COMP2200/COMP6200 - Data Science

Dr. Guanfeng Liu (Unit Convenor)

guanfeng.liu@mq.edu.au

Dr. Xuehui Fan (Unit Lecturer)

xuhui.fan@mq.edu.au

Dr. Xuyun (Sean) Zhang (Unit Lecturer)

xuyun.zhang@mq.edu.au

This teaching material was prepared and modified by Steve Cassidy, Sonit Singh, Guanfeng Liu, Yang Zhang, and Xuyun Zhang.

About COMP2200/COMP6200

- **COMP2200/6200 Unit Guide**

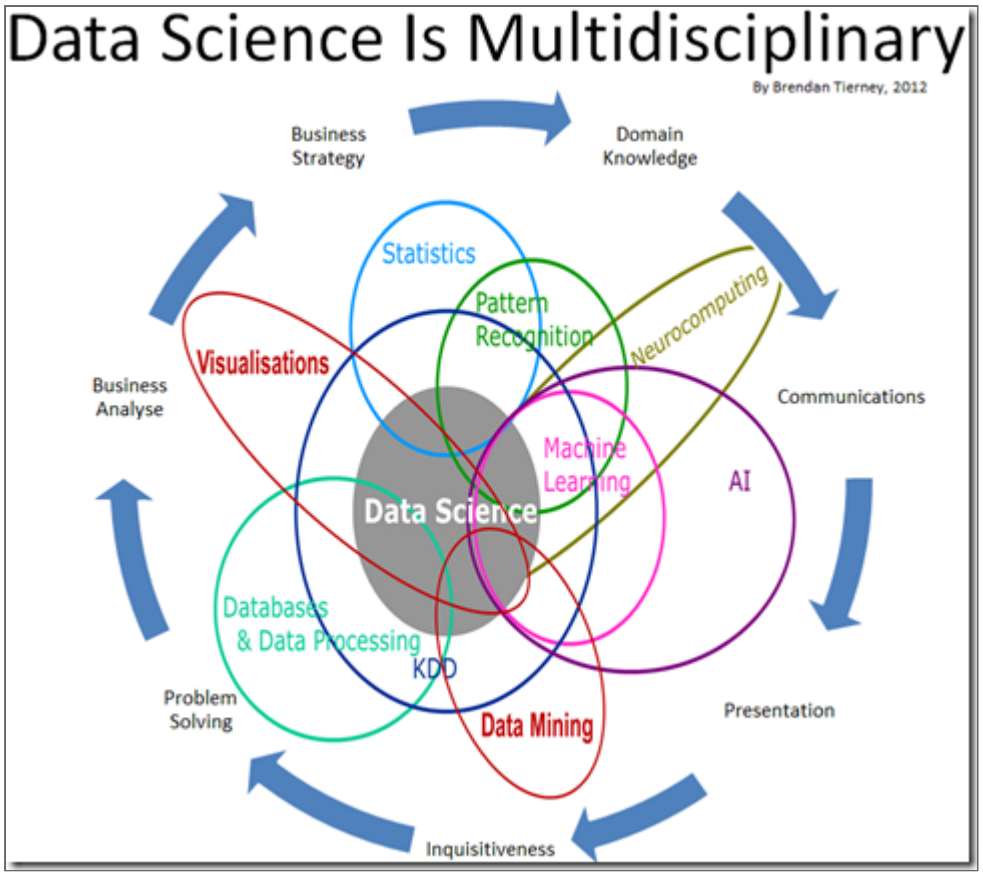
(Both COMP2200 and COMP6200 offerings are the same)

Data Science - Multidisciplinary

Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, psychology, and education.

Data Science - Disciplines

- **Computer Science**
  - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
- **Mathematics**
  - Mathematical Modeling
- **Statistics**
  - Statistical and Stochastic modeling, Probability.


Data Science Is Multidisciplinary
By Brendan Tierney, 2012
Business Strategy
Domain Knowledge
Communications
Presentation
Inquisitiveness
Problem Solving
Business Analyse
Statistics
Pattern Recognition
Neurocomputing
Visualisations
Machine Learning
AI
Data Science
Databases & Data Processing
KDD
Data Mining

What is your reason to choose this Data Science Unit?

How long will this Data Science Heat last?

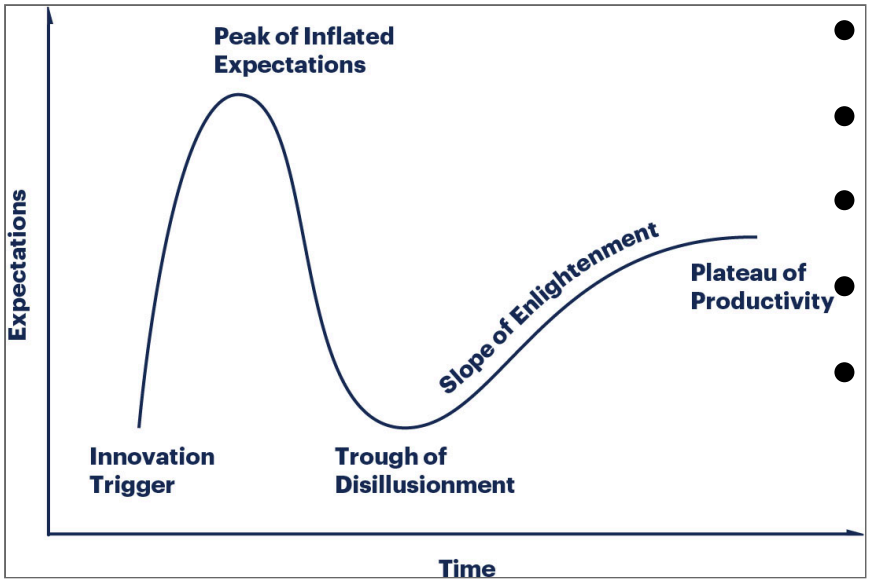
Data Science - Hype Cycle

When **new technologies** (e.g., Data Science, AI, Cloud Computing) make bold promises, how do you discern the hype from what's commercially viable? And when will such claims pay off, if at all?

