

# Real life use cases for Machine Learning

Heli Helskyaho

ITOUG Tech Days 2019

# Introduction, Heli

- \* Graduated from University of Helsinki (Master of Science, computer science), currently a doctoral student, researcher and lecturer (databases, Big Data, Multi-model Databases, methods and tools for utilizing semi-structured data for decision making) at University of Helsinki
- \* Worked with Oracle products since 1993, worked for IT since 1990
- \* Data and Database!
- \* CEO for Miracle Finland Oy
- \* Oracle ACE Director
- \* Ambassador for EOUC (EMEA Oracle Users Group Community)
- \* Listed as one of the TOP 100 influences on IT sector in Finland (2015, 2016, 2017, 2018)
- \* Public speaker and an author
- \* Winner of Devvy for Database Design Category, 2015
- \* Author of the book Oracle SQL Developer Data Modeler for Database Design Mastery (Oracle Press, 2015), co-author for Real World SQL and PL/SQL: Advice from the Experts (Oracle Press, 2016)



# Oracle SQL Developer Data Modeler for Database Design Mastery

Design, Deploy, and Maintain World-Class Databases  
on Any Platform

**Heli Helskyaho**  
Oracle ACE Director

Forewords by C.J. Date and Tom Kyte



# Real World SQL & PL/SQL

Advice from the Experts

Arup Nanda  
Brendan Tierney  
Heli Helskyaho  
Martin Widlake  
Alex Nuijten



# 500+ Technical Experts Helping Peers Globally

**ORACLE**<sup>®</sup>  
ACE Program



**ORACLE**<sup>®</sup>  
ACE Director



**ORACLE**<sup>®</sup>  
ACE



**ORACLE**<sup>®</sup>  
ACE Associate

---

### **3 Membership Tiers**

- Oracle ACE Director
- Oracle ACE
- Oracle ACE Associate

[bit.ly/OracleACEProgram](https://bit.ly/OracleACEProgram)

### **Connect:**

✉ [oracle-ace\\_ww@oracle.com](mailto:oracle-ace_ww@oracle.com)

Facebook.com/oracleaces

@oracleace



**Oracle**  
**Groundbreakers**

Nominate yourself or someone you know: [acenomination.oracle.com](https://acenomination.oracle.com)

# What is Machine Learning?

- \* An important part of Artificial Intelligence (AI)
- \* Machine learning (ML) teaches *computers* to learn from *experience* (*algorithms*)
- \* “field of study that gives computers the ability to learn without being explicitly programmed“ -- Arthur Samuel, 1959
- \* A systematic study of algorithms and systems that improve their *knowledge* or *performance* with *experience*

# Real life use cases for ML

- \* Spam filters, Log filters/alarms
- \* Data analytics
- \* Image recognition, Speech recognition
- \* Medical diagnosis
- \* Robotics
- \* Fraud protection/detection (credit card)
- \* Product / music / movie recommendation
- \* ...

# A simple example, Chatbot

\* Demo

# Real life use cases for ML

- \* Online shopping (Amazon, Search, recommendations)
- \* Voice-to-Text, Smart Personal Assistants (mobile services: "recipe for bread", "find the nearest grocery")
  - \* Siri, Google Assistant, Alexa, Echo, Cortana,...
- \* Facebook
- \* ...



# My real life use case

- \* The face recognition (demo)

# An example of something more complicated

# Example Facebook, References

Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong, Xiaodong Wang, “Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective”, Facebook, Inc.

X. He, J. Pan, O. Jun, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Quinero Candela, “Practical lessons from predicting clicks on ads at facebook,” in Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, ser. ADKDD’14. New York, NY, USA: ACM, 2014, pp. 5:1–5:9.

