Introduction to Artificial Intelligence and Machine Learning

Alice Oh
alice.oh@kaist.edu

AI, ML, and Related Areas

Machine Learning

VS

Big Data

Machine Learning is
one way to **analyze**, **understand**, and **predict**
Big Data

Machine Learning

VS

Data Mining

Machine Learning does not require **structured data**, while Data Mining does

Machine Learning

VS

Artificial Intelligence

Machine Learning
develops **data-dependent solutions** to the problems in Artificial Intelligence

Machine Learning

VS

Statistics

Machine Learning is **deeply rooted** in, but **expands** the practical limitations of Statistics

Major Problem Formulations in ML

- Supervised Learning
- Unsupervised Learning
- Representation Learning
- (Reinforcement Learning) — not covered today

car

car

car

car

training

testing

car?

car?

car?

Supervised Learning

car

car

car

car

training

testing

car?

car?

car?

Supervised Learning

Meaningful Patterns?

car

car

car

car

training

testing

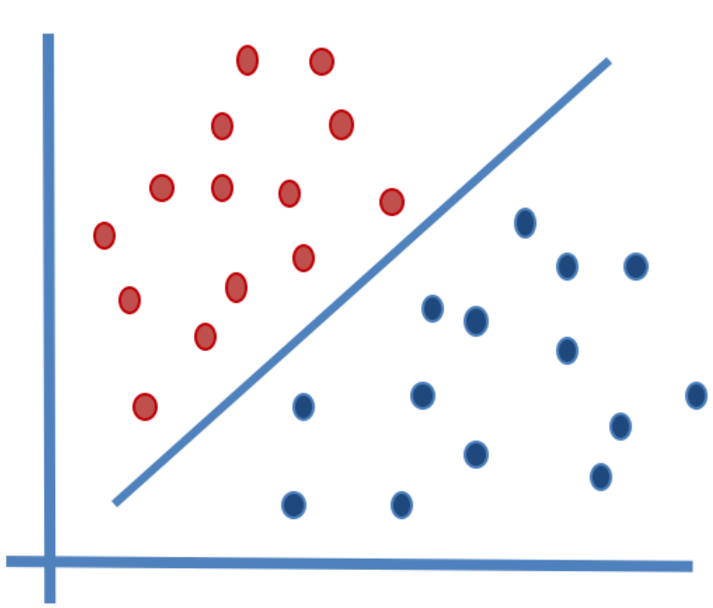
car?

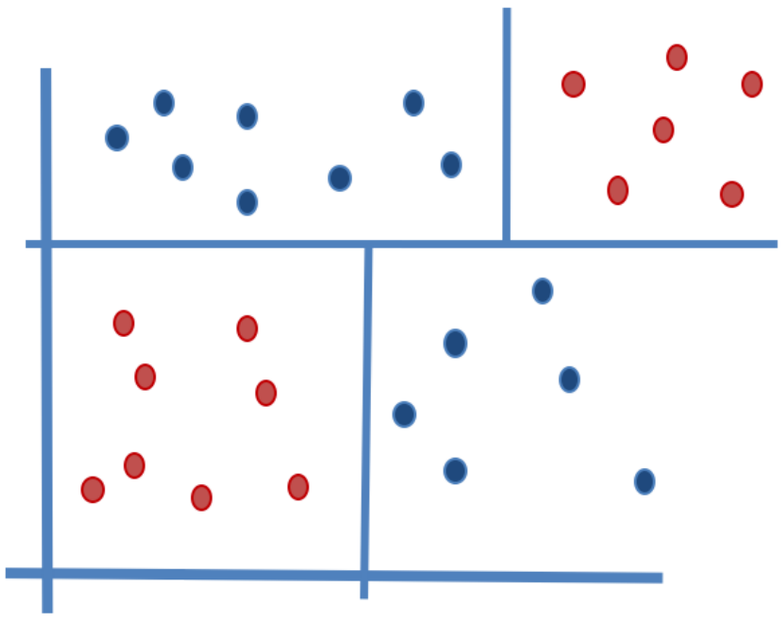
car?

car?

Supervised Learning

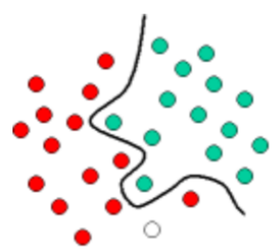
Design Features?





Supervised Learning

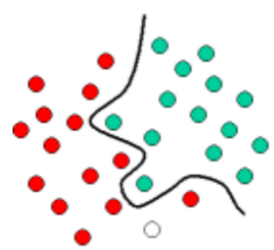
Classification



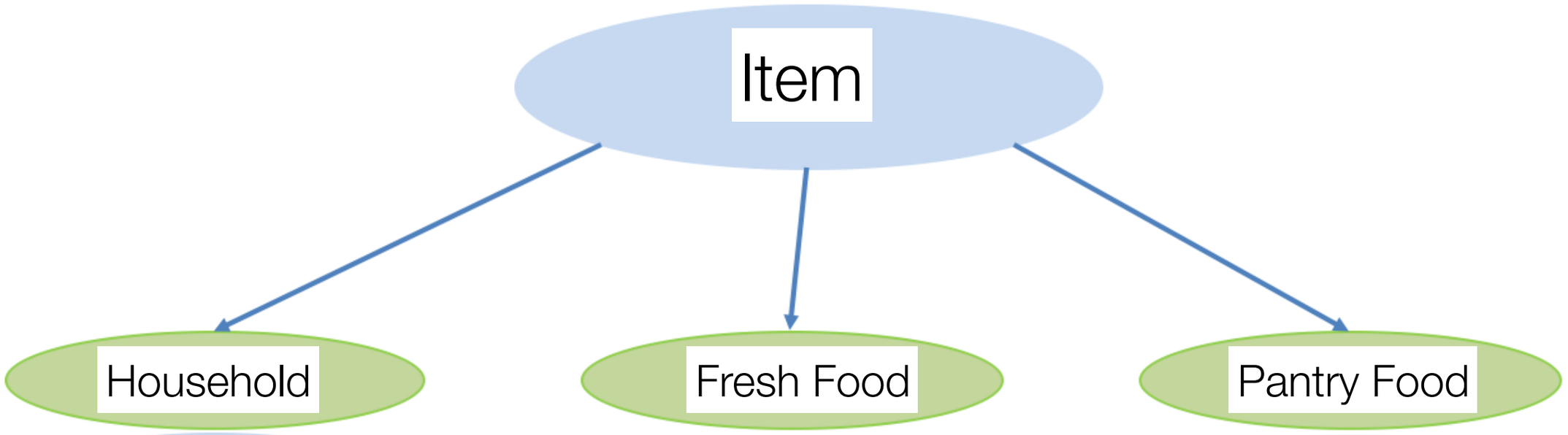
Transaction	Item	Price	Season	Store
1	Household	High	In-Season	Internet
2	Household	High	Off-Season	Internet
3	Fresh Food	High	In-Season	Offline
4	Pantry Food	High	In-Season	Offline
5	Pantry Food	Low	In-Season	Offline
6	Pantry Food	Low	Off-Season	Internet
7	Fresh Food	Low	Off-Season	Offline
8	Household	High	In-Season	Internet
9	Household	Low	In-Season	Offline
10	Pantry Food	Low	In-Season	Offline
11	Household	Low	Off-Season	Offline
12	Fresh Food	High	Off-Season	Offline
13	Fresh Food	Low	In-Season	Offline
14	Pantry Food	High	Off-Season	Internet
15	Pantry Food	High	Off-Season	???