> ***Gartner Hype Cycle*** *provide a graphic representation of the maturity and adoption of technologies and applications, and how they are potentially relevant to solving real business problems and exploiting new opportunities. Gartner Hype Cycle methodology gives you a view of how a technology or application will evolve over time, providing a sound source of insight to manage its deployment within the context of your specific business goals.*

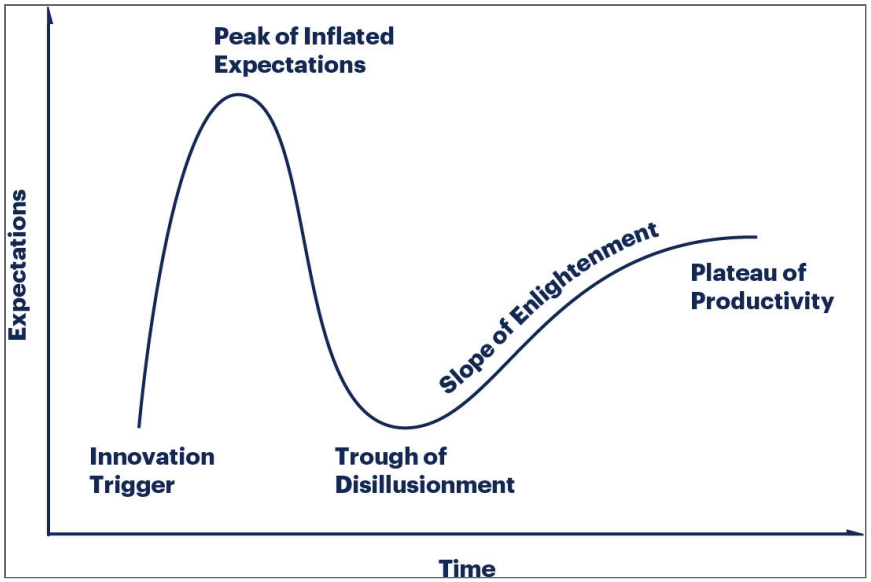
Data Science - Five Key Phases

Each Hype Cycle drills down into the five key phases of a technology's life cycle.



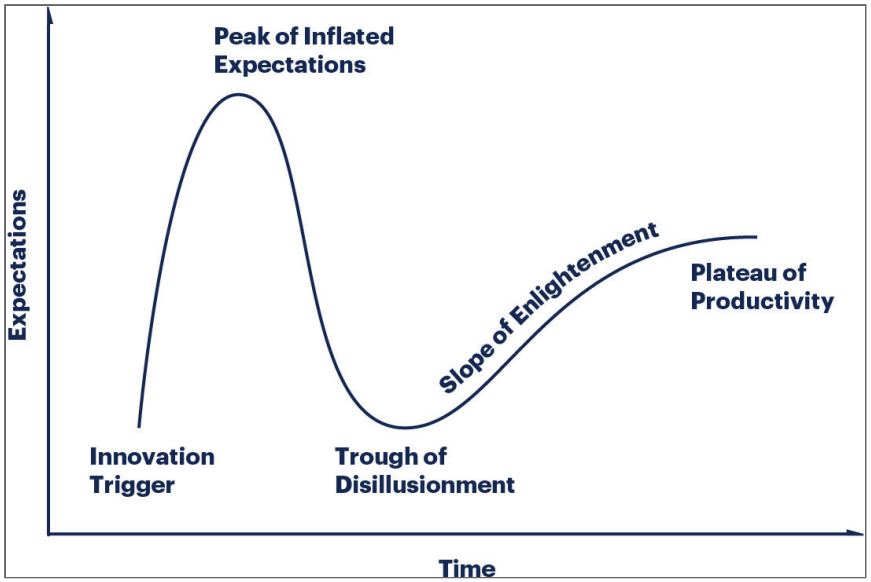
- Innovation Trigger
- Peak of Inflated Expectations
- Trough of Disillusionment
- Slope of Enlightenment
- Plateau of Productivity

1. Innovation Trigger



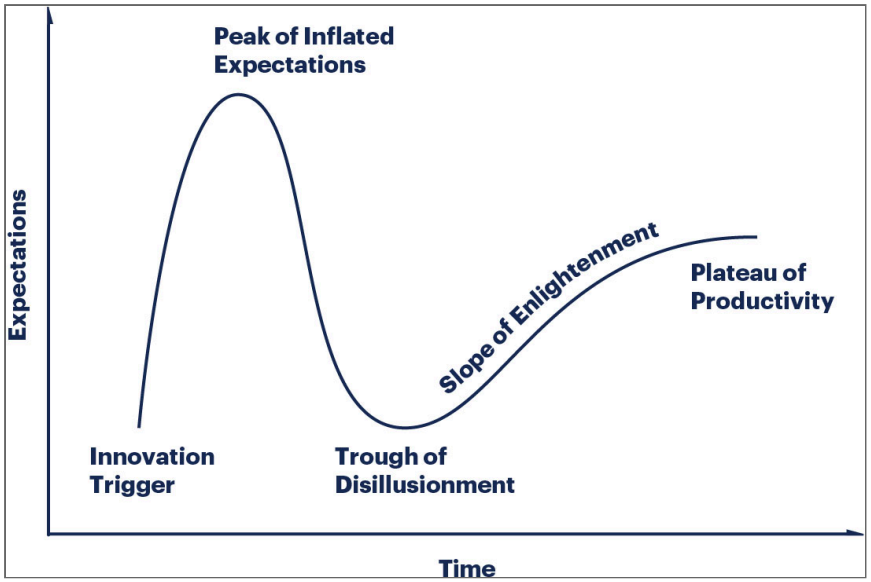
A potential technology breakthrough kicks things off. **Early proof-of-concept stories** and media interest trigger significant publicity. Often no usable products exist and commercial viability is unproven.

