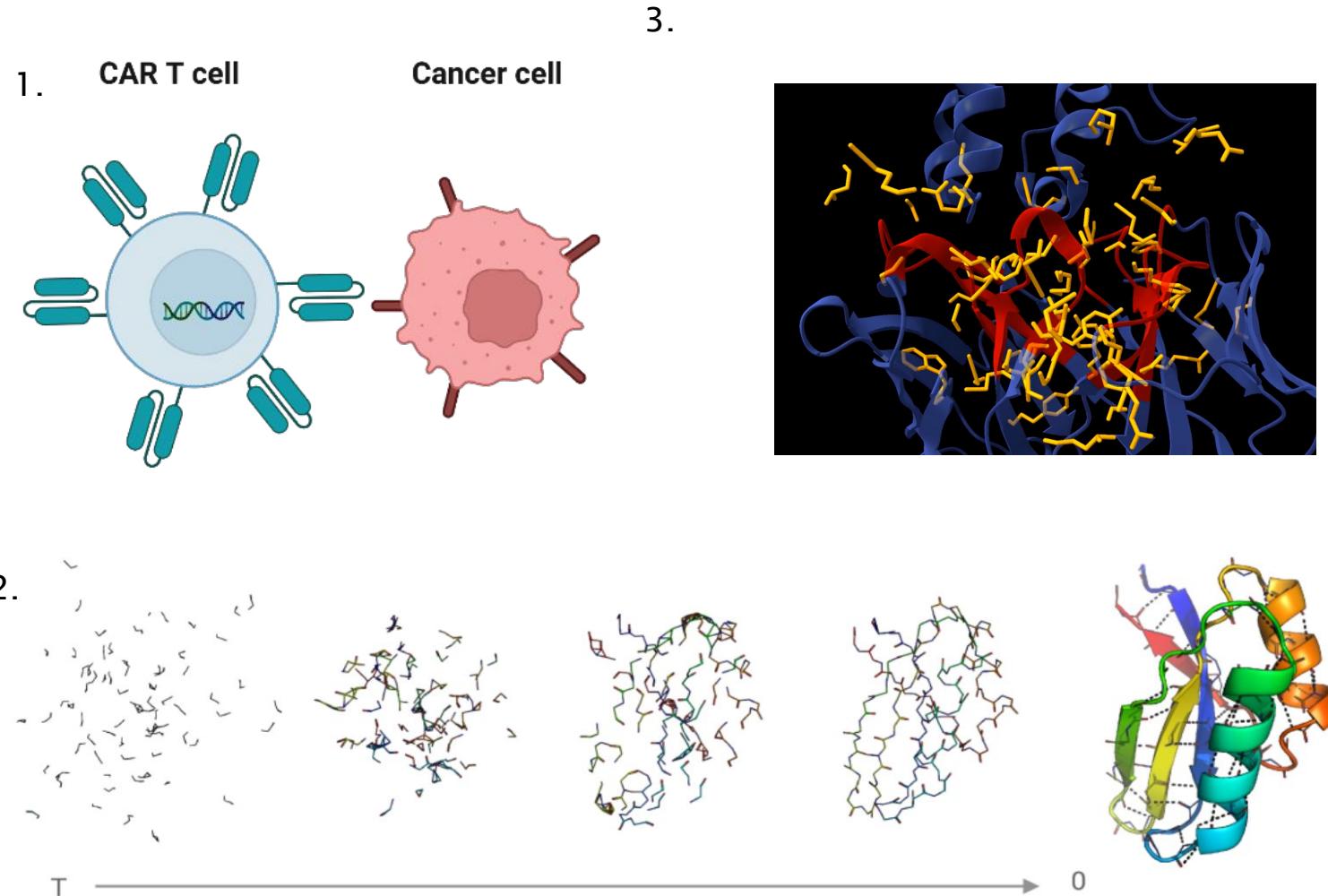
Results



Results



Conclusions



@muronglizi



[linkedin.com/in/
moritzschaef](https://linkedin.com/in/moritzschaef)



