# MULTIVARIATE STATISTICAL INFERENCE

*Mengqian LU*

# WHO AM I?

data mining, pattern recognition, prediction and forecasts

**Data Science**

**Statistics**

severe weather (floods, tornado etc.), atmospheric dynamics

**Earth Science**

**Atmospheric Science**

**Engineering**

**Water + Climate**

hydroclimatology, natural resources management, policy
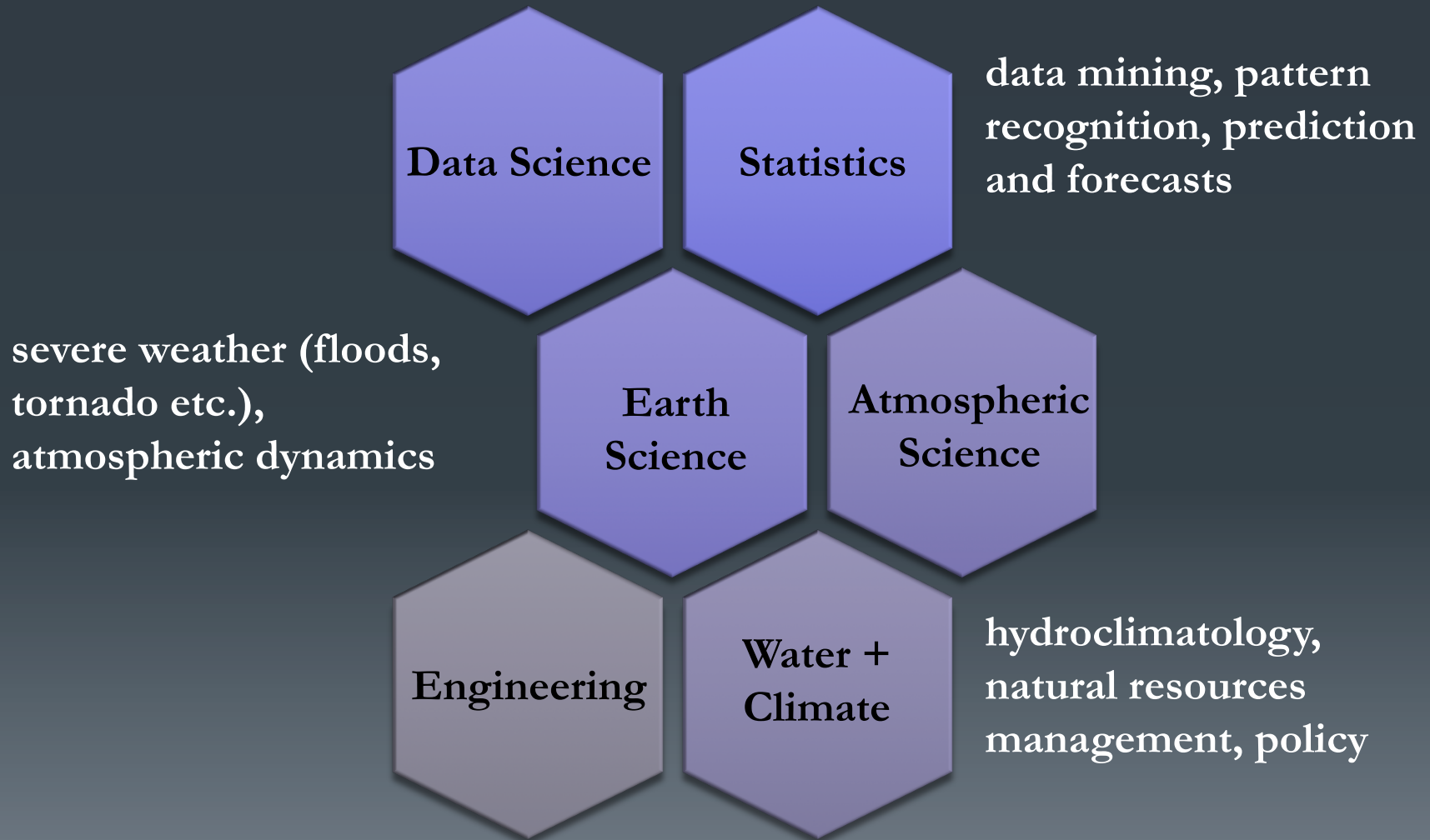
# WHO ARE YOU?

