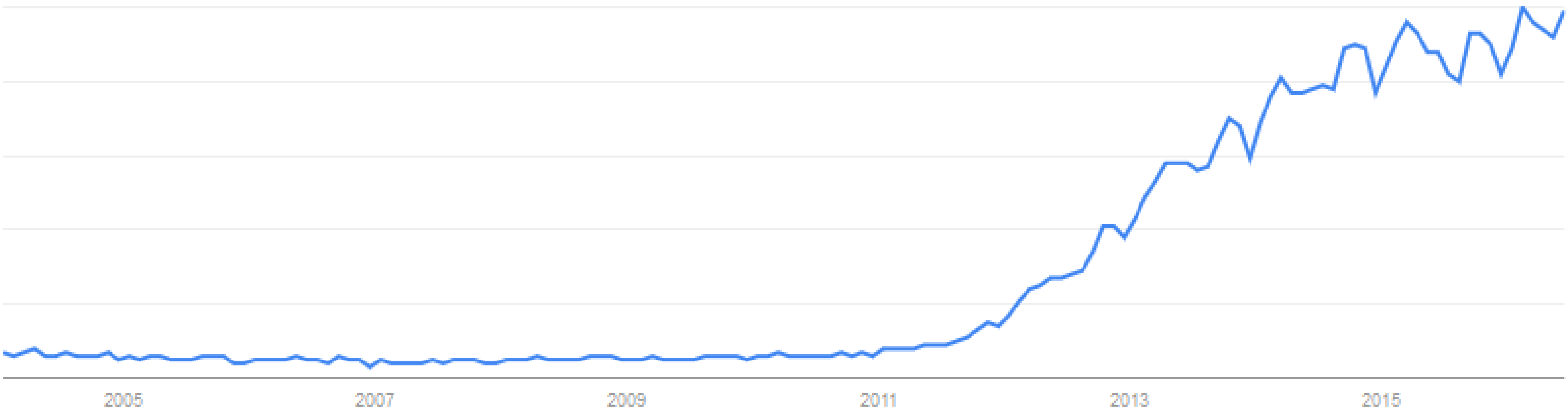
2016年度　統計関連学会　連合大会
金沢大学　2016年9月6-9日

The seven pillars of Data Science

Hideyasu SHIMADZU

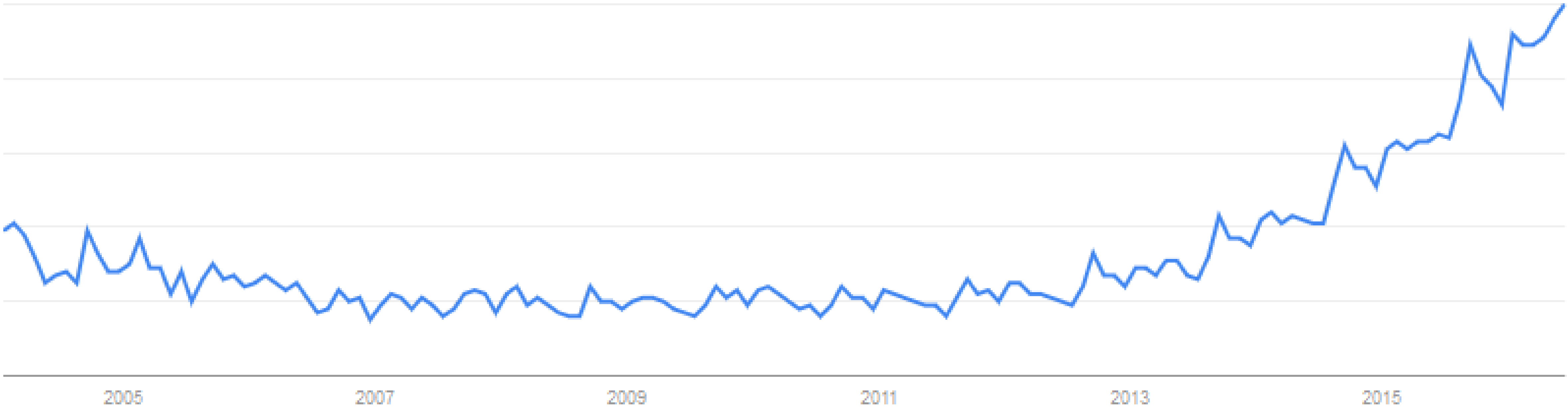
Department of Mathematical Sciences and
Centre for Data Science,
Loughborough University, UK

“Big Data”



