THE UNIVERSITY OF MELBOURNE

Data Analysis of Data Science Job Ads

with suggestions

Liuhua Peng

Lecturer, School of Mathematics and Statistics

Thanks !

Fiona Simpson

Careers & Industry Consultant Academic Engagement

Faculty of Science

The University of Melbourne

Peter Karutz

Senior Advisor Experiential Learning, Industry

Global Leadership and Employability, Academic Services

University Services

The University of Melbourne

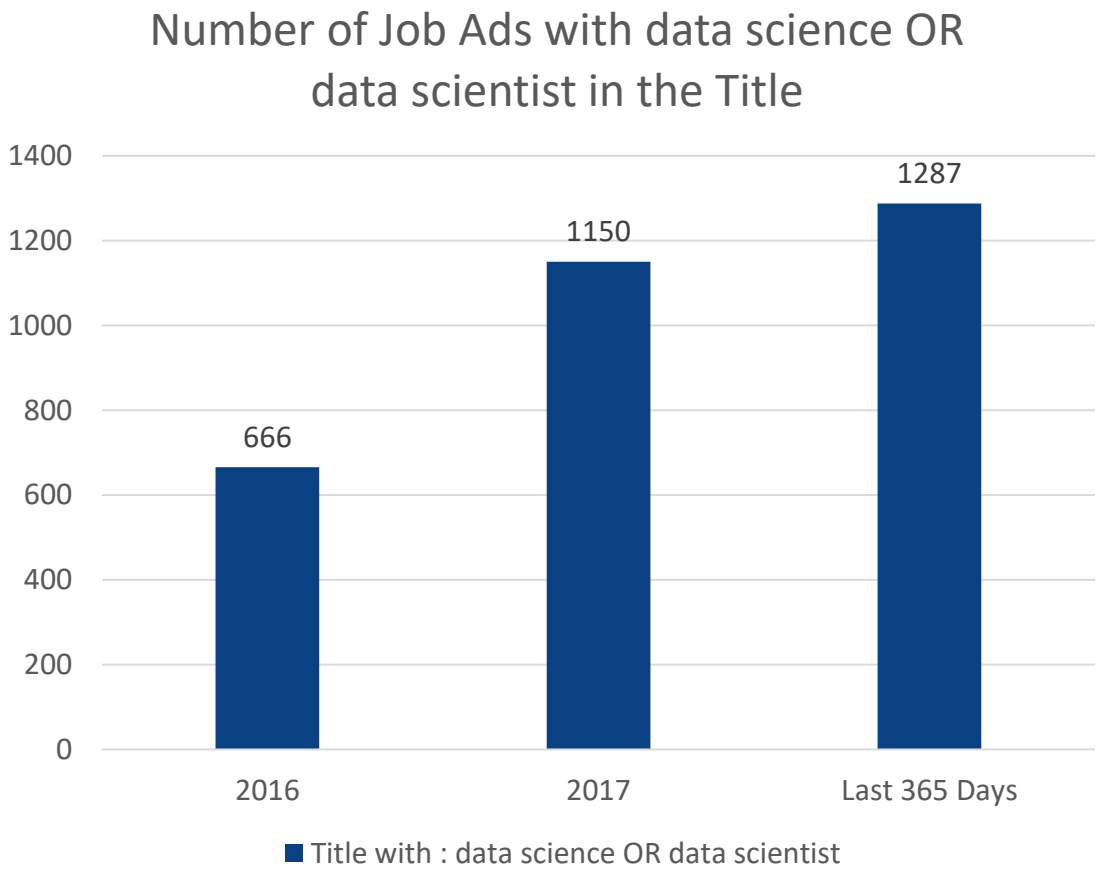
Kathy Ryan

Academic Services and Registrar

University Services

The University of Melbourne

An Increasing Market ! ! ! ! !



Last 365 Days: May 03, 2017 – May 02, 2018

- **Occupations**
  - Job Titles
  - Employers
- **Skills**
  - Baseline Skills
  - Software and Programming Skills
  - Specialized Skills
- **Education and Experience**
  - Degree
  - Experience
- **Salaries**