*What do you already know?*
*What is your expectation?*

[Survey](http://goo.gl/forms/0qy64czEQ9)

http://goo.gl/forms/0qy64czEQ9

# SOME ADMINISTRATIVE BITS

*COURSEWORKS@COLUMBIA*
*Piazza@COURSEWORKS*
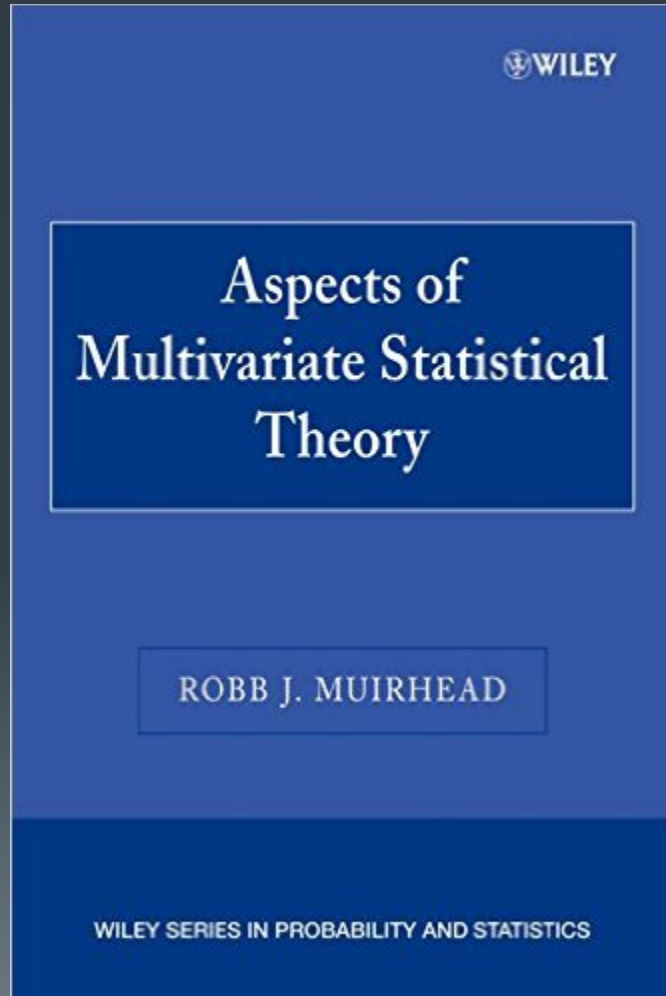
*GITHUB (https://github.com/***MRandomMax)**

*Office Hours: Friday 10AM – 11AM*

*TA: Haolei WENG*
*TA Office Hours: TBA*

# BOOKS

**AMST**

**IAMA**



Aspects of Multivariate Statistical Theory
ROBB J. MUIRHEAD
WILEY
WILEY SERIES IN PROBABILITY AND STATISTICS



