MIS772
Predictive Analytics

Model Deployment
Reflection on the lessons learnt

Refer to your textbook by Vijay Kotu and Bala Deshpande, *Data Science: Concepts and Practice*, 2nd ed, Elsevier, 2018.

Model Deployment

- **Linking development and deployment – closing the loop**
- **New data - new problems**
- **Transferring models into the production**
- **Examples: estimation, anomaly detection, text analytics**

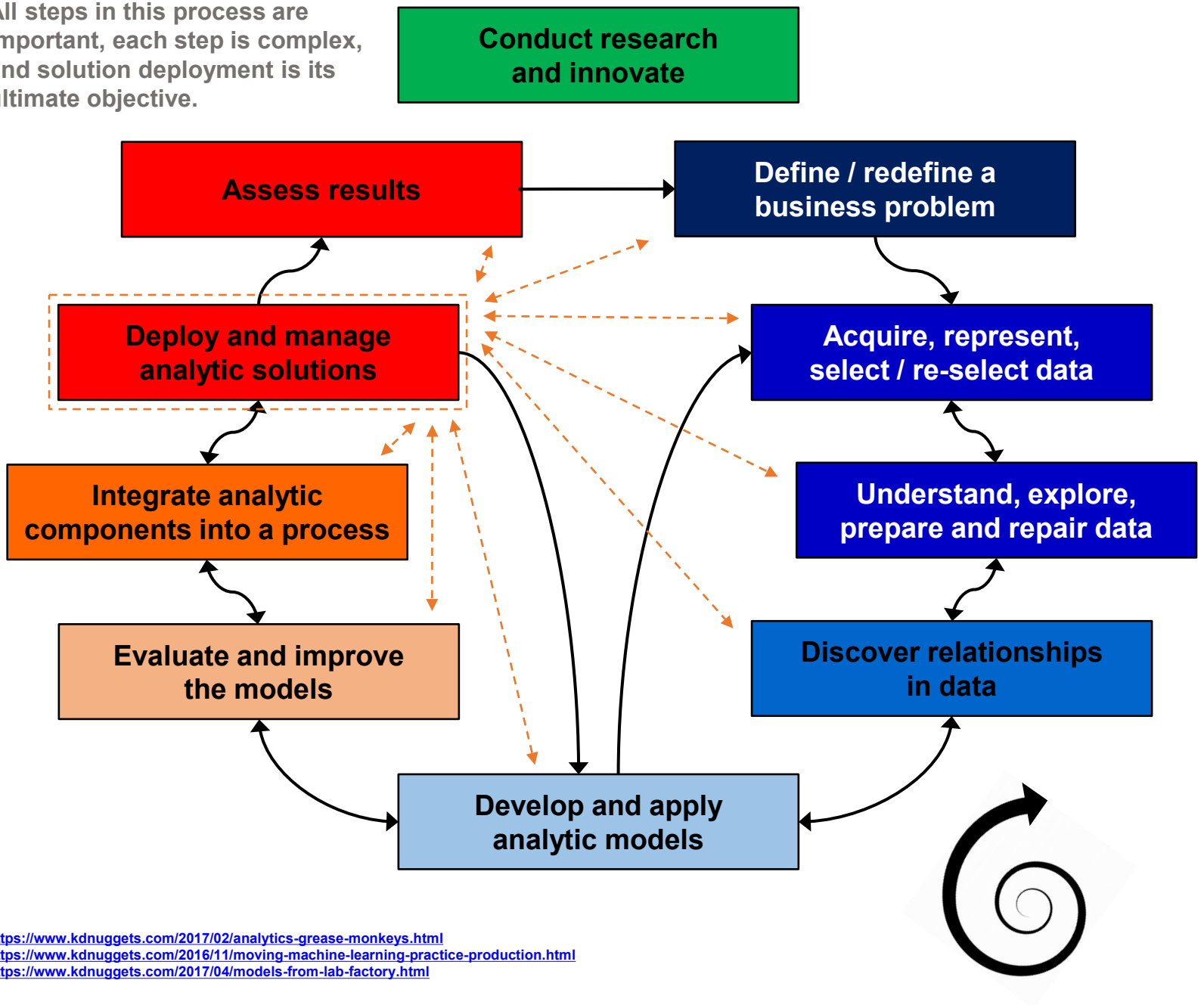
DEAKIN BUSINESS SCHOOL

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

Closing the Loop

All steps in this process are important, each step is complex, and solution deployment is its ultimate objective.



https://www.kdnuggets.com/2017/02/analytics-grease-monkeys.html
https://www.kdnuggets.com/2016/11/moving-machine-learning-practice-production.html
https://www.kdnuggets.com/2017/04/models-from-lab-factory.html

DEAKIN BUSINESS SCHOOL

Building Data Systems

From the organizational viewpoint, a strategy is required to develop capacity for the creation, deployment and maintenance of analytics processes and models.

From the data analysts' viewpoint, moving analytics from development into practice is very difficult

- The majority of problems solved in classes & tutorials are trivial as they are selected to convey fundamental concepts
- Often focus is on modelling and not end-to-end analytic process
- Models trained on small data sets do not scale up to real data sets
- Simplifying assumptions do not stand up in a complex world
- In practice, the effort goes not into development of models but into data pre-processing, model deployment and maintenance.
- We need skills and methods in model deployment (also known as application or scoring)

- Analytic processes are built to solve specific business problems
- Organizational infrastructure will have to support analytic solutions, e.g. digital platforms, digitalised products/services
- Analytic solutions will not work in isolation – will be integrated
- Past operation provides training data, future operation will use analytic models that will generate new and different data, i.e. a lifecycle approach
- All deployed models will change
- Development of analytics must focus on repeatable and automated deployment
- Everything must be tested
- Many organizations are not capable of developing all aspects of complex analytics
- Skills, tools, data and resources can nowadays be purchased, hired, pipelined and virtualised

https://www.kdnuggets.com/2016/06/building-data-systems-need.html
https://www.kdnuggets.com/2016/10/zementis-deep-learning-meets-deep-deployment.html
https://pages.dataiku.com/how-do-companies-build-production-ready-data-analytics-projects-0-0

Why all this fuss?