J. Dunn, “Introducing FBLeaRner flow: Facebook’s AI backbone,” May 2016, <https://fb.me/dunn2016>.

<https://code.facebook.com/posts/1072626246134461/introducing-fblearner-flow-facebook-s-ai-backbone/>

# Facebook's mission

- \* “Give people the power to build community and bring the world closer together.”
- \* Facebook connects more than two billion people as of December 2017
  - \* Could not be done without ML
  - \* *The massive amount of data* required by machine learning services presents challenges to Facebook's datacenters.
  - \* Several techniques are used to efficiently feed data to the models including *decoupling of data feed and training, data/compute co-location, and networking optimizations.*
  - \* *Disaster recovery planning* is essential
  - \* actively evaluating and prototyping *new hardware solutions* while remaining cognizant of game changing *algorithmic innovations*

# Facebook, some use cases for ML, the Products

- \* News Feed ranking
- \* Ads
- \* Search
- \* Sigma
- \* Lumos
- \* Facer
- \* Language Translation
- \* Speech Recognition

# News Feed

- \* ML is used for
  - \* ranking and personalizing News Feed stories
  - \* filtering out offensive content
  - \* highlighting trending topics
  - \* ranking search results, and much more.
- \* *General models are trained* to determine various user and environmental factors that should ultimately determine the rank order of content.
- \* *The model is used to generate a personalized set of the best posts, images, and other content to display from thousands of candidates, and the best ordering of this chosen content.*

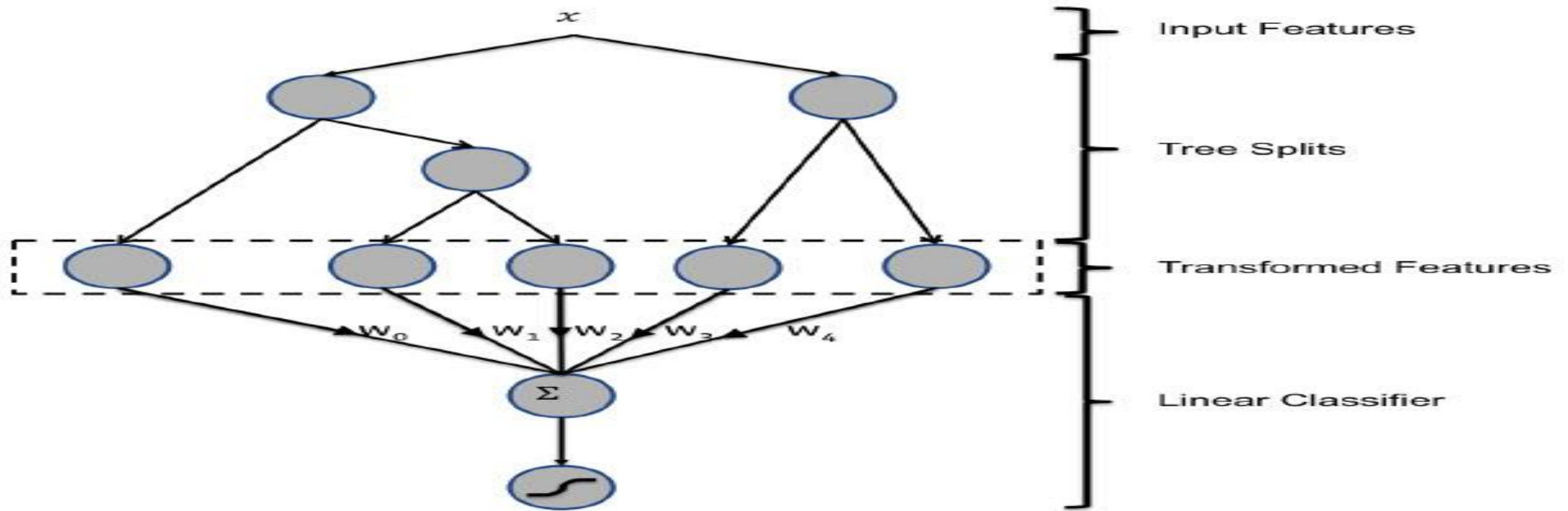
# Ads

- \* Online advertising allows advertisers to only bid and pay for measurable user responses, such as clicks on ads.
  - \* As a consequence, click prediction systems are *central* to most online advertising systems.
- \* General Ads models are trained to learn *how user traits, user context, previous interactions, and advertisement attributes* can be most predictive of the likelihood of clicking on an ad, visiting a website, and/or purchasing a product.
- \* Inputs are run through a trained model to immediately determine which ads to display to a particular Facebook user.

