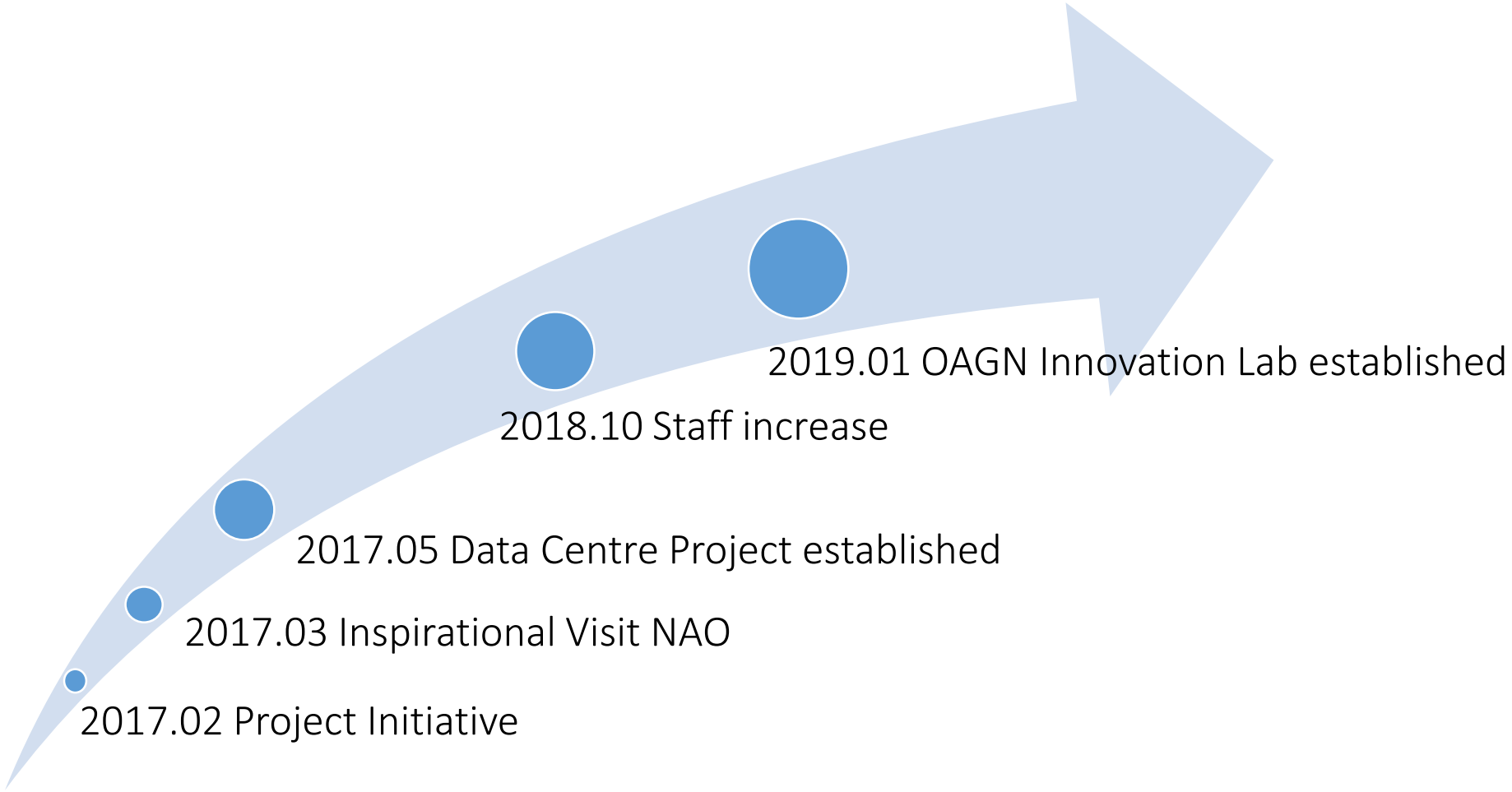
Data Science
&
Machine Learning

Musings and some examples from OAG Norway

Jan Roar
Chief Data Scientist
OAGN Innovation Lab

The OAGN Data Science Project



Purpose

1. Curating data for financial and performance audit
2. On-demand data analytics
3. Promoting data science and the use of machine learning at the OAGN
4. Experiment with new technologies and methods

