Data Science/Analytical Thinking for Enterprise Innovations

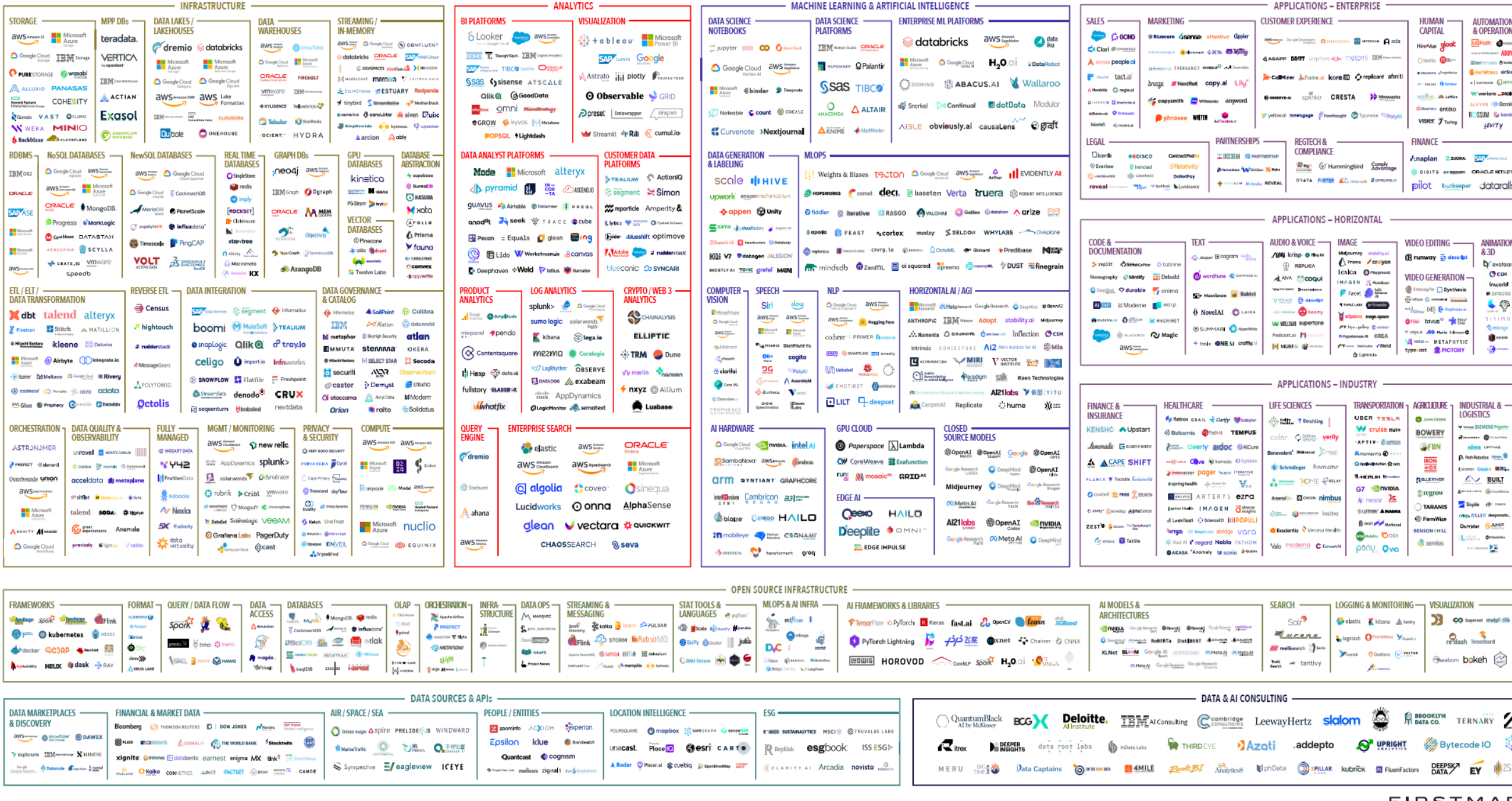
Longbing Cao | www.datasciences.org

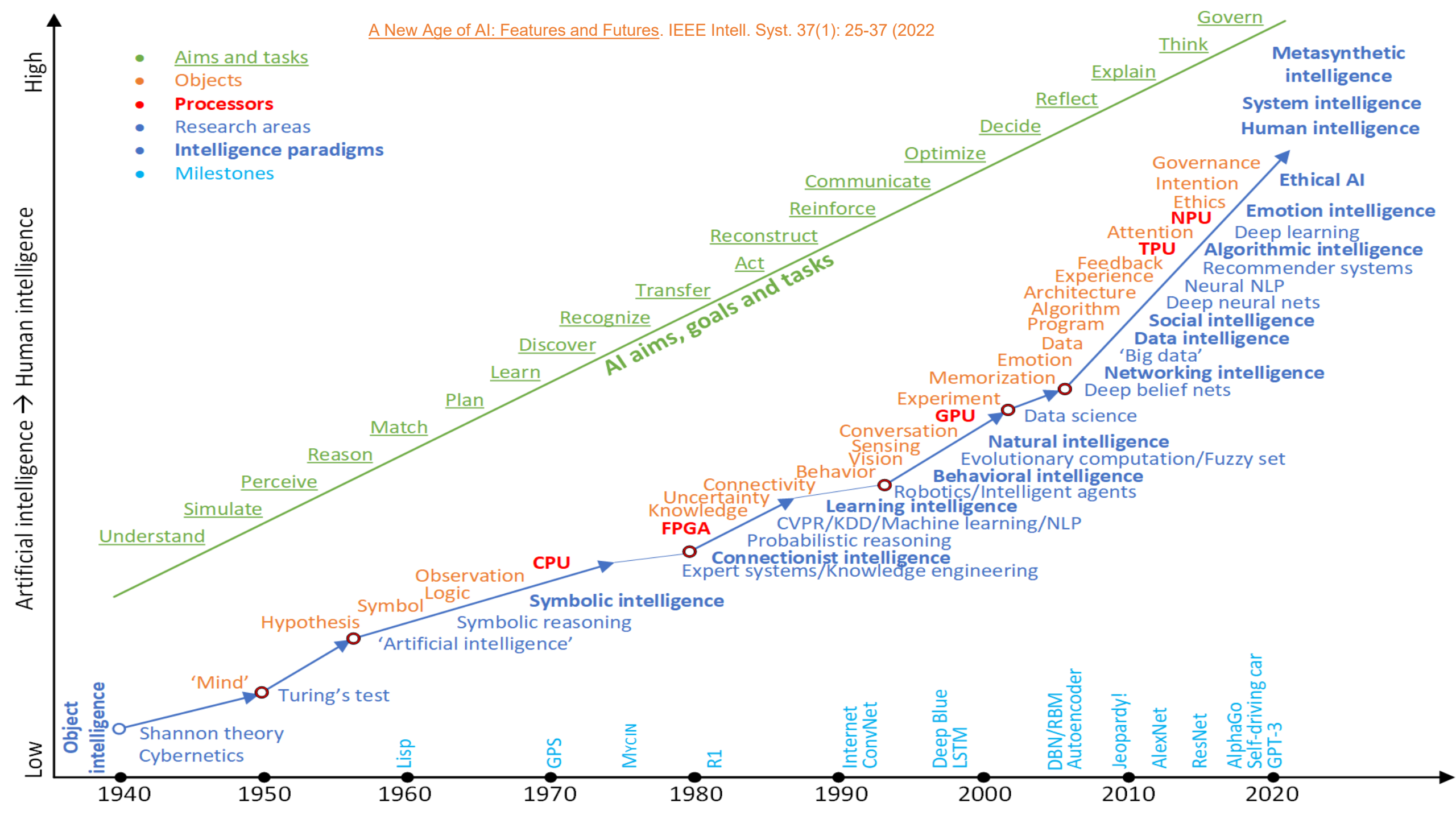
Declaration

- Please be noted not every cited material has been properly acknowledged in this presentation
- The author acknowledges the wisdom and contributions of all relevant contributors

- More information about the content discussed in this presentation is in the book: Data Science Thinking

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE




A New Age of AI: Features and Futures. IEEE Intell. Syst. 37(1): 25-37 (2022
Aims and tasks
Objects
Processors
Research areas
Intelligence paradigms
Milestones
Artificial intelligence → Human intelligence
High
Low
AI aims, goals and tasks
Understand
Simulate
Perceive
Reason
Match
Plan
Learn
Discover
Recognize
Transfer
Act
Reconstruct
Reinforce
Communicate
Optimize
Decide
Reflect
Explain
Think
Govern
Object intelligence
'Mind'
Turing's test
Shannon theory
Cybernetics
Hypothesis
Symbol
Logic
Observation
'Artificial intelligence'
Symbolic reasoning
Symbolic intelligence
CPU
FPGA
Knowledge
Uncertainty
Connectivity
Expert systems/Knowledge engineering
Connectionist intelligence
Probabilistic reasoning
CVPR/KDD/Machine learning/NLP
Learning intelligence
Behavior
Vision
Sensing
Conversation
Robotics/Intelligent agents
Behavioral intelligence
Experiment
GPU
Memorization
Emotion
Evolutionary computation/Fuzzy set
Natural intelligence
Data science
Deep belief nets
Networking intelligence
'Big data'
Data intelligence
Program
Algorithm
Architecture
Experience
Feedback
Deep neural nets
Social intelligence
Neural NLP
Attention
TPU
Recommender systems
Ethics
Intention
NPU
Deep learning
Governance
Algorithmic intelligence
Ethical AI
Emotion intelligence
Metasynthetic intelligence
System intelligence
Human intelligence
Lisp
GPS
Mycin
R1
Internet
ConvNet
Deep Blue
LSTM
DBN/RBM
Autoencoder
Jeopardy!
AlexNet
ResNet
AlphaGo
Self-driving car
GPT-3
1940
1950
1960
1970
1980
1990
2000
2010
2020

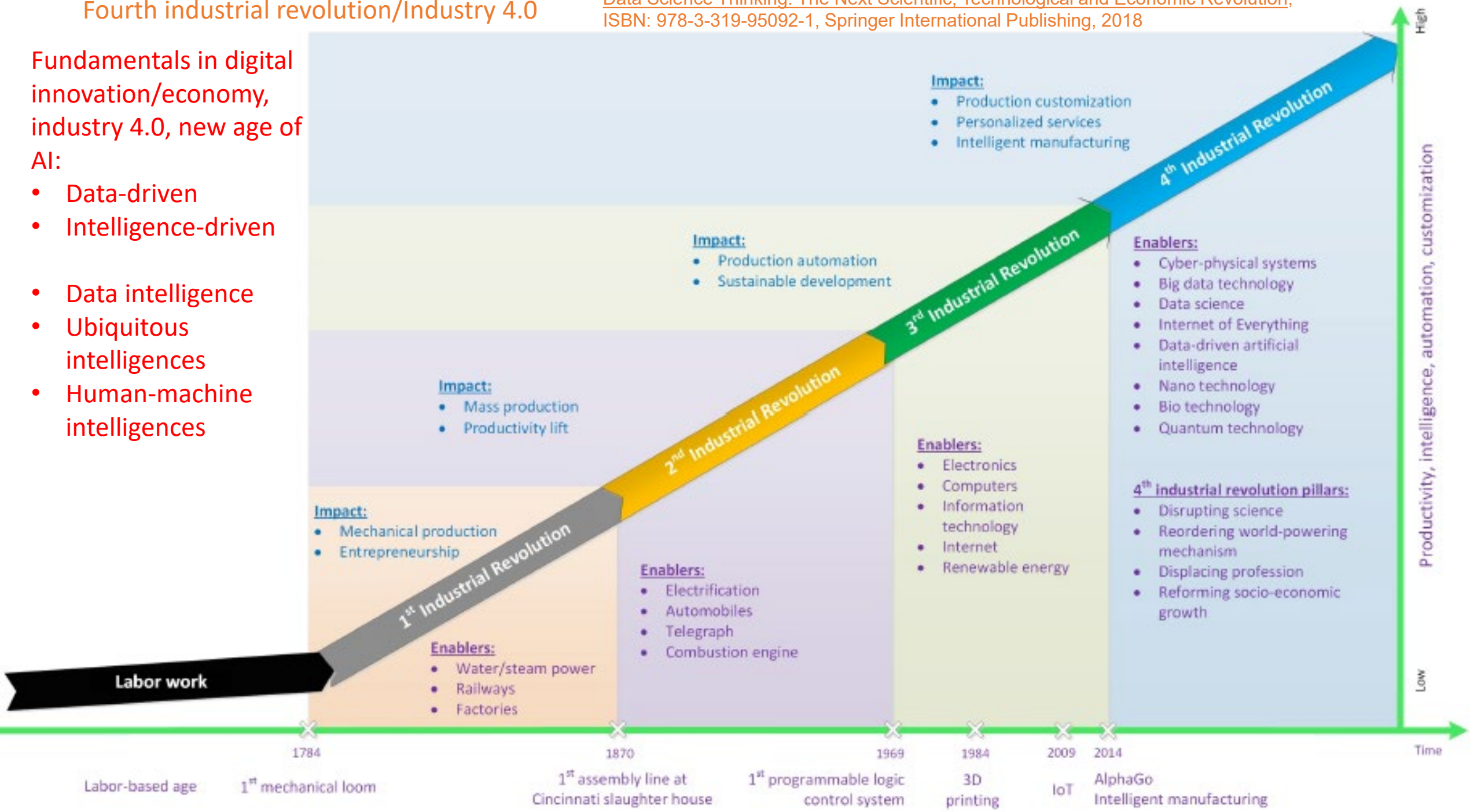
Fourth industrial revolution/Industry 4.0

Data Science Thinking: The Next Scientific, Technological and Economic Revolution,
ISBN: 978-3-319-95092-1, Springer International Publishing, 2018

Fundamentals in digital innovation/economy, industry 4.0, new age of AI:

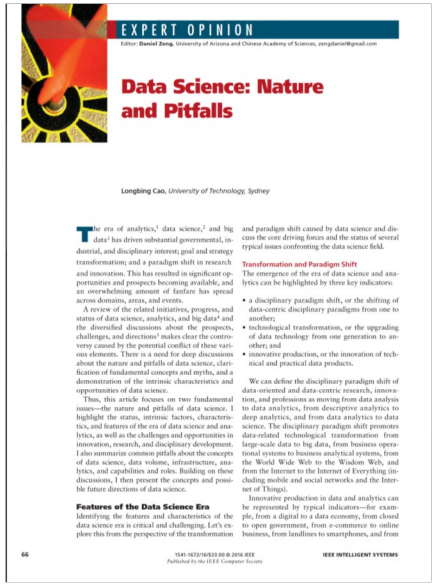
- Data-driven
- Intelligence-driven

- Data intelligence
- Ubiquitous intelligences
- Human-machine intelligences



Data Science Concepts

50 Years of data science: an immature discipline


EXPERT OPINION

Editor: Daniel Zeng, University of Arizona and Chinese Academy of Sciences, zengdaniel@gmail.com

Data Science: Nature and Pitfalls

Longbing Cao, *University of Technology, Sydney*

The era of analytics,[1] data science,[2] and big data[3] has driven substantial governmental, industrial, and disciplinary interest; goal and strategy transformation; and a paradigm shift in research and innovation. This has resulted in significant opportunities and prospects becoming available, and an overwhelming amount of fanfare has spread across domains, areas, and events.

A review of the related initiatives, progress, and status of data science, analytics, and big data[4] and the diversified discussions about the prospects, challenges, and directions[5] makes clear the controversy caused by the potential conflict of these various elements. There is a need for deep discussions about the nature and pitfalls of data science, clarification of fundamental concepts and myths, and a demonstration of the intrinsic characteristics and opportunities of data science.

Thus, this article focuses on two fundamental issues—the nature and pitfalls of data science. I highlight the status, intrinsic factors, characteristics, and features of the era of data science and analytics, as well as the challenges and opportunities in innovation, research, and disciplinary development. I also summarize common pitfalls about the concepts of data science, data volume, infrastructure, analytics, and capabilities and roles. Building on these discussions, I then present the concepts and possible future directions of data science.

Features of the Data Science Era

Identifying the features and characteristics of the data science era is critical and challenging. Let's explore this from the perspective of the transformation and paradigm shift caused by data science and discuss the core driving forces and the status of several typical issues confronting the data science field.

Transformation and Paradigm Shift

The emergence of the era of data science and analytics can be highlighted by three key indicators:

- a disciplinary paradigm shift, or the shifting of data-centric disciplinary paradigms from one to another;
- technological transformation, or the upgrading of data technology from one generation to another; and
- innovative production, or the innovation of technical and practical data products.

We can define the disciplinary paradigm shift of data-oriented and data-centric research, innovation, and professions as moving from data analysis to data analytics, from descriptive analytics to deep analytics, and from data analytics to data science. The disciplinary paradigm shift promotes data-related technological transformation from large-scale data to big data, from business operational systems to business analytical systems, from the World Wide Web to the Wisdom Web, and from the Internet to the Internet of Everything (including mobile and social networks and the Internet of Things).

Innovative production in data and analytics can be represented by typical indicators—for example, from a digital to a data economy, from closed to open government, from e-commerce to online business, from landlines to smartphones, and from

66 1541-1672/16$33.00 © 2016 IEEE Published by the IEEE Computer Society IEEE INTELLIGENT SYSTEMS

L. Cao. IEEE Intelligent Systems, 2016


DOI:10.1145/3015456

While it may not be possible to build a data brain identical to a human, data science can still aspire to imaginative machine thinking.

BY LONGBING CAO

Data Science: Challenges and Directions

WHILE DATA SCIENCE has emerged as an ambitious new scientific field, related debates and discussions have sought to address why science in general needs data science and what even makes data science a science. However, few such discussions concern the intrinsic complexities and intelligence in data science problems and the gaps in and opportunities for data science research. Following the literature review,[5,6,10–13,15,18] I offer a number of observations concerning big data and the data science debate. For example, discussion has covered not only data-related disciplines and domains like statistics, computing, and informatics but traditionally less data-related fields and areas like social science and business management as well. Data science has thus emerged as a new inter- and cross-disciplinary field. Although many publications are available, most (likely over 95%) concern existing concepts and topics in statistics, data mining, machine learning, and broad data analytics. This limited view demonstrates how data science has emerged from existing core disciplines, particularly statistics, computing, and informatics. The abuse, misuse, and overuse of the term "data science" is ubiquitous, contributing to the hype, and myths and pitfalls are common.[5] While specific challenges have been covered,[5,18] few scholars have addressed the low-level complexities and problematic nature of data science or contributed deep insight about the intrinsic challenges, directions, and opportunities of data science as an emerging field.

