

# Deteksjon av hvitvasking

Martin Jullum  
Seniorforsker  
Norsk Regnesentral

Analysesamling Skatteetaten, 15.11.23



# Agenda

Forskningssenteret BigInsight

Hvitvasking: Hva, hvordan og hvorfor?

Prosjekt 1: Maskinlæring for deteksjon av mistenkelige transaksjoner

Prosjekt 2: GNN for deteksjon av hvitvaskere



Journal of

**Money Laundering  
Control**

**Detecting money laundering  
transactions with  
machine learning**

Martin Jullum, Anders Løland and Ragnar Bang Huseby  
*Norwegian Computing Center, Oslo, Norway, and*

Geir Ånonsen and Johannes Lorentzen  
*DNB, Oslo, Norway*



Finding Money Launderers Using Heterogeneous Graph  
Neural Networks

Fredrik Johannessen<sup>\*1</sup> and Martin Jullum<sup>†2</sup>

<sup>1</sup>DNB, P.O. Box 1600, Sentrum, NO-0021 Oslo, Norway

<sup>2</sup>Norwegian Computing Center, P.O. Box 114, Blindern, NO-0314 Oslo, Norway

# Forskningsssenteret BigInsight



UiO : Universitetet i Oslo

UNIVERSITETET I BERGEN



Skatteetaten



Gjensidige



  
**BigInsight**  
2015 – 2023

100 mill. fra  
NFR + 50 mill.  
fra partnerne

## INNOVATION OBJECTIVES



Personalised  
marketing



Personalised  
health and  
patient safety



Personalised  
fraud  
detection



Sensor  
systems



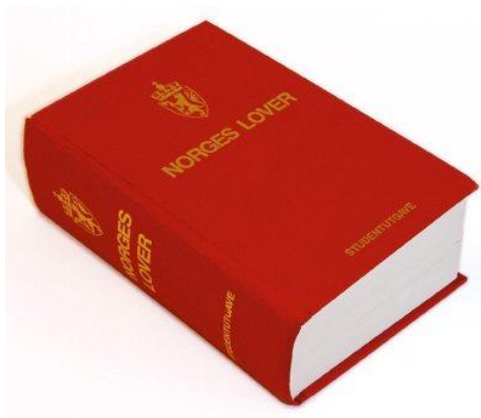
Forecasting  
power  
systems



Explaining AI

# Hvitvasking

## Hva, hvordan og hvorfor?



# Hvitvasking

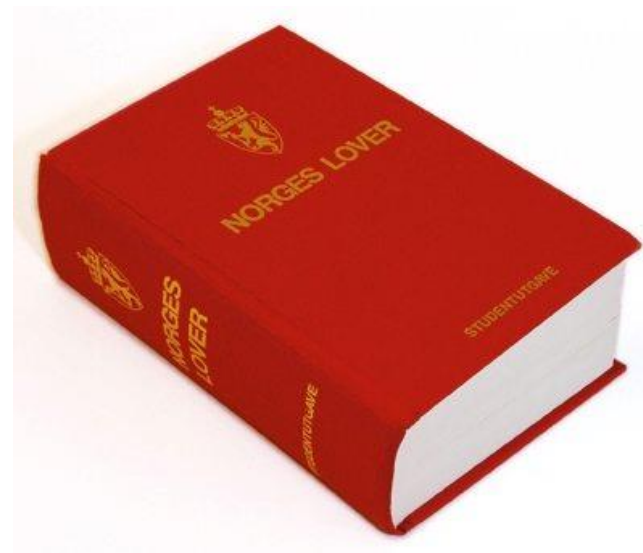
- ▶ Å få penger fra kriminell aktivitet til å se lovlig ut
- ▶ Eksempler
  - Kjøpe antikviteter med skitne penger – presenter som loftsfunn – selg lovlig
  - Inkluder midler fra kriminell aktivitet i egen lovlig virksomhet





# Hvitvasking

- ▶ Å få penger fra kriminell aktivitet til å se lovlig ut
- ▶ Eksempler
  - Kjøpe antikviteter med skitne penger – presenter som loftsfunn – selg lovlig
  - Inkluder midler fra kriminell aktivitet i egen lovlig virksomhet



- ▶ Hvitvaskingsloven: Alle finansielle foretak er pålagt å rapportere mistenkelige transaksjoner til Økokrim

