“The unauthorized attempts to access information […] were made on approximately 200,000 taxpayer accounts […]. The attempts were made using taxpayers’ personal information already obtained from ***sources outside the IRS****.*

[…]

Of the approximately 100,000 successful attempts …, only 13,000 possibly fraudulent returns were filed for tax year 2014, for which the ***IRS issued refunds totaling $39 million****. We are still determining how many of these returns were filed by actual taxpayers and which were filed using stolen identities.”*

Statement of IRS commissioner at a US senate hearing
June 2015

DARKNET
DIARIES

Synthetic Norway: How We Use Machine Learning to Generate a Synthetic Population

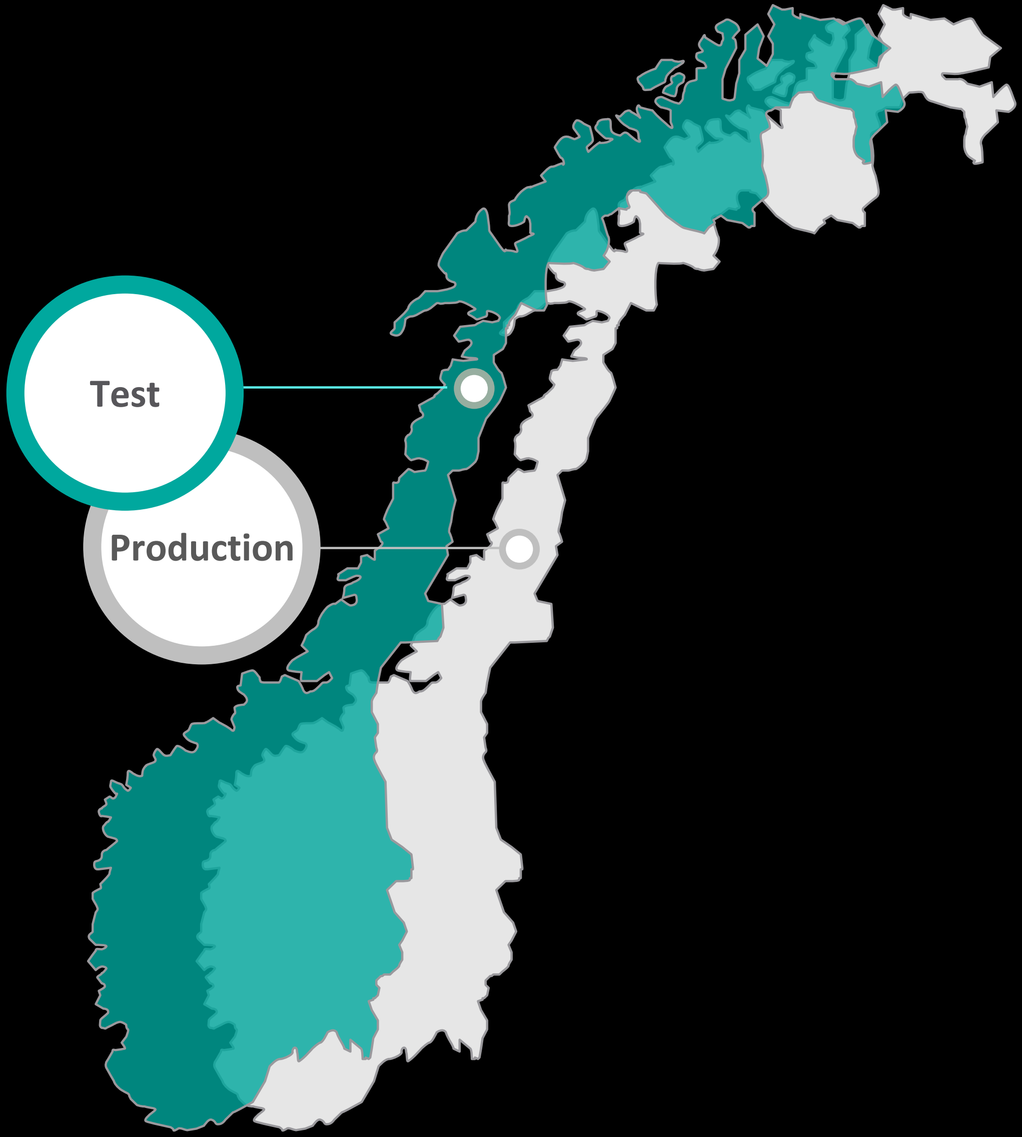
Razieh Behjati

Senior Consultant/CTO

Testify AS

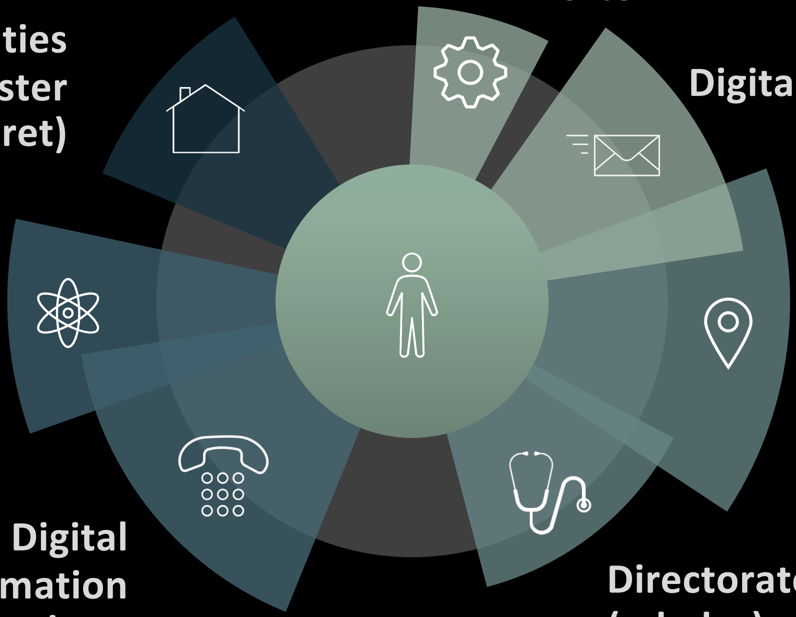
Testify smartere testing

The Norwegian Tax Administration


Test
Production

Statistically Representative

Alive = Dynamic



ID-Porten

Legal Entities
Register
(Enhetsregisteret)

Digital mailbox

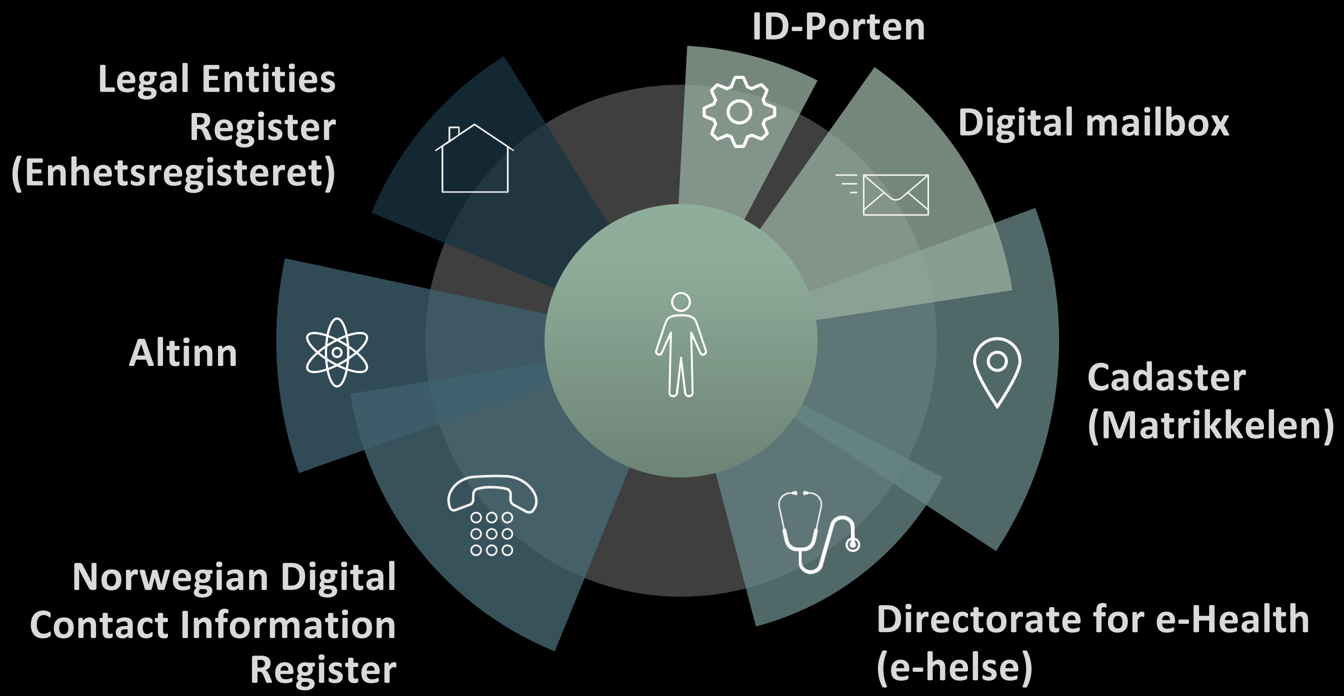
Altinn

Cadaster
(Matrikkelen)

Norwegian Digital
Contact Information
Register

Directorate for e-Health
(e-helse)

MF project



Legal Entities Register (Enhetsregisteret)

