Putting Data Science into Practice

-*It is NOT about the algorithms but about the data…*

Sune Askjær
Principal Data Scientist, PhD
Leap Beyond Group

Agenda

- My journey to become a Data Scientist
- Practical Data Science – Data is King!
- A use case story

LEAP|Beyond

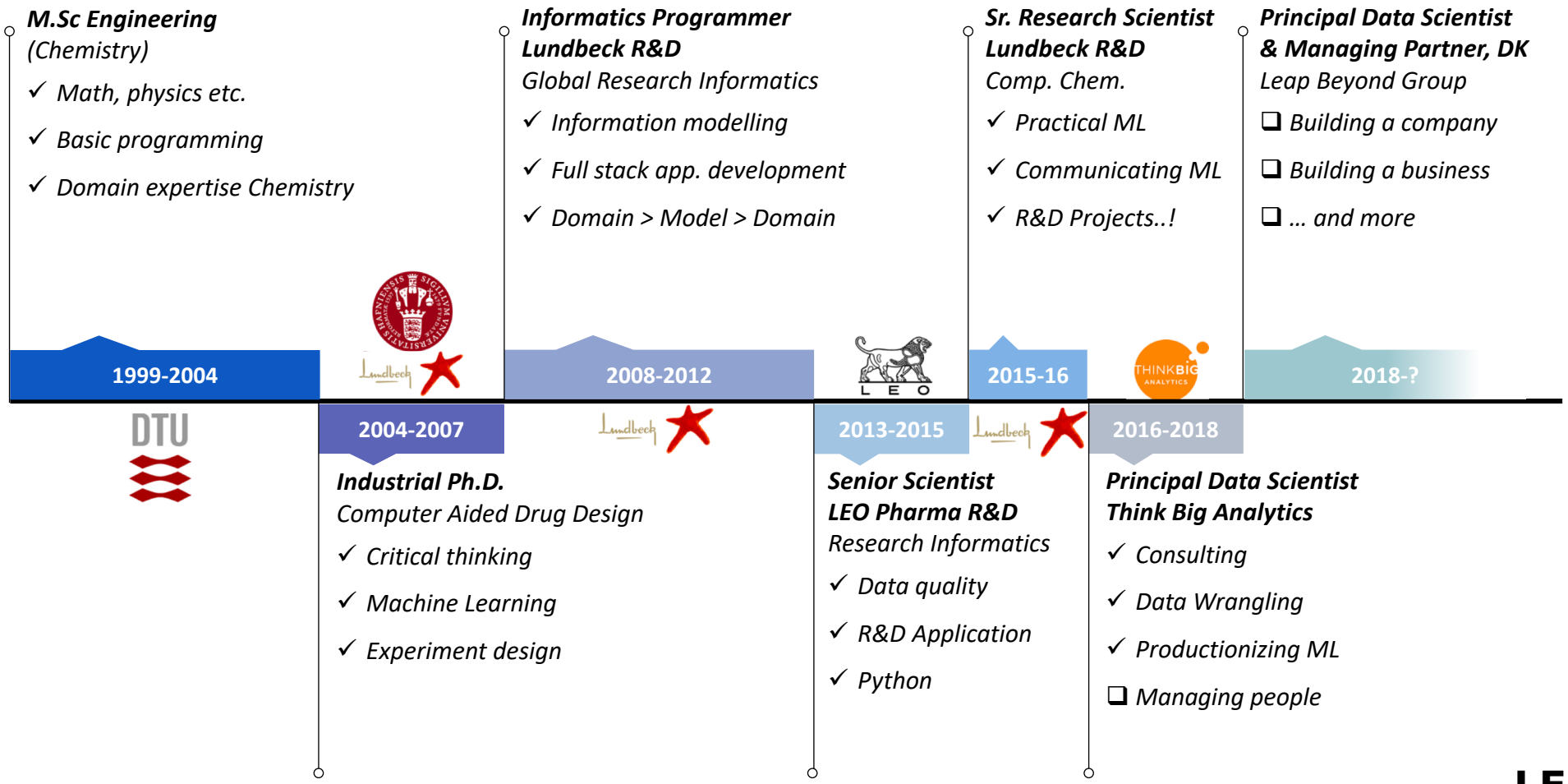
Agenda

- **My journey to become a Data Scientist**
- Practical Data Science – Data is King!
- A use case story

LEAP|Beyond

First, a bit of background…

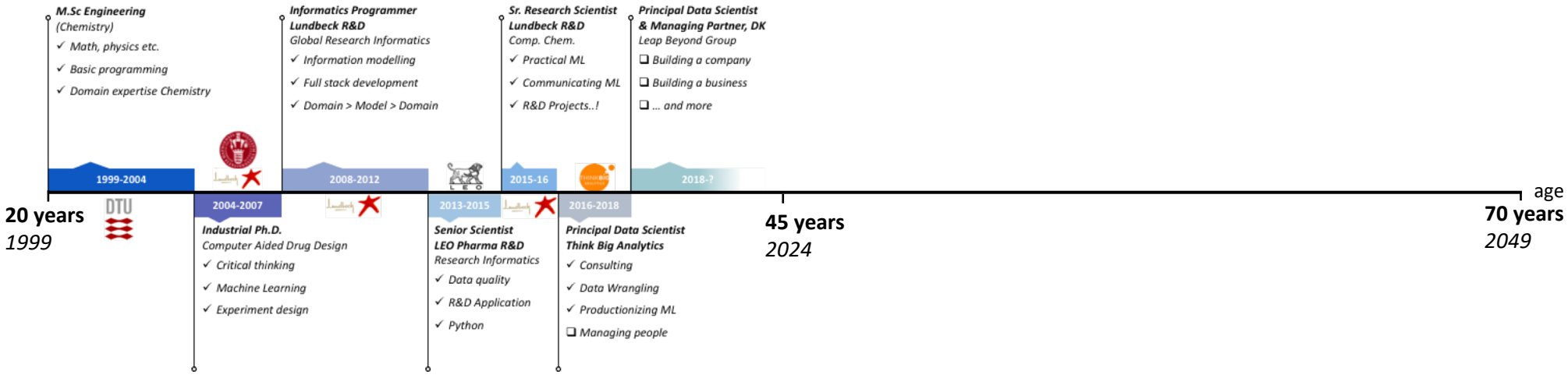
- How I became a consulting Data Scientist



