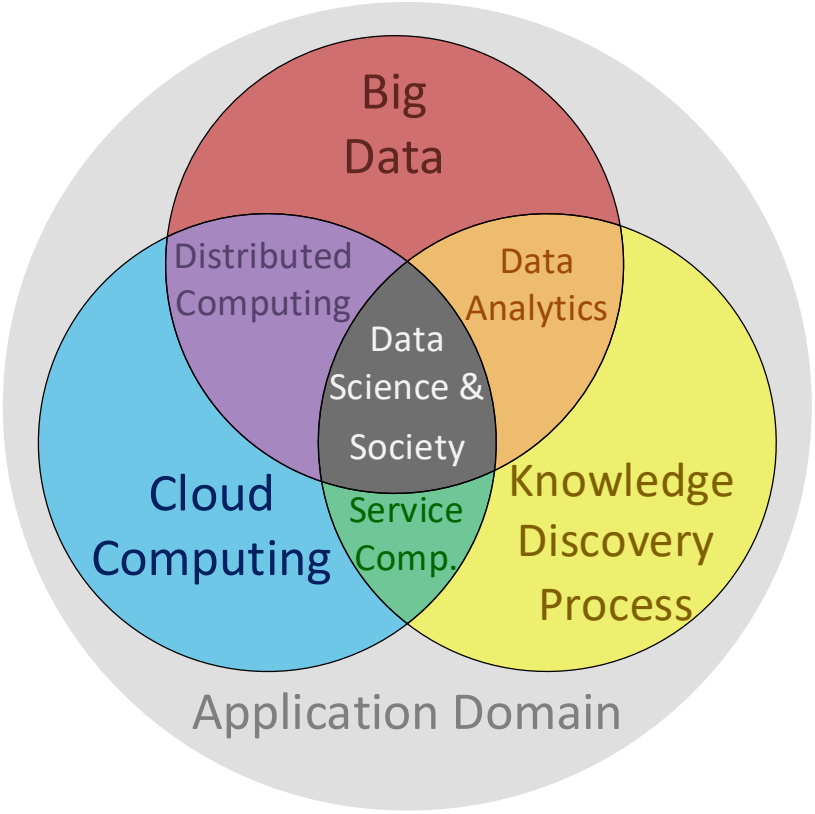
Synthesis & Trends

Final Lecture of Data Science & Society 2018

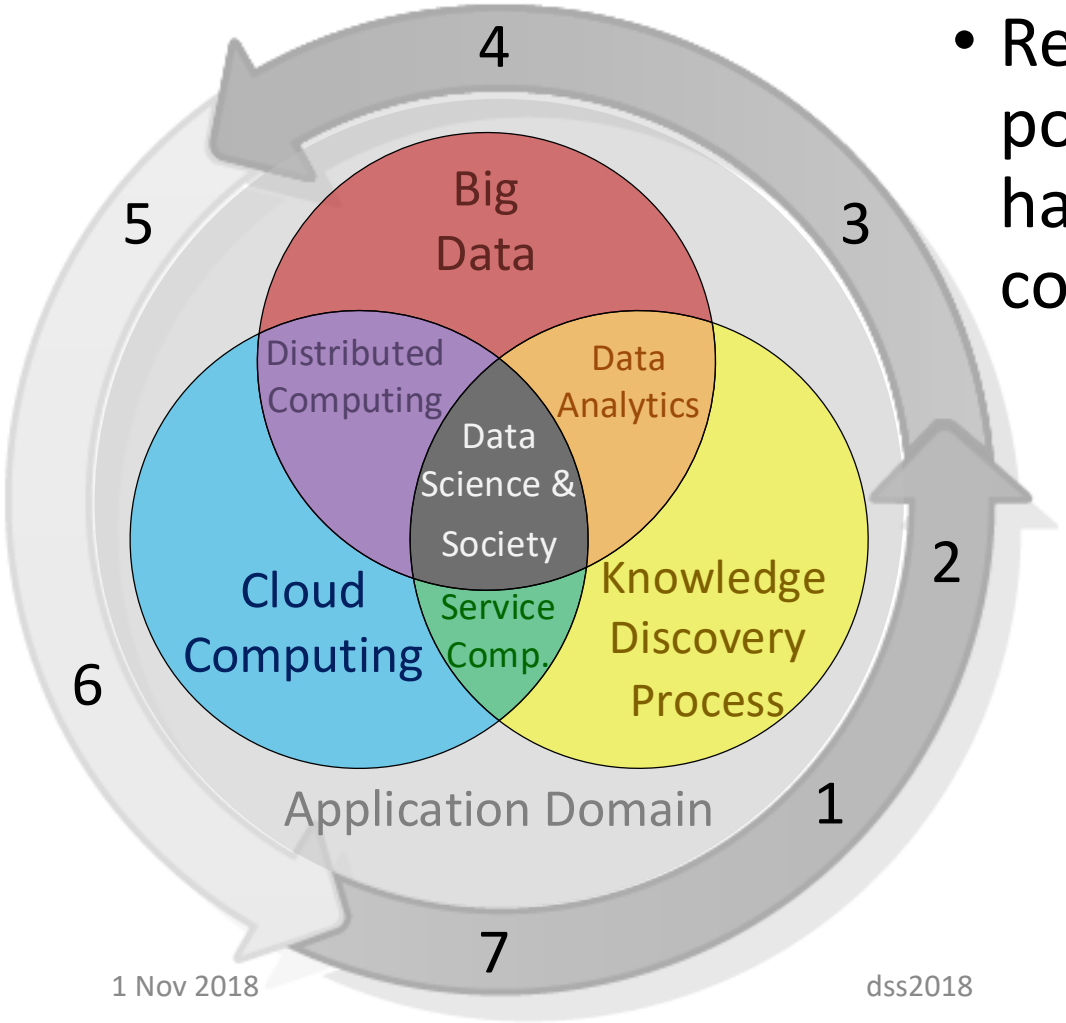


Dr. Marco Spruit :: PI Applied Data Science Lab

Agenda

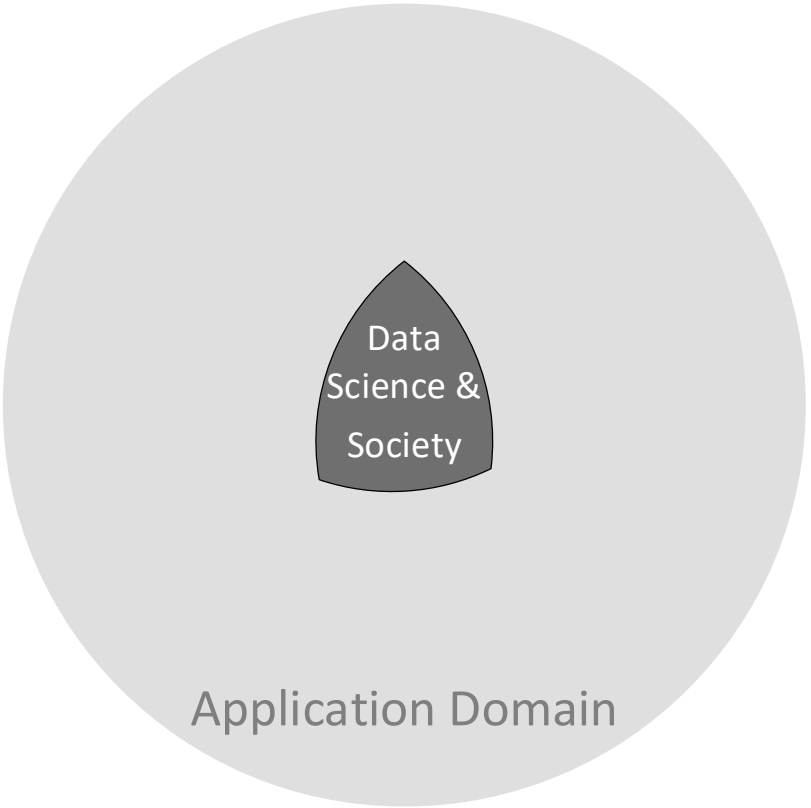
- Synthesis
  - What we have done in this course
- Trends in Data Science in 2018 and beyond
  - What has recently happened in the field
  - What can we expect to happen in the coming year
- Q/A final exam

Synthesis



- Revisiting and positioning what we have done in this course…

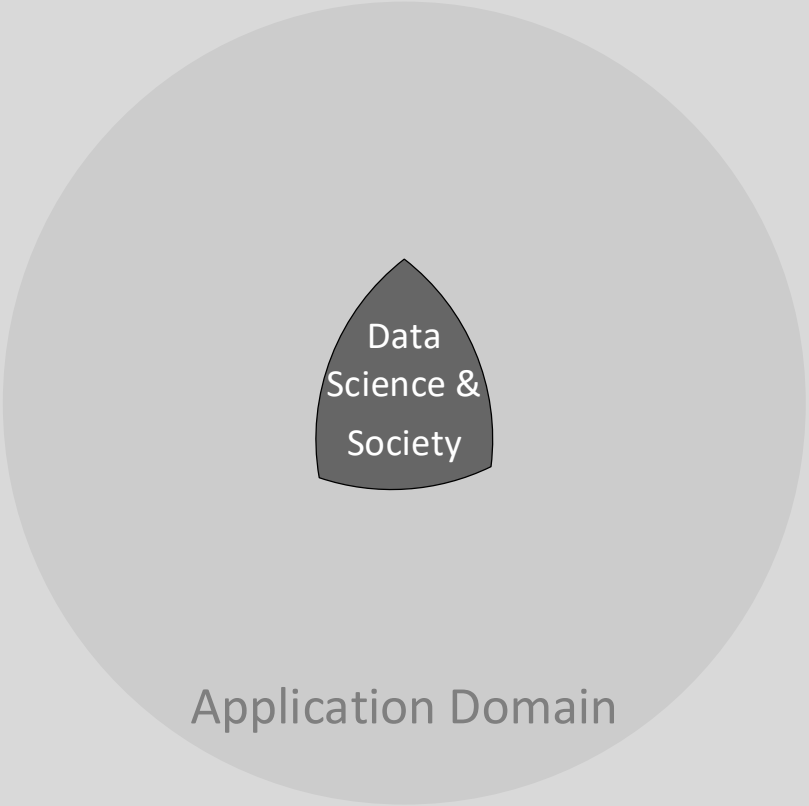
1. Data Science & *Society* focus



Application domains:

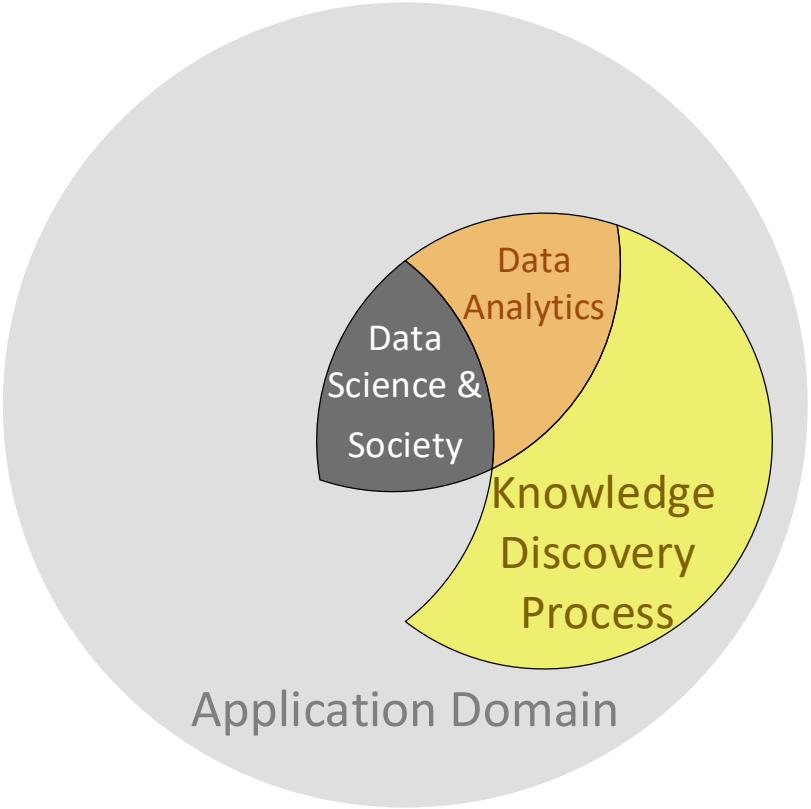
- Neonatology
- Business
- Epidemiology
- Geography
- Cell biology
- Ethics & Privacy
- Psychiatry