# Hvorfor er (anti-)hvitvask viktig?

 **REUTERS**  
BUSINESS NEWS

MARCH 19, 2020 / 8:14 PM / 5 MONTHS AGO

Sweden  
over E

**The Guardian**

Standard Chartered fined \$1.1bn for  
money-laundering and sanctions  
breaches

FINANCE

**FORTUNE**

A Money-Laundering  
Scandal

...ion fine

ABC

Justice and Crime

ing Mega-  
d the CEO  
t Bank

# Prosjekt 1: Maskinlæring for deteksjon av mistenkelige transaksjoner



dmlc  
**XGBoost**



Journal of

**Money Laundering  
Control**

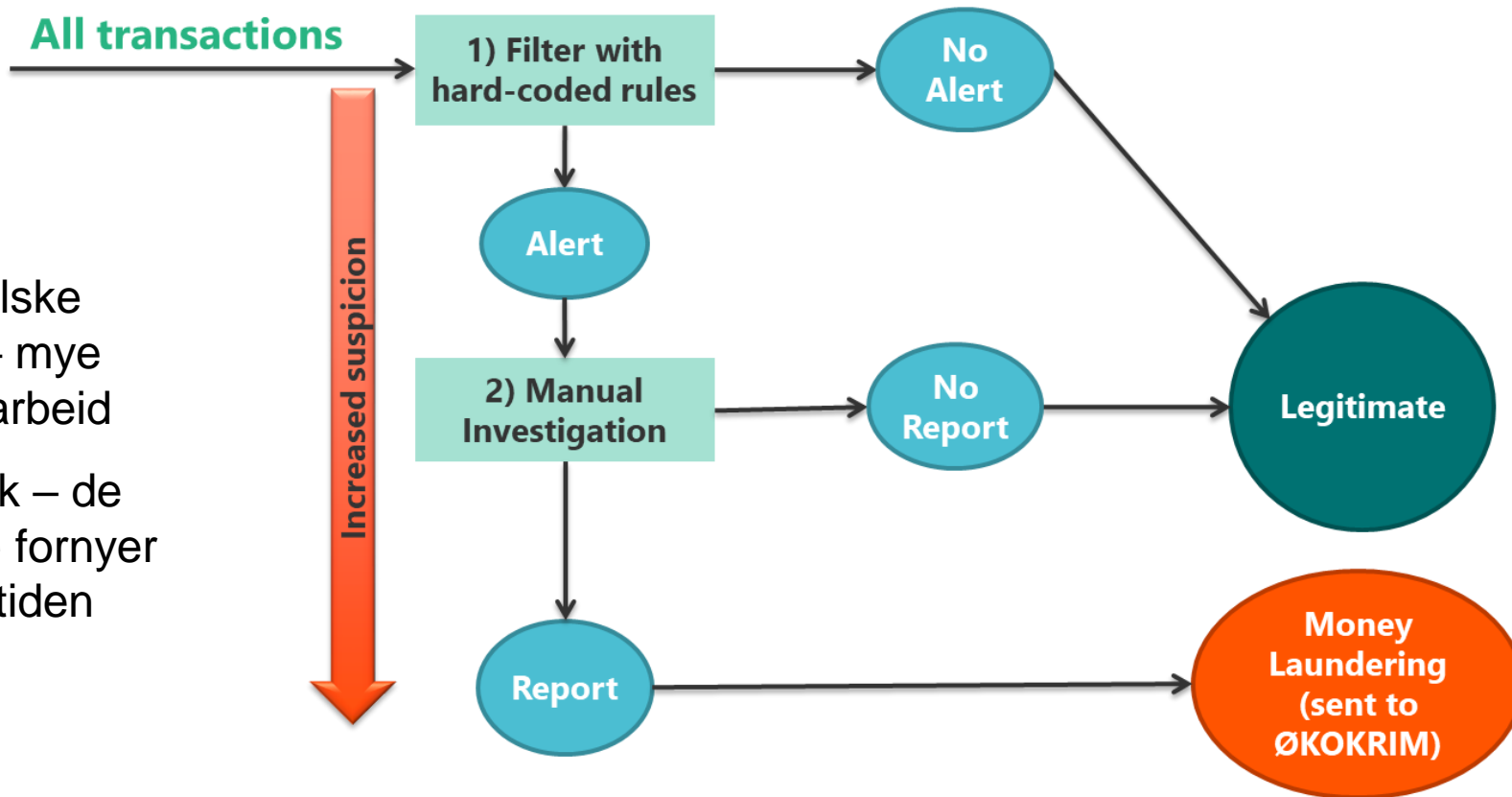
**Detecting money laundering  
transactions with  
machine learning**

Martin Jullum, Anders Løland and Ragnar Bang Huseby  
*Norwegian Computing Center, Oslo, Norway, and*

Geir Ånonsen and Johannes Lorentzen  
*DNB, Oslo, Norway*



# Typisk hvitvaskingssystem

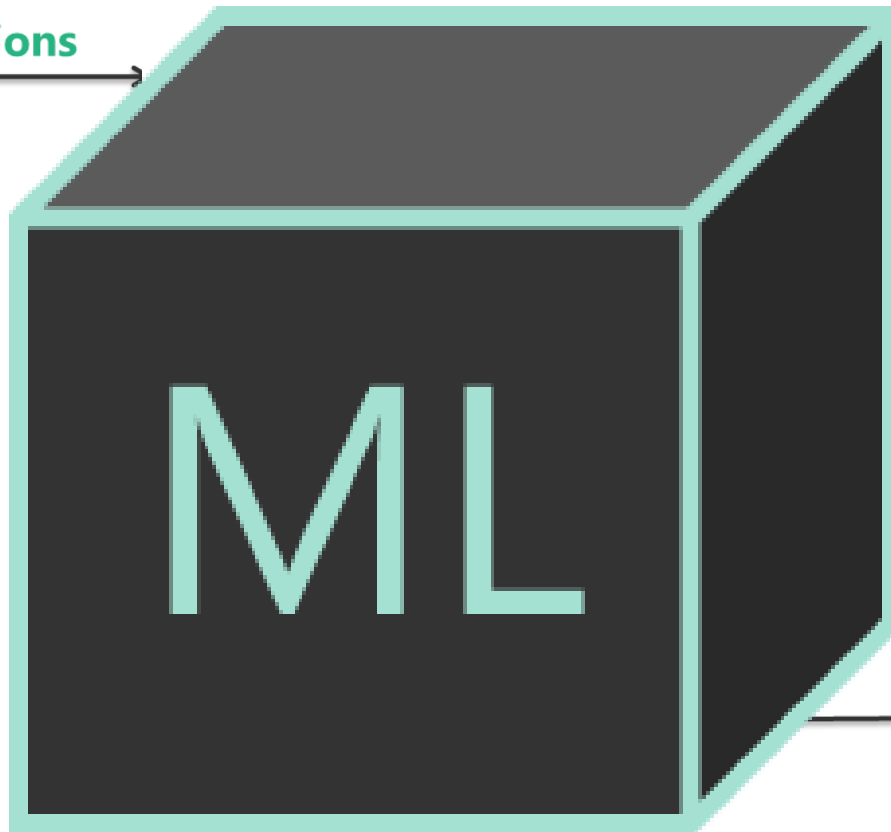


## Svakheter

- Mange falske positive – mye manuelt arbeid
- For statisk – de kriminelle fornyer seg hele tiden

