Artificial Intelligence: Strategies For Leading Business Transformation

Resource Library

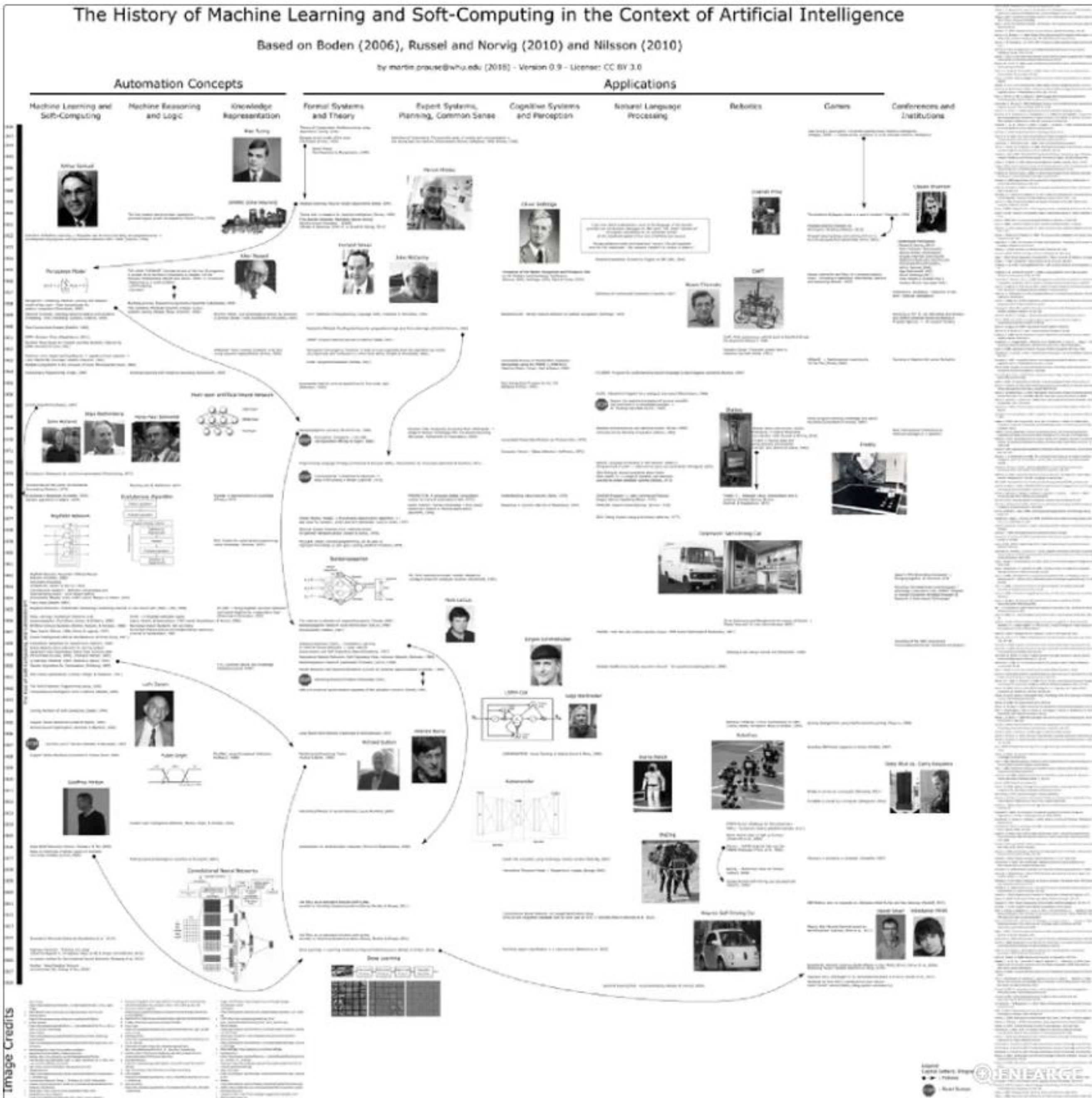
The Resource Library provides a rich and curated repository of information to help you delve deeper into specific areas of AI / ML.

- AI Trends and Landscape
- Select Algorithms and Applications
- Data Sources for Training and Validation
- Data Visualization Tools and Resources
- Modeling Software and Frameworks
- Developer Tools to Get Started
- Additional Training via Platform Companies
- Suggested Reading
- AI News and Community Sites
- A Primer on KNIME
- Select Use Cases & Frameworks in Action

Note: Links to few resources may require special access / subscriptions.

AI Trends: Looking Back

Explore the history of AI and automation dating back to the 1930s [Link]



A panel of AI experts share their insights on recent trends and future implications of AI in this report from Stanford University.

- Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller. "*Artificial Intelligence and Life in 2030.*" One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016.

This report is the first in a series to be issued at regular intervals as a part of the One Hundred Year Study on Artificial Intelligence (AI100). Starting from a charge given by the AI100 Standing Committee to consider the likely influences of AI in a typical North American city by the year 2030, the 2015 Study Panel, comprising experts in AI and other relevant areas focused their attention on eight domains they considered most salient: transportation; service robots; healthcare; education; low-resource communities; public safety and security; employment and workplace; and entertainment. In each of these domains, the report both reflects on progress in the past fifteen years and anticipates developments in the coming fifteen years. The report begins with a reflection on what constitutes Artificial Intelligence, and concludes with recommendations concerning AI-related policy. These recommendations include accruing technical expertise about AI in government and devoting more resources – and removing impediments – to research on the fairness, security, privacy, and societal impacts of AI systems.

Document: http://ai100.stanford.edu/2016-report.

The following links provide insights into a variety of business use cases and applications of artificial intelligence.

- Chui, M., Lund, S., Madgavkar, A., Mischke, J., & Ramaswamy, S. (2018). McKinsey Global Institute: Notes from the AI Frontier. Insights from Hundreds of Use Cases. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning

- McKinsey & Company: An Executive's Guide to AI. (2018). https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai

- Accenture. AI Explained: A Guide for Executives. (2018). https://www.accenture.com/_acnmedia/PDF-84/Accenture-AI-Explained-Overview.pdf

- Charlin, G., Gell, J., Bellefonds, N., Smith, T., Lui, V., Bellemare, J., Royston, J., & Sehili, C. (2018). Boston Consulting Group. Unlocking Growth in CPG with AI and Advanced Analytics. https://www.bcg.com/publications/2018/unlocking-growth-cpg-ai-advanced-analytics.aspx

The following are behind paywalls requiring subscription to access:

- Brant, K., & Austin, T. (2017). Gartner Hype Cycle for Artificial Intelligence, 2017. ID: G00314732. https://www.gartner.com/en/documents/3770467

- Sheil, B. (1987). Thinking about Artificial Intelligence, Harvard Business Review. https://hbr.org/1987/07/thinking-about-artificial-intelligence

- Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-Powered Organization, Harvard Business Review. https://hbr.org/2019/07/building-the-ai-powered-organization

AI Trends: Infrastructure Landscape

The AI landscape in constantly changing as new solutions emerge – the following list represents few of the current leaders:

Platform-as-a-Service (PaaS) providers offering end-to-end infrastructure products and services:

- Amazon Machine Learning: https://aws.amazon.com/machine-learning/
- Microsoft Machine Learning Studio and Cognitive Services:
  - https://azure.microsoft.com/en-us/services/machine-learning-studio/
  - https://azure.microsoft.com/en-us/services/cognitive-services/
- IBM Watson: https://www.ibm.com/watson/products-services
- Google Cloud AI: https://cloud.google.com/products/ai/
- Huawei HiAI: https://developer.huawei.com/consumer/en/hiai

Machine-Learning-as-a-Service (MLaaS) providers offering more specialized products and services:

- H20.ai, an open source ML platform: https://www.h2o.ai
- BigML, a ML platform that runs in the cloud or on-premise: https://bigml.com/
- Apache PredictionIO, open source ML server: http://predictionio.apache.org
- Floyd Hub, a Deep Learning specialized platform: https://www.floydhub.com/
- UC Berkeley's ML platform: http://mlbase.org/

In Module-1 we defined three machine learning problem types: supervised, unsupervised and reinforcement learning. Resources on the following pages provide deep dives into each area.