Data science promises new opportunities for scientific research, addressing, say, "What can I do now but could not do before, as when processing large-scale data?"; "What did I do before that does not work now, as in methods that view

key insights

- **Data science problems require systematic thinking, methodologies, and approaches to help spur development of machine intelligence.**
- **The conceptual landscape of data science assists data scientists trying to understand, represent, and synthesize the complexities and intelligence in related problems.**
- **Data scientists aim to invent data- and intelligence-driven machines to represent, learn, simulate, reinforce, and transfer human-like intuition, imagination, curiosity, and creative thinking through human-data interaction and cooperation.**

AUGUST 2017 | VOL. 60 | NO. 8 | COMMUNICATIONS OF THE ACM 59

L. Cao. Communications of the ACM, 2017


Data Science: Profession and Education

Longbing Cao
Advanced Analytics Institute, University of Technology Sydney, Australia

Keywords—Data science, advanced analytics, big data analytics, data science profession, data science courses, data science education

Advanced analytics, data science, and new-generation artificial intelligence (AI) represent probably the most promising areas and directions in today's and near-future's Information and Communications Technology (ICT) and Science, Engineering and Technology (SET) sectors and disciplines, where data science has become the major driving force of the new-generation AI. Data science and new-generation AI have attracted increasing interest from major governments, vendors and academia, with important initiatives launched by major countries such as the United States [1], China [2], and the European Commission [3].

However, despite the fact that the role of data scientists has been described as the sexiest job in the 21st century [4], [5], the qualifications and capabilities of a data scientist are not clearly defined; it is important yet undermined to define what makes the next-generation data scientists who can transform today's and future science, technology, innovation and economy [6].

On one hand, an increasing number of data science roles and courses are created by online course offers and traditional institutions [7], data science has clearly formed a new profession [8], [9], [10]. On the other, high-end data science employers (including major innovators, vendors, and enterprises) complain the limited availability of qualified data scientists to enable their strategic development and foster their future competitive advantages. These disclose the deviated status of existing professional and educational markets, the restricted benchmarking and accreditation of responsibilities and capabilities of data scientists, and the urgent need of standardizing and upgrading competencies and maturity of data science qualifications and education. This article intends to address these important issues, with an aim to contributing to the standardization and formalization of next-generation data science profession and education.

I. Data Science As A Profession

Data science has driven the emergence of a new profession: *data science profession*, or simply *data profession* [6]. Typical evidence for this data profession formation includes the increasing number of clearly divided, diversified, well-defined and predominant data-oriented roles and responsibilities, the increasingly clearly articulated and pursued qualifications, the increasingly clarified competency and capabilities of these roles, and their increasingly profound impact on driving and transforming the data research, innovation, economy and society.

A. Spectrum of Data Roles

Data roles broadly refer to job positions and responsibilities that are centered on, related to, or enabled by data and data-oriented (including data-driven, data-enabled, and data-based) tasks and agenda. Specifically, *data science roles* are those centered on data-driven discovery, innovation and practices. They have been built on the profession of software engineering, database administration, business intelligence, computing, enterprise application integration, business analysis, and statistical analysis.

There are various perspectives to structure and categorize data roles. In conducting enterprise data science, an enterprise data science team is often formed with different roles for fulfilling scientific and engineering tasks. Typically, a data science team consists of data roles that fulfill the responsibilities and tasks for undertaking (1) the *infrastructure* for data computing, analytics, and decision-support; (2) the engineering of data, software, applications, networking, communications, analytics, reporting, and decision-support; (3) the *discovery* for data-driven scientific exploration, modeling and optimization, including general and specific analytics, mining, learning, recognition, prediction and refinement; (4) the *decision strategy* for designing, implementing and evaluating data-driven decisions, strategies, and actions; (5) the *leadership* for planning, overseeing and governing hierarchical (e.g., from enterprise-level to groups and teams in the enterprise) visions, missions, plans and strategies; and (6) the *management* for administering data, resources, tasks, processes, and risk etc.

In Fig. 1, various roles and positions are listed that address the above areas of responsibilities in a corporate data science group [6]. Typical data science roles that differ from existing IT professionals are data planning strategist, analytical architect, data modeler, subject-specific analyst (e.g., behavior analyst, financial analyst, and visual analyst), discovery scientist, model operator, quality assurer, and decision strategist, etc. Interested readers are encouraged to read references such as [6] for more discussion on the various job positions, and in particular Chapter 10 Data Profession in [6] for discussion on the responsibilities associated with each of the above areas and those in Fig. 1.

B. Data Scientists and Engineers

Among the various data roles, two major ones are data scientists and data engineers. They constitute a collaborative

L. Cao. IEEE Intelligent Systems, 2019

D. Donoho, "50 Years of Data Science," 2015; http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf


1

Data Science: A Comprehensive Overview

LONGBING CAO, University of Technology Sydney, Australia

The twenty-first century has ushered in the age of big data and data economy, in which data DNA, which carries important knowledge, insights and potential, has become an intrinsic constituent of all data-based organisms. An appropriate understanding of data DNA and its organisms relies on the new field of data science and its keystone, analytics. Although it is widely debated whether big data is only hype and buzz, and data science is still in a very early phase, significant challenges and opportunities are emerging or have been inspired by the research, innovation, business, profession, and education of data science. This paper provides a comprehensive survey and tutorial of the fundamental aspects of data science: the evolution from data analysis to data science, the data science concepts, a big picture of the era of data science, the major challenges and directions in data innovation, the nature of data analytics, new industrialization and service opportunities in the data economy, the profession and competency of data education, and the future of data science. This article is the first in the field to draw a comprehensive big picture, in addition to offering rich observations, lessons and thinking about data science and analytics.

Additional Key Words and Phrases: Big data, Data Analysis, Data Analytics, Advanced Analytics, Big Data Analytics, Data Science, Data Engineering, Data Scientist, Statistics, Computing, Informatics

ACM Reference Format:
Longbing Cao, 2016. Data Science: A Comprehensive Overview. *Submitted to ACM Computing Surveys for Review* 1, 1, Article 1 (January 2016), 42 pages.
DOI: 0000001.0000001

1. INTRODUCTION

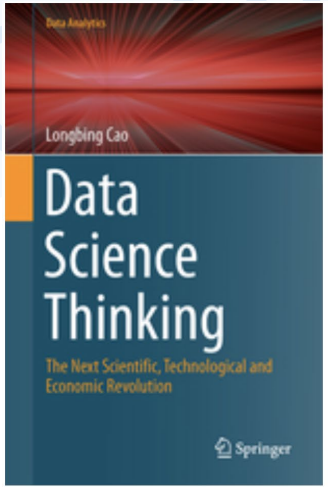
We are living in the age of big data, advanced analytics, and data science. The trend of "big data growth" [Laney 2001; CSC 2012; Beyer and Laney 2012; McKinsey 2011; Vesset et al. 2012], or "data deluge" [Hey and Trefethen 2003], has not only triggered tremendous hype and buzz, but more importantly presents enormous challenges which in turn bring incredible innovation and economic opportunities. Big data has attracted intensive and growing attention, initially from giant private data-oriented enterprise and lately from major governmental organizations and academic institutions. Typical examples include large data-centric projects in Google, Facebook and IBM, and strategic initiatives in the United Nations [UN 2010; USNSF 2012], EU [Commission 2014] and China [CN 2015].

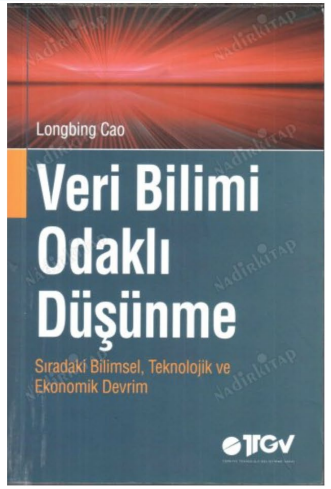
From the disciplinary development perspective, recognition of the significant challenges, opportunities and values of big data is fundamentally reshaping the traditional data-oriented scientific and engineering fields. It is also reshaping those non-traditional data engineering domains such as social science, business and management [Yiu 2012; Labrinidis and Jagadish 2012; Chen et al. 2012; Khan et al. 2014]. This reshaping and paradigm shifting is driven not just by data itself but all other aspects that could be created, transformed and/or adjusted by understanding, exploring and utilizing data.

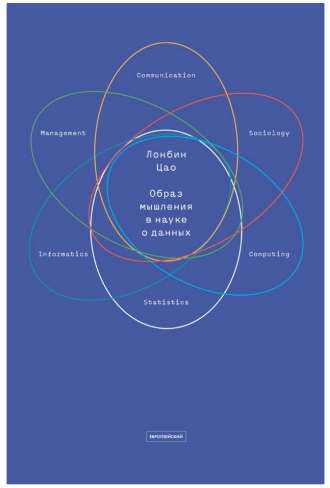
This work is partially supported by the Australian Research Council Discovery Grant, under grant DP130102691.
Author's addresses: L. Cao, Advanced Analytics Institute, University of Technology Sydney, Australia.

DOI: 0000001.0000001

L. Cao. ACM Computing Survey, 2017

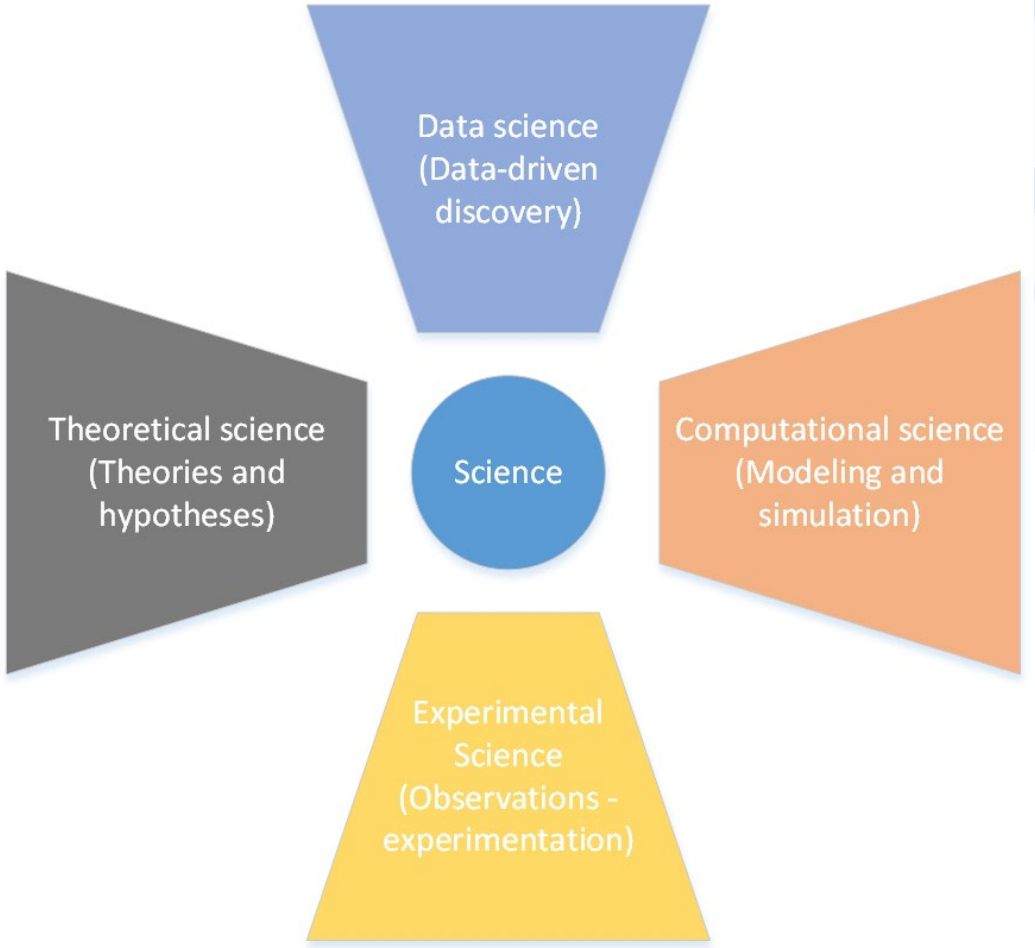






https://link.springer.com/content/pdf/10.1007%2F978-3-319-95092-1.pdf

Scientific paradigms – the era of data



Big data	Refers to data that are too large and/or complex to be effectively and/or efficiently handled by traditional data-related theories, technologies and tools.

What is data science?

Definition 2.2 (Data Science[1]). Data science is the science of data, or data science is the study of data.

Definition 2.3 (Data Science[2]). Data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, such as statistics, informatics, computing, communication, management and sociology, to study data and its domain employing data science thinking.

$$data\ science \overset{\text{def}}{=} \{statistics \cap informatics \cap computing \cap communication \cap sociology \cap management \mid data \cap domain \cap thinking\} \quad (2.1)$$

where "|" means "conditional on."

Data science: converting data to intelligence/wisdom

- Data intelligence
- Actionable intelligence
  - unique and valuable understanding, thinking, insights and expertise that can enable significantly better and smarter planning, decision-making and outcomes

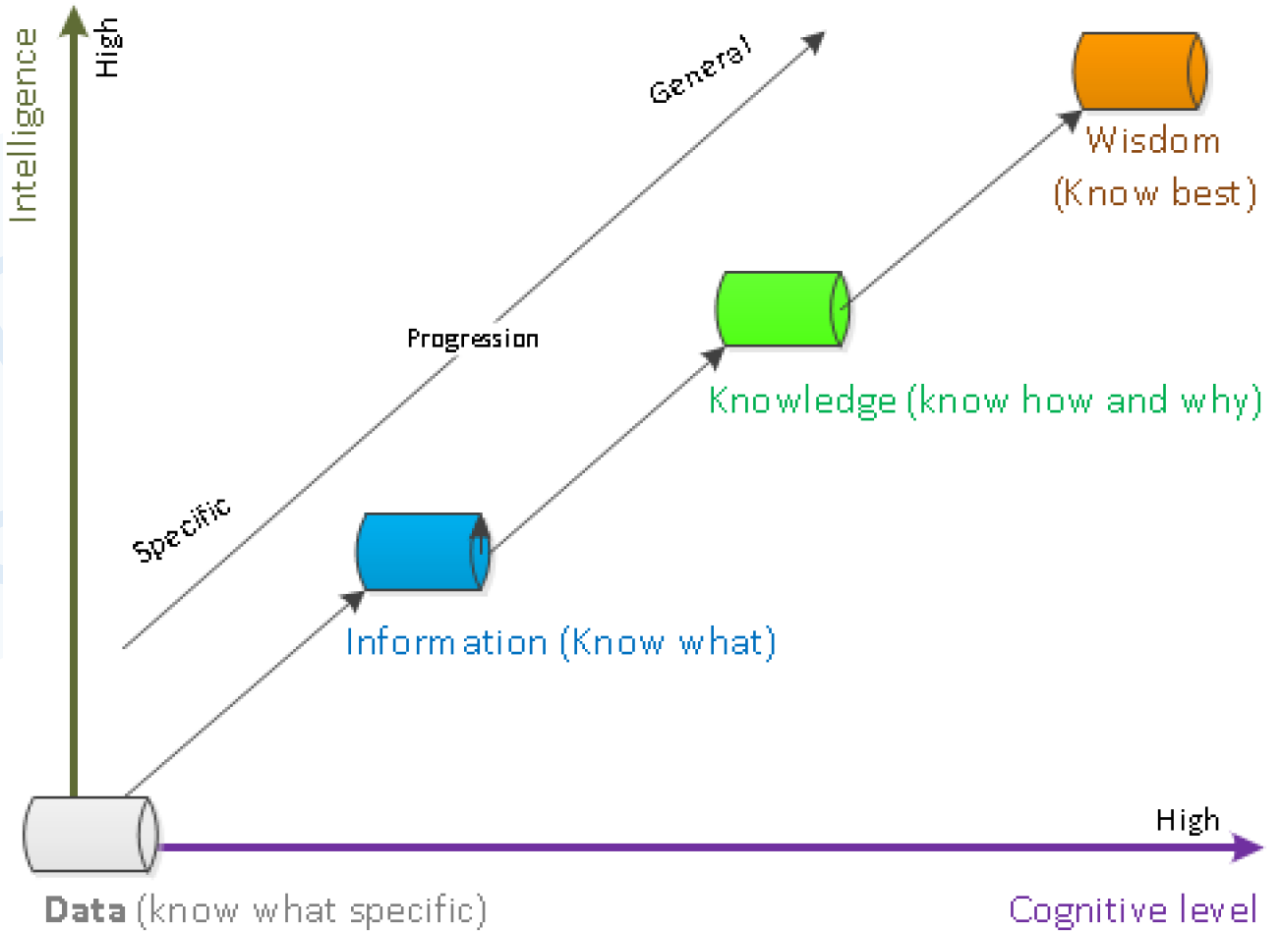
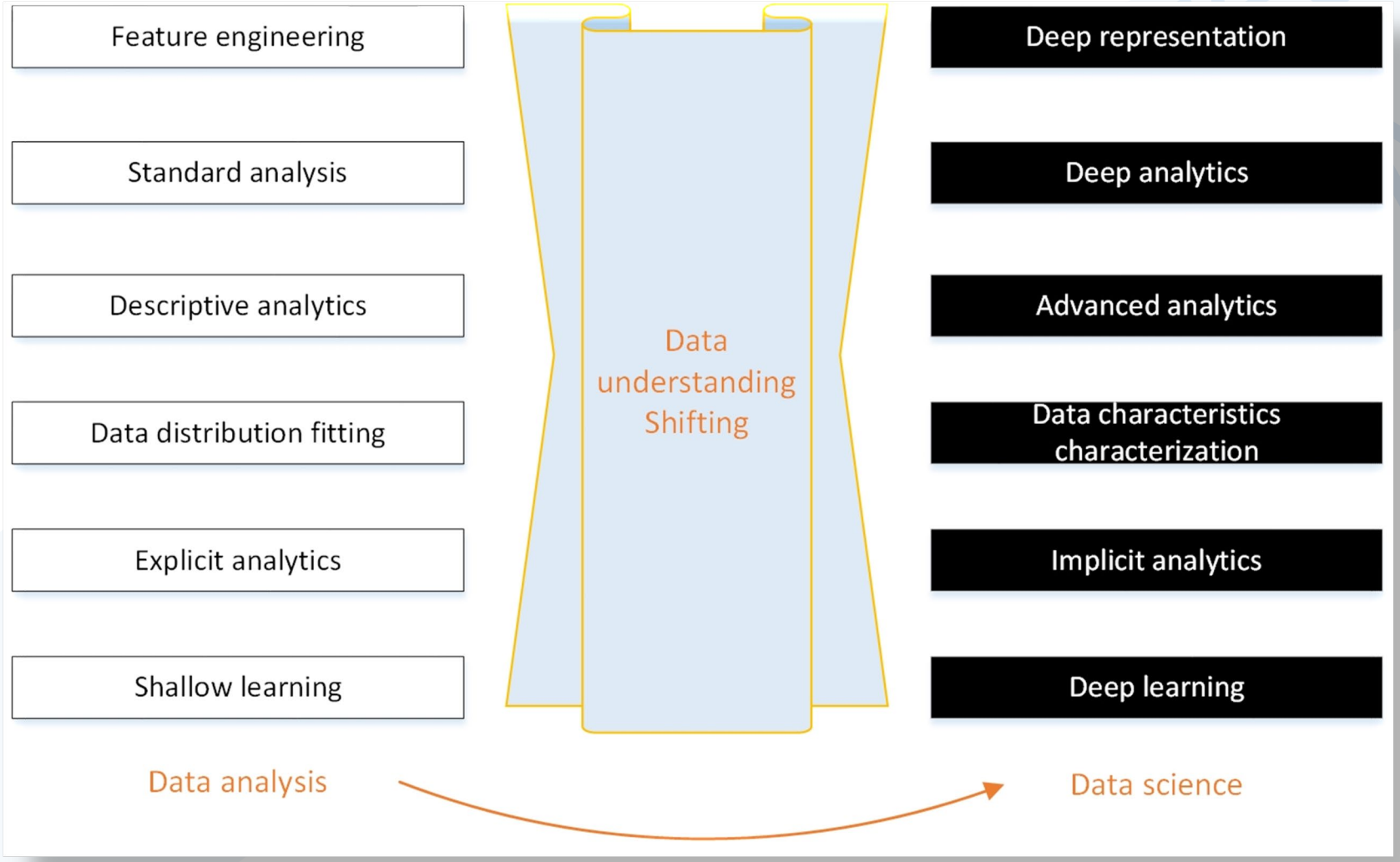