# Vår “løsning”

All transactions



Legitimate

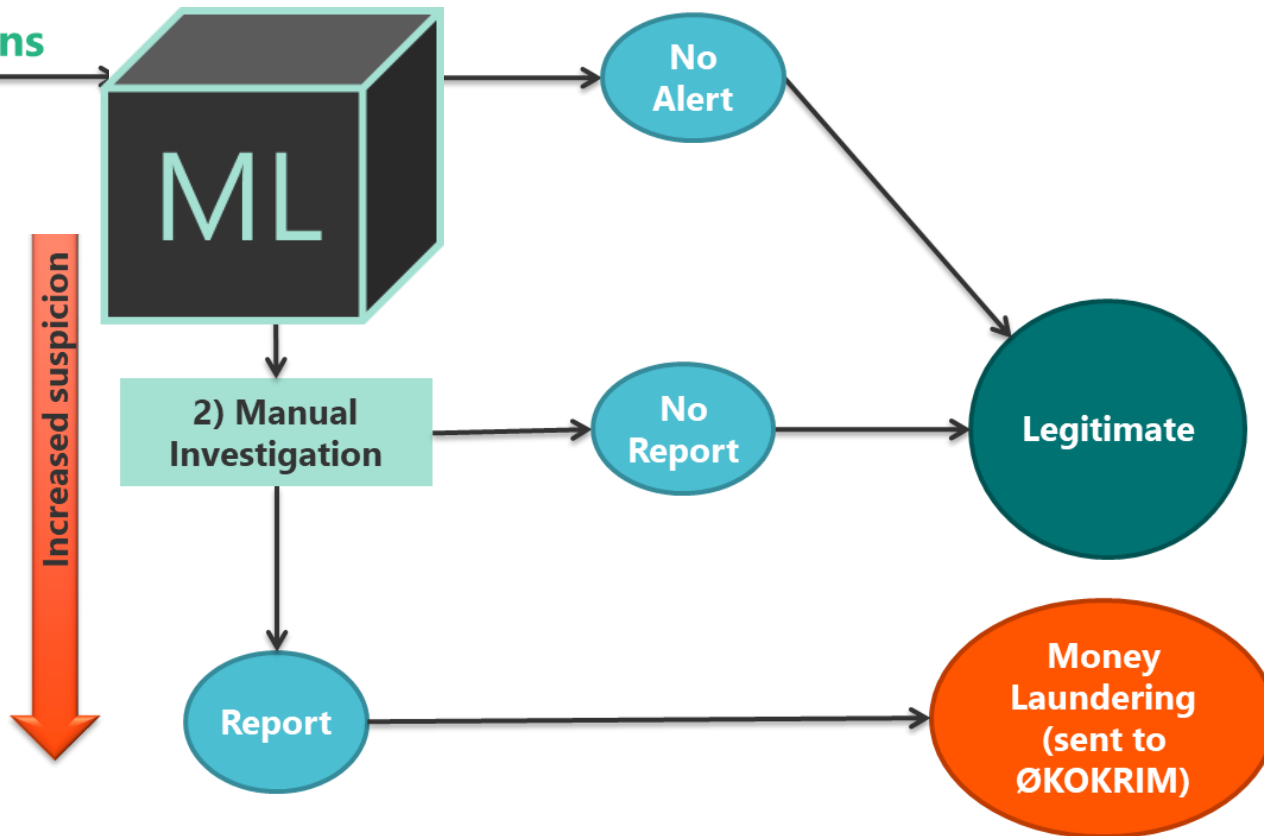
Money  
Laundering  
(sent to  
ØKOKRIM)

- Bytt ut hvitvaskingssystemet med en maskinlæringsmodell
- Typer data
  - transaksjons-historikk
  - kundedata
  - alarmer
  - manuelle saker

