Data Science for Art History:
Introduction to Cloud Computing

#DAHSS20

Harald Klinke

- Before: Local R with local data -> data analysis
- Last year: Tweet bot for public engagement
- This year: Covid19, all-online Summer School, we move to the cloud
- Where is the data? In the cloud
- Where is processing power? In the cloud

What we do

- Theory and practice.
- DAH is about experience.

- We talk about some principles and key terms
- We look at a few examples
- We try out stuff
- We create

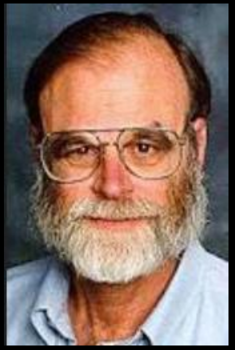
Wednesday **Intro to elements**	**Thursday** **Add cultural data**	**Friday** **Develop frontends**	**Saturday** **Prepare presentations**
Virtual coffee	Virtual coffee	Virtual coffee	Virtual coffee
Class*	**Class**	**Class**	**Class**
Experience	**Experience**	**Experience**	**Experience**
Discussion, prepare presentation, tasks for next day	**Discussion, prepare presentation, tasks for next day**	**Discussion, prepare presentation, tasks for next day**	**Discussion, prepare presentation**
Panel presentation	Panel presentation	Panel presentation	Public presentation

