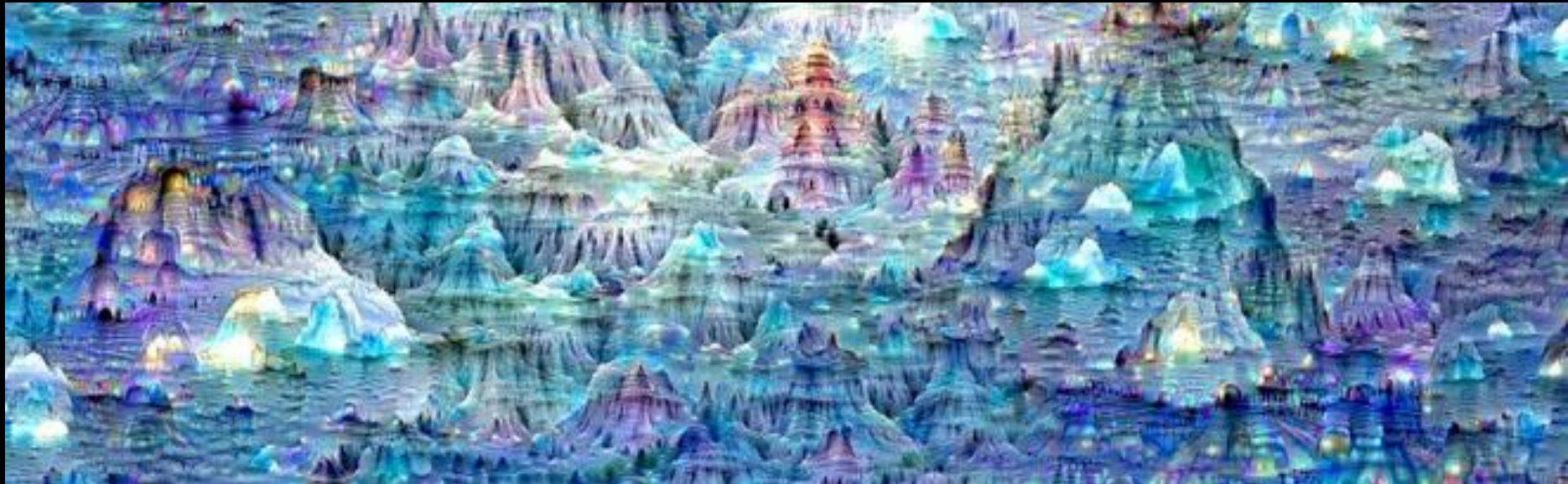


ADVANCED ANALYTICS AND DEEP LEARNING FOR BUSINESS

**Professor Rens Scheepers and
Assoc. Prof. Jacob Cybulski**

Dept of Info Sys and Bus Analytics

*Deakin Business School
Faculty of Business and Law
Deakin University*



WHAT IS ADVANCED ANALYTICS AND DEEP LEARNING

- **Advanced Analytics** is the autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools, typically beyond those of traditional business intelligence (BI), to discover deeper insights, make predictions, or generate recommendations (Gartner).
- Advanced analytic techniques include those such as data/text mining, machine learning, pattern matching, forecasting, visualization, semantic analysis, sentiment analysis, network and cluster analysis, multivariate statistics, graph analysis, simulation, complex event processing, neural networks.
- **Deep Learning** is a class of machine learning techniques which aim at building very large data mining models used for classification, estimation and clustering of data.
- **Neural Networks** are the most commonly used Deep Learning technique.
- Neural Networks consist of thousands of simpler models, called neurons, functionality of which is based on brain processes, which can be simulated with mathematical transformation of data.
- Special techniques have been developed to develop such large neural networks. As the networks are huge, the methods of neural network “training” are iterative.
- **GPUs**, the high-performance graphics cards, which have 1000s of processing cores, allow efficient creation and use of deep models.
- Deep learning packages, such as Tensorflow, TFLearn, Keras, MxNet, Caffe, CNTK, H2O, can be used from popular data analytics software, e.g. Anaconda, R / R Studio, RapidMiner, SAS, SPSS, Azure, etc.
- **Kaggle competitions** in data mining are being consistently won by international teams relying on deep learning solutions to competition problems.

SAMPLE APPLICATIONS

DEEP LEARNING, AI, MEDIA ANALYTICS

Traditional

- Game playing
- Weather/Climate prediction
- Disease diagnosis
- Image (Satellite) classification
- Image enhancement
- Face/Speech recognition
- Sound recovery
- CAT/MRI scan analysis
- Gravity (Astronomy) study
- Natural language processing
- Hand writing recognition
- Protein/Molecular analysis
- Drug design
- Brain mapping
- CCTV analysis
- Cyber attack detection
- Self-driving cars
- Robotics










Business








- Customer churn / risk analysis
- Demand forecasting
- Inventory analysis
- Stock market prediction
- Real-time sales analysis
- Credit rating analysis
- Insurance claim analysis
- Analysis of online user behaviour
- Prediction of real estate prices
- Inventory management
- Recommendation systems
- Fashion / style analytics
- Clothing, shoe, eyewear fitting
- Fraud and anomaly detection
- Financial auditing
- Classification of media releases
- Social media (text) sentiment analysis
- Visual (photo/video) sentiment analysis

