モード２科学としてのデータサイエンスと YOKOHAMA D-STEP

坂巻 顕太郎・汪 金芳

横浜市立大学

第 17 回 統計教育の方法論ワークショップ

Road Map

1. Data Science as Mode 2 Science
   - Science Revisited
   - From Data Analysis to Data Science
   - Teaching Data Science
     - Teaching Mode 2 Data Science at YCU
     - Data Science Club: Learning by Practicing
     - Data Science PBL for Juniors
   - Practical Data Science in Master's Program

2. YOKOHAMA D-STEP

Science: Oxford Dictionary

Definition (Science)

The intellectual and practical activity encompassing the systematic study of the structure and behaviour of the physical and natural world through observation and experiment.

1. a particular area of science.
2. a systematically organized body of knowledge on a particular subject.
3. archaic knowledge of any kind.

Science: Wikipedia

https://en.wikipedia.org/wiki/Science

Definition (Science)

Science (from the Latin word scientia, meaning "knowledge") is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.

Modern science is typically divided into three major branches.

1. natural sciences (e.g., biology, chemistry, and physics);
2. social sciences (e.g., economics, psychology, and sociology)
3. formal sciences (e.g., logic, mathematics, and theoretical computer science).

Disciplines that use existing scientific knowledge for practical purposes, such as engineering and medicine, are described as applied sciences.

Mode 1 Science

Gibbons, M. *et al.* (1994, pp.3-8) *The New Production of Knowledge: The Dynami*
and Research in Contemporary Societies. Sage Publications.

Definition (Mode 1 Science)

The essential ingredients of Mode 1 Science are

1. academic context;
2. disciplinary;
3. homogeneity;
4. autonomy;
5. traditional quality control(peer review),

Science: Tukey (1962)

John W Tukey (1962). The future of data analysis. *AMS*, 1-67.

4. Sciences, mathematics, and the arts. The extreme cases of science and art are clearly distinguished, but, as the case of the student who was eligible for Phi Beta Kappa because mathematics was humanistic and for Sigma Xi because it was scientific shows, the place of mathematics is often far from clear. There should be little surprise that many find the places of statistics and data analysis still less clear.

There are diverse views as to what makes a science, but three constituents will be judged essential by most, viz:

(a1) intellectual content,

(a2) organization into an understandable form,

(a3) reliance upon the test of experience as the ultimate standard of validity.

“Data Analysis” as Science

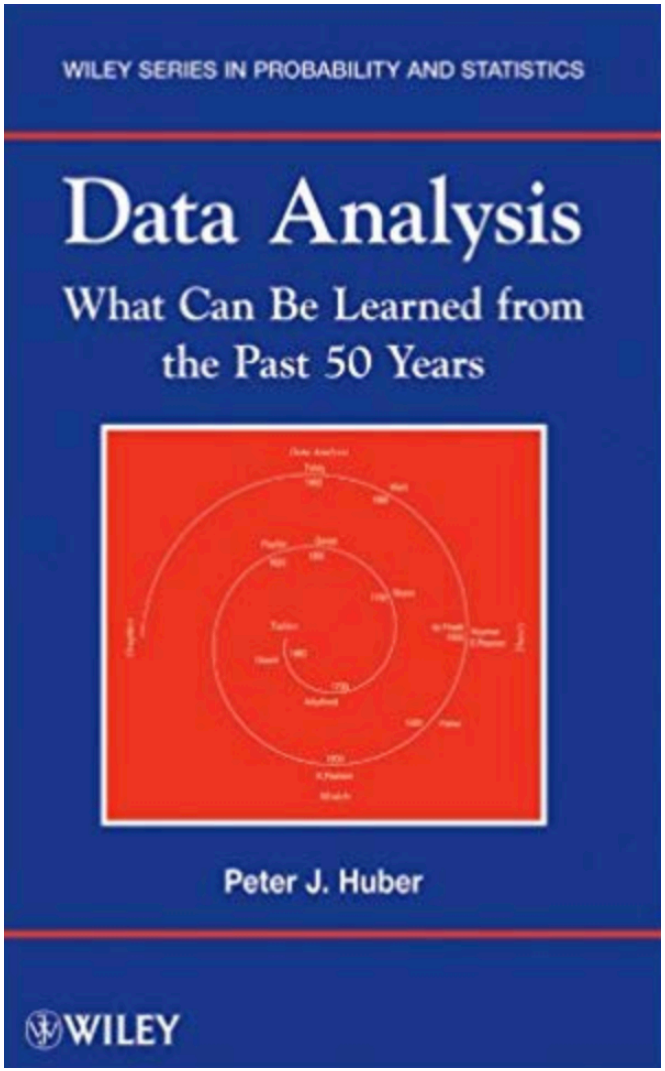
John W Tukey (1962). The future of data analysis. *AMS*, 1-67.

I. GENERAL CONSIDERATIONS

1. Introduction. For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. And when I have pondered about why such techniques as the spectrum analysis of time series have proved so useful, it has become clear that their “dealing with fluctuations” aspects are, in many circumstances, of lesser importance than the aspects that would already have been required to deal effectively with the simpler case of very extensive data, where fluctuations would no longer be a problem. All in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

Huber (2011) on Tukey (1962)

Peter J. Huber (2011). Data Analysis: What Can Be Learned From the Past 50 Yea



Huber on Tukey

Peter J. Huber (2011). Data Analysis: What Can Be Learned From the Past 50 Yea

Half a century ago, Tukey, in an ultimately enormously influential paper redefined our subject... [The paper] introduced the term "data analysis" as a name for what applied statisticians do, differentiating this term from formal statistical inference. But actually, as Tukey admitted, he "stretched the term beyond its philology" to such an extent that it comprised all of statistics.

Driving Forces on Tukey's "Data Analysis"

David Donoho (2015,
https://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf)

Tukey identified four driving forces in the new science:

1. The formal theories of statistics;
2. Accelerating developments in computers and display devices;
3. The challenge, in many fields, of more and ever larger bodies of data;
4. The emphasis on quantification in an ever wider variety of disciplines.

Science and Data Science

David M. Bleia and Padhraic Smythd (2917). PNAS, **114**(33), 8689–8692.

「統計的」、「計算的」「　人間的」の視点から
データサイエンスを俯瞰することが出来よう。
それぞれの視点がデータサイエンスを構成する不可欠な成分であるが、
これらの成分の有機的結合こそがデータサイエンスのエッセンスである。

PNAS PNAS PNAS

Science and data science

David M. Blei[a,b,c,1] **and Padhraic Smyth**[d,e]

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved June 16, 2017 (received for review March 15, 2017)

Data science has attracted a lot of attention, promising to turn vast amounts of data into useful predictions and insights. In this article, we ask why scientists should care about data science. To answer, we discuss data science from three perspectives: statistical, computational, and human. Although each of the three is a critical component of data science, we argue that the effective combination of all three components is the essence of what data science is about.

data science | statistics | machine learning

The term "data science" has attracted a lot of attention. Much of this attention is in business (1), in government (2), and in the academic areas of statistics (3, 4) and computer science (5, 6). Here, we discuss data science from the perspective of scientific research. What is data science? Why might scientists care about it?

Our perspective is that data science is the child of statistics and computer science. While it has inherited some of their methods and thinking, it also seeks to blend them, refocus them, and develop them to help them more effectively navigate and understand the contours of society, finding relevant sources to their work and identifying hard to spot patterns of language that suggest new interpretations and theories. Third, modern telescopes create digital sky surveys that have transformed observational astronomy, generating hundreds of terabytes of raw image data about billions of sky objects. A catalog of these objects, if available, would give astronomers an unprecedented window into the structure of the cosmos.

Causal Inference through Data Science (1/2)

July 12, 2

Data science is science's second chance to get causal inference right.

A classification of data science tasks

Miguel A. Hernán,[1,2] John Hsu[3,4], Brian Healy[5,6]

1. Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
2. Harvard-MIT Division of Health Sciences and Technology, Boston, MA
3. Mongan Institute, Massachusetts General Hospital, Boston, MA
4. Department of Health Care Policy, Harvard Medical School, Boston, MA
5. Department of Neurology, Harvard Medical School, Partners MS Center, Brigham and Women's Hospital, Boston, MA
6. Biostatistics Center, Massachusetts General Hospital, Boston, MA

Correspondence: Miguel Hernán, Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115; email: miguel_hernan@post.harvard.edu

Causal Inference through Data Science（2/2）

「データサイエンス」という用語の導入は、
データ解析の再定義を行う歴史的機運をもたらし、
観察データに基づく因果推論を推し進める自然な枠組みを与えよう。

Abstract

Causal inference from observational data is the goal of many data analyses in the health and social sciences. However, academic statistics has often frowned upon data analyses with a causal objective. The introduction of the term "data science" provides a historical opportunity to redefine data analysis in such a way that it naturally accommodates causal inference from observational data.

Like others before, we organize the scientific contributions of data science into three classes of tasks: description, prediction, and causal inference. An explicit classification of data science tasks is necessary to discuss the data, assumptions, and analytics required to successfully accomplish each task.

We argue that a failure to adequately describe the role of subject-matter expert knowledge in data analysis is a source of widespread misunderstandings about data science. Specifically, causal analyses typically require not only good data and algorithms, but also domain expert knowledge. We discuss the implications for the use of data science to guide decision-making in the real world and to train data scientists.

データ解析における現場知識の重要性
に対する認識の欠如は、
「データサイエンス」における幅広い誤解の源泉である。

Mode 2 Science

Gibbons, M. *et al.* (1994, pp.3-8) *The New Production of Knowledge: The Dynami and Research in Contemporary Societies*. Sage Publications.

Definition (Mode 2 Science)

The essential ingredients of production of knowledge by Mode 2 Science are

1. context of application;
2. transdisciplinary;
3. heterogeneity/diversity;
4. reflexivity/social accountability;
5. novel quality control source.