Fig. 2.1 Data-to-information-to-knowledge-to-intelligence-to-wisdom cognitive progression.
[2]Note: X-axis: the increase in cognitive level; Y-axis: the increase in intelligence

Data science thinking, Springer, 2018

Paradigm shift: Well-developed data analysis → Immature data science

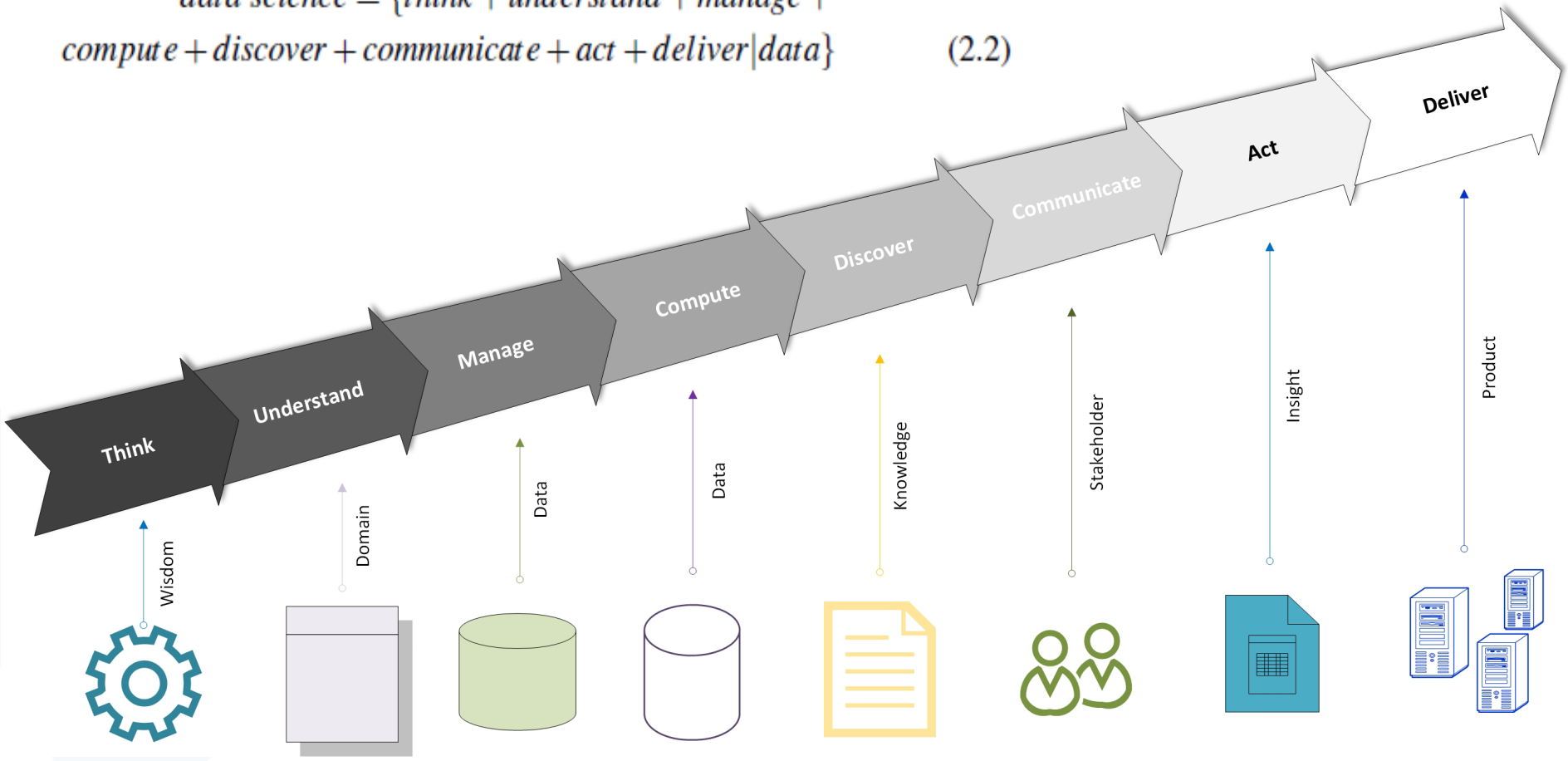


L. Cao. Data science thinking, Springer, 2018

Data science processes

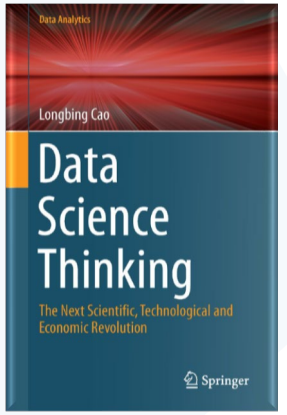
Definition 2.4 (Data Science[3]). From the *process* perspective, *data science* is a systematic approach to "thinking with wisdom", "understanding the domain", "managing data", "computing with data", "discovering knowledge", "communicating with stakeholders", "acting on insights", and "delivering products".

$$data\ science \overset{\text{def}}{=} \{think + understand + manage + compute + discover + communicate + act + deliver | data\} \quad (2.2)$$



- Not necessary 80:20 rule
- End-to-end solution
- Thinking, design, and actionability
- Embed data processing and feature engineering in learning

Data science/analytical thinking



Definition 3.1 (Data Science Thinking) Data science thinking refers to the perspective on the methodologies, process, structure, and traits and habits of the mind in handling data problems and systems.

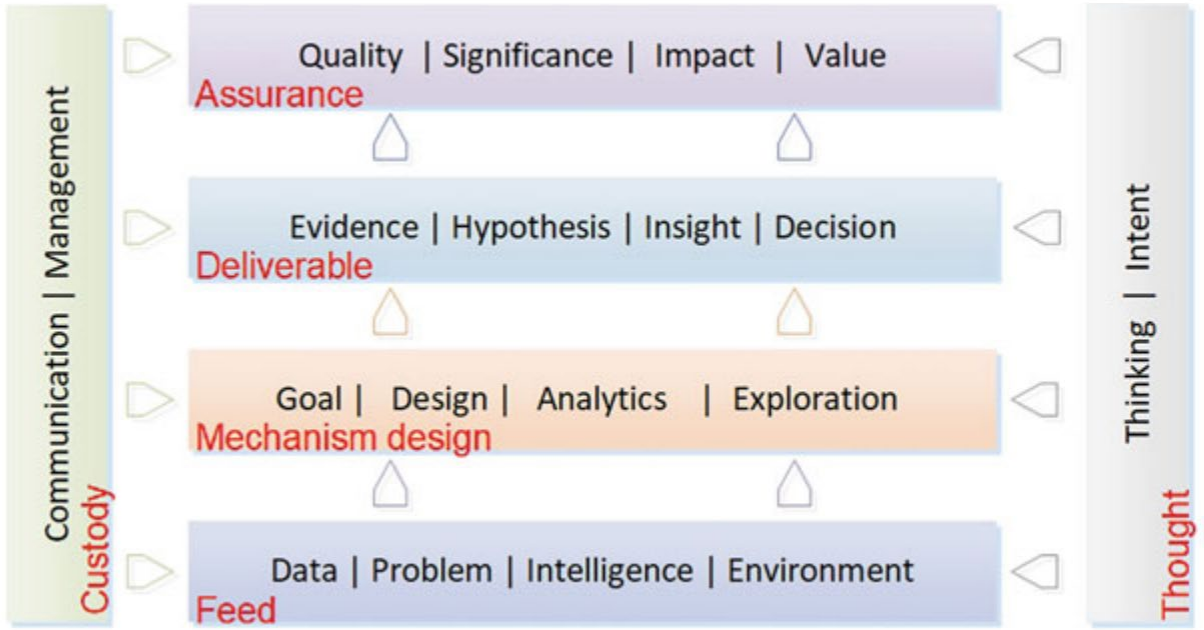


Fig. 3.5 Data science overview and structure

- Scientific thinking
- Data-oriented critical and creative thinking
- Individualized → holistic → systematic design
- From methodologies, processes, structures, designs to evaluation, deliverables
- Design thinking in analytics and learning paradigms and methods
- The best about what, how and why – insights and strategies

Gaps: complexities vs. capabilities/capacity

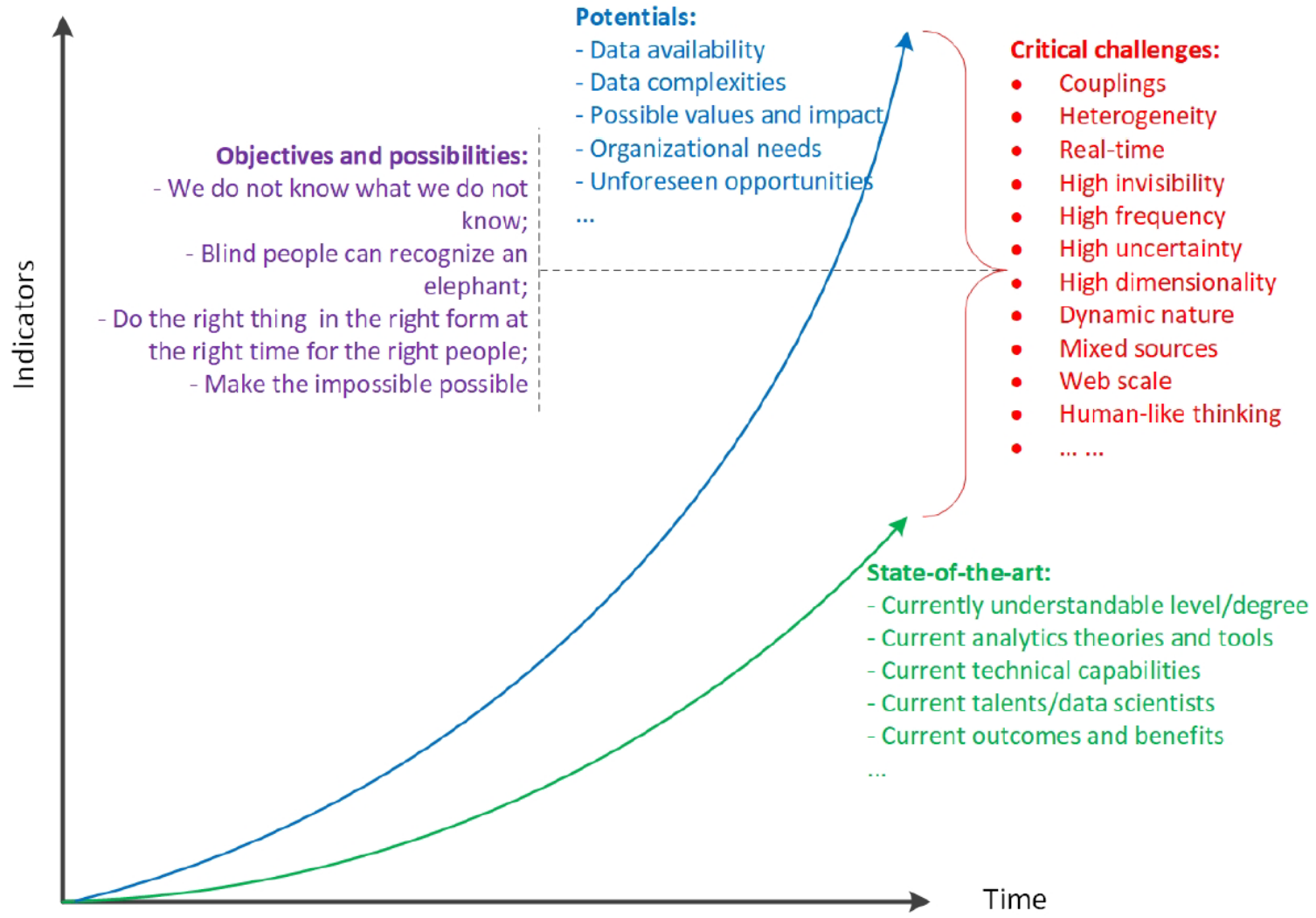


Fig. 5.1 Growing disciplinary gaps between data potential and disciplinary capabilities.

- Identify and explore the challenges and gaps
- Different designers deliver different results
- Data science thinking and capability make difference

Data science thinking, Springer, 2018

X complexities
X intelligences
X analytics

- E.g., deep learning on small, noisy, inconsistent, evolving case data
- Contextual factors: ethnic, social, cultural, persona, …
- Epidemiological knowledge on its transmission, …