*In spite of this
impressive list,
business applications
are still very few!*

KAGGLE COMPETITIONS IN DATA MINING

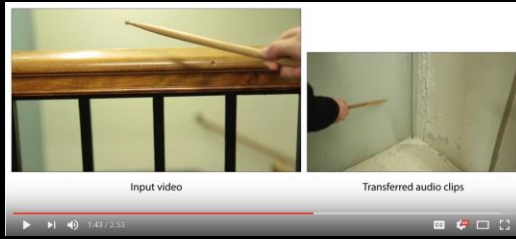
Current competitions (Sept 2017)

Active Competitions		
	Passenger Screening Algorithm Challenge Improve the accuracy of the Department of Homeland Security's threat recognition algorithms	3 months 198 teams 875 kernels \$1,500,000
	Zillow Prize: Zillow's Home Value Prediction (Ze...) Can you improve the algorithm that changed the world of real estate?	5 months 2263 teams 20778 kernels \$1,200,000
	Carvana Image Masking Challenge Automatically identify the boundaries of the car in an image	38 days 378 teams 985 kernels \$25,000
	Web Traffic Time Series Forecasting Forecast future traffic to Wikipedia pages	21 days 695 teams 4478 kernels \$25,000
	Personalized Medicine: Redefining Cancer Treat... Predict the effect of Genetic Variants to enable Personalized Medicine	43 days 828 teams 2939 kernels \$15,000
	NIPS 2017: Non-targeted Adversarial Attack Imperceptibly transform images in ways that fool classification models	42 days 400 kernels Swag
	NIPS 2017: Targeted Adversarial Attack Develop an adversarial attack that causes image classifiers to predict a specific target class	42 days 266 kernels Swag
	NIPS 2017: Defense Against Adversarial Attack Create an image classifier that is robust to adversarial attacks	42 days 274 kernels Swag
	ImageNet Object Detection Challenge Identify and label everyday objects in images	150 months Knowledge

Featured		All	Mine	Upvoted	Search datasets
17		All the news 143,000 articles from 15 American publications Andrew Thompson · updated 6 hours ago · journalism			156 downloads 3 comments
13		Cervical Cancer Risk Classification prediction of cancer indicators SURECOMMENDERS · updated 2 days ago			76 downloads 1 comment
513		Human Resources Analytics Why are our best and most experienced employees leaving prematurely? ludoben · updated 9 months ago · employment			28,846 downloads 89 comments
596		Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or genuine Andrea · updated 9 months ago · crime, finance			30,413 downloads 64 comments
10		#Charlottesville on Twitter A snapshot of American history in the making VincentLa · updated 2 days ago			59 downloads 1 comment
1		Austin 311 Calls 463k Public Complaints, 2013-17 Jacob Boysen · updated 2 days ago			4 downloads 0 comments
2		Crop Nutrient Database USDA data about crop nutrients in the U.S. Chris Crawford · updated 2 days ago · food and drink, science and culture, united states, pl...			6 downloads 0 comments

Past competitions
Very large data sets

SOME (VERY FAMOUS AND) RECENT DEEP LEARNING SYSTEMS



Adding sound to silent movies
<https://youtu.be/0FW99AQmMc8>



"little girl is eating piece of cake."



"baseball player is throwing ball in game."

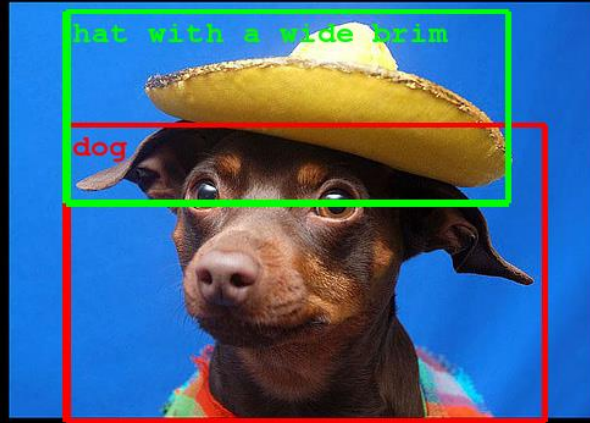


"a young boy is holding a baseball bat."



"a cat is sitting on a couch with a remote control."

Generation of image descriptions
<http://cs.stanford.edu/people/karpathy/deepimagesent/>



Understanding images
<https://research.googleblog.com/2014/09/building-deeper-understanding-of-images.html>

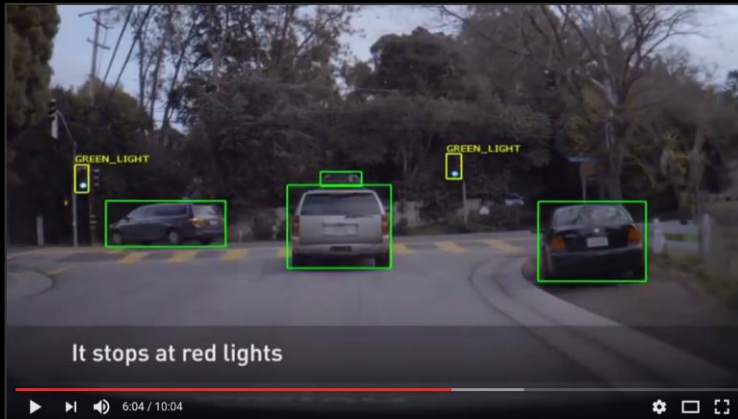


Deep image colorisation
<http://richzhang.github.io/colorization/>
<https://www.youtube.com/watch?v=eL5ilZgM89Q>