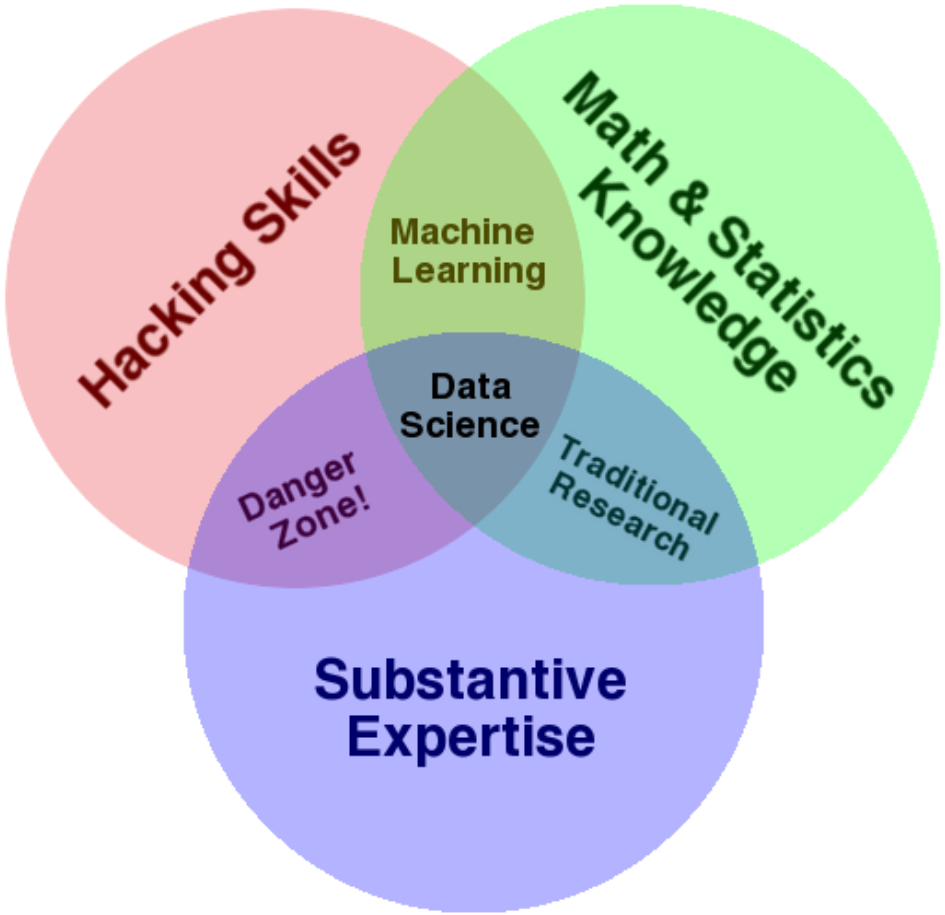
We automate the boring stuff,
so you can audit the exiting stuff!

Where we come from

- 2 political scientists
- 1 economist
- 1 sociologist
- 1 physicist

… we're not an IT project, we're a data science project

What is Data Science



You need both:

- Coding skills
- Statistical skills
- Audit skills

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The world is overflowing with data

This is difficult to manage for SAIs

Partly because of this

Law on The Office of the Auditor General of Norway § 12:

> «The OAGN can, without restrictions of confidentiality, demand any information, any disclosure or any document, and execute any examination deemed necessary by the OAGN to fulfil audit tasks.»

