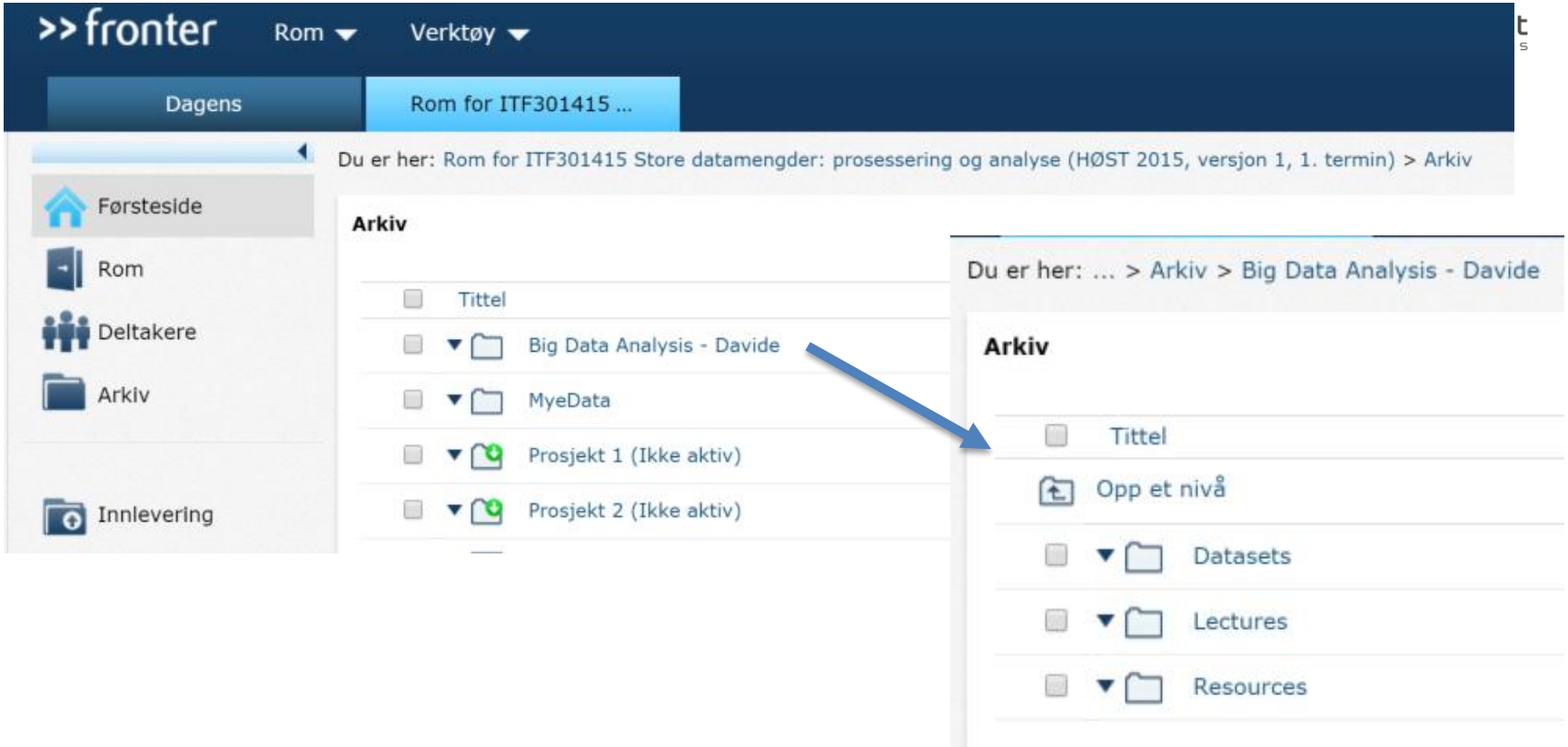
Last session (2015-10-08)

- Big Data facts
- Big Data Analysis Process
- Visualisation
- Machine Learning
- Examples

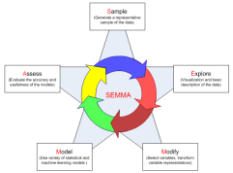
Today's session

- Big Data Analysis Process
- Hands-on

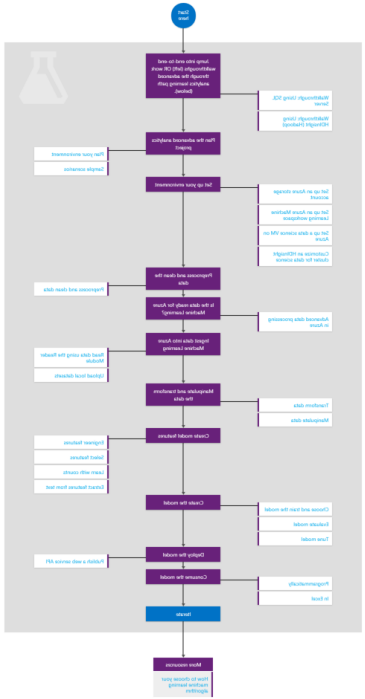


Big Data Analysis Process

- Formal "old school"
  - SEMMA
  - CRISP-DM
    - Spec uploaded to frontér

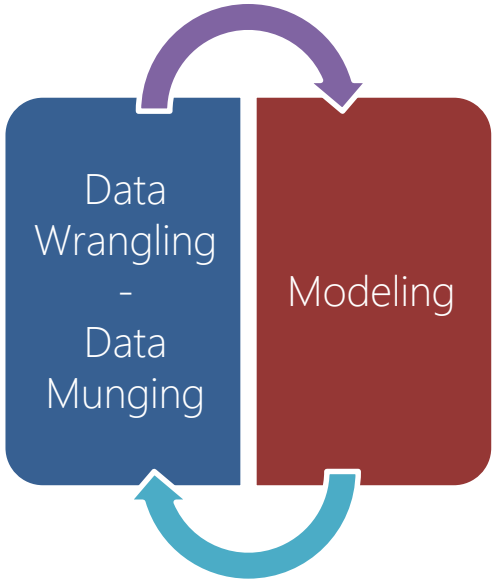


- Informal data science practice
  - Ad-hoc processes
  - Case dependent
  - Microsoft's suggested process: https://azure.microsoft.com/en-us/documentation/learning-paths/machine-learning-self-guided-predictive-analytics-training/



Big Data Analysis Process – Main Steps

- Data access
- Data pre-processing / cleaning
- Data transformation / manipulation
- Feature selection
- Feature extraction
- Feature engineering
- Model choice and training
- Model evaluation and tuning
- Model deployment



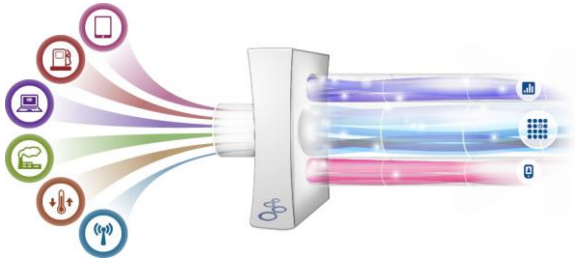
Big Data Analysis Process – Data Access

Data at Rest

- Inactive & unchanging data which is stored physically in any digital form
- Historical data

Data in Motion

- Streaming data
- Video feeds
- Message feeds (e.g. Twitter)
- Sensor data



Big Data Analysis Process – Data Access

- Data analysis (through machine learning) normally requires data at rest
- Streaming analytics can offer some limited "on-the-fly" analysis but usually does not include advaced modelling

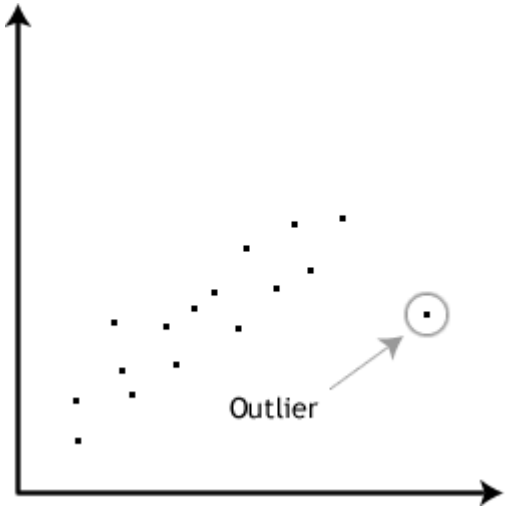
Data Pre-processing / Cleaning

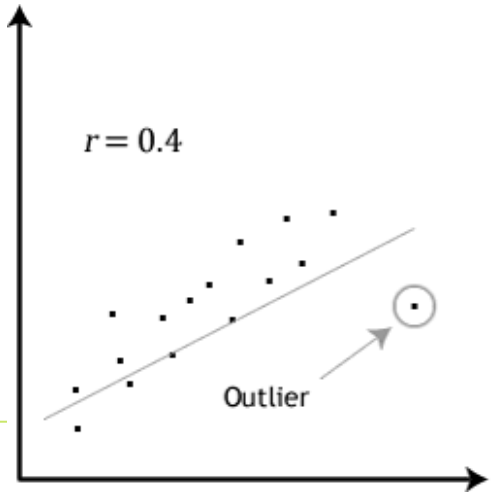
- Missing data
  - Remove row
  - Replace with fixed value
  - Replace with average (median) value
  - Replace with "most probable" value

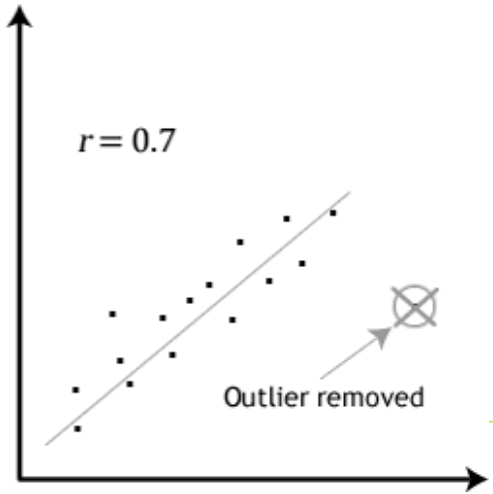
StartTime	Device	Resource	Time	Volume	Unit
12.08.2015 07:16	Ladetorget Moss 181	DO1	48	7,65	kwh
11.08.2015 21:32	Ladetorget Moss 181	DO1	15	10,68	kwh
11.08.2015 20:51	Ladetorget Moss 182	DO1	9	4,19	kwh
11.08.2015 18:58	Ladetorget Moss 181	DO1	8	5,74	kwh
11.08.2015 18:39	Ladetorget Moss 182	DO1	23	12,33	kwh
11.08.2015 18:16	Ladetorget Moss 181	DO1	32	10,66	kwh
11.08.2015 18:15	Ladetorget Moss 181	DO1	1	0	kwh
11.08.2015 17:58	Ladetorget Moss 181	DO2	17	11,05	kwh
11.08.2015 17:53	Ladetorget Moss 182	DO2	31	11,96	kwh
11.08.2015 17:44	Ladetorget Moss 181	DO1	10	3,26	kwh
11.08.2015 17:32	Ladetorget Moss 182	DO1	17	12,07	kwh
11.08.2015 16:28	Ladetorget Moss 181	DO1	6	3,95	kwh
11.08.2015 16:00	Ladetorget Moss 182	DO1	9	6,48	kwh
11.08.2015 15:44	Ladetorget Moss 185	DO2	3	0	kwh
11.08.2015 15:19	Ladetorget Moss 185	DO2	25	2,74	kwh
11.08.2015 14:35	Ladetorget Moss 186	DO1	3	0	kwh
11.08.2015 14:03	Ladetorget Moss 186	DO1	33	0	kwh
11.08.2015 12:53	Ladetorget Moss 182	DO2	27	12,46	kwh
11.08.2015 11:41	Ladetorget Moss 181	DO1	8	3,46	kwh
11.08.2015 11:25	Ladetorget Moss 186	DO2	2		NaN
11.08.2015 09:55	Ladetorget Moss 184	DO1	131	14,65	kwh
11.08.2015 07:35	Ladetorget Moss 185	DO2	445	7,18	kwh
11.08.2015 07:34	Ladetorget Moss 188	DO2	455	10,62	kwh
11.08.2015 07:34	Ladetorget Moss 185	DO2	1	0	kwh
11.08.2015 07:28	Ladetorget Moss 185	DO2	2		NaN
11.08.2015 02:13	Ladetorget Moss 181	DO2	11	6,58	kwh
10.08.2015 19:12	Ladetorget Moss 181	DO2	10	6,24	kwh
10.08.2015 18:17	Ladetorget Moss 182	DO2	10	6,21	kwh
10.08.2015 17:54	Ladetorget Moss 182	DO2	2	NaN	NaN
10.08.2015 15:10	Ladetorget Moss 182	DO1	6	3,74	kwh
10.08.2015 14:54	Ladetorget Moss 182	DO1	6	3,68	kwh
10.08.2015 14:27	Ladetorget Moss 181	DO2	20	11,16	kwh