Creation of "artistic" images from sketches and videos
<https://www.youtube.com/watch?v=fu2fzx4w3ml>
https://www.youtube.com/watch?v=FzvTLEB_3KY

SOME (VERY FAMOUS AND) RECENT DEEP LEARNING SYSTEMS



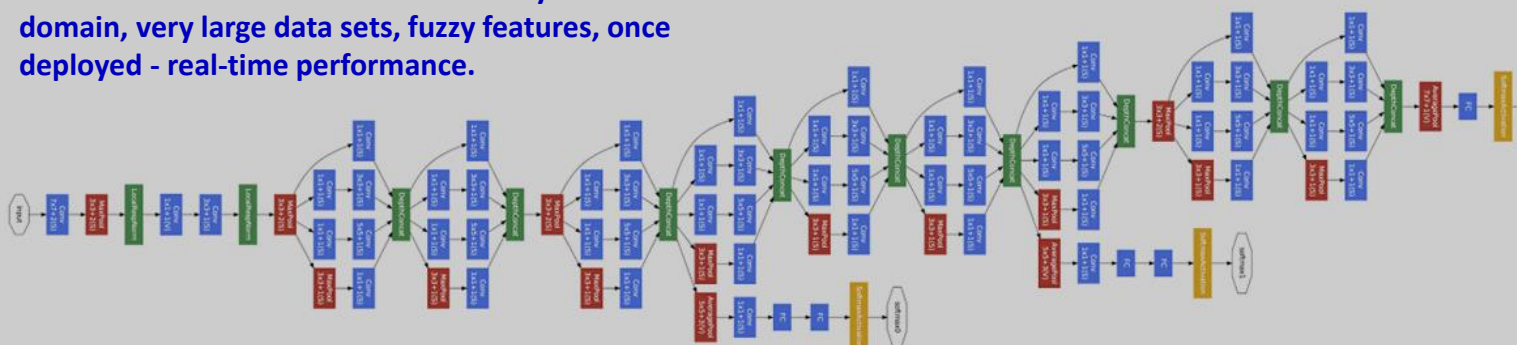
NVIDIA self-driving cars: Deep learning navigates streets, avoids obstacles, obeys traffic signs and rules
<https://www.youtube.com/watch?v=MF9NwOTLgE>



IBM Watson – Morgan movie trailer: Identifies movie clips that have emotional content for them to be included in a trailer
<https://www.youtube.com/watch?v=gJEzuYynaiw>

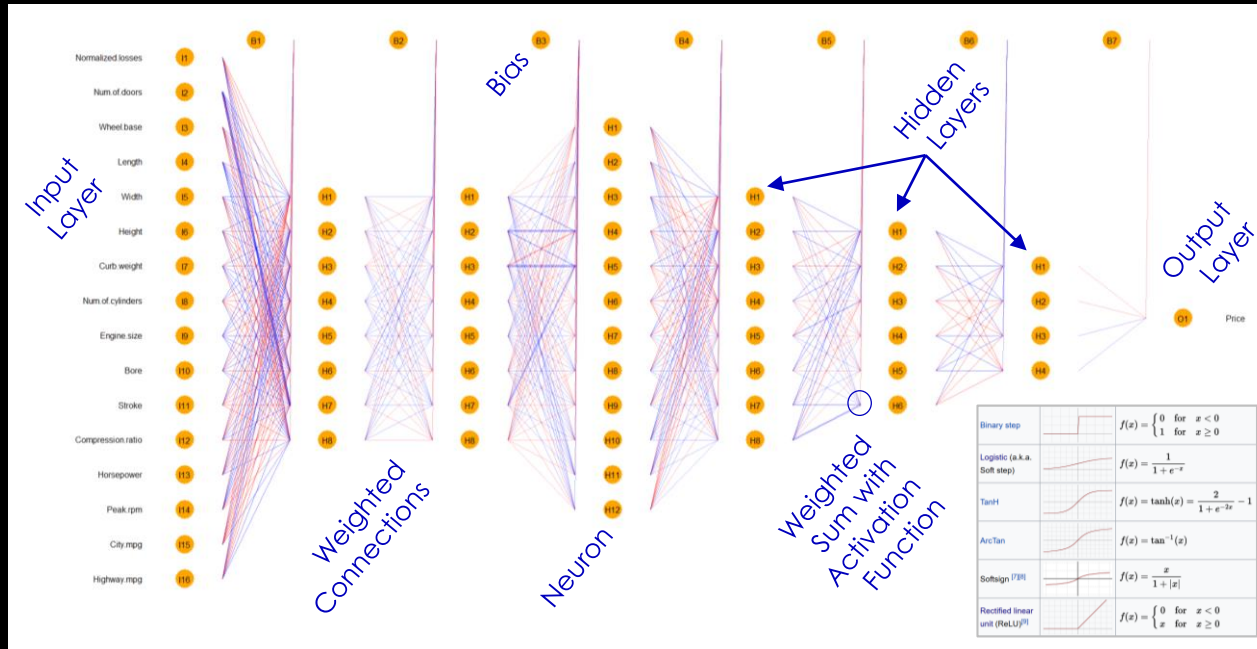
Google Inception network is used in image recognition. For example, it is able to identify a person in a photo (Admiral Grace Hooper) and the fact that she is wearing a uniform.
https://www.tensorflow.org/tutorials/image_recognition

Common features: until now exclusively in human domain, very large data sets, fuzzy features, once deployed - real-time performance.