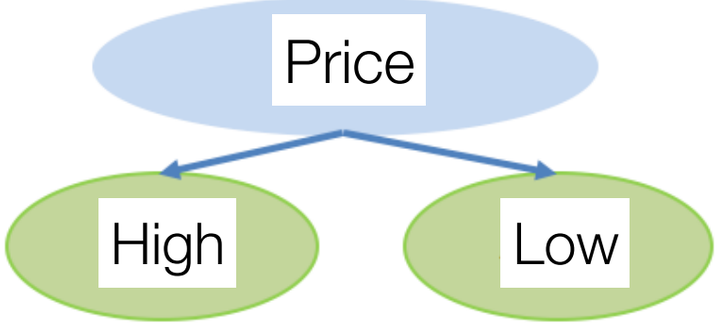
Supervised Learning

Decision Trees

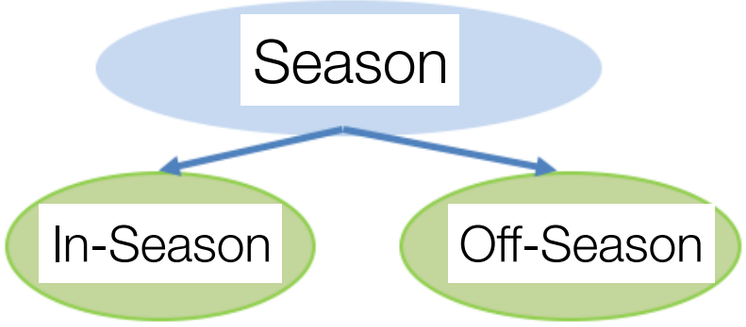


15 Pantry Food, High, Off-Season — Internet or Offline?





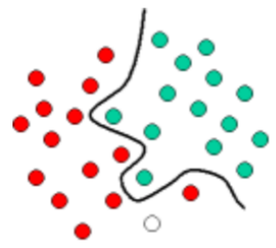
3,7,12,13 -- Offline



1,2,8 -- Internet 9,11 -- Offline

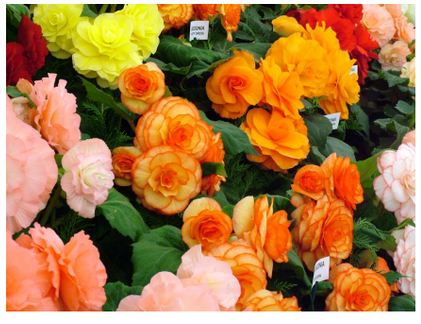
4,5,10 -- Offline 6,14 -- Internet

Supervised Learning | Decision Trees



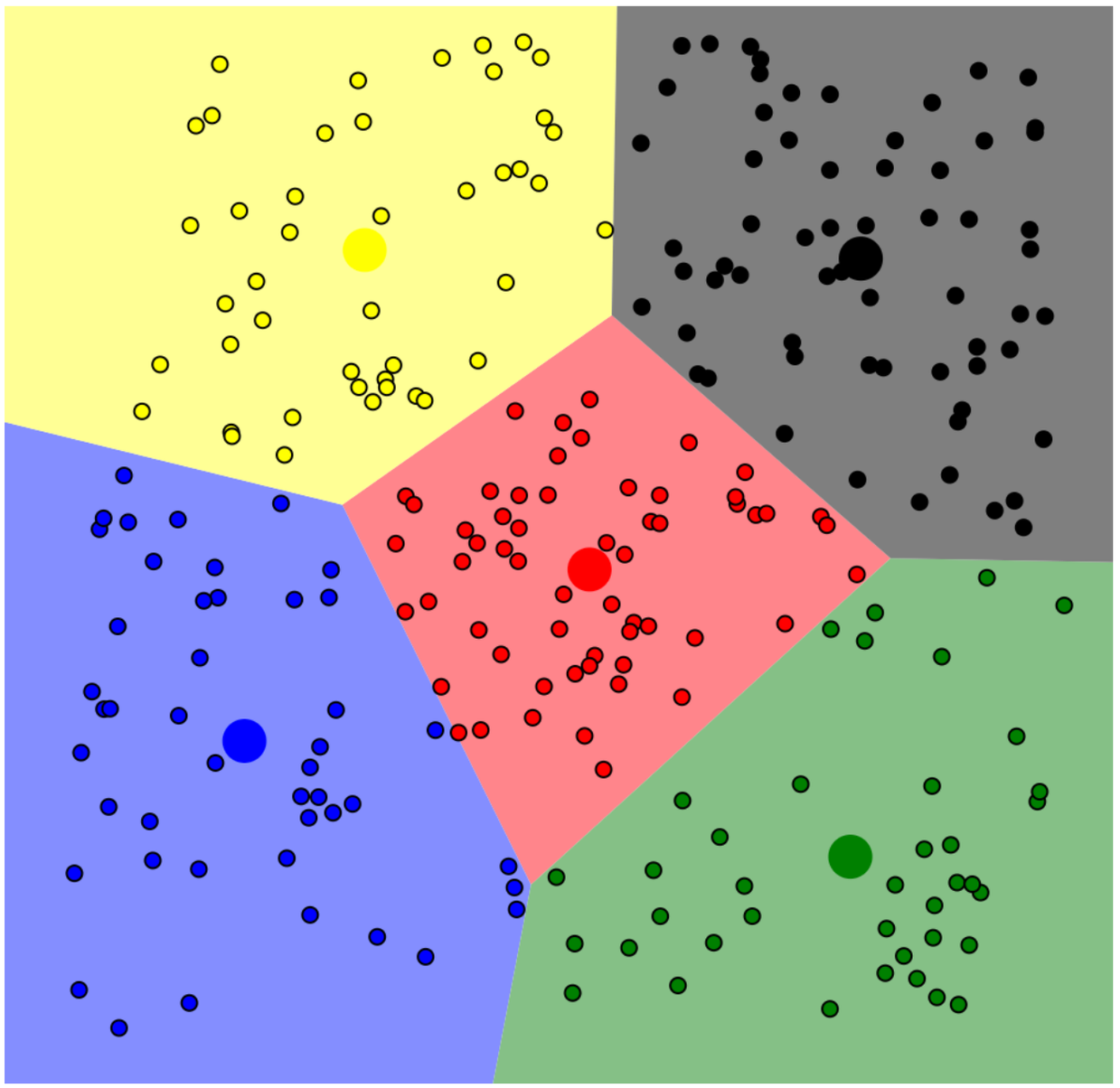
Unsupervised Learning

The Unlabelled Images



Unsupervised Learning

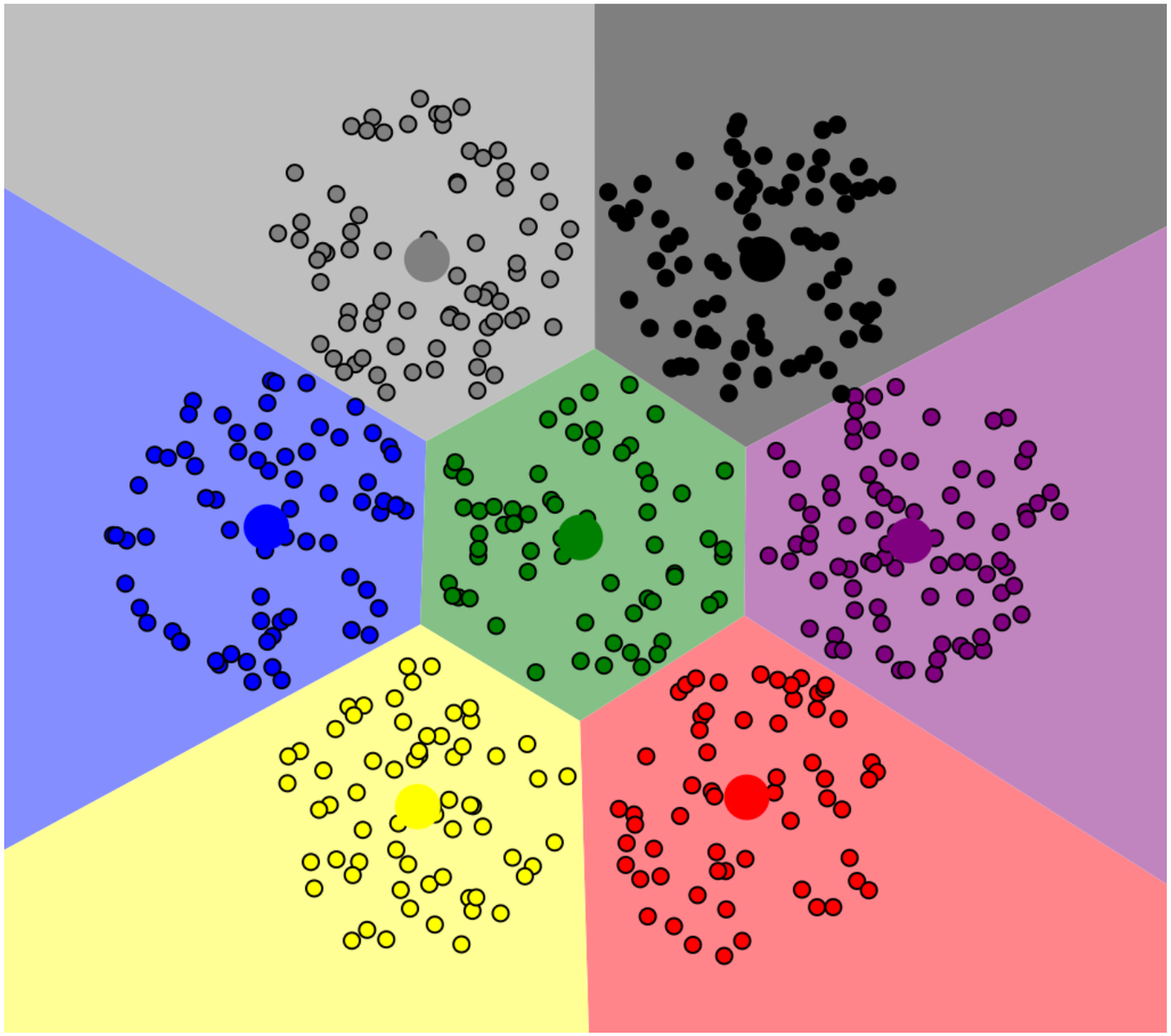