Data Pre-processing / Cleaning

- Outliers

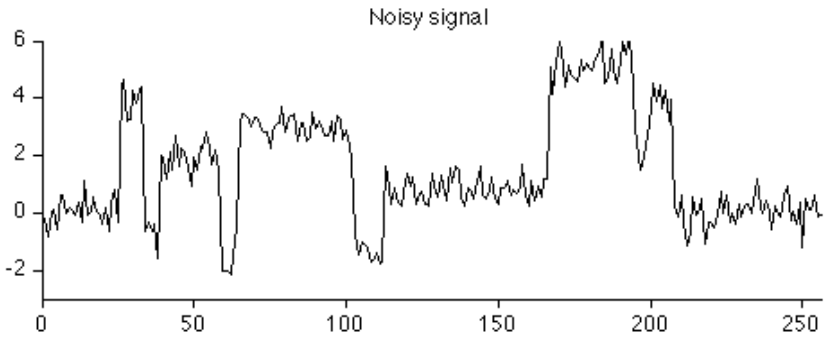


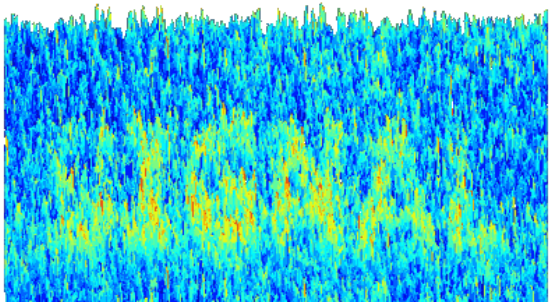


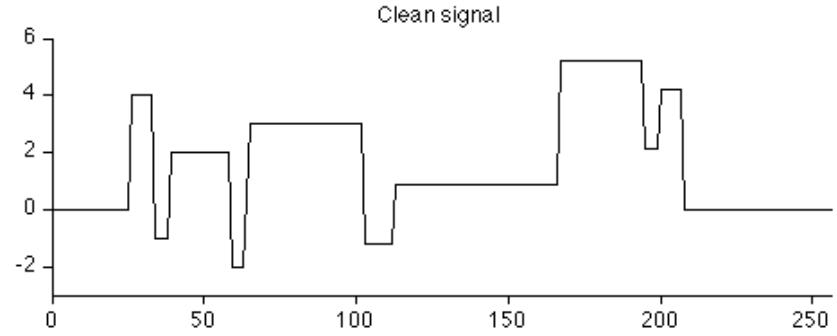


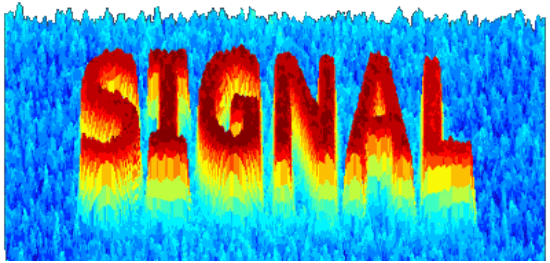
Data Pre-processing / Cleaning

- Noise



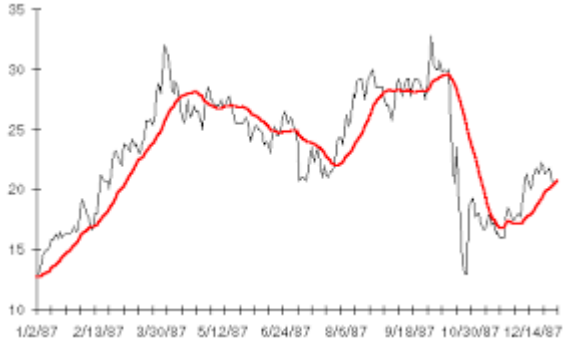


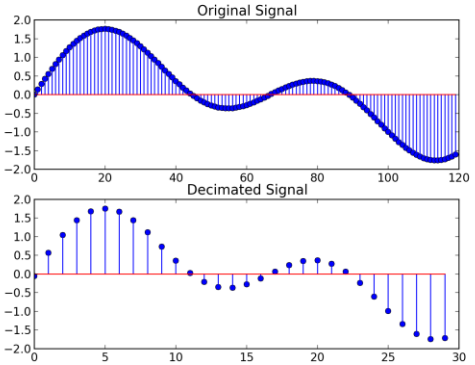


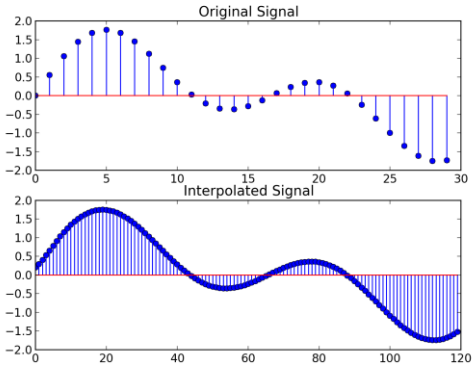


Data Transformation / Manipulation

- Filtering
  - Averaging (moving average)
  - …
- Scaling / Normalisation
- Sampling / Synchronisation
  - Down-sampling
    - Decimation
  - Up-sampling
    - Interpolation
- …







Feature Engineering

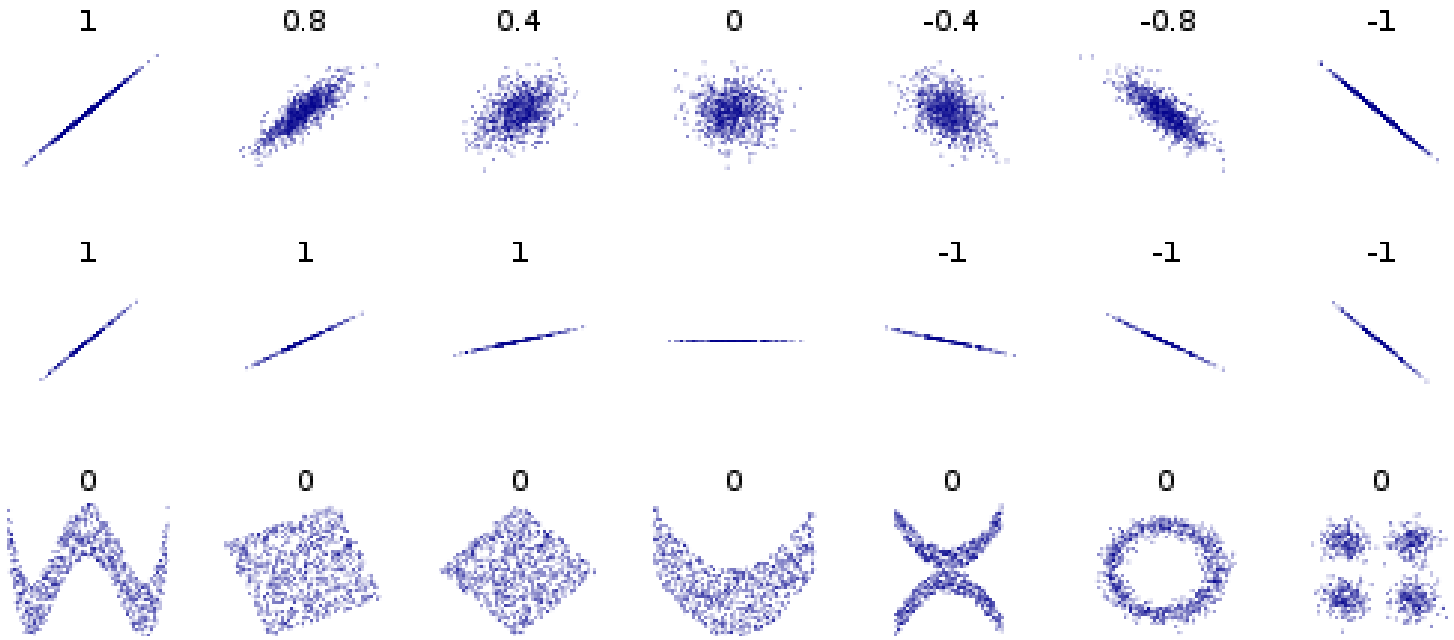
- Dimensionality Reduction
  - Feature Selection
  - Embedding
- Feature Extraction
- Feature Learning

Feature Selection

- Filter methods
  - E.g. Remove highly correlated variables

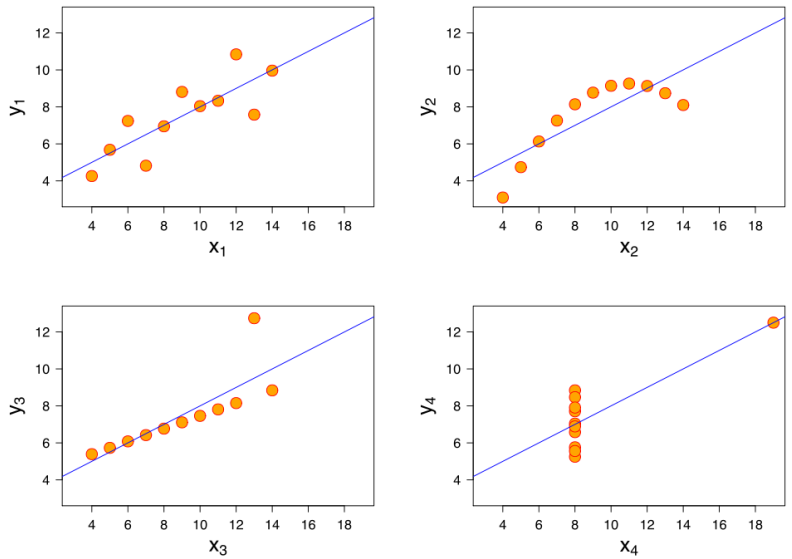
- Wrapper methods
  (Search for feature subset based on performance of the chosen data analysis algorithm)
  - Forward selection
    - Add best feature until satisfied
  - Backward elimination
    - Remove worst feature until satisfied
  - Global search
    - Genetic Algorithms
    - Boruta algorithm