ID-Porten

Digital mailbox

Altinn

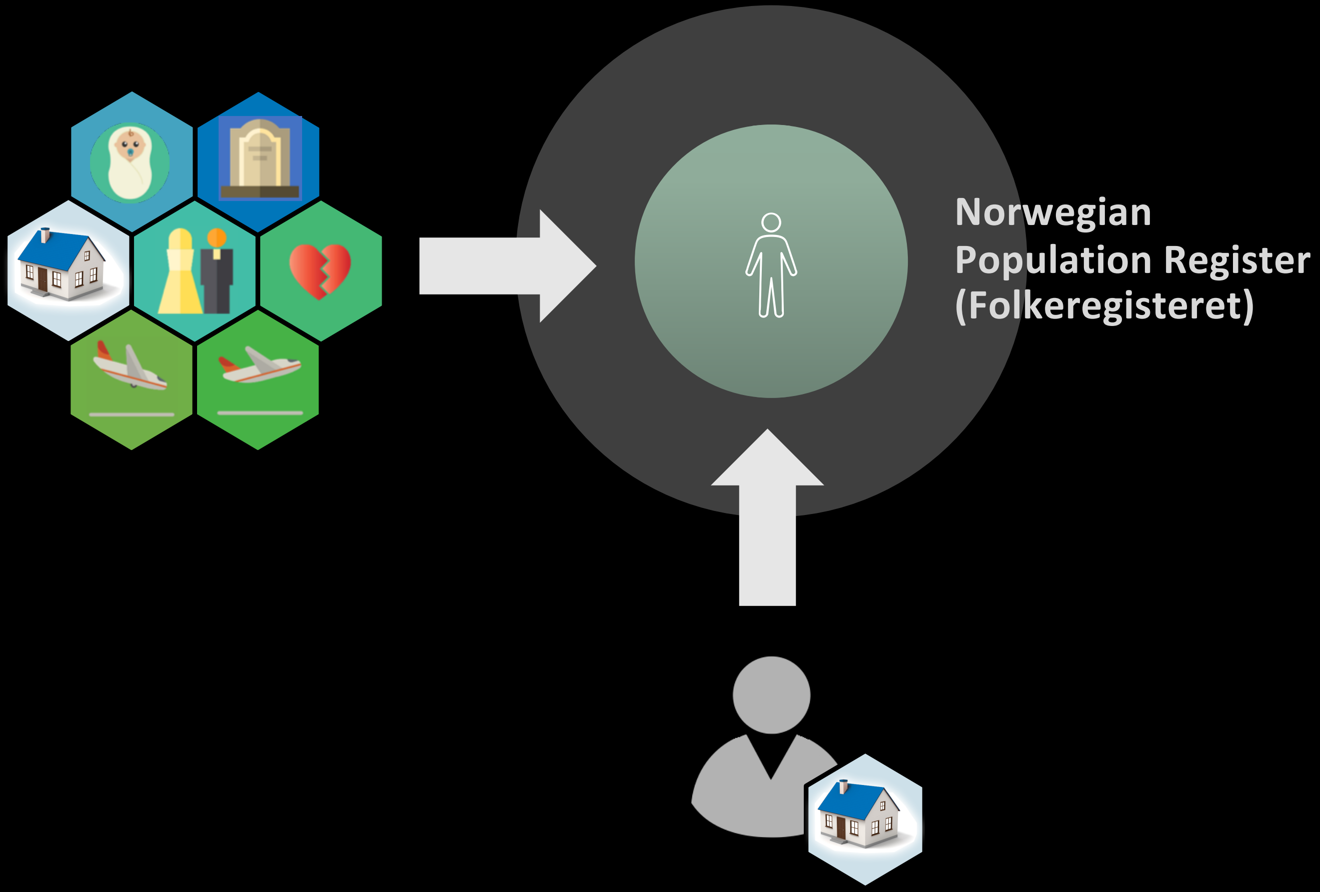
Cadaster (Matrikkelen)

Norwegian Digital Contact Information Register

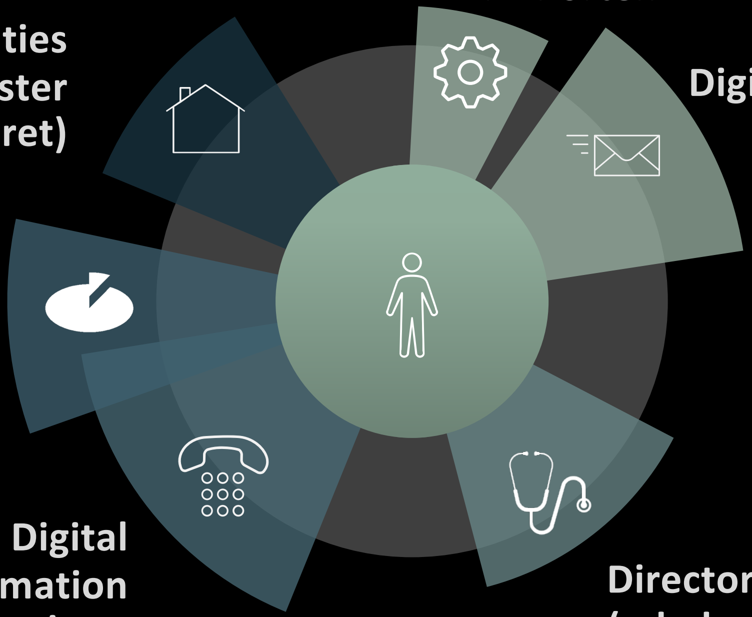
Directorate for e-Health (e-helse)

Production-like test environment

Data?



Norwegian
Population Register
(Folkeregisteret)



Legal Entities Register (Enhetsregisteret)

