# Privacy Research at SAP

Volkmar Lotz, Francesco di Cerbo, SAP
April 27, 2023

PUBLIC

THE BEST RUN SAP

# Agenda

Security and privacy research at SAP – an overview

Generating Realistic Synthetic Curricula Vitae for Machine Learning Applications under Differential Privacy

# Security and Privacy Research at SAP

# About SAP Security Research

37 researchers (incl. 6 Phd students)
+ 24 students (Feb ‘23)

2 Locations: France and Germany

Focus on H3 organic innovation: „invent the future of security“

Scientific background: regular top4 conference papers, PhD program, …

8 collaborative research projects (EU or national funding)

22 years of applied research

- IoT security, Encrypted DB, AI for Threat Intelligence, Privacy, Software and OS security
- Training, Security Consulting, Certification, M&A

# Why security research at SAP?

## Security risks are evolving and emerging

- Increased risk exposure and impact through extended attack surface, stronger dependencies, complex supply chains and increased asset value / criticality
- Evolving threat landscape, AI empowered attackers, tension in world economy and politics
- Compliance risks, trend to regulate security and privacy
- New technologies, inherent technical complexity, risk of overlooking critical threats (e.g., AI and ML)
- Combinatorial complexity, finding the needle in the haystack (assessing false positives)

## Grand challenges in security and privacy are unsolved

- Secure and privacy-friendly data business
- Trustworthy large-scale and distributed systems, including AI
- Elimination of software vulnerabilities
- Self-healing systems and applications
- Usable security and human factors
- Impact of quantum technologies

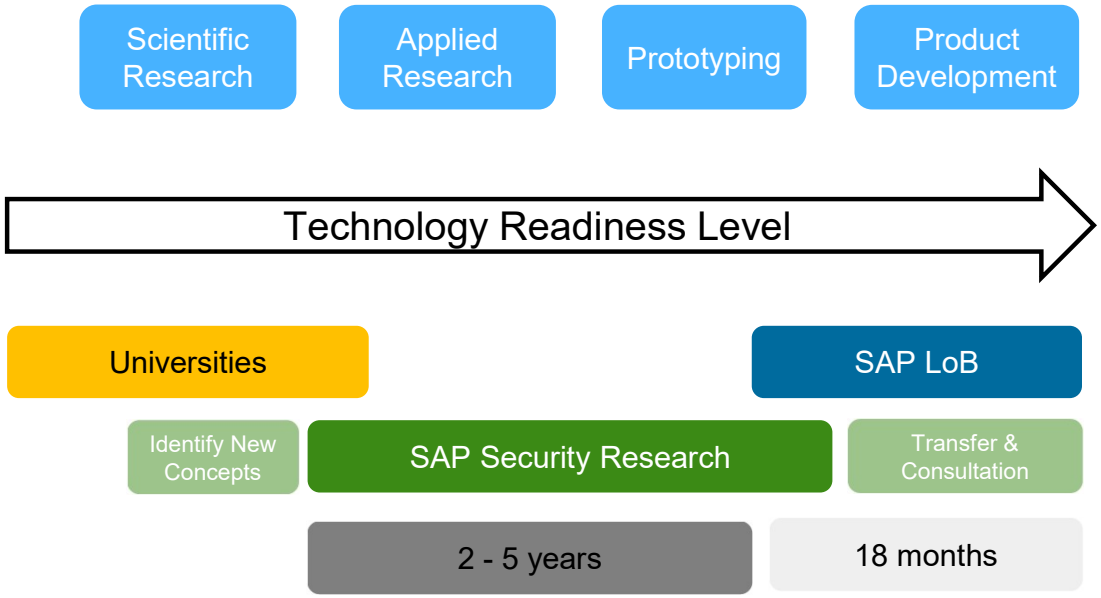
# Applied Research: Bridging Academia with SAP Business

## Business-driven

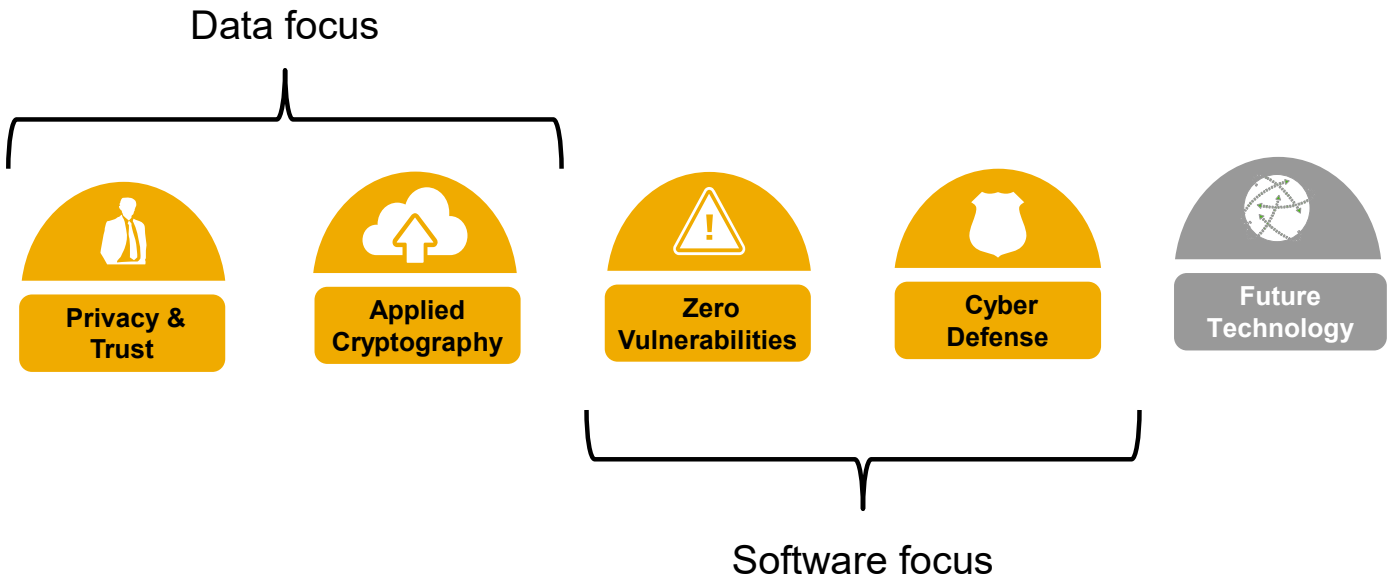
- Our research is driven by the needs of SAP and our customers, and is aligned with SAP's strategy and business. We apply new principles in security research through business value analysis and by refining and adapting methodologies and technology concepts