2. Peak of Inflated Expectations



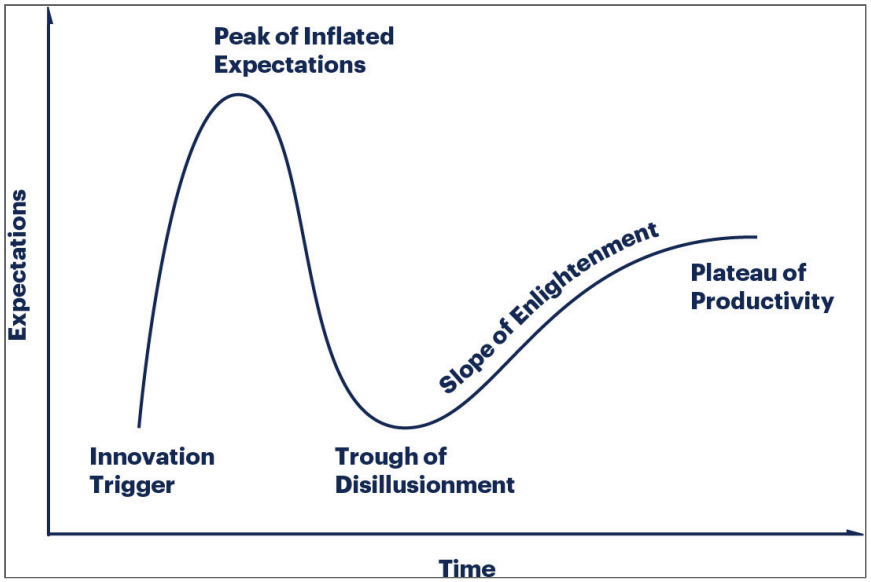
Early publicity produces **a number of success stories** — often accompanied by scores of failures. Some companies take action; many do not.

3. Trough of Disillusionment



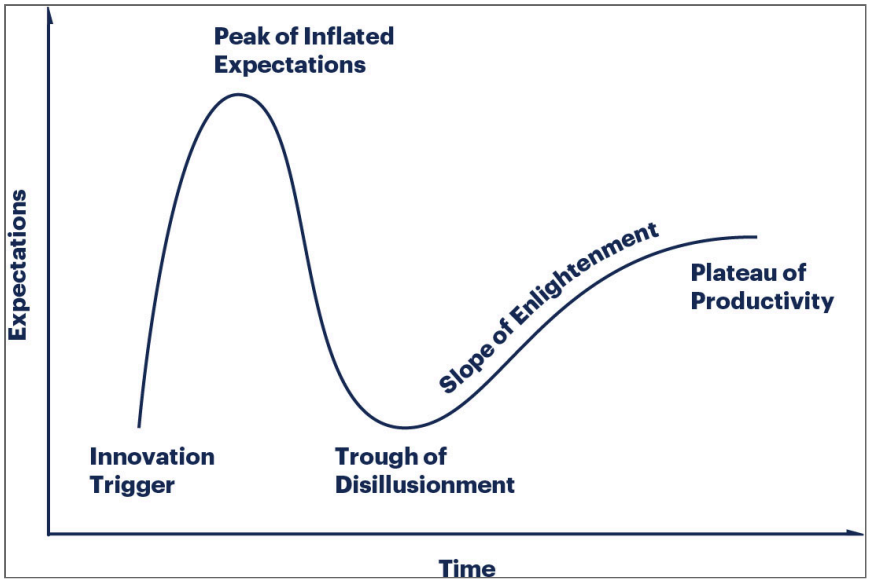
Interest wanes as experiments and implementations fail to deliver. Producers of the technology **shake out or fail**. Investments continue only if the surviving providers improve their products to the satisfaction of early adopters.

4. Slope of Enlightenment



More instances of how the technology can benefit the enterprise start to crystallize and become more widely understood. **Second- and third-generation products appear** from technology providers. More enterprises fund pilots; conservative companies remain cautious.