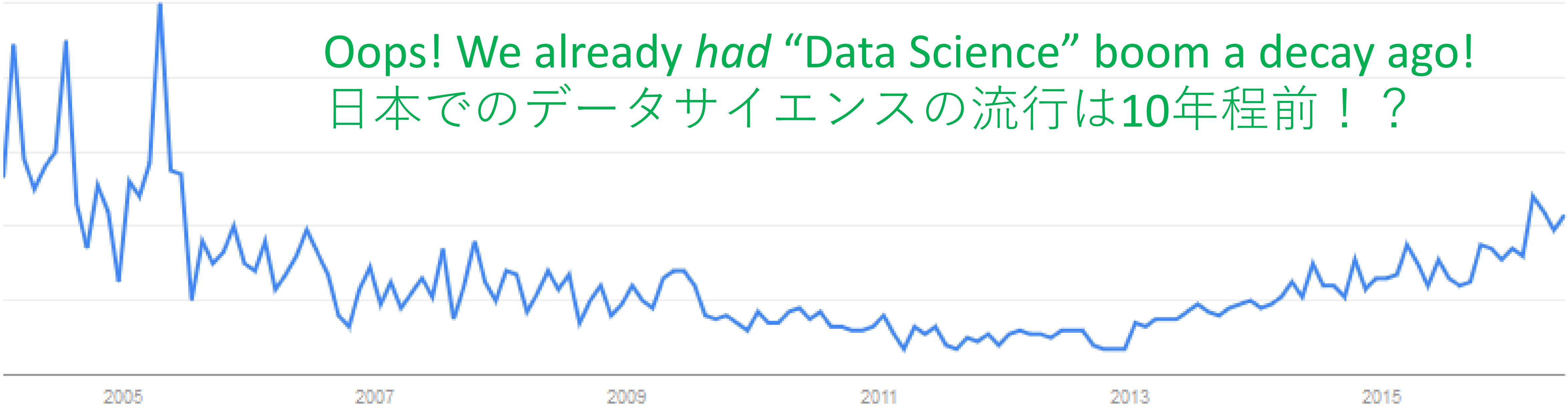
Google Trends

“Data Science”



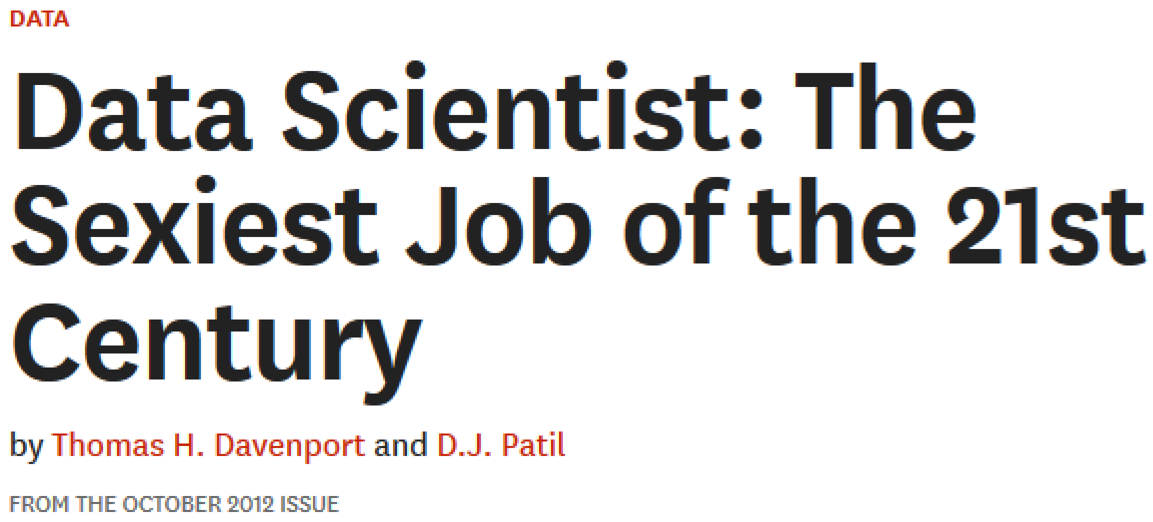
Google Trends

“データサイエンス” <- “Data Science” in Japanese



Oops! We already *had* “Data Science” boom a decay ago!
日本でのデータサイエンスの流行は10年程前！？

Google Trends



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

“If you torture the data long enough, it will confess.”

Ronald Coase

“Let data speak, never torture them.”

データを尋問するのではなく，データに語らせる

HBR.ORG
Harvard Business Review
OCTOBER 2012
46 The Big Idea
The True Measures Of Success
Michael J. Mauboussin
84 International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz
93 Leadership
What Ever Happened To Accountability?
Thomas E. Ricks
GETTING CONTROL OF
BIG DATA
How vast new streams of information are changing the art of management
PAGE 59

October 2012

Forbes

The World's 7 Most Powerful Data Scientists

By Tim O’Reilly (2011)