Data science: Challenges and directions, Communications of the ACM, 2017

Data science thinking, Springer, 2018

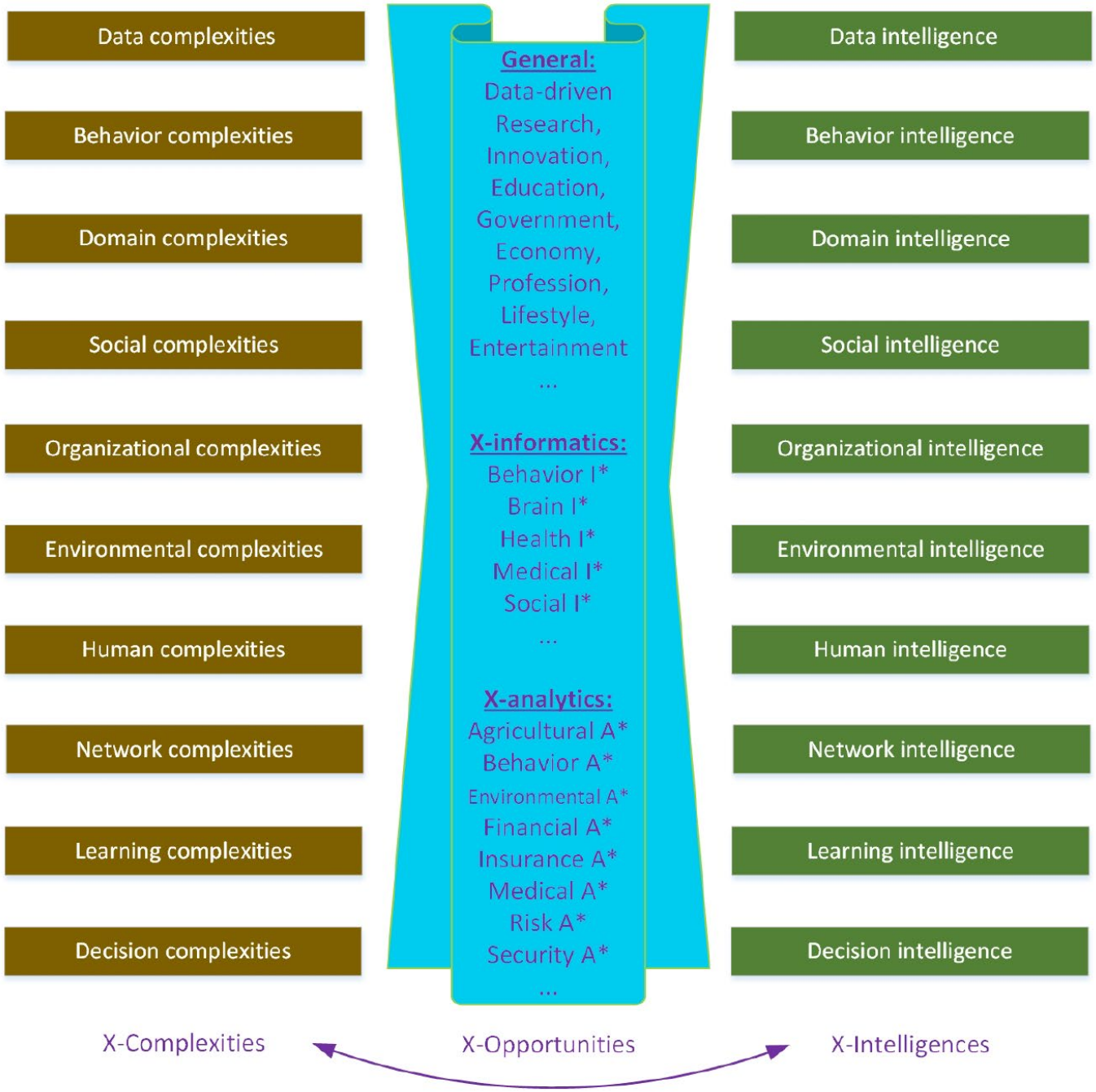
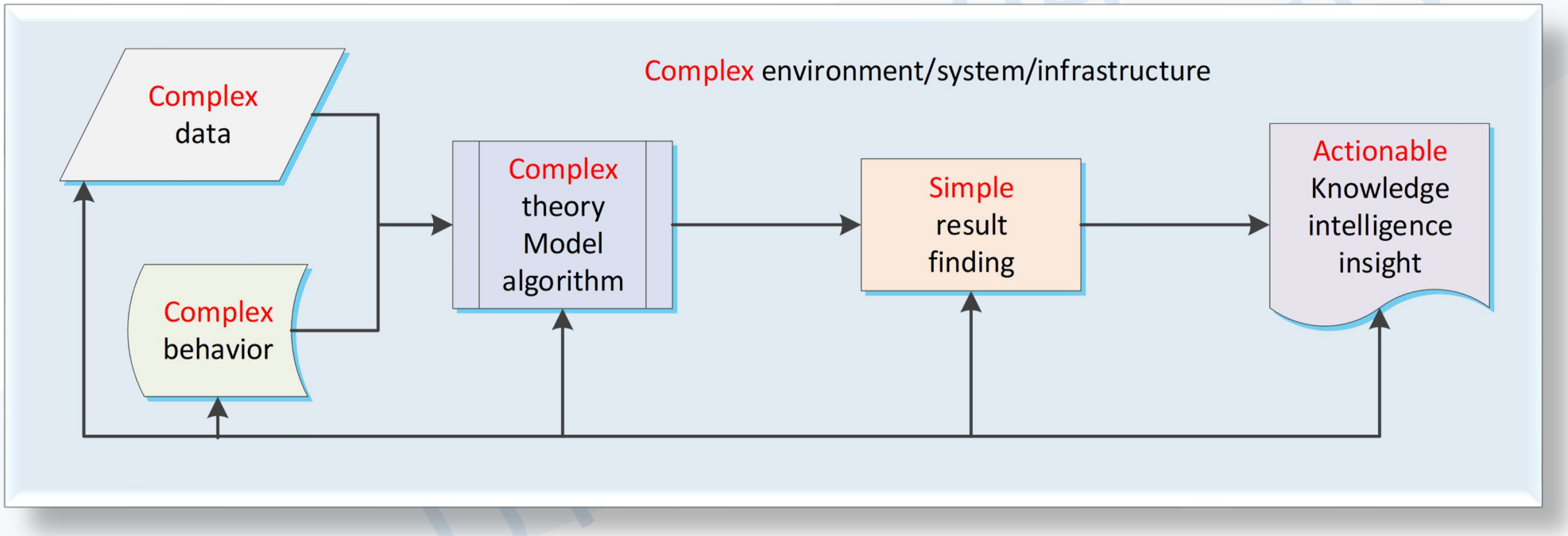


Fig. 1.3 The new X-generations: X-complexities, X-intelligence, and X-opportunities.

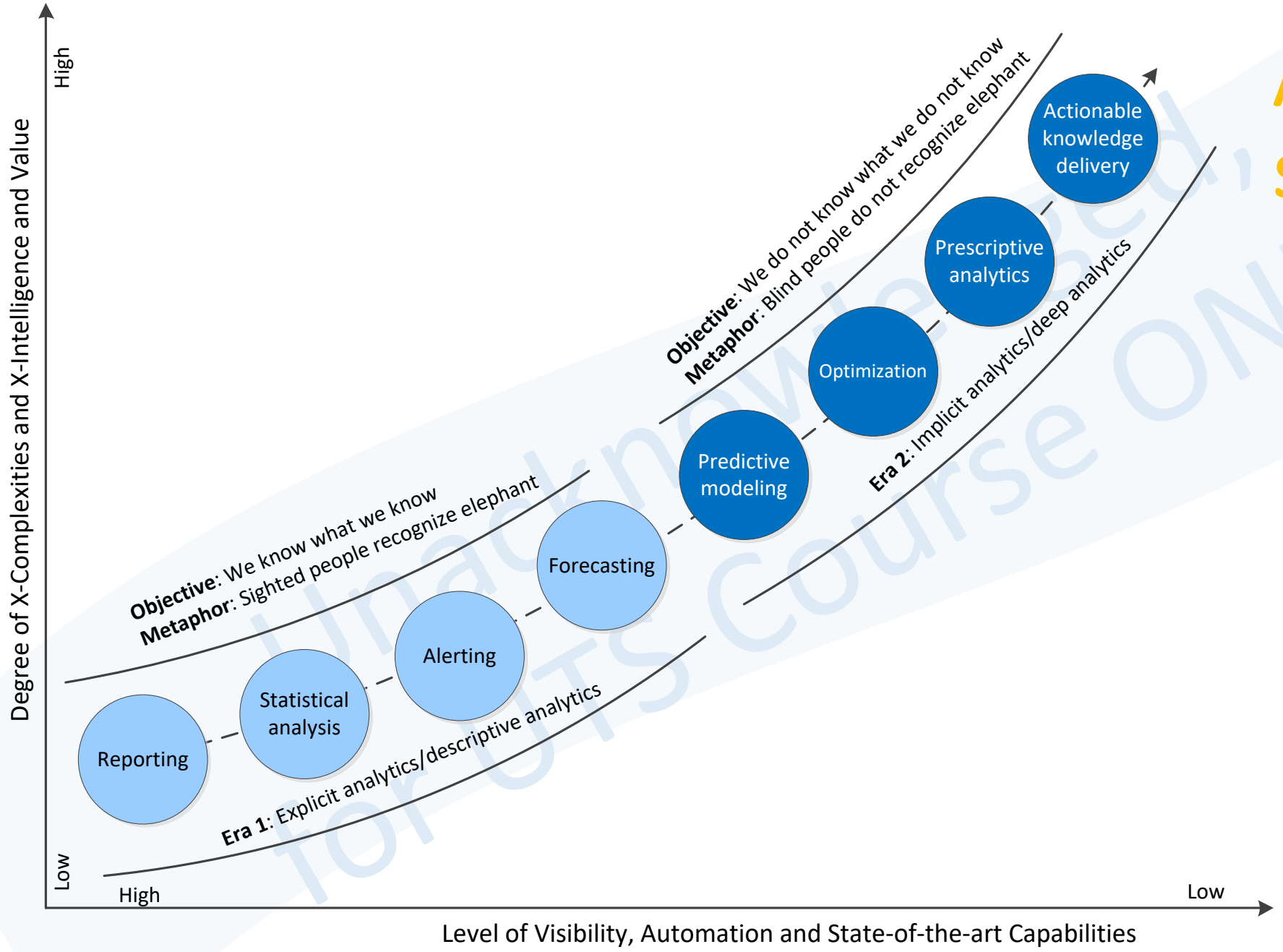
Complex real world
vs. often simple, specific solutions and results



- Enterprise data science is a complex process and system
- Understanding and quantifying complexities is a major task for new strategic and value proposition

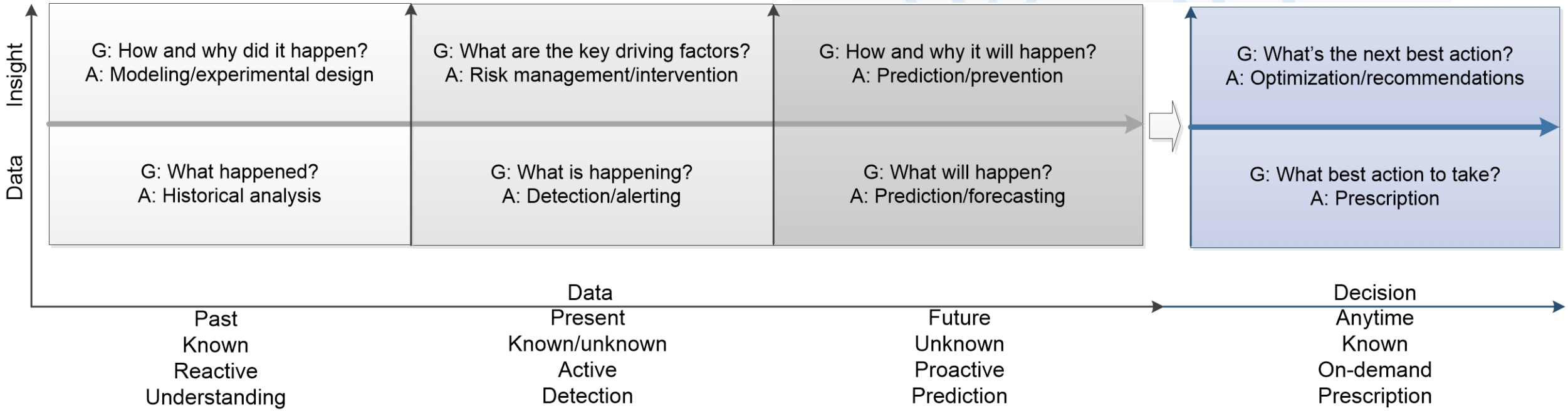
Enterprise Data Science

Analytics spectrum



- Present enterprise data science focus on explicit and descriptive analytics
- Enterprise data science shift to implicit and deep enterprise-wide analytics

Data →Insight/Decision



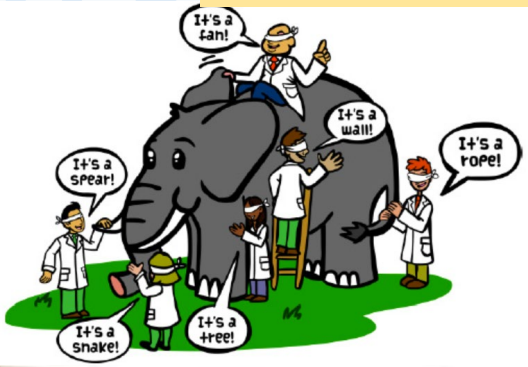
- What → why → how
- Lifelong: past → present → future
- Known → unknown → known →
- Reactive → active → proactive
- Understanding → detection → prediction → intervention

Data science thinking, Springer, 2018

Big data analytics era

Major effort:
Shallow learning
Descriptive analytics
Explicit analytics
Off-the-shelf analytics
…

Minor effort:
Deep analytics
Advanced analytics
Implicit analytics
Specialised analytics
…



Analytics Spectrum

Sighted people
recognize elephant

Blind people
recognize elephant

We do not know what we do not know: challenges, solutions, gaps, opportunities

Explore the unknowns

Fig. 3.7 Data science: The unknown world.

Data science: Challenges and directions, Communications of the ACM, 2017

Data science thinking, Springer, 2018

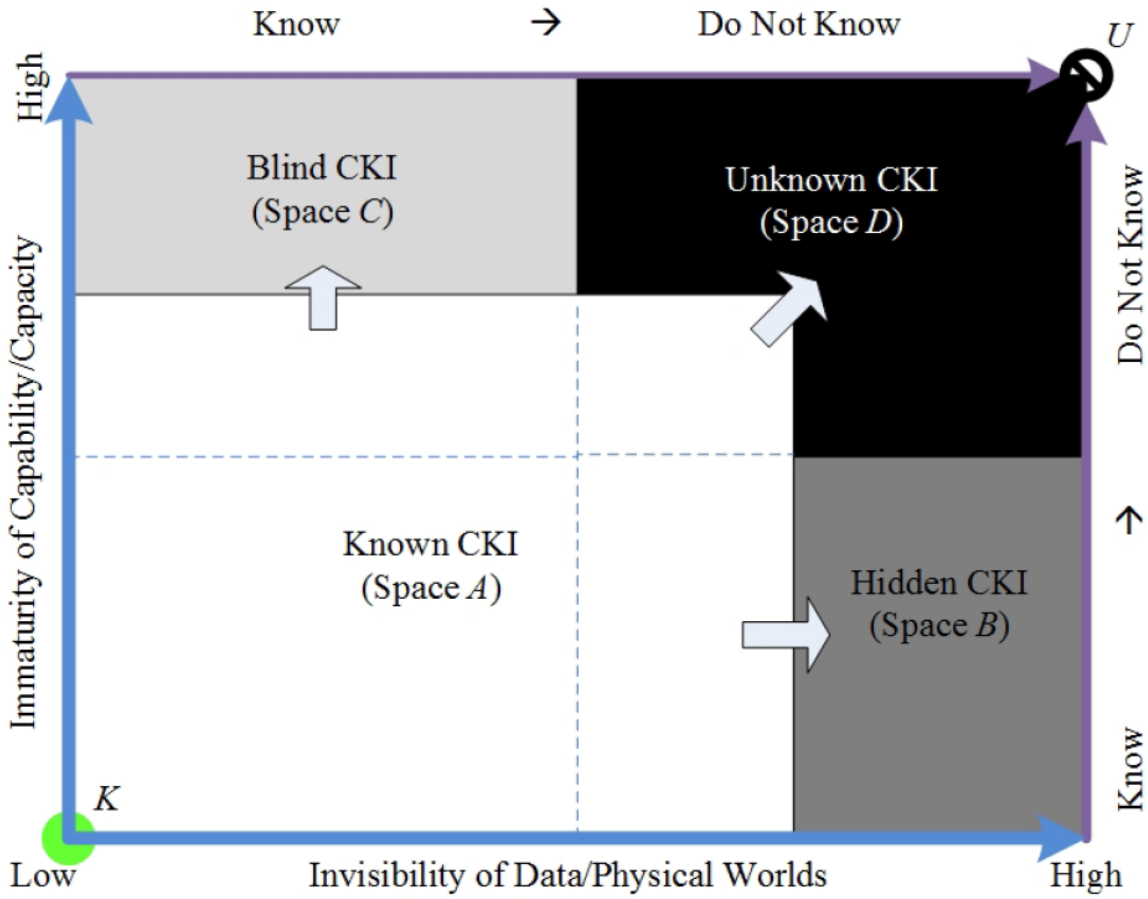
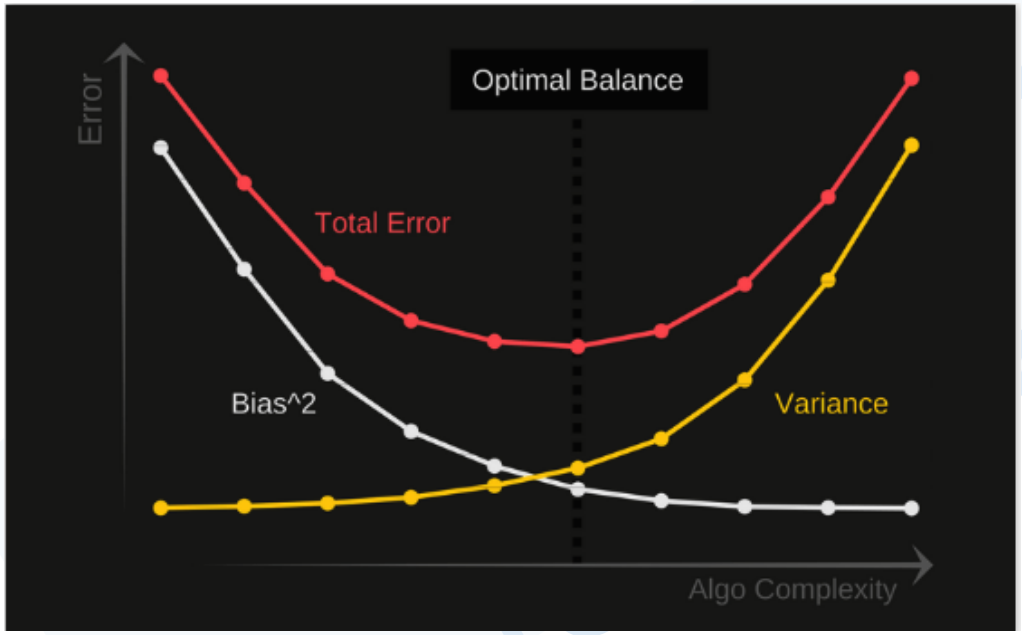


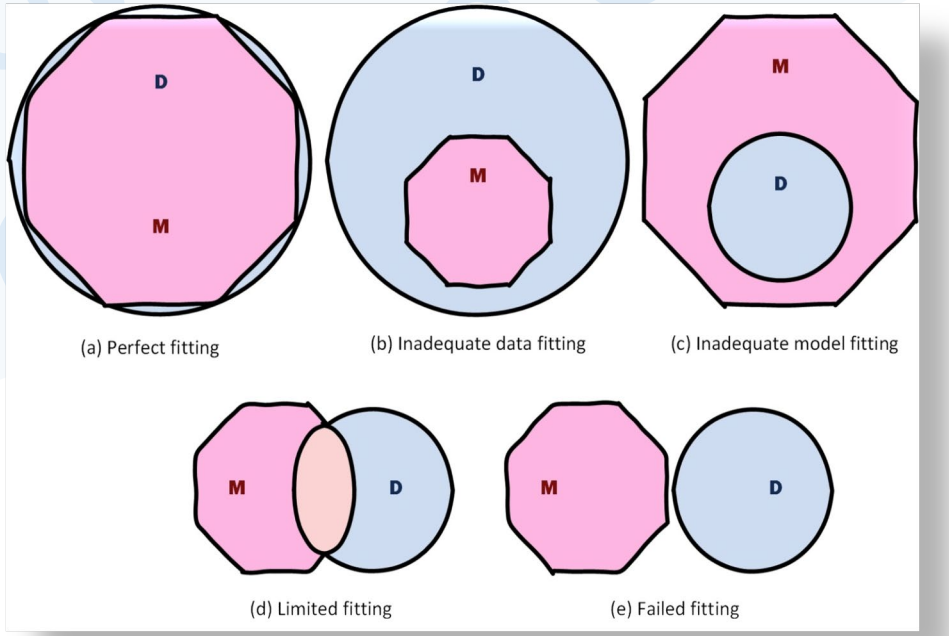
Fig. 3.6 Data science: Known-to-unknown research evolution.

