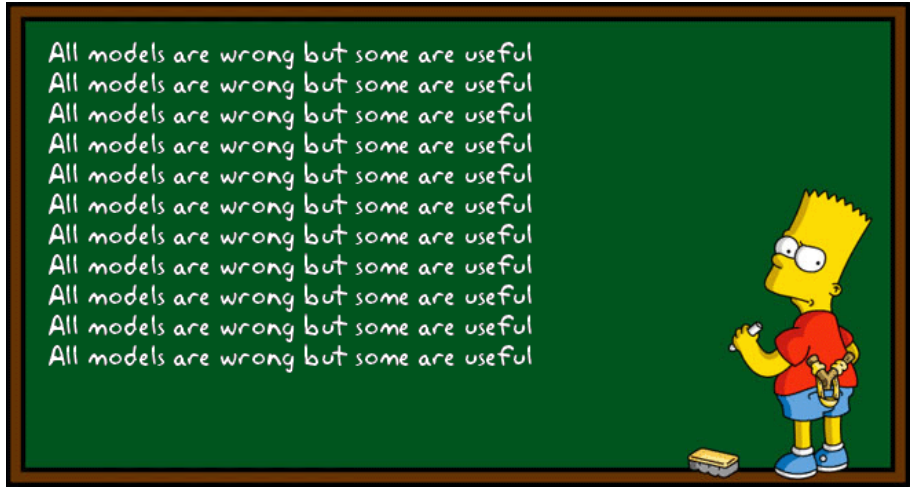
Lessons learned from 16 applied data science (meta) case studies