Supervised Learning

Learn with a labeled training set

Regression

Predict Values & Outcomes

What will the temperature be in Chicago tomorrow? How many units should we order next week?

Classification

Identify & Categorize

Is a transaction fraudulent? Is the negative customer feedback targeted at product, service or facilities?

Unsupervised Learning

Discover patterns in unlabeled data

Clustering

Discover Groupings

Which customer segments should I target? How should I organize my products catalog?

Dimensionality Reduction

Discover Rules/Associations

Which features explain my data's variability the best? How are the variables associated?

Reinforcement Learning

Learn based on feedback or rewards

Reinforcement

Act Based on Environment

How do machines play Chess or Go? How can robots better assist humans in a variety of tasks?

Select Algorithms – Regression

Regression models past relationships between independent variables and dependent variables to predict future values of the dependent variable.

- Regression algorithms: https://www.sciencedirect.com/science/article/pii/S0893608015001185
- Support Vector Regression: https://papers.nips.cc/paper/1238-support-vector-regression-machines
- Regression Trees: http://pages.stat.wisc.edu/~loh/treeprogs/guide/LohISI14.pdf
- Overview on timeseries regression: https://www.sciencedirect.com/topics/economics-econometrics-and-finance/arma-model

Select Algorithms – Classification

Classification techniques predict the class to which a data element belongs based on the data element's underlying attributes

- Support Vector machines: https://doi.org/10.1017/CBO9780511801389
- Naïve Bayes Classifier: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf
- Instance-based learning algorithms: https://link.springer.com/article/10.1007/BF00153759
- Random Forest for classification: https://link.springer.com/article/10.1023/A:1010933404324
- Deep Convolutional Neural Network for image classification: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks
- Recurrent neural network for sequential data: https://arxiv.org/abs/1506.00019
- Overview of deep neural networks: https://www.sciencedirect.com/science/article/pii/S0893608014002135
- Review of hierarchical classification algorithms: https://doi.org/10.2307/2981629

Select Algorithms – Clustering

Clustering algorithms discover relationships in data. Different sets of clusters may be determined from the same dataset depending on features used in models and their relative weights

- Overview of clustering methods:
  https://link.springer.com/chapter/10.1007/0-387-25465-X_15

- Review on k-means cluster algorithm:
  http://www.ijsret.org/pdf/121888.pdf

- DB-Scan clustering:
  https://dl.acm.org/citation.cfm?id=3068335

- FP-Growth clustering:
  https://dl.acm.org/citation.cfm?id=335372

- EM-Clustering:
  https://www.sciencedirect.com/science/article/pii/016794739290042E

- Self-Organizing Maps:
  https://link.springer.com/book/10.1007/978-3-642-88163-3

- Hierarchical Clustering:
  https://arxiv.org/abs/1105.0121

- Deep unsupervised clustering using autoencoders:
  https://arxiv.org/abs/1712.07788

Select Algorithms – Dimensionality Reduction

Dimension Reduction refers to the process of converting a set of data having vast dimensions (columns/features) into data with lesser dimensions ensuring that it conveys similar information concisely

- Latent Dirichlet Allocation: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- Principal Component Analysis: https://www.springer.com/gp/book/9780387954424
- Local Linear Embedding: https://link.springer.com/article/10.1007/s10462-010-9200-z
- Autoencoders: https://science.sciencemag.org/content/313/5786/504

Select Algorithms – Reinforcement Learning

Reinforcement learning algorithms take an action based on the environment, receiving a reward if the action brings the machine a step closer to maximizing the total rewards available. The algorithm optimizes for the best series of actions by correcting itself over time.

- Q-Learning:
  https://link.springer.com/article/10.1007/BF00992698

- Deep Q-Learning:
  https://arxiv.org/abs/1901.00137

- Deep Reinforcement Learning an Overview:
  https://arxiv.org/abs/1701.07274

- Dueling Network Architectures for Reinforcement Learning:
  https://arxiv.org/abs/1511.06581

- Actor-Critic Algorithms:
  https://papers.nips.cc/paper/1786-actor-critic-algorithms.pdf

- Continuous Control with Deep Reinforcement learning:
  https://arxiv.org/abs/1509.02971

Recommender Systems:

- Deep Learning based recommendation analysis:
  https://arxiv.org/pdf/1707.07435.pdf

- A review on recommender systems:
  https://arxiv.org/pdf/1811.11866.pdf

- Fast mining of association rules (Apriori Algorithm):
  http://www.cse.msu.edu/~cse960/Papers/MiningAssoc-AgrawalAS-VLDB94.pdf

Language Processing:

- Review on sentiment analysis:
  https://www.sciencedirect.com/science/article/pii/S0020025515002054

- Sentiment Analysis
  https://thesai.org/Publications/ViewPaper?Volume=8&Issue=6&Code=IJACSA&SerialNo=57

- Deep learning in machine translation:
  https://link.springer.com/chapter/10.1007/978-981-10-5209-5_6

- Trends in deep learning based natural language processing:
  https://arxiv.org/pdf/1708.02709.pdf

- Question Answering Dialog:
  https://arxiv.org/abs/1712.03316

Select Applications – Playing Games

Playing Games:

- Some Studies in Machine Learning Using the Game of Checkers: https://ieeexplore.ieee.org/document/5392560
- Playing Atari Games: https://deepmind.com/research/publications/playing-atari-deep-reinforcement-learning
- Head-up limit hold'em poker: https://science.sciencemag.org/node/624552.full
- Playing Doom: https://aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15130
- Mastering the esports game DOTA2: https://openai.com/five/
- Mastering the Game of Go: https://www.nature.com/articles/nature16961
- AlphaStar: Mastering the Real-Time Strategy Game StarCraft II: https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

Select Applications – Music and Images

Music and Image Generation:

- Generative adversarial nets:
  https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
- Image recognition using deep convolutional neural networks:
  https://link.springer.com/chapter/10.1007/978-3-319-63309-1_7
- Interactive image generation
  https://arxiv.org/abs/1905.03743
- Deep face recognition:
  https://arxiv.org/abs/1804.06655
- Deep learning for music generation:
  https://arxiv.org/abs/1709.01620?context=cs.LG
- Deep Jazz
  https://deepjazz.io/
- Deep learning for acoustic modelling:
  https://ieeexplore.ieee.org/abstract/document/7078992
- Speech enhancement using GANs:
  https://arxiv.org/abs/1703.09452
- Deep Fake assessment and detection:
  https://arxiv.org/abs/1812.08685
- Machine Recognition of Music Emotion:
  https://dl.acm.org/citation.cfm?id=2168754

Healthcare and Life Sciences Use Cases:

- Deep learning and medical image analysis: https://www.annualreviews.org/doi/abs/10.1146/annurev-bioeng-071516-044442

- Deep learning to improve breast cancer detection: https://www.nature.com/articles/s41598-019-48995-4

- Classification of skin cancer: https://www.nature.com/articles/nature21056

- Machine Learning in Genomic Medicine. https://ieeexplore.ieee.org/document/7347331

Select Applications – Financial Services

Banking and Financial Services Use Cases:

- Financial Trading:
  https://www.sciencedirect.com/science/article/pii/S0957417416000282

- Financial market forecasting:
  https://www.sciencedirect.com/science/article/pii/S0957417416302585

- Portfolio Optimization:
  https://arxiv.org/abs/0910.2276

- Stock market prediction:
  https://www.sciencedirect.com/science/article/pii/S0957417414004473

- Financial Accounting Fraud detection:
  https://arxiv.org/abs/1309.3944

Select Applications from other Industries:

- Stitch-Fix Algorithms Tour:
  https://algorithms-tour.stitchfix.com/
- Scaling ML at Uber with Michelangelo
  https://eng.uber.com/scaling-michelangelo/
- Artwork Personalization at Netflix
  https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76

Data Sources for Model Training

Publicly available data sources can help jumpstart model training and data exploration activities – these include data from various companies, natural sciences and simulations.