LEAP|Beyond

The race is long…

- Expect a 50 year long journey



LEAP|Beyond

Agenda

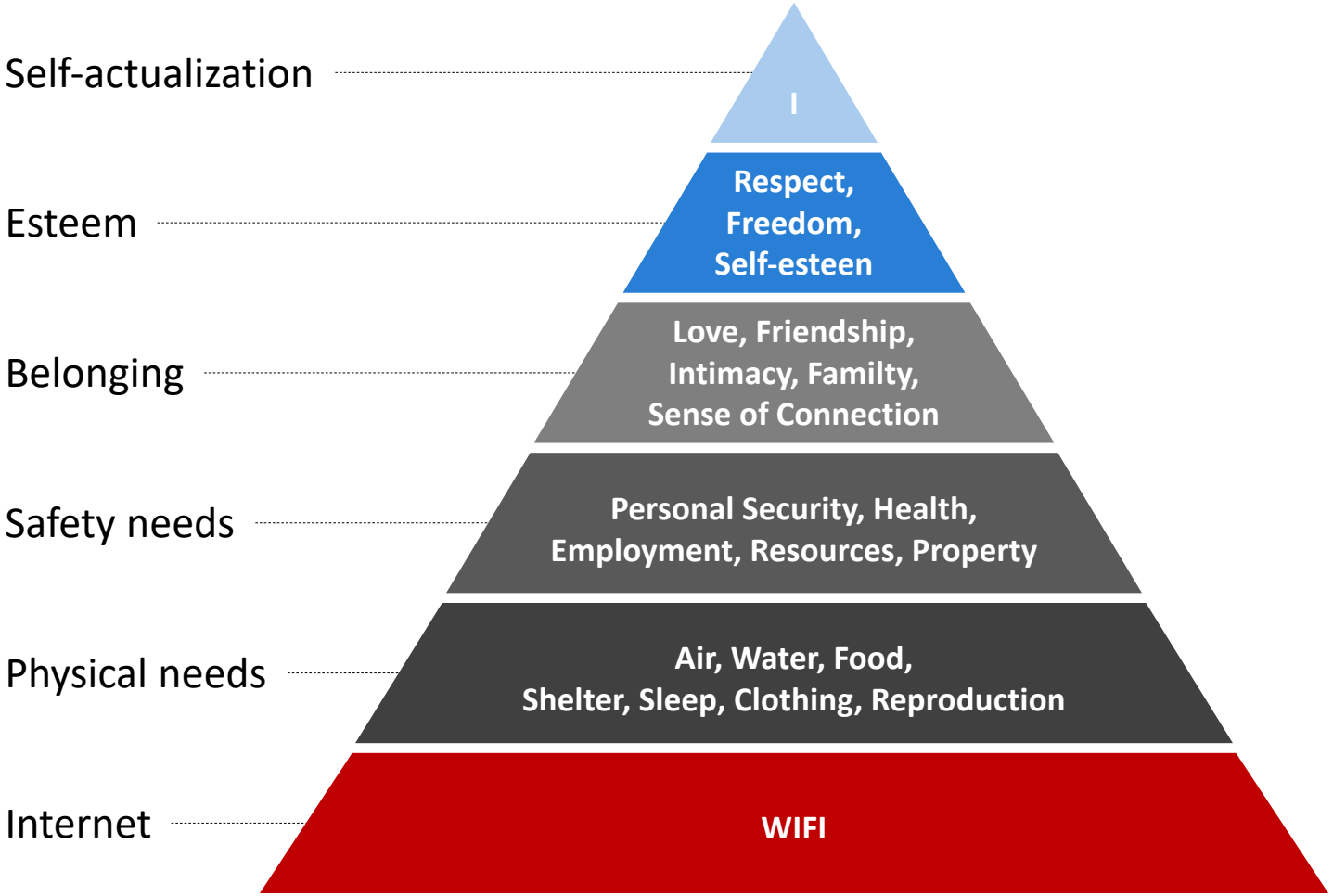
- My journey to become a Data Scientist
- **Practical Data Science – Data is King!**
- A use case story

LEAP|Beyond

Maslow’s heirarchy for needs:

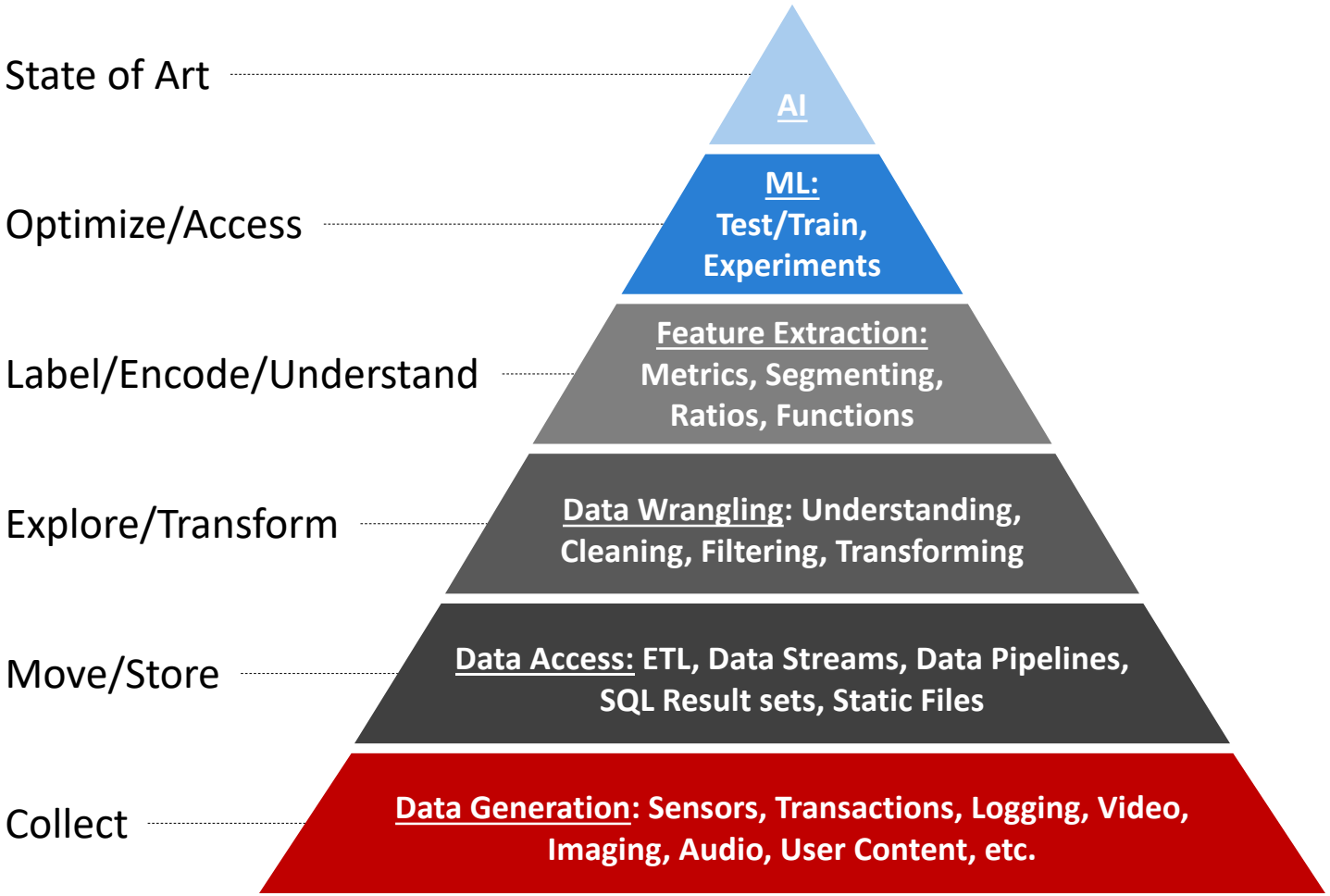
LEAP|Beyond

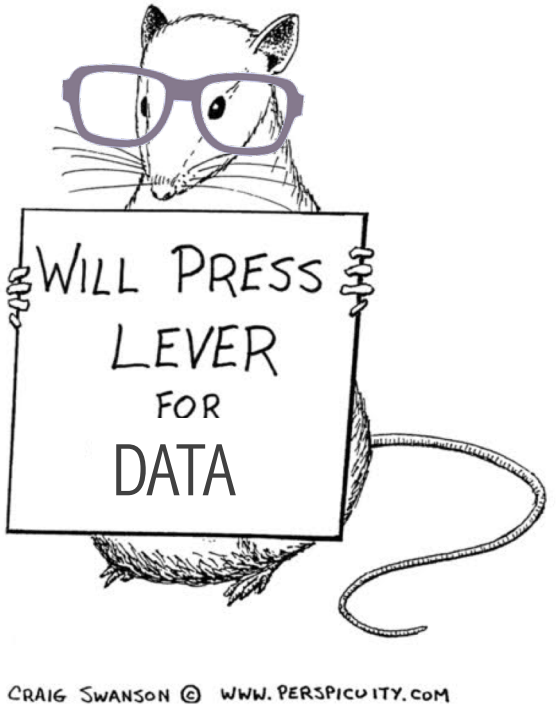
Maslow’s heirarchy for needs:



LEAP|Beyond

Data Science heirarchy for needs:





LEAP|Beyond

Yes, we have data!