# Vår “løsning”

## Mer realistisk oppsett!

All transactions

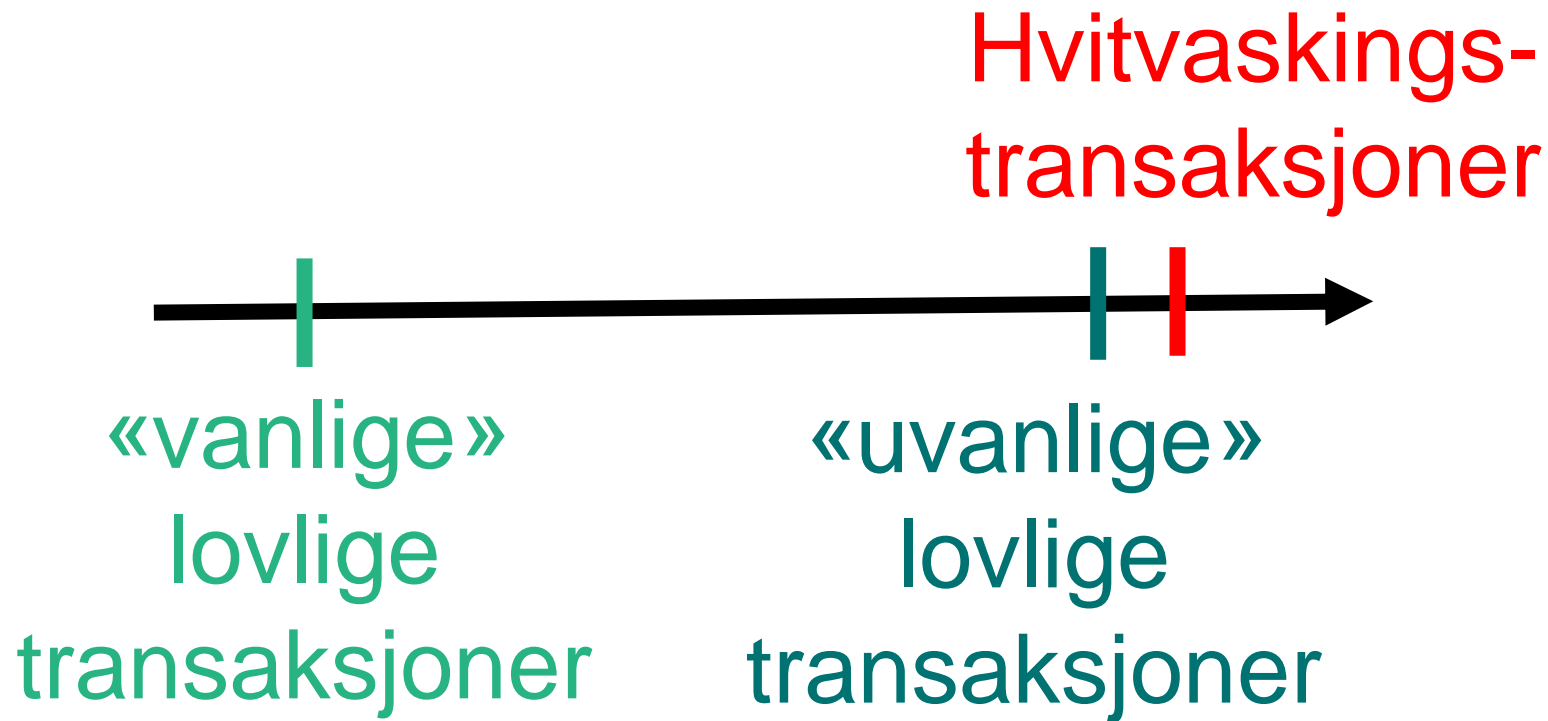


► Bytt ut  
hvitvaskingssystemet  
med en  
maskinlæringsmodell

► Typer data

- transaksjons-  
historikk
- kundedata
- alarmer
- manuelle saker

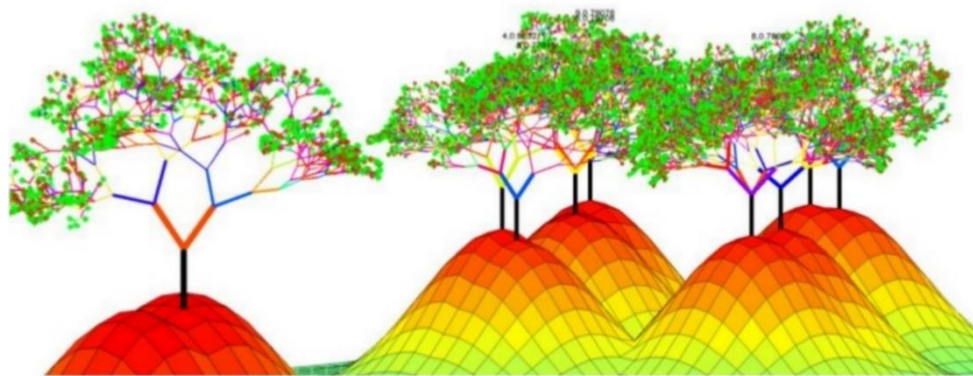
# Er dette så vanskelig da?



# Modellering

- Binær respons ( $Y$ ): Transaksjon sendt til Økokrim (Ja = 1, nei = 0)
- Vil predikere  $P(Y = 1 | \text{data relatert til aktuell transaksjon})$
- State of the art: **Gradient boosting** machines (GBM)
- **XGBoost** – veldig effektiv og fleksibel implementering av GBM basert på tre-modeller

dmlc  
**XGBoost**



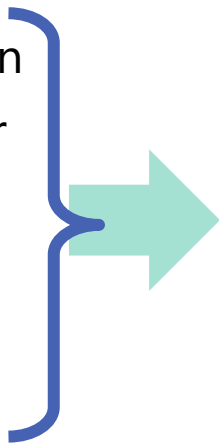


# Transformerings av rådata (feature engineering)

XGBoost krever numeriske tabelldata som input!

## Typer input data

- Spesifikk info om aktuell transaksjon
- Bakgrunnsinfo om sender/mottaker
- Sender/mottakers transaksjonshistorikk
- Tidligere rapporterte transaksjoner fra sender/mottaker



Y	X1	X2	X3	X4	X5	X6
1	0,453406	0,992838	0,734389	0,159918	0,397515	0,949952
0	0,274	0,654207	0,169886	0,493841	0,407112	0,939789
0	0,741897	0,855005	0,585788	0,366456	0,365123	0,57955
1	0,488119	0,465754	0,716517	0,493048	0,855049	0,632114
0	0,134458	0,762057	0,848194	0,098779	0,872603	0,063026
0	0,531914	0,998817	0,808215	0,060721	0,716595	0,35374
0	0,341509	0,8398	0,637808	0,48304	0,279987	0,730286
0	0,530306	0,463271	0,338713	0,986781	0,925251	0,272484
1	0,864123	0,652763	0,689599	0,080937	0,990294	0,364736
0	0,106812	0,900351	0,450224	0,143815	0,593244	0,020764

1716 kolonner (features)

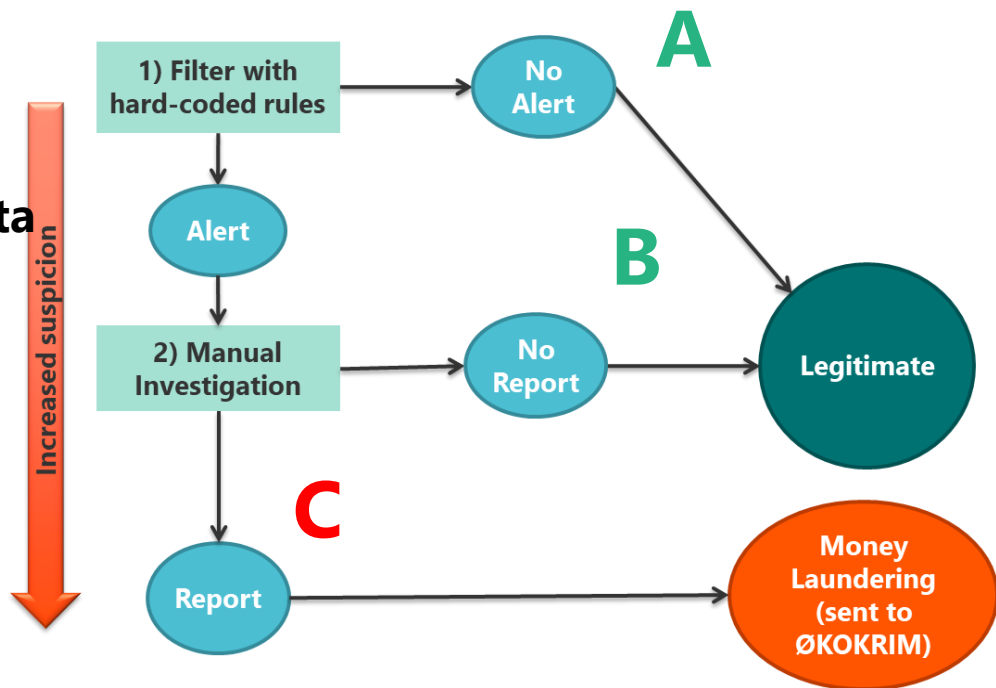
# Dataraffinering

## 2 år med modellerbare transaksjonsdata

- Alle transaksjoner som leder til
  - Rapportering (C)
  - Alarm, men ikke rapportering (B)
- Et tilfeldig utvalg normale transaksjoner (A)

## Dataraffinering

- Valgte #A = #B
- Kun én transaksjon fra hver manuelle inspeksjon (2)
- Ingen transaksjoner fra samme sender/mottaker to etterfølgende dager