# Predicting the Clicks

- \* The click prediction system needs to be *robust and adaptive*, and capable of learning from *massive* volumes of data.
- \* At Facebook they use a model which *combines decision trees with logistic regression*
- \* Based on their experience: the most important thing is to have *the right features* (those capturing historical information about the user or ad dominate other types of features) and *the right model*
- \* Measures: the accuracy of prediction





**Figure 1: Hybrid model structure. Input features are transformed by means of boosted decision trees. The output of each individual tree is treated as a categorical input feature to a sparse linear classifier. Boosted decision trees prove to be very powerful feature transforms.**

X. He, J. Pan, O. Jun, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Quinonero Candela, "Practical lessons from predicting clicks on ads at facebook," in Proceedings of the Eighth International Workshop on Data Mining for Online Advertising, ser. ADKDD'14. New York, NY, USA: ACM, 2014, pp. 5:1–5:9.

# Search

- \* Launches a *series* of distinct and specialized *sub-searches* to the various verticals, e.g., videos, photos, people, events, etc.
- \* A *classifier* layer is run atop the various search verticals to *predict which of the many verticals to search* (searching all possible verticals would be inefficient)
- \* The classifier and these search verticals consist of
  - \* an *offline* stage to *train* the models
  - \* and an *online* stage to *run the models* and perform the classification and search

# Sigma

- \* General *classification and anomaly detection framework* that is used for a variety of internal applications (site integrity, spam detection, payments, registration, unauthorized employee access, and event recommendations)
- \* Sigma includes *hundreds of distinct models running in production everyday*
  - \* each model is trained to detect anomalies (e.g. classify content)

# Lumos

- \* Extract high-level *attributes* and *embeddings* from *an image* and its *content*
  - \* That data can be *used as input* to other products and services
    - \* for example as it were text.

# Facer

- \* Facebook's *face detection and recognition framework*
- \* Given an image
  - \* *finds* all of the faces in that image
  - \* *runs a user-specific* facial-recognition algorithm to determine the likelihood of that face belonging to one of your top-N friends who have enabled face recognition
- \* This allows Facebook to suggest which of your friends you might want to tag within the photos you upload.

# Language Translation

- \* Service that manages *internationalization* of Facebook content
- \* Supports *translations* for more than 45 languages (as the source or target language)
  - \* supports more than 2000 translation directions
  - \* serves 4.5B translated post impressions every day
- \* Each language pair direction has its own model
  - \* multi-language models are being considered

# Speech Recognition

- \* Converts *audio streams into text*
- \* Provides automated captioning for video
- \* Most streams are English language
  - \* other languages will be available in future
- \* Additionally, non-language audio events are also detected with a similar system (simpler model).

# Algorithms Facebook uses for these services

Models	Services
Support Vector Machines (SVM)	Facer (User Matching)
Gradient Boosted Decision Trees (GBDT)	Sigma
Multi-Layer Perceptron (MLP)	Ads, News Feed, Search, Sigma
Convolutional Neural Networks (CNN)	Lumos, Facer (Feature Extraction)
Recurrent Neural Networks (RNN)	Text Understanding, Translation, Speech Recognition

TABLE I

MACHINE LEARNING ALGORITHMS LEVERAGED BY PRODUCT/SERVICE.

\* Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective,

Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong, Xiaodong Wang  
Facebook, Inc.