Groups of Similar Images



http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Unsupervised Learning

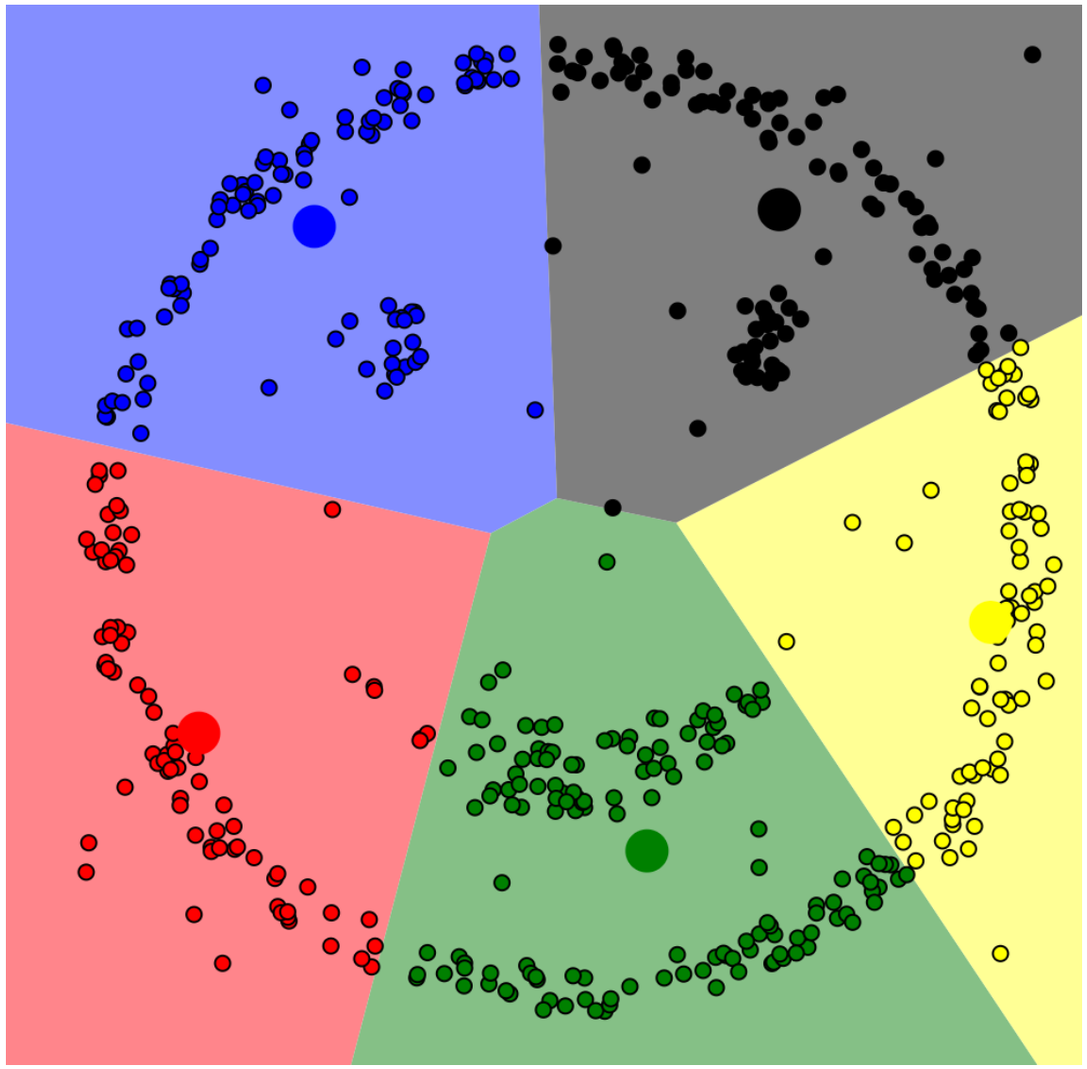
K-means clustering



http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Unsupervised Learning

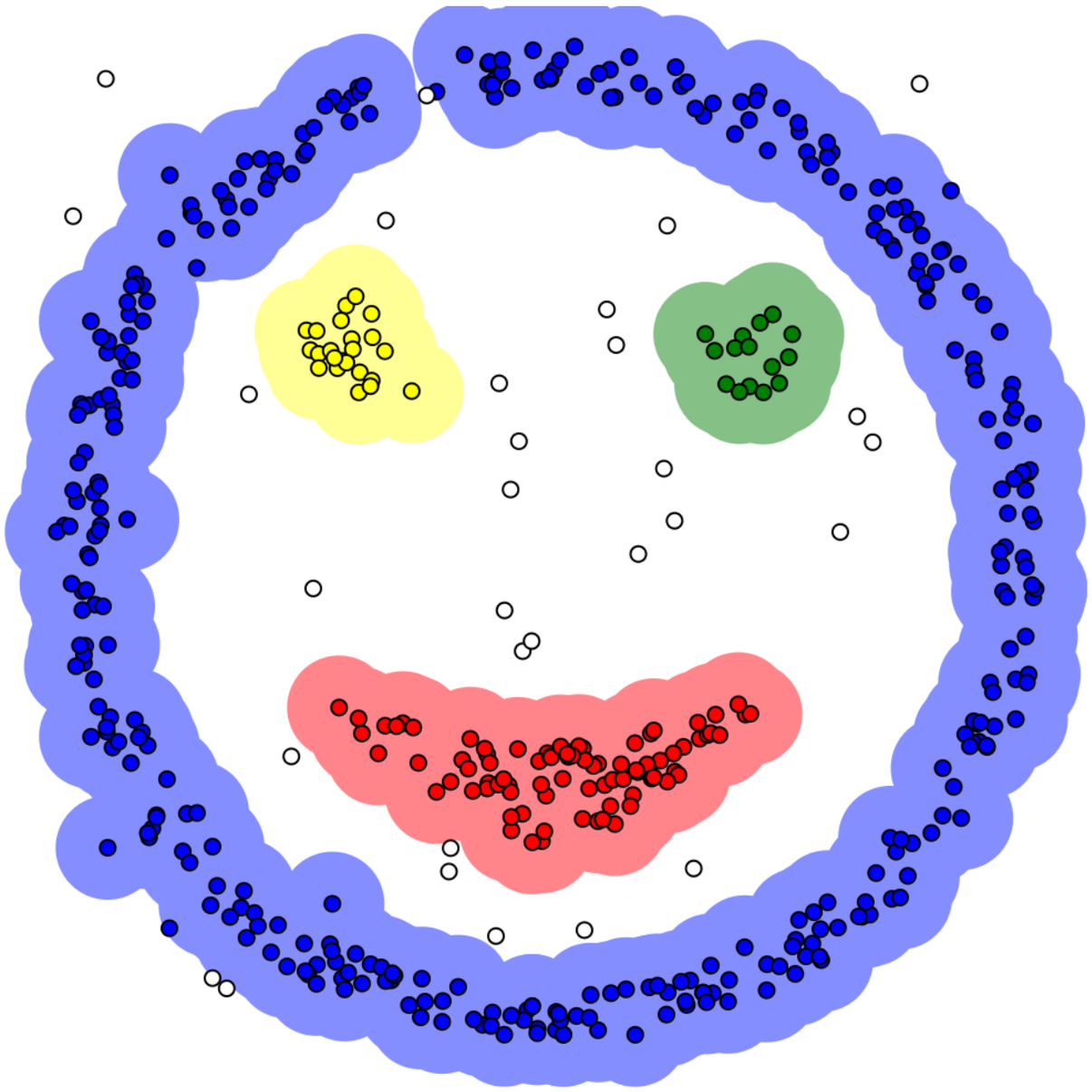
K-means clustering



http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Unsupervised Learning

K-means clustering



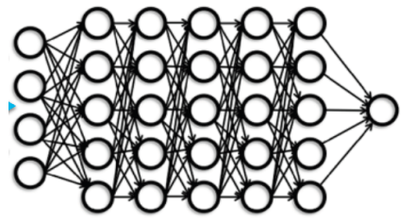
http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Unsupervised Learning

