Cloud Computing & Big Data

PARALLEL & SCALABLE MACHINE LEARNING & DEEP LEARNING

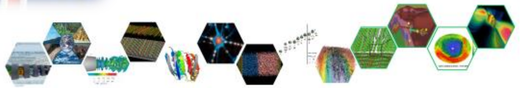
Ph.D. Student Chadi Barakat

School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland

Juelich Supercomputing Centre, Forschungszentrum Juelich, Germany

LECTURE 11

IHPC National Competence Center for HPC & AI in Iceland



@ProfDrMorrisRiedel @Morris Riedel @MorrisRiedel @MorrisRiedel

https://www.youtube.com/channel/UCWC4VKHmL4NZgFfKoHtANKg

morris@hi.is

Big Data Analytics & Cloud Data Mining

November 2, 2021

Online Lecture

EUROPEAN OPEN SCIENCE CLOUD

EOSC NORDIC

EuroHPC Joint Undertaking

RAISE Center of Excellence

EURO

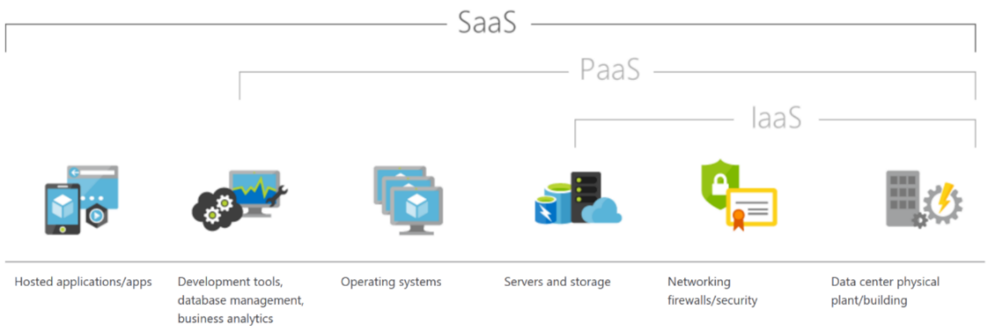
UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

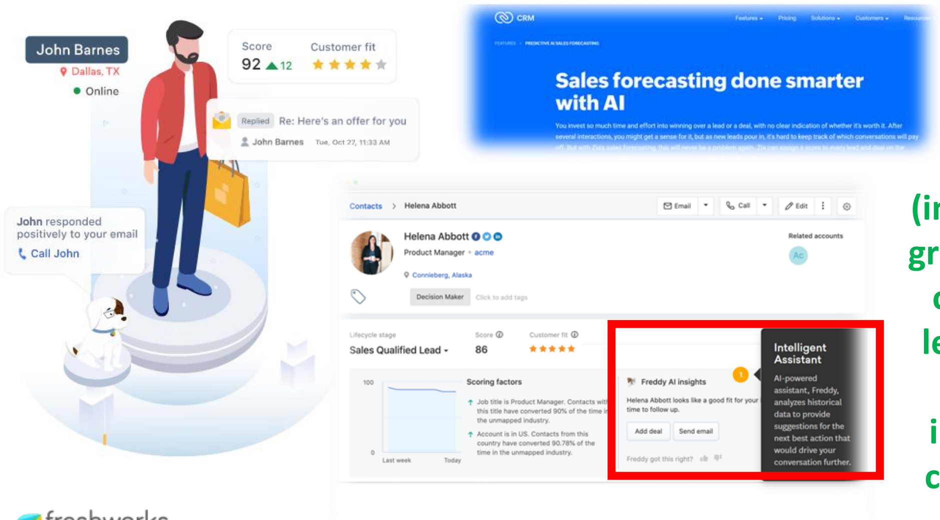
ADMIRE malleable data solutions for HPC

HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

Review of Lecture 10 –Software-As-A-Service (SAAS)



- SAAS Examples of Customer Relationship Management (CRM) Applications



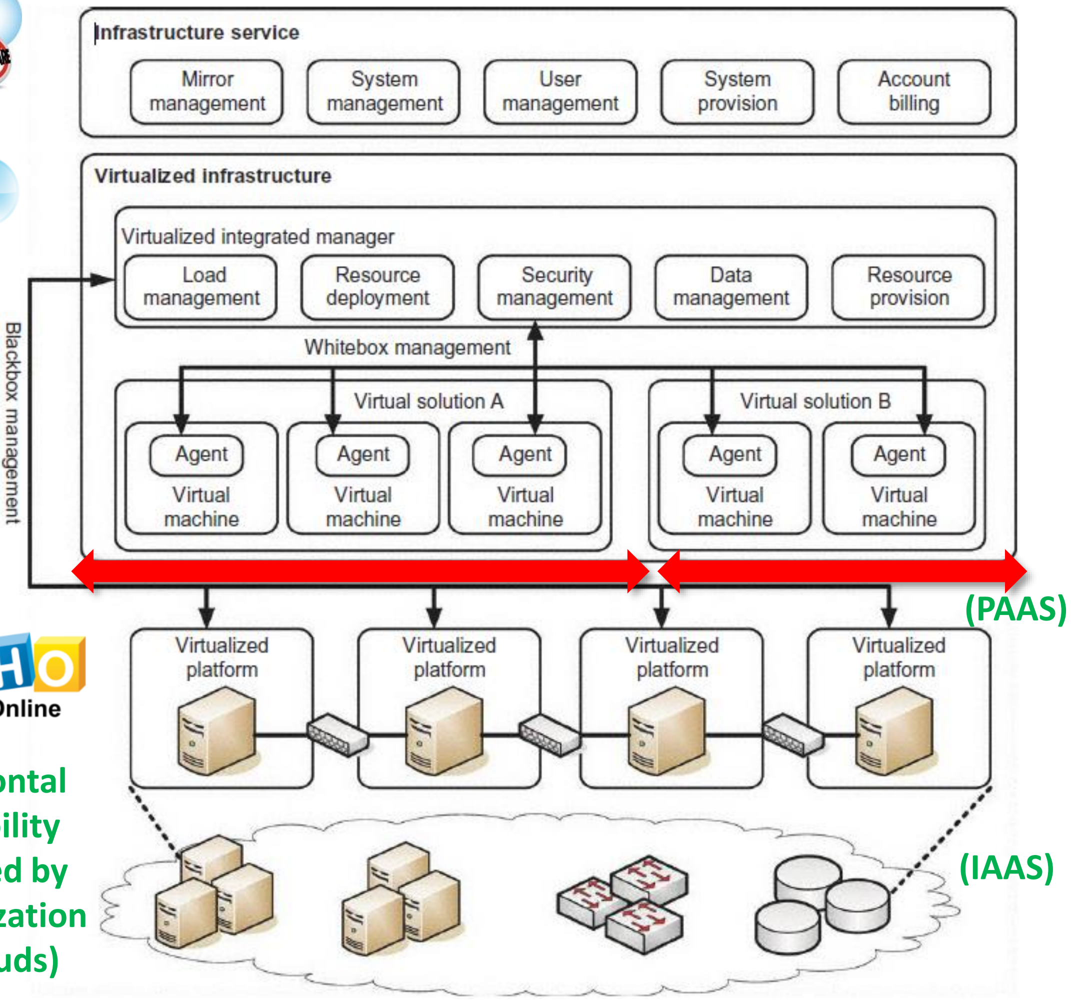
freshworks

freshworks CRM

(introducing a growing range of machine learning and artificial intelligence capabilities)

Work · Online

(horizontal scalability enabled by Virtualization in Clouds)



[5] AWS Sagemaker *[4] Freshworks Web page* *[3] ZOHO CRM Web page* *modfied from [2] Distributed & Cloud Computing Book* *[1] Microsoft Azure SAAS*

Outline of the Course

1. Cloud Computing & Big Data Introduction
2. Machine Learning Models in Clouds
3. Apache Spark for Cloud Applications
4. Virtualization & Data Center Design
5. Map-Reduce Computing Paradigm
6. Deep Learning driven by Big Data
7. Deep Learning Applications in Clouds
8. Infrastructure-As-A-Service (IAAS)
9. Platform-As-A-Service (PAAS)
10. Software-As-A-Service (SAAS)
11. Big Data Analytics & Cloud Data Mining
12. Docker & Container Management
13. OpenStack Cloud Operating System
14. Online Social Networking & Graph Databases
15. Big Data Streaming Tools & Applications
16. Epilogue

+ additional practical lectures & Webinars for our hands-on assignments in context

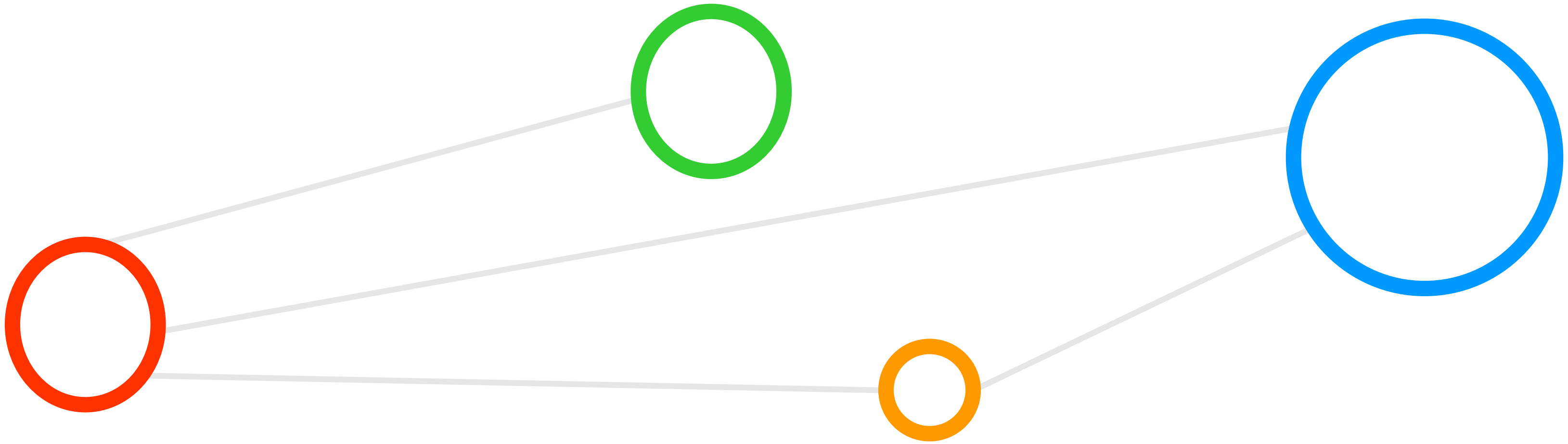
- Practical Topics
- Theoretical / Conceptual Topics

Lecture Outline

- Association Rule Mining in Big Data
  - What is Association Rule Mining
  - The On4Off project and the needs of the Retail Industry
  - The Apriori and FP-Growth algorithms
  - The challenges we face and how to overcome them
  - How this applies to the Cloud (Amazon SageMaker, Google Cloud ML, Azure Machine Learning…)
- Deep Learning for Label Generation
  - Refresher: what is Deep Learning?
  - The problem we're trying to solve in On4Off
  - CNNs, Residual Networks, and Transfer Learning
  - Dealing with bad pictures and sparse data: Data Augmentation
  - Colour detection and classification
- Data Mining in Healthcare

- **Promises from previous lecture(s):**
- ***Practical Lecture 0.1:*** **Lecture 11 will provide more insights into how the algorithm works & how they scale for big datasets in cloud computing environments**
- ***Practical Lecture 0.1:*** **Lecture 11 will provide insights about using the real dataset of the pieper perfume stores and its big data mining processing challenges**
- ***Practical Lecture 0.1:*** **Lecture 11 will provide more insights how to use configuration options in data mining algorithms to perform fine tuned data analysis**
- ***Practical Lecture 3.1:*** **Lecture 11 provides more details on using recommender engines & that are partly considered as data mining technique**
- ***Lecture 5:*** **Lecture 11 will provide more details on data analytics techniques using parallel computing for data mining applications in Clouds today**

Association Rule Mining in Big Data



Association Rule Mining (1)

- Methodology
  - Sometimes referred to as simply ‘Association Rules’
  - Used to discover unknown relationships hidden in datasets
  - Rules refer to a set of identified frequent itemsets that represent the uncovered relationships in datasets
  - Identify rules that will predict the occurrence of one or more items based on the occurrences of other items in the dataset

[6] Big Data Tips, Association Rules

- Approach
  - Unsupervised machine learning method
  - No direct guiding output data is given to find the patterns
  - Several algorithms exist to perform association rule mining (e.g., Apriori, FP Growth, MAFIA, etc.)