1. Larry Page, CEO, Google
2. Jeff Hammerbacher, Chief Scientist, Cloudera and
   DJ Patil, Entrepreneur-in-Residence, Greylock Ventures
3. Sebastian Thrun, Professor, Stanford University and
   Peter Norvig, Data Scientist, Google
4. Elizabeth Warren, Candidate, U.S. Senate (Massachusetts)
5. Todd Park, CTO, Department of Health and Human Services
6. Alex "Sandy" Pentland, Professor, MIT
7. Hod Lipson and Michael Schmidt,
   Computer Scientists, Cornell University

Data Science activities in the UK

- Alan Turing Institute (HQ: British Library): 2015-
  - Big Data (University of Cambridge)
  - Edinburgh Data Science (University of Edinburgh)
  - Oxford Internet Institute (University of Oxford)
  - Centre for Data Science (University College London)
  - Warwick Data Science Institute (University of Warwick)
- Data Science Institute (Imperial College London): 2014-
- Data Science Institute (Lancaster University): 2014-
- Centre for Data Science (Loughborough University): 2014-
- Leeds Institute for Data Analytics (University of Leeds): 2014-
- Institute for Analytics and Data Science (University of Essex): 2015-
- Data Science Institute (University of Manchester): 2015-

EPSRC
Engineering and Physical Sciences
Research Council

£42m ~ 56億円 for the initial 5 yrs

More and more!

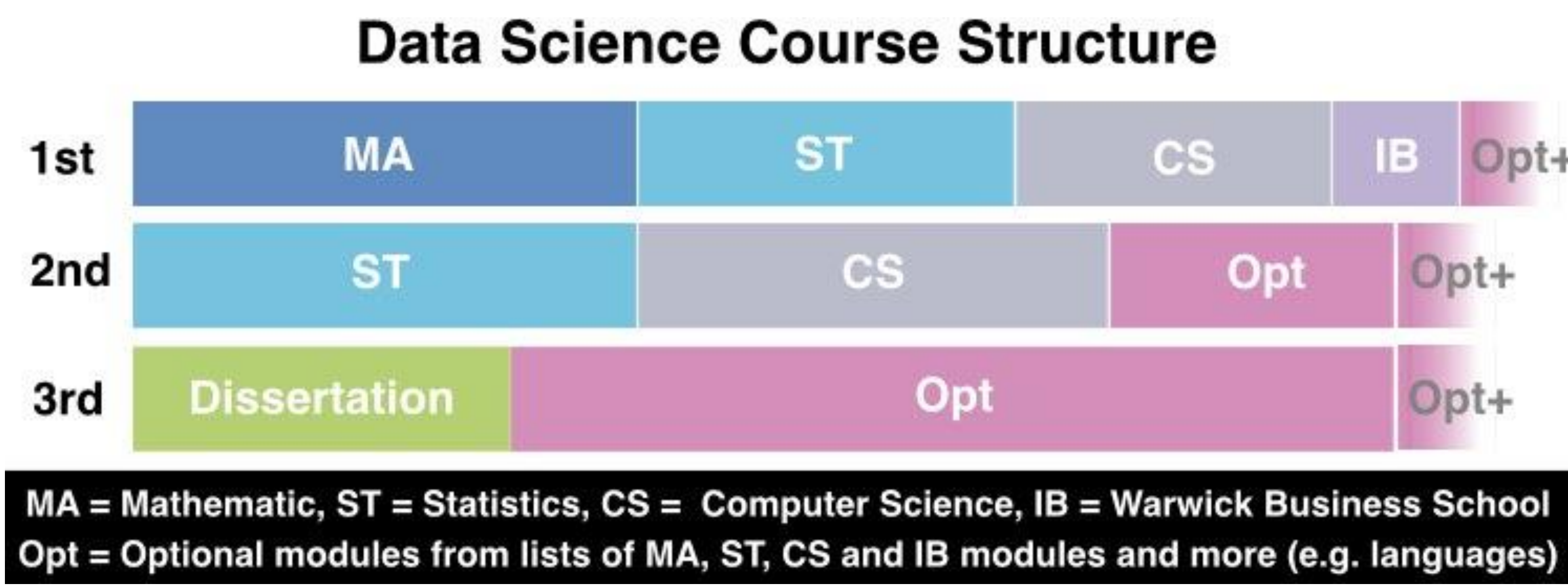
Data Science activities in the UK (cont.)

- 154 universities in the UK
- 35 universities offer Data Science related BSc courses for 2017 (30 universities for 2016) The Universities and Colleges Admissions Service (UCAS)
- 50 universities now offer Data Science related MSc courses degrees

More and more!

These courses are **jointly** offered by a group of departments: computer science, statistics, mathematics, subject matter disciplines.

BSc (Data Science)



Data Science Course Structure

1st	MA	ST	CS	IB	Opt+
2nd	ST	CS	Opt	Opt+	
3rd	Dissertation	Opt	Opt+		

MA = Mathematic, ST = Statistics, CS = Computer Science, IB = Warwick Business School
Opt = Optional modules from lists of MA, ST, CS and IB modules and more (e.g. languages)

1st year

Strong, general mathematical foundation. Programming, Data Structures, Probability and Exploratory Data Analysis.

