Master of DATA SCIENCE

Providing new perspectives to flourishing data science industry

Apply now for entry in September 2022

According to the recommendation of a columnist at Forbes, the HKU Master of Data Science programme was ranked as Top 10 among AI and data science master's courses for 2021 in the world.

TOP 10

Machine learning
Data Analytics
Advanced Statistical Modelling
Social Networks
Cloud Computing
IS THE PROGRAMME FOR YOU

• Jointly offered by Department of Statistics and Actuarial Science and Department of Computer Science
• The curriculum of Master of Data Science (MDASC) programme adopts a well-balanced and comprehensive pedagogy of both statistical as well as computational concepts and methodologies, underpinning applications that are not limited to business or a single field alone

World-class Rankings of HKU

Quacquarelli Symonds (QS)
QS World University Rankings by Subject 2021:
- #22 World Rankings 2022
- #3 Asia Rankings 2022

Times Higher Education (THE)
THE World Rankings 2022:
- #4 Asia Rankings 2021

Top-notch Scientists in the Faculty
Clarivate Analytics’ Essential Science Indicators 2019:
16.5% of our professoriate staff are the world’s Top 1% scholars

IS THE PROGRAMME FOR YOU

• Interdisciplinary and comprehensive curriculum
• Solid foundation in statistical and computational analyses
• Electives over a broad range of contemporary topics from Computer Science, Mathematics and Statistics
• Capstone project with real-life scenario

Where will this Programme Lead You

Transferable skills
- Up-to-date knowledge in data science helping to decipher the data and extract valuable information that can be used as a strategic part of critical decision-making
- Hands-on training in data science methodologies using powerful software, enhancing competency for data-scientists who require advanced computing and modelling skills
- Collaboration and communication of disciplinary knowledge in data science to specialists and the general public, and ability to appraise professional ethics

Targeted Taught Postgraduate Programmes Fellowships Scheme
MDASC is selected as an eligible programme under the University Grants Committee for Targeted Taught Postgraduate Programmes Fellowships Scheme. Selected local students admitted to the MDASC (full-time or part-time) in the academic year 2022-23 are eligible to apply (with terms and conditions apply).

Local offer recipients who wish to apply for the Fellowship Scheme should prepare a proposal on how they can contribute to the priority areas (i.e. Research and STEM) of Hong Kong after completing MDASC. Successful Fellowship Scheme applicants will each receive an award of HK$120,000.

Suoxinda Scholarship in Data Science
Two scholarship recipients, each receiving HK$20,000, would be selected from the students entering the MDASC programme based on academic merit and admissions interview performance.

Reimbursable Course(s) by Continuing Education Fund (CEF)
The following courses have been included in the list of reimbursable courses for CEF purposes:
- COMP7503 Multimedia technologies
- COMP7506 Smart phone apps development
- COMP7507 Visualisation and visual analytics
- STAT8017 Data mining techniques
- STAT8019 Marketing analytics

All CEF applicants are required to attend at least 70% of the courses before they are eligible for fee reimbursement under the CEF.

The mother programme (MDASC) of these courses is recognised under the Qualification Framework (QF Level 6)

Department of Statistics and Actuarial Science

Strongly tied with international professional bodies in statistics and actuarial science, the Department of Statistics and Actuarial Science (SAAS) enjoys a very high profile in both teaching and research. SAAS research areas span from classical areas of statistics, to a range of applied domains, and the rapidly developing areas of big data and artificial intelligence. In addition to the HKU-TCL Joint Research Centre for Artificial Intelligence, the SAAS Data Science Lab has been established to serve as platform for interdisciplinary research.