DB Scan

Representation Learning

Deep Neural Network



Science 28 July 2006:
Vol. 313 no. 5786 pp. 504–507
DOI: 10.1126/science.1127647

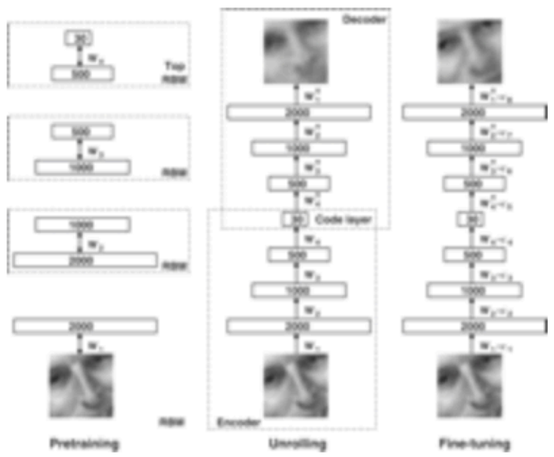
REPORT

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton*, R. R. Salakhutdinov

+ Author Affiliations

* To whom correspondence should be addressed; E-mail: hinton@cs.toronto.edu



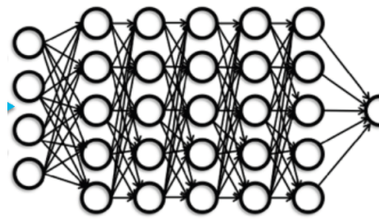
View larger version:
» In this page » In a new window
» Download PowerPoint Slide for Teaching

Fig. 1. .