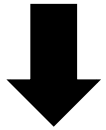
ID	F1	F2	F3	F4	F5	F6	F7	F8	L
R1	2	< 2	B	3.2	42		2	43.54	0
R2	4	< 2	A	5.2	58			42.87	
R3	5	2 - 4	A	4,8	74	5	999	44.12	1
R3	5	2 - 4	A	4,8	74		999	44.12	1
R4		8 - 12	H	4.3	64		2	43.78	0
	7	5 - 7	B	4.0	22		999	P	1
L	2	2 - 4	A		79		999	44.04	1
R6	0	> 12	G	3.9	"47		2	254843	0
…	…	…	…	…	…	…	…	…	…
Rn	5	8 - 12	A	5.3	52	5	2	42.77	0

LEAP|Beyond

Data Wrangling/Preprocessing:

- ✓ Removing duplicates
- ✓ Imputing missing values
- ✓ Imputing missing labels
- ✓ Labelling/Encoding ranges
- ✓ Labelling categories
- ✓ One-Hot Encoding
- ✓ Investigating anomalies
- ✓ Reducing cardinality
- ✓ Correcting mis-formatting
- ✓ Type casting
- ✓ Remove sparse columns
- ✓ Contamination of target variable
- ✓ Recode extreme values
- ✓ Deduce high covariance
- ✓ Correcting example labels
- ❑ Feature extraction

ID	F1	F2	F3	F4	F5	F6	F7	F8	L
R1	2	< 2	B	3.2	42		2	43.54	0
R2	4	< 2	A	5.2	58			42.87	
R3	5	2 - 4	A	4,8	74	5	999	44.12	1
R3	5	2 - 4	A	4,8	74		999	44.12	1
R4		8 - 12	H	4.3	64		2	43.78	0
	7	5 - 7	B	4.0	22		999	P	1
L	2	2 - 4	A		79		999	44.04	1
R6	0	> 12	G	3.9	"47		2	254843	0
…	…	…	…	…	…	…	…	…	…
Rn	5	8 - 12	A	5.3	52	5	2	42.77	0



ID	F1	F2	F3_A	F3_B	F3_O	F4	F5	F8	L
R1	2	1	0	1	0	3.2	42	43.54	0
R3	5	2	1	0	0	4.8	74	44.12	1
R4	5	4	0	0	1	4.3	64	43.78	0
?1	7	3	0	1	0	4.0	22	42.77	1
L	2	2	1	0	0	4.4	79	44.04	1
R6	0	5	0	0	1	3.9	47	45	0
…	…	…	…	…	…	…	…	…	…
Rn	5	4	1	0	0	5.3	52	42.77	0

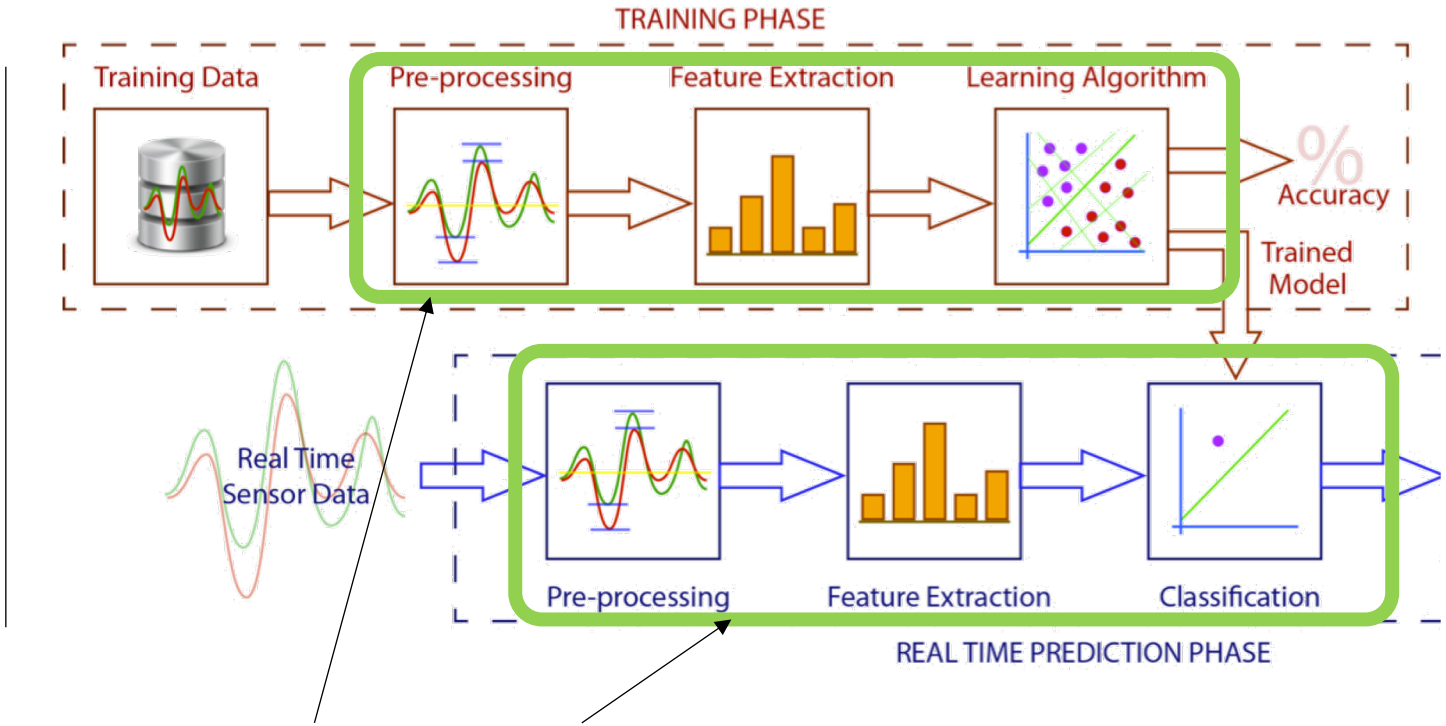
LEAP|Beyond

”*Train the way you fight*” – also in Machine Learning

Versus

LEAP|Beyond

Machine Learning Pipelines to the Rescue:



Same ML Pipeline used for training and production

LEAP|Beyond

Agenda

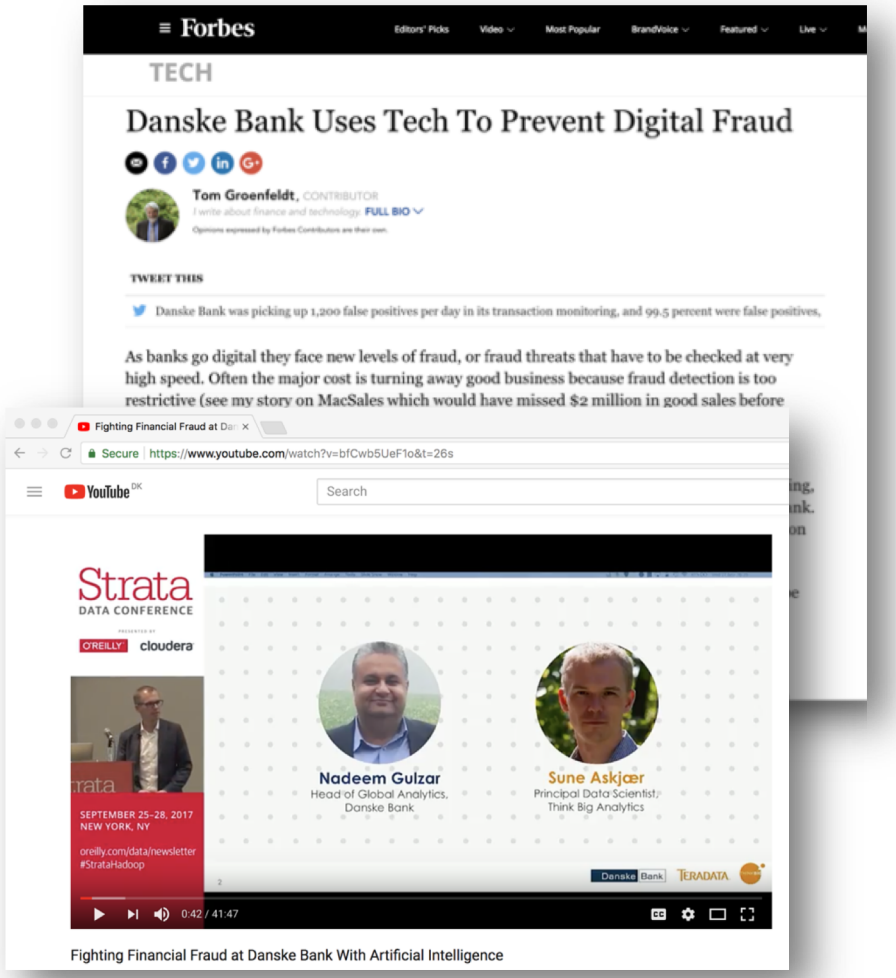
- My journey to become a Data Scientist
- Practical Data Science – Data is King!
- **A use case story**

LEAP|Beyond

Fraud Detection use case with Danske Bank

- *Note: Project done while at Think Big Analytics*

- ✓ Creating a >100m row training set
- ✓ Building ML pipeline
- ✓ Building production environment
- ✓ Deploying pipeline in production
- ✓ Due diligence and quality assurance
- ✓ Going beyond the traditional ML



LEAP|Beyond

Data Driven Approach to Fight Fraud

Challenges for Fraud Detection

Low Detection Rate
ONLY ~40%
of fraud cases are detected

Many false positives
99.5%
of cases are not fraud related

High Fraud Loss
Tens of Millions
€ lost each month

Fast evolving fraud sophistication

Ambitions for Fraud Project

- Danske Bank advanced analytics blueprint
- Reduce false-positives & Enhance fraud detection rate
- Data driven approach to real time scoring of transactions

Danske Bank TERADATA THINKBIG

Banking Anti-Fraud Solution

By leveraging the power of a thoughtful and strong data and analytics strategy, we unleash high impact business outcomes.

Data Modelling, Pipeline and Ingestion

- Organisation and silos of data
- Real-time data integration
- Security and Procedures: following existing bank procedures