Fitting gaps between model capacity vs. data potential

- Beyond fitting: data, model (parameter)
- Appropriate processes: sampling, validation etc.
- Appropriate thinking, theory, design, fitness, etc.



Total Error = Bias^2 + Variance + Irreducible Error



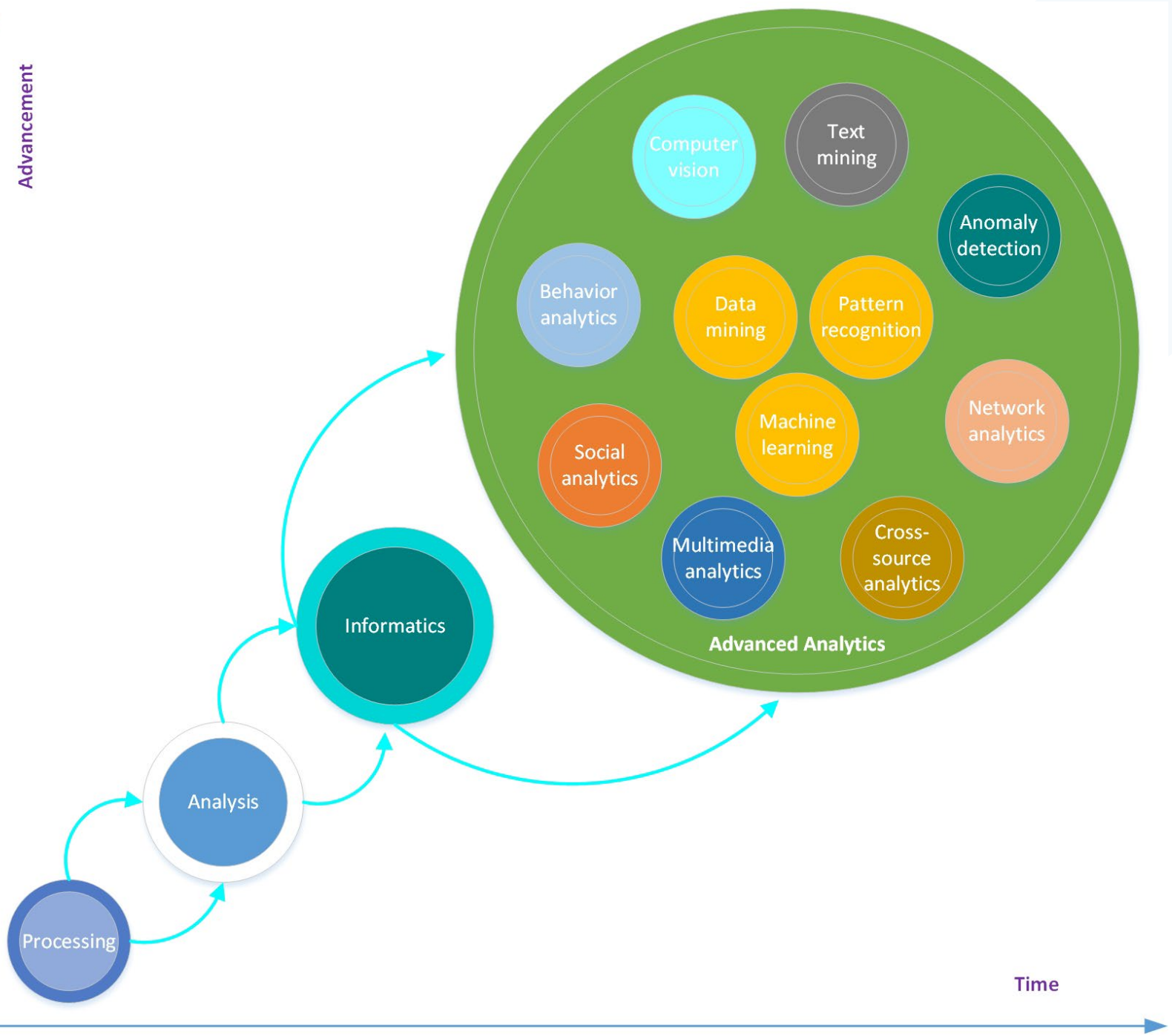
L. Cao, Data science thinking, Springer, 2018

- Proper processes: sampling, validation etc.
- Proper thinking, design, fitness, etc.

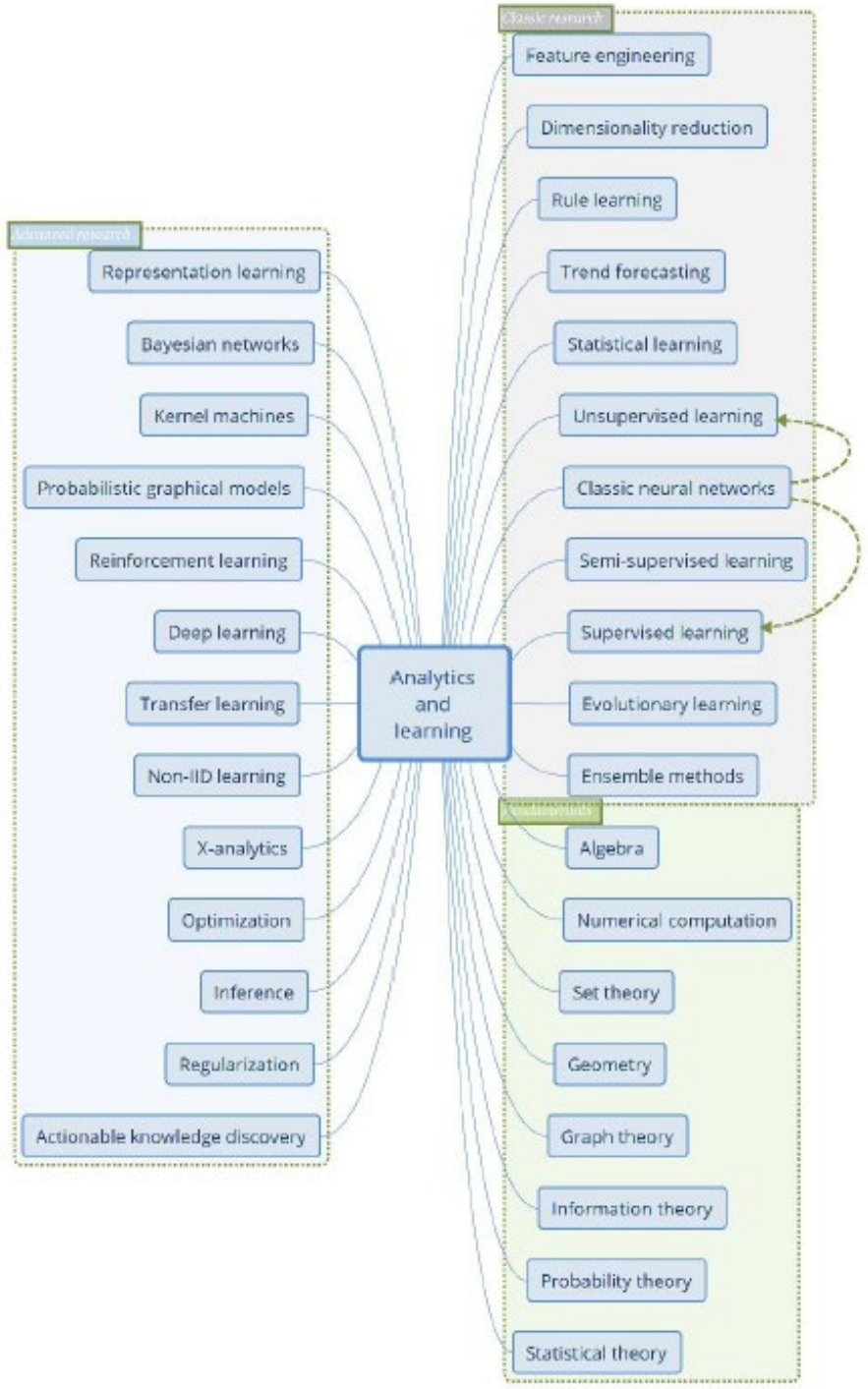
Data science thinking, Springer, 2018

Advanced analytics

- Transactions
- Demographics
- Behaviors
- Communications
- Interactions
- Networks
- Intent
- Emotion



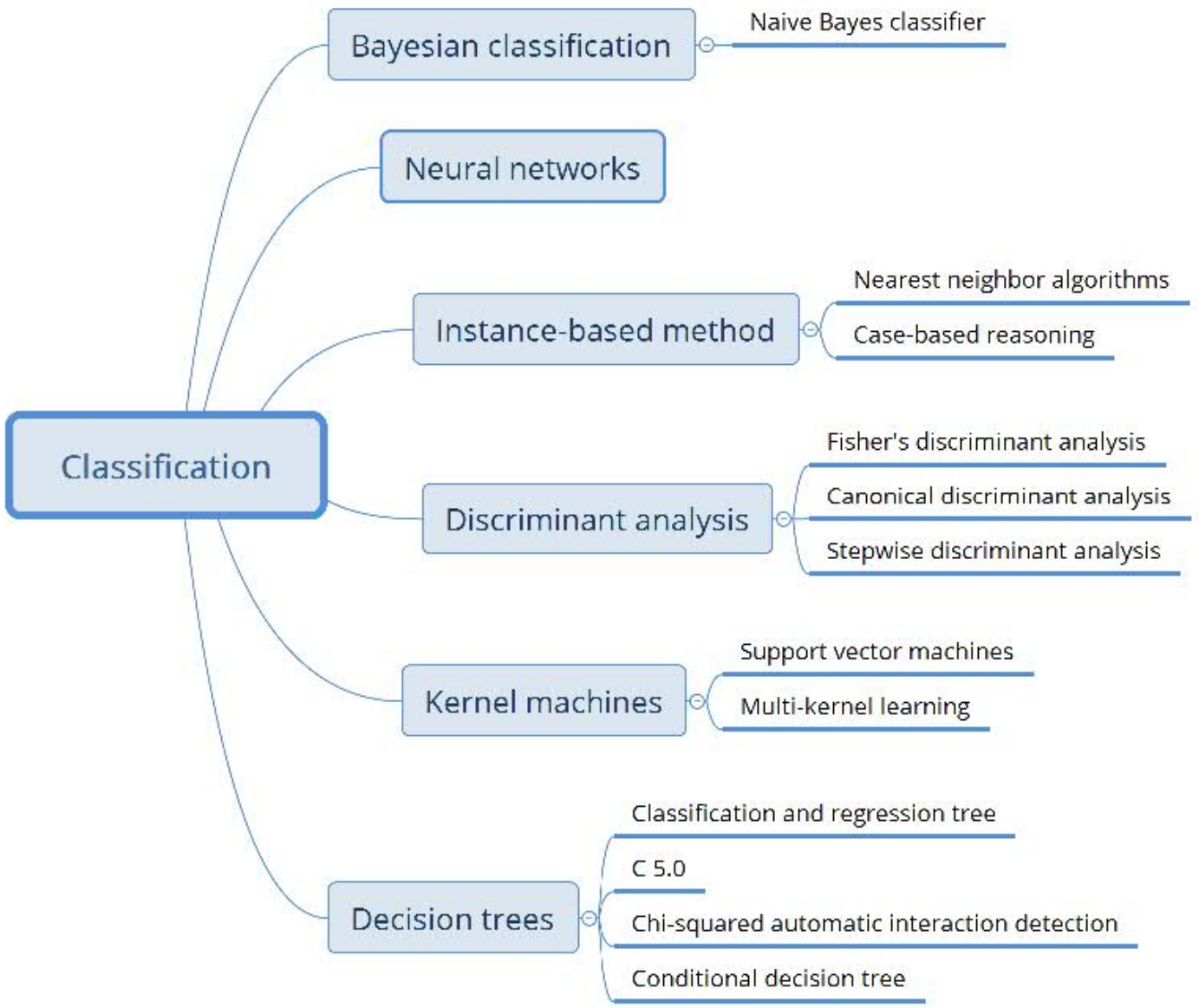
Analytics and learning techniques



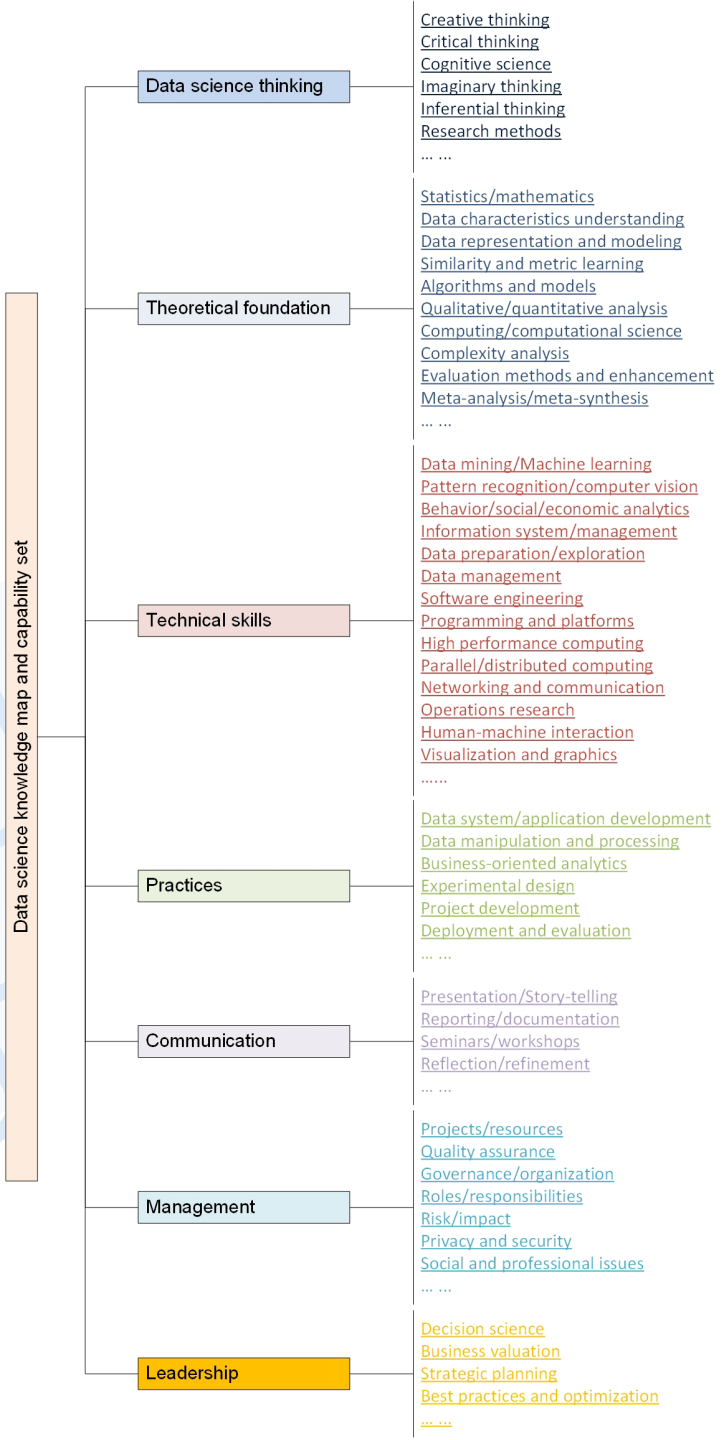
- Data science foundations
- Classic theories and methods
- Advanced theories and methods
- Data science thinking and design

Classifiers

- Popular shallow classifiers may not work well
- Pros and cons
- Fundamental assumptions for each method
- Different design methods
- Semi-supervised classification
- Unsupervised classification



Knowledge map & capability set of a qualified data scientist



While it is impossible to achieve everything as we want,

it is hoped that the MBA Innovation Studio will empower you with some of enterprise data science thinking, knowledge, experience, and readiness

Data scientists vs. data engineers

- Domain understanding
- Constraint understanding
- Data characteristics/complexities understanding
- Business-analytics problem transformation
- Data usage planning
- Analytics project building
- Data preparation
- Feature engineering
- Knowledge discovery
- Insight extraction
- Communications of results
- Analytics operationalization
- Solution marketing
- Project management

Data Scientists

- Data/system requirement engineering
- Data system selection, installation, maintenance, management
- Data acquisition, extraction, integration
- Data quality enhancement
- Data transformation
- Data matching, loading, and sharing
- Data exploration
- Enhance data competency
- Analytics programming
- Data discovery system development
- Data discovery performance enhancement
- Computational performance enhancement
- Data-related social issue management
- Data system risk management

Data Engineers

Responsibilities of Data Scientists vs. Data Engineers

- Everyone claims to be a data scientist or is doing data science
- So what about you?