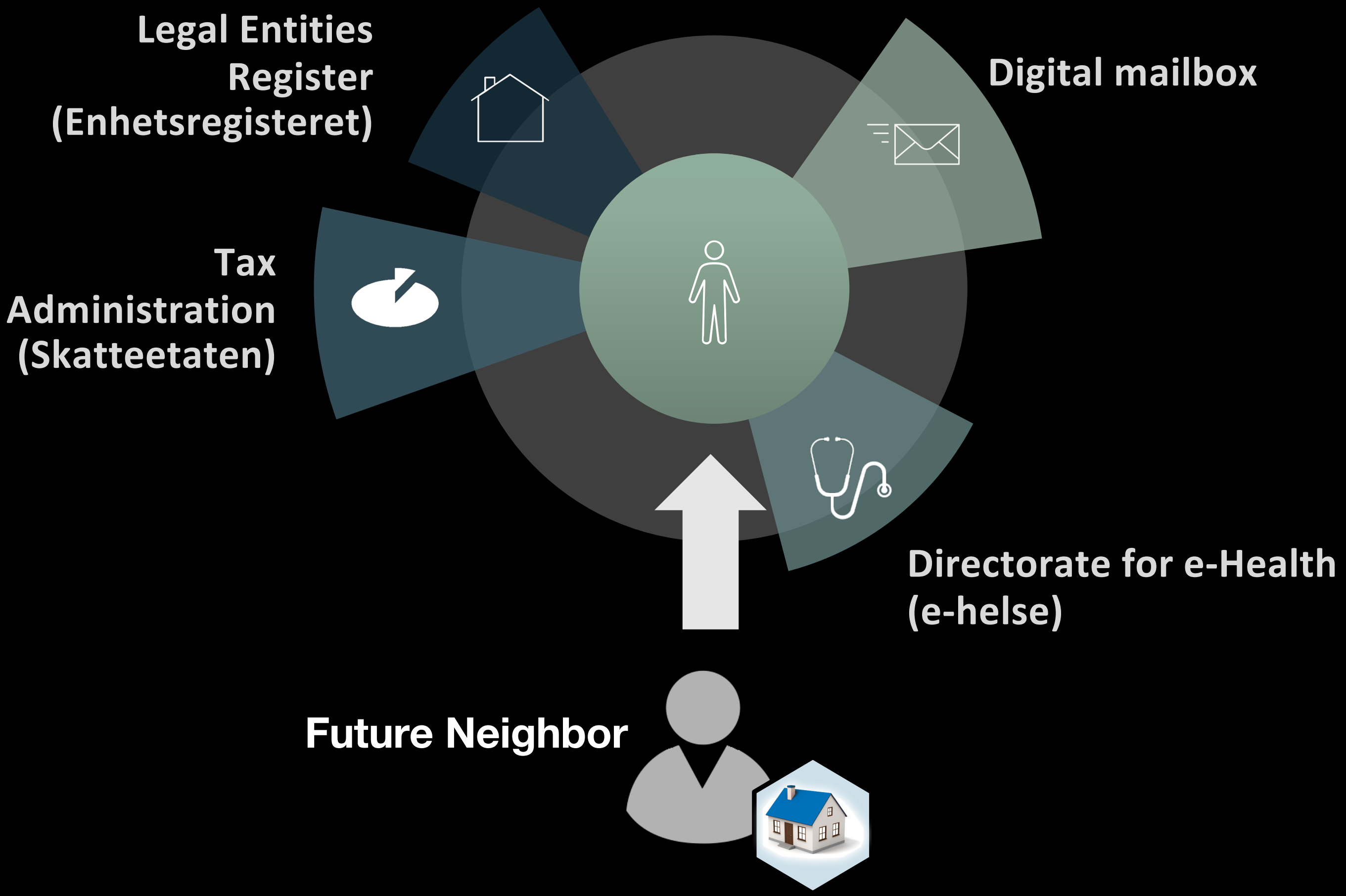
ID-Porten

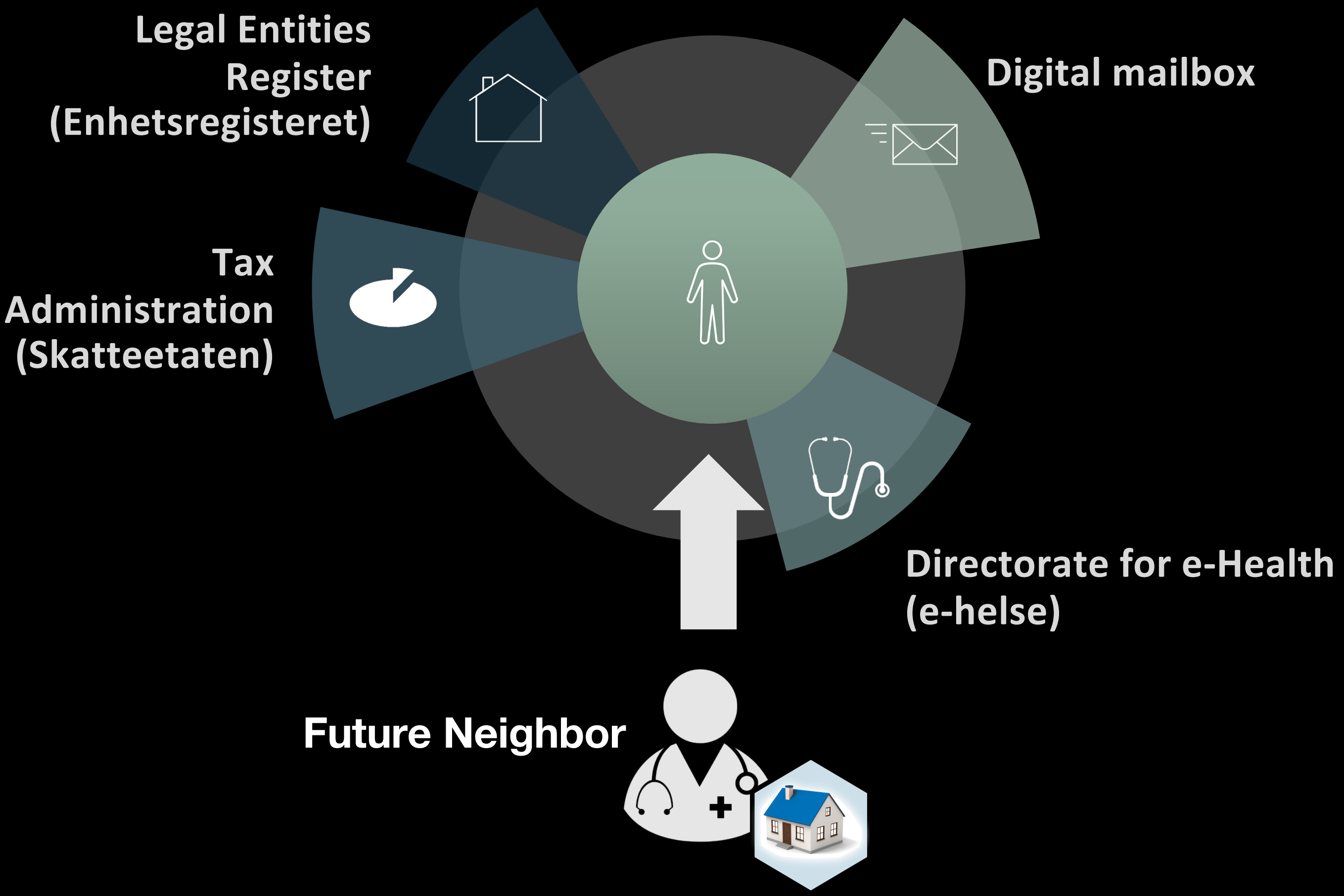
Digital mailbox

Directorate for e-Health (e-helse)

Norwegian Digital Contact Information Register

Tax Administration (Skatteetaten)


Legal Entities
Register
(Enhetsregisteret)
Digital mailbox
Tax
Administration
(Skatteetaten)
Directorate for e-Health
(e-helse)
Future Neighbor



Legal Entities Register (Enhetsregisteret)

Digital mailbox

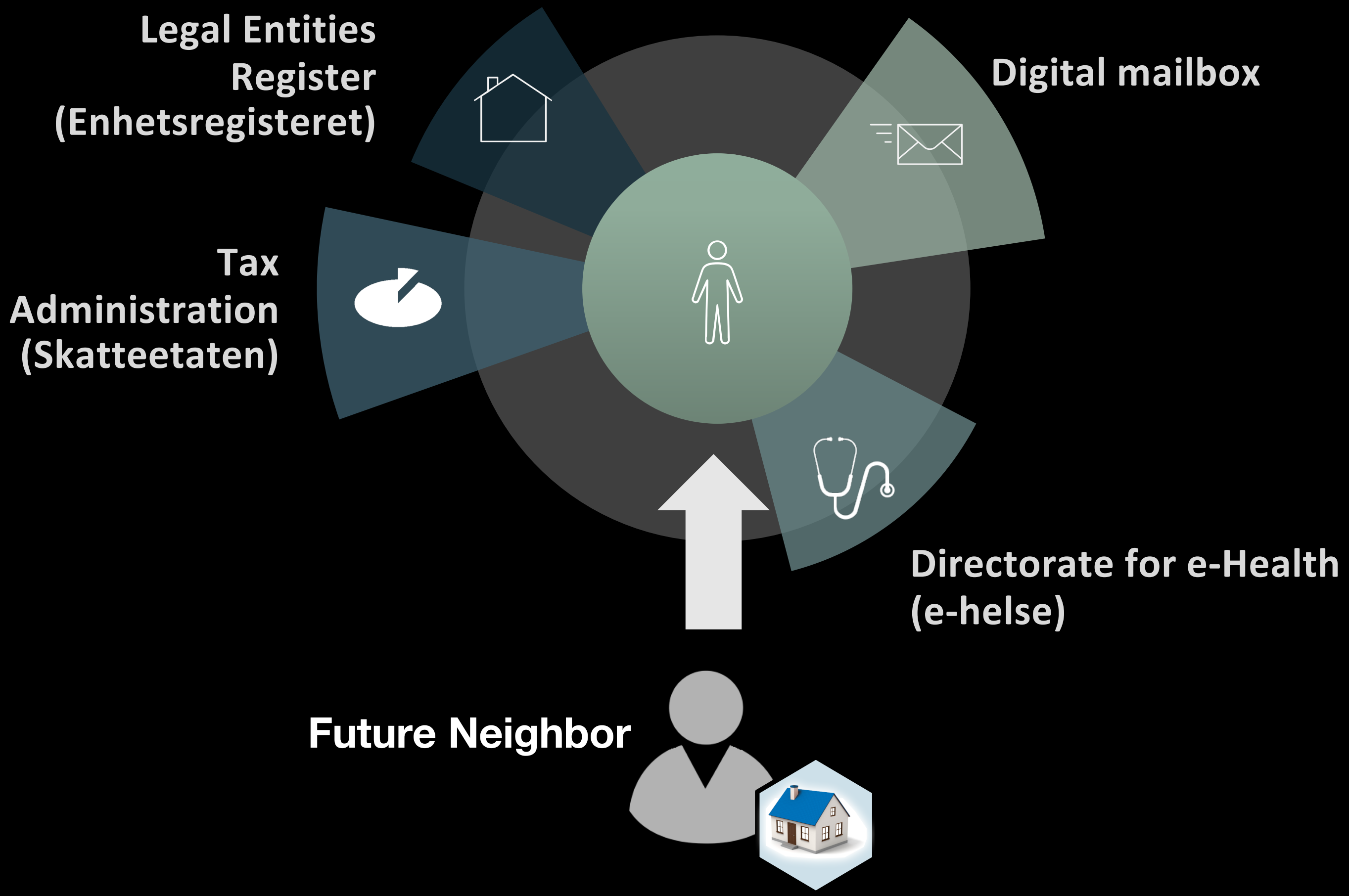
Tax Administration (Skatteetaten)

Directorate for e-Health (e-helse)

Future Neighbor

Why dynamic?

No life event = No test



Legal Entities Register (Enhetsregisteret)

Digital mailbox

Tax Administration (Skatteetaten)

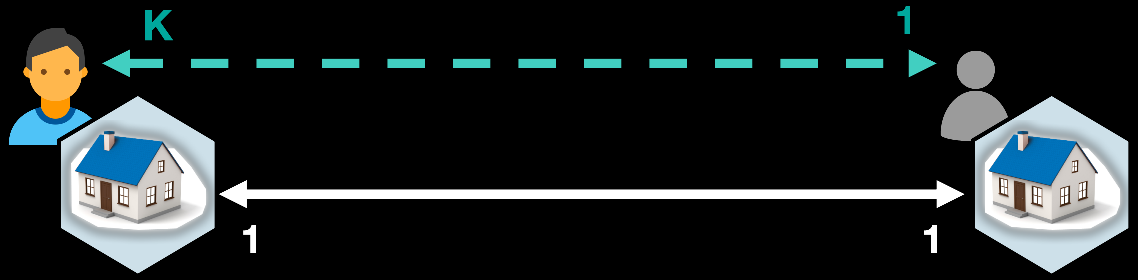
Directorate for e-Health (e-helse)

Future Neighbor

Why is representativeness important?