When creating a data model using past data:

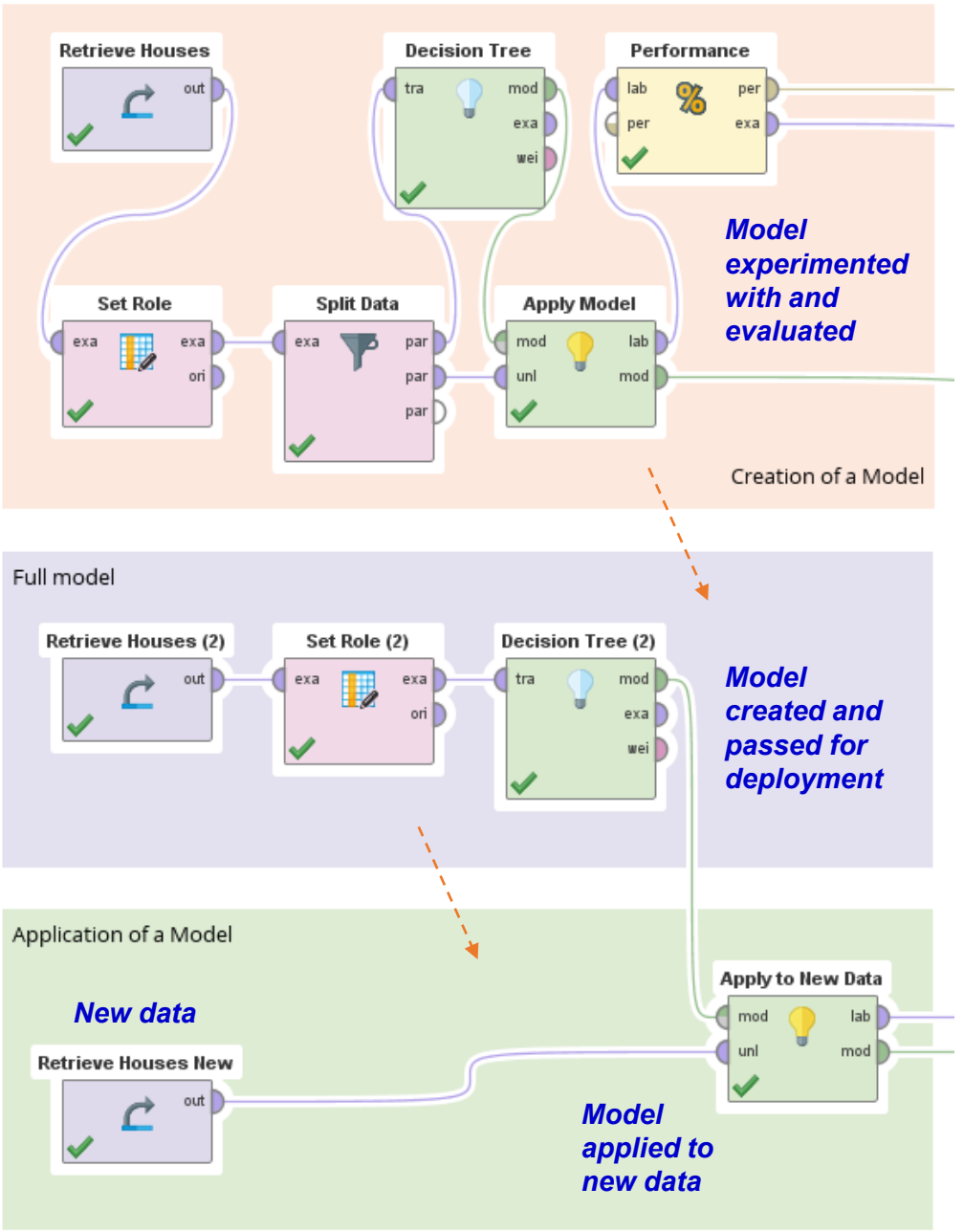
- Acquisition process may bias the collected sample
- Data statistics can be derived from the sample
- Data normalization and binning rely on the training data statistics
- If needed, creation of dummy variables is almost automatic, we assume (incorrectly) we know all possible nominal values
- Selection of attributes is based on relationships between a label and its predictors
- Text-attributes are derived from the training sample
- Transformations could assist model development
- Models are optimized for best outcome on validation data

When applying the model to the new / future data:

- New data comes from the population
- Statistics of the new data are less relevant (e.g. sample of 1)
- Normalisation and binning also rely on training data
- If missing values are unavoidable, the only strategy available is that developed for training data
- As described by the model, dummy variables are encoded according to training data, any new nominal values will cause errors
- Exactly the same attributes must be used in model application, and there is no label!
- New text will use terms different from those used in training sample
- Any new data transformation must be identical to that used in training