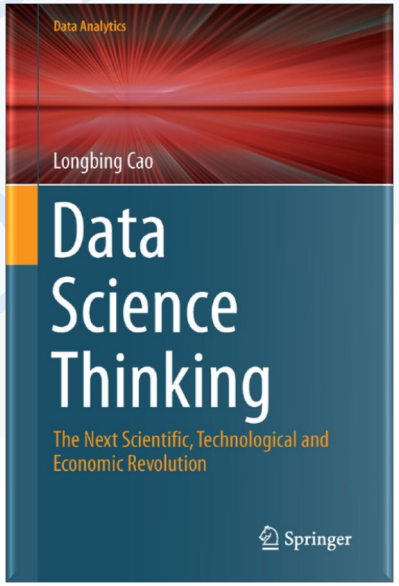
Data Economy & Case Studies

Data economy family

Case studies - discussion

Chapter 9: 27 domains

- Advertising
- Aerospace and astronomy
- Arts, creative design and humanities
- Bioinformatics
- Consulting services
- Ecology and environment
- E-commerce and retail
- Education
- Engineering
- Finance and economy

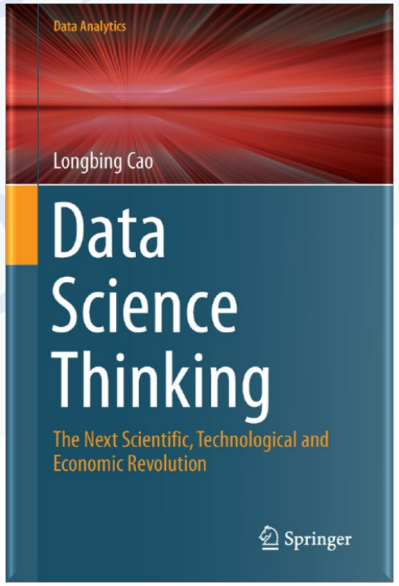


https://link.springer.com/content/pdf/10.1007%2F978-3-319-95092-1.pdf

Case studies - discussion

Chapter 9: 27 domains

- Gaming industry
- Government
- Healthcare and clinics
- Living, sports and entertainment
- Management, operations and planning
- Marketing and sales
- Medicine
- Physical-cyber-social society, networks
- Publishing and media

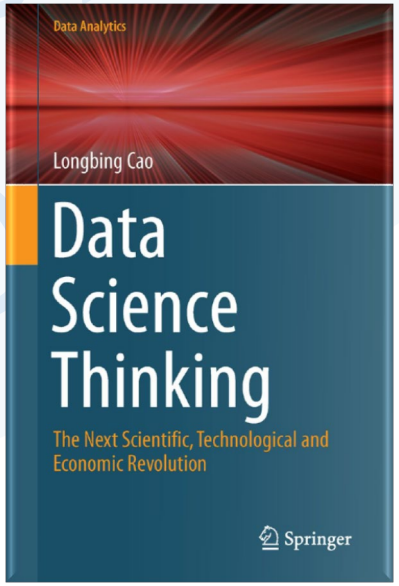


https://link.springer.com/content/pdf/10.1007%2F978-3-319-95092-1.pdf

Case studies - discussion

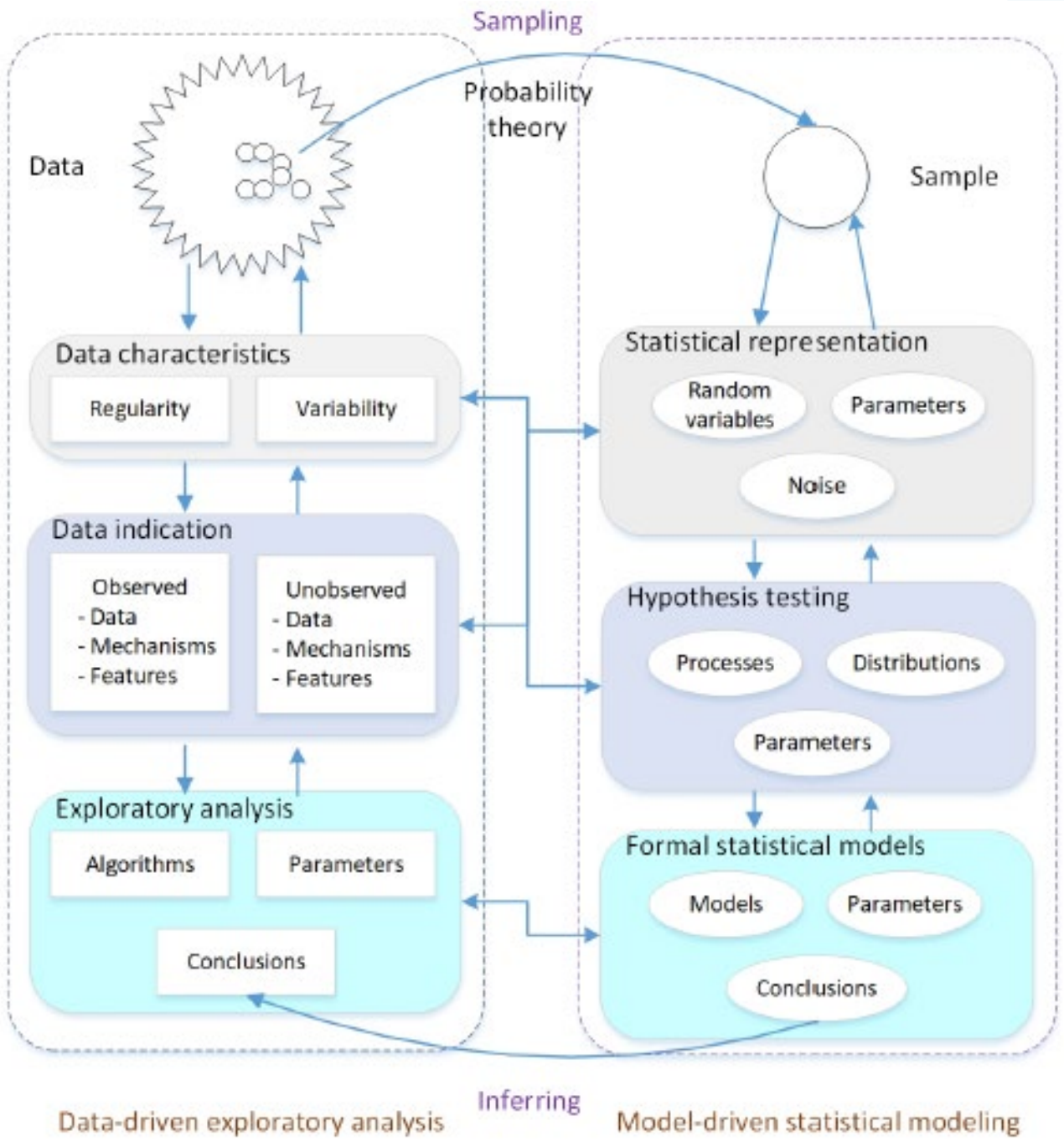
Chapter 9: 27 domains

- Recommendation services
- Science
- Security and safety
- Social sciences and problems
- Sustainability
- Telecom and mobile services
- Tourism and travel
- Transportation



https://link.springer.com/content/pdf/10.1007%2F978-3-319-95092-1.pdf

Domain, data + model-driven statistical inference



Evaluation - Knowledge actionability

- Interestingness & actionability

$$Int(p) = I(t_i(p), b_i(p))$$

$$Int(p) = t_o(\mathbf{x}, \widetilde{p}) \wedge t_s(\mathbf{x}, \widetilde{p}) \wedge b_o(\mathbf{x}, \widetilde{p}) \wedge b_s(\mathbf{x}, \widetilde{p})$$

$$Int(p) \rightarrow \hat{I}(\hat{t_o}(), \hat{t_s}(), \hat{b_o}(), \hat{b_s}())$$
$$= \alpha\hat{t_o}() + \beta\hat{t_s}() + \gamma\hat{b_o}() + \delta\hat{b_s}()$$

$$AKD^{e,\tau,m\in M} \longrightarrow O_{p\in P}(Int(p))$$
$$\rightarrow O(\alpha\hat{t_o}()) + O(\beta\hat{t_s}()) +$$
$$O(\gamma\hat{b_o}()) + O(\delta\hat{b_s}())$$

$$act(p) = O_{p\in P}(Int(p))$$
$$\rightarrow O(\alpha\hat{t_o}(p)) + O(\beta\hat{t_s}(p)) +$$
$$O(\gamma\hat{b_o}(p)) + O(\delta\hat{b_s}(p))$$
$$\rightarrow t_o^{act} + t_s^{act} + b_o^{act} + b_s^{act}$$
$$\rightarrow t_i^{act} + b_i^{act}$$

Table 3. Measurement of interest of data-driven versus domain-driven data mining.

Interest		Traditional Data-Driven	Domain-Driven
Technical	Objective	Technical objective *tech_obj*()	Technical objective *tech_obj*()
	Subjective	Technical subjective *tech_subj*()	Technical subjective *tech_subj*()
Business	Objective	—	Business objective *biz_obj*()
	Subjective	—	Business subjective *biz_subj*()
Integrative		—	Actionability *act*()

- Technical and business evaluation
- Objective and subjective evaluation
- Holistic/systematic evaluation

Knowledge Actionability: Satisfying Technical and Business Interestingness, International Journal of Business Intelligence and Data Mining (IJBIDM)

Many classic methods may not be actionable: Frequent pattern mining and association rule

Combined mining

Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns, WIREs Data Mining and Knowledge Discovery, 3(2): 140-155, 2013

- Combined sources/patterns/methods, etc.
  - Pair patterns

$$\mathcal{P}: \begin{cases} X_1 \to T_1 \\ X_2 \to T_2 \end{cases} \qquad \mathcal{P}: \begin{cases} X_\mathrm{p} \to T_1 \\ X_\mathrm{p} \wedge X_\mathrm{e} \to T_2 \end{cases}$$

  - Cluster patterns

$$\mathcal{P}: \begin{cases} X_1 \to T_1 \\ \cdots \\ X_k \to T_k \end{cases}$$

  - Derivative patterns

$$\mathcal{P}: \begin{cases} X_\mathrm{p} \to T_1 \\ X_\mathrm{p} \wedge X_{\mathrm{e},1} \to T_2 \\ X_\mathrm{p} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \to T_3 \\ \cdots \\ X_\mathrm{p} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \wedge \cdots \wedge X_{\mathrm{e},k-1} \to T_k \end{cases}$$

An Example of Combined Pattern Clusters

Clusters	Rules	X_p	X_e		T	Cnt	$Conf$ (%)	I_r	I_c	$Lift$	$Cont_\mathrm{p}$	$Cont_\mathrm{e}$	$Lift$ of $X_\mathrm{p} \to T$	$Lift$ of $X_\mathrm{e} \to T$
		demographics	arrangements	repayments										
$\mathcal{P}_1$	P_5	marital:sin	irregular	cash or post	A	400	83.0	1.12	0.67	1.80	1.01	2.00	0.90	1.79
	P_6	&gender:F	withhold	cash or post	A	520	78.4	1.00		1.70	0.89	1.89	0.90	1.90
	P_7	&benefit:N	withhold & irregular	cash or post & withhold	B	119	80.4	1.21		2.28	1.33	2.06	1.10	1.71
	P_8		withhold	cash or post & withhold	B	643	61.2	1.07		1.73	1.19	1.57	1.10	1.46
	P_9		withhold & vol. deduct	withhold & direct debit	B	237	60.6	0.97		1.72	1.07	1.55	1.10	1.60
	P_{10}		cash	agent	C	33	60.0	1.12		3.23	1.18	3.07	1.05	2.74
$\mathcal{P}_2$	P_{11}	age:65+	withhold	cash or post	A	1980	93.3	0.86	0.59	2.02	1.06	1.63	1.24	1.90
	P_{12}		irregular	cash or post	A	462	88.7	0.87		1.92	1.08	1.55	1.24	1.79
	P_{13}		withhold & irregular	cash or post	A	132	85.7	0.96		1.86	1.18	1.50	1.24	1.57
	P_{14}		withhold & irregular	withhold	C	50	63.3	2.91		3.40	2.47	4.01	0.85	1.38

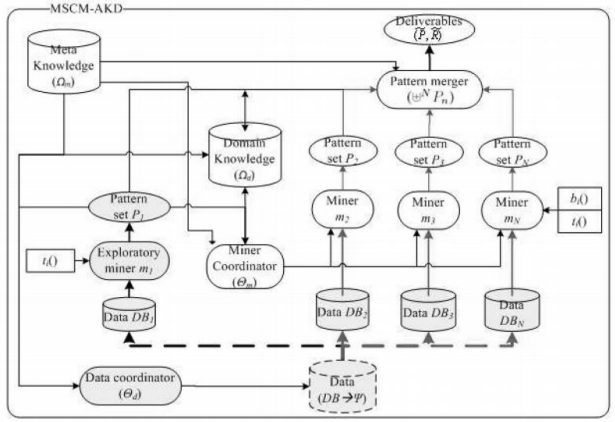


Fig. 5. Multisource + combined-mining-based AKD.

- Flexible Frameworks for Actionable Knowledge Discovery, IEEE Trans. on Knowledge and Data Engineering
- Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns

Impactful & impact-target collective behaviors

- Conditional impact ratio (*Cir*)

$$
\begin{aligned}
&Cir(Q\bar{T}|P) \\
&= \frac{Prob(Q\bar{T}|P)}{Prob(Q|P) \times Prob(\bar{T}|P)} \\
&= \frac{Prob(PQ \to \bar{T})/Prob(P)}{(Prob(PQ)/Prob(P)) \times (Prob(P \to \bar{T})/Prob(P))} \\
&= \frac{Prob(PQ \to \bar{T})/Prob(PQ)}{Prob(P \to \bar{T})/Prob(P)}.
\end{aligned}
$$

- Conditional Piatetsky-Shapiro's (P-S) ratio (*Cps*)

$$
\begin{aligned}
Cps(Q\bar{T}|P) &= Prob(Q\bar{T}|P) - Prob(Q|P) \times Prob(\bar{T}|P), \\
&= \frac{Prob(PQ \to \bar{T})}{Prob(P)} - \frac{Prob(PQ)}{Prob(P)} \times \frac{Prob(P \to \bar{T})}{Prob(P)}
\end{aligned}
$$

- Business impact and utility

$$\alpha_s = \sum b_i \times p_i \times v_i$$

$$\beta_s = \sum |b_i| \times \beta_i \times p_i \times v_i$$

$$SR = (R_p - R_f)/\sigma_p$$

$$TR = \frac{\sum_{i=1}^{u} \text{AskPrice}_i * \text{AskVolume}_i - \sum_{j=1}^{v} \text{BidPrice}_j * \text{BidVolume}_j}{\text{TotalInvestment}}$$

$$IR = \left(\sum_{i=1}^{n} (\text{Index}_{i+1} - \text{Index}_i)/\text{Index}_i\right)/n$$

Mining Impact-Targeted Activity Patterns in Imbalanced Data, IEEE Trans. on Knowledge and Data Engineering, 20(8): 1053-1066, 2008

$$
\begin{cases}
PLN \to T \\
PLN, DOC \to T \\
PLN, DOC, DOC \to T \\
PLN, DOC, DOC, DOC \to T \\
PLN, DOC, DOC, DOC, REA \to T \\
PLN, DOC, DOC, DOC, REA, IES \to T.
\end{cases}
$$

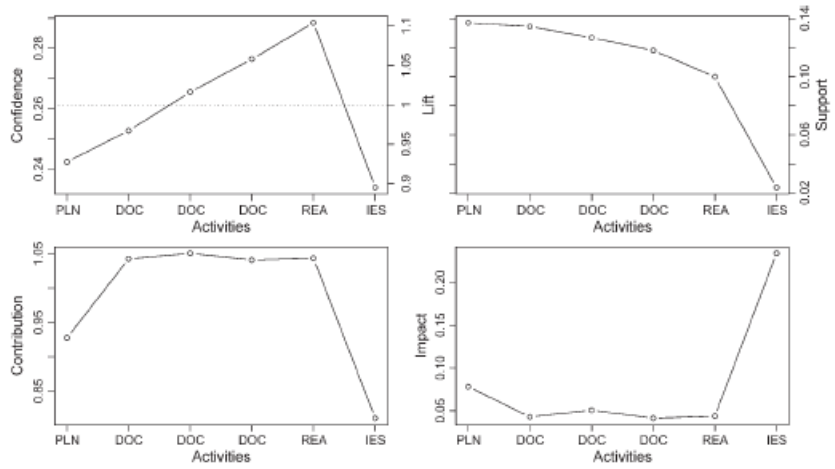
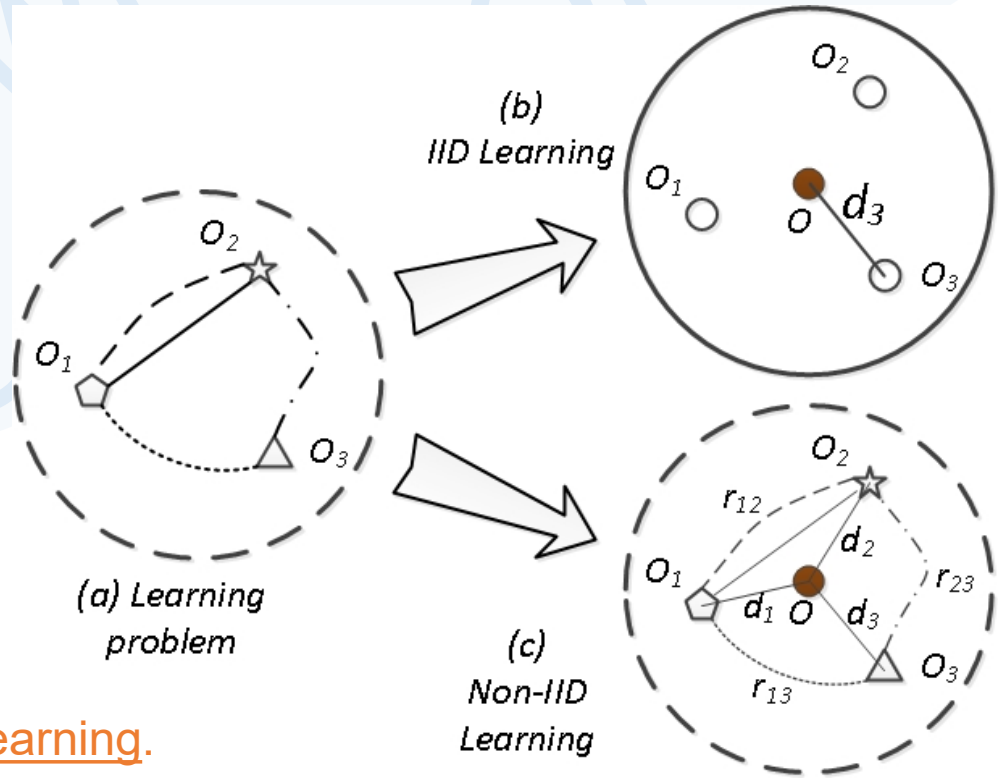


Fig. 3. Dynamic charts showing the dynamics of incremental cluster patterns.