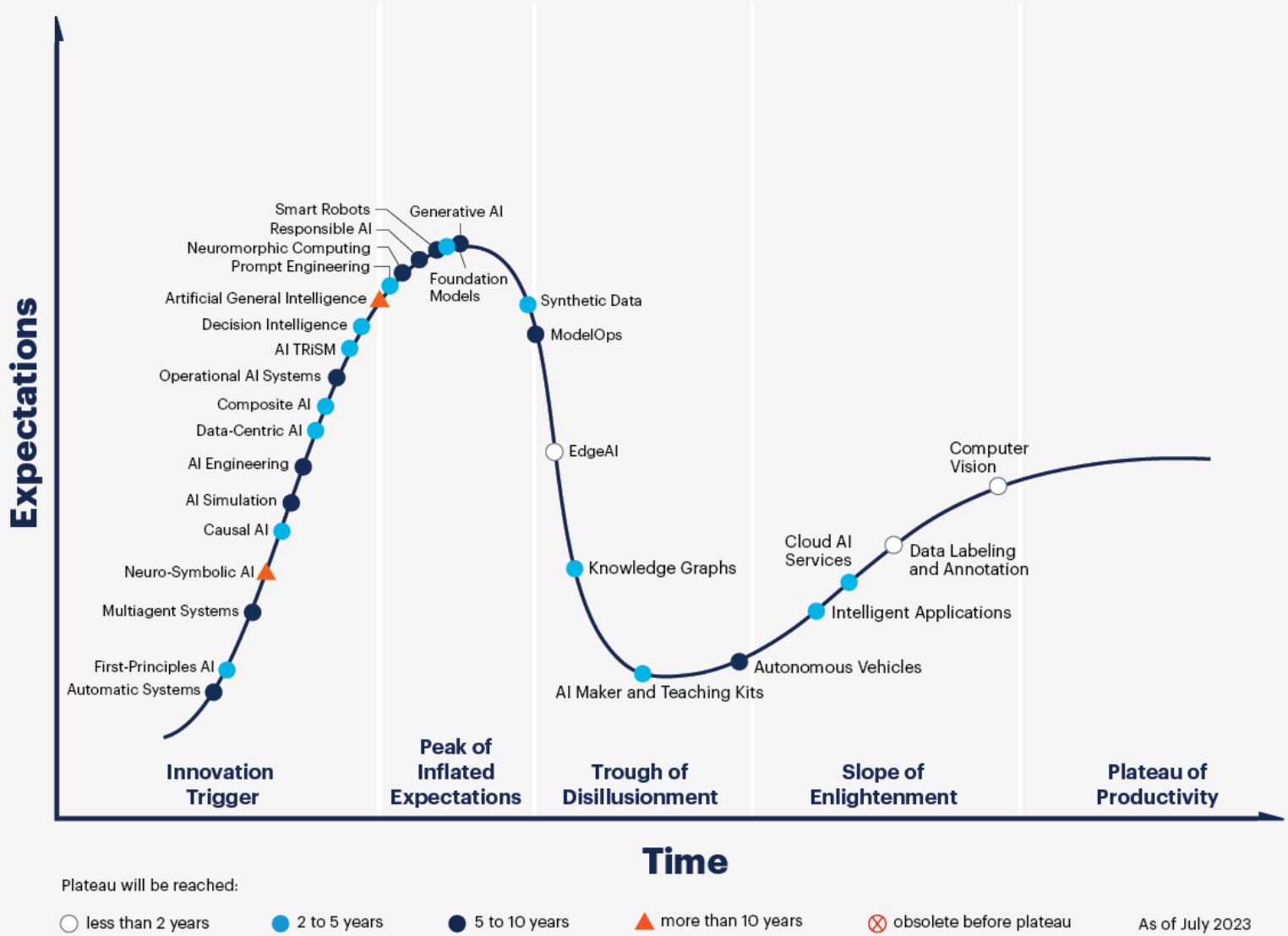
5. Plateau of Productivity



Mainstream adoption starts to take off. Criteria for assessing provider viability are more clearly defined. The technology's broad market applicability and relevance are clearly **paying off**.

Data Science - Gartner Hype Cycle

Hype Cycle for Artificial Intelligence, 2023



gartner.com

Source: Gartner

Gartner

Why Data Science Hype Cycles?

- Separate hype from the real drivers of a technology's commercial promise
- Reduce the risk of your technology investment decisions
- Compare your understanding of a technology’s business value with the objectivity of experienced IT analysts

Data Science Job Market

Data Science Jobs

> *Data science combines several disciplines, including statistics, data analysis, machine learning, and computer science. This can be daunting if you're new to data science, but keep in mind that different roles and companies will emphasize some skills over others, so you don't have to be an expert at everything.* ***Top 10 Careers in Data Science***

How do businesses find applicants?

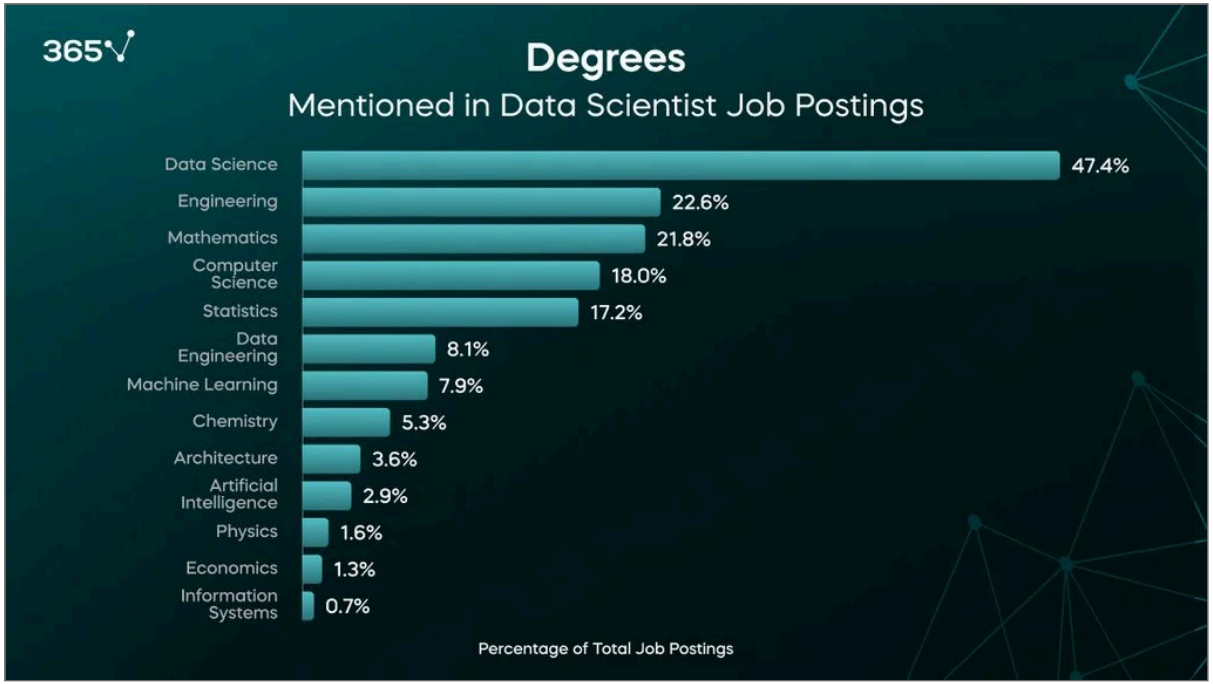
- Universities
- User Group Membership Lists
- Linkedin
- Technology Conference
- Venture Capitalist
- Host a Competition