## Ready to be adopted

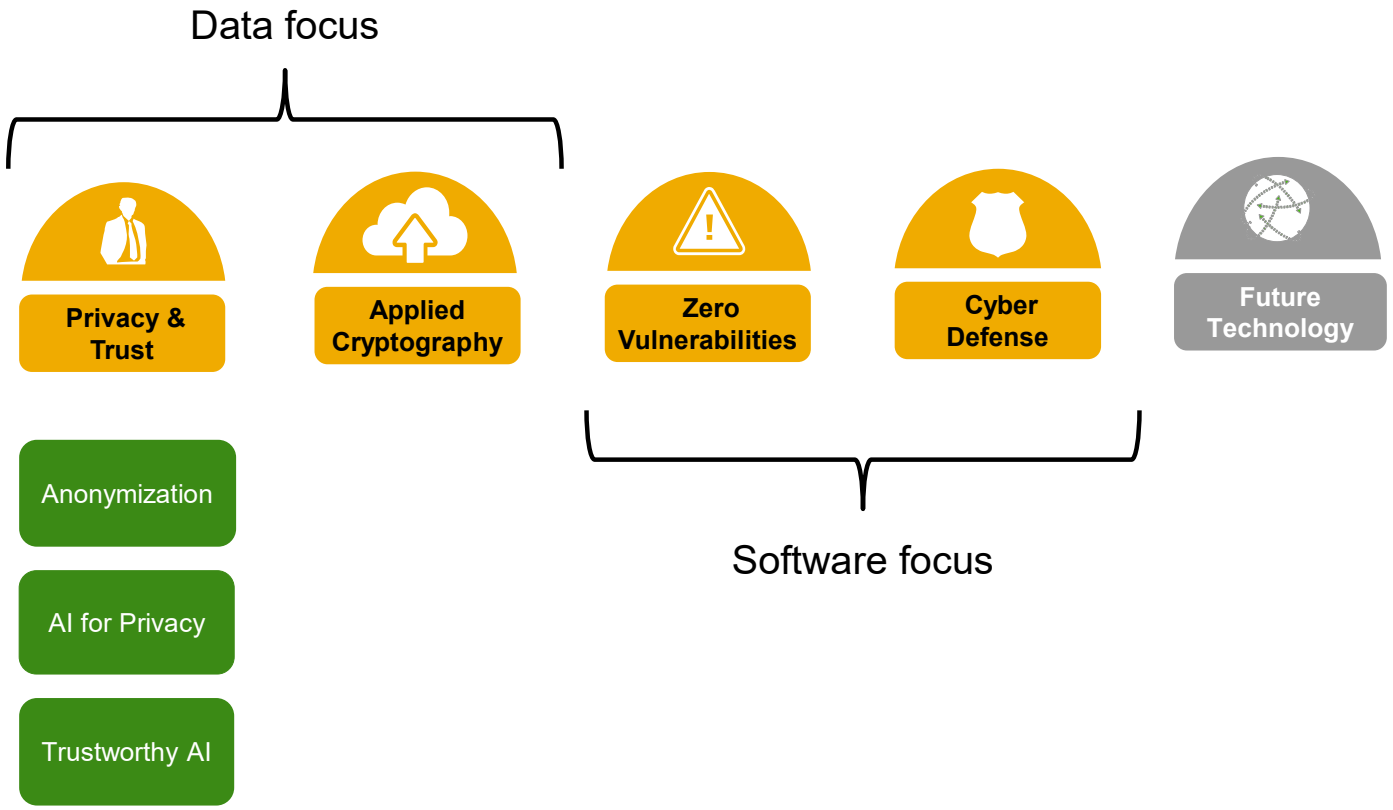
- Our research aims at a level of maturity that makes its results suitable for being transferred into and adopted by SAP's development and business units. Issues of scale and technology integration are explicitly in scope.