Correlation - limitations

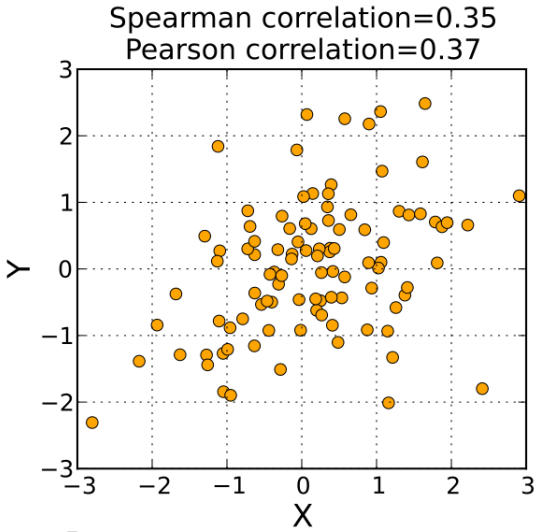


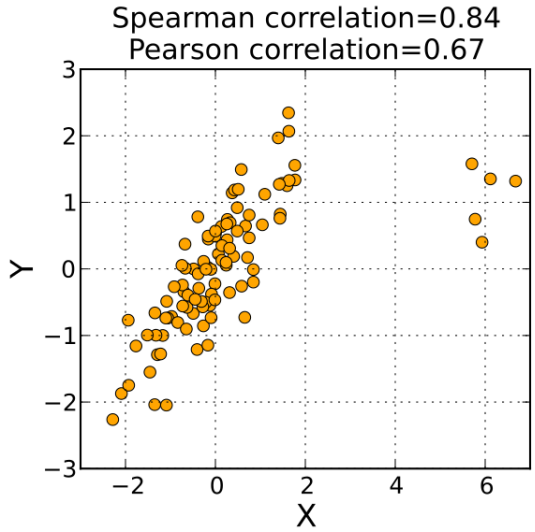
Correlation - limitations

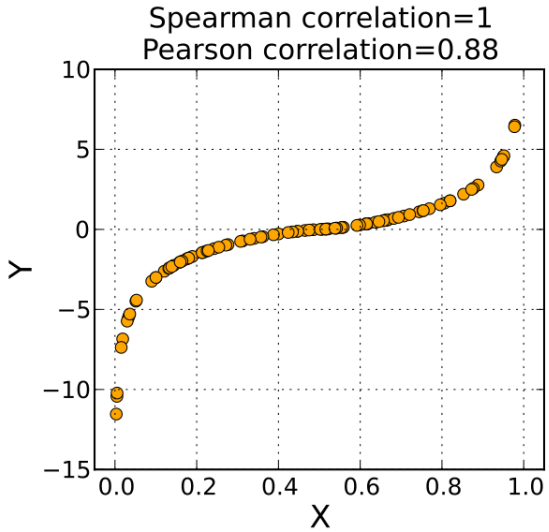
These cases have the same mean=7.5, variance=4.12, correlation=0.816 and regression line y=3+0.5x



Alternative correlation definitions



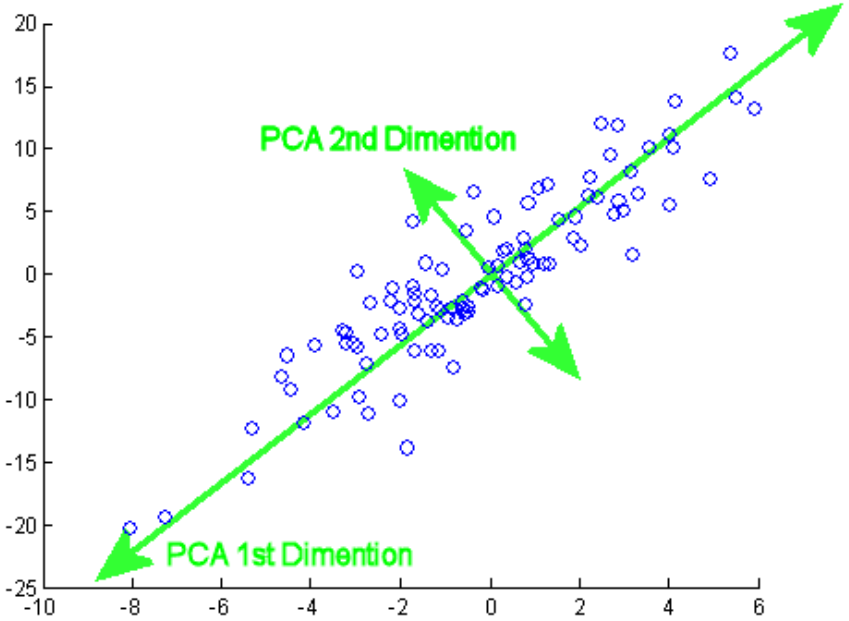




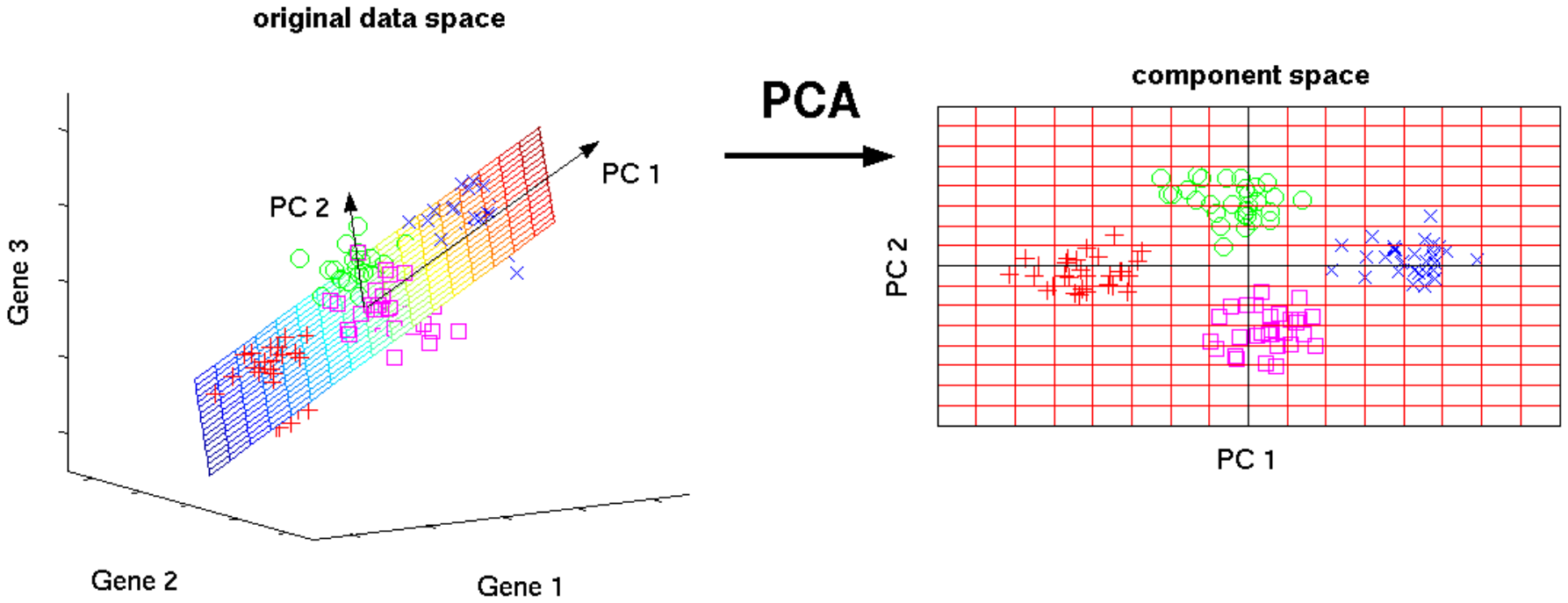
Dimensionality Reduction

- Linear
  - PCA – Principal Component Analysis
- Non-linear
  - Manifold learning
    - ISOMAP
    - t-SNE
    - ….