@moritzschaef
@qoto.org



mschaefer@cemm.at

Deep learning models applied to genomics data

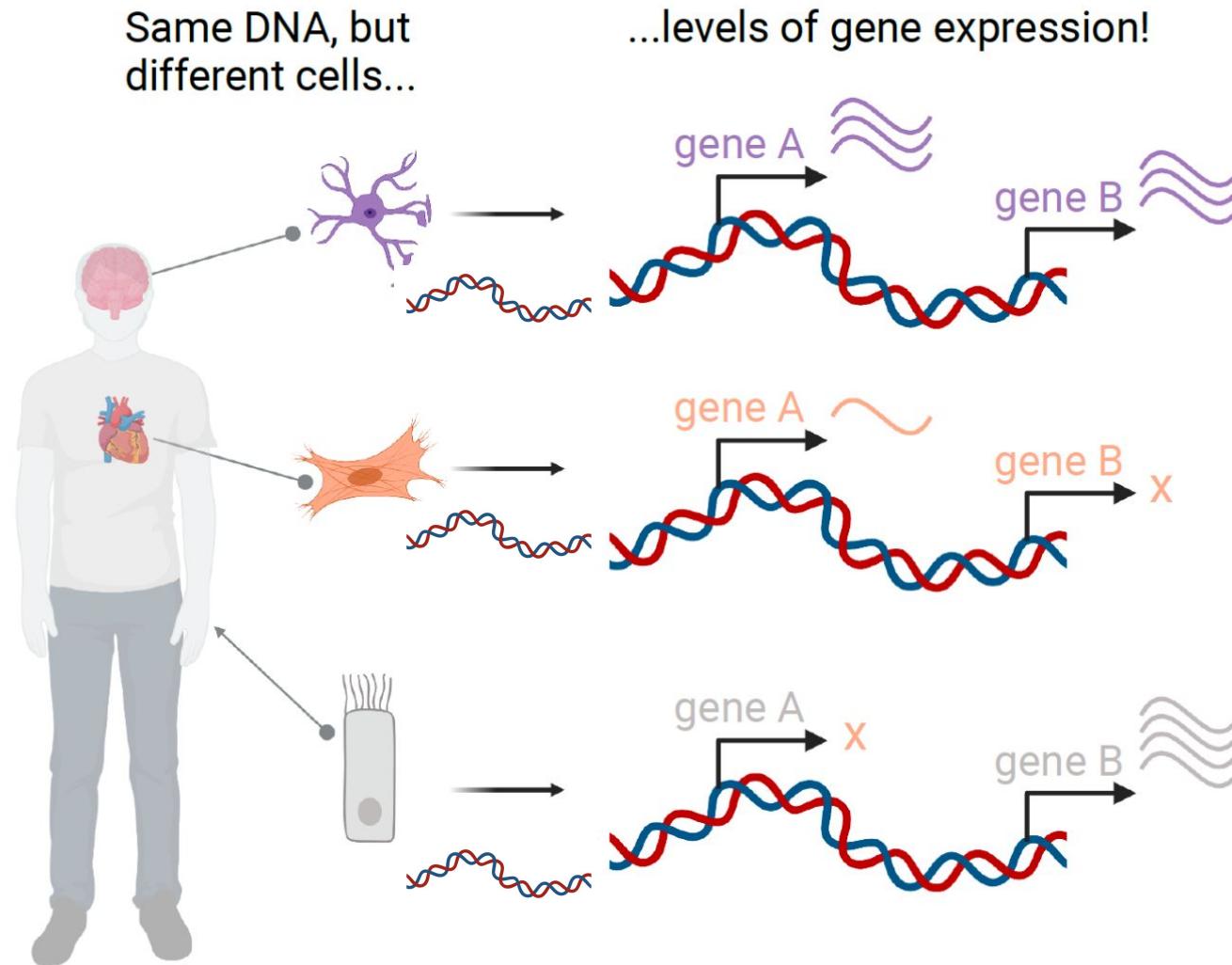
Daria Romanovskaya

PhD student @Bock lab/AI institute

 @daria_rom__

 dromanovskaya@cemm.at

How do we measure differences in cells?