GEARS & KNOBS OF DEEP LEARNING

DEEP NEURAL NETWORKS



https://en.wikipedia.org/wiki/Activation_function

- Neural networks take numeric variables on input and produce numeric or categorical variables on output
- The network consists of (great) many layers
- Each layer consists of neurons, each connected with all neurons of the previous layer via weighed edges
- Each neuron calculates a weighted sum of all values from the previous layer – similar to logistic regression
- A constant value, called bias, is added to the sum
- A non-linear activation function is finally applied to transform and scale the result
- The aim of neural network training is to identify the most suitable network architecture, the weights of the connections and biases from the set of input-output examples
- After training the neural network can predict the output from new, previously unknown inputs
- There exist many algorithms of neural network training and optimisation

HOW DOES IT WORK?

The screenshot displays the TensorFlow Neural Networks Playground interface. At the top, a navigation bar includes 'File', 'Edit', 'View', 'History', 'Bookmarks', 'Tools', and 'Help'. The browser address bar shows the URL: `playground.tensorflow.org/#activation=relu®ularization=L2&batchSize=10&dataset=spiral®Dataset=reg-plane&learningRate=0.03®ul`. Below the browser, a dark blue banner reads: "Tinker With a **Neural Network** Right Here in Your Browser. Don't Worry, You Can't Break It. We Promise."

The main interface features a control panel with the following settings:

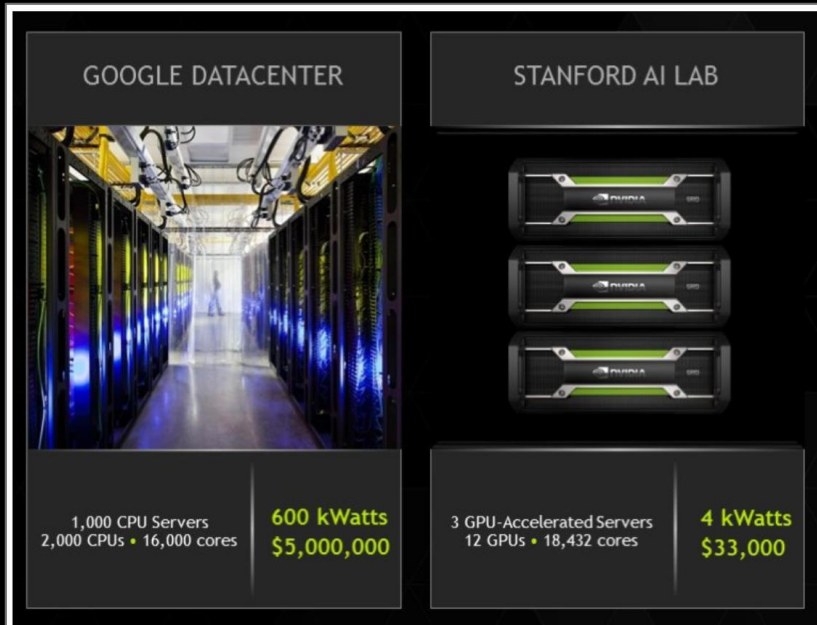
- Epoch: 002,312
- Learning rate: 0.03
- Activation: ReLU
- Regularization: L2
- Regularization rate: 0.003
- Problem type: Classification

The central area is divided into four sections:

- DATA:** Includes a dropdown for "Which dataset do you want to use?" (currently set to "Spiral"), a slider for "Ratio of training to test data: 50%", a "Noise: 5" slider, and a "Batch size: 10" slider. A "REGENERATE" button is located below these controls.
- FEATURES:** Lists input features: X_1 , X_2 , X_{12} , X_{22} , X_1X_2 , $\sin(X_1)$, and $\sin(X_2)$. A note states: "The outputs are mixed with varying weights, shown by the thickness of the lines."
- 4 HIDDEN LAYERS:** Shows a neural network diagram with 5 neurons in the first hidden layer, 7 neurons in the second, 5 neurons in the third, and 2 neurons in the fourth. A tooltip indicates: "This is the output from one neuron. Hover to see it larger."
- OUTPUT:** Displays "Test loss 0.056" and "Training loss 0.019" with a corresponding loss graph. A color-coded plot shows the spiral dataset. A legend indicates: "Colors shows data, neuron and weight values." Below the plot are checkboxes for "Show test data" and "Discretize output".

WHAT MAKES DEEP LEARNING WORK EFFICIENTLY? GPUs!

2013 – Google and Stanford AI Lab (\$\$\$)



Benchmarks with [BIDMach library](#) show that main classification algorithms run on a single instance with a GPU are faster than on a cluster of hundred CPU instances with distributed technologies such as SPARK.