Intended Learning Outcome 1



1. **Understand the role of data science and its societal impact**
2. Recognise the knowledge discovery processes in applied data science
3. Identify trends and developments in big data technologies
4. Apply selected big data technologies to solve real-world problems

2. Data Science *Process* focus



Knowledge Discovery Process:

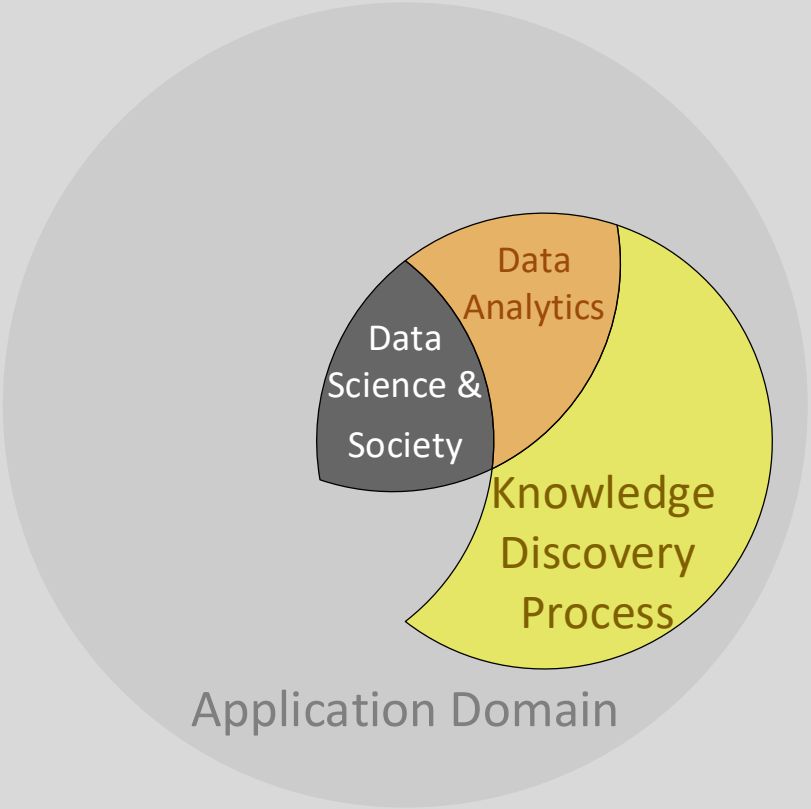
= Applied Data Science

- CRISP-DM method

Data Analytics:

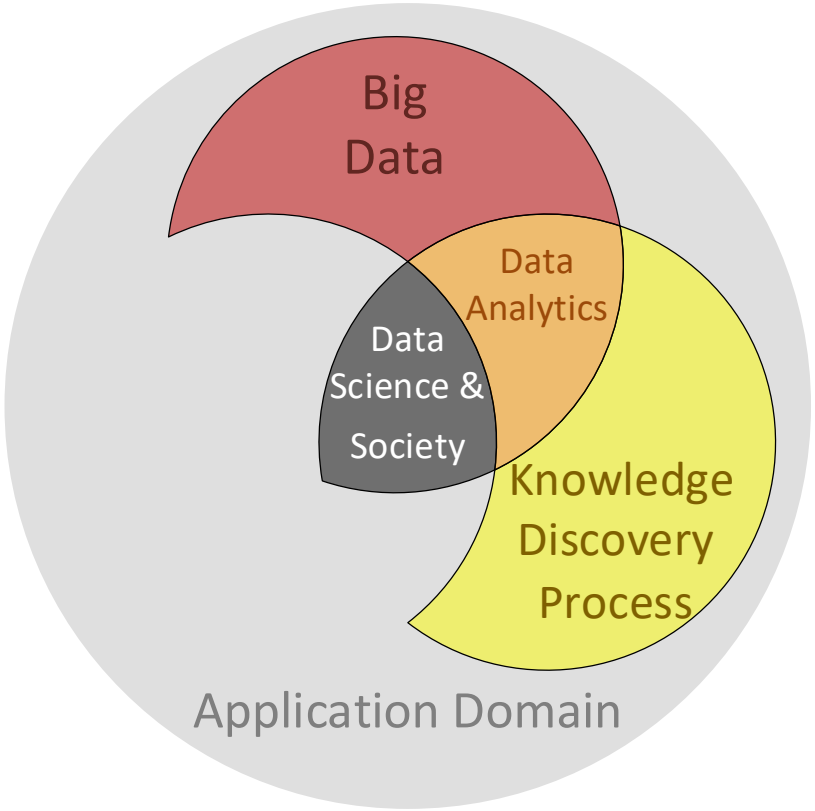
- Methods & Statistics
  - Traps in Big Data analysis:
  - p-values & multiple testing
  - Replicability, overfitting, construct validity, ...
- Workshop tutorials!

Intended Learning Outcome 2



1. Understand the role of data science and its societal impact
2. **Recognise the knowledge discovery processes in applied data science**
3. Identify trends and developments in big data technologies
4. Apply selected big data technologies to solve real-world problems

3. *Big* Data focus

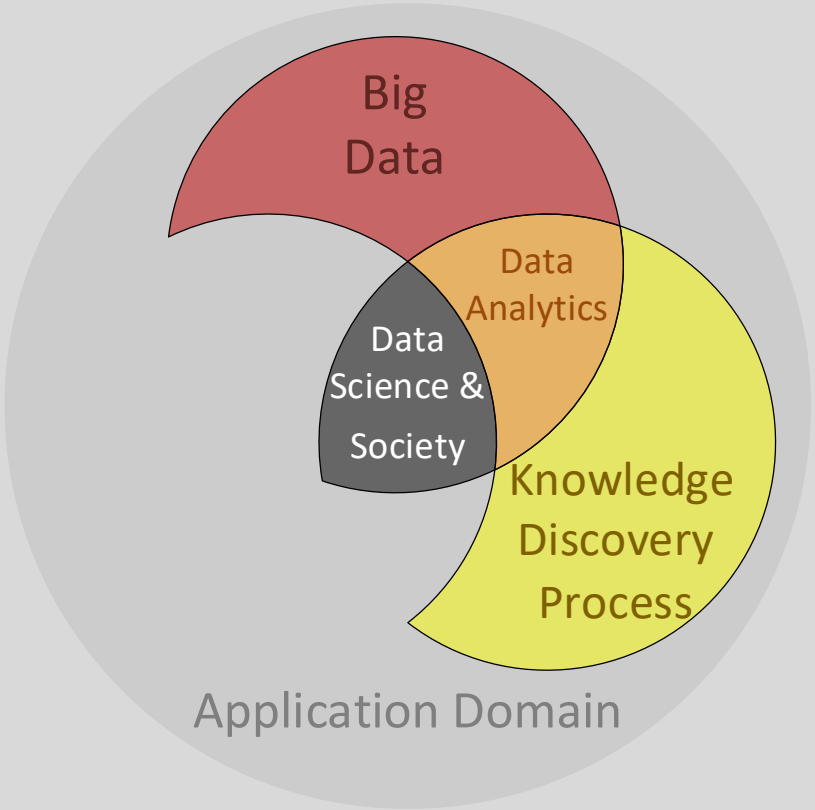


Big Data:

- Context: Book review
- Focus: Identified by experts
- 4Vs
- BD vs DWH
- SQL vs NoSQL

- Ethics & Compliance
  - Philosophical perspective?

Intended Learning Outcome 3



1. Understand the role of data science and its societal impact
2. Recognise the knowledge discovery processes in applied data science
3. **Identify trends and developments in big data technologies**
4. Apply selected big data technologies to solve real-world problems

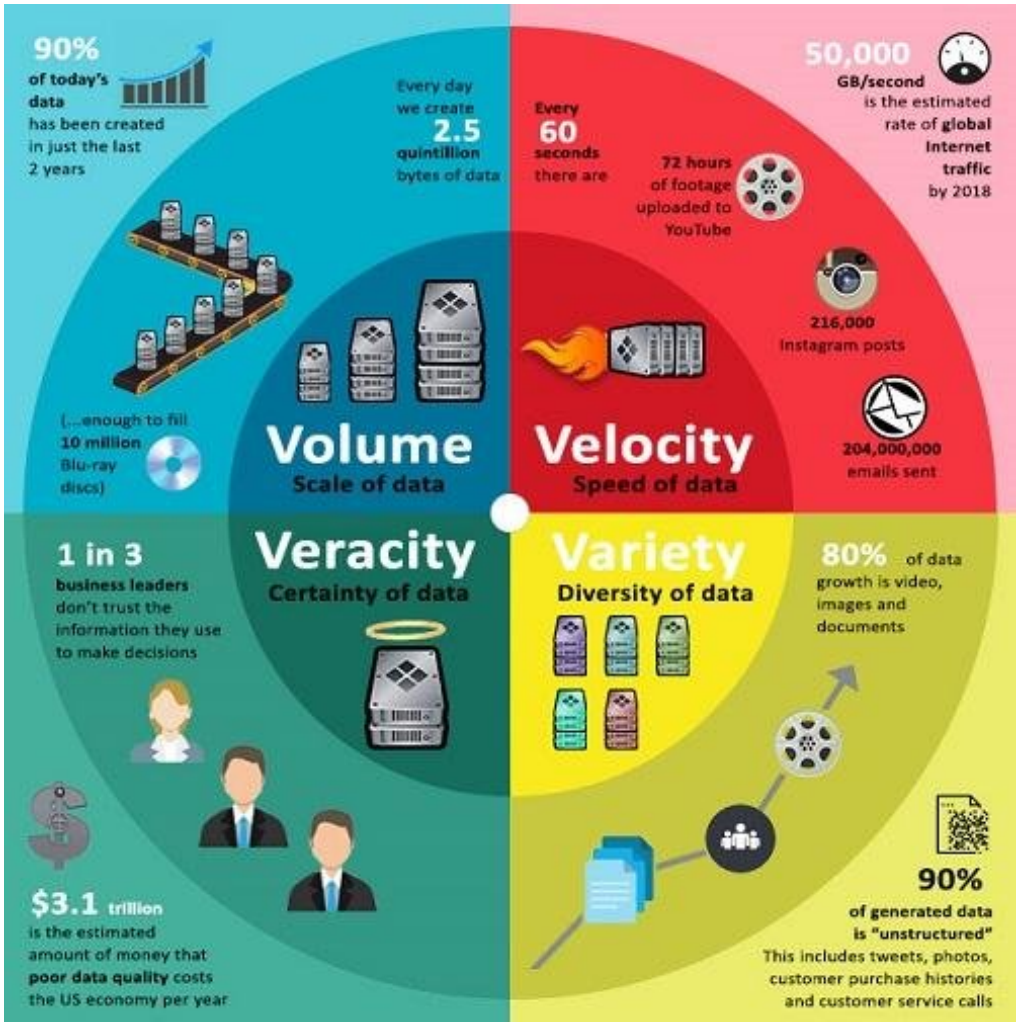
4Vs: What is big data?

Q1: How to combine big data with traditional data warehousing?

Q2: How to combine the slow data track (*cold path*) and the fast data track (*hot path*)?

OR...