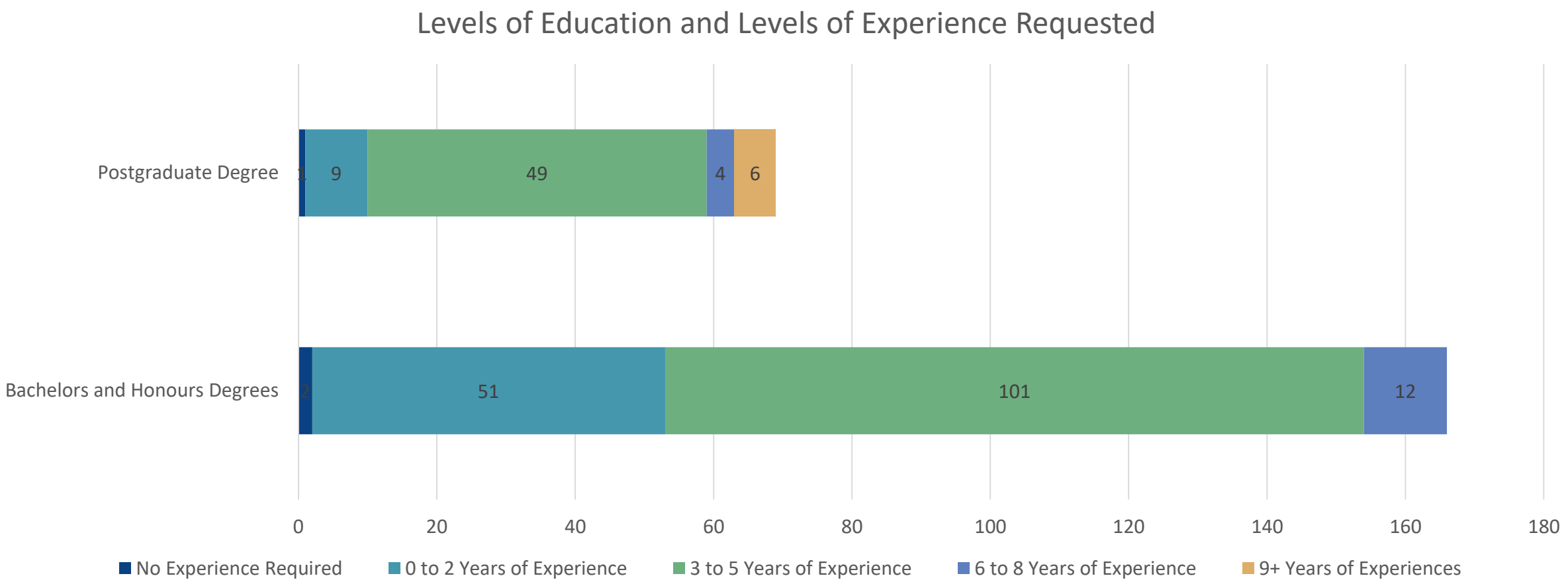
Trend in Advertised Occupations

Occupation (Job Title)	# Job Ads in 2016	# Job Ads in 2017	# Job Ads in Last 365 Days
Data Science	616	1030	1165
Data Analyst	13	38	32
Software Programmers	8	4	4
Database Administrator	5	9	17
University Lecturer	5	14	14
Software Engineer	3	7	8
Business Intelligence Analyst	2	9	8
Business Manager	1	2	4