- **The discovery of association rules fundamentally depends on the discovery of <u>frequent itemsets</u>**
- **Frequent itemset means finding sets of items that appear in (or are related to) many "baskets"**

Association Rule Mining (2)

[6] Big Data Tips, Association Rules

- Famous Example in Retail
  - Illustrating a rule based on a strong relationship between the sale of Diapers and the sale of Beer
  - Many customers who buy Diapers also buy Beer
  - Investigating the transactions to find those frequent itemsets seems to be easy
- Challenges
  - In real datasets millions or billions of transactions are searched
  - Transaction search across 100000 of different items that may identify 1000 of rules
- Algorithms Benefit
  - Automation of the process using association rule mining algorithms.
  - Rules help to identify new opportunities and ways for cross-selling products to customers

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

market basket transactions

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

The On4Off Project

- Commercial Environments
  - Large quantities of data are accumulated in databases from day-to-day operations
  - Lays the foundation for mining association rules: no data – no association rule mining!
- Retail Example
  - Customer purchase data are collected on a daily basis at the checkout counters of city stores or when shopping at online stores
  - Accumulated data items are often market basket transactions
- Motivation to Collect and Analyze Data
  - Managers of stores are interested in analyzing the collected data in order to learn the purchasing behaviour of customers
  - Enables a large variety of business-related applications based on the identified rules in the data (to be reviewed from store managers!)

[7] German ON4OFF project

[6] Big Data Tips, Association Rules

The Apriori Algorithm (1)

- The most commonly used algorithm for Association Rule Mining.
- For a given list of shopping baskets, the algorithm collects a list of all items and compares their frequency to the set minimum support value.
- After pruning the non-frequent items, the algorithm builds two-itemsets from the remaining items and compares their frequency to the minimum support value.
- Frequent two-itemsets (achieve minimum support) are then used to build three-itemsets (if possible) with the frequent one-itemsets, and so forth.

TID	Items
1	{Bread,Milk}
2	{Bread,Diapers,Beer,Eggs}
3	{Milk,Diapers,Beer,Cola}
4	{Bread,Milk,Diapers,Beer}
5	{Bread,Milk,Diapers,Cola}

Item	Count
Beer	3
Bread	4
Cola	2
Diapers	4
Milk	4
Eggs	1

2-Itemset	Count
{Beer,Bread}	2
{Beer,Diapers}	3
{Beer,Milk}	2
{Bread,Diapers}	3
{Bread,Milk}	3
{Diapers,Milk}	3

3-Itemset	Count
{Bread,Milk,Diapers}	3

Example Transactions and frequent Itemset generation in Apriori [8]

- **Apriori is the most commonly used algorithm for Association Rule Mining.**
- **It can create multi-itemsets by iteratively going through the transactions.**

The Apriori Algorithm (2)

[10] MLxtend Lib, Association Rules

- Using the Apriori algorithm to generate frequent itemsets for association rules is as simple as implementing functions from the MLxtend module. [9]
- Initially, a one-hot encoded list is created from the list of transactions, then it is fed to the algorithm along with a value for min_support (in this case 0.01).
- The generated candidate itemsets can range in size from 1 to K-1 where K is the total number of unique items.

```
[['HugoBoss', 'Chanel', 'Rituals', 'Armani', 'Biotherm', 'Lancome', 'Dior'],
 ['HugoBoss', 'Biotherm', 'Armani', 'Dior', 'Chanel', 'Clinique', 'Lancome'],
 ['Biotherm', 'Armani', 'Rituals', 'Clinique', 'Chanel'],
 ['Biotherm', 'Armani', 'Rituals', 'Clinique', 'Chanel'],
 ['Armani', 'Lancome', 'Chanel'],
 ['Rituals', 'HugoBoss', 'Chanel', 'Lancome', 'Clinique', 'Dior', 'Clinique'],
 ['Chanel', 'Lancome', 'Rituals', 'Biotherm'],
 ['Clinique', 'HugoBoss', 'Clinique', 'Lancome', 'Dior'],
 ['HugoBoss', 'Dior', 'Biotherm', 'Clinique', 'Chanel', 'Rituals'],
 ['Clinique', 'Chanel', 'Dior', 'Armani', 'Lancome', 'Rituals', 'Biotherm'],
 ['Biotherm', 'Armani', 'Rituals', 'Clinique', 'Chanel'],
 ['Clinique', 'Rituals', 'Armani', 'HugoBoss', 'Dior', 'Chanel', 'Clinique'],
 ['HugoBoss', 'Dior', 'Lancome', 'Clinique', 'Armani'],
 ['Clinique', 'Biotherm', 'Clinique', 'Armani'],
 ['HugoBoss'],
 ['Lancome', 'Dior', 'Clinique'],
 ['Armani', 'Clinique', 'HugoBoss'],
 ['HugoBoss', 'Clinique', 'Clinique'],
 ['Dior'],
 ['Clinique', 'Biotherm', 'Armani', 'HugoBoss'
 ['Armani', 'Lancome', 'Biotherm', 'HugoBoss']
 ['Biotherm',
  'Rituals',
  'Chanel',
  'HugoBoss',
  'Clinique',
  'Lancome',
  'Armani'],
```

	HugoBoss	Chanel	Rituals	Armani	Biotherm	Lancome	Dior	Clinique
0	True	True	True	False	True	True	True	True
1	True	True	True	True	True	True	True	False
2	True	True	True	True	False	False	False	True
3	True	True	True	True	False	False	False	True
4	True	False	True	False	False	False	True	False
5	False	False	True	True	True	True	True	True
6	False	True	True	False	False	False	True	True
7	False	False	False	True	True	True	True	False
8	False	True	True	True	True	True	False	True
9	True	True	True	True	True	False	True	True
10	True	True	True	True	False	False	False	True
11	True	False	True	True	True	True	False	True
12	True	False	False	True	True	True	True	False
13	True	True	False	True	False	False	False	False
14	False	False	False	False	False	True	False	False
15	False	False	False	True	True	False	True	False
16	True	False	False	True	False	True	False	False
17	False	False	False	True	False	True	False	False
18	False	False	False	False	True	False	False	False
19	True	True	False	True	True	True	False	True
20	True	True	False	False	False	True	True	False
21	True	True	True	True	False	True	True	True

	support	itemsets
0	0.476190	(HugoBoss)
1	0.501587	(Chanel)
2	0.438095	(Rituals)
3	0.701587	(Armani)
4	0.406349	(Biotherm)
...	...	...
241	0.012698	(Lancome, Chanel, Clinique, Dior, Biotherm, Ar...
242	0.012698	(Lancome, Clinique, Rituals, Dior, Biotherm, A...
243	0.015873	(Lancome, HugoBoss, Chanel, Rituals, Dior, Bio...
244	0.012698	(HugoBoss, Chanel, Clinique, Rituals, Dior, Bi...
245	0.015873	(Lancome, HugoBoss, Chanel, Clinique, Rituals,...

246 rows × 2 columns

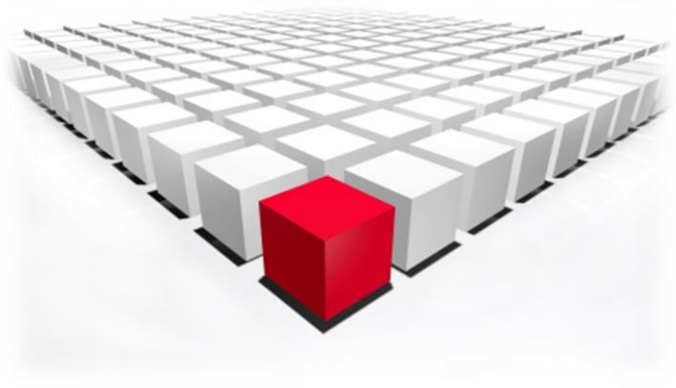
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(HugoBoss)	(Chanel)	0.476190	0.501587	0.323810	0.680000	1.355696	0.084958	1.557540
1	(Chanel)	(HugoBoss)	0.501587	0.476190	0.323810	0.645570	1.355696	0.084958	1.477891
2	(HugoBoss)	(Armani)	0.476190	0.701587	0.358730	0.753333	1.073756	0.024641	1.209781
3	(Clinique)	(HugoBoss)	0.438095	0.476190	0.266667	0.608696	1.278261	0.058050	1.338624
4	(Rituals)	(Chanel)	0.438095	0.501587	0.269841	0.615942	1.227986	0.050098	1.297754
...	...	...	...	...	...	...	...	...	...
375	(HugoBoss, Clinique, Rituals, Dior, Biotherm, ...	(Chanel)	0.012698	0.501587	0.012698	1.000000	1.993671	0.006329	inf
376	(HugoBoss, Clinique, Dior, Biotherm, Armani)	(Chanel, Rituals)	0.019048	0.269841	0.012698	0.666667	2.470588	0.007559	2.190476
377	(Lancome, HugoBoss, Clinique, Chanel, Rituals,...	(Armani)	0.025397	0.701587	0.015873	0.625000	0.890837	-0.001945	0.795767
378	(Lancome, HugoBoss, Clinique, Chanel, Dior, Ar...	(Rituals)	0.025397	0.438095	0.015873	0.625000	1.426630	0.004747	1.498413
379	(Lancome, HugoBoss, Clinique, Rituals, Dior, A...	(Chanel)	0.022222	0.501587	0.015873	0.714286	1.424051	0.004727	1.744444

380 rows × 9 columns

- **MLxtend is an extremely useful python library for implementing Apriori and other machine learning algorithms.**

The Apriori Algorithm (3)

- For a relatively small number of transactions, Apriori is an effective and thorough method of extracting candidate itemsets.
- Since it has to make several passes on the whole dataset, first to extract all the unique items and their support, then to iteratively analyse the interactions of each subsequent itemset with other items, the algorithm requires a lot of resources to run over larger datasets.
- All of this before even pruning and generating association rules.