Datasets that address a broad variety of use cases:

- Google's dataset search engine:
  https://toolbox.google.com/datasetsearch

- Amazon's open data sets:
  https://registry.opendata.aws/

- Microsoft research open data:
  https://msropendata.com/

- Public data set list (curated by Xiaming Chen and contributors):
  https://github.com/awesomedata/awesome-public-datasets

Datasets that serve specific use cases:

- Visual cognition:
  https://www.visualdata.io/

- Nvidia's simulation on autonomous driving:
  https://developer.nvidia.com/drive/drive-constellation

- Other publicly available data sets:
  - https://www.drivendata.org
  - https://www.crowdanalytix.com
  - https://www.kaggle.com

Data Sources for Model Validation

A critical aspect of optimizing ML algorithms is using test data to benchmark model performance – these datasets serve as a great testbed for teams to validate their ML models.

- Unity's simulation environment for reinforcement learning agents: https://unity3d.com/machine-learning
- Deepmind's experiments for behavioral reinforcement agents: https://github.com/deepmind/bsuite
- A platform to compare ML hardware, software and services: https://mlperf.org/
- Various datasets often used in ML research: https://keras.io/datasets/

  This includes:
  - CIFARx small image classification
  - IMDB movie reviews sentiment classification
  - Reuters newswire topics classification
  - MNIST database of handwritten digits
  - Fashion-MNIST database of fashion articles
  - Boston housing price regression dataset
- UC Irvine ML repository: https://archive.ics.uci.edu/ml/index.php
- LIME: A technique for explaining model predictions https://arxiv.org/abs/1602.04938

Data visualization tools and techniques help with exploration and presentation activities leading to a better understanding of the data.

- D3 – A JavaScript library for data visualization: https://d3js.org/
- Highcharts – Another popular JavaScript library for data visualization: https://www.highcharts.com/
- Tableau – a business intelligence tool: https://www.tableau.com/
- Vega – Interactive graphs: https://vega.github.io/vega/
- Leaflets – An open source JavaScript library for visualizing maps: http://leafletjs.com/
- Python-based libraries for visualizing data:
  - matplot: https://matplotlib.org
  - seaborn: https://seaborn.pydata.org
  - ggplot: http://ggplot.yhathq.com/

Programming Languages

- Python: A general programming language. Because of its wide support by many ML frameworks it is the number one choice for ML: https://www.python.org/
- The R Project: A statistical computing and graphics language: https://www.r-project.org/
- Octave: An Open source language similar to Matlab https://www.gnu.org/software/octave/
- Julia: A language for scientific research: https://julialang.org/

Modeling Software and Frameworks

Machine Learning Libraries

- Scikit-Learn – a general-purpose ML framework for Python: http://scikit-learn.org
- CRAN Package Repository for R: https://www.kdnuggets.com/2015/06/top-20-r-machine-learning-packages.html
- SciPy - Scientific and statistical software tools for Python:
  - https://www.scipy.org/
  - http://pandas.pydata.org/
  - http://www.numpy.org
  - http://deeplearning.net/software/theano/

Machine Learning Frameworks

- Google Tensorflow, a popular deep learning computing framework:
  https://www.tensorflow.org/

- Facebook PyTorch (integrated Caffee 2) for deep learning:
  https://pytorch.org/

- Microsoft's Cognitive Toolkit (CNTK):
  https://github.com/microsoft/CNTK

- Keras, a lightwide framework on top of TensorFlow, CNTK or Theano:
  https://keras.io/

- Apache Mahout, a distributed linear algebra framework:
  https://mahout.apache.org/

- MXNet, a general deep learning framework provided by the Apache Foundation:
  https://mxnet.apache.org/

- Natural Language Toolkit (NLTK), a framework for working with human language data:
  http://www.nltk.org/

- Open Source Computer Vision Library (OpenCV):
  https://opencv.org/

Developer Tools to get Started

Integrated Development Environments (IDEs) provide a base foundational to get started.

- Jupyter, a popular browser-based IDE:
  https://blog.jupyter.org/tagged/jupyter-notebook
- Anaconda, a package manager specialized in data science libraries:
  https://www.anaconda.com
- R-Studio, an IDE for R:
  https://rstudio.com/
- Spyder, an IDE for Python:
  https://www.spyder-ide.org
- PyCharm, an IDE for Python:
  https://www.jetbrains.com › pycharm
- Visual Studio:
  https://visualstudio.microsoft.com
- Sublime Text Editor:
  https://www.sublimetext.com

Additional Training Via Platform Companies

Many leading platform companies have opened their internal training resources to the public. Few such resources are listed below.

Google Cloud Platform (GCP) Training Catalog

- Focuses on developing skills to design, deploy and maintain Google Cloud technologies by providing an overview of the platform, tools and APIs.

Microsoft Machine Learning School | Machine Learning APIs

- Microsoft launched various programs and tracks to educate how its solutions can be used to solve machine learning problems

Amazon Web Services (AWS) Machine Learning Training

- Amazon opened the same machine learning courses used to train its engineers to all developers through AWS.

Facebook Facebook Guide to Machine Learning

- Facebook has a variety of training videos and documentation around the use of machine learning at the company and general conceptual introductions.

IBM IBM Watson Academy

- IBM offers a range of trainings including the use of its famous Watson library for Machine Learning and Artificial Intelligence.

Apple Machine Learning and Development at Apple

- Apple allows developers to implement machine learning models to integrate with all products including training on integrating outside systems like IBM's Watson.

Suggested Reading – Academia

- Boden, Margaret. *Mind as Machine: A History of Cognitive Science*. Oxford University Press, 2006.
- Chapelle, Olivier, Bernard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.
- Frankish, Keith and William Ramsey, eds. *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, 2014.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009.
- Kelleher, John D., Brian M. Namee, and Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press, 2015.
- Mitchell, Tom M. *Machine Learning.* McGraw-Hill series in computer science. New York: McGraw-Hill, 1997.
- Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach (3rd Edition)*. Upper Saddle River, New Jersey: Pearson Education, 2010.
- Sutton, Richard, and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2017.
- Witten, I. H., and Eibe. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. MORGAN KAUFMANN PUBLISHERS, 2005.

Suggested Reading – Business

- Agrawal, A., J. Gans, and A. Goldfarb. *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press, 2018.
- Aoun, Joseph. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. MIT Press, 2017.
- Brynjolfsson, Erik, and Andrew McAfee. *The Second Machine Age*. W. W. Norton & Company, 2014.
- Daugherty, P., and Wilson, H. J. *Human + Machine: Reimagining Work in the Age of AI*. Harvard Business Press, 2018.
- Ford, M. *Architects of Intelligence: The Truth about AI from the People Building It*. Packt Publishing, 2018.
- Kurzweil, Ray. *The Age of Intelligent Machines*. MIT Press, 1990.
- Lee, K. *AI Superpowers: China, Silicon Valley, and the New World Order*. Houghton Mifflin Harcourt, 2018.
- Walsh, M. *The Algorithmic Leader: How to be Smart When Machines are Smarter than You*. Page Two, 2019.

AI News and Community Sites

Bookmarking these side will keep you up to date on the latest technical developments

- DeepMind: https://deepmind.com/
- OpenAI: https://openai.com/
- Google's AI Blog: https://ai.googleblog.com/
- Uber's AI Blog: https://eng.uber.com/category/articles/ai/
- Airbnb Engineering Blog: https://medium.com/airbnb-engineering
- Netflix Research Blog: https://research.netflix.com/
- Facebook AI Blog: https://ai.facebook.com/
- Tencent AI: https://ai.qq.com/hr/weixin.shtml
- Data Science and Mining: https://www.kdnuggets.com/
- Data Science Central: https://www.datasciencecentral.com/
- Towards Data Science (aggregates relevant data science topics) https://medium.com/towards-data-science/data-science/home
- AI Report: https://aiindex.org/

The following resources will help you get started with KNIME – from installing the platform and few extensions, to building familiarity with the interface and workflow tools.