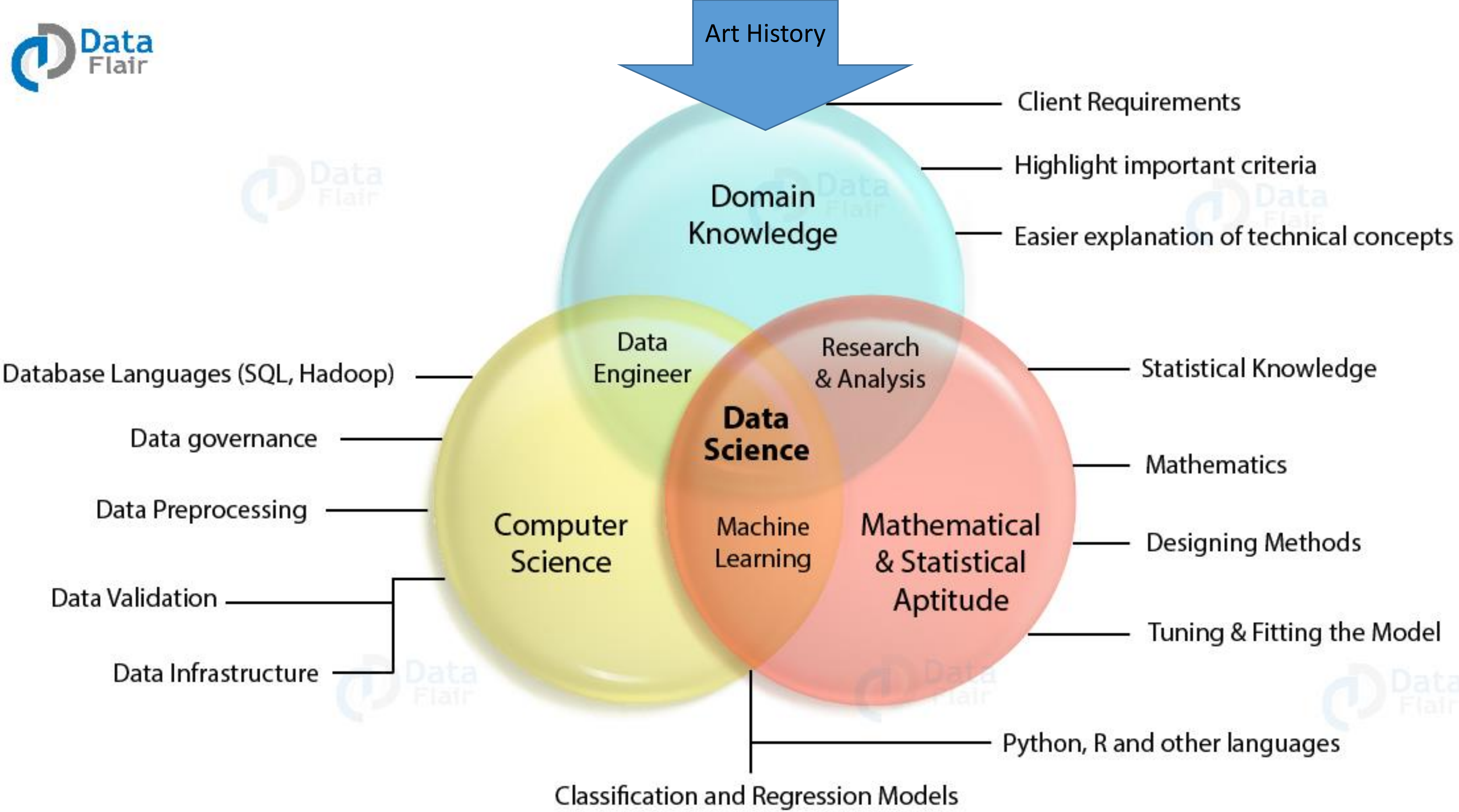
Digital Art History

- … is data-driven Art History

Data Science

- … is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data.
- … uses statistics, data analysis, machine learning and domain knowledge in order to understand and analyze actual phenomena with data.
- After (1) empirical, (2) theoretical and (3) computational, *data-driven* is considered to be the "fourth paradigm" of science
- Computer scientist Jim Gray asserted: “Everything about science is changing because of the impact of information technology" and the volume of new data being generated.





Art History

Domain Knowledge