PCA – Principal Componen Analysis



PCA – Principal Componen Analysis

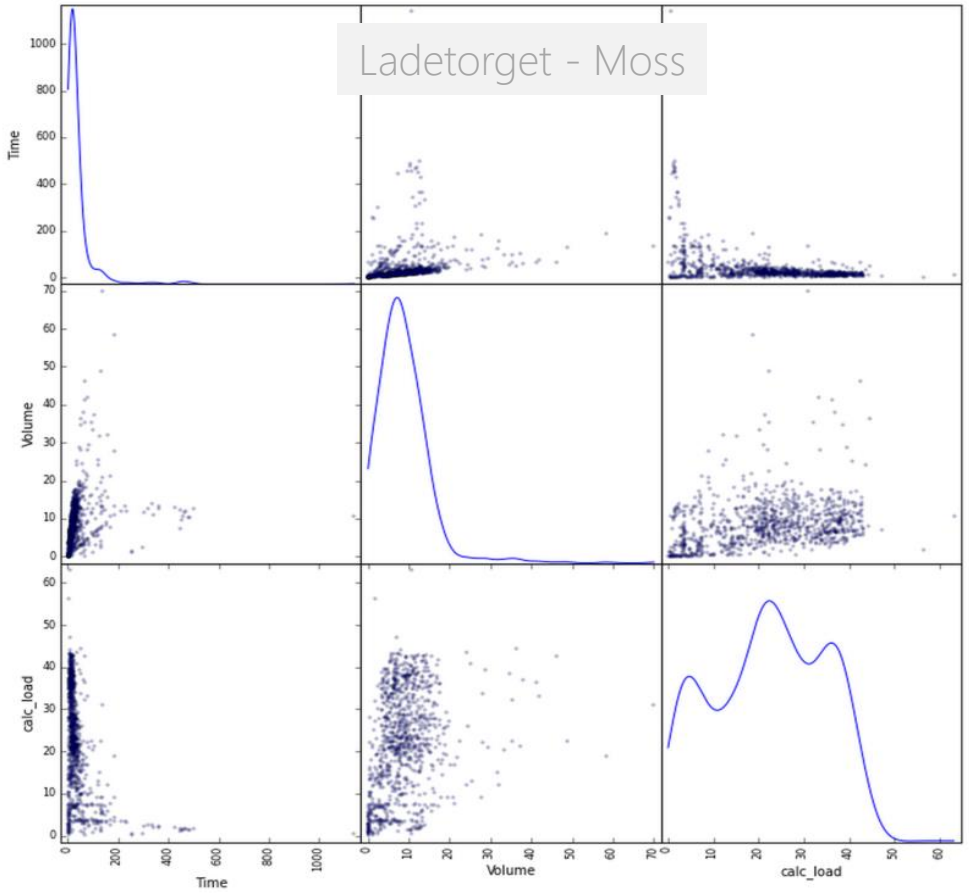


Handwritten digits – A Case for dimensionality reduction

1797 8x8 images = 64 feature vector

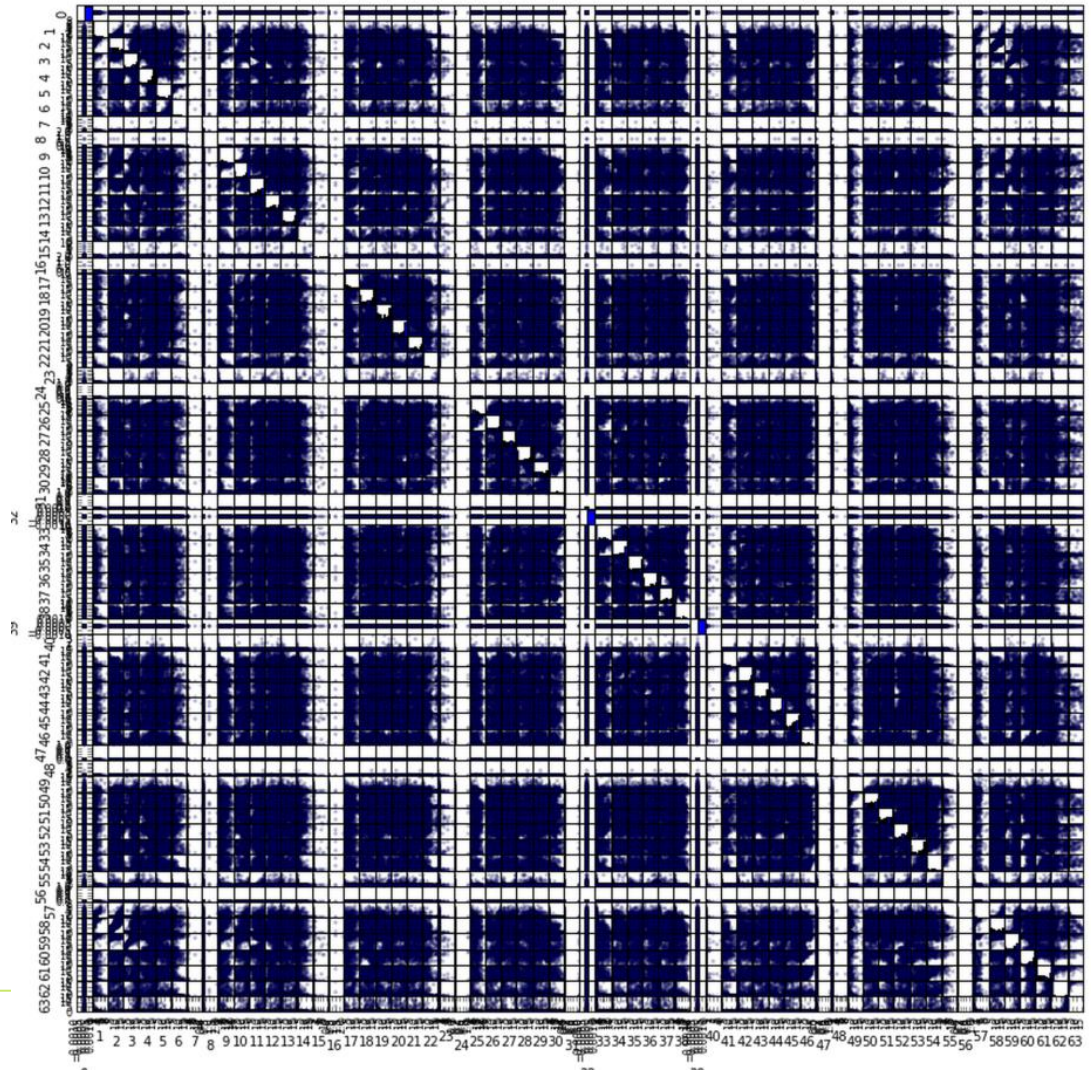
A selection from the 64-dimensional digits dataset

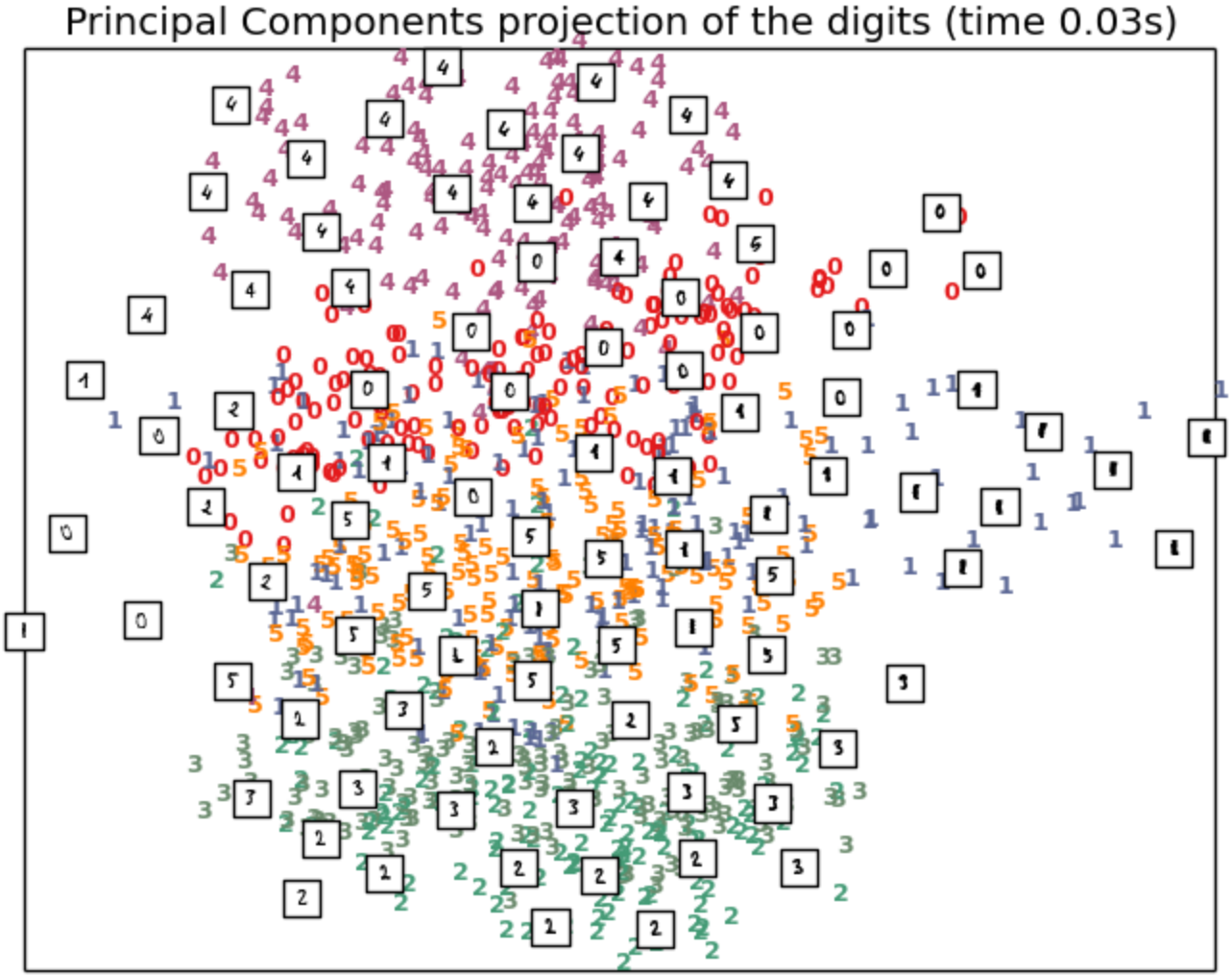
Scatter Matrix Plot - examples



Scatter Matrix of handwritten digits dataset

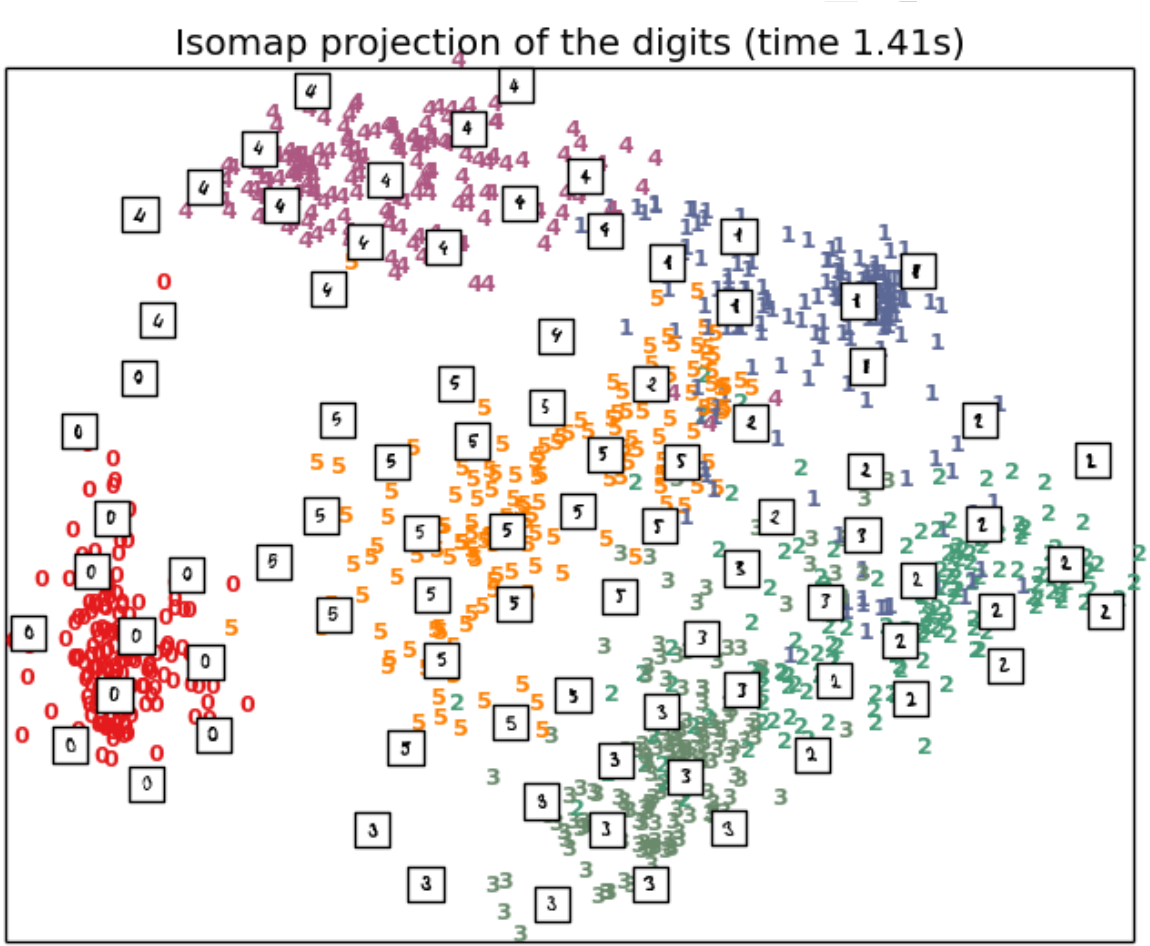
Unusable!