# How do they do all this at Facebook?

# FBLearner Platform

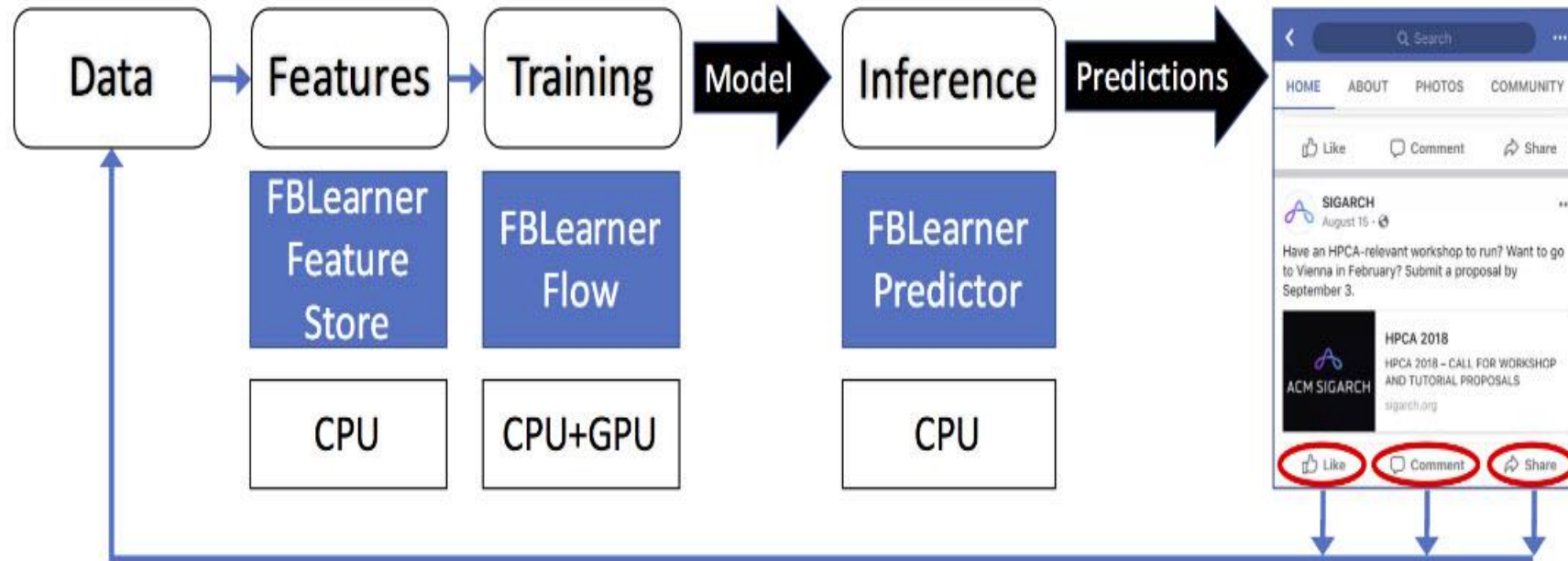


Fig. 1. Example of Facebook's Machine Learning Flow and Infrastructure.

\* Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective, Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong, Xiaodong Wang Facebook, Inc.

# FBLearner Feature Store

- \* The starting point for a ML modeling task is to *gather and generate features*.
- \* The Feature Store is a *catalog* of several feature generators
  - \* can be used both for training and real-time prediction
  - \* serves as a marketplace that multiple teams can use to share and discover features

# FBLearner Flow

- \* Facebook's machine learning *platform for model training*
- \* **Workflows:** A workflow is a single pipeline defined within FBLearner Flow and is the entry point for all machine learning tasks.
  - \* Each workflow performs a specific job, such as training and evaluation of a specific model.
  - \* Workflows are defined in terms of operators and can be parallelized.
- \* **Operators:** Operators are the building blocks of workflows
  - \* In FBLearner Flow, operators are the smallest unit of execution and run on a single machine.
- \* **Channels:** Channels represent inputs and outputs, which flow between operators within a workflow.
  - \* All channels are typed using a custom type system.
- \* Flow has tooling for experiment management.
- \* The user interface keeps track of all of the artifacts and metrics generated by each workflow execution or experiment.
  - \* The user interface makes it simple to compare and manage these experiments.