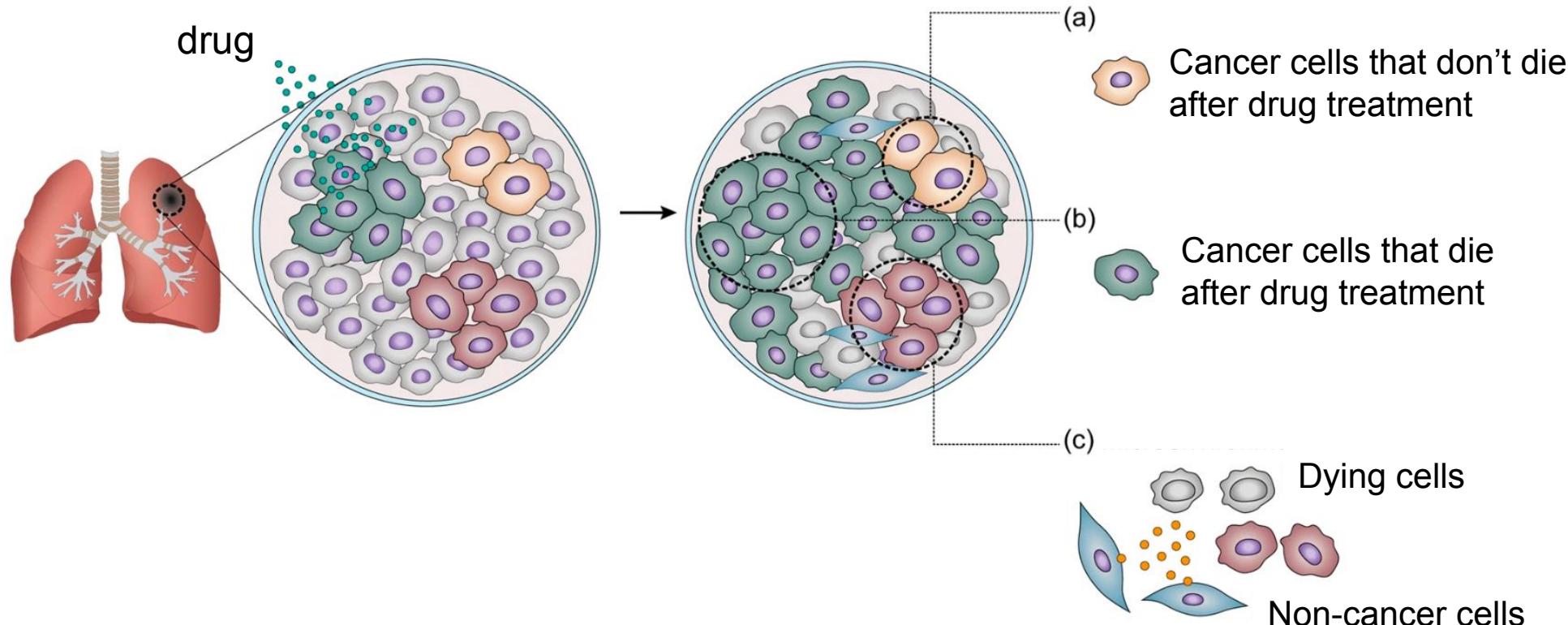


Feature matrix:

	geneA	geneB	geneZ
cell1	5	4	1
cell2	0	9	2
cell3	4	5	8
cell4	6	1	2
.....
cellN

With biorender.com

How understanding those differences could be applied in clinics?



- Groups (or cell types) present
- Key differences between them
- New characteristics and targets

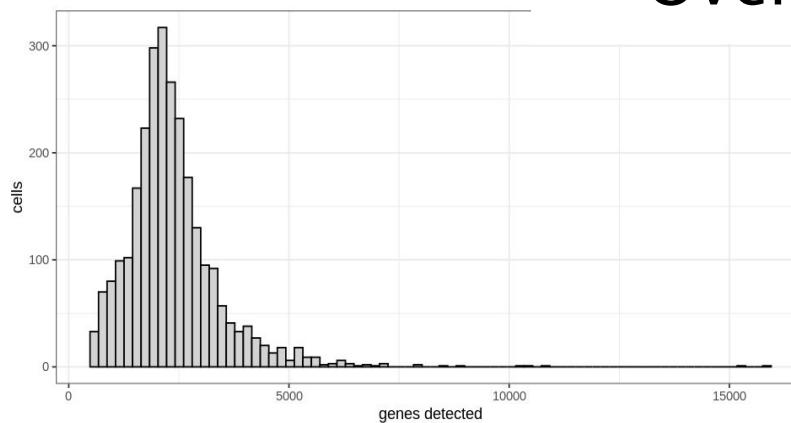
Hwang et al.:<https://doi.org/10.1038/s12276-018-0071-8>

Challenges in single-cell data analysis

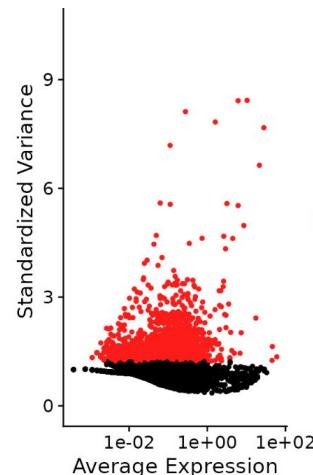
Within each dataset:

	geneA	GeneB	geneZ
cell1	5	4
cell2	0	9
cell3	4	5
cell4	6	1
.....			
cellN			

2700 cells

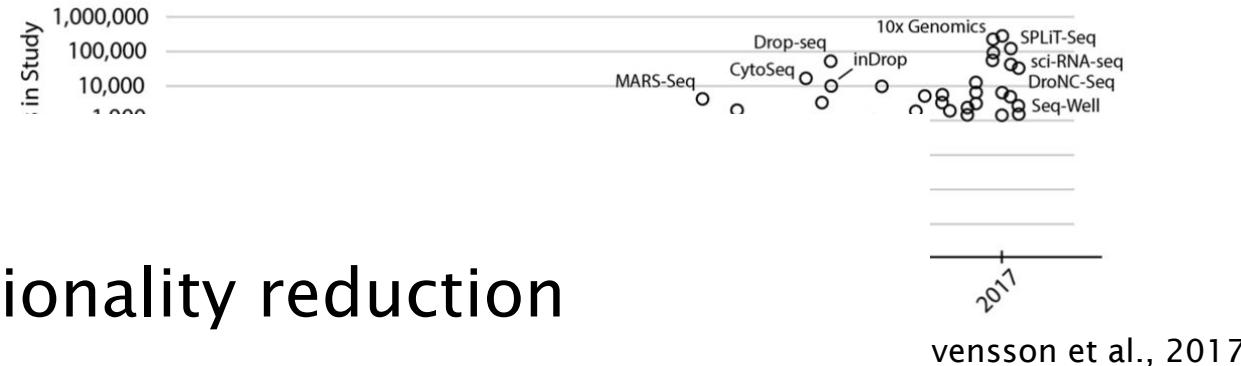


-> best described by regularized negative binomial regression

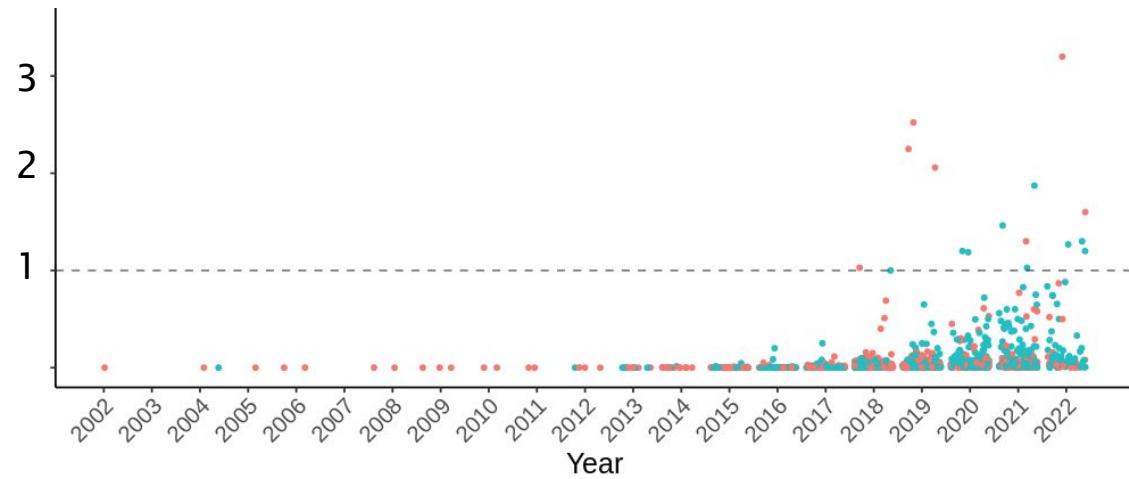


- Up to 36k features
- Sparse data
- Significance testing
- Feature selection
- Meaningful dimensionality reduction
- Overcoming sparsity and denoising the data

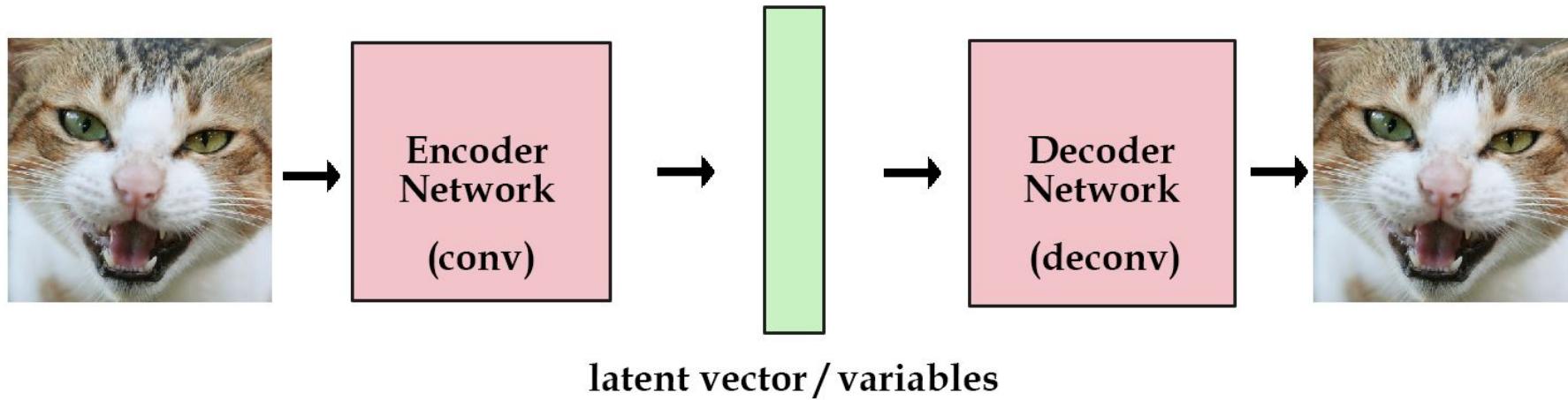
We live in the era of extremely large datasets:



vensson et al., 2017



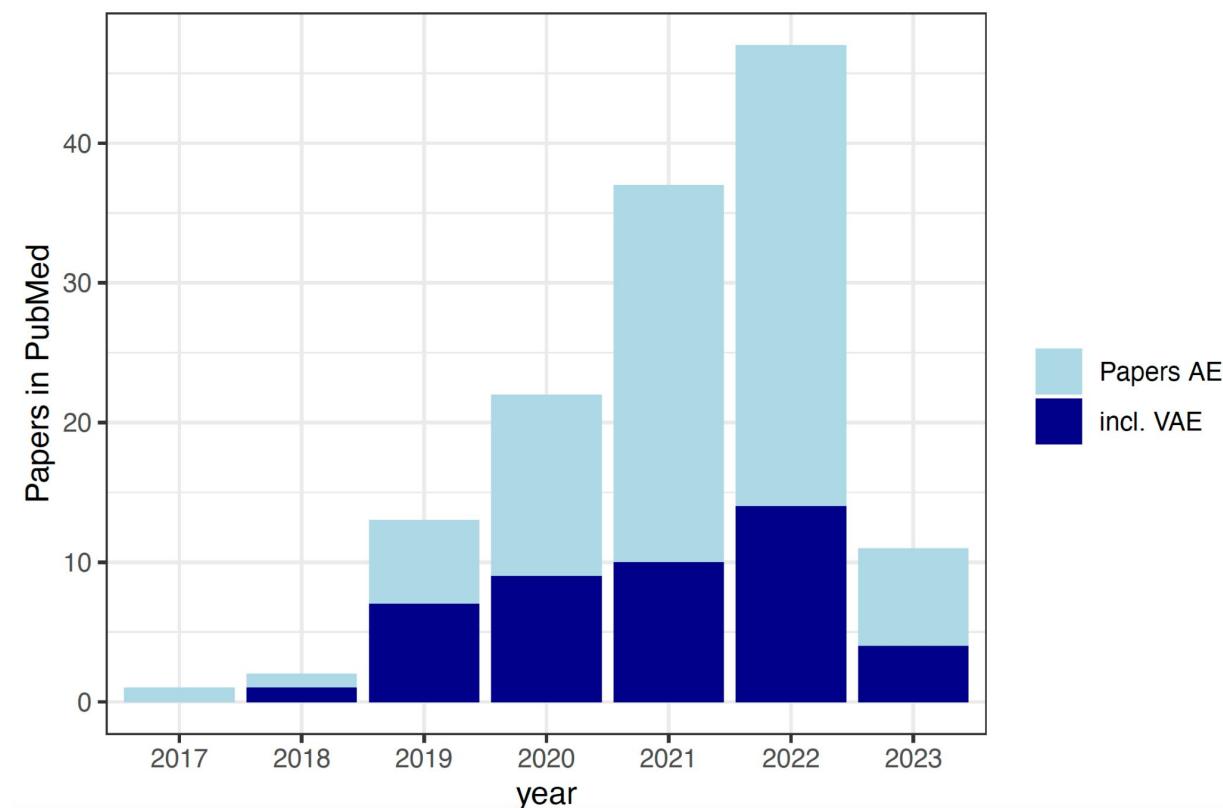
Autoencoders as a tool for analysing transcriptomic data



<https://kvfrans.com/variational-autoencoders-explained/>

<https://towardsdatascience.com/understanding-variational-autoencoders-vae-f70510919f73>

Autoencoders as a tool for analysing transcriptomic data



Article | Open Access | Published: 23 January 2019

Single-cell RNA-seq denoising using a deep count autoencoder

Gökçen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller & Fabian J. Theis

Nature Comm.

51k Accesses

Interpretable factor models of single-cell RNA-seq via variational autoencoders

Valentine Svensson , Adam Gayoso, Nir Yosef, Lior Pachter

Fast and precise single-cell data analysis using a hierarchical autoencoder

Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N. Luu & Tin Nguyen

Nat.