- CPU = Central Processing Unit
Makes you computer run
- GPU = Graphical Processing Unit
Displays graphics on your monitor
- CPUs are used in all computers
GPUs are used in all computers
- In the past, high-performance GPUs have been designed for gaming and specialist video, VR / AR applications
- NVIDIA released programmable GPU with 1000s of CUDA “cores”, each allowing parallel execution of a simple program
- NVIDIA GTX 1080 Ti GPU has: 3,500 cores
Your laptop CPU has: 4 to 16 cores
- Cost of NVIDIA GTX 1080 Ti GPU: A\$1,200
- Typical gaming computer can support up to 4 NVIDIA GPUs (1.2kWatts): 14,000 cores
- Total cost of each NVIDIA GPU-based high performance computer for deep learning is (Deakin Business School 2017): A\$14,000

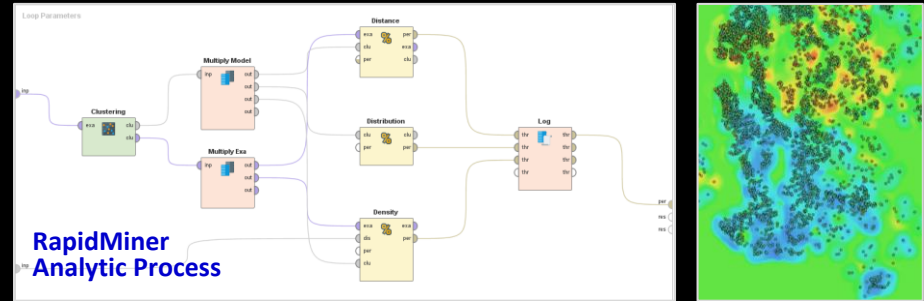
2017 – Amazon.com Lambda Deep Learning DevBox - with 4x NVIDIA GTX TITAN X 12GB, 1.2 kWatts, Ubuntu 14.04 LTS, CUDA, Caffe, Torch, and CuDNN (US\$14,899 + \$26.49 shipping)



INSPIRED AT ICM VISLAB, POLAND DEEP LEARNING AT DEAKIN

Example Project at ICM VisLab (2 wks)

- A German hospital required assistance with postoperative diagnosis of Achilles tendon injuries.
- They provided VisLab with 2000 CAT scans (in 7 planes) with additional information of previous diagnoses.
- VisLab staff experienced in Medicine, Maths and IT used this information to create a *deep learning classifier* of medical images, using UC Berkeley Caffe deployed on the National Supercomputer Infrastructure.
- The reported performance (98%) exceeded that of professional diagnosticians and lead to consulting contracts and publications.



Projects at Deakin – No longer dark science

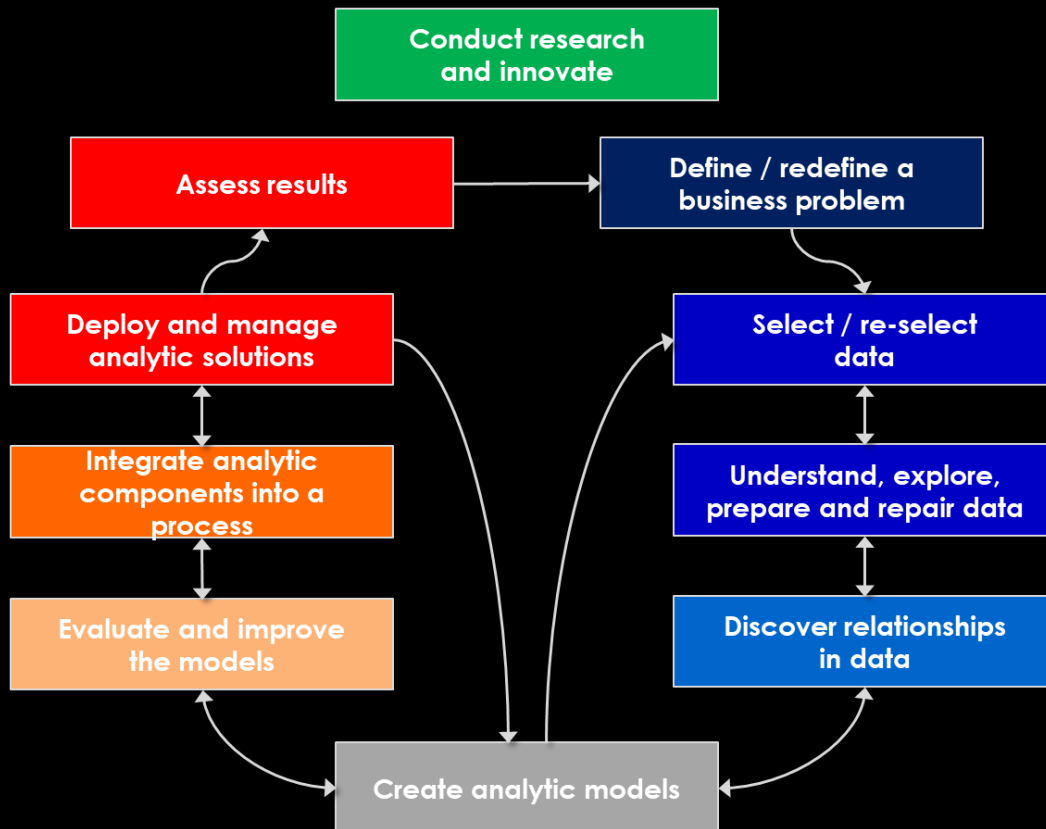
- DBS researchers and external partners will collaborate with DISBA staff to acquire and pre-process data, and then create, test and deploy *deep learning models*.
- The facility will rely on self-service analytics, possible via high-level analytic workflow tools, allowing researchers to focus on modelling of analytic solutions and interpretation of results via data visualization.
- All modelling tasks will be carried out in a dedicated lab, on high-capacity PCs, equipped with special purpose hardware and software to support deep learning tasks. Projects exceeding the lab capacity will be conducted using Deakin or external cloud services (paid for on a project-by-project basis).
- Projects resulting from collaboration between DBS and DISBA staff will result in joint publications, grants and HDR supervision.