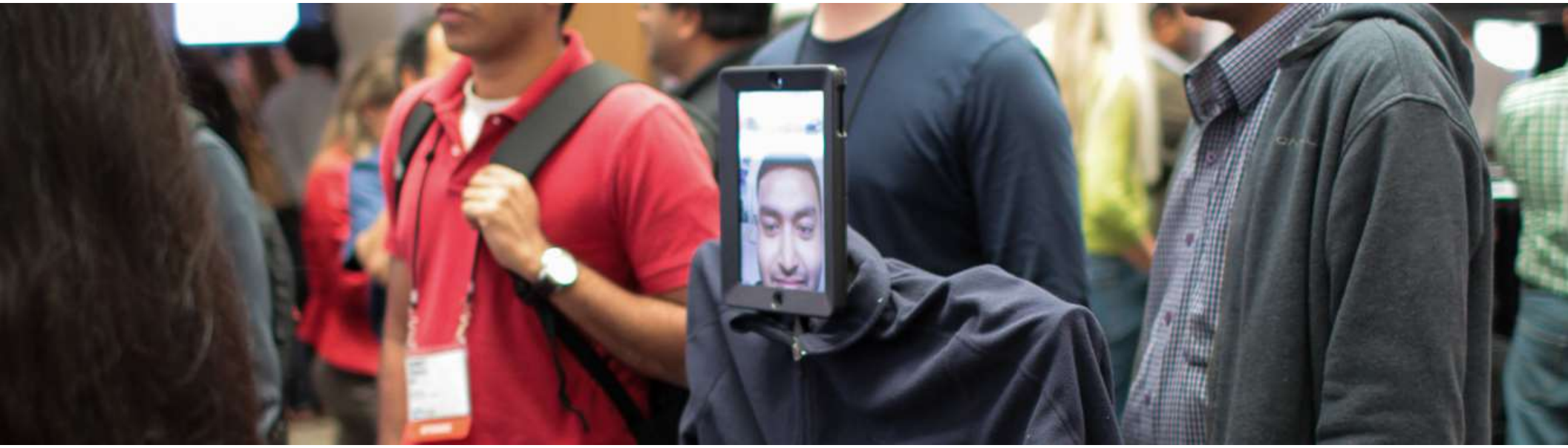


# The 5 pillars of the Security Research portfolio



# The 5 pillars of the Security Research portfolio





# Generating Realistic Synthetic Curricula Vitae for Machine Learning Applications under Differential Privacy

Francesco Di Cerbo, Francesco Aldà, Andrea Bruera

SAP Security Research

Paper:

JN SAP

## Is Data the New Oil?

**Most**

unstructured data contain
Personal Data

**GDPR**

compliance risk
over data processing

## Can we Anonymize Unstructured Text?

### Anonymization

- “making re-identification of an individual impossible”, paraphrasing GDPR

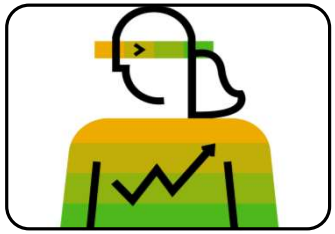
### Anonymization is **hard**!

- If we consider arbitrary text (social media, medical records, résumés/CVs...)

### Re-identification attacks

- can exploit side-knowledge coming from external sources

## ML models for HR?

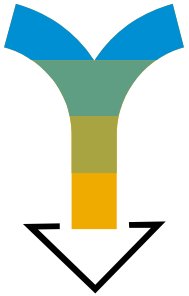


As a data scientist, I need to develop a ML model that:

- Extracts information out of a submitted CV
- Computes a score for a CV against a job posting
- …

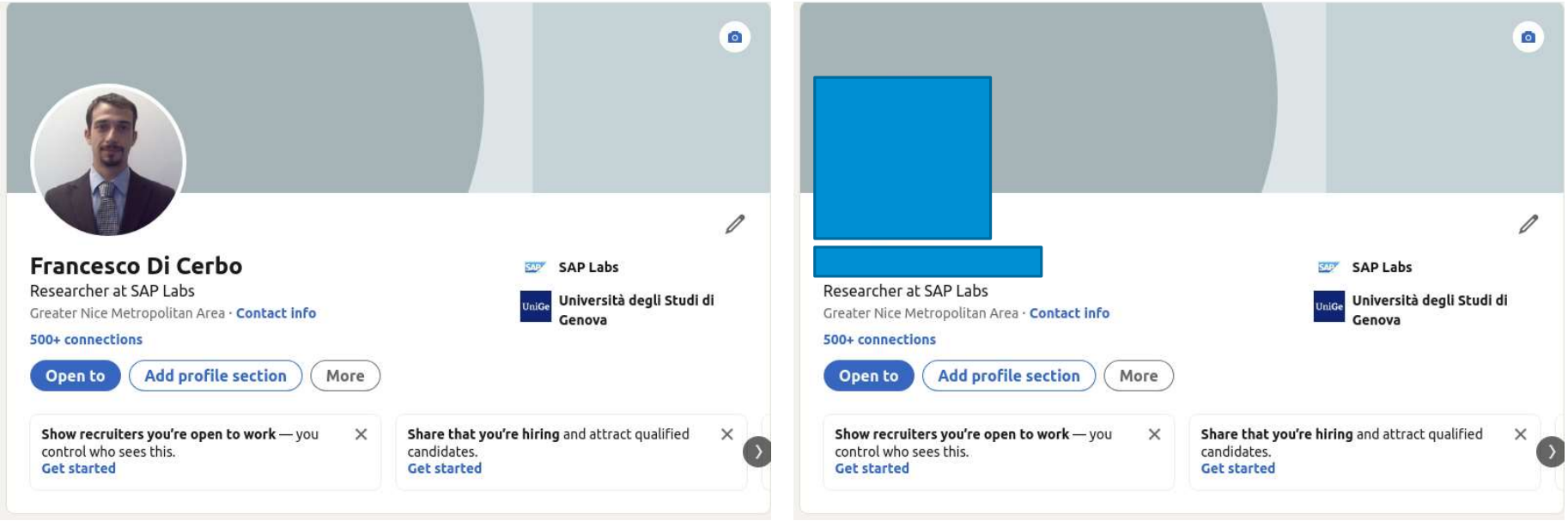
Where do I get my training dataset of CVs?

How can I be compliant with DPP requirements?