Mode 1 vs. Mode 2 Science

	Mode 1	**Mode 2**
Problem	Academic context	Context of application
Relation with other displines	Disciplinary	ransdisciplinary
Institute	Homogeneity	Heterogeneity
Goal	Autonomy	Reflexivity/social accountability (social accountability)
Criterion	Traditional quality control (peer review)	Novel quality control

1. Data Science as Mode 2 Science
   - Science Revisited
   - From Data Analysis to Data Science
   - Teaching Data Science
     - Teaching Mode 2 Data Science at YCU
     - Data Science Club: Learning by Practicing
     - Data Science PBL for Juniors
   - Practical Data Science in Master's Program

2. YOKOHAMA D-STEP

masters of the field. "Data analysts", even if professional statisticians, will have had far less exposure to professional data analysts during their training.

Three reasons for this hold today and can at best be altered slowly:

(c1) Statistics tends to be taught as part of mathematics.

(c2) In learning statistics *per se* there has been limited attention to data analysis.

(c3) The number of years of intimate and vigorous contact with professionals is far less for statistics Ph.D.'s than for physics (or mathematics) Ph.D.'s.

How to Teach DA: Tukey (1962)

John W Tukey (1962). The future of data analysis. *AMS*, 1-67.

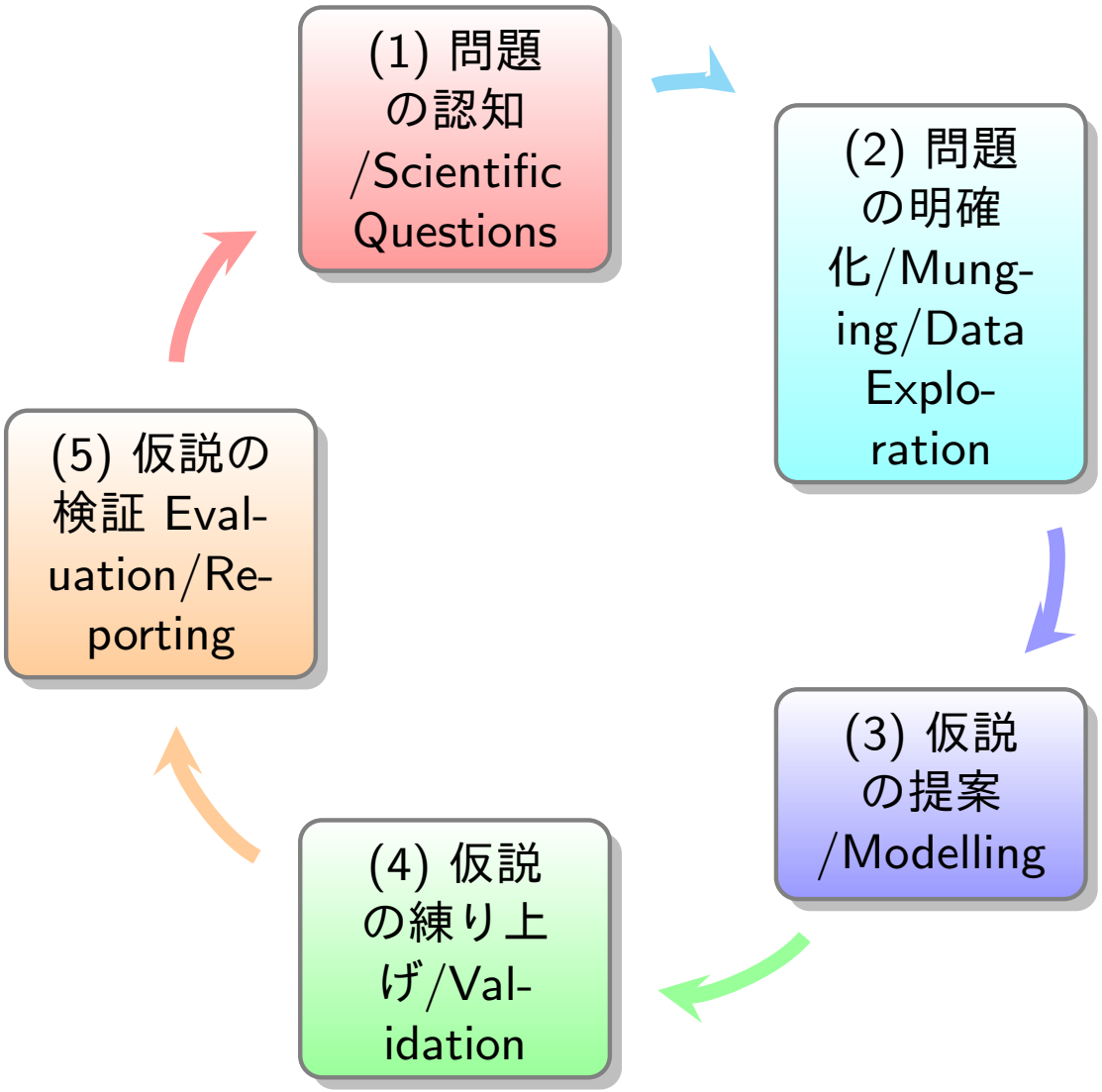
Data analysis, and the parts of statistics which adhere to it, must then take on the characteristics of a science rather than those of mathematics, specifically:

(b1) Data analysis must seek for scope and usefulness rather than security.

(b2) Data analysis must be willing to err moderately often in order that inadequate evidence shall more often *suggest* the right answer.

(b3) Data analysis must use mathematical argument and mathematical results as bases for judgment rather than as bases for proof or stamps of validity.

Teach DS Based on John Dewey's Structured Reflective Thinking



1. Data Science as Mode 2 Science
   - Science Revisited
   - From Data Analysis to Data Science
   - Teaching Data Science
     - Teaching Mode 2 Data Science at YCU
     - Data Science Club: Learning by Practicing
     - Data Science PBL for Juniors
   - Practical Data Science in Master's Program

2. YOKOHAMA D-STEP

Data Science Club: Learning by Practicing

YCUデータサイエンス倶楽部

目次

YCUデータサイエンス倶楽部とは

統計学・機械学習を初めとするデータサイエンスの基礎から、最先端までの研究やデータサイエンスがもたらす価値創造まで探求する研究会です。メンバーは横浜市立大学データサイエンス学部の学生から構成されます。他学部の学生はオブザーバーとして参加したい場合、必ず指導教員を通して倶楽部顧問までご連絡ください。

