EDEM

Master Data analytics | 3a Edición

Artificial Intelligence Intro I

Jose Peris Adsuara

Jose Peris Adsuara

Head of Ai @Bit2Me
Data analytics @Edem

EDEM
Escuela de Empresarios

- **CTO // Digital Product Manager @aiwannapay.com**
  aiwanna
  Sep 2019 – Present · 2 yrs 4 mos

  The most complete SaaS solution for HORECA channel. Scan, order and pay from your smartphone.
  Artificial Intelligence for Small and Medium enterprises. Developing AI algorithms and display augmented analytics. Reduce your costs, increase sales and automate your repetitive tasks

- **Senior Consultant**
  straiqr.ai GmbH · Part-time
  Jun 2021 – Present · 7 mos

  Straiqr is a German company that is changing fashion rules through Metaverse, nft´s, digital fashion & AI
  We create digital fashion for the digital generation.We believe in a world where digital garment and clothing design is authentic and true to you. STRAIQR Intelligent Shopping is ready - now - in an easy to navigate store, where accurate delivery and sustainability is a consideration at every step of the way.

- **Teacher // Machine Learning and AI consulting and for MÁSTER EN DATA ANALYTICS & MASTER FINANZAS**
  EDEM Escuela de Empresarios
  Apr 2019 – Present · 2 yrs 9 mos
  Valencia y alrededores, España

  -Coordinating AI and ML contents , team building
  -Machine Learning and Deep Learning teacher:
  -Master Data Analytics
  -Master Finanzas
  -EMBA Executive
  -Webinar " IA aplicada a pymes"
  -Online

- **Data scientist // CEO & co-founder**
  tailor
  Jul 2017 – Sep 2019 · 2 yrs 3 mos
  Valencia, Spain

  Our mision is to improve the online shopping experience in fashion industry.
  Tailored by Big Data is an spanish Startup born in Valencia wich is reasearching in AI.
  Developing algorithms for fashion and trend industry .Our project has been developed under Garaje de Lanzadera incubator program & Órbita accelerator program
  https://lanzadera.es/
  http://www.programaorbita.com
  http://www.tailoredbybigdata.com
  Deep Learning - Machine Learning - Computer vision

2021:
-Winner Star Starups contest “Star days” hackathon

2019:
-Included in PlayBook platform, powered by Plug and Play Ventures

2018:
-Selected for Lanzadera start-up accelerator program
-Selected for Programa Órbita start-up accelerator program
-Selected as speaker in VI Mobile e-commece Madrid

// Index

01// Dataset……04

02 // Algorithm development process……..11

02.1 Variables

02.2 Artificial Vision……..36

-Use cases:

+tailor

02.3 Learn by doing……..…..100

02.4 Tabular data……130

-Use cases:

+Bitcoin

Dataset

DATASETS

Big data

DATA TYPES

ETL: extract, transform & load

Output table - 2:1 - XLS Reader

File

Table "demo_sales.xls [Sales Data]" – Rows: 7085 | Spec – Columns: 11 | Properties | Flow Variables

Row ID	I ID	D Sales	D Cost	S Categ...	S Product	S Sale D...	S Quarter	I Year	S
Row0	2004	2,691.88	1,442	Spices	Saffron	12/28/08	Q4	2008	Jes
Row1	2005	910.51	639.12	Spices	Saffron	12/28/08	Q4	2008	Jes
Row2	2006	69.15	71.47	Fruit	Plums	12/28/08	Q4	2008	Jes
Row3	2007	22.62	22.21	Fruit	Plums	12/28/08	Q4	2008	Jes
Row4	2008	4,141.73	1,184.612	Fruit	Apples	12/28/08	Q4	2008	Jes
Row5	2009	509.53	508.293	Spices	Pepper	12/28/08	Q4	2008	Jes
Row6	2010	804.86	434.44	Spices	Saffron	12/28/08	Q4	2008	Kar
Row7	2011	49.37	24.275	Spices	Saffron	12/26/08	Q4	2008	Jim
Row8	2012	1,019.26	528.28	Spices	Saffron	12/26/08	Q4	2008	Jim
Row9	2013	12.6	8.033	Spices	Pepper	12/26/08	Q4	2008	Jim
Row10	2014	116.68	41.163	Fruit	Bananas	12/27/08	Q4	2008	Cat
Row11	2015	128.82	82.48	Fruit	Oranges	12/27/08	Q4	2008	Cat
Row12	2016	6,838.33	2,993.82	Spices	Ginger	12/27/08	Q4	2008	Cat
Row13	2017	1,161.78	385.452	Fruit	Lemons	12/27/08	Q4	2008	Cat
Row14	2018	1,011.56	545.28	Fruit	Bananas	12/25/08	Q4	2008	Eric
Row15	2019	3,435.28	1,849.92	Fruit	Pears	12/25/08	Q4	2008	Eric

DATA TYPES

Types of data:

Double: D, fractional parts “1.7854” 123

Integer: I , non fractional parts “7” 123

String: S , category “diabetes” category

Date or Time: Clock, “ 24-07-1997”

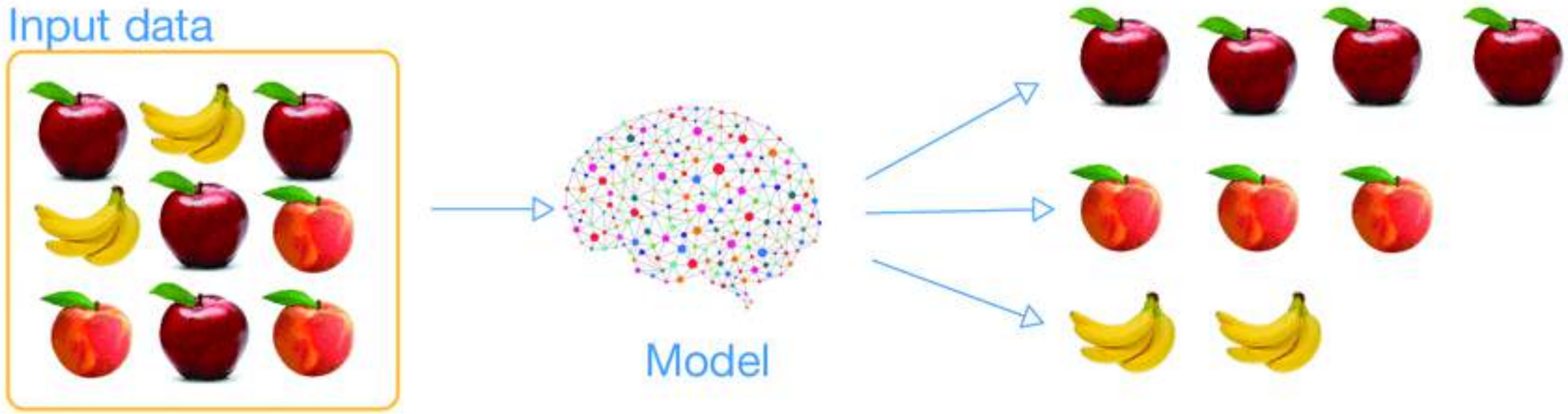
Dictionary: **[…],** “[24,3,52,48,76]”

Images: Img, “cat.png”

Document: “great food, excellent service”

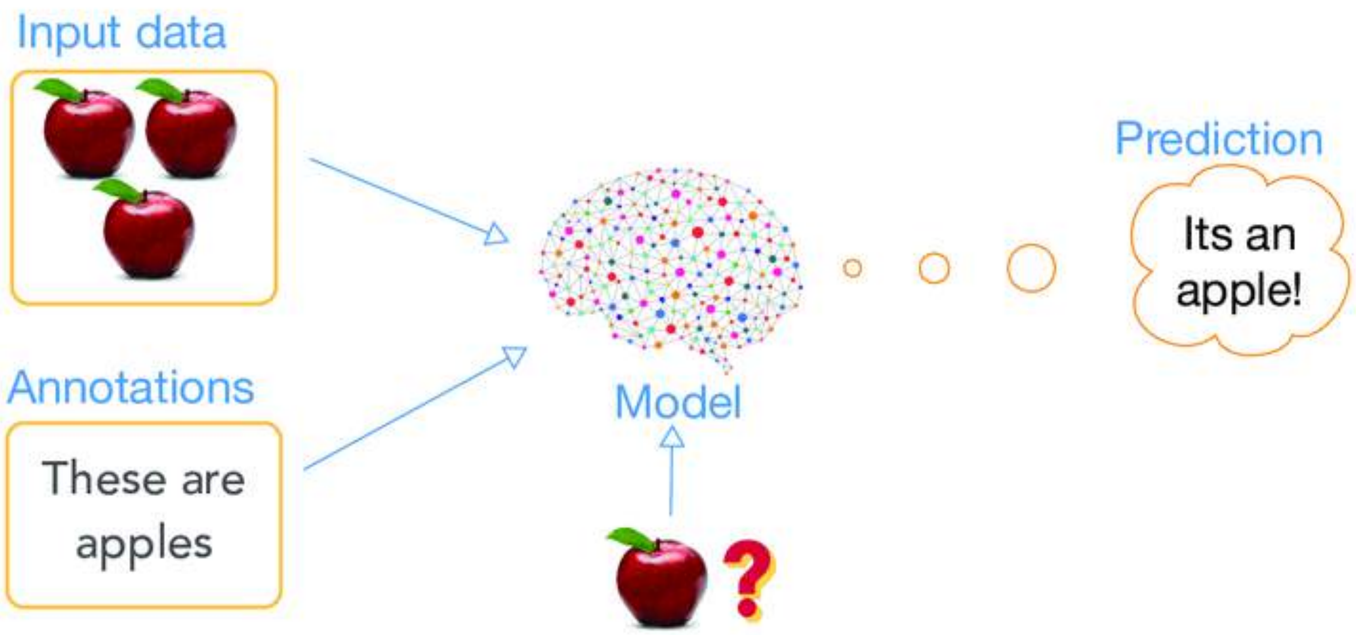
ML // EASY EXPLAINED

What is the input of an algorithm? **Unlabeled data** unsupervised learning



ML // EASY EXPLAINED

What is the input of an algorithm? **Labelled data** supervised learning



kaggle

Search

Sign In

Register

- Home
- Compete
- Data
- Code
- Communities
- Courses
- More

Datasets

Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating.

+ New Dataset

Search datasets

Filters

Datasets | Tasks | Computer Science | Education | Classification | Computer Vision | NLP | Data Visualization

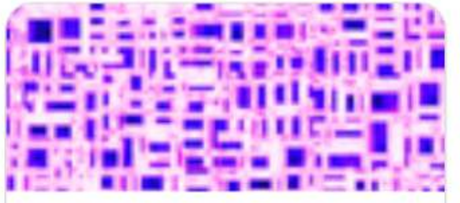
Trending Datasets

See All

Which oil is the best?

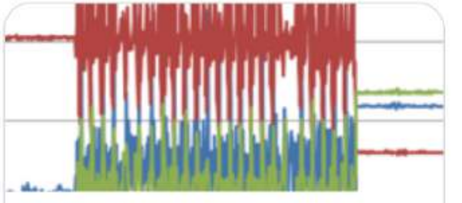
PandaBrenda · Updated 38 minutes ...

Usability 7.9 · 7 KB



Animals Detection Images Dataset

antoreepjana · Updated an hour ago



UNIUD DA-AD

MITEL-UNIUD · Updated 2 hours ago

Usability 7.6 · 2 MB

Petitions from change.org

Mohd Abdul Azeem · Updated 4 hou...

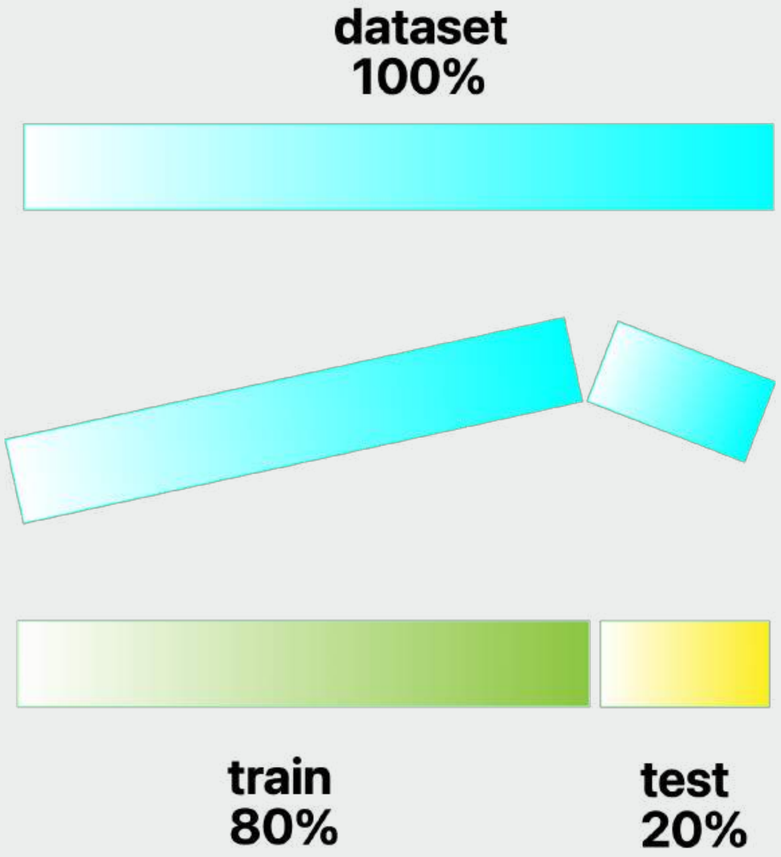
Kaggle

View Active Events

// Development process

Algorithm

Training & test: We divide our data set into training to learn and test to measure the accuracy of predictions



1 //

Predicts

Demand
Stock market
number of clients
Stock out
Energy consumption

2 //

CLassifies

Images
Fake news
Fraud detection
Diseases
Type of reservation
Sentiment analysis

3 //

Groups

Customer segmentation
Smart geolocation
Recommender systems
Online shopping intention

WORKFLOW

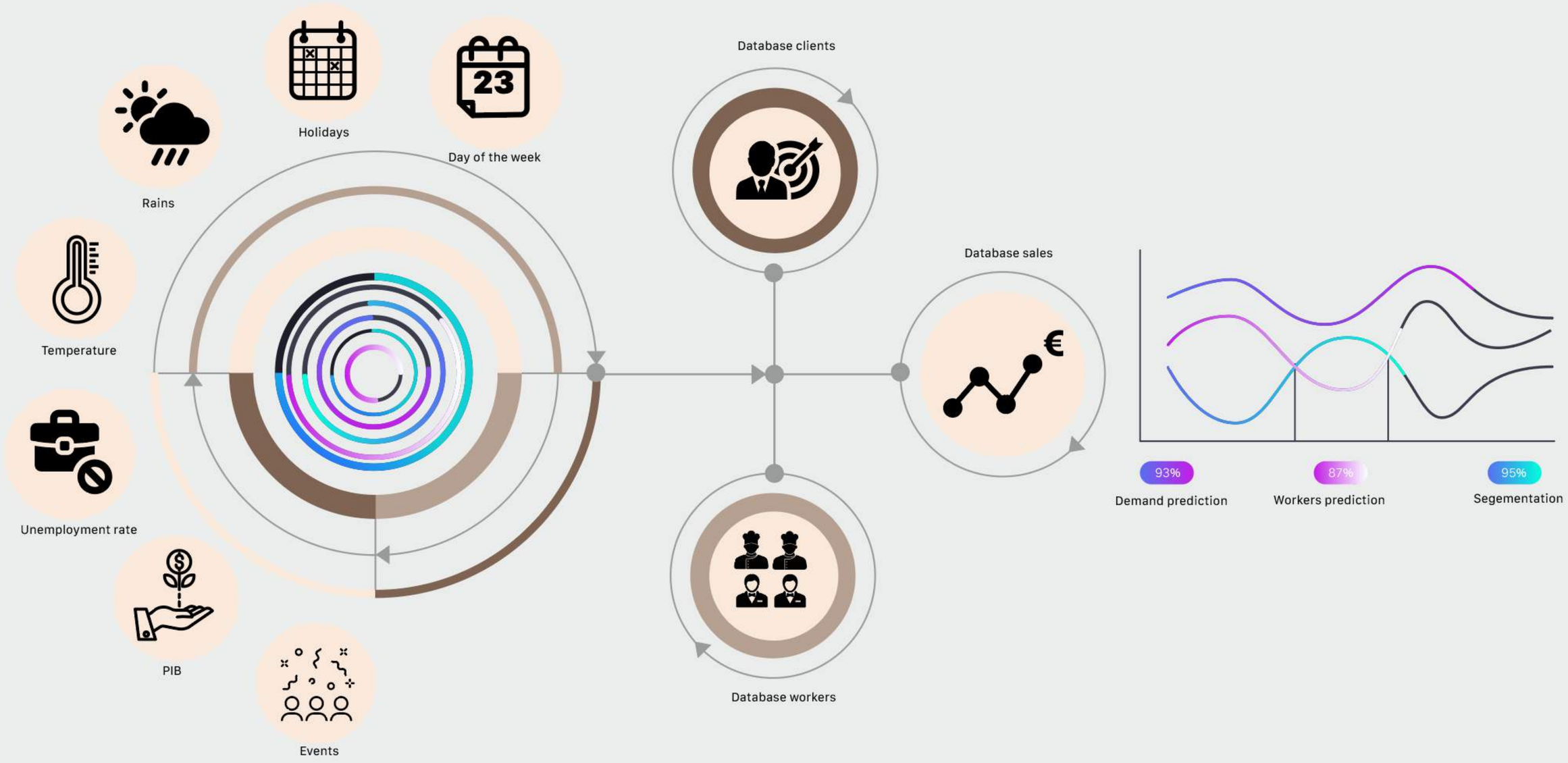
BUSINESS ANALYSIS
& DATA CAPTURE

CUSTOM AI / ML MODELS
DEVELOPMENT

MODEL DEPLOYMENT
& DASHBOARD VISUALIZATION

- Demand prediction
- Customer segmentation – Buyer persona
- Churn rate
- Market Basket Analysis
- E- Commerce
- Augmented analytics
- Use cases

DEMAND PREDICTION



DEMAND PREDICTION

60%

90%

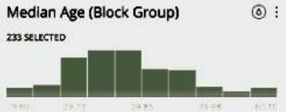
95%

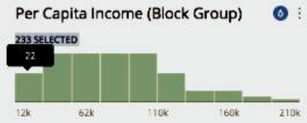
view as

Menu Id	Date	Sales
G1	2017-01-01T00:00:00	104
G1	2017-02-01T00:00:00	112
G1	2017-03-01T00:00:00	115
G1	2017-04-01T00:00:00	252
G1	2017-05-01T00:00:00	132
G1	2017-06-01T00:00:00	110
G1	2017-07-01T00:00:00	173
G1	2017-08-01T00:00:00	232
G1	2017-09-01T00:00:00	172
G1	2017-10-01T00:00:00	215
G1	2017-11-01T00:00:00	304
G1	2017-12-01T00:00:00	68
G2	2017-01-01T00:00:00	191

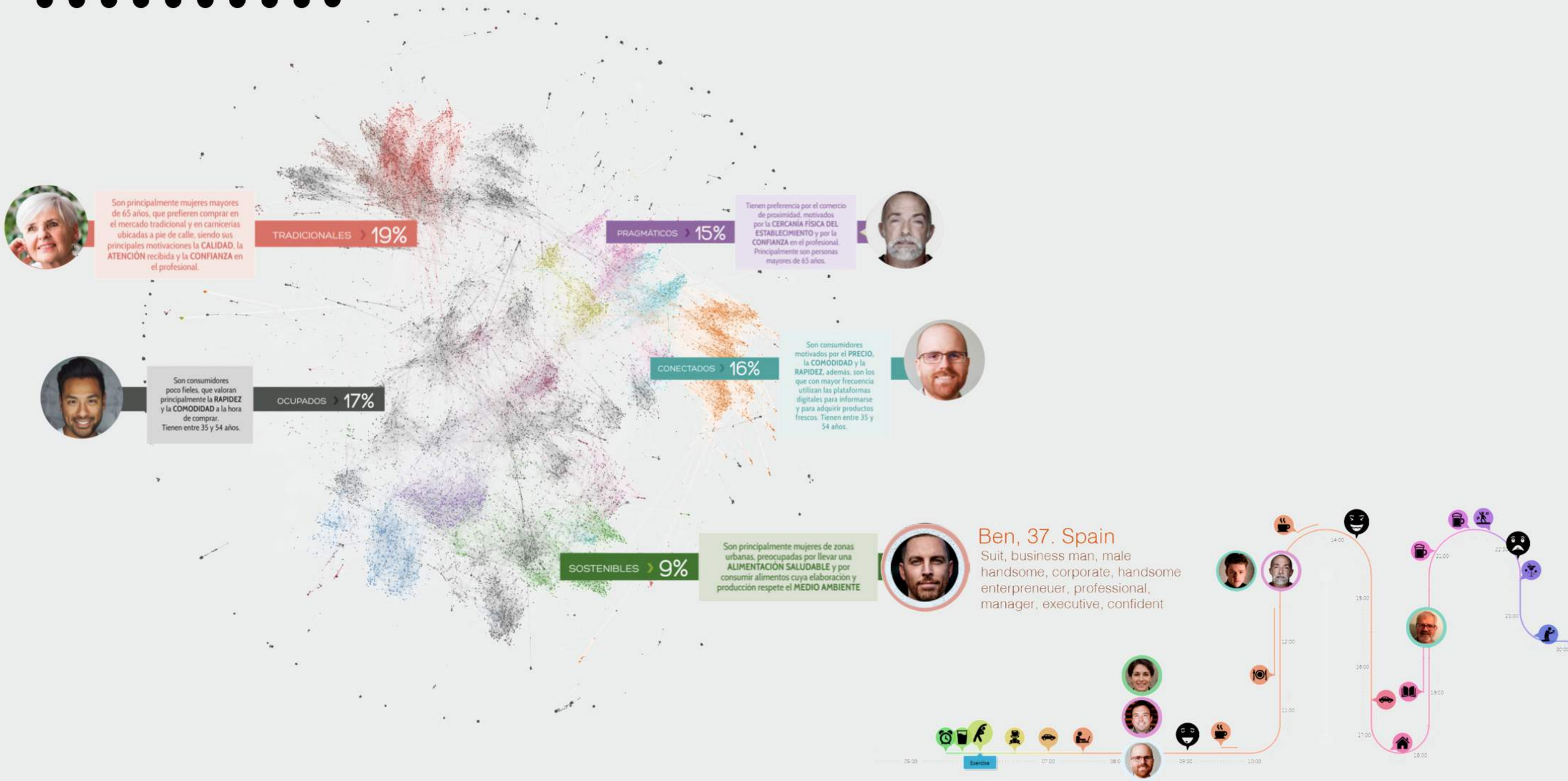
view as

Menu Id	Date	Sales	Weather	Holiday	Festival	Promotions
G1	2017-01-01T00:00:00	104	0	0	1	0
G1	2017-02-01T00:00:00	112	0	0	0	0
G1	2017-03-01T00:00:00	115	0	0	0	0
G1	2017-04-01T00:00:00	252	1	0	0	1
G1	2017-05-01T00:00:00	132	1	0	0	1
G1	2017-06-01T00:00:00	110	1	0	0	1
G1	2017-07-01T00:00:00	173	1	0	0	1
G1	2017-08-01T00:00:00	232	1	0	0	0
G1	2017-09-01T00:00:00	172	1	0	0	0
G1	2017-10-01T00:00:00	215	0	0	1	0
G1	2017-11-01T00:00:00	304	0	0	1	0
G1	2017-12-01T00:00:00	68	0	1	1	0





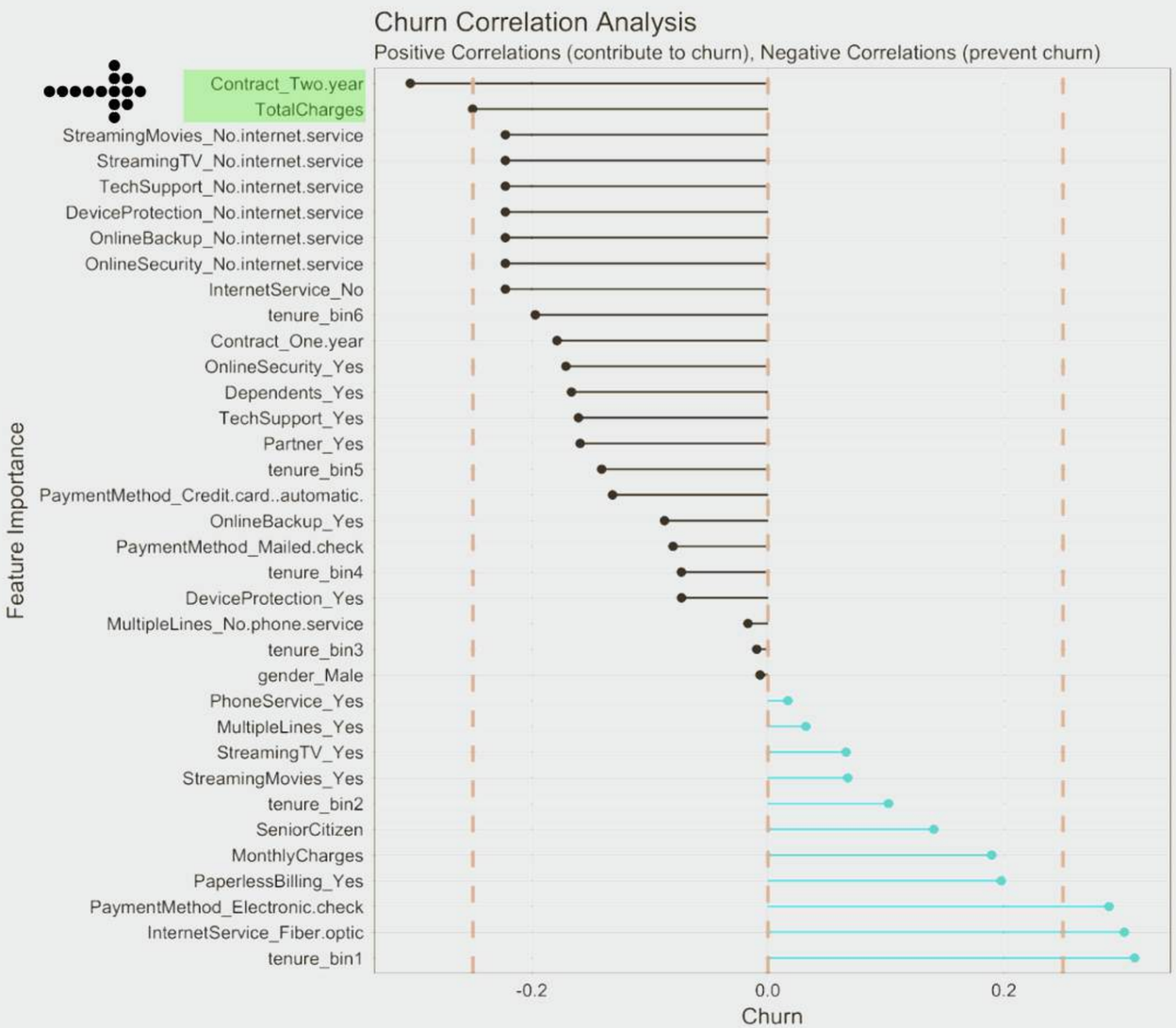
CUSTOMER SEGMENTATION

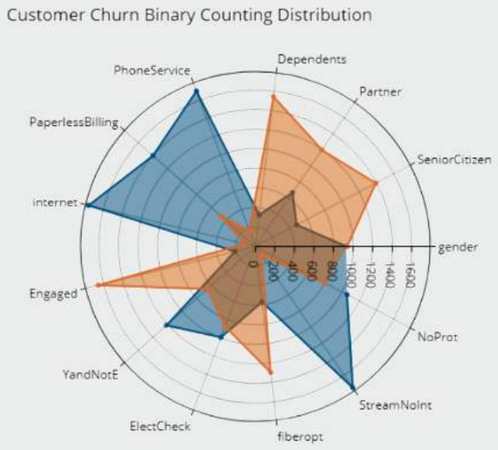


CHURN RATE

Id	Name	Gender	Churn risk	Spendings
230431	Adriana	Female	3%	$291
227555	Brenda	Female	5%	$92
94522	Brett	Male	5%	$103
230814	Roxanna	Female	5%	$136
97026	Donnie	Male	6%	$25
223626	Donnie	Male	6%	$108
84119	Mike	Male	6%	$390
222681	Adriana	Female	7%	$505
220861	Christopher	Male	7%	$362
220774	David	Male	7%	$7.657

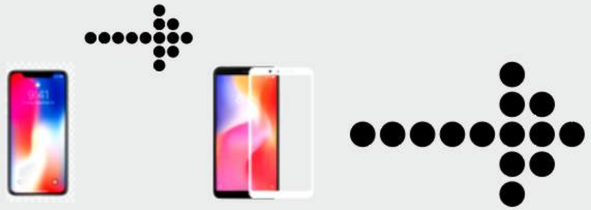
Id	Name	Gender	Churn risk	Spendings
	Greg	Male	95%	$167
40764	Roxanna	Female	88%	$69
210998	Jad	Male	82%	$387
74758	Shandra	Female	79%	$899
104108	Shandra	Female	79%	$155
92663	Susan	Female	78%	$79
226346	Donna	Female	76%	$120
216920	Jack	Male	76%	$421
234001	Janet	Female	76%	$145
72173	Wendy	Female	76%	$745
224742	Gigi	Female	75%	$232

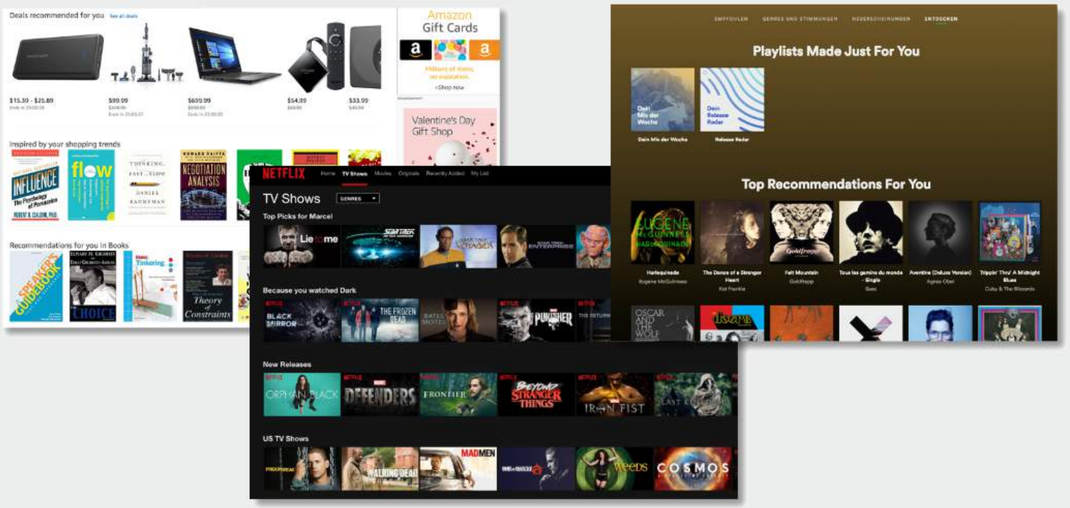




MARKET BASKET ANALYSIS

HISTORICAL TRANSACTIONS → ASSOCIATIVE RULE → RECOMMENDER SYSTEM





MARKET BASKET ANALYSIS

HISTORICAL TRANSACTIONS

ASSOCIATIVE RULES

INSIGHTS BETWEEN PRODUCTS

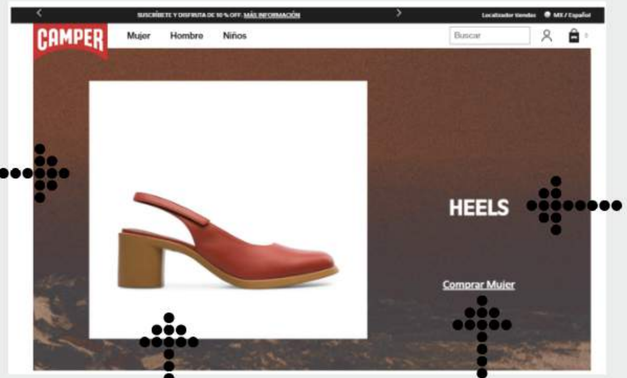
```
Rule: ['light cream'] -> ['chicken']
Support: 0.004532728969470737
Confidence: 0.29059829059829057
Lift: 4.84395061728395
====================================
Rule: ['mushroom cream sauce'] -> ['escalope']
Support: 0.005732568990801226
Confidence: 0.3006993006993007
Lift: 3.790832696715049
====================================
Rule: ['pasta'] -> ['escalope']
Support: 0.005865884548726837
Confidence: 0.3728813559322034
Lift: 4.700811850163794
====================================
Rule: ['herb & pepper'] -> ['ground beef']
Support: 0.015997866951073192
Confidence: 0.3234501347708895
Lift: 3.2919938411349285
====================================
Rule: ['tomato sauce'] -> ['ground beef']
Support: 0.005332622317024397
Confidence: 0.3773584905660377
Lift: 3.840659481324083
```

```
====================================
Rule: ['whole wheat pasta'] -> ['olive oil']
Support: 0.007998933475536596
Confidence: 0.2714932126696833
Lift: 4.122410097642296
====================================
Rule: ['pasta'] -> ['shrimp']
Support: 0.005065991201173177
Confidence: 0.3220338983050847
Lift: 4.506672147735896
====================================
Rule: ['frozen vegetables', 'chocolate'] -> ['shrimp']
Support: 0.005332622317024397
Confidence: 0.23255813953488375
Lift: 3.2545123221103784
====================================
Rule: ['ground beef', 'cooking oil'] -> ['spaghetti']
Support: 0.004799360085321957
Confidence: 0.5714285714285714
Lift: 3.2819951870487856
====================================
Rule: ['spaghetti', 'frozen vegetables'] -> ['ground beef']
Support: 0.008665511265164644
Confidence: 0.31100478468899523
Lift: 3.165328208890303
====================================
```

E-COMMERCE

CUSTOMIZED WEBSITE FOR EVERY USER BASED ON THEIR DATA

CHANGE BACKGROUND

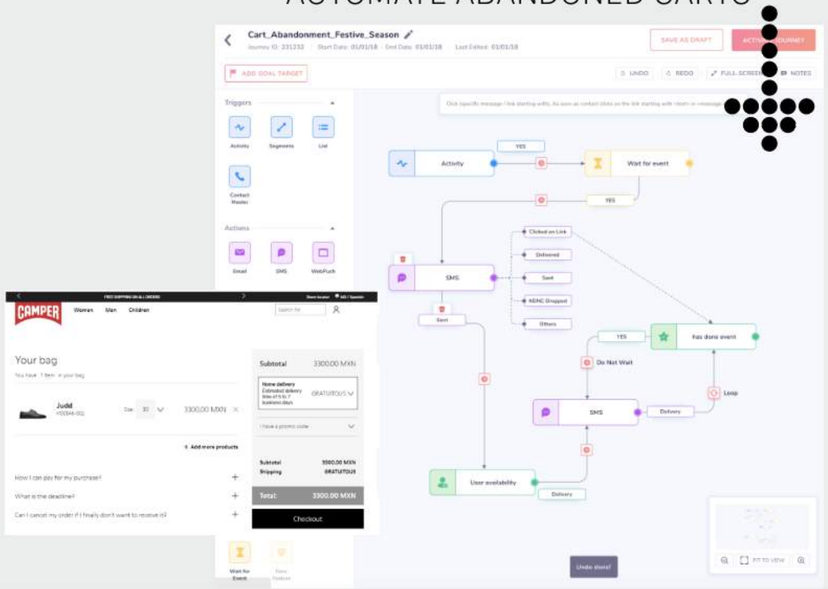


CHANGE TEXT

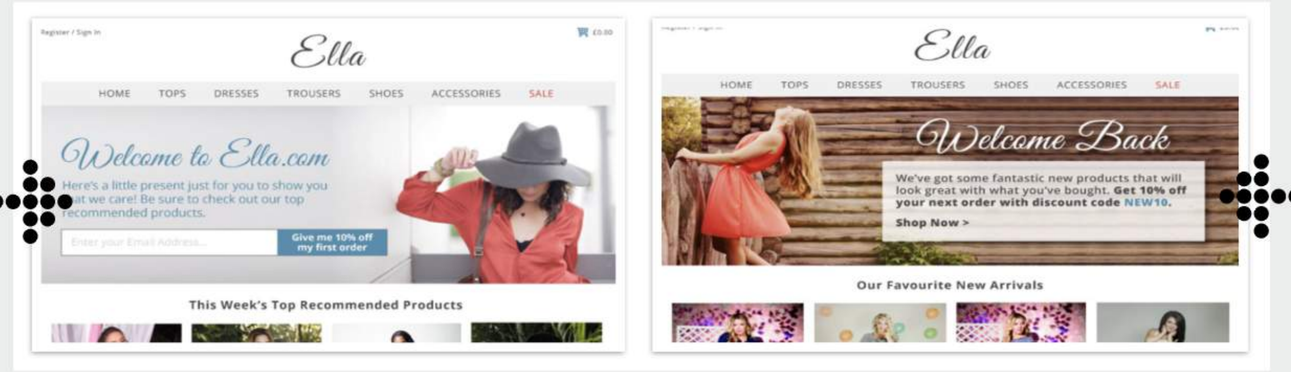
CHANGE BANNER

BANNER FOR USERS THAT SEARCHED IN "BOOTS" CATEGORY

AUTOMATE ABANDONED CARTS

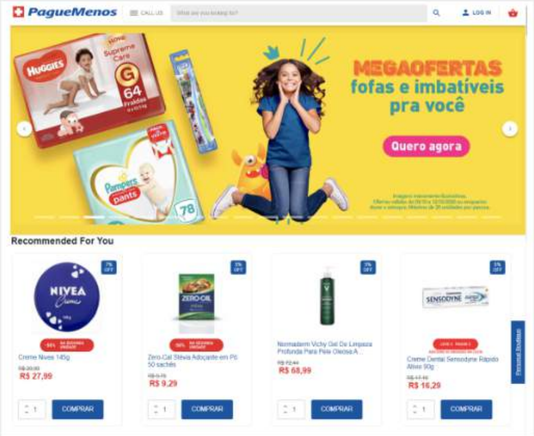


CUSTOMIZATION BASED ON SEX

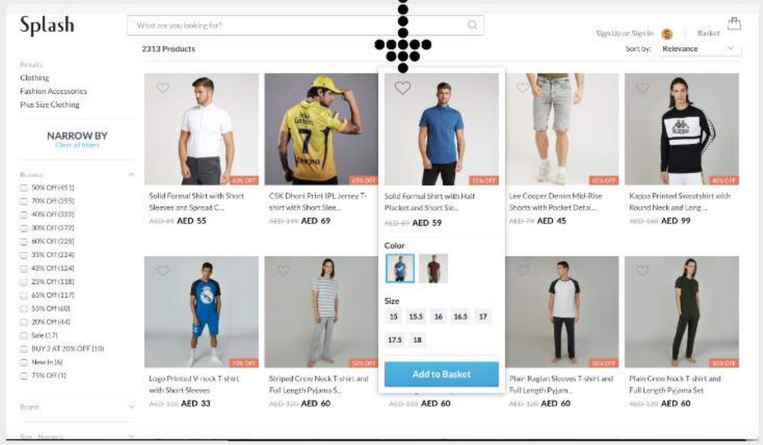


E-COMMERCE

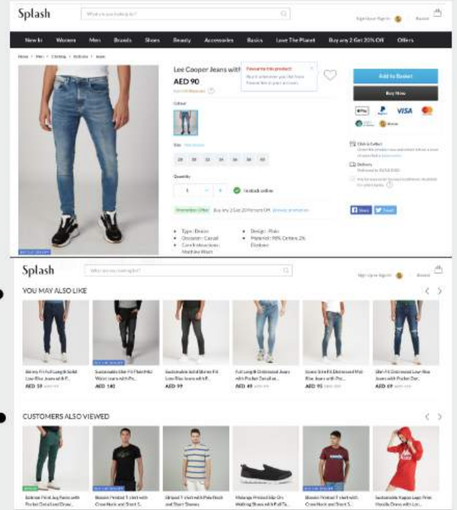
1. -CUSTOM RECOMMENDATIONS FOR YOU IN " HOME PAGE"



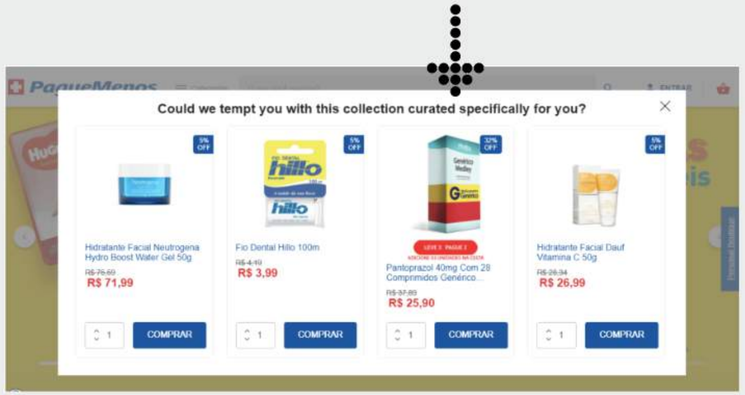
2. - BOUTIQUE PERSONAL



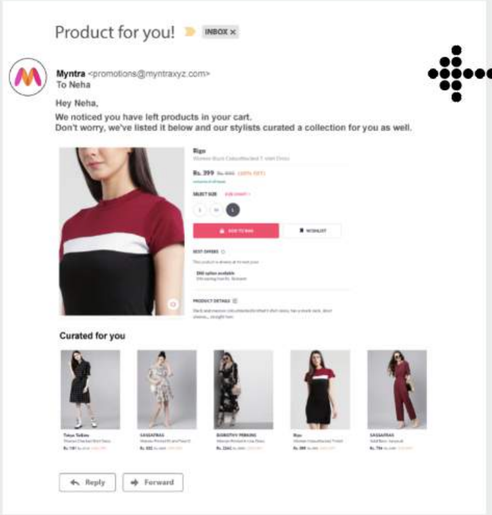
2. - RECOMMENDER SYSTEM WHILE PAYING:



3. - BUYER VS. BROWSER: CUSTOM OFFERS



5. - INTERACTIONS VÍA E-MAIL OR PUSH



WORKFLOW

Business analysis

Extract , transform and load

Variables analysis: business insights

clustering, PCA, correlation matrix

classificator type of clients

regressor demand prediction

classificator buyer/no buyer

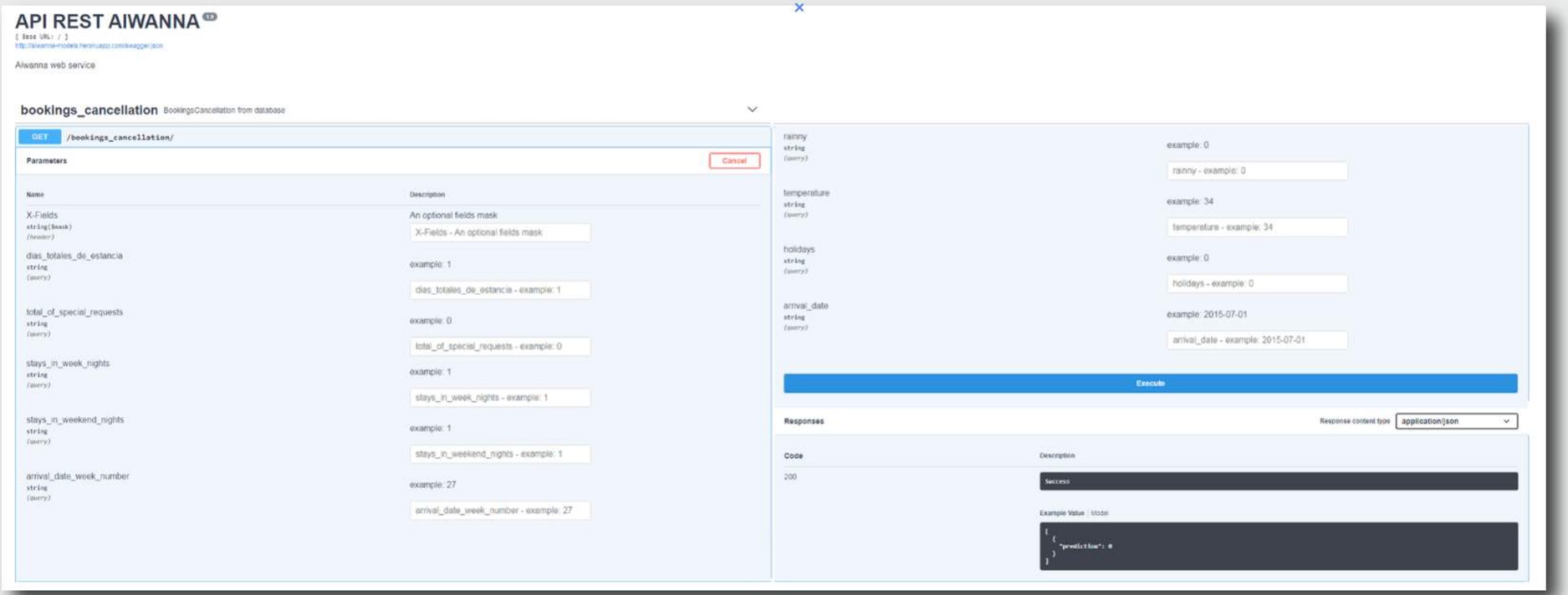
recommender system

classificator churn rate

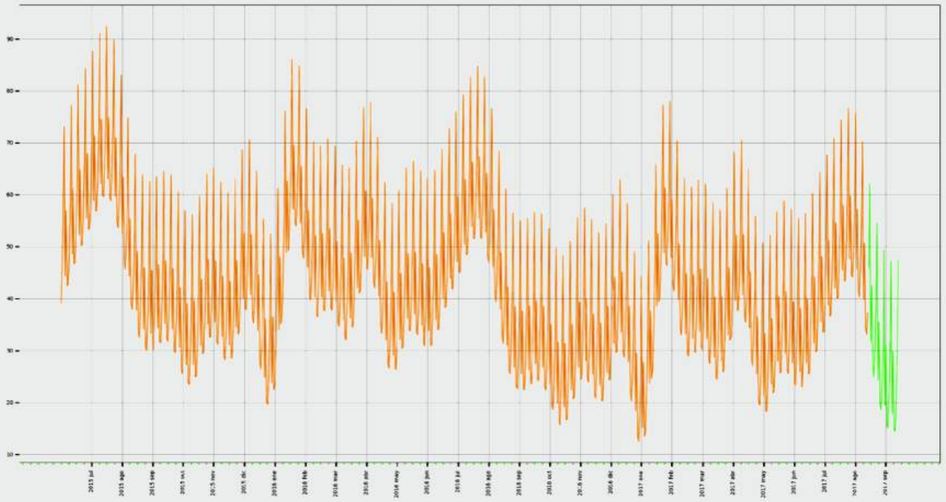
customer segementation

deployment

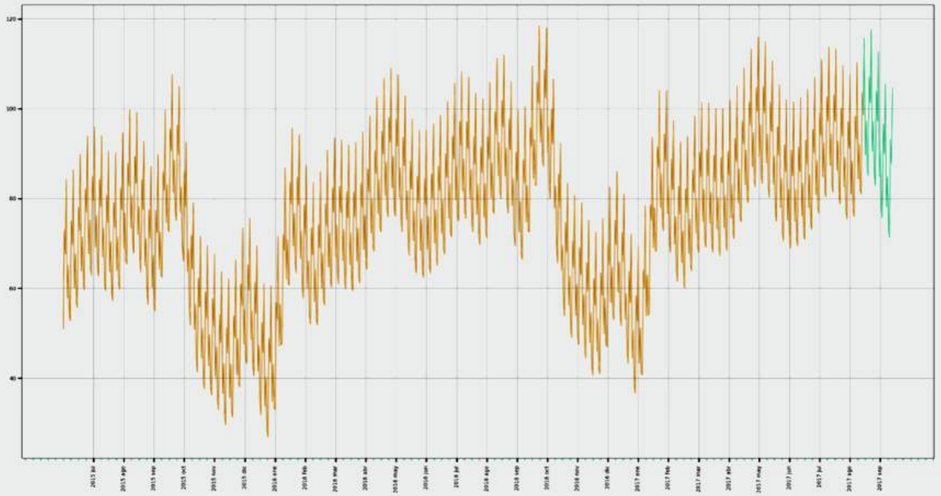
AUGMENTED ANALYTICS

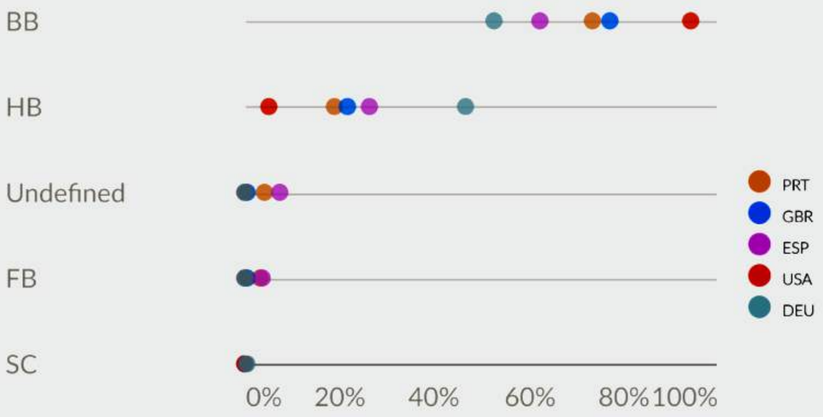


FORECASTING CLIENTES FROM PORTUGAL



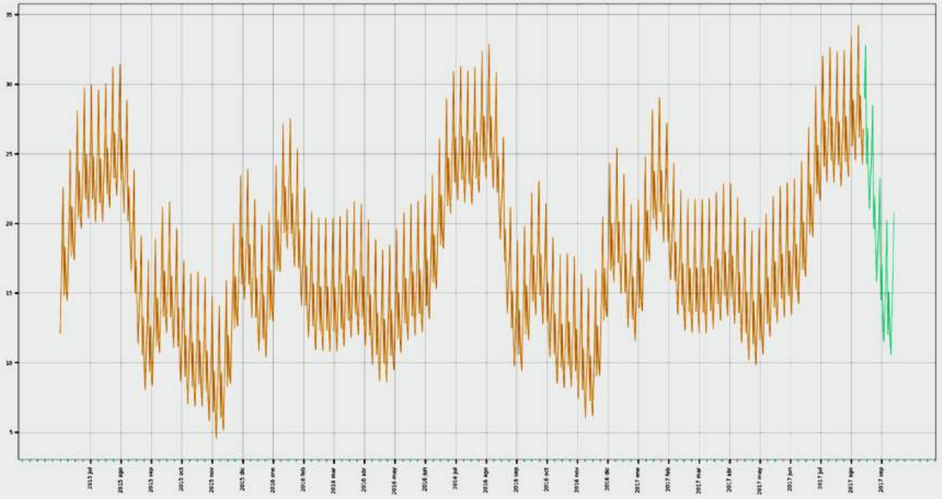
FORECASTING "BRED & BREAKFAST"

