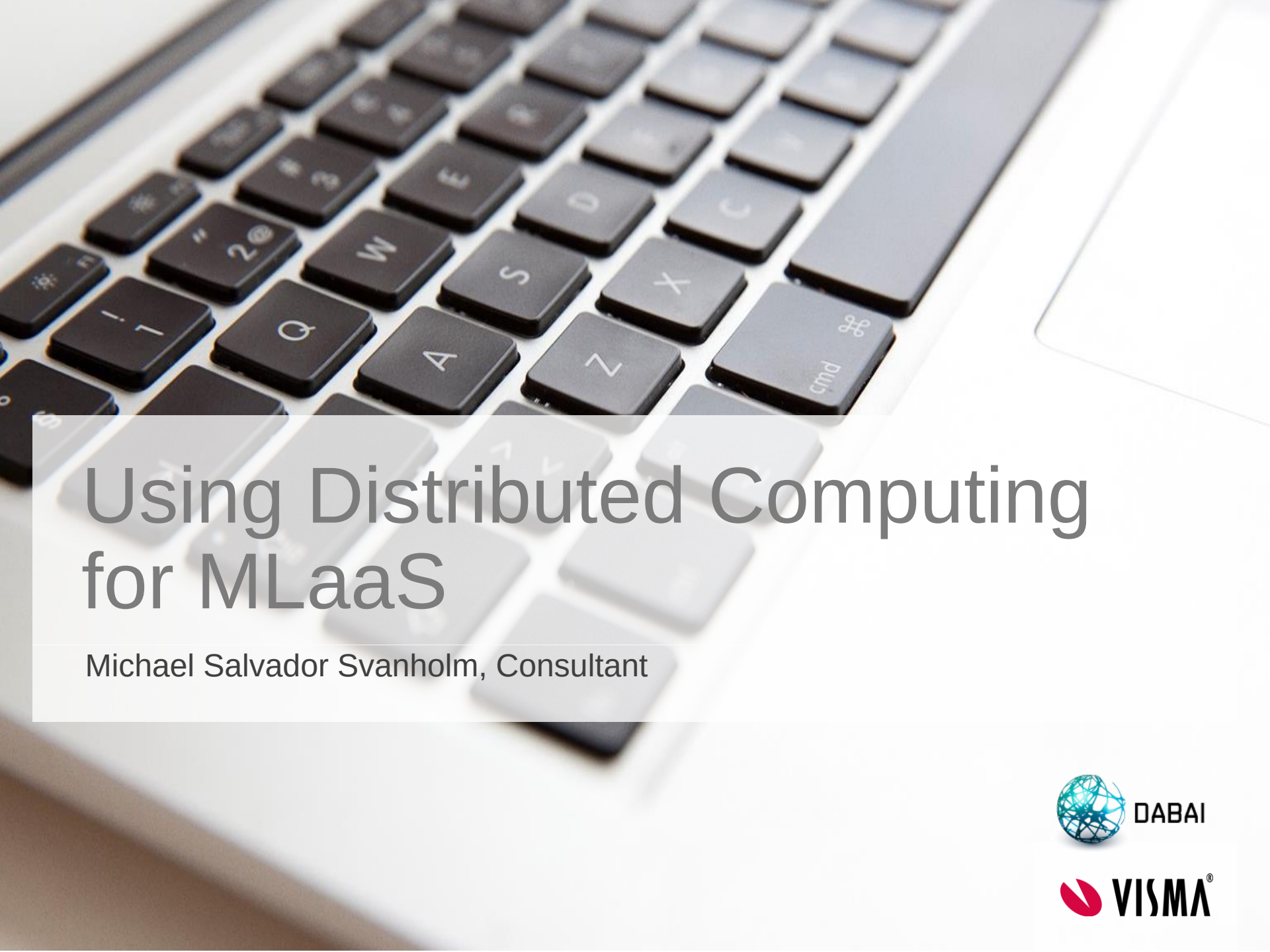


Using Distributed Computing for MLaaS

Michael Salvador Svanholm, Consultant

DABAI

VISMA

Distributed computing is a method to deliver results fast, when facing a growing amount of data

We have used Apache 

to distribute our Machine learning tools.

So far, we have created: Anomaly Detection and Classification.

DABAI

VISMA

Distributed computing is a method to deliver results fast, when facing a growing amount of data

We have used Apache 

to distribute our Machine learning tools.

So far, we have created: Anomaly Detection and Classification.

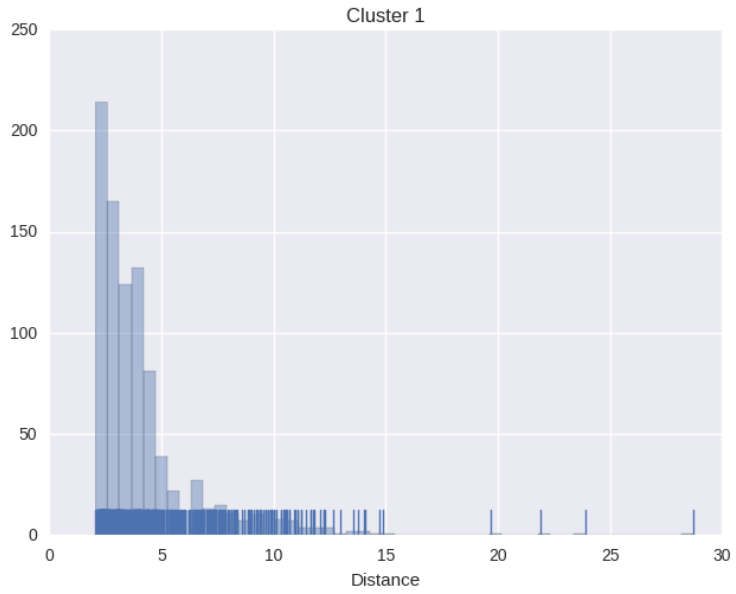
Ideally, clients can use these tools without help, if they "know" their own data.

DABAI

VISMA®

Anomaly Detection using K-means clustering can be used to clean data

On the other hand, anomalies can also be “data of interest” which means, that a lot of value can potentially come from examining them.

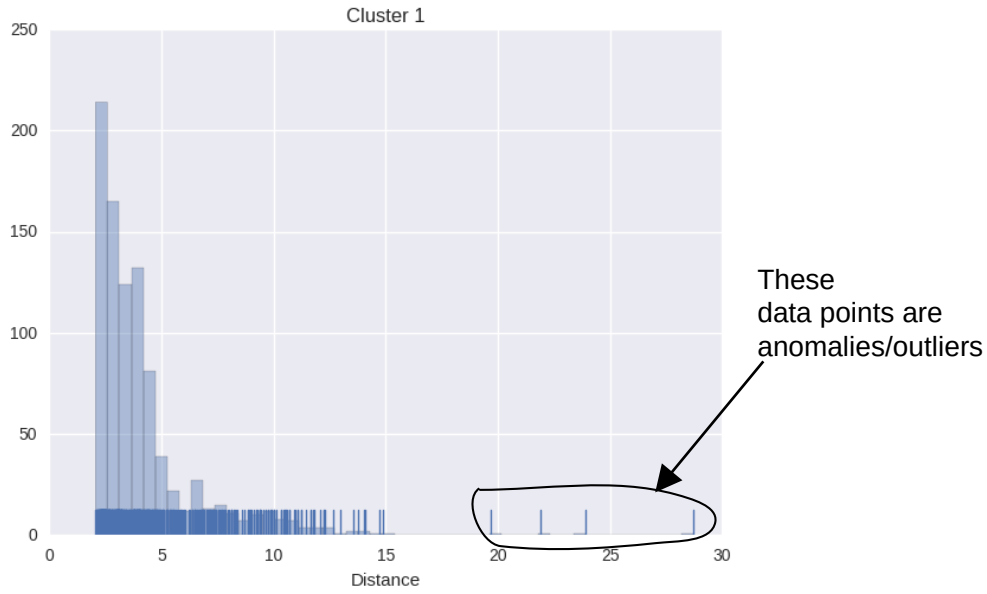


DABAI

VISMA®

Anomaly Detection using K-means clustering can be used to clean data

On the other hand, anomalies can also be “data of interest” which means, that a lot of value can potentially come from examining them.