Reading Data Science Job Descriptions

> *One important piece of advice for your job search is to read data science job descriptions carefully. This will enable you to apply to jobs you're already qualified for, or develop specific data skill sets to match the roles you want to pursue.*

Data Scientist Job Postings

The Data Scientist Job Market in 2024 [Research on 1,000 Job Postings]



Data Scientists

Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions


ANATOMY OF A
DATA SCIENTIST
SALARY
Average salary of data scientists is
$120,000/year
EDUCATION
88% of all data scientists have at least a Master's degree
46% of data scientists have a PhD
BENEFITS
Harvard Business Review called data science the "Sexiest Job of the 21st Century"
One of the fastest growing careers in the United States
94% of data science graduates have found jobs since 2011
SKILLS
Programming languages (R, Python, SQL, Hive, etc.)
Statistics
Multivariable calculus and linear algebra
Machine learning
Software engineering
Wrangle, visualize, and communicate data to management
RESPONSIBILITIES
Conduct research
Extract, clean, and analyze data from varied sources
Solve problems
Build automation tools
Communicate findings to management
CAREER POSSIBILITIES
The majority of data scientists work in the technology industry.
Other options include marketing, consulting, healthcare and pharmaceuticals, finance, government, gaming, and many more.
RESOURCES:
https://insidebigdata.com/2017/08/05/benefits-data-scientist-career/
https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm
https://blog.udacity.com/2014/11/data-science-job-skills.html
https://online.rutgers.edu/resources/infographics/what-can-you-do-with-a-career-in-data-science/?program=mi
THE COMPUTER MERCHANT, LTD.
THE IT STAFFING COMPANY

Four Types of Data Science Jobs

1. Data Analyst

There are some companies where being a data scientist is synonymous with being a data analyst.
Your job might consist of tasks like pulling data out of SQL databases, becoming an Excel or Tableau master, and producing basic data visualizations and reporting dashboards. You may on occasion analyze the results of an A/B test or take the lead on your company's Google Analytics account.

> *"A company like this is a great place for an aspiring data scientist to learn the ropes."*

Once you have a handle on your day-to-day responsibilities, a company like this can be a great environment to try new things and expand your

skillset.

2. Data Engineer

Some companies get to the point where they have a lot of traffic (and an increasingly large amount of data), and they start looking for someone to set up a lot of the data infrastructure that the company will need moving forward.
They're also looking for someone to provide analysis. You'll see job postings listed under both “Data Scientist” and “Data Engineer” for this type of position. Since you'd be (one of) the first data hires, heavy statistics and machine learning expertise is less important than strong software engineering skills.

> *“Mentorship opportunities for junior data scientists can be less plentiful at a company looking to leverage rapidly increasing amounts of data.”*

As a result, you’ll have great opportunities to shine and grow via trial by fire, but there will be less guidance and you may face a greater risk of

flopping or stagnating.

3. Machine Learning Engineer

There are a number of companies for whom their data (or their data analysis platform) is their product.
In this case, the data analysis or machine learning can be pretty intense.
This is probably the ideal situation for someone who has a formal mathematics, statistics, or physics background and is hoping to continue down a more academic path.

> *"Machine Learning Engineers often focus more on producing great data-driven products than they do answering operational questions for a company."*

Companies that fall into this group could be consumer-facing companies with massive amounts of data or companies that are offering a data-based service.

4. Data Science Generalist

A lot of companies are looking for a generalist to join an established team of other data scientists. The company you're interviewing for cares about data but probably isn't a data company.
It’s equally important that you can perform analysis, touch production code, visualize data, etc.

