Society of Actuaries

Asia-Pacific Actuarial Teaching Conference

Data Analytics Education at HKU

Kam C. Yuen

CAE Center of Actuarial Excellence
SOCIETY OF ACTUARIES

Statistics & Actuarial Science 統計及精算學系

July 9, 2018 (HKU)

WELCOME

to Hong Kong

and HKU!

From everywhere in the world!

Agenda

- Actuarial Teaching at HKU
- Sharing of Teaching Data Visualization
- Student Engagement with Data Mining
- Adaptation to the Big Data Era
- Further Development

Asia-Pacific Actuarial Teaching Conference

Statistics & Actuarial Science
統計及精算學系

1

Actuarial Teaching at HKU

with the first actuarial programme in Hong Kong

"

Good teachers know how to bring out the best in students.

– Charles Kuralt

First Actuarial Program in Hong Kong

Admitted talents with
- Quantitative skills
- Mathematical knowledge
- Business acumen
- Excellent academic background

Our Curiculum Structure

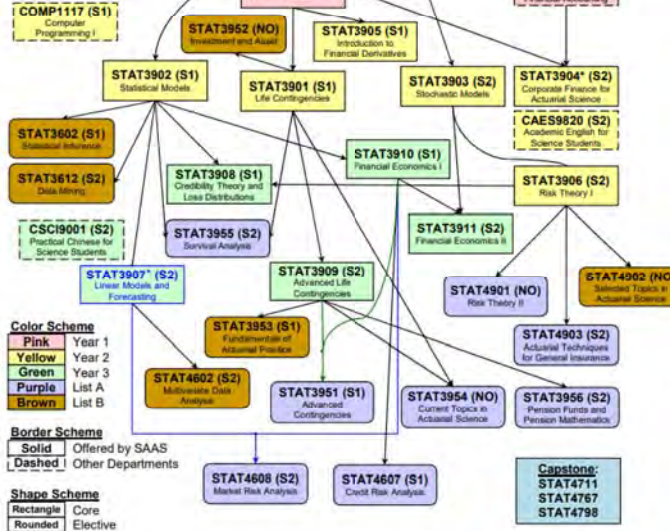Flow Chart of Disciplinary Courses for BSc(ActuarSc)

📌 Our Curiculum with SOA ASA Pathway

INTRODUCTORY I	INTRODUCTORY II	ACTUARIAL	ADVANCED	PROFESSIONALISM
VEE ECONOMICS ***ECON1210, ECON1220**	**EXAM** INVESTMENT AND FINANCIAL MARKETS **STAT3903, STAT3904, STAT3905, STAT3910**	**EXAM** LONG-TERM ACTUARIAL MATHEMATICS **STAT3901, STAT3903, STAT3906, STAT3909, STAT3956**	**e-LEARNING** FUNDAMENTALS OF ACTUARIAL PRACTICE **STAT3953**	**SEMINAR** ASSOCIATESHIP PROFESSIONALISM COURSE
VEE ACCOUNTING AND FINANCE ***ACCT1101, STAT3904**		**EXAM** SHORT-TERM ACTUARIAL MATHEMATICS **STAT3906, STAT3908, STAT4903**		
EXAM FINANCIAL MATHEMATICS **STAT2902, STAT3905**	**VEE** MATHEMATICAL STATISTICS ***STAT3902**	**EXAM** STATISTICS FOR RISK MODELING **STAT3907, STAT4608, STAT4602, STAT3612, STAT4904**	**PROJECT** PREDICTIVE ANALYTICS **STAT4904**	
EXAM PROBABILITY **STAT2901, STAT3902**				

YOU MUST COMPLETE 2 EXAMS BEFORE APPLYING TO VALIDATE YOUR EDUCATIONAL EXPERIENCE (EDUCATIONAL EXPERIENCES MAY HAVE OCCURRED PRIOR TO COMPLETING EXAMS).