Konstanz Information Miner (KNIME) is a modular environment, which enables easy visual assembly and interactive execution of a data pipeline. It is designed as a teaching, research and collaboration platform, which enables simple integration of new models, as well as access to a rich library of existing algorithms and methods. It allows the user to visually assemble and adapt the analysis flow using standardized building blocks, which are connected through pipes carrying data or models. An additional advantage of KNIME is the intuitive, graphical user interface.

1. KNIME

First, follow these instructions to install KNIME on Windows, Mac or Linux

https://www.knime.com/knime-introductory-course/chapter1/section1/installation-guide

Once installed, you will see a welcome page such as the one below:

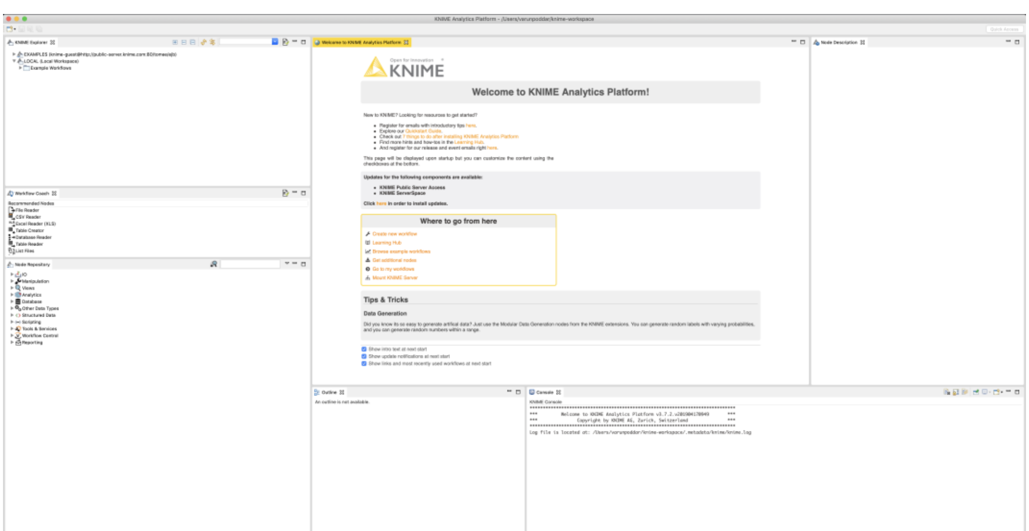


Figure 1 – KNIME Welcome Page

2. KNIME EXTENSIONS

Next, install few basic KNIME Extensions (extensions provide additional functionality such as access to and processing of complex data types, and the use of advanced algorithms)

https://www.knime.com/knime-introductory-course/chapter1/section1/install-knime-extensions

The basic KNIME Analytics Platform does not include many of the nodes that you see in various KNIME libraries and applications. Those nodes belong to “Extensions” packages and usually must be installed separately. Unless at installation time the package containing all free extensions was selected, you will need to install the KNIME Extensions now. You can do so by following the link “Get additional Nodes” in the “Welcome to KNIME” page. The bare minimum in terms of needed extensions is “KNIME & Extensions”, “KNIME Labs Extensions” and “KNIME Community Contributions –Other”.

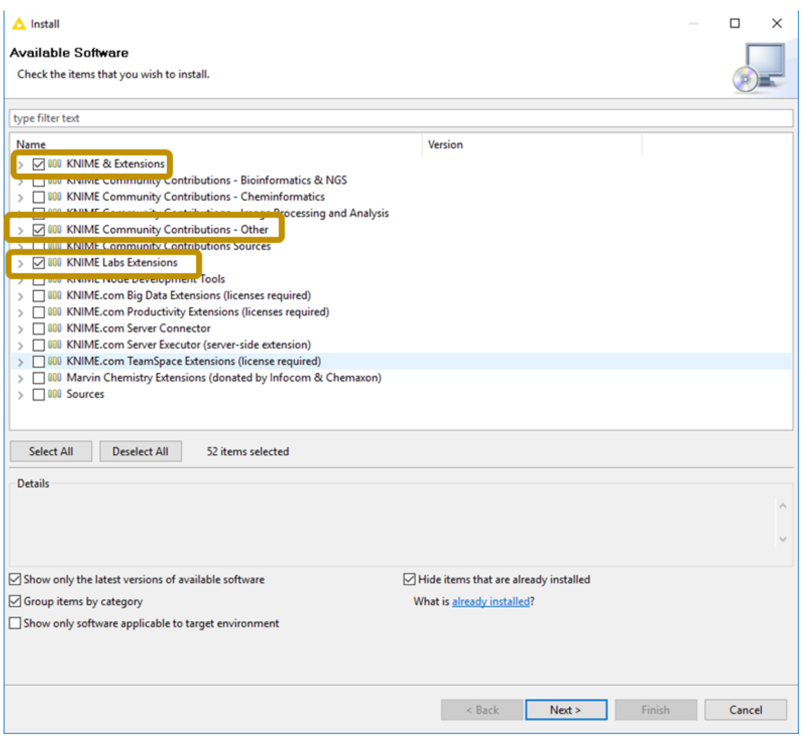


Figure 2 – KNIME Extensions

3. KNIME WORKBENCH

Now that you are ready to get started, get to Know the KNIME Workbench

https://www.knime.com/knime-introductory-course/chapter1/section1/knime-workbench

Figure 3 shows a screenshot of the standard KNIME workbench. In the center – the Workflow Editor – is where you design the workflow. A workflow is a series of steps such as reading in data from various sources and processes it in several, parallel analysis flows, consisting of preprocessing, modeling, and visualization. On the left are three helpful panels – the file explorer which lists various working files and models, a recommendation engine that guides users as they create a workflow, and a repository of nodes which are essentially steps in the workflow. Search results and help guides are shown in the right panels for selected items.

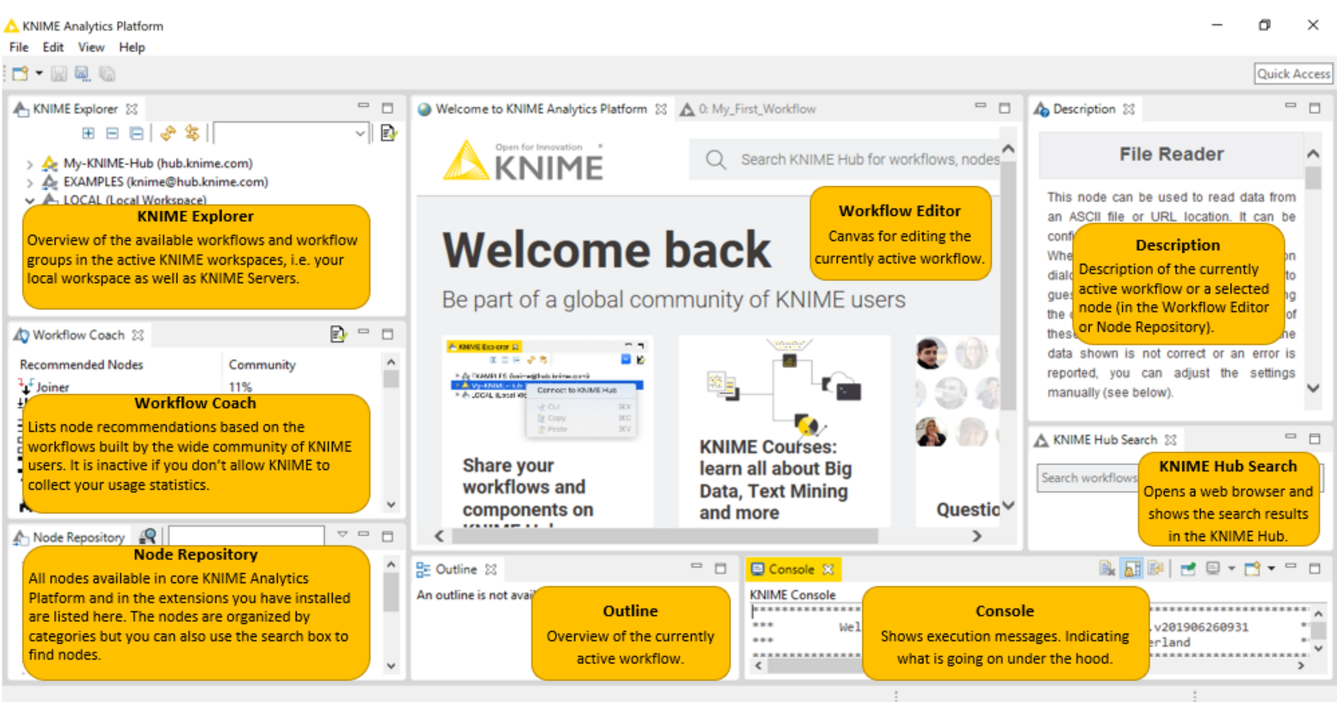


Figure 3 – KNIME Workbench (the interface)