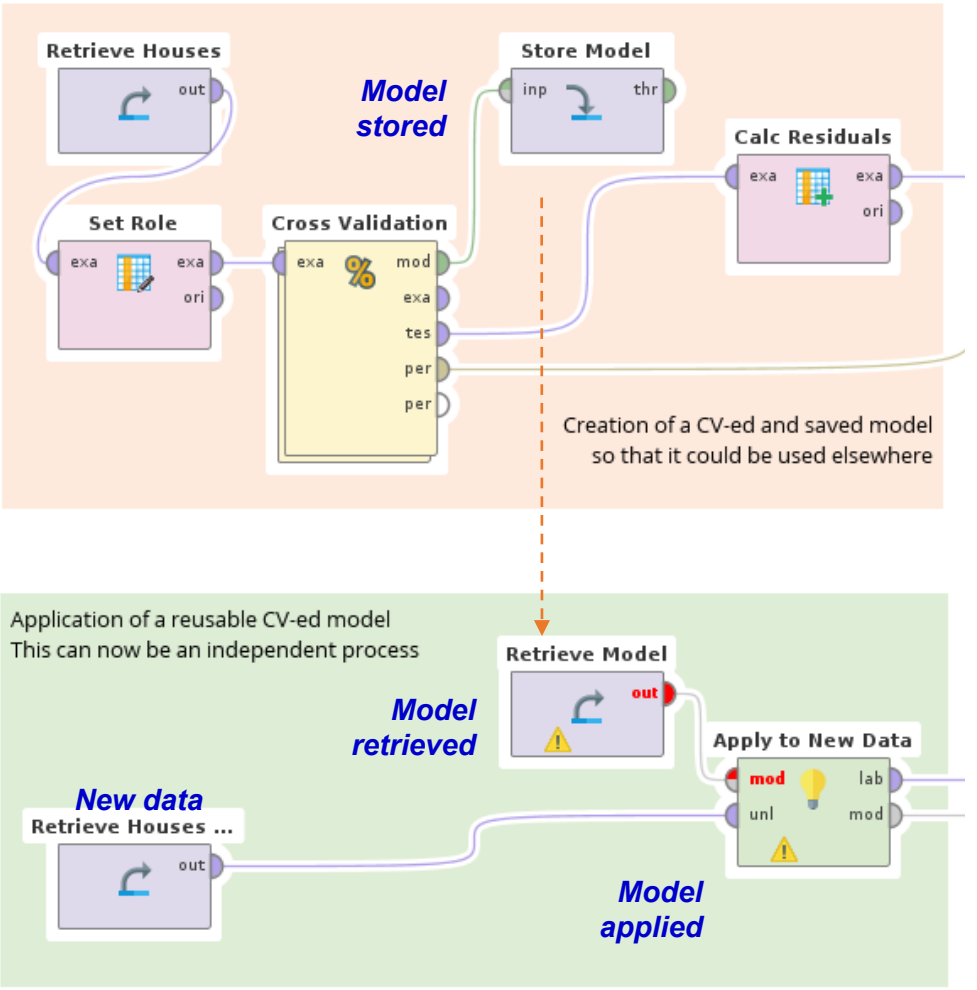
https://www.analyticsvidhya.com/blog/2016/05/8-reasons-analytics-machine-learning-models-fail-deployed/

- ***Deployment of a simple analytic process...***
- **The model is trained and evaluated using training and validation parts of the available data.**
- **The model can then be applied to new data.**
- **Its performance is estimated to be similar to that achieved in model validation.**
- **It is important to validate the model properly and possibly conduct its *honest test*.**



Saving Models and Applying Models

- Using the available data to develop, evaluate and apply a model in a single process is wasteful.
- Instead development and application should be split into separate analytic processes.
- In RM, cross-validation, provides performance estimates and the trained model.
- This model can then be *stored for later retrieval* and application to a new data set.
- The saved model can also be used across the analytic portfolio in many independently developed processes.



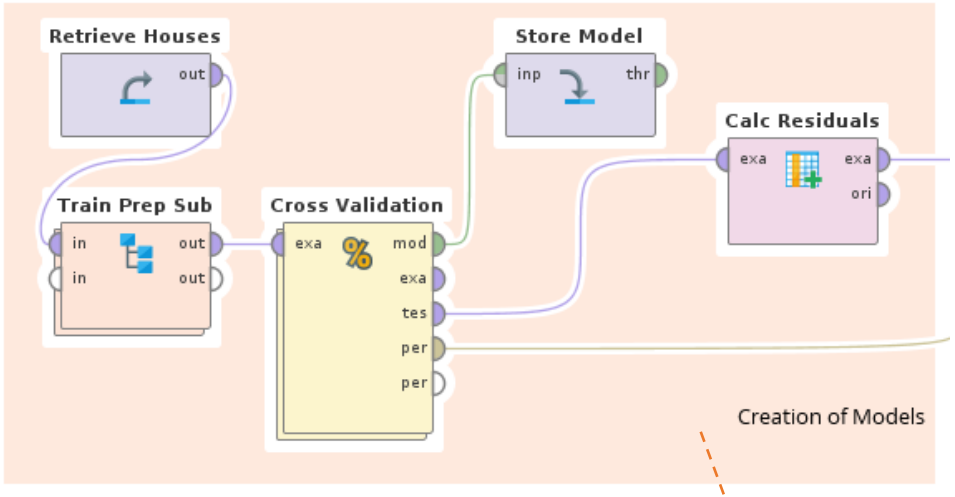
DEAKIN BUSINESS SCHOOL

Complex Deployment
More than Model Creation

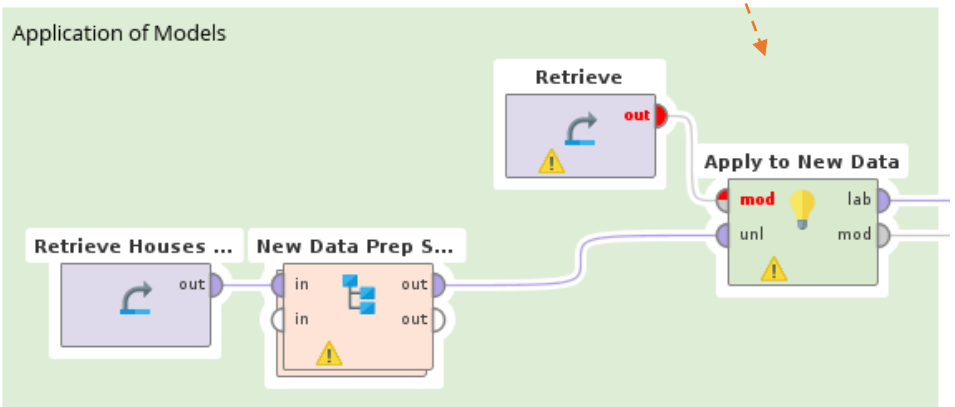
The majority of analytic processes require not only model development but also data pre-processing to facilitate creation of a model with optimum performance.