Principal Components projection of the digits (time 0.03s)

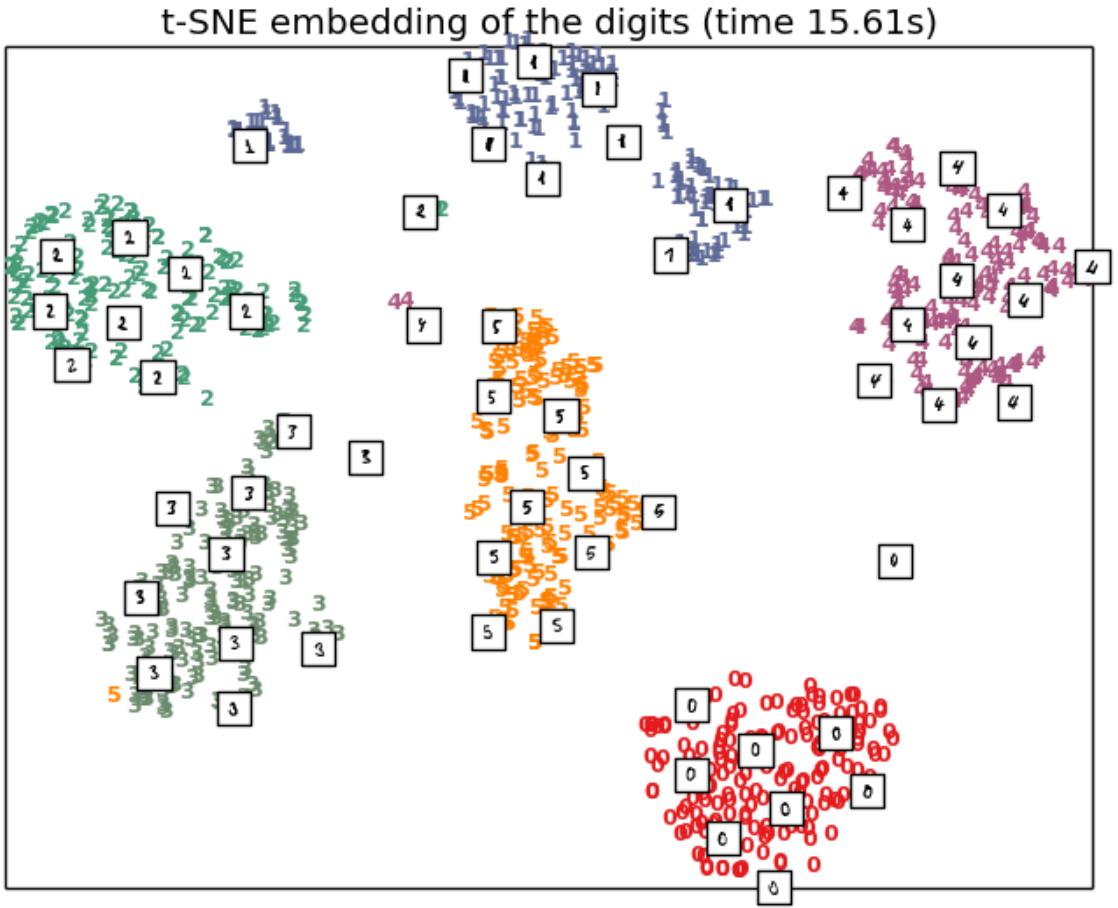
Isometric Mapping

Isomap seeks a lower-dimensional mapping which maintains geodesic distances between all points



t-distributed Stochastic Neighbor Embedding

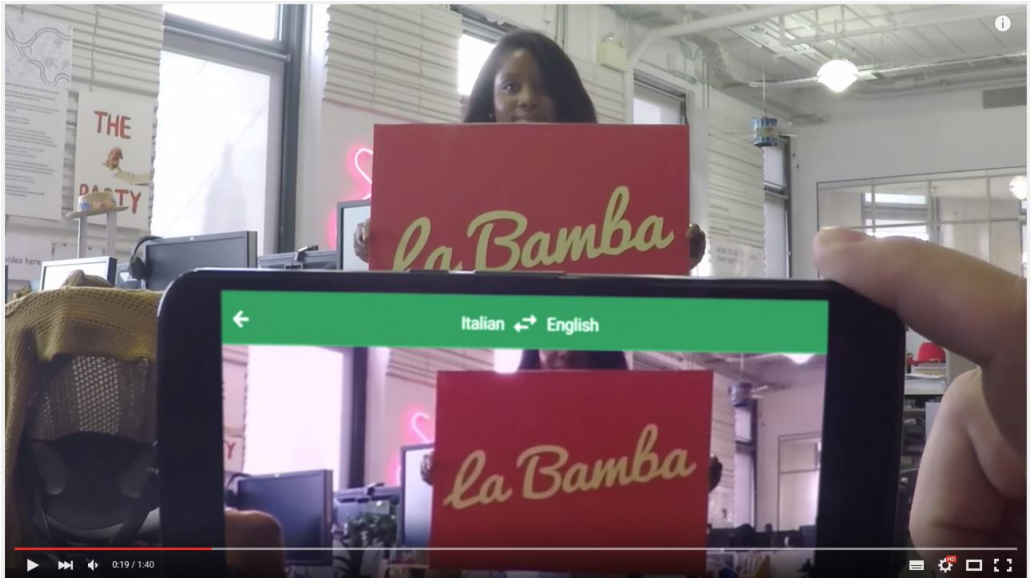
Models each high-dimensional object by a two-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points



How Google Translate Makes Signs Instantly Readable

https://www.youtube.com/watch?v=0zKU7jDA2nc

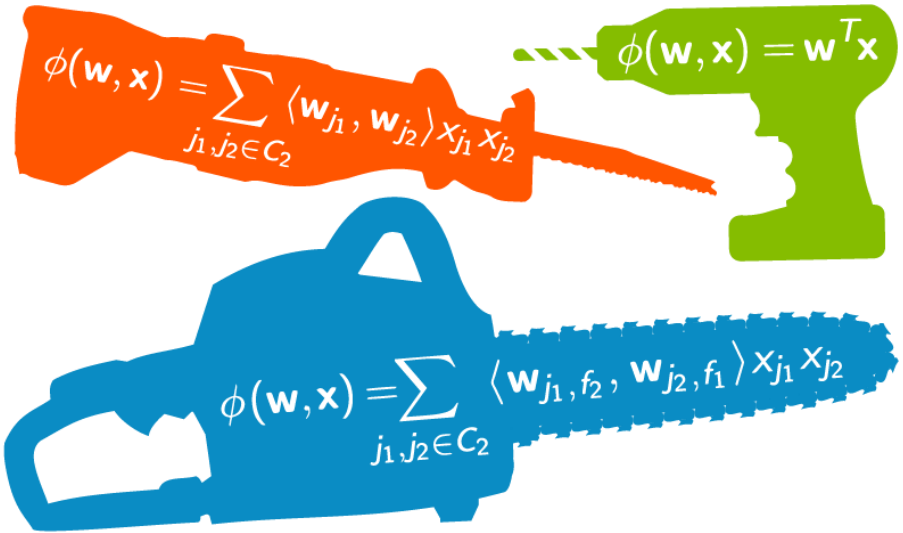
Google Translate vs. “La Bamba”



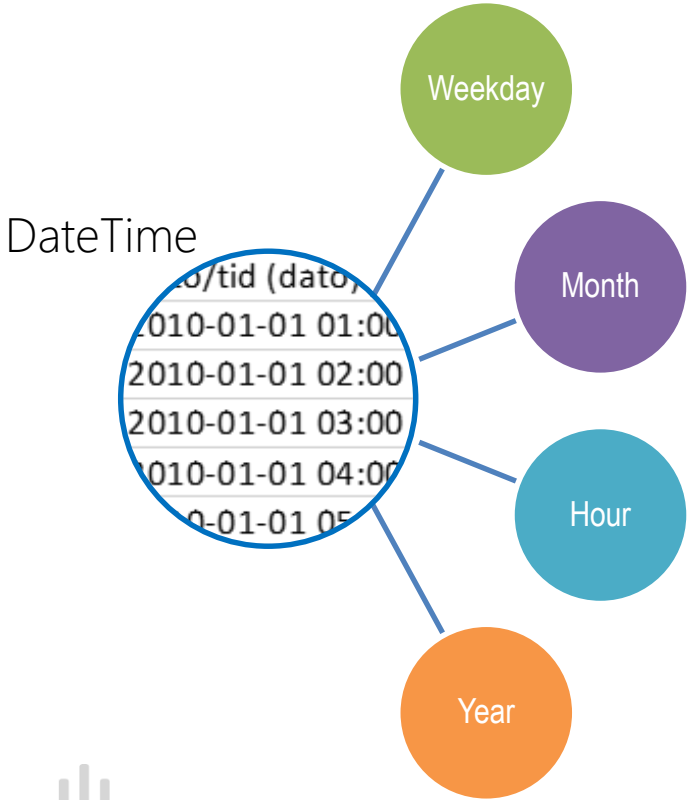
https://www.youtube.com/watch?v=06olHmcJjS0

Feature Engineering – Feature Extraction

- Build new features from existing ones
- Learn new features

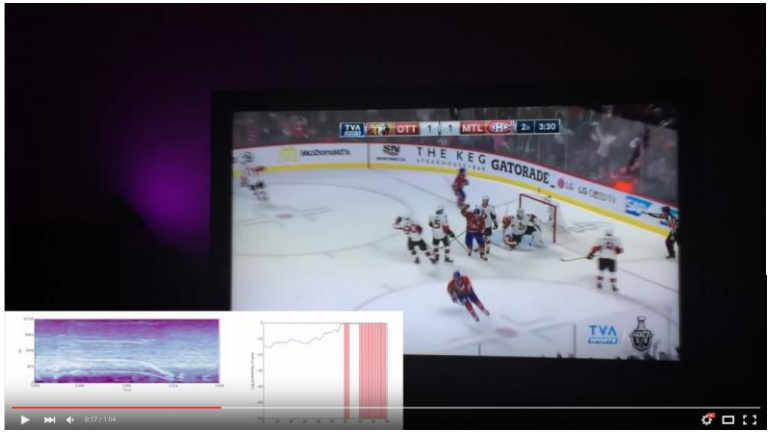


Simple Feature Extraction

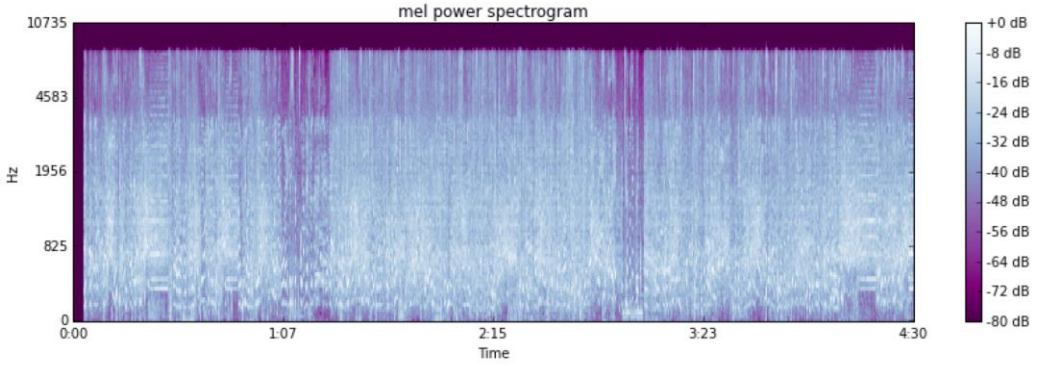


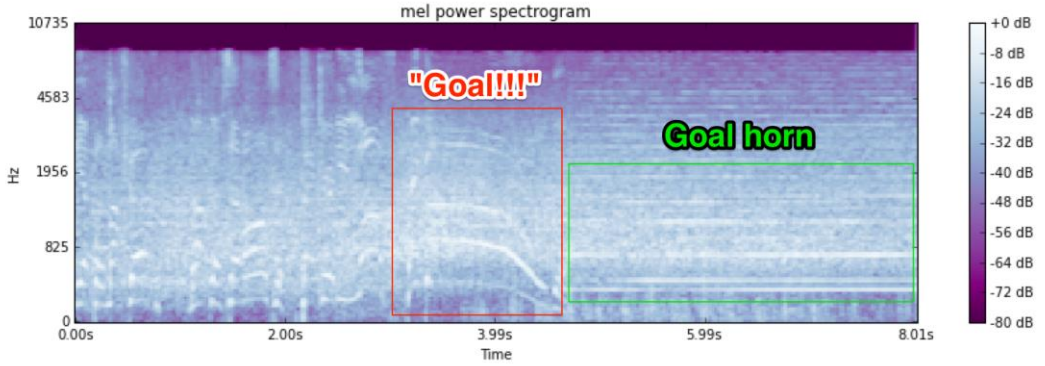
Dato/tid (dato)	Temp	kWh/h	Weekday	Month	Hour	Year
2010-01-01 01:00	-6,7	15280	6	1	1	2010
2010-01-01 02:00	-7,7	15008	6	1	2	2010
2010-01-01 03:00	-7,7	14896	6	1	3	2010
2010-01-01 04:00	-7,9	14712	6	1	4	2010
2010-01-01 05:00	-8,2	14576	6	1	5	2010
2010-01-01 06:00	-8,2	14720	6	1	6	2010
2010-01-01 07:00	-8,2	14928	6	1	7	2010
2010-01-01 08:00	-7,8	15040	6	1	8	2010