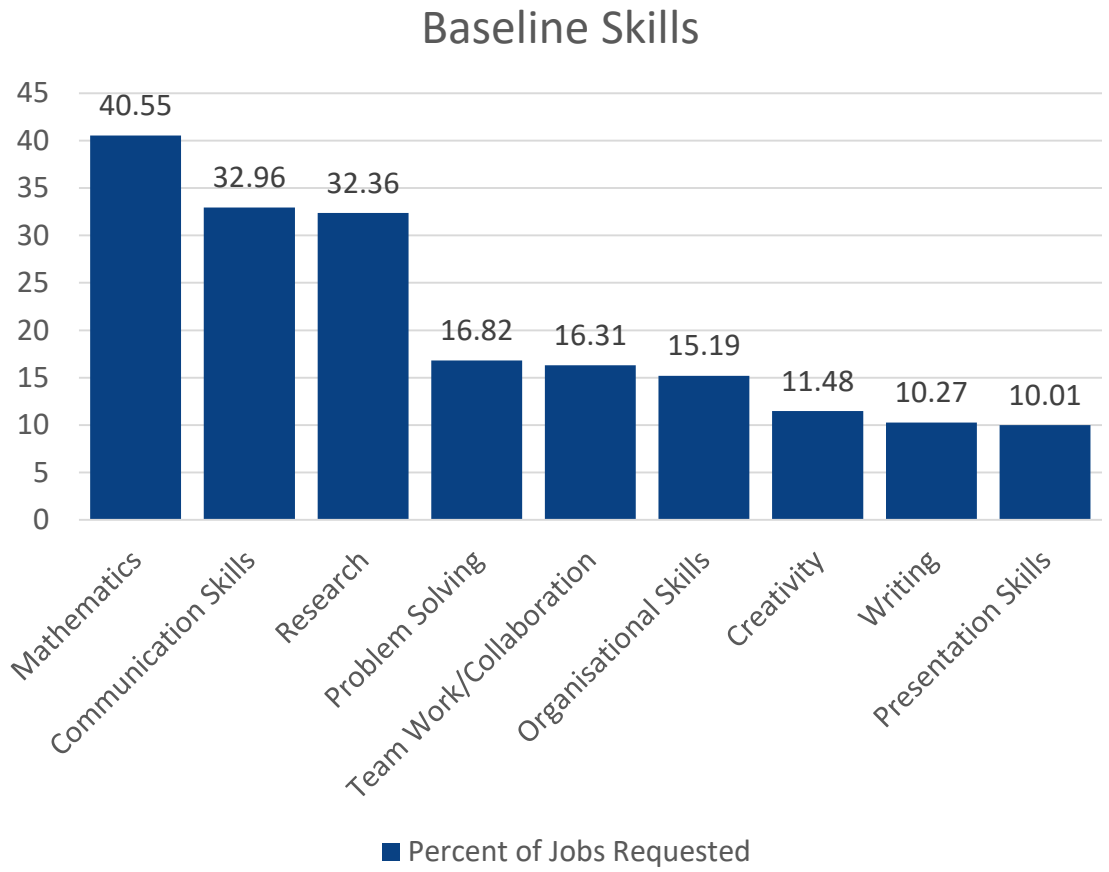
Employers

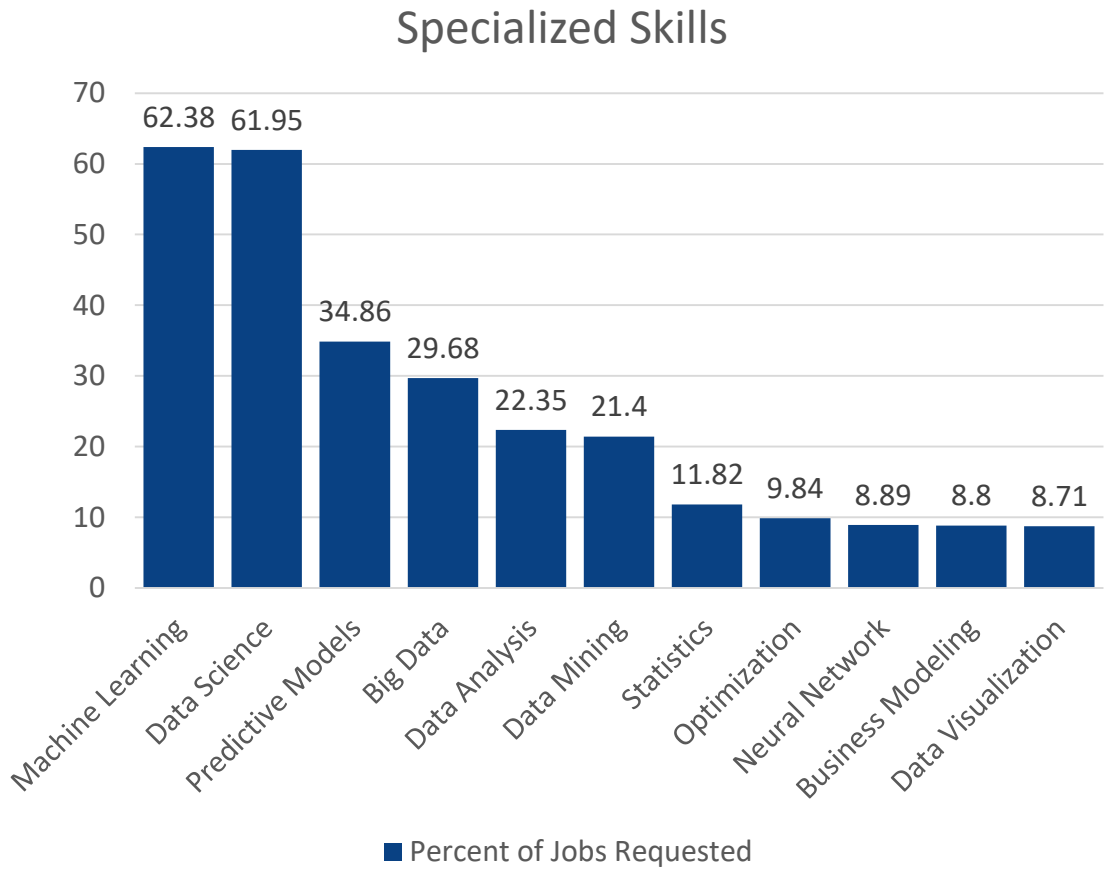

General Electric Company
Swinburne University of Technology
Monash University
General Assembly
Australia Group
Air New Zealand
Insurance
Amazon
Accenture
New South Wales Goverment
Spark New Zealand
BHP Billiton Limited
Woolworths
Suncrop Group Limited
Hatch Limited
Big Wednesday Digital
Crown Limited
Macquarie University
Gold Coast City Council
Sas Institute
IAG
IBM
Australian Taxation Office
Ernst and Young
Rangtech
Westpac
BUPA
Teradata Australia
Telstra
Optus
Expedia Incorporated
Readify Pty Ltd
Tabcorp Holdings Limited
National Australia Bank
The University of Sydney
Australian Department of Defence
Science New Zealand Inc
Australia Security Investments Commission
Queenslan University of Technology
Bureau of Meteorology