13k Citations

Deep generative modeling for single-cell transcriptomics

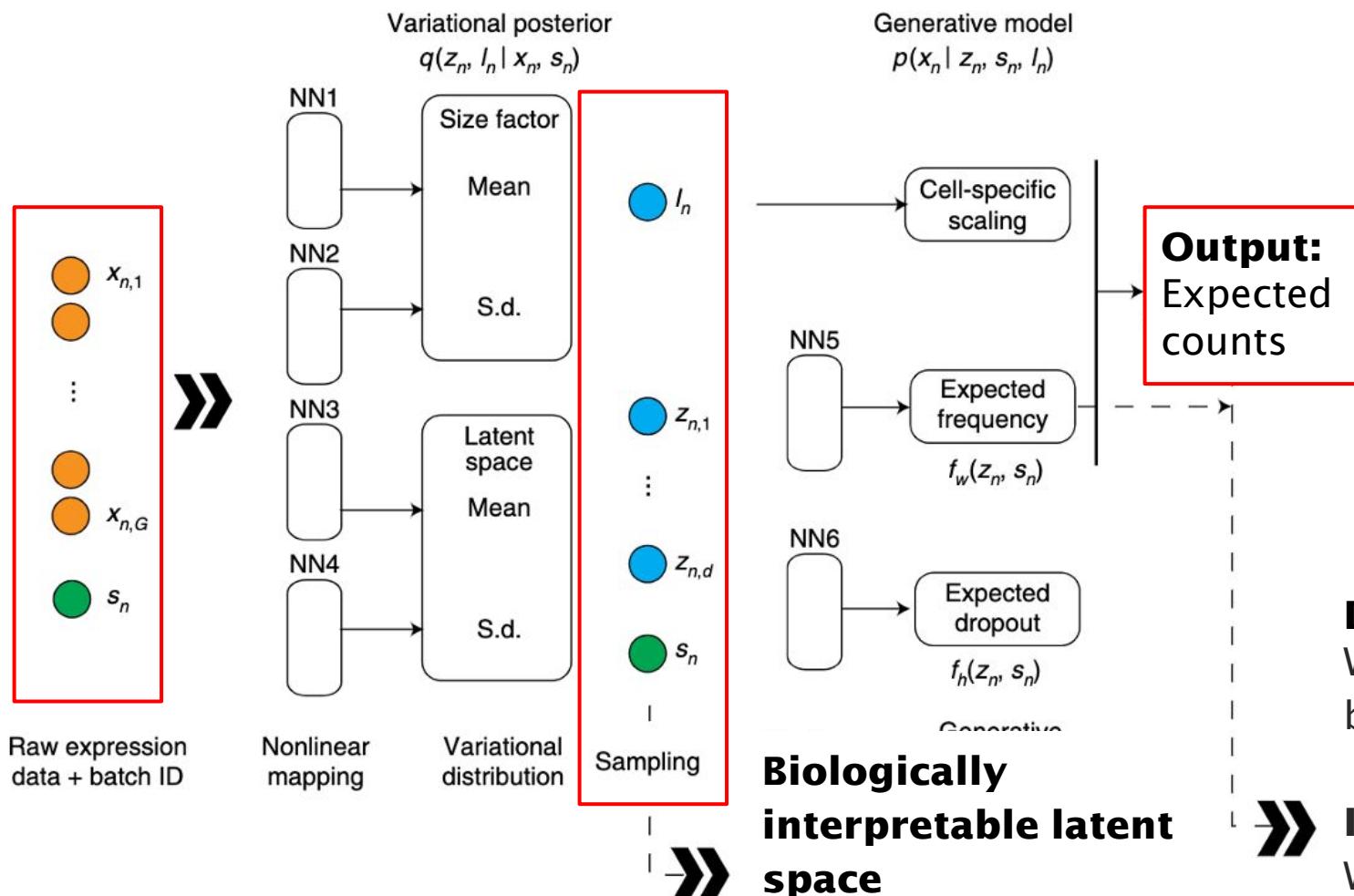
Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan & Nir Yosef