倶楽部長：十鳥大地（DS学部一年）

倶楽部顧問：汪 金芳（DS学部教授）

活動内容

2018年8月から次の統計学・機械学習の分野における名著を輪読する予定です。

統計的学習の基礎 ―データマイニング・推論・予測―

Trevor Hastie, Robert Tibshirani, Jerome Friedman

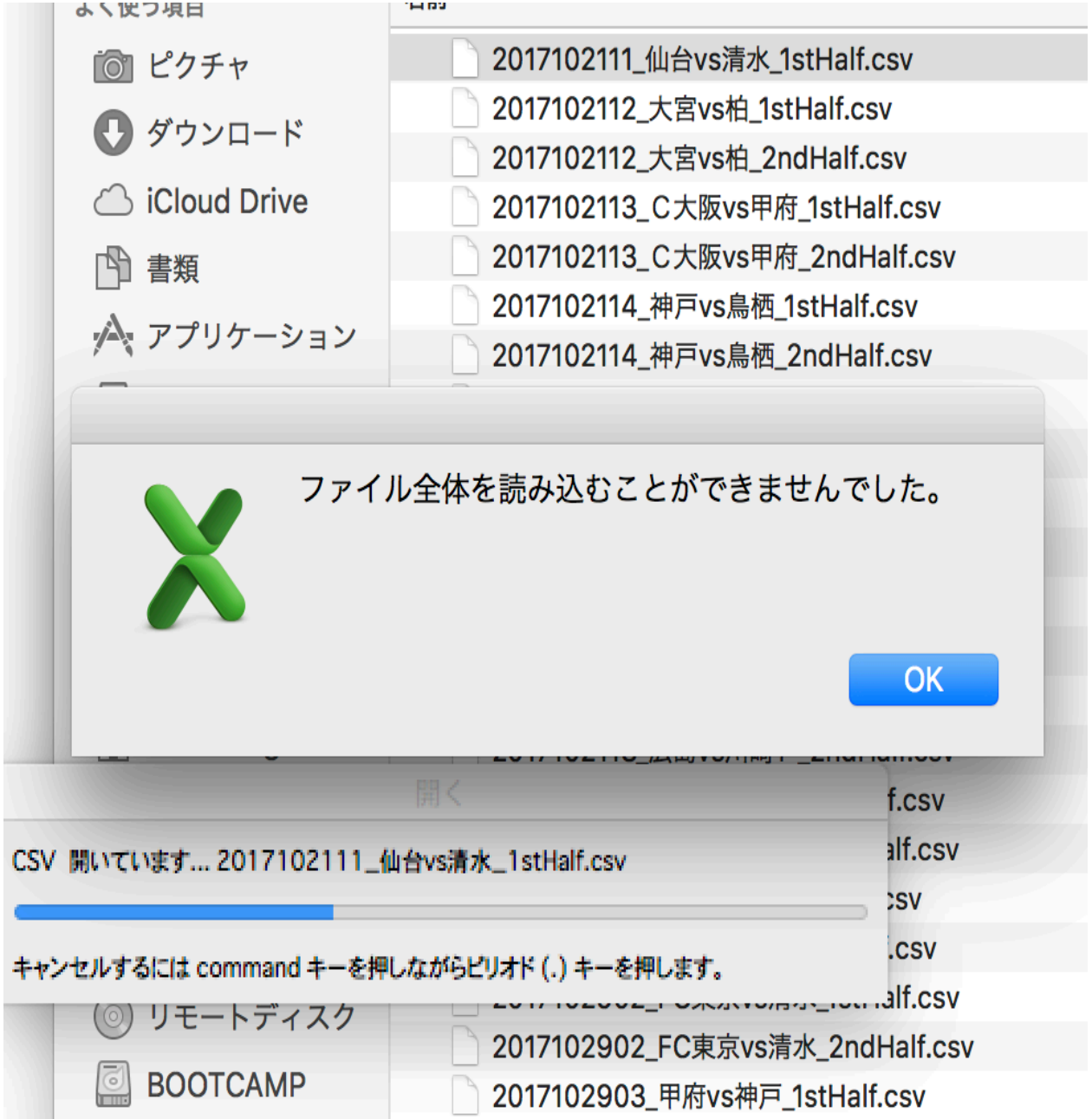
サッカーの勝利要因と予測について

横浜市立大学データサイエンス学部1年

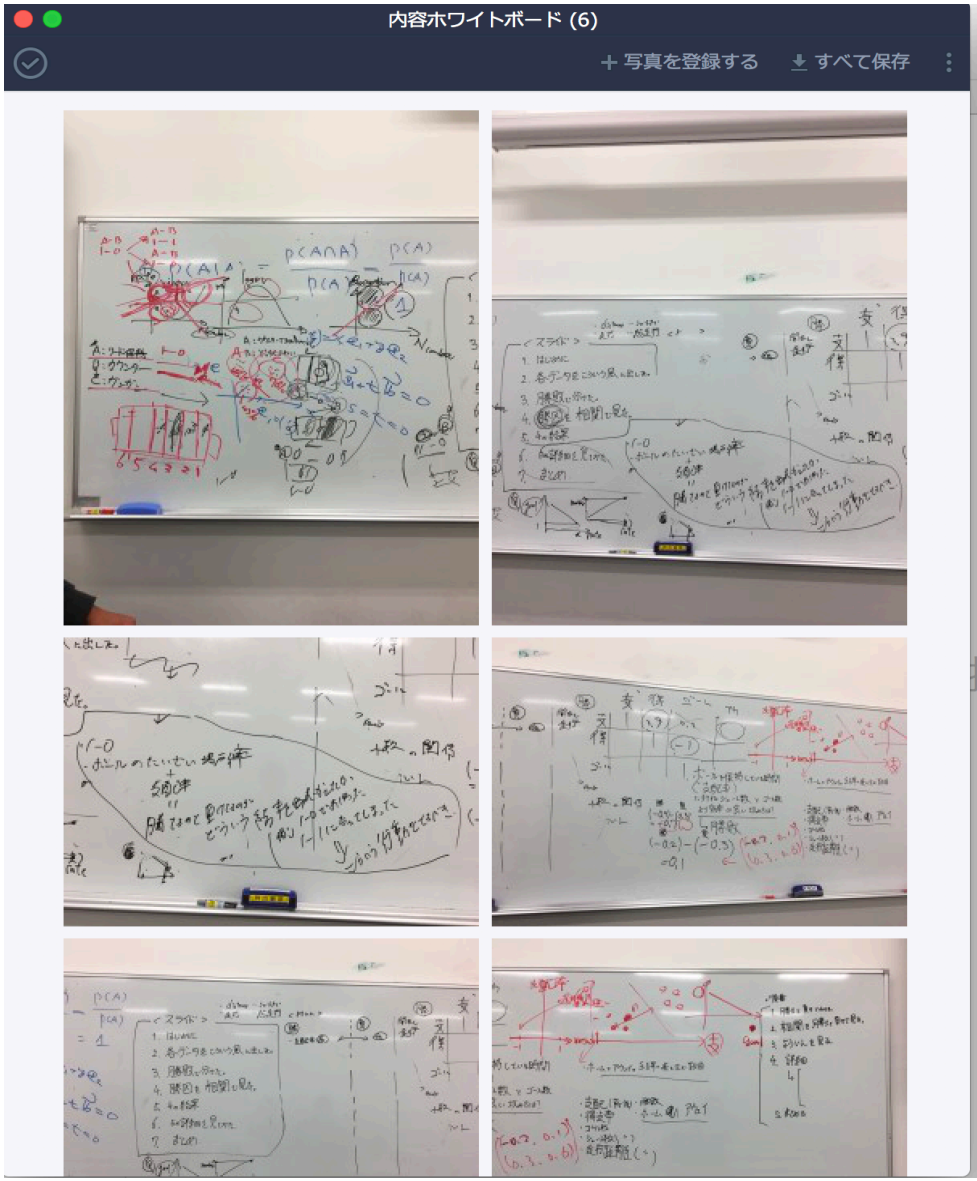
橘川 和司　千田 晟也　羽田野 佑奈
宇井 裕翔　長島 正悟　浦 優太　汪 金芳

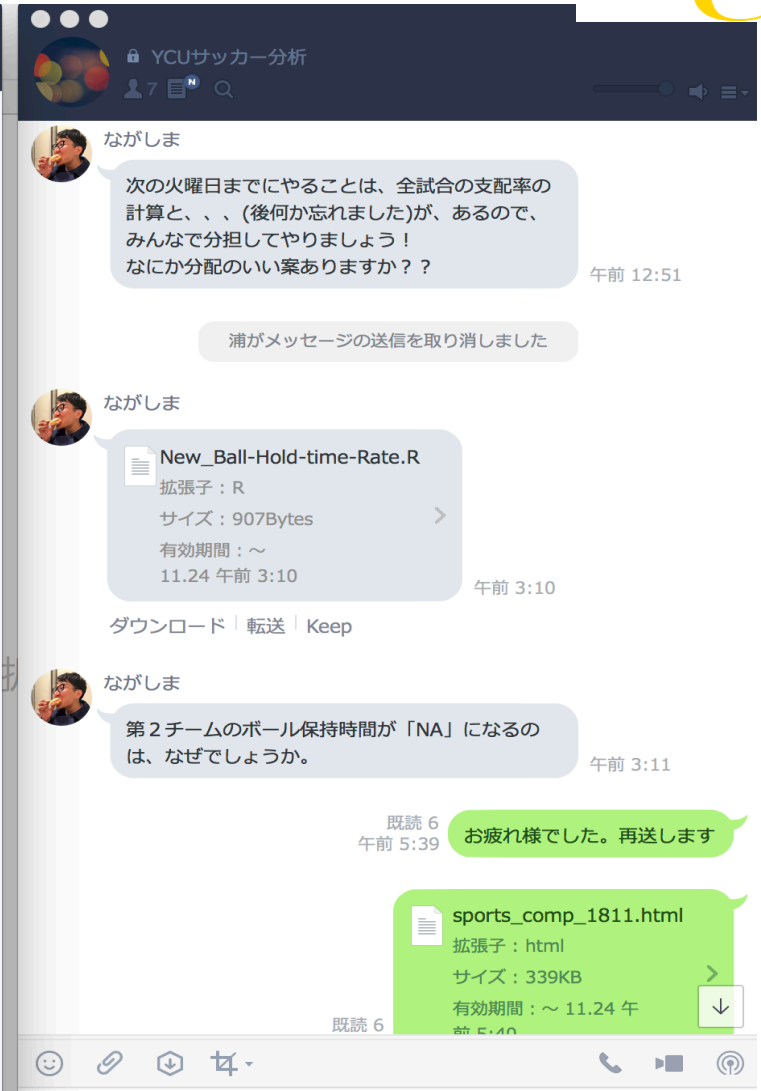
第8回スポーツデータ解析コンペティション：分析部門（サッカー）

Cann't Open File!



Most of Time Spent on “what to do”’’





5. 付録：Rのコード

```r
#走行距離の自動計算
#csvファイル名のベクトル
library("data.table")
#csvのファイルがたくさん入っているディレクトリ内で実行する
#ディレクトリの変更をやっておくこと！
Files=list.files()
#ボールタッチの試合情報_TBをcsvにして読み込む
#これも、ディレクトリ注意！
gameinfo=fread("GameInformation_TB.csv",header = T, sep = ',',
               encoding = 'UTF-8')
colnames(gameinfo)=c("gameId","date","half","homeId",
                     "homeName","homeNa","awayId",
                     "awayName","awayNa","atudiumId",
                     "studiumName","homeScore","homeResultId",
                     "awayScore","awayResultId",
                     "viewer","weather1","weather2",
                     "weather3","temp","humid",
                     "wind","grandState","grandSurface",
                     "startTime")
#チームIDだけ抜き出す
teamId=unique(subset(gameinfo,select = "homeId"))
colnames(teamId)="id"
#試合idとチームidの照らし合わせをする
idSet=subset(gameinfo,select = c("gameId","homeId","awayId"))

runDistanceNext=function(){
  #前後半に分かれているので、
  #全ファイル数の２分の１ずつに計測する
  for(i in 1:(length(Files)/2)){
    #IDセットからゲームIDを抜き出す
    gameID=idSet$gameId[i]
    #前半と後半のデータをそれぞれ読み込む
    firstHalf=fread(Files[2*i-1],header = T, sep = ',',
                    encoding = 'UTF-8')
    secondHalf=fread(Files[2*i],header = T, sep = ',',
                     encoding = 'UTF-8')
    #変数名を日本語から英語にする
    colnames(firstHalf) <- c("id", "frame","home","system",
                             "number","X","Y","speed")
    colnames(secondHalf) <- c("id", "frame","home","system",
                              "number","X","Y","speed")

    #ホームの出場選手一覧
    homePlayer=cbind(unique(subset(firstHalf,home==1,select = c("number")))
                     ,unique(subset(secondHalf,home==1,select = c("number"))))
    colnames(homePlayer)=c("firstNumber","secondNumber")
    #前後半で出場する選手の人数が違うので調整する
    if(length(unique(subset(firstHalf,home==1,select = c("number"))$number))!=
       length(unique(subset(secondHalf,home==1,select = c("number"))$number))){
```

```r
                              "number","X","Y","speed")

    #ホームの出場選手一覧
    homePlayer=cbind(unique(subset(firstHalf,home==1,select = c("number")))
                     ,unique(subset(secondHalf,home==1,select = c("number"))))
    colnames(homePlayer)=c("firstNumber","secondNumber")
    #前後半で出場する選手の人数が違うので調整する
    if(length(unique(subset(firstHalf,home==1,select = c("number"))$number))!=
       length(unique(subset(secondHalf,home==1,select = c("number"))$number))){}
    #NAとなる部分（人数の関係）ができた。
    #アウェイの出場選手一覧
    awayPlayer=cbind(unique(subset(firstHalf,home==2,select = c("number")))
                     ,unique(subset(secondHalf,home==2,select = c("number"))))
    colnames(awayPlayer)=c("firstNumber","secondNumber")
    #ホームの選手と同様に調整を行う
    if(length(unique(subset(firstHalf,home==2,select = c("number"))$number))!=
       length(unique(subset(secondHalf,home==2,select = c("number"))$number))){}

    #試合の途中結果と最終結果は
    Judge=subset(GameResult,gameId==gameID)$judge

    #前半終了時

    #選手の走行距離を入れるためのデータフレームを作る
    homeRunDistance=data.frame("firstNumber"=homePlayer$firstNumber,
                               "firstDistance"=rep(0,length(homePlayer$firstNumber)),
                               "secondNumber"=homePlayer$secondNumber,
                               "secondDistance"=rep(0,length(homePlayer$secondNumber)))
    awayRunDistance=data.frame("firstNumber"=awayPlayer$firstNumber,
                               "firstDistance"=rep(0,length(awayPlayer$firstNumber)),
                               "secondNumber"=awayPlayer$secondNumber,
                               "secondDistance"=rep(0,length(awayPlayer$secondNumber)))
    #走行距離のデータフレームを完成させる
    #ホーム前半
    for(j in 1:length(homePlayer$firstNumber)){}
    #ホーム後半
    for(j in 1:length(homePlayer$secondNumber)){}
    #アウェイ前半
    for(j in 1:length(awayPlayer$firstNumber)){}
    #アウェイ後半
    for(j in 1:length(awayPlayer$secondNumber)){}
    #完成

    #要約用データフレームを作る
    #チーム別の合計Summary｛総走行距離、延べ出場選手数、試合数、
    #一試合における走行距離、一試合における出場選手、選手一人における走行距離
    if(i==1){}
```