DABAI

VISMA

Detecting anomalies in the Danish Business Registry Data (CVR-data)

We found that some companies are anomalies, compared to others, on a subset of features in the CVR-data from the Danish Business Authority.

Prototypes that define this cluster

cvrNummer	distance	prediction	navn
21147206	2.066353568196321	1	MURERFIRMAET BENT KLAUSEN
19318478	2.0884896692133625	1	GR GRUPPEN
17693875	2.11493972553152	1	BOWL'N' FUN, HORSENS
50494810	2.1200173105370186	1	MURERMESTER ERIK ANDERSEN AF GILLELEJE
64139711	2.1228699304318637	1	KNUD JENSEN OG SØNNER. SLIMMINGE

Outliers in this particular cluster

cvrNummer	distance	prediction	navn
15504749	28.698234527997432	1	BLUE HORS
43352911	23.901648914674382	1	VASCO GROUP
12937806	21.92192643393095	1	SCHELENBORG GODS
75152914	19.65885816641198	1	RENGØRINGSSELSKABET AF 1984
10279488	14.896399168468225	1	MASAI CLOTHING COMPANY

DABAI

VISMA

Bankruptcy prediction using classification on the Danish Business Registry Data (CVR-data)

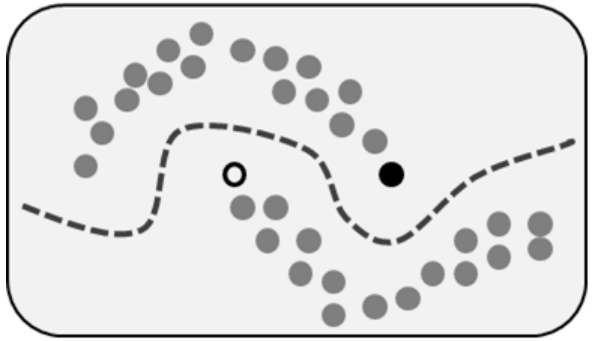
Our analysis shows that the latest amount of “årsværk” and number of “closed production units” are significant in respect to keeping a company from going bankrupt.

On the other hand, number of “open production units”, the second latest amount of “årsværk” are significant in respect to a company that has gone bankrupt.

What’s next?

Semi supervised learning:

We can use a few labeled points with unlabeled data.



Black/White data points: Labeled data.
Grey data points: Unlabeled data.

Created by: Techerin

DABAI

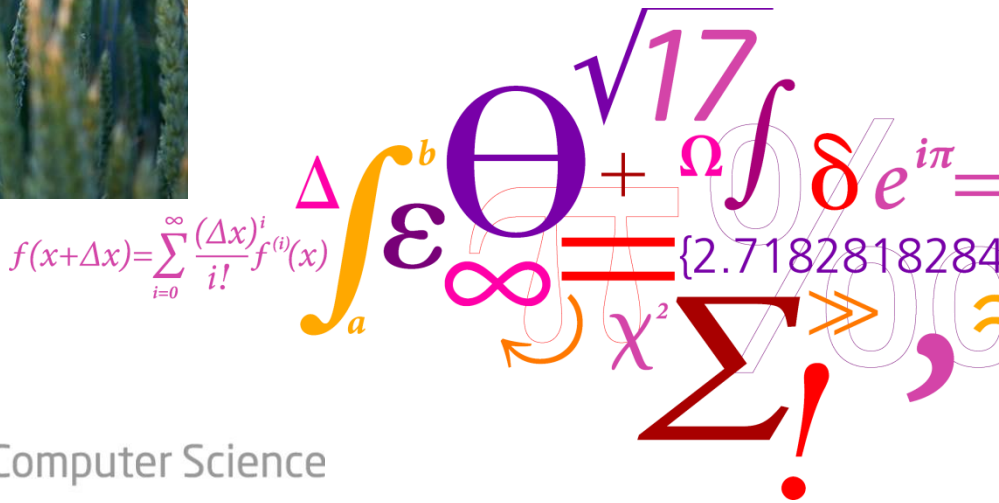
VISMA®


Thank you for
your attention
VISMA

Big Data in the Food Supply Chain

Methods for handling missing data

Niels Bruun Ipsen



DTU Compute
Department of Applied Mathematics and Computer Science

Setting

- Increased use of Big Data methods within the Food Supply Chain[1][2]

Setting

- Increased use of Big Data methods within the Food Supply Chain[1][2]
- Missing data reasons: corrupted, expensive, unknown

Setting

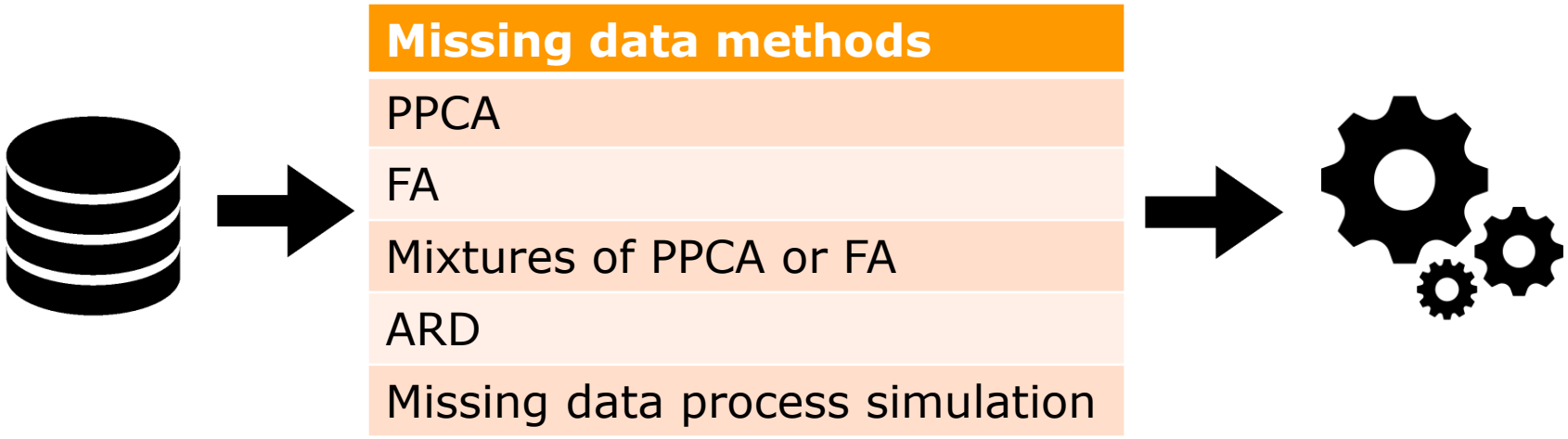
- Increased use of Big Data methods within the Food Supply Chain[1][2]
- Missing data reasons: corrupted, expensive, unknown
- Influence by missing data limits performance [3]

Setting

- Increased use of Big Data methods within the Food Supply Chain[1][2]
- Missing data reasons: corrupted, expensive, unknown
- Influence by missing data limits performance [3]
- How to handle missing data in a formal way in a Big Data context?

Setting

- Increased use of Big Data methods within the Food Supply Chain[1][2]
- Missing data reasons: corrupted, expensive, unknown
- Influence by missing data limits performance [3]
- How to handle missing data in a formal way in a Big Data context?