You will be spending most of your time in the workflow editor and the node repository. The workflow editor is where workflows are assembled. Workflows are made up of individual tasks, represented by nodes. In KNIME, individual tasks are represented by nodes. Nodes can perform all sorts of tasks, including reading/writing files, transforming data, training models, creating visualizations and many others. You create a workflow by dragging nodes from the Node Repository to the workflow editor, then connecting, configuring and executing them.

From the large variety of nodes in the node repository, you can select data sources, data preprocessing steps, model building algorithms, as well as visualization tools and drag them onto the Workflow Editor, where they can be connected to other nodes. KNIME is written in Java and its graphical workflow editor is implemented as an Eclipse plug-in. It is easy to extend through an open API and a data abstraction framework, which allows for new nodes to be quickly added in a well-defined way.

4. KNIME WORKFLOWS and NODES

Once you are familiar with the interface, understand more about workflows and nodes

https://www.knime.com/knime-introductory-course/chapter1/section2/workflow-coach

A workflow usually starts with a node that reads in data from some data source, which are usually text files, but databases can also be queried by special nodes. Imported data is stored in an internal table-based format consisting of columns with a certain (extendable) data type (integer, string, image, molecule, etc.) and an arbitrary number of rows conforming to the column specifications. These data tables are sent along the connections to other nodes that modify, transform, model, or visualize the data.

Modifications can include handling of missing values, filtering of column or rows, oversampling, partitioning of the table into training and test data and many other operators. Following these preparatory steps, predictive models with machine learning or data mining algorithms such as decision trees, Naive Bayes classifiers or support vector machines are built. For inspecting the results of an analysis workflow numerous view nodes are available, which display the data or the trained models in diverse ways.

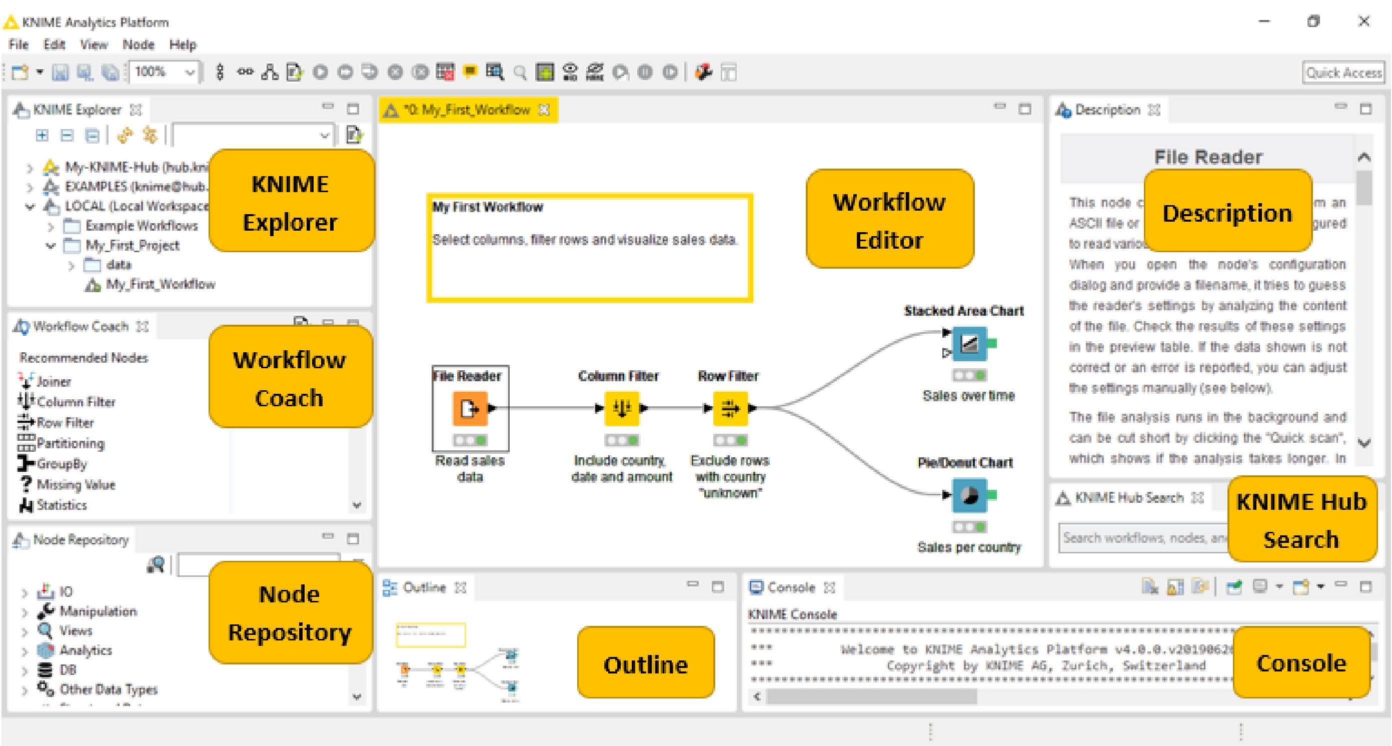


Figure 4 – KNIME Workflow Editor with a Simple Workflow

A "node" is a basic processing unit in the KNIME workflow – it performs a particular task converting inputs into outputs for subsequent nodes. The following ports, conventions and right-click options will help as you work with nodes.

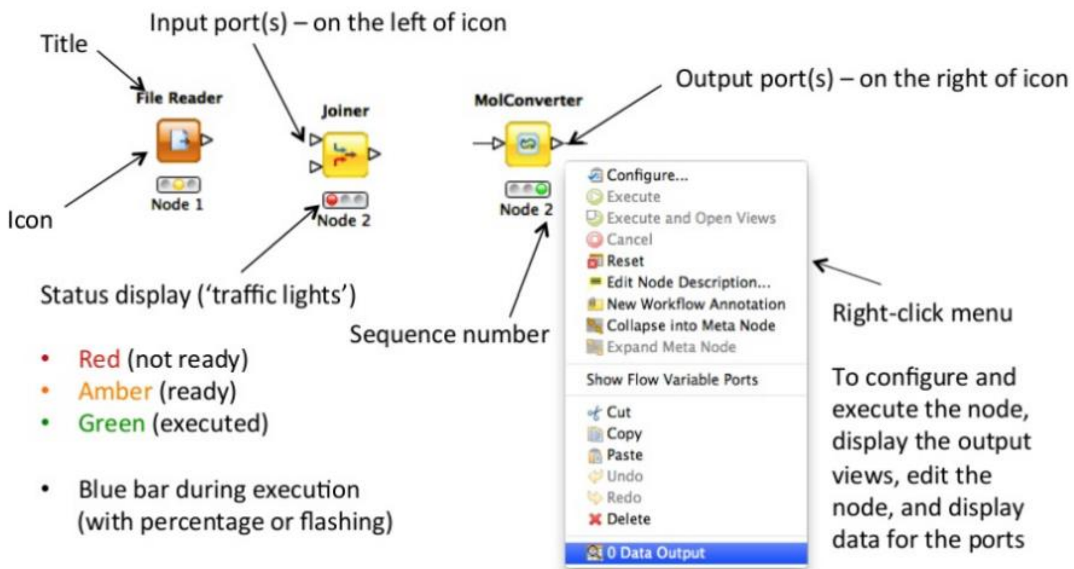


Figure 5 – Key Elements of a Node

5. ADDITIONAL RESOURCES

KNIME is a very popular open source platform. Its community of users have contributed a rich set of resources including tools, ML libraries and training material. Many of these are available through the links below.