Use R!
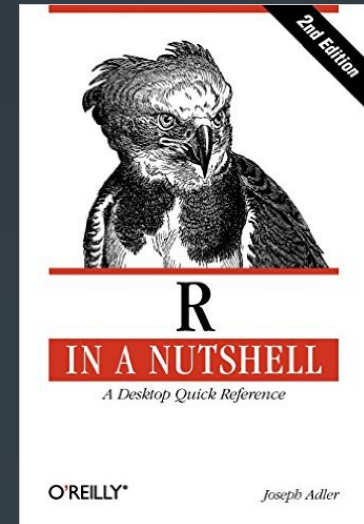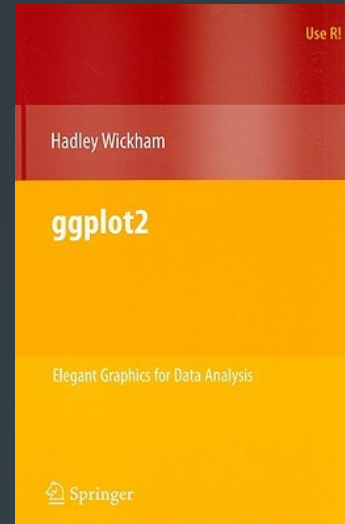Brian Everitt
Torsten Hothorn
An Introduction to Applied Multivariate Analysis with R
Springer
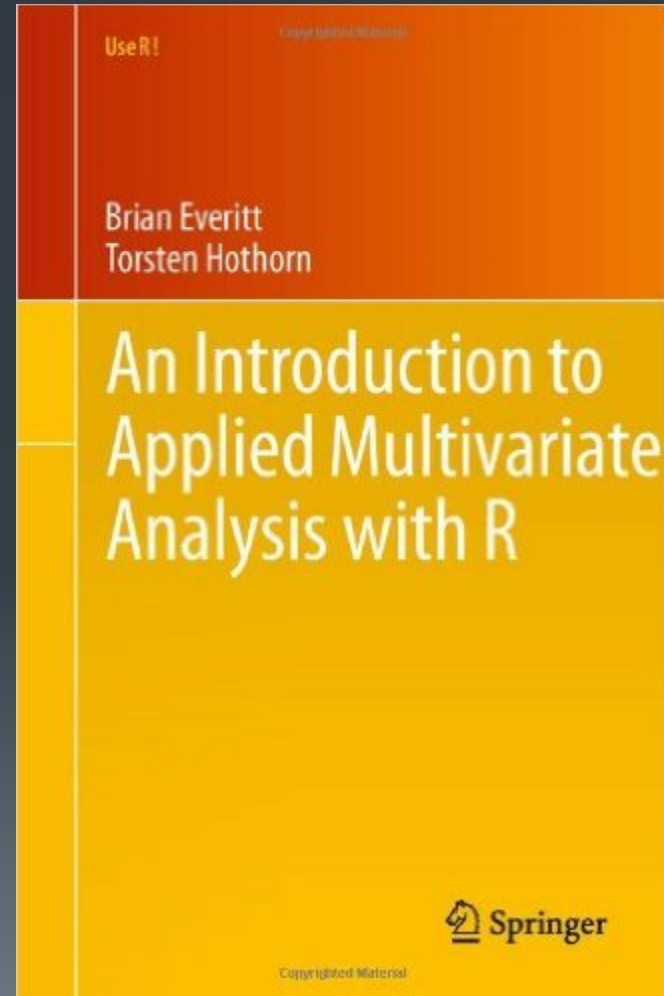
# MORE TEXTS & WEBSITES ON R