More Complex Feature Extraction



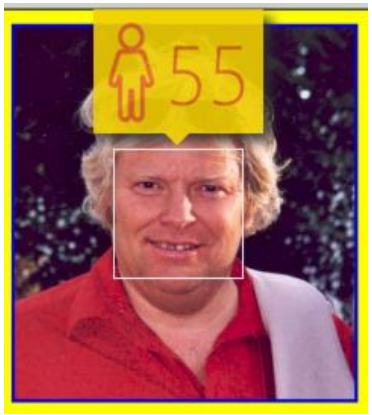
https://www.youtube.com/watch?v=tmh_eAq9yp4

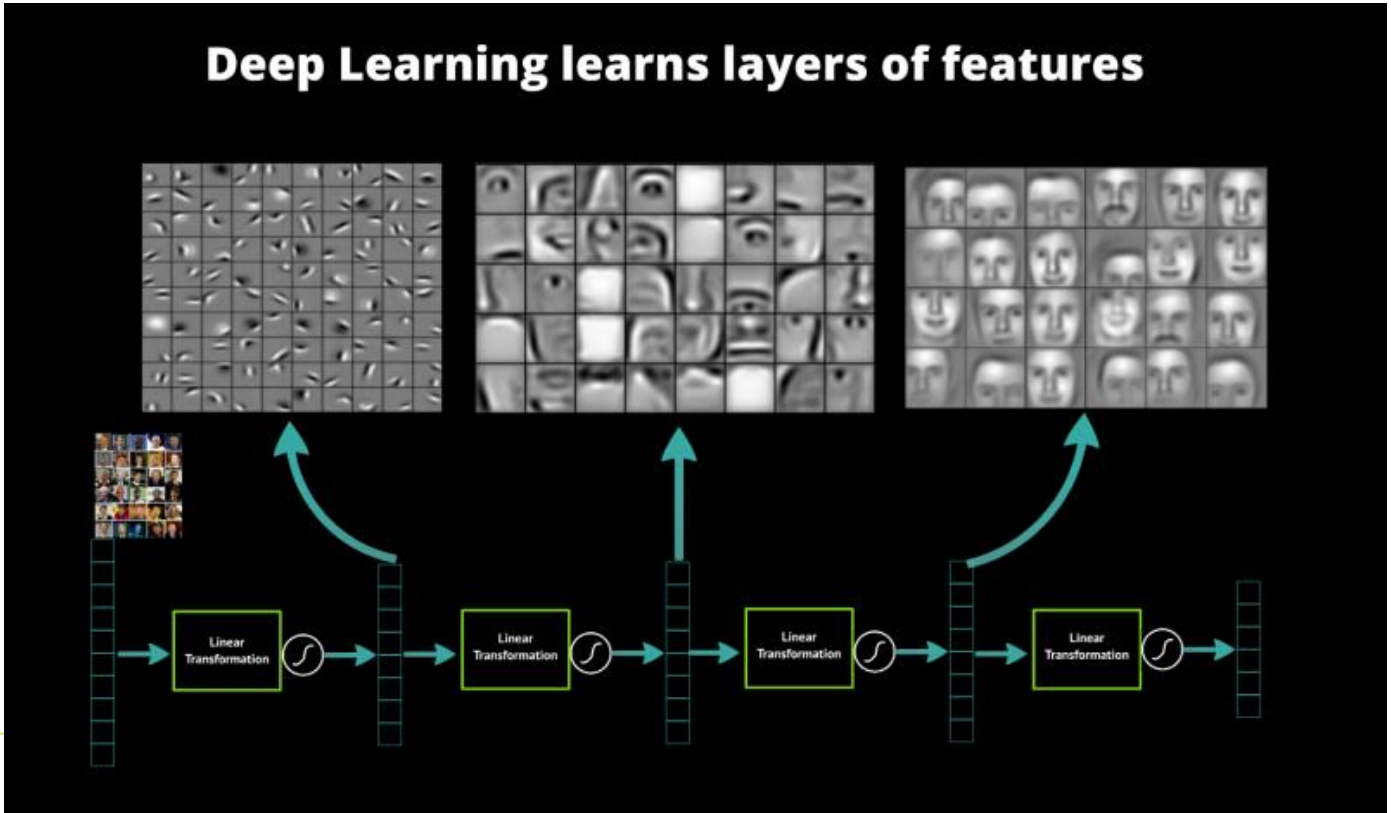


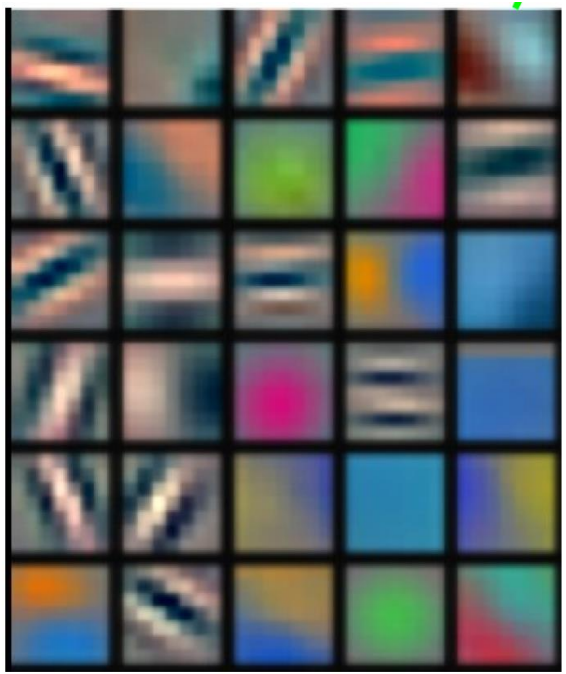


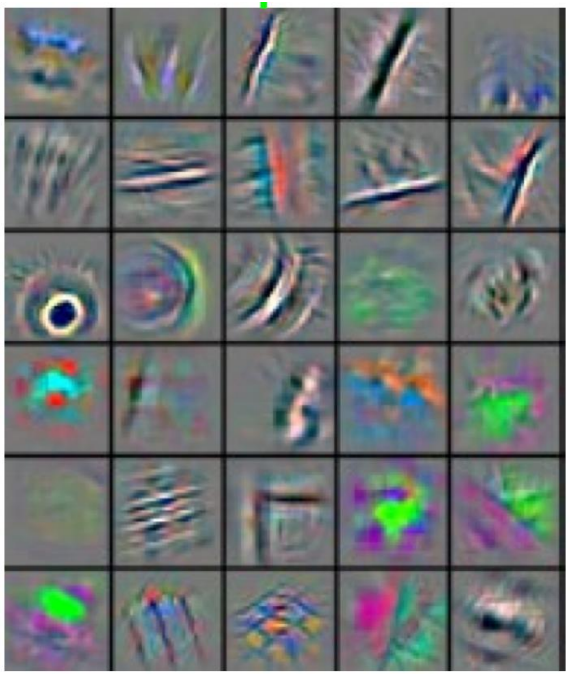
Feature Learning

- Use machine learning to create features useful for machine learning
- Deep Learning