70:21 runDistanceNext()

Data Science PBL for Juniors

Industry and Government Oriented Problems

NISSAN

マイナビ

帝国データバンク

FamilyMart

Rejoui

ANA ANA TRADING CORP, U.S.A.

OPEN YOKOHAMA

横浜市 City of Yokohama

NTTグループ

PBL実習(インターンシップ)実施計画テンプレート

- 実習の類型

□実績追体験型　□解決プロセス現在進行型　□課題未着手型　□課題理解型　□その他

- 実習テーマ

- 実習内容

【課題】

【内容詳細】

- 期間：　月　日（　曜日）～　月　日（　曜日）（　日）　[最短1週間～最長3週間]
- スケジュール

	第1週		第2週		第3週	
	AM	PM	AM	PM	AM	PM
1日目						
2日目						
3日目						
4日目						
5日目						

- 受け入れ人数：　　人　（グループワーク）
- 期待スキル

- 事前学習

2019.08.02

Practical Data Science in Master's Program

Curriculum of Master in Data Science

DS専攻 博士前期課程カリキュラム

■必修科目　□選択科目　※（）内は単位数

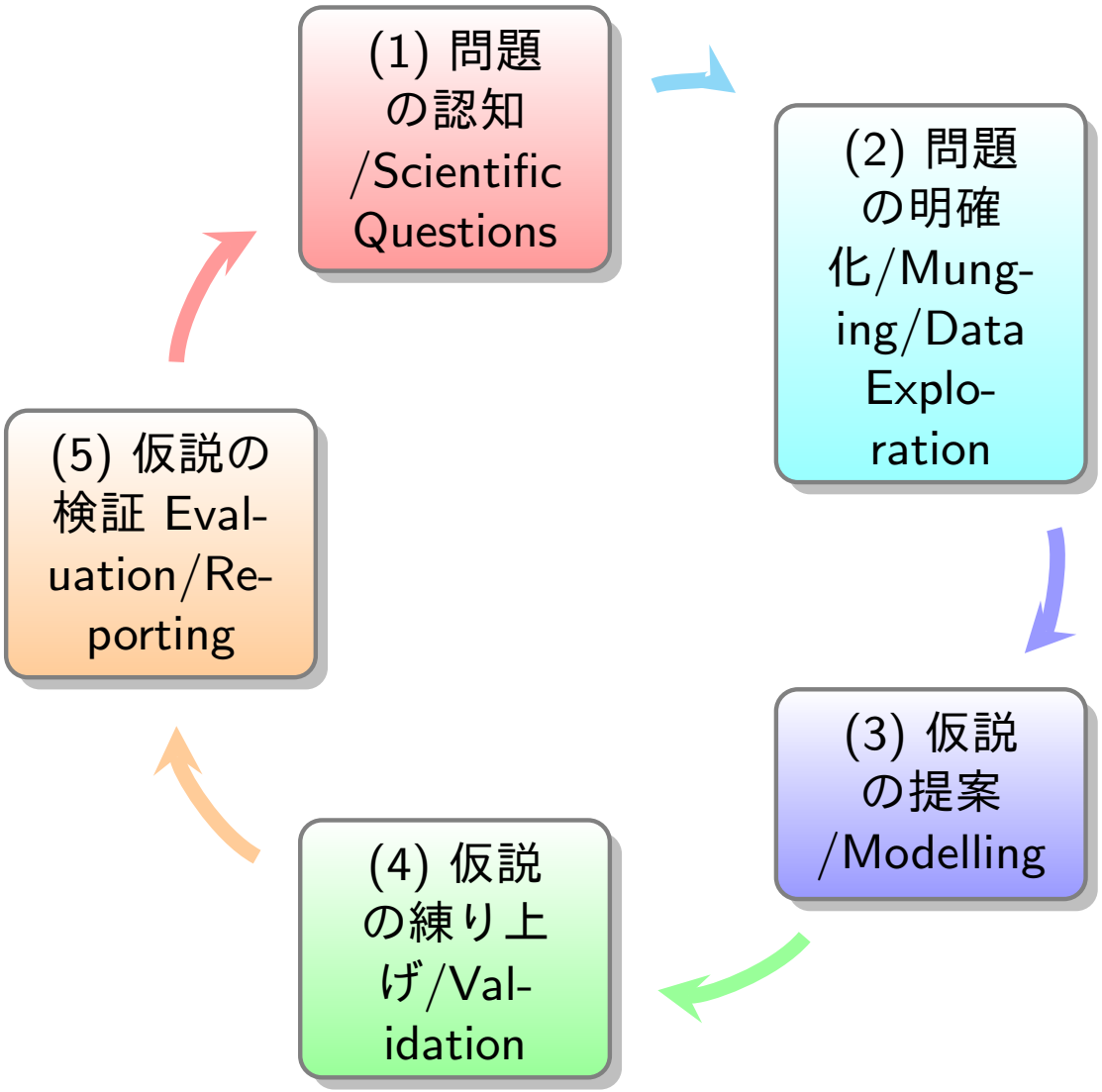
M1前期	M1後期	M2前期	M2後期
DS研究指導 I〜IV（計8／1年・2年）・修士論文（0）			
PDS I（2）	PDS II（2）	PDS III（2）	
統計学特論（2）	機械学習特論（2）		
デザイン思考特論（1）	データマンジング特論（2）		
応用倫理学（1）			
〈統計科学中心型〉			
多変量統計解析特論（2）	時系列データ解析特論（2）		
最適化の基礎と応用特論（2）	都市環境データ解析特論（2）		
実験計画と因果推論特論（2）	他専攻・他研究科科目（2）		
標本調査特論（2）			
〈計算機科学中心型〉			
クラウドコンピューティング特論（2）	ビッグデータ処理基盤特論（2）		
計算機統計学特論（2）	非構造化データ特論（2）		
プログラミング特論（2）	実験とシミュレーション特論（2）		
	データ可視化特論（2）		
	自然言語処理特論（2）		
データアナリティクス特別講義・データエンジニアリング特別講義・データサイエンス展開特別講義（集中講義等）（2）			

修了要件：30 単位

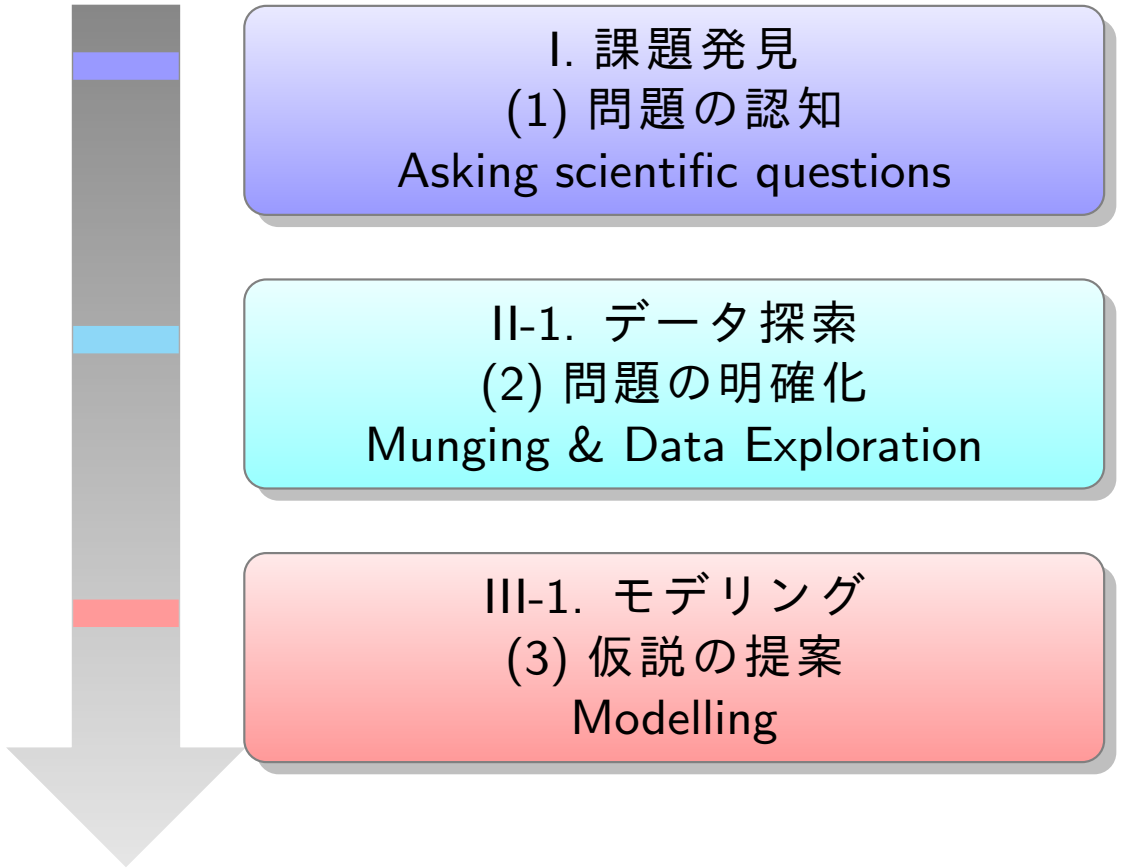
- 特別研究・特別演習：14単位　PDS(必修)6単位、ゼミ/修士論文(必修)8単位
- 共通科目：16単位 講義/演習(必修) 8単位、講義/演習(選択) 8単位以上