- Client Requirements
- Highlight important criteria
- Easier explanation of technical concepts

Data Engineer

Research & Analysis

Data Science

Computer Science

- Database Languages (SQL, Hadoop)
- Data governance
- Data Preprocessing
- Data Validation
- Data Infrastructure

Machine Learning

- Python, R and other languages
- Classification and Regression Models

Mathematical & Statistical Aptitude

- Statistical Knowledge
- Mathematics
- Designing Methods
- Tuning & Fitting the Model

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

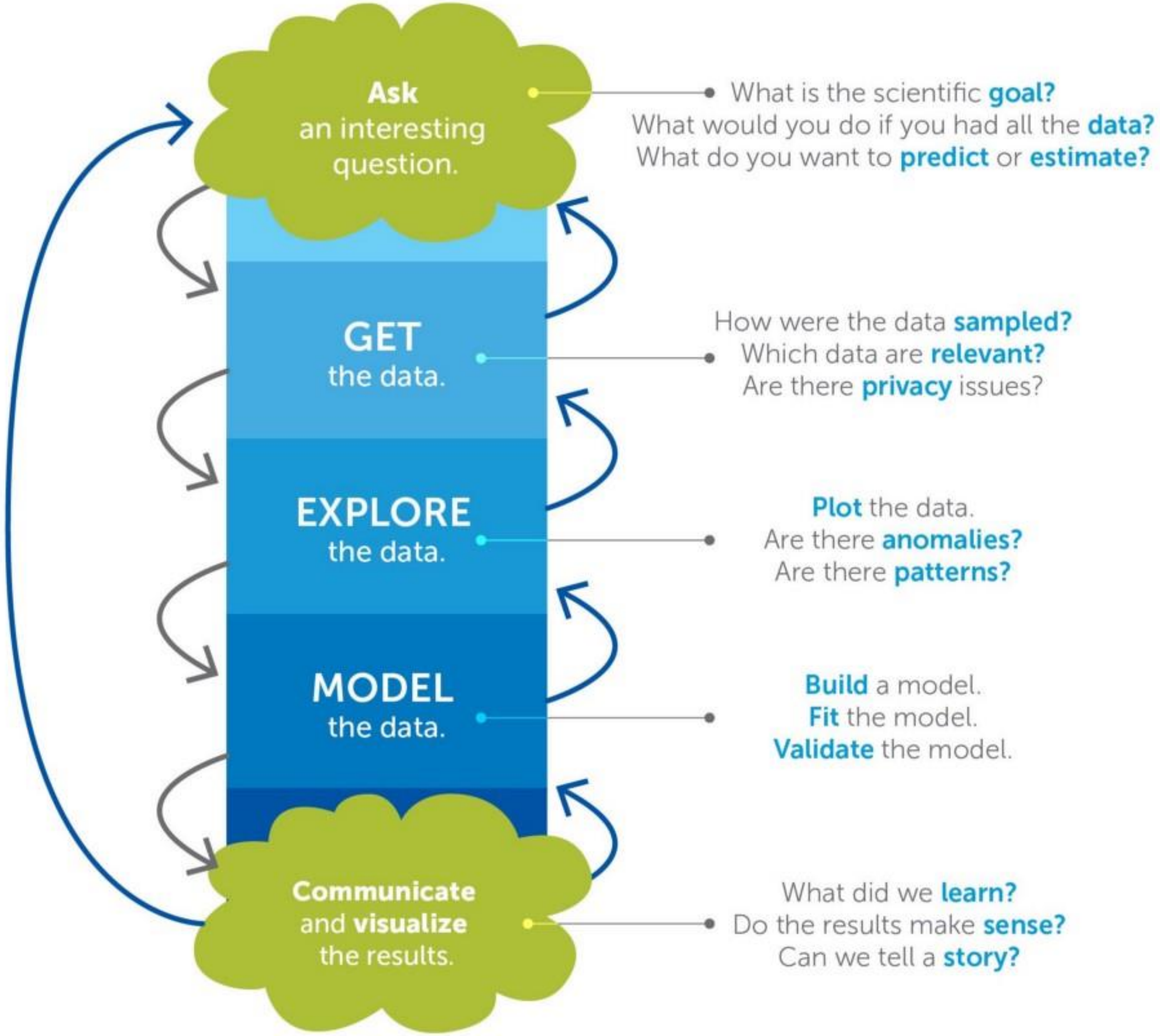
DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