*VEE approved courses

more related to data analytics

📌 Learning Diversity

- Tutorial
- Assignment
- Project
- Discussion
- Presentation
- Seminar
- eLearning

2

Sharing of Teaching Data Visualization

inside and outside classroom

"

An investment in knowledge pays the best interest.

– Benjamin Franklin

Course: Data Visualization

STAT3622 Data Visualization

This course will focus on how to work with statistical graphics, graphics that display statistical data, to communicate and analyze data. Students will learn a set of tools such as R to create these graphics and critically evaluate them.

Content: Grammar of graphics, visualizing patterns over time, visualizing relationship, visualizing spatial relationships, visualizing texts.

http://www.statsoft.org/teaching/stat3622/

The Data Science Venn Diagram
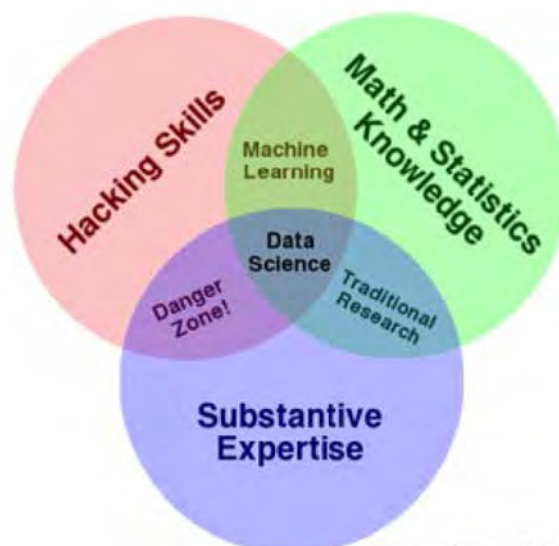
Drew Conway (2010)

Course Webpage

Class Schedule		Lecture Notes	Extra Materials
Lecture 1	Sep 4	**Introduction** (HTML | Rpres)	Link: RStudio and Markdown
Lecture 2	Sep 4, 7	**Data Science and Statistics** (HTML | Rpres)	1. CareerCast Best Jobs 2017
		Sexy DS job, DS Venn Diagram, DS Workflow	2. Hans Rosling's TED talk
Lecture 3	Sep 11, 14	**Exploratory Data Analysis** (HTML | Rpres)	
		EDA, R base plots, R:Lattice package	
Lecture 4	Sep 18, 21	**R Graphics with ggplot2** (HTML | Rpres)	
Lecture 5	Sep 25	**Data Manipulation** (HTML | Rpres)	Quiz 1 (PDF | Code | Dataset)
		R:dplyr package	
Lecture 6	Sep 28	**Animated Graphics in R** (HTML | Rpres)	ImageMagick
		R:magick package	
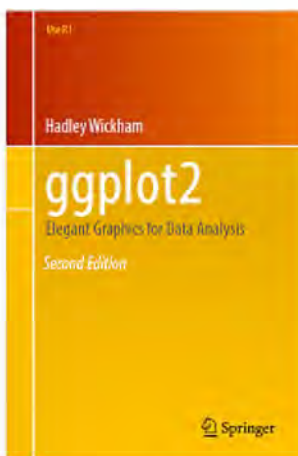
R Graphics

Springer 2016 (2nd edition)

- 40 years after Tukey (1977)

- R: base -> lattice -> ggplot2

- "ggplot2, started in 2005, is an attempt to take the good things about base and lattice graphics and improve on them with a strong underlying model" (Hadley Wickham)

- One of most commonly downloaded R packages, refreshed version 2.x.x since Dec.2015

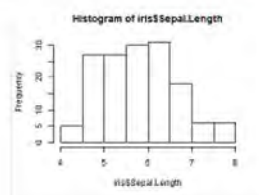- Based on the Grammar of Graphics by Wilkinson (2005; Springer 2ed)

R Graphics

Base, Lattice and ggplot2 styles (first impression)

```
hist(iris$Sepal.Length) # Base graphics
```