(my own highly informal translation)

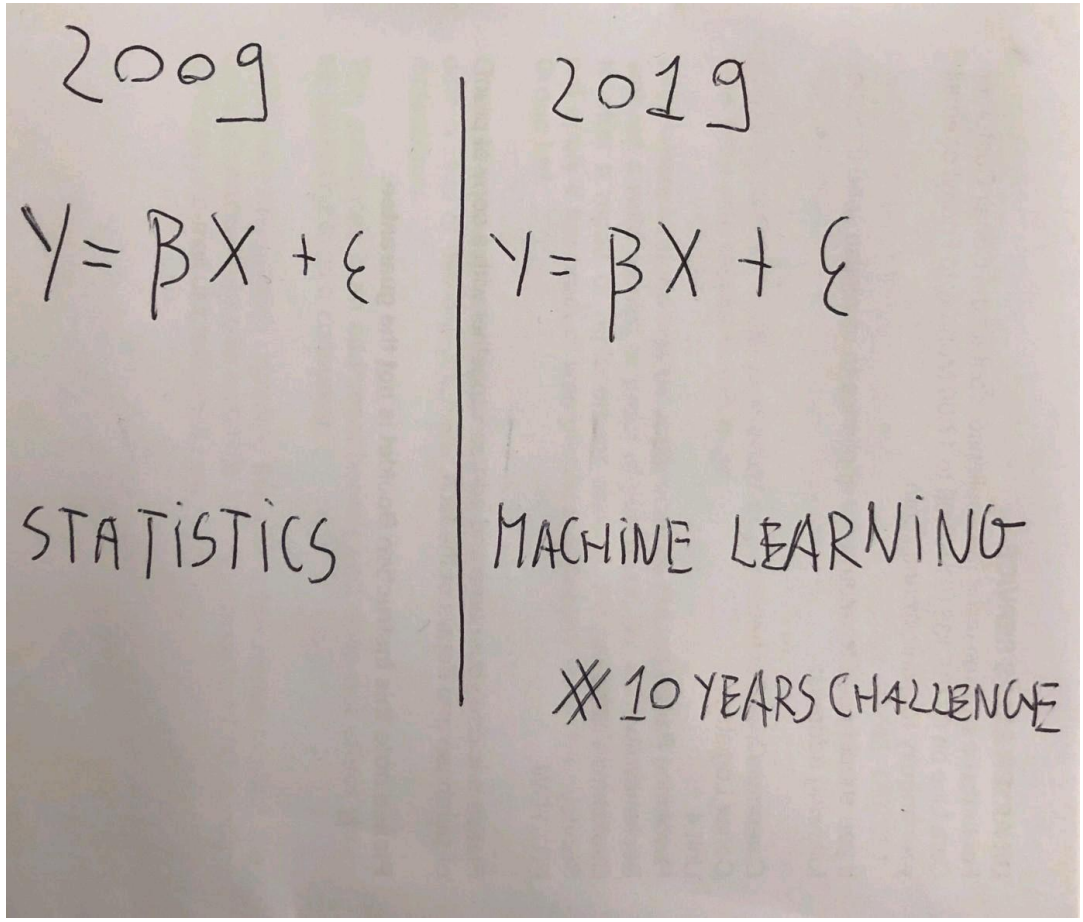
Data → Information → Knowledge → Wisdom

What’s all this about machine learning?

«Everybody» keeps fussing about
machine learning and artificial intelligence these days
… and you can get a bit worried…
… when seeing stuff like this…


Machine Learning Algorithms
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

However…



Found on LinkedIn

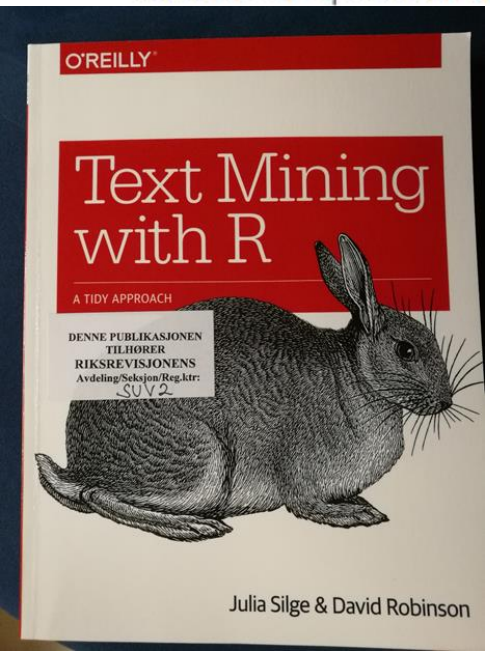
What’s the point – an example

- You have 10 000 pages of procurement contract documents (text)
- And you have regulations for procurement
- Then you want to compare the 10 000 pages and the regulations
- To uncover illegal purchases
- Can machines do this for us?