As the data scales up, the problem becomes more and more complex and the computations more costly

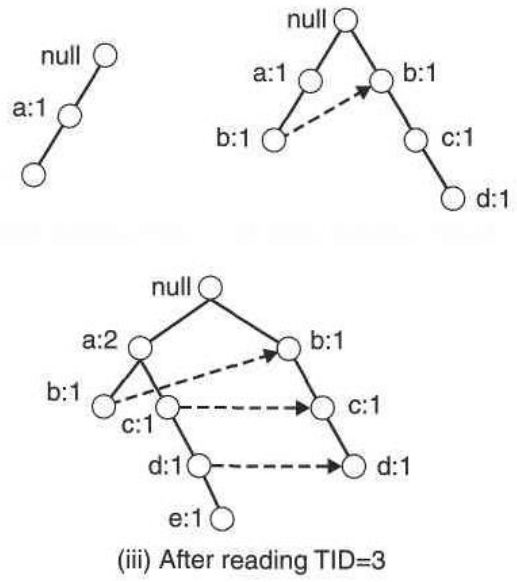
- **Take into consideration computation cost when planning your machine learning approach: Apriori may not be the best approach for larger datasets.**

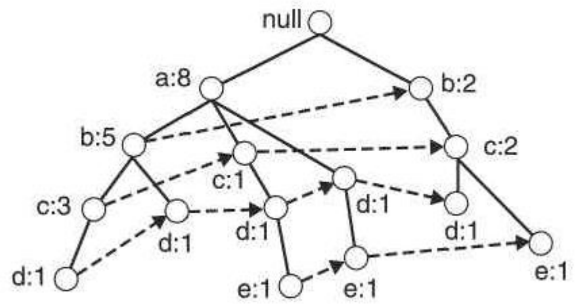
The FP-Growth Algorithm (1)

- Approach is radically different from Apriori.
- Instead of recursively scanning the transactions, the Frequent Pattern (FP)-growth algorithm generates first a list of 1-itemsets and sorts them by order of support, then builds a FP-Tree.
- The Tree encodes the information contained in the transactions and automatically highlights the most frequent itemsets.
- This data can be extracted directly from the tree instead of making multiple passes over the data.

- **Instead of doing multiple passes over the dataset, FP-Growth builds a Frequent Pattern Tree from which the frequent itemsets can be easily extracted.**

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}





Constructing the FP-Tree [8]

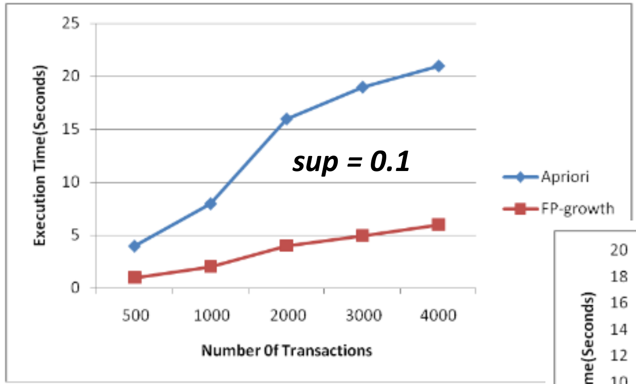
The FP-Growth Algorithm (2)

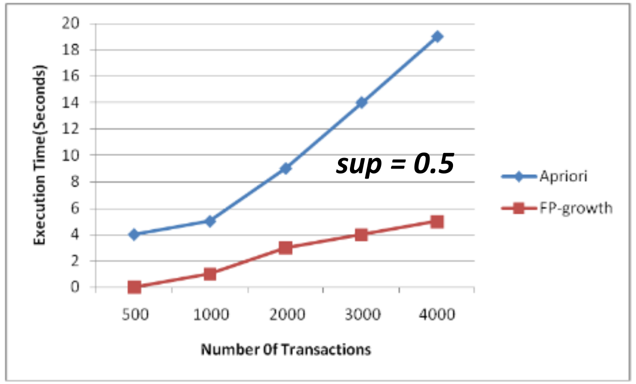
[10] MLxtend Lib, Association Rules

- Similarly to Apriori, the transactions are binarized before being fed into the algorithm [11].
- Generating the list of frequent itemsets is significantly faster in FP-Growth.

	support	itemsets
0	0.504762	(Lancome)
1	0.501587	(Chanel)
2	0.476190	(HugoBoss)
3	0.438095	(Clinique)
4	0.438095	(Rituals)
...	...	...
241	0.022222	(HugoBoss, Chanel, Clinique, Rituals, Dior, Bi...
242	0.012698	(HugoBoss, Chanel, Clinique, Dior, Biotherm, A...
243	0.012698	(HugoBoss, Clinique, Rituals, Dior, Biotherm, ...
244	0.012698	(HugoBoss, Chanel, Clinique, Rituals, Dior, Bi...
245	0.041270	(HugoBoss, Clinique, Rituals, Dior, Armani)

246 rows × 2 columns





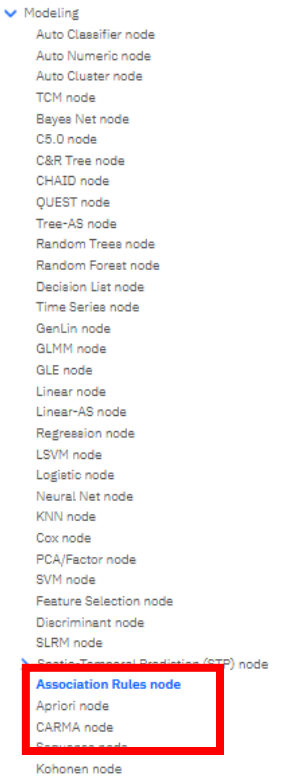
[12] FP-Growth vs Apriori speedup

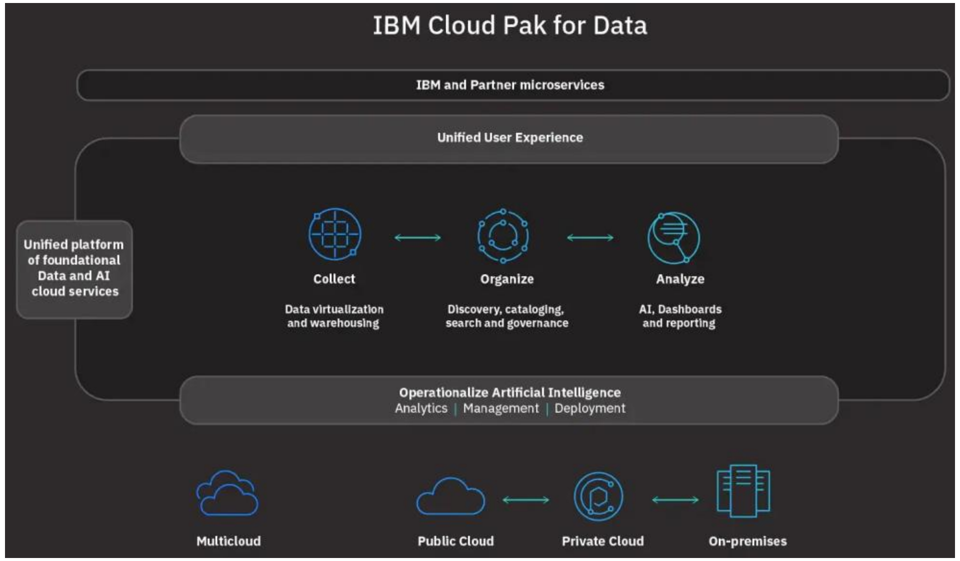
```
from mlxtend.frequent_patterns import apriori

%timeit -n 100 -r 10 apriori(df, min_support=0.6)

3.16 ms ± 66.3 µs per loop (mean ± std. dev. of 10 runs, 100 loops each)

%timeit -n 100 -r 10 apriori(df, min_support=0.6, low_memory=True)

3.43 ms ± 131 µs per loop (mean ± std. dev. of 10 runs, 100 loops each)

from mlxtend.frequent_patterns import fpgrowth

%timeit -n 100 -r 10 fpgrowth(df, min_support=0.6)

1.24 ms ± 38.7 µs per loop (mean ± std. dev. of 10 runs, 100 loops each)
```

[11] MLxtend Lib, FP-Growth

Example: IBM Cloud Pak for Data – SPSS Modeler & Association Rule Mining

- Approach & Association Rule Mining
  - Fully-integrated data & AI platform
  - Builds on RedHat OpenShift Platform (based on Kubernetes)
  - OpenShift enables to run the IBM Cloud Pak for Data platform on Clouds like IBM Cloud, AWS, Google Cloud, or MS Azure
  - Kubernetes is an open source, extensible container orchestrator for Cloud systems
  - IBM SPSS Modeler & data mining nodes





[26] IBM Cloud Pak for Data

[27] RedHat OpenShift

[28] Kubernetes

Apriori node

Last updated: Oct 23, 2020

The Apriori node discovers association rules in your data.

Association rules are statements of the form:

```
if antecedent(s) then consequent(s)
```

CARMA node

Last updated: Oct 23, 2020

The CARMA node uses an association rules discovery algorithm to discover association rules in the data.

Association rules are statements in the form:

```
if antecedent(s) then consequent(s)
```

Association Rules node

Last updated: Oct 23, 2020

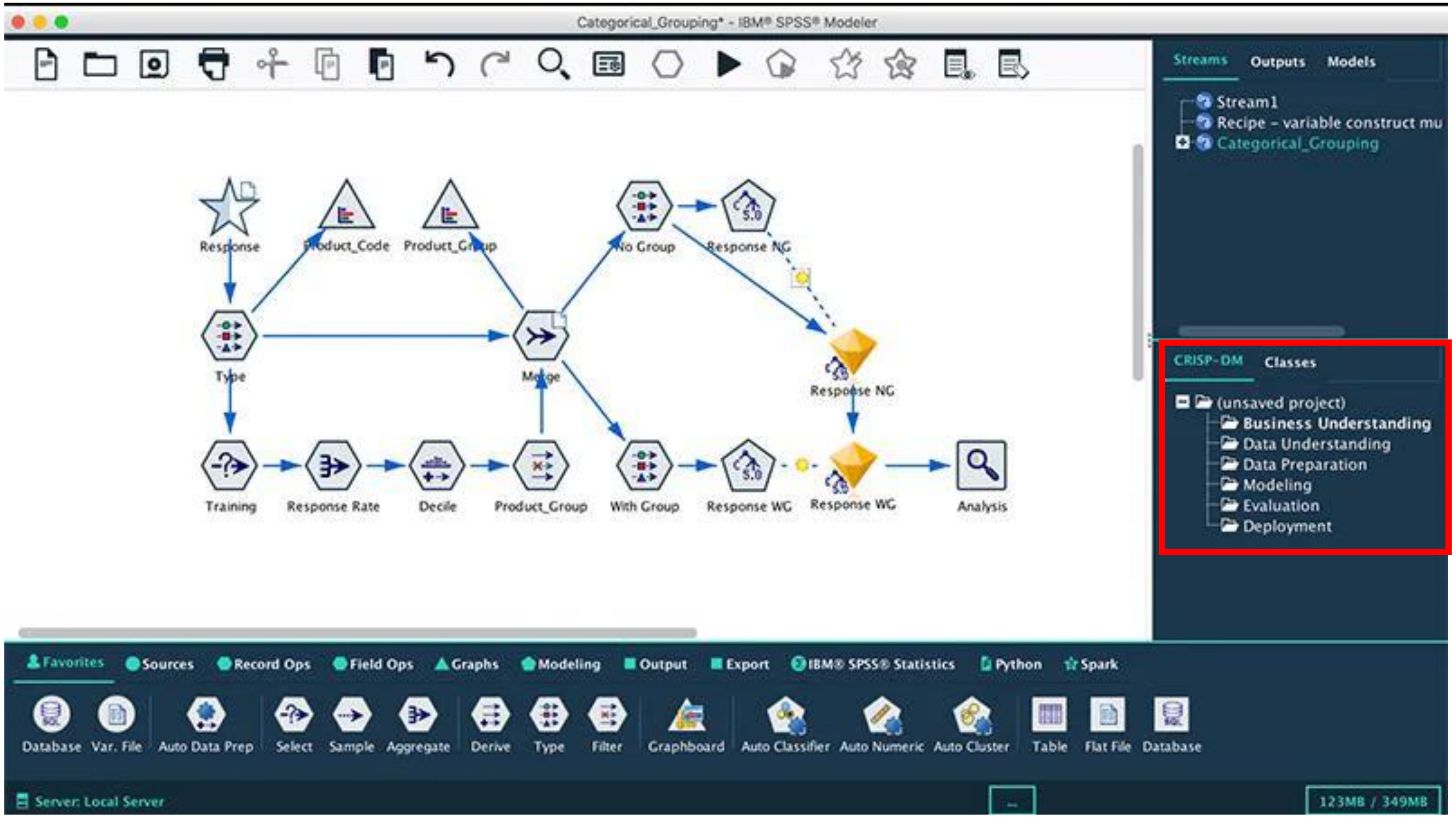
Association rules associate a particular conclusion (the purchase of a particular product, for example)