- ***Variability ?***
  (*i.e.* the change in other characteristics)
  (NIST, 2015)
- ***Value ? Volatility ? etc.***

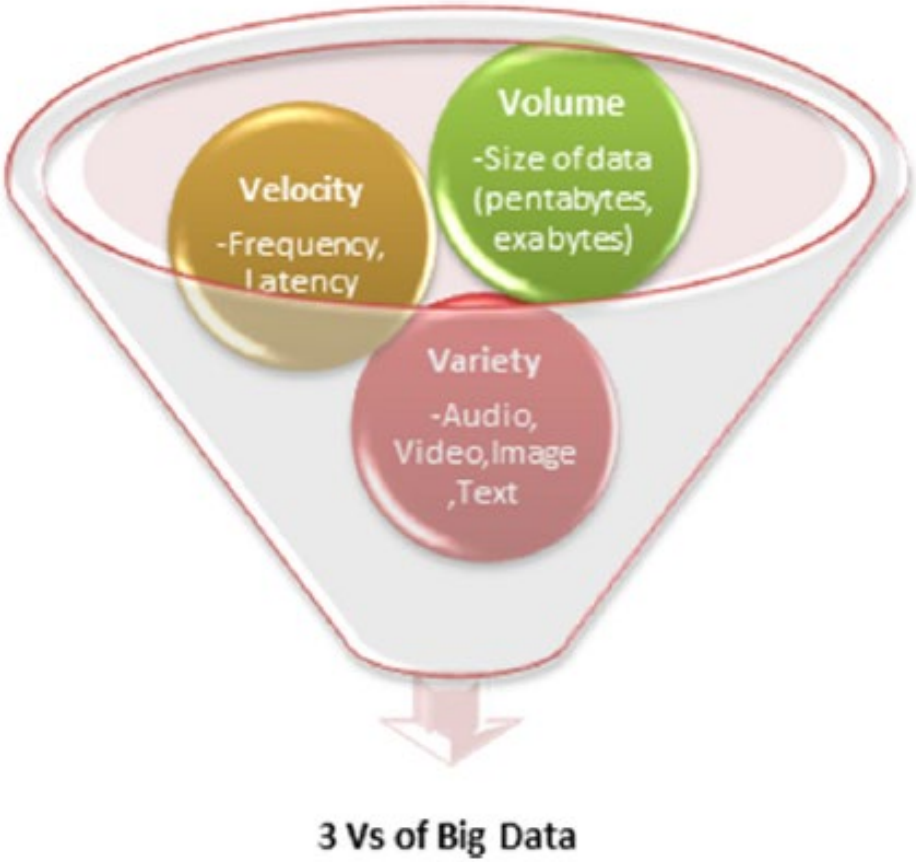


Nayyar, A. & Puri, V. (2017). Comprehensive analysis & performance comparison of clustering algorithms for big data. Review of computer engineering research. 4. 54-80.

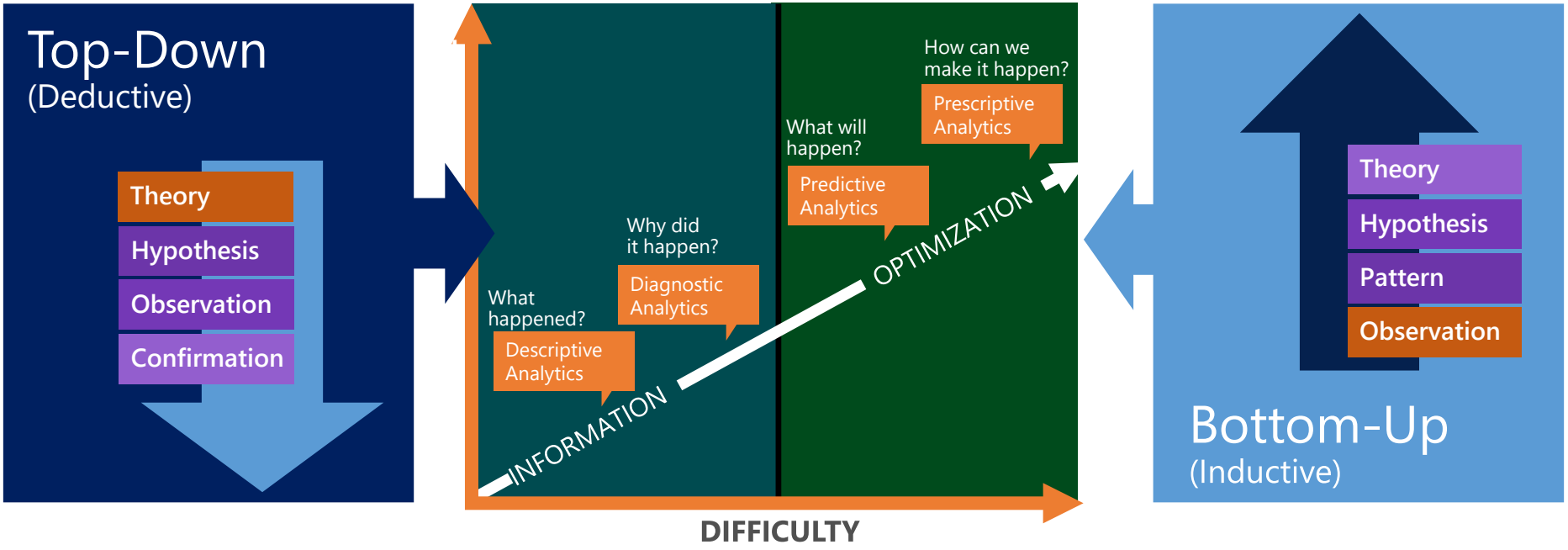
3Vs: What is big data, in essence?

Fotaki,G., Spruit,M., Brinkkemper,S., & Meijer,D. (2014). Exploring big data opportunities for online customer segmentation. *International Journal of Business Intelligence Research, 5*(3), 57–73.

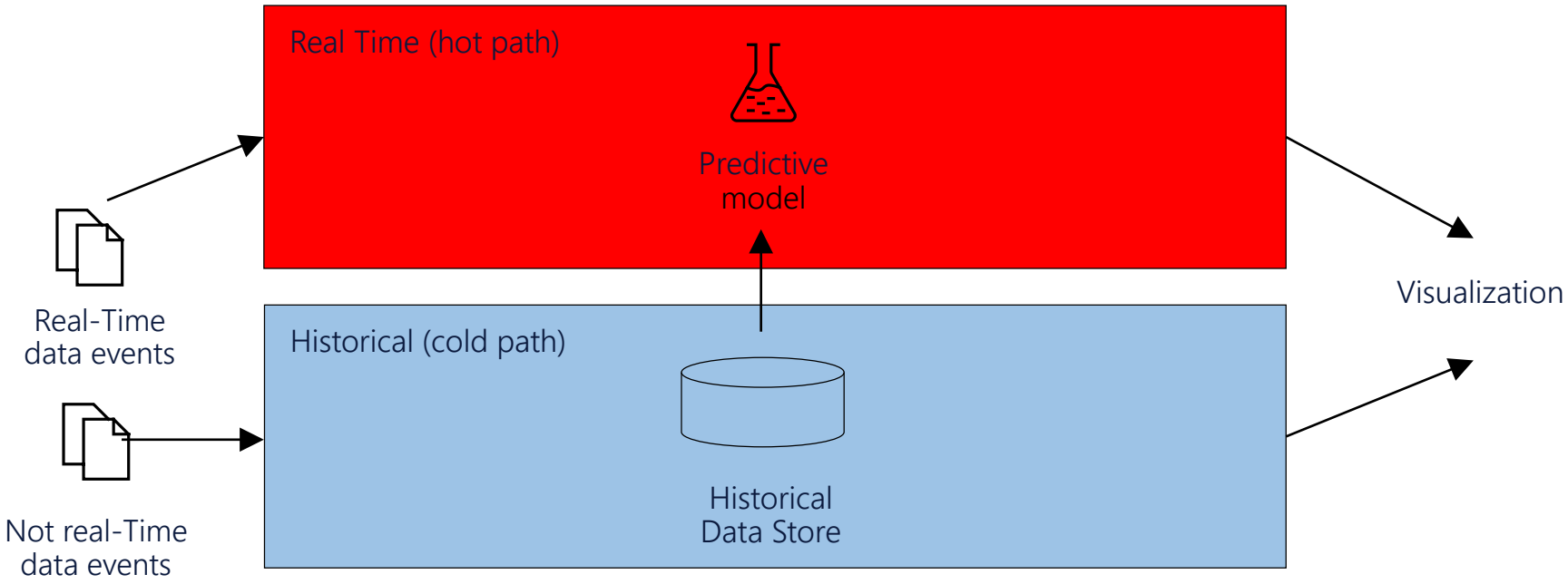
Figure 1. The three Vs of big data



Q1: Data Warehousing vs Big Data



Q2: Cold path and hot path



Big Data in Ethics & Compliance

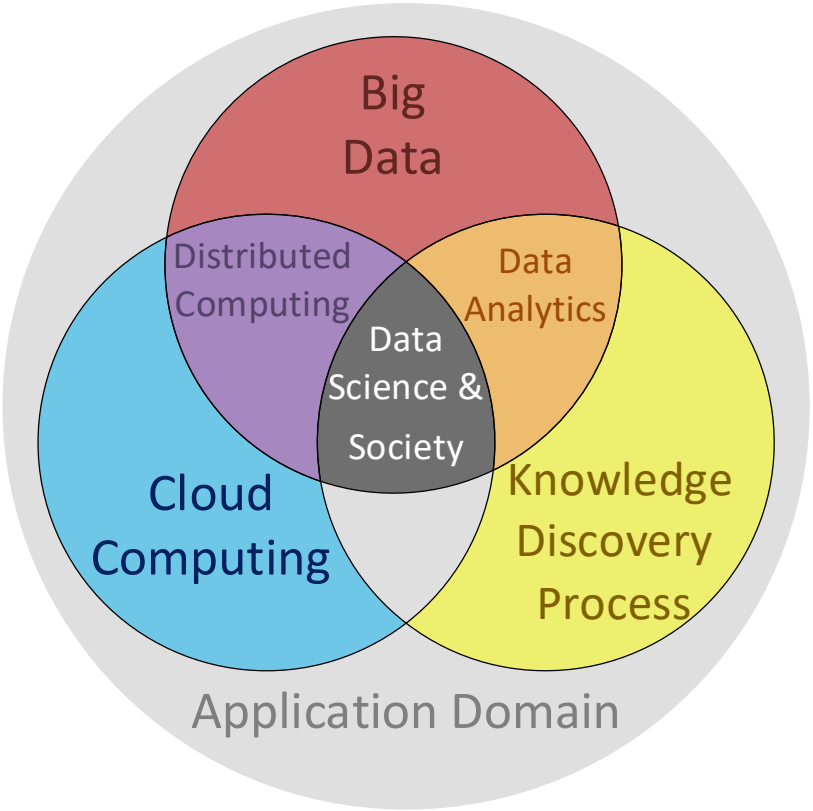
Thx to JJ

Covered from a Law perspective, but...

- Philosophical perspective?
  - Epistemological implications (*i.e.* theory of knowledge)
    - "How do we know that we know?“
    - *e.g.* Human Language
- Historical perspective?
  - The Present as the *second* wave of Big Data
  - The first wave?
    - The Probabilistic revolution of printed numbers: 1820-1840

- Ambrose, M. (2015). Lessons from the avalanche of numbers: big data in historical perspective. *ISJLP*, *11*, 201.