The **Data Science** Process



Ask an interesting question.

- What is the scientific **goal?**
- What would you do if you had all the **data?**
- What do you want to **predict** or **estimate?**

GET the data.

- How were the data **sampled?**
- Which data are **relevant?**
- Are there **privacy** issues?

EXPLORE the data.

- **Plot** the data.
- Are there **anomalies?**
- Are there **patterns?**

MODEL the data.

- **Build** a model.
- **Fit** the model.
- **Validate** the model.

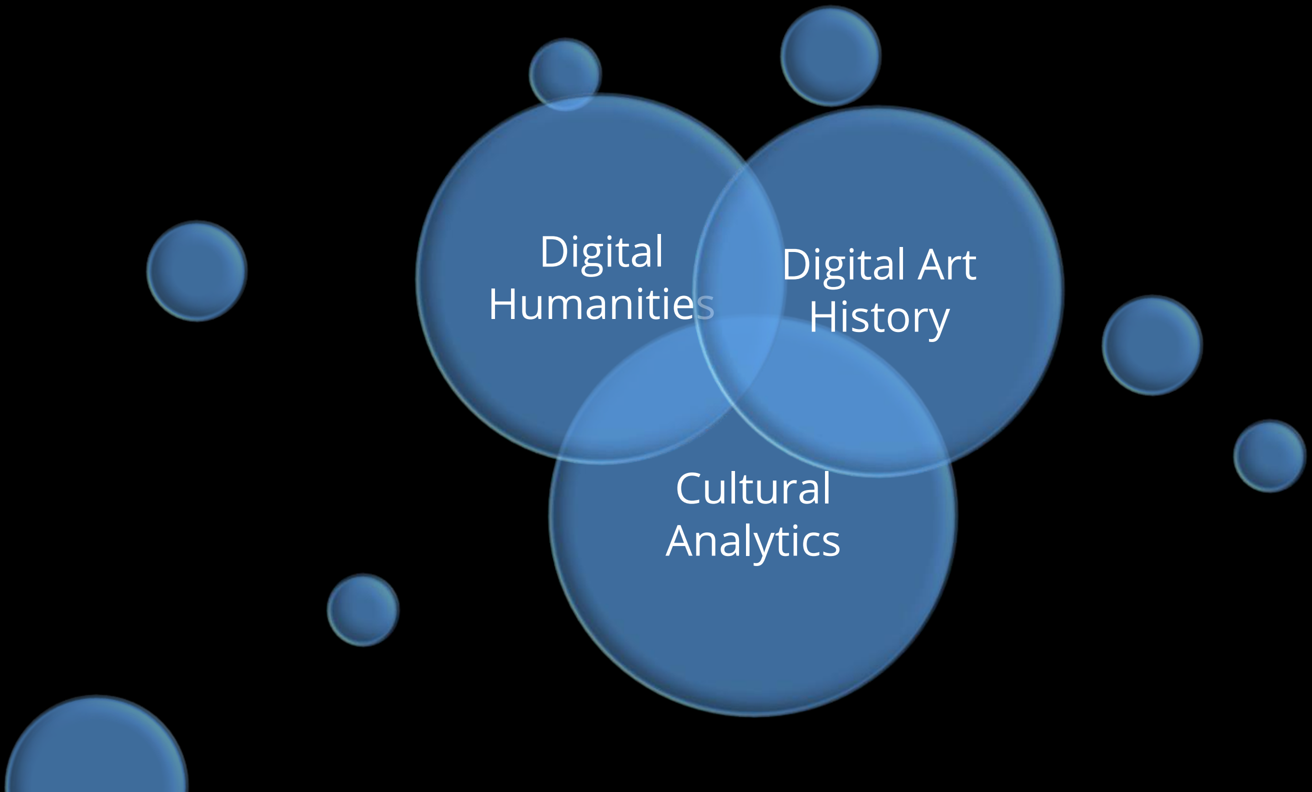
Communicate and **visualize** the results.