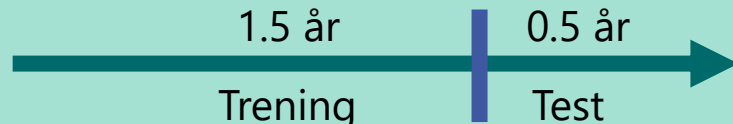
# Trening, testing og modellering

## Modellering

- 10-fold kryssvalidering
- Stoppekriterium (# boosting-iterasjoner): AUC
- Tuning: Tilfeldig + iterativt gridsøk
- Trening på GPU
- Endelig modell brukt til prediksjon på testdata:

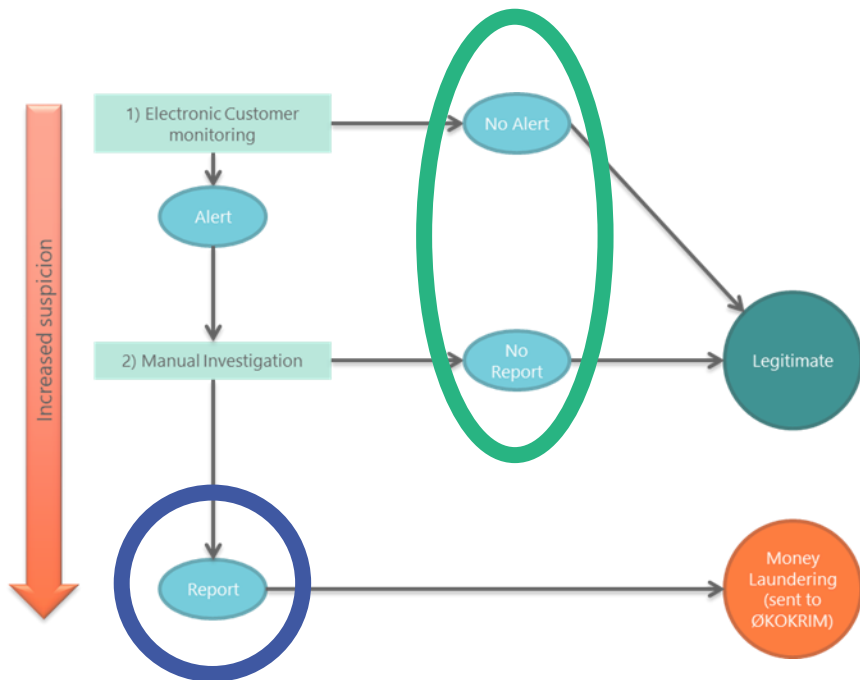
$$\hat{f}(x_{\text{test}}) = \frac{1}{10} \sum_{i=1}^{10} \hat{f}_{cv,-i}(x_{\text{test}})$$