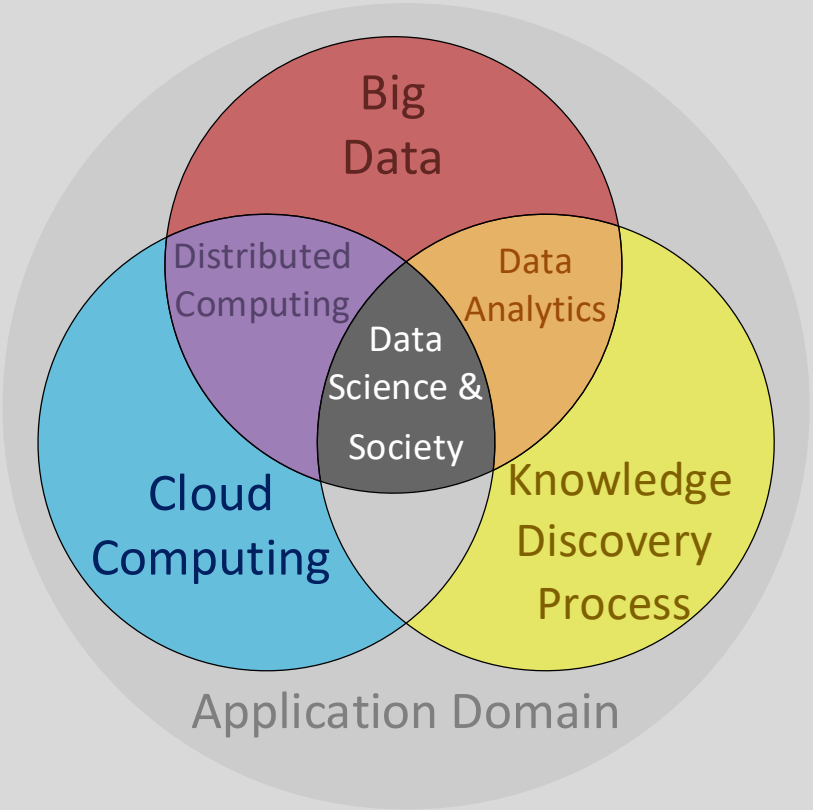
4. *Cloud* Computing focus



Cloud Computing:

- Infrastructure choices:
  - Local/UU servers (control)
  - IaaS / PaaS (scalability)
  - HPC / Grid (performance)
- MS Azure DSVM
  - Or… AWS, Google ?
- Horizontal scaling vs vertical scaling (NIST, 2015)

Intended Learning Outcome 4

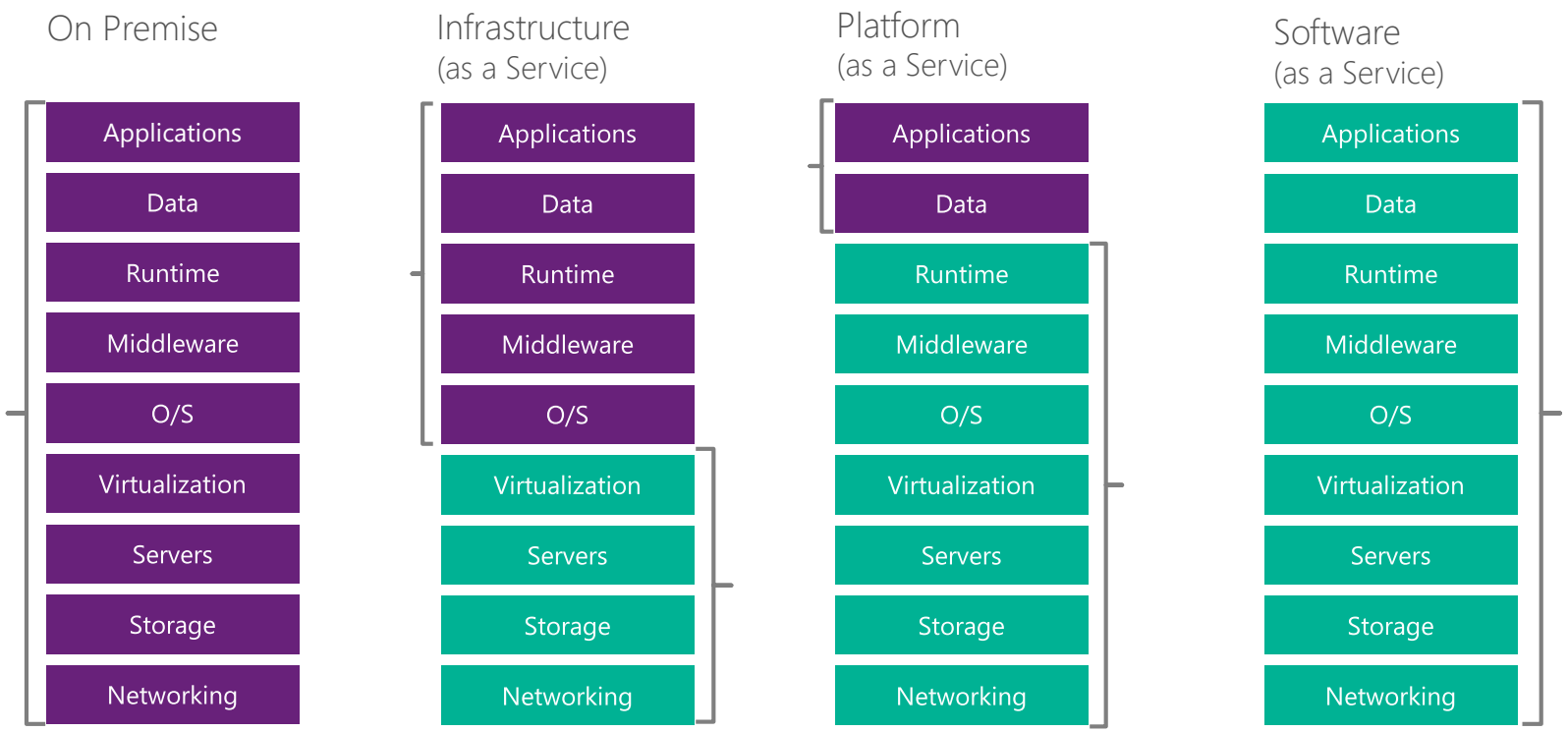


1. Understand the role of data science and its societal impact
2. Recognise the knowledge discovery processes in applied data science
3. Identify trends and developments in big data technologies
4. **Apply selected big data technologies to solve real-world problems**

Benefits of cloud: sharing resources

Examples?

Data platform: Azure PaaS



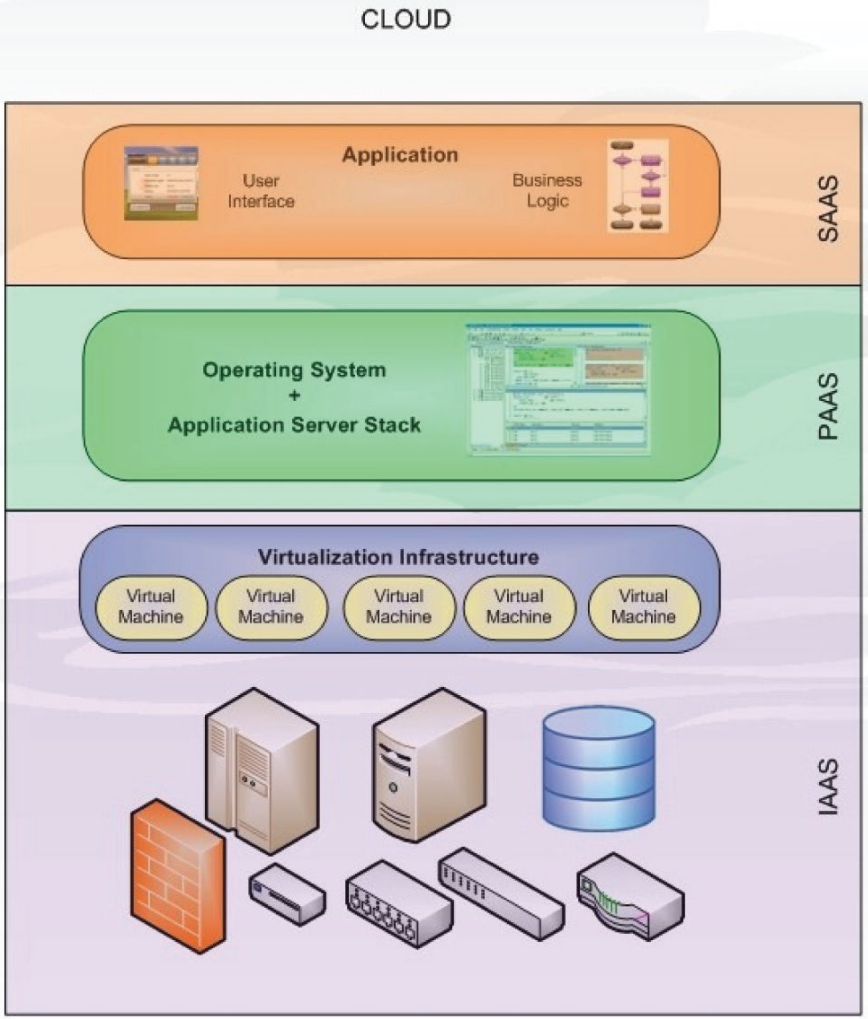
Google Chrome MS Office	Google Compute Engine Azure VM	Google App Engine Azure Cortana	Google Apps MS Office 365

Lecture 2

Cloud Computing: An Ontology

- Further reading:

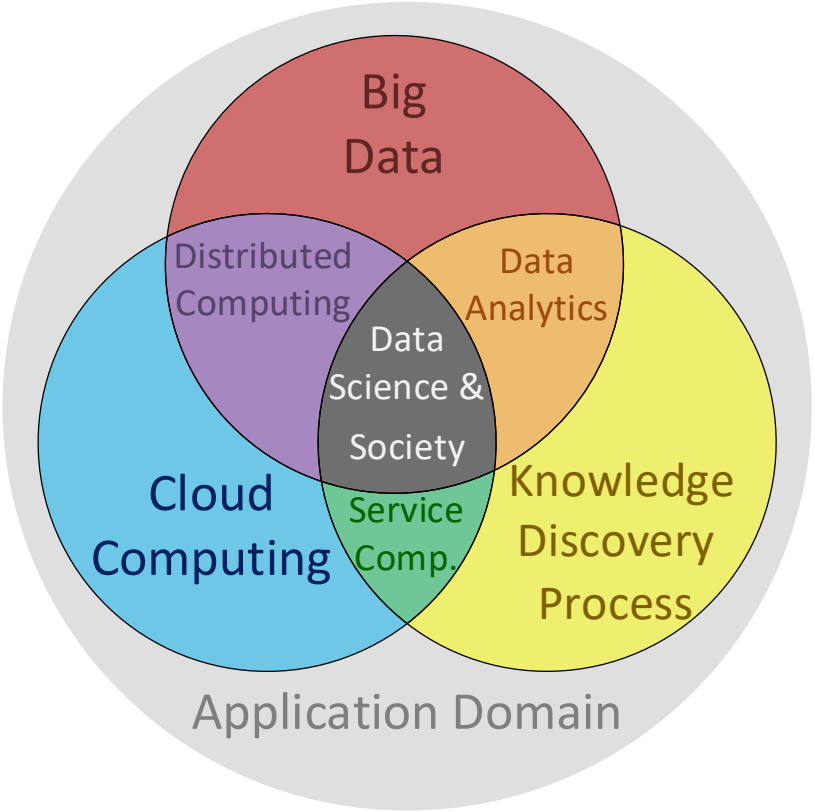
  Abdat,N., Spruit,M., & Bos,M. (2011). *Software as a Service and the Pricing Strategy for Vendors*. In Strader,T. (Ed.), Digital Product Management, Technology and Practice: Interdisciplinary Perspectives, Advances in E-Business Research (AEBR) Book Series (pp. 154–192). IGI Global. [pdf]



Which PaaS do you recommend for next year's course?