Mining Impact-Targeted Activity Patterns in Imbalanced Data, IEEE Trans. on Knowledge and Data Engineering

Beyond i.i.d. – Non-IID learning

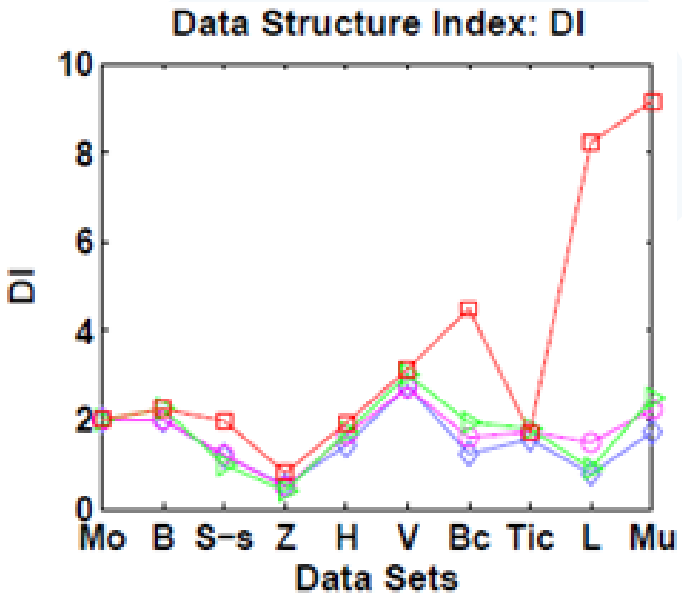
- Outcomes to be delivered by IID analytical/learning methods/algorithms on non-IID data could be:
  - incomplete
  - biased, or even
  - misleading



Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning.
IEEE Intell. Syst. 37(4): pp. 5-17, (2022)

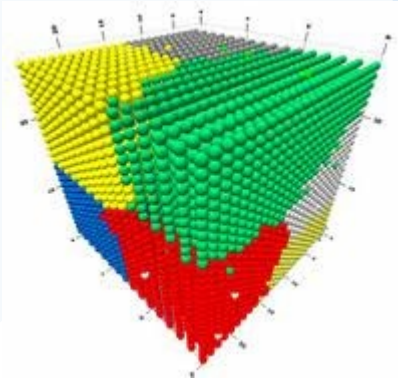
Coupled Attribute Similarity Learning on Categorical Data, IEEE Transactions on Neural Networks and Learning Systems, 26(4): 781-797 (2015)

IID K-means



acct_id	rade_date	rade_time	sec_code	ade_price	trade_vol	trade_dir	seat_code	trade_bal
210266501	20090106	112138	600331	5.63	200	B	51721	200
315726605	20090106	92500	600477	7.4	400	B	73061	2000
315726605	20090106	92500	600477	7.4	1200	B	73061	3200
315726605	20090106	145838	600477	7.64	1600	S	73061	1600
315726605	20090107	93952	600477	7.67	1600	B	73061	3200
315726605	20090106	92500	600547	48	400	B	73061	1200
315726605	20090106	95552	600547	49.14	200	S	73061	1000
315726605	20090106	95756	600547	49.1	200	S	73061	800
783486703	20090106	92500	600001	3.32	1000	B	46451	6000
783486703	20090106	92500	600001	3.32	1000	B	46451	7000

Clustering



Objective functions:

-K-means

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Note:
- X_j Individual objects only!

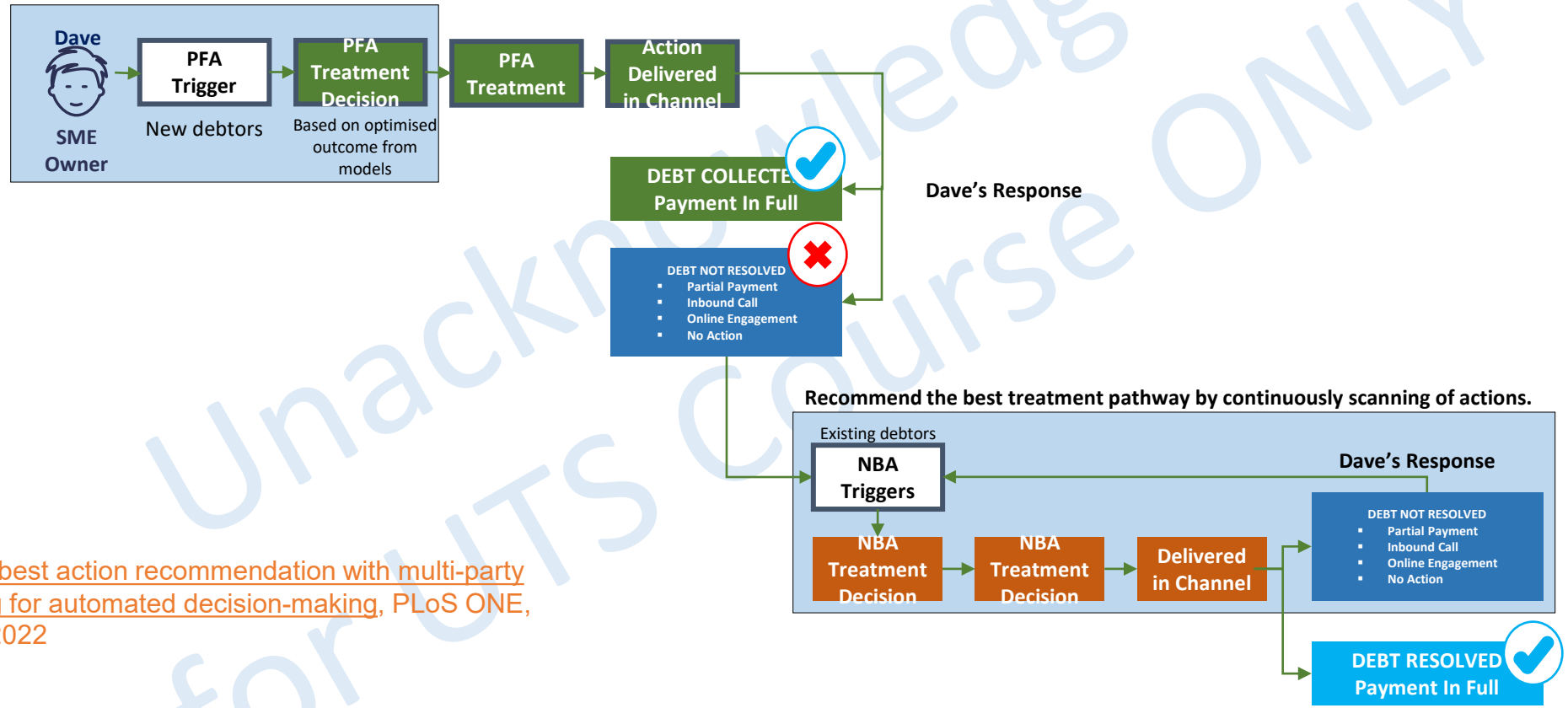
-FCM

$$J_{\mathrm{FCM}}(\boldsymbol{\mu}, \boldsymbol{A}) = \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^m ||\boldsymbol{x}_j - \boldsymbol{a}_i||^2$$

$$\sum_{i=1}^{c} \mu_{ij} = 1 \quad \text{for all } j \in J.$$

Question:
- How about X_{j1} and X_{j2} dependent?

Non-linear, Tailored Client Engagement



Personalized next-best action recommendation with multi-party interaction learning for automated decision-making, PLoS ONE, 17(1): e0263010, 2022

PNBA learning framework

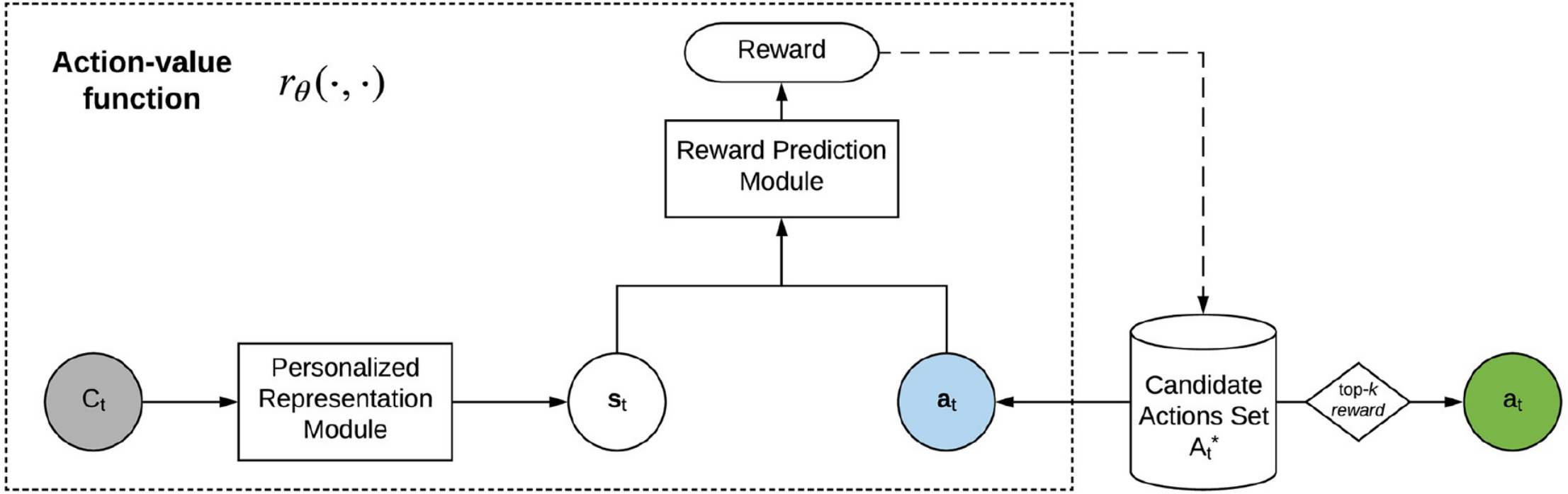
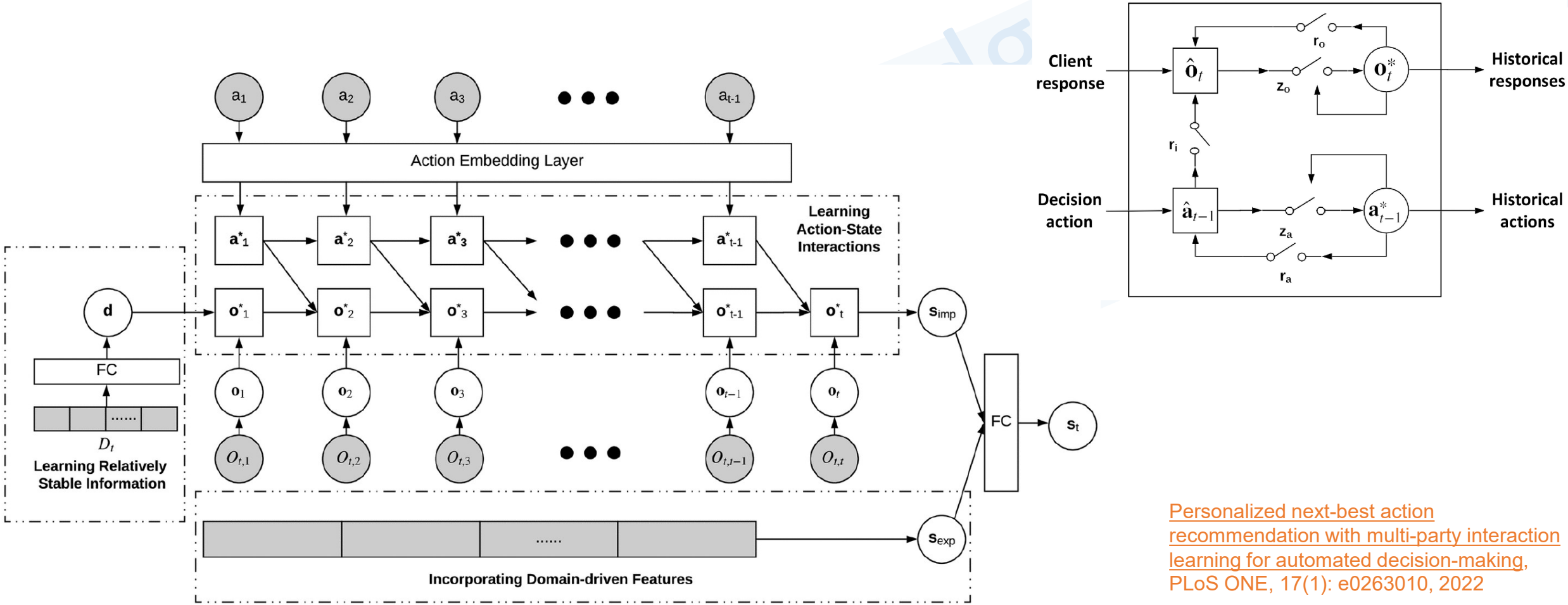


Fig 2. The framework for modeling the next-best action-oriented personalized decision-making.

Personalized next-best action recommendation with multi-party interaction learning for automated decision-making, PLoS ONE, 17(1): e0263010, 2022

Learn personalized interaction representation



Personalized next-best action recommendation with multi-party interaction learning for automated decision-making, PLoS ONE, 17(1): e0263010, 2022

Fig 3. A reinforced coupled recurrent network to learn personalized client representation.

PNBA: Case studies

- Non-Markovian NBA recommendation

Table 2. Average reward lift for 10 actions recommended by 11 deep models over the review measured by domain-driven debt collection rules.

Model	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Total_Avg	Action_Avg
CRN_IMB	5	4	3.0534	2.8752	6.8	2.1415	2.6984	3.3567	1.6772	2.9969	2.5569	3.4599
CRN	2.1957	3.5383	2.2068	2.6616	3.216	2.074	2.326	2.6277	1.7654	2.3425	2.1942	2.4954
WD	2.604	1.5992	2.0979	2.2798	3.2239	1.9824	2.2629	2.6967	0.9899	2.312	2.1089	2.2049
LSTM	0.9722	1.0987	0.9391	0.974	1.1272	1.0159	0.897	1.1097	1.1024	1.0847	1.0013	1.0321
WD_LSTM	2.0471	1.2731	1.9709	2.4755	2.2217	1.8129	2.0816	2.1909	1.1405	2.105	1.9198	1.9319
WD_Res_LSTM	1.7247	0.8219	1.7007	1.9816	2.4985	1.8164	1.9851	2.0921	0.8285	1.967	1.8488	1.7416
WD_Multi_LSTM	1.684	1.0468	1.6591	1.774	1.6924	1.7083	1.671	2.1678	1.2222	1.8098	1.7161	1.6435
GRU	0.5783	0.0865	0.9852	1.1201	1.5022	0.9154	0.861	0.9463	1.0347	1.0416	0.9345	0.9071
WD_GRU	1.0049	0.6397	1.3454	1.7369	2.1271	1.6489	1.6049	2.1562	0.665	1.6602	1.611	1.4589
WD_Res_GRU	1.4488	1.1333	1.7364	1.3479	2.2259	1.6932	1.7091	1.9582	1.2507	1.8869	1.7248	1.6391
WD_Multi_GRU	1.6329	1.8399	1.9114	1.7949	1.8781	1.8206	2.0276	1.7613	1.0508	2.2347	1.8959	1.7952
Δ_IMB	92.01%	117.40%	45.55%	16.15%	110.92%	8.03%	19.25%	24.47%	34.10%	29.62%	21.24%	56.92%
Δ	-15.68%	92.31%	5.19%	7.52%	-0.25%	4.62%	2.79%	-2.56%	41.15%	1.32%	4.04%	13.18%

Sampling non-IIDness in deep networks

- Sampling non-IIDnesses
- Distributional vulnerability
- High confidence prediction on out-of-distribution samples

Label vs distributional fitting

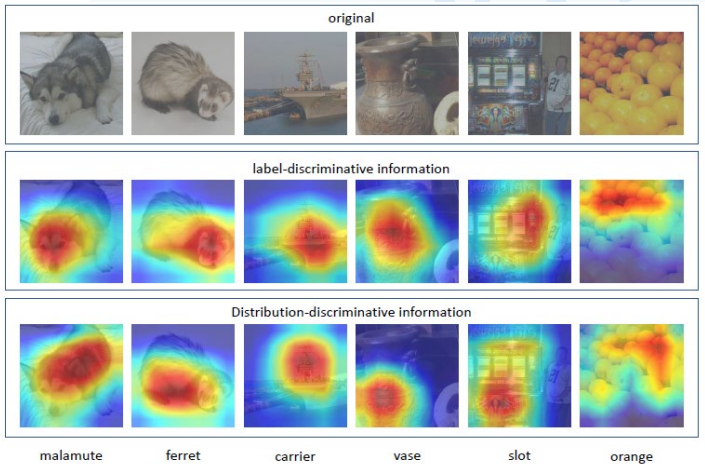


Fig. 6.8 DRL: Heat maps of Grad-CAM for label- and distribution-discriminative representations.
Red regions correspond to high scores for class, while blue regions correspond to low scores. The figure is best viewed in color.

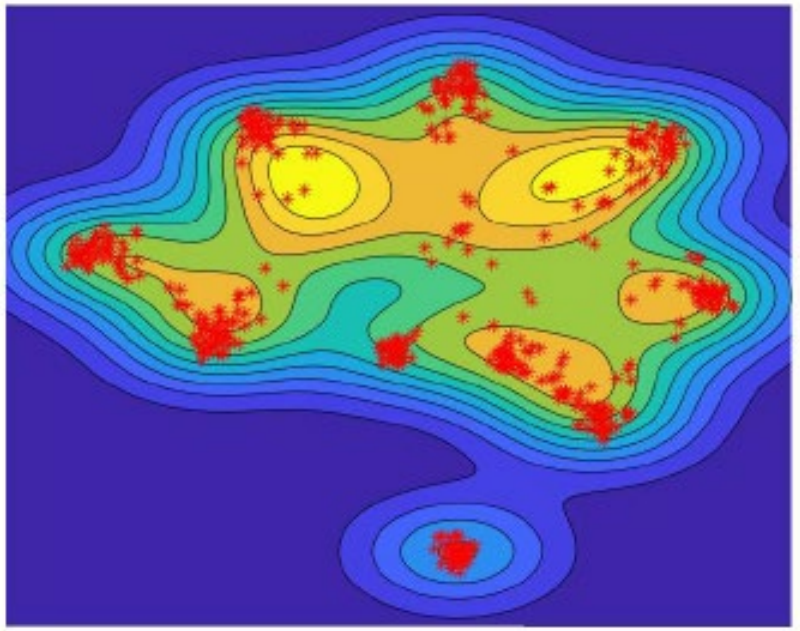
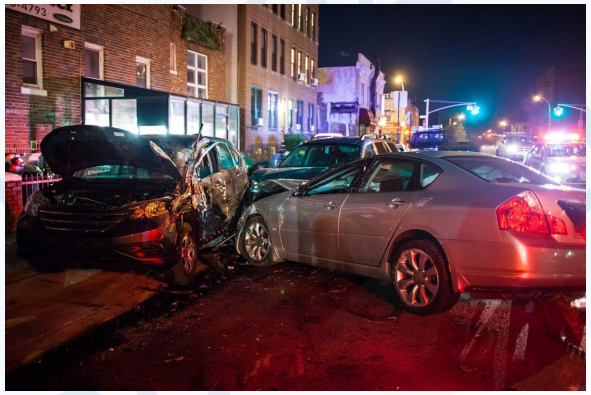


Fig. 1. The heat map of prediction confidence by ResNet18 on CIFAR10. The embedding results are constructed by t-SNE [7]. Red points correspond to training in-distribution samples. Yellow regions correspond to high confidence for predictions, while blue regions correspond to low confidence. ResNet18 assigns high-confidence predictions on samples located in the regions outside the training in-distribution samples, i.e., out-of-distribution samples. It shows ResNet18 does not discriminate between in- and out-of-distribution samples. The figure is best viewed in color.