# FBLearner Flow

- \* The platform consists of three core components:
  - \* *an authorship and execution environment for custom distributed workflows*
  - \* *an experimentation management UI for launching experiments and viewing results*
  - \* *numerous predefined pipelines for training the most commonly used machine learning algorithms at Facebook.*

# FBLearner Predictor

- \* Facebook's *internal inference engine* that uses the models trained in FBLearner Flow to provide predictions in real time.
  - \* Can be used
    - \* as a multitenancy service
    - \* or as a library that can be integrated in product specific backend services
  - \* Is used by multiple product teams at Facebook, many of which require low latency solutions.
- \* The direct integration between Flow and Predictor also helps with
  - \* running online experiments
  - \* managing multiple versions of models in productions

# Frameworks for deep learning

- \* Two distinct but synergistic frameworks for deep learning at Facebook:
  - \* PyTorch, which is optimized for *research*
  - \* Caffe2, which is optimized for *production*

# PyTorch

- \* PyTorch is the framework for AI *research* at Facebook which enables rapid experimentation
  - \* Flexibility
  - \* Debugging
  - \* Dynamic neural networks
- \* Not optimized for production and mobile deployments (Python)
- \* When research projects produce valuable results, *the models need to be transferred to production.*
  - \* Traditionally, rewriting the training pipeline in a product environment with other frameworks.



# Caffe2

- \* Facebook's in-house *production* framework
  - \* For training and deploying large-scale machine learning models
- \* Focuses on several key features required by products:
  - \* Performance
  - \* cross-platform support
  - \* coverage for fundamental machine learning algorithms (convolutional neural networks (CNNs), recurrent networks (RNNs), and multi-layer perceptrons (MLPs)) and up to tens of billions of parameters

# Open Neural Network Exchange, ONNX

- \* Different tools are better for different subset of problems and have varying tradeoffs on flexibility, performance, and supported platforms . As a result, there should be a way to *exchange trained models between different frameworks or platforms*.
- \* ONNX is a format to represent deep learning models in *a standard way* to enable interoperability across different frameworks and vendor-optimized libraries.
- \* ONNX is designed as an *open specification*
- \* Within Facebook, ONNX is used for transferring research models from the PyTorch environment to high-performance production environment in Caffe2.
  - \* ONNX provides the ability to automatically capture and translate static parts of the models.
  - \* An additional toolchain facilitates transfer of dynamic graph parts from Python by either mapping them to control-flow primitives in Caffe2 or reimplementing them in C++ as custom operators.

# Caffe2 and PyTorch projects are merging

Caffe2 and PyTorch join forces to create a Research + Production platform  
PyTorch 1.0:

[https://caffe2.ai/blog/2018/05/02/Caffe2\\_PyTorch\\_1\\_0.html](https://caffe2.ai/blog/2018/05/02/Caffe2_PyTorch_1_0.html)