2nd year

Statistical topics in considerable depth, Algorithms, Databases, Software Engineering. Optional modules: Artificial Intelligence, Linear Statistical Modelling etc.

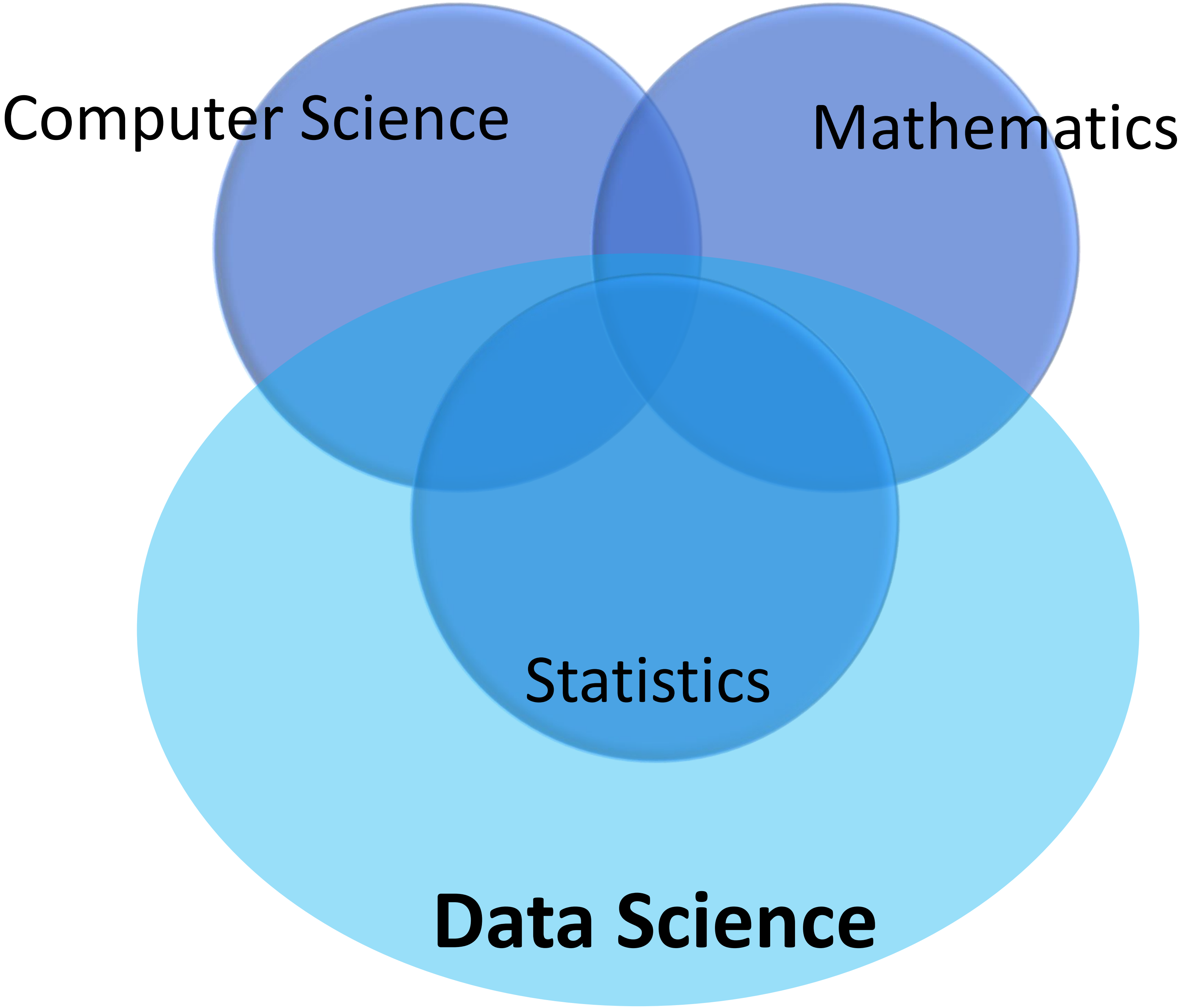
3rd year

Data Science Project; Optional modules: Machine Learning, Bayesian Forecasting etc.