For example, the rule

```
beer <= cannedveg & frozenmeal (173, 17.0%, 0.84)
```

➢ **Lecture 12 will provide more insights about the importance of containers such as Singularity or Docker and introduces Kubernetes**

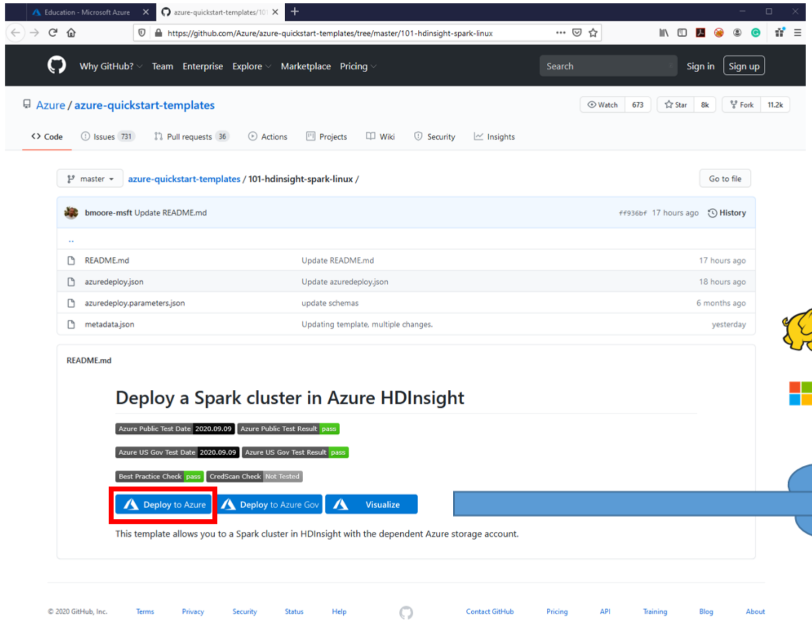
SPSS Modeler – Overview Example – Modeling with Nodes in a Pipeline

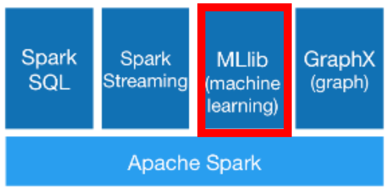


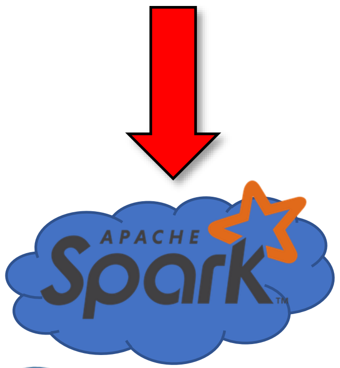
[29] SPSS Modeler

MS Azure Cloud Example – Using FP-GROWTH via Apache Spark MLlib

- Apache Spark (cf. Lecture 2)
  - MLlib implements FP-GROWTH
  - Scalable solution for Big Data



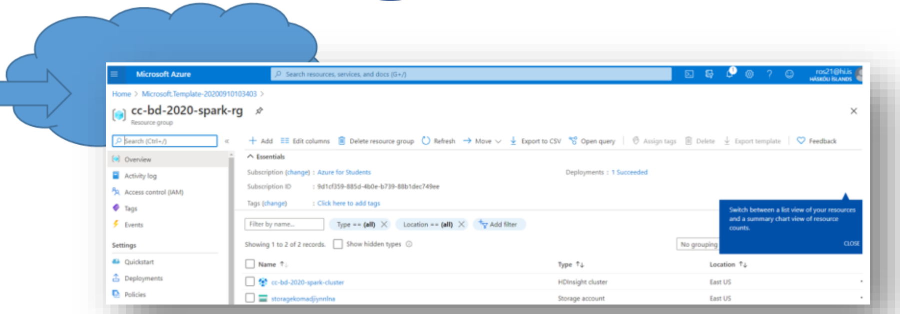




HDInsight

Microsoft Azure

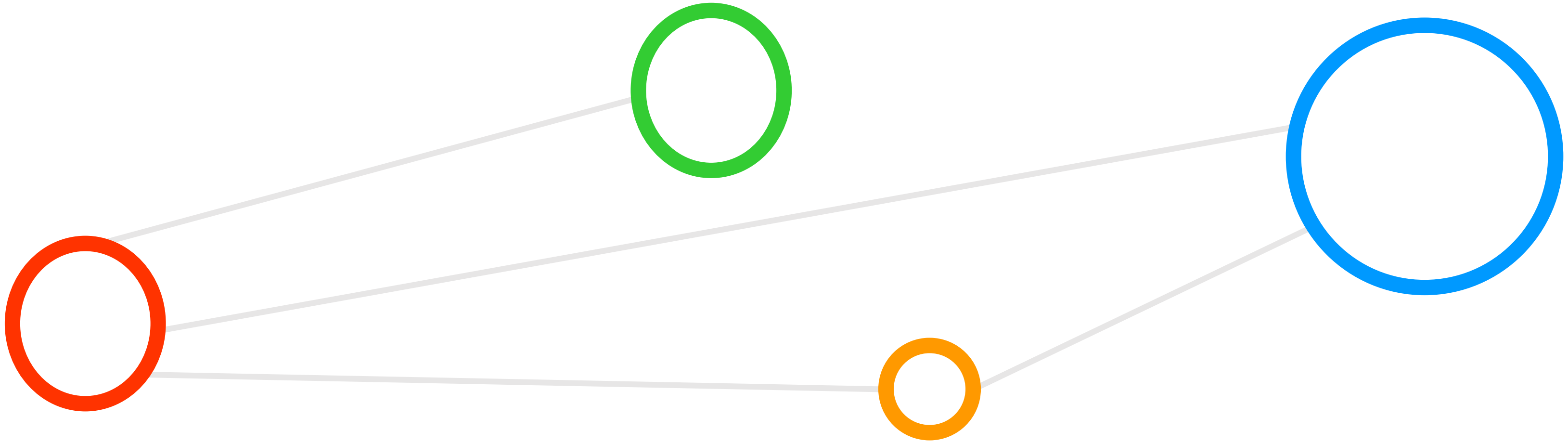
```python
from pyspark.ml.fpm import FPGrowth

df = spark.createDataFrame([
    (0, [1, 2, 5]),
    (1, [1, 2, 3, 5]),
    (2, [1, 2])
], ["id", "items"])

fpGrowth = FPGrowth(itemsCol="items", minSupport=0.5, minConfidence=0.6)
model = fpGrowth.fit(df)

# Display frequent itemsets.
model.freqItemsets.show()

# Display generated association rules.
model.associationRules.show()

# transform examines the input items against all the association rules and summarize the
# consequents as prediction
model.transform(df).show()
```



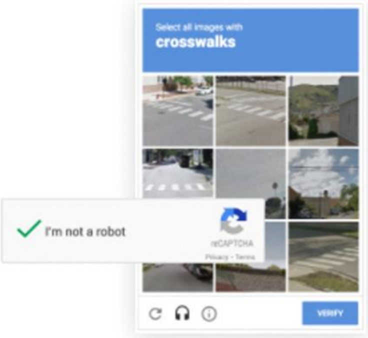
[30] Azure Portal Hub *[31] Apache Spark*

[32] Microsoft Azure HDInsight Service

Deep Learning for Label Generation

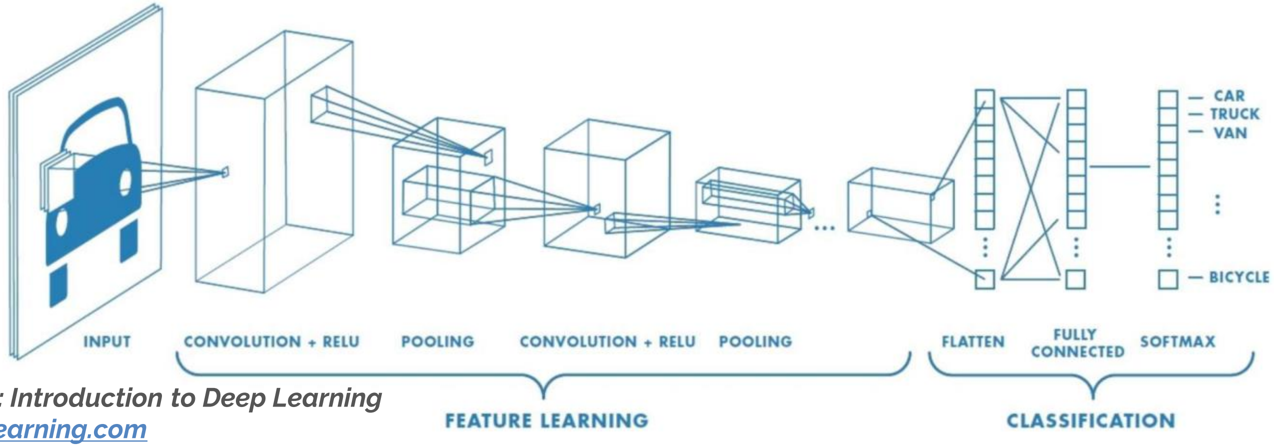


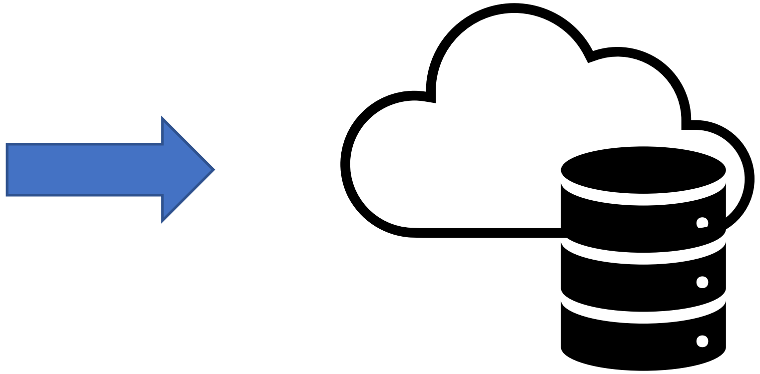
Cloud Data Mining – A Different Perspective



[13] Google reCaptcha

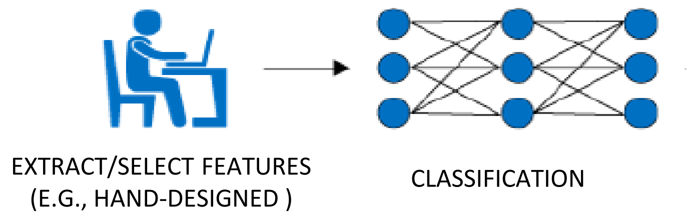
- Deep Learning has applications in the cloud
- Predominantly not user-facing.
- Main aims include:
  - Mining image databases to generate labels/tags for products.
  - Improving Natural Language Processing by learning over open-access literature.
  - Enriching databases to improve search and recommendations.