Histogram of iris$Sepal.Length

```
library(ggplot2)
ggplot(data=iris,
       aes(x=Sepal.Length,
           fill=Species)) +
  geom_histogram(bins=8)
```
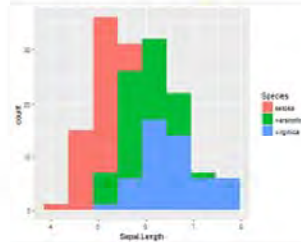


```
library(lattice)
histogram(data=iris, ~Sepal.Length|Species,
          layout=c(3,1))
```


Case Study 1: Plots with Maps

World Map in R

```
library(maps)
map('world', fill=TRUE, border='darkgray', col='gray', bg='white')
```

World Map by ggplot2

```
library(ggplot2)
thismap = map_data('world')
ggplot(thismap, aes(long, lat, group=group)) +
  geom_polygon(fill='yellow', colour="green")
```
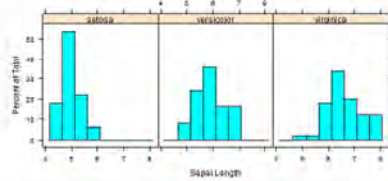
Hong Kong SAR Map

```
ggplot(hkmap, aes(long, lat, group=group, fill=id)) +
  geom_polygon(show.legend = F)
```

Case Study 1: Plots with Maps

Hong Kong Regions and Districts

```
wikidata = read.csv("HK18Districts.csv")
wikidata
```

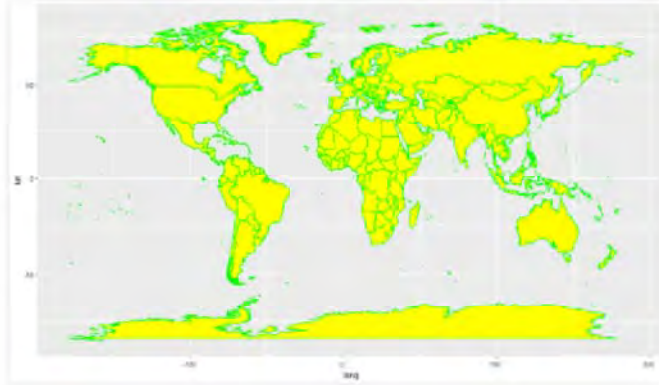|	District	Population	Area	Density	Region	Code
1	Central and Western	244600	12.44	19983.92	HK	CW
2	Eastern	574500	18.56	31217.67	HK	EA
3	Southern	269200	38.85	6962.68	HK	SO
4	Wan Chai	150900	9.83	15300.10	HK	WC
5	Sham Shui Po	390600	9.35	41529.41	KL	SS
6	Kowloon City	405400	10.02	40194.70	KL	KC
7	Kwun Tong	641100	11.27	56779.05	KL	KU
8	Wong Tai Sin	426200	9.30	45645.16	KL	WT
9	Yau Tsim Mong	318100	6.99	44864.09	KL	YT
10	Islands	146900	175.12	825.14	NT	IS
11	Kwai Tsing	507100	23.34	21503.86	NT	KI
12	North	310800	136.61	2220.19	NT	NO
13	Sai Kung	448600	129.65	3460.08	NT	SK
14	Sha Tin	648200	68.71	9433.85	NT	ST
15	Tai Po	307100	136.15	2220.35	NT	TP
16	Tsuen Wan	303600	61.71	4887.38	NT	TW
17	Tuen Mun	495900	82.89	5889.38	NT	TM
18	Yuen Long	607200	138.46	4297.99	NT	YL

- HK population by districts: https://en.wikipedia.org/wiki/Districts_of_Hong_Kong

http://www.statsoft.org/wp-content/uploads/2017/09/Lecture9_Maps.html

Case Study 2: HK House Prices

http://www.statsoft.org/wp-content/uploads/2017/09/Lecture10_HousePrice.html

2 Temporal and Spatial Variations

```
n = dim(DataX)[1]
m = dim(DataX)[2]-1

matplot(DataX[, -1], type="l", lty=1, col=rainbow(m),
        xaxt="n", ylab="", xlim=c(1,n+10))
axis(1,at=1:n,labels=format(DataX$Date, "%Y%m"))
text(n, tail(DataX[,-1], 1), names(DataX)[-1], p=4, cex=0.6)
title(main="Home Prices for Selected Estates in Hong Kong", xlab="Ti
"Unit Price (net area)")
```