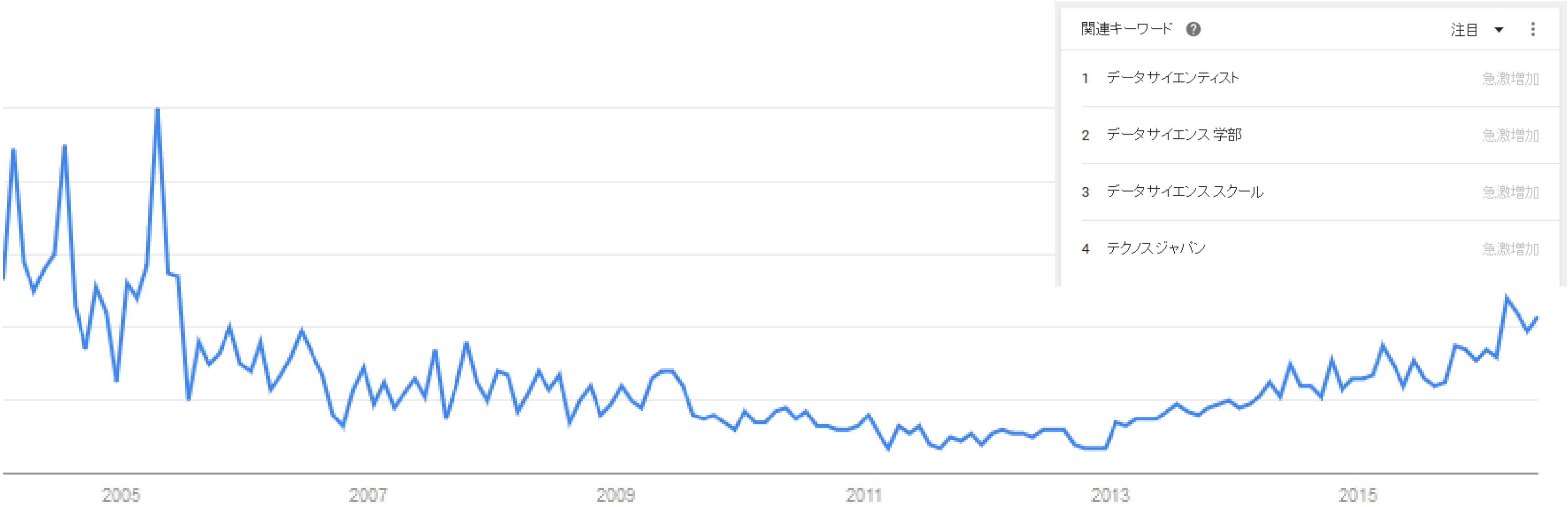
Lancaster University

MSc (Data Science) for 1 year taught course

- Data Mining
- Data Science Fundamentals
- Programming for Data Scientists
- Statistical Inference
- Statistical Methods and Modelling
- Likelihood Inference
- Generalised Linear Models
- Elements of Distributed Systems
- Systems Architecture and Integration
- Applied Data Mining


Computer Science
Mathematics
Statistics
Data Science

“データサイエンス” <- “Data Science” in Japanese



関連キーワード ?　　注目 ▼

1	データサイエンティスト	急激増加
2	データサイエンス 学部	急激増加
3	データサイエンス スクール	急激増加
4	テクノスジャパン	急激増加

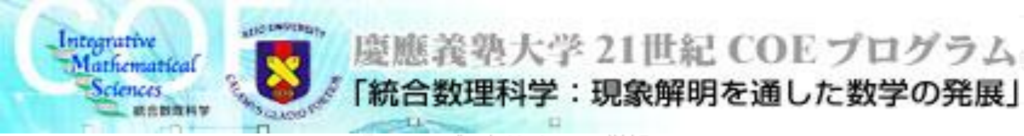
Google Trends

“データサイエンス” <- “Data Science” in Japanese

1996 第64回日本統計学会 (JSS)
共通テーマ 「データサイエンス I, II, III」

2013 International Conference (RSS)
Contributed Session: Data Science

2013 Joint Statistical Meeting (ASA)
Speaking Clearly About Data Scientists



慶應義塾大学 21世紀 COE プログラム
「統合数理科学：現象解明を通した数学の発展」

データサイエンスによる現象の数理(2003-)

Cherry Bud Workshop series

- *Data Science* and System Reduction (2004)
- Quantitative Risk Management (2005)
- Building Models from Data (2006)
- Interaction through Data (2007)
- Discovery through *Data Science* (2008)

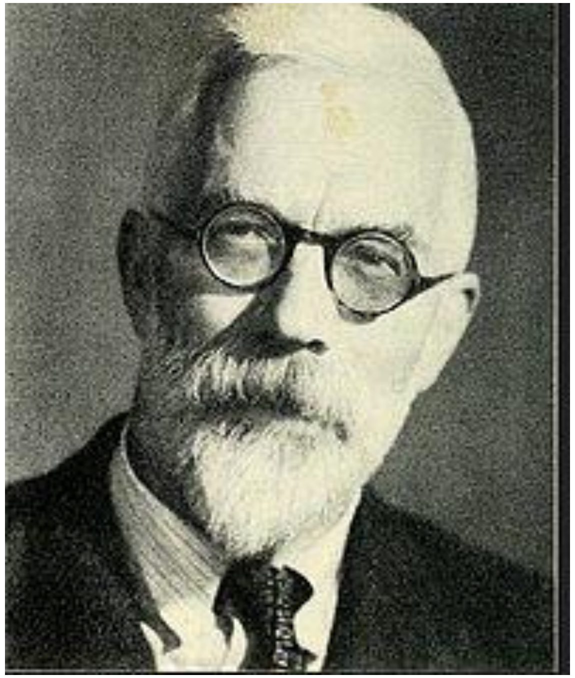
Google Trends

温故知新
Dicipulus est prioris posterior dies—Today is the scholar of yesterday.