Zurich University of Applied Sciences

zh aw

fdal 2018 conference on industrial applied data science, Lugano, Oct 18-19, 2018

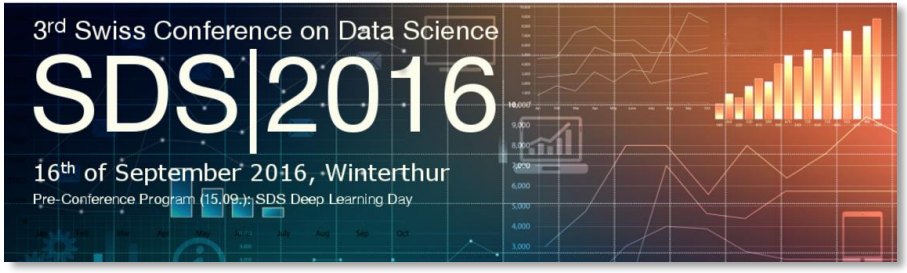
Kurt Stockinger & Thilo Stadelmann



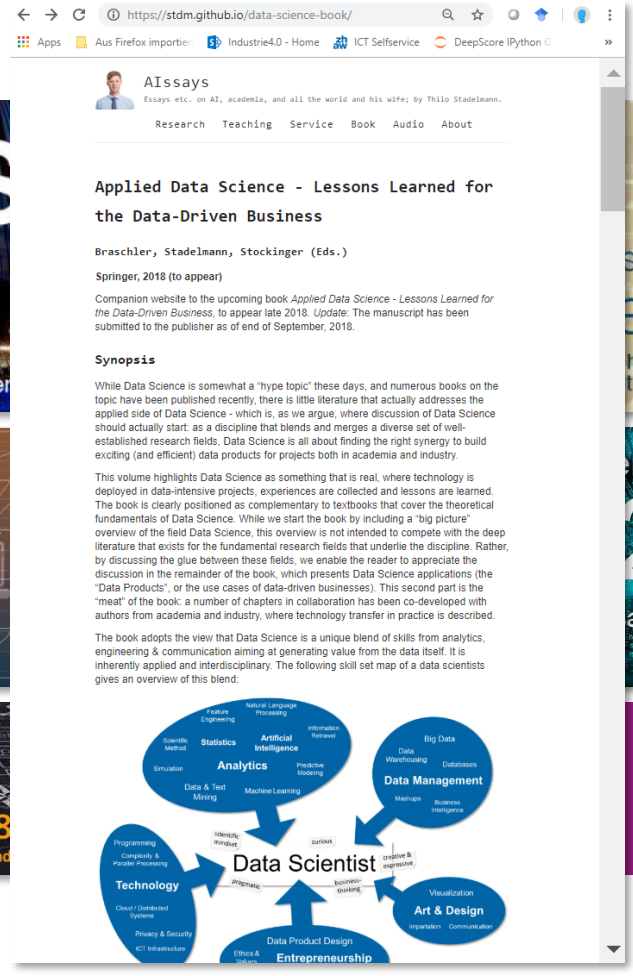
datalab
www.zhaw.ch/datalab

Collecting lessons learned from half a decade of data science

Zurich University of Applied Sciences

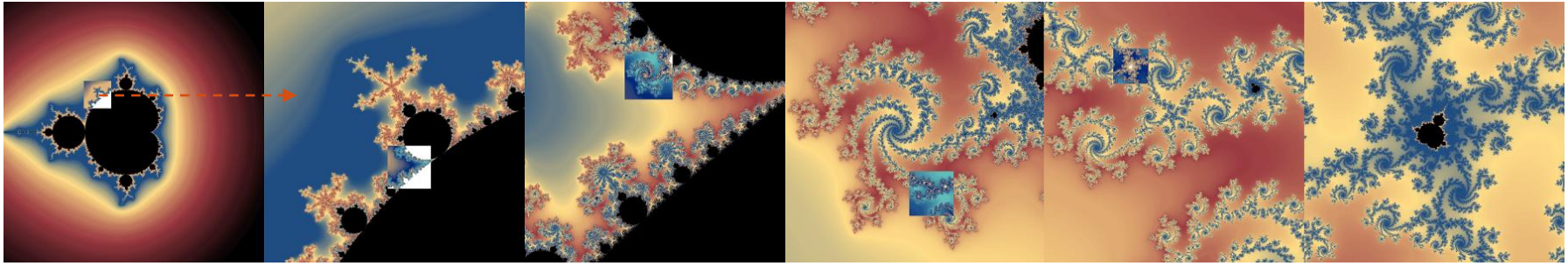


Collecting lessons learned from half a decade of data science



Agenda

- The study
- Checklist: Eight commandments
- Inspiration: methodology, technology, innovation, education



The study

16 contributions, spanning much of data science

Taxonomy	Discussed in chapters															
Main focus	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Fundamentals of Data science	x	x	x													
Methodology or algorithm				x	x	x	x	x	x	x			x			x
Tool							x		x						x	
Application		x	x							x	x	x	x	x	x	x
Survey or tutorial				x	x			x			x					
Stages in knowledge discovery process	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Data recording	x							x	x	x				x	x	x
Data wrangling	x				x			x			x	x			x	
Data analysis	x			x	x	x	x	x	x	x	x	x	x	x		x
Data visualization and/or interpretation	x		x	x			x				x	x				x
Decision making	x		x				x			x			x	x		
Competence area	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Technology								x	x					x	x	x
Analytics				x	x	x			x	x	x	x	x	x		x
Data Management						x	x		x		x	x		x	x	x
Entrepreneurship		x	x					x								
Communication							x					x				

Taxonomy	Discussed in chapters															
Data modalities	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Numerical data	x	x	x	x		x	x	x	x	x	x	x	x			
Text					x	x	x			x	x					
Images	x			x												x
Audio				x												
Time series	x		x	x		x					x					
Transactional data							x			x	x	x				
Open data						x										
Application domain	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Research	x		x	x	x	x		x						x		
Business		x			x	x		x	x				x			
Biology				x												x
Health	x		x			x					x				x	x
eCommerce and retail		x					x			x		x				
Finance		x											x			
IT							x									
Industry and manufacturing				x					x							
Services	x	x		x							x					

✓ Eight commandments

✓ Eight commandments

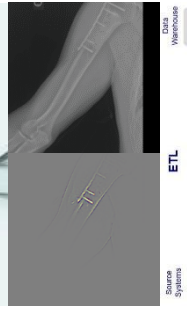
1. DO: **embrace interdisciplinarity**, seek knowledge exchange