Natural Language Processing

What it is and why it matters

Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.



Source:

https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

NLP in practice – from speech to text

Google Cloud Why Google Solutions Products Pricing Getting started

AI & Machine Learning Products

Cloud Speech-to-Text

Speech-to-text conversion powered by machine learning and available for short-form or long-form audio.

GO TO CONSOLE

View documentation for this product.

Powerful speech recognition

Google Cloud Speech-to-Text enables developers to convert audio to text by applying powerful neural network models in an easy-to-use API. The API recognizes 120 languages and variants to support your global user base. You can enable voice command-and-control, transcribe audio from call centers, and more. It can process real-time streaming or prerecorded audio, using Google's machine learning technology.

Convert your speech to text right now

Select a language and click "Start Now" to begin recording

Input type
Microphone File upload

Language
English (United States)

Speaker diarization BETA
Off

Speakers
1 speaker

Punctuation

Google Home

Algorithms – Economic, but Effective?

- government have started using machine learning, for example in awarding grants
  - even for decisions involving discretionary considerations

- we will need performance auditors that can look into “the black box of algorithms”

- to evaluate and conclude on decisions made by machines

If you understand this,
you have understood machine learning

$$\mathrm{Y} = f(x) + e$$

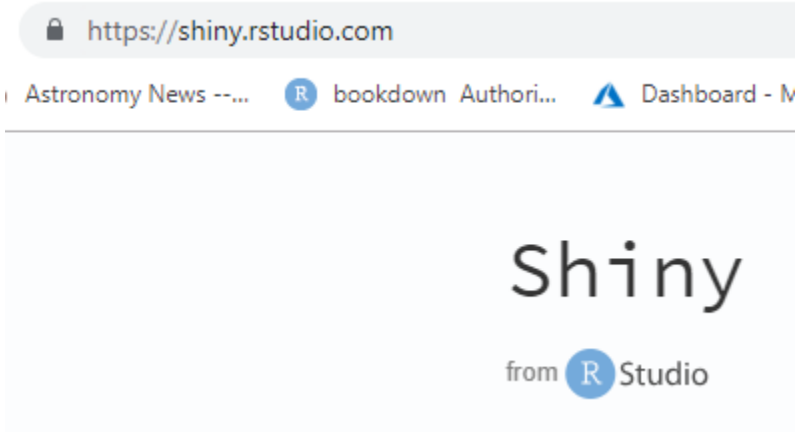
Some concrete examples

Example: Webapp for DEA-analysis

- DEA = Data Envelopment Analysis
- Benchmark analysis based on Output vs. Input
- Example:
  - District courts – how many judges/office clerks vs. how many cases handled
- Productivity Analysis

DEA – efficiency in the form of productivity

- DEA an important method for performance audit, but
- Most software for DEA is old and/or user unfriendly
- Several packages in R for doing DEA # but not easy to use
- So: we wrote a Shiny app for DEA
- We named it pioneeR



We’re open!

pioneeR is open source (free to use and free to edit the source code)!