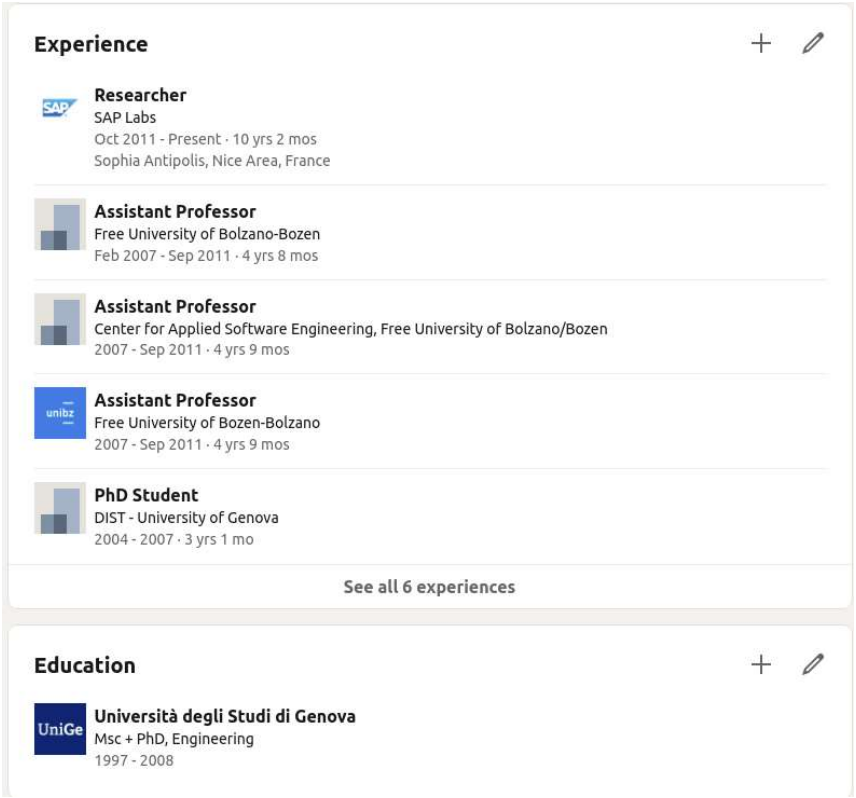
Can I anonymize CVs?

# Re-Identification Risk Assessment Example: CVs/ Résumés (1/2)




Is this sufficient?

# Re-Identification Risk Assessment Example: CVs/ Résumés (2/2)



**Experience**

**Researcher**
SAP Labs
Oct 2011 - Present · 10 yrs 2 mos
Sophia Antipolis, Nice Area, France

**Assistant Professor**
Free University of Bolzano-Bozen
Feb 2007 - Sep 2011 · 4 yrs 8 mos

**Assistant Professor**
Center for Applied Software Engineering, Free University of Bolzano/Bozen
2007 - Sep 2011 · 4 yrs 9 mos

**Assistant Professor**
Free University of Bozen-Bolzano
2007 - Sep 2011 · 4 yrs 9 mos

**PhD Student**
DIST - University of Genova
2004 - 2007 · 3 yrs 1 mo

See all 6 experiences

**Education**

**Università degli Studi di Genova**
Msc + PhD, Engineering
1997 - 2008

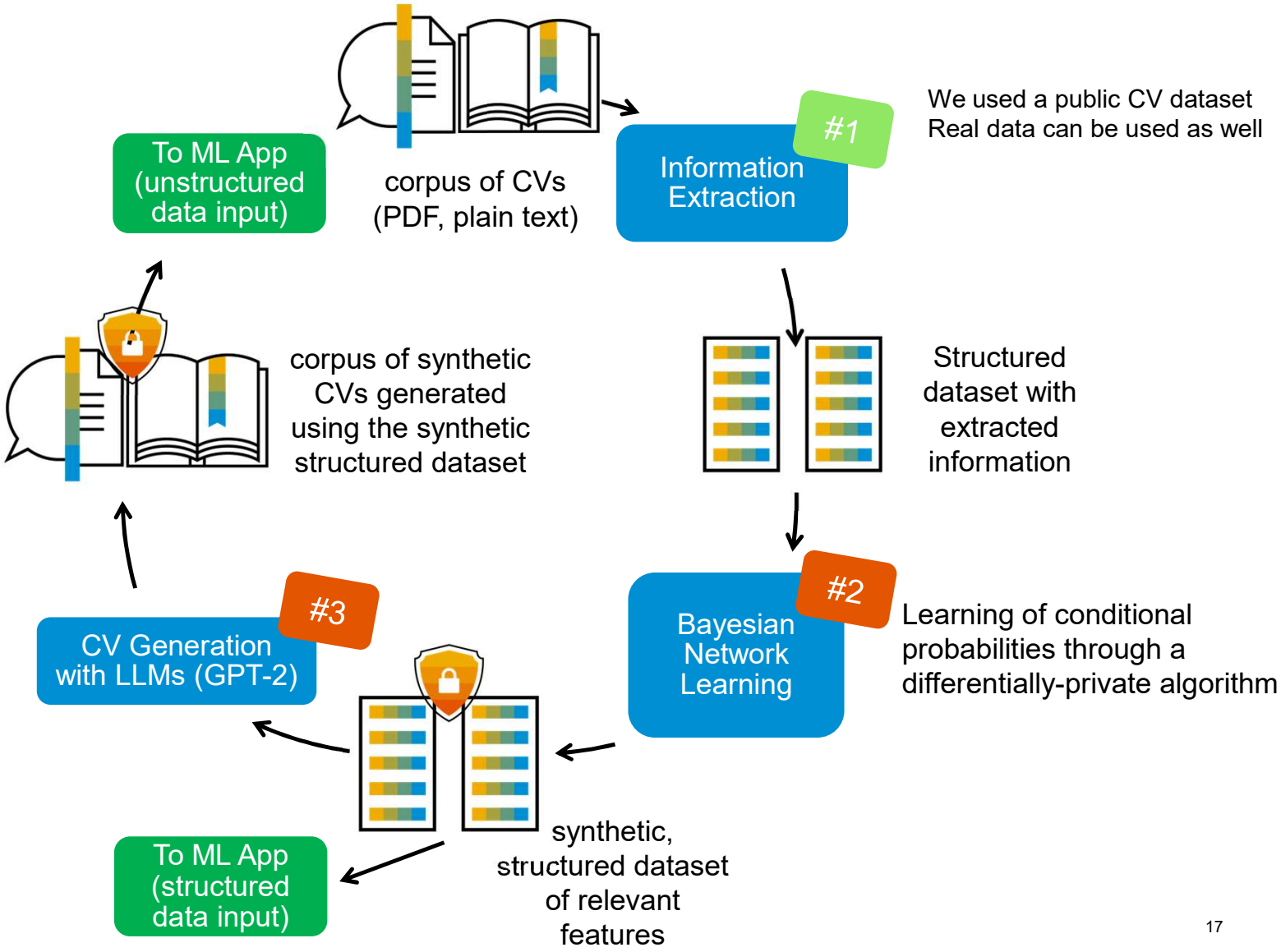
Is this sufficient?
**NO!**
Uniqueness of education & work experience