- What did we **learn?**
- Do the results make **sense?**
- Can we tell a **story?**

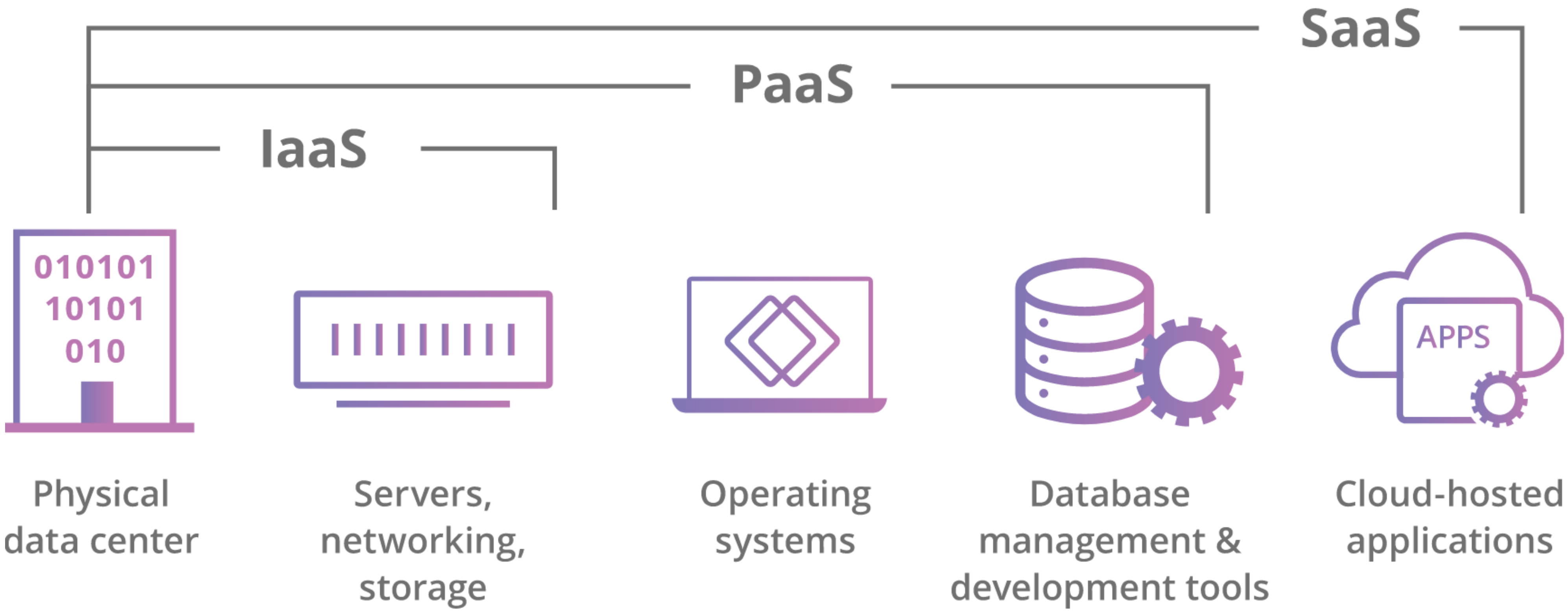
Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course http://cs109.org/.