Pretraining consists of learning a stack of restricted Boltzmann machines (RBMs), each having only one layer of feature detectors. The learned feature activations of one RBM are used as the "data" for training the next RBM in the stack. After the pretraining, the RBMs are "unrolled" to create a deep autoencoder, which is then fine-tuned using backpropagation of error derivatives.

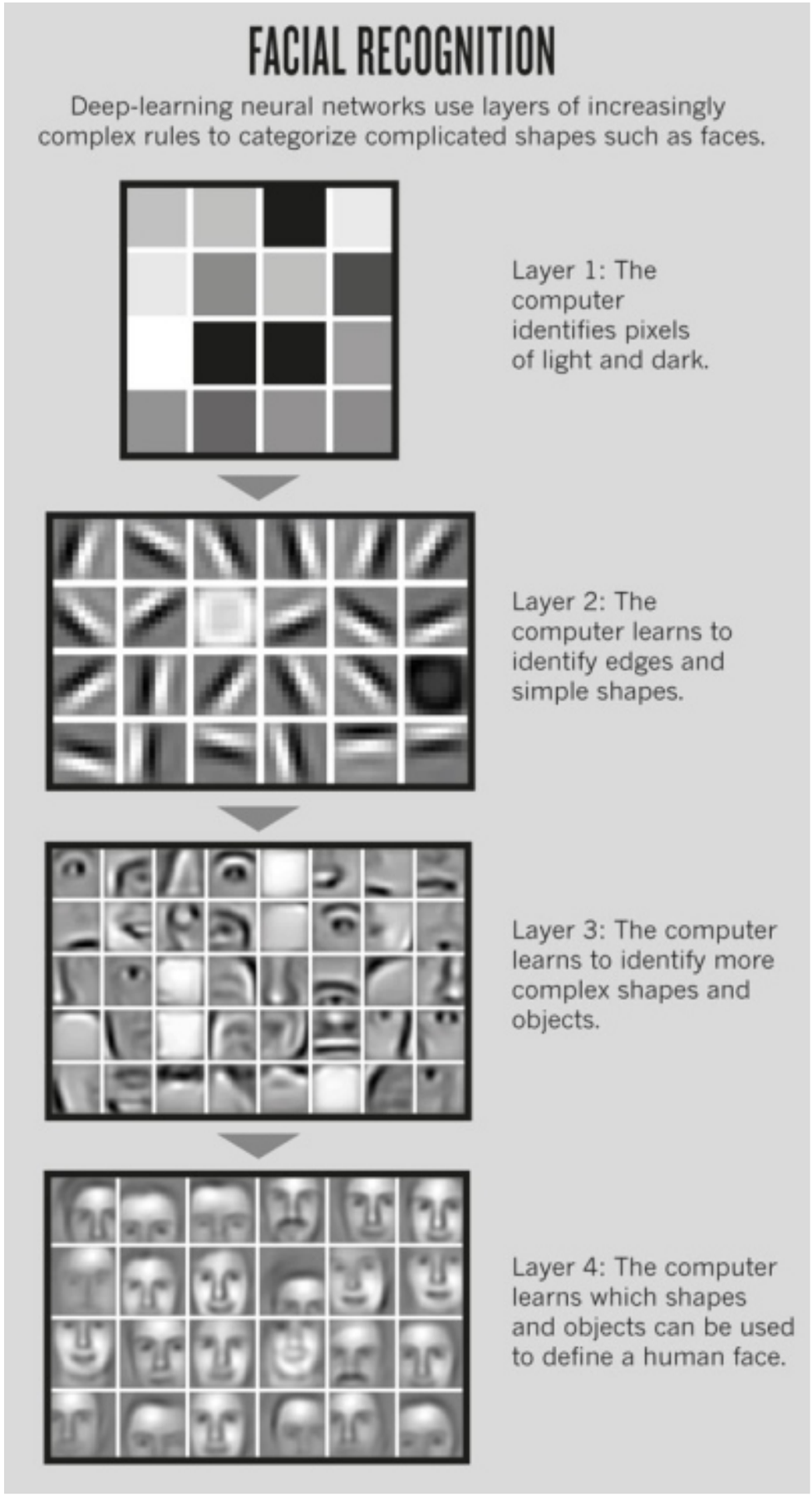
Representation Learning

Hinton & Salakhutdinov
Science, 2006



Representation Learning

Different Levels of Abstraction (Andrew Ng, Nature 2014)



Why now?

- Complexity of model is very high — requires huge datasets, advanced hardware with fast processors, large memory, high I/O speed

- Model is prone to overfitting — advanced algorithms to overcome overfitting are needed

- Parameter estimation is difficult — clever modifications to the model, as well as advanced algorithms to estimate parameters more quickly

Data & Infrastructure

THE MAN BEHIND THE GOOGLE BRAIN: ANDREW NG AND THE QUEST FOR THE NEW AI

Stanford professor Andrew Ng, the man at the center of the Deep Learning movement.
Photo: Ariel Zambelich/Wired

Famous AI Systems

May 12, 1997

IBM Chess Machine Beats Humanity's Champ

By BRUCE WEBER

NEW YORK -- In brisk and brutal fashion, the IBM computer Deep Blue unseated humanity, at least temporarily, as the finest chess playing entity on the planet on Sunday, when Garry Kasparov, the world chess champion, resigned the sixth and final game of the match after just 19 moves, saying, "I lost my fighting spirit."