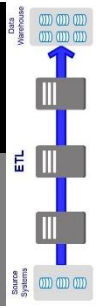
✓ Eight commandments

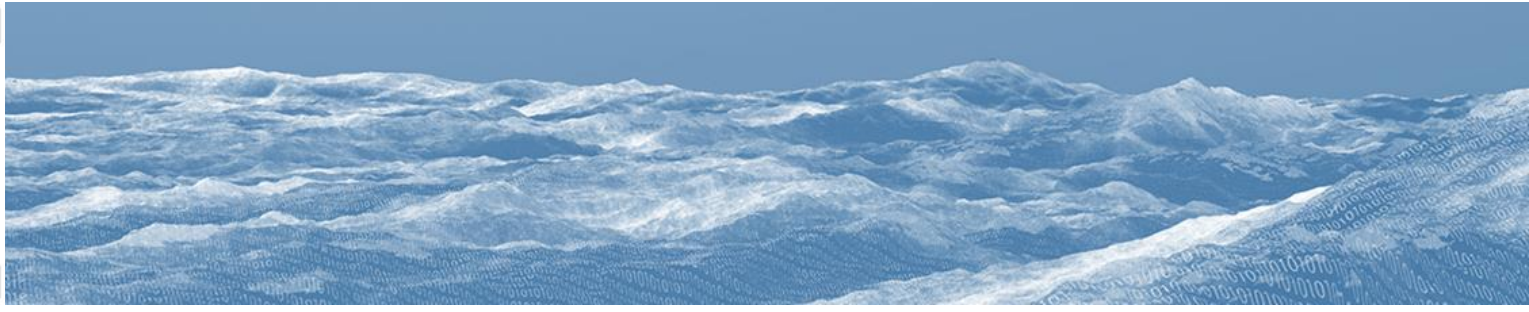
1. DO: **embrace interdisciplinarity**, seek knowledge exchange
2. DO: **build trust** by data usage transparency & security provisions

✓ Eight commandments

1. DO: **embrace interdisciplinarity**, seek knowledge exchange
2. DO: **build trust** by data usage transparency & security provisions
3. DO: **cherish data wrangling**, ideally automate it → it's the basis for analysis





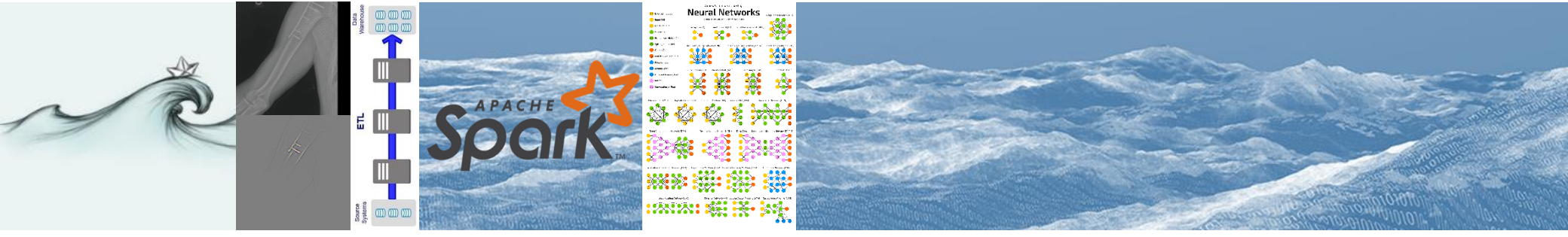


✓ Eight commandments

1. DO: **embrace interdisciplinarity**, seek knowledge exchange
2. DO: **build trust** by data usage transparency & security provisions
3. DO: **cherish data wrangling**, ideally automate it → it's the basis for analysis
4. DO: **leverage stream processing** tools for real time big data analysis