- 学部4年生対象の早期履修科目は、特論科目を5科目10単位とする

PDS Based on John Dewey's Structured Reflective Thinking

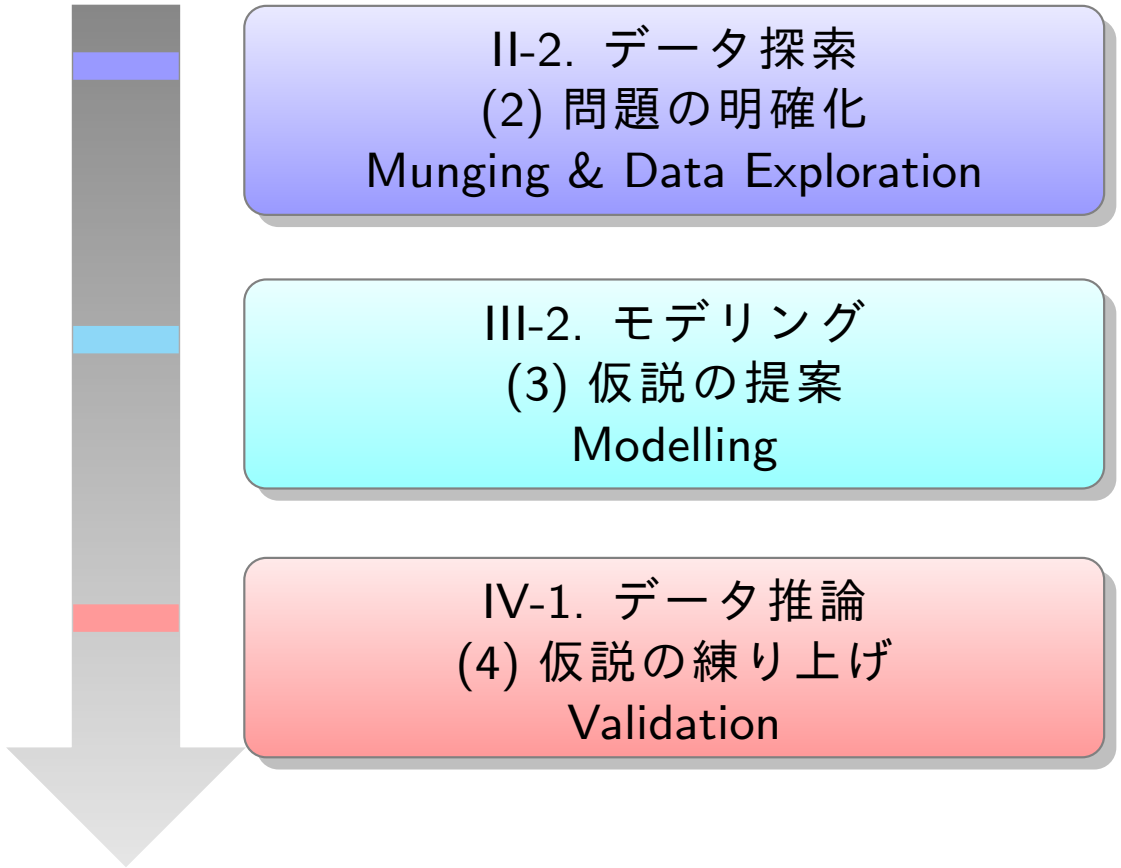


Practical Data Science I



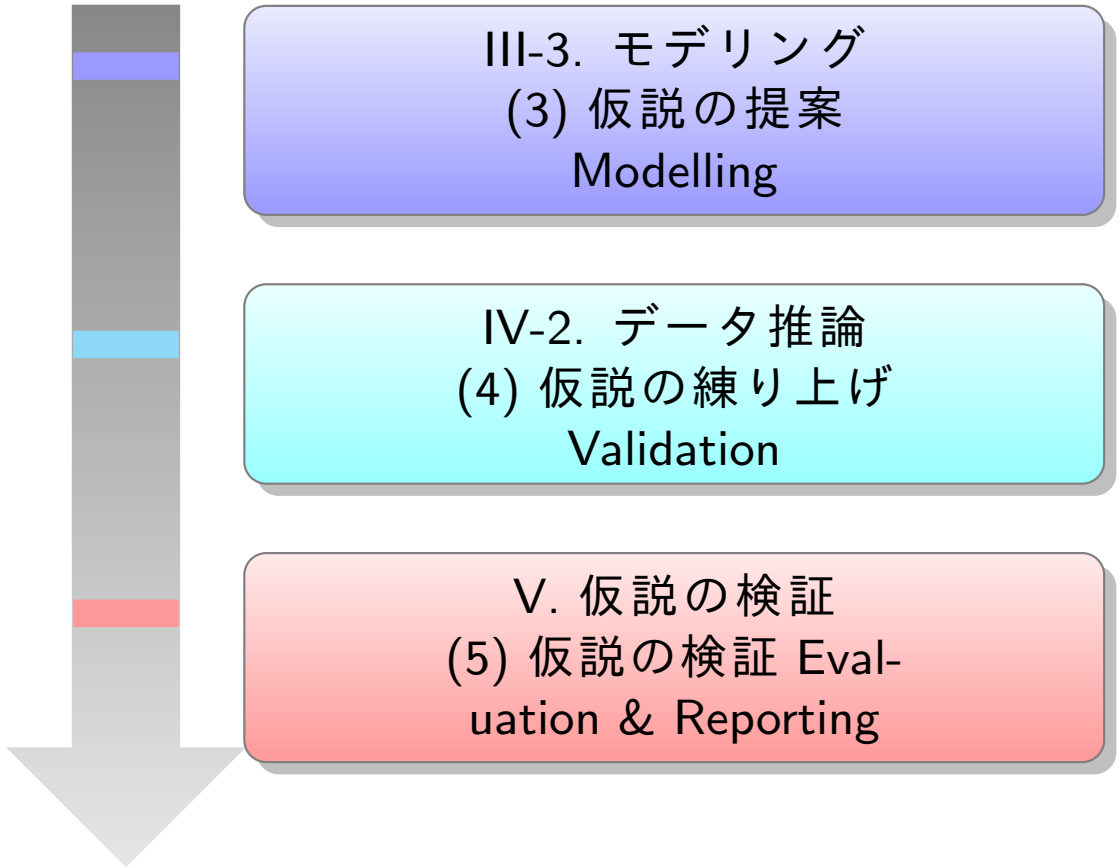
M1 前期 2 単位 (Semester 1, 2 credits)

Practical Data Science II



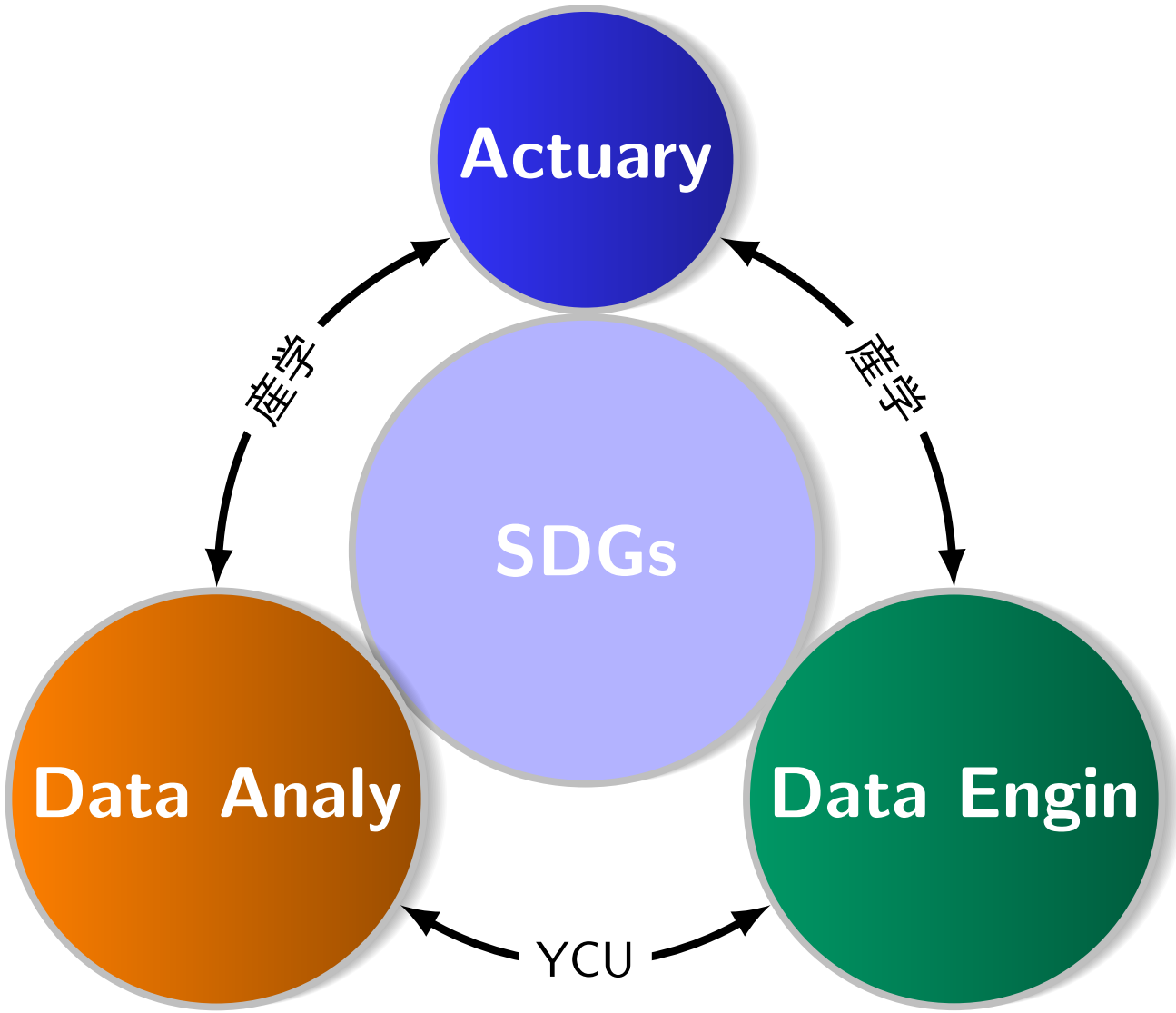
M1 後期 2 単位 (Semester 2, 2 credits)