Degree and Experience

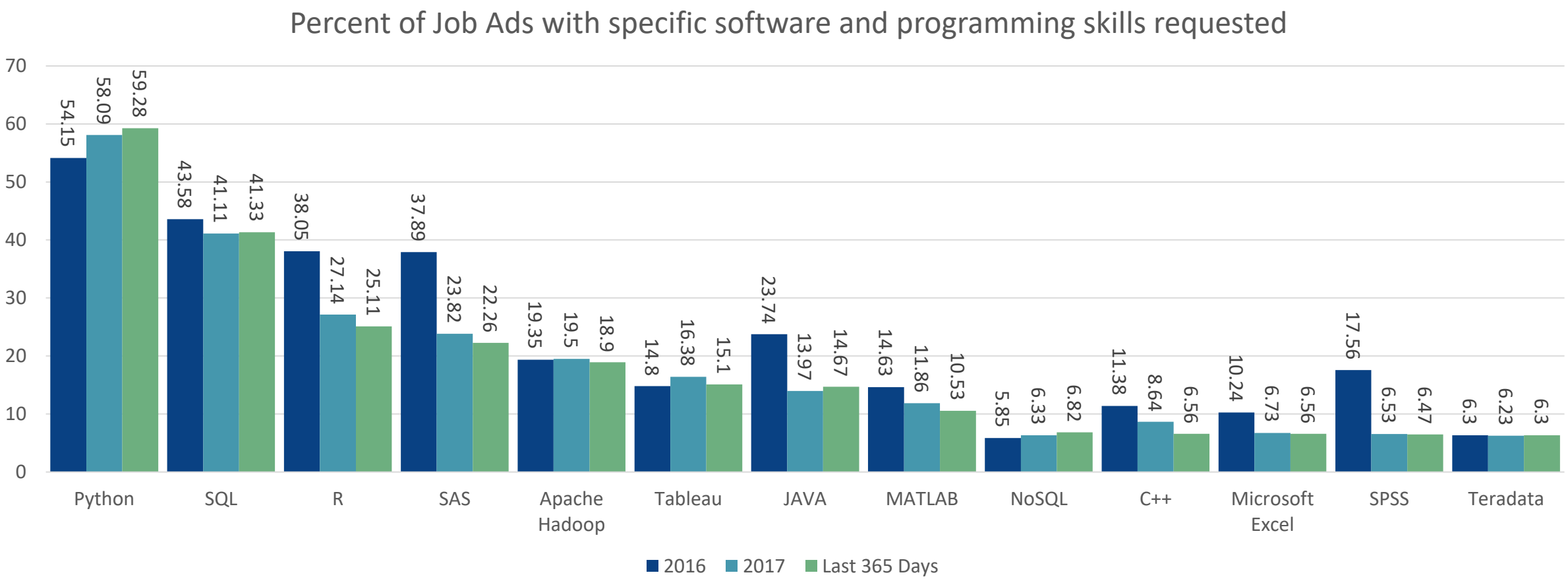


Top Skills: Baseline and Specialized Skills



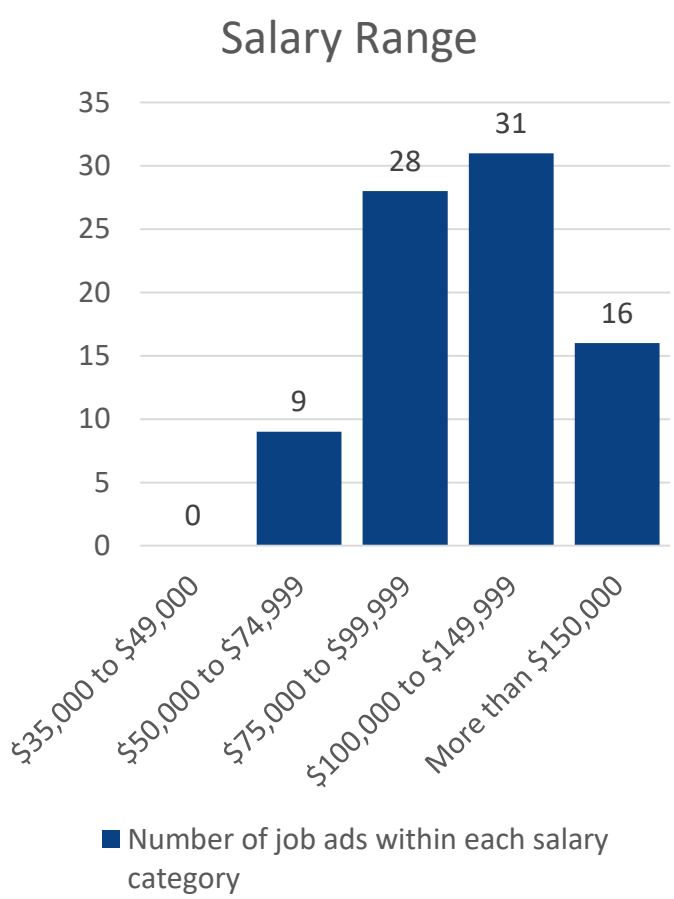


Top Skills: Software and Programming Skills



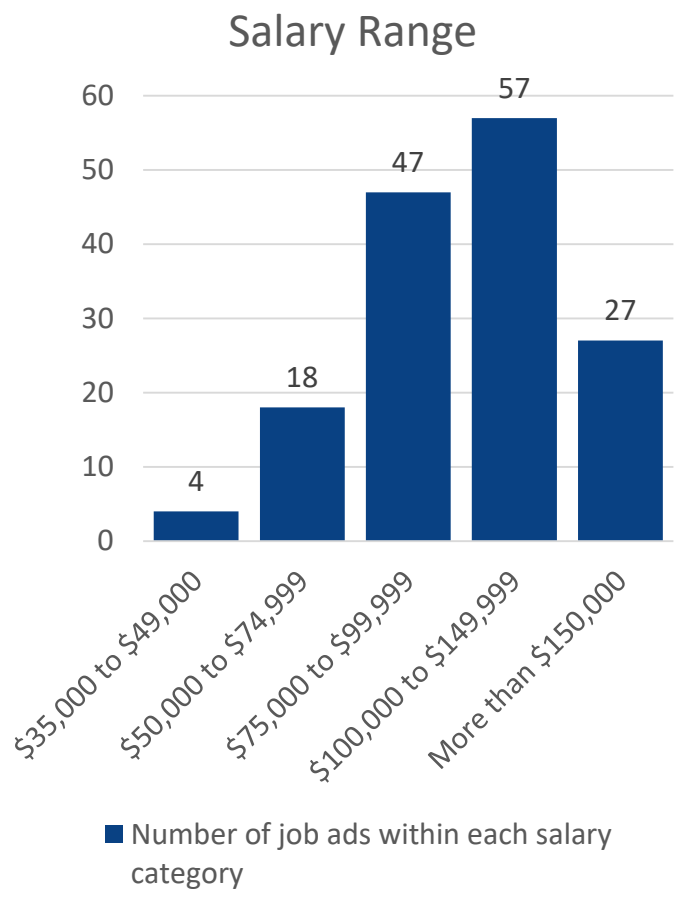
Salaries ! ! ! ! ! ! ! ! !

2016



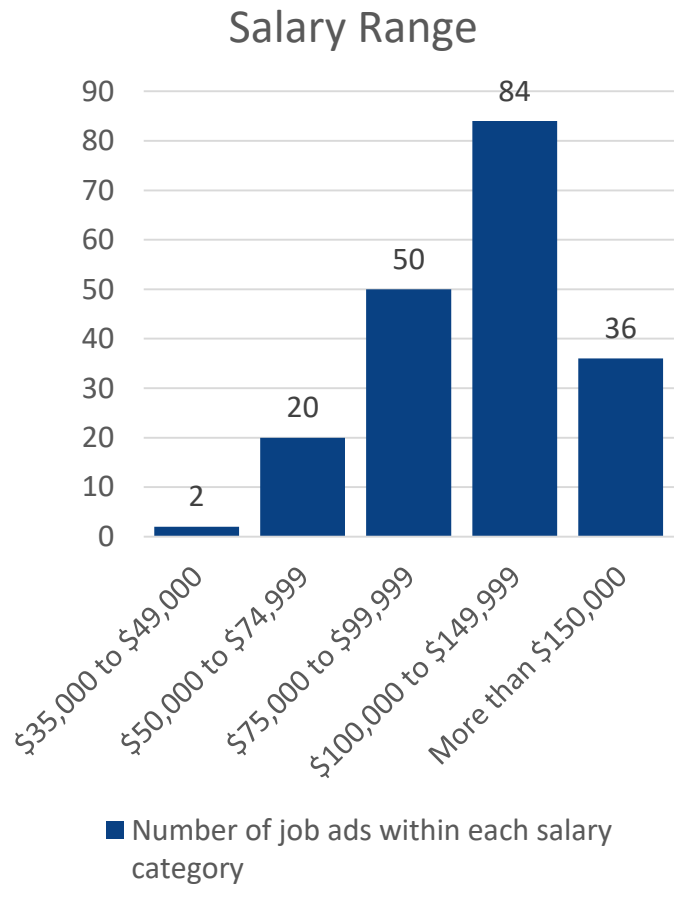
Mean salary: $110,000
Median salary: $100,000