Model Management Framework

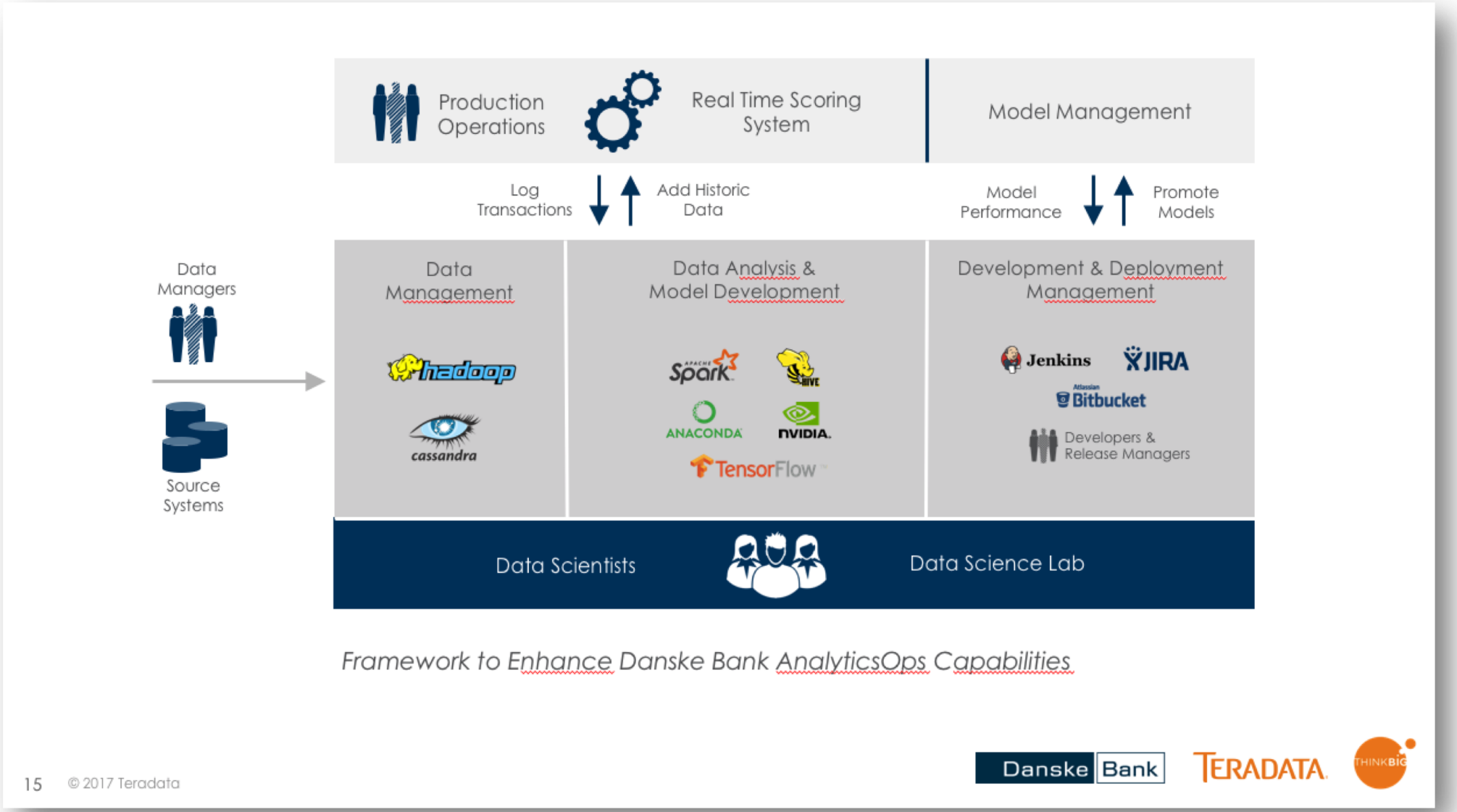
- Multiple models running in production at the same time
- Mix of traditional and advance deep learning methods
- AnalyticsOps: Deploying machine learning models in production

Machine Learning and Artificial Intelligence

- Hard to operationalize insights
- Availability of analytic capabilities/ skills and data
- Interpreting the results of machine learning models