An answer to the desperate call for experts in processing complex digital data, SAAS jointly offers the MDASC programme with the Department of Computer Science, teaching students how to analyse data and formulate data-driven strategies.
WHAT YOU WILL LEARN

Those whose interest in high-level analytical skills straddles the disciplinary divide between statistics and computational analytics

Those who wish to pursue further study in the field of data science after studying science, social sciences, engineering, medical sciences, information systems, computing and data analytics in their undergraduate studies

Examples of backgrounds of admitted students:
- CEO
- Director
- Vice-President
- Head of Global Markets
- Principal Application Analyst
- Senior Analyst Programmer
- Business Analytics Manager
- Compliance Manager
- Software Development Engineer
- Quantitative Researcher
- Solution Developer
- Teacher

Hear from our graduates

Kwok Fung CHOW, Class of 2021
Senior Quantitative Developer, Nautilus Technology Limited

MDASC is a comprehensive master degree in theory in statistics to real world examples in computer science. With the elective subjects, I could not only study spatial data analysis for interest to know more about how to handle geographical data, but also could take the course for data visualisation for work to illustrate the data analysis in an explainable and attractive way using Tableau to target audiences. With this programme, I enhance my knowledge in statistical theory and learn more about the latest data science topics, such as cloud computing, deep learning, text mining and so on. It would be a good programme for those who would like to know more about data science or work as a data scientist. After finishing the programme, I am now more confident in explaining my models to my colleague and have a more clear direction on improving the accuracy, speed for current models and even deploy more complex machine learning models for my daily task.

Kai WANG, Class of 2021
Machine Learning Engineer, ByteDance

The MDASC programme has a well-balanced and flexible curriculum in Statistics and Computer Science. As a machine learning engineer, I work with large volume of data every day, which requires strong background in big data and machine learning techniques. The programme equipped me with these core skills and built me a solid foundation for my career pursuit. Moreover, the flexibility of the curriculum and the diverse background of classmates help students to pursue any career path related to data, such as finance, data analyst, data scientist, etc. For me, the MDASC journey is a rewarding one, and I believe it will be a rewarding journey for anyone who believes the power of data.

Programme structure

Design of curriculum (72 credits)

Compulsory Courses (36 credits)
- COMP7404 Computational intelligence and machine learning (6 credits)
- DASC7001 Statistical inference for data science (6 credits)
- DASC7104 Advanced database systems (6 credits)
- DASC7606 Deep learning (6 credits)
- STAT7102 Advanced statistical modelling (6 credits)
- STAT8003 Time series forecasting (6 credits)

Disciplinary Electives (24 credits)
with at least 12 credits from List A and 12 credits from List B

List A
- COMP7005 Advanced topics in data science (6 credits)
- COMP7305 Cluster and cloud computing (6 credits)
- COMP7409 Machine learning in trading and finance (6 credits)
- COMP7503 Multimedia technologies (6 credits)
- COMP7506 Smart phone apps development (6 credits)
- COMP7507 Visualisation and visual analytics (6 credits)
- COMP7906 Introduction to cyber security (6 credits)
- FITE7410 Financial fraud analytics (6 credits)
- ICOM6044 Data science for business (6 credits)

List B
- COMP7606 Computer-based simulation (6 credits)
- STAT6008 Advanced statistical inference (6 credits)
- STAT6013 Financial data analysis (6 credits)
- STAT6015 Quantitative risk management (6 credits)
- STAT6016 Spatial data analysis (6 credits)
- STAT6019 Current topics in statistics (6 credits)
- STAT7008 Programming for data science (6 credits)
- STAT7017 Data mining techniques (6 credits)
- STAT7019 Marketing analytics (6 credits)
- STAT7036 Statistical methods for network data (3 credits)
- STAT8007 Natural language processing and text analytics (3 credits)

Capstone requirement (12 credits)

DASC7600 Data science project (12 credits)

Remarks:
1. Students who have completed the same courses in their previous studies in HKU, e.g. Master of Statistics or Master of Science in Computer Science, may be permitted to select up to 24 credits of disciplinary electives from either List A or List B above if they are not able to find any untaught options from either of the lists of disciplinary electives.
2. The programme structure will be reviewed from time to time and is subject to change.