ANALYTIC PROCESS

Data analytics is a complex process, which requires many inter-related activities, which need to be streamlined and rigorous.

Data analytics for research, to be effective and efficient, requires a streamlined business-like process.

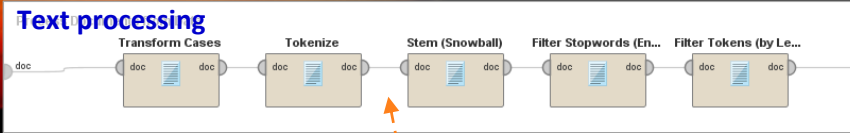
Data analytics for business, to be respectable and reproducible, needs scientific rigour.



- Define a business problem
- Select data
 - Structured and/or unstructured
 - What to predict (label)
 - What are the predictors (attributes)
- Explore and understand data
 - Statistics
 - Distribution
 - Relationships
- Build the model
- Evaluate model performance
 - Training performance
 - Hold-out validation
 - Cross-validation
- Integrate the model with enterprise systems
- Deploy validated model
 - Use the validated model
 - Predict labelled attribute
 - Account for possible error
- As the world changes assess the model results and its performance – a new model may be needed!

TEXT ANALYTICS PROJECT IN PROPENSITY ANALYSIS

This text mining model aims to create new variables from text and then use them to predict passenger views on quality of meals, entertainment, seats, crew and other services.

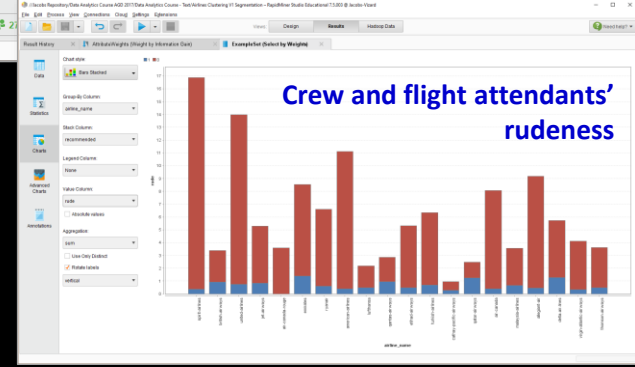
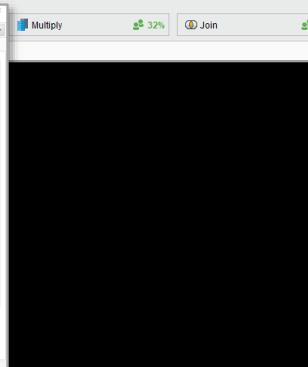
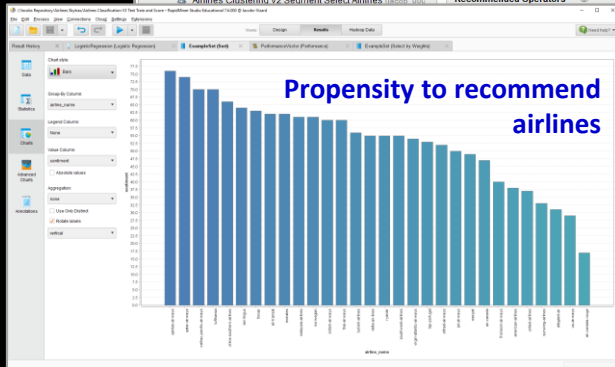


The main screenshot shows the RapidMiner Studio interface with a workflow for 'Airlines Classification V3 Text Train and Score'. The workflow is divided into several stages:

- Training data prep:** Includes 'Airlines Train', 'Filter Examples', 'Select Text Only', and 'Set Role' operators.
- Predictive model creation and validation:** Involves 'Process Train Docs', 'Weight by Informa...', 'Select by Weights', and 'Validation' operators.
- Model application to new data:** Uses 'Process New Docs', 'Select by Weights (2)', and 'Apply Model (2)' operators.
- New data prep:** Includes 'Airlines Score', 'Filter Examples (1)', 'Select Attributes', and 'Set Role (2)' operators.
- Sentiment calculation:** A sub-workflow containing 'Filter Examples (2)', 'Sort', 'Aggregate', 'Generate Attributes', and 'Rename' operators.