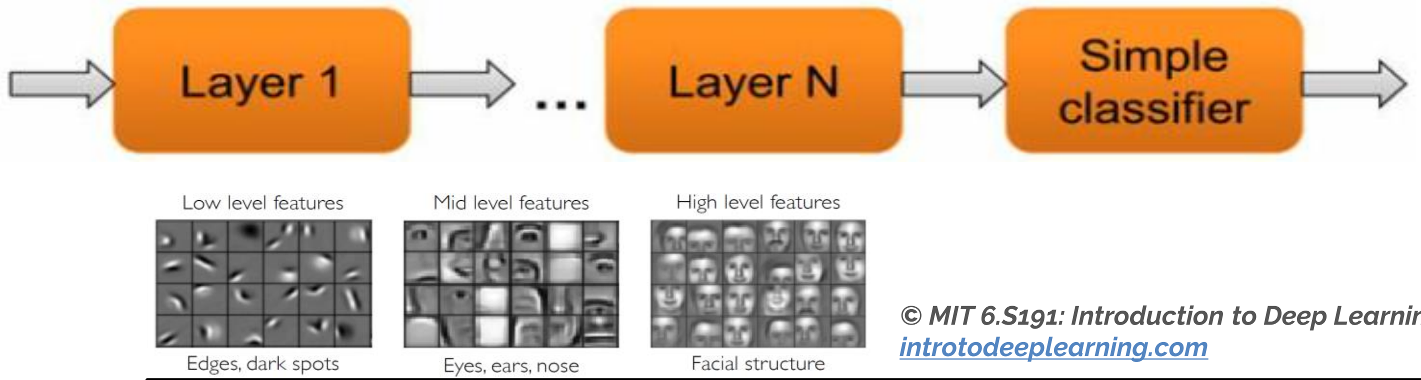


Deep Learning, a Refresher

- **Shallow learning**: learning networks that usually have at most one to two layers
  - They compute linear or nonlinear functions of the data (**often hand-designed features**)



- **Deep learning:** means a deeper network with many layers of non-linear transformations
  - No universally accepted definition of how many layers constitute a "deep" learner

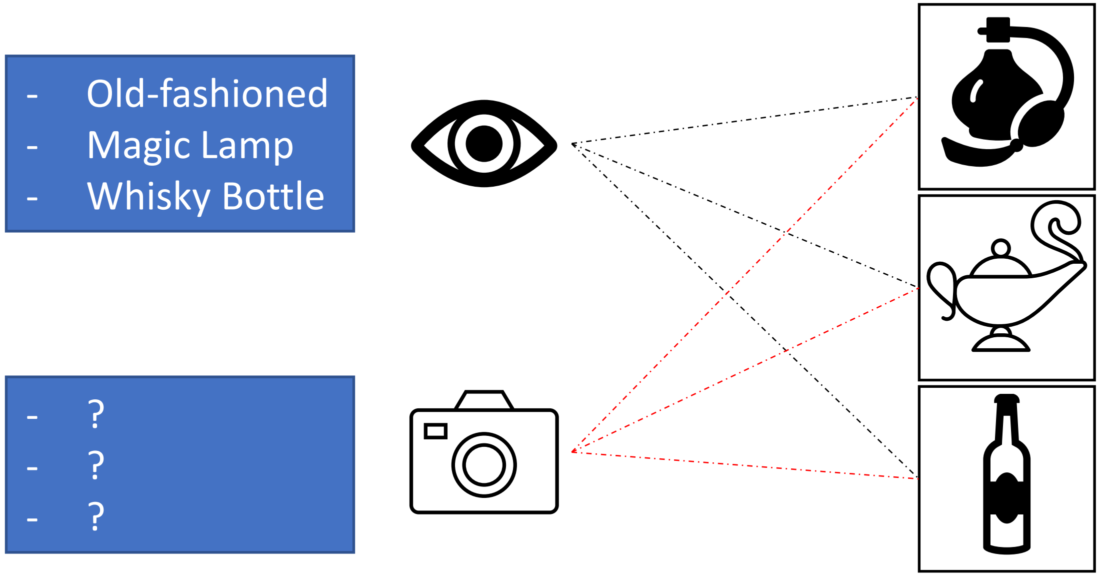


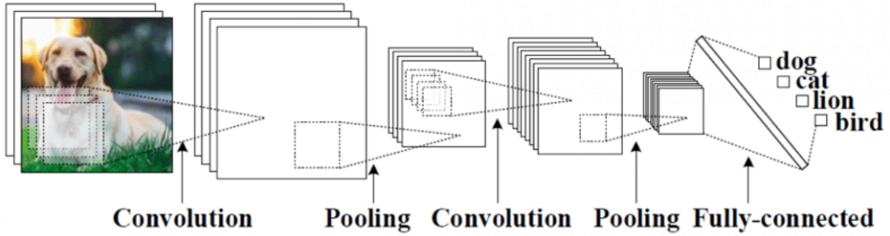
© MIT 6.S191: Introduction to Deep Learning
introtodeeplearning.com

Refer to previous Lecture 6 for a more thorough description of Deep Learning theory and methods.

Defining the Problem

- Can’t remember the name of a perfume:
  - What did it look like?
  - What was its colour?
  - What specific features did it have?
- Could we train a Neural Network to output this type of labels for a database of perfume pictures?
- Do we have enough images to properly train?
- How do we generate the labels?
- Do we have the computing resources to perform this process?





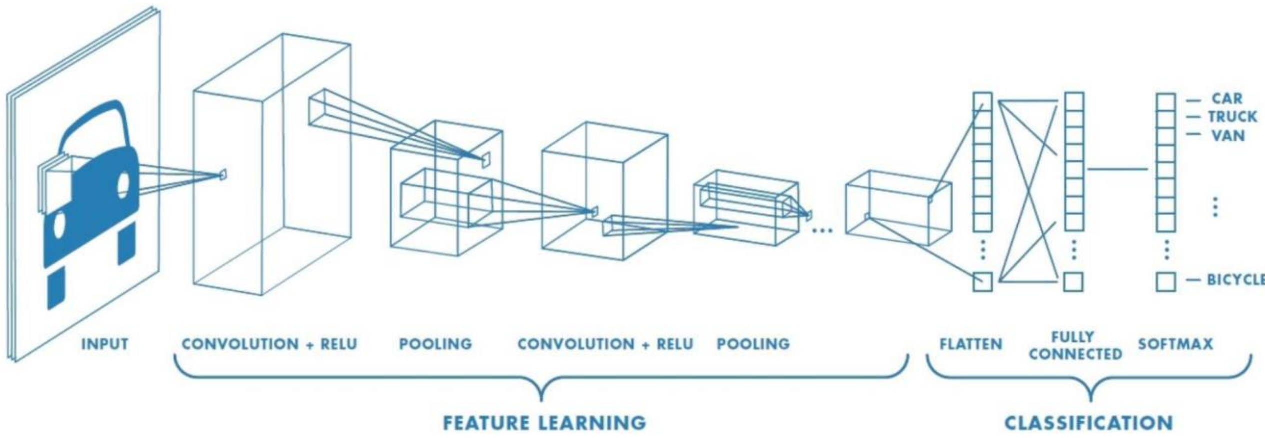
Convolutional Neural Network

[14] Gunnar Carlsson

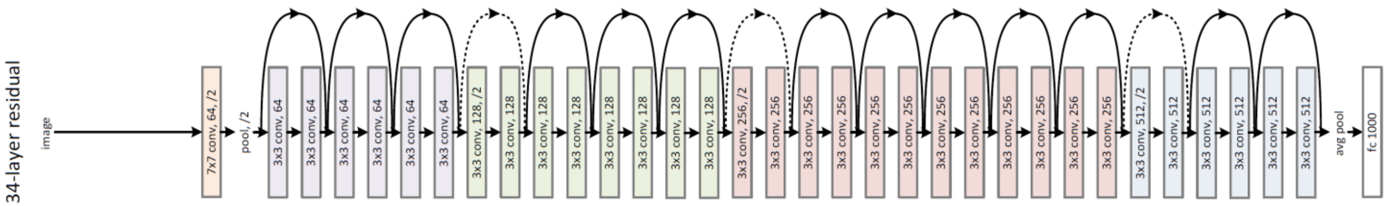
Step 1: Understand the problem.
Step 2: Understand your data.
Refer to the CRISP-DM documentation [15]

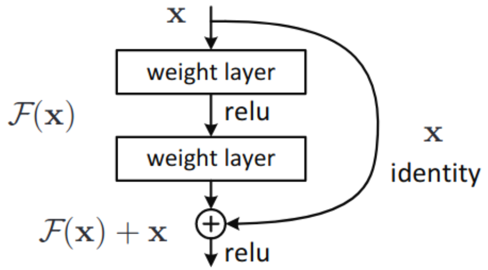
CNNs and Residual Networks

- Convolutional Neural Networks are well adapted to learning from image data.
- Should we build our own network and train it ourselves?
  - Consider the number of images available and the labels you have.
- Pretrained Networks are available which have very high accuracy at object detection.
- ResNet50 winner of the ILSVRC 2015 classification task.
- 1000 classes from the ImageNet [17] Database (animals, objects, shapes).



© MIT 6.S191: Introduction to Deep Learning
introtodeeplearning.com





ResNet [16]

IMAGENET

ImageNet [17]

Refer to previous Lectures 6 and 7 for more information about CNNs and Residual Networks.

Practical Example

- Predicting with ResNets is as easy as loading the correct modules, and then passing your image through the model.

```
model = ResNet50(input_shape=(224,224,3))
```

```
predictions = pd.DataFrame()
imgin = cv2.imread('products/0084.jpg')
imgin = cv2.cvtColor(imgin, cv2.COLOR_BGR2RGB)
image = np.reshape(imgin, newshape=(1,224,224,3))
preds = model.predict(image)
predictions = pd.DataFrame(decode_predictions(preds, top=10))
```

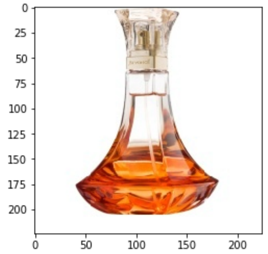
- ResNet50, 101, and 152 are part of the Keras package for Python [19]
- The main predictions are straightforward and are descriptive of exactly what is in the picture, but that's not what we're looking for.
- The secondary predictions show a bit more uncertainty which is (in a way) what we're looking for.

predictions

	0	1	2	3	4	5
0	(n04399382, teddy, 0.668742)	(n03476684, hair_slide, 0.2513311)	(n03188531, diaper, 0.011113512)	(n04162706, seat_belt, 0.010227377)	(n06359193, web_site, 0.009432958)	(n02834397, bib, 0.008623936)