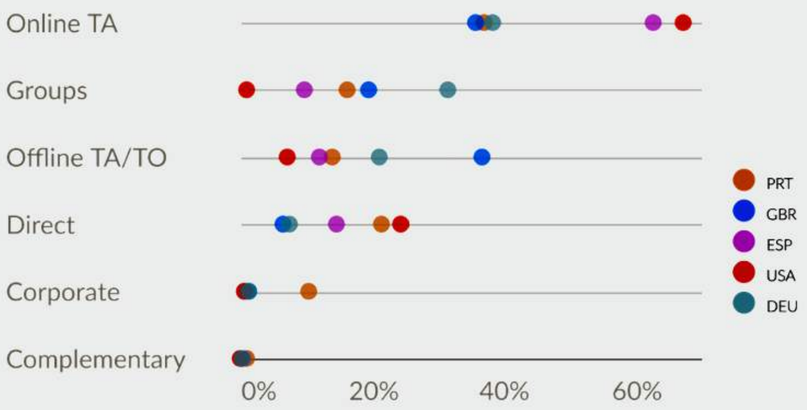




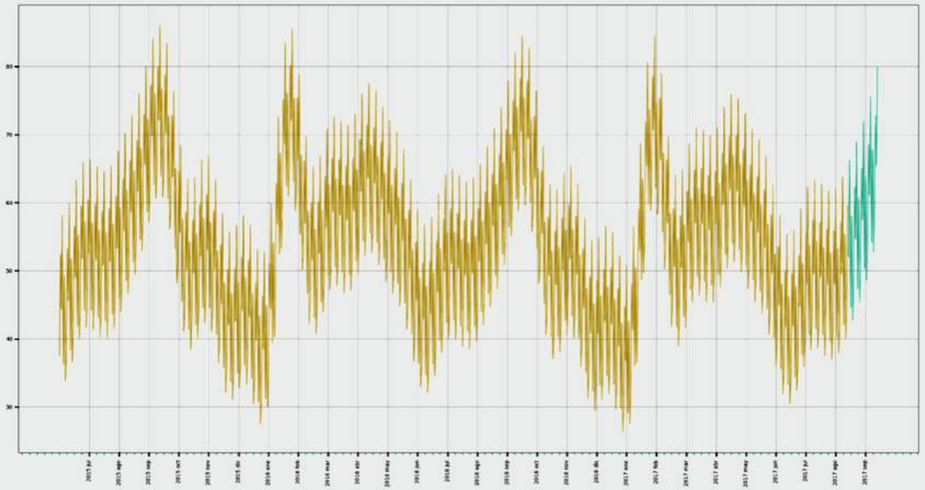
AUGMENTED ANALYTICS

FORECASTING SALES CHANNEL "DIRECT"





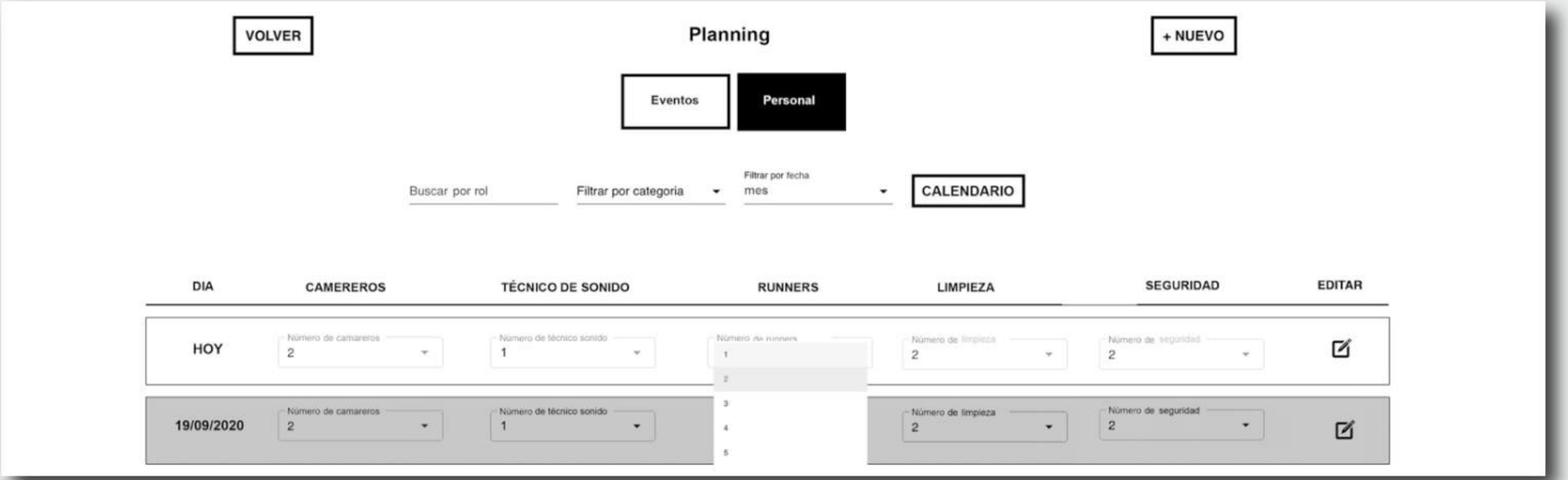
FORECASTING "TYPE OF ROOM"

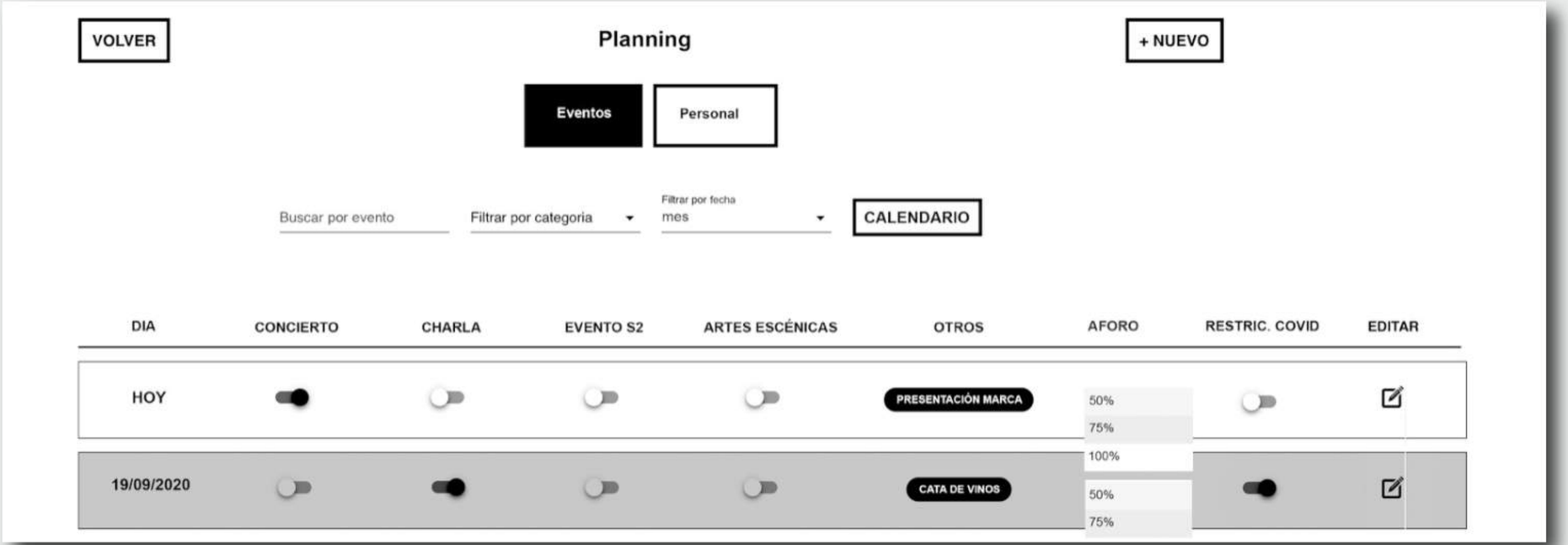


StaysInWeekendNights

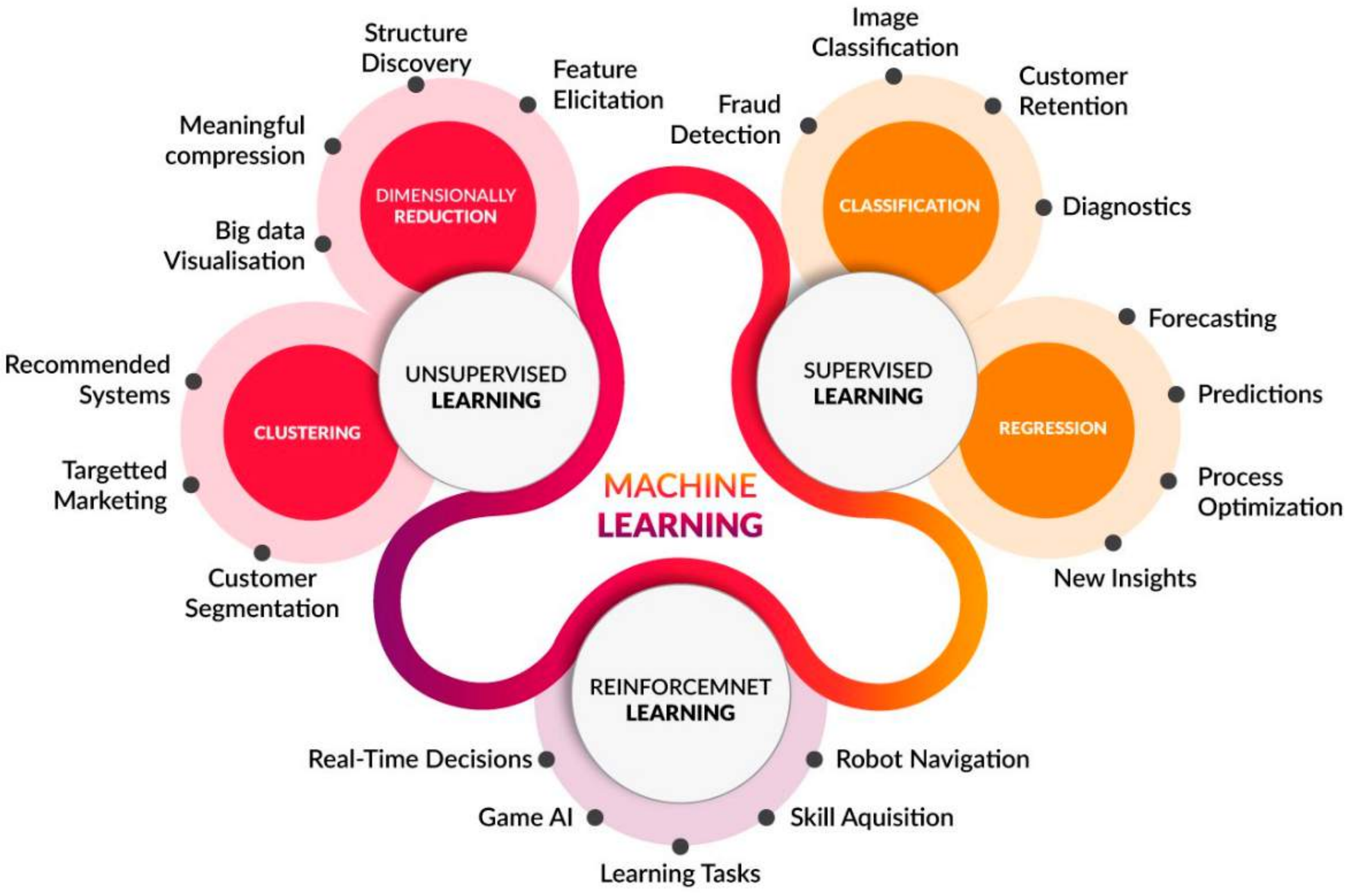
ReservedRoomType	Reserved...	0	1	2	3
	A	9371	5881	6972	414
	B	3	-	-	-
	C	242	229	394	25
	D	2085	1351	3239	238
	E	1416	1011	2040	170
	F	358	220	449	33
	G	480	356	670	41
	H	219	143	211	15
	L	5	1	-	-
	P	2	-	-	-

DATA CAPTURE



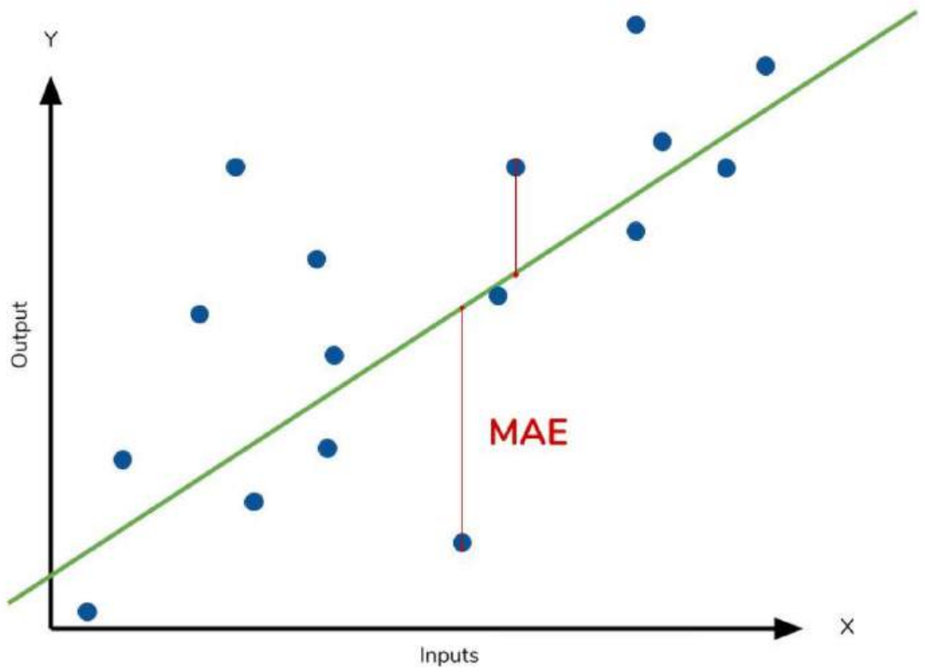


ALGORITHM // Development


MACHINE LEARNING
UNSUPERVISED LEARNING
DIMENSIONALLY REDUCTION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
CLUSTERING
Recommended Systems
Targetted Marketing
Customer Segmentation
SUPERVISED LEARNING
CLASSIFICATION
Image Classification
Fraud Detection
Customer Retention
Diagnostics
REGRESSION
Forecasting
Predictions
Process Optimization
New Insights
REINFORCEMNET LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

ALGORITHM // Evaluation

Regression evaluation:



Classification evaluation:

Test Results	Yes	No
Positive	**True Positive (TP):** Person with coronavirus tested positive	**False Positive (FP):** Person without coronavirus tested positive
Negative	**False Negative (FN):** Person with coronavirus tested negative	**True Negative (TN):** Person without coronavirus tested negative

Evaluate classification

Use cases

// Development process

Algorithm

artificial vision

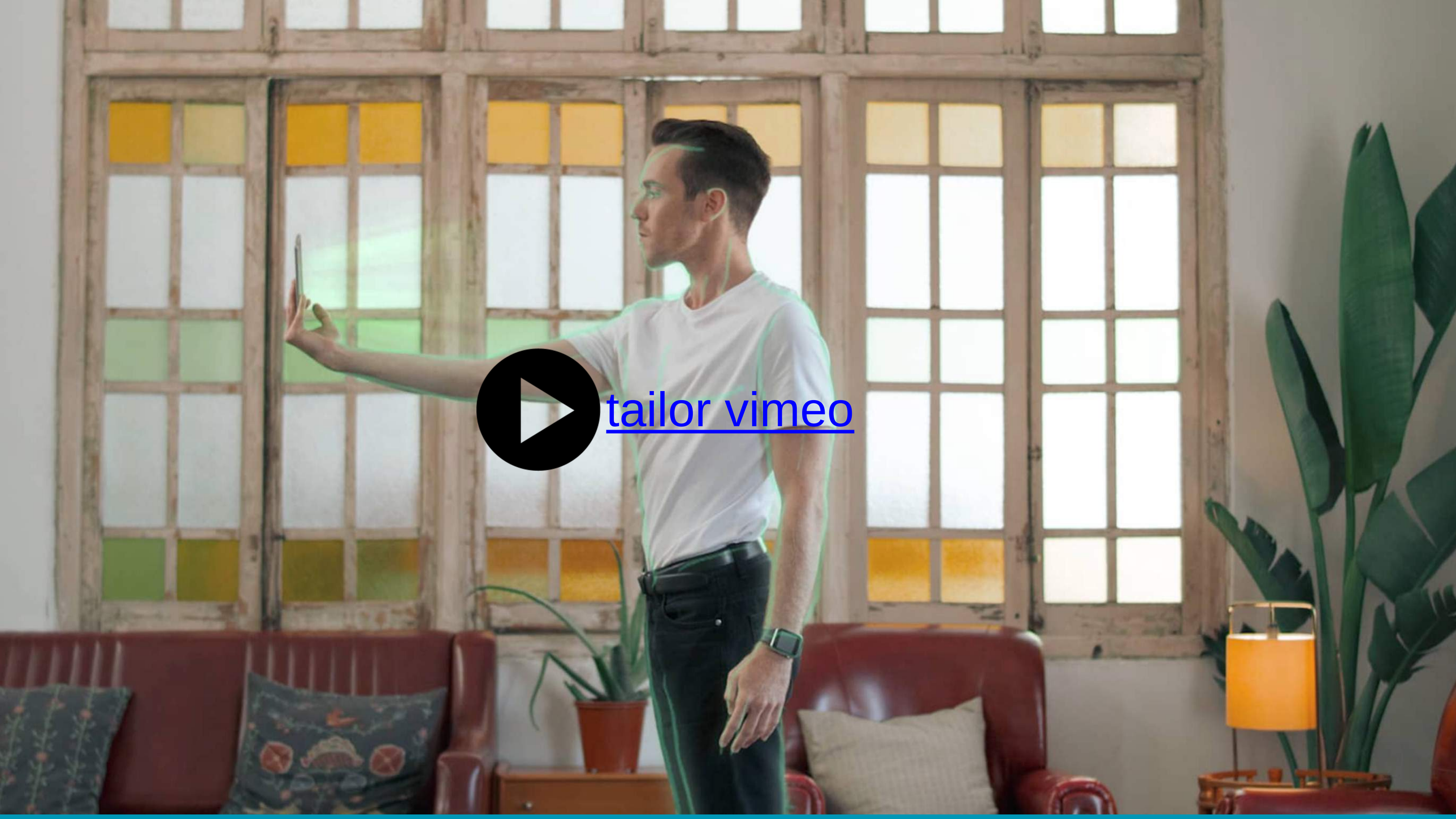
Business problem statement

- State a predictive problem to improve client business value.
- Know the impact of predictive result.
- This is should be in design challenge

How might we (goals) so that (impact/outcome)?

Example:

How to target promotion or offer to potential customer based on previous customer?


tailor vimeo

Define **variables**

- Define variable (column) that will be needed for predict the label
- The variable should be consistent in measurement (Example: Currency is Rupiah or Dollar?)
- The variable should be consistent in specifications (Example: NPWP or KTP?)
- Variable should be associated with unit analysis, vice versa

Example:

What kind of product that customer might to buy?

Variable : Product pair, Order ID, Frequency

Use Case tailor

Define **label**

- Define Label that will be predicted from the variable
- The label came after variable, because it's predictive
- Label should be associated with unit analysis, vice versa

Example:

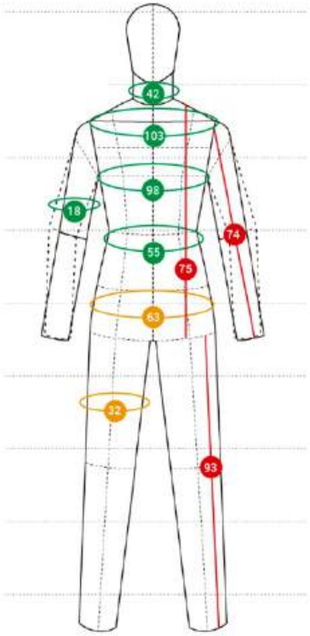
How might we choose Supplier/Vendor so that we get fair price for new supplier?

Variable: Lead time (aggregate), cost (aggregate), services, insurance policy, profile

Label: Average price

USE CASE // tailor

The Human Body



Set the **threshold**

- Either rank (relative / absolute) or certain score
- Threshold should be set for achieve success criteria
- Should be adjust by business condition
- Threshold should be associated with unit analysis
- It should be associated with Label, vice versa

Example:

How much demand on the several periods ahead, so we can restock properly?

Success Criteria: Minimize the occurrence of understock and overstock phenomenon

Threshold: Stock amount less than next certain time unit demand

Success criteria

- Should be either minimize or maximize
- It's should be derived by statistical criteria
- The statistical criteria should be linearly correlated with business impact
- If represent by money, human resources or time, it's better
- It should be associated with label, vice versa

Example:

When does the right time to do restocking for particular product and particular distribution center?

Success Criteria: Minimize understock & overstock phenomenon

Papers

Multi-Person Pose Estimation using Part Affinity Fields

Zhe Cao, Shih-En Wei, Tomas Simon, Yaser Sheikh
Carnegie Mellon University

THE ROBOTICS INSTITUTE

Paper Analysis

What **useful info** could we extract from a paper?

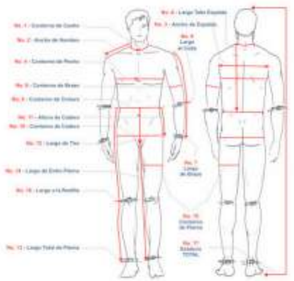
Learn to read a paper

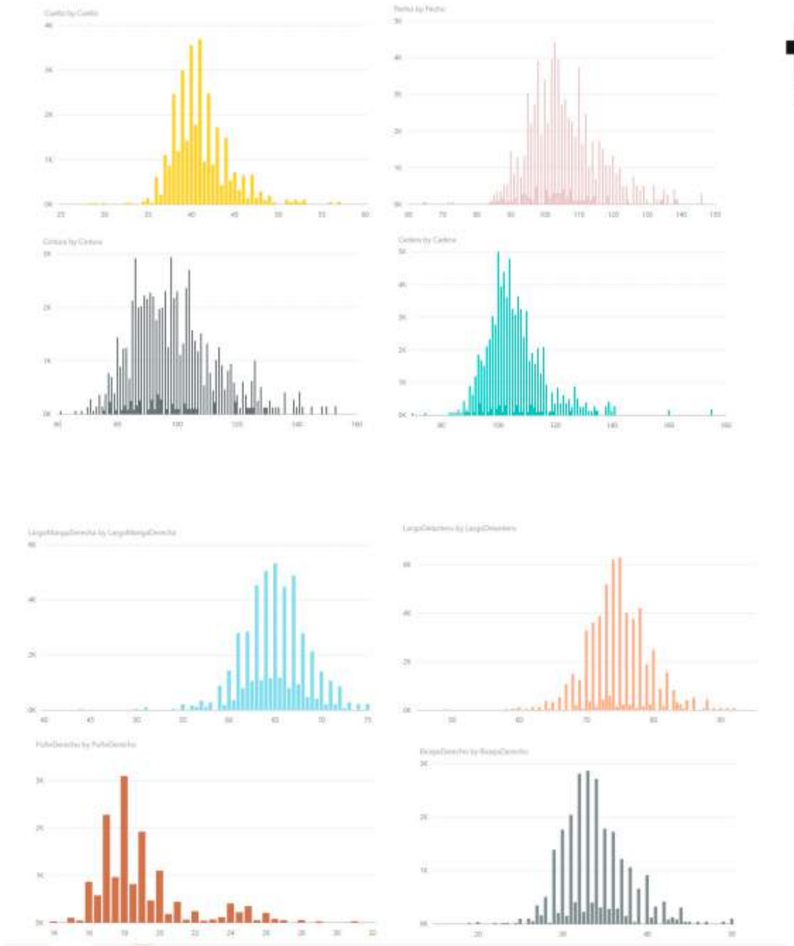
AI PROJECT

3 // DATA PROJECT

USE CASE // tailor

Visualice

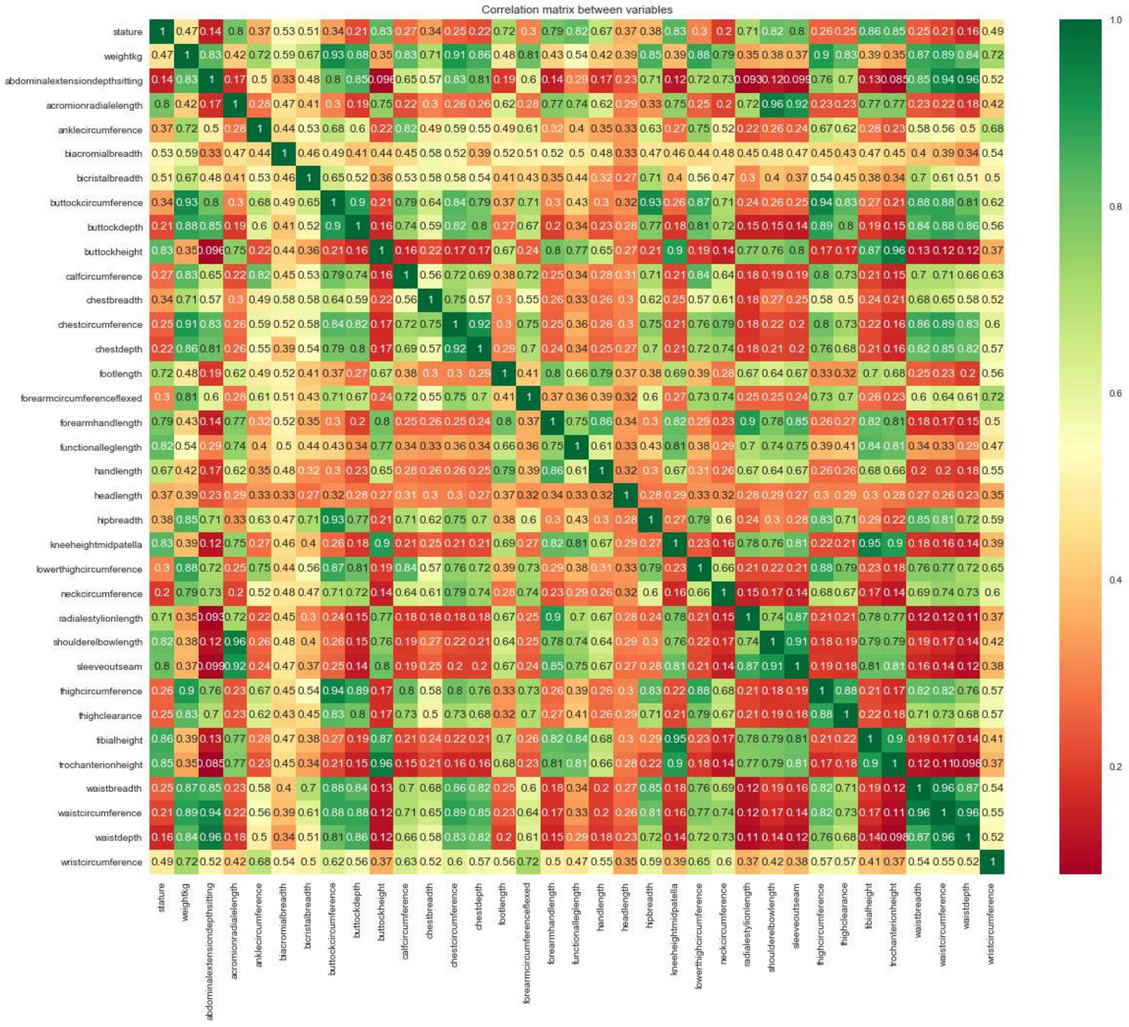




tailor

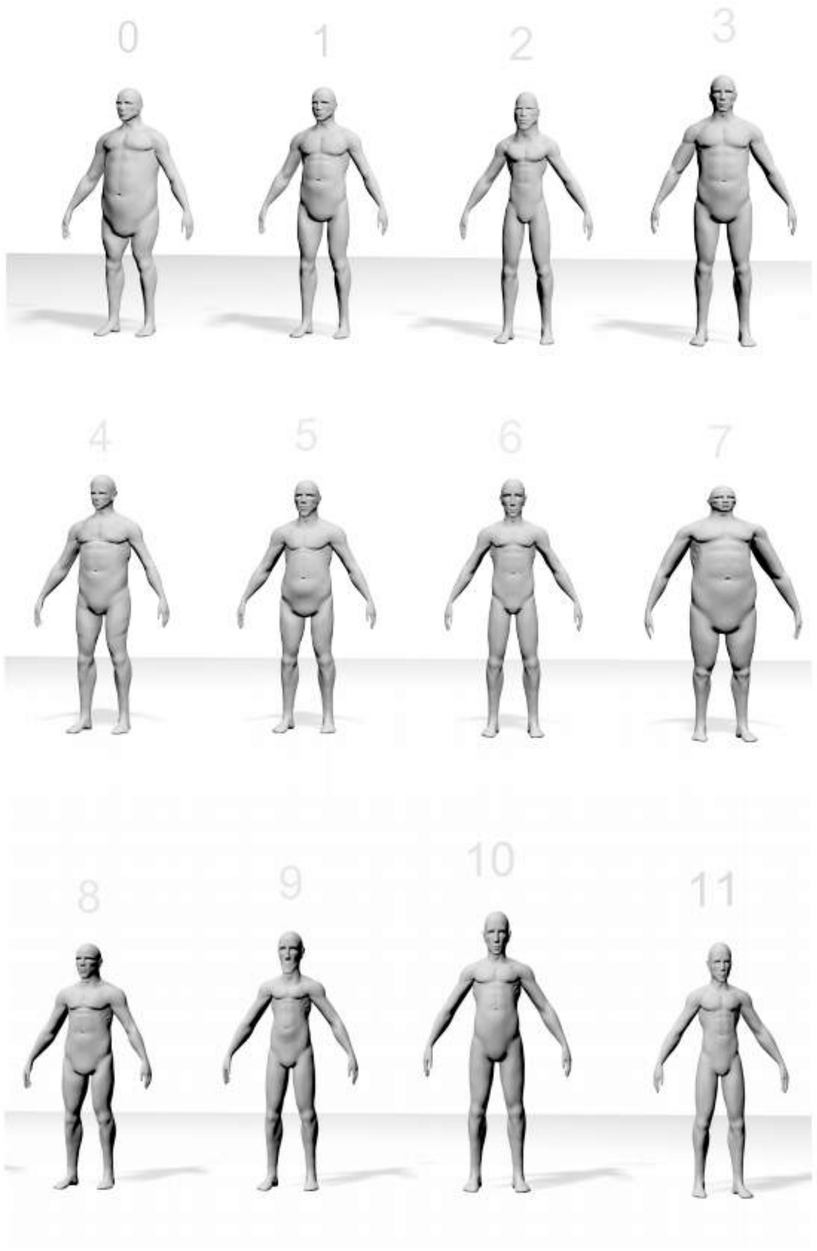
USE CASE // tailor

Analize



USE CASE // tailor

Clustering tailor



tailor

USE CASE // tailor

Clustering tailor

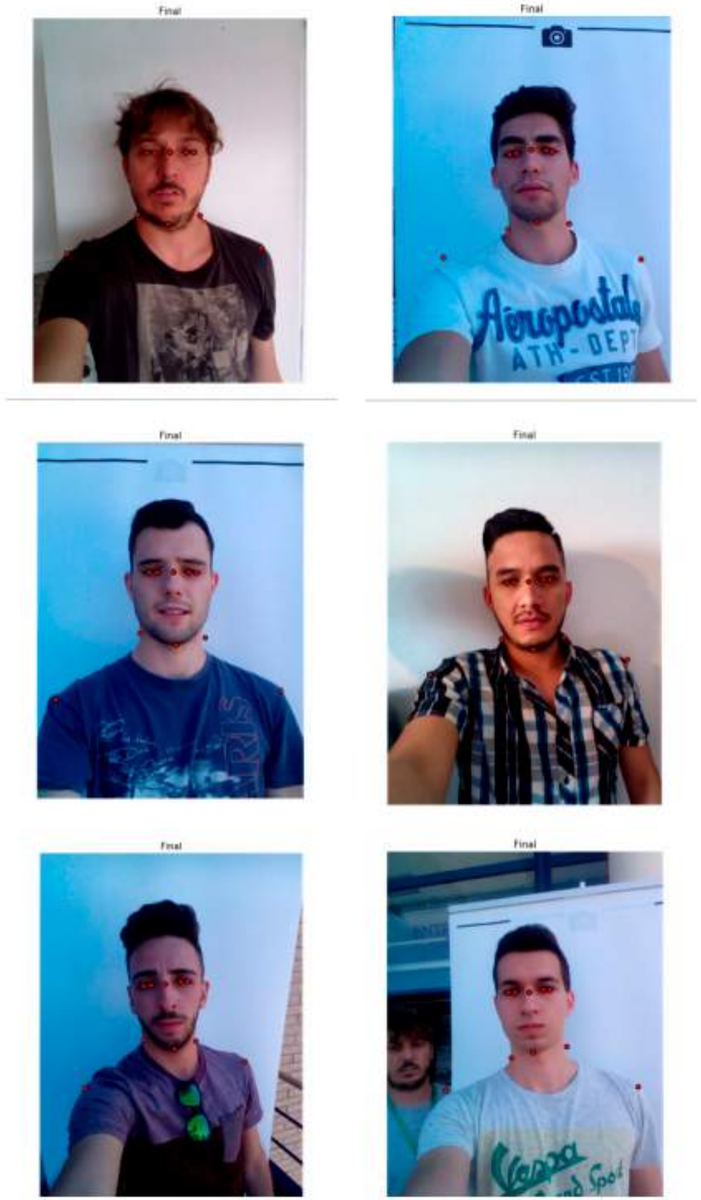
	Cuello	Pecho	Cintura	Cadera	AnchoEspalda	LargoDelantero	LargoEspalda	LargoMangaDerecha	LargoMangaIzquierda	PuñoDerecho	PuñoIzquierdo	BicepsDerecho	BicepsIzquierdo
5	30,9	67,0	63,4	71,4	34,7	53,8	53,8	48,4	48,4	14,8	14,8	22,2	22,2
4	37,7	91,8	81,0	93,3	43,3	69,7	69,6	62,1	62,1	17,1	17,4	29,4	29,4
0	39,3	98,1	87,3	99,2	45,3	73,6	73,6	65,1	65,1	17,7	18,2	31,9	31,9
2	40,3	103,6	96,4	103,0	45,4	71,6	71,5	62,9	62,8	18,7	19,2	33,1	33,1
1	41,0	105,4	95,8	106,6	47,3	77,9	78,0	68,5	68,5	18,9	19,4	34,3	34,3
7	42,2	110,9	105,9	109,2	46,9	74,5	74,5	64,1	64,1	19,8	20,3	34,8	34,8
3	44,1	118,0	114,3	115,6	48,9	78,2	78,2	65,9	65,9	20,6	21,1	37,0	37,0
8	46,2	127,4	127,4	125,4	50,7	80,6	80,4	65,3	65,3	22,0	22,4	40,1	40,1
6	48,9	142,2	147,6	145,3	54,4	85,8	85,4	65,6	65,6	23,3	23,6	45,1	45,1