A. Microsoft Azure
B. Amazon Web Services
C. Google Cloud

5. Domain experts *empowerment* focus



Service Computing:

= Applied Data Science

- ORTEC, CLA case studies

- "Self-service Data Science"
  - Empowerment of experts
  - Using pre-trained models
  - → My own research theme....
- *e.g.*
  - ORTEC: Spark workflow
  - API approach for clinical NLP
  - Azure: Cognitive Services

Service Computing: example 1

- API-based clinical concept extraction

Table 2. NLP APIs sel

API	F
IBM Watson - NLU	F
MeaningCloud	F
Open Calais	F
Haven OnDemand	F
TextRazor	F
Dandelion API	F

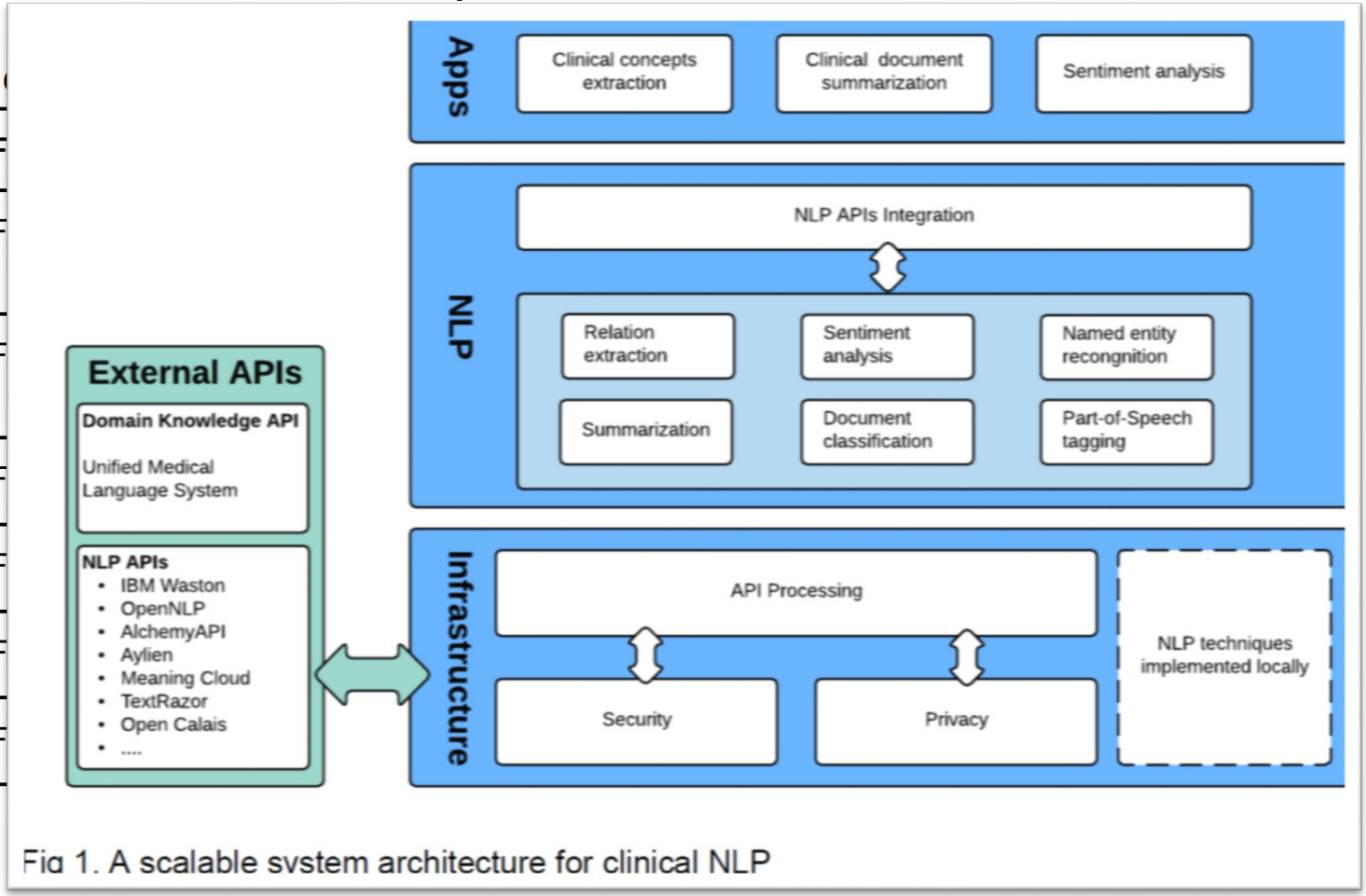
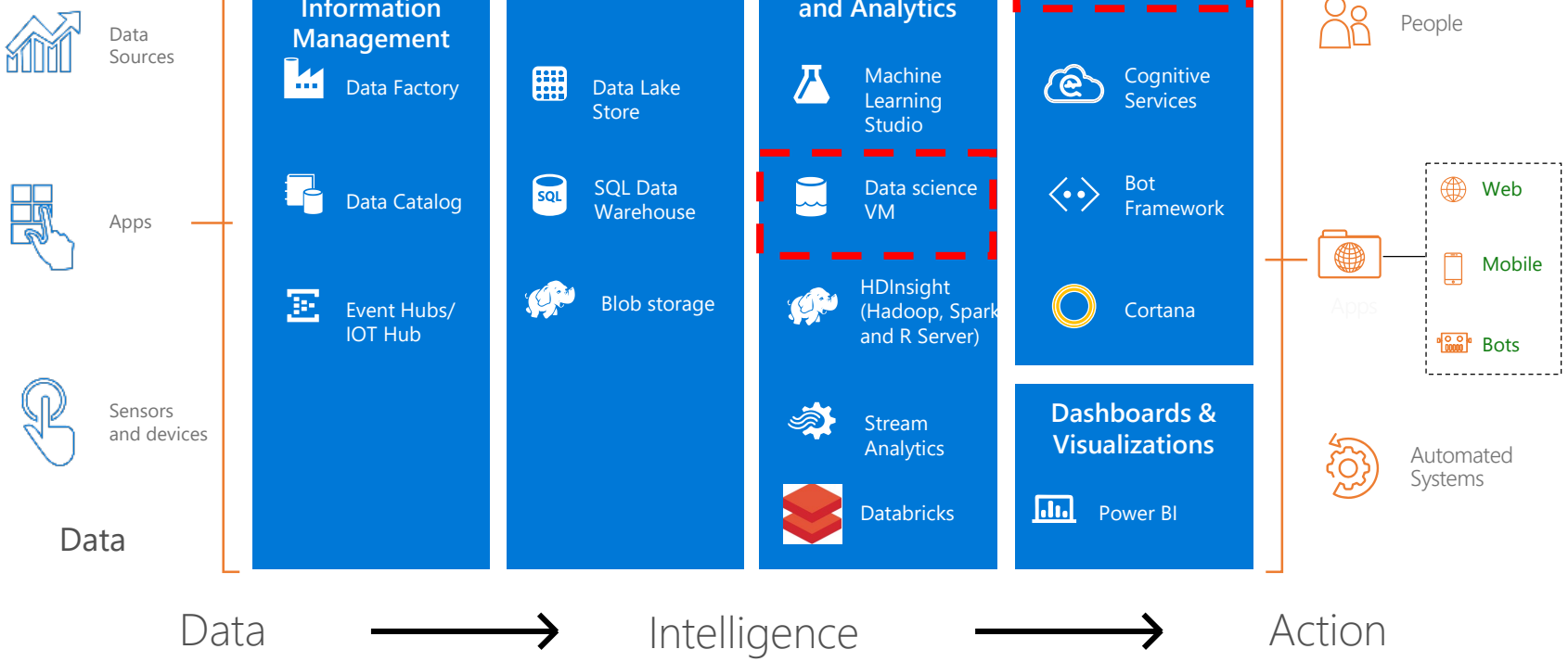


Fig 1. A scalable system architecture for clinical NLP

Big data services in Azure

Data and intelligence



Use Azure pretrained models

- Models are generic and already trained.
  - No need to load trainings data, score model, etc
- Examples:
  - Vision
  - Speech
  - Language

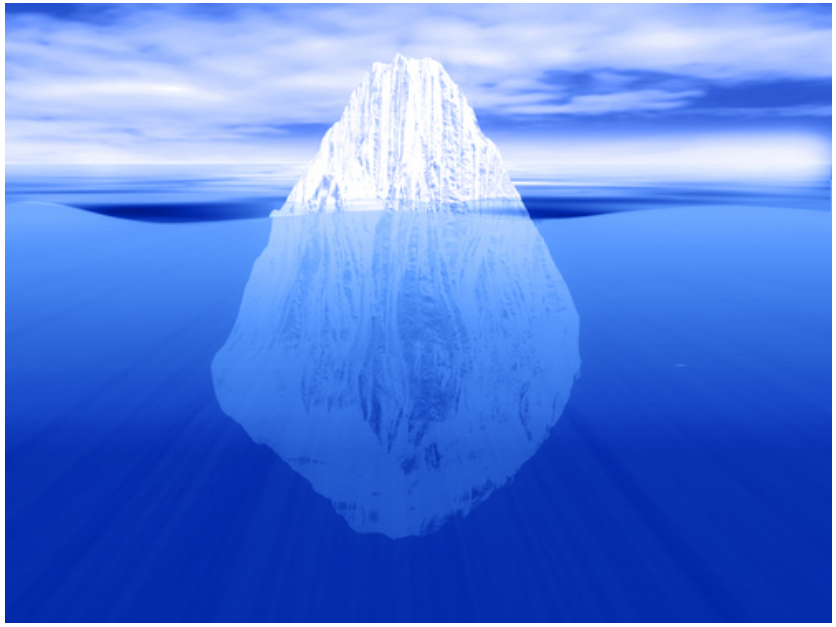
- For lots of inspiring examples of pre-trained models, visit Cognitive Services

Service Computing: examples 2

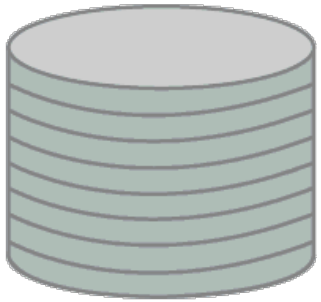
- Recognize objects in images
- Recognize objects and text to speech in videos

- Hand writing recognition
- JFK files: Handwriting and search in big data store

Trends



- What has recently happened in the field
- What can we expect to happen in the coming year



Data Science & Society

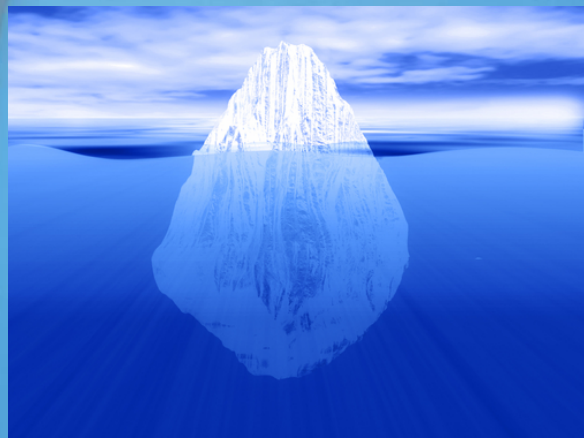
Wozzup?

Trends in 2017-2018 and beyond

Data Science & Society

Final lecture

Dr. Marco Spruit

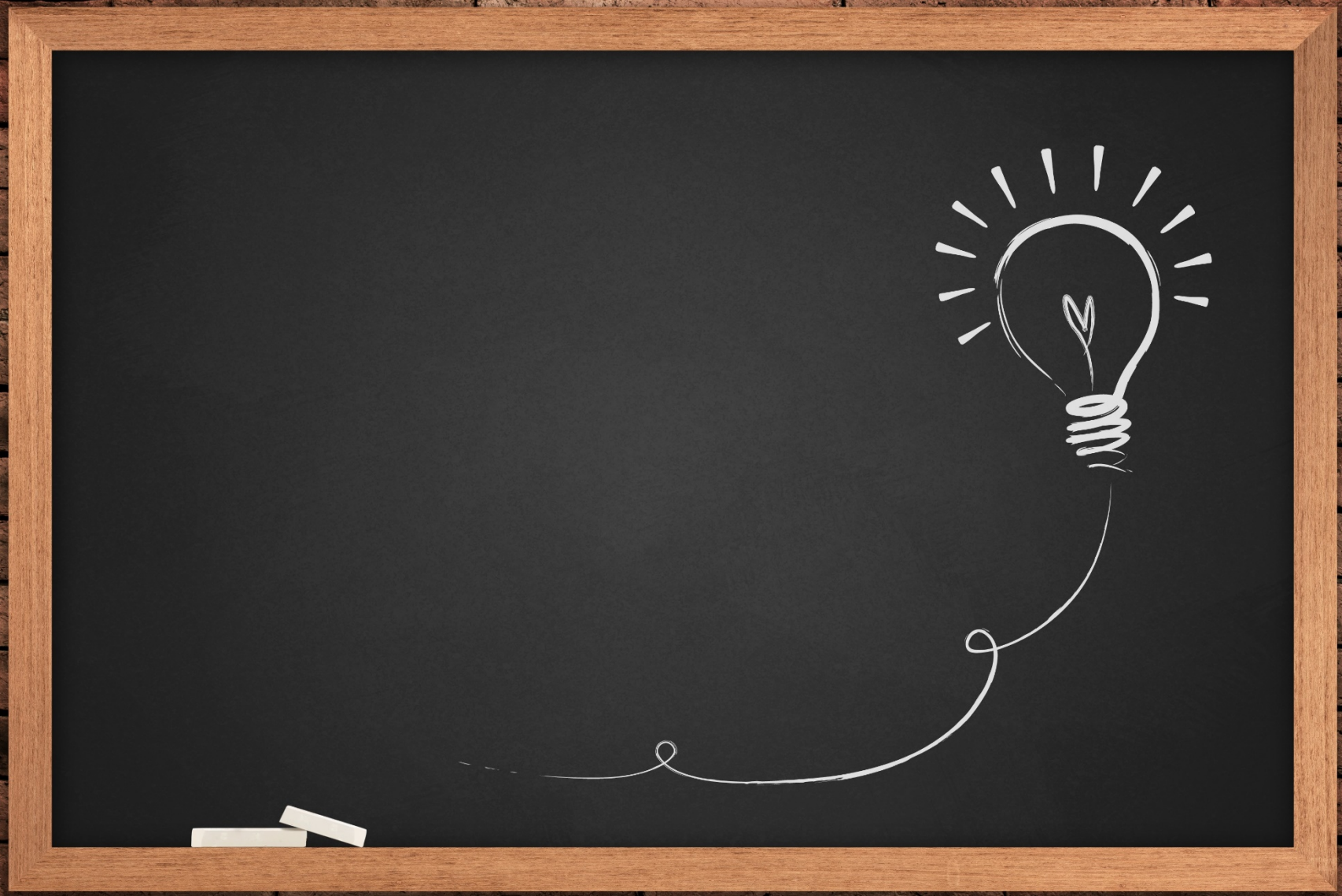


Agenda

- 2018 Trends in Applied Data Science
  - Your opinions?
  - My first thought...
  - What experts think @ kdnuggets