Practical Data Science III

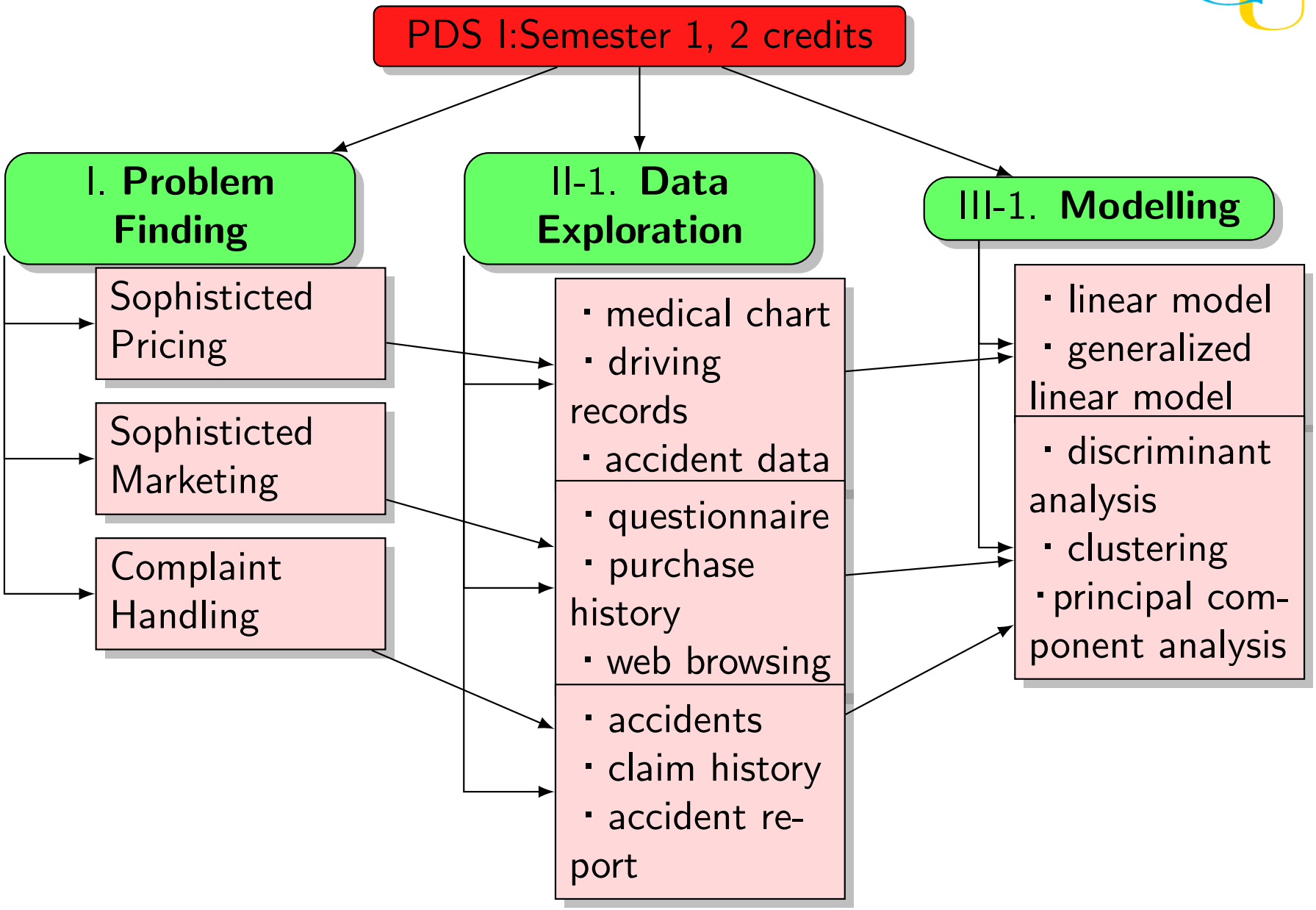


M2 前期 2 単位 (Semester 3, 2 credits)