Project Outline

- Probabilistic PCA
  - Subspace estimation
  - Posterior probability distribution
  - Robustness

Project Outline

- Probabilistic PCA
  - Subspace estimation
  - Posterior probability distribution
  - Robustness
- Generalization
  - Factor Analysis, mixtures

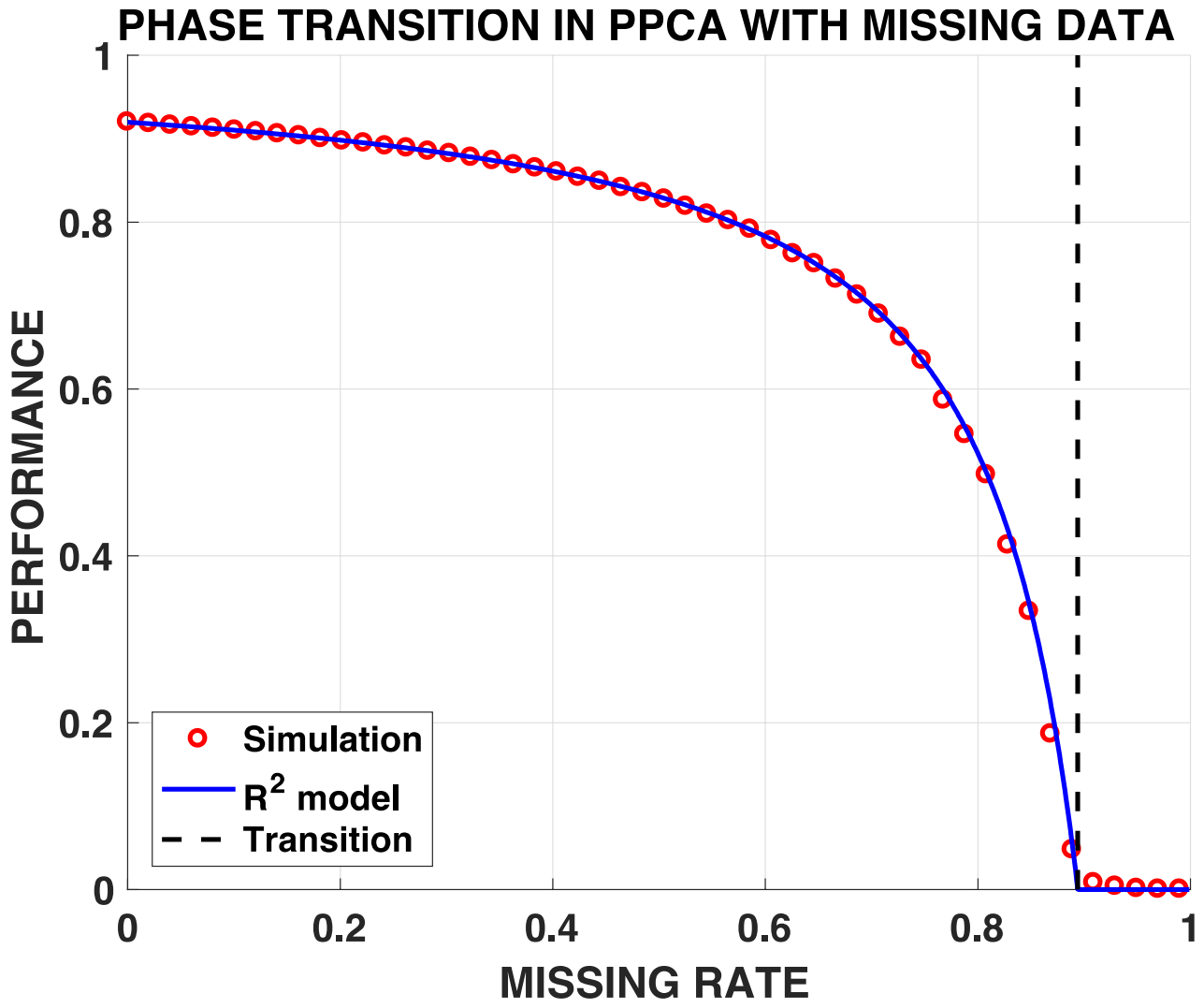
Project Outline

- Probabilistic PCA
  - Subspace estimation
  - Posterior probability distribution
  - Robustness
- Generalization
  - Factor Analysis, mixtures
- Automation
  - Automatic Relevance Determination, MLaaS

Project Outline

- Probabilistic PCA
  - Subspace estimation
  - Posterior probability distribution
  - Robustness
- Generalization
  - Factor Analysis, mixtures
- Automation
  - Automatic Relevance Determination, MLaaS
- Process estimation


PHASE TRANSITION IN PPCA WITH MISSING DATA
PERFORMANCE
MISSING RATE
0
0.2
0.4
0.6
0.8
1
n = 2500
n = 5000
n = 7500
n = 10000

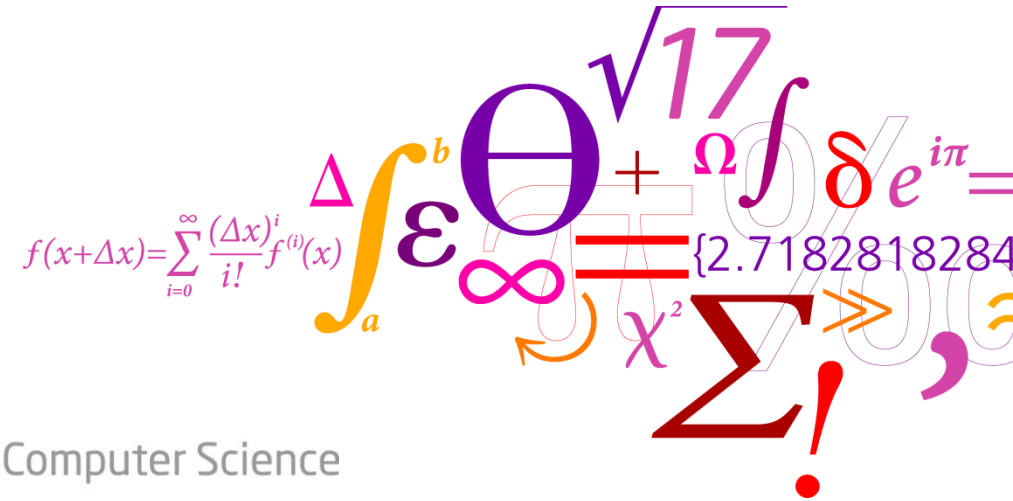

PHASE TRANSITION IN PPCA WITH MISSING DATA
PERFORMANCE
MISSING RATE
0
0.2
0.4
0.6
0.8
1
Simulation
R2 model
Transition

Thank you

[1] Lokers, Rob, et al. "Analysis of Big Data technologies for use in agro-environmental science."

[2] Marvin, Hans JP, et al. "A holistic approach to food safety risks: Food fraud as an example."

[3] Anagnostopoulos, Christos, and Peter Triantafillou. "Scaling out big data missing value imputations: pythia vs. godzilla."