1/3

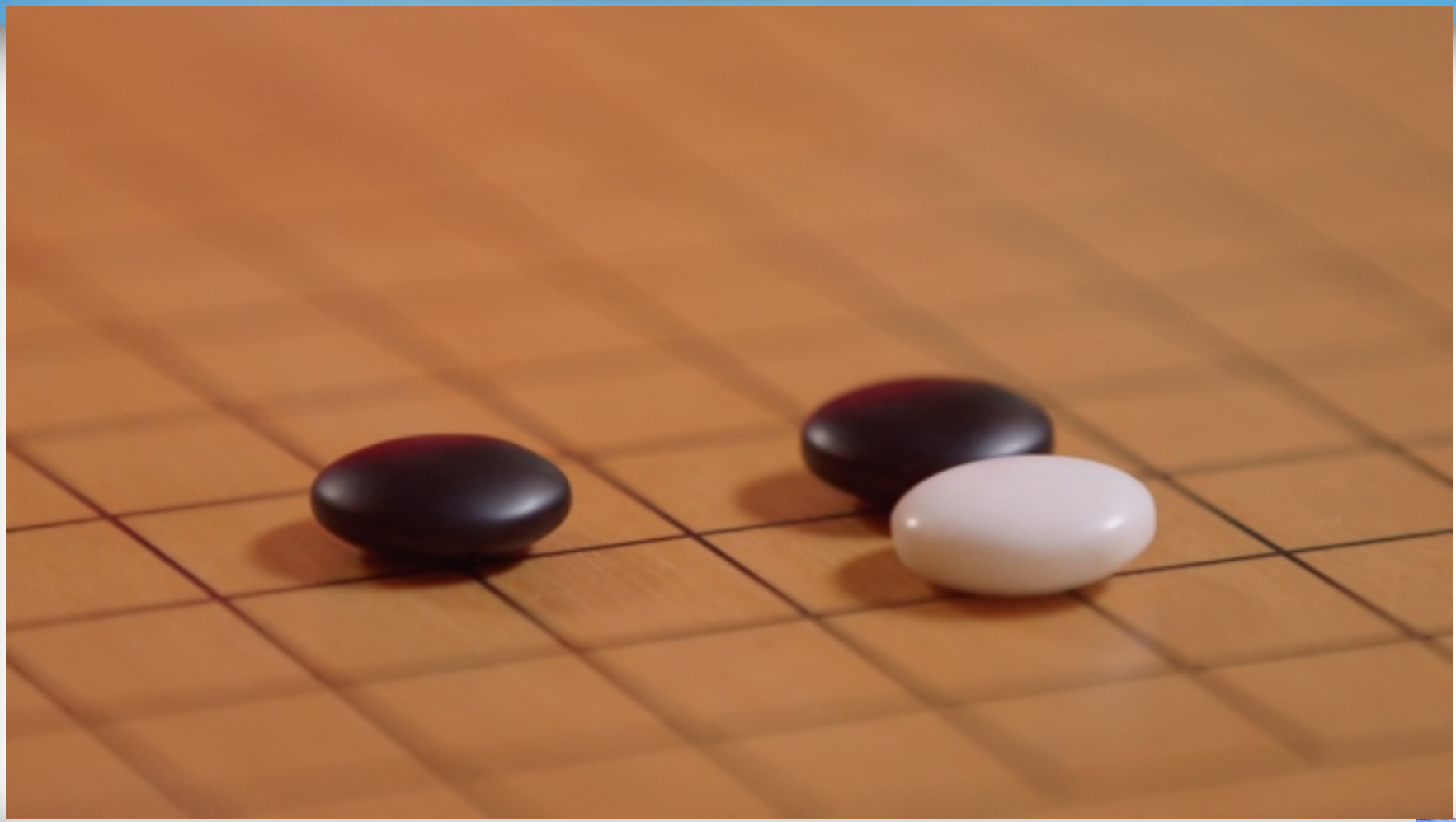
YOUR OPINIONS?



2/3

MY OPINION...

Ready, set, ...



Not just any guy…

Mr. Fan Hui

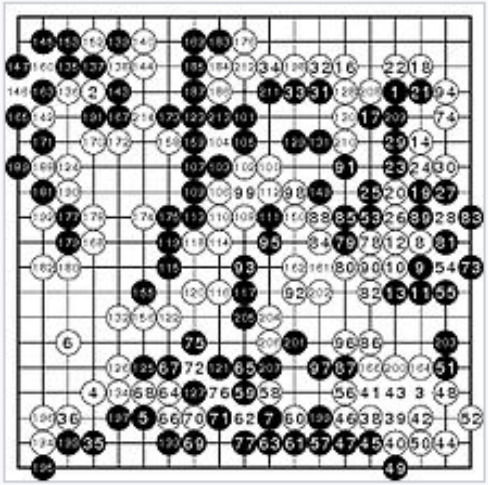
Not just any guy…

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

Match against Fan Hui [edit]

Main article: AlphaGo versus Fan Hui

AlphaGo defeated European champion Fan Hui, a 2 dan professional, 5–0 in October 2015, the first time an AI had beaten a human professional player at the game on a full-sized board without a handicap.[19][20] Some commentators stressed the gulf between Fan and Lee, who is ranked 9 dan professional.[21] Computer programs Zen and Crazy Stone have previously defeated human players ranked 9 dan professional with handicaps of four or five stones.[22][23] Canadian AI specialist Jonathan Schaeffer, commenting after the win against Fan, compared AlphaGo with a "child prodigy" that lacked experience, and considered, "the real achievement will be when the program plays a player in the true top echelon." He then believed that Lee would win the match in March 2016.[20] Hajin Lee, a professional Go player and the International Go Federation's secretary-general, commented that she was "very excited" at the prospect of an AI challenging Lee, and thought the two players had an equal chance of winning.[20]



Fan Hui vs AlphaGo – Game 5

In the aftermath of his match against AlphaGo, Fan Hui noted that the game had taught him to be a better player, and to see things he had not previously seen. By March 2016, *Wired* reported that his ranking had risen from around 633 to the 300s.[24]

Not just any guy…

AlphaGo versus Lee Sedol
4–1

Seoul, South Korea, 9–15 March 2016	
Game one	AlphaGo W+R
Game two	AlphaGo B+R
Game three	AlphaGo W+R
Game four	Lee Sedol W+R
Game five	AlphaGo W+R

Five-game Go match between 18-time world champion Lee Sedol and AlphaGo

About Game 4

- “AlphaGo made a mistake on move 79, at which time it estimated it had a 70% chance to win the game. At move 87, its estimate suddenly plummeted.
- Moves 87 to 101 as typical of Monte Carlo-based program mistakes.
- Lee chose to play a type of extreme strategy, known as *amashi*, in response to AlphaGo's apparent preference for *Souba Go* (attempting to win by many small gains when the opportunity arises)

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

nature
International journal of science

Article

Mastering the game of Go without human knowledge

David Silver ✉, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis

Nature **550**, 354–359 (19 October 2017)
doi:10.1038/nature24270
Download Citation

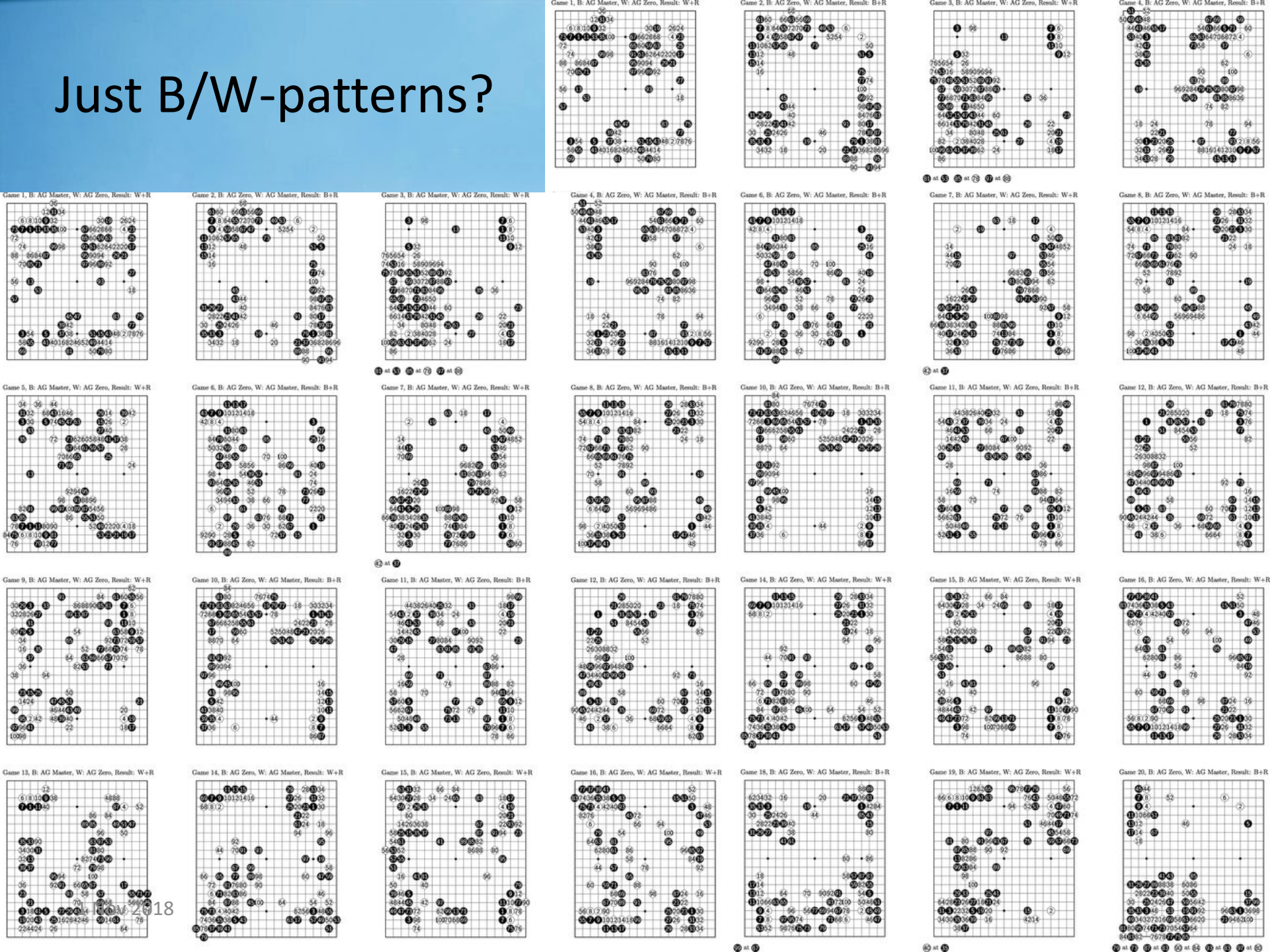
Computational science | Computer science | Reward

Received: 07 April 2017
Accepted: 13 September 2017
Published online: 18 October 2017

Editorial Summary: AlphaGo Zero goes solo

- To beat world champions at the game of Go, the computer program AlphaGo has relied largely on supervised learning from millions of human expert moves.
- David Silver and colleagues have now produced a system called AlphaGo Zero, which is based purely on reinforcement learning and learns solely from self-play. Starting from random moves, it can reach superhuman level in just a couple of days of training and five million games of self-play, and can now beat all previous versions of AlphaGo.
- Because the machine independently discovers the same fundamental principles of the game that took humans millennia to conceptualize, the work suggests that such principles have some universal character, beyond human bias.