As data pre-processing transforms the original data attributes, to apply the newly developed model, the identical transformations must also be applied to all new data.

Data pre-processing models need to be stored



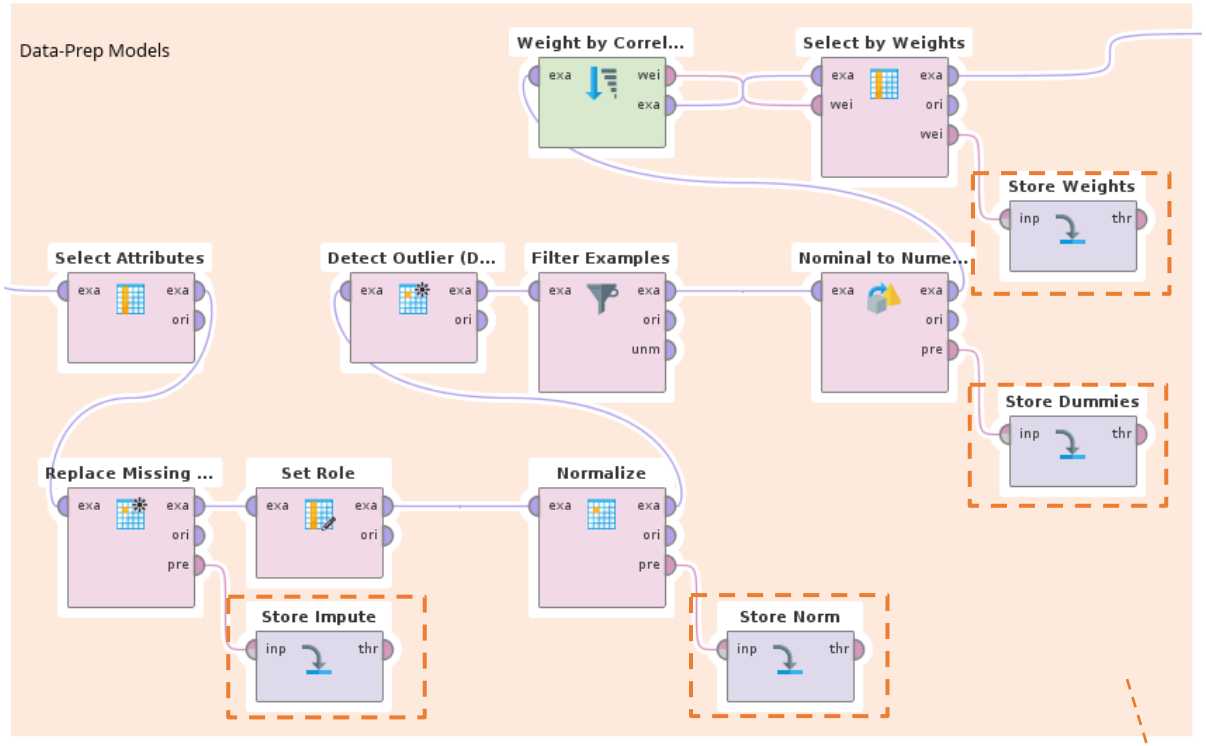
Data pre-processing models need to retrieved and applied to newly acquired data



The complexity of deployment shifts from model development to data pre-processing

DEAKIN BUSINESS SCHOOL

Capturing and Application of Pre-processing Models

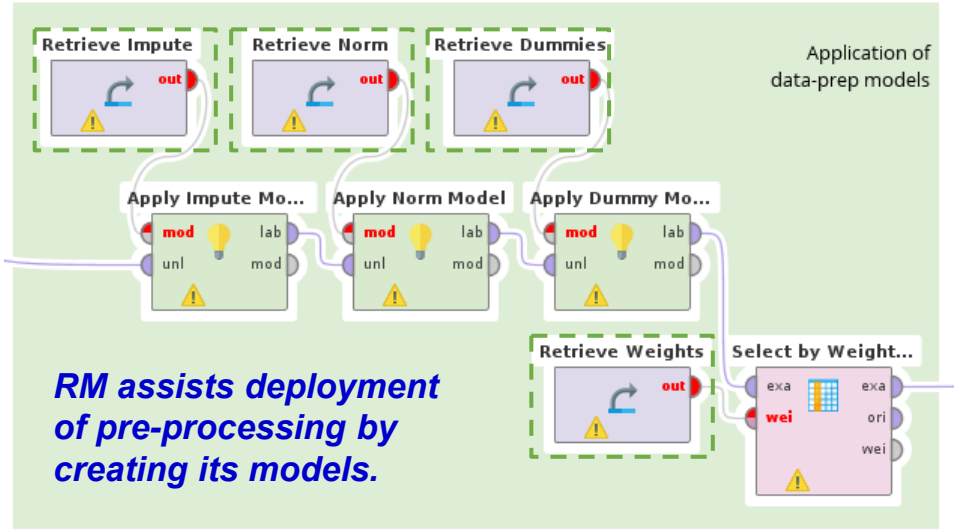


All pre-processing relies on many options, which must be applied <u>in exactly the same way to future data.</u>

These options must be selected with deployment in mind – this means be prepared for the unexpected and uncertain!