> *“Some of the most important ‘data generalist’ skills are familiarity with tools designed for ‘big data,’ and experience with messy, ‘real-life’ datasets.”*

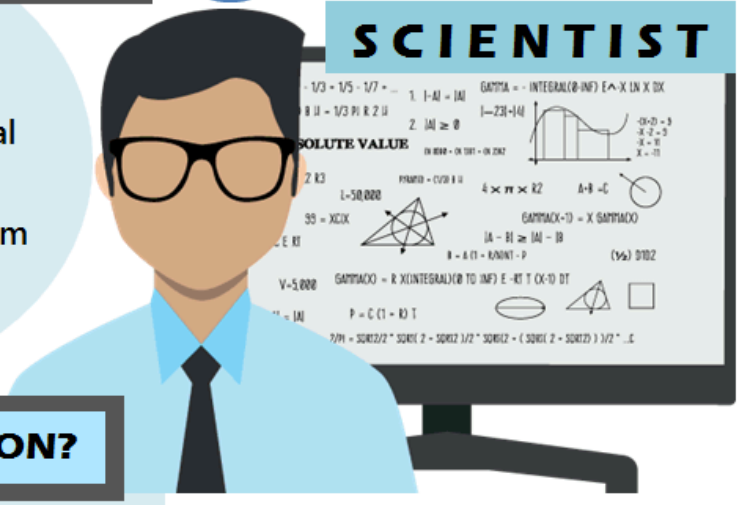
Generally, these companies are either looking for generalists or they’re looking to fill a specific niche where they feel their team is lacking, such as data visualization or machine learning.

What do Data Scientists do?

DATA SCIENTIST

WHO AM I?

I am a part analyst & part artist. I use my analytical and technical abilities to extract meaning / insights from massive data sets.



WHAT DO I DO?

1. I cleanse existing raw data & build models to predict future data.
2. I go beyond merely collecting and reporting data, to look at data from multiple angles & give meaning to it.
3. I identify the correct business problem(s) & offer solutions (via visualizations, reports or blogs) by best applying the data.

WHAT DO I RELY ON?

1. Analytics
2. Predictive Models
3. Statistical Analysis & Modeling
4. Data Mining
5. Sentiment Analysis
6. What-if Analysis

THE PROCESS I FOLLOW

- **Define Problem**
- **Structure Data**
- **Use Programming Language**

WHAT DO I EARN?

After oil & gas geologists, mine is the 2nd highest paid job in the world!

100,000 to 150,000

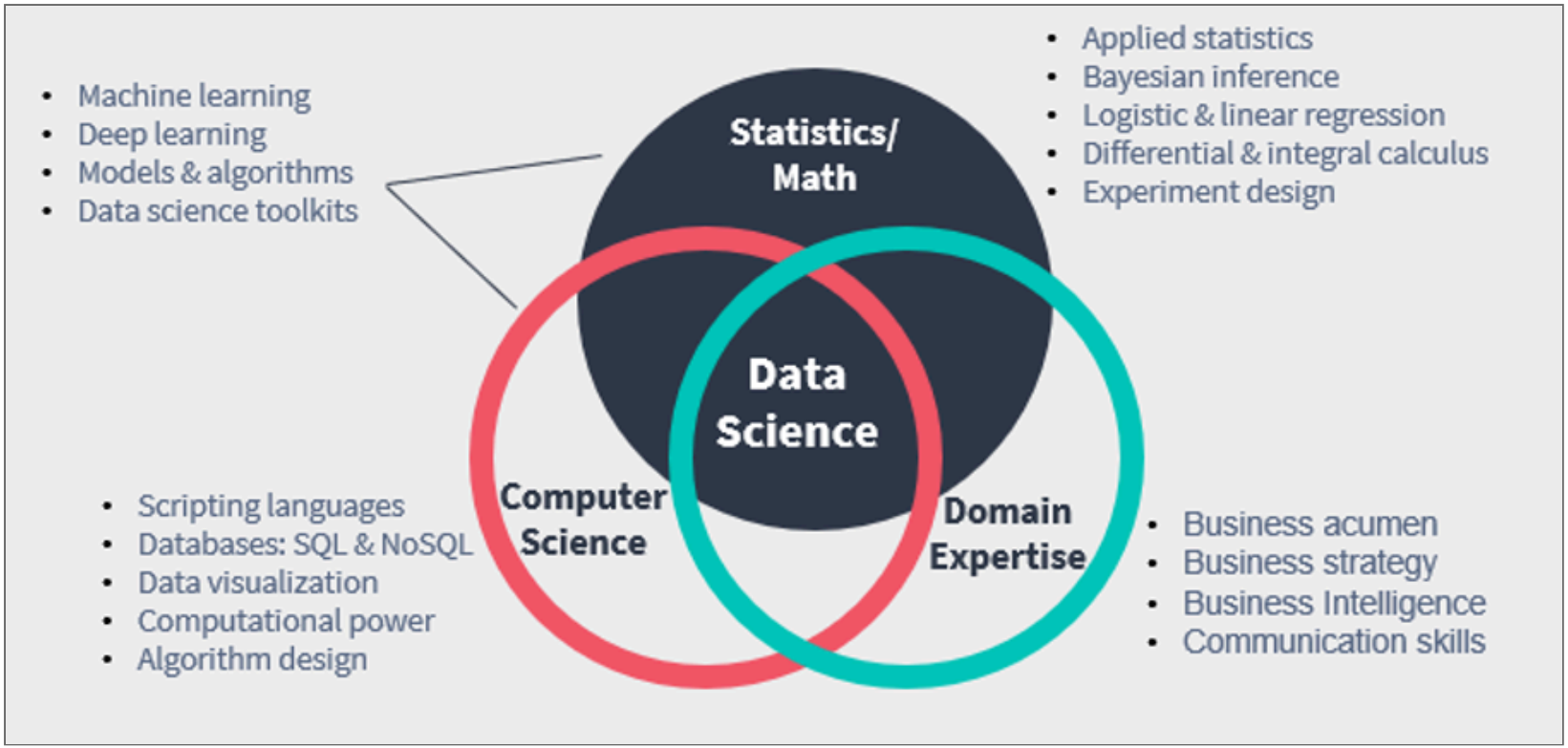
HOW DO I HELP ORGANIZATIONS TODAY?