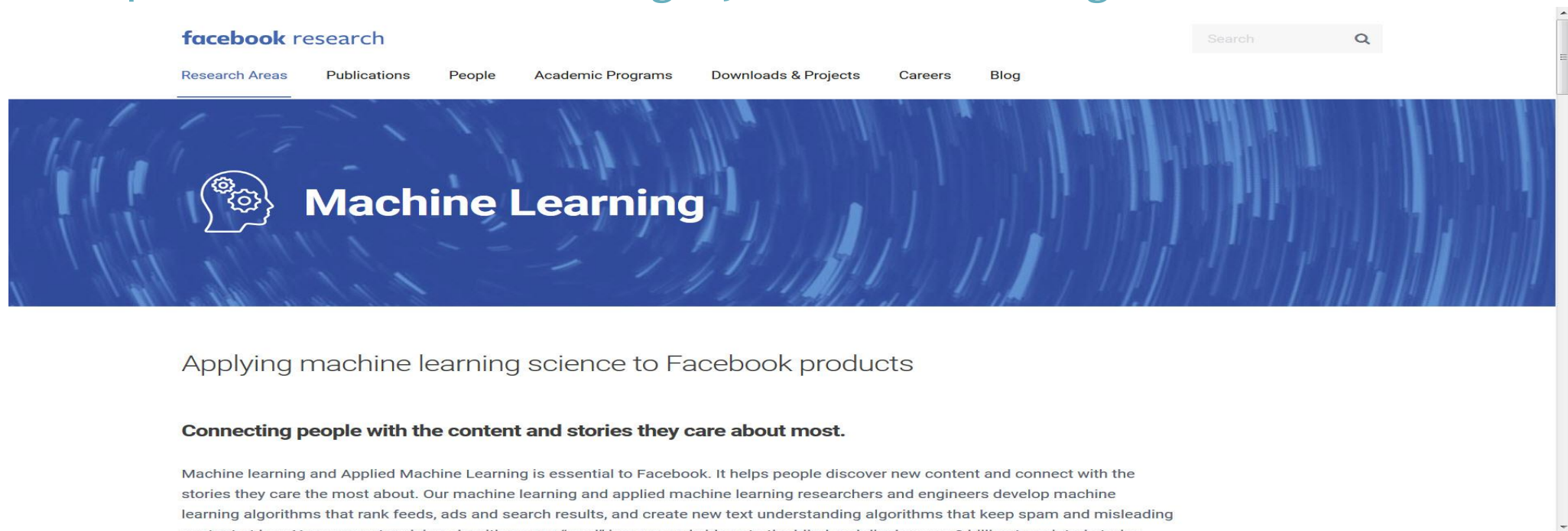
“We realized that in order to deliver the best user experience, it makes sense to combine the beneficial traits of Caffe2 and PyTorch into a single package and enable a smooth transition from fast prototyping to fast execution. It’d also improve our developer efficiency by more easily utilizing a shared set of tools.”

# The success factors 1/2

- \* success is predicated on the availability of extensive, high-quality **data**
- \* complex preprocessing logic is applied to ensure that *data is cleaned and normalized* to allow efficient transfer and easy learning
- \* The *ability to rapidly process* and feed these data to the training machines is important for ensuring that we have fast and efficient offline training.
- \* These impose very high resource requirement especially on storage, network, and CPU.
- \* actively evaluating and prototyping new hardware solutions while remaining cognizant of game changing algorithmic innovations

# Facebook, research

\* <https://research.fb.com/category/machine-learning/>



The screenshot shows the Facebook Research website. At the top, there is a navigation bar with the text "facebook research" and a search box. Below the navigation bar, there are several menu items: "Research Areas", "Publications", "People", "Academic Programs", "Downloads & Projects", "Careers", and "Blog". The main content area features a large blue banner with a white icon of a head with gears and the text "Machine Learning". Below the banner, there is a paragraph of text: "Applying machine learning science to Facebook products". This is followed by a bolded sub-heading: "Connecting people with the content and stories they care about most." and a paragraph of text: "Machine learning and Applied Machine Learning is essential to Facebook. It helps people discover new content and connect with the stories they care the most about. Our machine learning and applied machine learning researchers and engineers develop machine learning algorithms that rank feeds, ads and search results, and create new text understanding algorithms that keep spam and misleading content at bay. New computer vision algorithms use 'food' images and videos to the blind and display over 2 billion translated stories..."

# The success factors 2/2

- \* *Knowing what to measure to know what to improve*

# What to measure?

- \* Number of positives, number of negatives, number of true positives, number of false positives, number of true negatives, number of false negatives
- \* Portion of positives, portion of negatives
- \* Class ratio
- \* Accuracy, Error rate
- \* ROC curve, coverage curve,
- \* ...
- \* It all depends on how we define the performance for the answer to our question (experiment): *experimental objective*

# Facebook

- \* “we noticed that the largest improvements in accuracy often came from *quick experiments, feature engineering, and model tuning* rather than applying fundamentally different algorithms”
- \* An engineer may need to attempt hundreds of experiments before finding a successful new feature or set of hyperparameters.