## Out-of-time testing

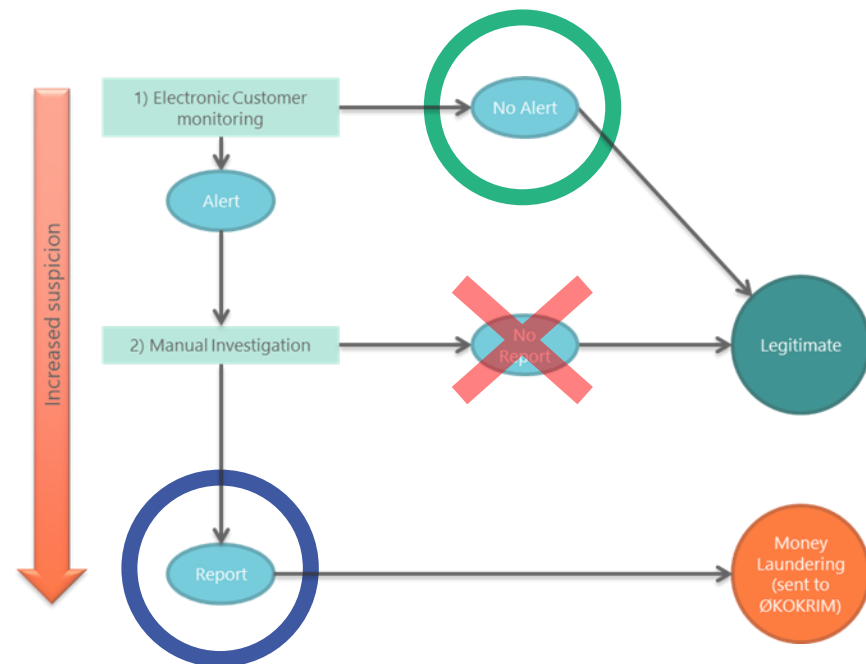


# 2 treningsscenarier

Alle datatyper

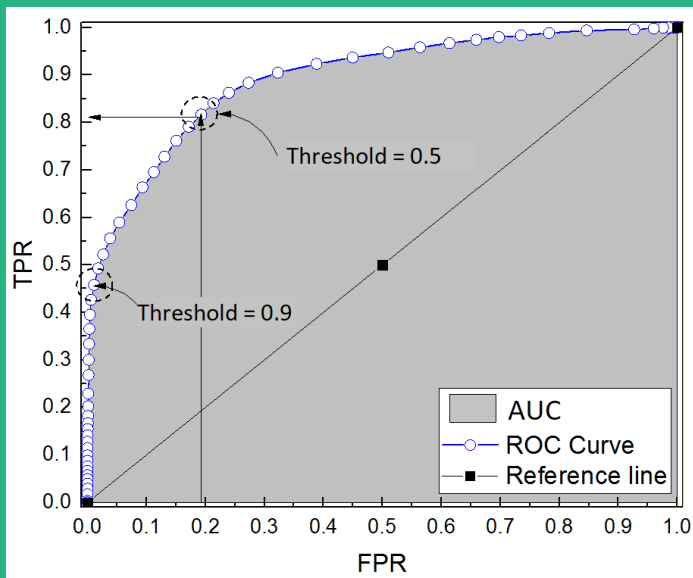


Uten urapporterte transaksjoner



# Evaluering av scenariene

Rangering:  
AUC



Sannsynligheter:  
Brier score

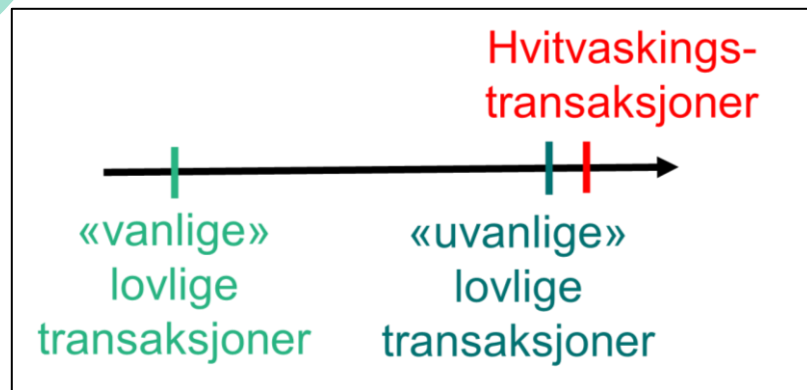
$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{p}_i)^2$$



# Sammenligning av scenarier

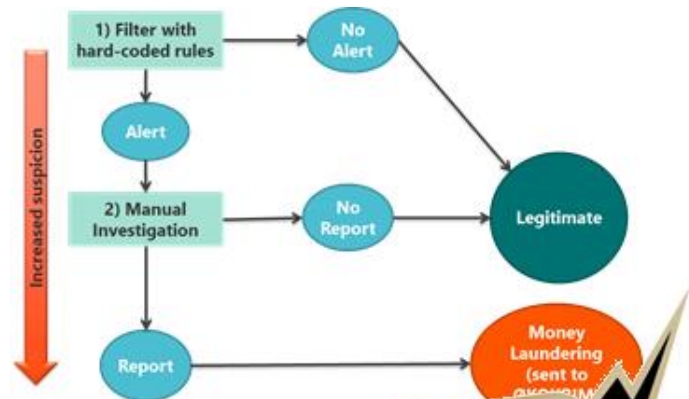
	Alle datatyper	Uten urapporterte transaksjoner
AUC	0.907	0.852
Brier	0.025	0.340

**MYE bedre!**



# ML vs dagens AML system

- Vanskelig å sammenligne
- **PPP = Proportion of Positive Predictions:**  
Andel transaksjoner som må kontrolleres for å finne 95% av de rapporterte transaksjonene

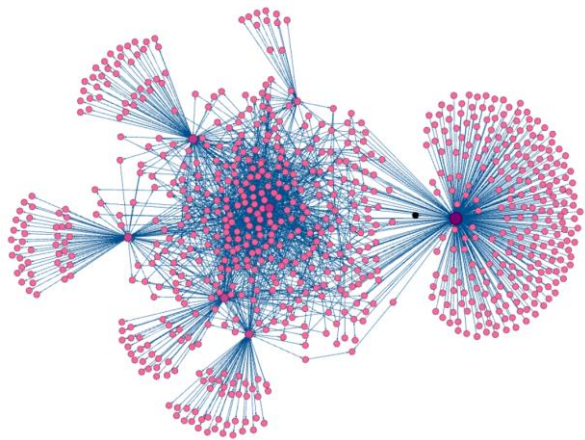


	ML (alle datatyper)	Dagens system
PPP	31.5 %	48.9 %

# Begrensninger

- ▶ Vi bruker egentlig ikke **transaksjonsnettverket**
  - **Hvem** du sender/mottar penger til/fra
- ▶ Informasjon fra sosiale/profesjonelle nettverk blir ikke utnyttet
- ▶ Mange variabler – vanskeligere å sette i produksjon
- ▶ Modellen lærer kun fra det som allerede er rapportert

# Prosjekt 2: GNN for deteksjon av hvitvaskere



Finding Money Launderers Using Heterogeneous Graph  
Neural Networks

Fredrik Johannessen<sup>\*1</sup> and Martin Jullum<sup>†2</sup>

<sup>1</sup>DNB, P.O. Box 1600, Sentrum, NO-0021 Oslo, Norway

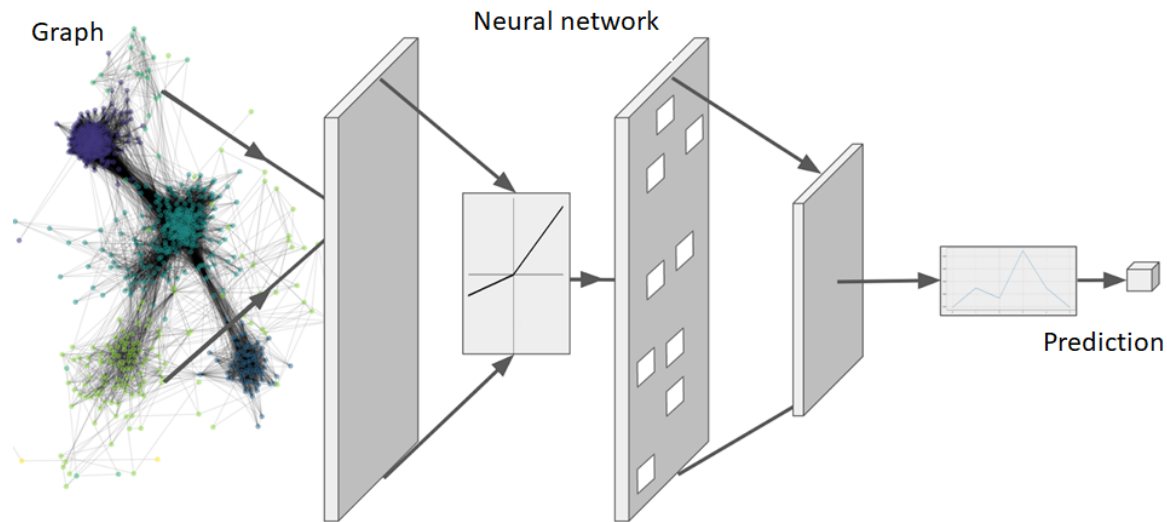
<sup>2</sup>Norwegian Computing Center, P.O. Box 114, Blindern, NO-0314 Oslo, Norway



PyG

# Bakgrunn

- ▶ GNN: Graph Neural Network
  - Klasse metoder som bygger prediktive modeller direkte på grafdata
- ▶ Prosjektet startet som en masteroppgave