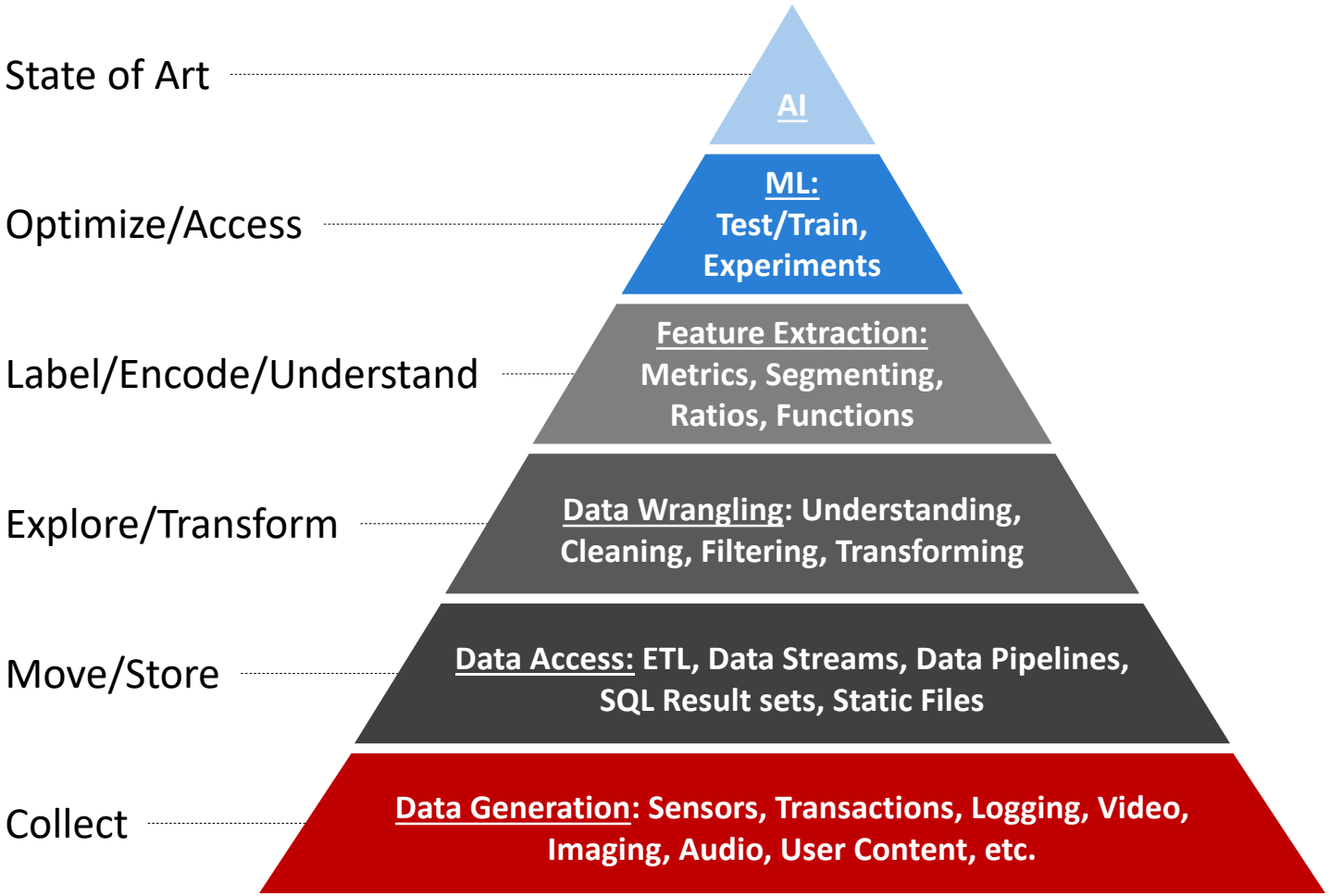
Danske Bank TERADATA THINKBIG

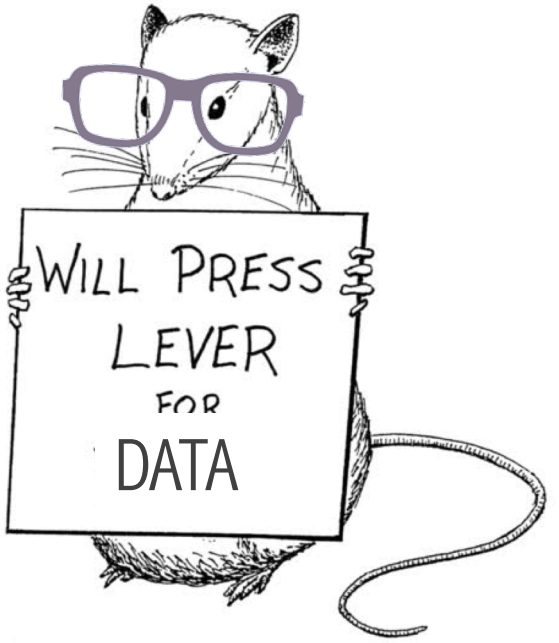
Slide from public presentation @ www.slideshare.net/RonBodkin/fighting-financial-fraud-at-danske-bank-with-artificial-intelligence



Slide from public presentation @ www.slideshare.net/RonBodkin/fighting-financial-fraud-at-danske-bank-with-artificial-intelligence

Data Science heirarchy for needs:





LEAP|Beyond

Deep Learning Opportunity

Current models can only catch ~70% of all fraud cases

Traditional ML models view transactions atomically

Often missed fraud transactions are part of a series

Capturing correlation across many features

Danske Bank TERADATA THINKBIG

LEAP|Beyond

Slide from public presentation @ www.slideshare.net/RonBodkin/fighting-financial-fraud-at-danske-bank-with-artificial-intelligence

Three Deep Learning Architectures to Deliver Value

ConvNet

- Designed for spatial correlated features, but by transforming transactions into a 2D image, we can learn temporal correlated features.
- Deeper ConvNet allows learning more complex & general features.

Goal: Learn kernels from temporal & static features to gain insight into the characteristics of fraud.

LSTM

- Learn temporal information and classify if the sequence of transactions contains fraud.
- Shares knowledge across learning time.

Goal: Learn transaction patterns within a window. Two solutions can be tested: flag fraud or predict next transaction and define an error.

Auto-Encoders

- Learn how to generate normal transactions, potentially large volumes of non-fraud data.
- AE provide a low level representation of the data.

Goal: Build a model that learns how to generate non-fraud data. To detect fraud, define a reconstruction error rate for the fraud cases