Compulsory Courses
COMP7404 Computational intelligence and machine learning
This course will teach a broad set of principles and tools that will provide the mathematical, algorithmic and philosophical framework for tackling problems using Artificial Intelligence (AI) and Machine Learning (ML). AI and ML are highly interdisciplinary fields with impact in different applications, such as, biology, robotics, language, economics, and computer science. AI is the science and engineering of making intelligent machines, especially intelligent computer programs, while ML refers to the changes in systems that perform tasks associated with AI. Ethical issues in advanced AI and how to prevent learning algorithms from acquiring morally undesirable biases will be covered. Topics may include a subset of the following: problem solving by search, heuristic (informed) search, constraint satisfaction, games, knowledge-based agents, supervised learning, unsupervised learning; learning theory, reinforcement learning and adaptive control and ethical challenges of AI and ML.

DASC7101 Statistical inference for data science
Computing power has revolutionised the theory and practice of statistical inference. Reciprocally, novel statistical inference procedures are becoming an integral part of data science. By focusing on the interplay between statistical inference and methodologies for data science, this course reviews the main concepts underpinning classical statistical inference, studies computer-intensive methods for conducting statistical inference, and examines important issues concerning statistical inference drawn upon modern learning technologies. Contents include classical frequentist and Bayesian inferences, computer-intensive methods such as the EM algorithm, the bootstrap and the Markov chain Monte Carlo, large-scale hypothesis testing, high-dimensional modeling, and post-model-selection inference.

DASC7104 Advanced database systems
The course will study some advanced topics and techniques in database systems, with a focus on the aspects of database systems design & algorithms and
WHAT YOU WILL LEARN

big data processing. Traditional topics include: query optimisation, physical database design, transaction management, crash recovery, parallel databases. It will survey the recent developments in selected areas such as NoSQL databases and big data management systems.

DASC7606 Deep learning
Machine learning is a fast growing field in computer science and deep learning is the cutting edge technology that enables machines to learn from large-scale and complex datasets. Ethical implications of deep learning and its applications will be covered and the course will focus on how deep neural networks are applied to solve a wide range of problems in areas such as natural language processing, and image processing. Other applications such as financial predictions, game playing and robotics may also be covered. Topics covered include linear and logistic regression, artificial neural networks and how to train them, recurrent neural networks, convolutional neural networks, generative models, deep reinforcement learning and unsupervised feature learning.

STAT7102 Advanced statistical modelling
This course introduces modern methods for constructing and evaluating statistical models and their implementation using popular computing software, such as R or Python. It will cover both the underlying principles of each modelling approach and the model estimation procedures. Topics from: (i) Linear regression models; (ii) Generalised linear models; (iii) Model selection and regularisation; (iv) Kernel and local polynomial regression; selection of smoothing parameters; (v) Generalised additive models; (vi) Hidden Markov models and Bayesian networks.

STAT8003 Time series forecasting
A time series consists of a set of observations on a random variable taken over time. Such series arise naturally in climatology, economics, finance, environmental research and many other disciplines. In addition to statistical modelling, the course deals with the prediction of future behaviour of these time series. This course distinguishes different types of time series, investigates various representations for them and studies the relative merits of different forecasting procedures.

Disciplinary Electives

COMP7105 Advanced topics in data science
This course will introduce selected advanced computational methods and apply them to problems in data analysis and relevant applications.

COMP7305 Cluster and cloud computing
This course offers an overview of current cloud technologies, and discusses various issues in the design and implementation of cloud systems. Topics include cloud delivery models (SaaS, PaaS, and IaaS) with motivating examples from Google, Amazon, and Microsoft; virtualisation techniques implemented in Xen, KVM, VMWare, and Docker; distributed file systems, such as Hadoop file system; MapReduce and Spark programming models for large-scale data analysis, networking techniques in hyper-scale data centers. The students will learn the use of Amazon EC2 to deploy applications on cloud, and implement a SPARK application on a Xen-enabled PC cluster as part of their project.