**Department of Mathematics**  
University of Oslo

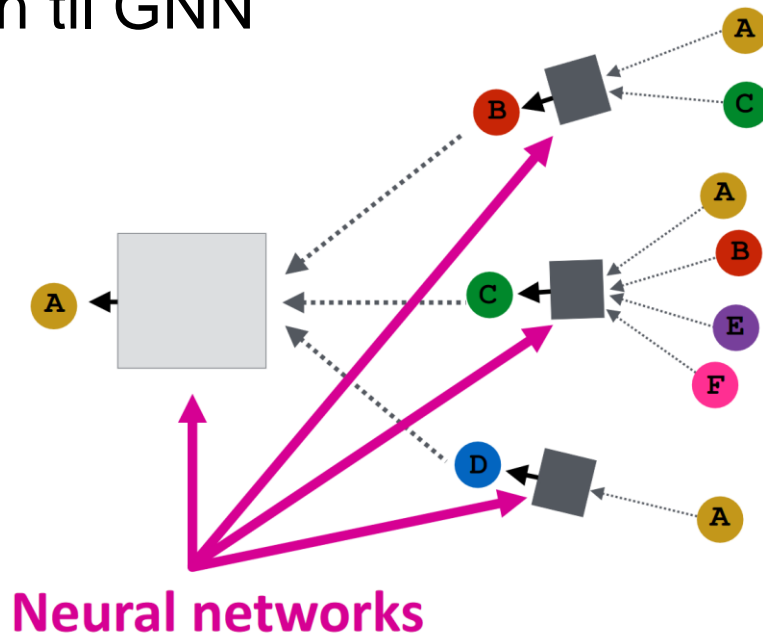
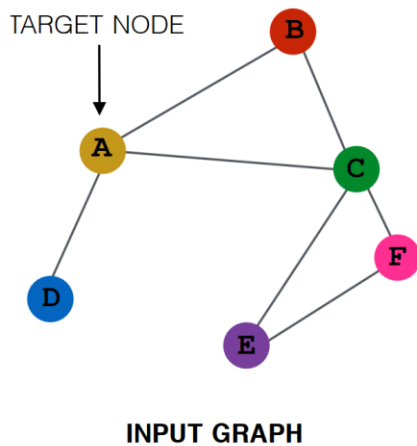
**Finding Money Launderers  
Using Heterogeneous Graph  
Neural Networks**

Fredrik Johannessen  
Master's Thesis, Spring 2022



# Message passing

- Message passing er kjerneideen til GNN

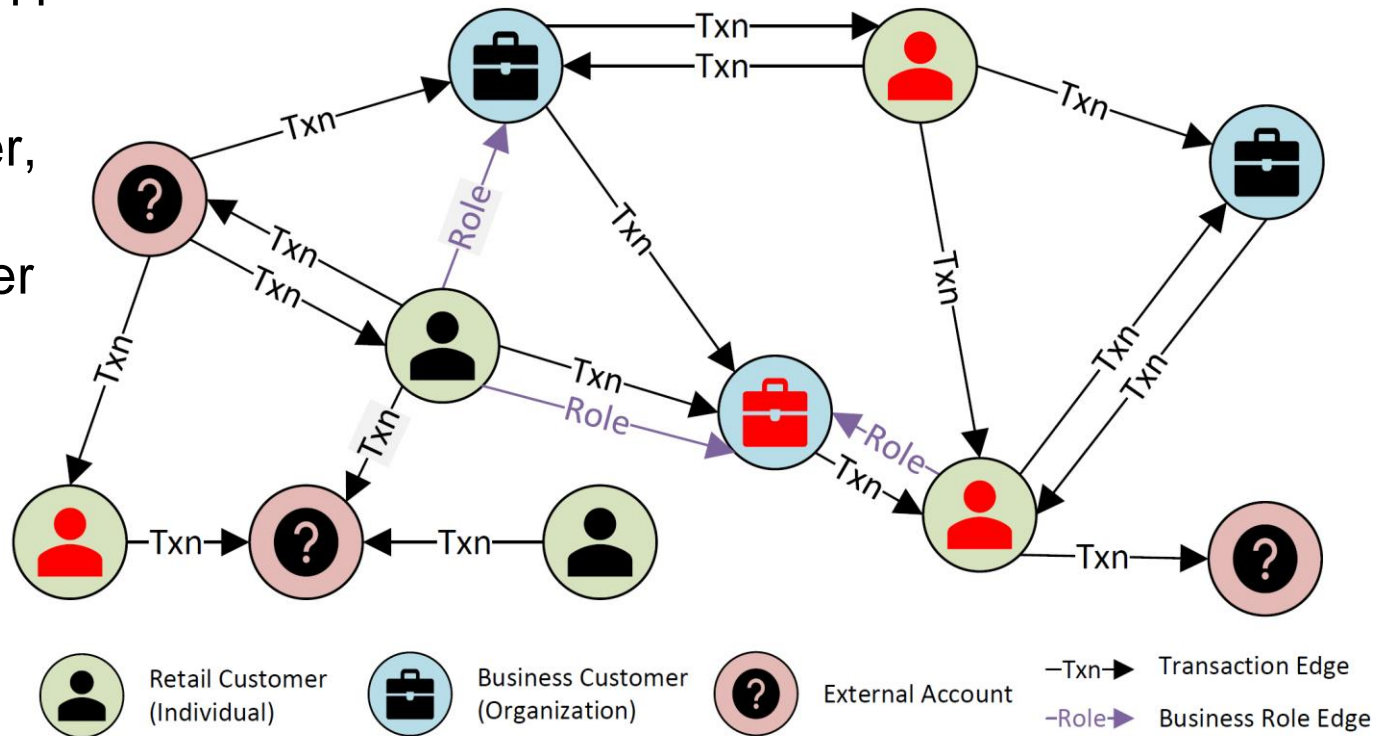


- Aggregeringsparametere deles på tvers av noder – tillater generalisering til nye noder

# Vårt hvitvaskingsnettverk

Grafen er heterogen i både

- **Noder:** individer, bedrifter, eksterne kontoer
- **Kanter:** transaksjoner, roller