- Increase data accuracy
- Develop strategies
- Improve operational efficiency
- Reduce costs
- Mitigate risks
- Offer personalized products/services

Data Scientists Skills

- Knowledge in Stats, Data Mining, and Machine Learning
- Open source tools such as Python
- Data Visualisation
- Data warehousing and architecture
- Coding skills
- Soft skills like communication, teamwork, ethical factors, etc.

Data Scientist Skill Set



What is Data Science?

What is Data Science (NYU)

> *At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them. With such automated methods turning up everywhere from genomics to high-energy physics, data science is helping to create new branches of science, and influencing areas of social science and the humanities.*

What is Data Science (Investopedia)

> *A field of Big Data which seeks to provide meaningful information from large amounts of complex data. Data Science combines different fields of work in statistics and computation in order to interpret data for the purpose of decision making*

What is Data Science (Berkeley)

> *There is significant and growing demand for data-savvy professionals in businesses, public agencies, and nonprofits. The supply of professionals who can work effectively with data at scale is limited, and is reflected by rapidly rising salaries for data engineers, data scientists, statisticians, and data analysts.*

Defining Data Science - Science

- How do we characterise Science and Scientific methods?
- What is the alternative to or opposite of Science?
- Can we trust scientific results? Are they more trustworthy than the alternatives?
- What are the most important characteristics of Science?
- How does this all relate to **Data Science**?

Logical Positivism describes the mainstay of the scientific method. **Positivism** means that conclusions are based on observations and evidence. **Logical** means that deductions are made according to the rules of logic. Together these describe a methodology of research that draws conclusions based on logical deductions from the foundations of observation and evidence.

Here's a Science Checklist from **Berkeley** - a classroom activity to help kids understand what science is:

Defining Data Science - Data

Data Science clearly involves Data. What will we need to learn about to become a Data Scientist?

- Data is generated from a range of sources: instruments, surveys, human behaviour, models, simulations
- The size of data varies: hundreds, thousands, millions of observations (when is it Big Data?)
- How data is made available: files in various formats, streams of data over the network, databases, the web
- What is data: not just simple observations and measurements, text as data, metadata as data

***IBM estimates that 90 percent of the data in the world today has been created in the past two years.* source Berkeley**

You will learn some data wrangling skills:

- Reading different kinds of file format
- Cleaning data:
    - filling in missing values,
    - making values consistent (Australia, Aus, Oz)
    - merging data sources
    - selecting rows and columns
    - generating aggregate data

Defining Data Science - Analysis Methods

Once we have our data, one goal of DS is to perform some kind of analysis to try to support a hypothesis or tell a story. Analysis methods include:

- summary statistics
- data smoothing, dimensionality reduction, feature selection
- visualisation
- statistical hypothesis testing

This will draw on you basic statistical knowledge (STAT1170) and extend it in a few areas.

Defining Data Science - Modelling and Prediction

A big part of Data Science is about going beyond the observations that you have in your data to be able to predict future outcomes or make judgements about unseen data based on past observations.
Modelling is the use of past data to build a computational model of a system. That model can then be used in different ways: to classify unseen data (is it spam) attach labels to data (face recognition, image labelling) or predict future trends (where will the stock market go next).

Modelling Techniques

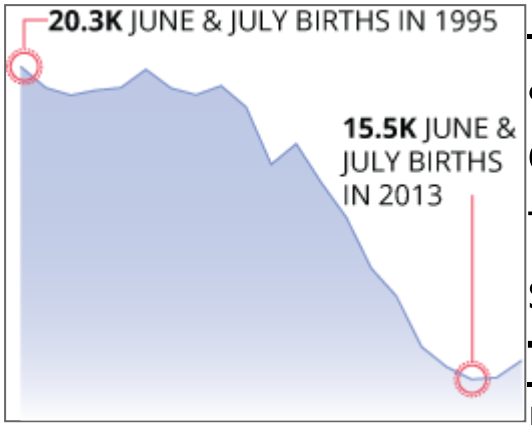
We use many techniques for building models, we'll learn a few in this unit:

- **Linear Regression** works for simple systems where the relationship between observeable variables is linear, it can be extended to deal with more complex relationships
- **Logistic Regression** allows us to predict categorical outcomes (spam or ham, good investment or bad)
- **Clustering** algorithms find patterns in data and show us groupings that might not be obvious
- **Classification** algorithms allow us to label unseen data based on past observations

Machine Learning

All of these are really examples of **Machine Learning** to some degree. They rely on seeing past examples of data that has known outcomes and enable us to build models that can predict outcomes for new observations. The difference between the models that we will see and full on Deep Learning models is one of complexity.
The harder the problem you are trying to build a model of (eg. speech recognition, language understanding) the more input parameters you generally have and the more **degrees of freedom** in your model are needed to account for the complex system you are learning. Many degrees of freedom requires more data to train and longer, more complex training process.

Defining Data Science - Story Telling



The Timing of Baby Making from The Pudding is a nice example of a story told with supporting data. The Pudding is a great example of a website that publishes regular data supported stories and showcases DS methods.

The Guardian Data Blog also publishes data based stories.

Charting the COVID-19 Spread in Australia did analyis of COVID-19 cases per state in Australia.

Defining Data Science - Interpretation and Reproducibility

Using DS methods we can take large amounts of data and find patterns, display summaries, create visualisations, draw conclusions. So what does it mean to be responsible in this context?

