Introduction

1. Introduction
    - Course Materials
    - Time-table
    - Software
    - References
    - Getting Help

2. Data Science
    - What is Data Science?
    - Why Data Science?
    - How?

1. Introduction
   - Course Materials
   - Time-table
   - Software
   - References
   - Getting Help
2. Data Science
   - What is Data Science?
   - Why Data Science?
   - How?

Canvas

- All course materials (lecture notes, tutorial questions and tutorial solutions, datasets, etc.) will be uploaded to the workbin of Canvas.
  `canvas.nus.edu.sg`

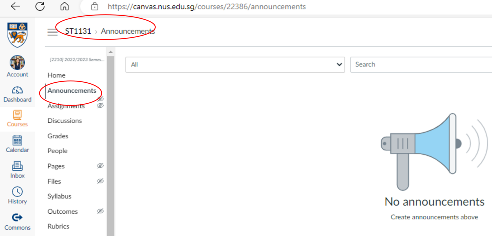
Canvas

- All course materials (lecture notes, tutorial questions and tutorial solutions, datasets, etc.) will be uploaded to the workbin of Canvas.
  `canvas.nus.edu.sg`

- All announcements will be made through Canvas.

Canvas

- All course materials (lecture notes, tutorial questions and tutorial solutions, datasets, etc.) will be uploaded to the workbin of Canvas. `canvas.nus.edu.sg`
- All announcements will be made through Canvas.



https://canvas.nus.edu.sg/courses/22386/announcements

ST1131 › Announcements

No announcements

Create announcements above

Lectures

- There are pre-recorded lectures and live lectures.

Lectures

- There are pre-recorded lectures and live lectures.

- Before attending any live lecture, please finish all the previous lectures for better understanding.

Lectures

- There are pre-recorded lectures and live lectures.
- Before attending any live lecture, please finish all the previous lectures for better understanding.
- All the live lectures are recorded and will be uploaded to Canvas –> Videos/Panopto.

1. Introduction
   - Course Materials
   - Time-table
   - Software
   - References
   - Getting Help
2. Data Science
   - What is Data Science?
   - Why Data Science?
   - How?

Lectures & Tutorials

Class	Days	Time	Venue
Lecture	Tue & Fri	8 – 10 AM	LT 28
Tutorial	Fri	9 - 9:45 AM	LT 28 [1]

- Tutorial attendance is graded.

[1] Starts from Week 3

1. Introduction
    - Course Materials
    - Time-table
    - Software
    - References
    - Getting Help

2. Data Science
    - What is Data Science?
    - Why Data Science?
    - How?

R

- We are using R in this course. Either RGui or RStudio is accepted.

R

- 4.1.0 is the oldest version that is accepted.

R

- 4.1.0 is the oldest version that is accepted.

- The test/exam will require the use of R and will contain R output. Hence you must be familiar with the routines that we call during the lectures/tutorials.

R

- 4.1.0 is the oldest version that is accepted.

- The test/exam will require the use of R and will contain R output. Hence you must be familiar with the routines that we call during the lectures/tutorials.

- You **will be tested** on how to use R to produce the output (numerical/graphical).

1. Introduction
    - Course Materials
    - Time-table
    - Software
    - References
    - Getting Help

2. Data Science
    - What is Data Science?
    - Why Data Science?
    - How?

Recommended Book

- *An Introduction to Statistical Learning with Applications in R*
  2nd edition
  Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.

Recommended Book

- *An Introduction to Statistical Learning with Applications in R*
  2nd edition
  Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data
  2015
  EMC Education Services

1. Introduction
   - Course Materials
   - Time-table
   - Software
   - References
   - Getting Help

2. Data Science
   - What is Data Science?
   - Why Data Science?
   - How?

Asking Questions

- Post your question(s) to Canvas/Discussions.

Asking Questions

- Post your question(s) to Canvas/Discussions.

- Send email to the lecturer via `pham.kimcuc@nus.edu.sg`. Put “DSA1101” somewhere in the title of your email.

Asking Questions

- Post your question(s) to Canvas/Discussions.
- Send email to the lecturer via `pham.kimcuc@nus.edu.sg`. Put "DSA1101" somewhere in the title of your email.
- Book a time slot for consultation.