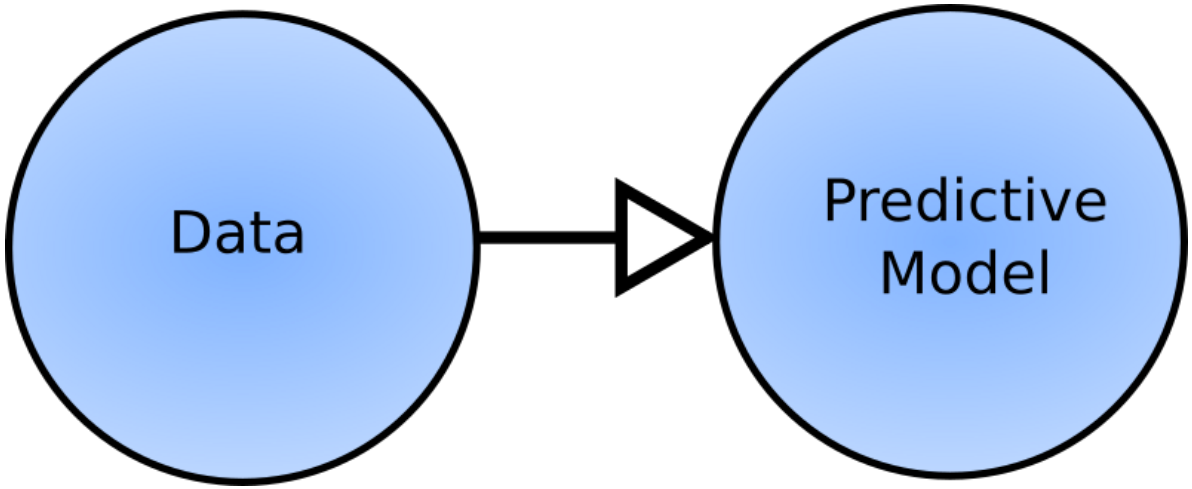


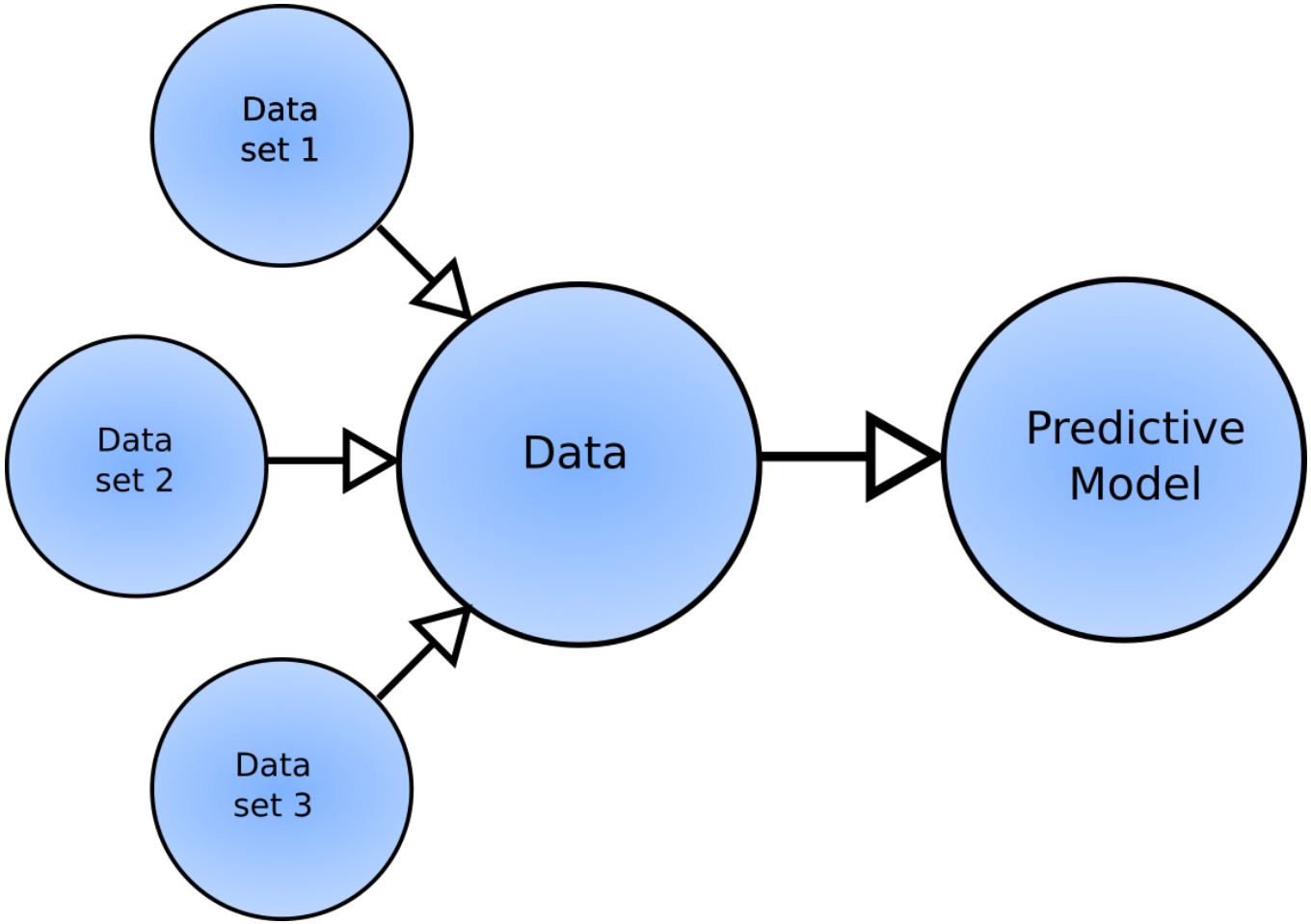
DTU Compute
Department of Applied Mathematics and Computer Science

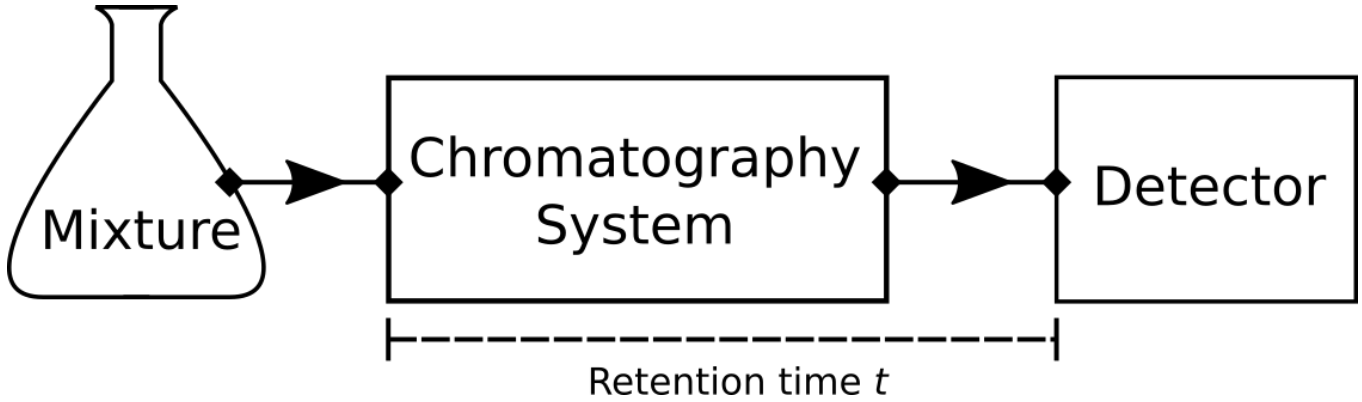
Integrating Big Data in Food

Philip Johan Havemann Jørgensen, Ph.d. student

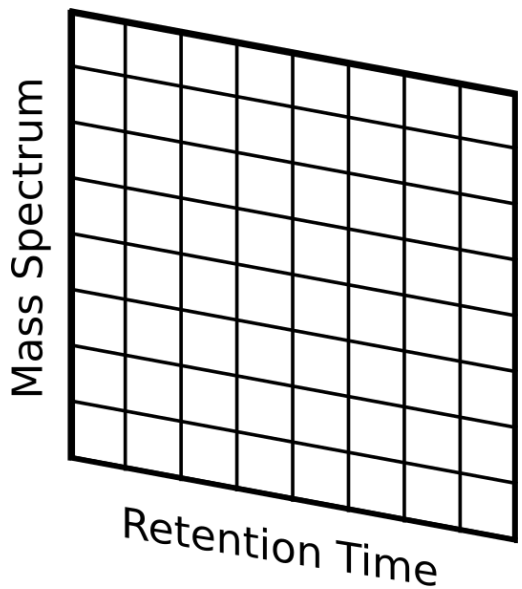
DTU Compute
Institut for Matematik og Computer Science