You can find us on GitHub*:

github.com/Riksrevisjonen/pioneeR

*The largest collection of open source software on the Internet

Takeaway:

- Program code is universal

- As such very easy to share

Search: 19 000 PDFs – what to do

- We have 19 000 PDF documents from Norwegian hospital boards
- 25 hospital boards, 12-14 board meetings a year, last 5 years)
- All documents published on the web (25 different websites)

RiksRevisjonen Helsedokun × | Elasticsearch: RESTful, Dist ×

Sikker | https://www.elastic.co/products/elasticsearch

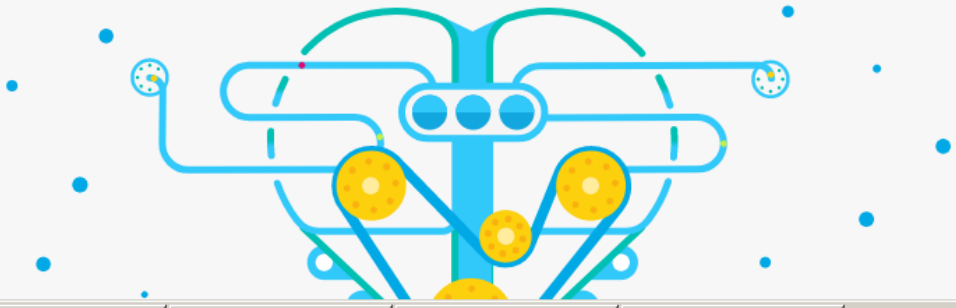
Apper | Ove Haugland Jakobse | hack4no | Riksreven Sl | Google | 18.196.12.166 | Riksreven (riksreven) | Overview - Microsoft Te | Helsedokumentsøk | RiksRevisjonen Helsedo