Topics

1. Introduction to R programming

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours
    - Decision trees

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours
    - Decision trees
    - Regression analysis

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours
    - Decision trees
    - Regression analysis
    - Naïve Bayes

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours
    - Decision trees
    - Regression analysis
    - Naïve Bayes
4. Diagnostics of classifiers

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours
    - Decision trees
    - Regression analysis
    - Naïve Bayes
4. Diagnostics of classifiers
5. Model validation

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours
    - Decision trees
    - Regression analysis
    - Naïve Bayes
4. Diagnostics of classifiers
5. Model validation
6. Unsupervised learning

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours
    - Decision trees
    - Regression analysis
    - Naïve Bayes
4. Diagnostics of classifiers
5. Model validation
6. Unsupervised learning
    - k-means clustering

Topics

1. Introduction to R programming
2. Introduction to basic probability and statistics
3. Supervised learning
    - k-nearest neighbours
    - Decision trees
    - Regression analysis
    - Naïve Bayes
4. Diagnostics of classifiers
5. Model validation
6. Unsupervised learning
    - k-means clustering
    - Association rules

1. Introduction
    - Course Materials
    - Time-table
    - Software
    - References
    - Getting Help

2. Data Science
    - What is Data Science?
    - Why Data Science?
    - How?

1. Introduction
    - Course Materials
    - Time-table
    - Software
    - References
    - Getting Help

2. Data Science
    - What is Data Science?
    - Why Data Science?
    - How?

We are generating more data than ever

Patient
Healthcare
Data
Family History
Exam Records
Patient Symptoms
Admission History
Treatment Records
Ambulatory Data
Prescription History
Affiliated Physicians

Fragmented healthcare data impedes collaboration and patient-centered care

The Data Deluge

`https://explodingtopics.com/blog/big-data-stats`

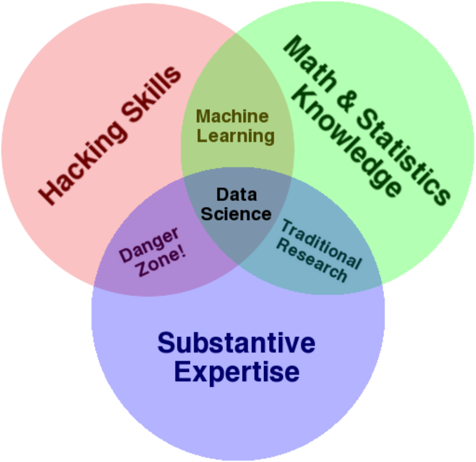
- High volume of data
- Complexity of data types and structures
- Speed of new data creation and growth

What is Data Science?

“The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data.”

Hal Varian on data in McKinsey commentary

What makes a data scientist

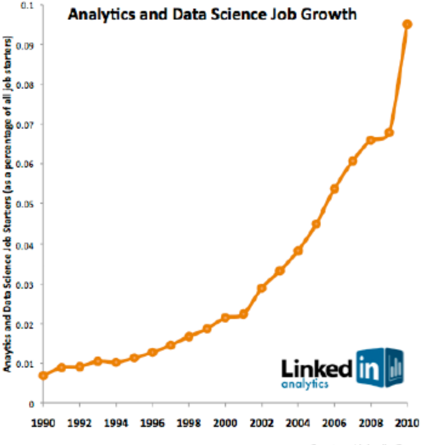


Drew Conway's Data Science Venn Diagram

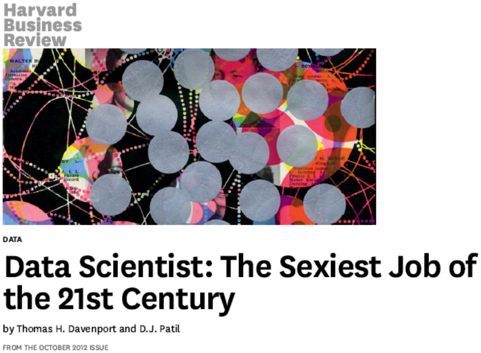
1. Introduction
    - Course Materials
    - Time-table
    - Software
    - References
    - Getting Help

2. Data Science
    - What is Data Science?
    - Why Data Science?
    - How?

A Decade ago...



Courtesy LinkedIn Corp.