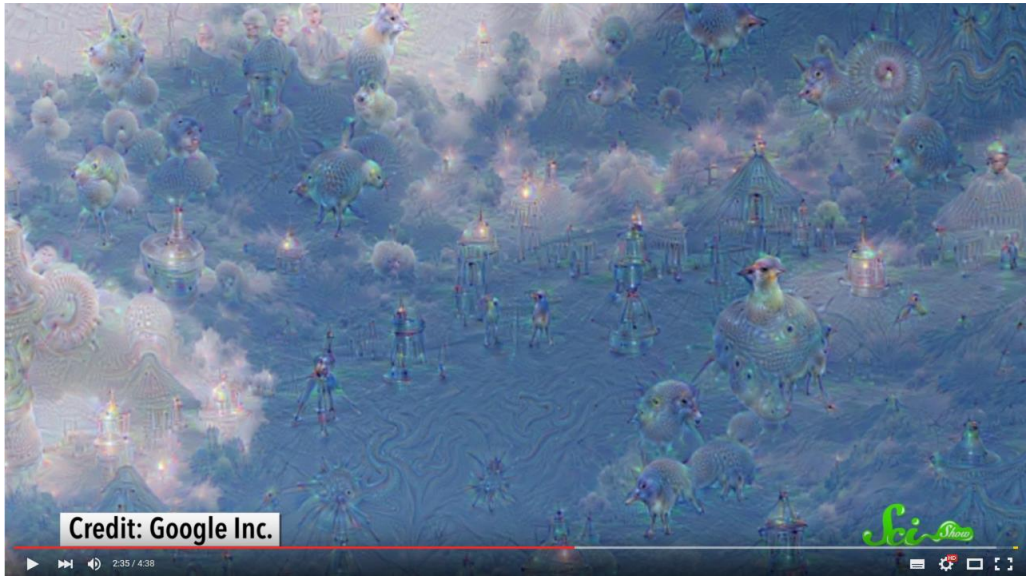




Machine Learning & Deep Neural Networks Explained

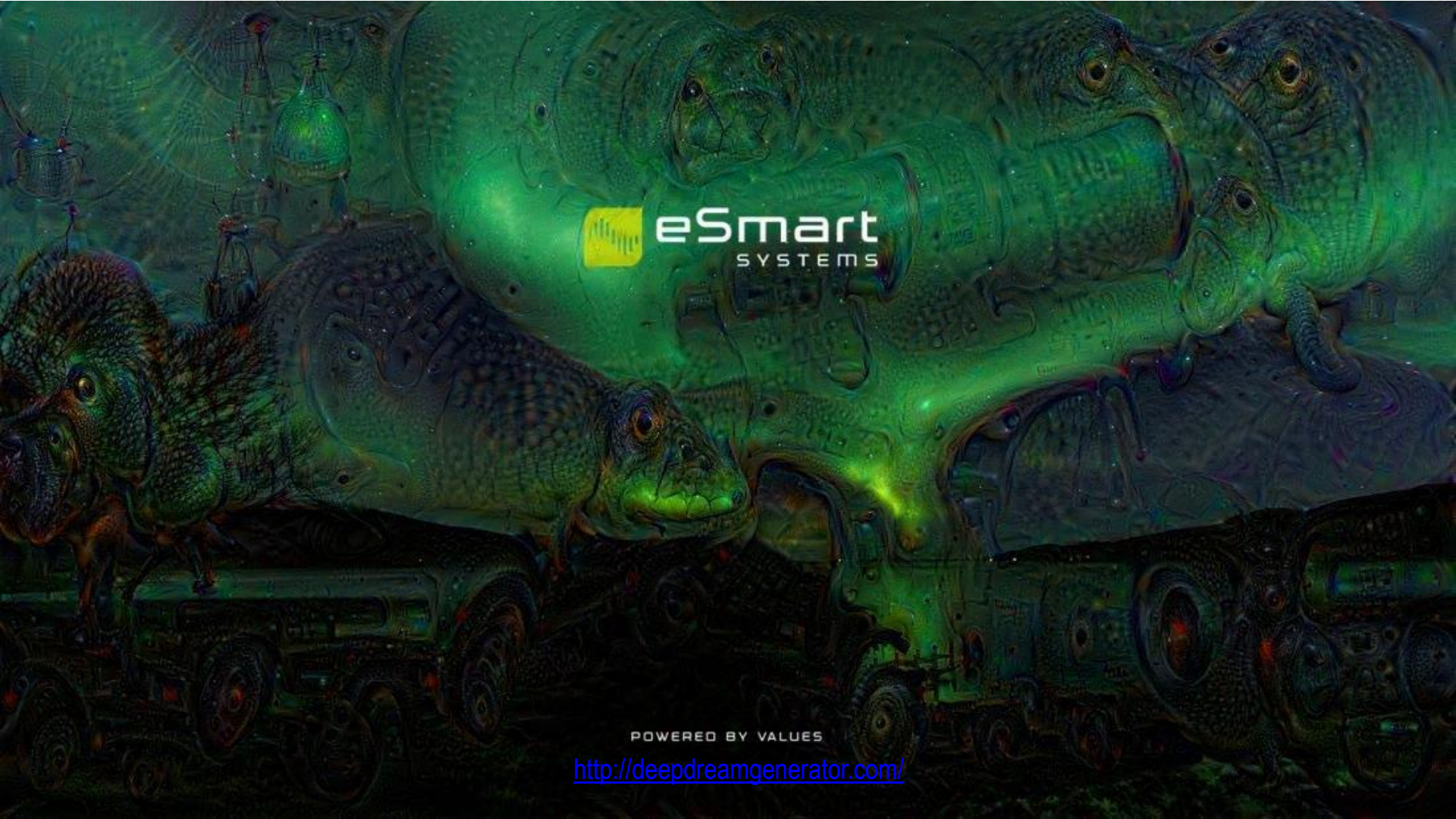
https://www.youtube.com/watch?v=bHvf7Tagt18

DeepDream: Inside Google's 'Daydreaming' Computers



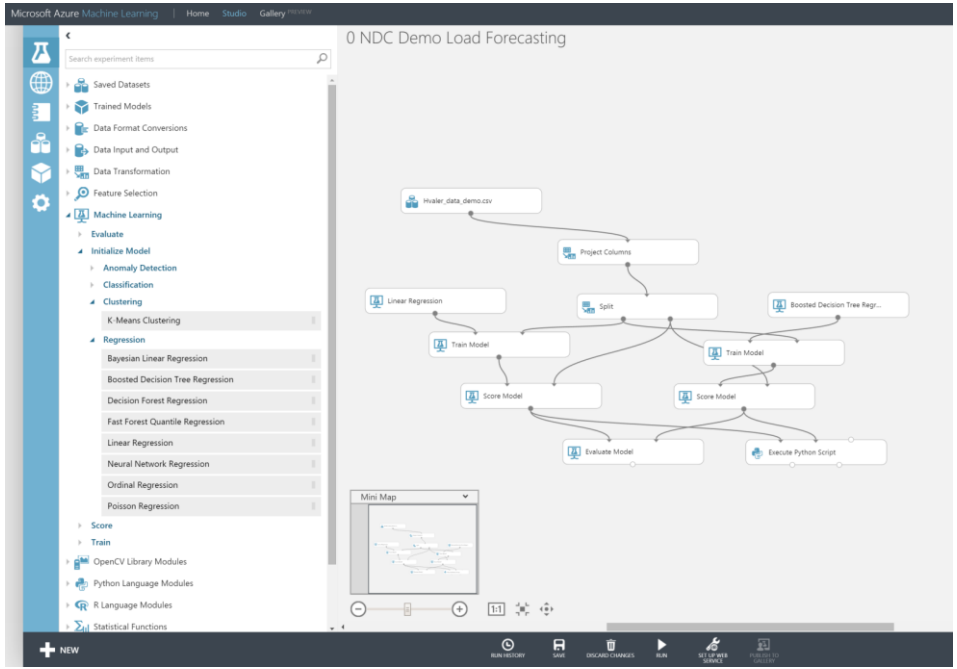
https://www.youtube.com/watch?v=3hnWf_wdgzs


eSmart
SYSTEMS
POWERED BY VALUES


eSmart
SYSTEMS
POWERED BY VALUES
http://deepdreamgenerator.com/

Hands-on session

- https://studio.azureml.net



Microsoft Azure Educator Grant Award

- Microsoft has made available 1 Microsoft Azure 12-month, $250/month Educator Pass
- 40- 6 month, $100/month Student Passes
- at no cost

- Codes can be redeemed at www.microsoftazurepass.com

Forskningsparken/ Oslo Science Park - Auditoriet

Torsdag November 05, 2015

Agenda for dagen:

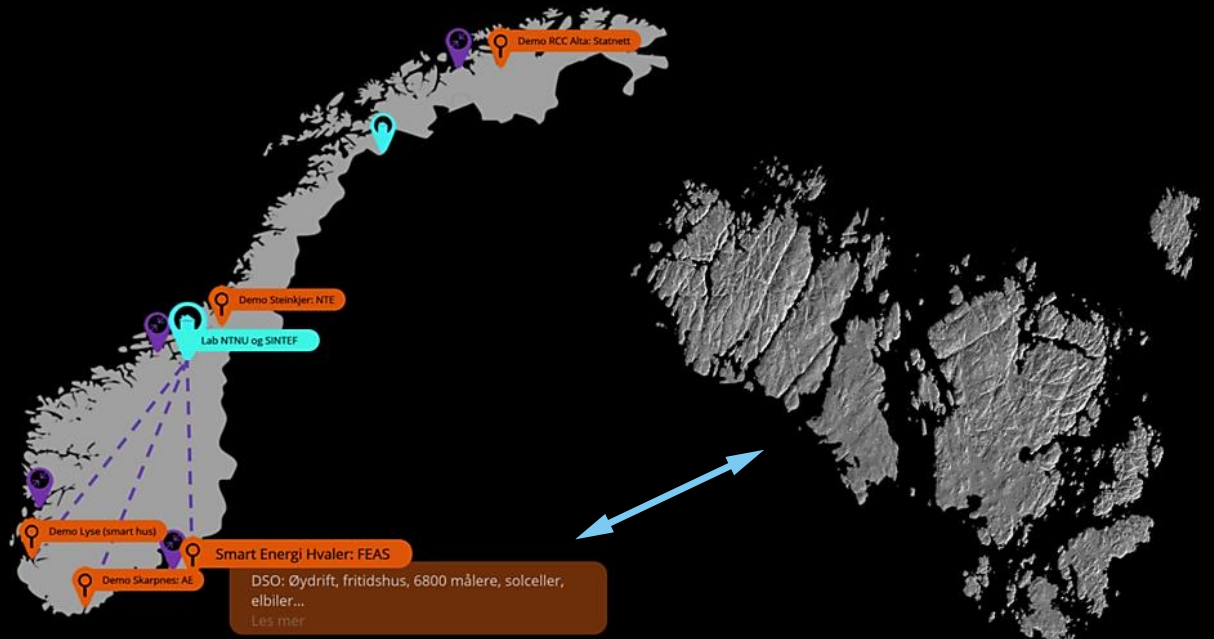
8.30 – 9.00 Welcome; registration and coffee
9:00 – 10:00 Overview of Microsoft Analytics Platform: Data Lake + Kona, etc
10:00 – 10:30 Q&A/Break
10:30 – 11:30 Building Big Data Applications Using Azure HDInsight Service
11:30 – 12:30 Q&A / LUNSJ
12:30 – 13:30 Power BI overview
13:30 – 14:00 Q&A/Break
14:00 – 15:00 End-to-End Analytics Solution: Real-World Scenario & Demo

Deltakelse er gratis – og vi spanderer mat og drikke!

For påmelding kan du følge linken til eventsiden.

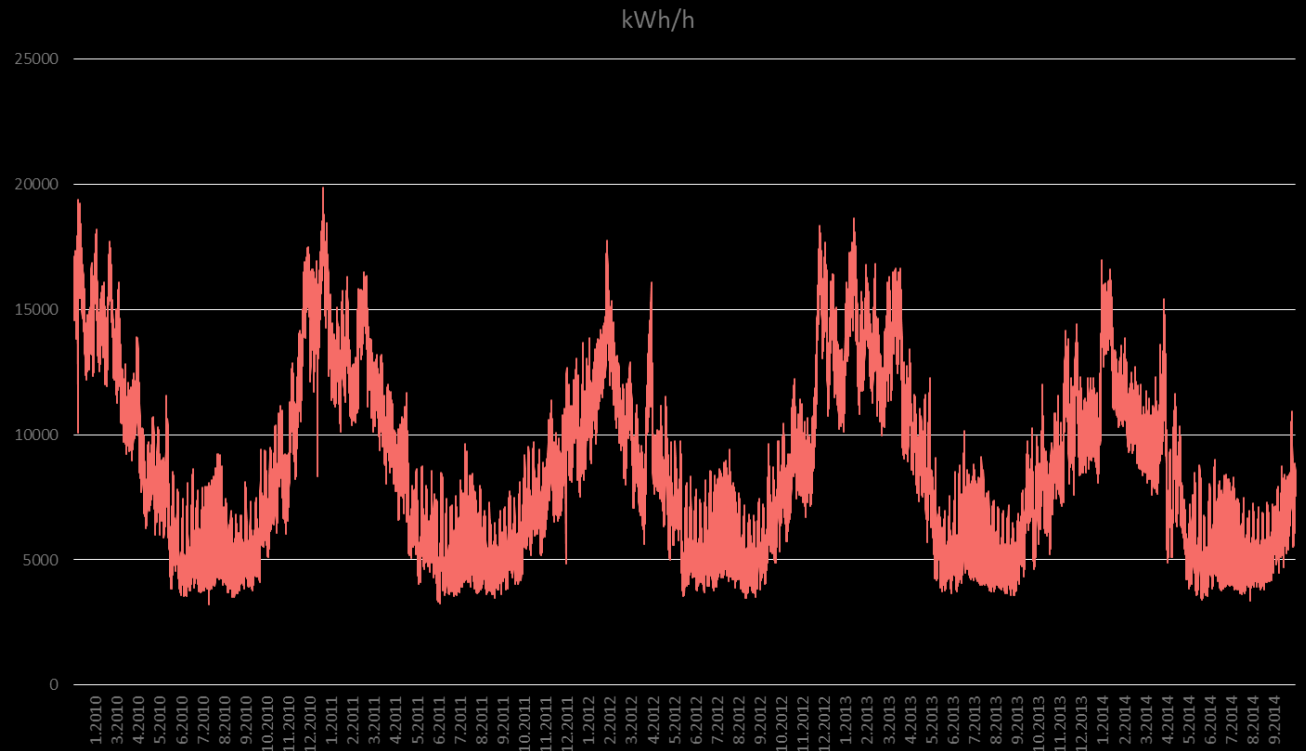
http://blogs.msdn.com/b/dpenorway/archive/2015/10/13/big-data-ekspertene-kommer-til-oslo.aspx