Home Prices for Selected Estates in Hong Kong
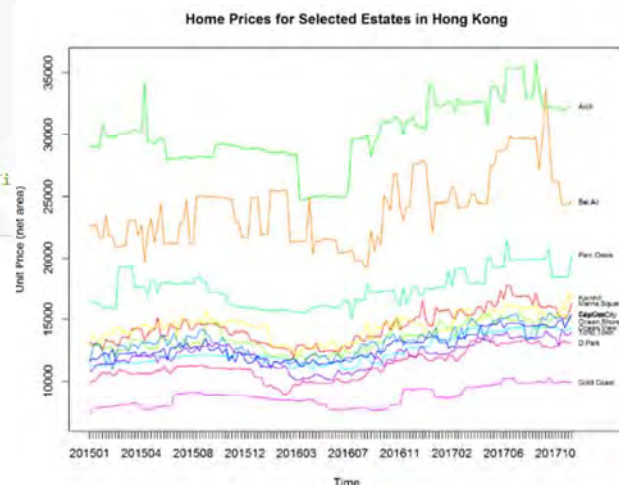
Case Study 3: Global Top Brands

http://www.statsoft.org/wp-content/uploads/2017/09/Lecture11_TopBrand.html
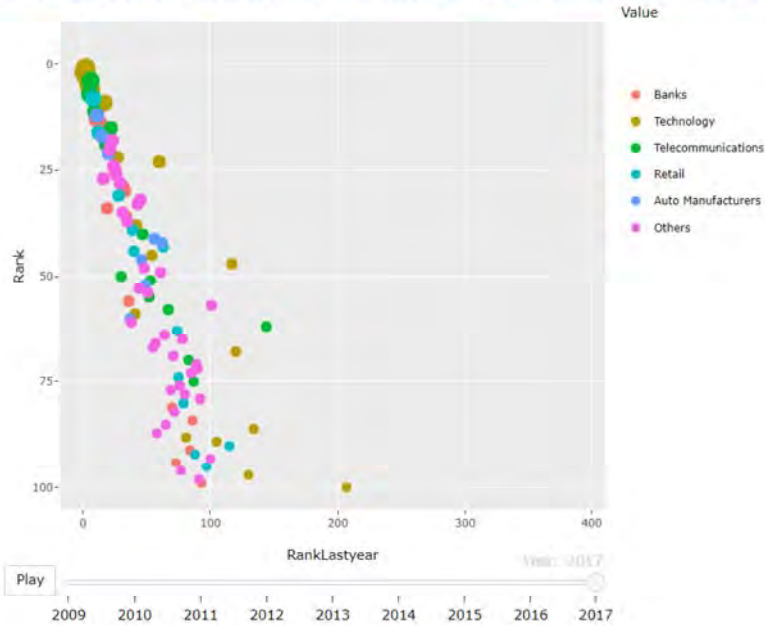
Case Study 4: AIEQ Case Study

http://www.statsoft.org/wp-content/uploads/2017/09/Lecture12_AIEQ.html

3

Student Engagement with Data Mining

to get the best
learning experience

"

Experience is a hard teacher because she gives the test first, the lesson afterward.

– Vernon Law

STAT3612 Data Mining

With an explosion in information technology in the past decade, vast amounts of data appear in a variety of fields such as finance, customer relations management and medicine. The challenge of understanding these data with the aim of creating new knowledge and finding new relationships among data attributes has led to the innovative usage of statistical methodologies and development of new ones. In this process, a new area called data mining is spawned. This course provides a comprehensive and practical coverage of essential data mining concepts and statistical models for data mining.