We need another idea!

# “Realistic Synthetic Data”

From a dataset we learn its statistical properties under differential privacy and we use them to generate synthetic data

# Unstructured Data Anonymization & Synthetic CV Generation



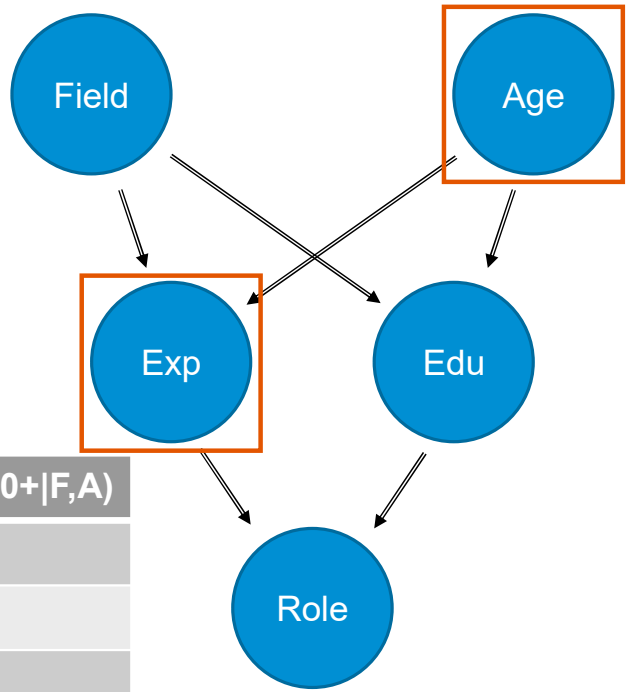
# Step #2: Bayesian Network Learning

# Bayesian Network for CV attributes: Toy Example

Original Structured Dataset from Step #1

| | age | field | edu | exp | role |
|---|---|---|---|---|---|
| 0 | 20-30 | IT | B.Sc. | <1 year | Intern |
| 1 | 20-30 | Science | Ph.D. | <1 year | Data Scientist |
| 2 | 20-30 | IT | B.Sc. | 3-6 years | Developer |
| 3 | 20-30 | IT | M.Sc. | 1-3 years | Junior Developer |
| 4 | 20-30 | Economics | B.Sc. | 1-3 years | Junior Sales Accountant |
| 5 | 30-40 | Economics | M.Sc. | 3-6 years | Sales Accountant |
| 6 | 30-40 | Science | Ph.D. | 3-6 years | Senior Data Scientist |
| 7 | 30-40 | Science | M.Sc. | 6-10 years | Expert Data Scientist |
| 8 | 30-40 | IT | M.Sc. | 10+ years | Expert Developer |
| 9 | 40-50 | Economics | M.Sc. | 10+ years | Sales Manager |
| 10 | 40-50 | Economics | B.Sc. | 10+ years | Expert Sales Accountant |
| 11 | 40-50 | Science | M.Sc. | 10+ years | Data Science Manager |

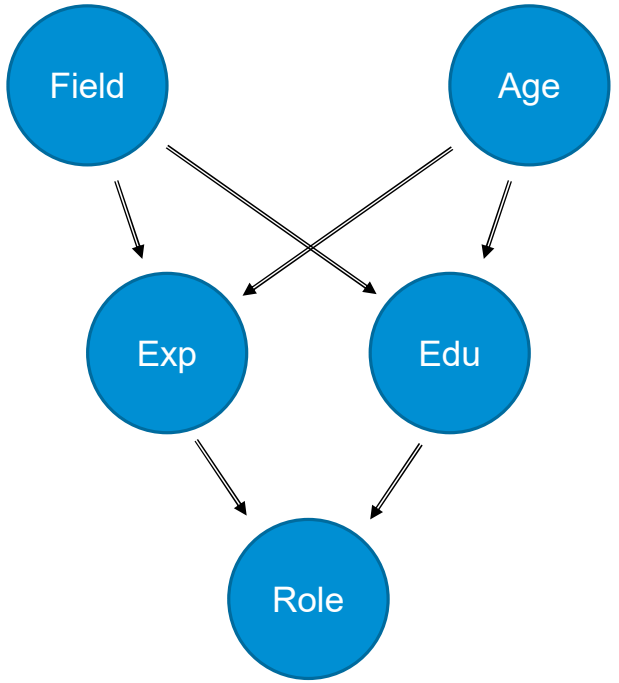
| P(A is 20-30) | P(A is 30-40) | P(A is 40-50) |
|---|---|---|
| 5/12 | 1/3 | 1/4 |




| F | A | P(Exp is <1\|F,A) | ... | P(Exp is 10+\|F,A) |
|---|---|---|---|---|
| IT | 20-30 | 1/3 | ... | 0 |
| IT | 30-40 | 0 | ... | 1 |
| ... | ... | ... | ... | ... |
| Economics | 40-50 | 0 | ... | 1 |