COMP7409 Machine learning in trading and finance
The course introduces our students to the field of Machine Learning, and help them develop skills of applying Machine Learning, or more precisely, applying supervised learning, unsupervised learning and reinforcement learning to solve problems in Trading and Finance.

This course will cover the following topics. (1) Overview Machine Learning and Artificial Intelligence, (2) Supervised Learning, Unsupervised Learning and Reinforcement Learning, (3) Major algorithms for Supervised Learning and Unsupervised Learning with applications to Trading and Finance, (4) Basic algorithms for Reinforcement Learning with applications to optimal trading, asset management, and portfolio optimisation, (5) Advanced methods of Reinforcement Learning with applications to high-frequency trading, cryptocurrency trading and peer-to-peer lending.

COMP7503 Multimedia technologies
This course presents fundamental concepts and emerging technologies for multimedia computing. Students are expected to learn how to develop various kinds of media communication, presentation, and manipulation techniques. At the end of course, students should acquire proper skill set to utilise, integrate and synchronise different information and data from media sources for building specific multimedia applications.

Topics include media data acquisition methods and techniques; nature of perceptually encoded information; processing and manipulation of media data; multimedia content organisation and analysis; trending technologies for future multimedia computing.

COMP7506 Smart phone apps development
Smart phones have become very popular in recent years. The number of smart phone users worldwide today surpasses three billion and is forecast to further grow by several hundred million in the next few years. Smart phones play an important role in mobile communication and applications. Smart phones are powerful as they support a wide range of applications (called apps). Most of the time, smart phone users just purchase their favorite apps wirelessly from the vendors. There is a great potential for software developer to reach worldwide users.

This course aims at introducing the design issues of smart phone apps. For example, the smart phone screen is usually much smaller than the computer monitor. We have to pay special attention to this aspect in order to develop attractive and successful apps. Various modern smart phone apps development environments and programming techniques (such as Java for Android phones, and Swift for iPhones) will also be introduced to facilitate students to develop their own apps.

Students should have basic programming knowledge.

COMP7507 Visualisation and visual analytics
This course introduces the basic principles and techniques in visualisation and visual analytics, and their applications. Topics include human visual perception; color; visualisation techniques for spatial, geospatial and multivariate data, graphs and networks; text and document visualisation; scientific visualisation; interaction and visual analysis.

Topics include media data acquisition methods and techniques; nature of perceptually encoded information; processing and manipulation of media data; multimedia content organisation and analysis; trending technologies for future multimedia computing.

COMP7906 Introduction to cyber security
The aim of the course is to introduce different methods of protecting information and data in the cyber world, including the privacy issue. Topics include introduction to security; cyber attacks and threats; cryptographic algorithms and applications; network security and infrastructure.

FITE7410 Financial fraud analytics
This course aims at introducing various analytics techniques to fight against financial fraud. These analytics techniques include: descriptive analytics, predictive analytics, and social network learning. Various data set will also be introduced, including labeled or unlabeled data sets, and social network data set. Students learn the fraud patterns through applying the analytics techniques in financial frauds, such as, insurance fraud, credit card fraud, etc.

Key topics include: Handling of raw data sets for fraud detection; Applications of descriptive analytics, predictive analytics and social network analytics to construct fraud detection models; Financial Fraud Analytics challenges and issues when applied in business context.

ICOM6044 Data science for business
The emerging discipline of data science combines statistical methods with computer science to solve problems in applied areas. In this case we focus on how data science can be used to solve business problems especially those in electronic commerce. By its very nature e-commerce is able to generate large amounts of data and data mining methods are quite helpful for managers in turning this data into knowledge which in turn can be used to make better decisions. These data sets and their accompanying quantitative methods have the potential to dramatically change decision making in many areas of business. For example, ideas like interactive marketing, customer relationship management, and database marketing are pushing companies to utilise the information they collect about their customers in order to make better marketing decisions.