Content: Data pre-processing, classification and regression trees, credit scoring, kNN classifier, cluster analysis and neural networks.

http://www.statsoft.org/teaching/stat3612/

📌 Course Webpage

STAT3612 | Statistics Me×		www.statsoft.org/teaching/stat3612/	
Lecture 3	Jan 30	**Machine Learning** (HTML | Rmd)	
		Stat3612 Landscape	
Lecture 4	Feb 2	**Linear Regression** (HTML | Rmd)	Tutorial 3 (HTML | Rmd)
		LM, LSE, Model Inference	
		Diagnostics, Variable Selection	
Lecture 5	Feb 6-9	**Basis Expansion** (HTML | Rmd)	Tutorial 4 (HTML | Rmd)
		Nonparametric regression, Feature representation,	
		Piecewise linear fitting, Regression Splines, GAM	
Lecture 6	Feb 13-23	**Regularization** (HTML | Rmd)	Tutorial 5 (HTML | Rmd)
		Smoothing spline, Ridge Regression,	
		Lasso, Elastic Net, Sparse Modeling	
Lecture 7	March 2-13	**Classification** (HTML | Rmd)	Tutorial 6 (HTML | Rmd)
		Logistic/Softmax regression	Assignment 1 (Due: March 15)
		LDA/QDA, kNN, Confusion matrix, ROC curve	Dataset: BitCoin.csv
Lecture 8	March 20-23	**Tree-based Methods** (HTML | Rmd)	Tutorial 7 (HTML | Rmd)
		Classification and Regression Trees	**Test 1 on April 3**
		Bagging, Random Forest, GBM	
Lecture 9	March 27	**Support Vector Machines** (HTML | Rmd)	Tutorial 8 (HTML | Rmd)
		Separating Hyperplane, Maximal Margin, Kernel trick	Assignment 2 (Due: April 11)
		LIBSVM, Hyperparameter optimization	**Group Project Announcement**

Students have the chance to be exposed to various data analytics techniques
- Regression
- Machine Learning
- Classification
- Support Vector Machine
- Principal Component Analysis

With the use of R

STAT3612 Data Mining

www.statsoft.org/wp-content/uploads/2018/stat3612/Lecture11_Unsupervised?#8

Hierarchical clustering

```
DataX = iris[,-5]
hc = hclust(dist(DataX), method="complete")
plot(hc)
```

Cluster Dendrogram

dist(DataX)
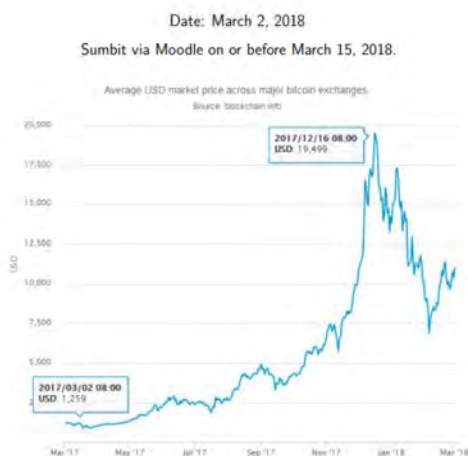hclust (*, "complete")

StatSoft.org

8 / 18

A good learning practice is to work on some practical problems…

Assignments are designed to let students
- Familiarize with data handling
- Practise the process of analyzing data
- Build models to fit the data
- Experience the practical analysis for real life problem

STAT3612 Assignment 1: Bitcoin Prices

Date: March 2, 2018

Sumbit via Moodle on or before March 15, 2018.

Average USD market price across major bitcoin exchanges.
Source: blockchain.info

The bitcoin prices have been increasing in the past year but also encountered dramatic drops from the peak price at 16 December last year. In this assignment, you are directed to perform the trend analysis of daily bitcoin prices based on STAT3612 modeling methods.