Sir Ronald Fisher

Analyses of the vast amount of data accumulated from the "Classical Field Experiments“.

- 1919 Rothamsted Research
- 1921 *Studies in Crop Variation*
- 1925 *Statistical Methods for Research Workers*



John Tukey

The future of data analysis (Tukey 1962)

“For a long time I have thought I was a statistician ... I have had cause to wonder and to doubt. ... 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.”

Statistics is not enough to cover things!!

John Tukey (cont.)

The future of data analysis (Tukey 1962)

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 in an understandable form,

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

Mathematics cannot be a science but Data Analysis can be a new science!

-> Data Science ought to be ***science***!!

- 赤池 (1998) 時系列解析の方法
- 柴田 (2000) データサイエンスのすすめ
- 柴田 (2001) データリテラシー
- 林(2001) データの科学
- J. Tukey (1962) The future of Data Analysis
- T. Speed (1986) Questions, answers and Statistics
- C. Wu (1997) “Statistics = Data Science?"
- W. Cleveland (2001) Data science: an action plan for expanding the technical areas of the field of statistics
- P. Diggle (2015) Statistics: a data science for the 21st century
- ASA (2015) ASA Statement on The Role of Statistics in Data Science
- D. Donoho (2015) 50 years of Data Science

赤池(1998) データを用いて必要な情報を作り出すこと．モデルは仮説の表現複雑であり，基本的な知的活動．これによりデータに意味が生じ，情報が創造される．

林(2001) 複雑であいまいな現象について，データを中心に据えて物を見ていこうとするものである．これがデータの科学の根本理念である．

柴田 (1984-)
データの上流から下流まで

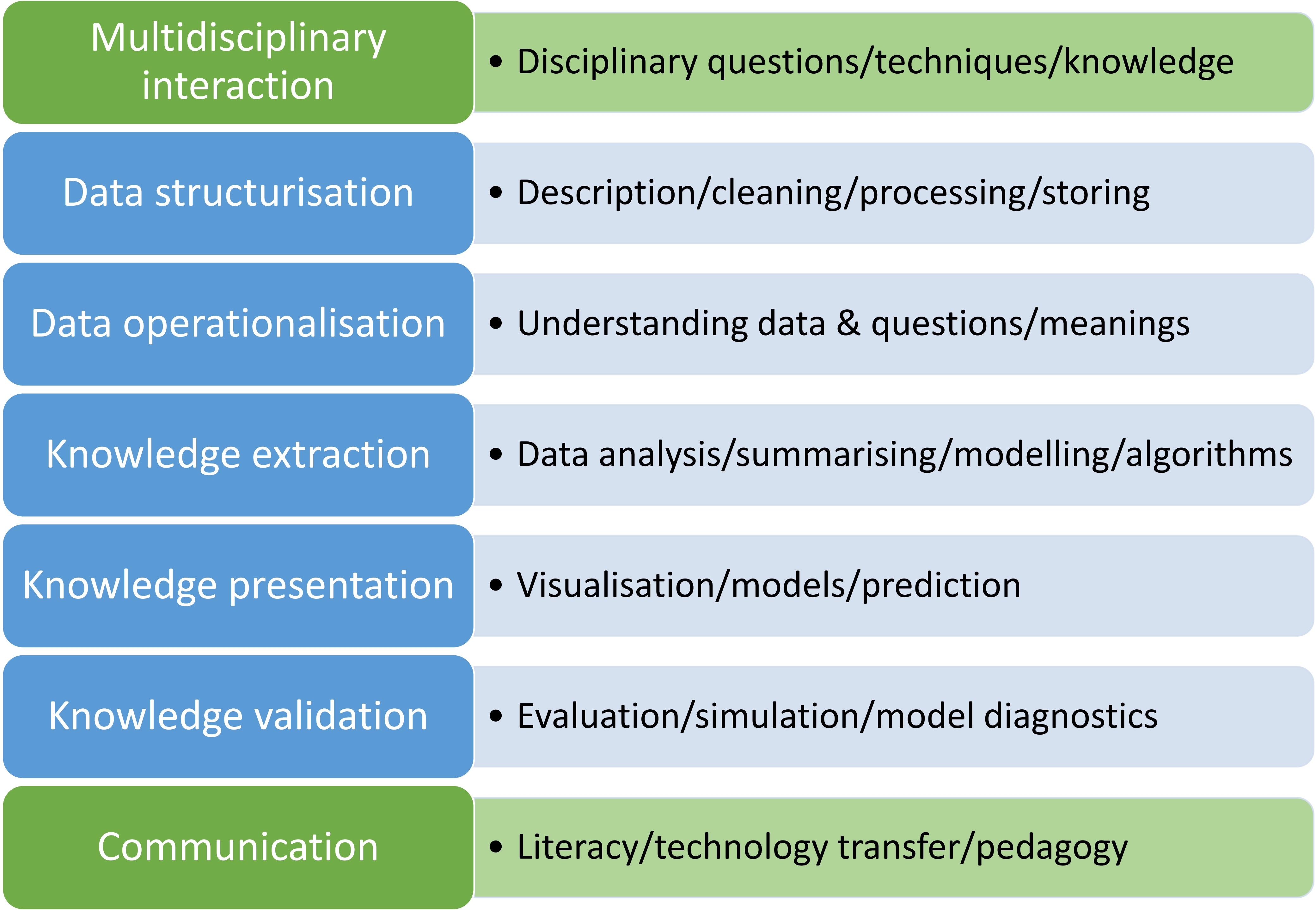
- Data collection
- Data description/clearing
- Data browsing/visualisation
- Data modelling
- Model validation