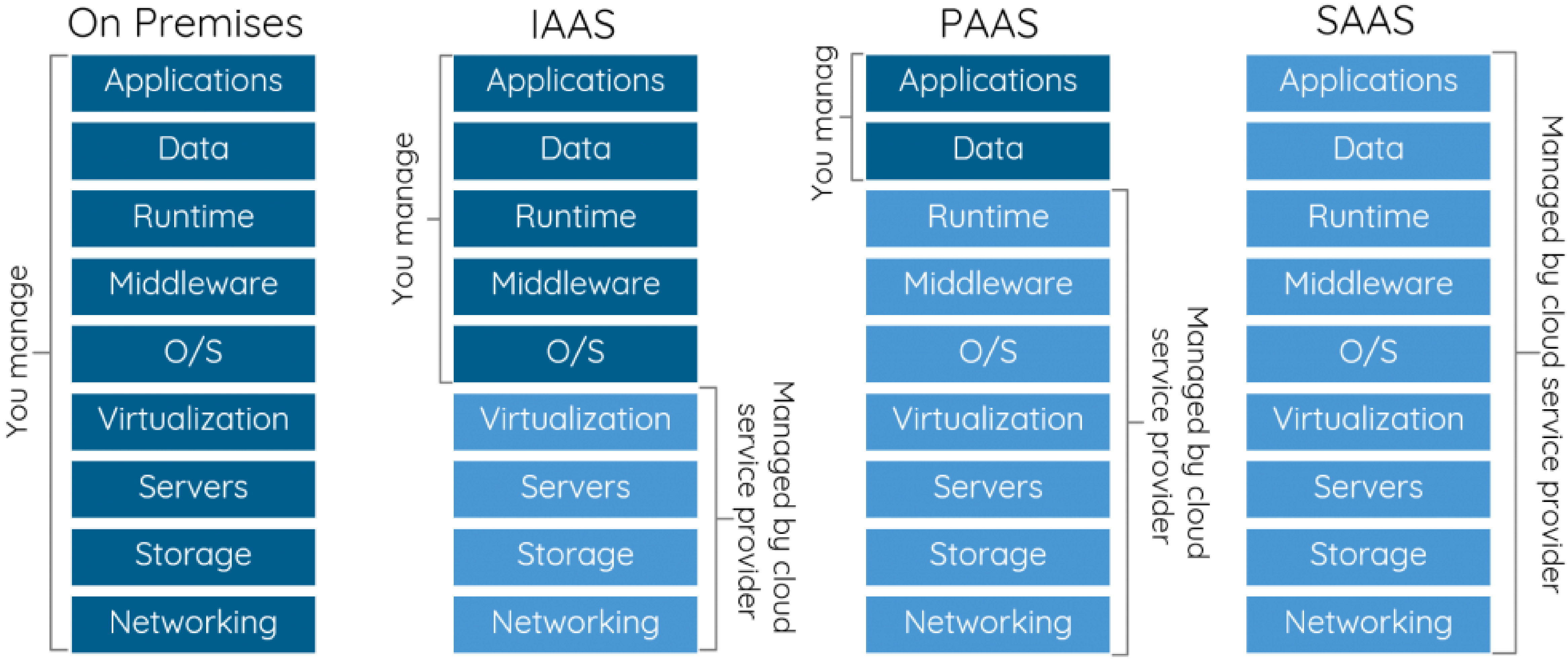
Cultural Analytics

- … refers to the use of computational, visualization, and big data methods for the exploration of contemporary and historical cultures.