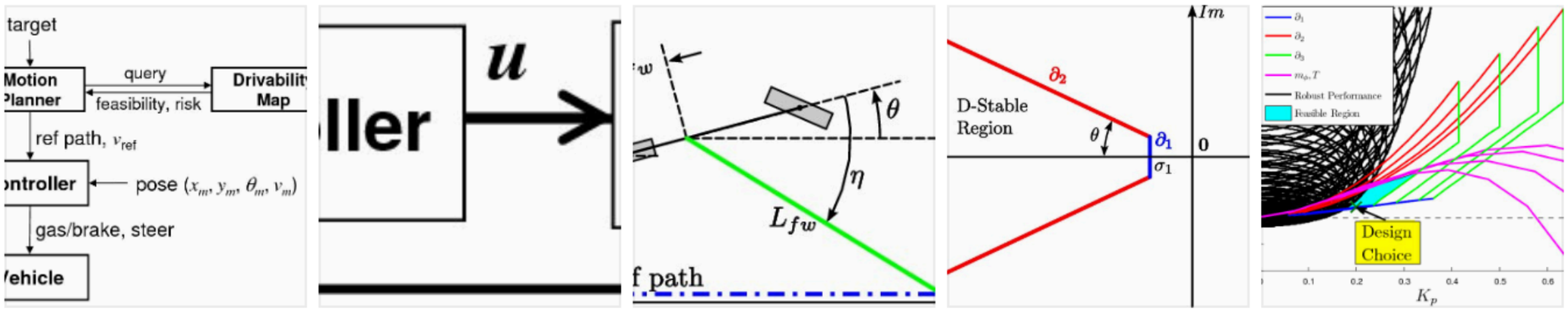
DARPA Urban Challenge 2007

http://www.youtube.com/watch?v=IULI63ERek0

DARPA Urban Challenge: Team MIT

Source: http://researchgate.net

Source publication



Motion Planning in Complex Environments Using Closed-loop Prediction

Article | Full-text available | Aug 2008

Yoshiaki Kuwata · Justin Teo · Sertac Karaman · [...] · Jonathan How

This paper describes the motion planning and control subsystems of Team MIT's entry in the 2007 DARPA Grand Challenge. The novelty is in the use of closed-loop prediction in the framework of Rapidly-exploring Random Tree (RRT). Unlike the standard RRT, an input to the controller is sampled, followed by the forward simulation using the vehicle model...