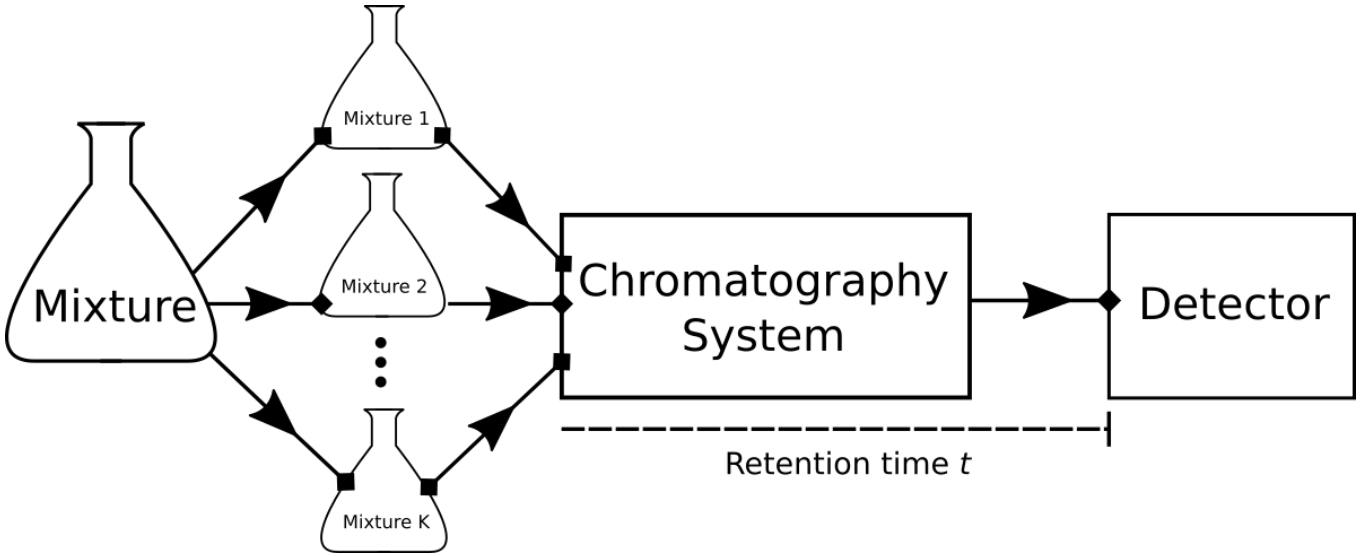

Data
Predictive
Model


Data
set 1
Data
set 2
Data
set 3
Data
Predictive
Model

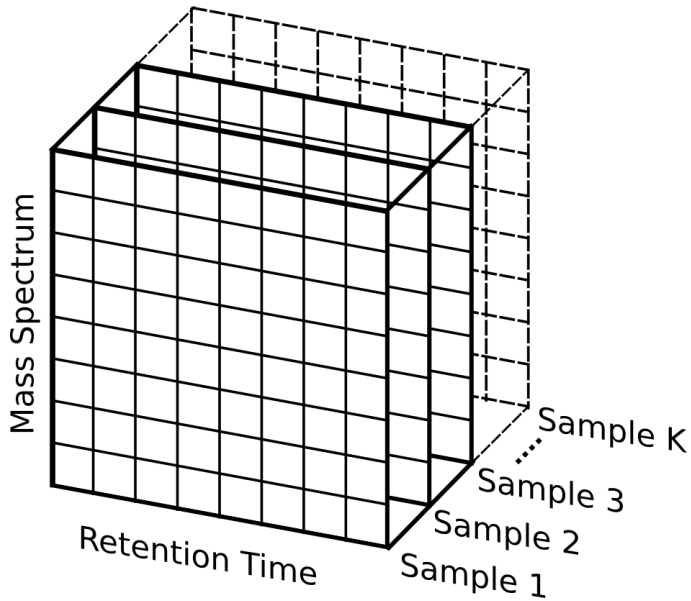


Measurements for mass spectrum $\times$ retention time

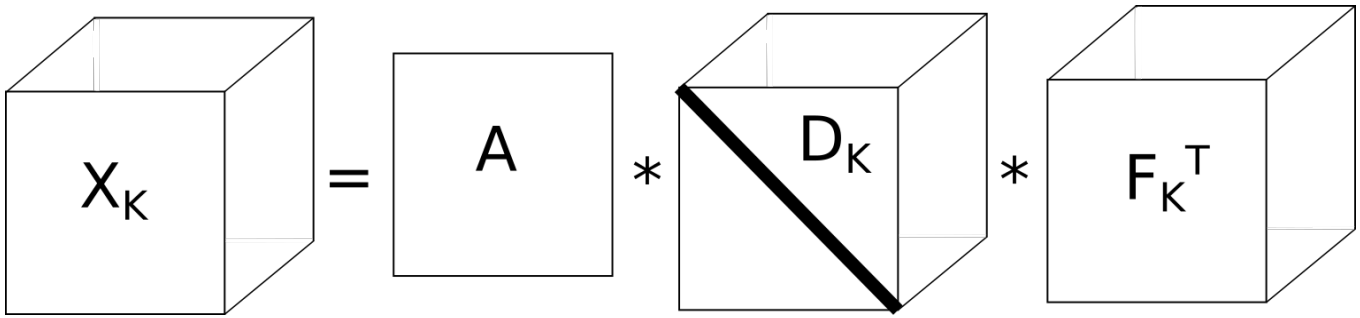




Measurements for mass spectrum $\times$ retention time $\times$ **samples**

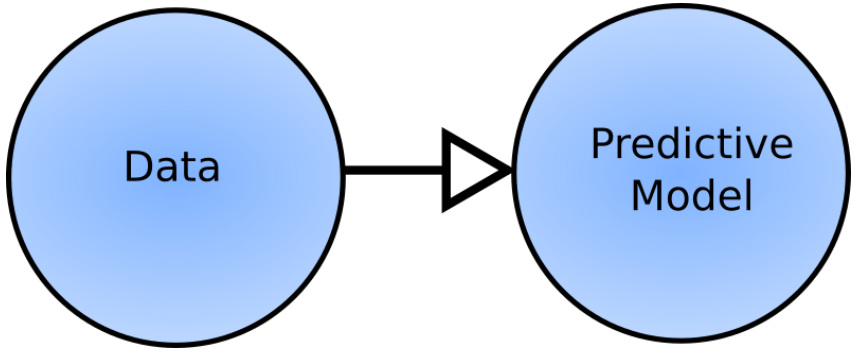


Tensor Factorization (Parafac2):



$$\mathbf{X_k} = \mathbf{A}\mathbf{D}_k\mathbf{F}_k^T$$

Key challenge: Determining the correct number of components (Trying to use a probabilistic formulation to solve it)



- Capturing relations in multimodal data
  - Data Fusion
- Improving Predictive Analysis
  - Transfer Learning/Domain Adaptation

DTU

Thank you!

DABAI

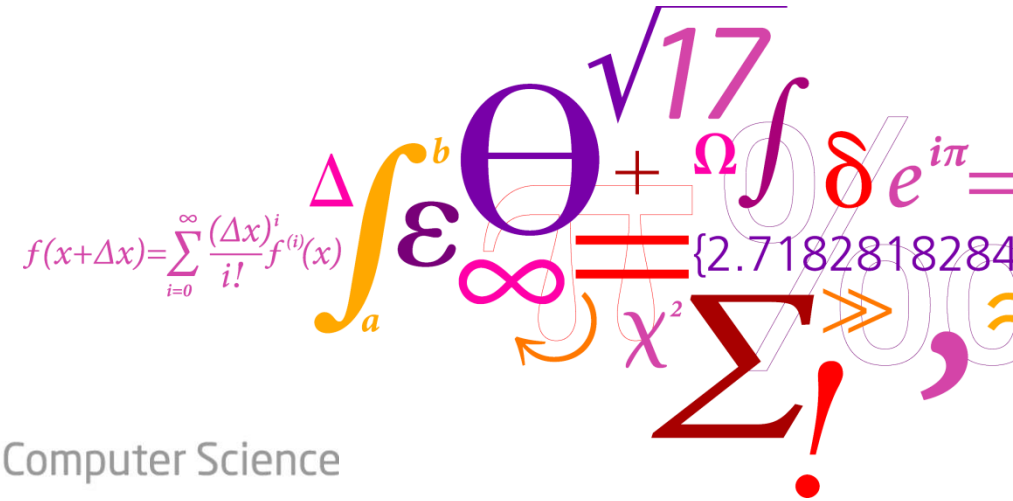
DTU

Knowing Nothing

- Computers and Semantics in Text

Jeppe Nørregaard

PhD Student with Lars Kai Hansen as supervisor



DTU Compute
Department of Applied Mathematics and Computer Science

People interact with computers

People interact with computers … and other people

People interact with computers … and other people

Doesn’t know what it’s selling

Motivations

Imagine a computer that…

- "knew" Wikipedia

Fact Sheet

U.S. Army Europe

Public Affairs Office
Tel: 0611-143-537-0005, FAX: 0611-705-3049 DSN: 314-537-0005
www.eur.army.mil | usarmyeurope.contact@mail.mil

Atlantic Resolve

Armored and Aviation Brigade Rotations Overview:

- Nine-month rotations scheduled into the foreseeable future
- Enhances the deterrence capabilities, increases ability to respond to potential crises and defend our allies and partners in the European community
- Remains under U.S. command
- Focuses on strengthening capabilities, sustaining readiness through bilateral and multinational training and exercises