2017



Mean salary: $109,000
Median salary: $100,000

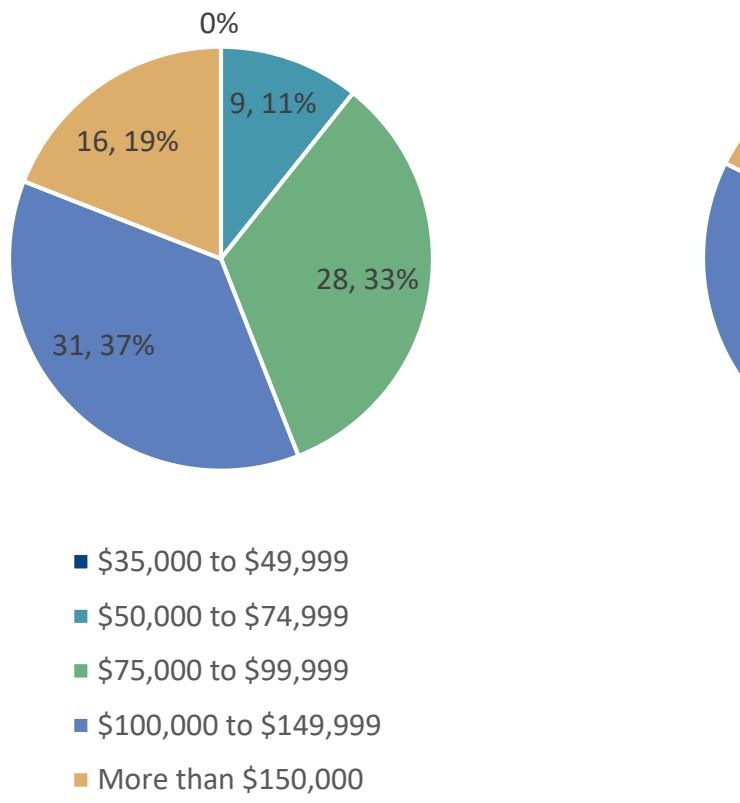
Last 365 Days



Mean salary: $112,000
Median salary: $108,000

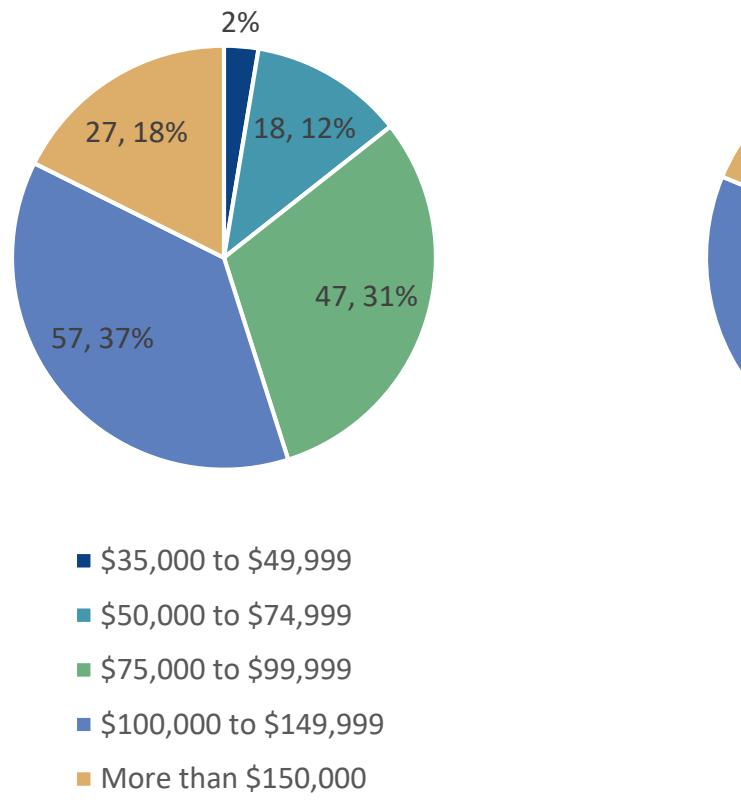
Salaries ! ! ! ! ! ! ! ! !

2016



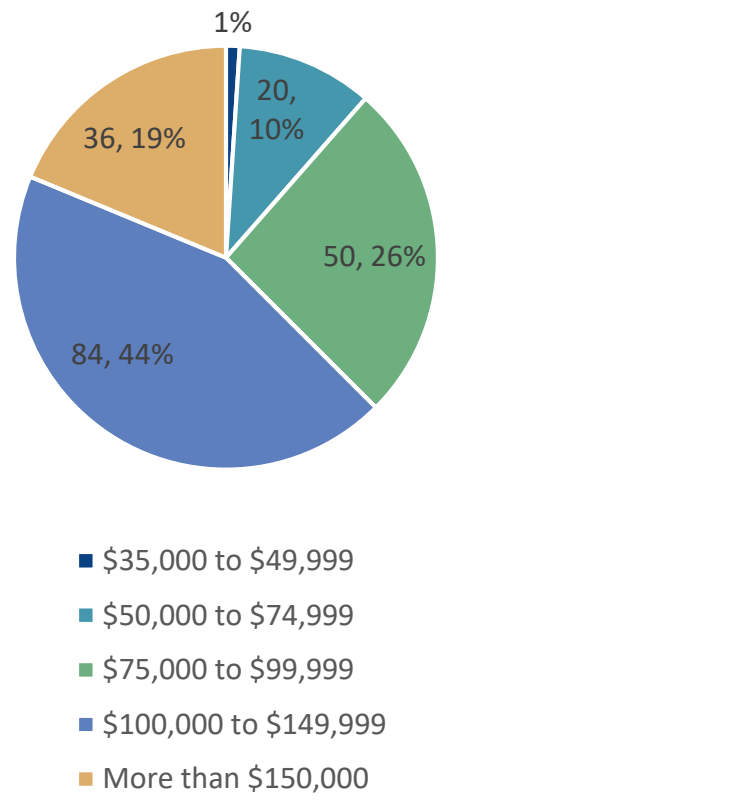
Mean salary: $110,000
Median salary: $100,000