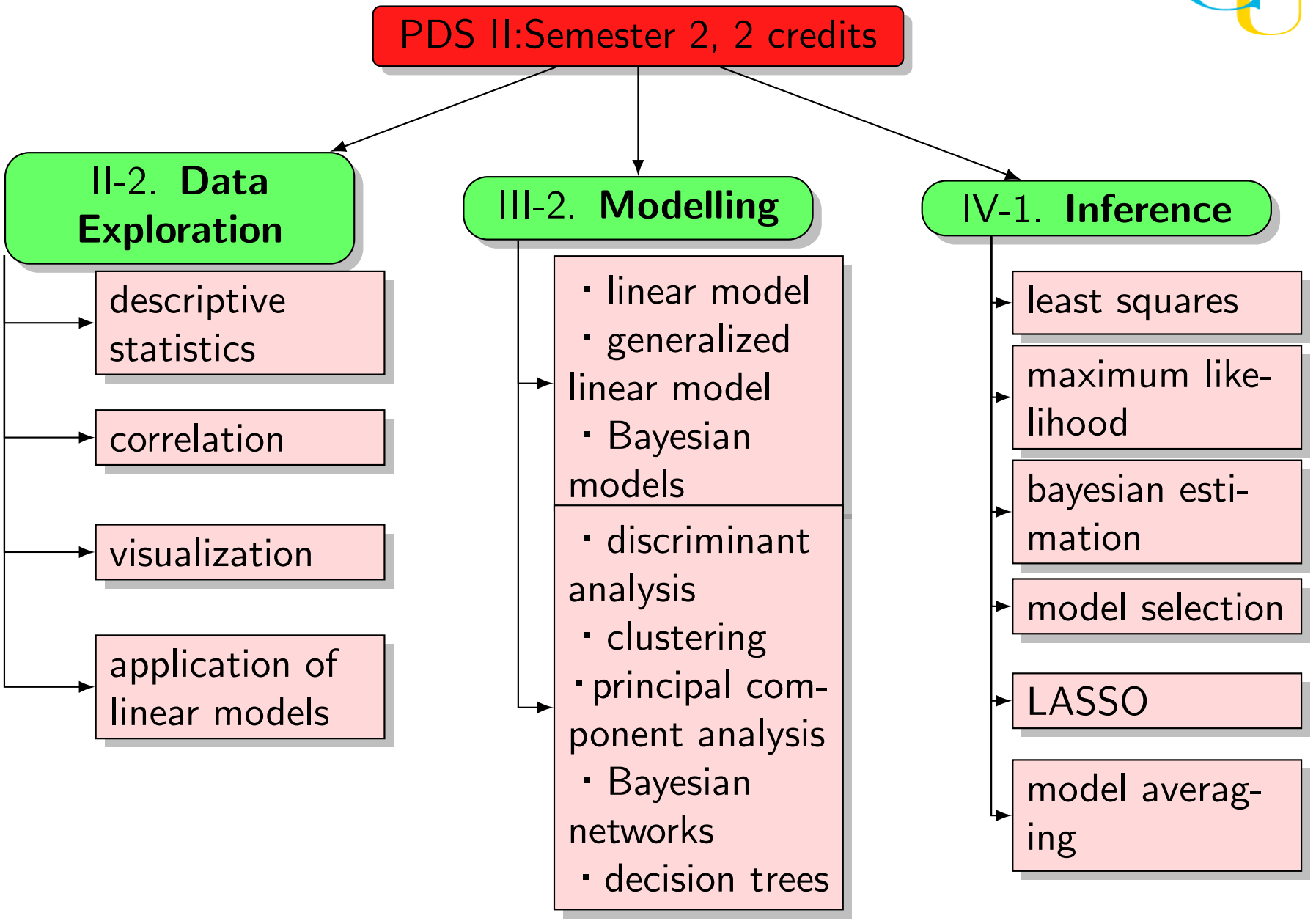
Data Science for Sustainable Insurance



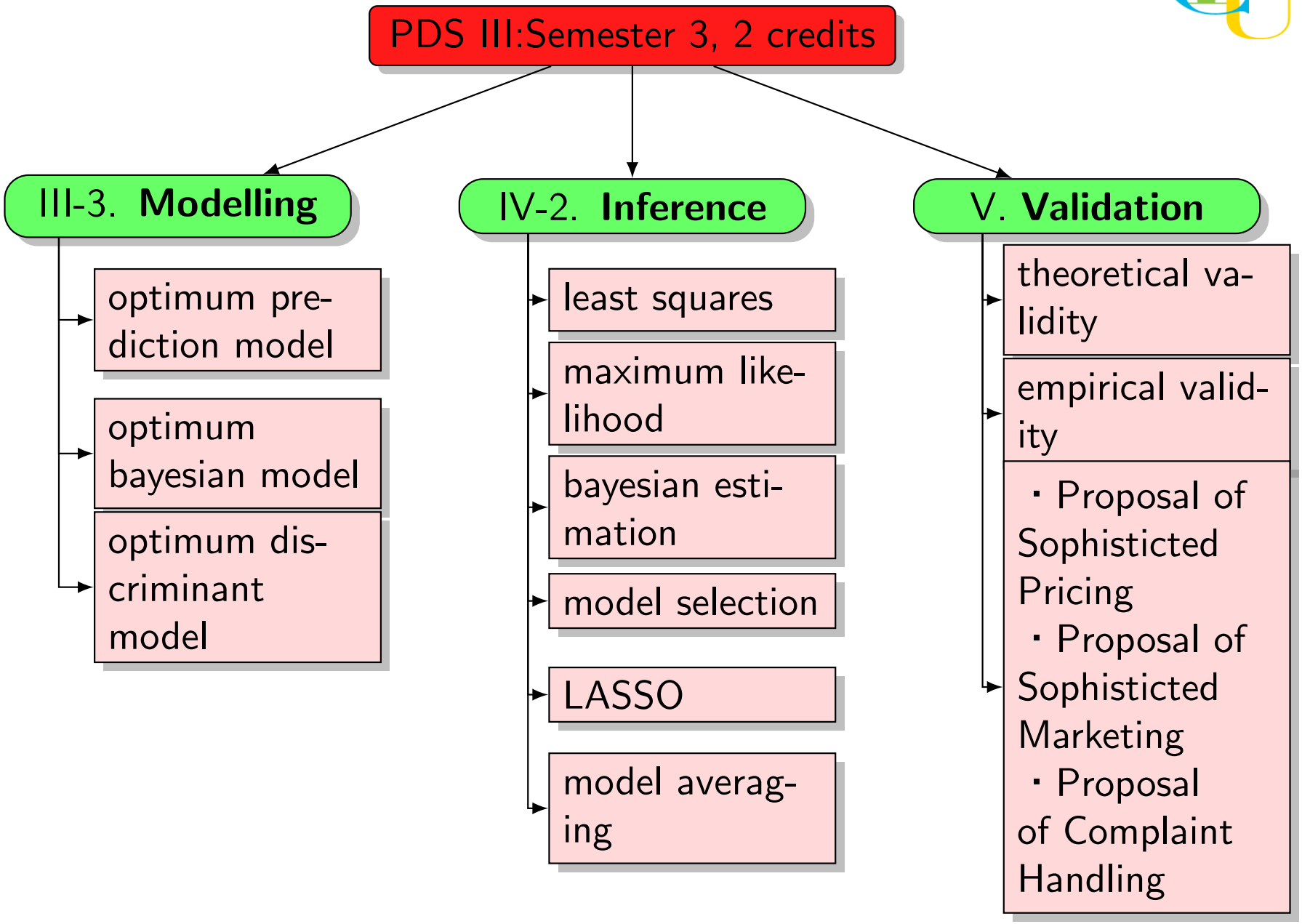
PDS for Sustainable Insurance (I)



PDS for Sustainable Insurance (II)



PDS for Sustainable Insurance (III)



文理融合・実課題解決型データサイエンティスト育成

YOKOHAMA D-STEP

Data Scientist Educational Program

D-STEPの取り組み

横浜市立大学 データサイエンス推進センター
坂巻顕太郎

sakamaki@yokohama-cu.ac.jp

横浜市立大学
YOKOHAMA CITY UNIVERSITY

2019.02.19 プレスリリース 大学

データサイエンティスト育成プログラムの募集開始！文部科学省「超スマート社会の実現に向けたデータサイエンティスト育成事業」始動

横浜市立大学は、東京理科大学、明治大学と３大学共同で「文理融合・実課題解決型データサイエンティスト育成※１」事業を実施します。これは、平成30年度文部科学省「超スマート社会の実現に向けたデータサイエンティスト育成事業※２」に採択された事業で、横浜市内のみならず我が国の社会的、経済的課題をデータサイエンスで解決できる人材育成のプラットフォームを構築し、5か年で約200人の高度データサイエンティスト、約800名のデータエキスパートを産学官連携のもとで実践的に育成することを目指すものです。特に、データサイエンティストが社会活動の中で直面する［課題設定⇒データ収集・分析⇒新たな価値創造⇒実装］という一連のプロセスへの対応を念頭に置いたPBL※３を行い、受講生の社会展開力を涵養することに力点を置いています。
本事業では、2018年に横浜市立大学に開設されたデータサイエンス学部の知見を生かして、大学間、産業界、行政と連携して実社会から求められるデータサイエンティストを育成するために、3つの教育プログラムを開設し、本年2月19日より受講生の募集を開始します。

横浜市立大学 YOKOHAMA CITY UNIVERSITY

プログラム

受講生の保有知識・スキルを基に、独り立ち可能なデータサイエンティストを要請。 また、データサイエンティストを社会・企業に広げ、データサイエンティストの活躍の場を創るビジネスパーソン文化まで促成

A 課題発見・解決型　データサイエンティスト育成コース

データサイエンスに係る一連の流れを"習得したサイエンティスト"を輩出

・データサイエンス分野"非"専攻分野者が対象
・データサイエンスの基礎知識から発展までを習得
・通年のPBLで課題の発見・解決までの一連の流れを学習

B データ分析型　データサイエンティスト育成コース

データサイエンス分野で自走できる"習熟したサイエンティスト"を輩出

・データサイエンス分野専攻者が対象
・データサイエンスに関する高度な理論・方法を習得
・半期のPBLで実課題解決から価値創出までを経験

C 社会人（データエキスパート）育成コース

文化を創りサイエンティストを使いこなせる"エキスパート"を輩出

・短期間でデータサイエンスに関する基礎知識や活用手法を習得
・現場で活かせる軽微な分析素地を醸成
・データサイエンティストの活躍の場、活用方法を理解

なお、本プログラムは、国が定める大学等における履修証明制度に該当する 「履修証明プログラム」となります。「履修証明プログラム」は、学校教育法第 １０５条及び学校教育法施工規則第１６４条の規定に基づき、大学が特別に社会 貢献等を趣旨として、主として社会人向けに体系的な教育プログラムを開設し、 その修了者（１２０時間以上の履修を完了する者）に対し、学長名の履修証明を交付するものです。

https://d-step.yokohama/

横浜市立大学
YOKOHAMA CITY UNIVERSITY

講義概要（A, Bコース）

Aコース: 課題発見・解決型データサイエンティスト育成コース
Bコース: データ分析型データサイエンティスト育成コース

授業科目	A	B	開講時期	単位	授業科目	A	B	開講時期	単位
統計学基礎	◎		e-learning	1	統計・機械学習モデリング	◎	◎	後期	1
応用線形代数	◎		e-learning	1	データマンジング	○	○	後期	2
データ分析基礎	◎		e-learning	1	最適化と計算機科学	○	○	後期	2
ITセキュリティ・情報倫理	◎	◎	e-learning	1	非構造化データ特論	○	○	後期	1
多変量データ解析	○	○	前期	2	実験とシミュレーション特論	○	○	後期	1
標本調査法	○	○	前期	1	プログラミング特論	○	○	後期	1
欠測データ解析	○	○	前期	2	計算機統計学特論	○	○	前期集中	1
ノンパラメトリック法	○	○	前期	2	時系列データ解析特論	○	○	前期集中	1
最適化理論	○	○	前期	2	都市環境データ解析特論	○	○	前期集中	1
数理ファイナンス*	○	○	前期	2	データ可視化特論	○	○	前期集中	1
デザイン思考	◎	◎	前期集中	1	数理医学	○	○	集中	2

◎: 必修科目, ○: 選択科目, △: 選択必修科目

授業科目	A	B	開講時期	単位
政策課題解決PBL	◎	△	前期	2
ビジネス課題解決PBL	◎	△	後期	2

https://www.yokohama-cu.ac.jp/academics/ds/syllabus.pdf

横浜市立大学 YOKOHAMA CITY UNIVERSITY

政策課題解決PBL

講義概要	本講義では、個票データ（擬似データ）を用いて、グループワークを通じた実践的な形で問題解決の一連のプロセスを学習する。ビジネス経験、データ分析経験が豊富な講師陣が、各グループのアウトプットに対するフィードバックを提供する。問題解決における他者との協働を身に着けるだけでなく、模擬インタビューやプレゼンテーションを実施し、問題解決に必要な能力の素地を総合的に養う。

授業計画	1	講義①（データ分析における心構え）
	2	講義②（ロジカルシンキング）
	3	課題説明
	4	講義③（データ分析）
	5	講義④（インタビューの心得・インタビューによる情報収集）
	6	グループワーク（課題設定・データ分析）
	7	グループワーク（データ分析・発表資料作成）
	8	中間発表
	9	講義⑤（解決策提言に向けたストーリー構築）
	10	グループワーク（中間発表でのレビューを踏まえた再検討）
	11	グループワーク（データ分析・発表資料作成）
	12	グループワーク（データ分析・発表資料作成）
	13	最終発表リハーサル
	14	最終発表に向けた資料のアップデート
	15	最終発表