- **KNIME Learning Hub** https://www.knime.com/learning-hub
- **KNIME TV Channel** https://www.youtube.com/user/KNIMETV
- **KNIME Cheat Sheets** https://www.knime.com/learning/cheatsheets

AI Canvas.
Empowering the Visually Impaired using AI.

TURKCELL

Business Problem
According to Turkey's legal rules, Telco operators in Turkey have to send a voice message that contains personal pricing and tariff details to their disabled customers upon request. So, Turkcell needed an automated and scalable solution to send voice messages to visually impaired subscribers.

Business Value
Turkcell wanted to mitigate legal risk, avoid compliance penalties levied by Turkey's legal institution, and create a better customer experience. They aspired to build a solution using which visually impaired subscribers could learn about their bills and tariff changes conveniently just by opening voice messages.

Objective Function
Key goal is to convert given text into human like speech with a high degree of accuracy. Also, the speech voice that is auto generated should not be a disturbing robotic voice but portray a friendly, human personality and tone.

Modeling Approach
Turkcell needed a platform that converts given text into voice files. So, AI / Deep Learning based Text-To-Speech (TTS) algorithms were a good fit. Turkcell applied Tacotron's TTS framework. A single speaker's 40 hours of voice recording was used for training. The model was then deployed to make real-time text to speech conversions.

Model Training
The model generates human like voices in real time. Turkcell provides a synthesizer web service. The platform is designed to convert text messages into speech files at the rate of 5 texts per second (TPS). As necessary, additional servers are added to increase TPS processing capacity. The TTS Model is re-trained on-demand only if needed.

Customer Value
Initially, Turkcell's disabled subscribers were informed of tariff changes and bills via text messages. Most of the time they could not read or understand the message, thereby failing to act on it timely. As a result of this voice message, visually impaired subscribers can easily follow their tariff and bill details just by opening the voice message.

Data Strategy
Core data entities include the Message Catalog, Customer Tariffs and Invoices, Subscriber Requests, Professional Voice Recording (for Text-to-Speech model training) and Tariff Charging Systems. Externally batched voice records and message logs are used for additional training and testing. Turkcell contracted with a voice actor and produced 40+ hours of voice data records to build the initial training set and preprocessing rules. This training data was used to build a TTS model, instrumented on Tacotron RNN algorithm. TTS doesn't need explicit subscribers data and consent. Turkey has GPDR like personal data protection rules and Turkcell works within those guidelines in using the subscribers' data to train and refresh these AI models.

AI Canvas.
Subscribers Listen to any App in the Dergilik App.

TURKCELL

Business Problem
Dergilik is Turkcell's magazine app that includes a lot of magazines and newspapers on different topics such as home design, technology, sports, and business. Turkcell was building AI capabilities in other parts of its business, and wanted to extend application to provide automated voice-generation for Dergilik's magazines/newspapers based content.

Business Value
Turkcell extended the AI platform that was built to provide its visually impaired subscribers with accurate tariff information to its magazine app. Allowing customers to listen to any content in the Dergilik app increased customer discovery of new content, drove greater engagement and created a better experience for all users (not just the visually impaired).

Objective Function
Key goal is to convert given text into human like speech with a high degree of accuracy. Also, the speech voice that is auto generated should not be a disturbing robotic voice but portray a friendly, human personality and tone.

Modeling Approach
Turkcell needed a platform that converts given text into voice files. So, AI / Deep Learning based Text-To-Speech (TTS) algorithms were a good fit. Turkcell applied Tacotron's TTS framework. A single speaker's 40 hours of voice recording was used for training. The model was then deployed to make real-time text to speech conversions.

Model Training
The model generates human like voices in real time. Turkcell provides a synthesizer web service. The platform is designed to convert content into speech files at-scale, with server processing capacity scaled to meet demand. The TTS Model is re-trained on-demand only if needed.

Customer Value
Turkcell customer research surfaced that a significant part of its subscriber base prefers to listen to Dergilik contents instead of reading it, especially while travelling. This service offers another way for subscribers to engage with the app, and gives visually impaired customers access to content they could not consume otherwise.

Data Strategy
Core data entities include the Dergilik (Turkcell's magazine app) Contents, Subscriber Requests, Professional Voice Recording (for model training) and Tariff Charging Systems. Externally batched voice records and message logs are used for additional training and testing. Turkcell contracted with a voice actor and produced 40+ hours of voice data records to build the initial training set and preprocessing rules. This training data was used to build a TTS model, instrumented on Tacotron RNN algorithm. TTS doesn't need explicit subscribers' data and consent. Turkey has GPDR like personal data protection rules and Turkcell works within those guidelines in using the subscribers' data to train and refresh these AI models.

AI Canvas.
Predict Most-Liked Photos to Share.

Business Problem
Lifebox is Turkcell's Cloud based backup & storage app. Users can save contacts, photos, videos and files securely. One big use case is users sharing photos with their network. Turkcell wanted to extend this capability by providing users with smart recommendations that intelligently picked photos that are more likely to garner likes on social media platforms such as Instagram.

Business Value
Within the Lifebox app, users can select a set of pictures that they want to share on Instagram and Photo-Pick will predict which photos will get more likes on Instagram. This increases stickiness with Turkcell's service, differentiates it from competing offerings and harnesses the viral nature of social media posts, particularly photo-sharing.

Objective Function
Key objective is to score each photo in a range of 1-5. Generated scores must be distinctive in order to recommend the best (most share-worthy) photo to subscribers.

Modeling Approach
Tensorflow framework and Keras libraries are used to implement the AI model. Turkcell trained the model using 100,000+ publicly available photos from Instagram. Turkcell used mean squared error to evaluate the model's success.

Model Training
The model generates scores for each image in real-time. The platform is designed to calculate scores for each requested image at the rate of 30 per second. The model is manually re-trained only if needed.

Customer Value
Turkcell subscribers store their photos on Lifebox and share them continuously on Instagram. One choice that frequently perplexes their subscriber is which photo is the best one to post to Instagram. This service takes that decision-making off the user's mind, intelligently recommending one that is more likely to get noticed and liked.

Data Strategy
Core data entities include external data from Public Instagram Accounts (for training image scoring model) and Subscriber Requests. Since public domain data from Instagram was used, special consent of subscribers was not needed. Turkey has GPDR like personal data protection rules and Turkcell worked within those guidelines in using the subscribers' data to train and refresh these AI models.

AI Canvas.
Energy Optimization at the Learning Mill.

Business Problem
Noodle.AI helped a client accurately predict hourly energy consumption at a steel mill, within their melt shop. They went on to create the first Learning Mill – one that continuously senses, learns and improves its processes. Their objective was to increase profit per mill hour while meeting increasing demand for high quality steel at predictable times, all while conserving the amount of energy consumed.

Business Value
Trading revenue from selling energy back to the grid can reduce the overall energy cost incurred by the steel mill by about 5-10%. Moreover, it makes the transmission grid more efficient as the regional transmission organization can redeploy unused energy elsewhere thereby reducing carbon emissions. This also supports real-time and day-ahead energy trading / hedging strategies.

Objective Function
The solution has three objectives: (1) Predict energy per production batch, (2) Time duration of each production batch, (3) Profile energy consumption over time in a production batch. The production schedule drives the prediction models to generate hourly energy profiles for the next 2 days.

Modeling Approach
Regression techniques are used for the first two parts of the problem and clustering for the third part. Noodle.AI used 8 months of historical data for training and 1 month for testing the initial model.

Model Training
Predictions are generated every 15 mins using the latest production schedule. These are used as inputs to day-ahead and real-time energy trades. Models are re-tuned quarterly to keep learning new information from the mill such as new product types. The solution can easily be deployed to other similar steel mills.