Donoho (2015)

- Data Exportation and Preparation
- Data Representation and Transformation
- Computing with Data
- Data Modelling
- Data visualisation and presentation
- Science about Data Science

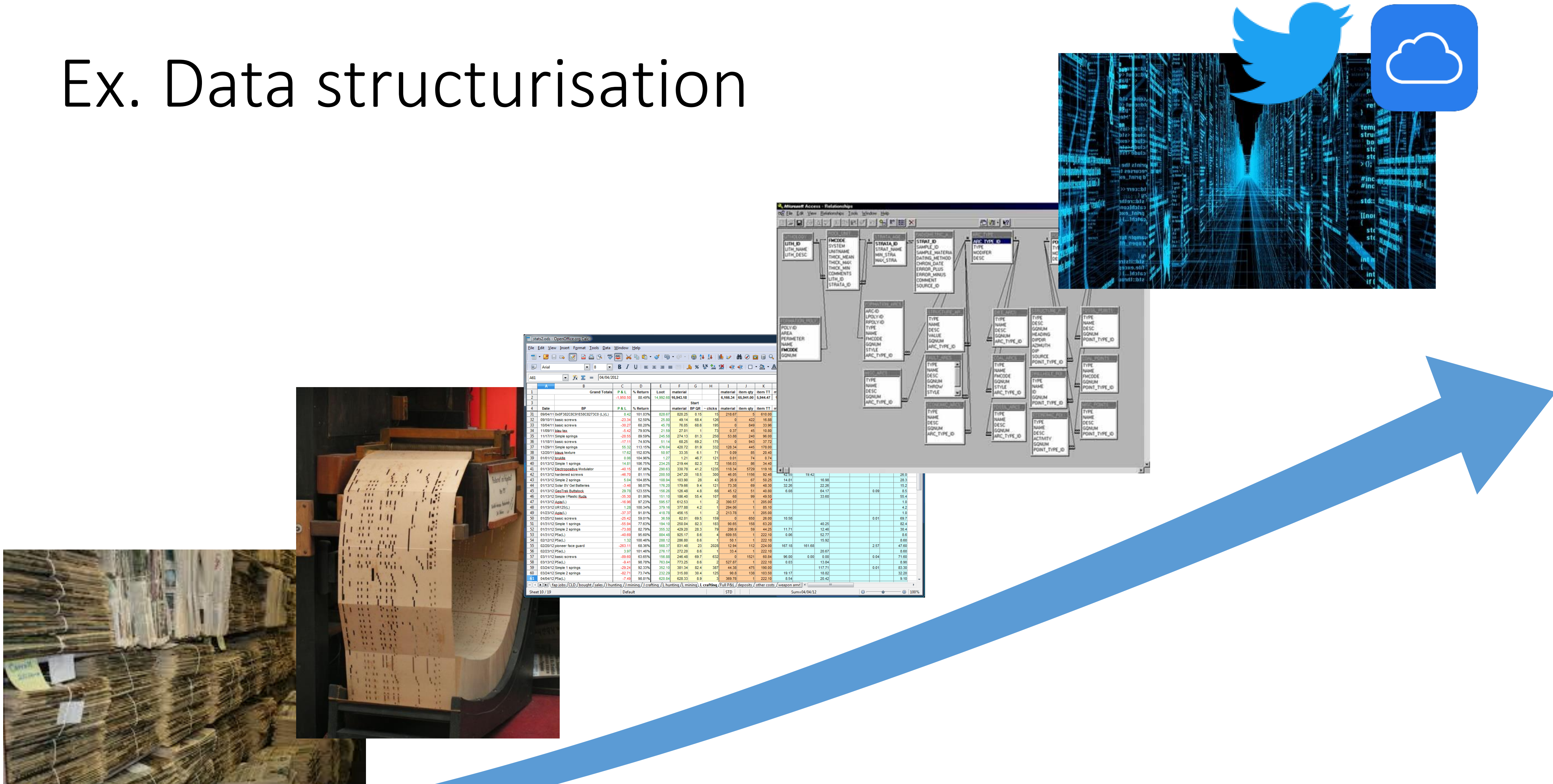
Cleveland (2001)