Nature Methods 15, 1053–1058 (2018) | Cite this article

58k Accesses | 483 Citations | 178 Altmetric | Metrics

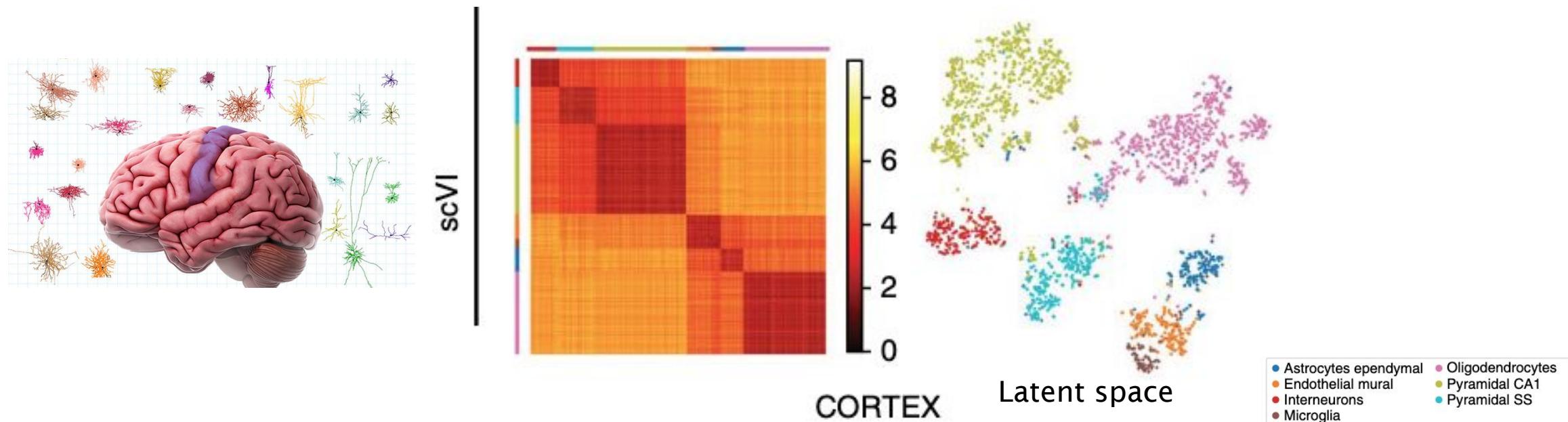
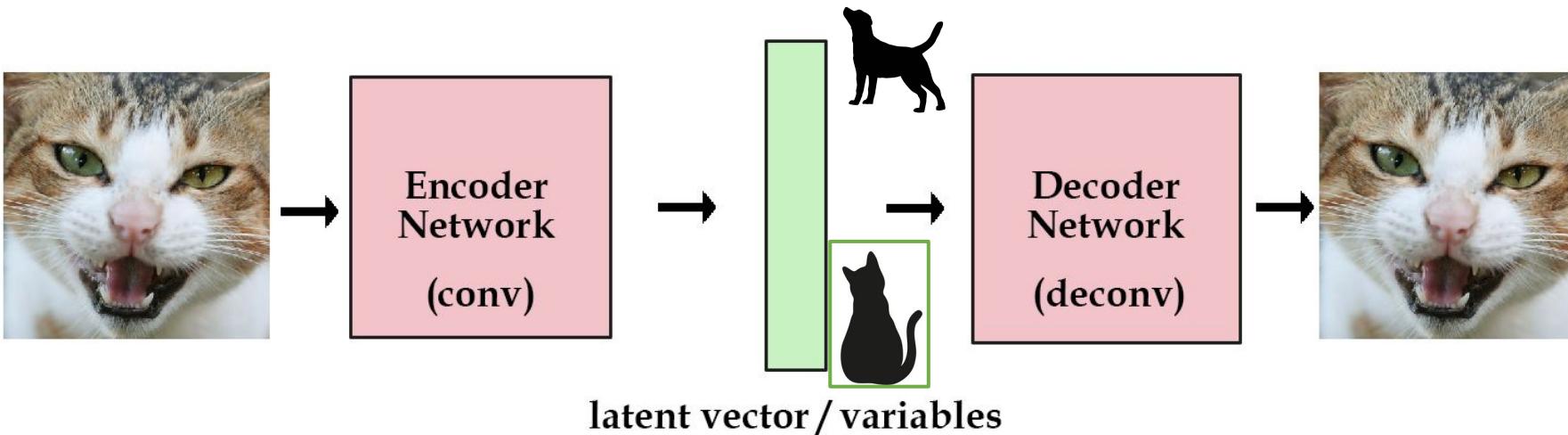
scVI model (not only) for dimensionality reduction in transcriptomics data

Input:
gene
expression
matrix



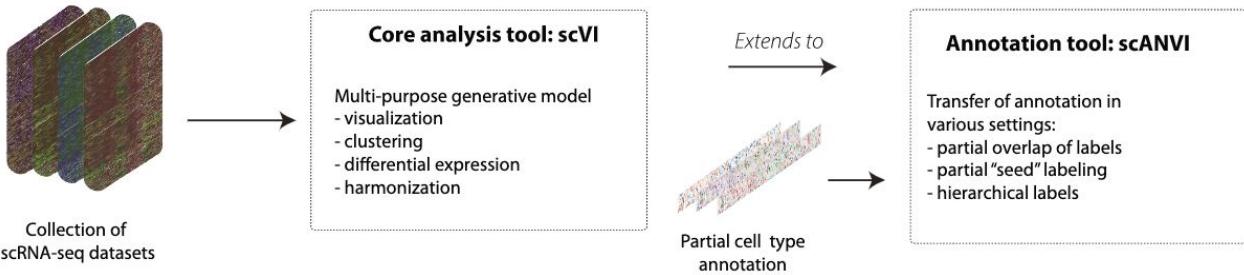
https://docs.scvi-tools.org/en/stable/user_guide/models/scvi.html