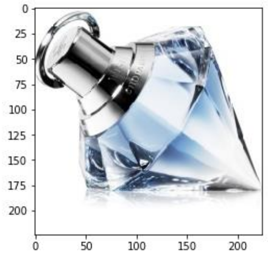
predictions

	0	1	2	3	4	5
0	(n03916031, perfume, 0.9999505)	(n03690938, lotion, 1.5465565e-05)	(n04380533, table_lamp, 1.1884465e-05)	(n02948072, candle, 3.7579064e-06)	(n03314780, face_powder, 3.1802883e-06)	(n03666591, lighter, 1.3714757e-06)



predictions

	0	1	2	3	4	5
0	(n03916031, perfume, 0.5255641)	(n02815834, beaker, 0.22582252)	(n07892512, red_wine, 0.08552182)	(n04522168, vase, 0.04217156)	(n03950228, pitcher, 0.026605058)	(n03443371, goblet, 0.02558193)



predictions

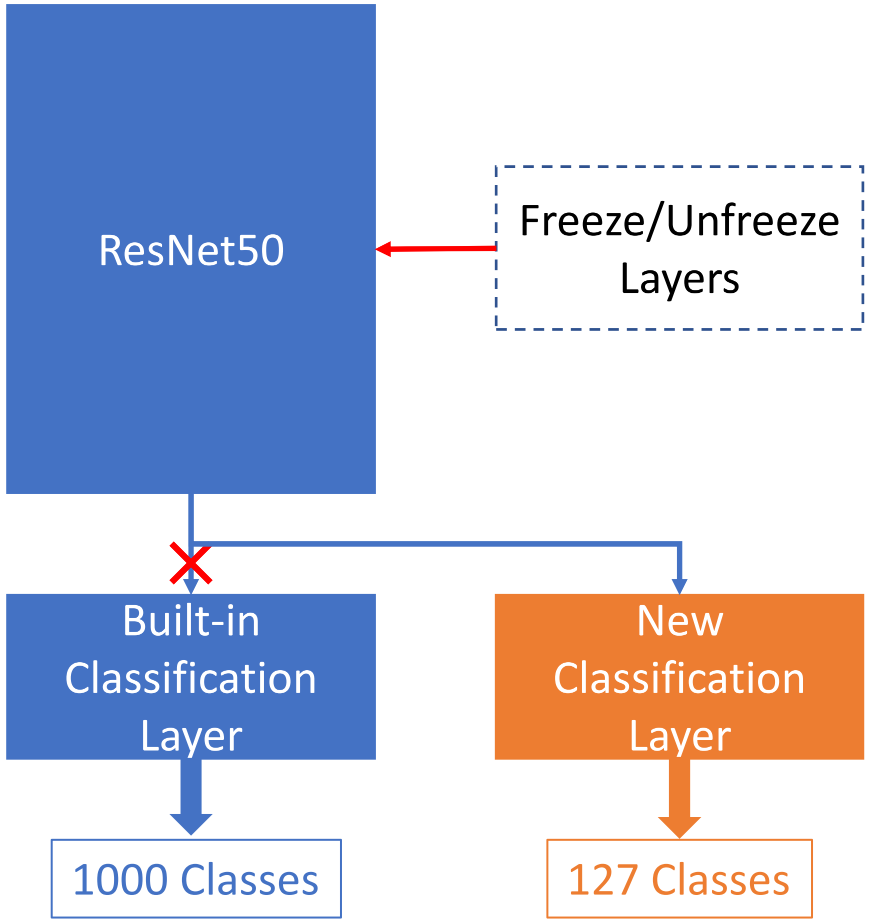
	0	1	2	3	4	5
0	(n03916031, perfume, 0.6627372)	(n04286575, spotlight, 0.14629175)	(n03062245, cocktail_shaker, 0.110806584)	(n03676483, lipstick, 0.012394556)	(n03483316, hand_blower, 0.009056245)	(n04557648, water_bottle, 0.009022212)

python™ *[18] Python*

[19] Keras

Transfer Learning

- Outputs of the Pre-trained Network don't necessarily fit our desired outputs.
- Adapt the Network to our classification problem:
  - Remove the output layer and attach our own classification layer with desired number of classes.
  - Freeze/unfreeze ResNet50 layers depending on how much of the original knowledge we want to keep.
  - Prepare our Training and Testing data with individual labels. (Quality of the labels defines the quality of our training)
- Perform training and adjust the approach depending on your outputs and your testing results.



Transfer Learning applied to ResNet50

Transfer Learning is the process of adapting a pre-trained network to a new task [20]. The new task has a common aspect with the original task but different outputs.

Computational Cost of Training

- Running any pre-trained network or a given set of pictures is not inherently costly since all the calculation have already been done.
- A Deep Neural Network as complex as ResNet50 has around 25 million trainable parameters.
- Training on a normal machine is near impossible.
- HPC is the solution where possible.
- Otherwise, seek Cloud resources.

```
conv5_block3_3_conv (Conv2D)    (None, 7, 7, 2048)   1050624     conv5_block3_2_relu[0][0]
conv5_block3_3_bn (BatchNormali (None, 7, 7, 2048)   8192        conv5_block3_3_conv[0][0]
conv5_block3_add (Add)          (None, 7, 7, 2048)   0           conv5_block2_out[0][0]
                                                                 conv5_block3_3_bn[0][0]
conv5_block3_out (Activation)   (None, 7, 7, 2048)   0           conv5_block3_add[0][0]
avg_pool (GlobalAveragePooling2 (None, 2048)         0           conv5_block3_out[0][0]
predictions (Dense)             (None, 1000)         2049000     avg_pool[0][0]
==============================================================================================
Total params: 25,636,712
Trainable params: 25,583,592
Non-trainable params: 53,120
```

[23] Google Cloud

[1] Microsoft Azure

[5] AWS

Dealing with Bad Pictures and Sparse Data

- The success of your learning problem is defined by the quality of your data (image size, image quality, data size, label accuracy, etc.)
- What to do when you have a limited number of images?
  - Get more images!
    - If we could, we wouldn't be having this problem…
  - Train over the same images.
    - NO! Just no!
  - Generate your own images.
    - Sounds crazy, but it works!
- Data Augmentation:
  - Applied during training.
  - Rotates, flips, translates, and crops the image before feeding it to the network.
  - Grows the size of your data without negatively affecting your training.

Data augmentation is one solution to the problem of sparse data that doesn't negatively affect training results.

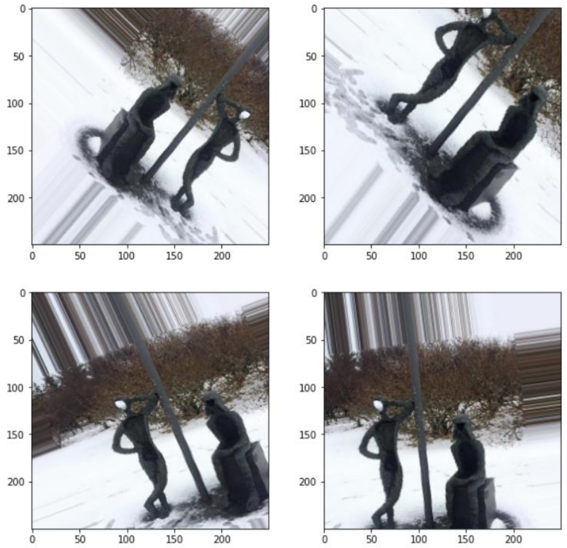
```
[7]: datagen = ImageDataGenerator(rotation_range=45,
                                  width_shift_range=0.2,
                                  height_shift_range=0.2,
                                  shear_range=0.2,
                                  zoom_range=0.2,
                                  horizontal_flip=True,
                                  fill_mode='nearest')
```

```
[5]: img
```

[5]:

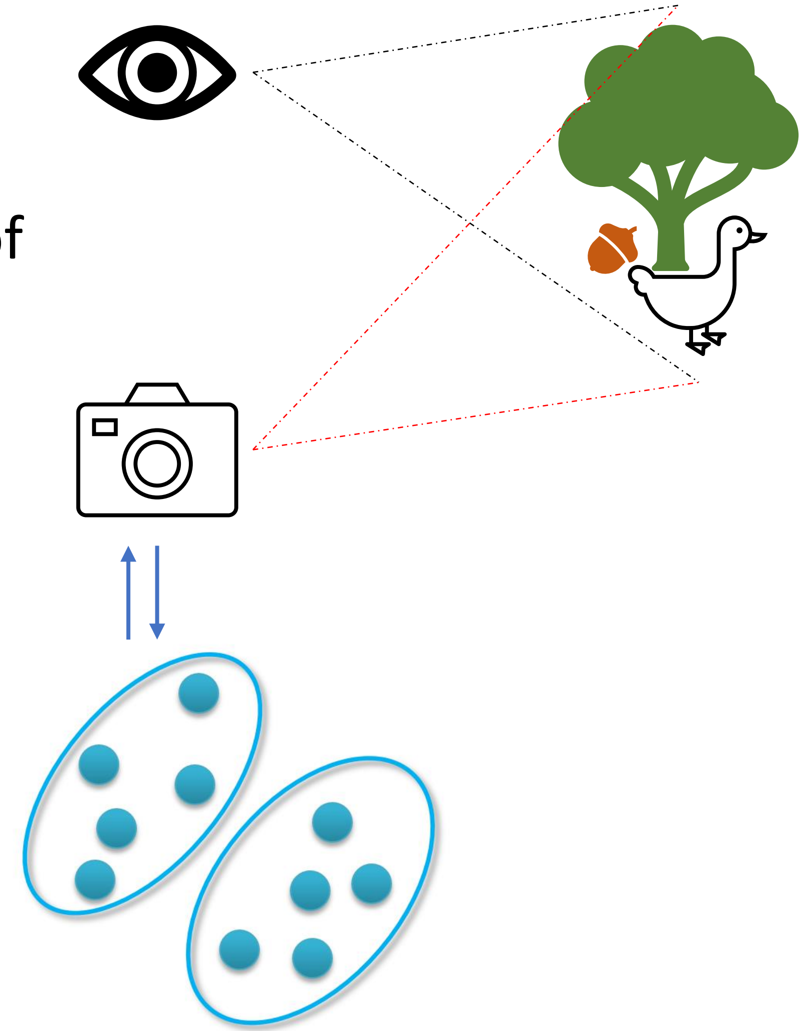
```
[6]: x = image.img_to_array(img)
     x = x.reshape((1,)+x.shape)
```

```
[13]: i=0
      plt.figure(figsize=(10,10))
      for batch in datagen.flow(x,batch_size=1):
          plt.subplot(2,2,i+1)
          plt.imshow(image.array_to_img(batch[0]))
          i+=1
          if i%4==0:break
```



Colour Detection and Classification

- Humans, like other animals are adapted to recognizing colour contrasts to find food in vegetation ➔ millions of years of evolution.
- How do we transfer that knowledge to computers?
- We could train a neural network, but the objective is to simplify.
- Simpler solution is clustering/unsupervised learning.



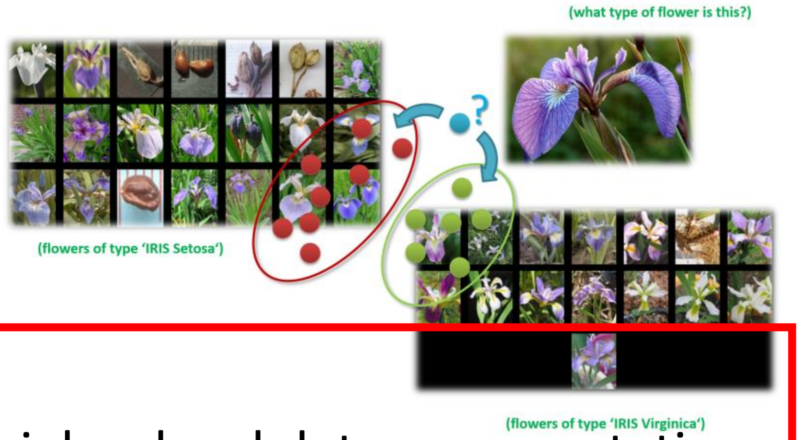
- Occam's Razor: the simplest solutions are often the best.
- Your job is to find out what solution is the simplest/lightest.
- Experience and a proper understanding of the problem are necessary.