Demo Case – 6800 Smart Meters on Hvaler



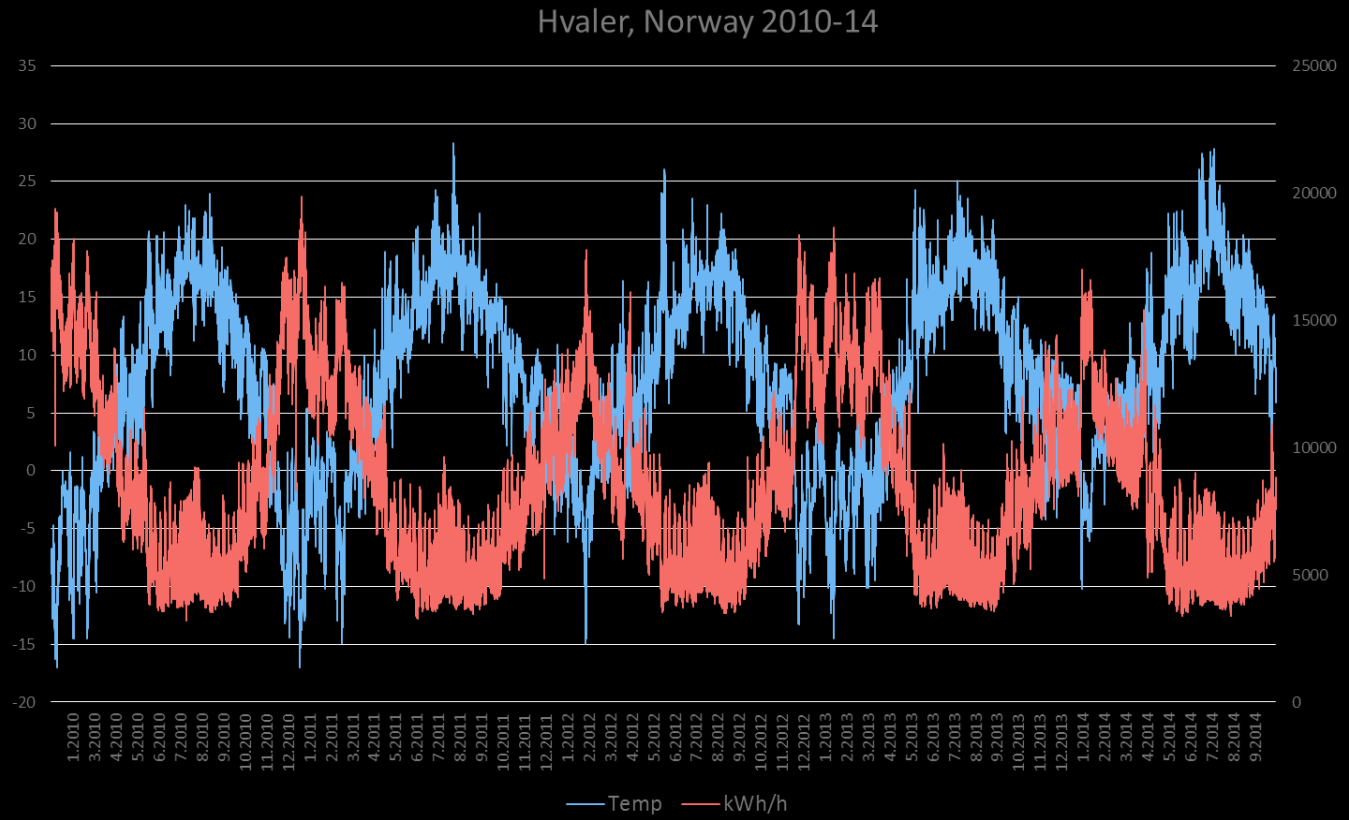
eSmart SYSTEMS

Consumption in Hvaler 2010-2014



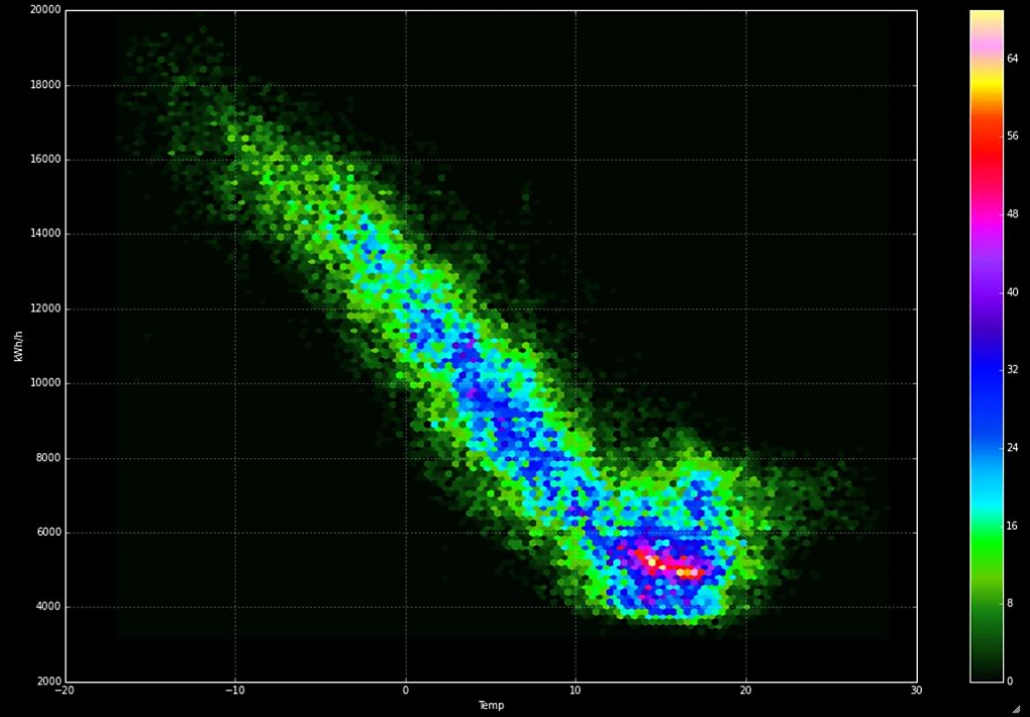
eSmart SYSTEMS

Consumption vs. Temperature

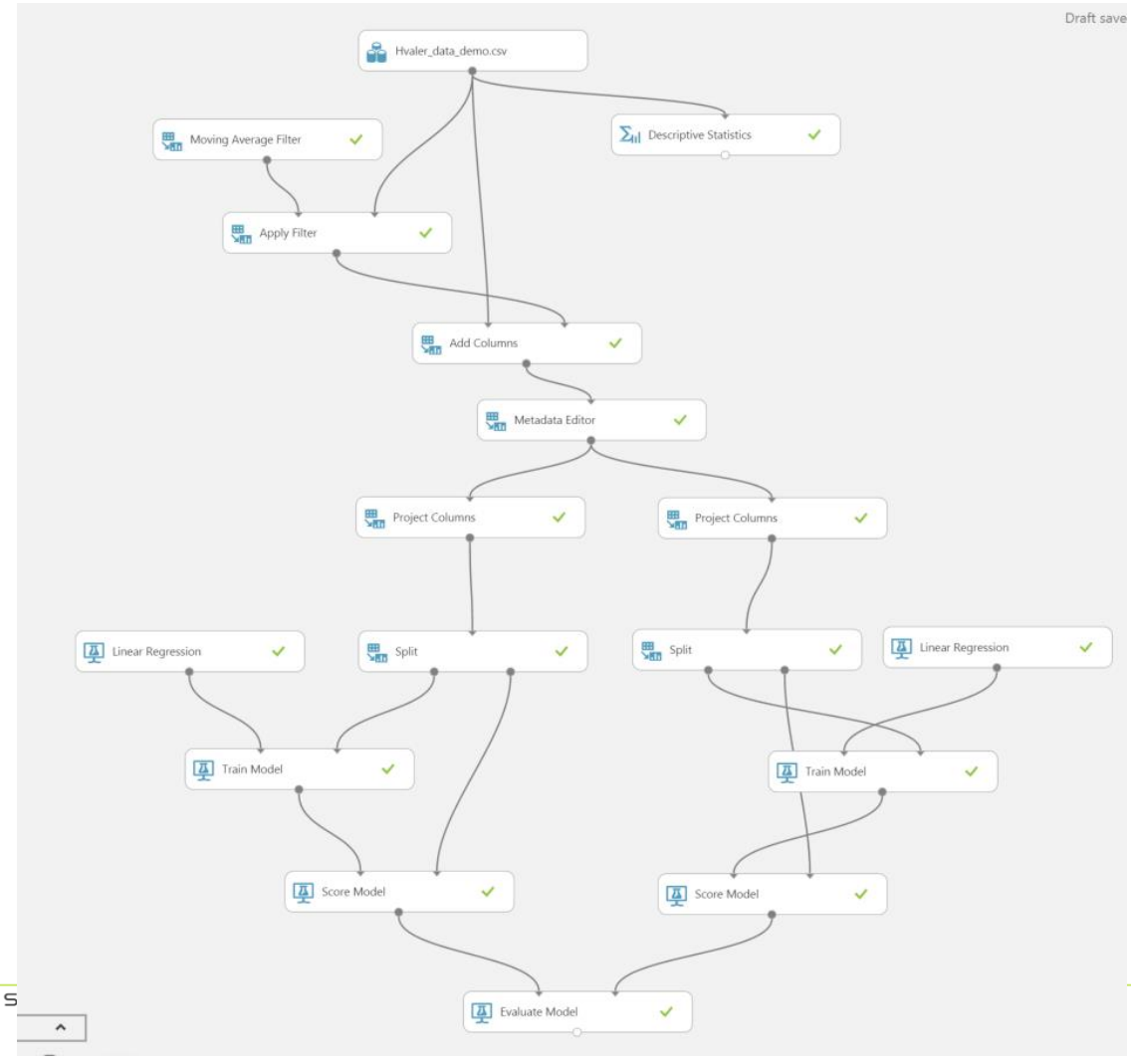


eSmart SYSTEMS

Consumption vs. Temperature



eSmart SYSTEMS


Draft save
Hvaler_data_demo.csv
Moving Average Filter
Descriptive Statistics
Apply Filter
Add Columns
Metadata Editor
Project Columns
Project Columns
Linear Regression
Split
Split
Linear Regression
Train Model
Train Model
Score Model
Score Model
Evaluate Model

Home assignments

by 2015-10-22 (next Thursday)

- Experiment with Azure ML Studio
  - Come with questions at the next lesson
- Explore the available Azure ML documentation
- Bring PC to next lesson (required, minimum one per group)