Quick-R: http://www.statmethods.net/
R-bloggers: http://www.r-bloggers.com/
Github: http://github.com/
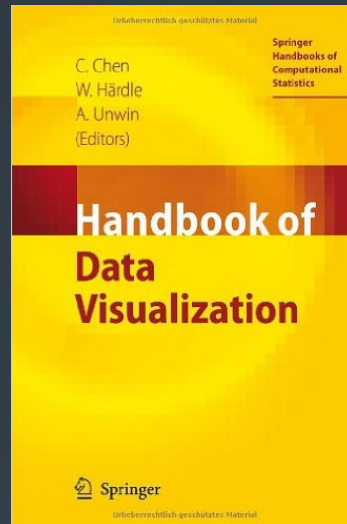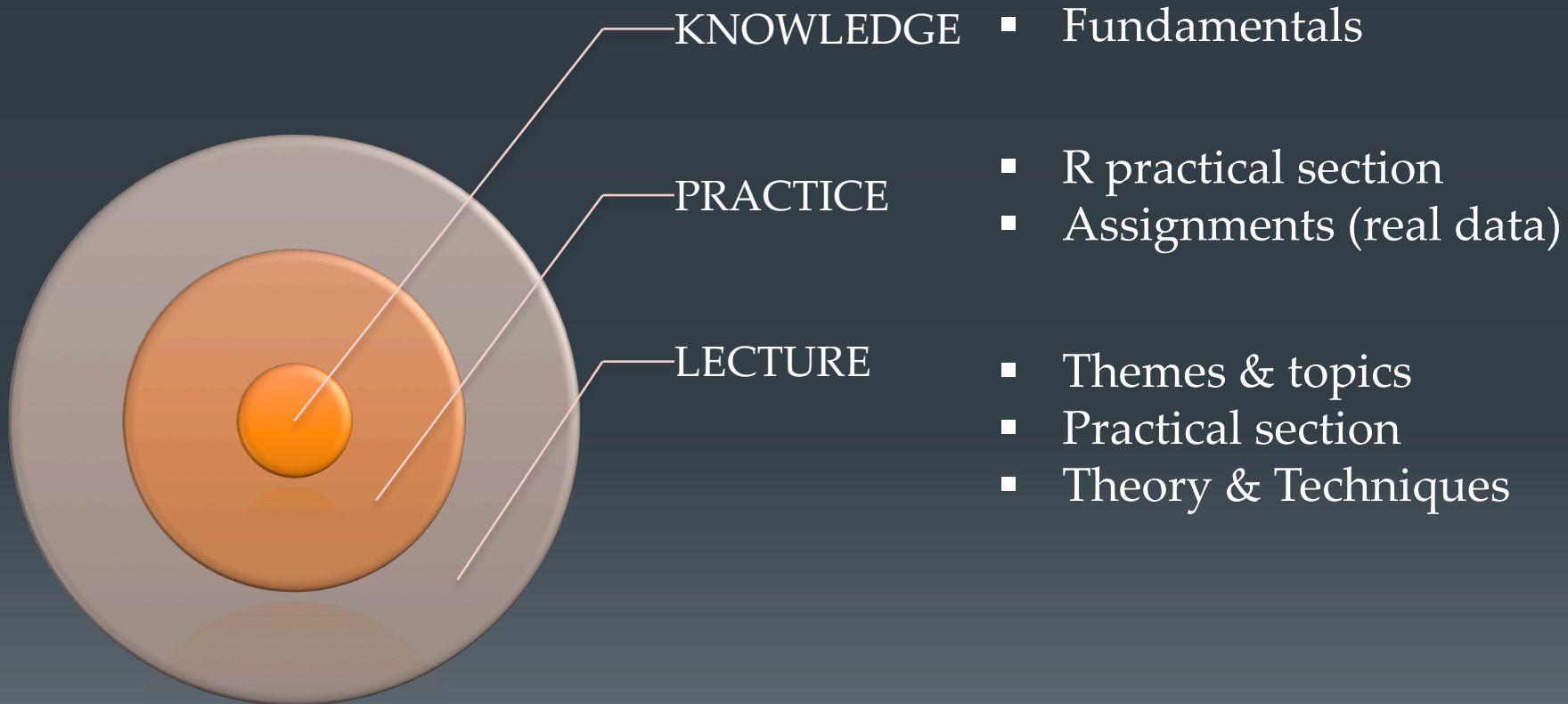
# WORD CLOUD

Time-lagged spatial structure

Hierarchical clustering

Patterns recognition

Signal detection

Cluster analysis

Partitional clustering

Spatiotemporal

Exploratory factor analysis

Principal component analysis

MSI

Canonical correlation analysis

Dimension reduction

Multidimensional scaling

Correlation networks

Confirmatory factor analysis

Independent Component Analysis

# THE PLAN

KNOWLEDGE

- Fundamentals

PRACTICE

- R practical section
- Assignments (real data)

LECTURE

- Themes & topics
- Practical section
- Theory & Techniques

# EVERY STEP COUNTS...

❑ Practice interactively and collaboratively

❑ No matter what is your level of coding, analysis or modeling, DO participate in all the coding – the fast way to learn