View

DARPA Urban Challenge: Team MIT

Source: http://researchgate.net

IBM Research Watson 2011

http://www.youtube.com/watch?v=FC3IryWr4c8

Computer Wins on 'Jeopardy!': Trivial, It's Not

By **JOHN MARKOFF** FEB. 16, 2011

See how this article appeared when it was originally published on NYTimes.com

IBM Watson Jeopardy | February 2011

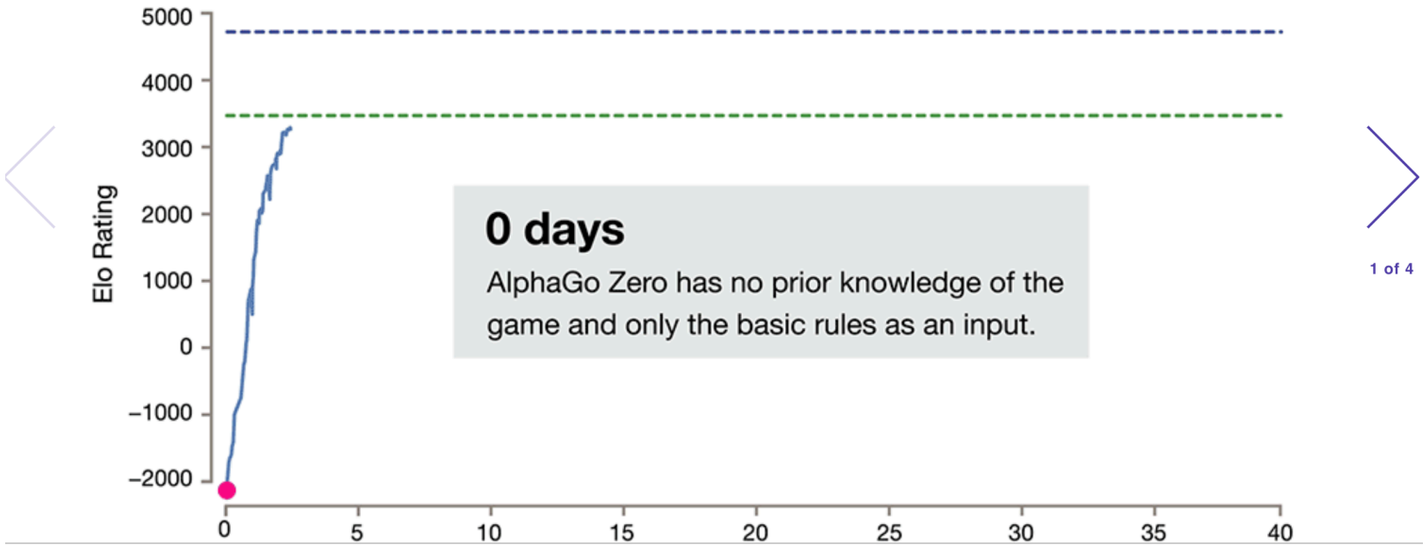
DeepMind AlphaGo 2016
vs. Lee Sedol

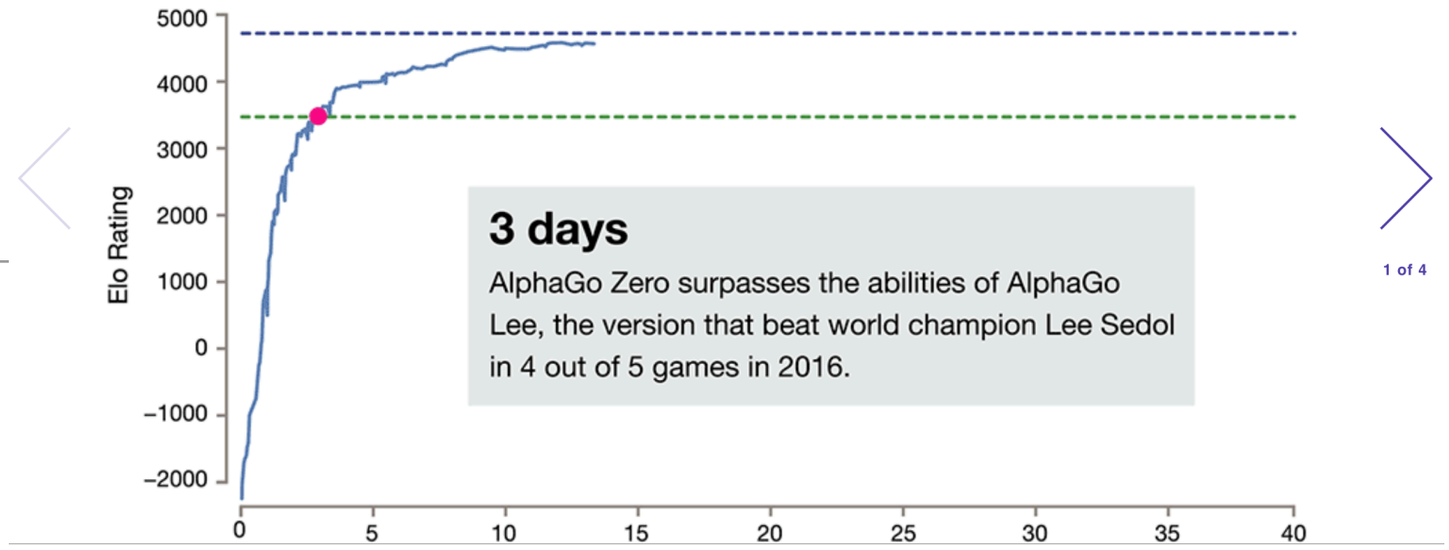
DeepMind AlphaGo Zero 2017

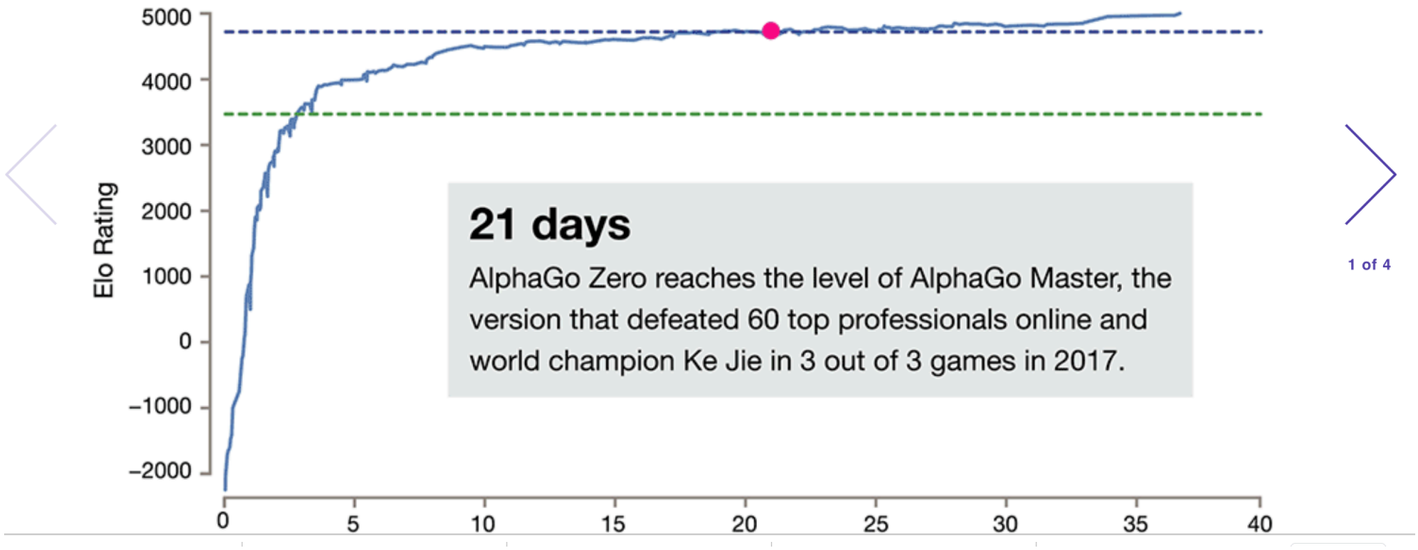
https://youtu.be/9xlSy9F5WtE

AlphaGo Zero

Trains without any data
Image Source:
www.deepmind.com

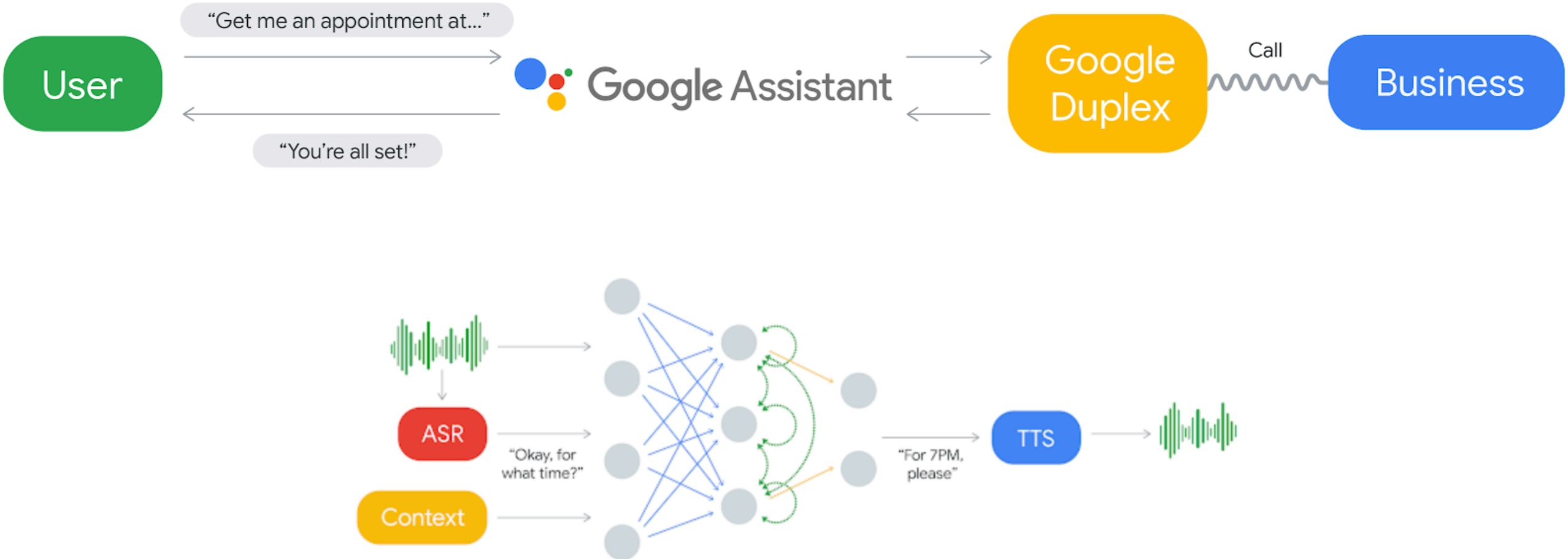






Google Duplex 2018

https://youtu.be/D5VN56jQMWM

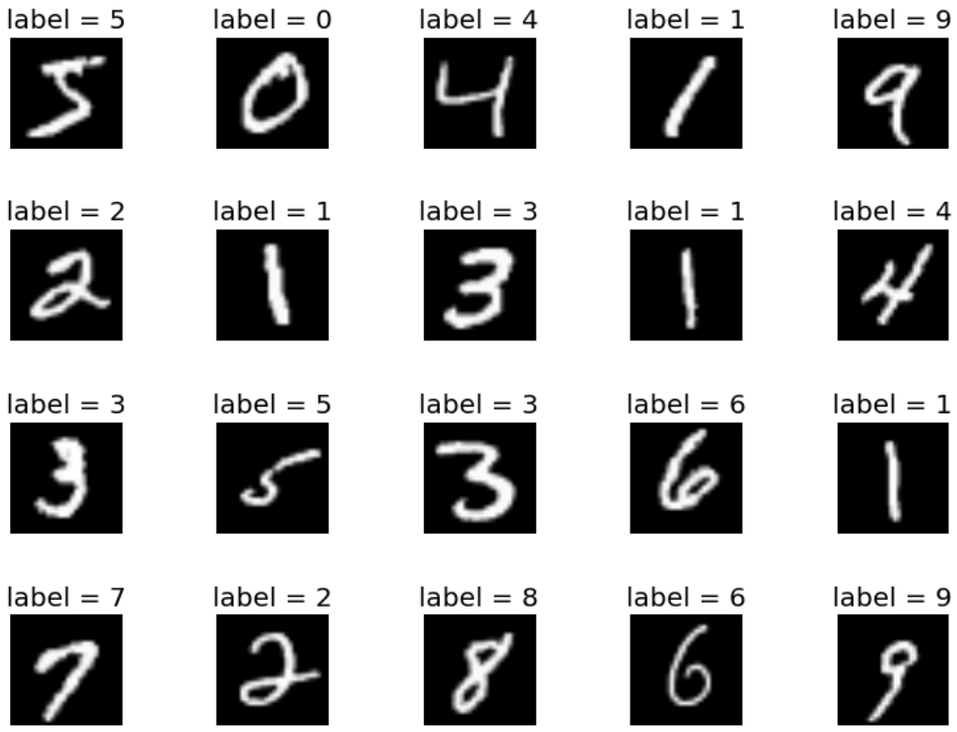


Google Duplex

Source: ai.googleblog.com

Major Domains within Artificial Intelligence & Corresponding Datasets

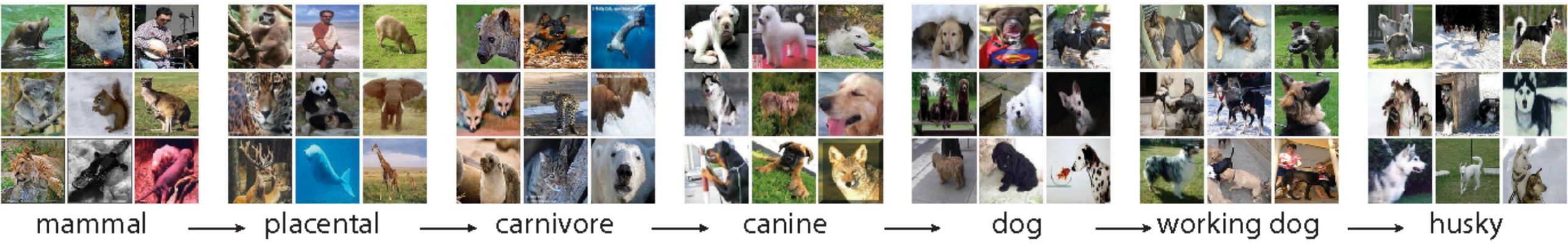
Visual Intelligence

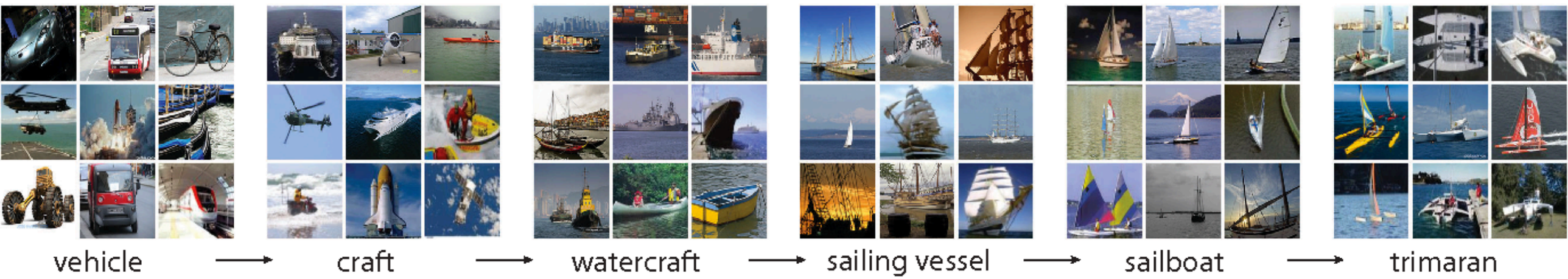


Source: https://corochann.com/mnist-dataset-introduction-1138.html

MNIST
http://yann.lecun.com/exdb/mnist/

Source: https://en.wikipedia.org/wiki/MNIST_database





Source: ImageNet: A large-scale hierarchical image database, Deng, et al., CVPR 2009

Visual Intelligence

ImageNet

Language Intelligence

SQUAD Dataset
Stanford Question Answer Dataset
Rajpurkar, et al. 2018

Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

Europarl Corpus

From Wikipedia, the free encyclopedia

The **Europarl Corpus** is a corpus (set of documents) that consists of the proceedings of the European Parliament from 1996 to the present. In its first release in 2001, it covered eleven official languages of the European Union (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish).[1] With the political expansion of the EU the official languages of the ten new member states have been added to the corpus data.[1] The latest release (2012)[2] comprised up to 60 million words per language with the newly added languages being slightly underrepresented as data for them is only available from 2007 onwards. This latest version includes 21 European languages: Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavic (Bulgarian, Czech, Polish, Slovak, Slovene), Finno-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek.[1]