- Infrastructure-as-a-Service (IaaS): e.g. Amazon web services
- Platform-as-a-Service (PaaS): e.g. Google App Engine
- Software-as-a-Service (SaaS): e.g. Google docs


SaaS
PaaS
IaaS
010101
10101
010
APPS
Physical data center
Servers, networking, storage
Operating systems
Database management & development tools
Cloud-hosted applications



On Premises

You manage
- Applications
- Data
- Runtime
- Middleware
- O/S
- Virtualization
- Servers
- Storage
- Networking

IAAS

You manage
- Applications
- Data
- Runtime
- Middleware
- O/S

Managed by cloud service provider
- Virtualization
- Servers
- Storage
- Networking

PAAS

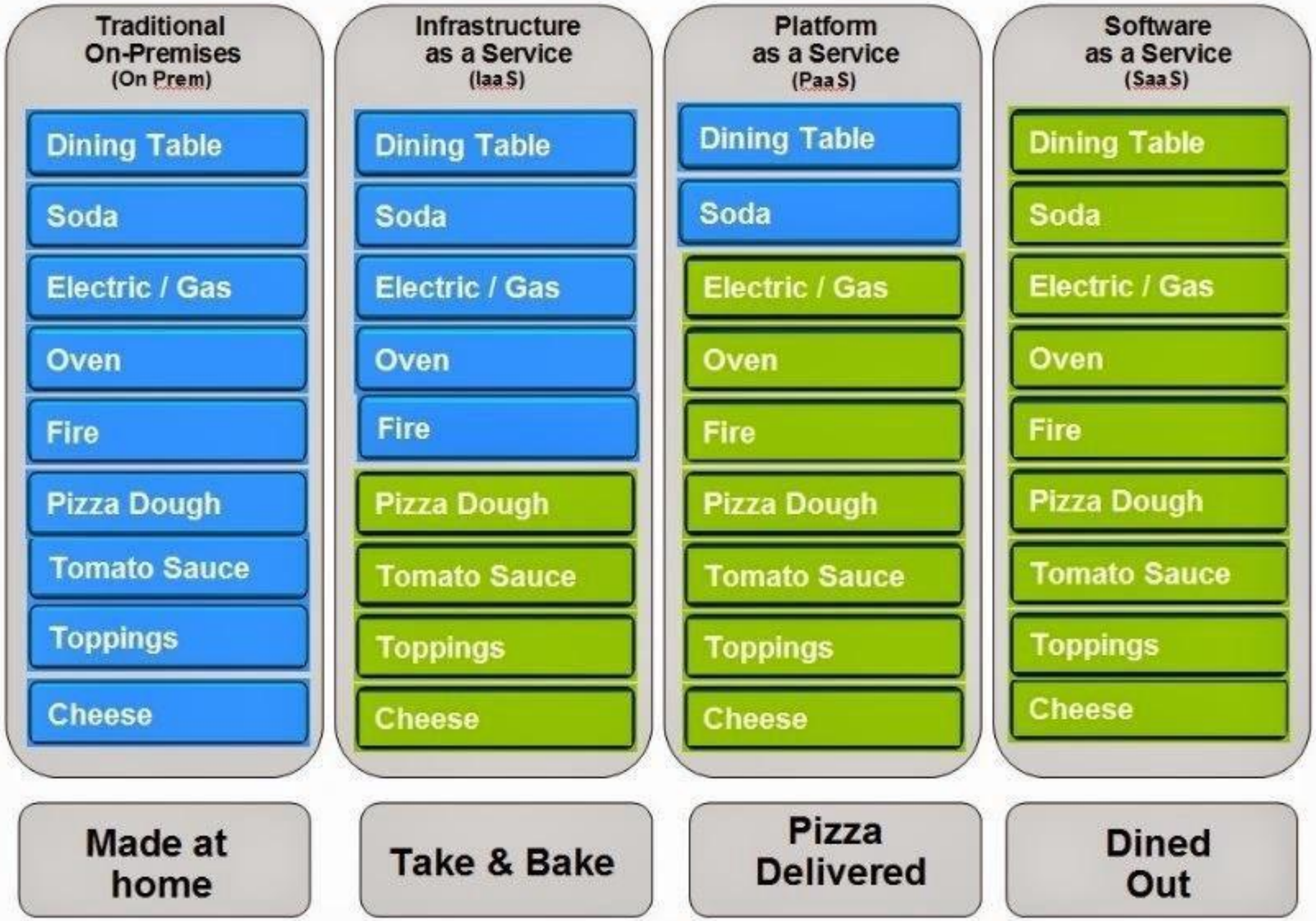
You manag
- Applications
- Data

Managed by cloud service provider
- Runtime
- Middleware
- O/S
- Virtualization
- Servers
- Storage
- Networking