This course focuses on how data science methods can be applied to solve managerial problems in marketing and electronic commerce. Our emphasis is developing a core set of principles that embody data science: empirical reasoning, exploratory and visual analysis, and predictive modeling. We use these core principles to understand many methods used in data mining and machine learning. Our strategy in this course is to survey several popular techniques and understand how they
WHAT YOU WILL LEARN

map into these core principles. These techniques are illustrated with case studies. However, the emphasis is not on the software for implementing these techniques, but on understanding the inputs and outputs of these techniques and how they are used to solve business problems.

**STAT6008 Advanced statistical inference**
This course covers the advanced theory of point estimation, interval estimation and hypothesis testing. Using a mathematically-oriented approach, the course provides a formal treatment of inferential problems, statistical methodologies and their underlying theory. It is suitable in particular for students intending to further their studies or to develop a career in statistical research. Contents include: (1) Decision problem – frequentist approach: loss function; risk; decision rule; admissibility; minimaxity; unbiasedness; Bayes’ rule; (2) Decision problem – Bayesian approach: prior and posterior distributions, Bayesian inference; (3) Estimation theory: exponential families; likelihood; sufficiency; minimal sufficiency; completeness; UMVU estimators; information inequality; large-sample theory of maximum likelihood estimation; (4) Hypothesis testing: uniformly most powerful (UMP) test; monotone likelihood ratio; UMP unbiased test; conditional test; large-sample theory of likelihood ratio; confidence set; (5) Nonparametric inference; bootstrap methods.

**STAT6013 Financial data analysis**
This course aims at introducing statistical methodologies and their underlying theory. Financial applications and statistical techniques for marketing decision making are intertwined in all lectures. Contents include: recent advances in modern portfolio theory, copula, market microstructure, stochastic volatility models and high frequency data analysis.

**STAT6015 Advanced quantitative risk management**
This course covers statistical models and methods of risk management, especially of Value-at-Risk (VaR). Contents include: Value-at-risk (VaR) and Expected Shortfall (ES); univariate models (normal model, log-normal model and stochastic process model) for VaR and ES; models for portfolio VaR; time series models for VaR; extreme value approach to VaR; back-testing and stress testing.

**STAT6016 Spatial data analysis**
This course covers statistical concepts and tools involved in modelling data which are correlated in space. Applications can be found in many fields including epidemiology and public health, environmental sciences and economics. Spatial data analysis topics include: (1) Outline of three types of spatial data: point-level (geostatistical), areal (lattice), and spatial point process. (2) Model-based geostatistics: covariance functions and the variogram; spatial trends and directional effects; intrinsic models; estimation by curve fitting or by maximum likelihood; spatial prediction by least squares, by simple and ordinary kriging, by trans-Gaussian kriging. (3) Areal data models: introduction to Markov random fields; conditional, intrinsic, and simultaneous autoregressive (CAR, IAR, and SAR) models. (4) Hierarchical modelling for univariate spatial response data, including Bayesian kriging and lattice modelling. (5) Introduction to simple spatial point processes and spatio-temporal models. Real data analysis examples will be provided with dedicated R packages such as geoR.

**STAT6019 Current topics in statistics**
This course includes two modules. The first module, Causal Inference, is an introduction to key concepts and methods for causal inference. Contents include 1) the counterfactual outcome, randomised experiment, observational study; 2) Effect modification, mediation and interaction; 3) Causal graphs; 4) Confounding, selection bias, measurement error and random variability; 5) Inverse probability weighting and marginal structural models; 6) Outcome regression and propensity score; 7) The standardisation and the parametric g-formula; 8) G-estimation and the structural nested model; 9) Instrumental variable method; 10) Machine learning methods for causal inference; 11) Other topics as determined by the instructor. The second module, Posterior Inference and Simulation, cover topics from: 1) Large-sample properties of posterior distribution; 2) Langevin dynamics and Hamiltonian MCMC; 3) Sequential Monte Carlo methods; 4) Approximation Bayesian computation; 5) Variational Bayesian methods; 6) Other topics as determined by the instructor.