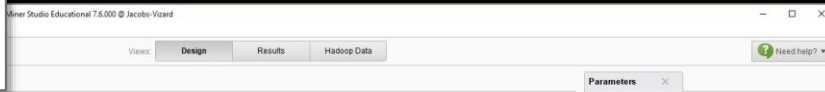
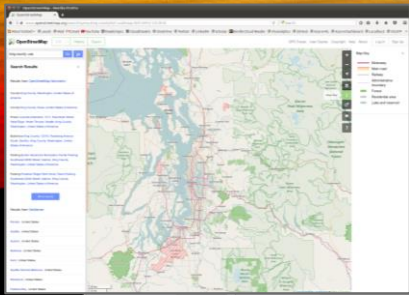
Annotations on the right side of the interface include:

- 'Jacob Cybulski Predictive Text Analytics Model with New Text Scoring' and 'Try: Different classifiers with different parameters'.
- 'Text Mining with Sentiment / Propensity Analysis'.
- 'Passenger groups' with a corresponding radar chart showing sentiment scores across various categories.
- 'Gradient Boosted Trees' help text, including a synopsis: 'Executes GBT algorithm using H2O 3.4'.



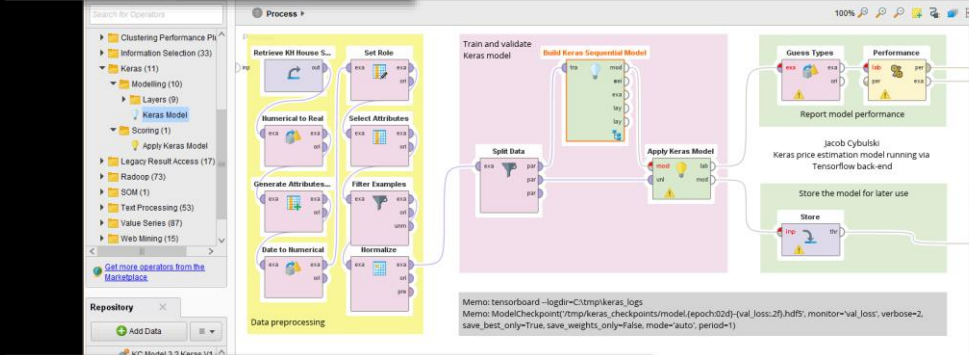
DEEP LEARNING PROJECT KING COUNTY REAL-ESTATE

Problem Statement / Contextualisation



Modelling Analytic Process

- Acquiring data
- Cleaning data
- Model training
- Model validation
- Model optimisation
- Data visualisation
- Reporting results

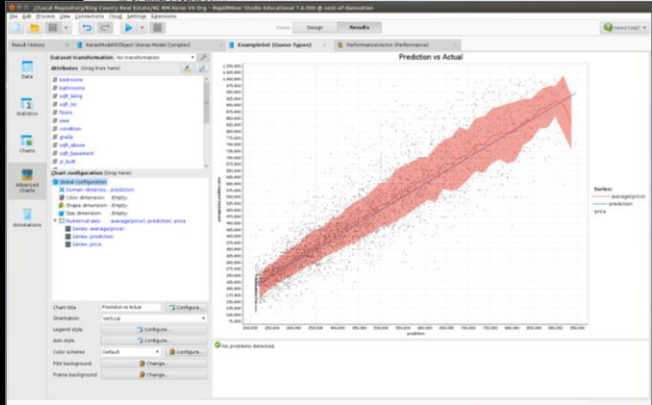
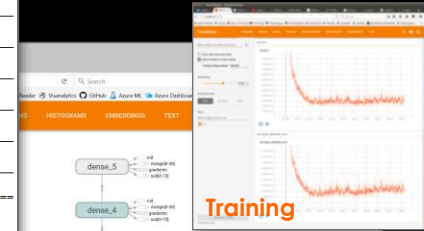


KerasModelIOObject

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 50)	1050
dropout_1 (Dropout)	(None, 50)	0
dense_2 (Dense)	(None, 30)	1530
dropout_2 (Dropout)	(None, 30)	0
dense_3 (Dense)	(None, 20)	620
dropout_3 (Dropout)	(None, 20)	0
dense_4 (Dense)	(None, 10)	210
dense_5 (Dense)	(None, 1)	11

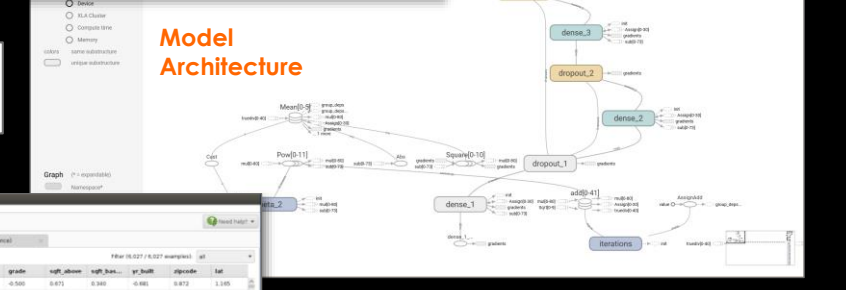
Total params: 3,411
Trainable params: 3,421
Non-trainable params: 0

Model Parameters



PerformanceVector
PerformanceVector:
root_mean_squared_error: 76966.348 +/- 0.000
absolute_error: 54851.875 +/- 53991.578
correlation: 0.921

Performance



Interpretation of Results

This deep learning model allows effective prediction of real-estate prices in King County, Washington, USA.