2017



Mean salary: $109,000
Median salary: $100,000

Last 365 Days

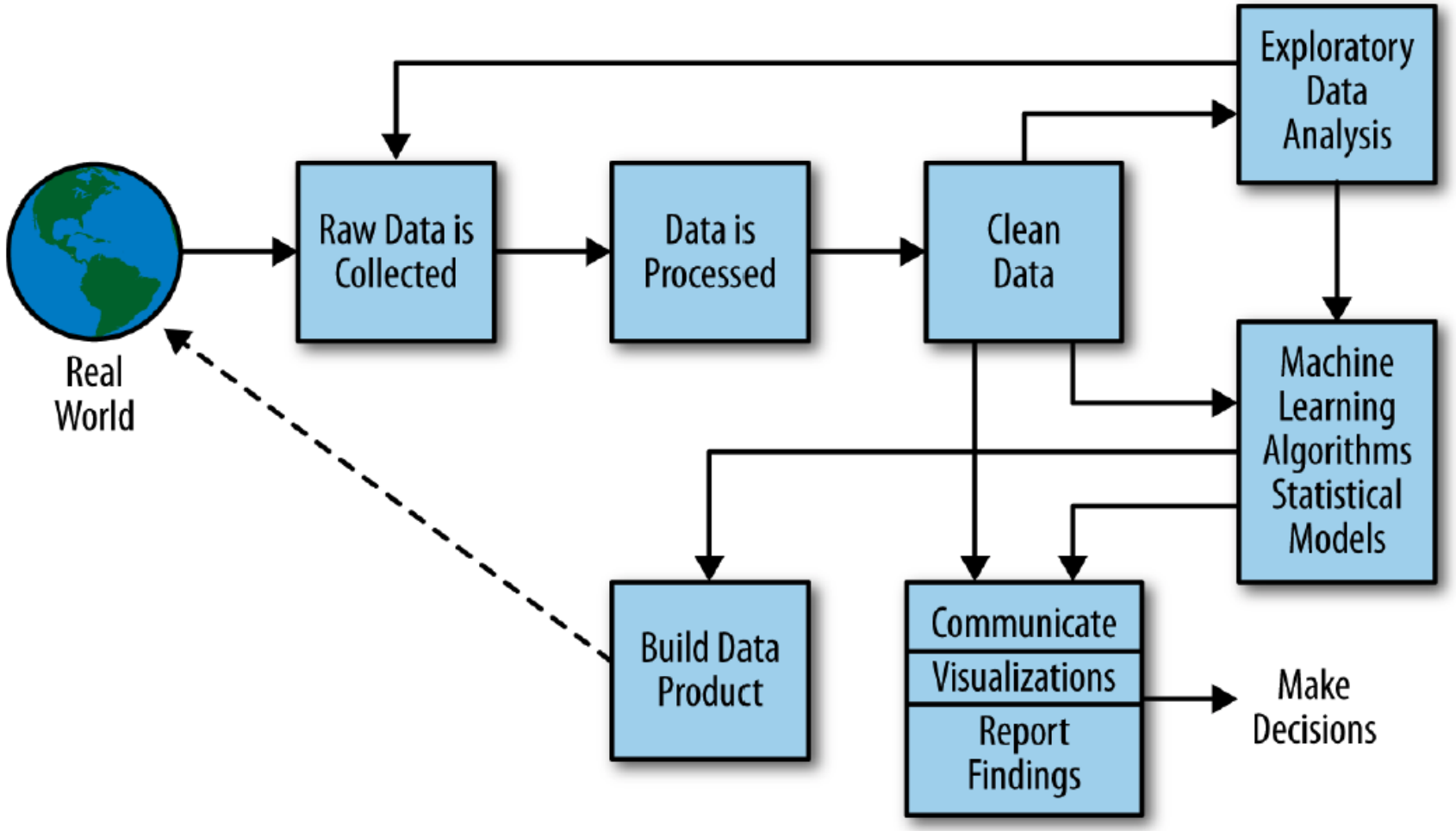


Mean salary: $112,000
Median salary: $108,000

THE UNIVERSITY OF MELBOURNE

Suggestions and Tips

The Data Science Process

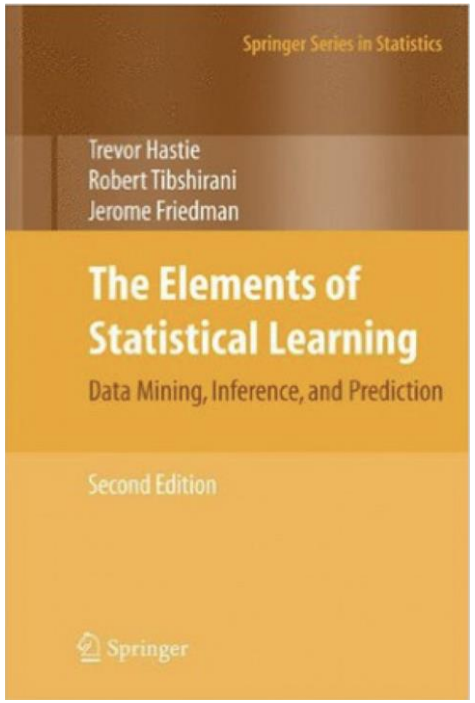


Schutt and O'Neil (2013), *Doing Data Science*. O'Reilly.