# Oracle SQL Developer, Data Miner

- \* Oracle SQL Developer is a free tool from Oracle
- \* Has an add-on called Data Miner
- \* *Advanced analytics* (Data Miner uses that) is a **licensed product** (in the EE database separately licensed, in the Cloud: Database Service either High Performance Package or Extreme Performance Package)

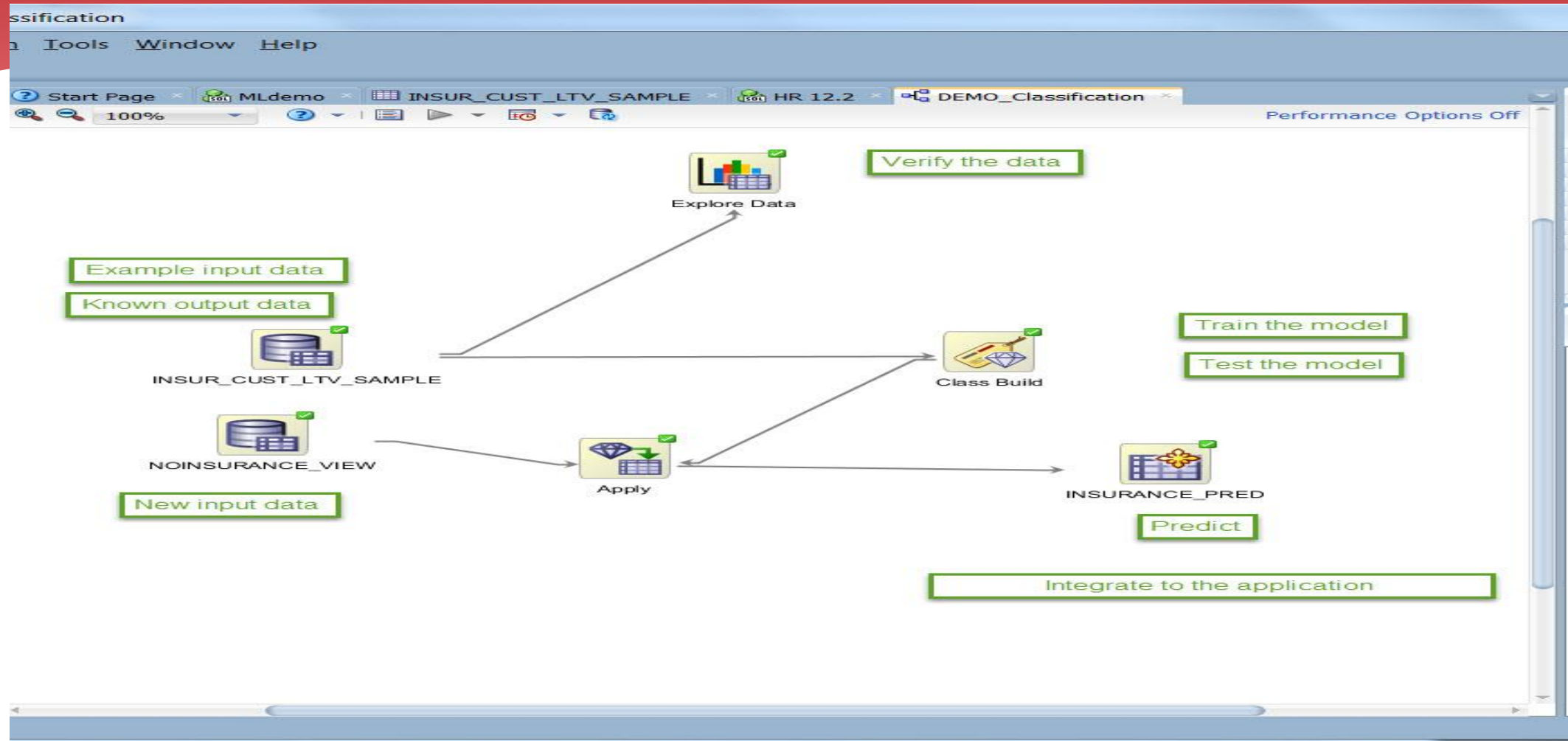
- \* Oracle Data Miner GUI Installation Instructions