**STAT7008 Programming for data science**
In the big data era, it is very easy to collect huge amounts of data. Capturing and exploiting the important information contained within such datasets poses a number of statistical challenges. This course aims to provide students with a strong foundation in computing skills necessary to use R or Python to tackle some of these challenges. Possible topics to be covered may include exploratory data analysis and visualisation, collecting data from a variety of sources (e.g. excel, web-scraping, APIs and others), object-oriented programming concepts and scientific computation tools. Students will learn to create their own R packages or Python libraries.

**STAT8017 Data mining techniques**
With the rapid developments in computer and data storage technologies, the fundamental paradigms of classical data analysis are mature for change. Data mining techniques aim at helping people to work smarter by revealing underlying structure and relationships in large amounts of data. This course takes a practical approach to introduce the new generation of data mining techniques and show how to use them to make better decisions. Topics include data preparation, feature selection, association rules, decision trees, bagging, random forests and gradient boosting, cluster analysis, neural networks, introduction to text mining.

**STAT8019 Marketing analytics**
This course aims to introduce various statistical models and methodology used in marketing research. Special emphasis will be put on putting analytics and statistical techniques for marketing decision making including market segmentation, market response models, consumer preference analysis and conjoint analysis. Contents include market response models, statistical methods for segmentation, targeting and positioning, statistical methods for new product design.

**STAT8306 Statistical methods for network data**
The six degree of separation theories that human interactions could be easily represented in the form of a network. Examples of networks include router networks, the World Wide Web, social networks (e.g. Facebook or Twitter), genetic interaction networks and various collaboration networks (e.g. movie actor collaboration network and scientific paper collaboration network). Despite the diversity in the nature of sources, the networks exhibit some common properties. For example, both the spread of disease in a population and the spread of rumors in a social network are in sub-logarithmic time. This course aims at discussing the common properties of real networks and the recent development of statistical network models. Topics may include common network measures, community detection in graphs, preferential attachment random network models, exponential random graph models, models based on random point processes and the hidden network discovery on a set of dependent random variables.

**STAT8307 Natural language processing and text analytics**
The textual data constitutes an enormous proportion of unstructured data which is characterised as one of “V’s in Big Data. The logical and computational reasonings are applied to transform large collection of written resources to structured data for use in further analysis, visualisation, integration with structured data in database or warehouse, and further refinement using machine learning systems. This course introduces the methodology of text analytics. Topics include natural language processing, word representation, text categorisation and clustering, topic modelling and sentiment analysis. Students are required to possess basic understanding of Python language.

**Capstone Requirement**
**DASC7600 Data science project**
students will be required to carry out independent work on a major project under the supervision of individual staff member. A written report is required.
Through merging the strengths of the two departments, we aim to equip students with state-of-the-art computational skills and the frontier of high-performance analytics. Students will be well-prepared for wide-ranging job opportunities in data science when they graduate.

Programme Director and Head of Department of Statistics and Actuarial Science

Professor G S YIN
MA Temple; MSc, PhD N Carolina

Admissions
Requirements
1. A Bachelor’s degree with honours, or an equivalent qualification;
2. Applicants should have taken at least 1 university or postsecondary certificate course in each of the following 3 subjects or related areas
   ◊ calculus and algebra
   ◊ computer programming
   ◊ introductory statistics
3. Fulfill the University Entrance Requirements.

How to apply
Main Round Deadline: 12 noon, December 15, 2021 (GMT+8)
Clearing Round Deadline: 12 noon, January 31, 2022 (GMT+8)

Online application
admissions.hku.hk/tpg/

Further Information
Programme details

Enquiries
Department of Statistics and Actuarial Science
Tel: (852) 3917 4152 Email: mdasc@hku.hk