Step 1. (20%) Download the historical daily bitcoin prices for the last year from

STAT3612 Assignment 2: Modeling BitCoin Price Trend

Date: March 29, 2018

Sumbit via Moodle on or before April 11, 2018.

In Assignment 1, you have tried the piecewise linear modeling of the bitcoin prices with sequentially adding knots. In this assignment, try the knot selection by the lasso technique.

Step 1. (25%) Download the historical daily bitcoin prices as in Assignment 1 (click here). For the 365 days, set the equally spaced knots with every 10 days away (i.e. $10, 20, 30, \ldots$). For each knot, construct the basis function of the following form:

$$\phi_j(t) = \begin{cases} 0 & \text{if } t \leq \tau_j \\ x - \tau_j & \text{if } \tau_j < t \leq \tau_{j+1} \\ \tau_{j+1} - \tau_j & \text{if } t > \tau_{j+1} \end{cases}$$

for each knot τ_j. Then, plot the first 5 bases and the last 5 bases on the same chart.

Step 2. (25%) Perform lasso with piecewise linear bases and BitCoin prices, by using R:glmnet for varying regularization parameters λ. Plot the solution paths: a). along $\|\beta\|_1$, and b). along $\log(\lambda)$.

Step 3. (25%) Based on the results in Step 2, choose an appropriate model with half number

A group project work is a significant component in this course

It aims at

▷ Integrating the analytics techniques taught in the course

▷ Promoting teamwork and presentation skills

STAT3612 Group Project: Data Science Challenge in Predicting Credit Card Delinquency

Release date: March 13, 2017

Problem description:

This project is to build a statistical or machine learning model for predicting the 60-day delinquency in a credit card portfolio. The training dataset consists of 6655 cards owned by 2000 customers together with the following attributes:

- **Age**: card duration since the origination
- **FICO**: credit score of the customer
- **Line**: credit line of limit of card
- **Balance**: outstanding balance of card
- **Payment**: minimum required payment of card

These varaibles are observed when the cards are all "current" (i.e. heathy without delinquency). The target response of interest is their follow-up performance in the next 12 months as measured by **FlagD60**, an indicator for 60 days past due (i.e. payment delinquency).

The model performance will be assessed based on a test data that consists of 1000 additional customers, in total 3306 cards. We use the single-valued ranking criteria:

$$F = \frac{2 \cdot P \cdot R}{P + R}, \quad \text{with} \quad P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}$$

based on TP (true positive), FP (false positive) and FN (false negative) rates.

This is a competition among students, to arouse their interest with active participation.

Each group can make one and only one submission of their prediction every 3 days.

They are required to give an in-class presentation with peer assessment.

The final report is submitted with detailed methodology and results.

Group Project Leader Bo... ×

www.statsoft.org/wp-content/uploads/2017/01/LeaderBoard.html

Group Project Leader Board

STAT3612 Data Mining
April 19, 2017 (Last Update)

- 1 Submission Timeline
- 2 Leader Board
- 3 Group Details

1 Submission Timeline

Competition period: March 20 – April 18 (One and only one submission every 3 days)

Ten-submission timeline: **March** 20-22, 23-25, 26-28, 29-31, **April** 1-3, 4-6, 7-9, 10-12, 13-15, 16-18

2 Leader Board

Group	X1st	X2nd	X3rd	X4th	X5th	X6th	X7th	X8th	X9th	X10th	MaxF1	Rank
Flying Learners	0.1600	0.2282	0.2733	0.2772	0.2800	0.2009	0.3237	0.2929	0.3695	0.4089	0.4089	1
Team6	NA	NA	NA	0.2436	NA	0.1562	0.2909	0.2776	0.0665	0.3919	0.3919	2
Team13	NA	NA	NA	0.0146	0.3263	0.2088	0.2133	0.3505	NA	0.2200	0.3505	3
Team2	0.2468	0.2282	0.2898	NA	0.1373	0.2959	0.2868	0.1546	0.3283	0.3434	0.3434	4
AlphaGuess	0.1146	0.1132	0.2211	0.2051	0.2051	0.2447	0.2938	0.2933	0.3049	0.3342	0.3342	5
Team15	0.1604	NA	0.1579	0.1474	0.2162	NA	0.1983	0.2300	0.3252	0.3048	0.3252	6
Team10	0.1977	0.2183	0.2009	0.2986	0.3081	0.0812	0.2896	0.2931	0.3160	NA	0.3160	7
1min=59s	NA	0.0662	0.1319	0.2522	0.2523	0.3152	0.2991	0.2948	0.2984	0.3140	0.3152	8