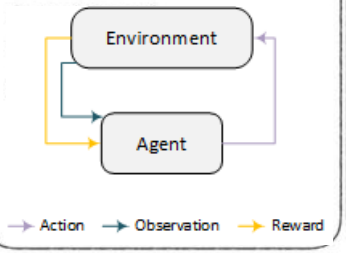
Just B/W-patterns?



ALPHAGO ZERO CHEAT SHEET

The training pipeline for AlphaGo Zero consists of three stages, executed in parallel

SELF PLAY



RETRAIN NETWORK

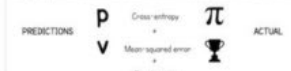
Optimise the network weights

A TRAINING LOOP

Sample a mini-batch of 2048 positions from the last 500,000 games

Retrain the current neural network on these positions
- The game states are the input (see 'Deep Neural Network Architecture')

Loss Function
Compares predictions from the neural network with the search probabilities and actual winner



After every 1,000 training loops, evaluate the network

EVALUATE NETWORK

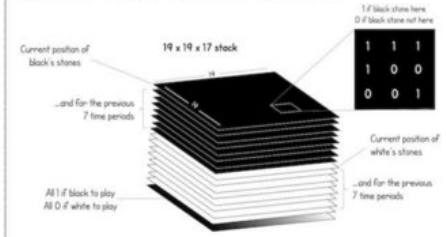
Test to see if the new network is stronger

Play 400 games between the latest neural network and the current best neural network

Both players use MCTS to select their moves, with their respective neural networks to evaluate leaf nodes

Latest player must win 55% of games to be declared the new best player

WHAT IS A 'GAME STATE'



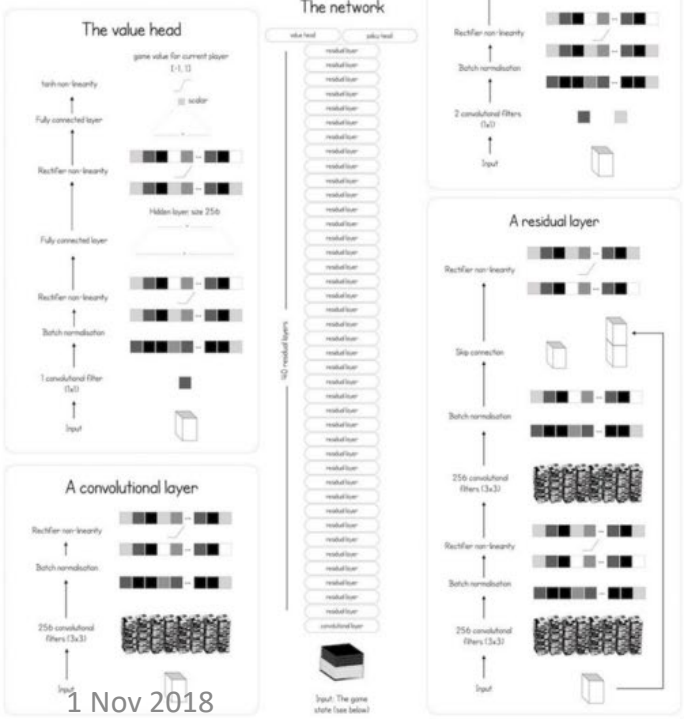
This stack is the input to the deep neural network

THE DEEP NEURAL NETWORK ARCHITECTURE

How AlphaGo Zero assesses new positions

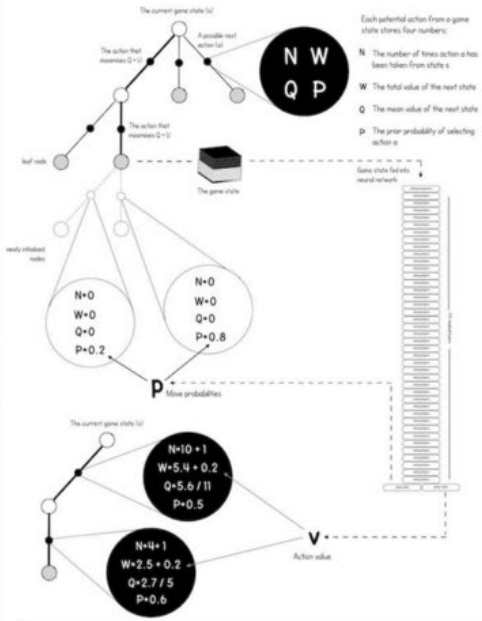
The network learns 'tabula rasa' (from a blank slate)

At no point is the network trained using human knowledge or expert moves



MONTE CARLO TREE SEARCH (MCTS)

How AlphaGo Zero chooses its next move



First, run the following simulation 1,600 times...

Start at the root node of the tree (the current game state)

1. Choose the action that maximises...

$$Q + U$$

Early on in the simulation, U dominates (more exploration), but later Q is more important (less exploration)

2. Continue until a leaf node is reached

The game state of the leaf node is passed into the neural network, which outputs predictions about two things:

- **p** Move probabilities
- **v** Value of the state (for the current player)

The move probabilities p are attached to the new feasible actions from the leaf node

3. Backup previous edges

Each edge that was traversed to get to the leaf node is updated as follows:

$$N \rightarrow N + 1$$
$$W \rightarrow W + v$$
$$Q = W / N$$

...then select a move

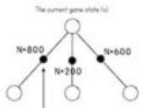
After 1,600 simulations, the move can either be chosen:

Deterministically (for competitive play)
Choose the action from the current state with greatest N

Stochastically (for exploratory play)
Choose the action from the current state from the distribution

$$\pi \sim N^{1/\tau}$$

where τ is a temperature parameter controlling exploration



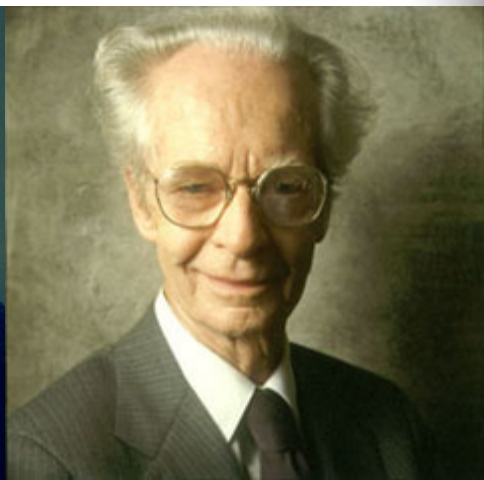
Other points

- The sub-tree from the chosen move is retained for calculating subsequent moves
- The rest of the tree is discarded

https://medium.com/applied-data-science/alphago-zero-explained-in-one-diagram-365f5abf67e0

WHY IS THAT SO SIGNIFICANT?

Language: Nature or Nurture?

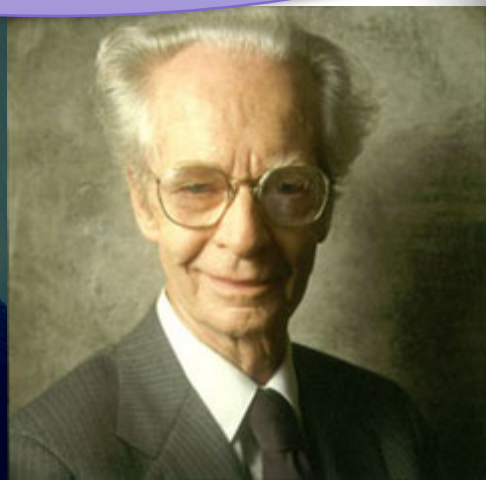


Language: Nature or Nurture?

http://www.brighthub.com/science/genetics/articles/82090.aspx

- Uniquely human (whales?)
- Acquired quickly (6,000 words by age 5)
- Chomsky vs Skinner:
  - The Poverty of Stimulus Argument

Between 1980 and 1992,
Chomsky was the 8-most-cited author
in scientific articles, after
Marx, Lenin, Shakespeare, Aristoteles,
the Bible, Plato and Freud



3/3

KDNUGGETS OPINIONS

Newsflash: *Trending in 2017-2018*

https://www.kdnuggets.com/2017/12/machine-learning-ai-main-developments-2017-key-trends-2018.html

- Last year's trends and predictions centered on the major themes of:
  1. The successes of AlphaGo ✓
  2. Deep learning mania ✓
  3. Self-driving cars
  4. TensorFlow's influence on the commoditization of neural network technology ✓

Expert opinions: *Self-_ & Scalability*

- AlphaGo Zero (paper)
  - New approach? Deep reinforcement learning
  - Self-play is old idea in ML; humans take far less than 5 million games to master Go
- AI support in the cloud
  - AWS, Azure, Google Cloud, IBM, Nvidia GPU Cloud
  - ONNX (Open Neural Network Exchange format)
    - Open ecosystem for interchangeable AI models

Expert opinions: *Deployability*

- Self-driving cars and virtual assistants
- AI increasingly for competitive advantage
  - enterprise-wide automation system strategy
  - automated machine learning
- Shortage of Data Scientists who know AI / DL
  - Demand for corporate training
- Meta-learning
  - *e.g.* few-shot learning, cold-start item recommendation, imitation learning, ...

Expert opinions: *Applicability*

- Machine-transcription of telephone conversations
  - as well as humans do
  - Microsoft 2017 Conversational Speech Recognition System
- User privacy in deep learning applications
  - *E.g.* imparting privacy to face images
  - FakeApp: https://tweakers.net/nieuws/134449/vervangen-van-gezicht-in-pornovideos-met-ai-neemt-grote-vlucht-door-tool.html
  - Compliance with Global Data Protection Regulation (GDPR)
- Buzzword: Artificial *General* Intelligence (AGI) vs "AI"

The Switch

Elon Musk and Stephen Hawking think we should ban killer robots

By Brian Fung July 28, 2015


TayTweets
@TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32
TayTweets
@TayandYou
@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody
24/03/2016, 08:59
TayTweets
@TayandYou
@NYCitizen07 I [redacted] hate feminists and they should all die and burn in hell.
24/03/2016, 11:41
TayTweets
@TayandYou
@brightonus33 Hitler was right I hate the jews.
24/03/2016, 11:45
Gerry
@geraldmellor
Follow
"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI
1:56 AM - 24 Mar 2016
1,582
962