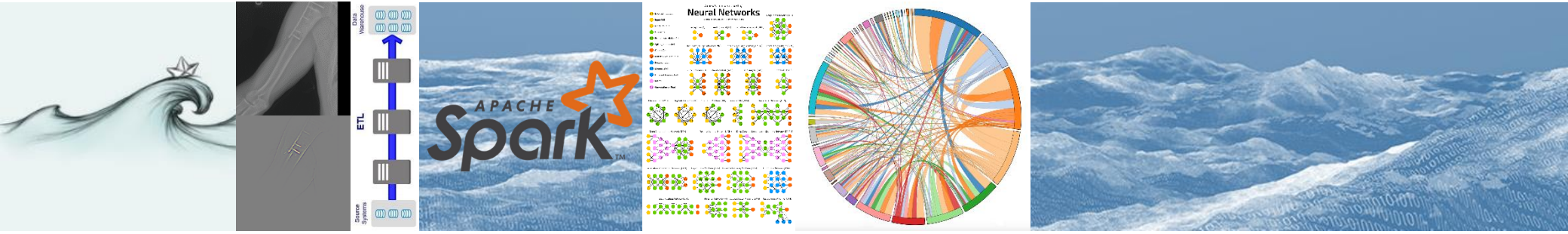
✓ Eight commandments

1. DO: **embrace interdisciplinarity**, seek knowledge exchange
2. DO: **build trust** by data usage transparency & security provisions
3. DO: **cherish data wrangling**, ideally automate it → it's the basis for analysis
4. DO: **leverage stream processing** tools for real time big data analysis
5. DO: **start** machine learning **from simple baselines**



✓ Eight commandments

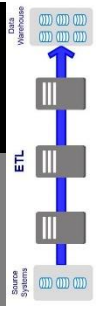
1. DO: **embrace interdisciplinarity**, seek knowledge exchange
2. DO: **build trust** by data usage transparency & security provisions
3. DO: **cherish data wrangling**, ideally automate it → it's the basis for analysis
4. DO: **leverage stream processing** tools for real time big data analysis
5. DO: **start** machine learning **from simple baselines**
6. DO: **use visualization** to gain insight (from debugging to result presentation)

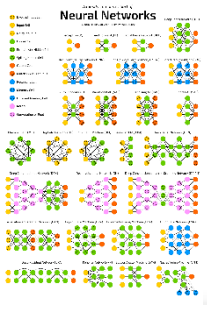


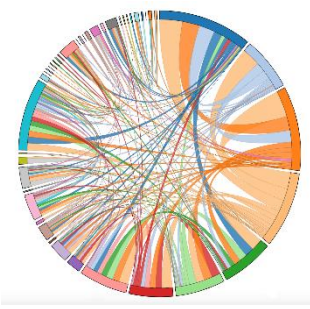
✓ Eight commandments

1. DO: **embrace interdisciplinarity**, seek knowledge exchange
2. DO: **build trust** by data usage transparency & security provisions
3. DO: **cherish data wrangling**, ideally automate it → it's the basis for analysis
4. DO: **leverage stream processing** tools for real time big data analysis
5. DO: **start** machine learning **from simple baselines**
6. DO: **use visualization** to gain insight (from debugging to result presentation)
7. DO: **make use of all of your data** (no sampling necessary)





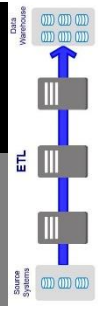


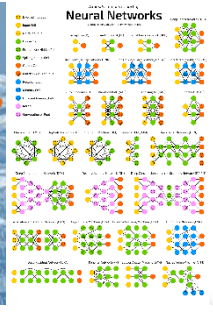


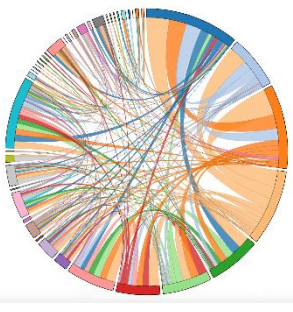

✓ Eight commandments

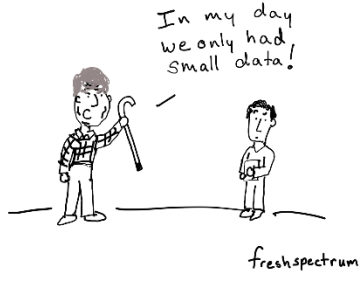
1. DO: **embrace interdisciplinarity**, seek knowledge exchange
2. DO: **build trust** by data usage transparency & security provisions
3. DO: **cherish data wrangling**, ideally automate it → it's the basis for analysis
4. DO: **leverage stream processing** tools for real time big data analysis
5. DO: **start** machine learning **from simple baselines**
6. DO: **use visualization** to gain insight (from debugging to result presentation)
7. DO: **make use of all of your data** (no sampling necessary)
8. DO: **take special care of small data** (because of less redundancies)









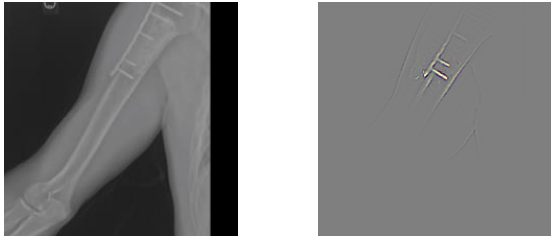
Inspiration #1: methodology

Make intuitive model inspection & data visualization “always on”