Variance in data

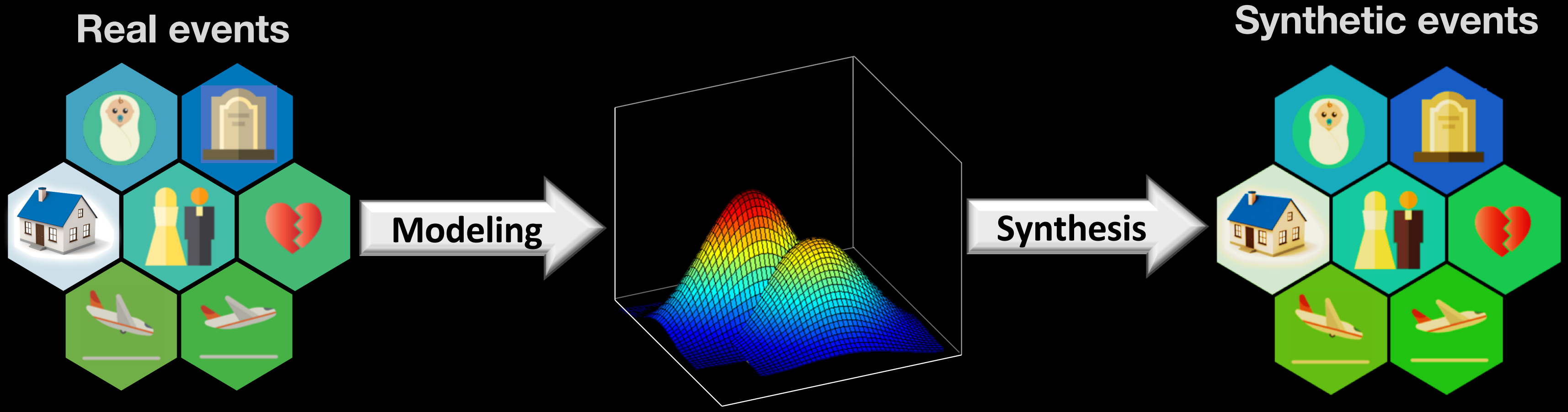
Why not use anonymized data?



Dynamic

Statistically Representative

How?



Real events

Modeling

Synthesis

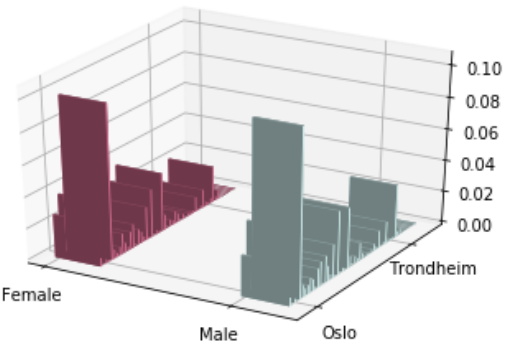
Synthetic events

Statistical representation of life events

Birth

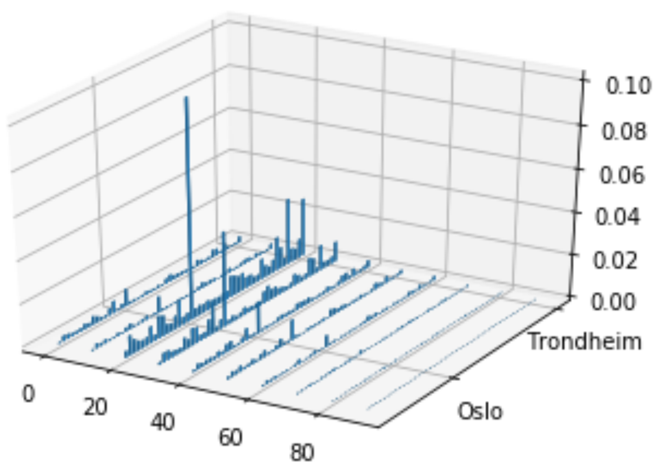
$P(Male) \sim P(Female) \sim 0.5$

$P(Birth\ Place|Female)$



$P_{birth}(Gender \cap Birth\ Place)$

Relocation



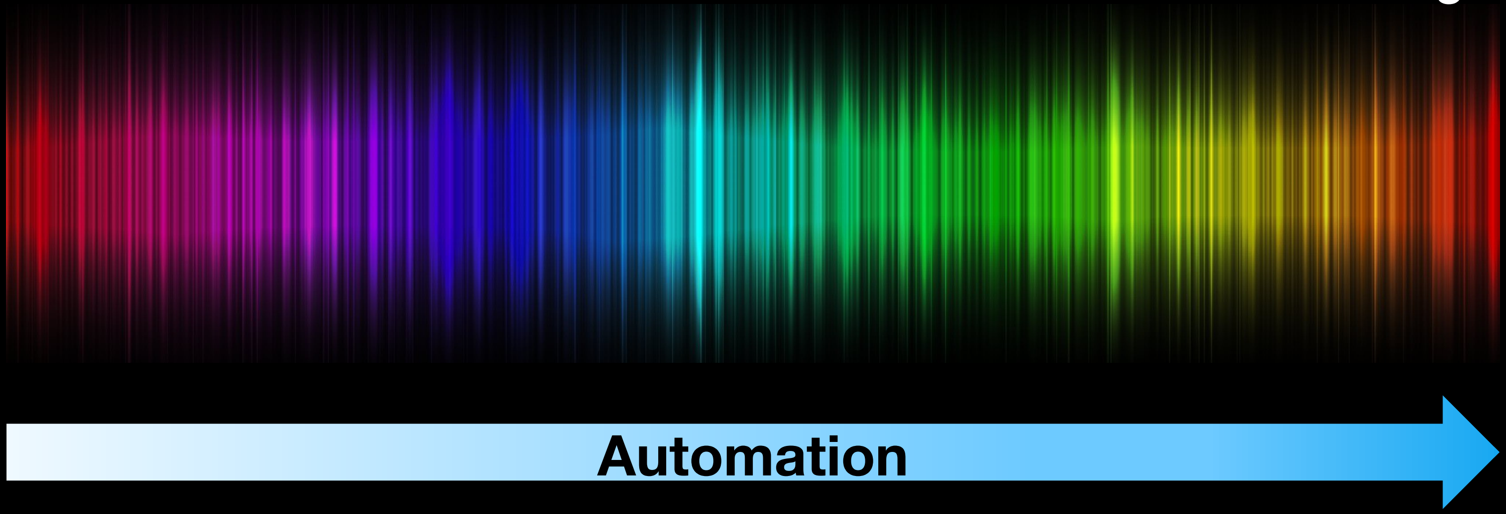
$P_{relocation}(Age)$

$P_{relocation}(Age \cap New\ City)$

A spectrum of algorithms

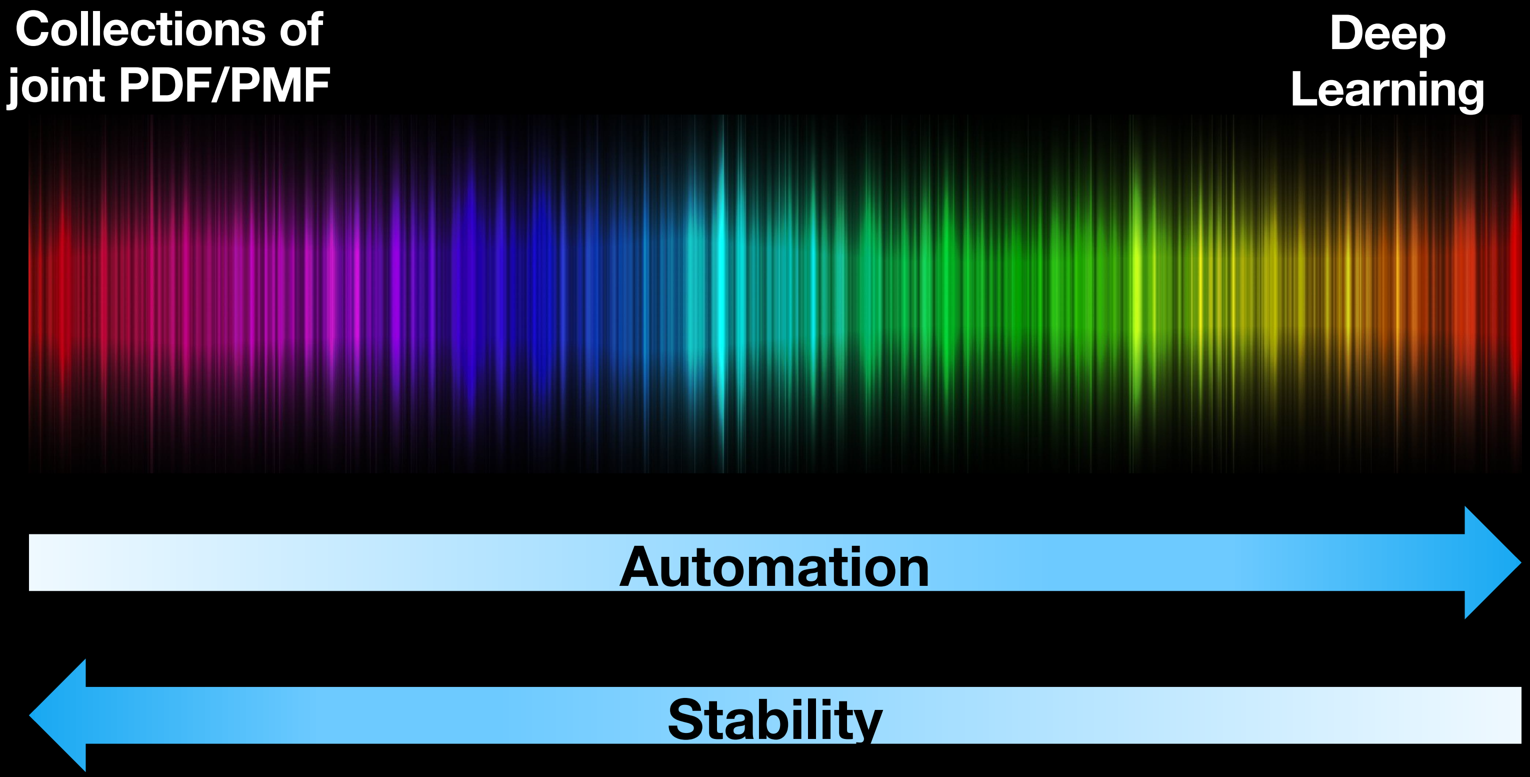
Collections of joint PDF/PMF

Deep Learning



Automation

A spectrum of algorithms



Collections of joint PDF/PMF

Deep Learning

Automation

Stability

Use of Deep Learning

How is deep learning relevant?