<http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/odmrinstallation-2080768.html>

- \* Tutorial

<http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/12c/BigDataDM/ODM12c-BDL4.html>

# Oracle SQL Developer demo



# Chapter 10



## **Real World SQL & PL/SQL**

Advice from the Experts

Arup Nanda  
Brendan Tierney  
Heli Helskyaho  
Martin Widlake  
Alex Nuijten

*Oracle  
Press*

# Oracle R Enterprise

- \* a component of the Oracle Advanced Analytics Option (payable option)
- \* open source R statistical programming language in an Oracle database

# Chapter 11



## **Real World SQL & PL/SQL**

Advice from the Experts

Arup Nanda  
Brendan Tierney  
Heli Helskyaho  
Martin Widlake  
Alex Nuijten

*Oracle  
Press*

# Predictive Queries in Oracle 12c

- \* Predictive Queries enable you to build and score data quickly using the in-database data mining algorithms
- \* Predictive Queries can be
  - \* built using Oracle Data Miner
  - \* written using SQL

# Chapter 12



## **Real World SQL & PL/SQL**

Advice from the Experts

Arup Nanda  
Brendan Tierney  
Heli Helskyaho  
Martin Widlake  
Alex Nuijten

*Oracle  
Press*

# SQL, Demo (Oracle Autonomous DW)

## MyFirstNotebook



```
%script
BEGIN
  DBMS_DATA_MINING.DROP_MODEL(model_name => 'StudentEnrollment');
END;
/
```

FINISHED

PL/SQL procedure successfully completed.

Took 3 sec. Last updated by HELI at January 28 2019, 3:05:23 PM.

```
-- DEMO
-- dataset Studentenrolment

--- Preparations
-- create training set
CREATE TABLE Student_training_data
AS SELECT * FROM StudentEnrollment
WHERE ORA_HASH (Student_ID, 99, 5) < 65;
```

FINISHED

Updated 9129 row(s).

Took 1 sec. Last updated by HELI at January 23 2019, 9:12:18 PM.

```
-- create testing set
CREATE TABLE Student_testing_data
AS SELECT * FROM StudentEnrollment
WHERE ORA_HASH(Student_ID, 99, 5) >= 65;
```

FINISHED



# And so many more languages to learn...

- \* Python
  - \* R
  - \* C/C++
  - \* Java
  - \* JavaScript
  - \* Julia, Scala, Ruby, Octave, MATLAB, SAS
- 
- \* <https://medium.com/towards-data-science/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7>

# The future and now!

- \* AI and machine learning is here and it's the future
- \* So many interesting areas to learn
- \* Pick your area and **START LEARNING!**

# Conclusion

- \* ML can be used "everywhere":
  - \* Spam filters
  - \* Log filters (and alarms)
  - \* Data analytics
  - \* Image recognition
  - \* Speech recognition
  - \* Medical diagnosis
  - \* Robotics
  - \* Chatbots
  - \* ...

# Conclusion

- \* Facebook uses ML "everywhere"
  - \* News Feed ranking
  - \* Ads
  - \* Search
  - \* Sigma
  - \* Lumos
  - \* Facer
  - \* Language Translation
  - \* Speech Recognition

# Conclusion

- \* You can use ML "everywhere"
  - \* Start small and when you learn more do more
  - \* Define a Task and let ML solve it
  - \* Machines are not taking our jobs but helping us to do more interesting things
  - \* With ML we can understand our data better and make better decisions

# THANK YOU!

QUESTIONS?

Email: [heli@miracleoy.fi](mailto:heli@miracleoy.fi)

Twitter: [@HeliFromFinland](https://twitter.com/HeliFromFinland)

Blog: [Helifromfinland.com](http://Helifromfinland.com)