elastic Products Cloud Services Customers Learn downloads contact EN

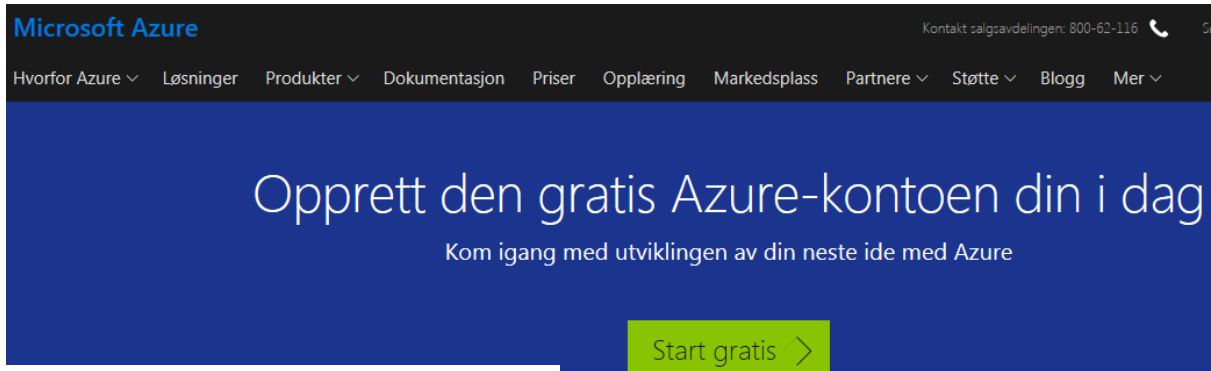
Elasticsearch Download As A Service

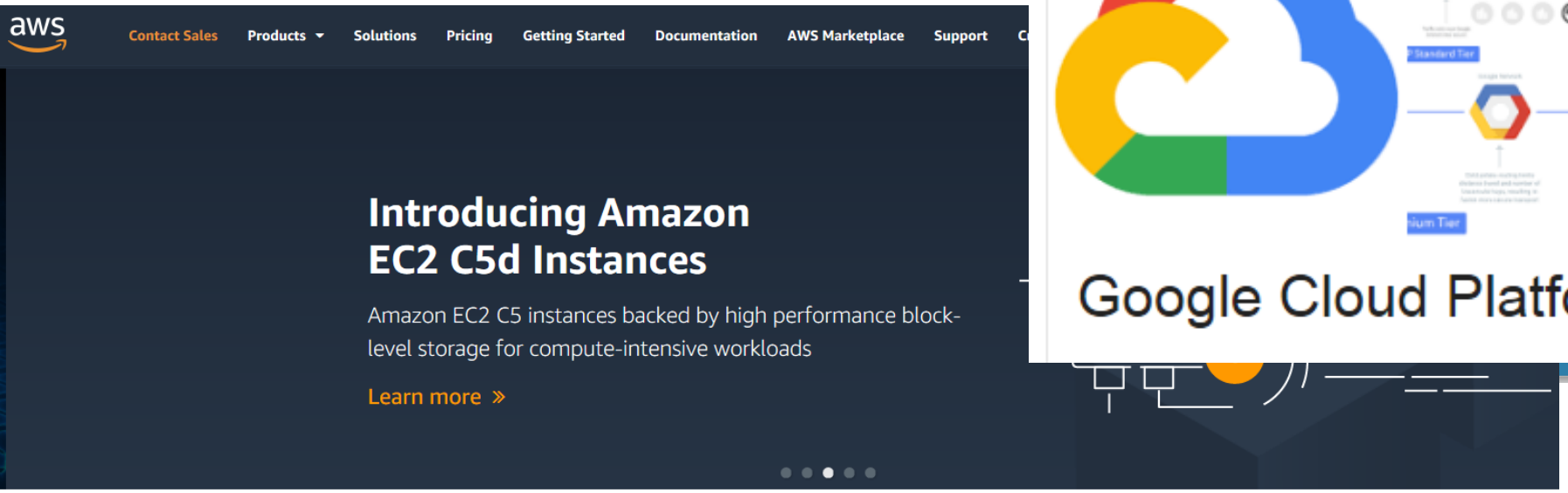
The Heart of the Elastic Stack

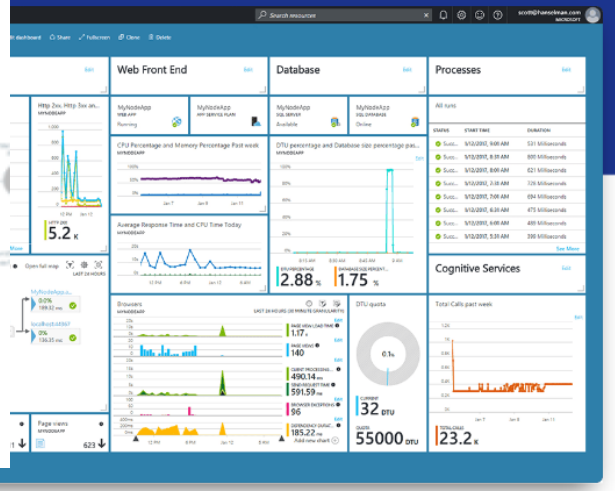
Elasticsearch is a distributed, RESTful search and analytics engine capable of solving a growing number of use cases. As the heart of the Elastic Stack, it centrally stores your data so you can discover the expected and uncover the unexpected.