The data that makes up the corpus was extracted from the website of the European Parliament and then prepared for linguistic research.[1] After sentence splitting and tokenization the sentences were aligned across languages with the help of an algorithm developed by Gale & Church (1993).[1]

The corpus has been compiled and expanded by a group of researchers led by Philipp Koehn at the University of Edinburgh. Initially, it was designed for research purposes in statistical machine translation (SMT). However, since its first release it has been used for multiple other research purposes, including for example word sense disambiguation. EUROPARL is also available to search via the corpus management system Sketch Engine.[3]

Source: https://en.wikipedia.org/wiki/Europarl_Corpus

Europarl Corpus | Machine Translation

gText United Nations, Department for General Assembly and Conference Management

United Nations Parallel Corpus

Introduction

The United Nations Parallel Corpus v1.0 is composed of official records and other parliamentary documents of the United Nations that are in the public domain. These documents are mostly available in the six official languages of the United Nations. The current version of the corpus contains content that was produced and manually translated between 1990 and 2014, including sentence-level alignments.

The corpus was created as part of the United Nations commitment to multilingualism and as a reaction to the growing importance of statistical machine translation (SMT) within the Department for General Assembly and Conference Management (DGACM) translation services and the United Nations SMT system, Tapta4UN.

The purpose of the corpus is to allow access to multilingual language resources and facilitate research and progress

Source: https://conferences.unite.un.org/UNCorpus

UN Parallel Corpus | Machine Translation

Welcome to GLUE

The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems. GLUE consists of:

- A benchmark of nine sentence- or sentence-pair language understanding tasks built on established existing datasets and selected to cover a diverse range of dataset sizes, text genres, and degrees of difficulty,
- A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language, and
- A public leaderboard for tracking performance on the benchmark and a dashboard for visualizing the performance of models on the diagnostic set.

GLUE Benchmark | http://www.gluebenchmark.com

Thank you

Alice Oh
alice.oh@kaist.edu