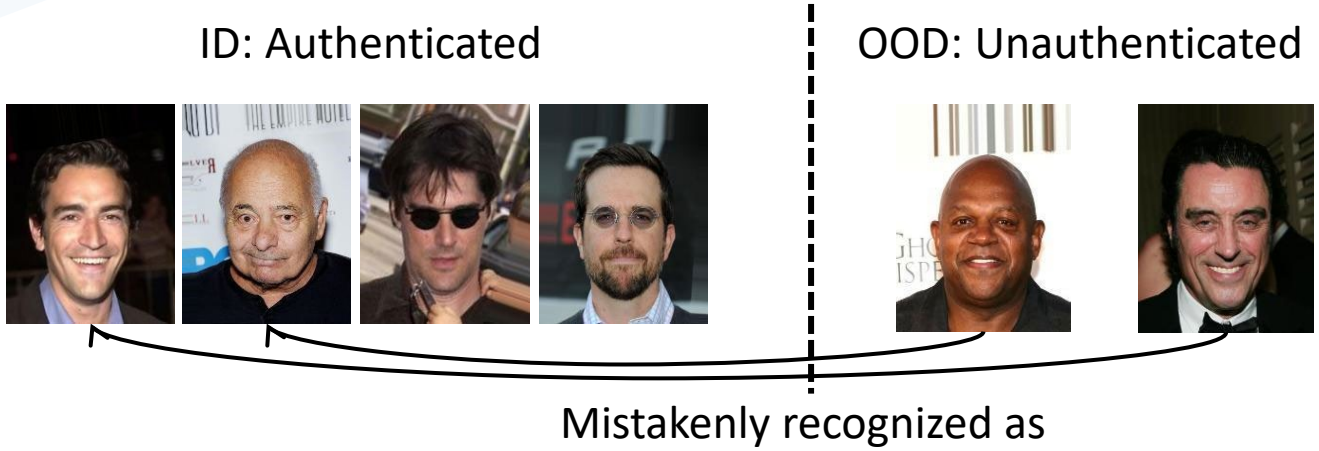
Zhilin Zhao, Longbing Cao, Kun-Yu Lin. Out-of-distribution Detection by Cross-class Vicinity Distribution of In-distribution Data, arXiv:2206.09385

Various vulnerabilities of deep networks

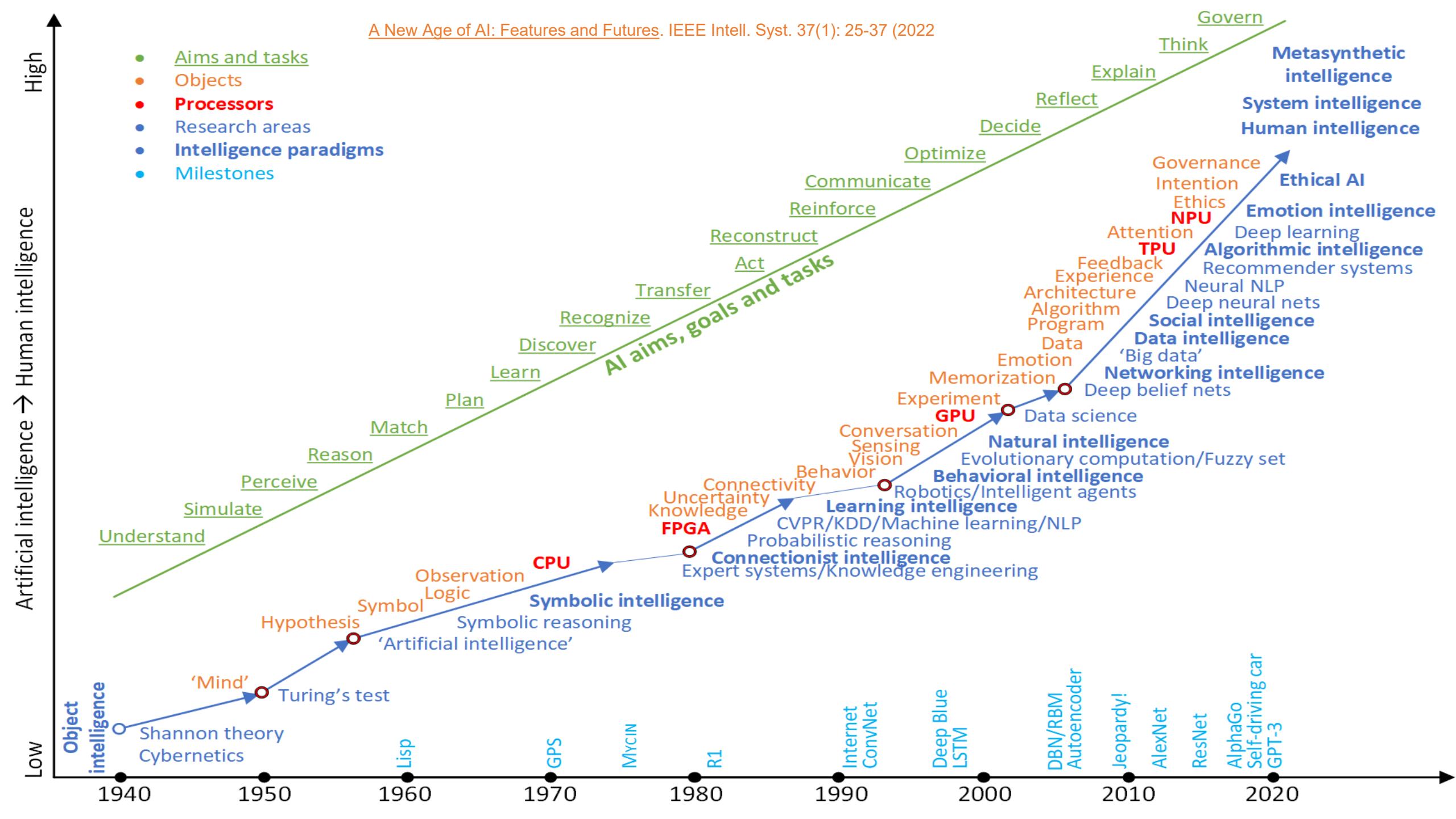
- Network, distribution, domain… vulnerability
  - Face recognition: incorrectly recognize a stranger as a person authenticated by the system
  - Driverless car: a high-confidence action at an unknown situation, which should be passed to the human driver for handling, may cause a serious accident
  - Deep fake



https://fee.org/media/27691/car_crash.jpg?anchor=center&mode=crop&width=1200&rnd=131672371490000000

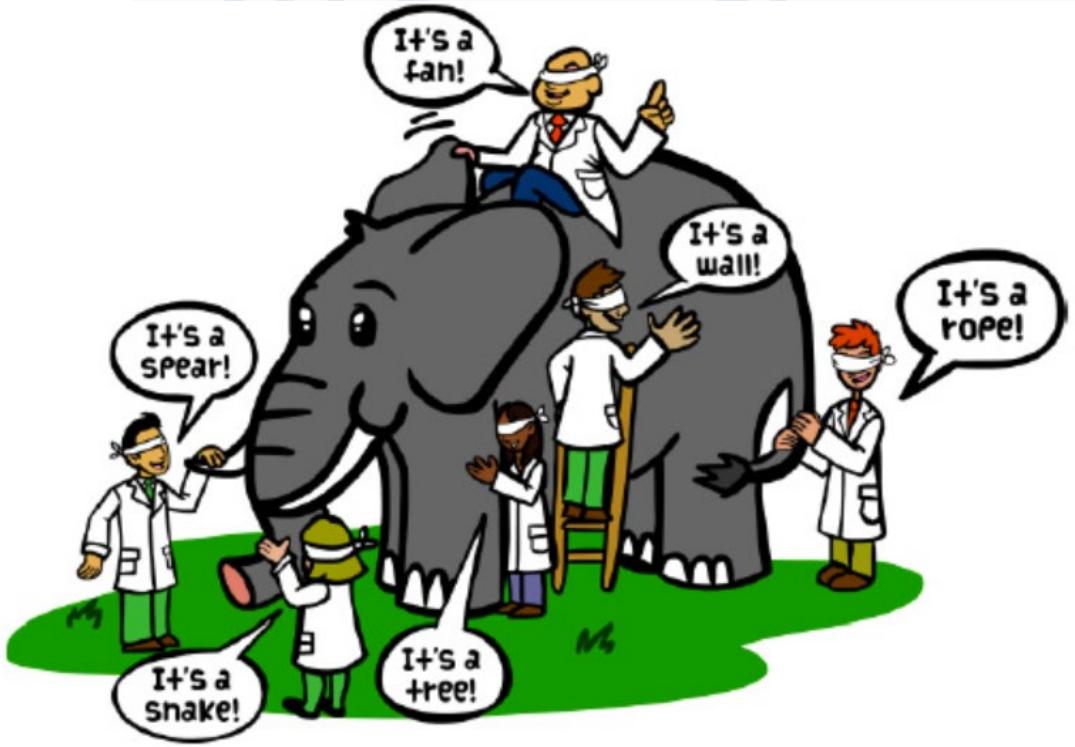


Concluding Remarks

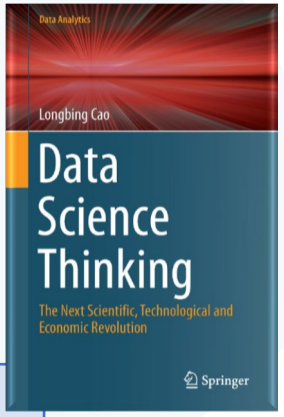

A New Age of AI: Features and Futures. IEEE Intell. Syst. 37(1): 25-37 (2022
Aims and tasks
Objects
Processors
Research areas
Intelligence paradigms
Milestones
Artificial intelligence → Human intelligence
High
Low
AI aims, goals and tasks
Understand
Simulate
Perceive
Reason
Match
Plan
Learn
Discover
Recognize
Transfer
Act
Reconstruct
Reinforce
Communicate
Optimize
Decide
Reflect
Explain
Think
Govern
Object intelligence
'Mind'
Turing's test
Shannon theory
Cybernetics
Hypothesis
Symbol
Logic
Observation
'Artificial intelligence'
Symbolic reasoning
Symbolic intelligence
CPU
FPGA
Knowledge
Uncertainty
Connectivity
Expert systems/Knowledge engineering
Connectionist intelligence
Probabilistic reasoning
CVPR/KDD/Machine learning/NLP
Learning intelligence
Behavior
Vision
Sensing
Conversation
Robotics/Intelligent agents
Behavioral intelligence
Experiment
GPU
Memorization
Emotion
Evolutionary computation/Fuzzy set
Natural intelligence
Data science
Deep belief nets
Networking intelligence
'Big data'
Data intelligence
Program
Algorithm
Architecture
Experience
Feedback
Deep neural nets
Social intelligence
Neural NLP
Attention
TPU
Recommender systems
Ethics
Intention
NPU
Deep learning
Governance
Algorithmic intelligence
Ethical AI
Emotion intelligence
Metasynthetic intelligence
System intelligence
Human intelligence
Lisp
GPS
Mycin
R1
Internet
ConvNet
Deep Blue
LSTM
DBN/RBM
Autoencoder
Jeopardy!
AlexNet
ResNet
AlphaGo
Self-driving car
GPT-3
1940
1950
1960
1970
1980
1990
2000
2010
2020

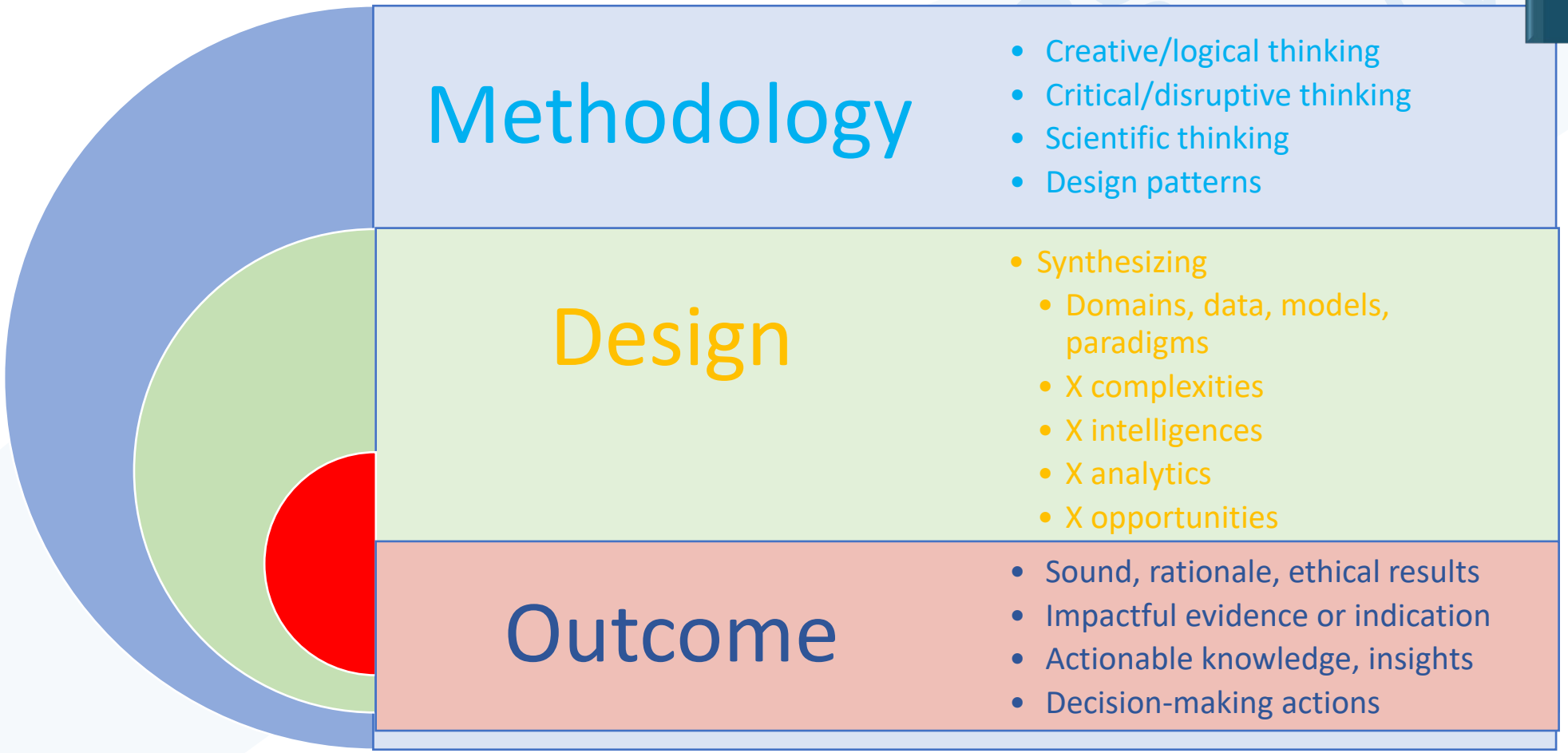
We do not know what we do not know

- How can AI enable blind people to tell a genuine story about elephant?
  - Beyond IID thinking
  - Couplings between parts
  - Heterogeneities between parts
  - From touching/representation → analysis → reasoning/inference → summarization/integration
  - Local – global picture (known → unknown)/optimization



Data science thinking matters





References

www.datasciences.org

- David Donoho. 50 years of data science.
- Longbing Cao: Data Science Thinking, Springer, 2018. https://www.springer.com/us/book/9783319950914
- Data Science: A Comprehensive Overview. ACM Computing Surveys, 50(3), 43:1-42, 2017. https://arxiv.org/abs/2007.03606
- Data Science: Nature and Pitfalls. IEEE Intelligent Systems, Volume: 31, Issue: 5, 66-75, 2016. https://arxiv.org/abs/2006.16964
- Data Science: Challenges and Directions. Communications of the ACM, Vol. 60 No. 8, Pages 59-68, 2017. https://arxiv.org/abs/2006.16966
- Data Science: Profession and Education. IEEE Intelligent Systems, 34(5): 35-44, 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3755747

https://link.springer.com/content/pdf/10.1007%2F978-3-319-95092-1.pdf

References

www.datasciences.org

- Data Mining for Business Applications, ISBN: 978-0-387-79420-4, Springer, 2008
- AI in Finance: Challenges, Techniques and Opportunities, SSRN, 1-40, 2021
- Data science and AI in FinTech: An overview. Int. J. Data Sci. Anal. 12(2): 81-99, 2021
- Effective detection of sophisticated online banking fraud on extremely imbalanced data, World Wide Web, 16:449–475, 2013
- Copula-Based High Dimensional Cross-Market Dependence Modeling, DSAA2017 Research Track, 734-743, 2017
- An effective contrast sequential pattern mining approach to taxpayer behavior analysis, World Wide Web 19(4): 633-651, 2016
- Personalized next-best action recommendation with multi-party interaction learning for automated decision-making, PLOS One, 17(1): e0263010, 2022
- Table2Vec-automated universal representation learning of enterprise data DNA for benchmarkable and explainable enterprise data science. Sci Rep 11, 23957, 2021
- Combined mining: Analyzing object and pattern relations for discovering and constructing complex yet actionable patterns. WIREs Data Mining Knowl. Discov. 3(2): 140-155, 2013
- Domain Driven Data Mining, ISBN: 978-1-4419-5737-5, Springer, 2010