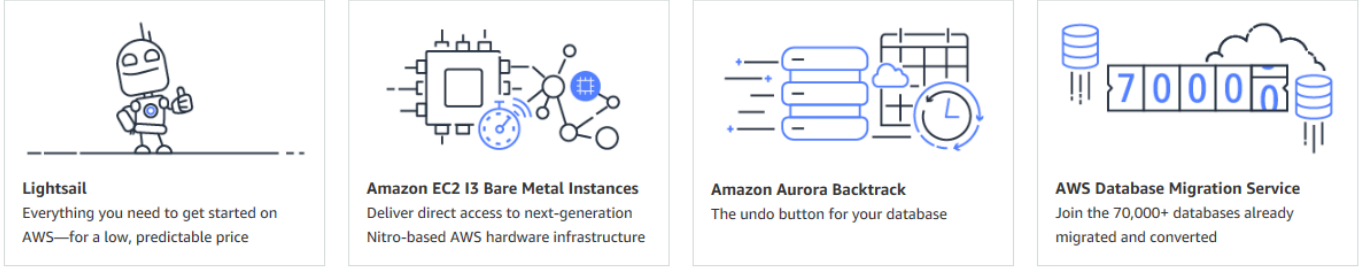


Cloud Platform: The big 3







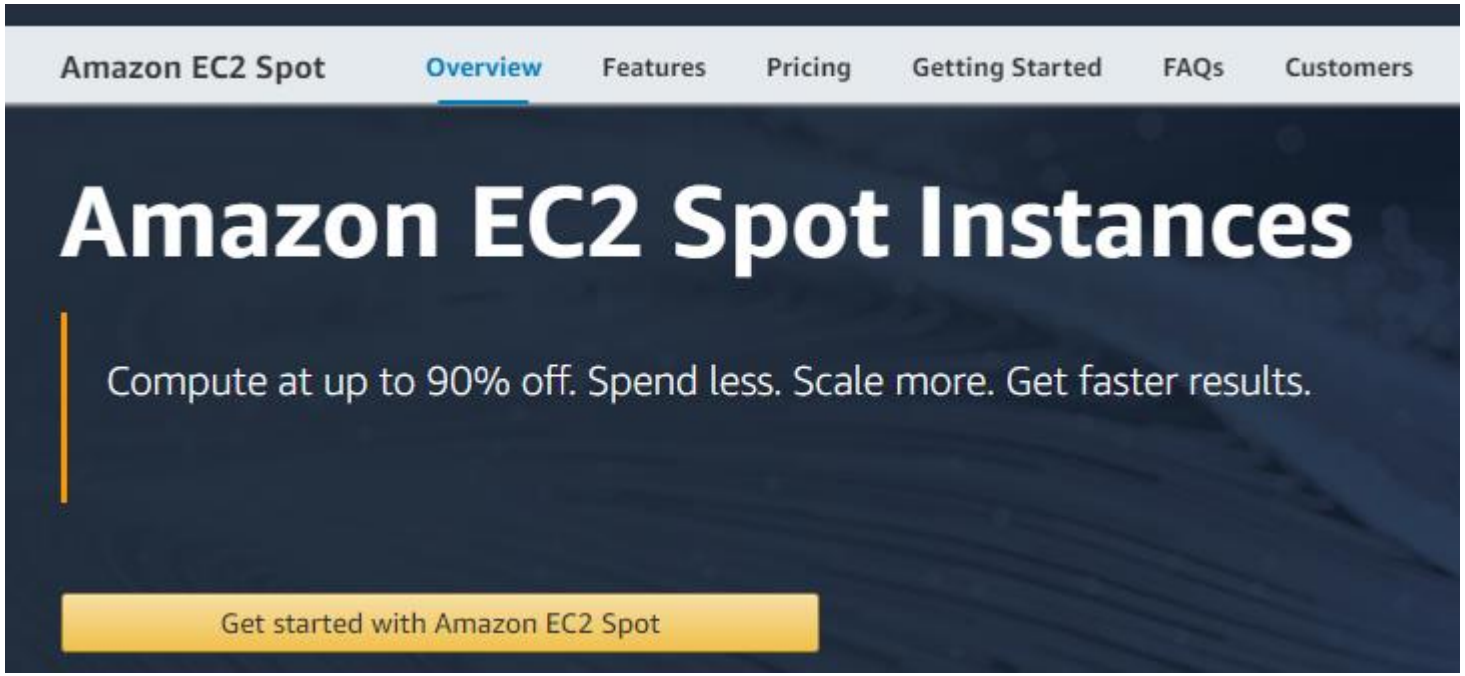


Cloud computing

- You need heavy computing machinery to run machine learning on large datasets

- Buy an on-prem machine for $ 10 000?

- Or rent one in the cloud for $ 50 a week?



http:/[redacted]8787/

RStudio

Fil Rediger Vis Favoritter Verktøy Hjelp

Astronomy News -- Scien... bookdown Authoring Bo... Dashboard - Microsoft Az... general Riksrevisjonen Ny... general SAI_Data_Sci_Coll... Google http--forum-sider-startside Microsoft Demos Ove Haugland Jakobsen -