SOA Case Study Challenge

2016 Student Case Study Challenge

The Society of Actuaries (SOA) is dedicated to advancing education and research for the actuarial profession. As part of its mission, the SOA recognizes the importance of partnering with colleges and universities to help develop future actuaries. With this in mind, the SOA launched its first-ever, Student Case Study Challenge, which provides an opportunity for teams of students to apply their actuarial skills on a real world problem. Read the details of the case study [PDF], which focuses on quantifying healthcare industry adverse disruptors.

SOA Announces Recipients of Case Study Challenge

The Society of Actuaries (SOA) announced the recipients of the Student Case Study Challenge:

- First place, the University of Waterloo with a $5,000 grant. (Winning Submission [PDF])
- Second Place, the University of Hong Kong with a $3,000 grant. (Winning Submission [PDF])
- Third place, two recipients: The University of Iowa (Winning Submission [PDF]) and Georgia State University (Winning Submission [PDF]) with $2,000 grants each.

SOA Case Study Challenge

2017 Student Case Study Challenge

The Society of Actuaries (SOA) is dedicated to advancing education and research for the actuarial profession. As part of its mission, the SOA recognizes the importance of partnering with colleges and universities to help develop future actuaries. With this in mind, the SOA is holding a Student Case Study Challenge, which provides an opportunity for teams of students to apply their actuarial skills on a real world problem. Read the details of the case study [PDF], which focuses on the coastal development of an island. Additional data is in the Akua Island Student Data Spreadsheet [XLS].

SOA Announces Recipients of Case Study Challenge

The Society of Actuaries (SOA) announces the recipients of the Student Case Study Challenge:

- First place, the Chinese University of Hong Kong with a $5,000 grant. (Winning Submission [PDF])
- Second Place, Ball State University with a $3,000 grant. (Winning Submission [PDF])
- Third place, the University of North Carolina at Chapel Hill with a $2,000 grant. (Winning Submission [PDF])

The finalists included a team from the University of Iowa (Winning Submission [PDF]), another team from Ball State University (Winning Submission [PDF]), and a team from Macquarie University (Winning Submission [PDF]).

Semi-finalists, in alphabetical order, are:

- Australian National University
- Boston University
- California State University - Fullerton
- Columbia University
- Georgia State University
- Indiana University Northwest
- University of Calgary
- University of Central Florida
- University of Hong Kong
- Université du Québec à Montréal

4

Adaptation to the Big Data Era

for future actuarial direction

"

Give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime.

— Maimonides

- Master of Data Science

- Bachelor of Science (Decision Analytics)

- New Bachelor Programme:
  - ▷ BASc in Applied Artificial Intelligence

Asia-Pacific Actuarial Teaching Conference

5

**Further
Development**

*Statistics &
Actuarial
Science
統計及精算學系*

> *Education is the key to success in life, and teachers make a lasting impact in the lives of their students.*
>
> *— Solomon Ortiz*

New Actuarial Science Course

New SOA exams
- SRM: Statistics for Risk Modeling
- PA: Predictive Analytics

New course:
- STAT4904 Statistical Learning for Risk Modelling
- To make sense of the vast and complex data sets that have emerged in insurance and finance, it is essential to have a firm understanding of the basic statistical modelling and prediction techniques. This course introduces some useful predictive analytics techniques, such as principal component analysis, naïve Bayes classification, decision tree models, and cluster analysis. The R programming language will be used for actual implementation

THANKS!