Text synthesis

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016

Text synthesis

The quick brown fox …

Word	P(word | sentence, context)
a	0.0001
book	0.0001
is	0.01
jumps	0.8

Text synthesis

The quick brown fox …

Word	P(word | sentence, ~~context~~) Text to translate
a	0.0001
book	0.0001
is	0.01
jumps	0.8

Life event = Sentence

Female born in Oslo on 11.09.2019 …

Female born in Oslo on 11.09.2019 …

0111092019F0301…

Character level language modeling

Character = Token in vocabulary

P(char | char-sequence)

0111092019F0301…

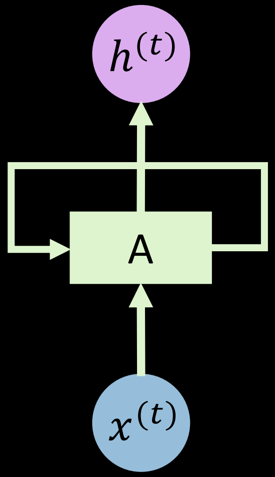
Character level language modeling

Character = Token in vocabulary

P(char | char-sequence)

0111092019F0301…

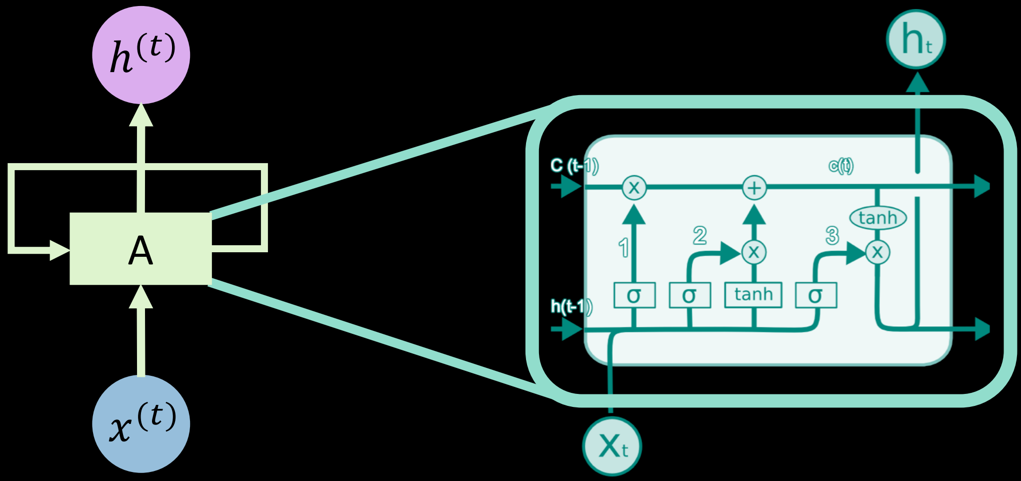
Long Short-Term Memory (LSTM)



https://www.deeplearningbook.org/, Chapter 10

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

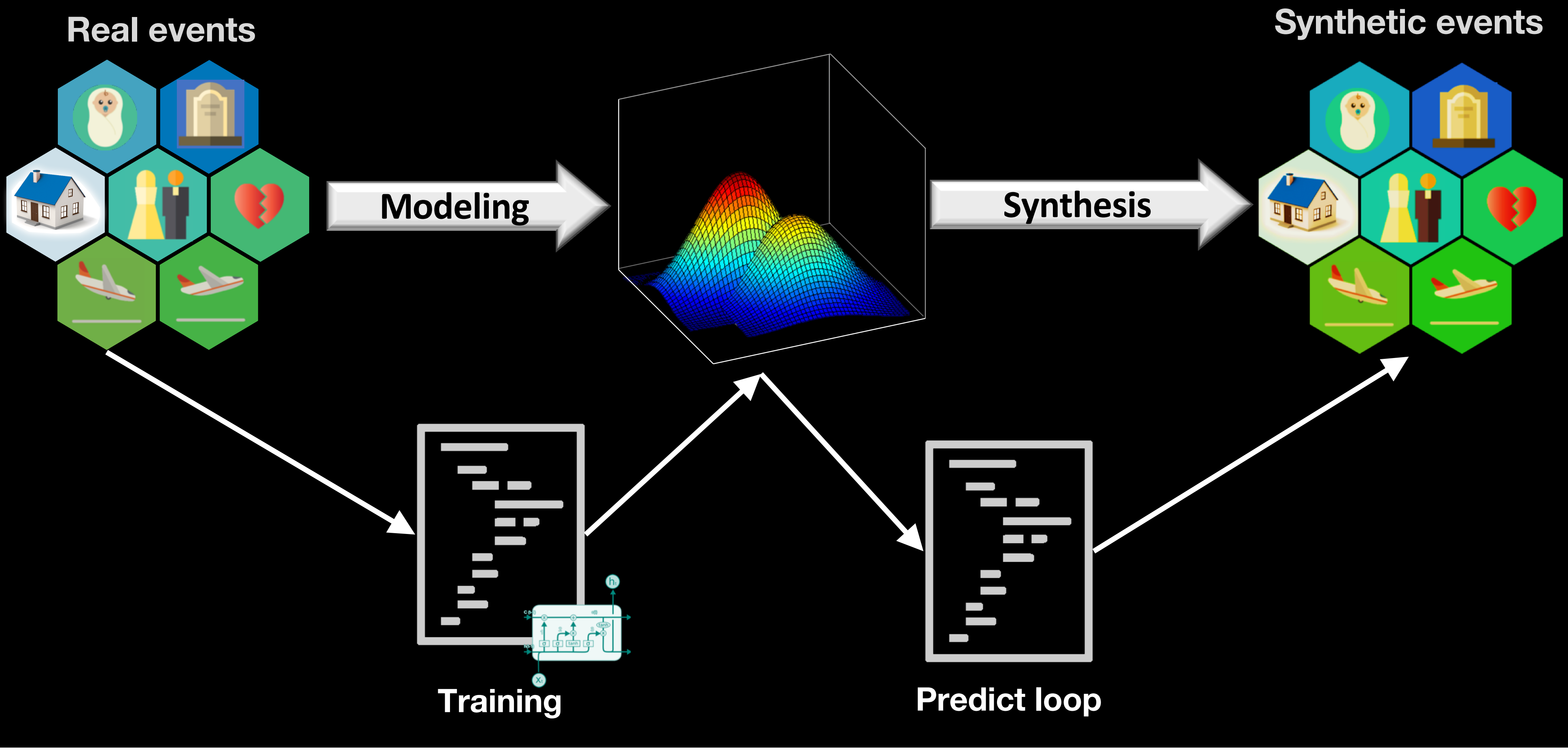
Long Short-Term Memory (LSTM)