4th Infantry Division, Mission Command Element:

- Based in Baumholder, Germany, and has been the regionally aligned division headquarters for Europe since 2015
- Oversees rotational units, tactical headquarters for U.S. land forces
- Provides U.S. Army Europe a division-level command and control capability

3rd Armored Brigade Combat Team, 4th Infantry Division Overview:

- Will begin arriving in January 2017 from Colorado and will bring ~3,500 personnel, 87 tanks, 18 Paladins; 419 Humvee variants; 144 Bradley fighting vehicles (446 tracked vehicles, 907 wheeled vehicles, 650 trailers)
- Beginning of armored brigade continuous presence and back-to-back rotations of U.S. troops and equipment

Fact Sheet

U.S. Army Europe

Public Affairs Office
Tel: 0611-143-537-0005, FAX: 0611-705-3049 DSN: 314-537-0005
www.eur.army.mil | usarmyeurope.contact@mail.mil

Atlantic Resolve

Armored and Aviation Brigade Rotations Overview:

- Nine-month rotations scheduled into the foreseeable future
- Enhances the deterrence capabilities, increases ability to respond to potential crises and defend our allies and partners in the European community
- Remains under U.S. command
- Focuses on strengthening capabilities, sustaining readiness through bilateral and multinational training and exercises

4th Infantry Division, Mission Command Element:

- Based in Baumholder, Germany, and has been the regionally aligned division headquarters for Europe since 2015
- Oversees rotational units, tactical headquarters for U.S. land forces
- Provides U.S. Army Europe a division-level command and control capability

3rd Armored Brigade Combat Team, 4th Infantry Division Overview:

- Will begin arriving in January 2017 from Colorado and will bring ~3,500 personnel, 87 tanks, 18 Paladins; 419 Humvee variants; 144 Bradley fighting vehicles (446 tracked vehicles, 907 wheeled vehicles, 650 trailers)
- Beginning of armored brigade continuous presence and back-to-back rotations of U.S. troops and equipment

March, 27th

16:36 Politics
Lavrov urges Europe to work harder towards implementing Minsk deal

16:00 Ukraine
Ukrainian command covers missile shortages by investigation commission's aircrash

15:32 Ukraine
Yarosh demands to forbid entry for Ukrainians working in Russia

15:03 Politics
Kiev answers Donbass humanitarian program by pushing Republics away

14:25 DPR Situation Reports
Donetsk Defense: Situation Report, 03/27/2017

13:52 LPR Situation Reports
Lugansk Defense: Situation Report,

US sends 3,600 tanks against Russia - Massive NATO deployment underway

Wednesday, January 4, 2017 - 16:43

The NATO war preparation against Russia, 'Operation Atlantic Resolve', is in full swing. 2,000 US tanks will be sent in coming days from Germany to Eastern Europe, and 1,600 US tanks is deployed to storage facilities in the Netherlands. At the same time, NATO countries are sending thousands of soldiers in to Russian borders.

According to US Army Europe, 4,000 troops and 2,000 tanks will arrive in three US transport ships to Germany next weekend. From Bremerhaven, US troops and huge amount of military material, will be transported to Poland and other countries in Central and Eastern Europe.

USA is sending to Russian borders 3rd Brigade of the 4th Infantry Division. Overall, more than 2,500 pieces of cargo are shipped to Germany, where those will be unloaded in the period January 6-8. US military material and troops will continue to Poland by rail and military convoy's. Massive US military deployment should be ready by January 20.

"Some 900 cars with military materiel will be transported by train from Bremerhaven to Poland. There are also about 600 pieces of freight that will be transported by train to Poland from the military training ground at Bergen-Hohne. Nearly 40 vehicles will travel directly by road from Bremerhaven to Poland," told Bundeswehr press office.

> *"Three years after the last American tanks left the continent, we need to get them back," said Lieutenant General Frederick "Ben" Hodges, commander of US forces in Europe.*

Donetsk city:
16:46:17
March 27th, 2017

Editorial

DPR to prevent new hostilities' round aspired by Kiev – Zakharchenko

Mar 27, 2017 | POLITICS, DEFENCE, EDITORIAL

DPR, LPR succeed in information security – Finnish journalist

March, 27th

16:36 Politics
Lavrov urges Europe to work harder towards implementing Minsk deal

16:00 Ukraine
Ukrainian command covers missile shortages by investigation commission's aircrash

15:32 Ukraine
Yarosh demands to forbid entry for Ukrainians working in Russia

15:03 Politics
Kiev answers Donbass humanitarian program by pushing Republics away

14:25 DPR Situation Reports
Donetsk Defense: Situation Report, 03/27/2017

13:52 LPR Situation Reports
Lugansk Defense: Situation Report,

US sends 3,600 tanks against Russia - Massive NATO deployment underway

Wednesday, January 4, 2017 - 16:43

The NATO war preparation against Russia, 'Operation Atlantic Resolve', is in full swing. 2,000 US tanks will be sent in coming days from Germany to Eastern Europe, and 1,600 US tanks is deployed to storage facilities in the Netherlands. At the same time, NATO countries are sending thousands of soldiers in to Russian borders.

According to US Army Europe, 4,000 troops and 2,000 tanks will arrive in three US transport ships to Germany next weekend. From Bremerhaven, US troops and huge amount of military material, will be transported to Poland and other countries in Central and Eastern Europe.

USA is sending to Russian borders 3rd Brigade of the 4th Infantry Division. Overall, more than 2,500 pieces of cargo are shipped to Germany, where those will be unloaded in the period January 6-8. US military material and troops will continue to Poland by rail and military convoy's. Massive US military deployment should be ready by January 20.

"Some 900 cars with military materiel will be transported by train from Bremerhaven to Poland. There are also about 600 pieces of freight that will be transported by train to Poland from the military training ground at Bergen-Hohne. Nearly 40 vehicles will travel directly by road from Bremerhaven to Poland," told Bundeswehr press office.

> *"Three years after the last American tanks left the continent, we need to get them back," said Lieutenant General Frederick "Ben" Hodges, commander of US forces in Europe.*

Donetsk city:
16:46:17
March 27th, 2017

Editorial

DPR to prevent new hostilities' round aspired by Kiev – Zakharchenko
Mar 27, 2017 | POLITICS, DEFENCE, EDITORIAL

DPR, LPR succeed in information security – Finnish journalist

Fake News

~3.500 personnel == 3.600 tanks ?

Motivations

Imagine a computer that…

- "knew" Wikipedia

Motivations

Imagine a computer that…

- "knew" Wikipedia
- could fact check news

Motivations

Imagine a computer that…

- "knew" Wikipedia
- could fact check news
- perhaps a little Turing test?

We are currently working on

Giving computers their own memory

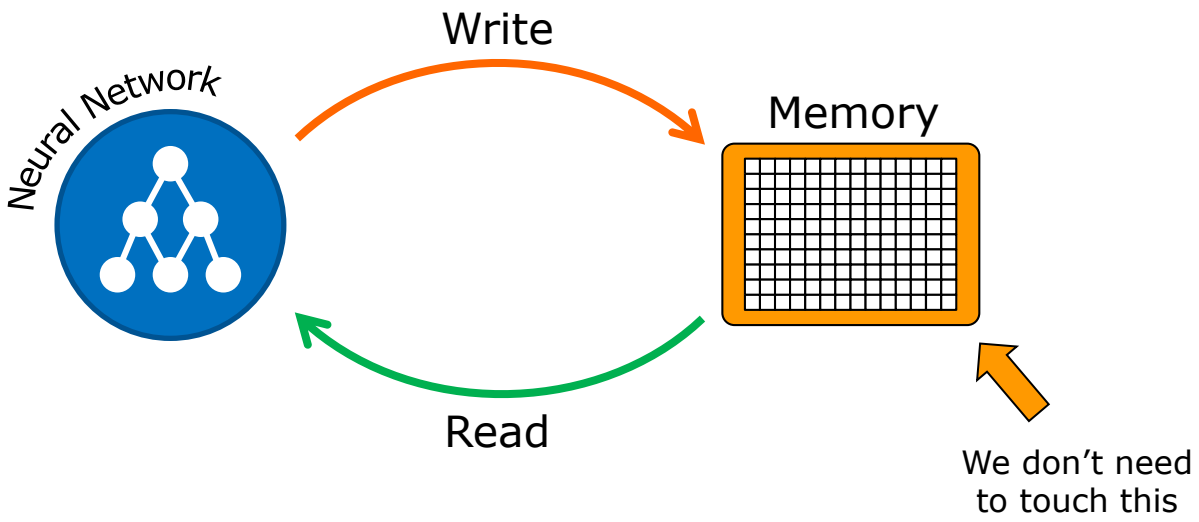
Exam time!