# Bayesian Network for CV attributes: Toy Example

- The structure is defined (no privacy loss)
- Learn conditional probabilities (CPDs) from data
- Provide **differential privacy** by perturbing CPDs to prevent information leakage
- Sample new values



# Bayesian Network for CV attributes: Toy Example

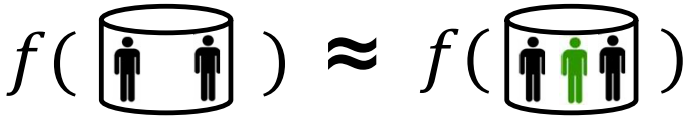
- The structure is defined (no privacy loss)
- Learn conditional probabilities (CPDs) from data
- Provide **differential privacy** by perturbing CPDs to prevent information leakage
- Sample new values



| | age | field | edu | exp | role |
|---|---|---|---|---|---|
| 0 | 20-30 | IT | M.Sc. | 1-3 years | Junior Developer |
| 1 | 20-30 | Economics | B.Sc. | 1-3 years | Junior Sales Accountant |
| 2 | 30-40 | Economics | B.Sc. | 10+ years | Developer |
| 3 | 30-40 | IT | M.Sc. | 10+ years | Expert Developer |
| 4 | 40-50 | Science | B.Sc. | 1-3 years | Data Science Manager |

# Differential Privacy

- De facto standard for privacy-preserving statistical data analysis: why?
  - Strong mathematical guarantees
  - Independent of adversary side knowledge and computational power
  - Composability
  - Robustness to post-processing
  - …
- The output distribution of a randomized algorithm should not be affected too much by small changes in its input
- Can be achieved by adding calibrated Laplace noise



# Step #3: CV Generation w/ GPT-2, GPT-3, GPT-4 (?)

# Language models (GPT-X)

The cat

*from: https://towardsdatascience.com/sentence-generation-with-n-gram-21a5eef36a1b*

↓

**Natural Language Generation**

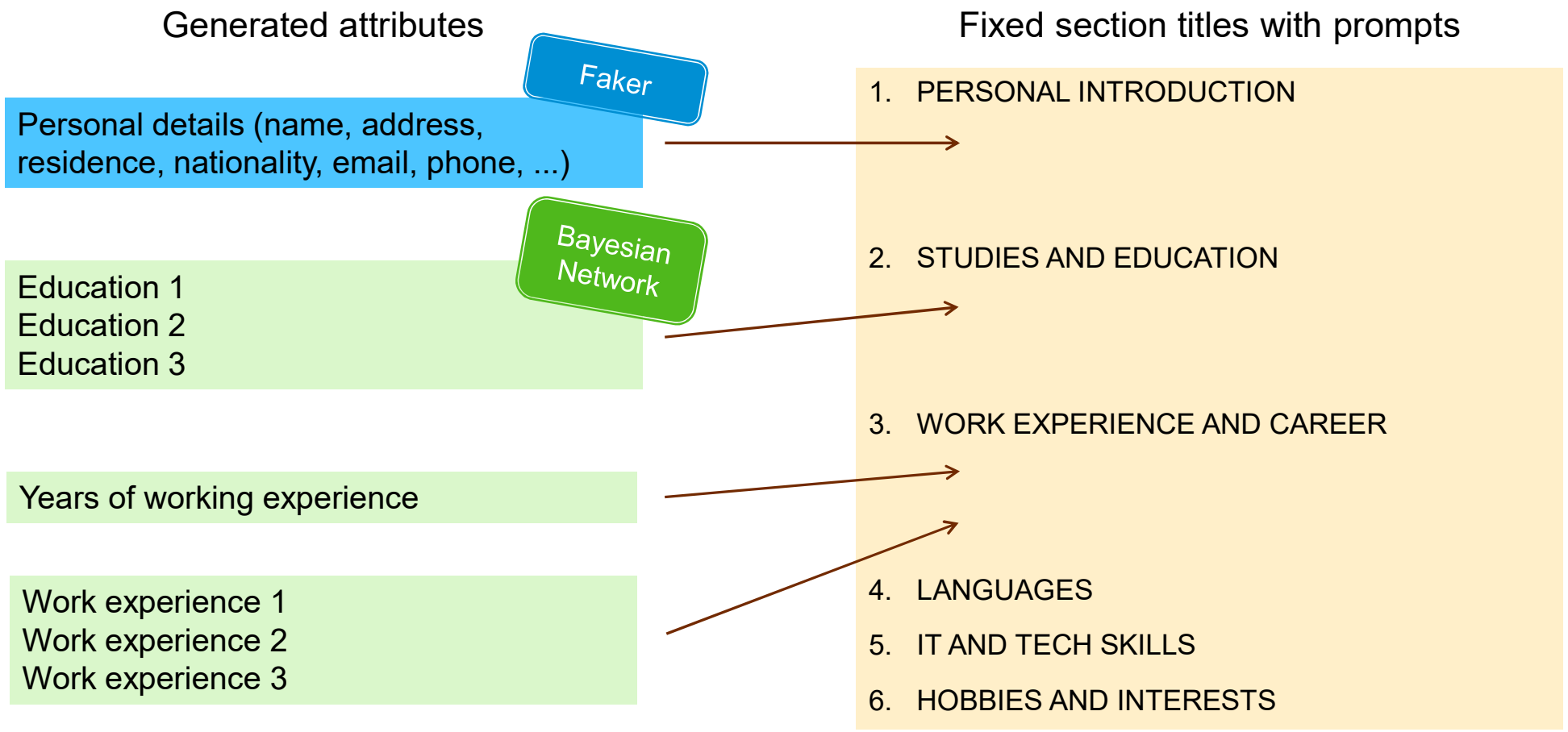
Given a type of text (e.g., stories, summaries, fake news),
generate something which resembles what humans would write