Harvard Business Review

DATA

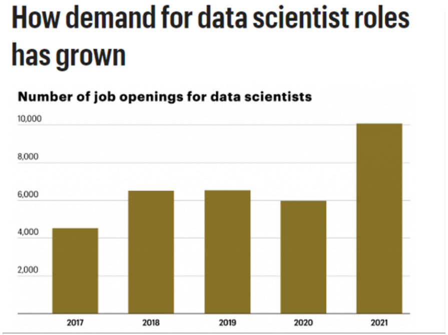
Data Scientist: The Sexiest Job of the 21st Century

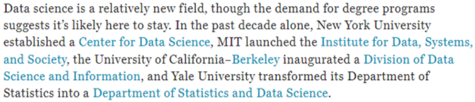
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

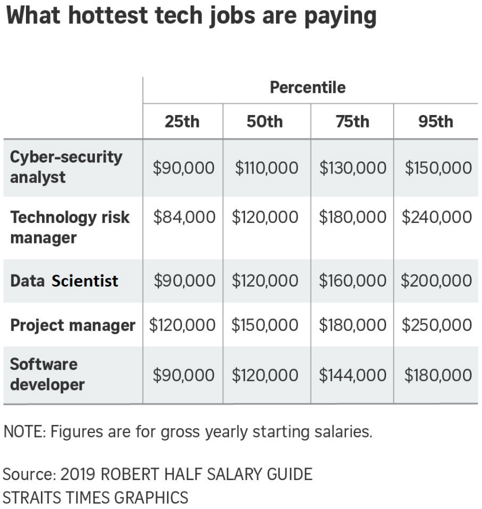
The rise in demand for data science talent via Linkedin analytics (left) and Harvard Business Review (right)

In 2022




Data science is a relatively new field, though the demand for degree programs suggests it's likely here to stay. In the past decade alone, New York University established a Center for Data Science, MIT launched the Institute for Data, Systems, and Society, the University of California–Berkeley inaugurated a Division of Data Science and Information, and Yale University transformed its Department of Statistics into a Department of Statistics and Data Science.

https://fortune.com/education/articles/glassdoors-no-3-best-job-in-the-u-s-has-seen-job-growth-surge-480/



What hottest tech jobs are paying

	Percentile			
	25th	50th	75th	95th
Cyber-security analyst	$90,000	$110,000	$130,000	$150,000
Technology risk manager	$84,000	$120,000	$180,000	$240,000
Data Scientist	$90,000	$120,000	$160,000	$200,000
Project manager	$120,000	$150,000	$180,000	$250,000
Software developer	$90,000	$120,000	$144,000	$180,000

NOTE: Figures are for gross yearly starting salaries.

Source: 2019 ROBERT HALF SALARY GUIDE
STRAITS TIMES GRAPHICS

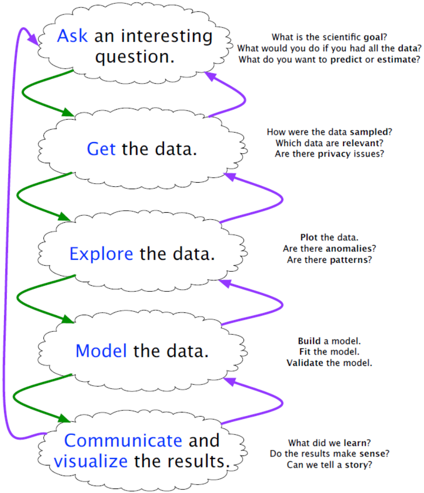
Smart Nation
SINGAPORE

- Urban data
- Health data
- Transport data
- Financial data
- Open data from public agencies

“A Smart Nation is one where people are empowered by technology to lead meaningful and fulfilled lives. Through harnessing the power of networks, data and info-comm technologies, we seek to improve living, create economic opportunity and build a closer community.”

1. Introduction
    - Course Materials
    - Time-table
    - Software
    - References
    - Getting Help

2. Data Science
    - What is Data Science?
    - Why Data Science?
    - How?


Ask an interesting question.
Get the data.
Explore the data.
Model the data.
Communicate and visualize the results.

What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?

How were the data sampled?
Which data are relevant?
Are there privacy issues?

Plot the data.
Are there anomalies?
Are there patterns?

Build a model.
Fit the model.
Validate the model.

What did we learn?
Do the results make sense?
Can we tell a story?

The data science process by Hanspeter Pfister, Joe Blitzstein and Verena Kaynig (`http://cs109.org`)

Our Interest is in Population

Definition 1

The **population** is the total set of subjects in which we are interested. A **sample** is the subset of the population for whom we have, or plan to have, data for. This subset is often randomly selected.

Our Interest is in Population

Definition 1

The **population** is the total set of subjects in which we are interested. A **sample** is the subset of the population for whom we have, or plan to have, data for. This subset is often randomly selected.

How we select our sample affects what population we can generalize the results to.