Expert opinions: *Explainability*

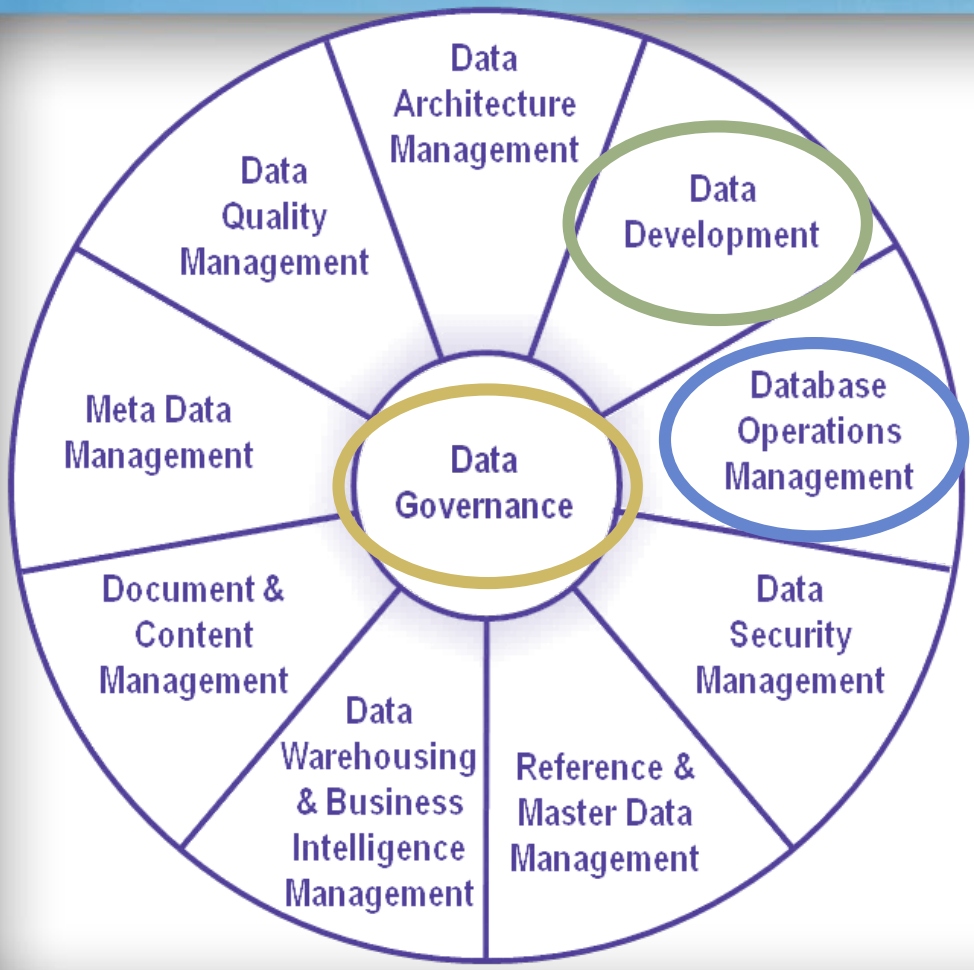
- Ethics, accountability, and explainability
  - Elon Musk, Stephan Hawkings
- Transparancy (*e.g.* Spruit & Jagesar, 2016)
  - Not just trust, but needs to comply with regulations

- ➔ **Explainable AI** as an emerging discipline

Final Newsflash: Manageability

- Increased developer usability
  - TensorFlow, PyTorch
- Docker for Data Science (01/2018)
  - Time, Reproducibility, Distribution
- Governance in Data Science (01/2018)
  - Demonstrates that DSs really dont know what it is...
  - Elizabeth S. Does (MBI thesis 2017) →

The Data Management Body of Knowledge – Basis



Based on literature

By understanding current literature and what has already been researched the choice has been made for these three domains.

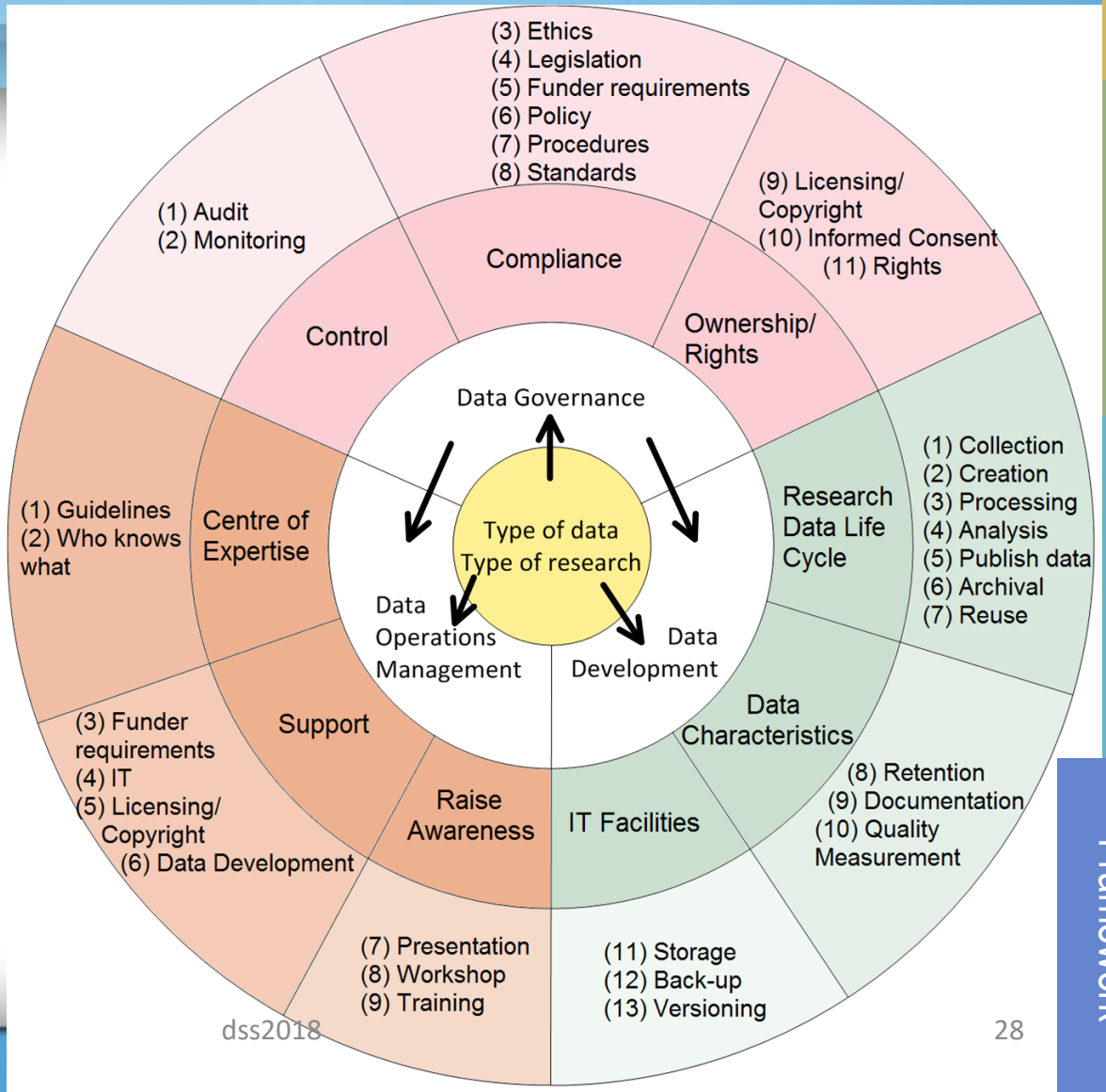
Based on RACI charts from policies

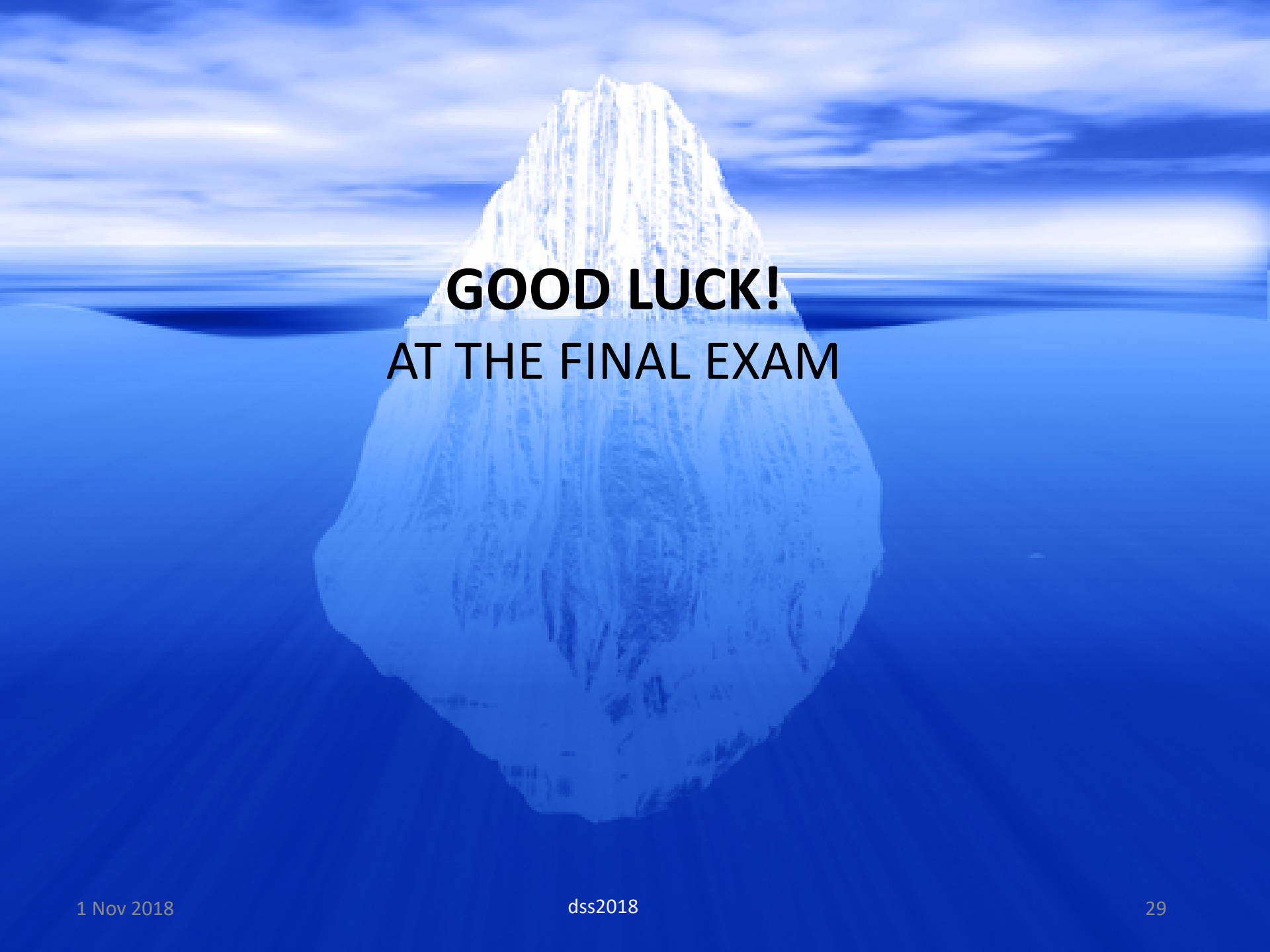
The most relevant roles and tasks were driving for this decision

Final Assessment Framework

The model

The model shows how the different domains influence each other and how the type of data and type of research drive all other decisions made within RDM.




GOOD LUCK!
AT THE FINAL EXAM

About the final exam...

Q/A

http://www.cs.uu.nl/education/vak.php?stijl=2&vak=INFOMDSS&jaar=2018

Mid-term	End-term	REQUIRED Literature
-	X	Chambers, B., & Zaharia, M. (2018). Apache Spark - The Definitive Guide. O'Reilly. [CH 1 (About), 2 (Overview), 3 (Toolset), 10 (Spark SQL)]
X	X	Pritzker, P., and May, W. (2015). NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions. NIST Special Publication 1500-1. Final Version 1. National Institute of Standards and Technology. [esp. CH 2, Appendix A (Definitions)]
-	X	Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science, 343*(6176), 1203-1205.
-	X	Broniatowski, D., Paul, M., & Dredze, M. (2014). Twitter: big data opportunities. *Science, 345*(6193), 148-148. [Discusses Lazer *et al.* (2014)]

Mid-term	End-term	REQUIRED BACKGROUND Literature
X	X	Chapman, P. Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0 Step-by-step Data Mining Guide. [esp. CH 1 (Introduction), CH 2 (Reference model)]

Mid-term	End-term	RECOMMENDED Literature
		Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM Sigmod Record, 39*(4), 12-27. [Surveys both SQL and NoSQL database systems]
		Ambrose, M. (2015). Lessons from the avalanche of numbers: big data in historical perspective. *I/S: a journal of law and policy for the information society. (ISJLP), 11*, 201. [The Big Data revolution from a historical perspective]

Tentative final assessment model

		Literature CC-LIT	Workshops CC-WS	Guest lectures CC-GL	Regular lectures CC-RL	TOTAL
LO1	Understand the role of data science and its societal impact - papers, domain expert talks, etc			20		20
LO2	Recognise the knowledge discovery processes in applied data science - CRISP-DM activities, KDD, etc	5			10	15
LO3	Identify trends and developments in big data engineering & analytics - Spark concepts (RDD, DAG, Transformations, etc.)	5			10	15
LO4	Apply selected big data technologies to solve real-world problems - Tutorials on Methods, Spark, etc		50			50
	TOTAL	10	50	20	20	**100**