https://www.deeplearningbook.org/, Chapter 10

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Other algorithms



Real events

Modeling

Synthesis

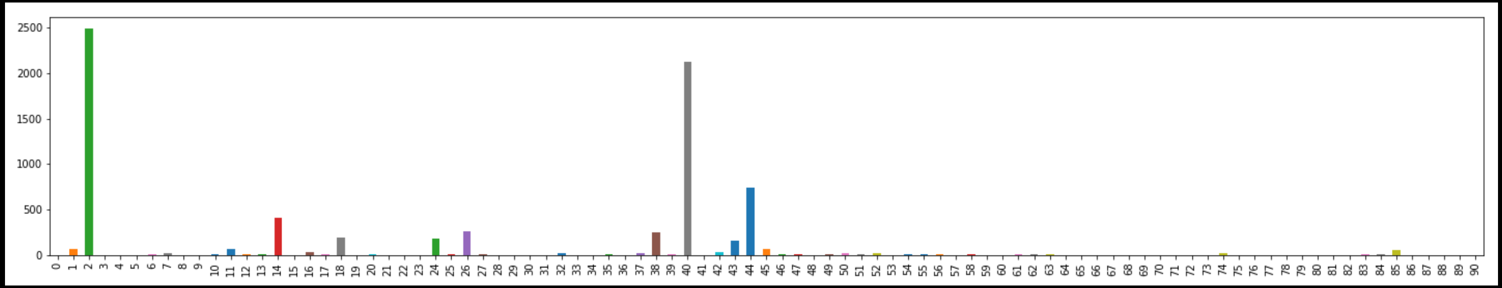
Synthetic events

Training

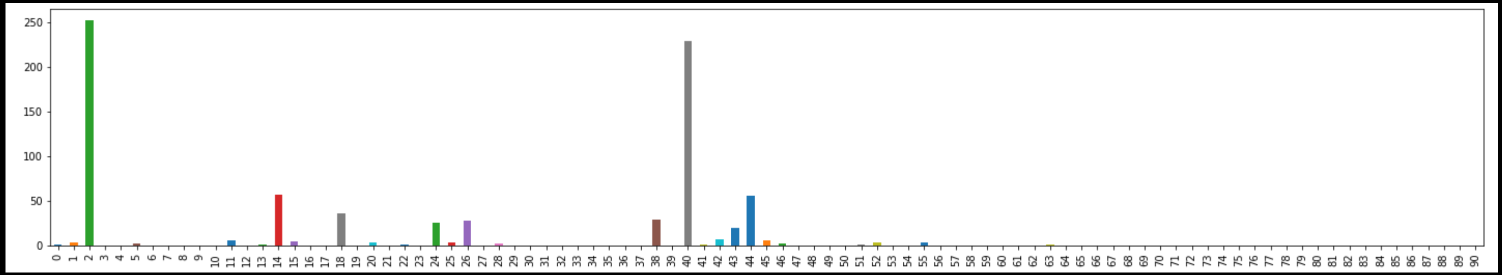
Predict loop

Distribution of event types

Original



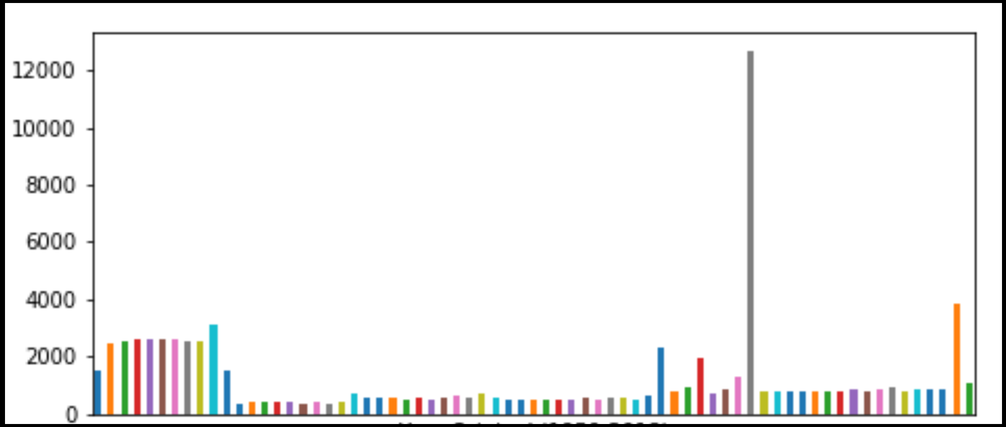
Synthetic



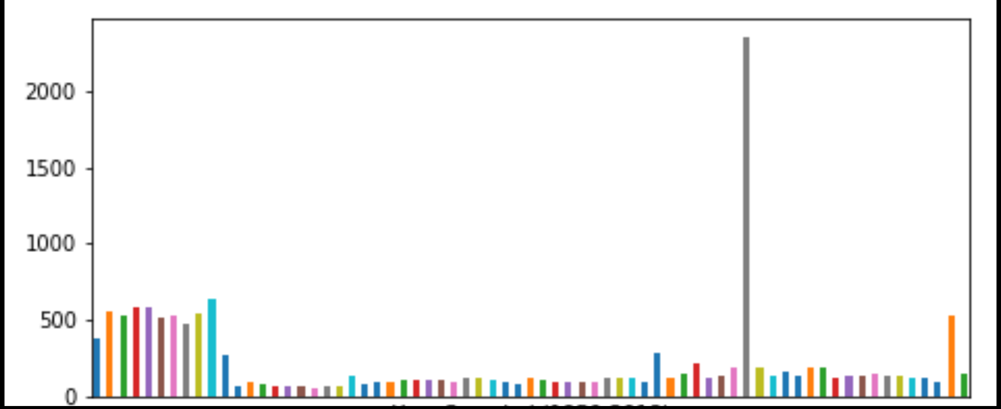
Distribution of the birthyears

Year

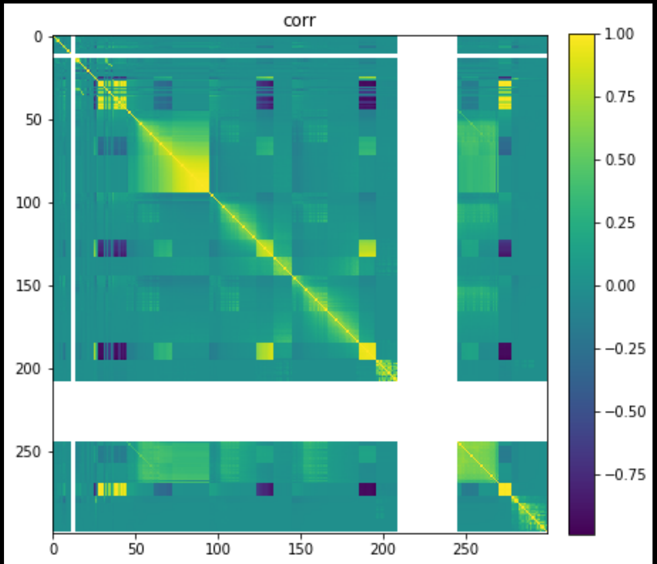