All knowledge in the universe

Exam time!

All knowledge you need

Differentiable Neural Computers[0]



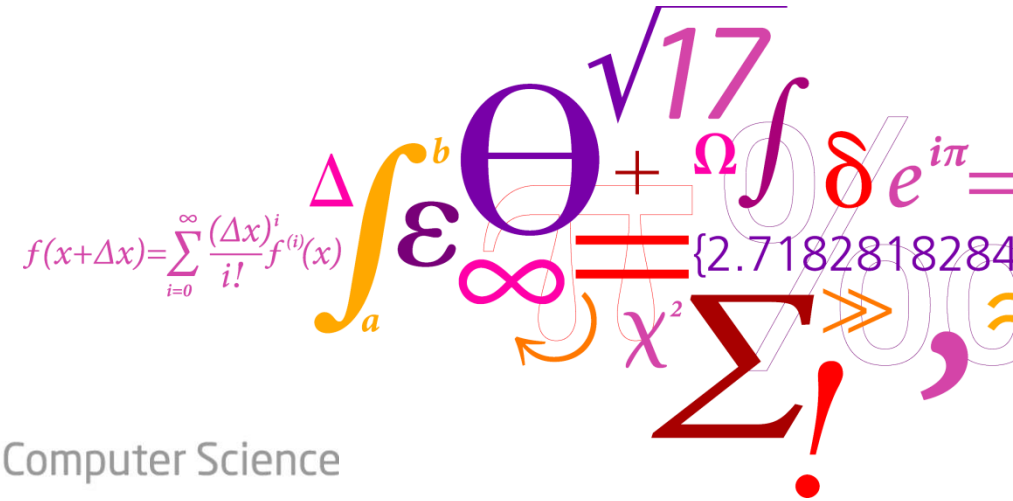
[0] Graves, Alex, et al. "Hybrid computing using a neural network with dynamic external memory." *Nature* 538.7626 (2016): 471-476.

DABAI

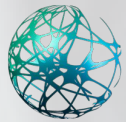
DTU

Thank You

Jeppe Nørregaard



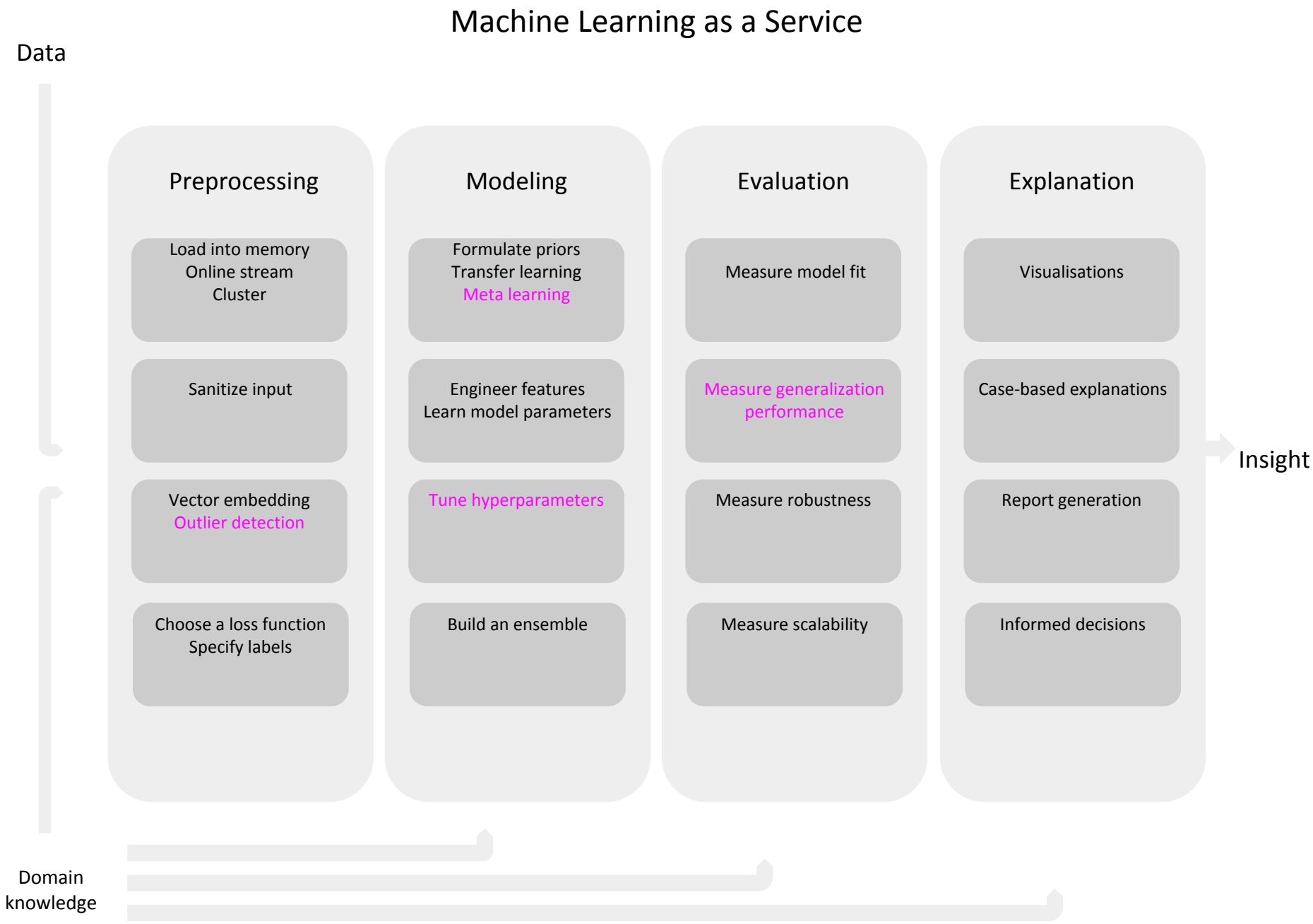
DTU Compute
Department of Applied Mathematics and Computer Science