Customer Value
There did not exist any tool to support energy trading. The new software application acts as a central destination to compute and display all necessary information (like forecasted energy usage and optimal trade amount) for determining energy trades and selling energy back to the grid during periods of low energy consumption.

Data Strategy
The primary data entities are from the manufacturing execution systems at the steel mill which include sensor data (50,000 sensors were embedded in the mill) and operational data, such as, steel grades, scrap recipes, temperature changes, crew information, heat sequence, etc. All data is batched and ingested at 15-min intervals. A secure VPN tunnel is usedto transport data from the steel mill systems into Noodle.AI's compute infrastructure to ensure data protection. All primary data is owned by the steel mill and the results are for internal use only. Secondary data sources are National Oceanic and Atmospheric Administration (NOAA) for weather information that is collected once a day.

AI Canvas.
Signature Verification for Fraud Prevention.

Business Problem
Approximately 2 millions checks were cashed per day in Brazil in 2017. A trained analyst has less than 15 seconds to decide if a check is fraudulent. This leads to fewer signature verifications and contributes to significantly higher rate of fraud.

Business Value
Enabling automated signature verification increases the number of signatures evaluated with higher consistency and accuracy rates. Analysts only evaluate when the model is unsure and there are exceptions, freeing up their time for other value-added activities. Also, models can create a rank-ordered queue of transactions for fraud units to investigate.

Objective Function
Key objective is to recognize individual signatures as forged or genuine based on both forged and genuine versions of same person's signature during training. The loss function produced by current weights is propagated backward through the neural network to update weights, adjusting iteratively to reduce the loss function.

Modeling Approach
This model uses Siamese neural networks to evaluate new signature images in a pairwise manner against the on-file image. The pairing with the highest score according to the verification network is awarded the highest probability for the one-shot task (verified or not verified), accepting those signatures that are above a threshold.

Model Training
The model learns image representations via a supervised metric-based approach with Siamese neural networks, then reuses that network's features for one-shot learning. As the network learns from its mistakes for every training iteration, it adjusts weights iteratively to reduce the loss function.

Customer Value
Bank resources are deployed towards investigating exceptions or managing customers as opposed to verifying signatures. Additionally, customer satisfaction is improved by protecting individual accounts from fraud and not challenging innocent transactions.

Data Strategy
The data entities central to this use case are the customer identity record, check transaction details, and signature images. Leveraging a process called transfer learning, the team initialized the neural network with parameter values learned from a different dataset of 300+ signature images. The model was initialized with pretrained weights with just the fully-connected layers at the end being trained and the convolutional layers kept fixed throughout training. Online verification leveraged an electronic signing system which provided data such as the pen's position, azimuth/altitude angle, and pressure at each time-step of the signing. Offline verification relied on 2D visual (pixel) data acquired from scanning signed documents.

AI Canvas.
Ad Targeting Using ML.

Business Problem
Advertisers on platforms such as Facebook only bid and pay for measurable user responses making click prediction systems critical for any online advertising system. Facebook has over 750 million daily active users and over 1 million active advertisers, making the task of predicting such clicks near impossible without effective use of machine learning techniques.

Business Value
Highly targeted segments allow advertisers to reach the most relevant audience on Facebook, thereby driving better reach, conversion and return on advertising spend. Accurate click through rate prediction is essential to the success of online ad bidding and auctions. Consequently, these factors are critical to growth in advertising revenue.

Objective Function
Facebook serves ads tied to specific user demographics and interests, with the objective of maximizing likelihood that the user clicks on that ad. Key modeling considerations included that the model be robust, adaptive, learn from massive volumes of data balancing memory, latency and compute time.

Modeling Approach
Various modeling approaches are used, off which most pertinent to this specific objective is a combination of decision trees with logistic regression. Number of trees vary from 1 to 2000 with predictions of models trained on one full day of data tested on the next day's data.

Model Training
Tests have shown that prediction accuracy degrades as the delay between training and test sets increases. So, Facebook retrains models daily. Given compute trade-offs, boosted decision trees are trained daily, while the linear classifier is trained online near real-time i.e. directly as the labelled ad impressions arrive.

Customer Value
Facebook users benefit from a more relevant ad and a tailored experience that is tied to their search queries and interests. Businesses are able to reach a more targeted, valuable segment thereby generating a higher return on their ad spend.

Data Strategy
Just one day of Facebook ads impression data is massive, so Facebook applies techniques (uniform subsampling, negative down sampling) to reduce training volume data and select fewer, higher priority features. The data used for modeling can be categorized as contextual or historical. Contextual information relates to the context in which an ad is shown such as device, current page that user is on, local time of day, and day of week. Historical information is related to previous interaction for the ad or user, such as cumulative number of clicks on the ad, click through rate in prior week, or average click through rate of the user. Facebook studies have shown that top 10 features are responsible for half of total feature importance, while last 300 features contribute less than 1% feature importance. Historical features provide considerably more explanatory power, while contextual features are critical to handle the cold start problem and provide a reasonable click through rate prediction.

AI Canvas.
Medical Device Compliant Handling with RPA.

Business Problem
The FDA requires medical device manufactures to record complaints about the devices in a structured format or face significant fines and penalties. This results in cumbersome processes and significant overhead for medical device companies, and presents an opportunity to drive efficiencies using robotic process automation.

Business Value
Automation of the complaint process reduces errors and the number of employees who are required. Fines in this area were $1.38M over the past 5 years. An average employee in this area earned $85,000 in salary and benefits. Using RPA, medical device companies can drive significant cost savings.

Objective Function
The RPA process has three fundamental decision points: (1) If Complaint, open Complaint Item (2) If Investigation required, open Investigation (3) Review for reportability and reports if needed. The outcome of these decisions is that the RPA Bot automatically opens a case and files a report with the FDA in situations that demand it.

Modeling Approach
The RPA bot was trained by the company's employees themselves with input from IT, compliance and legal staff. The bot that was created took 5 days to create and deploy. Once it began working in the company's environment, machine learning analytics were finalized after three months of data gathering and fed back to retrain the bot.

Model Training
Load balancing with strict policies addresses infrastructure scaling and real-time needs. Analytics are refreshed as needed when regulations change. The bots provide a log of all actions. If mistakes are made, the system is taught to learn from the bots' mistakes and prevent similar mistakes in the future.

Customer Value
Device manufacturers and end customers benefit from: (1) Reduced time spent on complaints & adverse events (2) Greater accuracy via automated data entry (3) Tracked product safety issues and optimized customer support (4) Self-managed email box and communications with automated reporting

Data Strategy
The sources of data are the complaint themselves, the Unique Device Identification codes, the Device History Records, and the design history files. This data is pulled in real-time from respective databases, and streamed live to the RPA bot. The bots are trained to access information in a method that is compliant with all state and federal laws. If a bot is working on data from another country, it is trained to be compliant with that country's laws also. The bot interprets intent using natural language processing for complaints that are submitted using paper forms. Situations for which the bot IDKs (results in an "I Don't Know" response) are sent to a queue that is reviewed by a human, who not only follows up on that situation but also trains the bot to handle the exception if possible.

AI Canvas.
Intelligent Conversion Rate Optimization.

Business Problem
Experimentation has traditionally been limited to A/B testing as the principal tool for validating/rejecting conversion and personalization hypotheses. However, most A/B tests do not produce positive results, and most companies do not have the necessary resources or traffic levels to run the number of A/B tests required to see a consistent ROI for money spent on website optimization.

Business Value
Ascend improves ROI of experimentation by increasing both test velocity and win-rate without increasing manual resources. This improves full-funnel optimization driving higher conversion rates and sales revenue. It also creates efficiencies for marketing teams by providing an automated system for massive multivariate optimization that finds subtle combinations of variables that will increase conversion.

Objective Function
Ascend uses adaptive evolutionary optimization to test the impact of a large set of individual changes, using an evolutionary process that efficiently searches through the space of all combinations of these changes (100s to millions of designs) to predict winning designs that have the highest likelihood of increasing conversion.