Examples of pre-processing:

- **Dealing with missing values**
- **Data normalization**
- **Transforming attribute types**
- **Selection of attributes**
- **Etc.**

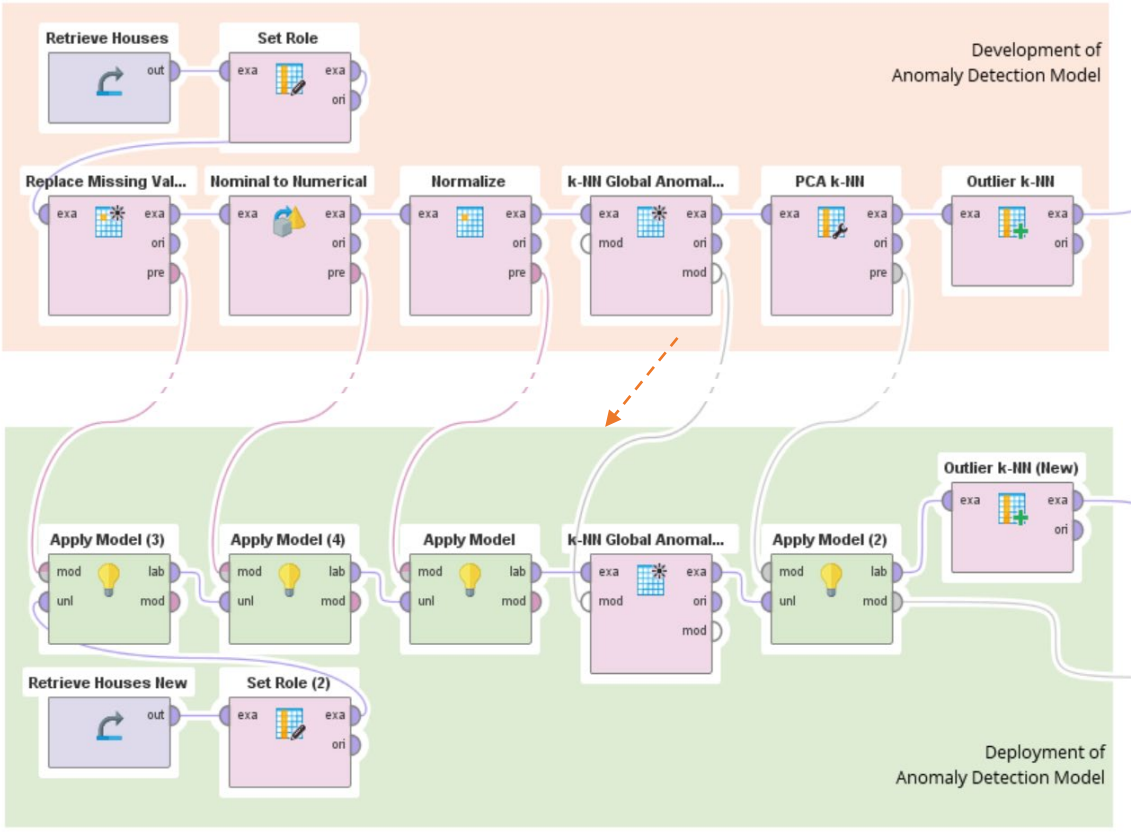


RM assists deployment of pre-processing by creating its models.

DEAKIN BUSINESS SCHOOL

Deployment of Anomaly Detection

- **The following example illustrates deployment of anomaly detection and PCA models.**

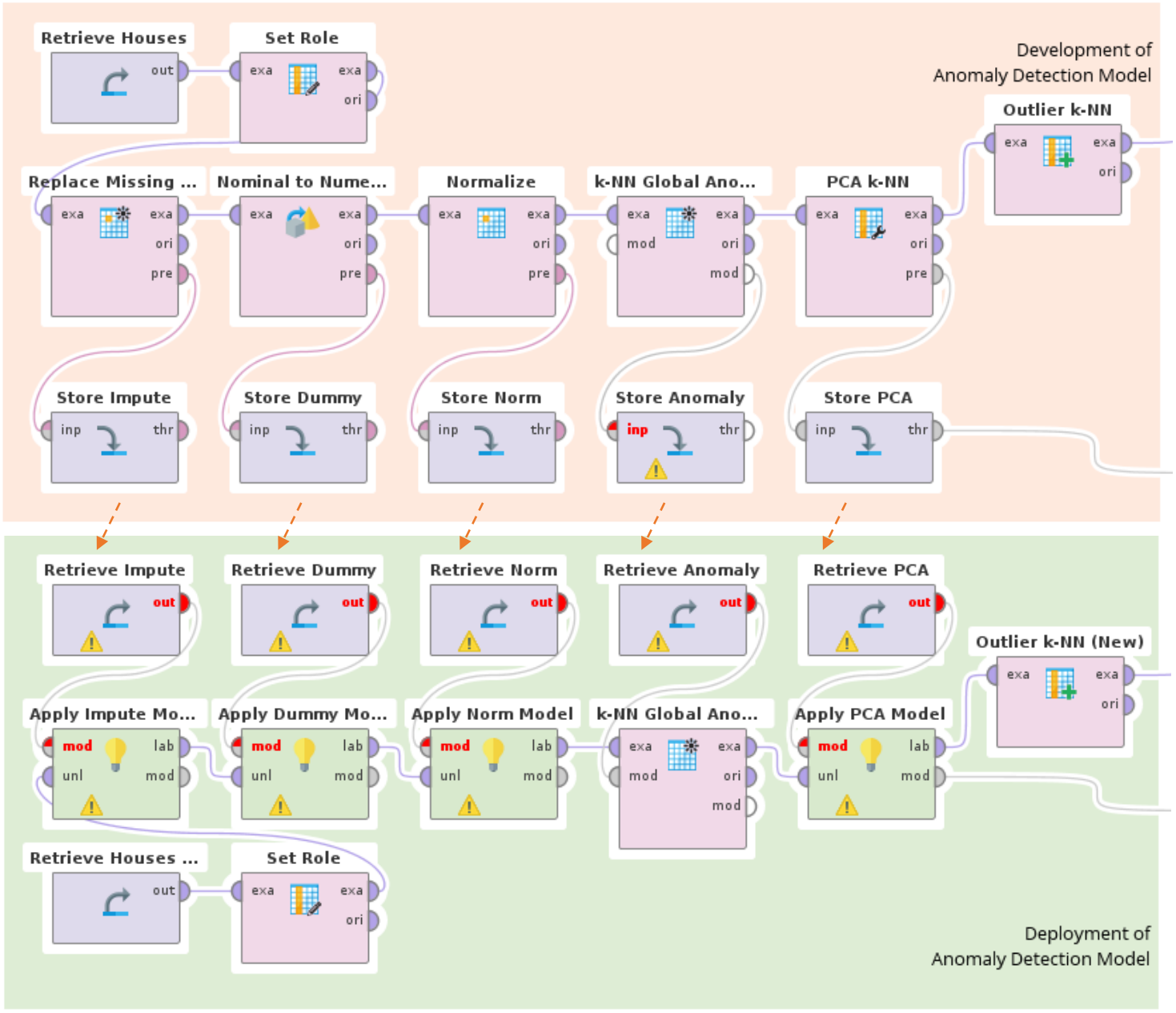


Deployment models span the entire spectrum of analytic tasks, i.e.