Defining Data Science - Interpretation and Reproducibility

The phrase **'Lies, Damn Lies and Statistics'** is a comment on the way that statistics are used to mislead people. Show a graph or a table of numbers that appears to back up your argument and you sound more convincing. For an in-depth treatment of this in the context of DS see: **Calling Bullshit In the Age of Big Data** (we may draw on some of their presentations and activities this semester).
So how do we protect against this kind of deception?

Go back to our Science discussion - conclusions should be based on **observations of evidence**. So if I present a graph, I should also present the sources of data that I used to generate the graph.
Even with the same data, I can't fully understand your results without knowing what methods you used to analyse the data. For example, did you use all of the data or did you exclude some observations? Did you smooth, scale or otherwise manipulate the data? What classification, clustering or modelling methods did you use?

Reproducibility

Reproducibility is an increasingly important goal in Science and Data Science is playing a big role in helping to achieve it. Scientists are concerned that many published studies could not be repeated by other groups because of this (see **Replication Crisis on Wikipedia**). Part of the response is an increased awareness of the need to make research data openly available and to publish the software and methods that was used to generate the results. See **Open Science Foundation**, **Zenodo**, **Figshare** as examples of services that have been established to help scientists share research data and software.

Data Reasoning in a Digital World

Graphs can be a powerful way to share data with the public. But not all graphs are created equal. Choosing the wrong graph type, mislabeling axes, or using an inappropriate or inconsistent scale can affect the way data appears—which can lead readers to misinterpret the data.

Source

How to spot misleading graphs

5 Ways Writers Use Misleading Graphs To Manipulate You

How Fermi Estimation can help you identify Bullshit in this information age.

- Physicist Enrico Fermi Was a Master of Guesstimation
- Order of magnitude estimates of the expected answer
- Useful skill when evaluating your answers to problems
- Needs some knowledge of quantities in the world:
    - Sizes of things
    - Population of countries
    - Frequency of events
- Understand orders of magnitude - is it closer to 10, 100, 1000?

Can you estimate number of piano tuners in Sydney?

Assumptions

In any Fermi problem, we first lay out what it is we need to know, then list some assumptions:

- How often pianos are tuned?
- How long it takes to tune a piano?
- How many hours a year the average piano tuner works?
- The number of pianos in Sydney?

Assumptions

- **Assumption 1**: The average piano owner tunes his piano once a year. (Just a guess since the average piano owner isn't tuning only one time every ten years, nor ten times a year. One time a year seems like a reasonable guesstimate.)
- **Assumption 2**: It takes 2 hours to tune a piano. A guess. Maybe it's only 1 hour, but 2 is within an order of magnitude, so it's good enough.
- **Assumption 3**: Let's assume 40 hours a week. 40 hours a week x 50 weeks approximately 2,000-hour work year.
- **Assumption 4**: 2 pianos for every 100 people (Again a guess to include schools and institutions with pianos)
- **Assumption 5**: Sydney population approximately 5 million

Source:WIRED

Let Estimate

1. There are 5 million people in Sydney.
2. There are 2 pianos for every 100 people.
3. There are 100,000 pianos in Sydney (10^7 / 10^2 = 10^5)
4. Pianos are tuned once a year.
5. It takes 2 hours to tune a piano.
6. Piano tuners work 2,000 hours a year.
7. In one year, a piano tuner can tune 1,000 pianos (2,000 hours per year ÷ 2 hours per piano).
8. It would take 100 tuners to tune 100,000 pianos (100,000 pianos ÷ 1,000 pianos tuned by each piano tuner).

So approximately **100 piano tuners in Sydney**

How much toilet paper do we use every year in Australia?

Assumptions

- Average person uses 1 roll per week
- 1 sheet is approximately 10 cm long
- Australia's population approximately 25 million
- Earth's circumference approximately 40,000 Kilometre.

Source: Fermi Problems: from toilet paper to housing the world

Let's calculate

1. Average person consumes 50 rolls per year (1 roll per week x 50 weeks per year)
2. Length of 1 roll = 30m (300 sheets x 10 cm per sheet)
3. Average person consumption per year = 50 x 30m = 1.5 Kilometre
4. Total consumption in Australia per year = 25 million x 1.5 Kilometre = approximately 40 million Kilometre
5. 40 million Kilometre / 40, 000 Kilometre = 1000

So we can wrap our Earch 1,000 times with 1 year consumption of toilet paper in Australia

Technology

You will be learning to use **Python** for this analysis with the associated statistical and graphical tools. It is also common to see the **R language** used for Data Science - you'll see that if you do the Statistics units in the DS major.
We'll use Git and **Github** to host and share our work and track changes.
We're using **Jupyter Notebooks** to carry out analysis and generate documented presentations of results.

Jupyter Notebooks

The **Jupyter** is in part a response to this need to provide more easily replicated analysis. You'll see how it enables us to create a single document that contains both the commentary on the data and analysis and the code that generates the results. Sharing a notebook as part of the product of your research can enhance the reproducibility of your work, which makes it better science.

Github

The motivation for using Github is twofold.

- Using Git to track changes provides another trace of evidence for what we did
- Might help someone understand the train of thought in the experiment
- Using Github to **publish** the repository makes it available to all
- Others can clone the repository and run **your code**

Acknowledgements

- **Fermi Problems: from toilet paper to housing the world**
- **Calling Bullshit in the Age of Big Data**
- **5 Ways Writers Use Misleading Graphs To Manipulate You**