❑ *The earlier you fail, the faster you grow*

ASK FOR HELP

# THE EVALUATION

| Course Evaluation: | |
|---|---|
| Participation: | 10% |
| Assignment: | 30% |
| Midterm: | 25% |
| Final Exam: | 35% |

# Why Multivariate Approach?

*The World is Multivariate*

❑ Data Scientists or Researchers now are dealing with many variables of interest, and it's getting more and more complex

❑ The space defined by all/most/some variables matter

# THE ESSENCE…

❑ To recognize the inherent structure of "the space" through application and interpretation of a variety of statistical methods

# Why Multivariate Approach?

❑ Observations: Univariate ➔ "The Space"

❑ Multiple response outcomes

  ▪ *The target: Univariate ➔ A Space*

*But, sometimes no outcome variable(s)*

# THE PURPOSES…

### Detect & Explore

EDAV

- Determine structure
- Extract information e.g. The Survey
- Correlational

### Explain & Predict

- Causality
- Target outcomes e.g. your weather prediction

# THE PURPOSES…

## Prediction & Explanation

- The goal in most research is to predict
- Then what are the best predictors?
    - e.g. Extract from "the Space" by Principal component analysis
- However, determining variable importance can be a suspect endeavor
    - Deemed statistically significant may not have a physical meaning, nor be reproducible
    - Also has to do with the sample

# THE PURPOSES…

**Detect & Explore**

- Another goal is to find the underlying structure, latent variable
  - e.g. Observed behaviors like Giddiness, Silliness, Irrationality, Possessiveness and Misunderstanding reduced to the underlying construct of 'Love'

- Typical approaches involve dimension reduction (PCA, MDS), classification (cluster analysis) and reducing variables (factor analysis)

# Initial examination of data is important

Checklist:
1. Types of variables: nominal/categorical, ordinal, continuous (interval or ratio)
2. Missing values (complete-case analysis or available-case analysis) or outliers (transformation, log or √)

**@IAMA1.3**
**Read IAMA Ch1**

# Initial examination of data is important

Checklist:
1. Types of variables: nominal/categorical, ordinal, continuous (interval or ratio)
2. Missing values (complete-case analysis or available-case analysis) or outliers (transformation, log or √)
3. <u>Sample vs. Population</u>
   - Generalize to real world from sample – the purpose of inferential analyses
   - Avoid sample-specific result!

# Multivariate Distributions

❑ Describe the underlying structure of a vector of random variables

❑ Derive marginal properties of the individual variable

❑ Describe relationships between variables

❑ Inference based on a sample

# A QUICK REVIEW

*Let $x_{ij}$ be $j^{th}$ variable ($j=1,\ldots,p$) on the $i^{th}$ observation*

*X is the n x p matrix:*

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{p} \end{pmatrix}$$

## Univariate Statistics:

*Sample mean:* $\quad \bar{x}_j = \dfrac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad j = 1,\ldots,p$

*Sample variance:* $\quad s_j^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2 \qquad j = 1,\ldots,p$

# Bivariate Statistics:

*Sample covariance:*

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \left( x_{ij} - \overline{x}_j \right) \left( x_{ik} - \overline{x}_k \right) \quad j = 1, \ldots, p; k = 1, \ldots, p$$

$$s_j^2 = s_{jj}$$

*Sample correlation coefficient:*

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}} \sqrt{s_{kk}}} \quad j = 1, \ldots, p; k = 1, \ldots, p$$

## Properties of the Correlation Coefficient

*… If $|r_{jk}| = 1$, there is constants (a,b) that $x_{ij} = a + b x_{ik}$*

*… the value of $r_{jk}$ does not change w/ linear transformation of variable*

# AFTER CLASS

1. Complete the [survey](survey)

2. Install [R](R) and [Rstudio](Rstudio)

3. Read AMST Ch1 and IAMA Ch1, Syllabus