- **Data pre-processing**
- **Dimensionality reduction**
- **Clustering**
- **Anomaly detection**
- **Text processing**
- **Predictive models**
- **Data visualization models**
- **…**

DEAKIN BUSINESS SCHOOL

Deployment of Anomaly Detection



Resulting development and deployment models
(a lot of models are being created)

DEAKIN BUSINESS SCHOOL

Development and Deployments of Text Models

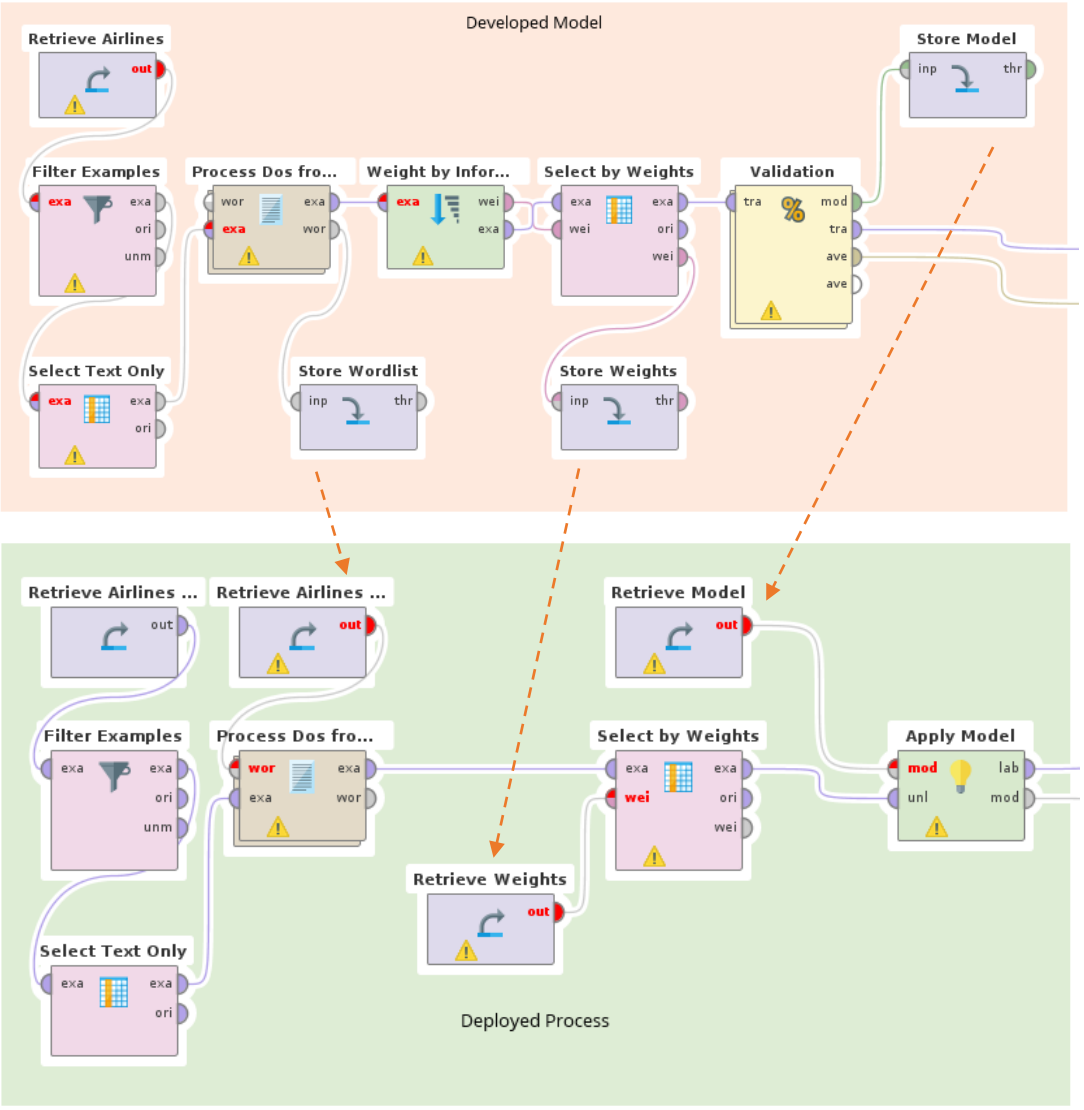
Deployment of text mining models is much more complex.

Text representation in TF-IDF vector space requires access to the wordlist created with the original text.

Dimensionality reduction after text parsing also relies on the previously weighed and selected attributes.

Both need to be saved, and then retrieved and re-applied to pre-process the new data.

All other models are deployed in the way discussed previously.