Original



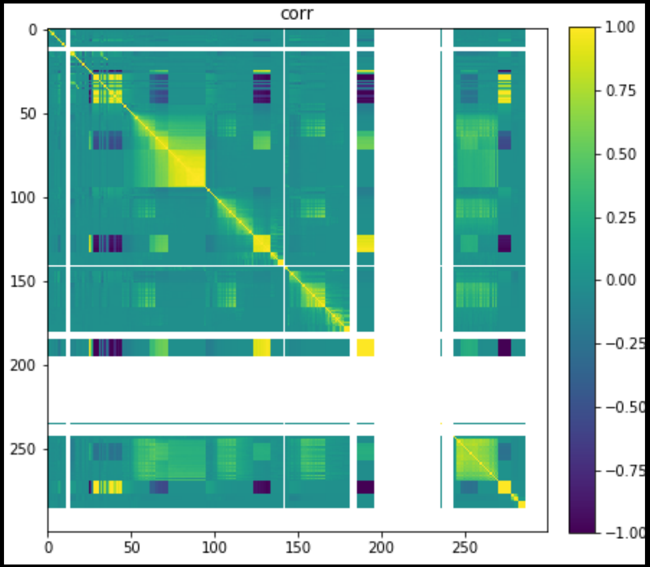
Synthetic



Comparison of correlation matrices

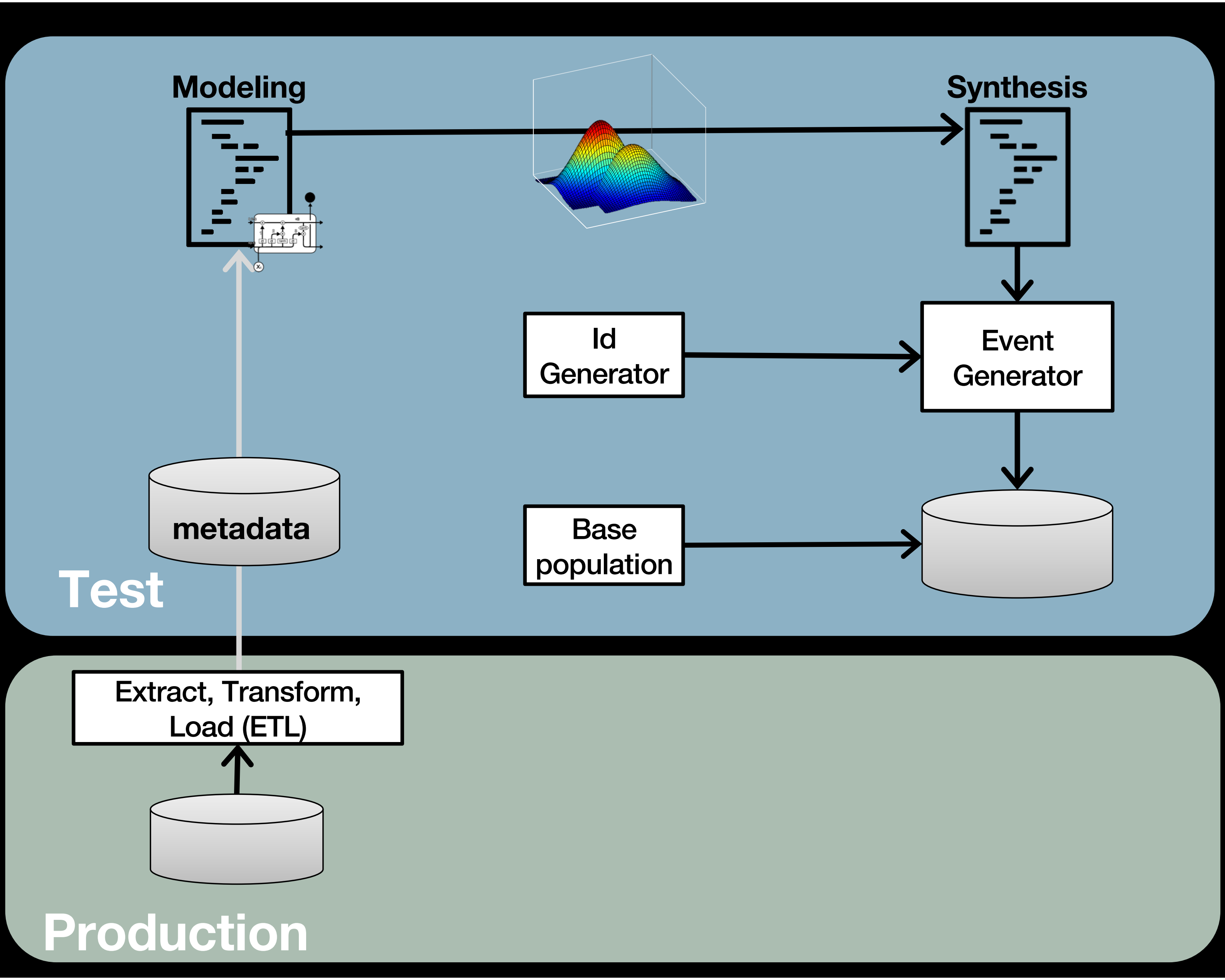


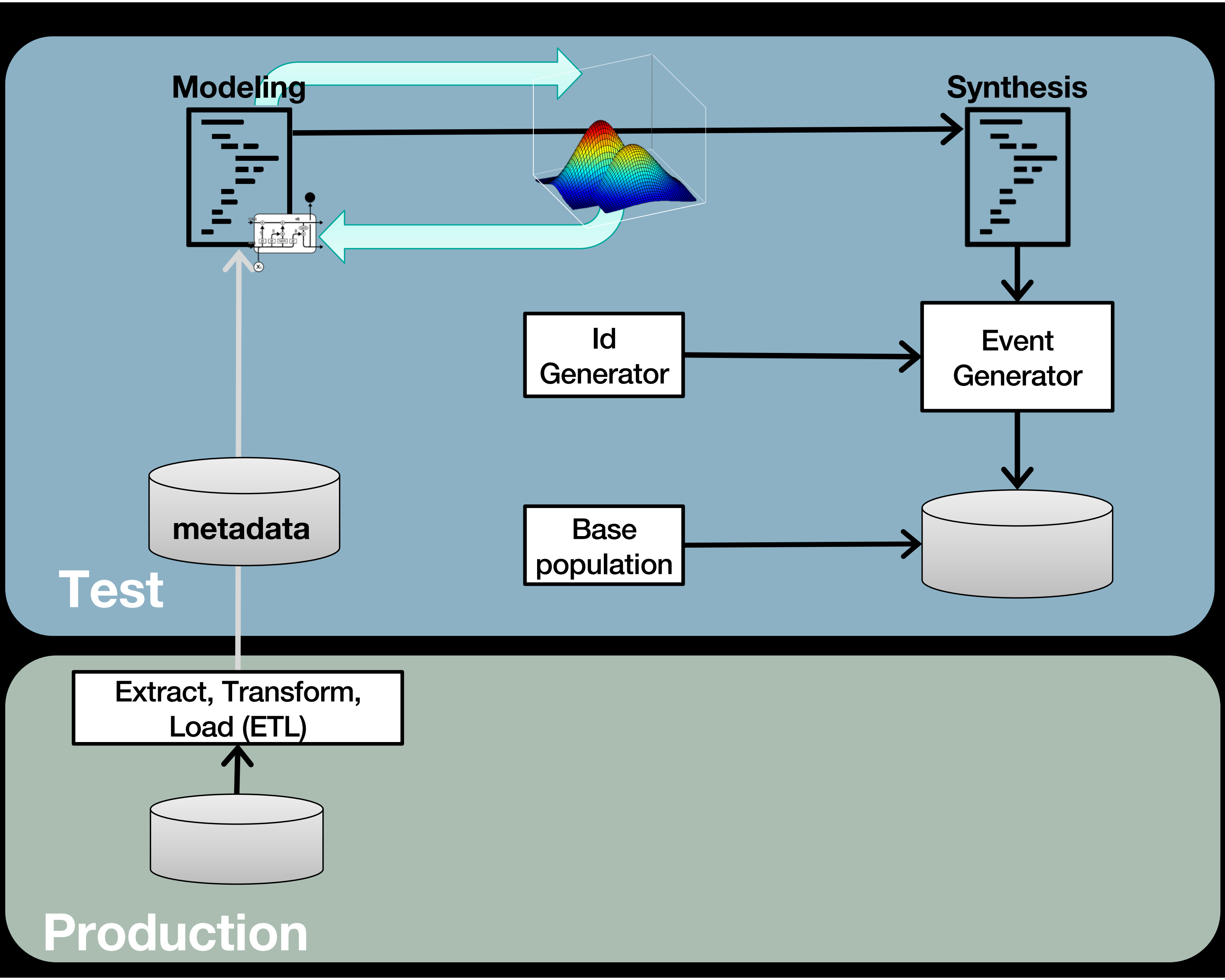
Original

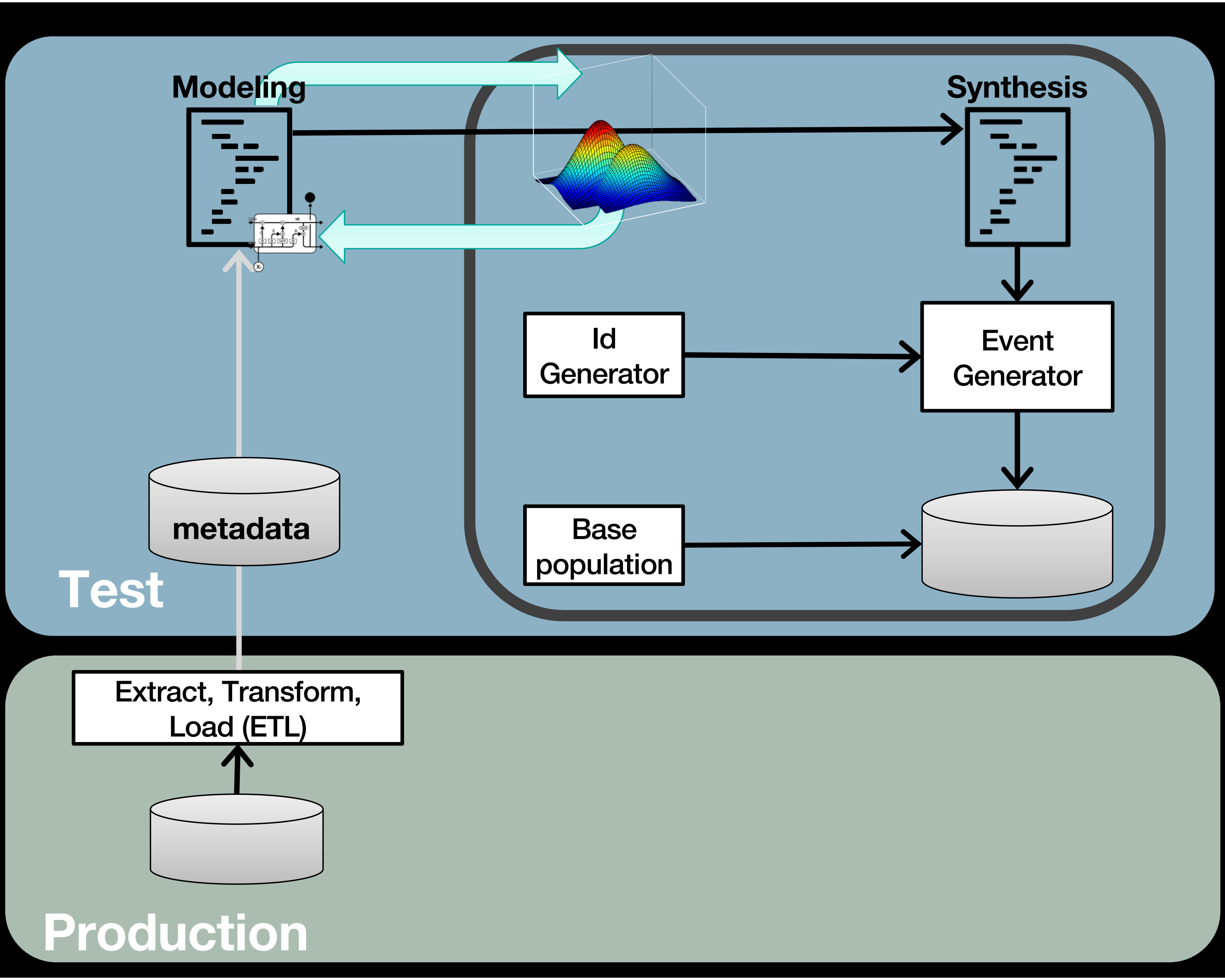


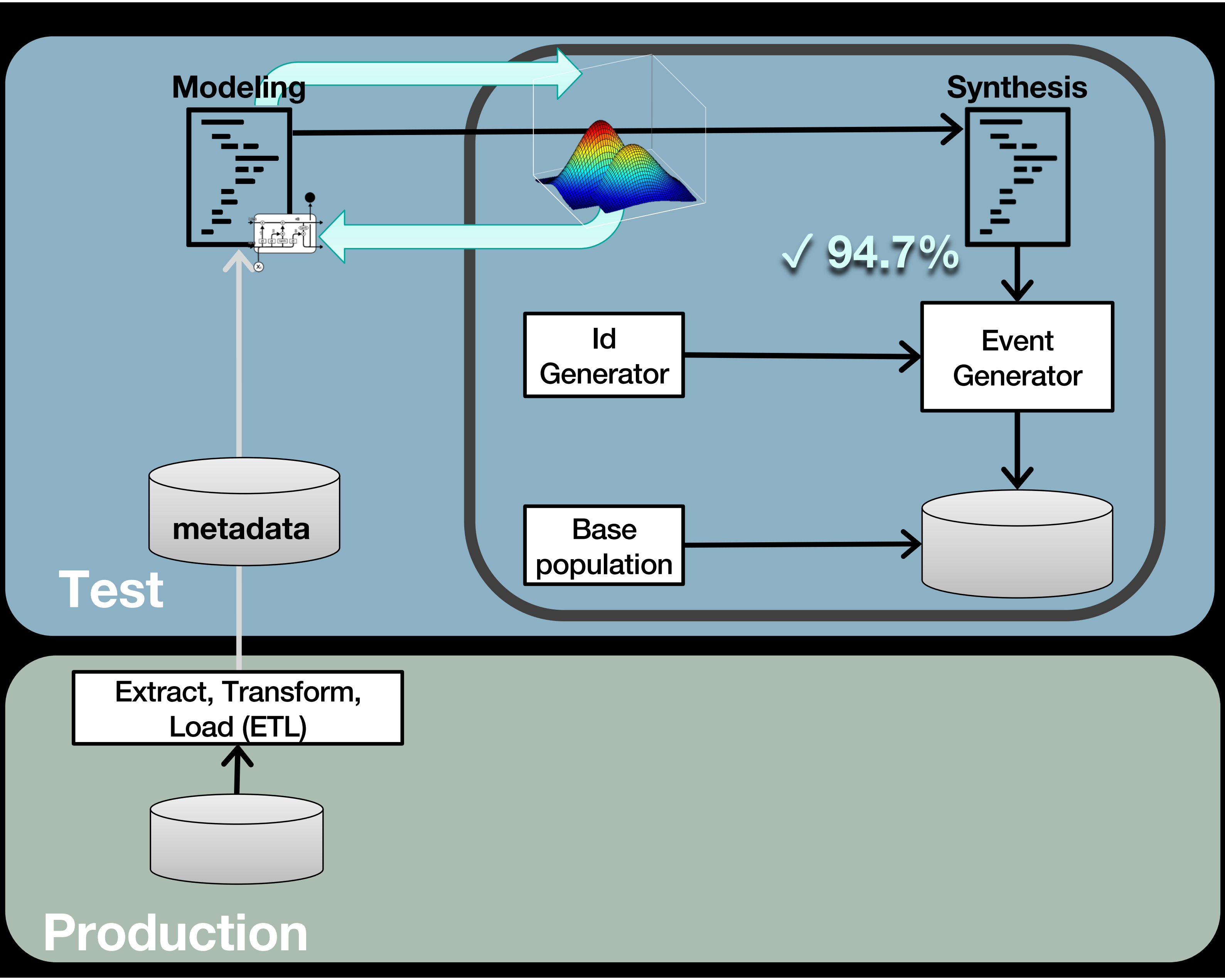
Synthetic

Machine Learning is only a small part of the solution


Modeling
Synthesis
Id Generator
Event Generator
metadata
Base population
Test
Extract, Transform, Load (ETL)
Production