Automating unsupervised learning

DABAI

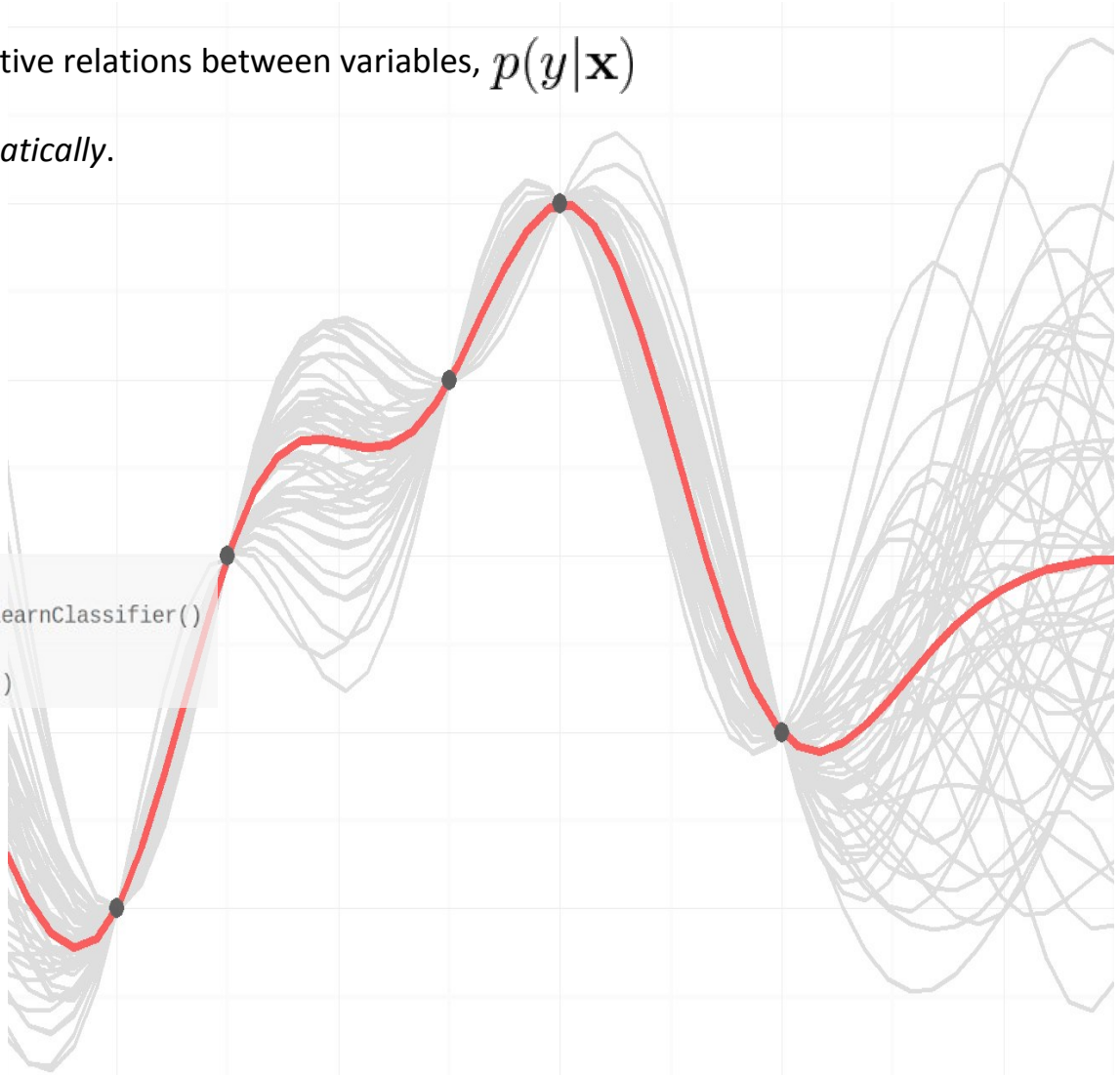
Frans Zdyb


Machine Learning as a Service
Data
Preprocessing
Load into memory
Online stream
Cluster
Sanitize input
Vector embedding
Outlier detection
Choose a loss function
Specify labels
Modeling
Formulate priors
Transfer learning
Meta learning
Engineer features
Learn model parameters
Tune hyperparameters
Build an ensemble
Evaluation
Measure model fit
Measure generalization performance
Measure robustness
Measure scalability
Explanation
Visualisations
Case-based explanations
Report generation
Informed decisions
Insight
Domain knowledge

Supervised learning finds predictive relations between variables, $p(y|\mathbf{x})$

There are systems that do this *automatically*.

```
>>> import autosklearn.classification
>>> cls = autosklearn.classification.AutoSklearnClassifier()
>>> cls.fit(X_train, y_train)
>>> predictions = cls.predict(X_test, y_test)
```



Auto-sklearn[1]
a wrapper around the scikit-learn, uses

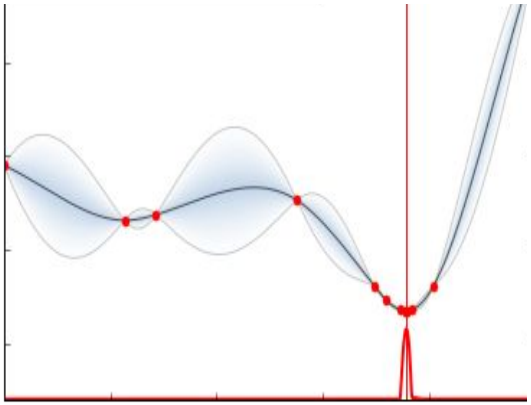
meta-learning,
Bayesian optimization and
ensemble building

to outperform the state-of-the-art on the ChaLearn AutoML Challenge.
Classification works really well. Regression is coming along nicely.

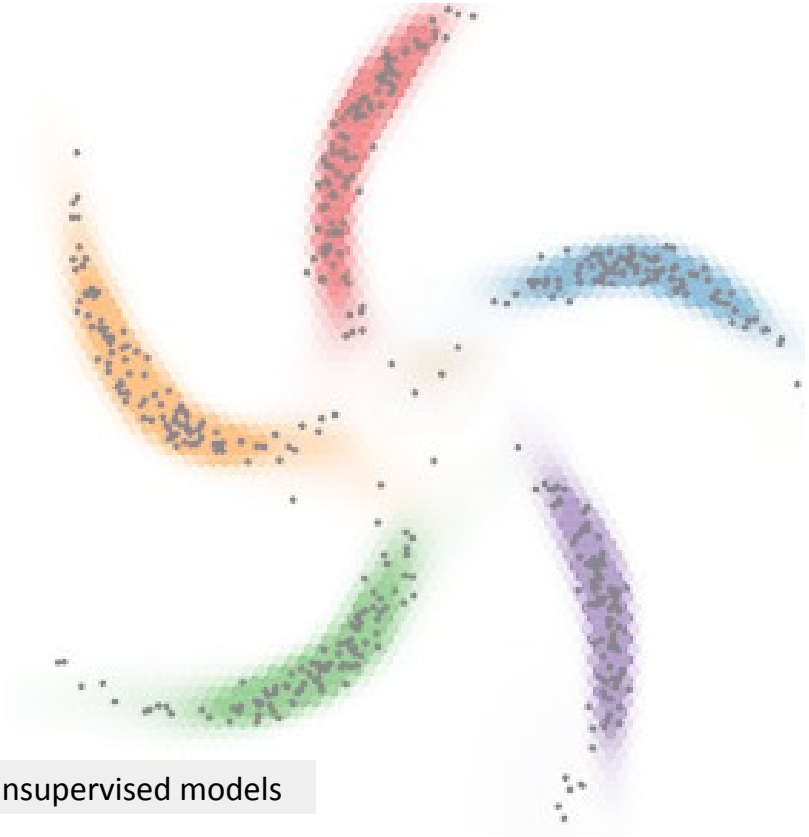
1 “Efficient and Robust Automated Machine Learning”, Hutter et al., 2015

Unsupervised learning finds generalizable dependencies between variables, $p(y, \mathbf{x})$

Automating it is largely *unexplored territory*.



Bayesian Optimization with Gaussian Process



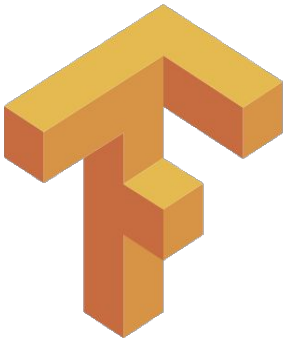
Hypothesis: We can use Bayesian Optimization to tune unsupervised models

- Generalize to unseen data
- Robust to different training sets
- Detect outliers
- Aid in supervised learning

Python + Numpy + Scipy

TensorFlow

for distributed numerical computing and automatic differentiation

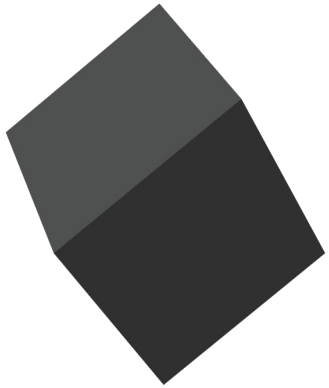


Edward[2]

for probabilistic modeling, built on top of TensorFlow

- Graphical models
- Neural networks
- Bayesian non-parametrics

- Variational Inference
- MCMC



GPyOpt[3]

for Bayesian Optimization

- Easy to use
- Parallel
- Up to date

2 Edward: A library for probabilistic modeling, inference, and criticism, 2016,
3 GPyOpt: A Bayesian Optimization framework in python, 2016,

`edwardlib.org`
`sheffieldml.github.io/GPyOpt/`


Thank you!