Tensorflow Dashboard

- Model training
- Model Validation

Deployment of Analytic Process

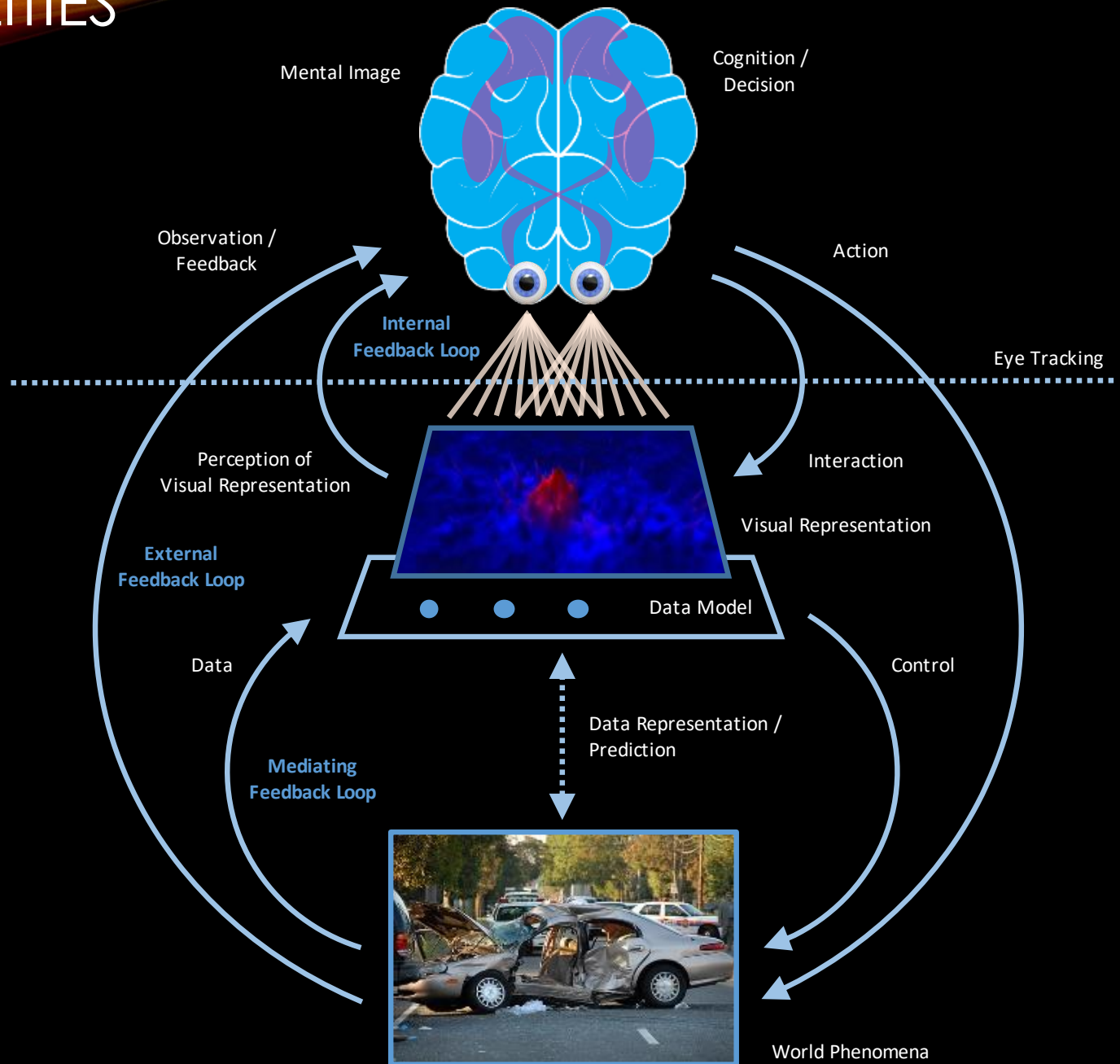
- Application (possibly on a server)
- Results (possibly shared via server)

OTHER FACILITIES

The lab will also provide other facilities to support analytics-related research.

These will include:

- Remote-control cameras for observation studies,
- VR/AR for immersive data visualisation,
- Eye tracking equipment to study the impact of visual representation on collaborative problem-solving and decision-making.





DEAKIN BUSINESS SCHOOL ADVANCED ANALYTICS AND DEEP LEARNING

- Rens Scheepers
- Jacob Cybulski
- Bardo Fraunholz
- Lemai Nguyen
- Dilal Saundage
- Lasitha Dharmasena
- Scott Salzman
- Ali Tamaddoni
- Mory Namvar

We are currently working towards the development of the capacity to deploy deep learning solutions to be used effectively by our colleagues and business partners.

The initial applications and research will include commercial image classification, social media sentiment analysis, analysis of financial reports and stock market predictions.

Interactive data visualization will assist exploration of data and interpretation of results.

THANK YOU