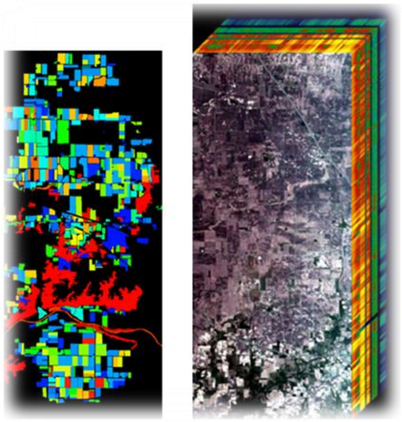
Learning Approaches – What means Learning from data? (reminder from Lecture 1)

- The basic meaning of learning is ‘to use a set of observations to uncover an underlying process‘
- The three different learning approaches are supervised, unsupervised, and reinforcement learning

[21] Image sources: Species Iris Group of North America Database, www.signa.org

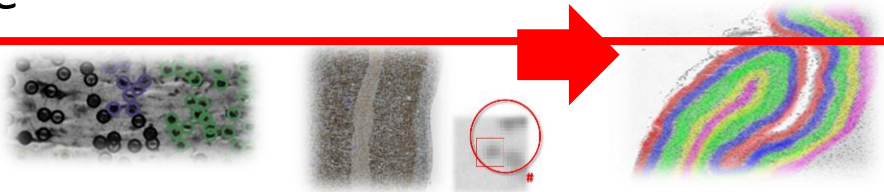
- Supervised Learning
  - Majority of methods follow this approach in this course
  - Example: credit card approval based on previous customer applications



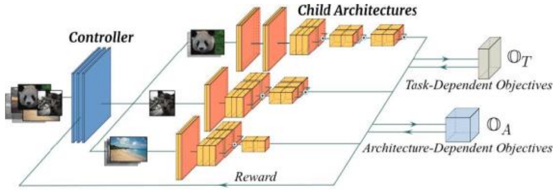


- Unsupervised Learning
  - Often applied before other learning → higher level data representation
  - Example: Coin recognition in vending machine based on weight and size

[22] A.C. Cheng et al., ‘InstaNAS: Instance-aware Neural Architecture Search’, 2018



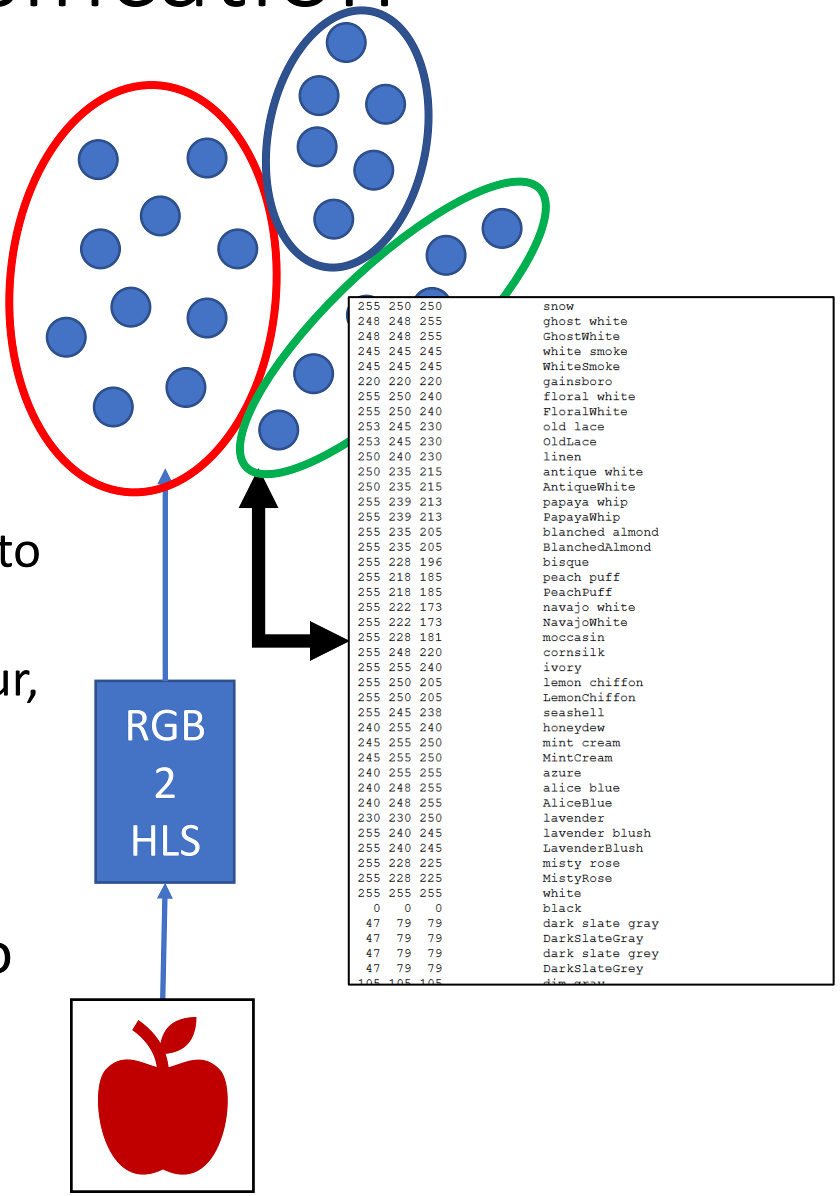
- Reinforcement Learning
  - Typical ‘human way‘ of learning
  - Example: Toddler tries to touch a hot cup of tea (again and again)



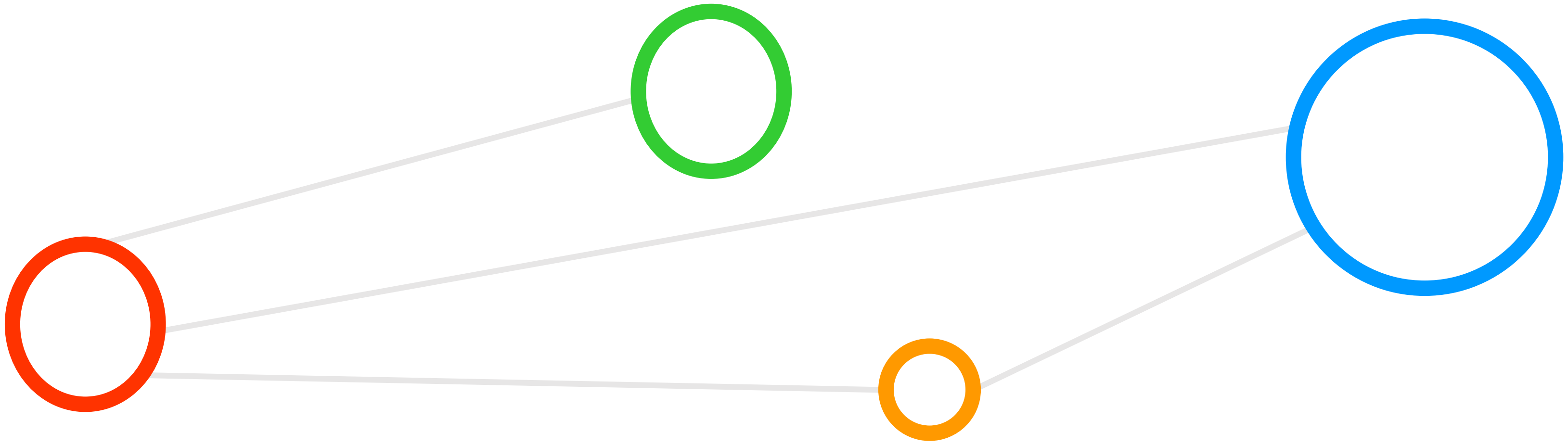
➢ **Refer to Lecture 1 for a refresher about the types of learning.**

Colour Detection and Classification

- Different techniques to clustering results.
- All related to distance between point (cosine, Euclidean, etc.)
- Our clustering will be on the distance between colours:
  - D((R,G,B), (r,g,b)) depends too much on the interaction between the three components from one picture before even comparing to the other picture.
  - Solution is to use HLS where only 1 component defines the colour, and the other two define how bright and how concentrated it is, respectively.
- Every Linux machine has a rgb.txt file with basic colour names and RGB values. We convert these values to HLS to get the final labels.



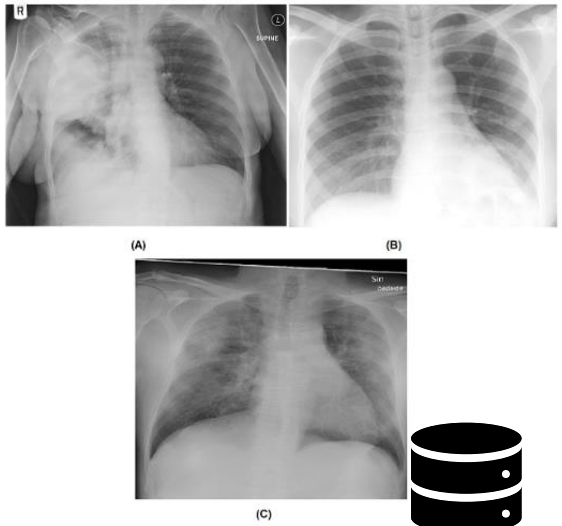
Data Mining in Healthcare



How this Fits into my Ph.D.

- Classification Rule Mining is an effective way of finding relationships between several features at a time, and there's potential in applying it to medical information systems to detect multi-parameter interactions.
- Ex: combinations of several drugs that lead to greater or reduced effect, interactions between drug intake and levels of physiological parameters.
- Deep Neural Networks are currently being adapted to analysing chest X-rays for COVID-19 diagnosis [24] but can also be used to browse through other open-source datasets (Ex: MIMIC-3 database[25]) to uncover new information from ICU patient information.

[24] Chest X-Rays from COVID-NET dataset

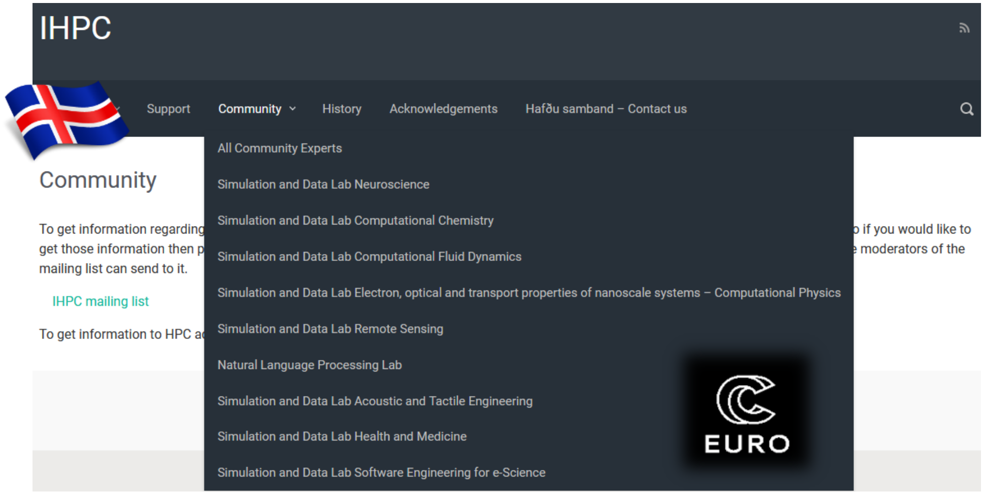


MIMIC

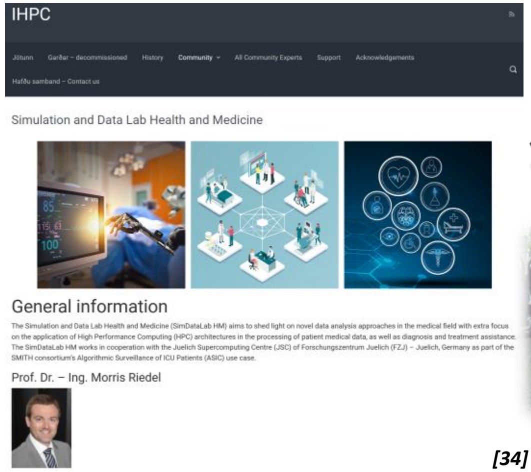
[25] MIMIC Critical Care Database

Establishing International Connections

- Research done in Germany benefits the international community through cooperation.
- Set up a centre for collaboration and innovation using HPC in Iceland.
- One part is Health and Medicine, but also establishes labs for Computational Fluid Dynamics, Remote Sensing, and Neuroscience, among others.