5 millioner noder  
9 millioner kanter

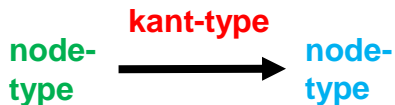
# Eksisterende GNN-rammeverk

- ▶ De fleste metodene er utviklet for homogene grafer (GCN, GraphSage, GAT, ...)
- ▶ R-GCN (Schlichtkrull et al., 2018) er en utvidelse av GCN som håndterer flere kanttyper
  - Bruker én message passing funksjon per kanttype
  - Håndterer **ikke** flere node-typer
  - Håndterer **ikke** forklaringsvariabler for kantene
- ▶ MPNN (Gilmer et al., 2017) er et rammeverk som forrener mange homogene GNN-metoder

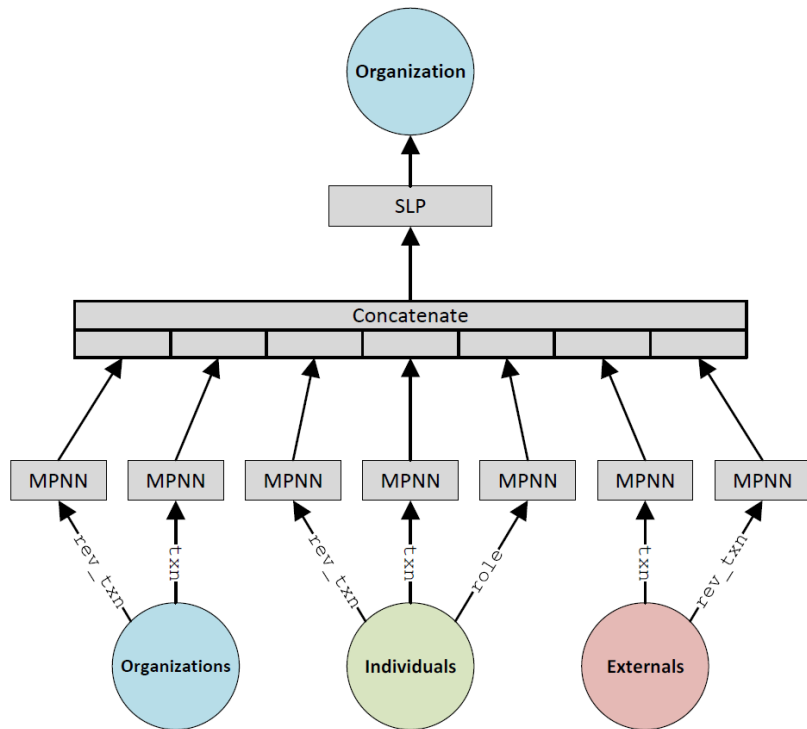
# Vår modell: H-MPNN

► **H-MPNN:** Utvider MPNN til å håndtere heterogene grafer:

- En MPNN-modell per kombinasjon av

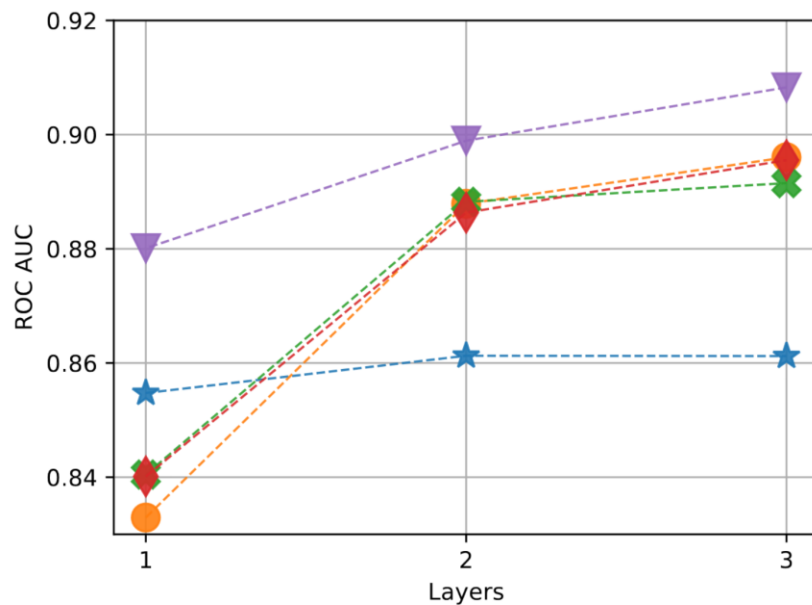
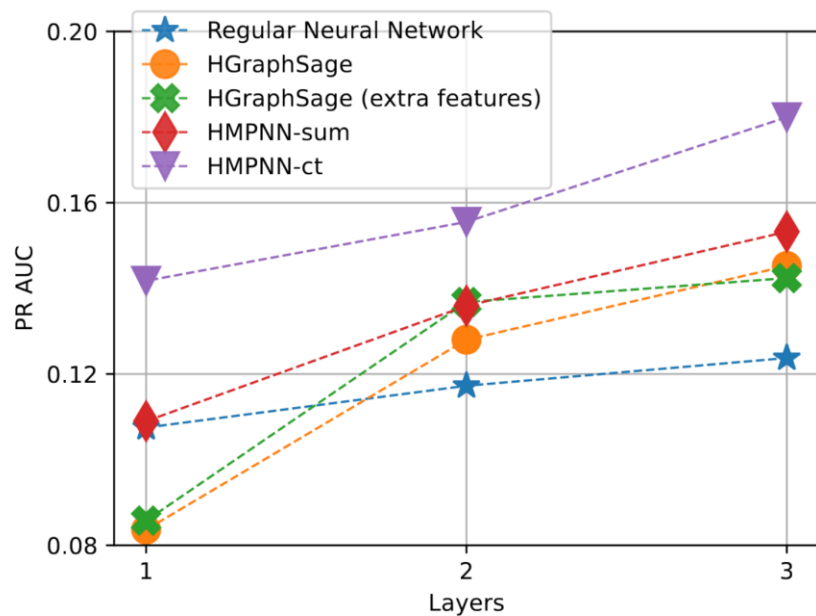


- 2 alternativer for aggregering til endelig embedding:
  - Sum av MPNN-embeddings
  - Trener ekstra SLP-lag med alle MPNN-embeddings som input



PyG

# Resultater





Takk for oppmerksomheten



**Martin Jullum – [martinjullum.com](http://martinjullum.com) – [jullum@nr.no](mailto:jullum@nr.no)**