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Console Terminal

```
                 from httpuv.cpp:12:
/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include/boost/smart_ptr/shared_ptr.hpp: In member function ‘boost::shared_ptr<T>
& boost::shared_ptr<T>::operator=(std::auto_ptr<_Up>&&)’:
/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include/boost/smart_ptr/shared_ptr.hpp:578:38: warning: ‘template<class> class st
d::auto_ptr’ is deprecated [-Wdeprecated-declarations]
         this_type( static_cast< std::auto_ptr<Y> && >( r ) ).swap( *this );
                                      ^~~~~~~~~
In file included from /usr/include/c++/7/bits/locale_conv.h:41:0,
                 from /usr/include/c++/7/locale:43,
                 from /usr/include/c++/7/iomanip:43,
                 from httpuv.cpp:5:
/usr/include/c++/7/bits/unique_ptr.h:51:28: note: declared here
   template<typename> class auto_ptr;
                            ^~~~~~~~~
gcc -std=gnu99 -I"/usr/share/R/include" -DNDEBUG -Ilibuv/include -pthread -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/Rcpp/in
clude" -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include" -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/later/inclu
de"   -fvisibility=hidden -DSTRICT_R_HEADERS -fpic  -g -O2 -fdebug-prefix-map=/build/r-base-3U0YWo/r-base-3.5.3=. -fstack-protector-
strong -Wformat -Werror=format-security -Wdate-time -D_FORTIFY_SOURCE=2 -g  -c md5.c -o md5.o
g++ -std=gnu++11 -I"/usr/share/R/include" -DNDEBUG -Ilibuv/include -pthread -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/Rcpp/
include" -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include" -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/later/inc
lude"   -fvisibility=hidden -DSTRICT_R_HEADERS -fpic  -g -O2 -fdebug-prefix-map=/build/r-base-3U0YWo/r-base-3.5.3=. -fstack-protecto
r-strong -Wformat -Werror=format-security -Wdate-time -D_FORTIFY_SOURCE=2 -g -c mime.cpp -o mime.o
gcc -std=gnu99 -I"/usr/share/R/include" -DNDEBUG -Ilibuv/include -pthread -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/Rcpp/in
clude" -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include" -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/later/inclu
de"   -fvisibility=hidden -DSTRICT_R_HEADERS -fpic  -g -O2 -fdebug-prefix-map=/build/r-base-3U0YWo/r-base-3.5.3=. -fstack-protector-
strong -Wformat -Werror=format-security -Wdate-time -D_FORTIFY_SOURCE=2 -g  -c set_invalid_parameter.c -o set_invalid_parameter.o
g++ -std=gnu++11 -I"/usr/share/R/include" -DNDEBUG -Ilibuv/include -pthread -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/Rcpp/
include" -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include" -I"/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/later/inc
lude"   -fvisibility=hidden -DSTRICT_R_HEADERS -fpic  -g -O2 -fdebug-prefix-map=/build/r-base-3U0YWo/r-base-3.5.3=. -fstack-protecto
r-strong -Wformat -Werror=format-security -Wdate-time -D_FORTIFY_SOURCE=2 -g -c socket.cpp -o socket.o
In file included from /home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include/boost/smart_ptr/shared_ptr.hpp:28:0,
                 from /home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include/boost/shared_ptr.hpp:17,
                 from http.h:5,
                 from socket.h:4,
                 from socket.cpp:1:
/home/rstudio/R/x86_64-pc-linux-gnu-library/3.5/BH/include/boost/smart_ptr/detail/shared_count.hpp:355:33: warning: ‘template<class
> class std::auto_ptr’ is deprecated [-Wdeprecated-declarations]
     explicit shared_count( std::auto_ptr<Y> & r ): pi_( new sp_counted_impl_p<Y>( r.get() ) )
                                 ^~~~~~~~~
In file included from /usr/include/c++/7/memory:80:0,
```

Environment History Connections

Import Dataset

Global Environment

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Home

Name

.Rhistory

R

Summary: Some success factors

- We were given full freedom to experiment
- Recruited people from audit, not IT
- There were no detailed planning — only the “what”, not the “how”
- Full support from top management
- Short development cycles # agile development
- Prioritise solutions to long standing issues

www.saireports.org

A final snack before we adjourn