[33] IHPC Web Page



SMITH
Smart Medical Information Technology for Healthcare

Deep Learning Models for ARDS

[34] IHPC SimDataLab Health & Medicine Web Page

Big Data and Cloud Computing in the News

- Recently published articles concerning the application of Deep Neural Networks to analyse big data.
- Data collected and compiled over decades was mined and used.
- The networks were trained on local and remote machines (DeepMind having access to hardware on Google servers).
- Applied to new problems they were almost as accurate as the current experimental standards (X-ray Crystallography).

Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4] ✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4] ✉

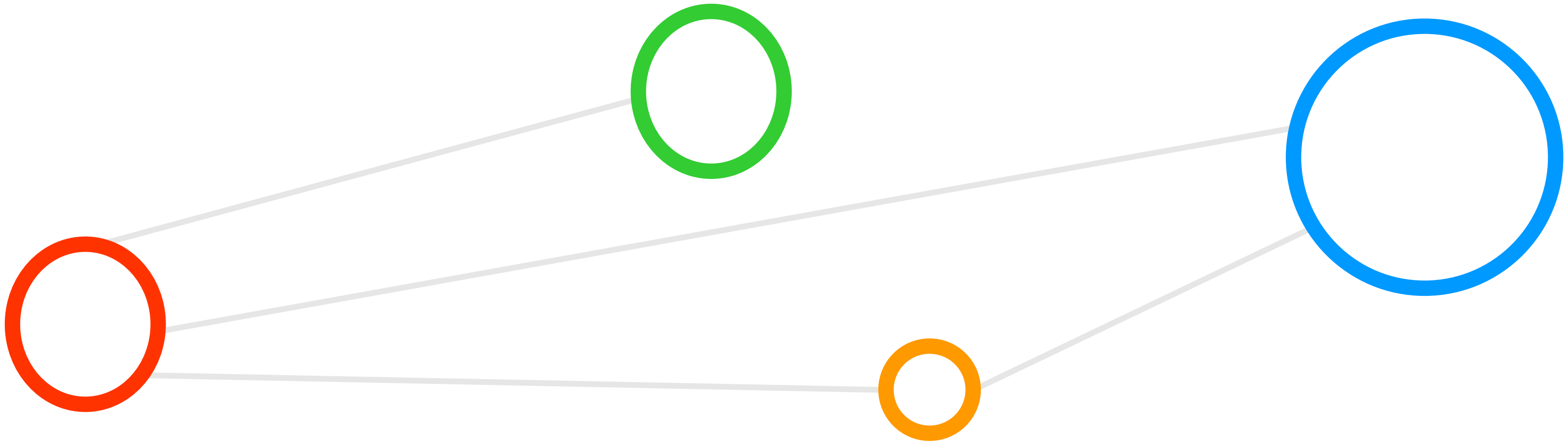
Science *RESEARCH ARTICLES*

Cite as: M. Baek *et al.*, *Science* 10.1126/science.abj8754 (2021).

Accurate prediction of protein structures and interactions using a three-track neural network

Minkyung Baek[1,2], Frank DiMaio[1,2], Ivan Anishchenko[1,2], Justas Dauparas[1,2], Sergey Ovchinnikov[3,4], Gyu Rie Lee[1,2], Jue Wang[1,2], Qian Cong[5,6], Lisa N. Kinch[7], R. Dustin Schaeffer[6], Claudia Millán[8], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[9,10], Andy DeGiovanni[12], Jose H. Pereira[12], Andria V. Rodrigues[12], Alberdina A. van Dijk[13], Ana C. Ebrecht[13], Diederik J. Opperman[14], Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K. Rathinaswamy[18], Udit Dalwadi[19], Calvin K. Yip[19], John E. Burke[18], K. Christopher Garcia[9,10,11,20], Nick V. Grishin[6,21,7], Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23*]

Lecture Bibliography



Lecture Bibliography (1)

- [1] Microsoft Azure Cloud, 'What is SAAS'
  Online: https://azure.microsoft.com/en-us/overview/what-is-saas/
- [2] K. Hwang, G. C. Fox, J. J. Dongarra, 'Distributed and Cloud Computing', Book
  Online: http://store.elsevier.com/product.jsp?locale=en_EU&isbn=9780128002049
- [3] ZOHO Customer Relationship Management Web page
  Online: https://www.zoho.com/crm/
- [4] Freshworks Customer Relationship Management Web page
  Online: https://www.freshworks.com/crm/
- [5] AWS Amazon Sagemaker Service
  Online: https://aws.amazon.com/sagemaker
- [6] Big Data Tips, Association Rules
  Online: http://www.big-data.tips/association-rules
- [7] German ON4OFF Retail AI Project
  Online: https://www.on-4-off.de/
- [8] P.-N. Tan, M. Steinbach, V. Kumar, 'Introduction to Data Mining', Book
- [9] MLxtend Library, Apriori
  Online: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/
- [10] MLxtend Library, Association Rules
  Online: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/
- [11] MLxtend Library, FP-Growth
  Online: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth/
- [12] M.S. Mythili, A.R. Mohamed Shanavas, 'Performance Evaluation of Apriori and FP-Growth Algorithms', *International Journal of Computer Application*, vol. 79, no. 10, October 2013
  Online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.1361&rep=rep1&type=pdf
- [13] Google reCaptcha
  Online: https://www.google.com/recaptcha/about/

Lecture Bibliography (2)

- [14] G. Carlsson, Topological Data Analysis
  Online: https://www.ayasdi.com/using-topological-data-analysis-understand-behavior-convolutional-neural-networks/
- [15] P. Chapman, "The CRISP-DM User Guide," 1999.
  Online: https://s2.smu.edu//~mhd/8331f03/crisp.pdf
- [16] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", *29th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, June 2016, doi:10.1109/CVPR.2016.90
- [17] ImageNet
  Online: http://image-net.org/index
- [18] Python
  Online: https://www.python.org/
- [19] Keras
  Online: https://keras.io/api/applications/
- [20] F. Chollet, “Transfer Learning and Fine-Tuning”
  Online: https://keras.io/guides/transfer_learning/
- [21] Species Iris Group of North America Database, Online:
  http://www.signa.org
- [22] Cheng, A.C, Lin, C.H., Juan, D.C., InstaNAS: Instance-aware Neural Architecture Search, Online:
  https://arxiv.org/abs/1811.10201
- [23] Google Cloud
  Online: https://cloud.google.com/
- [24] L. Wang, A. Wong, 'COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images’
  Online: https://arxiv.org/abs/2003.09871
- [25] MIMIC Critical Care Database
  Online: https://mimic.physionet.org/

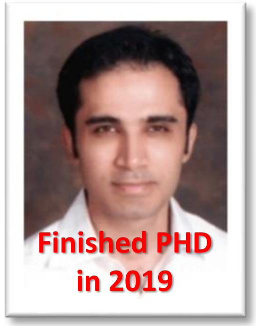
Lecture Bibliography (3)

- [26] IBM Cloud Pak for Data
  Online: https://www.ibm.com/products/cloud-pak-for-data
- [27] Redhat OpenShift
  Online: https://www.openshift.com/
- [28] Kubernetes Container Orchestration Tool
  Online: https://kubernetes.io/
- [29] IBM SPSS Modeler
  Online: https://www.ibm.com/products/spss-modeler
- [30] Microsoft Azure Portal Hub
  Online: https://portal.azure.com/#home
- [31] Apache Spark Web page
  Online: http://spark.apache.org/
- [32] Microsoft Azure HDInsight Service
  Online: https://azure.microsoft.com/en-us/services/hdinsight/
- [33] IHPC Web Page
  https://ihpc.is/
- [34] IHPC SimDatalab Health and Medicine Webpage
  https://ihpc.is/simulation-and-data-lab-health-and-medicine/

Acknowledgements – High Productivity Data Processing Research Group

PD Dr.
G. Cavallaro

PD Dr.
A.S. Memon



PD Dr.
M.S. Memon

PhD Student
E. Erlingsson

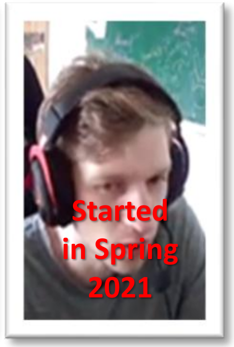
PhD Student
S. Bakarat

PhD Student
R. Sedona

PhD Student
P. H. Einarsson

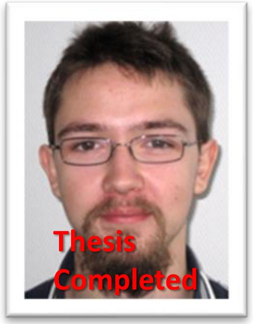
PhD Student
S. Sharma

PhD Student
M. Aach

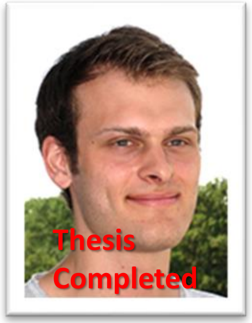


PhD Student
D. Helmrich

Dr. M. Goetz
(now KIT)

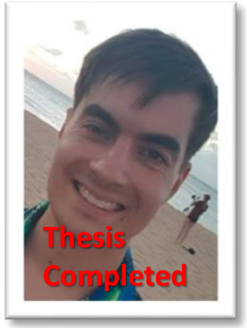


MSc M.
Richerzhagen
(now other division)



MSc
P. Glock
(now INM-1)

MSc
C. Bodenstein
(now
Soccerwatch.tv)



MSc G.S.
Guðmundsson
(Landsverkjun)

PhD Student
Reza

PhD Student
E. Sumner

This research group has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 763558 (DEEP-EST EU Project) and grant agreement No 951740 (EuroCC EU Project) & 951733 (RAISE EU Project)


services
measurement
within
fusion
new
funding
Policy-based
concepts
device
Analysis
forms
cross-disciplinary
resources
Centre
climate
Computer
expertise
Approaches
construction
Fusion
year
like
Computational
Test
E.g
Computing
PB
community
disciplines
results
Enable
Modelling
different
Cross-Disciplinary
methods
big
nodes
Environment
EU
High
centers
Key
scientific
important
Science
etc
DKRZ
Use
storage
basis
Infrastructure
files
model
international
simulations
ENES
technologies
formats
physical
SMAQ
engineering
References
brain
increasing
via
EUDAT
IPCC
Big
Project
Data
MapReduce
create
processing
Common
TB
Methods
computing
Using
European
databases
GB
Supercomputing
Work
images
codes
analysis
stored
Resources
analyze
Large
Summary
data-intensive
Service
Scientific
levels
often
DLCL
Level
Research
external
use
Matters
Concrete
performance
JSC
computational
research
e.g
science
HPC
large-scale
Juelich
datasets
system
management
acquisition
compute
Landscape
System
Climate
modelling
AR5
Hardware
distributed
cores
access
hundreds
Nutshell
Services
Technologies
Structure
current
network
DLCLs
Understanding
structures
Earth
directory
General
Germany
Simulation
project
Provide
NASA
Energy
systems
day
Processing
Health