Modeling Approach
Ascend uses evolutionary algorithms that are inspired by biological evolution (ex: natural selection). Websites “evolve” by testing several elements in real time. Early “winners” are bred together to form a new generation of “offspring” websites that inherit all the best features and combinations of design elements (i.e. best genetics) to increase conversions.

Model Training
The AI relies on Bayesian and traditional statistical techniques learning over time which combination of elements are effective. It gradually focuses the search around most promising designs, and automatically generates new experiments based on a combination of high-performing hypothesis, continually searching for higher performance.

Customer Value
AI-powered evolutionary testing allows website managers and digital marketers to test several elements on their website to find the right combination for highest conversions. They can do this faster than with other testing processes, scale their to manage different web properties with less resource time and spend more time ideating than doing statistics.

Data Strategy
Data is driven by the search space, which is the total number of possible solutions in an experiment, and is determined by multiplying the number of variations of each element. For instance, if you have 1 landing page but want to test 2 headlines, 3 versions of an image, 4 variations of the CTA and 5 different background colors, the search space would be 120 different candidates. These design combinations are tested iteratively over an “evolution run” of 6-8 ”generations”. Other data sources that act as input include the website’s traffic volume (no. of users that visit the website), website conversions (sales velocity), clickstream data (web browsing and interaction data) and any available demographic information.

AI Canvas.
Fraudulent and Overpayment Claims Prediction.

Business Problem
Healthcare Fraud and Claims Overpayments in US is a $100B Dollar industry. It poses a significant challenge to Payer industry since they incur billions of dollars to identify and recover this leakage. Some of this cost is also passed to the patient population directly leading to increasing cost. The business wants to identify the overpayment and the fraudulent claims at a scale and at a high speed leveraging power of Advanced Analytics and AI. Early identification of such claims lead direct operational savings.

Business Value
The current method of overpayment detection is post-payment and it leads to huge cost of recovery, and less than optimum success. The client wants to identify such claims by means of early detection and reduce losses. The client organization pays $1B+ annually in overpayments and fraud, and it has the recovery cost of 10%. Thus, the potential opportunity is of $100M+.

Objective Function
Develop analytical application to identify claims overpayments at adjudication time. Dependent variable is overpaid claim and/or fraud indicator. Features include base statistics, frequency based, concentration based, velocity based, amount based and geospatial based variables.

Modeling Approach
Started with Logistic Regression technique and then moved to Random Forest as the eventual technique. Initial models were built in R, ported to Python and finally deployed on H2O as the production platform.

Model Training
Develop model and rules performance tracking dataset which acts as a primary input for operational dashboard to executives. This dataset was used as a feedback mechanism to improve models and bring in higher business benefit.

Customer Value
Increases claims processing speed and scale (50,000 claims processed for overpayment tagging every 15 minutes). Results in significant savings from fraud and overpayment. Allows teams to incorporate various rules as a part of pre-pay process to aid early detection and prevention of fraudulent cases.

Data Strategy
Core data entities are Member Demographics, Membership, Claims, Provider Contract, Pre-Authorization, Doctors Visits, Diagnosis Codes, Drugs, etc. We had near real-time claims streaming to this analytical application with other reference datasets developed in the batch mode (daily). Created re-usable and scalable Data Quality framework; Developed Yarn log parser: to capture the execution times & detect performance bottlenecks in-process.

AI Canvas.
Predictive Maintenance at a Wind-farm.

Business Problem
Wind-farm operators spend $3 Billion to $4 Billion annually for operations and maintenance expenses. Unplanned downtimes increase operating costs, reduce availability and hurt energy production yield. With the efficient use of sensors and real-time machine learning models, operators can predict and prevent failure but deploying such systems at-scale is challenging.

Business Value
Uptake’s AI based SaaS predictive maintenance software can cut such operating expenses by up to 30% annually. Operators can increase their turbine productivity by quickly discovering and addressing underperforming turbines, preventing failures before they occur and improving annual energy production. Further, operators can produce more power when prices are optimal, ensuring maximized revenue as well.

Objective Function
The goal of Uptake’s platform goes beyond just predictive maintenance. The output is a user friendly dashboard showing actionable insights such as asset utilization, availability, fuel efficiency (where applicable), lost opportunities (in $ terms), and GADS availability.

Modeling Approach
First Uptake uses historical data from client’s wind farm to train machine learning models. One ML algorithm used here is PCA-based Anomaly Detection. Once client is convinced of the value, ML models are run on live data and continually tweaked/trained to improve accuracy.

Model Training
The algorithms are augmented by acquisition of Asset Performance Management, the world’s largest database of asset types, machine failures and preventative maintenance strategies. This allows Uptake to know know the 58,000+ ways machines can fail, and how to cost-effectively prevent those failures from happening.

Customer Value
Once Uptake has trained models, which are validated on one windfarm, these models can onboard new windfarms with very low marginal cost. This allows Uptake to offer a competitive Predictive Maintenance subscription to clients, and clients to drive greater operational efficiencies.

Data Strategy
If a machine provides telemetry data itself then Uptake uses that otherwise Uptake can install its own Edge device that transmits data via cellular or satellite bands. For windmills Uptake uses telemetry data such as turbine speed, temperature, oil level, vibrations, weather forecasts, etc. Uptake also ingests maintenance records, often filled out by hand by technicians. These are first cleaned, organized and pre-processed using AI techniques so that Uptake has accurate labels to train ML models on.

Thank You To All Our Contributors

Many individuals were instrumental in research and development of AI / ML use cases featured throughout this program. Our team would like to recognize the following individuals for sharing their experiences, products and services.

Automation Anywhere
Bill Lauritzen

Axtria
Snehamoy Mukherjee, Ardaman Kohli

Bahwan CyberTek
Rajasekaran Panchatcharam, Sivarama Krishnan Balasubramanian, Kavitha KK

Delhivery
Kshitij Chopra, Kabir Rustogi

EXL
Vivek Jetley, Nagendra Shishodia

Fooda
Jason Stulberg, Patrick Rice

Fruit of the Loom
Beth Rogers, Cameron White, Cody Bowlds

Infosys
Vinay Agarwal, Kisha Gupta, Shreyans Gautam, Akshat Mathur, Sumithra Sivaramakrishnan

Microsoft
Marc Ashbrook, Ella Ben-Tov, Sedhu Krishnamurthy, Roger Longbotham, Willie Mason, Shaaz Nasir, Jay Natarajan, Jaime Perena

Nio
Abhishek Singhal

Noodle.AI
Mike Godwin

Oracle
Chung Wu

Teradata
Chad Meley, Mary E Gros

Turkcell
Mehmet Buyukzincir, Caner Canak, Semih Kumluk, Sezen Felekoglu

Other
Manoj Bapat, Anup Dhalwani, Karthik Kotha, Alastair MacPhail, Suraj Madnani, Alok Prasad, Martin Prause, Amit Sahdev

Thank You To All Our Contributors

Additionally, we would like to thank the following companies for sharing their research, applications, approaches and insights.

Accenture
Airbnb
Amazon
Amsterdam Intensive Care Unit
Amtrak
American Express
Arago
ASI Mining
Asos
AT&T
Axonify
Ayasdi
Blue Prism
Cerebri.AI
Clover Group
Course5
Danske Bank
DeepCode
Deloitte Digital
DHL
eBay
Enlitic
Espoo
Fabula
FarmBeats
Fluidly
Gartner
GE Healthcare
Google
Harley Davidson
HealthScape
HireVue
IBM Watson
iNaturalist
Intuit
Lavazza
LawGeex

Maitland Group
Manthan
McKinsey & Company
Netflix
NorthPointe
Ocado
Pinterest
Port of Rotterdam
Qopius
QualSights
Qventus
Reliance Jio
Rio Tinto
Rolls Royce
Salesforce
SayKara
Sentient (Evolv)
SigmaIQ
SigTuple
Stitch Fix
Tata Consultancy Services
Textio
Tiger Analytics
Tricentis
Uber
UiPath
Unilever
Uniqlo
Uptake
vMock
Voyage
Wade & Wendy
Whitepages Pro
Xero
Zebra
ZestFinance
Zipline