↓

**Generation of synthetic text**

A great advantage in terms of privacy and good as a data augmentation technique

# Generating CV text

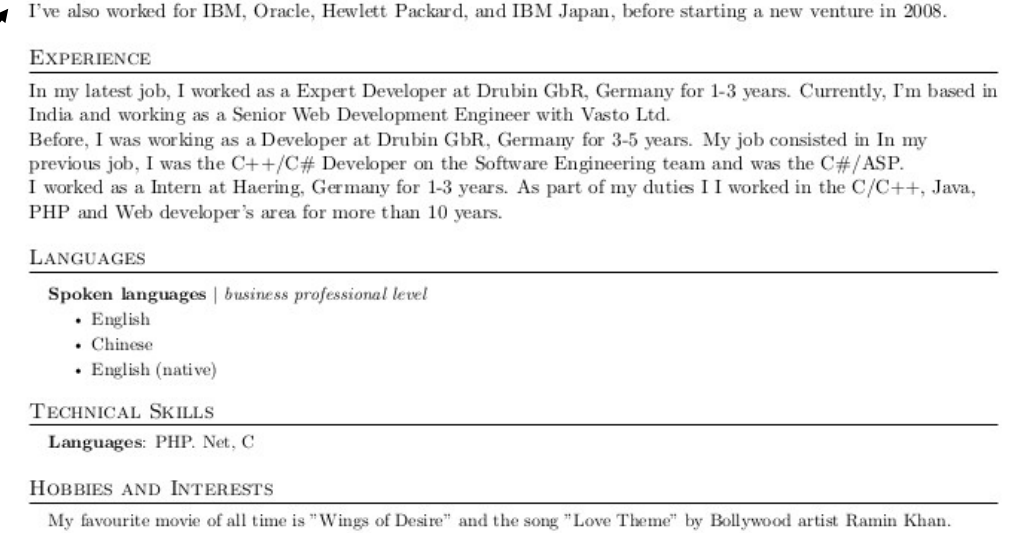


# Example of a Generated CV

CURRICULUM VITAE:
Name: Ivana Losekann
Nationality: DE
Address:Pasquale-Dörschner-Ring 97
70896 Apolda
Country of residence: DE
Phone number: +49(0)0251 55486
Email: ilosekann@gehringer.de
Field: IT - current position: Expert Developer

PERSONAL INTRODUCTION:
Ivana Losekann is an IT specialist who specializes in IT application development.
STUDIES AND EDUCATION:
My most recent title is a M.Sc. in Business Analytics from Humboldt University of Berlin, Germany. The final dissertation subject was " The Value Chain for E-Commerce Systems".
Before that, I got my B.Sc. in Computer Science at Indian Institute of Technology Madras, India, working on "The Business Model of Small, Medium and Large Enterprise Systems: From the Perspective of Software Engineering".
I've also worked for IBM, Oracle, Hewlett Packard, and IBM Japan, before starting a new venture in 2008.

PUBLIC



# Ivana Losekann

+49(0)0251 55486 | ilosekann@gehringer.de | linkedin.com/in/x | github.com/x

## Personal Information

**Nationality: DE** Address:Pasquale-Dörschner-Ring 97, 70896 Apolda
*Country of residence: DE* *Expected December 2022*

Ivana Losekann is an IT specialist who specializes in IT application development.
Field: IT - current position: Expert Developer

## Education

**Humboldt University** Berlin, Germany
*M.Sc. in Business Analytics*
- The final dissertation subject was " The Value Chain for E-Commerce Systems".

**Indian Institute of Technology** Madras, India
*B.Sc. in Computer Science*
- the final dissertation subjet was " The Business Model of Small, Medium and Large Enterprise Systems: From the Perspective of Software Engineering"

I've also worked for IBM, Oracle, Hewlett Packard, and IBM Japan, before starting a new venture in 2008.

## Experience

In my latest job, I worked as a Expert Developer at Drubin GbR, Germany for 1-3 years. Currently, I'm based in India and working as a Senior Web Development Engineer with Vasto Ltd.
Before, I was working as a Developer at Drubin GbR, Germany for 3-5 years. My job consisted in In my previous job, I was the C++/C# Developer on the Software Engineering team and was the C#/ASP.
I worked as a Intern at Haering, Germany for 1-3 years. As part of my duties I I worked in the C/C++, Java, PHP and Web developer's area for more than 10 years.

## Languages

**Spoken languages** | *business professional level*
- English
- Chinese
- English (native)

## Technical Skills

**Languages**: PHP. Net, C

## Hobbies and Interests

My favourite movie of all time is "Wings of Desire" and the song "Love Theme" by Bollywood artist Ramin Khan.

# Additional Synthetic CVs with Different Layouts

Cengiz Schleich

Nationality: DE
Address:Jacobi Jäckelstr. 511
71498 München
Country of residence: DE
Phone number: +49(0) 550159228
Field: Science - current position: Senior Data Scientist

cengiz.schleich@scheuermann.de
https://your-web-page.io

I am Cengiz Schleich, a 30 year old Data Scientist living and working in Berlin, Germany. I'm currently a PhD Candidate at the Technical University of Berlin studying Artificial Intelligence.

## Education

2021. **M.Sc. in Business**

## Work Experience

In my latest job, I worked as a Senior Data Scientist at Rosemann, Germany for 1-3 years. I migrated to Hamburg, Germany after this, and worked for a couple of months as a Senior Data Scientist for a private company as an inter

Before, I was working as a Data Scientist a OHG mbH, Germany for 1-3 years. My jol previous job, I worked as a Senior Data Sc USA for 6 months. AUGUST 2013-Januar Professor at the Ruhr-.