Biological signal is retained in the scVI latent space



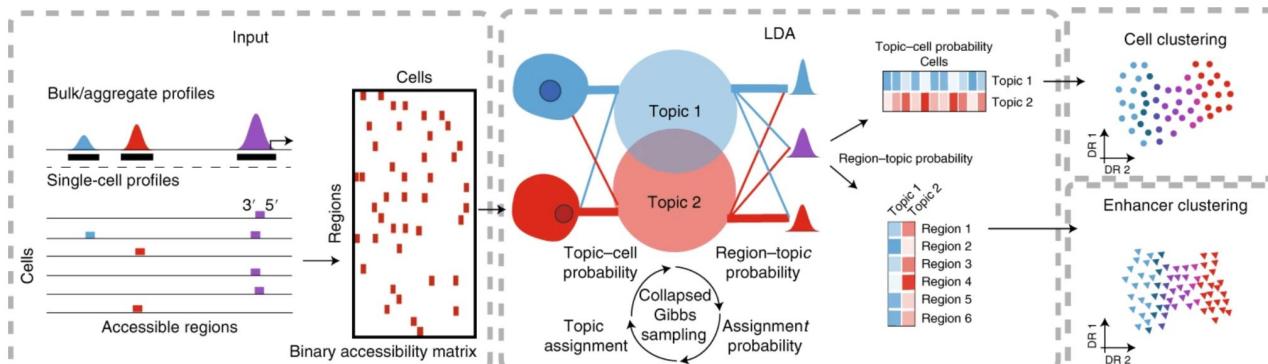
Wider application of deep learning methods in biomedical data analysis

Semi-supervised approach for leveraging existent cell-type annotation



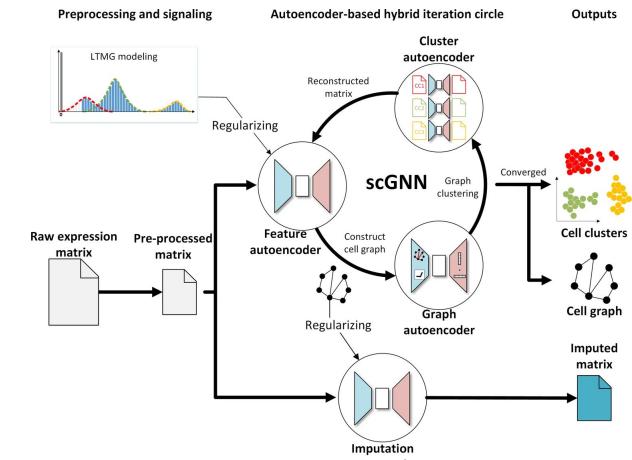
Xu et al.: <https://doi.org/10.15252/msb.20209620>

Topic modelling for dim. reduction in epigenetic data



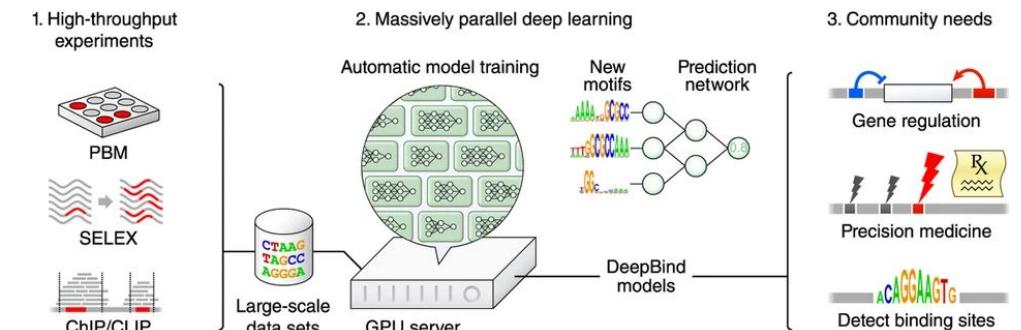
González-Blas: <https://doi.org/10.1038/s41592-019-0367-1>

Graph autoencoders



Wang et al.: <https://doi.org/10.1038/s41467-021-22197-x>

Convolutional neural networks to predict binding sites



Alipanahi et. al: <https://doi.org/10.1038/nbt.3300>

Thank you!

Talk 1: Machine learning for molecular medicine

Talk 2: Reasoning in large language models

Talk 3: Applying diffusion models to protein design

Talk 4: Deep learning models applied to genomics data

Prof. Christoph Bock, Prof. Matthias Samwald,

Dr. Moritz Schaefer, Daria Romanovskaia

AI Institute at the Medical University of Vienna

18:30 Intro

18:50 Synthetic Data - What is it, how to get it, and how can it help you

Paul Tiwald, mostly.ai

19:40 Announcements & Job Openings

Break & Buffet

20:20 AI in BioMedicine

Introduction and overview of the AI Institute at MedUni Vienna

Christoph Bock (Institute Director)

Scientific reasoning in large language models Matthias Samwald (Group Leader)

Applying diffusion models to protein design Moritz Schaefer (PostDoc)

How does deep learning help analyzing molecular data for precision medicine?

Daria Romanovskaia (PreDoc)



Next meetup:

February 28 @ A1

Musika! Fast Infinite Waveform Music Generation

Marco Pasini, JKU

Neural Program Synthesis - An Overview

Gabriele Libardi

Hot Papers: Language Models & Prompting

Michael Pieler

Send us announcements
& job openings!
contact@vdlm.at