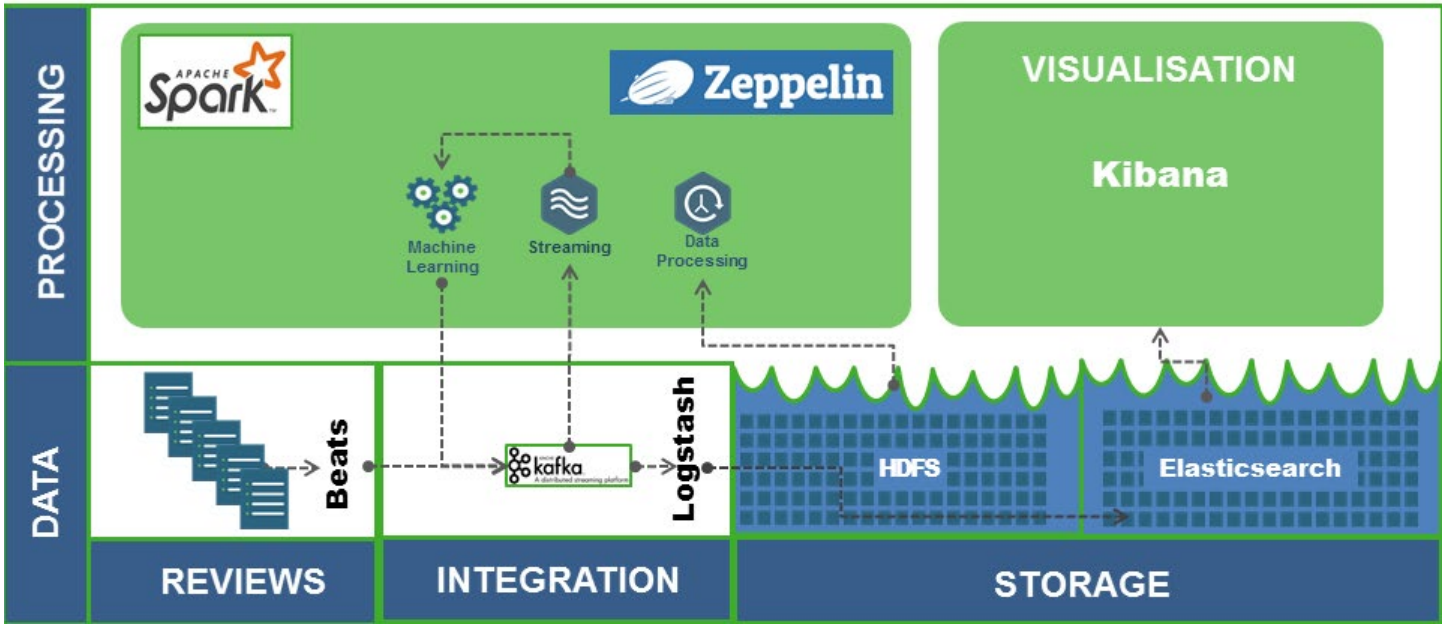
Row No.	id	prediction(recommended)	confidence(1)	confidence(0)	text	airlin	anoth	apolog	attent	avoid	becaus	book	cabin
1	6	1	0.658	0.342	sarajevo frankfurt l...	0.065	0	0	0	0	0	0	0
2	10	1	0.952	0.048	flight time econom...	0	0	0	0	0	0	0	0.306
3	15	1	0.876	0.124	istanbul ljubljana ...	0	0	0	0	0	0	0	0.061
4	70	1	0.999	0.001	return plenti legro...	0	0	0	0	0	0	0	0

DEAKIN BUSINESS SCHOOL

Deployment In-the-Large

- Regardless of your analytics platform the analytic process is not likely be deployed on this development platform.
- The analytic process will be integrated with an enterprise system, e.g.
  - RM models will run on the server
  - SAS models deploy to SAS Base
  - R models will be called via plugins
  - Python readily integrates with Python enterprise infrastructure
- There is also a tendency to deploy analytic solutions in the cloud, as part of a fully distributed architecture
- New tools for real-time analysis of data streams become widespread

Typical stream analytics architecture



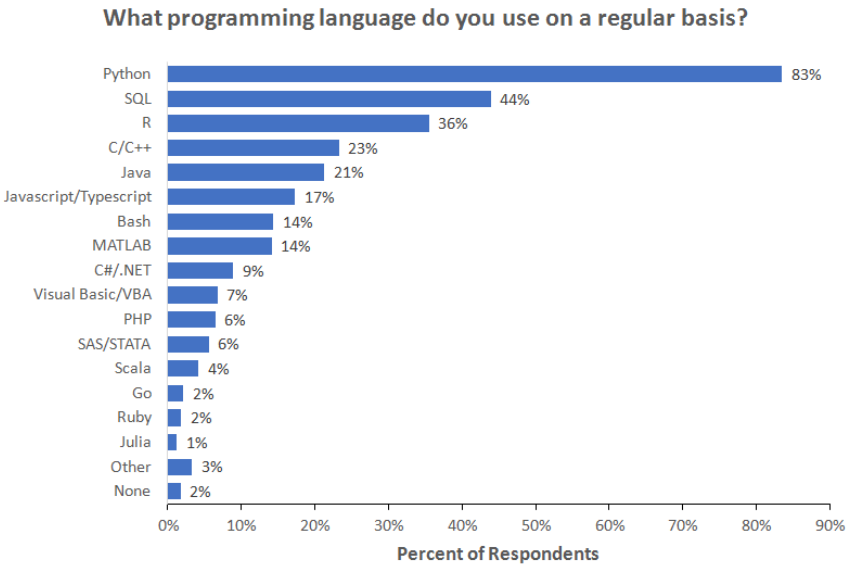
https://www.shi-gmbh.com/real-time-streaming-analytics-2/