Modeling
Synthesis
Id Generator
Event Generator
metadata
Base population
Test
Extract, Transform, Load (ETL)
Production


Modeling
Synthesis
Id Generator
Event Generator
Base population
metadata
Test
Extract, Transform, Load (ETL)
Production



Modeling

Synthesis

✓ 94.7%

Id Generator

Event Generator

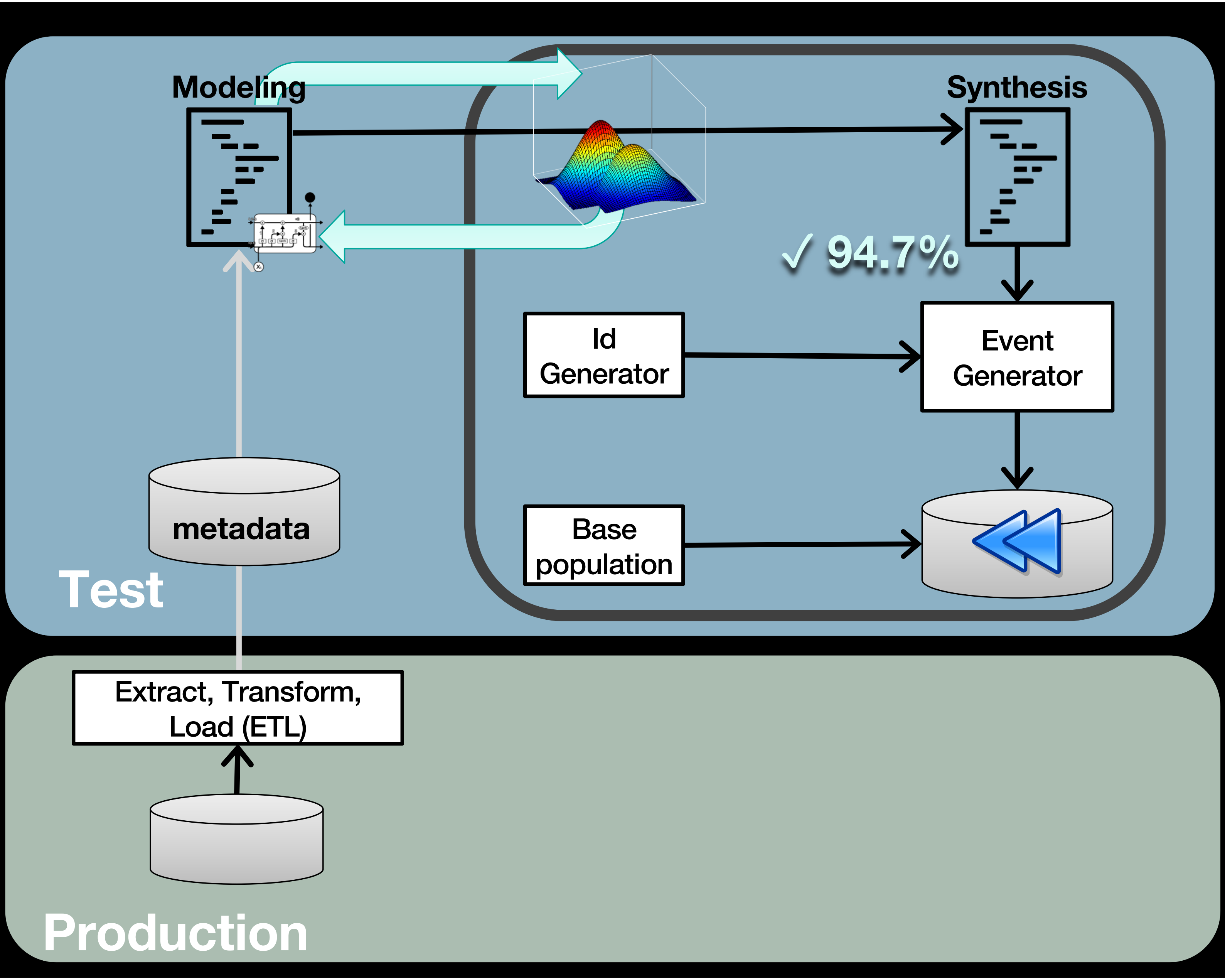
Base population

metadata

Test

Extract, Transform, Load (ETL)

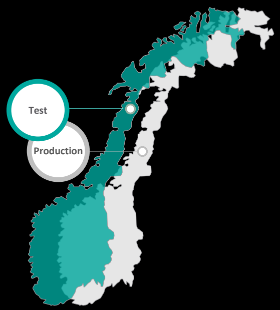
Production

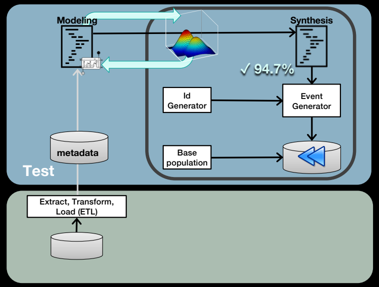

Modeling
Synthesis
✓ 94.7%
Id Generator
Event Generator
Base population
metadata
Test
Extract, Transform, Load (ETL)
Production

Current Status

- More than half a million people
- About 2000 new events every day
- About 100 new people every day

When is machine learning suitable for data generation?




Modeling

Synthesis

✓ 94.7%

Id Generator

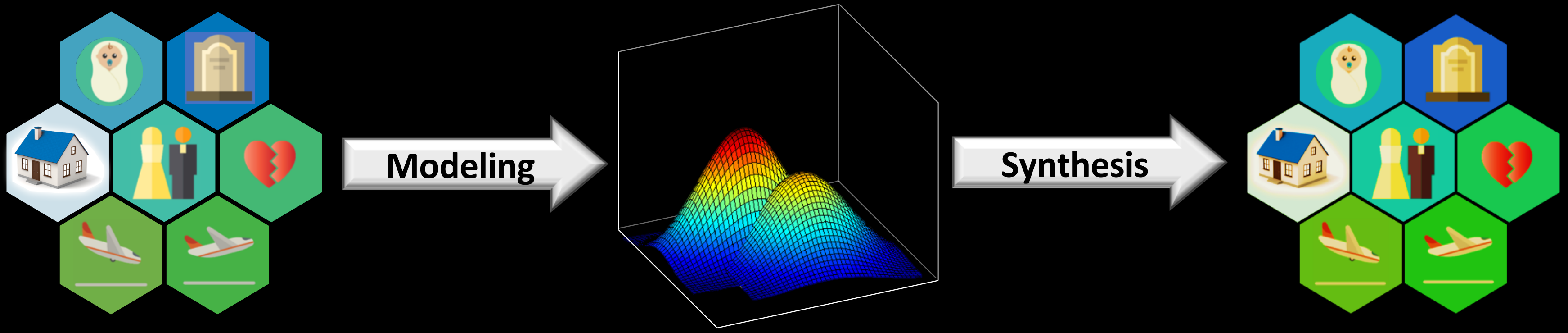
Event Generator

Base population

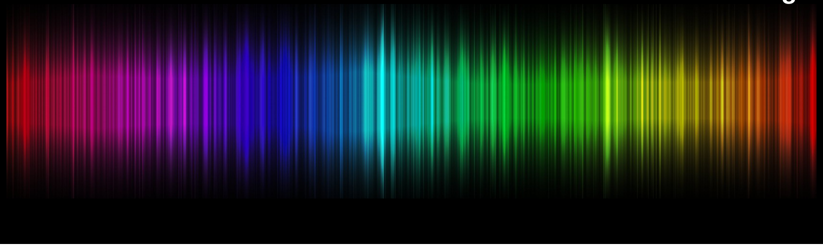
metadata

Test

Extract, Transform, Load (ETL)


Modeling

Synthesis


Collections of joint PDF/PMF

Deep Learning

Margrethe Bedregal

Marianne Rynning

Erik Arisholm

Atle Myklebost

Martin K. Gran

Chao Tan

Razieh Behjati

Tobias Lund-Melcher

Arne Asphjell

Knut Botheim

Rikard Eriksen

Viveca Liodden

Rune Myrdal

Lauritz Møllersen

Joachim Lous

Gaute Lote

Gisle Austefjord

Stein Petter Tokvam

Andreas Aubell

Testify
smartere testing

The Norwegian
Tax Administration

The Research Council
of Norway

Contact us!

razieh.Behjati@gmail.com

https://no.linkedin.com/in/rbehjati

@RBehjati

margrethe.bedregal@skatteetaten.no

erik.arisholm@testify.no