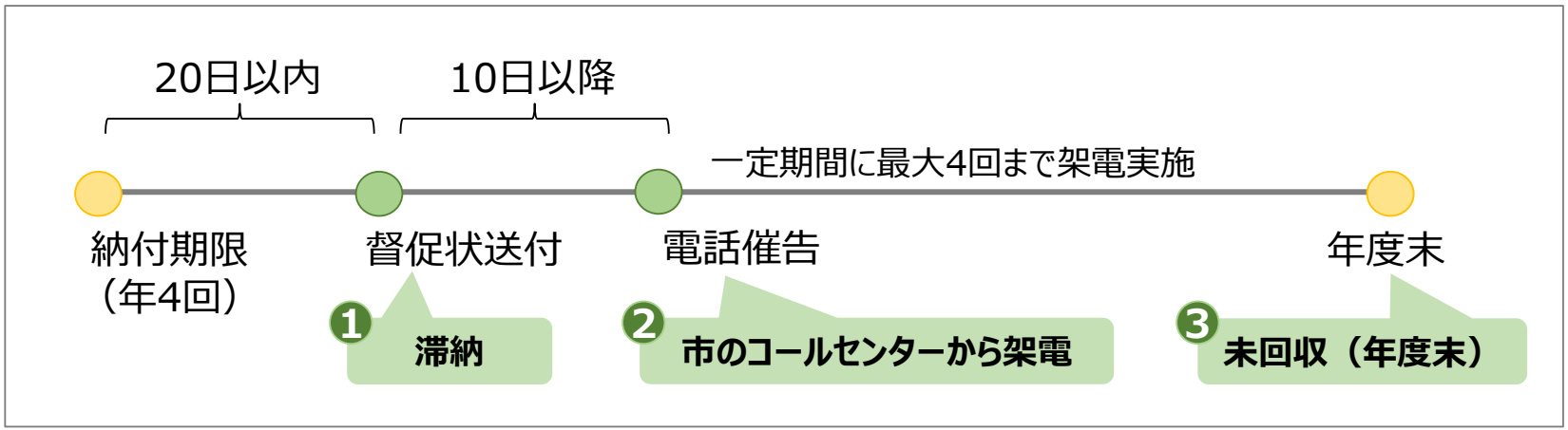
プロジェクトゴール

・**ゴール：今年の滞納未回収額を10%削減**

- A市の納税課では、「1年で滞納未回収額を10%減」という目標値を設定しています。
- 今年度中の回収額増加を狙うため、6月末までに施策をつくる必要があります。
- 大きな社会環境変化は見込まれないため、何も施策を行わなかった場合、今年度も約50億円の未回収が発生するものと想定しています。

横浜市立大学 YOKOHAMA CITY UNIVERSITY

督促業務



① 滞納とは、納税者が納付期限までに納税しないために督促状を送達したにもかかわらず、完納されないことをいう。督促状が送達されていることが滞納処分の原則的な前提条件となる。

② 督促状送付から一定期間後に、市のコールセンターから電話による納税の催告を行う。
A市では、電話が繋がらなかった場合、最大4回まで架電するルールとする。コールセンターでは市民税以外も担当しており、現状の架電回数を増加することは難しい。

③ 未回収は、電話による督促を実施しても年度末までに納付されなかった分を指す。

横浜市立大学
YOKOHAMA CITY UNIVERSITY

ビジネス課題解決PBL

- 現実の課題に対してデータから具体的な施策を提案
  - 2019年度は ベイスターズ、資生堂、ハレックスとの協同

- 講義内容
  - 企業と課題の共有
    - 課題とデータの説明（初回、中間レビュー）
  - 課題解決のためのデザイン
    - アウトカム / 解析単位の設定と説明変数の考案
  - データ加工
    - SQLによるデータ抽出や説明変数の加工など
  - 基本的な分析手法の適用
    - Rによる分析 （glm, step, glmnetなど）
  - 機械学習手法による高精度な予測
    - 各種パッケージとクロスバリデーション
  - 施策の立案とプレゼンテーション
    - 分析結果から施策を立案

横浜市立大学 YOKOHAMA CITY UNIVERSITY

課題とデータの説明

資生堂ジャパン

横浜DeNAベイスターズ

ハレックス

- 中間レビュー
  - 受講生によるビジネス課題解決策のプレゼン
  - 協力企業から提案事項に対する有益なコメントを得て，提案をさらにブラッシュアップ

横浜市立大学 YOKOHAMA CITY UNIVERSITY

[文部科学省] 超スマート社会の実現に向けたデータサイエンティスト育成事業
(事業責任者:山中 竹春)

横浜市立大学

自治体向け データエキスパート 育成コース開講


田栗 正隆


山中 竹春


坂巻 顕太郎


西内 啓

EBPM のためのデータ利活用方法

2016 年に施行された官民データ活用推進基本法では、多様かつ大量の情報を活用することにより超少子高齢社会における諸課題を解決するため、地方公共団体を含む様々な主体が保有するデータの適正かつ効果的な利用を推進しています。
基本理念は、データ活用により、自立的で個性豊かな地域社会の形成などを図り、根拠（エビデンス）に基づく政策立案や評価を行うことです。この理念に基づいて官民データを十分に活用するには、データ活用の知識や技術を修得することが重要になります。

自治体職員のために 首都圏初のデータサイエンス学部が 短期集中講座を開講

本セミナーでは、地方自治体の職員を対象に、1) エビデンスを理解するために必要な知識とデータ分析、2) すでに蓄積されているエビデンスを探索するために必要な文献検索の方法、3) 課題解決のためのデータ分析、の 3 つについて具体例に基づいた講義と演習を行います。これにより、エビデンスに基づいた政策立案や評価の効果的かつ効率的な推進に必要である、データ活用に関する基本的な素養の修得が期待されます。

詳細はこちらから https://d-step.yokohama

2020年 1月29日(水)・30日(木)

会場:神奈川産業振興センター第2会議室
(地下鉄関内駅 7番出口より徒歩2分・JR関内駅北口より徒歩5分・JR桜木町駅より徒歩7分・みなとみらい線馬車道駅 3番出口より徒歩7分)

主催:公立大学法人横浜市立大学

- 定　員: 50名
- 対　象: 自治体職員
- 受講料: 無料
- 持ち物: ノートPC
- 講師陣: 田栗　正隆（横浜市立大学）
  坂巻顕太郎（横浜市立大学）
  西内　啓（横浜市立大学/株式会社データビークル）
  など
- 事業責任者: 山中　竹春（横浜市立大学）

プログラム

Part. 1 2020年1月29日(水)9:30 – 12:00 エビデンスとデータ活用

どのような人が健康に関する課題があるかをデータから探索することと、スポーツジムなどに通いやすくする政策が健康づくりに寄与するかをデータから考えることは異なります。この違いを理解するには、エビデンスとは何か、データ分析で何ができるか、などを知ることが重要です。本パートでは、データの読み解きに関する講義からエビデンスやデータ活用の理解を促進することを目的とします。

Part. 2 2020年1月29日(水)13:00 – 17:00 既存のエビデンスを探索するための文献検索

課題や解決策に関する知見が既に存在するかどうかを事前に文献などから検討しなければ、それらの重要さやデータに基づく検討が必要かどうかの判断はできません。課題に対する知見（仮説）なしにビッグデータを解析することで Google Flu Trends がインフルエンザの流行予測に失敗したことが顕著な例です。特定の分野（業界や領域など）に関する知識の積み重ねはデータサイエンスを実装するうえでは重要になります。本パートでは、文献検索に関する講義と演習により、政策立案や評価をエビデンスに即して行うために必要となる事前の知見の収集の理解を目的とします。

Part. 3 2020年1月30日(木)9:30 – 17:30 データに基づく課題の検討

データ分析の目的に応じて、用いる分析手法や分析結果の解釈は異なります。例えば、課題解決策をデータから評価する際、解決策を実施した集団と実施していない集団のデータを単純に比較するのか、年齢や性別などの特徴を考慮して評価するのかは、どのようなデータを評価に用いるかによって異なります。本パートでは、データ分析の基礎的な演習を行い、データハンドリングやデータ分析に対する理解を深めることで、実際のデータ分析とエビデンスの関係を理解することを目的とします。

<注意事項>
ノート PC はデータ分析に用います。Excel（または csv ファイルを閲覧できるソフトウェア）、RStudio がインストールされていることが必要となります。RStudio のインストール、解析に必要なパッケージのインストールについては別途ご案内いたします。
※注　RStudio とパッケージのインストールには管理者権限が必要となるため、個人が所有する PC か、管理者アカウントで必要なソフトウェアとパッケージのインストールが完了した PC をご持参ください。管理者アカウントが使用できるかどうかは各自でご確認ください。

お問い合わせ先
公立大学法人 横浜市立大学 D-STEP 事務局
〒236-0027 神奈川県横浜市金沢区瀬戸 22-2
☎ 045-787-8906
E-mail: d_step@yokohama-cu.ac.jp
https://d-step.yokohama

文理融合・実課題解決型データサイエンティスト育成
Yokohama D-STEP
Data Scientist Educational Program

横浜市立大学

横浜市立大学 YOKOHAMA CITY UNIVERSITY

D-STEPの取り組み

- A, Bコース
  - データサイエンティストを育成するための通年コース
  - 統計や機械学習，デザインに関する講義
    - on-siteだけでなくonlineでも提供
  - Project Based Learningによるグループ演習
    - 実課題解決にデータサイエンスをどう生かすかを学ぶ機会

- Cコース
  - データサイエンスに関わる非専門家が必要な基礎的な知識を提供
    - 専門家と非専門家のコミュニケーションを円滑にする
  - データサイエンティストへの入門
    - A, Bコースへの接続

横浜市立大学 YOKOHAMA CITY UNIVERSITY

D-STEPの取り組み

- A, Bコース
  - データサイエンティストを育成するための通年コース
  - 統計や機械学習，デザインに関する講義
    - on-siteだけでなくonlineでも提供
  - Project Based Learningによるグループ演習
    - 実課題解決にデータサイエンスをどう生かすかを学ぶ機会

- Cコース
  - データサイエンスに関わる非専門家が必要な基礎的な知識を提供
    - 専門家と非専門家のコミュニケーションを円滑にする
  - データサイエンティストへの入門
    - A, Bコースへの接続

横浜市立大学 YOKOHAMA CITY UNIVERSITY