USE CASE // tailor

first dataset 150 img

USE CASE // tailor

first results



tailor

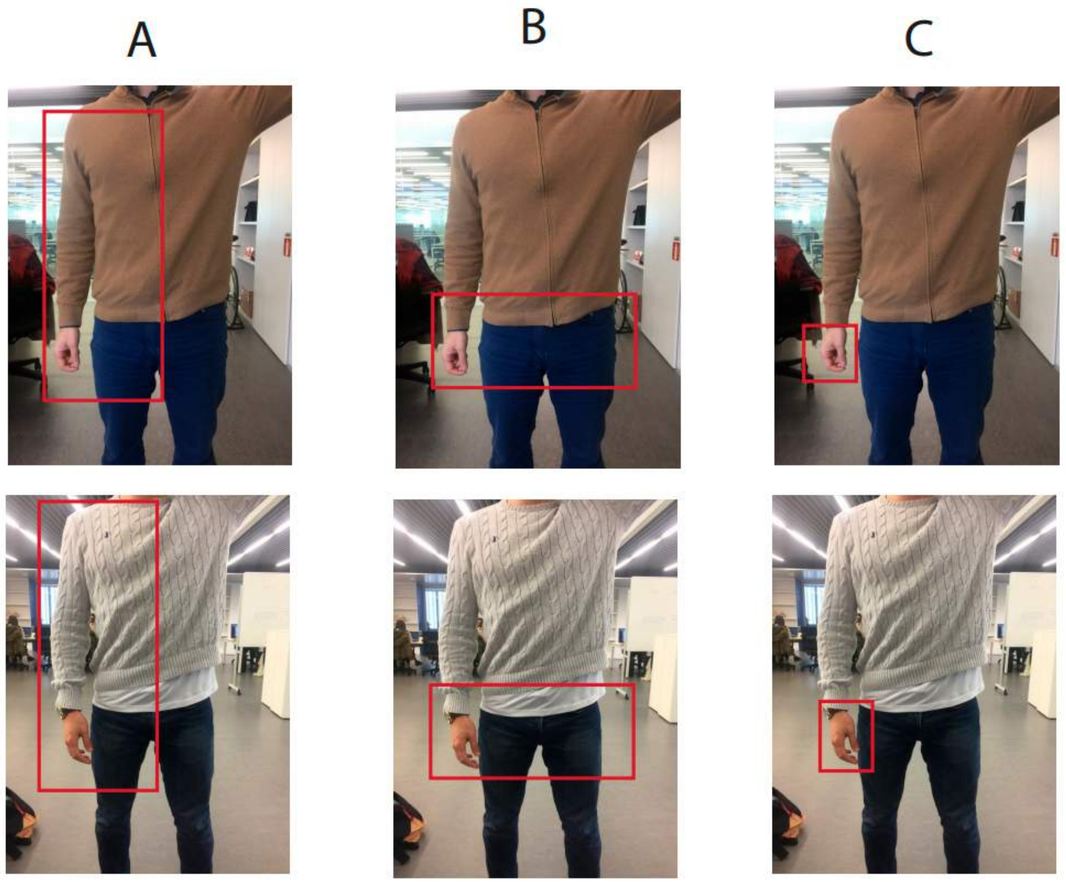
USE CASE // tailor

Research distance classificator

tailor

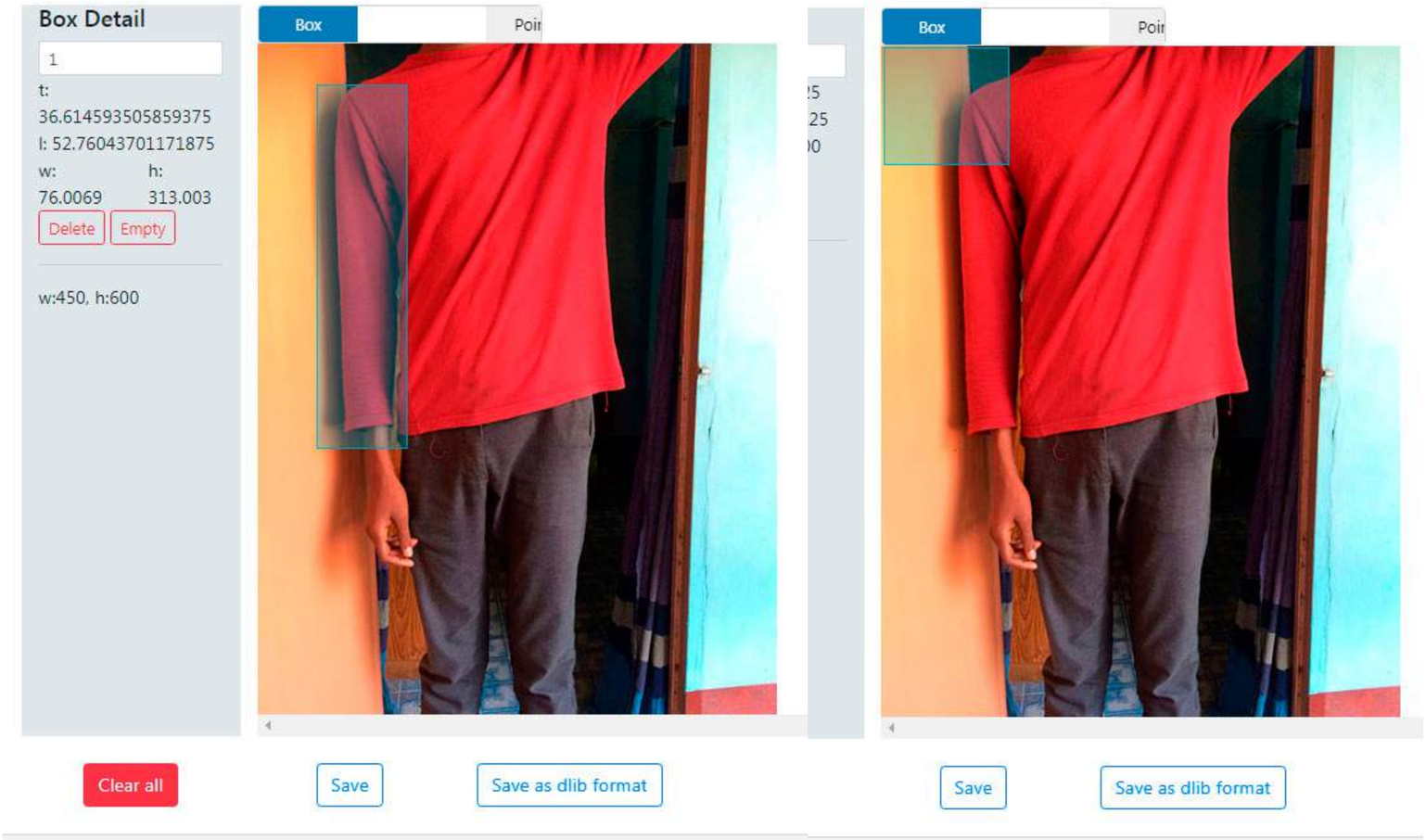
USE CASE // tailor

Research object detector



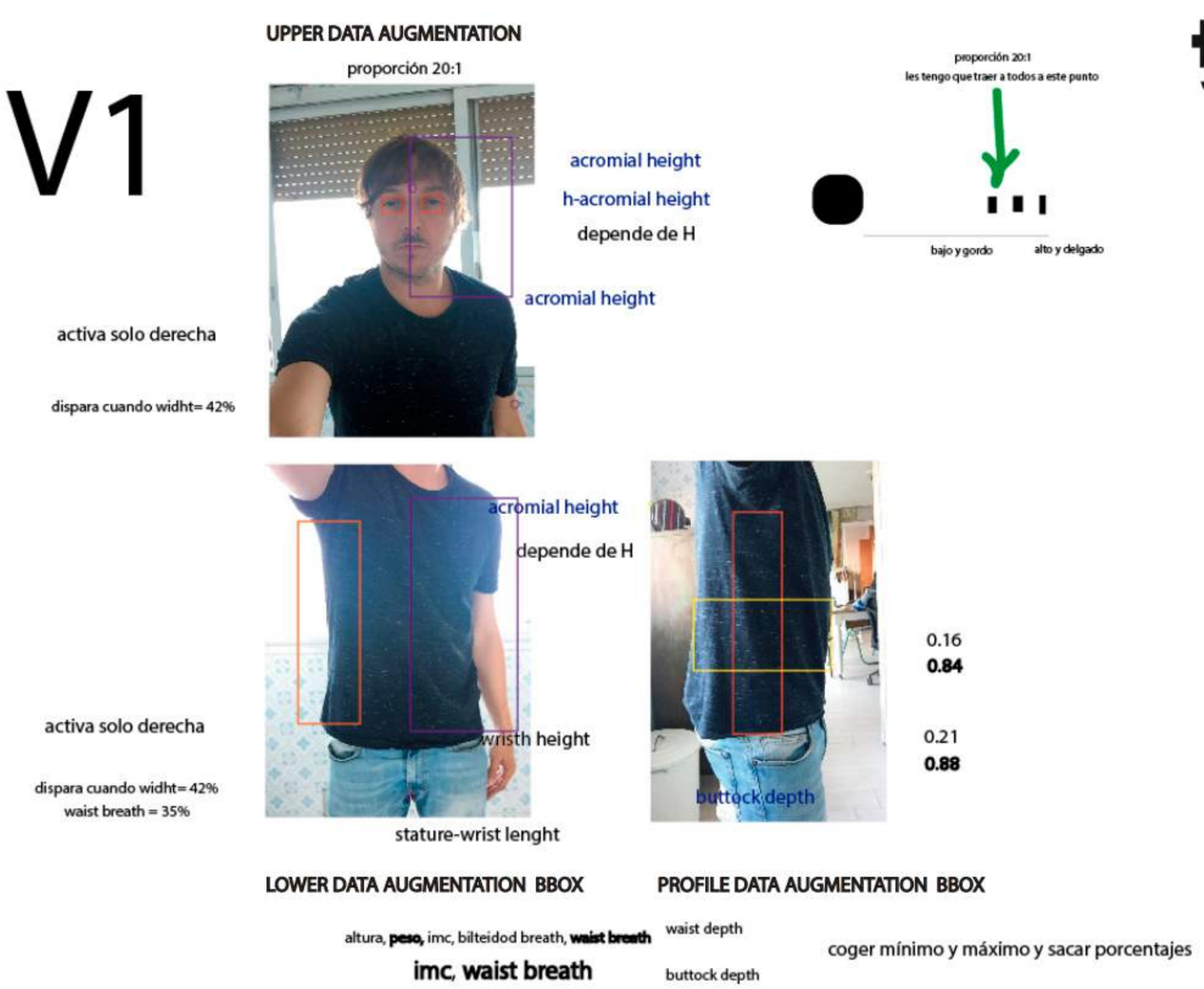
USE CASE // tailor

Research labeling



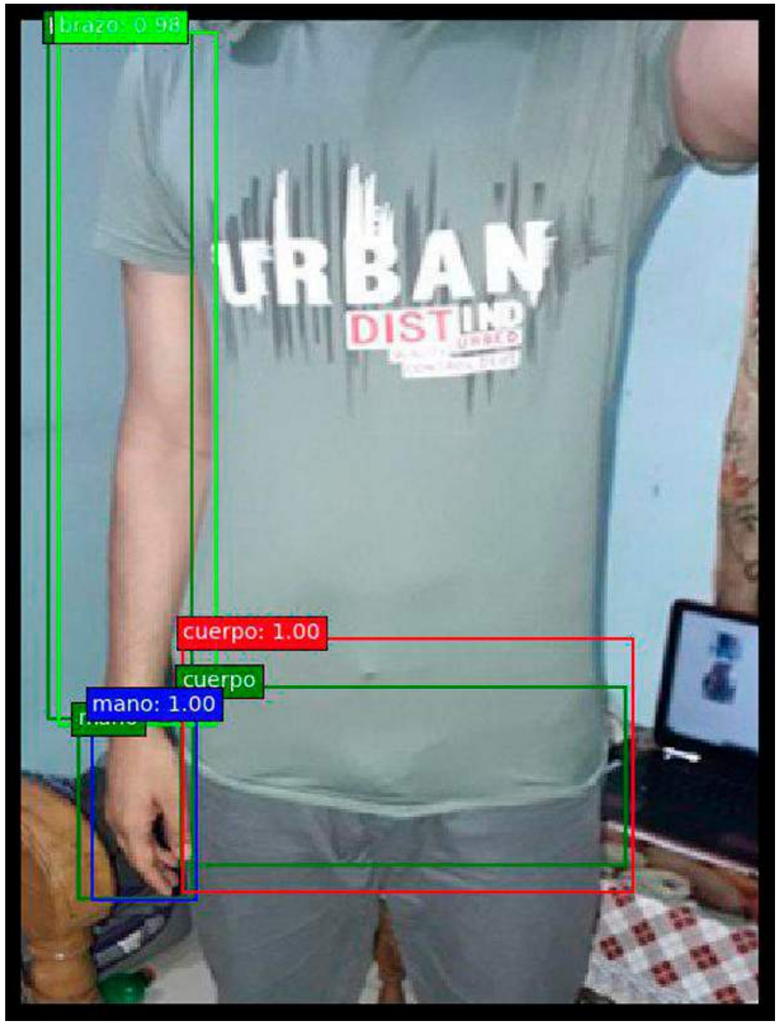
USE CASE // tailor

Research labeling



USE CASE // tailor

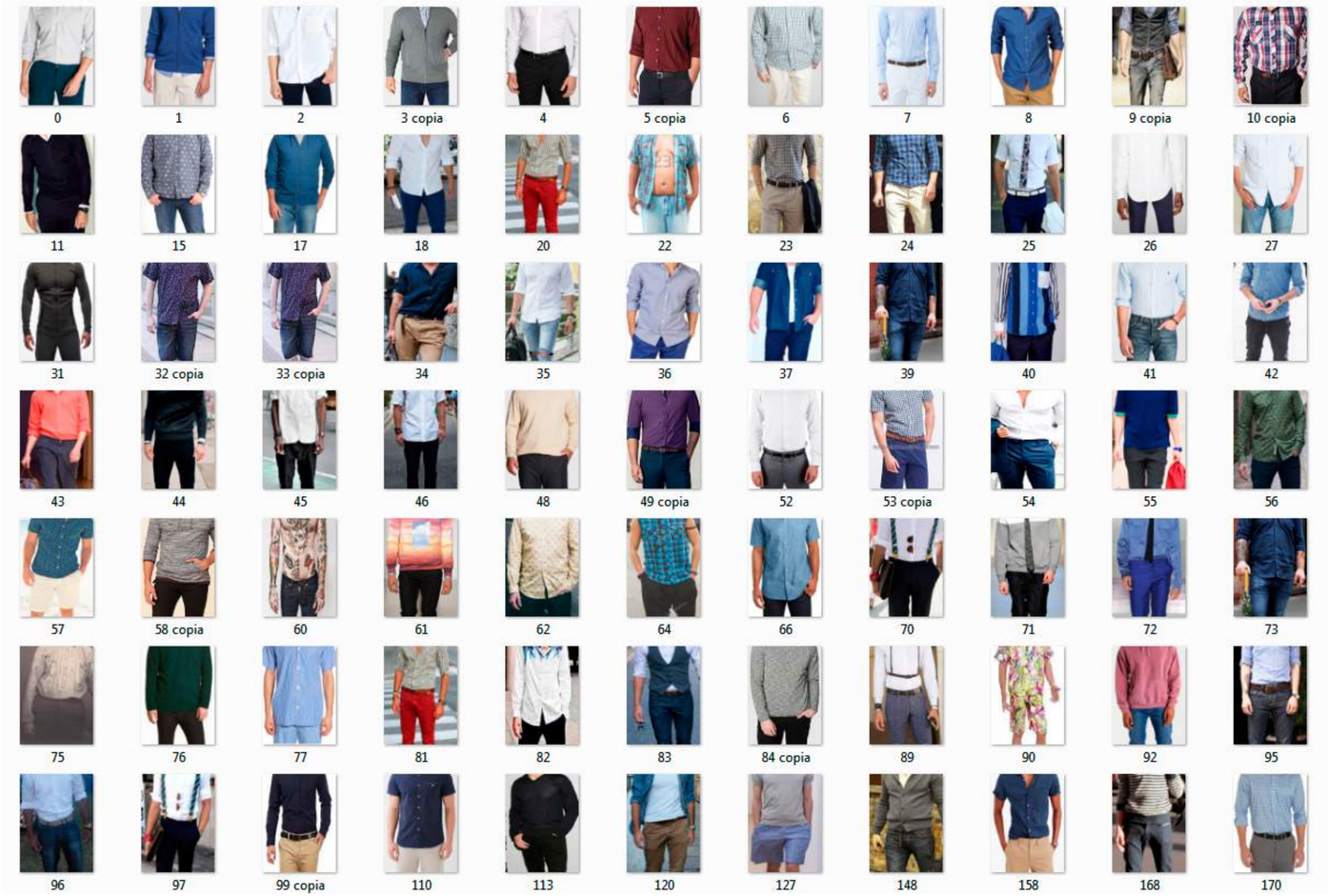
Research labeling



tailor

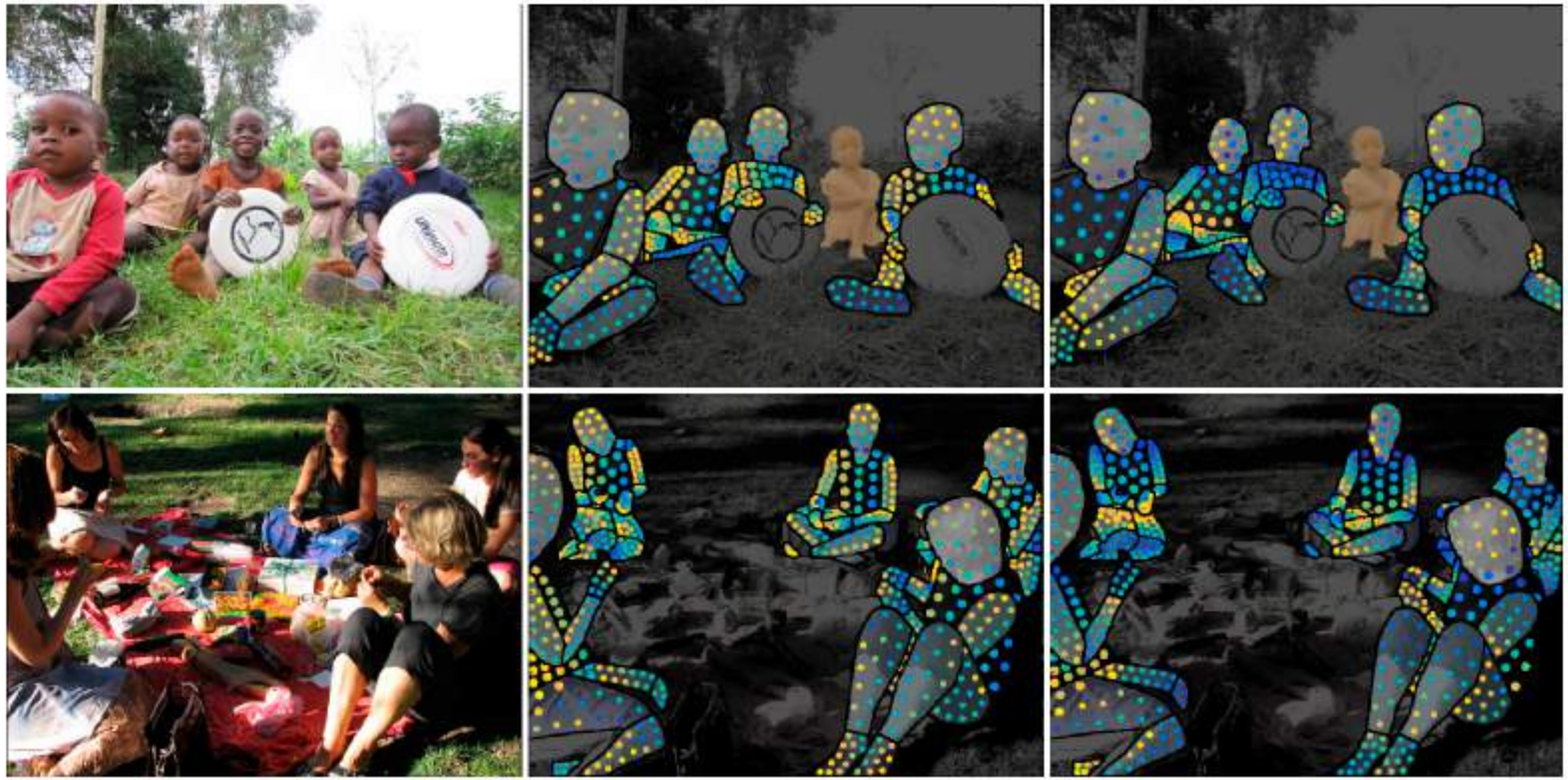
USE CASE // tailor

Second dataset



USE CASE // tailor

COCO dataset



USE CASE // tailor

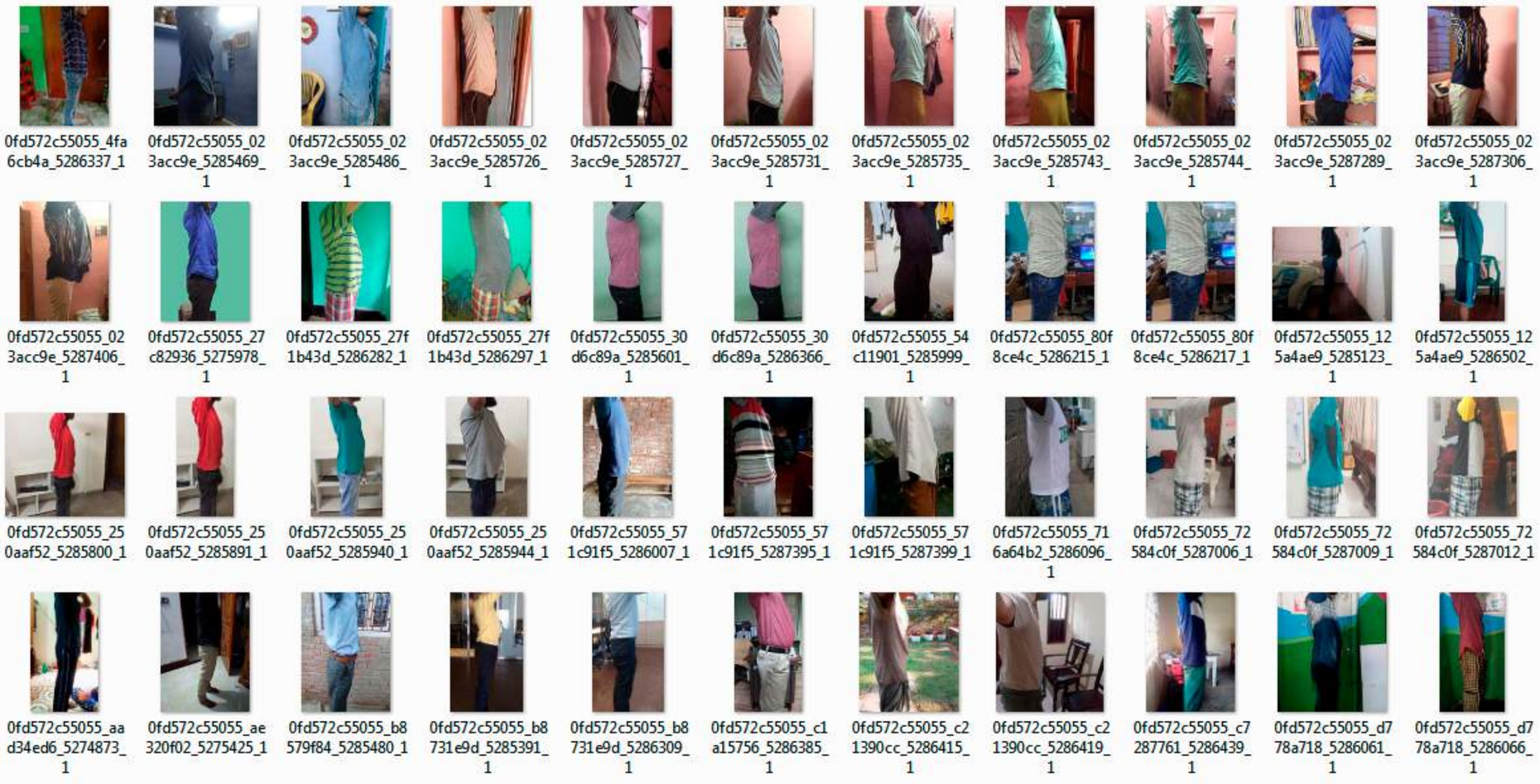
tailor dataset

USE CASE // tailor

tailor dataset

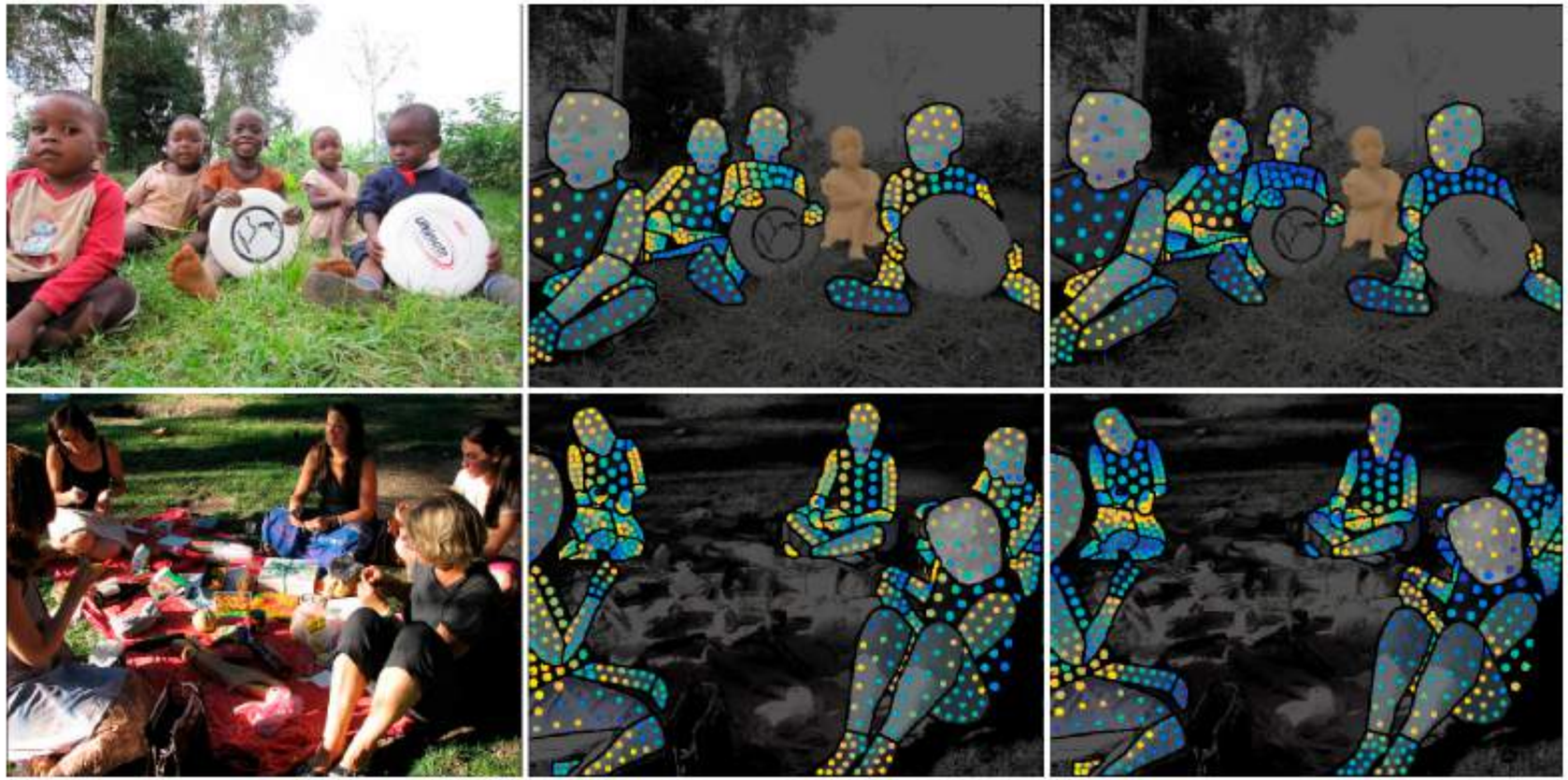
USE CASE // tailor

tailor dataset



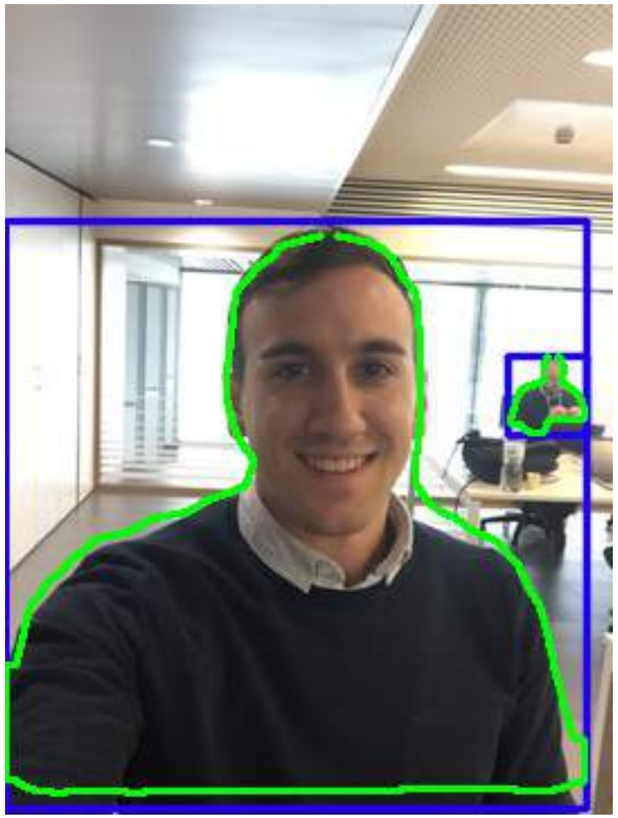
USE CASE // tailor

COCO dataset



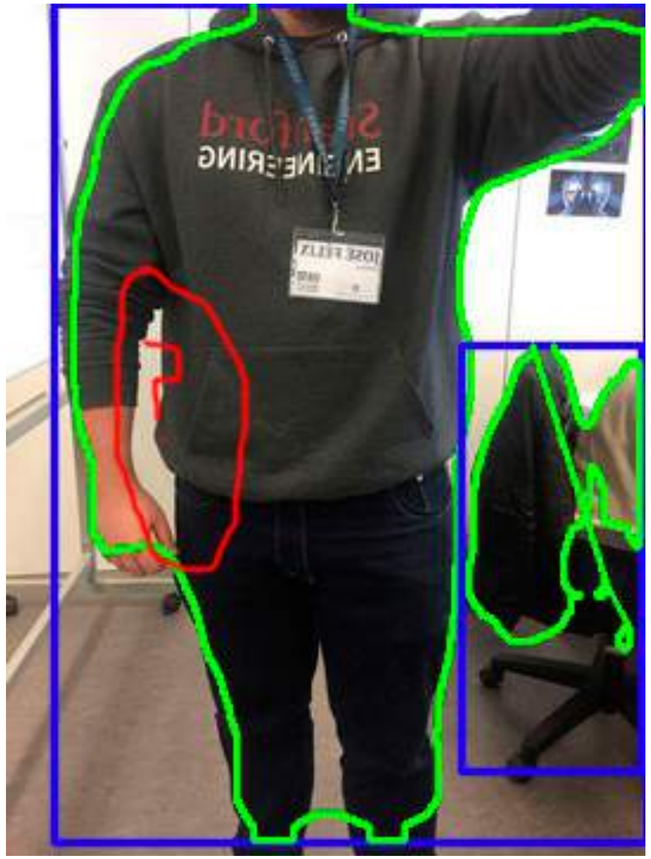
USE CASE // tailor

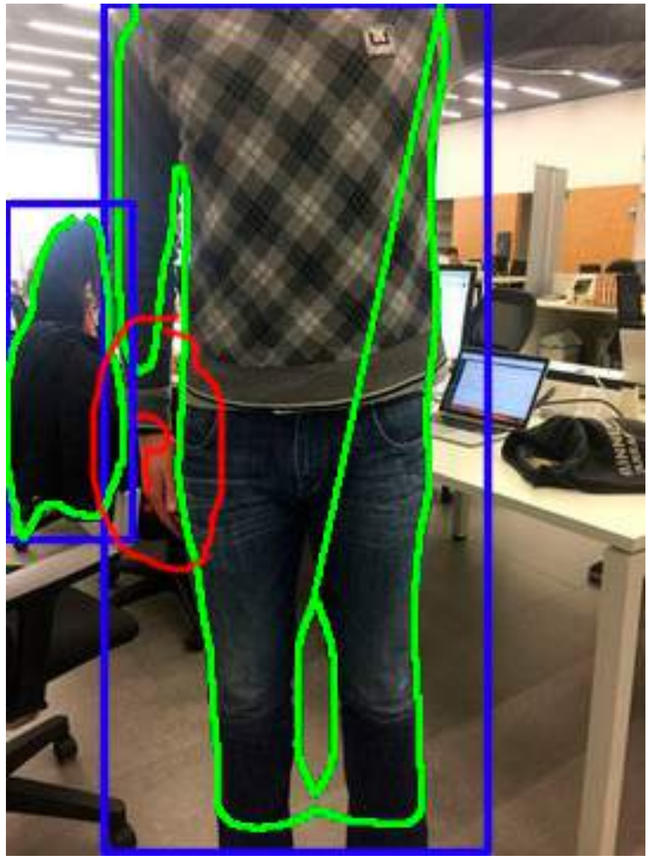
COCO + tailor results



USE CASE // tailor

Research distance classificator



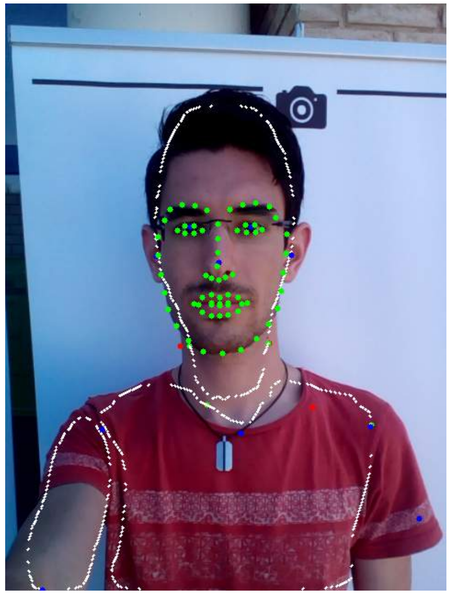
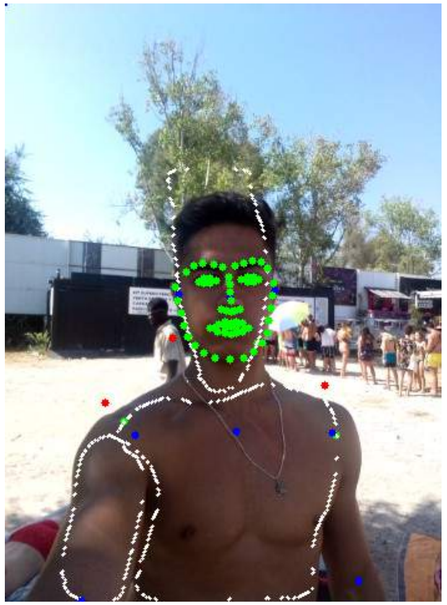
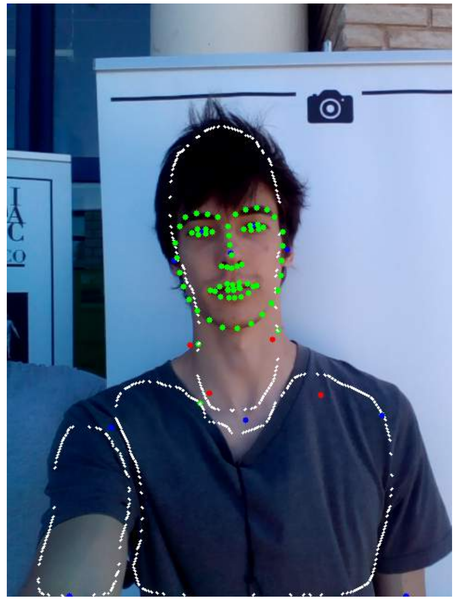
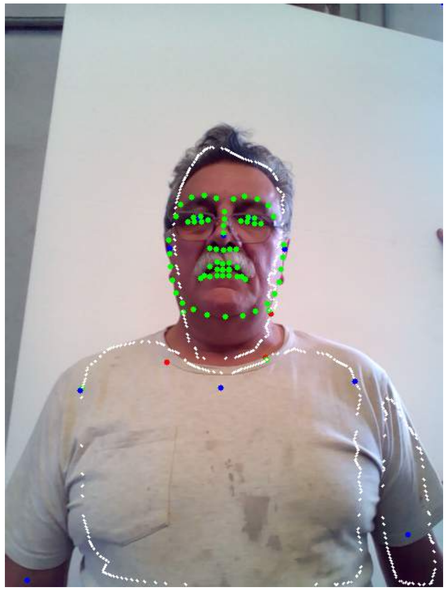
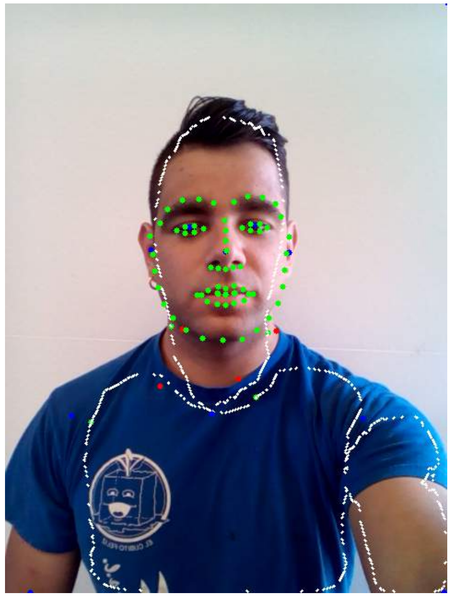