- Multidisciplinary investigation (25%)
- Models and methods for data (20%)
- Computing with data (15%)
- Pedagogy (15%)
- Tool evaluation (5%)
- Theory (20%)



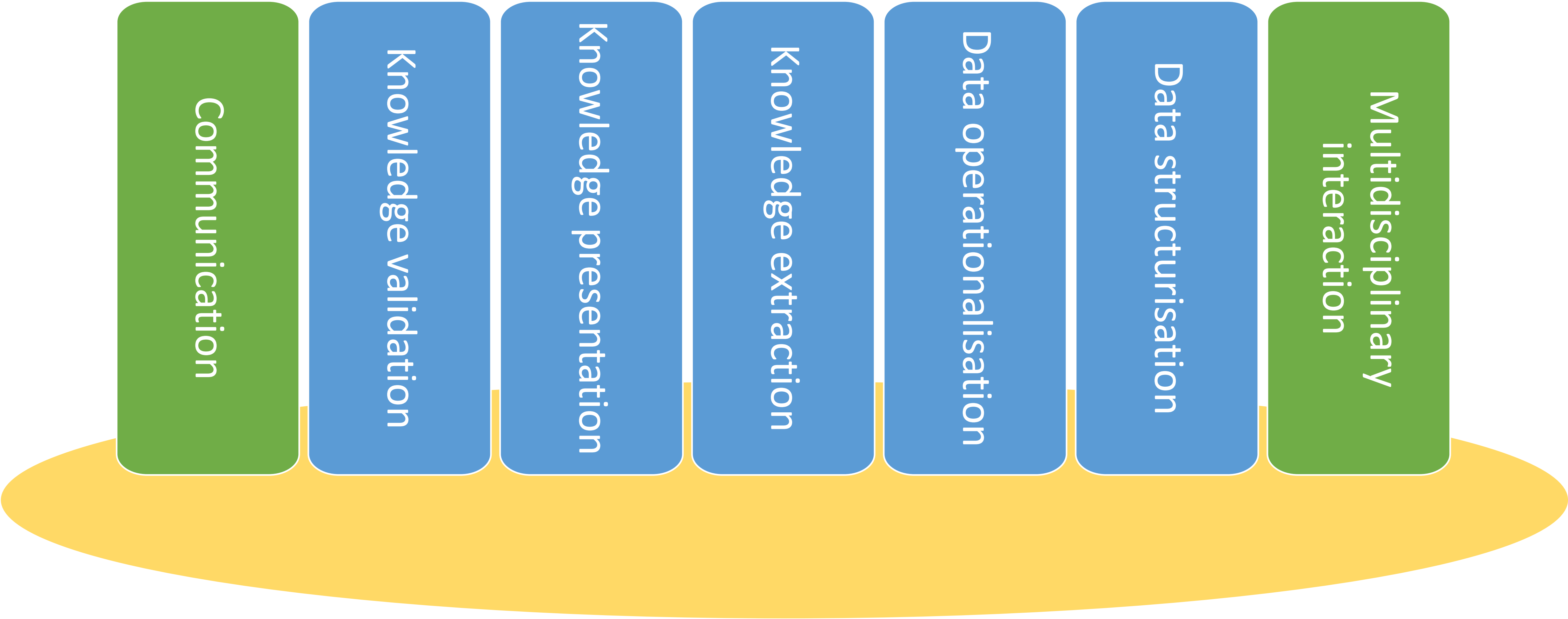
Multidisciplinary interaction	• Disciplinary questions/techniques/knowledge
Data structurisation	• Description/cleaning/processing/storing
Data operationalisation	• Understanding data & questions/meanings
Knowledge extraction	• Data analysis/summarising/modelling/algorithms
Knowledge presentation	• Visualisation/models/prediction
Knowledge validation	• Evaluation/simulation/model diagnostics
Communication	• Literacy/technology transfer/pedagogy

Ex. Data structurisation



Each pillar has expanded itself over time!

The seven pillars of Data Science



Communication

Knowledge validation

Knowledge presentation

Knowledge extraction

Data operationalisation

Data structurisation

Multidisciplinary interaction

Thank you very much for your attention!

Any questions/comments welcomed.

Hideyasu SHIMADZU