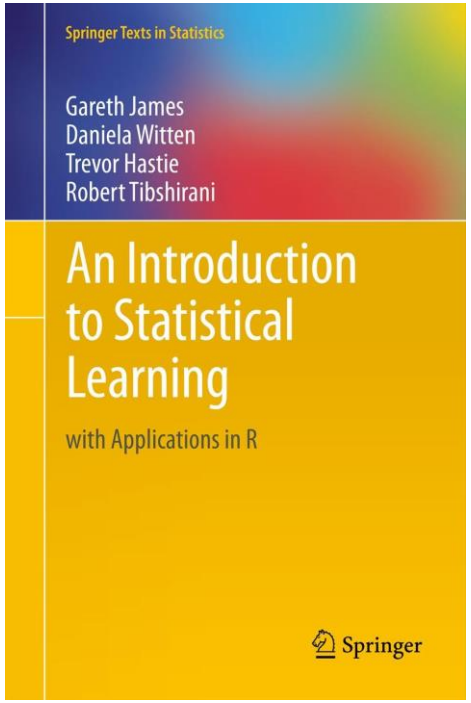
Useful Resources

1. Kaggle https://www.kaggle.com/
2. Coursera https://www.coursera.org/
3. Amazon Web Services https://aws.amazon.com/

The Elements of Statistical Learning
https://web.stanford.edu/~hastie/ElemStatLearn/

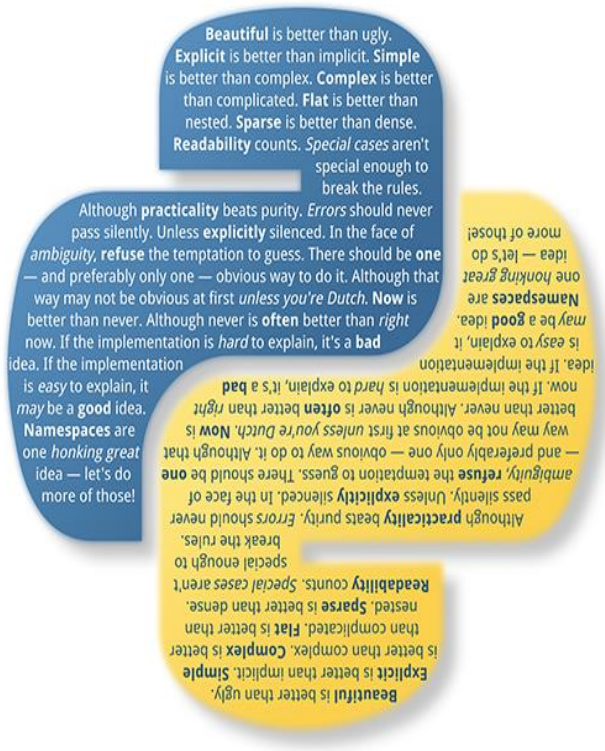


An Introduction to Statistical Learning with Applications in R
http://www-bcf.usc.edu/~gareth/ISL/



Top Python Data Analysis Libraries

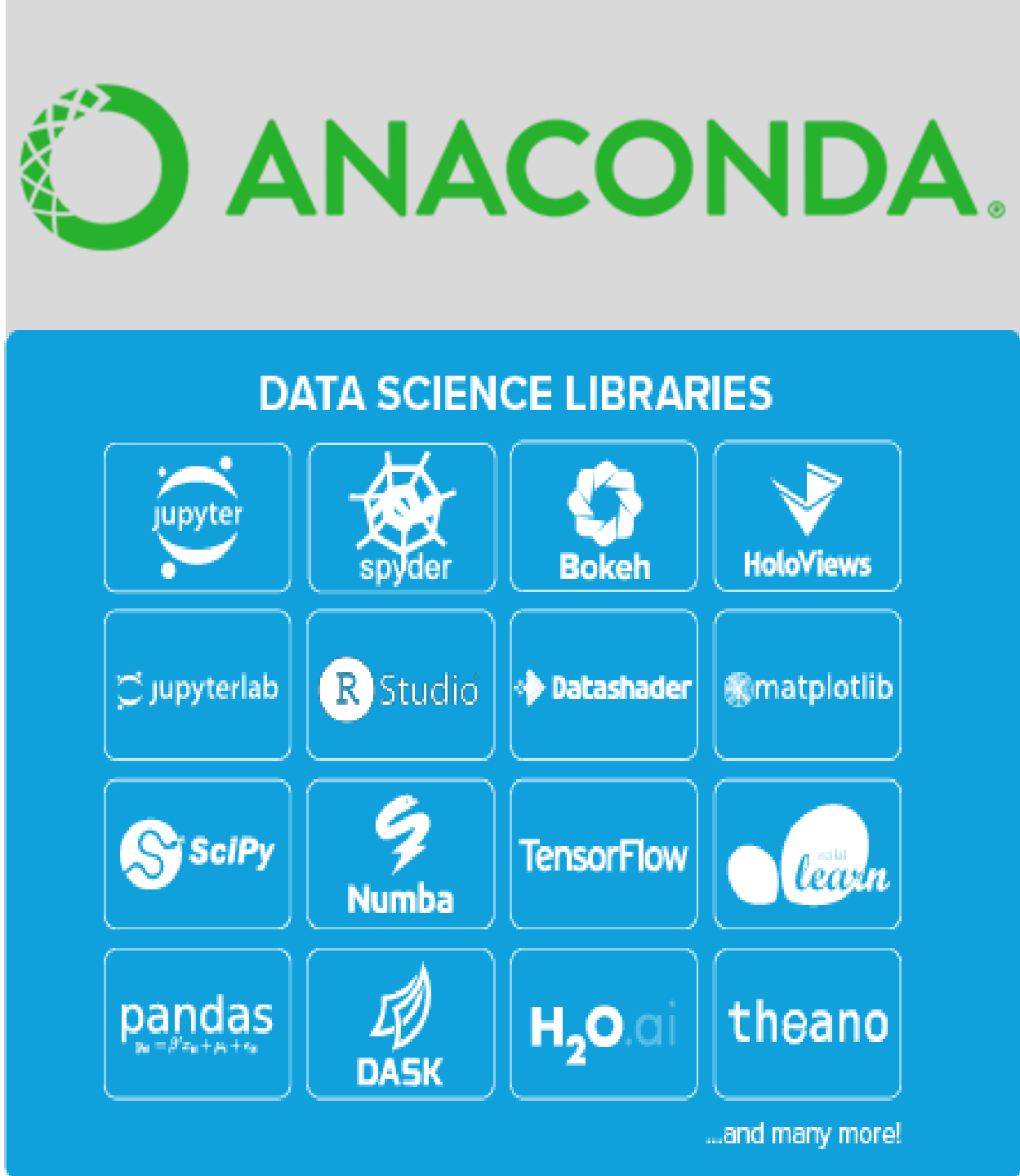
1. Numpy http://www.numpy.org/
2. Scipy https://www.scipy.org/
3. Pandas https://pandas.pydata.org/
4. Matplotlib https://matplotlib.org/
5. Seaborn https://seaborn.pydata.org/
6. Bokeh https://bokeh.pydata.org/
7. Scikit-learn http://scikit-learn.org/
8. Theano http://deeplearning.net/software/theano/
9. Tensorflow https://www.tensorflow.org/
10. BeautifulSoup4, Urllib2, Selenium, Scrapy