USE CASE // tailor

Segmentation

USE CASE // tailor

Face landmarks + segmentation

USE CASE // tailor

segmentation

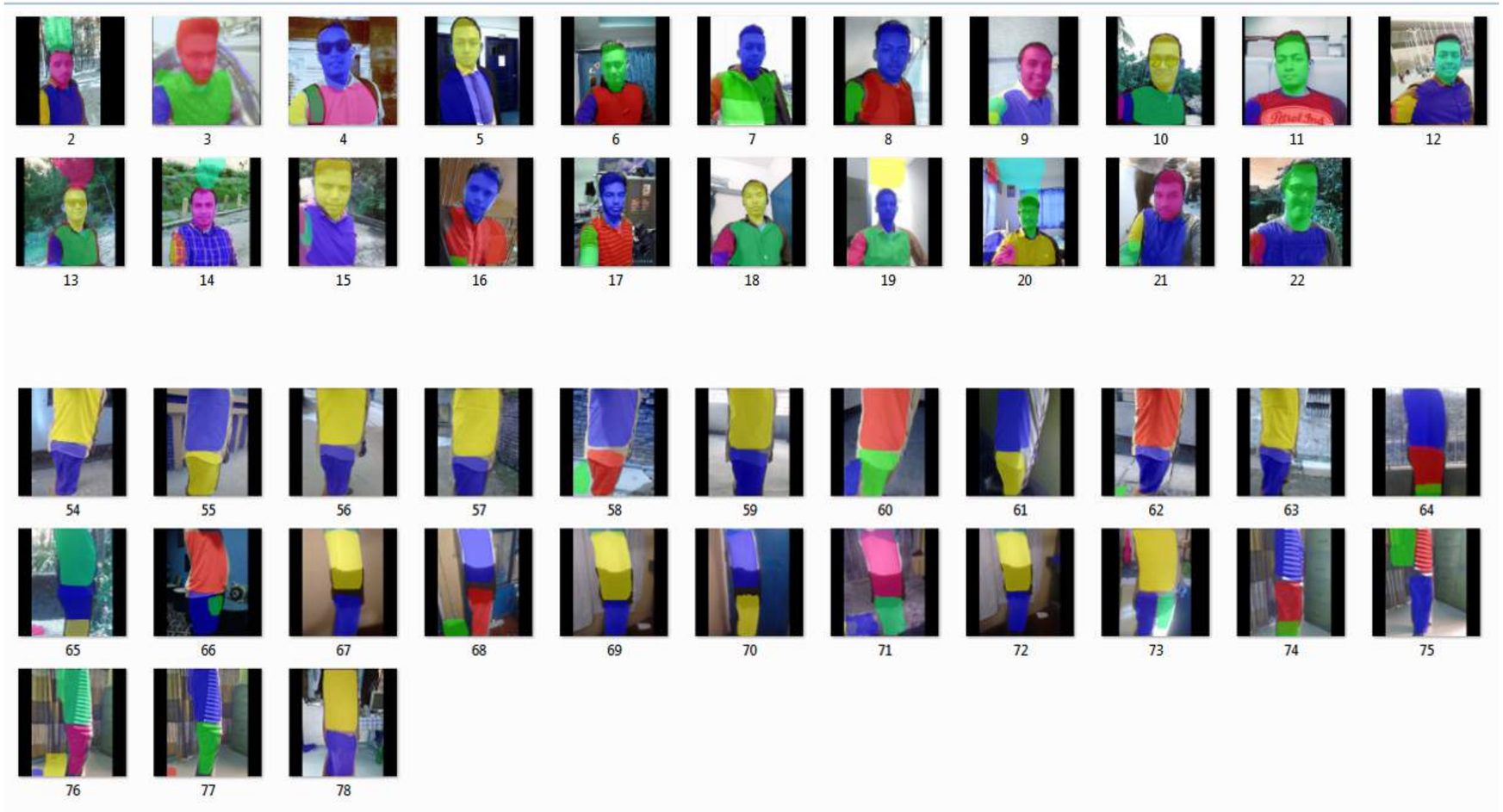
tailor

USE CASE // tailor

segmentation

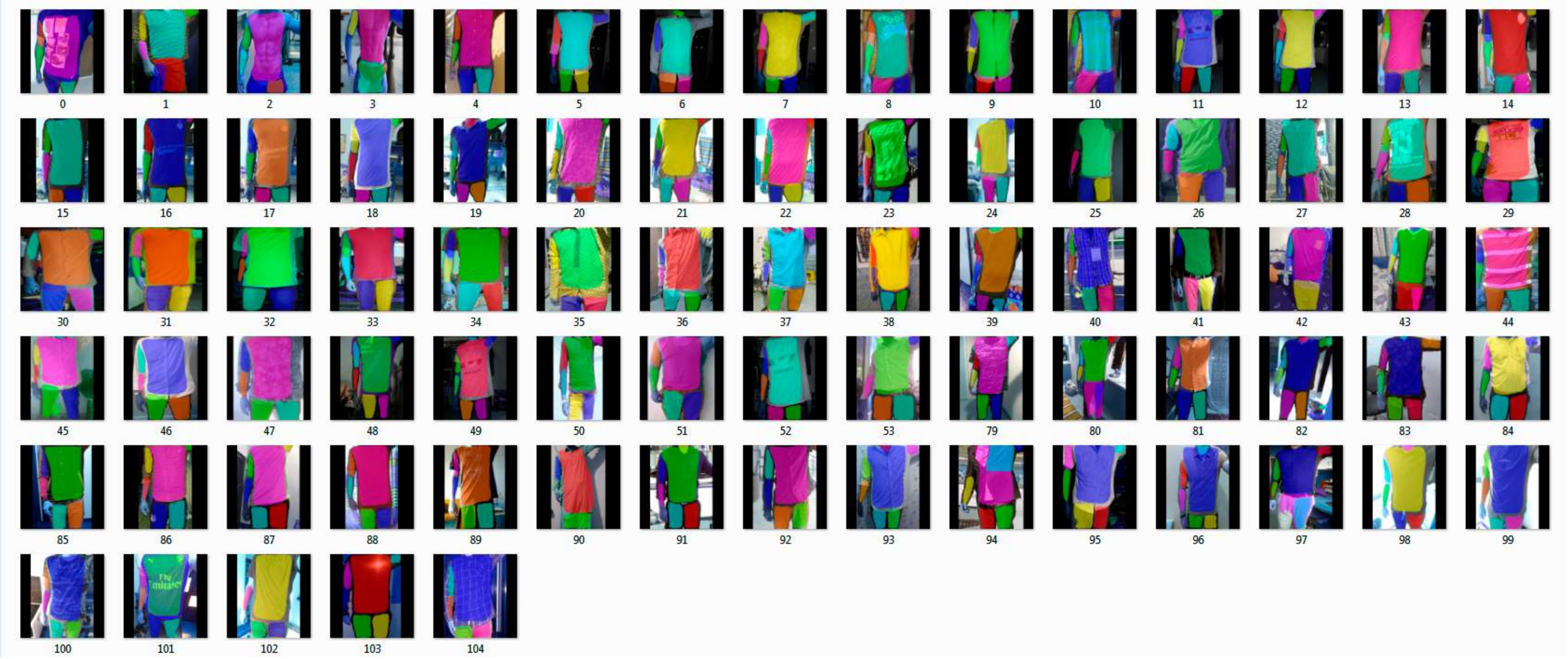
USE CASE // tailor

segmentation



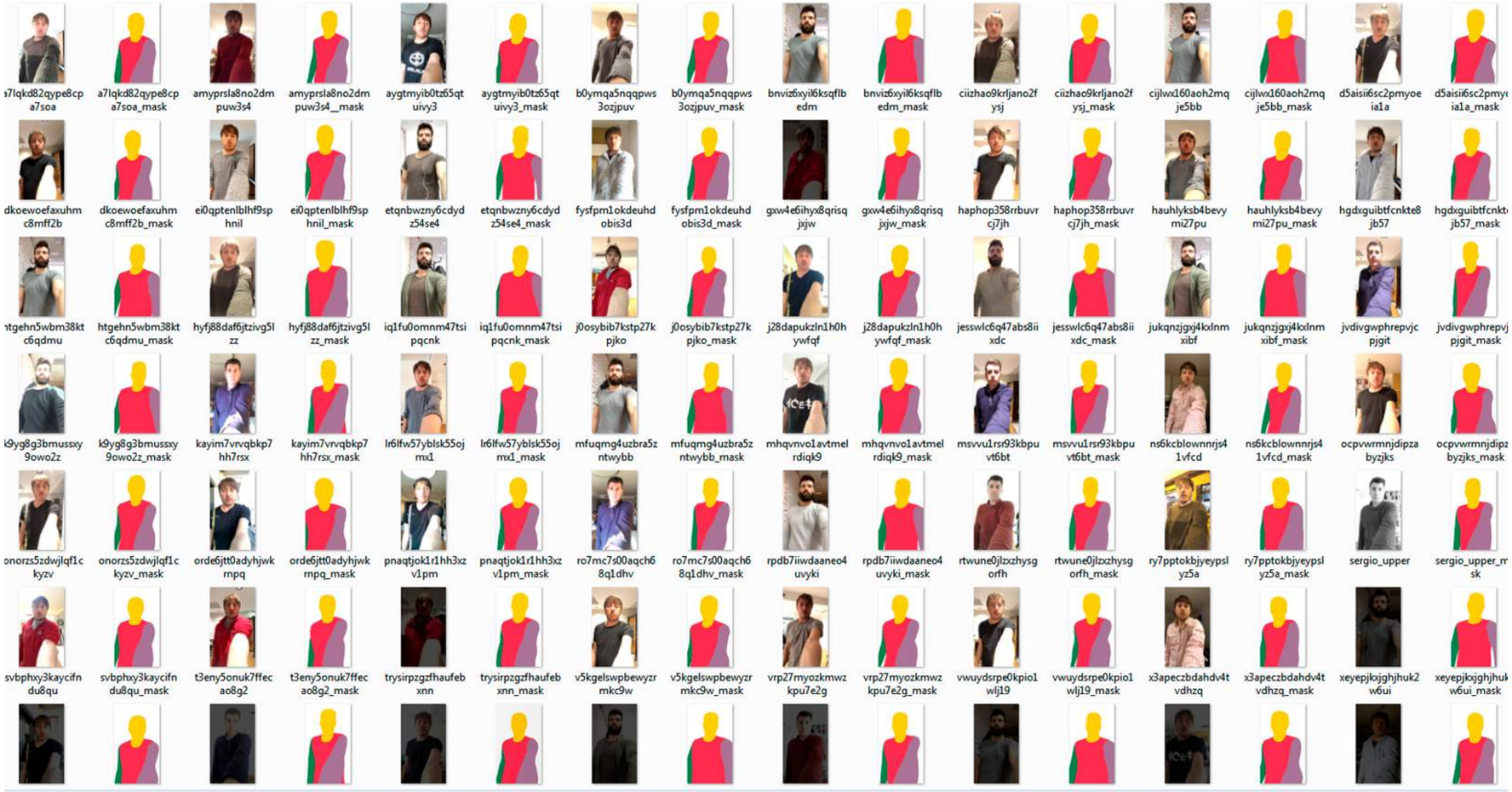
USE CASE // tailor

segmentation



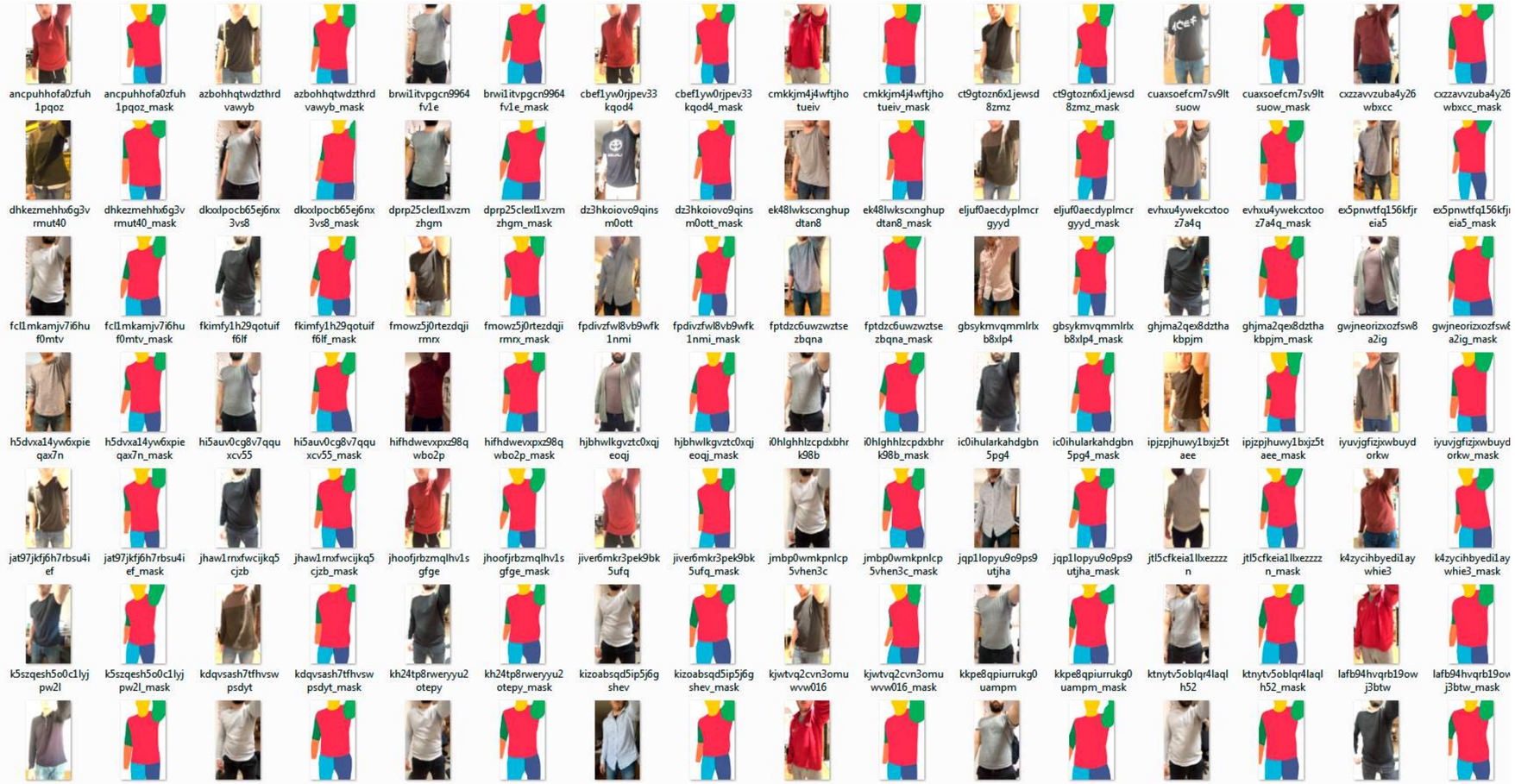
USE CASE // tailor

Labeling segmentation upper



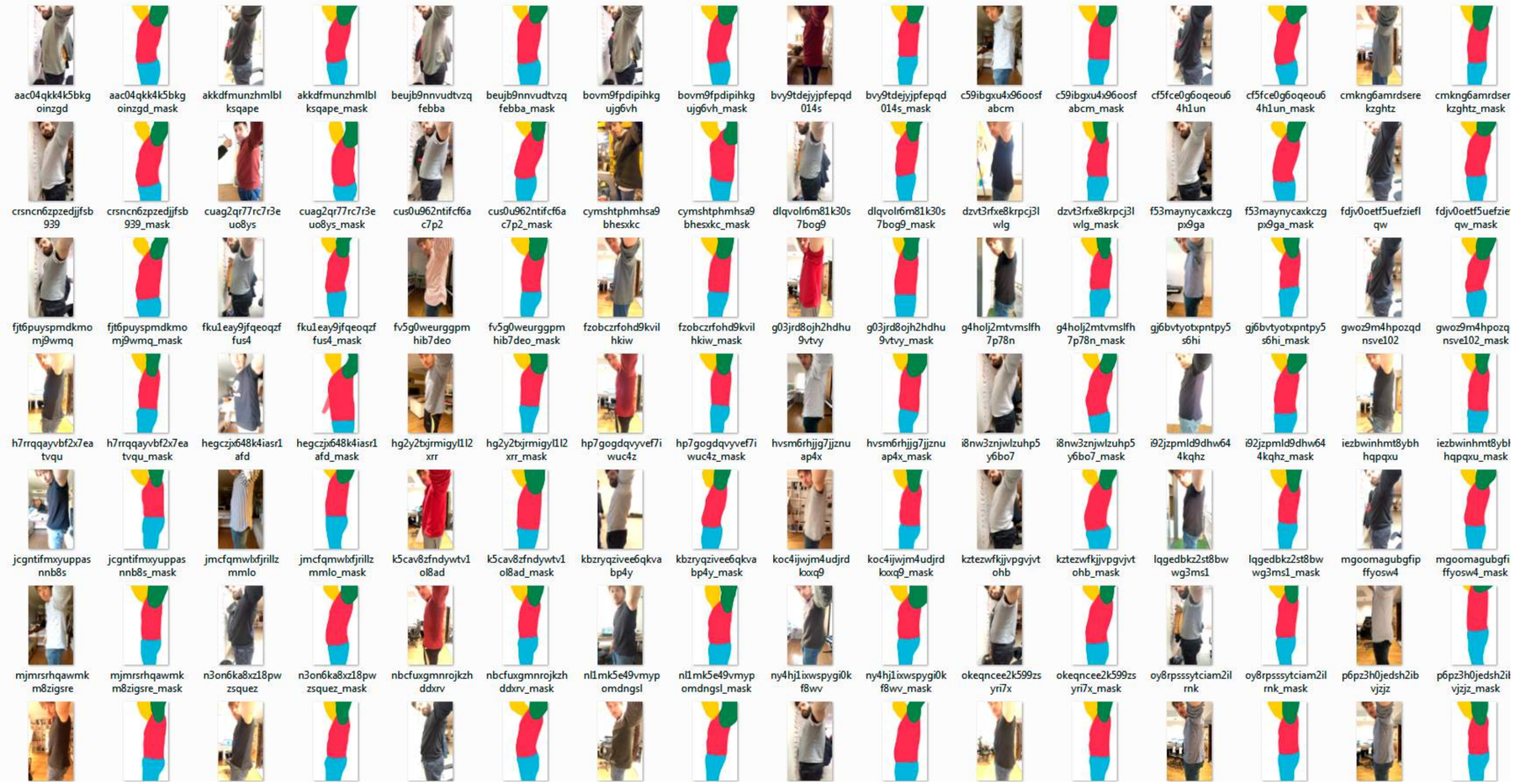
USE CASE // tailor

Labeling segmentation lower



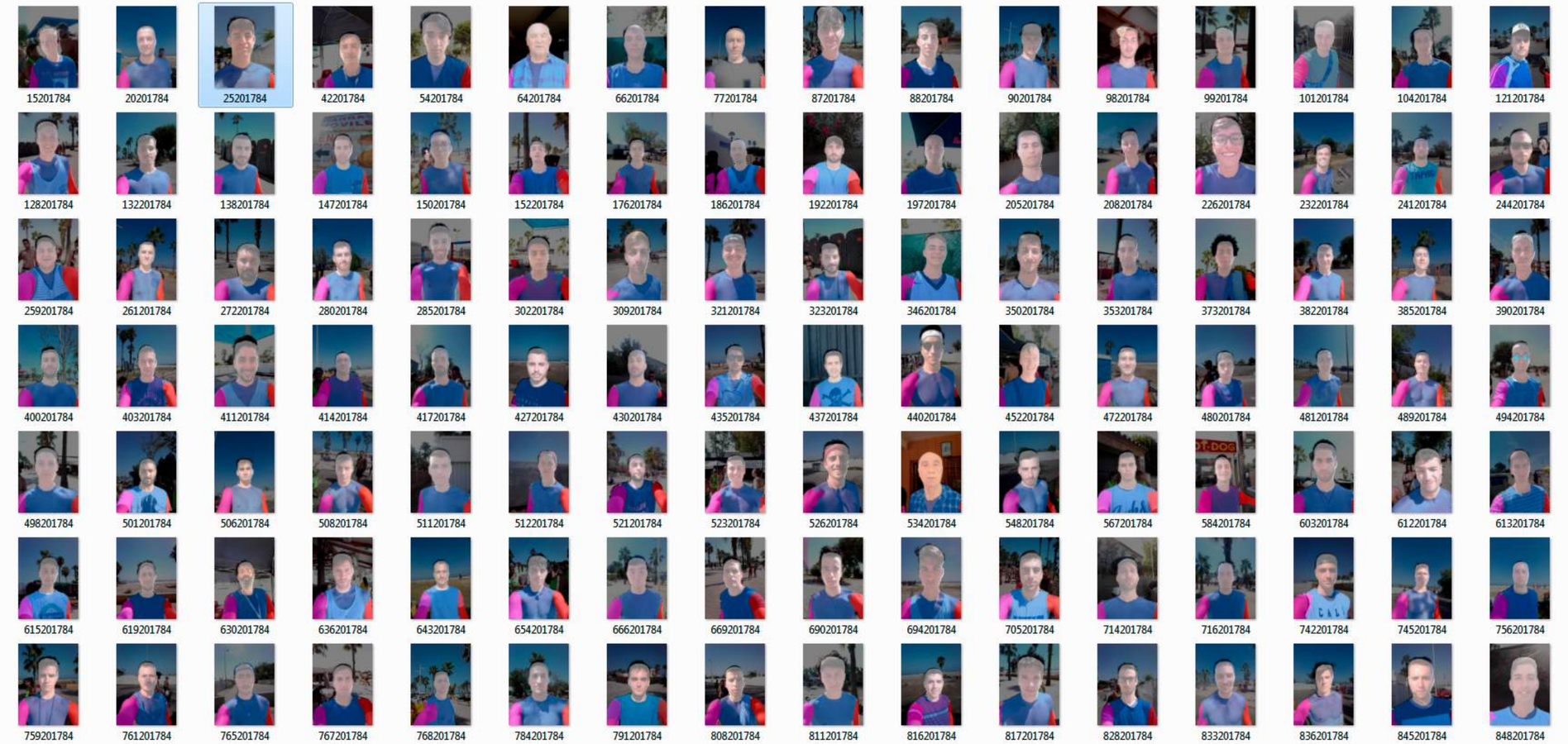
USE CASE // tailor

Labeling segmentation profile



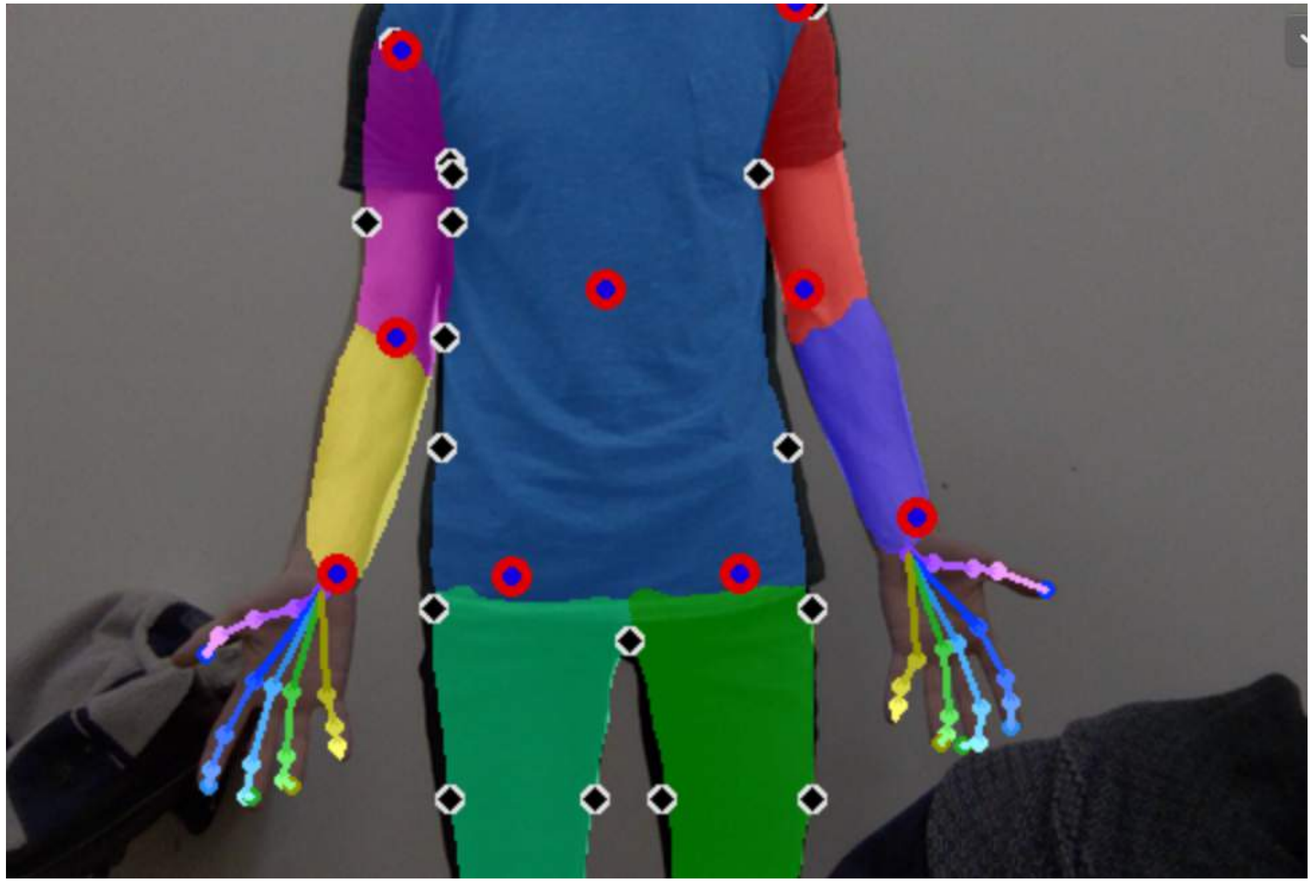
USE CASE // tailor

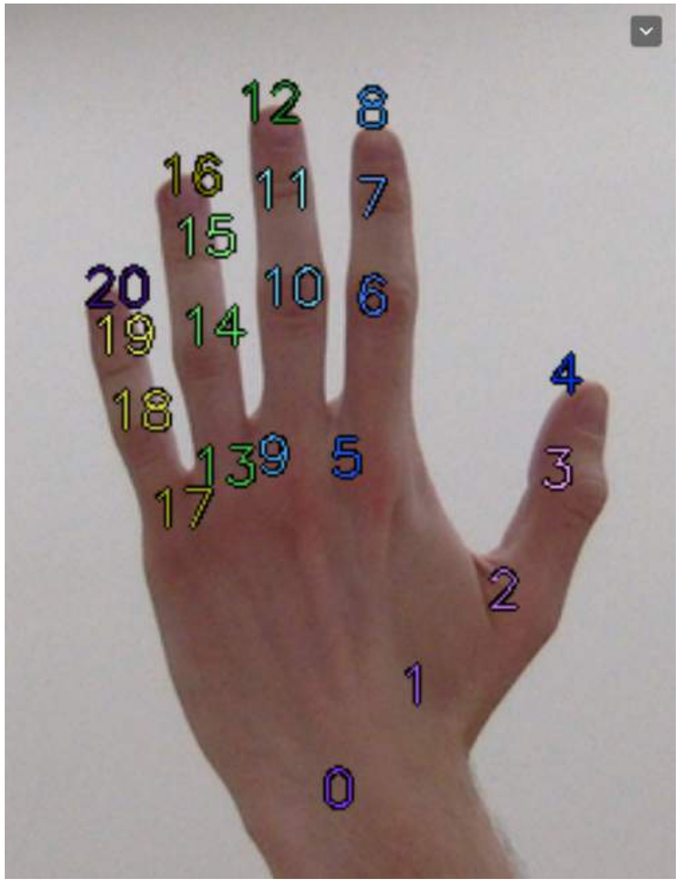
Use Case tailor



USE CASE // tailor

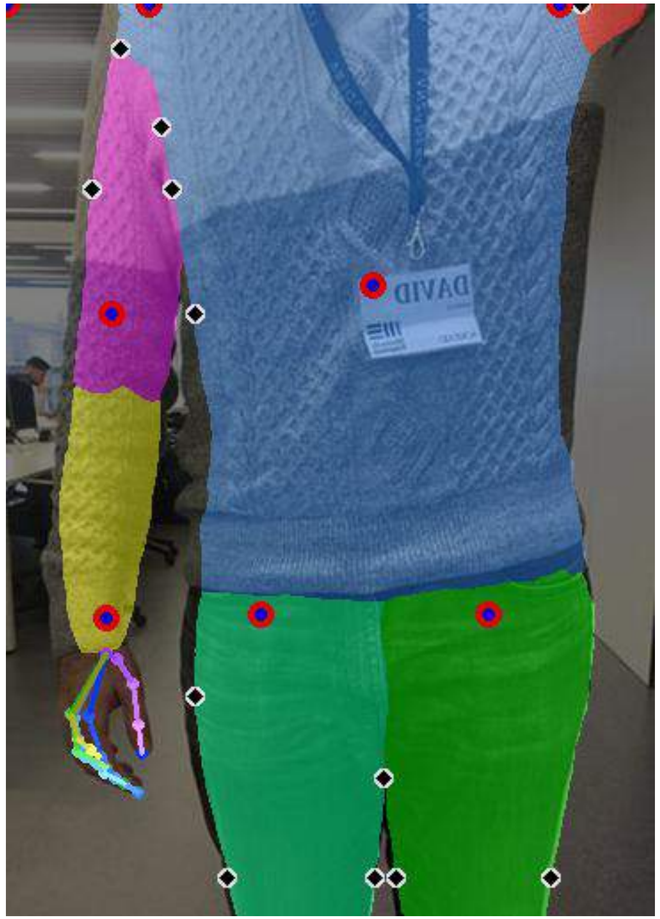
Hands landmarks





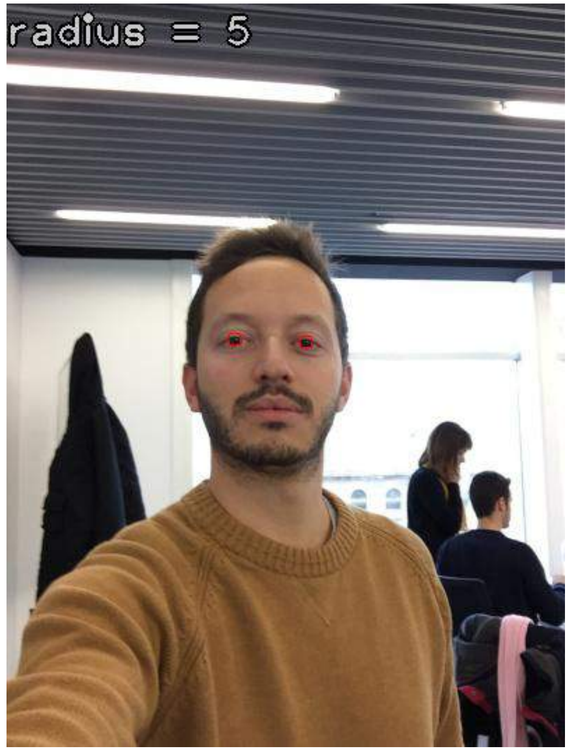
USE CASE // tailor

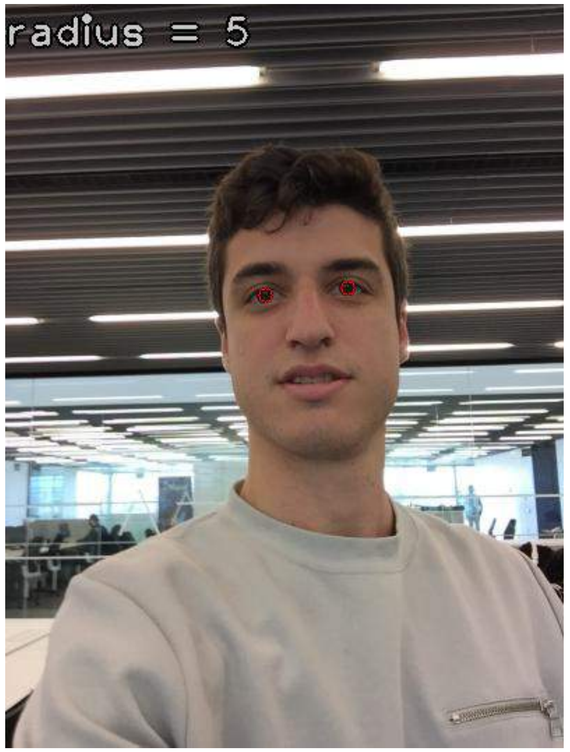
Hands landmarks

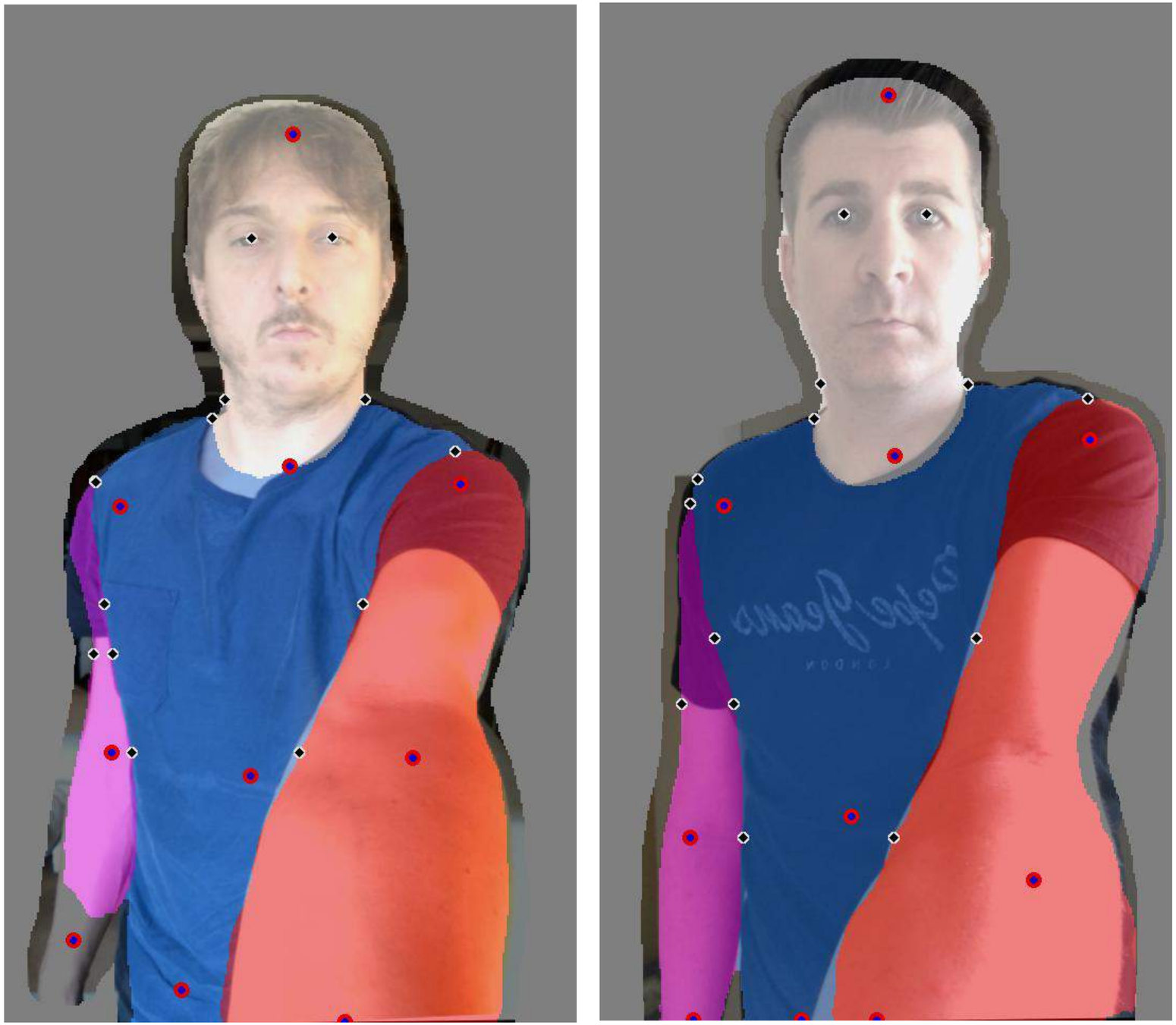


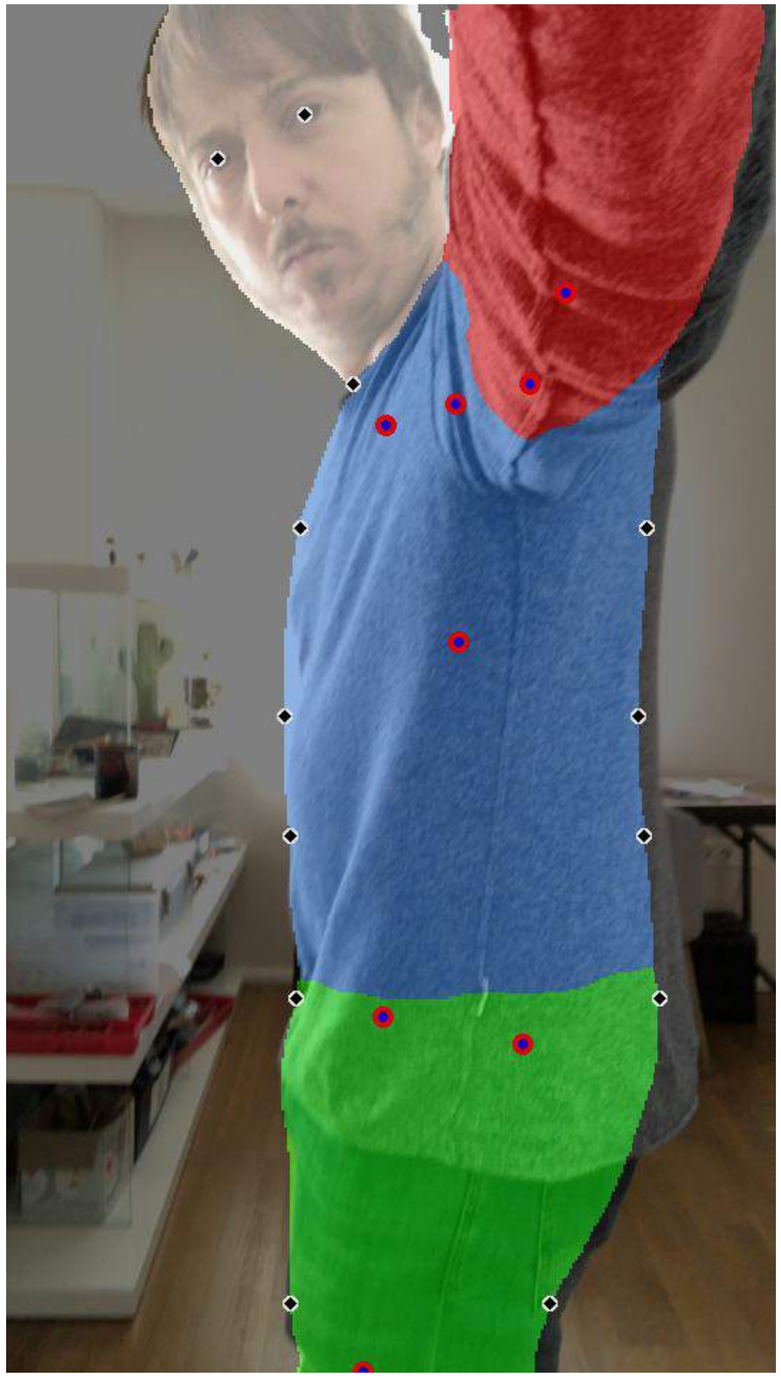
USE CASE // tailor

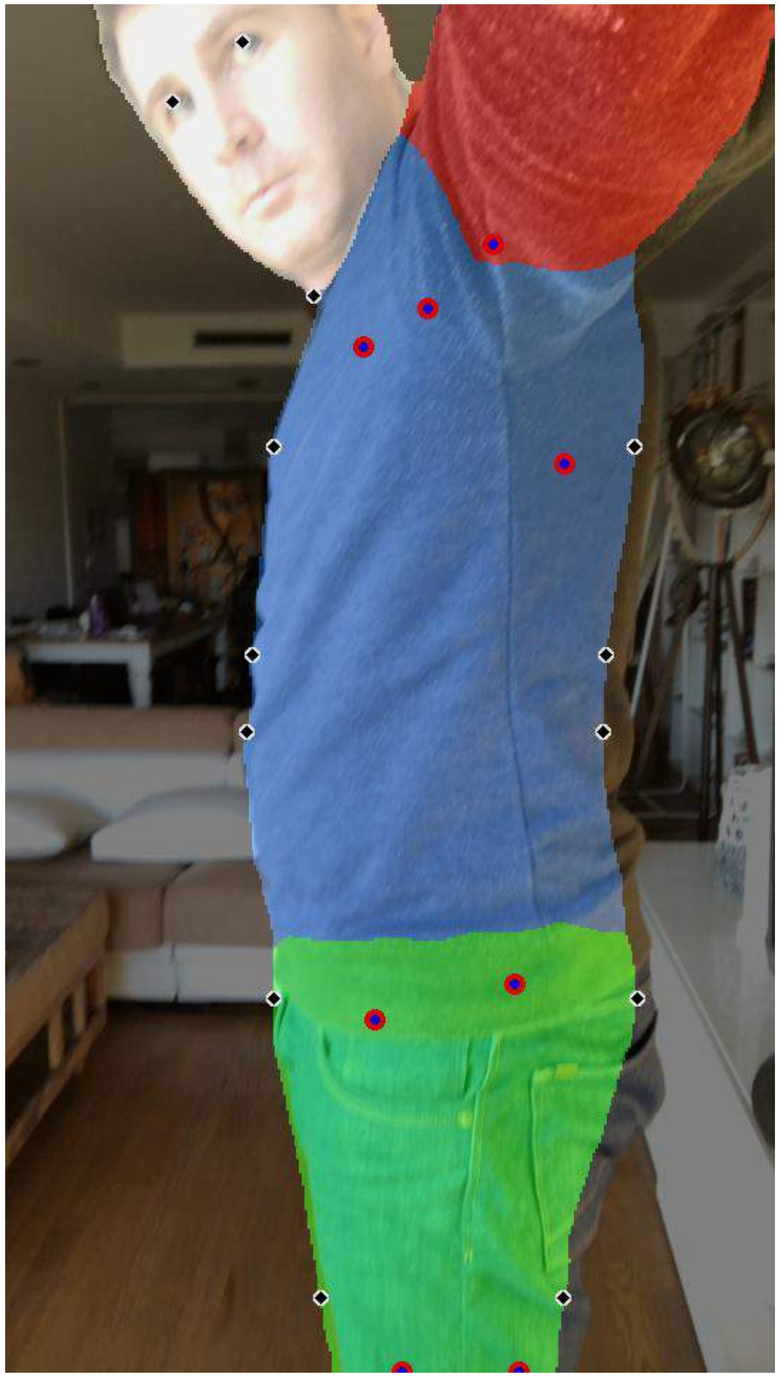
Eyes computer vision











USE CASE // tailor

Linear regression chest

x

BREATH

(23) CHEST BREADTH*

The maximum horizontal breadth of the chest at the level of the chest point anterior landmark is measured with a beam caliper. The participant stands erect, looking straight ahead with the heels together and the weight distributed evenly on both feet. The participant places both hands on the hips and takes a deep breath and holds it. The tissue is compressed with the beam caliper, and then the participant lowers the arms. The measurement is taken at maximum inspiration.

PERCENTILES				
FEMALES			MALES	
CM	IN		CM	IN
23.20	9.13	1ST	25.00	9.84
23.40	9.21	2ND	25.40	10.00
23.80	9.37	3RD	25.60	10.08
24.10	9.49	5TH	26.10	10.28
24.60	9.69	10TH	26.60	10.47
25.00	9.84	15TH	27.00	10.63
25.40	10.00	20TH	27.30	10.75
25.70	10.12	25TH	27.70	10.91
25.90	10.20	30TH	27.90	10.98
26.20	10.31	35TH	28.20	11.10
26.40	10.39	40TH	28.40	11.18
26.60	10.47	45TH	28.70	11.30
26.80	10.55	50TH	28.90	11.38
27.00	10.63	55TH	29.10	11.46
27.30	10.75	60TH	29.40	11.57
27.50	10.83	65TH	29.60	11.65
27.80	10.94	70TH	29.90	11.77
28.10	11.06	75TH	30.10	11.85
28.50	11.22	80TH	30.40	11.97
28.90	11.38	85TH	30.80	12.13
29.40	11.57	90TH	31.30	12.32
30.20	11.89	95TH	32.00	12.60
30.60	12.05	97TH	32.60	12.83
30.90	12.17	98TH	32.90	12.95
31.40	12.36	99TH	33.50	13.19

* In ANSUR this measurement was taken (in males) at the level of thelion (nipple). This change was

z

DEPTH

(25) CHEST DEPTH*

The horizontal distance between the chest point anterior landmark and the back at the same level is measured with a beam caliper. The participant stands erect, looking straight ahead. The shoulders and upper extremities are relaxed. The measurement is taken at the maximum point of quiet respiration.

PERCENTILES				
FEMALES			MALES	
CM	IN		CM	IN
18.90	7.44	1ST	19.70	7.76
19.70	7.76	2ND	20.20	7.95
19.90	7.83	3RD	20.60	8.11
20.50	8.07	5TH	21.10	8.31
21.30	8.39	10TH	22.00	8.66
21.90	8.62	15TH	22.70	8.94
22.50	8.86	20TH	23.10	9.09
22.80	8.98	25TH	23.60	9.29
23.20	9.13	30TH	23.90	9.41
23.60	9.29	35TH	24.30	9.57
23.90	9.41	40TH	24.60	9.69
24.20	9.53	45TH	25.00	9.84
24.50	9.65	50TH	25.30	9.96
24.90	9.80	55TH	25.60	10.08
25.30	9.96	60TH	26.00	10.24
25.60	10.08	65TH	26.40	10.39
26.00	10.24	70TH	26.70	10.51
26.60	10.47	75TH	27.20	10.71
27.00	10.63	80TH	27.70	10.91
27.60	10.87	85TH	28.10	11.06
28.40	11.18	90TH	28.80	11.34
29.60	11.65	95TH	29.80	11.73
30.10	11.85	97TH	30.40	11.97
30.70	12.09	98TH	30.90	12.17
32.20	12.68	99TH	31.60	12.44

* In ANSUR this measurement was taken (in males) at the level of thelion (nipple). This change was

$\bar{y}$

CIRCUMFERENCE

(24) CHEST CIRCUMFERENCE*

The maximum horizontal circumference of the chest at the level of chest point, anterior is measured with a tape. The participant stands erect, looking straight ahead. The shoulders and upper extremities are relaxed. The measurement is taken at the maximum point of quiet respiration.

PERCENTILES				
FEMALES			MALES	
CM	IN		CM	IN
78.40	30.87	1ST	87.20	34.33
79.50	31.30	2ND	89.40	35.20
80.70	31.77	3RD	90.70	35.71
82.40	32.44	5TH	92.20	36.30
84.50	33.27	10TH	94.80	37.32
86.30	33.98	15TH	96.70	38.07
87.70	34.53	20TH	98.30	38.70
88.90	35.00	25TH	99.60	39.21
90.00	35.43	30TH	100.90	39.72
90.90	35.79	35TH	102.10	40.20
92.00	36.22	40TH	103.20	40.63
92.90	36.57	45TH	104.30	41.06
94.00	37.01	50TH	105.60	41.57
95.00	37.40	55TH	106.70	42.01
96.10	37.83	60TH	107.80	42.44
97.50	38.39	65TH	109.00	42.91
98.70	38.86	70TH	110.40	43.46
99.90	39.33	75TH	111.70	43.98
101.50	39.96	80TH	113.30	44.61
103.30	40.67	85TH	115.10	45.31
105.70	41.61	90TH	117.20	46.14
109.40	43.07	95TH	120.70	47.52
111.80	44.02	97TH	123.30	48.54
114.00	44.88	98TH	125.10	49.25
116.10	45.71	99TH	127.70	50.28

USE CASE // tailor

Linear regression waist

x

BREATH

z

DEPTH

$\bar{y}$

CIRCUMFERENCE

(86) WAIST BREADTH

The horizontal breadth of the waist at the level of omphalion is measured with a beam caliper. The participant stands erect, looking straight ahead. The heels are together with the weight distributed equally on both feet. The measurement is taken at the maximum point of quiet respiration.

PERCENTILES				
FEMALES			MALES	
CM	IN		CM	IN
23.20	9.13	1ST	25.70	10.12
23.90	9.41	2ND	26.20	10.31
24.30	9.57	3RD	26.60	10.47
25.00	9.84	5TH	27.20	10.71
26.00	10.24	10TH	28.20	11.10
26.60	10.47	15TH	28.90	11.38
27.20	10.71	20TH	29.50	11.61
27.60	10.87	25TH	30.10	11.85
28.10	11.06	30TH	30.70	12.09
28.50	11.22	35TH	31.20	12.28
28.90	11.38	40TH	31.60	12.44
29.40	11.57	45TH	32.10	12.64
29.80	11.73	50TH	32.50	12.80
30.20	11.89	55TH	33.00	12.99
30.60	12.05	60TH	33.50	13.19
31.10	12.24	65TH	34.00	13.39
31.50	12.40	70TH	34.50	13.58
32.10	12.64	75TH	35.00	13.78
32.70	12.87	80TH	35.60	14.02
33.40	13.15	85TH	36.20	14.25
34.40	13.54	90TH	37.10	14.61
35.60	14.02	95TH	38.50	15.16
36.70	14.45	97TH	39.50	15.55
37.50	14.76	98TH	40.20	15.83
38.50	15.16	99TH	41.30	16.26

(88) WAIST DEPTH

The horizontal distance between the anterior and posterior omphalion landmarks is measured with a beam caliper. The participant stands erect, looking straight ahead. The heels are together with the weight distributed equally on both feet. The measurement is taken at the maximum point of quiet respiration.

PERCENTILES				
FEMALES			MALES	
CM	IN		CM	IN
15.90	6.26	1ST	17.40	6.85
16.30	6.42	2ND	17.90	7.05
16.50	6.50	3RD	18.30	7.20
16.90	6.65	5TH	18.80	7.40
17.60	6.93	10TH	19.60	7.72
18.20	7.17	15TH	20.20	7.95
18.60	7.32	20TH	20.70	8.15
19.00	7.48	25TH	21.20	8.35
19.40	7.64	30TH	21.60	8.50
19.80	7.80	35TH	22.10	8.70
20.10	7.91	40TH	22.50	8.86
20.50	8.07	45TH	22.90	9.02
20.90	8.23	50TH	23.40	9.21
21.30	8.39	55TH	23.90	9.41
21.70	8.54	60TH	24.30	9.57
22.20	8.74	65TH	24.80	9.76
22.60	8.90	70TH	25.50	10.04
23.20	9.13	75TH	26.10	10.28
23.80	9.37	80TH	26.70	10.51
24.50	9.65	85TH	27.40	10.79
25.60	10.08	90TH	28.40	11.18
27.10	10.67	95TH	30.00	11.81
28.20	11.10	97TH	30.90	12.17
28.90	11.38	98TH	31.80	12.52
30.00	11.81	99TH	32.90	12.95

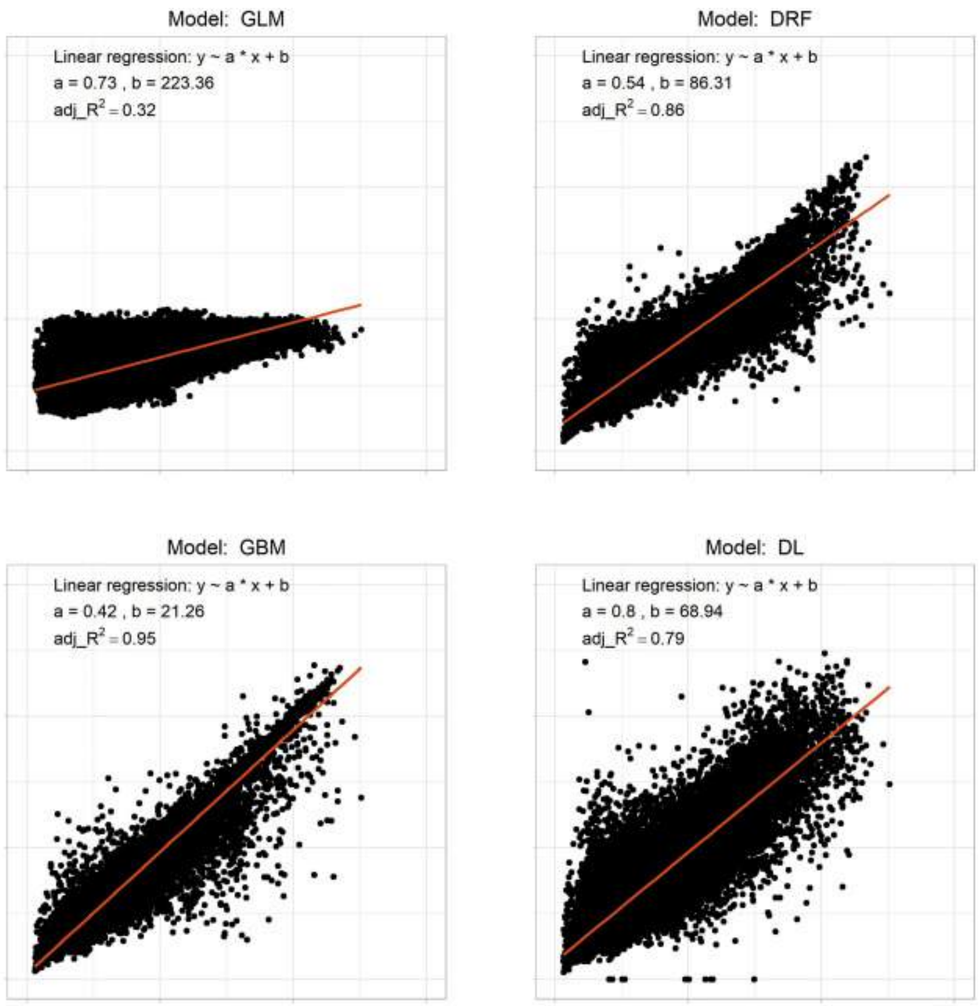
(87) WAIST CIRCUMFERENCE (OMPHALION)

The horizontal circumference of the waist, passing over all omphalion landmarks, is measured with a tape. The participant stands erect, looking straight ahead. The heels are together with the weight distributed equally on both feet. The measurement is taken at the maximum point of quiet respiration.

PERCENTILES				
FEMALES			MALES	
CM	IN		CM	IN
66.50	26.18	1ST	72.40	28.50
68.00	26.77	2ND	74.00	29.13
69.60	27.40	3RD	75.20	29.61
71.00	27.95	5TH	76.80	30.24
73.90	29.09	10TH	79.70	31.38
76.00	29.92	15TH	82.00	32.28
77.70	30.59	20TH	83.80	32.99
79.00	31.10	25TH	85.70	33.74
80.30	31.61	30TH	87.50	34.45
81.40	32.05	35TH	89.10	35.08
82.70	32.56	40TH	90.70	35.71
84.00	33.07	45TH	92.00	36.22
85.20	33.54	50TH	93.70	36.89
86.50	34.06	55TH	95.20	37.48
87.80	34.57	60TH	96.70	38.07
89.50	35.24	65TH	98.30	38.70
90.70	35.71	70TH	100.00	39.37
92.50	36.42	75TH	101.60	40.00
94.30	37.13	80TH	103.50	40.75
96.40	37.95	85TH	106.00	41.73
99.50	39.17	90TH	108.60	42.76
104.00	40.94	95TH	113.10	44.53
106.90	42.09	97TH	115.70	45.55
109.00	42.91	98TH	117.90	46.42
112.10	44.13	99TH	121.40	47.80

USE CASE // tailor

Linear regression 2D – 3D



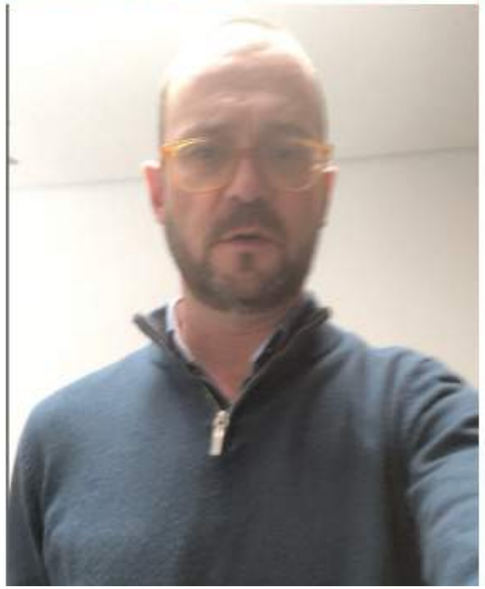
USE CASE // tailor

Use Case tailor

JSON Datos sin procesar Cabeceras

Guardar Copiar Contraer todo Expandir todo


```
code:                    "200"
items:
  cuello:
    cuello_upper:        171.00138887817232
    cuellocm:            11.46583181701228
    cuellopx:            136.80111110253787
  hombro_codo:
    antebrazouppcm:      21.328997113044558
    antebrazoupppx:      254.48049041134766
    hombre_codo_upper:   318.10061301418455
  image:
    url:                 "https://res.cloudinary.com/tailored/image/upload/v1544440647/dataset_ios/upper_body/au0uoo7h5i0t1pllxpfs.jpg"
  ojos:
    ojos_px:             14.317437752703926
  pecho:
    pecho_upper:         403.6087214122113
    pechouppcm:          27.06241013567865
    pechoupppx:          322.8869771297691
  pernera_a:
    a_cm:                23.745559230153397
    a_px:                283.31297181738785
    pernera_upper_a:     354.1412147717348
  pernera_b:
    b_cm:                8.918048828075925
    b_px:                106.4030074762927
    pernera_upper_b:     133.00375934536586
  praxiteles:
    praxiuppcm:          7.327685918596744
    praxiupppx:          174
message:                 ""
```

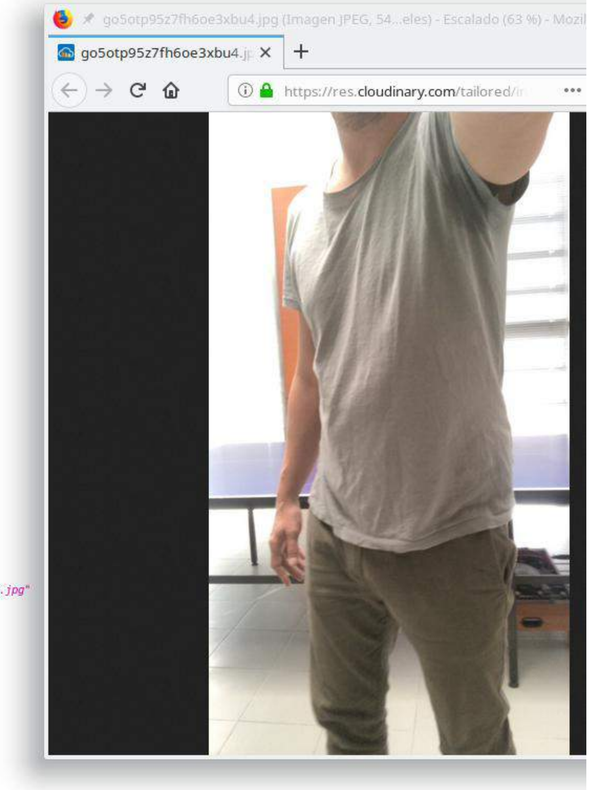
USE CASE // tailor

Use Case tailor

JSON Datos sin procesar Cabeceras

Guardar Copiar Contraer todo Expandir todo


```
code:                                   "200"
items:
  cadera:
    cadelocm:                           22.14176979090676
    cadelopx:                           262.63404196714487
    cadera_lower:                       328.2925524589311
  cadera_esqueleto_cadera:
    cadcadlocm:                         8.093428726931597
    cadcadlopx:                         96
    cadcadlower:                        120
  cadera_esqueleto_cintura:
    cadcinlocm:                         14.70306218725907
    cadcinlopx:                         174.4
    cadcinlower:                        218
  cintura:
    cintulo:                            571.5055093846375
    cintulocm:                          38.545325893777516
    cintulopx:                          457.2044075077101
  cuello_cadera:
    cuellocadlowcm:                     42.89517225273747
    cuellocadlower:                     636
    cuellocadlowpx:                     508.8
  hombro_cadera:
    homcadlocm:                         31.50855065027587
    homcadlopx:                         373.73787605753853
    homcadlower:                        467.17234507192313
  hombro_codo:
    antebrazolocm:                      16.726419368991966
    antebrazolopx:                      198.4
    antebrazolower:                     248
  image:
    url:                                "https://res.cloudinary.com/tailored/image/upload/v1544440435/dataset_ios/lower_body/go5otp95z7fh6oe3xbu4.jpg"
  largo_brazo:
    brazolocm:                          29.06287286235023
    brazolopx:                          344.7285308761084
    brazolower:                         430.9106635951355
  ojos:
    ojos_px:                            14.233769628026975
  pecho:
    pecholocm:                          16.846195339814727
    pecholopx:                          199.82071964638703
    pecholower:                         249.77589955798376
  pernera:
    pernelocm:                          -3.308224962865956
    pernelopx:                          -39.2404266659351
    pernelower:                         -49.050533332418865
message:                                ""
```

USE CASE // tailor

Use Case tailor

JSON Datos sin procesar Cabeceras

Guardar Copiar Contraer todo Expandir todo


```
code:                          "200"
items:
  cadera:
    cadeprocm:                 16.608789768819722
    cadeprofile:               247.70345173210646
    cadepropx:                 198.16276138568517
  cadera_equeleto_cadera:
    cadcadprocm:               41.50463304006854
    cadcadprofile:             619
    cadcadpropx:               495.20000000000005
  cadera_equeleto_cintura:
    cadcinprocm:               25.1441637964874
    cadcinprofile:             375
    cadcinpropx:               300
  cintura:
    cintuprocm:                17.920716697982638
    cintuprofile:              267.2695268825086
    cintupropx:                213.8156215060069
  cuello_cadera:
    cuellocadprocm:            41.50463304006854
    cuellocadprofile:          619
    cuellocadpropx:            495.20000000000005
  hombro_cadera:
    homcadprocm:               0
    homcadprofile:             0
    homcadpropx:               0
  image:
    url:                       "https://res.cloudinary.com/tailored/image/upload/v1544440646/dataset_ios/profile_body/glddoxiddympn4n952xc.jpg"
  ojos:
    ojos_px:                   14.317437752703926
  pecho:
    pechoprocm:                18.039862318539956
    pechoprofile:              269.04646438858845
    pechopropx:                215.23717151087078
message:                       ""
```

tailor

TRANSITIONS:

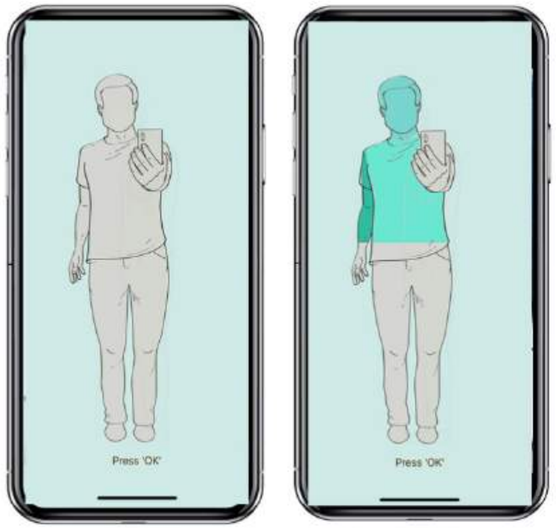
STEP1

OK

STEP2

OK

STEP3



OK

USE CASE // tailor

Preshooting real time

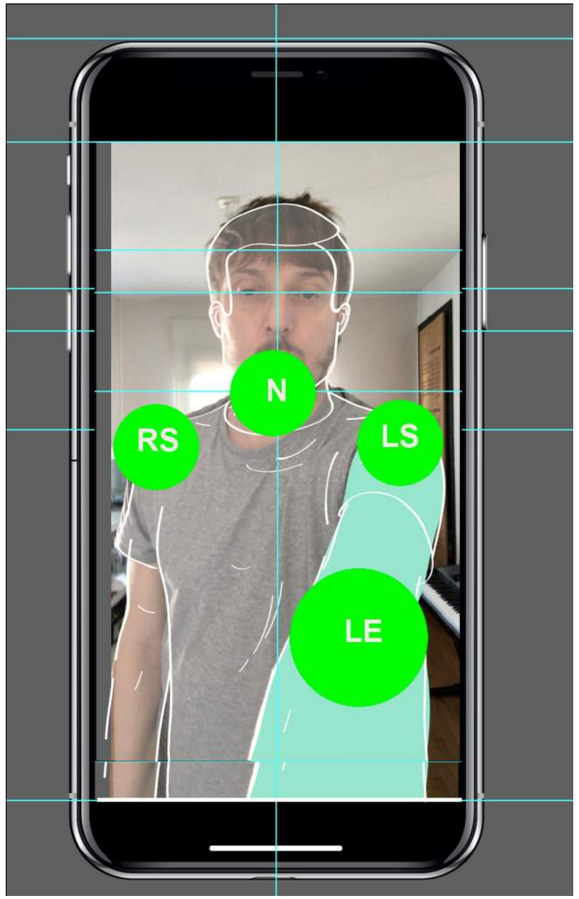
tailor

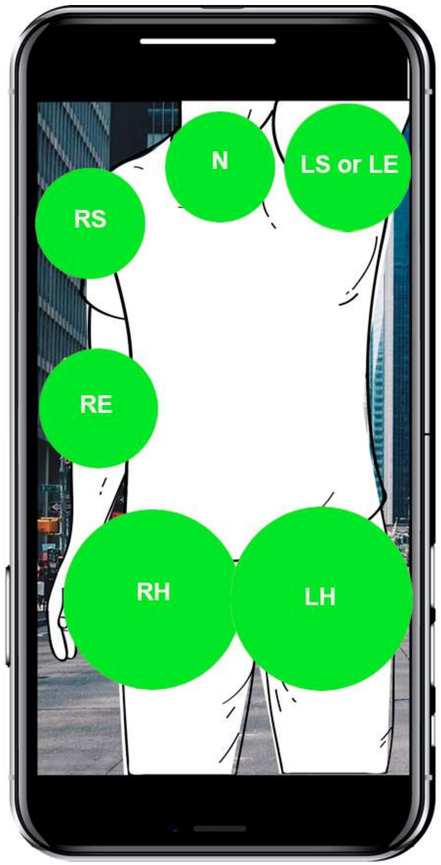
USE CASE // tailor

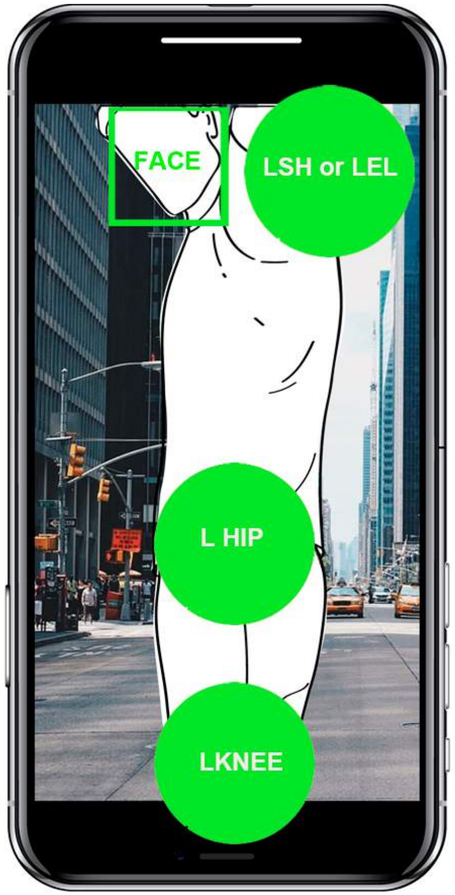
Preshooting real time

USE CASE // tailor

Preshooting real time

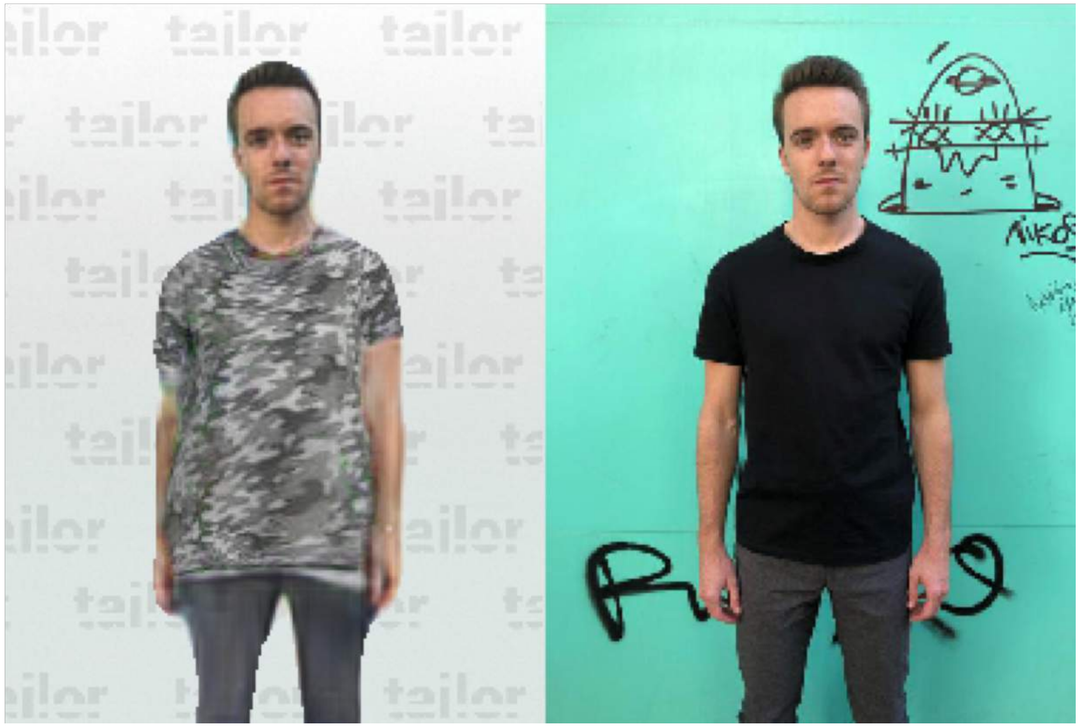






USE CASE // tailor

GAN´s : fitting room

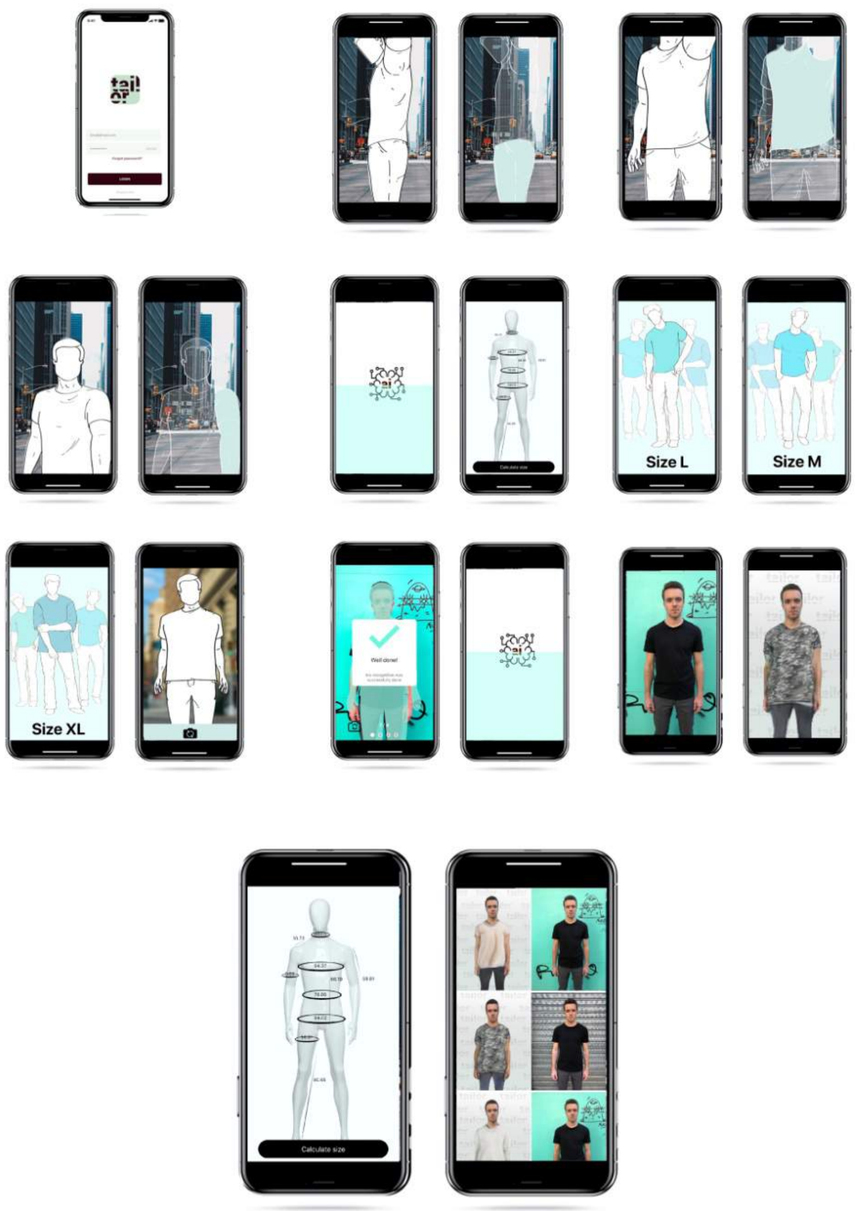


USE CASE // tailor

GAN´s : fitting room

USE CASE // tailor

Use Case tailor



tailor

1 // tailor components:

tailor // components

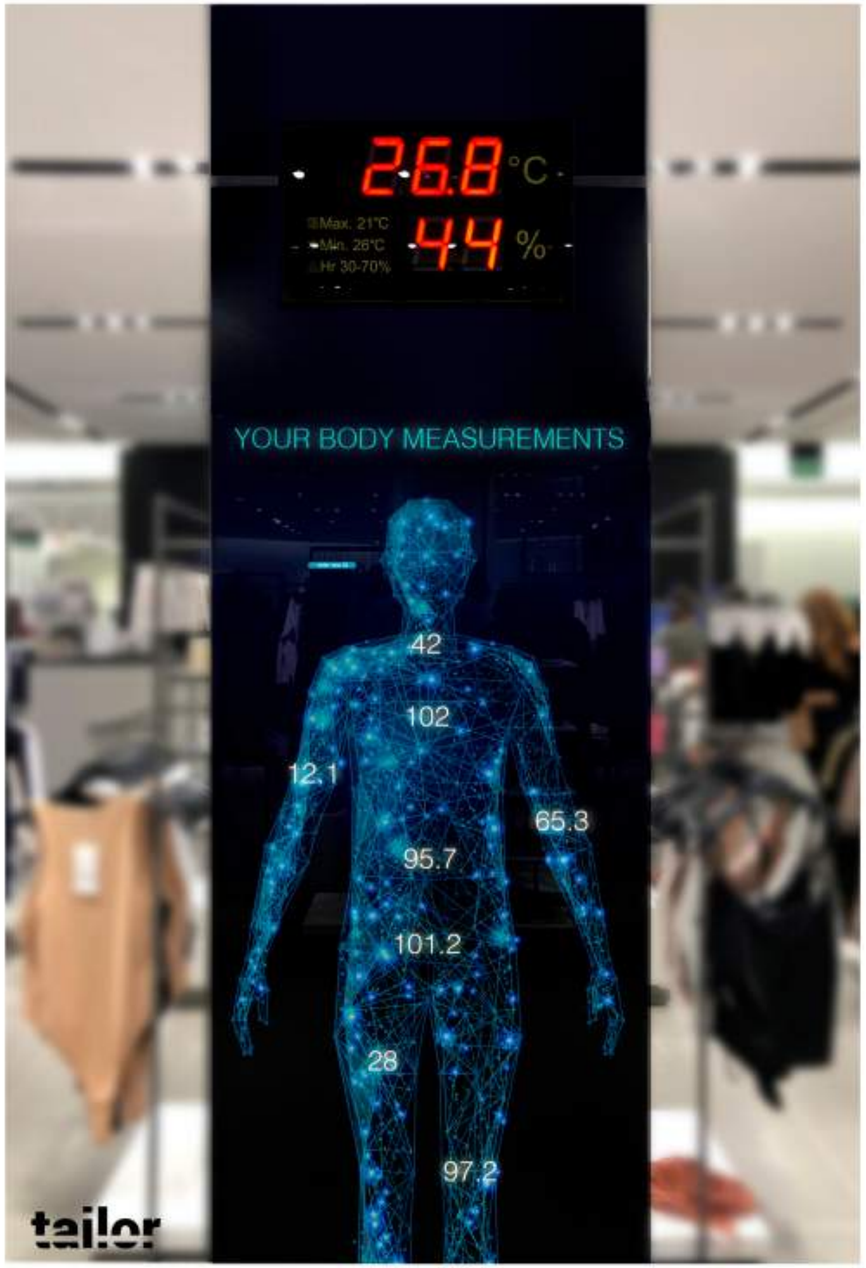
	description	quantity
POST-PROCESSING / / HD	Body segmentation upper body	1
	Body segmentation lower body	1
	Body segmentation profile	1
	Upper body pose	1
	Lower body pose	1
	Profile pose	1
	Background removal segmentation	1
	Iris detector	1
	Hand landmarks	1
PRE-SHOOTING // REAL TIME	Upper body pose real time lite	1
	Lower body pose real time lite	1
	Profile pose real time lite	1
	SVM	1
	Linear regressions	20
	API	1
	clustering	1
	assembly + landmarks parameterization	1
POC // FITTING ROOM	Try clothes on user	1
APP / / IOS	UI backend	1
	UI frontend	1
DATA // GATHERED AND LABELLED BY US	Body segmentation upper body images + labels	
	Body segmentation lower body images + labels	
	Body segmentation profile images + labels	
	Upper body pose	
	Lower body pose	
	Profile pose	
	upper body dataset measured by tailors	
	Body measurements	

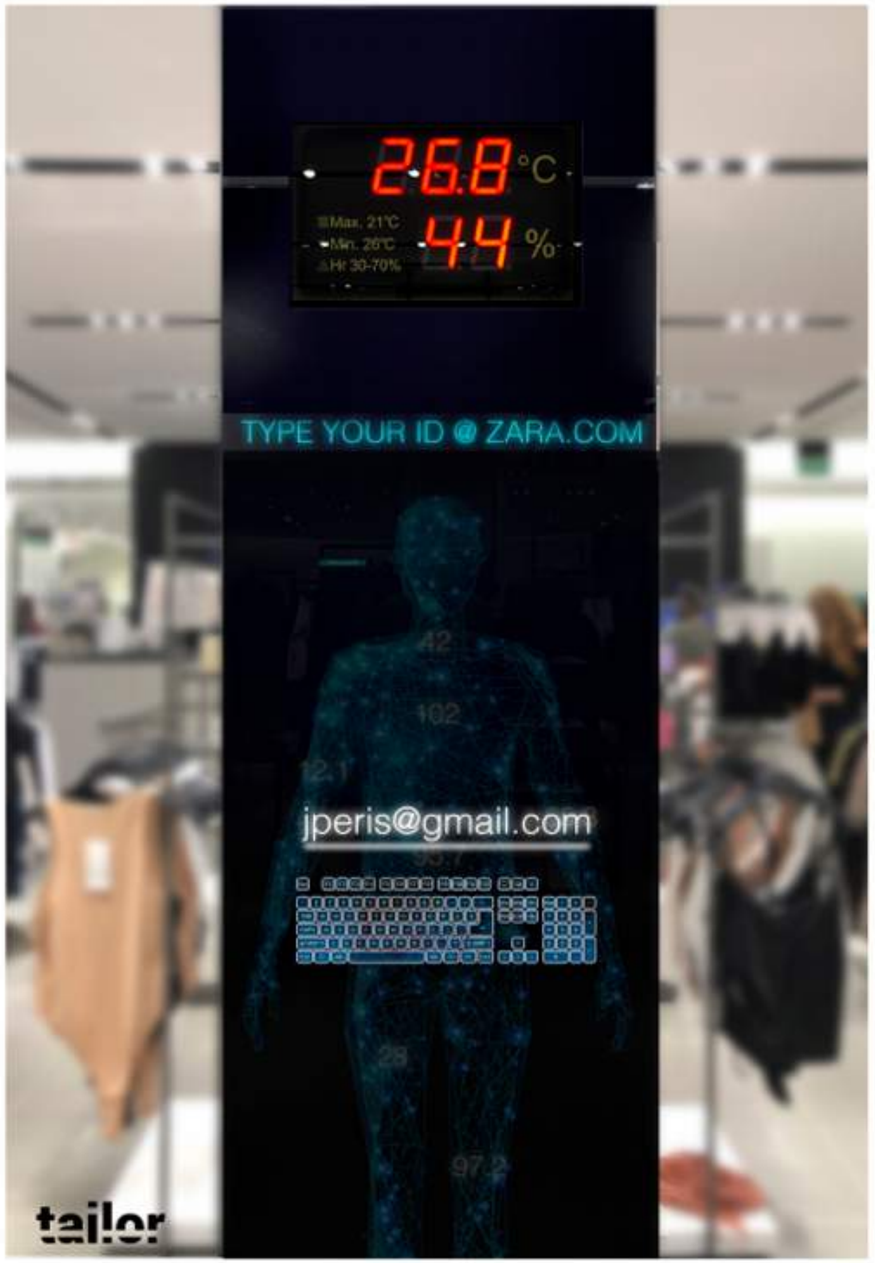
26.7°C
44%
SMART SHOP @ ZARA.COM
tailor

26.7°C
44%
SMART SHOP @ ZARA.COM
tailor

26.7
44
SMART SHOP @ ZARA.COM
tailor

26.7°C
44%
SMART SHOP @ ZARA.COM
tailor

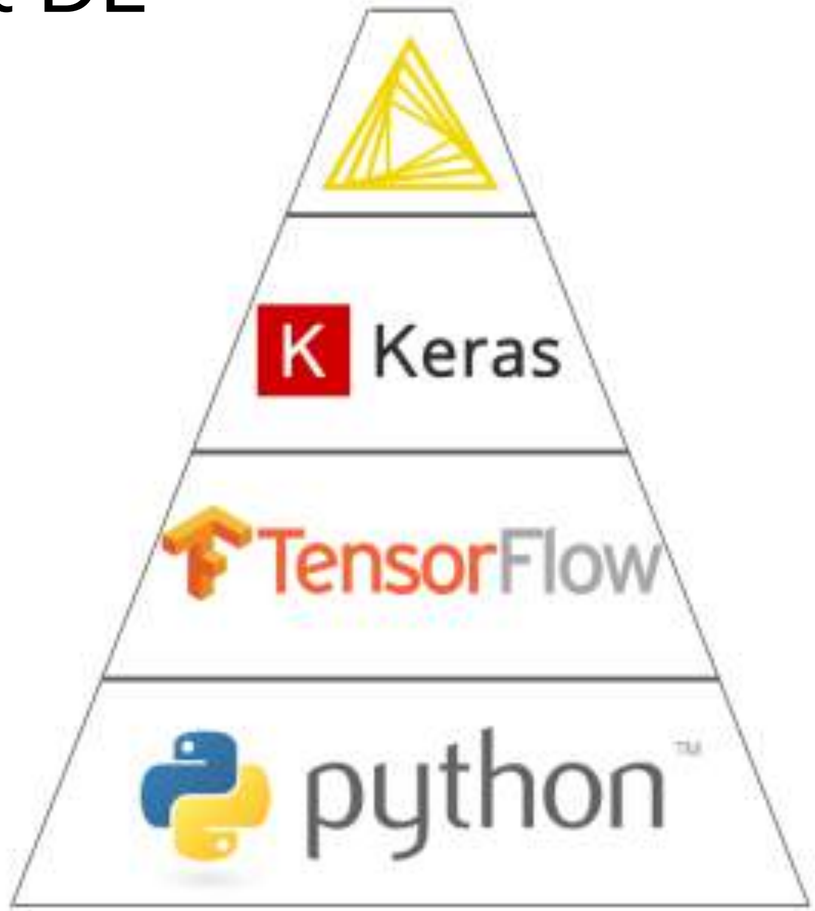

26.8 °C
44 %
YOUR BODY MEASUREMENTS
42
102
12.1
65.3
95.7
101.2
28
97.2
tailor


26.8 °C
44 %
TYPE YOUR ID @ ZARA.COM
jperis@gmail.com
tailor

Learn by Doing

TECHNOLOGIES

High & low level ML & DL



HIGH LEVEL : -No coding skills required
-ML & DL skills required
-Fast prototyping
-Better for understanding business goal

nodes o snippets:

rapidminer

Open for Innovation KNIME FREE

Microsoft Azure

IBM Watson

DataRobot

windows user interface:

WEKA The University of Waikato FREE

bigml

framework: Keras FREE fast.ai FREE

backend: TensorFlow FREE PyTorch FREE

languages python FREE R FREE

LOW LEVEL : -coding skills required
-ML & DL skills required
-Better for production

TECHNOLOGIES

AI project

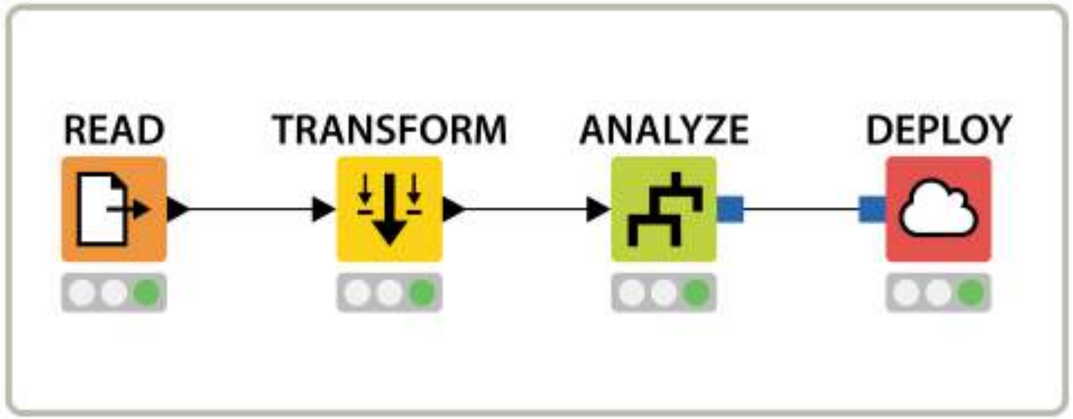
INTRO TO KNIME

Installation & register

Open for Innovation
KNIME

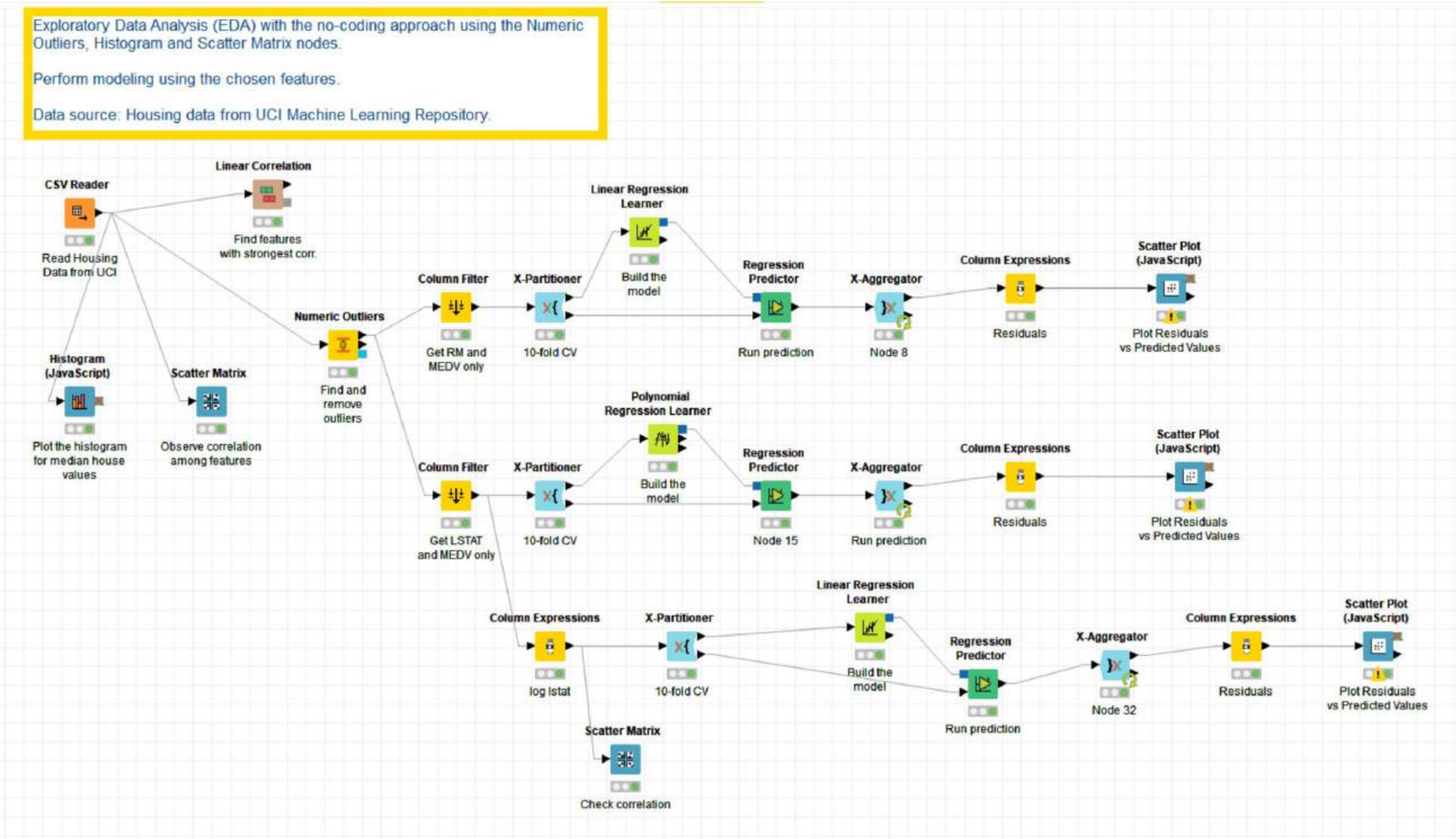
https://www.knime.com/

node concept



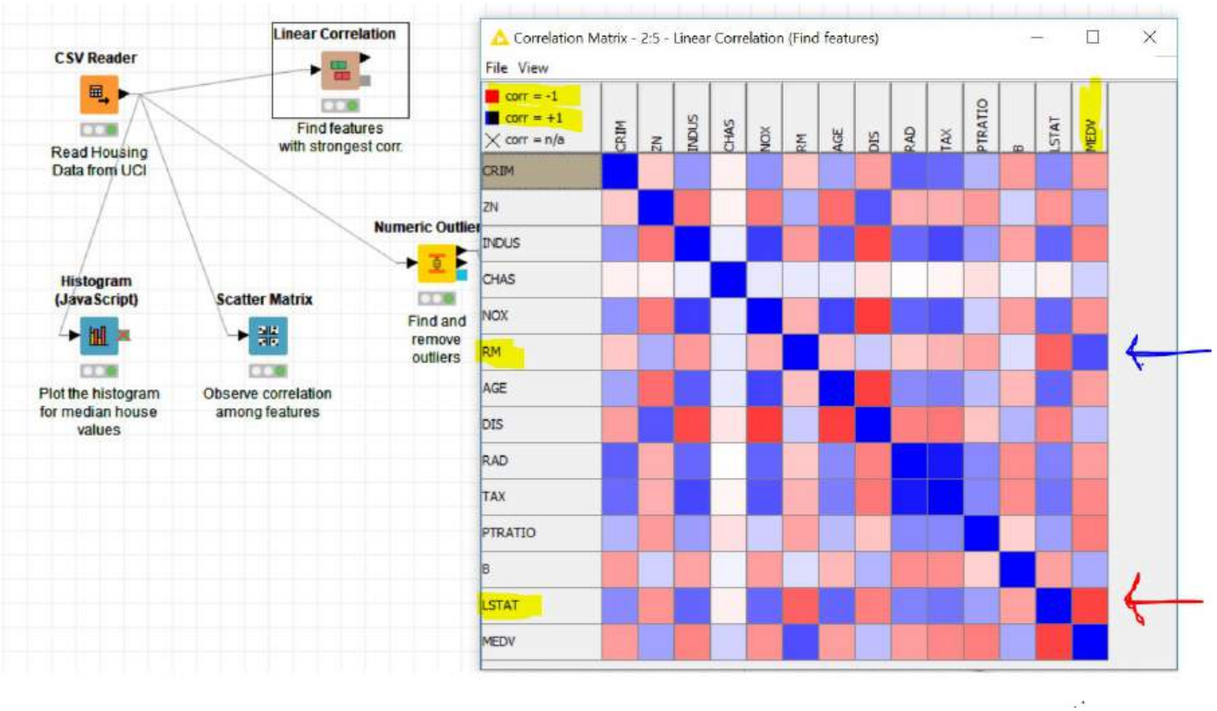
INTRO TO KNIME

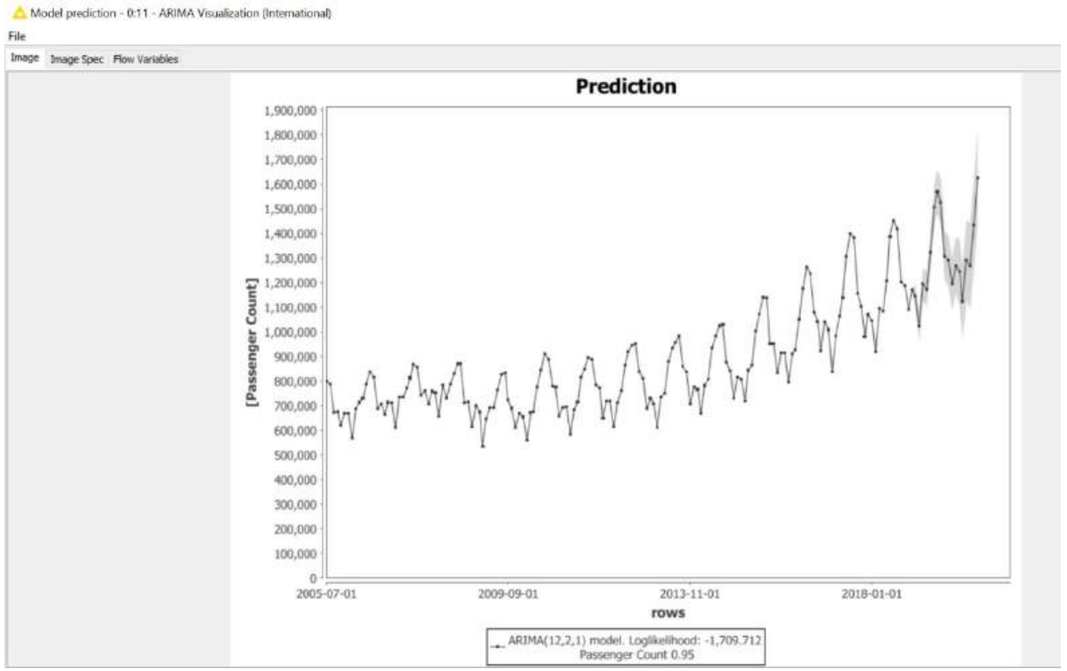
workflow



INTRO TO KNIME

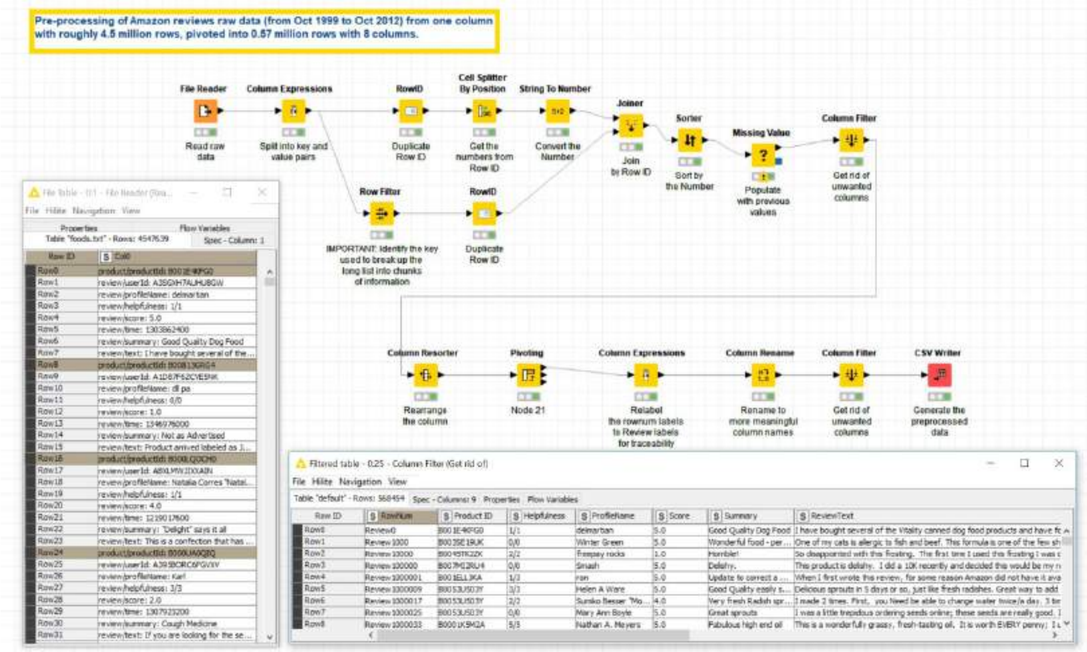
plot

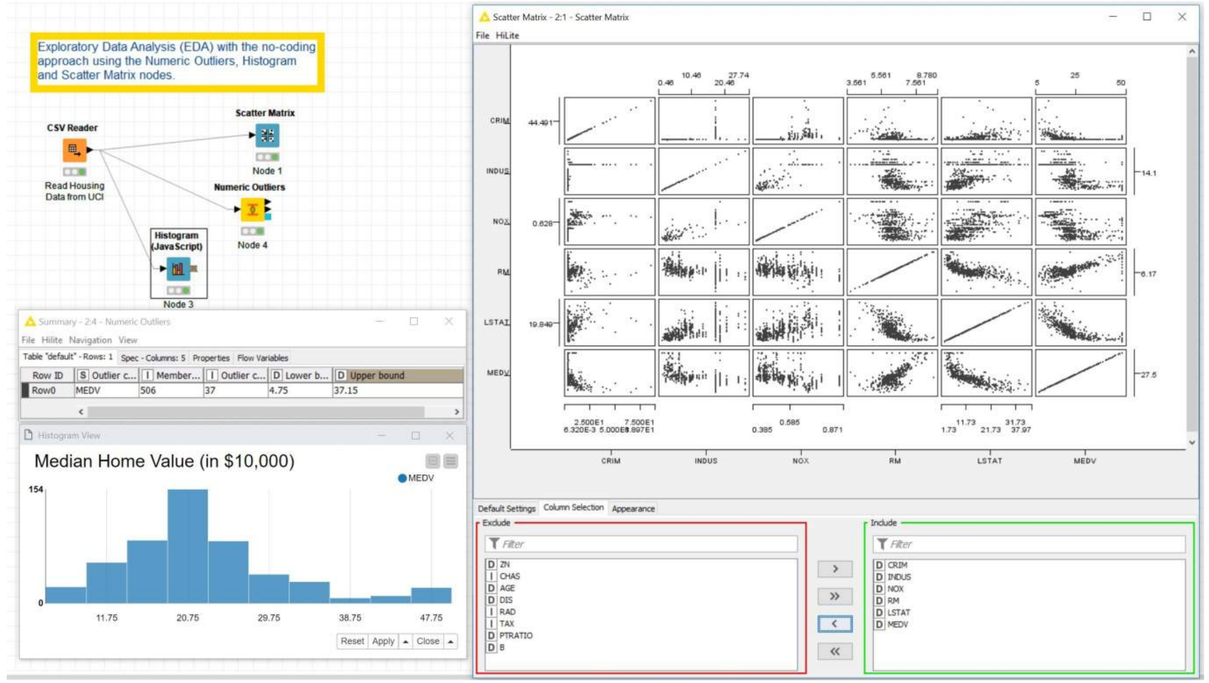


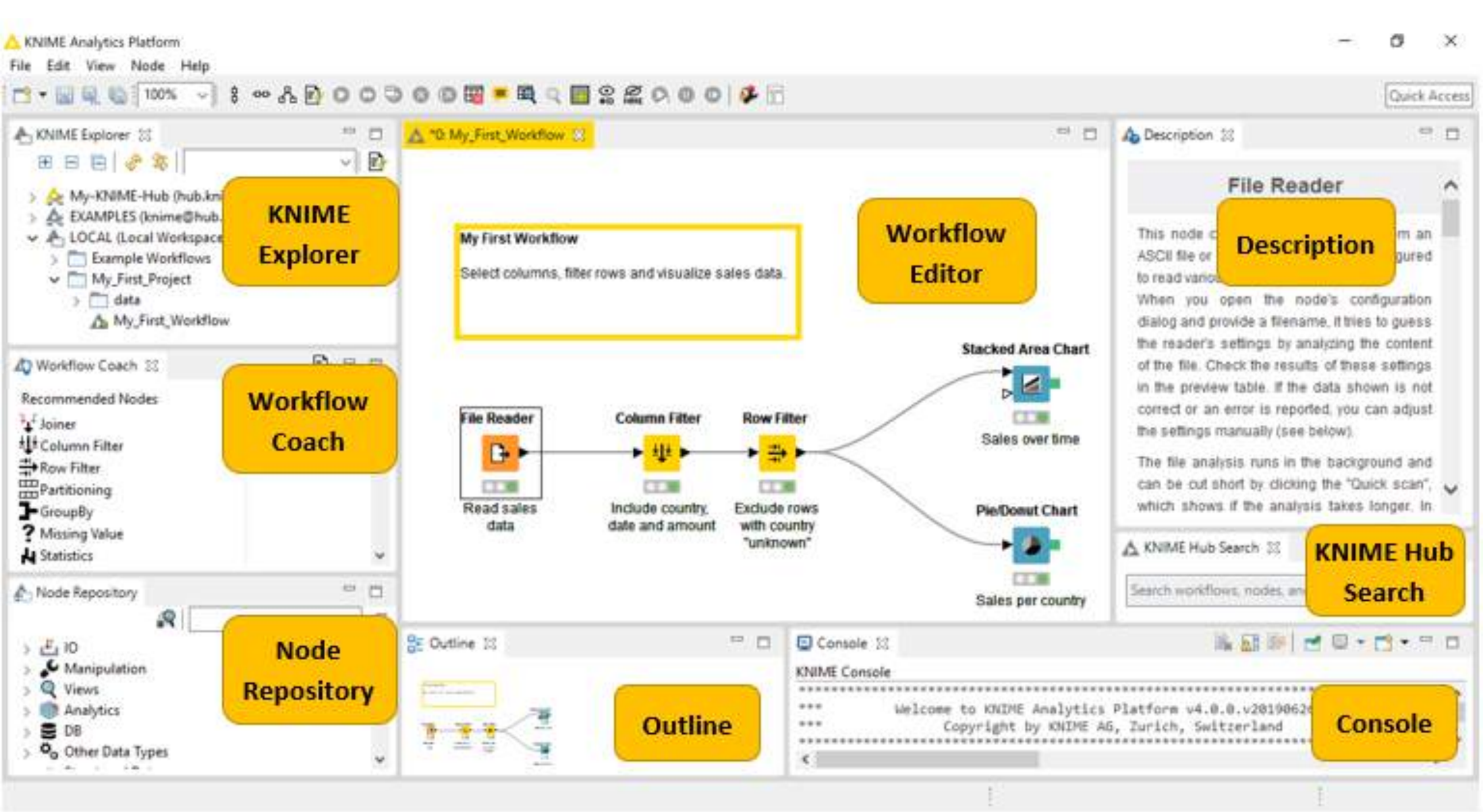


INTRO TO KNIME

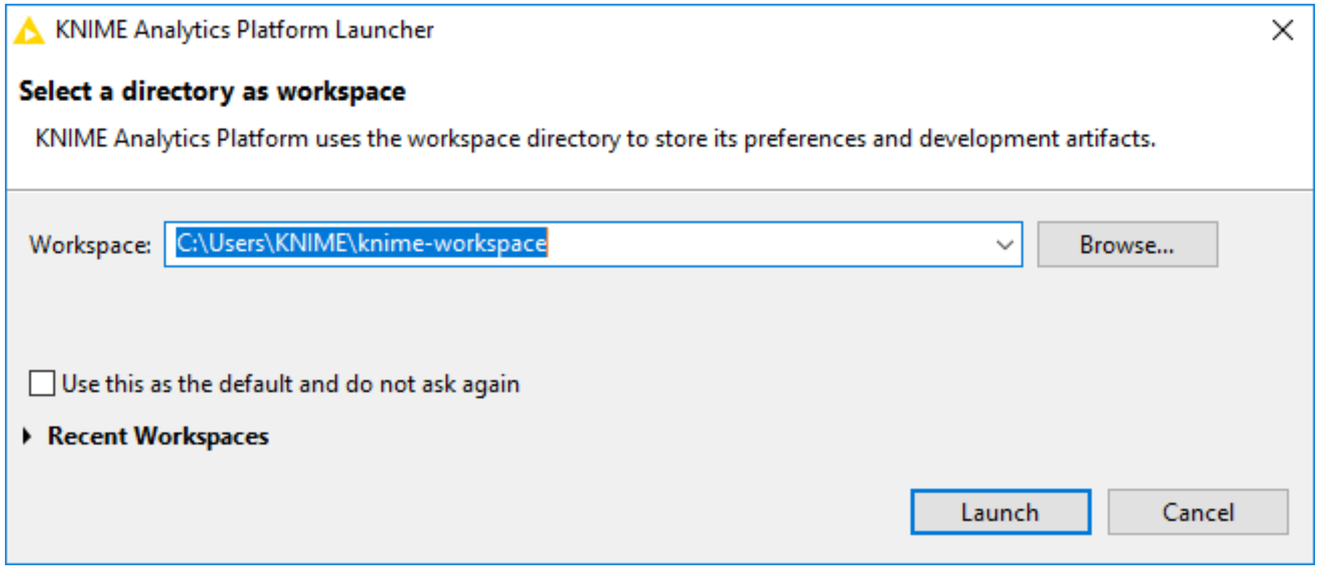
plot



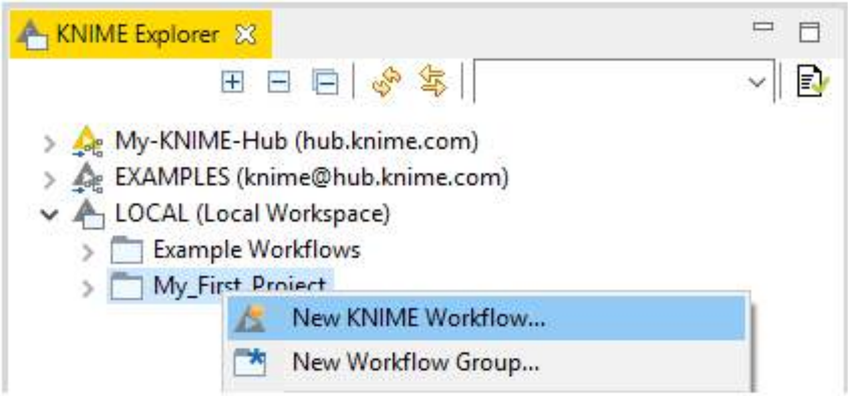


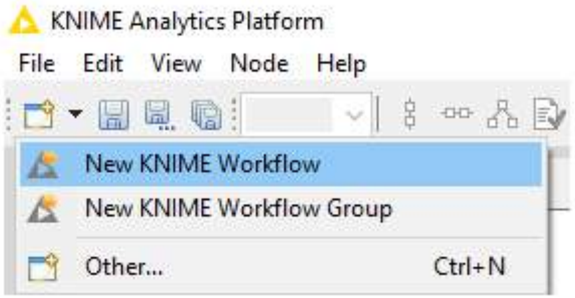

KNIME Analytics Platform
File Edit View Node Help
100%
Quick Access
KNIME Explorer
My-KNIME-Hub (hub.kn
EXAMPLES (knime@hub.
LOCAL (Local Workspace
Example Workflows
My_First_Project
data
My_First_Workflow
KNIME Explorer
Workflow Coach
Recommended Nodes
Joiner
Column Filter
Row Filter
Partitioning
GroupBy
Missing Value
Statistics
Workflow Coach
Node Repository
IO
Manipulation
Views
Analytics
DB
Other Data Types
Node Repository
*0: My_First_Workflow
My First Workflow
Select columns, filter rows and visualize sales data.
Workflow Editor
File Reader
Read sales data
Column Filter
Include country, date and amount
Row Filter
Exclude rows with country "unknown"
Stacked Area Chart
Sales over time
Pie/Donut Chart
Sales per country
Description
File Reader
Description
When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).
The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In
KNIME Hub Search
KNIME Hub Search
Outline
Outline
Console
KNIME Console
Welcome to KNIME Analytics Platform v4.0.0.v20190626
Copyright by KNIME AG, Zurich, Switzerland
Console

workspace

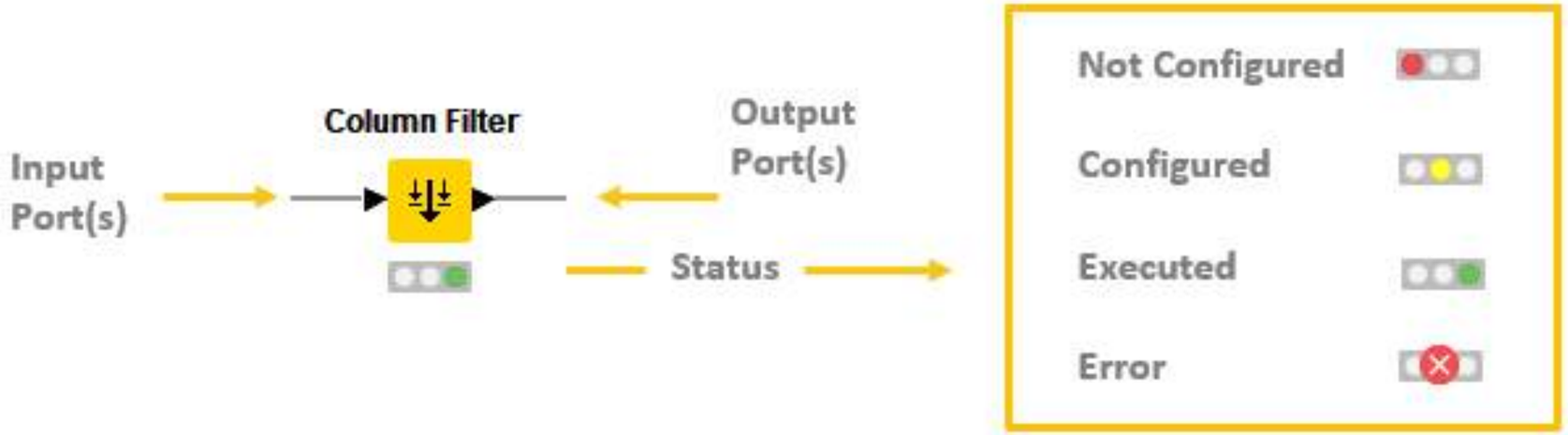

KNIME Analytics Platform Launcher
Select a directory as workspace
KNIME Analytics Platform uses the workspace directory to store its preferences and development artifacts.
Workspace:
C:\Users\KNIME\knime-workspace
Browse...
Use this as the default and do not ask again
Recent Workspaces
Launch
Cancel

Knime **workflow**

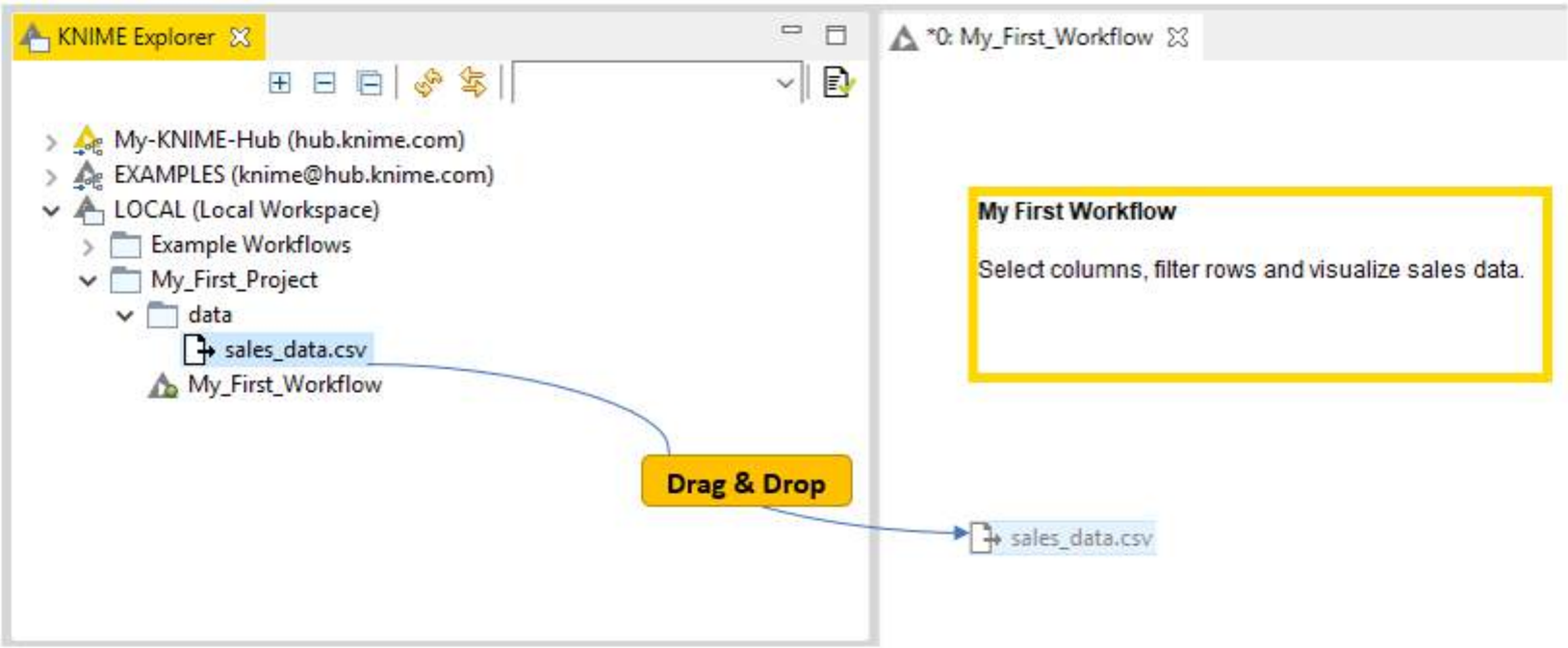




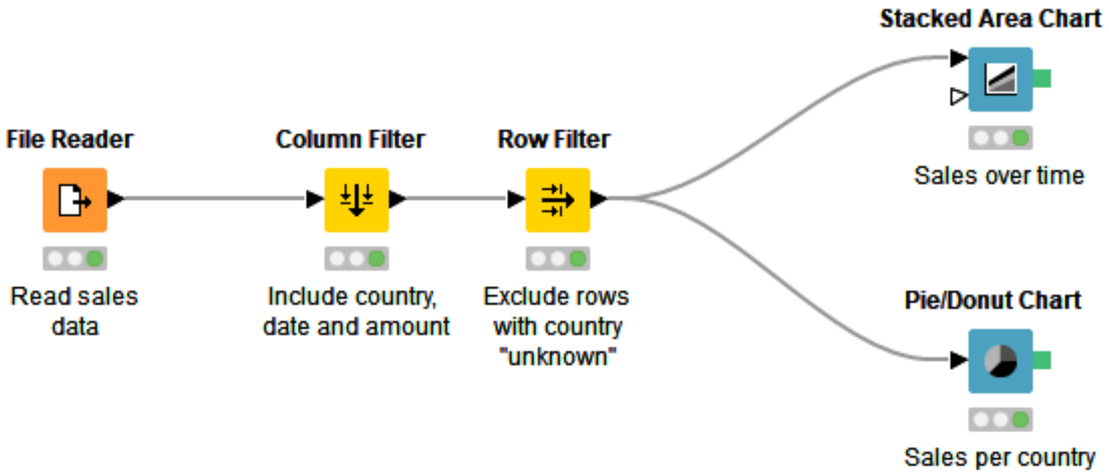
node status


Column Filter
Input Port(s)
Output Port(s)
Status
Not Configured
Configured
Executed
Error