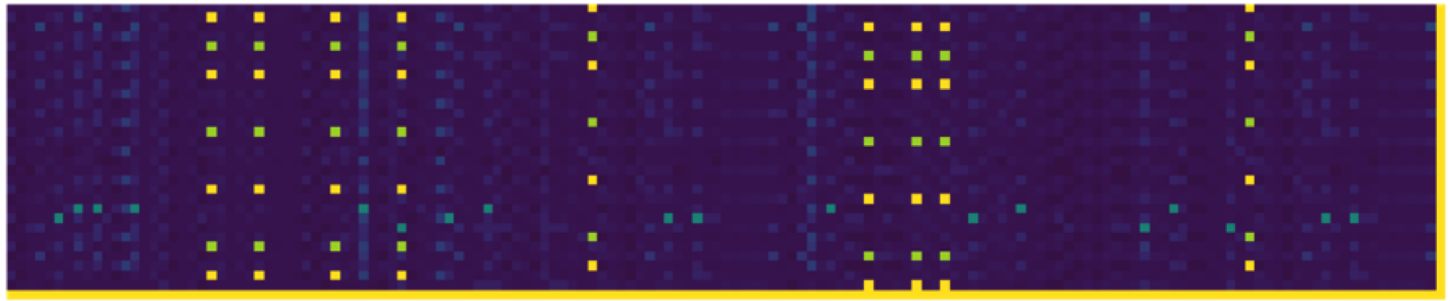
Danske Bank TERADATA THINKBIG

LEAP|Beyond

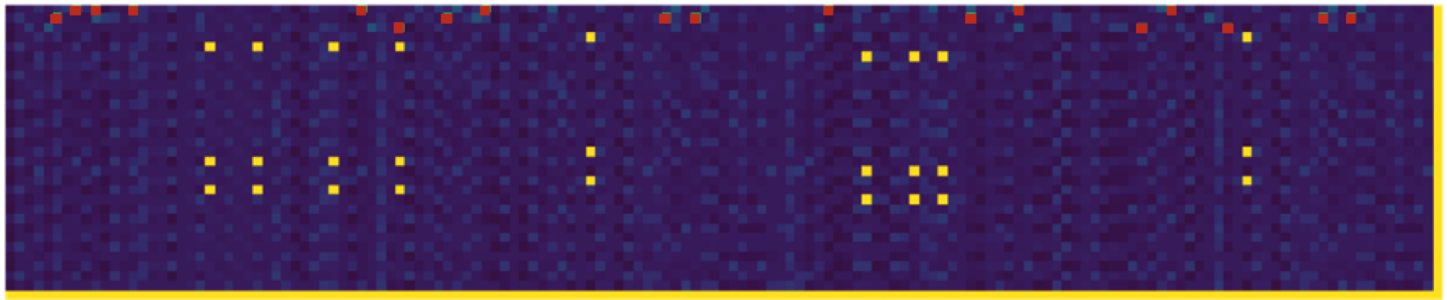
Slide from public presentation @ www.slideshare.net/RonBodkin/fighting-financial-fraud-at-danske-bank-with-artificial-intelligence

2D Transaction Image Example

Non-fraud Transaction Image



Fraud Transaction Image

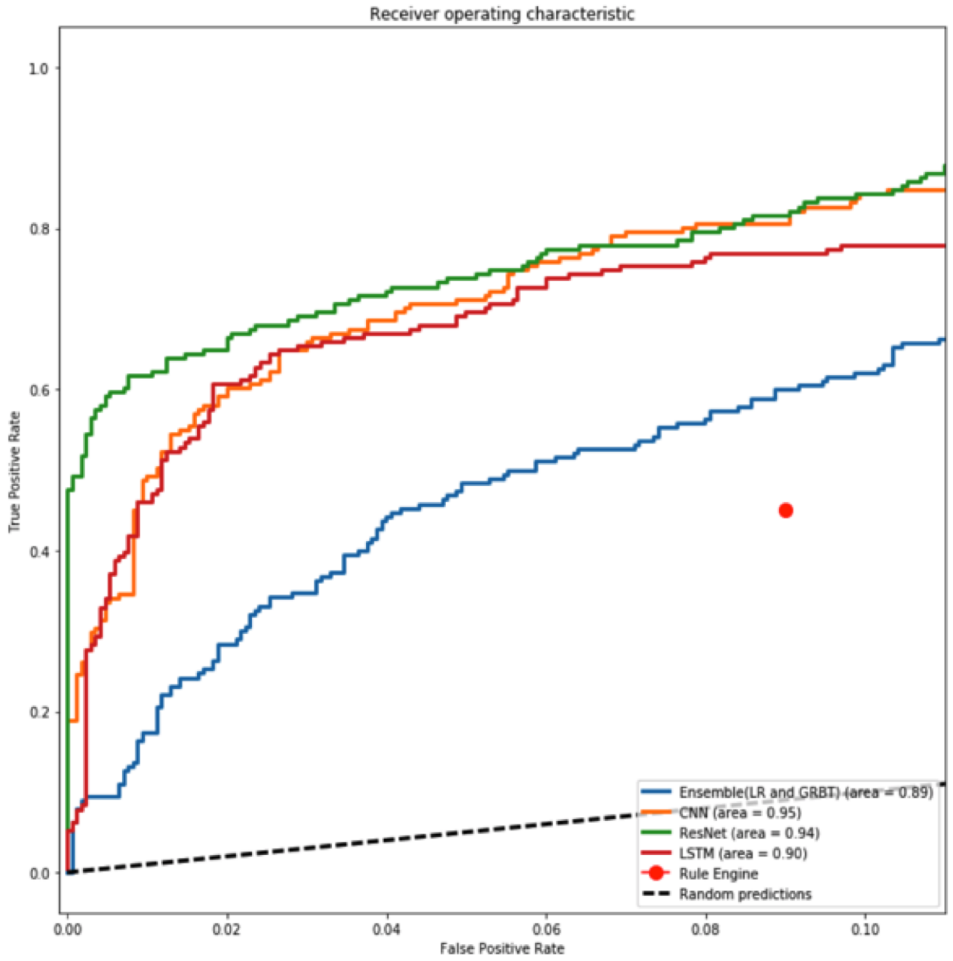


X-axis: features, Y-axis: time

Danske Bank TERADATA THINKBIG

LEAP|Beyond

Slide from public presentation @ www.slideshare.net/RonBodkin/fighting-financial-fraud-at-danske-bank-with-artificial-intelligence



Deep Learning First Results

on the fraud verification dataset

Comparison of the three deep learning models and the traditional machine learning ensemble model.

- Ensemble model (AUC 0.89)
- LSTM (AUC 0.90)
- ResNet (AUC 0.94)
- ConvNets (AUC 0.95)

Danske Bank TERADATA THINKBIG

LEAP|Beyond

Slide from public presentation @ www.slideshare.net/RonBodkin/fighting-financial-fraud-at-danske-bank-with-artificial-intelligence

LEAP|Beyond