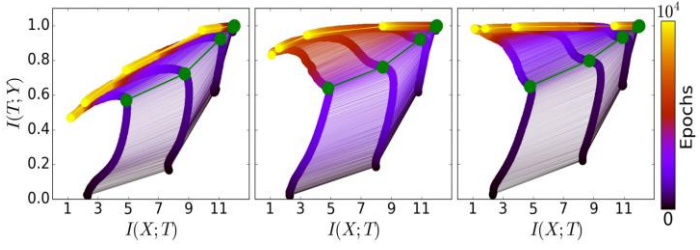
- **Building trust** with stakeholders

negative X-ray

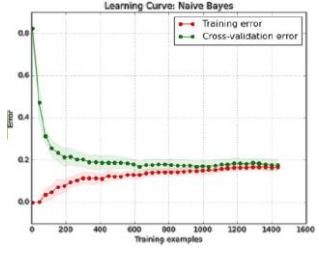
positive X-ray



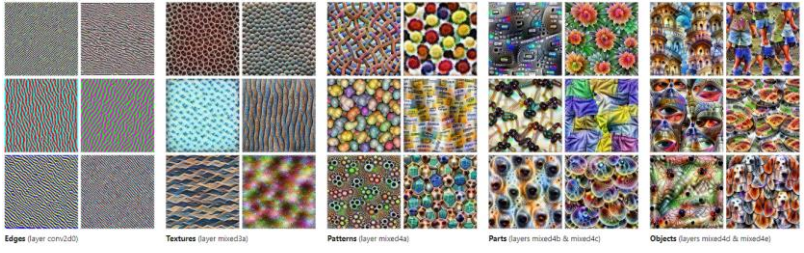
- **Debugging** capabilities for researchers & developers



DNN training on the Information Plane



a learning curve

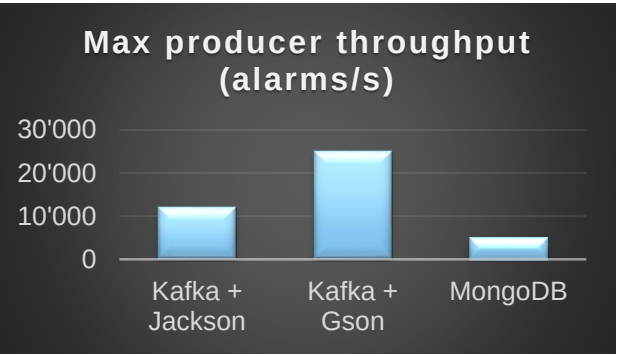


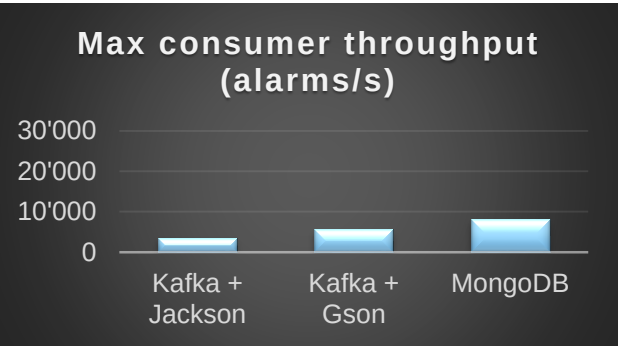
feature visualization

Inspiration #2: technology

Understand influences on big data system performance

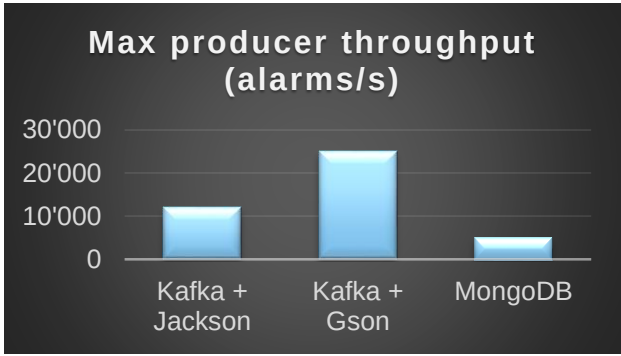
- Modern big data systems make parallel programming easy
- However, the complex distributed components need careful performance analysis & tuning to arrive at state of the art results:

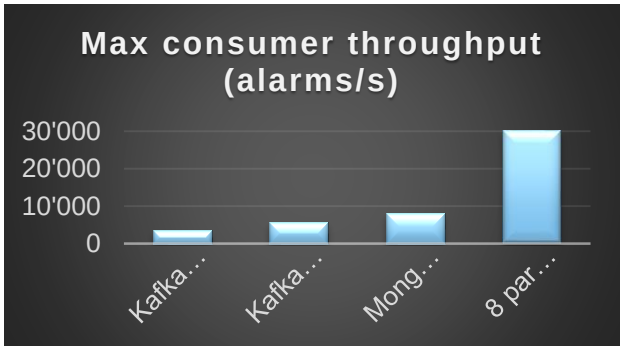




Configuring the Kafka Direct Stream in Apache Spark with proper settings…

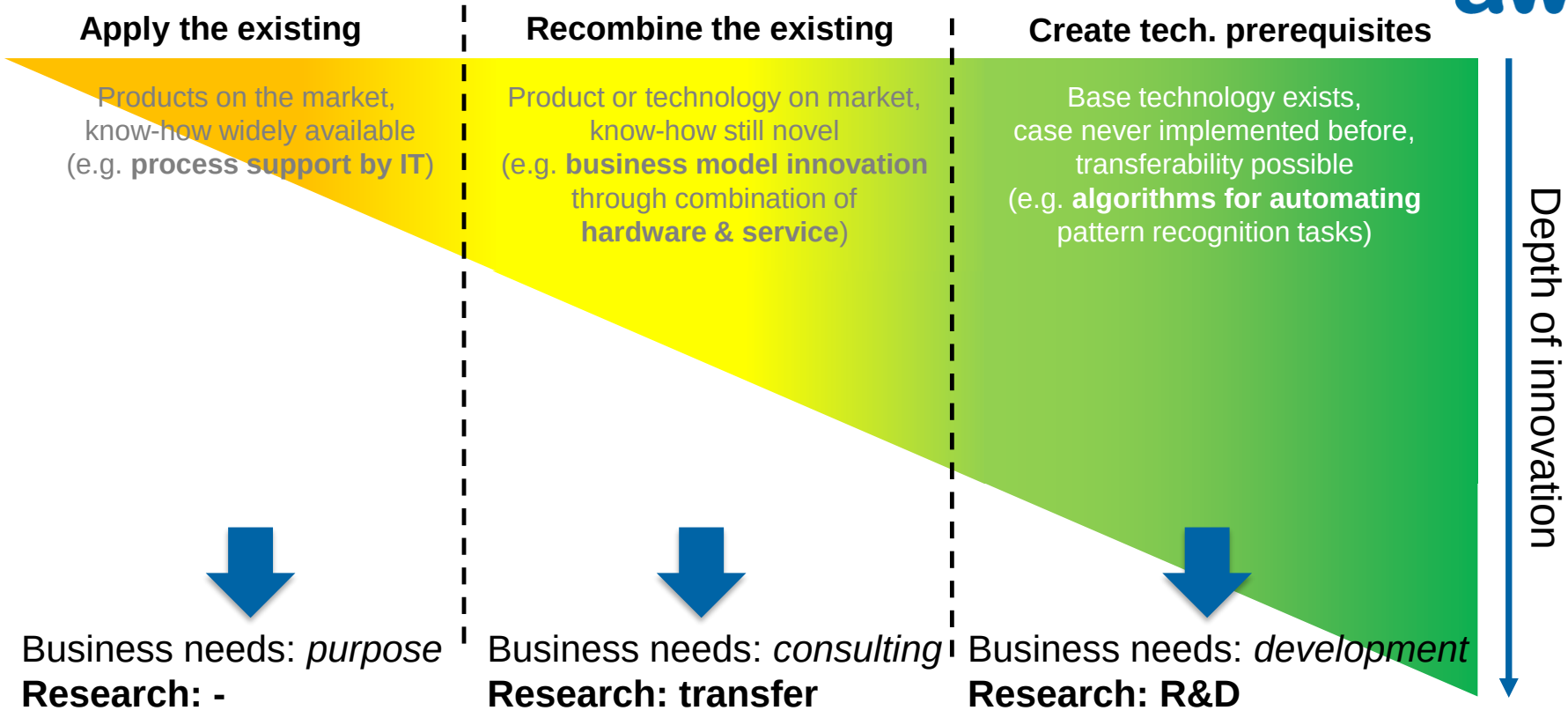
(num partitions = num cores)