DEAKIN BUSINESS SCHOOL

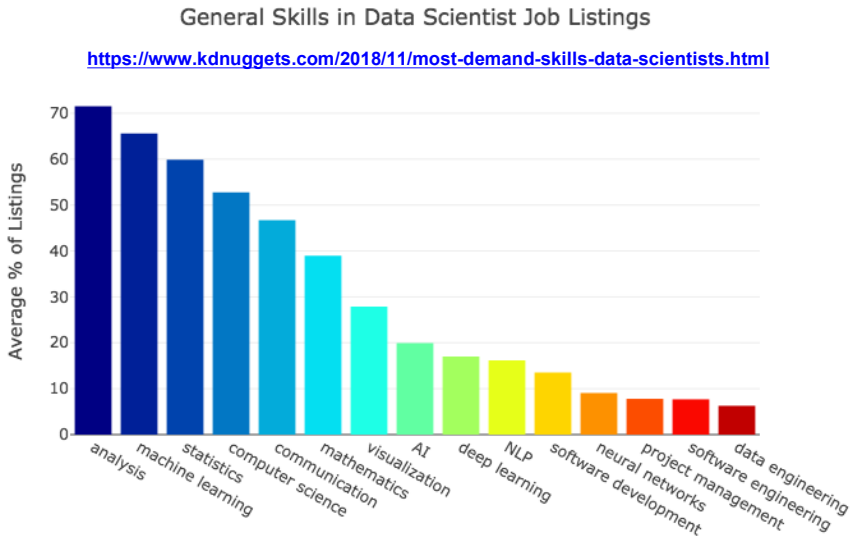
Skills for Data Analysts In-the-Large In-the-Future

You must invest in learning new skills!!!

2018-2020+

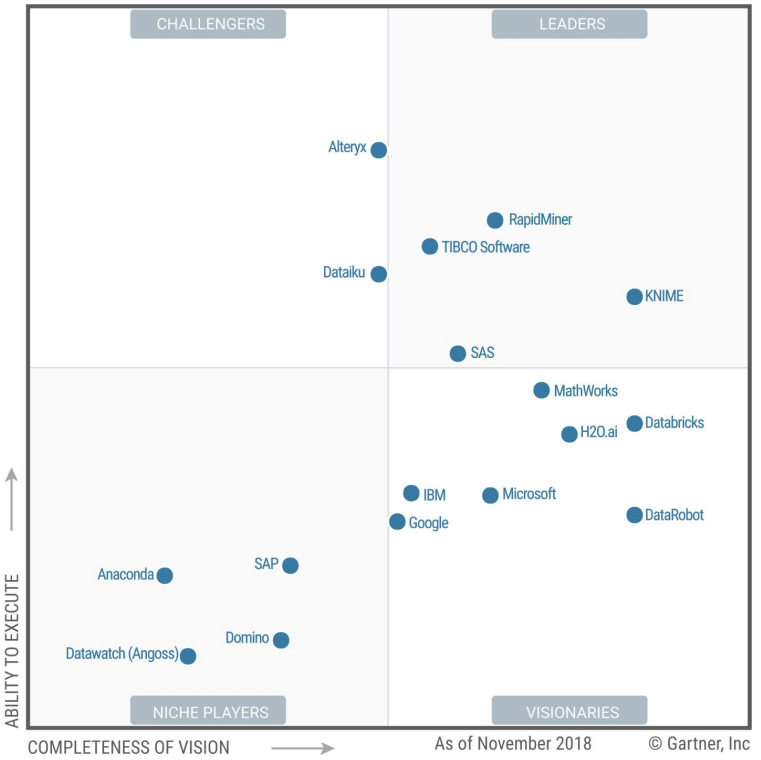


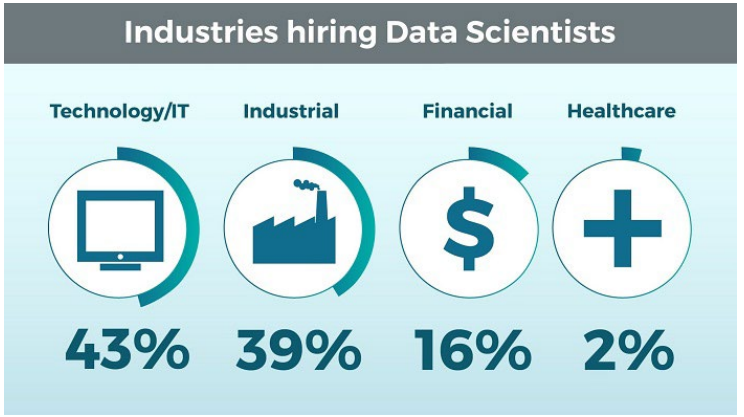
https://www.kaggle.com/kaggle/kaggle-survey-2018



https://www.kdnuggets.com/2018/11/most-demand-skills-data-scientists.html

https://rapidminer.com/resource/gartner-magic-quadrant-data-science-platforms/





https://www.kdnuggets.com/2019/03/typical-data-scientist-2019.html

DEAKIN BUSINESS SCHOOL

Summary / Review

- From the organizational viewpoint, enterprise needs a strategy to develop capacity for the creation and maintenance of analytics
- From the data analysts' viewpoint, moving analytics from development into practice is very difficult
- Problems solved in classes & tutorials are trivial
- Focus on analytic process, not just modelling
- Data pre-processing is costly and takes a lot of effort
- Models trained on small data sets do not necessarily scale up
- Simplifying assumptions do not stand up to the real world

- The following tasks need to be considered when deploying analytic processes:
  - data stats,
  - dimensionality reduction,
  - dealing with missing values,
  - attribute weighing,
  - attribute selection,
  - attribute transformation,
  - text parsing,
  - clustering,
  - anomaly detection,
  - data visualization,
  - optimisation parameters,
  - etc…

- *In some environments this process is purely manual, e.g. in Python or R.*

DEAKIN BUSINESS SCHOOL

Summary / Review

- What is the relationship between development and deployment of analytics?
- Why is deployment difficult?
- What deployment strategies are used by enterprises?
- How can a model be developed, evaluated and then deployed for use?
- What are the most common data pre-processing tasks that need to be considered when deploying a predictive model
- What text mining tasks must be deployed for the models to work on new texts?
- Why do normalization of training data and new data produce different results?
- Why word lists produced in model development must also be used by deployed models?
- Why term vectors produced with training data and those resulting from new data are incompatible?
- Why PCA-assisted visualization of data for training data set is not the same as that for new data?
- What tasks can be assisted with RapidMiner pre- and post-processing models.
- What skills are needed for the future of data analytics?