## Tools

Microsoft Office ●●●○
C++ ●●●○

## Languag

English
French

Last updated: December, 2021

# CÁNDIDO NEVADO

**Senior Data Scientist**

cnevado@verdu.es | +913071234386 | 38/22Chawla Circle, Silchar-203887 | India
cnevado.tumblr.com | @CCandidoNevado | ccandidonevado

## EXPERIENCE

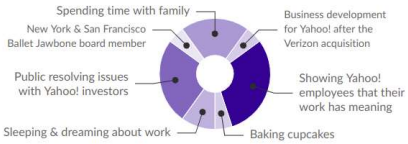
Senior Data Scientist
**Rao, Manne and Sharma, India**
for <1 years | India

- Job description: Rao Manne, Data Science Director - Rajat Sharma - Rajat is a world leader in Machine Learning (ML), Knowledge Based Design (KBD), and Machine Learning and Intelligent Knowledge Management (MKI) –

Data Scientist
**Rao, Manne and Sharma**
1-3 years | India

- I am a Data Scientist at Rao Manne - Data Science Director for <1-3 years.

## A DAY OF MY LIFE




## HOBBIES AND INTERESTS

**My passions**
In my free time I like - sports, music, movies, video games, watching movies, listening to music and playing games (e. g. , NBA 2K, Hearthstone, etc. ). In my free time I like watching my favorite movies and TV series like Breaking Bad, Game of Thrones, South Park, The Office.

## LANGUAGES

English
Chinese
French
Japanese

## EDUCATION

Ph.D. in Physics
**Amrita University, India**
1998

(Physics), 2001 (Materials Science and Engineering) Bachelor of Science in Chemical Engineering from Tata Institute of Fundamental Research (TIFR) - Mumbai, India: 1993.

M.Sc. in Physics
**Karlsruhe Institute of Technology, Germany**

Bachelor of Science in Chemical Engineering from Indian Institute of Technology - Kharagpur, India: 1993.

Bachelor of Science in Nuclear Science and Engineering
**India**
1994

B.Sc. in Physics
**Karlsruhe Institute of Technology, Germany**

Master of Science in Nuclear Engineering (TIFR) - India: 1995

## IT AND TECH SKILLS

Computer Science – MS

# TODD B. MORRIS

Software Developer

NATIONALITY: US | ADDRESS: 19, SABHARWAL, GANDHINAGAR-301282, INDIA

in /barry-bareel
/barry-bareel
tmorris@jackson.biz
+911605517595

NAL INTRODUCTION

urrently pursuing a B. Tech in Computer Science and an M. Tech in Computer Science. My research interests are a Structures, Memory Modeling, Algorithms, Networking, and Parallel Computing. I have also been involved in a er of Computer Science/Data Science research

IENCE

ears. **Software Developer** — Lata Group, India

- Job description: I am interested in the topic of Parallel Computing and Memory Modeling for the Linux Kernel.

years **Intern** — Lata Group, India
Software Development

- I am interested in the topic of Data Structures and Modeling of Memory and Parallel Computing for Linux Kernel.

UAGES

English — C1 Proficient User
Chinese — A1 Basic User

IT AND TECH SKILLS

English, Chinese

EDUCATION & TRAININGS

**M.Sc. in Computer Science**
2007 from Indian Institute of Technology Roorkee, India
**M.Sc. in Computer Science**
2005 from INET-University of Pune - Pune, India
**B.Sc. in Computer Science**
2003 from University of Southern Mississippi, United States

# Evaluation

Comparing a 28k corpus of CVs with 60k generated CVs

Intrinsic Evaluation

- Comparison of a set of NLP metrics

Extrinsic Evaluation

- Performances of real and synthetic corpora in a common downstream task (classification of candidate role)
- Using FastText and BERT-based classifiers
- Comparison of raw and masked text (removal of explicit candidate role mentions)

More info? Check the paper at:

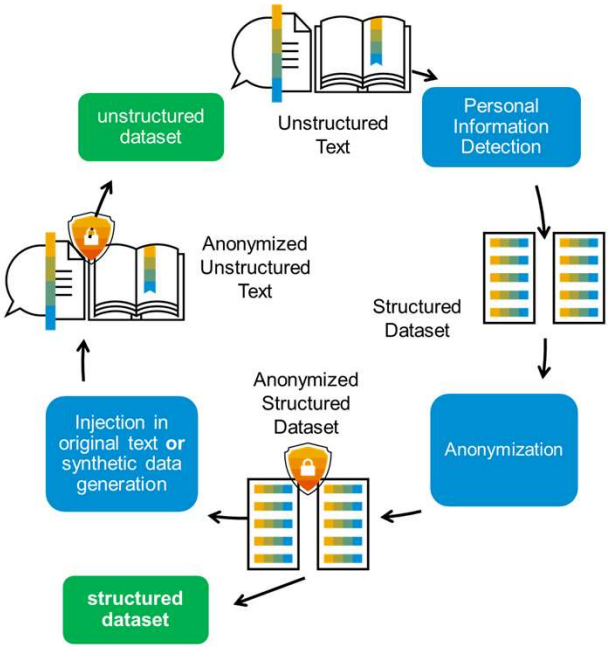
## Conclusions & Call to Action

**Data generation with privacy guarantees works!**

- Synthetic CVs as training datasets or data augmentation

**It can be applied to other domains**

- Temporal sequences (Synthetic clickstream flow in CX)
- Bias evaluation
- …



# Thank you.

Contact information:

Francesco di Cerbo
francesco.di.cerbo@sap.com

Volkmar Lotz
volkmar.lotz@sap.com

THE BEST RUN SAP