Inspiration #3: innovation

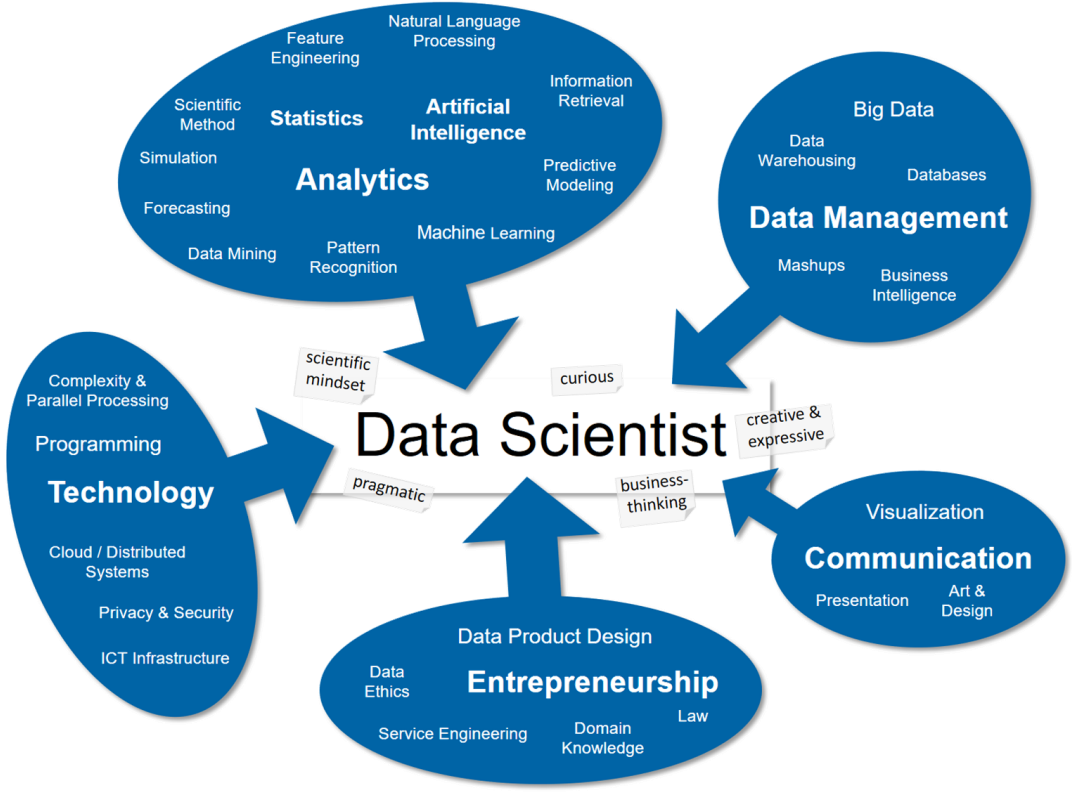
Use networks of experts to leverage different levels of innovation



Inspiration #4: education

Build interdisciplinary skills & experience on top of solid foundation

- Disciplinary **bachelor establishes foundation** in a constituting field
- Data science education imparts **core methods, tools, and project experience**



Conclusions

- **Crucial digital innovation needs to happen at the level of society**: how do we deal with the opportunities *“making sense of data”* is giving us?

Swiss Alliance for Data-Intensive Services

datalab
www.zhaw.ch/datalab

On me:
- Prof. AI/ML, head ZHAW Datalab, board Data+Service Alliance
- thilo.stadelmann@zhaw.ch
- +41 58 934 72 08
- https://stdm.github.io/

On the topics:
- Data science @ ZHAW: www.zhaw.ch/datalab
- Data science in CH: www.data-service-alliance.ch
- Applied data science book: https://stdm.github.io/data-science-book/

➔ Happy to answer questions & requests.