python™

Anaconda

1. **Anaconda** is a free and open source distribution of the **Python** and **R** for data science.
2. It aims to simplify package management and deployment.
3. Use **Jupyter Notebook** as IDE
4. Anaconda Distribution is used by over 6 million users, and it includes more than 250 popular data science packages suitable for Windows, Linux, and MacOS.
5. https://www.anaconda.com/download/



High Performance Computing in R

1. *Rcpp, RcppArmadillo, RcppEigen*: integrate R with C++
2. Parallel Computing in R:

   - ✓ *parallel, foreach*: execute the for loop in parallel
   - ✓ *SNOW, doSNOW*
   - ✓ *H2o*: facilitates machine learning (e.g. random forests, GBM) in a parallel environment
   - ✓ *Rhadoop, RHIPE*: interface between R and Hadoop
   - ✓ *Rmpi, pbdMPI*: interface MPI in R
   - ✓ *rgpu, gcbd, OpenCl*: GPU programming
   - ✓ *data.table, ff, bigmemory*: large memory and out-of-memory data

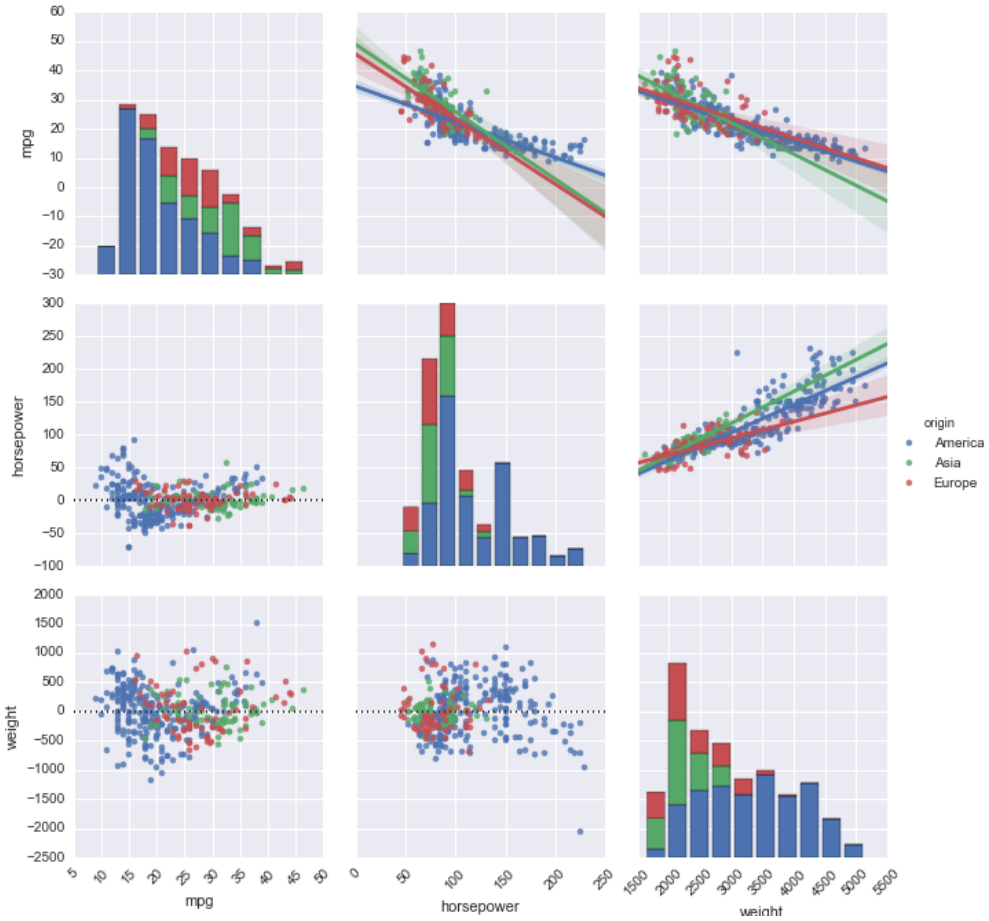
https://cran.r-project.org/web/views/HighPerformanceComputing.html

Useful Packages in R/Rstudio

1. *tidyverse* https://www.tidyverse.org/
2. *rvest*
3. *dplyr, tidyr, stringr, purrr*
4. *lubridate*
5. *ggplot2, maptools, rgdal*
6. *leaflet*
7. *caret, e1071, kernlab, nnet, rpart, xgboost, tensorflow*

https://cran.r-project.org/web/views/MachineLearning.html

8. *shiny* http://shiny.rstudio.com/
9. *knitr*



THE UNIVERSITY OF MELBOURNE

Thank you

Questions?

Liuhua Peng

Lecturer, School of Mathematics and Statistics