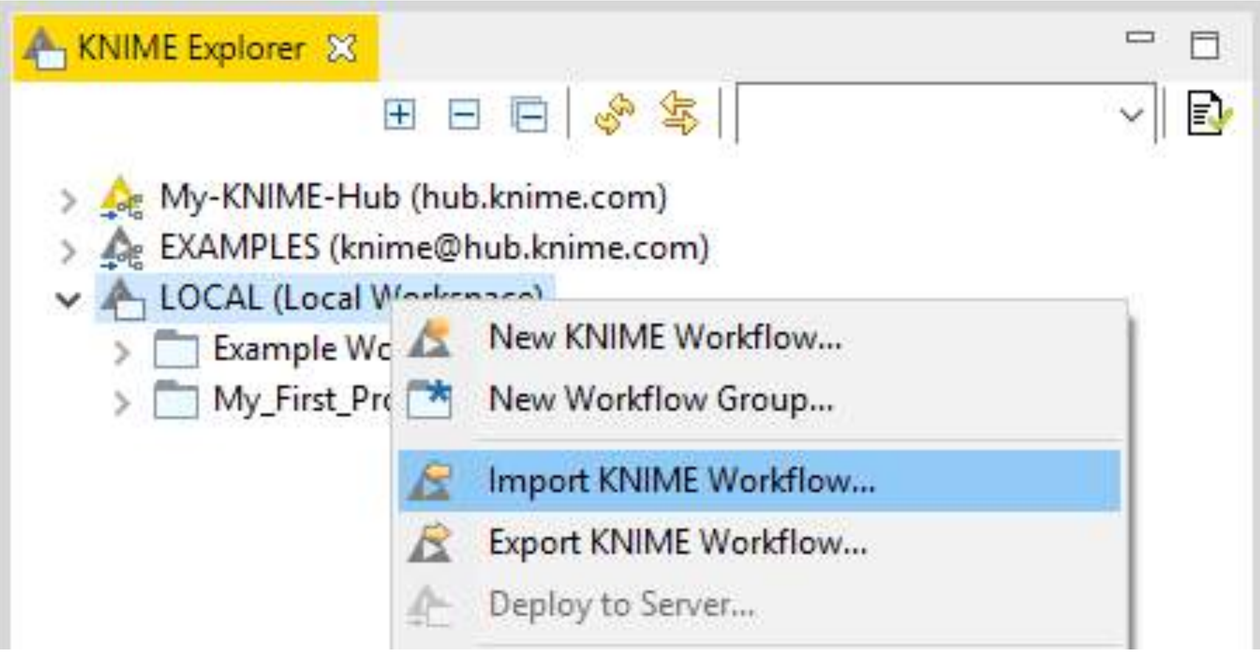
Versatile with all kinds data



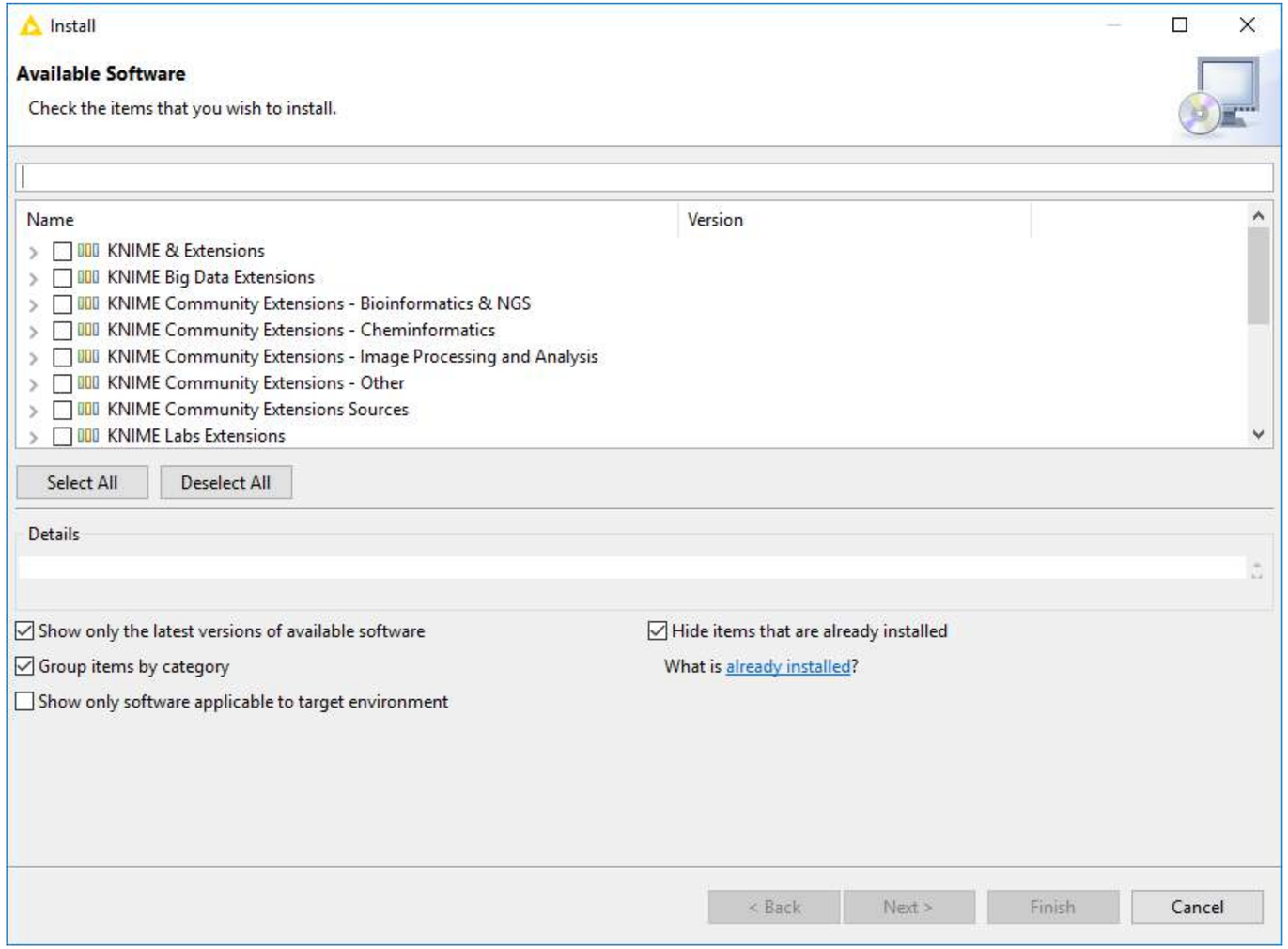
Versatile with all kinds data



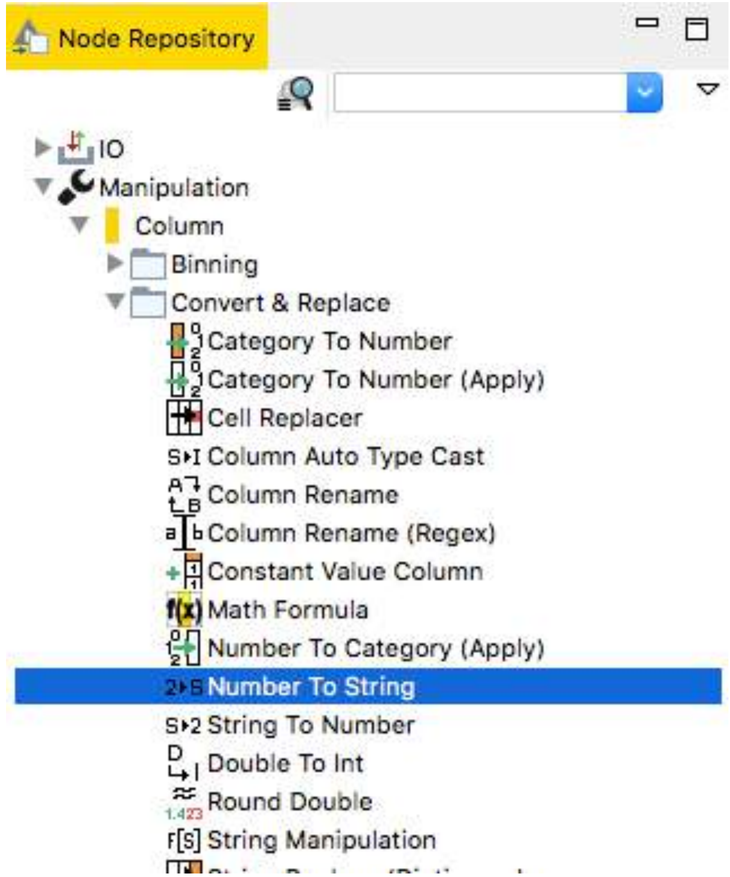
Import workflow



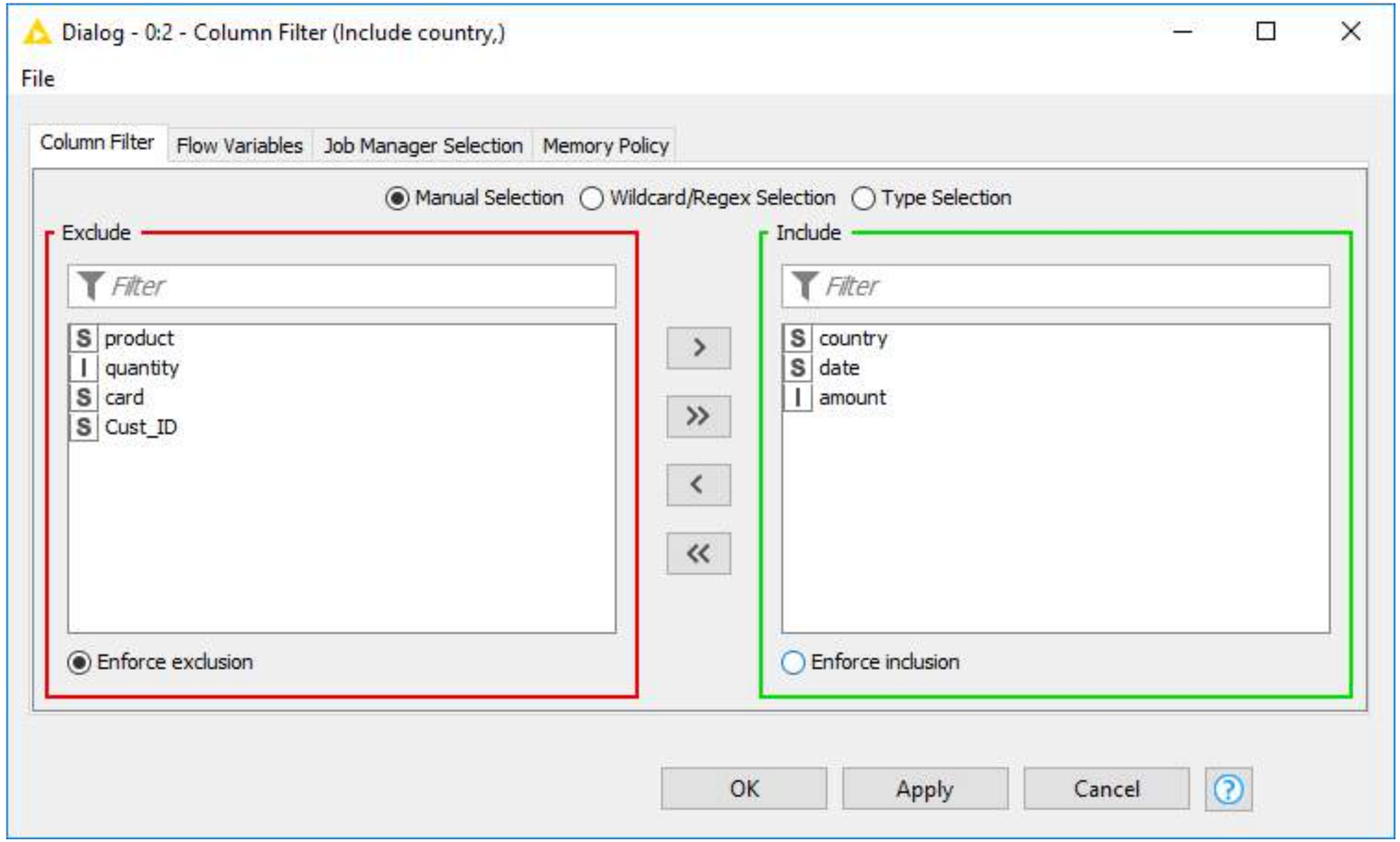
Add **extensions**



node **repository**



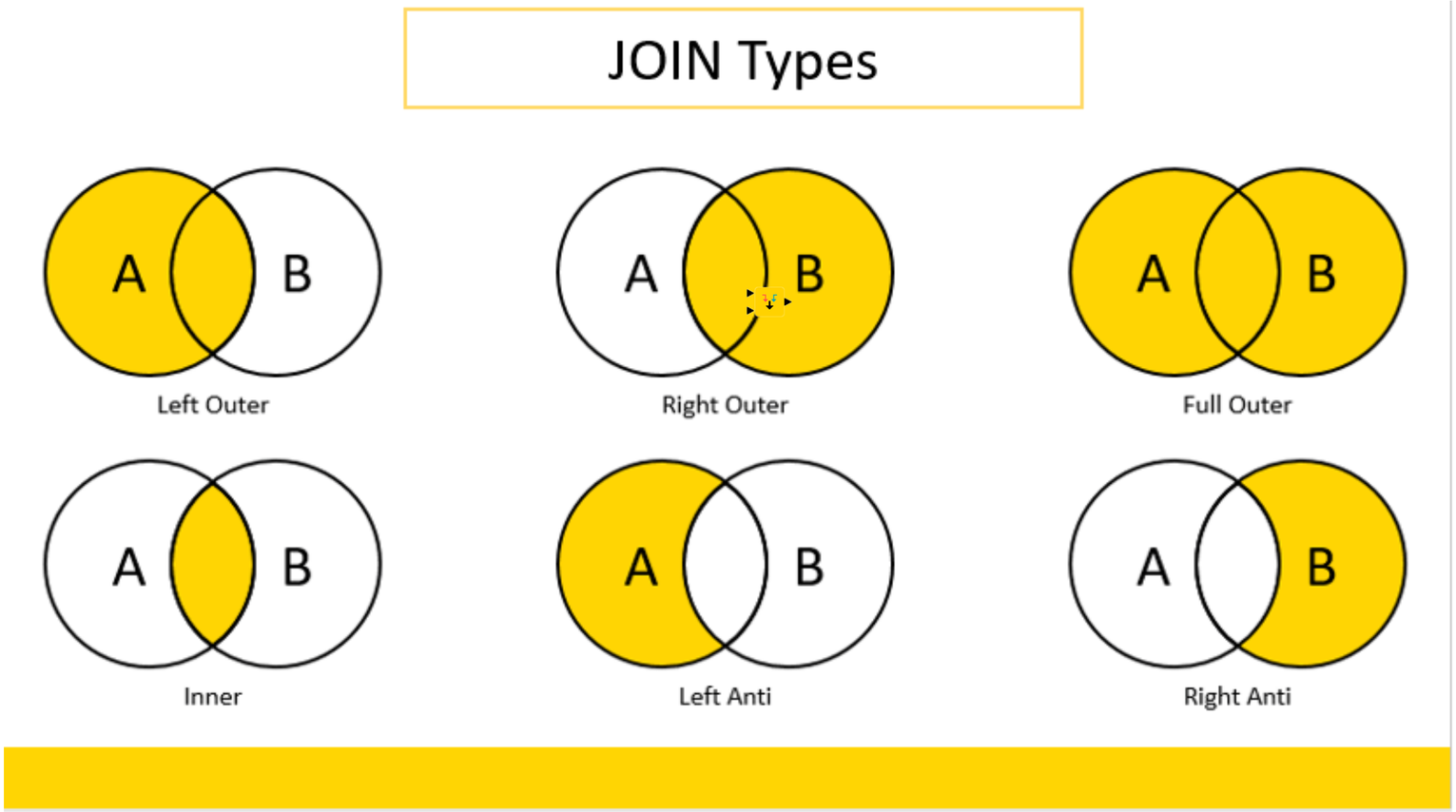
exclude / include **variables**



node **pit**

NodePit
BETA
We know 3943 nodes, 2118 workflows and 1939 users. Start mining!
NodePit Space
Share your workflows with everyone at any time – from within KNIME!

node **joiner**


JOIN Types
A
B
Left Outer
A
B
Right Outer
A
B
Full Outer
A
B
Inner
A
B
Left Anti
A
B
Right Anti

Cheat Sheet: Building a KNIME Workflow for Beginners

Open for Innovation® KNIME

Getting started with KNIME Analytics Platform

- Read through the installation guide at knime.com/installation
- Check out the 7 things you should do after installing KNIME Analytics Platform at knime.com/blog/seven-things
- Take the E-Learning Course at knime.com/knime-introductory-course
- Browse the workflows on the public EXAMPLES Server available in the KNIME Explorer

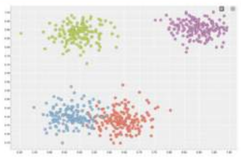
Understanding the traffic light system:

- Not configured: Node is not yet configured and cannot be executed with its current settings
- Configured: Node has been correctly configured and may be executed at any time
- Executed: Node has been successfully executed and results can be viewed and used in downstream nodes

EXPLORE

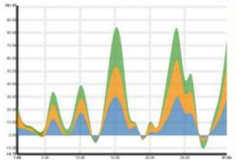
(All visualizations are interactive)

Scatter Plot: Represents input data rows as points in a two dimensional plot. Input dimensions (columns) on the x-y axis plot and graphical properties can be changed in the configuration window or interactively in the node view.



Sunburst Chart: Displays categorical columns through a hierarchy of rings. Each ring is sliced according to the nominal values in the corresponding column and to the selected hierarchy. This is a powerful chart for multivariate analysis.

Stacked Area Chart: Plots multiple numerical data columns on top of each other using the previous line as the base reference. The areas in between lines are colored for easier comparison. This chart is commonly used to visualize trending topics.



Line Plot

Line Plot: Plots numerical values in data columns (y-axis) against values in a reference column (x-axis). Data points are connected via colored lines. If the reference column on the x-axis contains sorted time values, the line plot graphically represents the evolution of a time series.

Color Manager

Color Manager: Assigns a color property to each input row based on the row's value in a selected column. This color property affects the graphical representation in the upcoming views.

Pie Chart

Pie Chart: Visualizes one aggregated metric for different data partitions with colored slices on a circle where the areas are proportional to the metric values. The partitions are defined by a categorical column.

Data Explorer

Data Explorer: Provides an interactive view to summarize the statistics of the input data via statistical measures and histograms - for both numerical and nominal columns.

Box Plot

Box Plot: Visualizes numeric columns using the quartile statistics. Watch out for the points at the end of the whiskers - they might mark outliers!

Bar Chart

Bar Chart: Visualizes one or more aggregated metrics for different data partitions with rectangular bars where the heights are proportional to the metric values. The partitions are defined by a categorical column.

ANALYZE

Decision Tree

Decision Tree: The Learner node trains a C4.5 or a CART decision tree. The configuration window includes options for pruning, early stopping, information measures, splitting values, and more. Both the Learner and the Predictor node provide an interactive view where the decision tree is displayed together with the input data propagation.

k-Means

k-Means: Implements the k-Means clustering algorithm. Number of clusters must be set prior to node execution. This node builds the clusters. The Cluster Assigner node finds the closest cluster and assigns it to the input data row. Being an unsupervised algorithm, this node pair doesn't follow the classic Learner - Predictor scheme.

Logistic Regression

Logistic Regression: The Learner node trains a logistic regression model to predict categorical target values. The configuration window includes options for solver, input feature choice, regularization functions to avoid overfitting, & more.

Scorer

Scorer: Calculates a number of performance measures such as accuracy, F1-score, or Cohen's Kappa, to quantify the quality of a classifier.

Numeric Scorer

Numeric Scorer: Calculates a number of numerical error measures, such as root mean squared error, mean absolute error, or R^2, to quantify the quality of a numerical predictor model.

ROC Curve

ROC Curve: Displays the Receiver Operating Characteristic (ROC) curve of a classifier working on a binary class problem. One of the two classes is arbitrarily chosen as the positive class and the ROC curve is built on the probabilities/scores produced for that class on the input data set.

Integrations to many open source data analytics tools are also available. Some use the KNIME node GUI (H2O, Weka, Keras, Spark MLlib). Others offer nodes with a development environment for scripting and debugging (R, Python, Java).

READ

File Reader

File Reader: Reads all text files, particularly character separated files, such as CSV files. The File Reader is the workhorse for reading text data.

Excel Reader (XLS)

Excel Reader (XLS): Reads content from sheets in Excel files (XLS, XLSX). Sheet and cells to be read can be defined in the configuration window.

Table Creator

Table Creator: Allows users to manually create a data table in its configuration window as a data sheet. Data cells can be copied and pasted in the sheet. Perfect for generating small data sets.

Model Reader

Model Reader: Reads machine learning models generated with any of the Learner nodes. Models are usually saved after training and reused in deployment.

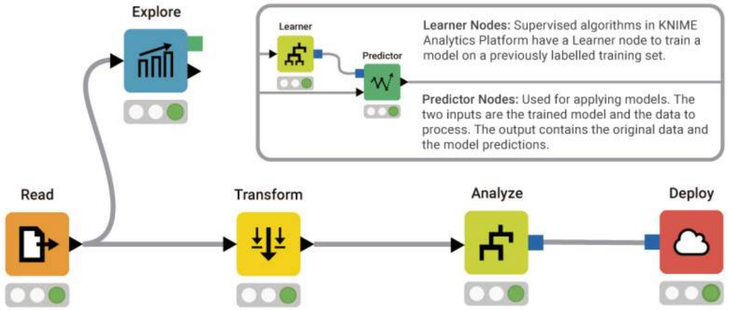
Table Reader

Table Reader: Reads data from a .table file. .table files are organized using a KNIME proprietary format, including the full file structure and are optimized for space and speed - providing maximum performance with minimum configuration!

Google Sheets Reader

Google Sheets Reader: Reads data from a Google Sheet file. Authentication occurs on the Google site. Google credentials are not saved within the KNIME workflow.

knime:// protocol: References a file path relative to some key location of the current KNIME installation like *knime://knime.workflow/../<filename>* or *knime://<knime.-server.mountpoint>/<path>/<filename>*



Learner

Predictor

Learner Nodes: Supervised algorithms in KNIME Analytics Platform have a Learner node to train a model on a previously labelled training set.

Predictor Nodes: Used for applying models. The two inputs are the trained model and the data to process. The output contains the original data and the model predictions.

TRANSFORM

GroupBy

GroupBy: Groups the rows of a table by the unique values in selected columns and calculates aggregation and statistical measures for the defined groups. Despite its simple name, it offers powerful functionality and has many unsuspected usages. For example - row deduplication.

Pivoting

Pivoting: Extends the aggregation functionality of the GroupBy node by creating an output data table with columns and rows for the unique values in selected input columns. Note: the unique values of the grouping column become rows and the unique values of the pivoting column become columns.

Rule Engine

Rule Engine: Applies a set of rules to each row of the input data table. All Rule Engine operators are also available in the Column Expressions node.

Partitioning

Partitioning: Splits data into two subsets according to a sampling strategy. This node is generally used to produce a training and a test set to train and evaluate a machine learning model.

Row Filter

Row Filter: Filters rows in or out from the input data table according to a filtering rule. The filtering rule can match a value in a selected column or numbers in a numerical range.

Math Formula

Math Formula: Implements a number of math operations across multiple input columns, from simple sum and average, to logarithms and exponentials. All Math Formula operators are also available in the Column Expressions node.

String to Date&Time

String to Date&Time: Converts values in a String column into Date&Time values. The Date&Time format contained in the String values can be manually defined or auto guessed.

Cell Splitter

Cell Splitter: Splits values in a selected column into two or more substrings, as defined by a delimiter match. Delimiter is a set character, such as a comma, space, or any other character or character sequence.

Column Filter

Column Filter: Filters columns in or out from the input data table according to a filtering rule. Columns to be retained can be manually picked or selected according to their type, or of a regex expression matching their name.

Column Rename

Column Rename: Assigns new names and types to selected columns, as configured in the dialog.

Joiner

Joiner: Joins rows from two data tables based on common values in one or more key columns. The most common join types are possible: inner join, left outer join, right outer join, and full outer join.

Sorter

Sorter: Sorts the table in ascending or descending order based on the values of a chosen column. In addition, it is possible to sort based on multiple columns.

Concatenate

Concatenate: Merges vertically two data tables, by piling up cells in columns with the same name. Cells in uncommon columns are filled with missing values. The Concatenate (Optional in) node merges vertically up to four data tables.

Missing Value

Missing Value: Defines a strategy to deal with missing values in the input data table - either globally on all columns, or individually for each single column.

String Manipulation

String Manipulation: Performs operations on String values in columns, such as combining two or more Strings together, extracting one or more substrings, trimming blank spaces, and so on. All operators are also available in the Column Expressions node.

DEPLOY

Data to Report

Data to Report: Marks the data table to be exported to BIRT - a partially open source reporting tool integrated within KNIME. When switching from KNIME to BIRT, the marked data sets are imported into BIRT. The Image To Report node marks the input images to be exported to BIRT.

Excel Writer (XLS)

Excel Writer (XLS): Writes the input data table to a sheet in an Excel file (XLS or XLSX).

Table Writer

Table Writer: Writes the input data table to a file using the .table KNIME proprietary format. This format includes the full file structure and is optimized for space and speed. Including the table structure in the file is a great advantage - especially when exchanging data files among users.

CSV Writer

CSV Writer: Writes the input data table to a CSV file.

Google Sheets Writer

Google Sheets Writer: Writes the input data table into a Google Sheet file. Authentication occurs on the Google site. Google credentials are not saved within the KNIME workflow.

Send to Tableau Server

Connectors to Tableau: Export input data table into a Tableau file or server for reporting.

Resources

- **KNIME Forum:** Join our global community and engage in conversations at forum.knime.com
- **KNIME Books:** More tips, ideas, and lessons from knime.com/knimepress
- **KNIME Events:** Take a course, attend a workshop, or join a meetup at knime.com/events
- **KNIME Blog:** Engaging topics, challenges, industry news, and knowledge nuggets at knime.com/blog
- **Workflow Hub:** Browse our example workflows and/or share your own workflows. Show appreciation for others by adding ratings, or comments at workflows.knime.com
- **More Guides:** Still using SAS or Excel? Transition to KNIME Analytics Platform with these handy guides at knime.com/knimepress
- **KNIME Server:** For team-based collaboration, automation, management, and deployment check out KNIME Server at knime.com/server

© 2018 KNIME AG. All rights reserved. The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.

KNIME AG Technoparkstrasse 1, 8005 Zurich, Switzerland / info@knime.com / www.knime.com

Cheat Sheet: Machine Learning with KNIME Analytics Platform

SUPERVISED LEARNING

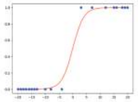
Supervised Learning: A set of machine learning algorithms to predict the value of a target class or variable. They produce a mapping function (model) from the input features to the target class/variable. To estimate the model parameters during the training phase, labeled example data are needed in the training set. Generalization to unseen data is evaluated on the test set data via scoring metrics.

CLASSIFICATION

Classification: A type of supervised learning where the target is a class. The model learns to produce a class score and to assign each vector of input features to the class with the highest score. A cost can be introduced to penalize one of the classes during class assignment.

Decision Tree: Follows the C4.5 decision tree algorithm. These algorithms generate a tree-like structure, creating data subsets, aka tree nodes. At each node, the data are split based on one of the input features, generating two or more branches as output. Further splits are made in subsequent nodes until a node is generated where all or almost all of the data belong to the same class.

Logistic Regression: A statistical algorithm that models the relationship between the input features and the categorical output classes by maximizing a likelihood function. Originally developed for binary problems, it has been extended to problems with more than two classes (multinomial logistic regression).



Naive Bayes Learner

Naive Bayes: Based on Bayes' theorem and assuming statistical independence between input features (thus "naive"), this algorithm estimates the conditional probability of each output class given the vector of input features. The class with the highest conditional probability is assigned to the input data.

SVM Learner

Support Vector Machine (SVM): A supervised algorithm constructing a set of discriminative hyperplanes in high-dimensional space. In addition to linear classification, SVMs can perform non-linear classification by implicitly mapping their inputs into high-dimensional feature spaces, where the two classes are linearly separable.

K Nearest Neighbor

k-Nearest Neighbor (kNN): A non-parametric method that assigns the class of the *k* most similar points in the training data, based on a pre-defined distance measure. Class attribution can be weighted by the distance to the *k*-th point and/or by the class probability.

NUMERIC PREDICTION & CLASSIFICATION

RProp MLP Learner

Artificial Neural Networks (ANN, NN): Inspired by biological nervous systems, Artificial Neural Networks are based on architectures of interconnected units called artificial neurons. Artificial neurons' parameters and connections are trained via dedicated algorithms, the most popular being the Back-Propagation algorithm.

Keras Dense Layer

Deep Learning: Deep learning extends the family of ANNs with deeper architectures and additional paradigms, e.g. Recurrent Neural Networks (RNN). The training of such networks, has been enabled by recent advances in hardware performance as well as parallel execution.

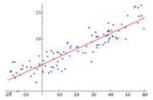
H2O Generalized Linear Model Learner

Generalized Linear Model (GLM): A statistics-based flexible generalization of ordinary linear regression, valid also for non-normal distributions of the target variable. GLM uses the linear combination of the input features to model an arbitrary function of the target variable (the link function) rather than the target variable itself.

NUMERIC PREDICTION

Numeric Prediction: A type of supervised learning for numeric target variables. The model learns to associate one or more numbers with the vector of input features. Note that numeric prediction models can also be trained to predict class scores and therefore can be used for classification problems too.

Linear/Polynomial Regression: Linear Regression is a statistical algorithm to model a multivariate linear relationship between the numeric target variable and the input features. Polynomial Regression extends this concept to fitting a polynomial function of a pre-defined degree.



Regression Tree: Builds a decision tree to predict numeric values through a recursive, top-down, greedy approach known as recursive binary splitting. At each step, the algorithm splits the subsets represented by each node into two or more new branches using a greedy search for the best split. The average value of the points in a leaf produces the numerical prediction.

TIME SERIES ANALYSIS

Time Series Analysis: A set of numeric prediction methods to analyze/predict time series data. Time series are time ordered sequences of numeric values. In particular, time series forecasting aims at predicting future values based on previously observed values.

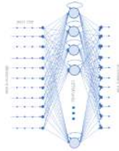
ARIMA Learner

Auto-Regressive Integrated Moving Average (ARIMA): A linear Auto-Regressive (AR) model is constructed on a specified number *p* of past values; data are prepared by a degree of differencing *d* to correct non-stationarity; while a linear combination - named Moving Average (MA) - models the *q* past residual errors. All ARIMA model parameters are estimated concurrently by various algorithms, mostly following the Box–Jenkins approach.

Lag Column, Linear Regression Learner

ML-based TSA: A numeric prediction model trained on vectors of past values can predict the current numeric value of the time series.

Long Short Term Memory (LSTM) Units: LSTM units produce a hidden state by processing *m* x *n* tensors of input values, where *m* is the size of the input vector at any time and *n* the number of past vectors. The hidden state can then be transformed into the current vector of numeric values. LSTM units are suited for time series prediction as values from past vectors can be remembered or forgotten through a series of logic gates.



ENSEMBLE LEARNING

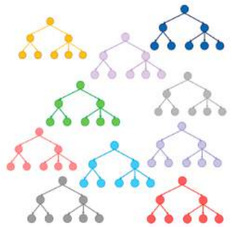
Ensemble Learning: A combination of multiple models from supervised learning algorithms to obtain a more stable and accurate overall model. Most commonly used ensemble techniques are Bagging and Boosting.

BAGGING

Bagging: A method for training multiple classification/regression models on different randomly drawn subsets of the training data. The final prediction is based on the predictions provided by all the models, thus reducing the chance of overfitting.

Tree Ensemble of Decision/Regression Trees: Ensemble model of multiple decision/regression trees trained on different subsets of data. Data subsets with less or equal rows and less or equal columns are bootstrapped from the original training set. Final prediction is based on a hard vote (majority rule) or soft-vote (averaging all probabilities or numeric predictions) on all involved trees.

Random Forest of Decision/Regression Trees: Ensemble model of multiple decision/regression trees trained on different subsets of data. Data subsets with the same number of rows are bootstrapped from the original training set. At each node, the split is performed on a subset of *sqrt(x)* features from the original *x* input features. Final prediction is based on a hard vote (majority rule) or soft-vote (averaging all probabilities or numerical predictions) on all involved trees.



Prediction Fusion

Custom Ensemble Model: Combining different supervised models to form a custom ensemble model. The final prediction can be based on majority vote as well as on the average or other functions of the output results.

XGBoost Tree Ensemble Learner

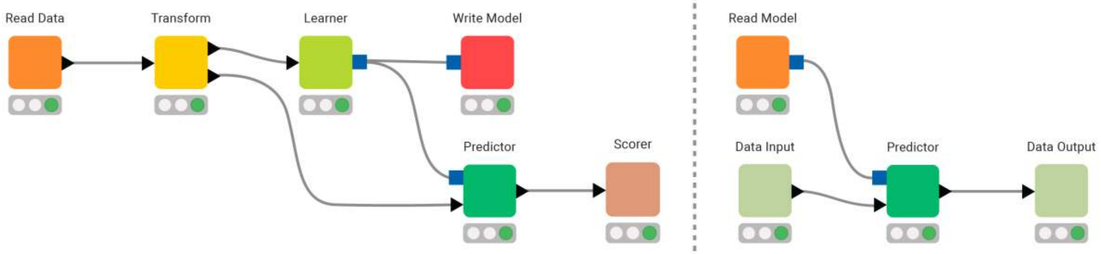
XGBoost: An optimized distributed library for machine learning models in the gradient boosting framework, designed to be highly efficient, flexible, and portable. It features regularization parameters to penalize complex models, effective handling of sparse data for better performance, parallel computation, and more efficient memory usage.

BOOSTING

Boosting: A method for training a set of classification/regression models iteratively. At each step, a new model is trained on the prediction errors and added to the ensemble to improve the results from the previous model state, leading to higher accuracy after each iteration.

Gradient Boosted Regression Trees: Ensemble model combining multiple sequential simple regression trees into a stronger model. The algorithm builds the model stagewise. At each iteration, a simple regression tree is fitted to predict the residuals of the current model, following the gradient of the loss function. This leads to an increasingly accurate and complex overall model. The same regression trees can also be used for classification.

DEPLOYMENT



TRAINING

EVALUATION

Evaluation: Various scoring metrics for assessing model quality - in particular, a model's predictive ability or propensity to error.

Scorer

Confusion Matrix: A representation of a classification task's success through the count of matches and mismatches between the actual and predicted classes, aka true positives, false negatives, false positives, and true negatives. One class is arbitrarily selected as the positive class.

Accuracy Measures: Evaluation metrics for a classification model calculated from the values in the confusion matrix, such as sensitivity and specificity, precision and recall, or overall accuracy.

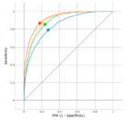
Cross Validation

Cross-Validation: A model validation technique for assessing how the results of a machine learning model will generalize to an independent dataset. A model is trained and validated *N* times on different pairs of training set and test set, both extracted from the original dataset. Some basic statistics on the resulting *N* error or accuracy measures gives insights on overfitting and generalization.

Numeric Scorer

Numeric Error Measures: Evaluation metrics for numeric prediction models quantifying the error size and direction. Common metrics include RMSE, MAE, or R^2. Most of these metrics depend on the range of the target variable.

ROC Curve: A graphical representation of the performance of a binary classification model with false positive rates on the *x*-axis and true positive rates on the *y* axis. Multiple points for the curve are obtained for different classification thresholds. The area under the curve is the metric value.



Resources

- **E-Books:** Learn even more with the KNIME books. From basic concepts in "KNIME Beginner's Luck", to advanced concepts in "KNIME Advanced Luck", through to examples of real-world case studies in "Practicing Data Science". Available for purchase at knime.com/knimepress
- **KNIME Blog:** Engaging topics, challenges, industry news, and knowledge nuggets at knime.com/blog
- **KNIME Hub:** Search, share, and collaborate on KNIME workflows, nodes, and components with the entire KNIME community at hub.knime.com
- **KNIME Forum:** Join our global community and engage in conversations at forum.knime.com
- **KNIME Server:** The enterprise software for team-based collaboration, automation, management, and deployment of data science workflows as analytical applications and services. Visit knime.com/server for more information.

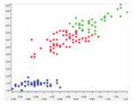
UNSUPERVISED LEARNING

Unsupervised Learning: A set of machine learning algorithms to discover patterns in the data. A labeled dataset is not required, since data are ultimately organized and/or transformed based on similarity or statistical measures.

CLUSTERING

Clustering: A branch of unsupervised learning algorithms that groups data together based on similarity measures, without the help of labels, classes, or categories.

k-Means: The *n* data points in the dataset are clustered into *k* clusters based on the shortest distance from the cluster prototypes. The cluster prototype is taken as the average data point in the cluster.



Hierarchical Clustering: Builds a hierarchy of clusters by either collecting the most similar (agglomerative approach) or separating the most dissimilar (divisive approach) data points and clusters, according to a selected distance measure. The result is a dendrogram clustering the data together bottom-up (agglomerative) or separating the data in different clusters top-down (divisive).

DBSCAN: A density-based non-parametric clustering algorithm. Data points are classified as core, density-reachable, and outlier points. Core and density-reachable points in high density regions are clustered together, while points with no close neighbors in low-density regions are labeled as outliers.

SOTA Learner

Self-Organizing Tree Algorithm (SOTA): A special Self-Organizing Map (SOM) neural network. Its cell structure is grown using a binary tree topology.

Fuzzy c-Means

Fuzzy c-Means: One of the most widely used fuzzy clustering algorithms. It works similarly to the k-Means algorithm, but it allows for data points to belong to more than one cluster, with different degrees of membership.

RECOMMENDATION ENGINES

Recommendation Engines: A set of algorithms that use known information about user preferences to predict items of interest.

Association Rule Learner

Association Rules: The node reveals regularities in co-occurrences of multiple products in large-scale transaction data recorded at points-of-sale. Based on the a-priori algorithm, the most frequent itemsets in the dataset are used to generate recommendation rules.

Spark Collaborative Filtering Learner (MLlib)

Collaborative Filtering: Based on the Alternating Least Squares (ALS) technique, it produces recommendations (filtering) about the interests of a user by comparing their current preferences with those of multiple users (collaborating).

Knime Data manipulation 1.1

https://hub.knime.com/knime/spaces/Examples/latest/02_ETL_Data_Manipulation/01_Filtering/02_Column_Filter

https://hub.knime.com/knime/spaces/Examples/latest/02_ETL_Data_Manipulation/01_Filtering/03_Row_Filtering

https://hub.knime.com/knime/spaces/Examples/latest/02_ETL_Data_Manipulation/03_Joining_and_Concatenating/03_JoinConcatenate_Examples

https://hub.knime.com/knime/spaces/Examples/latest/02_ETL_Data_Manipulation/02_Aggregations/06_Basic_Examples_for_Using_the_GroupBy_Node

https://hub.knime.com/knime/spaces/Examples/latest/02_ETL_Data_Manipulation/00_Basic_Examples/01_Example_for_Standard_Preprocessing

02_ETL_basics

Knime Data manipulation 1.5

Row ID	S product	S country	S date	I quantity	I amount	S card	S Cust_ID
Row0	prod_1	China	12.12.2009	1	35	Y	Cust_2
Row1	prod_2	Germany	01.02.2011	1	40	Y	Cust_1
Row2	prod_3	USA	17.03.2010	1	80	Y	Cust_3
Row3	prod_1	China	28.06.2010	10	350	Y	Cust_5
Row4	prod_2	Germany	31.03.2010	5	200	Y	Cust_4
Row5	prod_3	USA	20.08.2009	20	1600	?	Cust_3
Row6	prod_1	USA	11.10.2010	2	70	Y	Cust_6
Row7	prod_2	Germany	22.11.2009	15	600	N	Cust_1
Row8	prod_3	Germany	13.01.2010	1	80	?	Cust_4
Row9	prod_1	USA	04.07.2009	2	70	Y	Cust_3
Row10	prod_2	USA	20.01.2010	2	80	?	Cust_6
Row11	prod_3	Germany	14.09.2010	2	160	?	Cust_1
Row12	prod_1	Brazil	17.07.2010	5	175	?	Cust_7
Row13	prod_2	USA	07.07.2010	12	480	?	Cust_3
Row14	prod_4	unknown	12.12.2008	1	3	?	Cust_8
Row15	prod_3	China	02.01.2011	8	640	?	Cust_2
Row16	prod_1	Germany	20.03.2011	11	385	N	Cust_4
Row17	prod_2	USA	22.02.2010	6	240	?	Cust_6
Row18	prod_3	China	10.05.2009	2	160	?	Cust_2
Row19	prod_1	Germany	06.03.2011	10	350	?	Cust_4
Row20	prod_2	Germany	22.06.2010	6	240	?	Cust_1
Row21	prod_3	Brazil	08.06.2009	15	1200	?	Cust_7
Row22	prod_1	Germany	02.12.2009	1	35	Y	Cust_1
Row23	prod_2	Germany	11.02.2011	1	40	?	Cust_4
Row24	prod_3	China	12.03.2010	1	80	?	Cust_5
Row25	prod_1	China	28.08.2010	10	350	N	Cust_2

In this exercise you need to:
1)Read the file sales2008-2011.csv including a number of contracts with different customers. Each customer is identified by its unique Customer ID. Each contract is identified by its unique RowID.

2) Convert the contract dates from String to DateTime type using the String to Date/Time node with format dd.MM.yyyy (if you open the file, you will see that this is the format used for the date column).

3) For each country, extract the active customers. That is find out which customers are active in which countries

4) For each customer, then calculate:
- -Date of first contract
- -Total amount of money throughout all contracts
- -Countries where customer has signed contracts
- -The number of days between the first and the last purchase

5) Extract only those customers active in the **USA** and who paid in total **> 5000 $** and those customers active in **Germany** and who paid between **1000 and 2000$** in total. Collect them together in a single table.

6) For the selected customers, extract all their contracts from the original dataset and join the customer summary information to each contract data row.

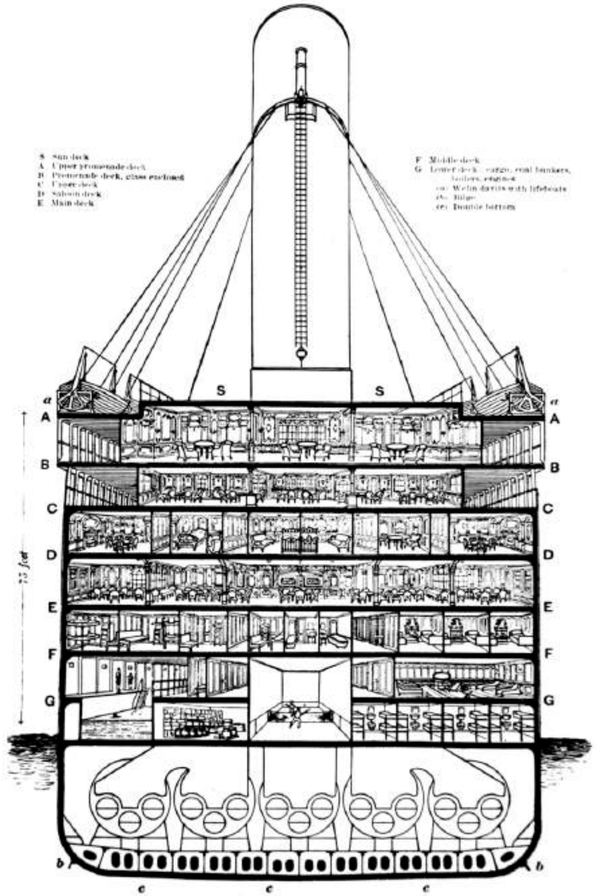
7) Write the resulting table to a CSV file in a "data" folder located in the workflow folder

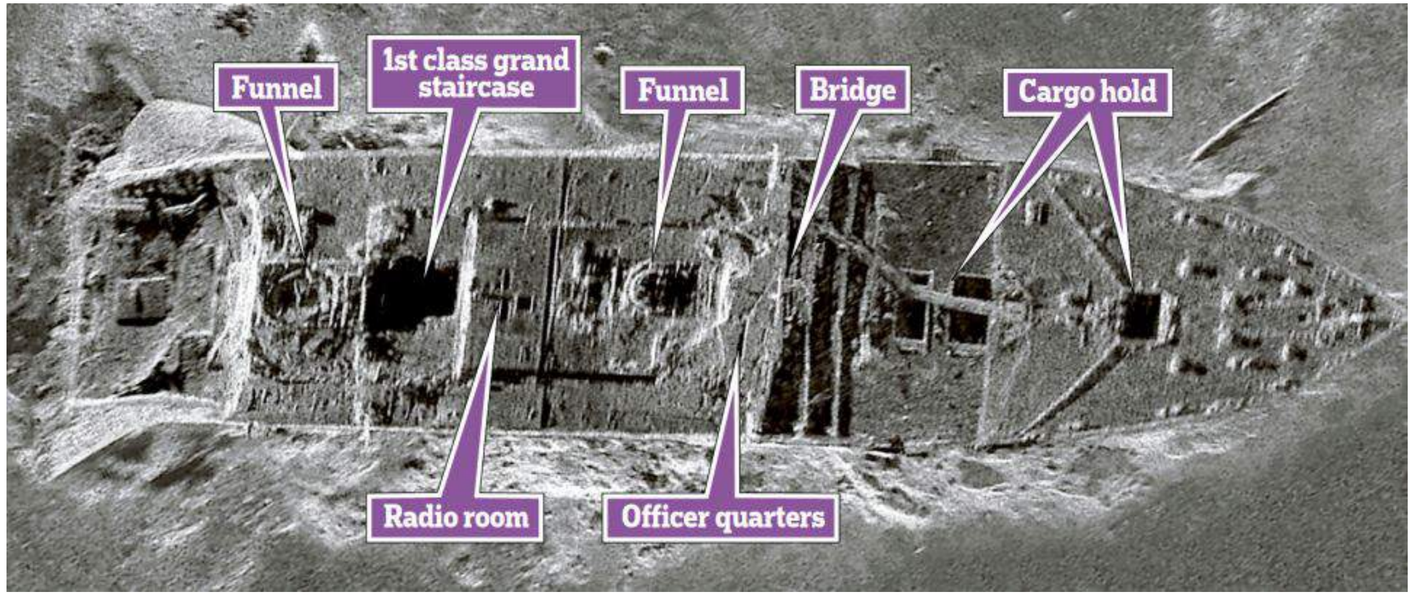
KNIME

titanic survived or not survived?

KNIME

titanic survived or not survived?





KNIME

titanic survived or not survived?

data always talks about:

-people

-events

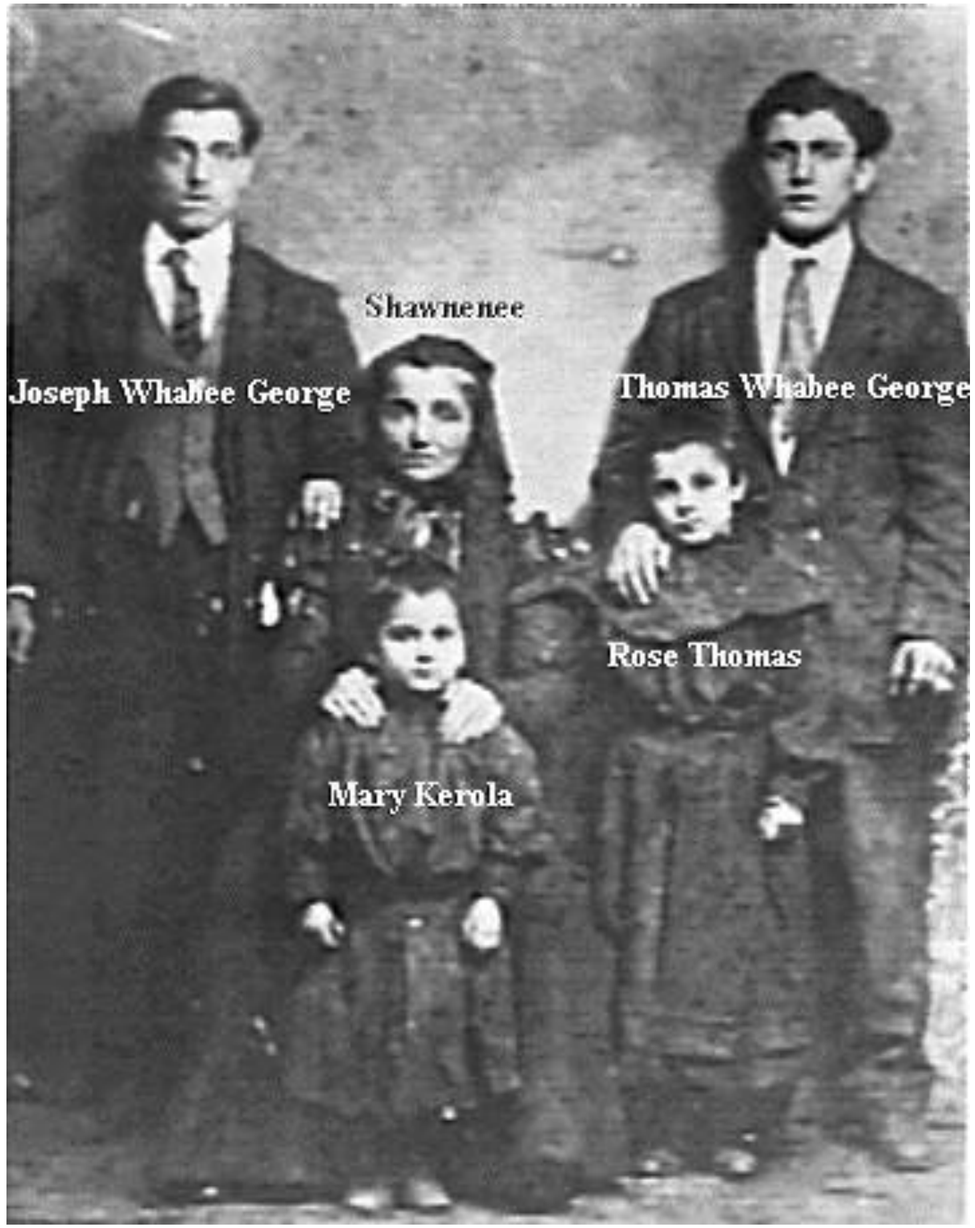
-patterns

-correlations

-stories

-variables

-connections



Knime titanic use case 1.6

VARIABLE DESCRIPTIONS

Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Knime titanic use case 1.6

Business understanding	Data understanding	Data preparation	Modeling	Evaluation
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results
Assess situation	Describe data	Clean data	Generate test design	Review process
Determine data mining goals	Explore data	Construct data	Build model	Determine next steps
Produce project plan	Verify data quality	Integrate data	Assess model	
		Format data		

Titanic data

// Development process

Algorithm

tabular data

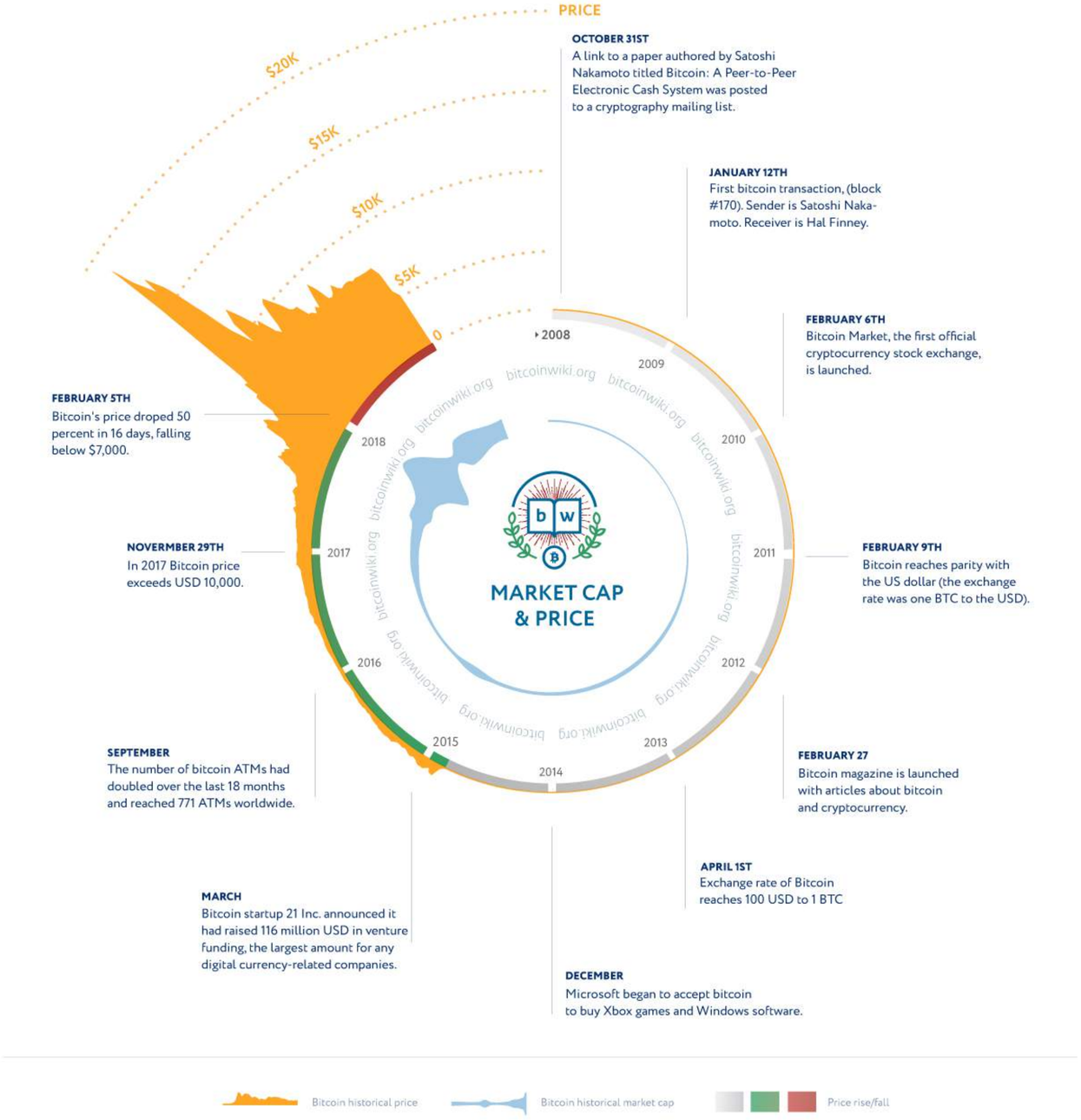
ML // Crypto Demand prediction

1. Investment Predictions: bitcoin

ML // Crypto Demand prediction

Intro // History of bitcoin prices

- 18 August 2008, the domain name bitcoin.org was registered.
- November 6th 2010, Bitcoin share capital reaches 1 million USD. Its exchange rate on MtGox reaches USD 0.50 per BTC.
- June 2nd 2011, USD to BTC rate is 10 USD to the coin. For 6 days, the Bitcoin value is fixed at 31.91 USD on MtGox.
- February 28th 2013, Bitcoin exchange rate surpasses 31.91 USD for the first time for th last 601 days.
- April 1st,2013 Exchange rate of Bitcoin reaches 100 USD to 1 BTC.
- January,2015 Coinbase raised 75 million USD as part of a Series C funding round, smashing the previous record for a bitcoin company.
- February ,2015 Bitcoin price reached USD 262.
- January 2017,After the rally for most of the second half of 2016, bitcoin broke the USD 1,000 mark for the first time in 3 years.
- June 12th 2017, Bitcoin exchange rate exceeds USD 3000 to the BTC.
- November 29th 2017, Bitcoin price exceeds USD 10,000.
- December 18th 2017, Bitcoin reaches a record high, but does not reach USD 20,000.
- December 28th 2017, The price of bitcoins fell after South Korea announced additional measures to regulate bitcoin trading, including the potential closure of exchanges, amo the volatile movements in the world's third largest cryptocurrency market.
- October 31st 2018, USD 6,300, on the 10 year anniversary of Bitcoin, price holds steac above USD 6,000 during a period of historically low volatility.



START DEMO

BLOCKCHAIN

DATA | Welcome to Blockchain Demo 2.0!

PREVIOUS HASH 0

HASH 000dc75a315c77a1f9c98fb6247d03dd18ac52632d7dc6a9920261d8109b37cf

GENESIS BLOCK on Tue, 17 Oct 2017 19:53:20 GMT

604

DATA | hola Mundo

PREVIOUS HASH 000dc75a315c77a1f9c98fb6247d03dd18ac52632d7dc6a9920261d8109b37cf

HASH 000bb3670a715503569def7f9c5400e7622a142bc187eef45f3760ca580cfae7

BLOCK #1 on Thu, 08 Apr 2021 09:22:50 GMT

7753

blockchaindemo.io

Network Nodes

Bitcoin transactions

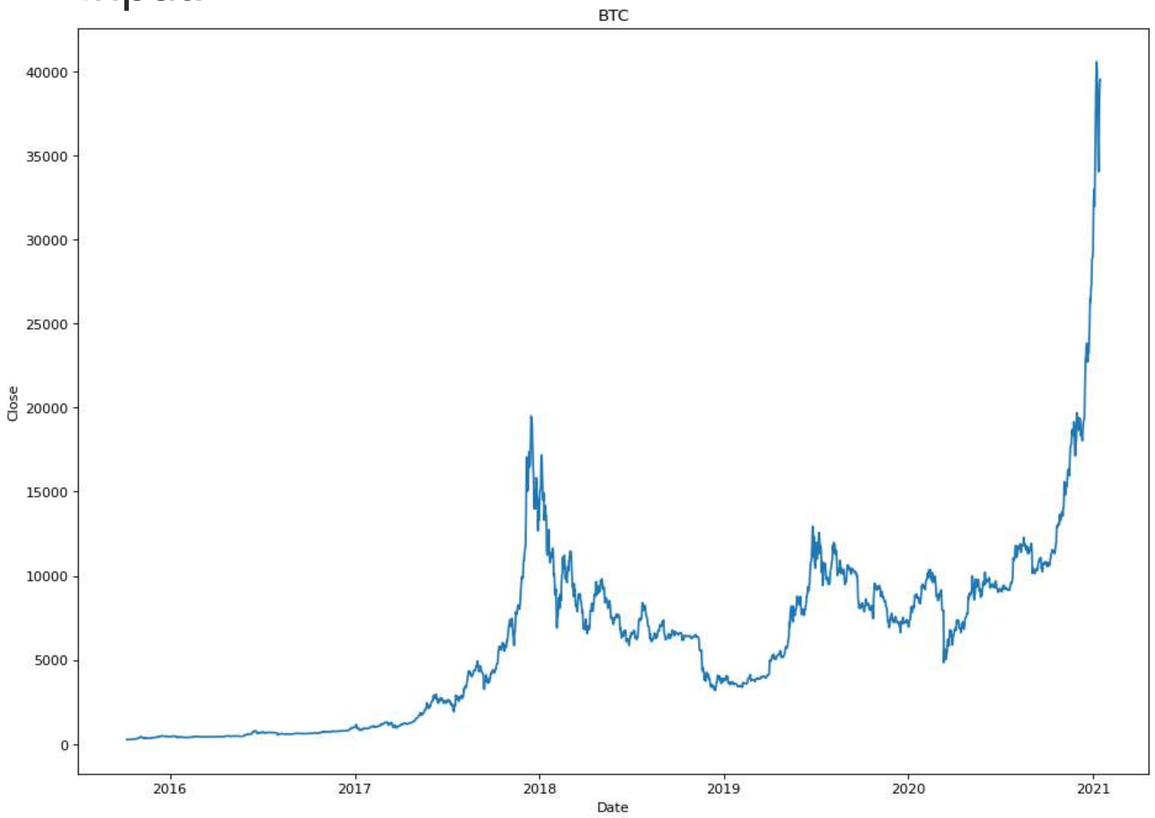
Mined transactions

Node map

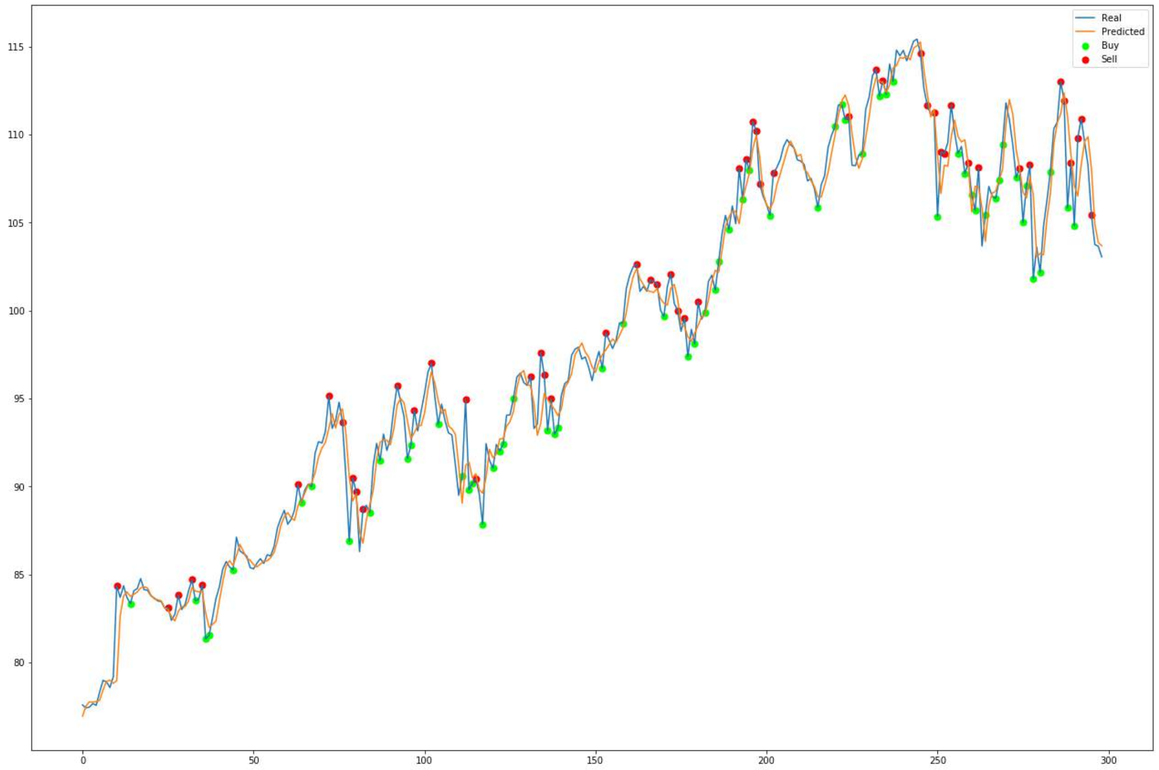
ML // Crypto Demand prediction

1 day // 5 day // 30 day Bitcoin pricing forecasting

Input:



Output:



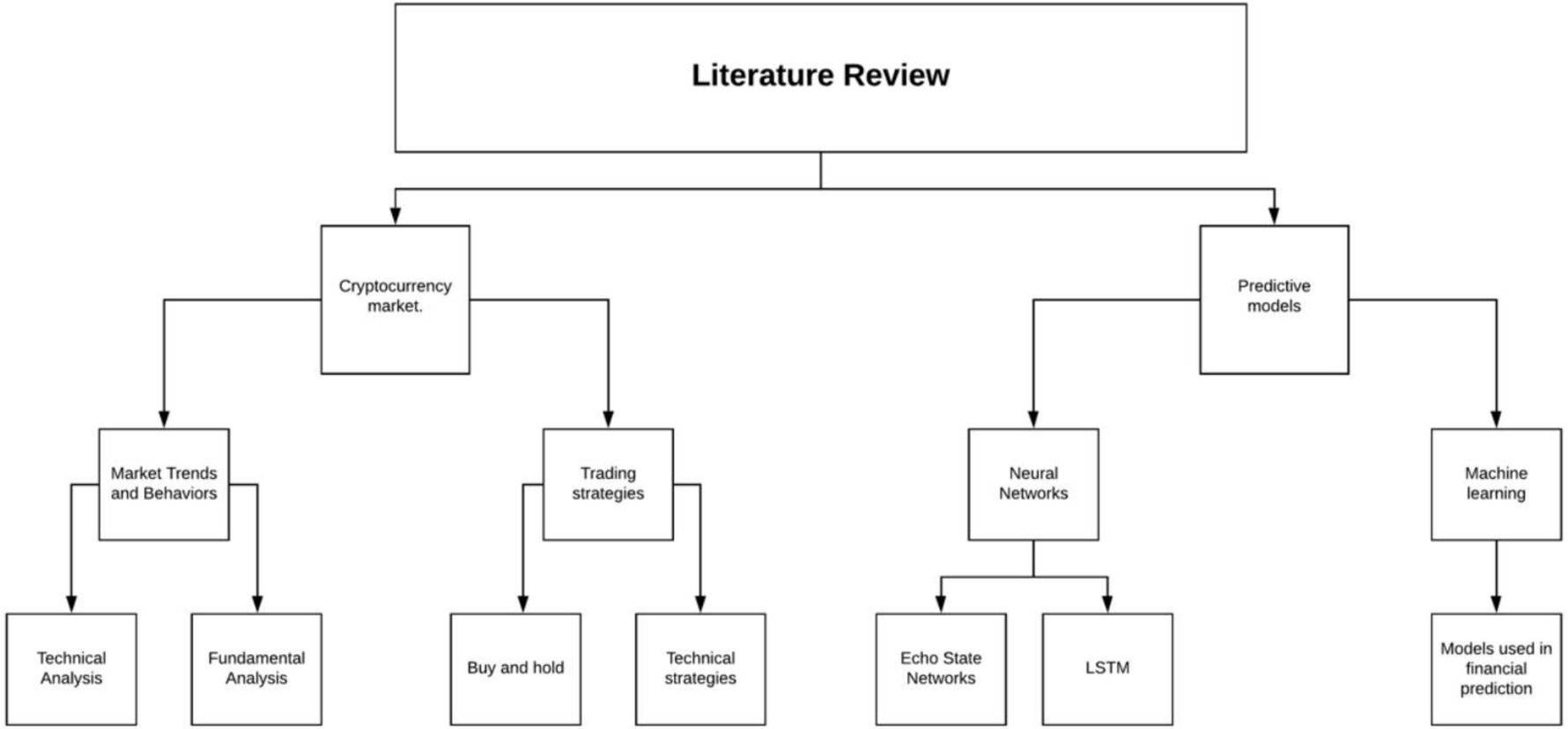
ML // Crypto Demand prediction

1.1.1 // Data collection

- This research is a collective set of data that is measurable by mathematical expressions and quantitative methods. The mathematical models will consist of multiple machine learning algorithms which will test the best accuracy and evaluation when forecasting the crypto-price.
- The data comprises of Bitcoin acquired from **Yahoo Finance API,** which is a platform which provides cryptocurrency markets data and insights. **From 27/10/2015 up to the 25/08/2020**
- The analysis will use a deductive reasoning approach on the secondary data from the cryptocurrency Bitcoin to develop a Quantitative predictive model, to predict the return price, for the **short (1-day), medium (5-day) and long (30-day) .**

ML // Crypto Demand prediction

1.1.2 // State of Art



ML // Crypto Demand prediction

2.1.1 // Data Preparation

The historical data is collected from Yahoo Finance for each currency consists of the daily features: **Open, Close, High, Low, Adjusted Close and Volume.**

Each of these features is used to **apply 80 of the most popular technical indicators**, which will then be explored and the cross-validation of the most key variables will be saved for prediction. These will be examined further and used within the machine learning models.

	TA-Lib code	Description
Overlapping	BBAND WIDTH, BBAND UPPER SIGNAL	Bollinger Bands
	EMA	Exponential Moving Average
	DEMA	Double Exponential Moving Average
	MIDPRICE	Midpoint Price over period
	SAR	Parabolic SAR
	SMA3, SMA5, SMA10, SMA20	Simple Moving Average
	TEMA	triple exponential moving average
	TRIMA	Triangular Moving Average
	WMA	Weighted Moving Average
Momentum indicators	ADX14, ADX20	Average Directional Movement Index
	APO	Absolute Price Oscillator
	DX	Directional Movement Index
	MACD, MACDSIGNAL, MACDHIST	Moving Average Convergence/Divergence
	MINUS_DI	Minus Directional Indicator
	MINUS_DM	Minus Directional Movement
	MOM1, MOM3, MOM5, MOM10	Momentum
	PLUS DI	Plus Directional Indicator
	PLUS DM	Plus Directional Movement
	PPO	Percentage Price Oscillator
	ROC	Rate of change
	ROCP	Rate of change Percentage
	ROCR	Rate of change ratio
	ROCR100	Rate of change ratio 100 scale
	RSI5, RSI10, RSI14	Relative Strength Index.
	SLOWK, SLOWD	Stochastic
	FASTK, FASTD	Stochastic Fast
	TRIX	1-day Rate-Of-Change (ROC) of a Triple Smooth EMA
Volatility indicators	ATR	Average True Range
	NATR	Normalized Average True Range
	TRANGE	True Range
Pattern recognition	CDLSHORTLINE	Short Line Candle
	CDLTASUKIGAP	Tasuki Gap
Cycle indicators	HT DCPERIOD	Hilbert Transform - Dominant Cycle Period HT

ML // Crypto Demand prediction

2.1.1 // Data Preparation

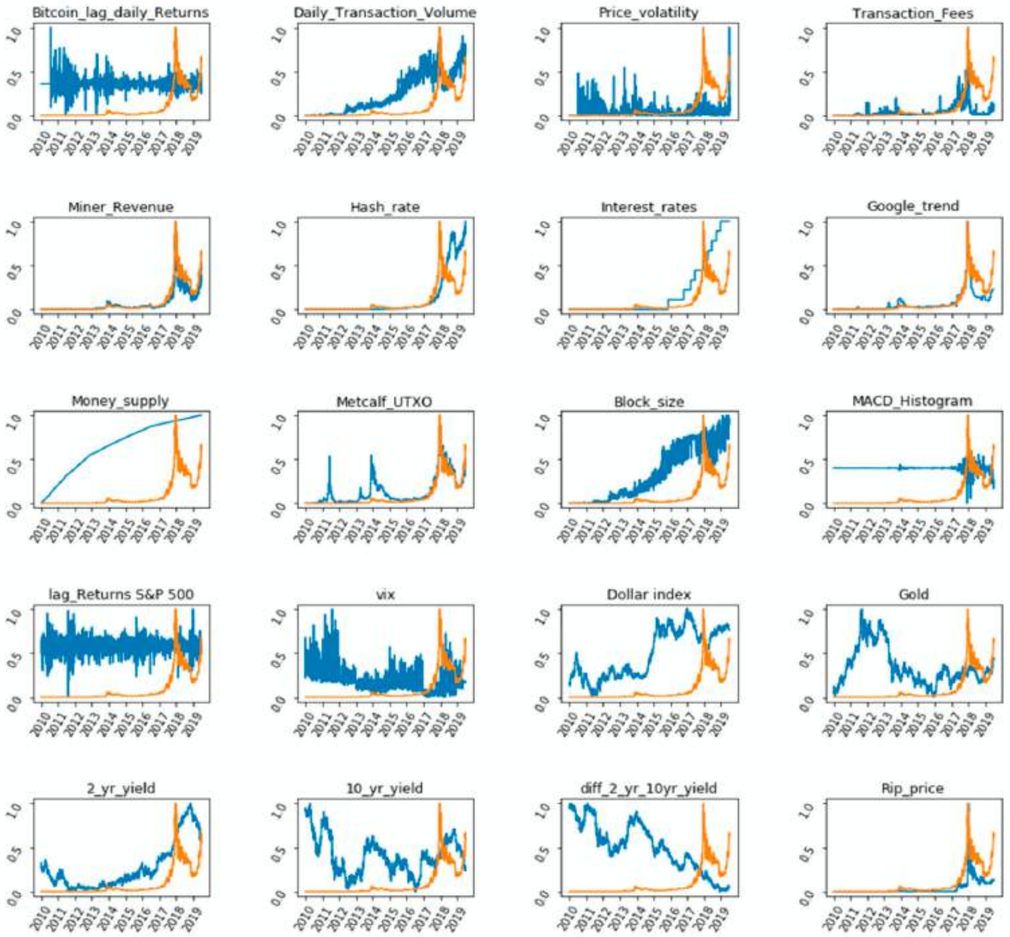


Table A1. Definition of variables and data source.

Features	Definition	Source
Bitcoin Price	Bitcoin prices.	https://charts.bitcoin.com/btc/
BTC Price Volatility	The annualized daily volatility of price changes. Price volatility is computed as the standard deviation of daily returns, scaled by the square root of 365 to annualize, and expressed as a decimal.	https://charts.bitcoin.com/btc/
BTC Miner Revenue	Total value of Coinbase block rewards and transaction fees paid to miners. Historical data showing (number of bitcoins mined per day + transaction fees) * market price.	https://www.quandl.com/data/BCHAIN/MIREV-Bitcoin-Miners-Revenue
BTC Transaction Volume	The number of transactions included in the blockchain each day.	https://charts.bitcoin.com/btc/
Transaction Fees	Total amount of Bitcoin Core (BTC) fees earned by all miners in 24-hour period, measured in Bitcoin Core (BTC).	https://charts.bitcoin.com/btc/
Hash Rate	The number of block solutions computed per second by all miners on the network.	https://charts.bitcoin.com/btc/
Money Supply	The amount of Bitcoin Core (BTC) in circulation.	https://charts.bitcoin.com/btc/
Metcalfe-UTXO	Metcalfe's Law states that the value of a network is proportional to the square of the number of participants in the network.	https://charts.bitcoin.com/btc/
Block Size	Miners collect Bitcoin Core (BTC) transactions into distinct packets of data called blocks. Each block is cryptographically linked to the preceding block, forming a 'blockchain.' As more people use the Bitcoin Core (BTC) network for Bitcoin Core (BTC) transactions, the block size increases.	https://charts.bitcoin.com/btc/
Google Trends	This is the month-wise Google search results for the Bitcoins.	https://trends.google.com
Volatility (VIX)	VIX is a real-time market index that represents the market's expectation of 30-day forward-looking volatility.	http://www.cboe.com/products/vix-index-volatility/vix-options-and-futures/vix-index/vix-historical-data
Gold price Level	Gold price level.	https://www.quandl.com/data/WGC/GOLD_DAILY_USD-Gold-Prices-Daily-Currency-USD
US Dollar Index	The U.S. dollar index (USDX) is a measure of the value of the U.S. dollar relative to the value of a basket of currencies of the majority of the U.S.' most significant trading partners.	https://finance.yahoo.com/quote/DX-Y.NYB/history?period1=1262332800&period2=1561878000&interval=1d&filter=history&frequency=1d
US Bond Yields	2-year / short-term yields.	https://www.quandl.com/data/USTREASURY/YIELD-Treasury-Yield-Curve-Rates
US Bond Yields	10-year/ long term yields.	https://www.quandl.com/data/USTREASURY/YIELD-Treasury-Yield-Curve-Rates
US Bond Yields	Difference between 2 year and 10 year/ synonymous with yield inversion and recession prediction	https://www.quandl.com/data/USTREASURY/YIELD-Treasury-Yield-Curve-Rates
MACD	MACD=12-Period EMA −26-Period EMA. We have taken the data of the MACD with the signal line. MACD line = 12-day EMA Minus 26-day EMA Signal line = 9-day EMA of MACD line MACD Histogram = MACD line Minus Signal line	
Ripple Price	The price of an alternative cryptocurrency.	https://coinmarketcap.com/currencies/ripple/historical-data/?start=20130428&end=20190924
One Day Lagged S&P 500 Market Returns	Stock market returns.	https://finance.yahoo.com/quote/%5EGSPC/history?period1=1230796800&period2=1568012400&interval=1d&filter=history&frequency=1d
Interest Rates	The federal funds rate decide the shape of the future interest rates in the economy.	http://www.fedprimerate.com/fedfundsrate/federal_funds_rate_history.htm#current

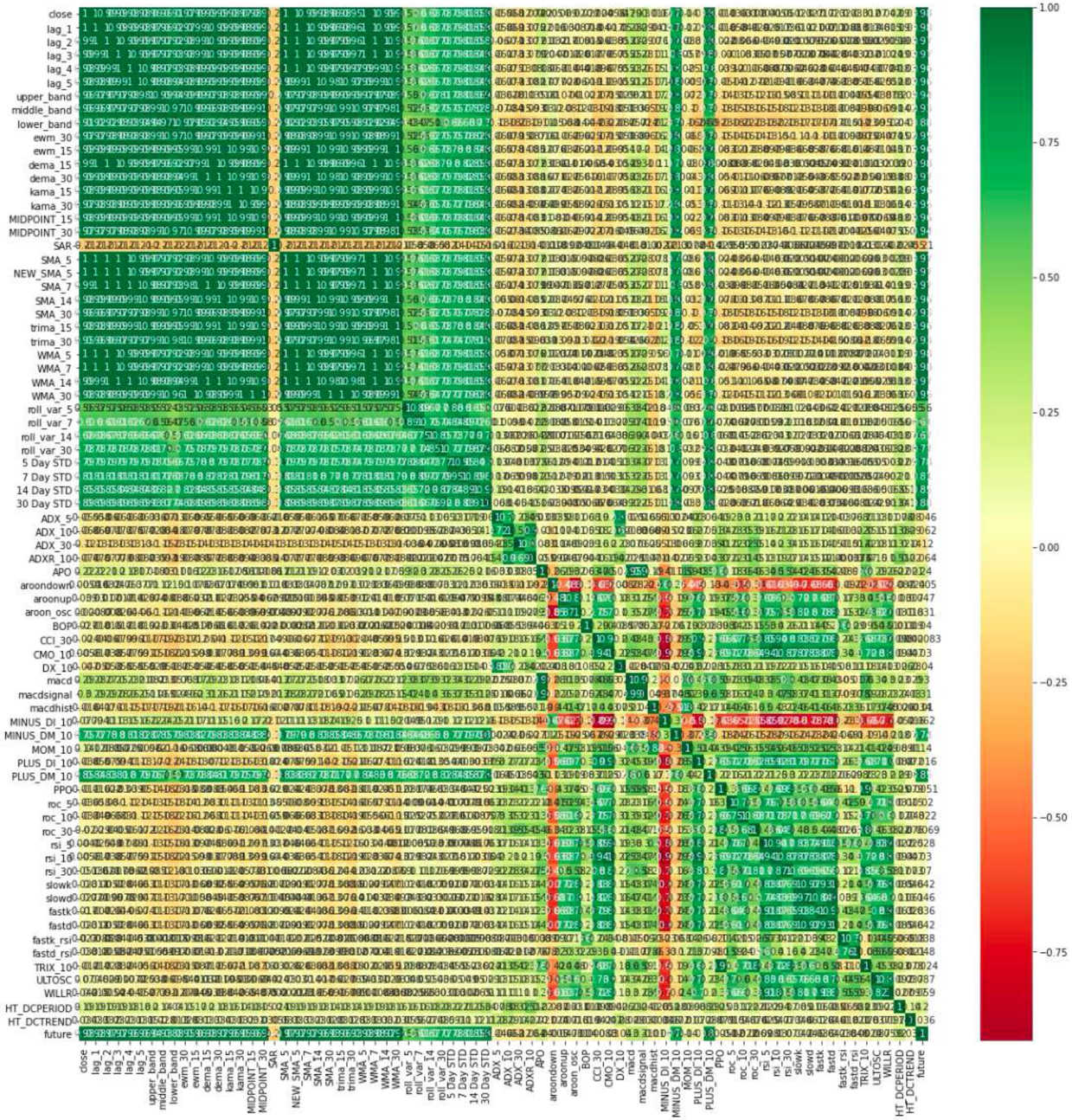
ML // Crypto Demand prediction

2.1.2 // Feature selection

80 Technical indicators

With **‘close price’** as the main factor, correlation analysis was complete to reduce the number of variables that would be entering the model and therefore reduce noise.

We just will work with those features with a clear linear correlation (> 0.7) with our target variable: “close” (green or red intense)

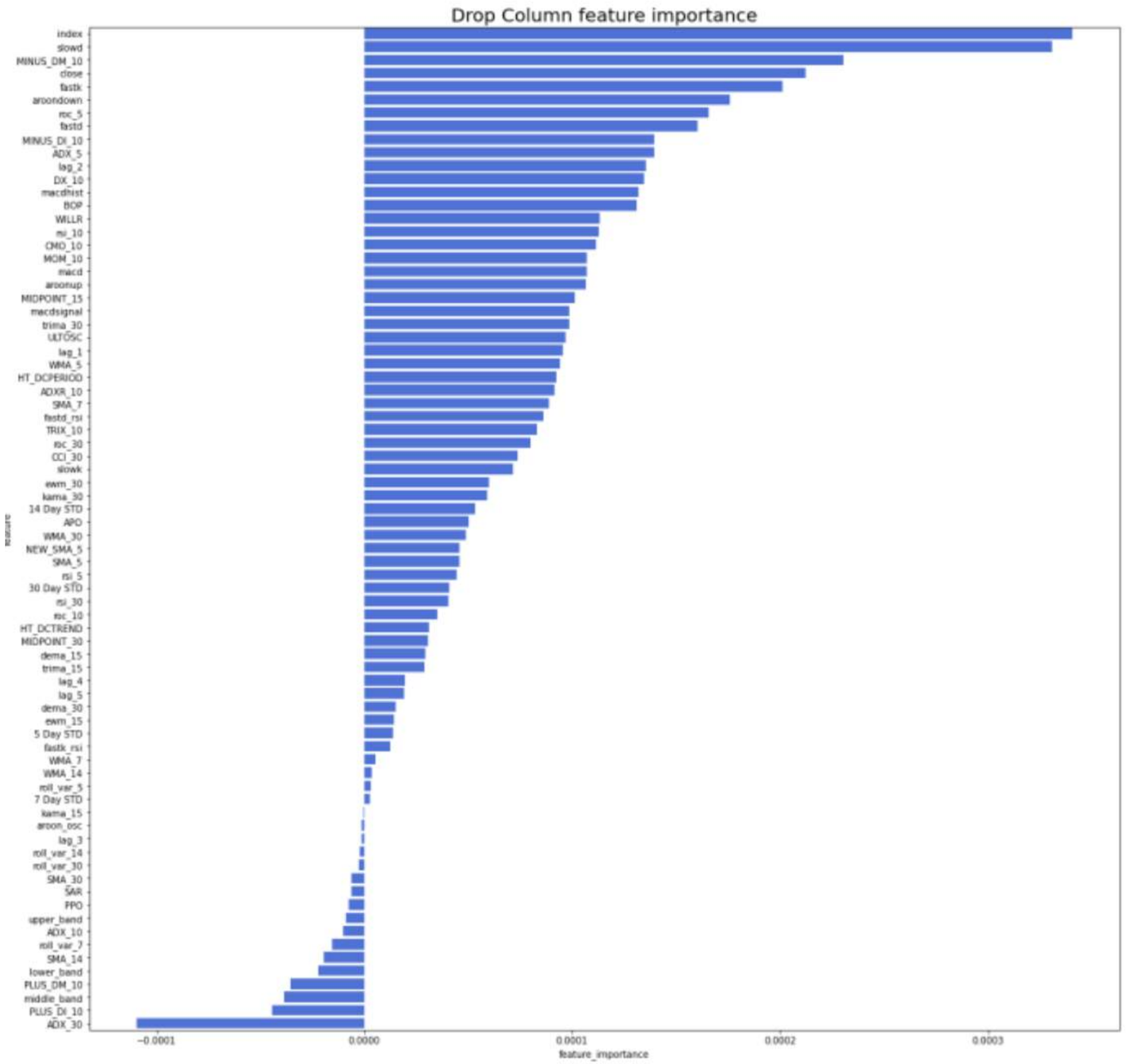


ML // Crypto Demand prediction

2.1.3 // Feature importance

Using a random forest regressor model as a baseline to **analyze the impact of features on the model produces** interesting results

Negative importance in this instance means that removing a given feature from the model will improve its performance



ML // Crypto Demand prediction

2.1.4 // Feature extraction

80 Technical indicators => 36 features

Feature importance rank table, that for BTC, the top-ranking indicator, other than closing, is t**he simple moving average of 5 day**s

Feature importance rank table, are the features which will be used in the machine learning and deep learning model, these features will not be used in the Echo State network, due to its strength as a one-dimensional modelling tool.

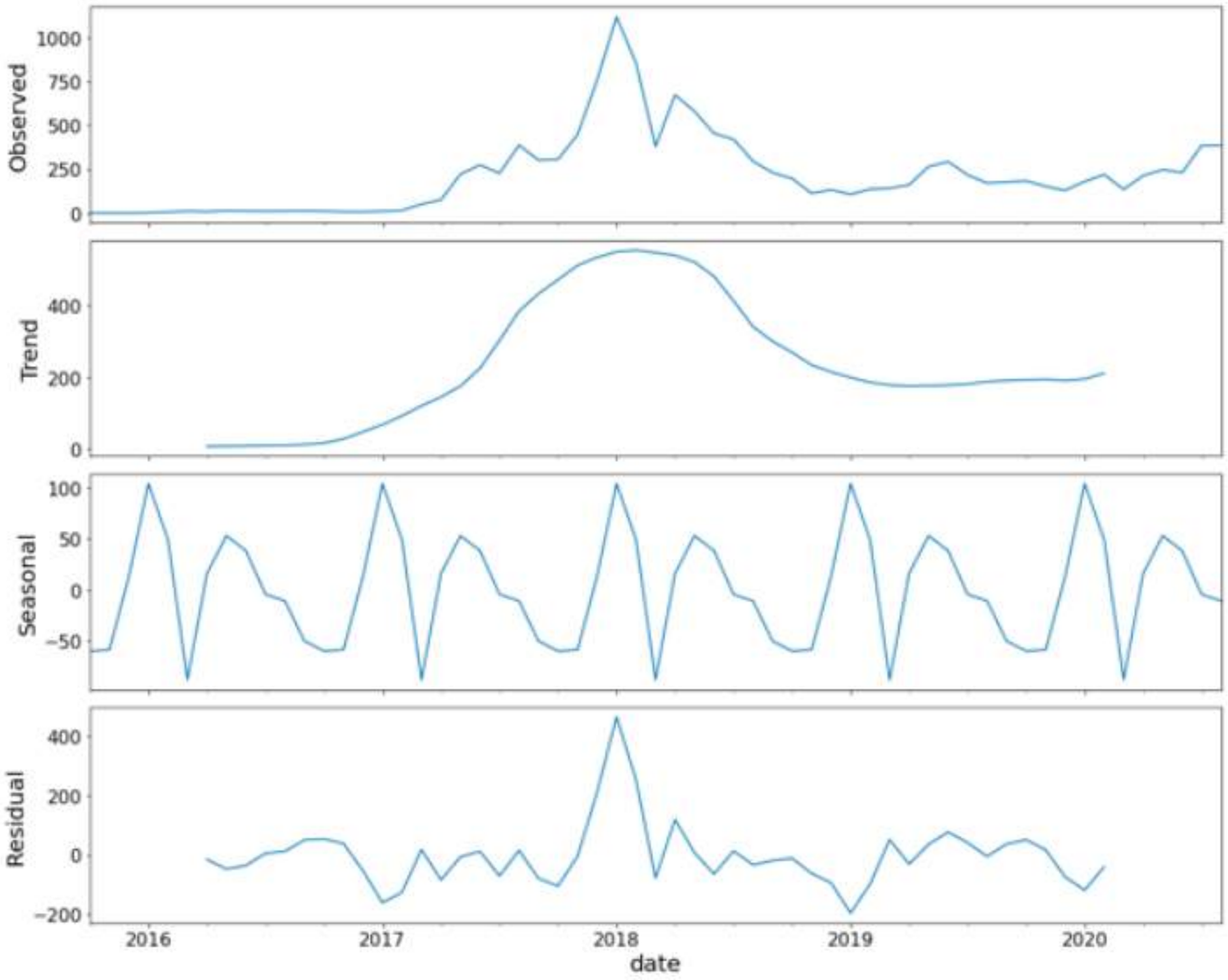
Indicator	Description	BTC - rank	ETH - rank
Std_Dev	Standard deviations, of [5, 7, 14, 30] days	[38,33,44,43]	[38,35, 19, 20]
ADX	The average directional movement over [5, 10, 30] days	[- ,- , 57]	[2, 50, 8]
ADXR_10	Average directional movement index rating over 10 days	[51]	[11]
APO	Absolute price oscillator	45	16
aroon_osc	Aroon Oscillator, [overall, upward and downward] motion	[13 , 36, 15]	[9, 7,5]
BOP	Balance of power	47	3
CCI_30	Commodity Channel index	7	6
CMO_10	Change Momentum Oscillator	9	4
DEMA	Double Exponential Moving Average [15, 30 day]	[56, 55]	-
DX_10	Directional Movement Index	52	15
EWM	Exponentially weighted moving average [15 days]	53	-
slow/fast[d, k]	Stochastic oscillators	[37, 50]	[33,39]
fastd_rsi	Stochastic RSI	[5,14]	[31,36]
HT_DCPERIOD	Hilbert Transform - Dominant Cycle Period	-	45
Kama	Kaufman Adaptive Moving Average (30)	54	-
lag_3	Closing from 3 days previous	-	46
Lower_B	Bollinger bands - lower band	-	47
MACD	Moving Average Convergence Divergence	12	10
MACD_hist	Moving average history over the slow period of 30 days with a signal period of 5 days.	49	21
MACD_signal	Directional signal of MACD	20	1
Midpoint	Midpoint over period. [30 days]	28	52
MINUS_DI_10	Minus Directional Indicator	26	26
MINUS_DM_10	Minus Directional Movement	23	12
MOM_10	Momentum	8	17
PLUS_DI_10	Plus Directional Indicator	11	43
PLUS_DM_10	Plus Directional Movement	2	18
PPO	Percentage Price Oscillator	6	27
ROC	Rate of change :	[21, 18, 25]	[32, 14, 24]
Roll_var	Rolling variable of [5, 7, 14, 30]days	[41, 40, 18, 25]	[49, 44, 41, 37]
RSI	Relative Strength Index [5, 10, 30]	[31, 46, 34]	[22, 13, 28]
SMA	Simple Moving Average [5,7,14,30]	[1, 48,32,30]	[56,59,- ,55]
TRIMA	Triangular Moving Average over 30 days	29	-
TRIX	1-day Rate-Of-Change (ROC) of a Triple Smooth EMA of 10 days	16	40
ULTOSC	Ultimate Oscillator	19	25
WILLR	Williams' %R	23	-
WMA	Weighted moving average	[16, 29, 17, 29]	[53, 54, 58, 48]

2.1.5 // Exploratory Data Analysis

Seasonality of BTC

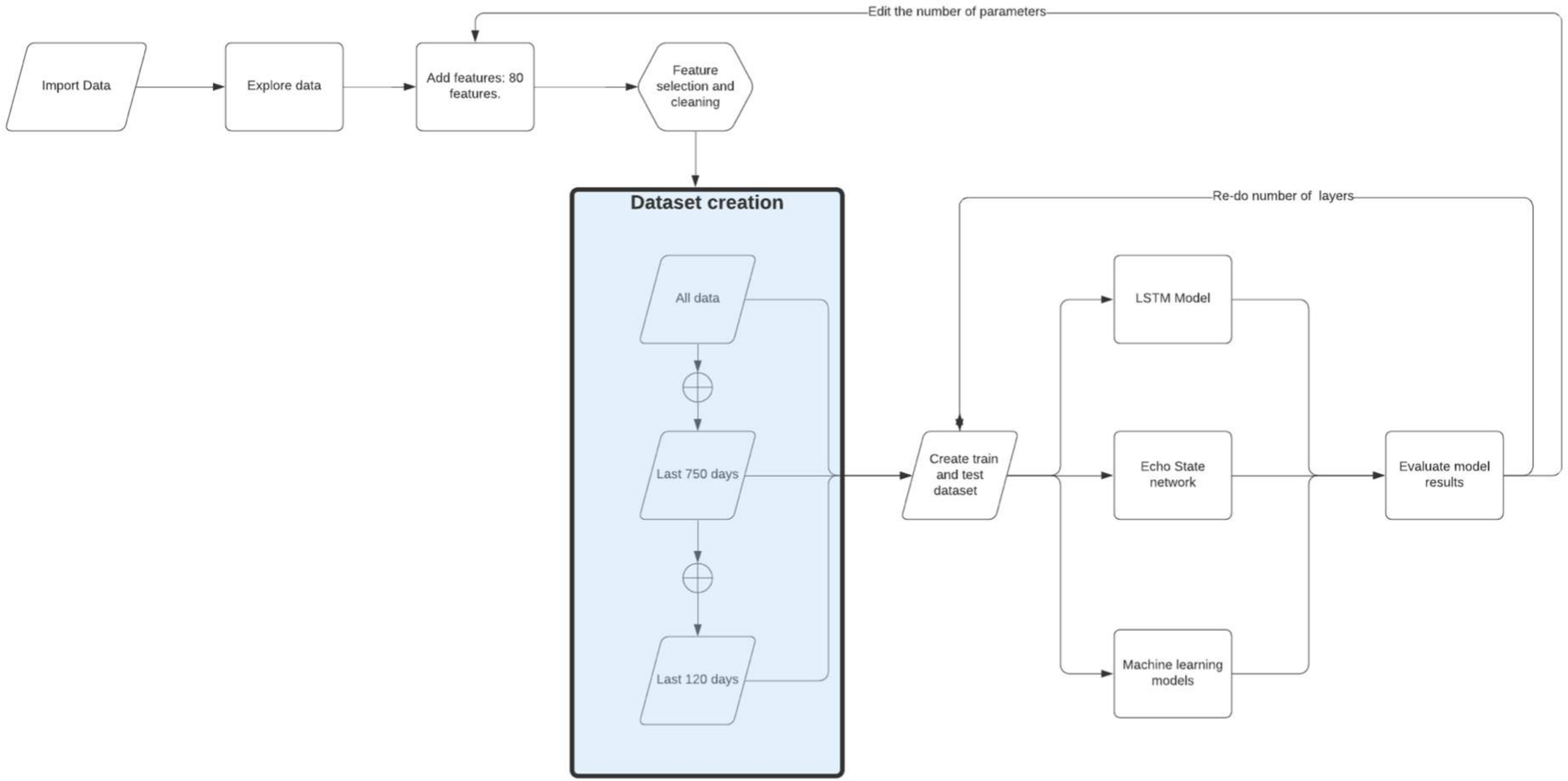
BTC is non-stationary. It can be seen from Figure 4.4 Rolling means smoothing (5-day) [BTC, ETH] and Figure 4.5 Rolling mean smoothing (30-day) [BTC, ETH], that the rolling mean is not an appropriate method to transform the dataset, this is due to how rapidly cryptocurrencies can change within one month.

As an example, on 20th March 2020 Bitcoin was worth $6,483.74, by 20th April the value has risen to $8,773.11.



ML // Crypto Demand prediction

3.1 // Choosing between models



ML // Crypto Demand prediction

4.1 // Training

ML // Crypto Demand prediction

5.1 // Evaluation

This study will present findings from several predictive models will be created, a comparison of these models will be made using **MSE and an accuracy calculation R^2**

	Test	BTC_MAE	BTC_MSE	BTC_RMSE	BTC_Score (R^2)
1 day	Linear regression	224.445225	130703.372	361.529213	0.94585718
	Bayesian Ridge	235.626552	131372.13	362.452936	0.94558015
	Elastic Net CV	533.21515	419161.78	647.427049	0.8263656
	Random Forest	263.323552	156112.453	395.110685	0.93533168
	XGBoost regression	309.834892	212341.594	460.805376	0.91203921
	NN regression	1821.77864	5683007.05	2383.90584	-1.3541398

	Test	BTC_MAE	BTC_MSE	BTC_RMSE	BTC_Score (R^2)
5 day	Linear regression	707.0275	880569.888	938.386854	0.63682719
	Bayesian Ridge	699.328325	884022.489	940.224701	0.63540324
	Elastic Net CV	693.468395	792622.69	890.293598	0.67309919
	Random Forest	719.96298	870510.72	933.01164	0.64097589
	XGBoost regression	686.251131	863182.218	929.076002	0.64399838
	NN regression	1777.81479	4719417.23	2172.42197	-0.9464259

ML // Crypto Demand prediction

5.1 // Evaluation

This study will present findings from several predictive models will be created, a comparison of these models will be made using **MSE and an accuracy calculation R^2**

	Test	BTC_MAE	BTC_MSE	BTC_RMSE	BTC_Score (R^2)
30 day	Linear regression	2023.01004	5773491.37	2402.80906	-1.308744
	Bayesian Ridge	1888.14072	5256411.69	2292.68657	-1.1019706
	Elastic Net CV	1743.14349	4577157.47	2139.42924	-0.8303457
	Random Forest	1841.44096	4563210.46	2136.16724	-0.8247684
	XGBoost regression	1884.94446	5512659.52	2347.90535	-1.2044408
	NN regression	3531.84716	17536760.2	4187.69151	-6.012722

ML // Crypto Demand prediction

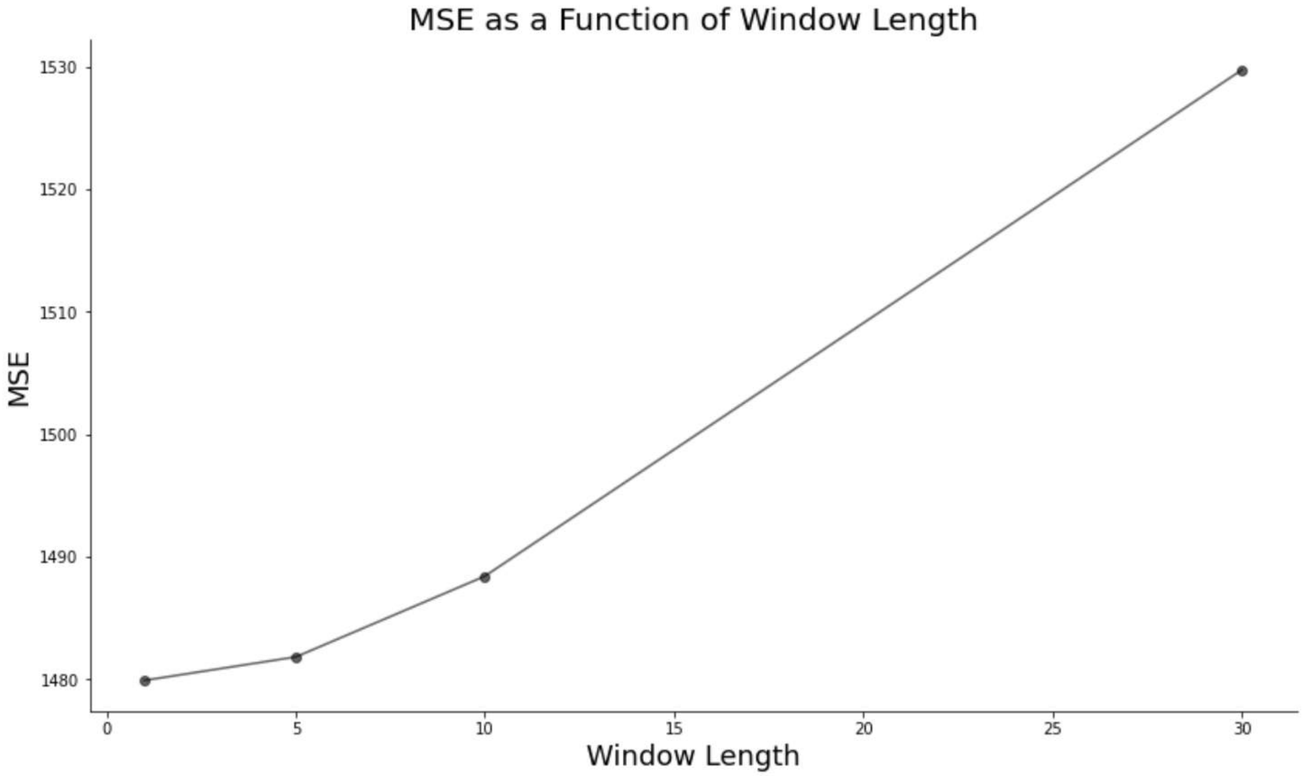
5.2 // Evaluation

This research confirms that the **chaotic data** although can have positive results for short- and medium-term prediction, for long-term prediction, technical analysis based- prediction is not sufficient.

ML // Crypto Demand prediction

6.1 // Parameter tuning

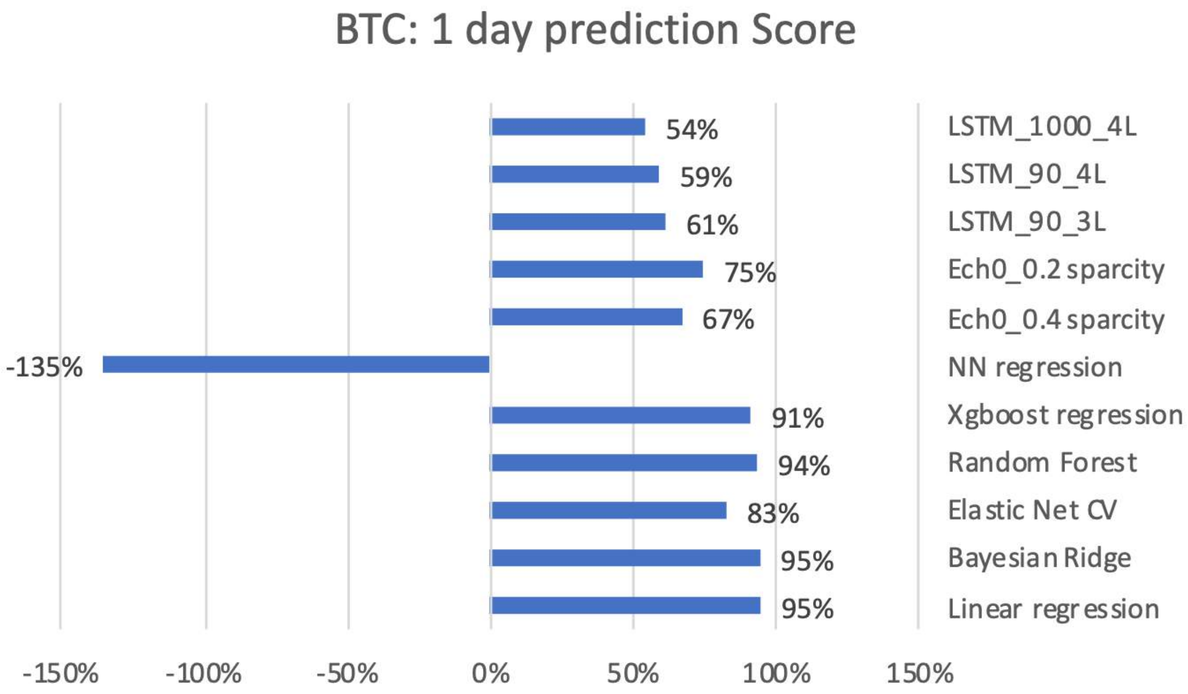
We tried several methods to optimize the models and get better accuracy. But the more you increase time window length the more the error still gets bigger.

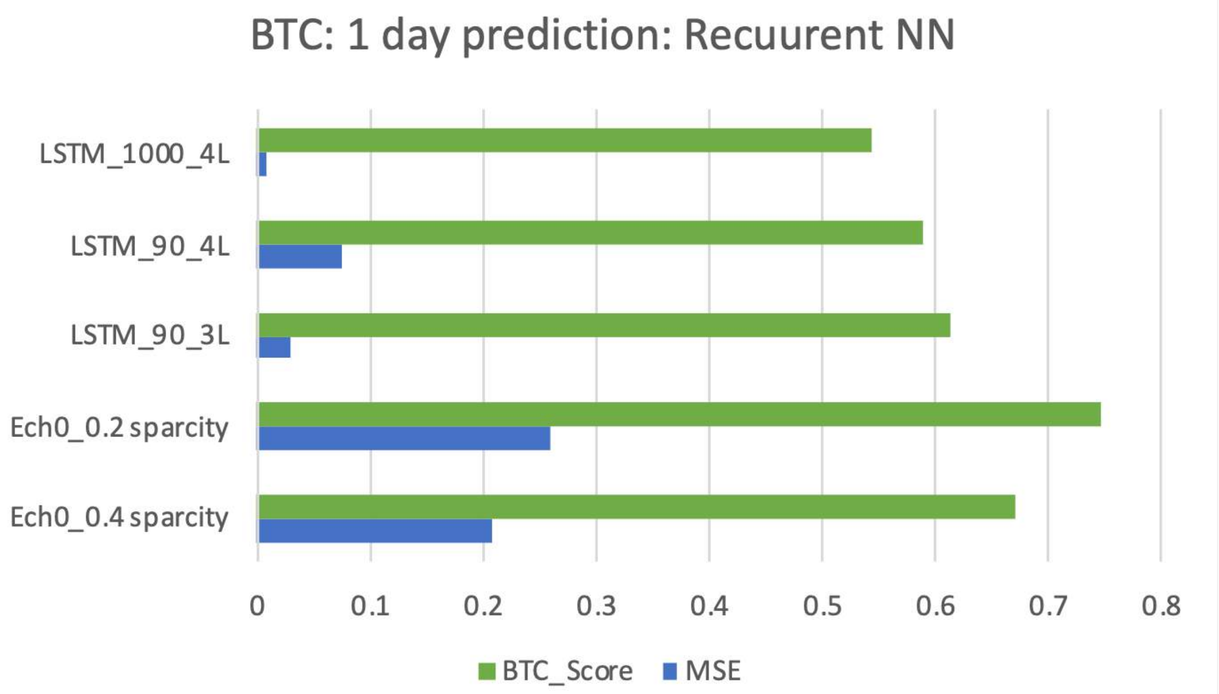


ML // Crypto Demand prediction

7.1 // Predictions

1 day predictions

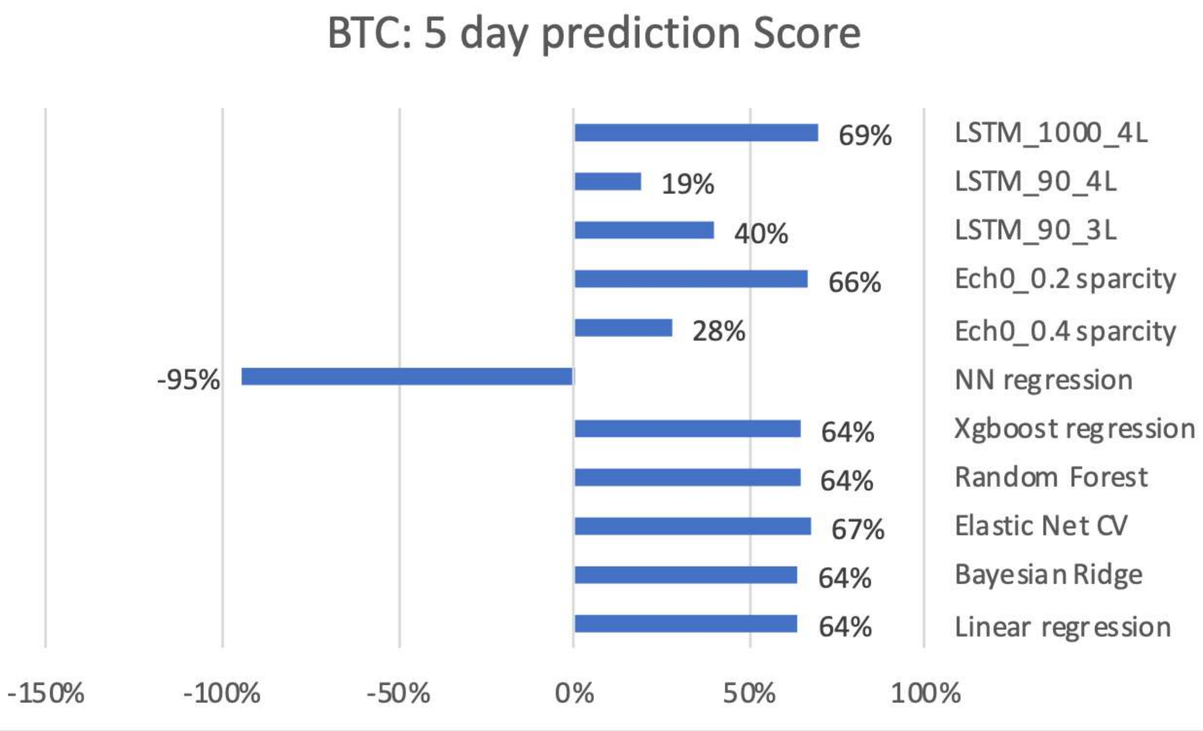


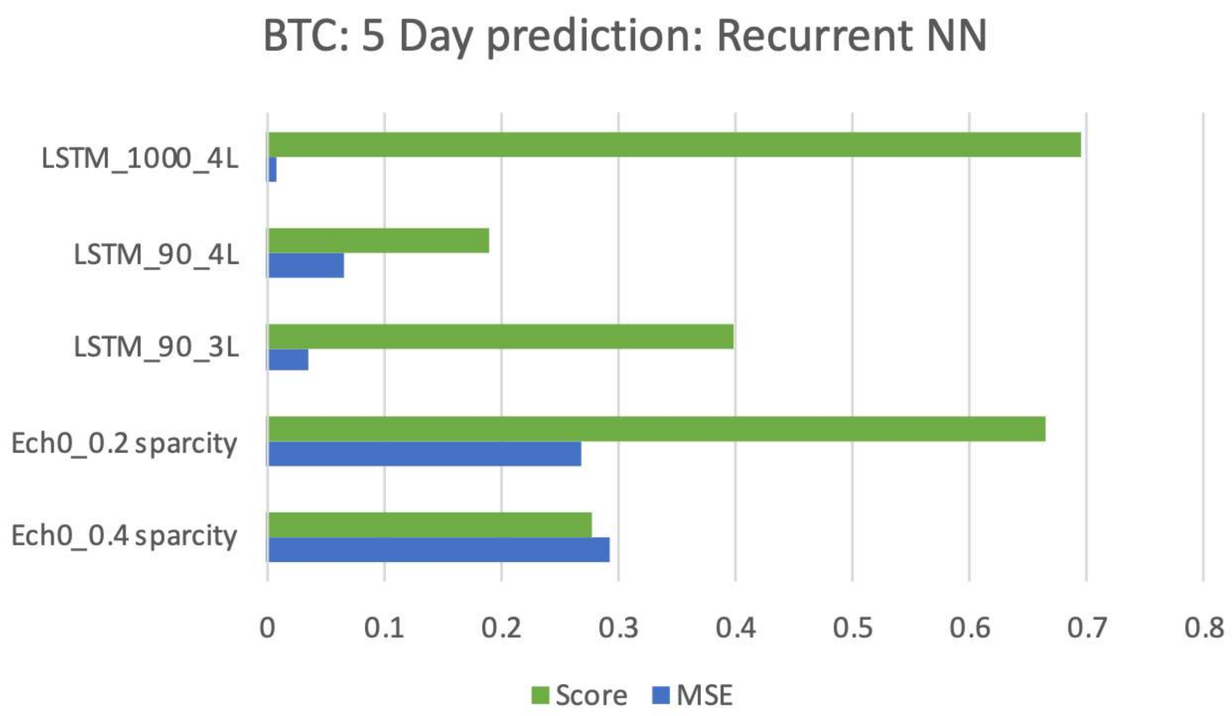


ML // Crypto Demand prediction

7.1 // Predictions

5 days predictions

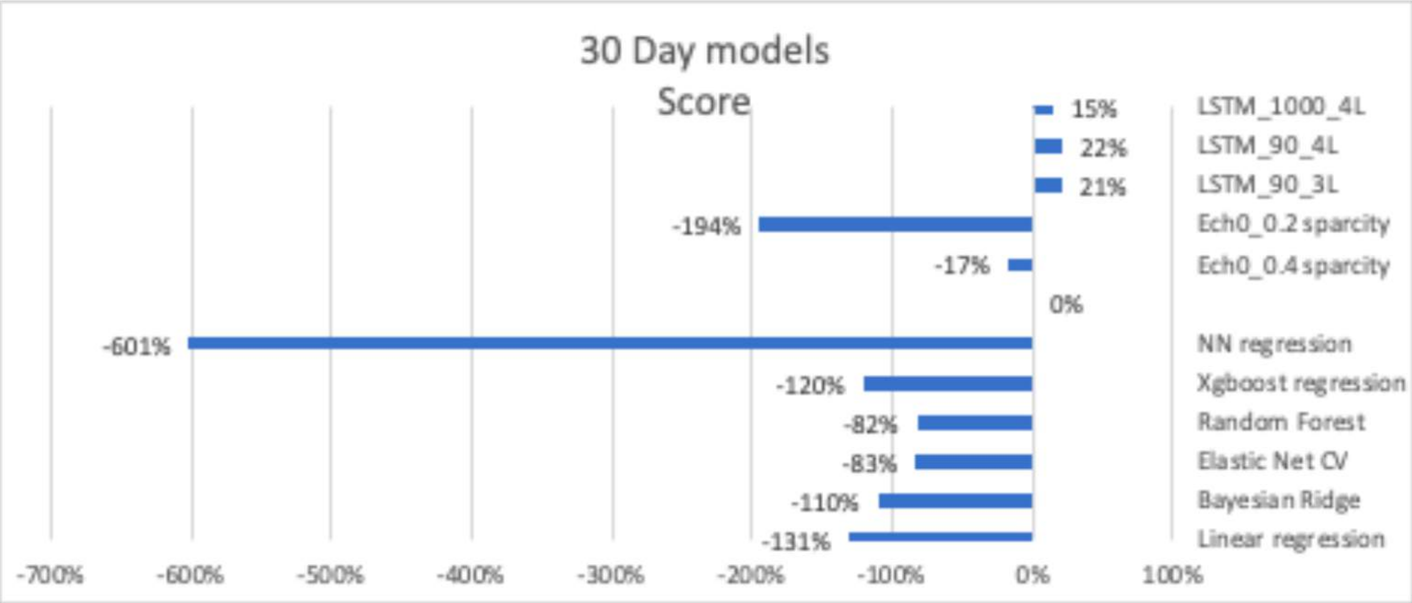


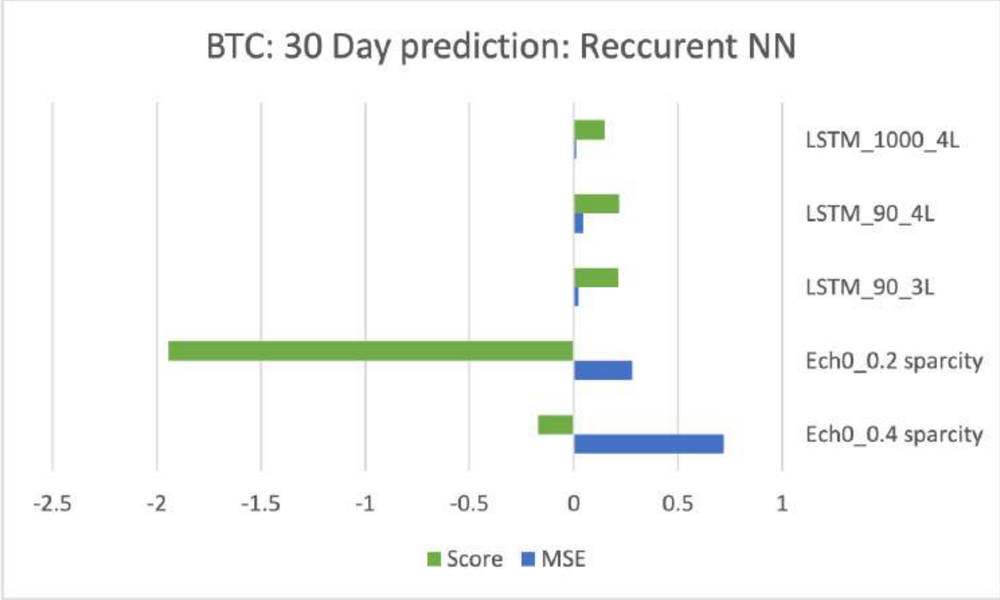


ML // Crypto Demand prediction

7.1 // Predictions

30 days predictions





8 // Conclusions

Using exploratory analysis to provide the technical indicators and tuning the hyperparameters within each network, can the short-, medium- and long-term returns of Bitcoin be predicted with only using technical analysis?

Null Hypothesis: Bitcoin, cannot be predicted with Machine learning and Neural Network models, to a degree of accuracy for short-, medium-(closing price of the week) and long-term(closing price of the month) price.

Alternate Hypothesis: Bitcoin short-, medium- and long-term direction of pricing, can be predicted by only using technical analysis with Machine learning and Neural Network.

This research confirms that the **chaotic data** although can have positive results for short- and medium-term prediction, for long-term prediction, technical analysis based- prediction is not sufficient.

Why?

1. Whales

ML // Crypto Demand prediction // **whales**

The whale metaphor holds well—these are the largest holders (holders, in bitcoin-speak) of BTC, and most of them **surface from time to time.**

At the other end of the spectrum, the Winklevoss twins, Tyler and Cameron, live in the limelight, and through their well-respected Gemini cryptocurrency exchange, specialize in the trading of digital currency, both bitcoin and altcoin, for institutional investors. They are probably the most famous crypto investors, especially after the purchase of **$11 million in bitcoin in 2013** following a successful lawsuit against Mark Zuckerberg and Facebook. It's estimated that at one time, the Winklevoss twins owned 1% of all bitcoin mined.

🐋 Humpback Whale = 5000 BTC or more
🐳 Whale = 1000 - 5000 BTC
🦈 Shark = 500 - 1000 BTC
🐬 Dolphin = 100 - 500 BTC
🐟 Fish = 50 - 100 BTC
🐙 Octopus = 10 - 50 BTC
🦀 Crab = 1 - 10 BTC
🦐 Shrimp = less than 1 BTC

- Pantera Capital
- Bitcoins Reserve
- Binary Financial
- Coin Capital Partners
- Falcon Global Capital
- Fortress
- Bitcoin Investment Trust
- Global Advisors Bitcoin Investment Fund

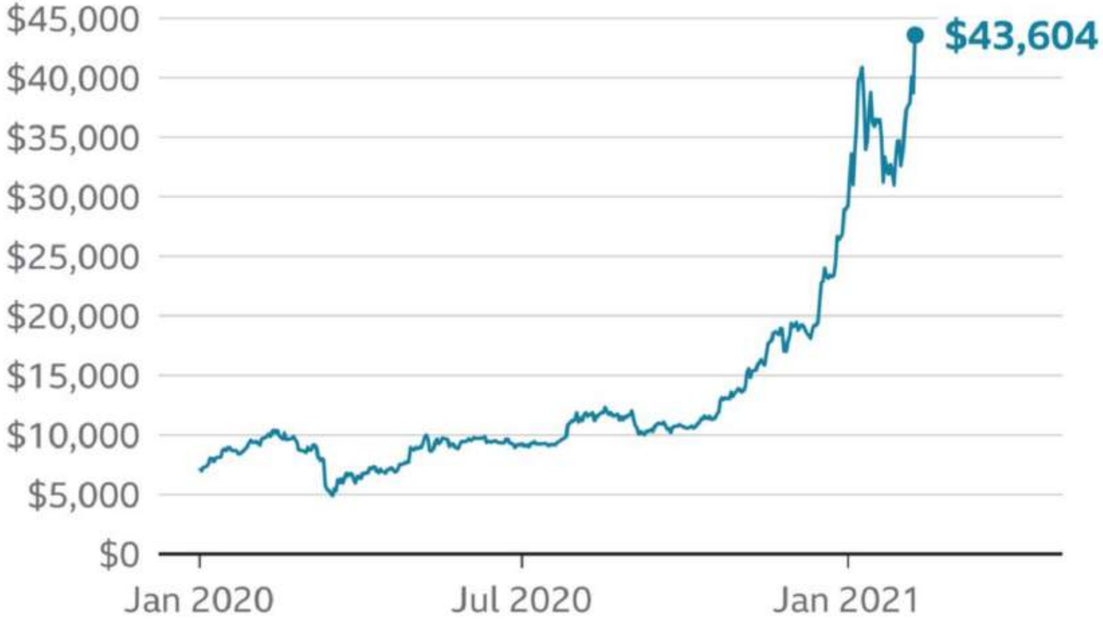
ML // Crypto Demand prediction // **whales**

Elon Musk's Tesla buys $1.5bn of Bitcoin causing currency to spike

8 February | Comments

Bitcoin jumps as Tesla buys currency

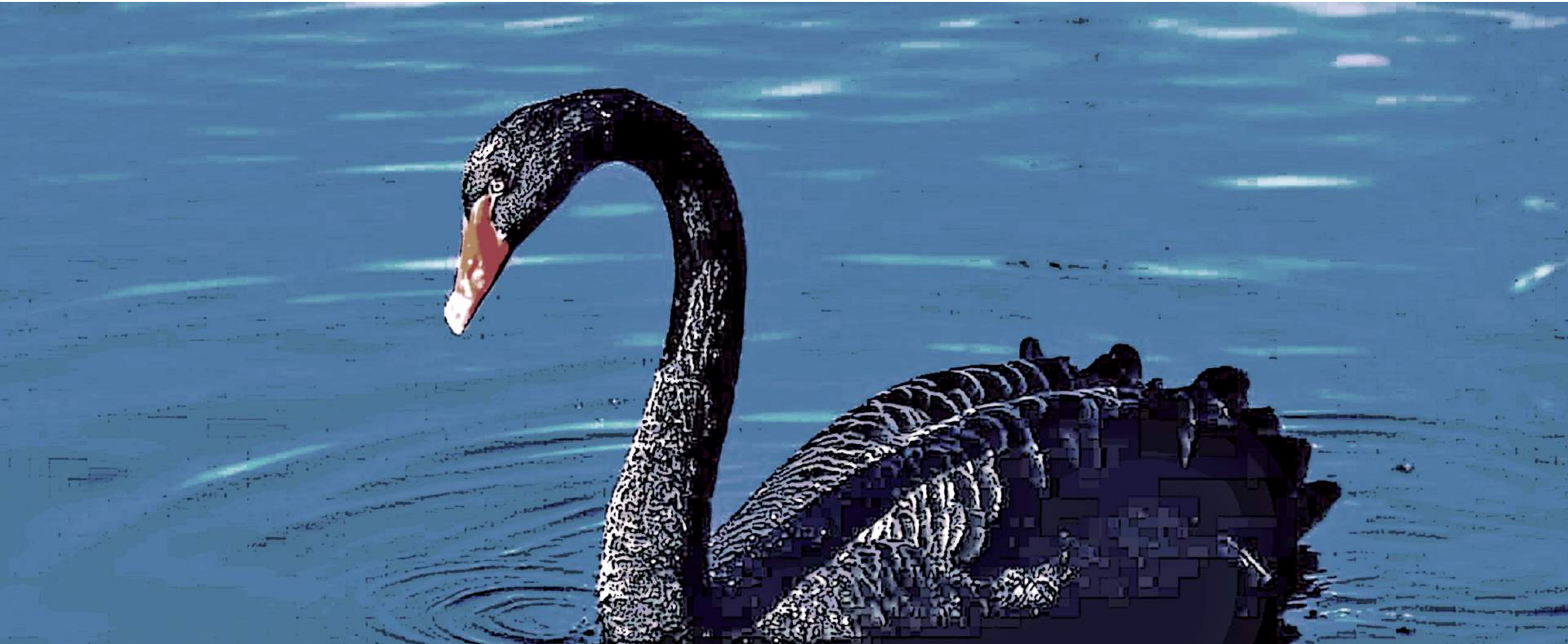
Tesla purchased $1.5bn of Bitcoin and announced it will start accepting the cryptocurrency as payment



Source: Bloomerg. Last update 08 February 2021, 15:00 GMT

BBC

1. Black Swan



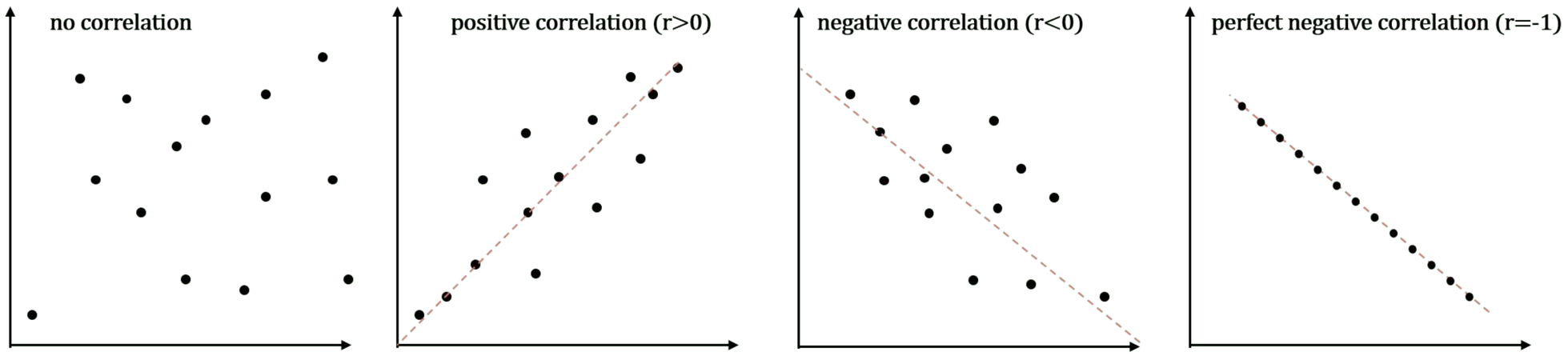
ML // Crypto Demand prediction // **black swan**

That unexpected event is what Taleb calls a **black swan**, and it comes from the fact that explorers in the 14th century believed all swans were white. One day, however, they came upon a black swan and that worldview was shattered.
Black swans can be both positive and negative, but a key point to remember is that positive black swans take time to show their effect while negative black swans can happen very rapidly. This can no doubt be seen with cryptocurrencies as well.

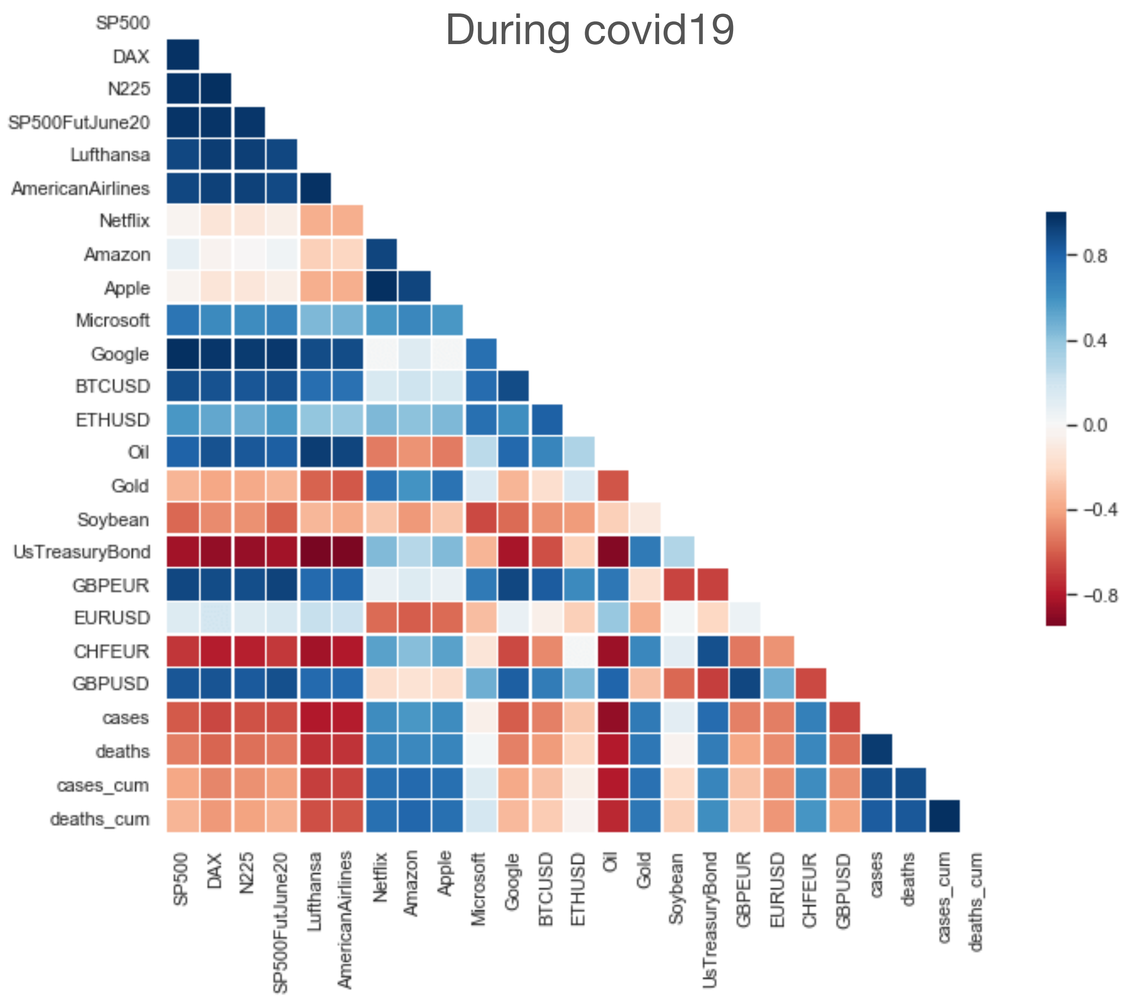
The Turkey Fallacy is essentially this: **Things work until they don't**. What we learn from the past turns out to be either irrelevant or completely false, and at worst looking at the past as an example of the future can be incredibly misleading.

For instance, online services **such as Amazon or Netflix have seen a sharp rise in the number of customers and have been able to increase their ratings**. But how strong are these developments correlated? In this blog post I will try to answer this question by creating a correlation matrix on COVID-19 and different financial assets.

How is the correlation coefficient interpreted?



ML // Crypto Demand prediction // **black swan**



Stock Market Indices / COVID-19

Let us start with the pairs of Stock market indices and COVID-19 data. As one could expect, the heatmap signals a negative correlation between the indices (DAX, S&P500, NIKI) and COVID-19. Consequently, when the number of cases rises, stock market indices tend to fall in value. If we look exactly, the total number of new cases seems to be more correlated than the number of total cases (cases_cum) or deaths (deaths_cum). In addition, one can observe that the stock market indices are correlated among each other.

Stock Market Indices / Online Service Provider Stocks

When we compare the stock markets with the shares of online service providers, the situation is heterogen. The shares of Microsoft and Google are positively correlated with the overall development of the markets. On the other hand, the shares of Netflix, Amazon and Apple are hardly correlated with the market development.

Stock Market Indices / Airline Stocks

There is a strong positive correlation between airline stocks and the stock market indices. This means, that airline stocks tend to follow the general market direction. This can also be attributed to the fact that airlines are particularly affected by the pandemic.

Stock Market Indices / Crypto-Currencies

Next, we compare Cryptocurrencies with the stock market indices. The results are surprising. BTC-USD correlates surprisingly strong positive with the general development of the stock markets. For ETH-USD and the markets, however, the correlation is only slightly positive.

COVID-19 / Currency Exchange Rates

The correlation between exchange rates and COVID-19 cases is not uniform. Of the four exchange rates considered, GBP/EUR and EUR/USD and GBP/USD show a slightly negative correlation. An exception is CHF/EUR, which shows a positive correlation to COVID-19 cases. This means, with the number of COVID-19 cases, the value of the Swiss Franc has increased against the EURO.

Treasury Bonds / Resources

Let us now look at the coefficients of resources and US Treasury Bonds. There is a strong negative correlation with the oil price and a strong positive correlation with the gold price.

Crypto-Currencies / Resources

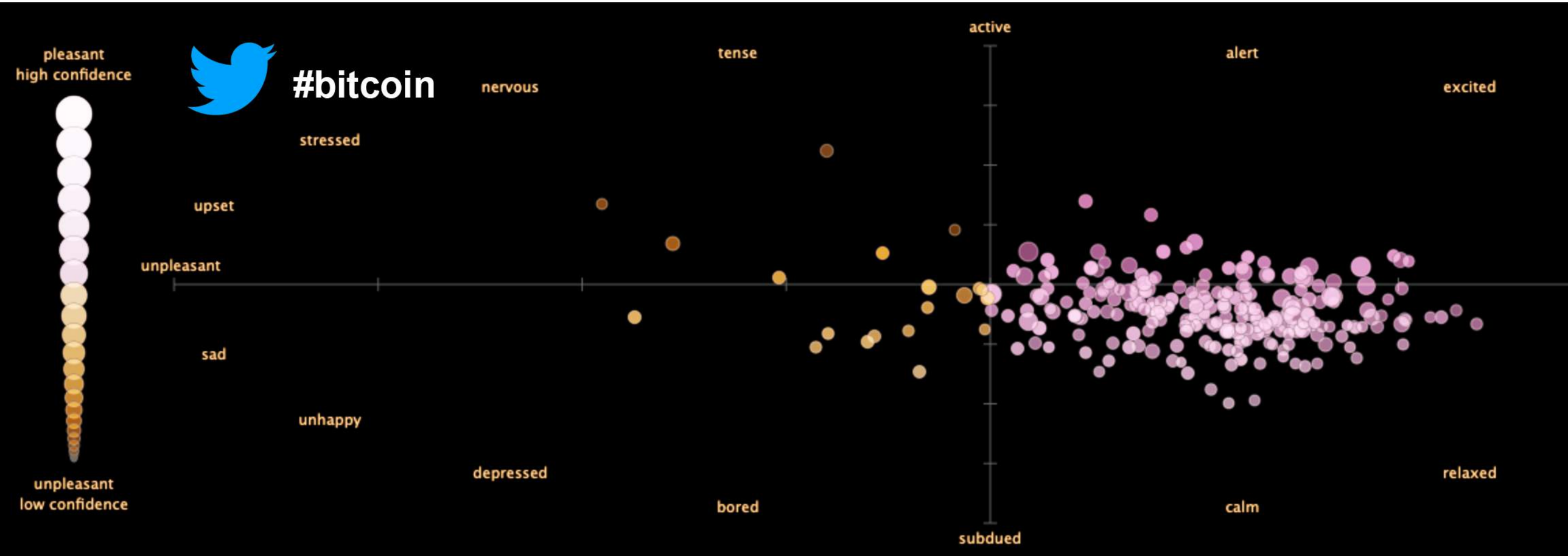
Finally, let us consider the coefficients of resources and cryptocurrencies. The only thing that is noticeable here is that BTCUSD correlates with the oil price. Based on the absence of a correlation with gold, it can be said that BTC-USD is not a comparable crisis currency and follows different rules. ETH-USD, on the other hand, is relatively uncorrelated with other resources.

Wars, Viruses, and Excessive Valuations

Event	Start of Sell-off/Previous Peak	Size of Sell-off	Duration of Sell-off (Trading Days)	Duration of Recovery (Trading Days)
Israel Arab War/Oil Embargo	October 29, 1973	-17.1%	27	1475
Iranian Hostage Crisis	October 5, 1979	-10.2%	24	51
Black Monday	October 13, 1987	-28.5%	5	398
First Gulf War	January 1, 1991	-5.7%	6	8
9/11 Attacks	September 10, 2001	-11.6%	6	15
SARS	January 14, 2003	-14.1%	39	40
Global Financial Crisis	October 9, 2007	-56.8%	356	1022
Intervention in Libya	February 18, 2011	-6.4%	18	29
Brexit Vote	June 8, 2016	-5.6%	14	9
COVID-19*	February 19, 2020	-29.5%	19	N/A (ongoing)

ML // Crypto Demand prediction // **twitter**

1. Sentiment analysis



ML // Crypto Demand prediction // **twitter**

Before Musk did anything, the value of bitcoin was roughly $ 32,200. By changing his 'bio' on Twitter and putting "#bitcoin" the purchase fever was unleashed and in just one hour that value has increased $ 5,000 to $ 37,400.

ML // Crypto Demand prediction // **twitter**

ML // Crypto Demand prediction // **twitter**

Inside this loop, we send our request for tweets within the 60-minute window — and then extract the information we want and append to our dataframe.
The result is a dataframe containing **~17K tweets** containing the word ‘tesla’ from the past seven days.

```
df.head()
```

	created_at	id	lang	text
0	2020-10-11T15:46:19.000Z	1315317815268847618	und	https://t.co/h2ho91JIft
1	2020-10-11T15:46:17.000Z	1315317808142708737	tr	@sanbel25 Tesla\n#CanYaman\n#ÖzgeGürel\n#BayYa...
2	2020-10-11T15:46:11.000Z	1315317784222609408	en	RT @maxH6294: @CannaFrom RETWEET\n\n #THRIV...
3	2020-10-11T15:46:11.000Z	1315317783874465792	fi	@miikaaulio @Olkurola @pvesterbacka @VentureBe...
4	2020-10-11T15:46:10.000Z	1315317780590153728	es	Protector eléctrico general, marca: TESLA; par...

ML // Crypto Demand prediction // **twitter**

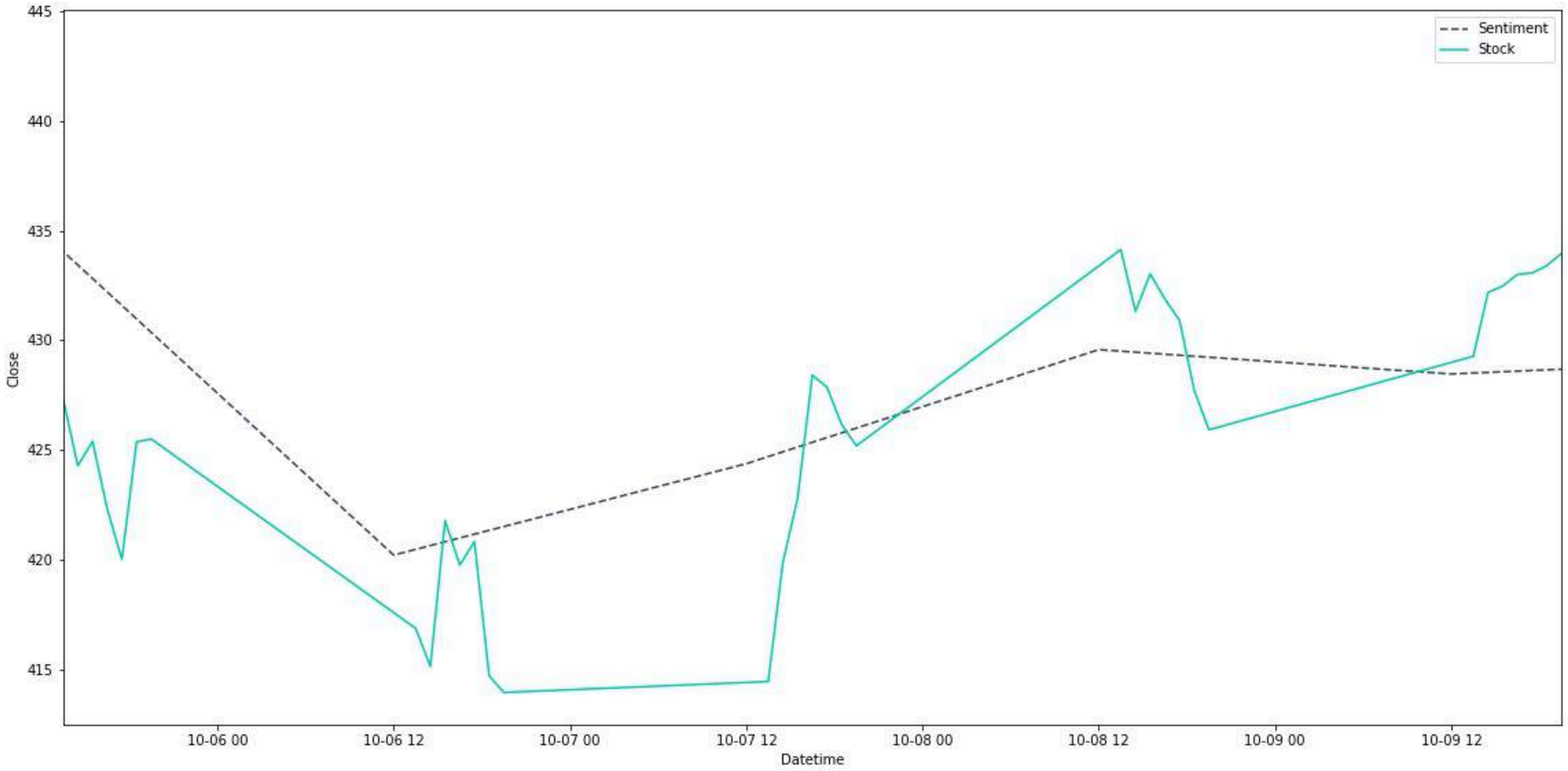
```
tweets.head()
```

	created_at	id	text	probability	sentiment
0	Sun Oct 11 10:23:39 +0000 2020	1315236612863795202	Tesla has installed a large supercharging station in Germany	0.997872	POSITIVE
1	Sun Oct 11 10:23:38 +0000 2020	1315236610464796673	Dealerships & billions of dollars of debt is the Albatross no one talks about. Software income is how Tesla is doing it today, and marginal cost of Software improvement is 0. Noone has the engineering/ innovation moat to make owners pay 2000 - 7000 $ a pop at zero marginal cost	0.511004	POSITIVE
2	Sun Oct 11 10:23:38 +0000 2020	1315236609340530688	Tesla *not up	0.999808	NEGATIVE
3	Sun Oct 11 10:23:37 +0000 2020	1315236606006059008	: Tesla One step closer to living full time in my car	0.998911	POSITIVE
4	Sun Oct 11 10:23:24 +0000 2020	1315236550960005124	: "If you want to find the secrets of the universe, think in terms of energy, frequency and vibration." ~ Nikola Tesla	0.999144	POSITIVE

TSLA stock prices Monday-Friday. The sentiment (originally scored from -1 to +1 has been multiplied to accentuate +ve or -ve sentiment, and centered on the average stock price value for the week.

ML // Crypto Demand prediction // **twitter**

It's clear that the Twitter sentiment and stock price are correlated during this week. Of course, a larger timestamp would provide greater confidence — but this provides us with an initial positive outcome to investigate further.



Gracias ;)