SAAS

Managed by cloud service provider
- Applications
- Data
- Runtime
- Middleware
- O/S
- Virtualization
- Servers
- Storage
- Networking

Pizza as a Service



Traditional On-Premises (On Prem)	Infrastructure as a Service (Iaa S)	Platform as a Service (Paa S)	Software as a Service (Saa S)
🟦 Dining Table	🟦 Dining Table	🟦 Dining Table	🟩 Dining Table
🟦 Soda	🟦 Soda	🟦 Soda	🟩 Soda
🟦 Electric / Gas	🟦 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas
🟦 Oven	🟦 Oven	🟩 Oven	🟩 Oven
🟦 Fire	🟦 Fire	🟩 Fire	🟩 Fire
🟦 Pizza Dough	🟩 Pizza Dough	🟩 Pizza Dough	🟩 Pizza Dough
🟦 Tomato Sauce	🟩 Tomato Sauce	🟩 Tomato Sauce	🟩 Tomato Sauce
🟦 Toppings	🟩 Toppings	🟩 Toppings	🟩 Toppings
🟦 Cheese	🟩 Cheese	🟩 Cheese	🟩 Cheese
Made at home	Take & Bake	Pizza Delivered	Dined Out

🟦 You Manage 🟩 Vendor Manages

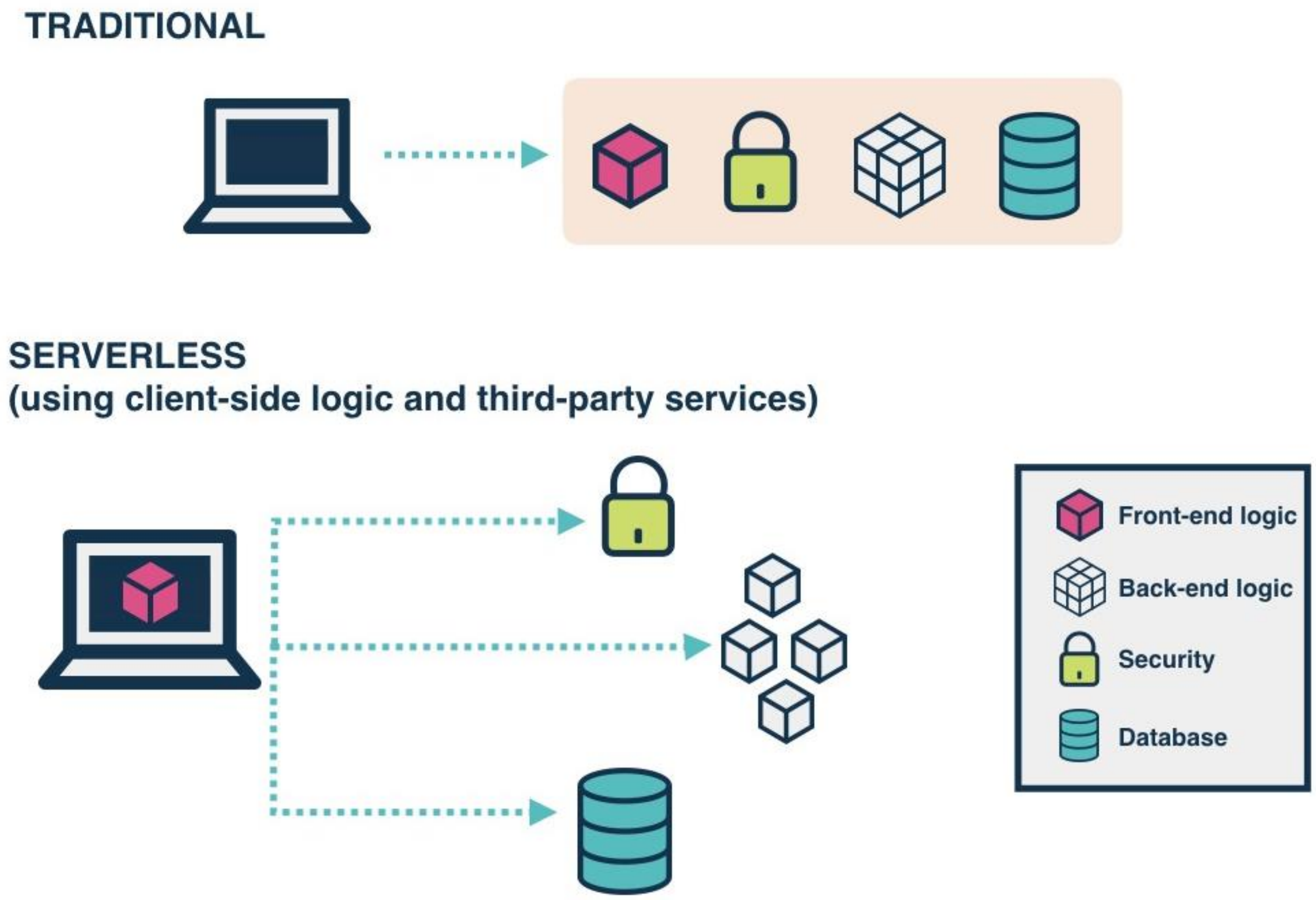
Pizza as a Service 2.0

http://www.paulkerrison.co.uk

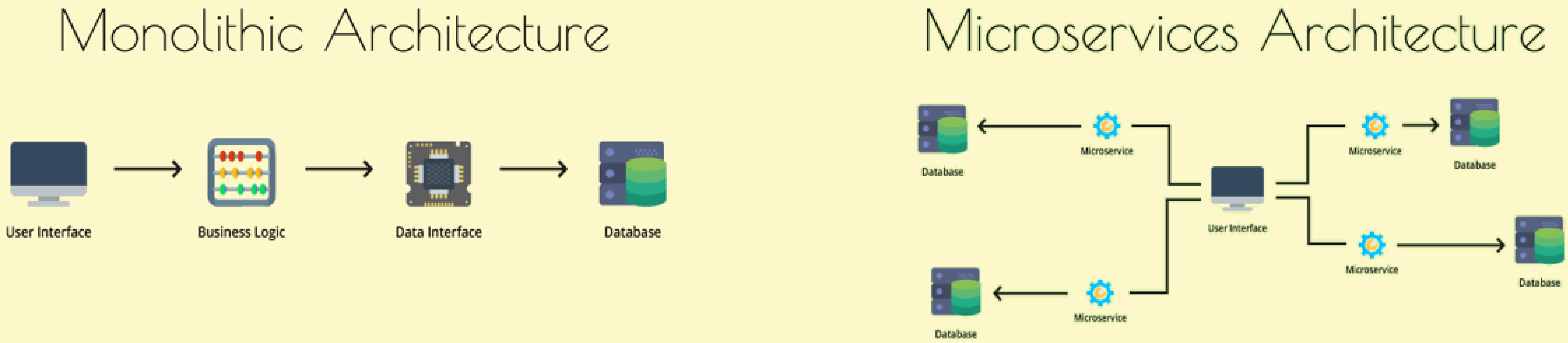
Tradition On-Premises (legacy)	Infrastructure as a Service (IaaS)	Containers as a Service (CaaS)	Platform as a Service (PaaS)	Function as a Service (FaaS)	Software as a Service (SaaS)	
🟦 Conversation	🟦 Conversation	🟦 Conversation	🟦 Conversation	🟦 Conversation	🟦 Conversation	Configuration
🟦 Friends	🟦 Friends	🟦 Friends	🟦 Friends	🟦 Friends	🟩 Friends	Functions
🟦 Beer	🟦 Beer	🟦 Beer	🟦 Beer	🟩 Beer	🟩 Beer	Scaling...
🟦 Pizza	🟦 Pizza	🟦 Pizza	🟩 Pizza	🟩 Pizza	🟩 Pizza	Runtime
🟦 Fire	🟦 Fire	🟩 Fire	🟩 Fire	🟩 Fire	🟩 Fire	OS
🟦 Oven	🟩 Oven	🟩 Oven	🟩 Oven	🟩 Oven	🟩 Oven	Virtualisation
🟦 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas	Hardware
Homemade	Communal Kitchen	Bring Your Own	Takeaway	Restaurant	Party	

🟦 You Manage 🟩 Vendor Manages

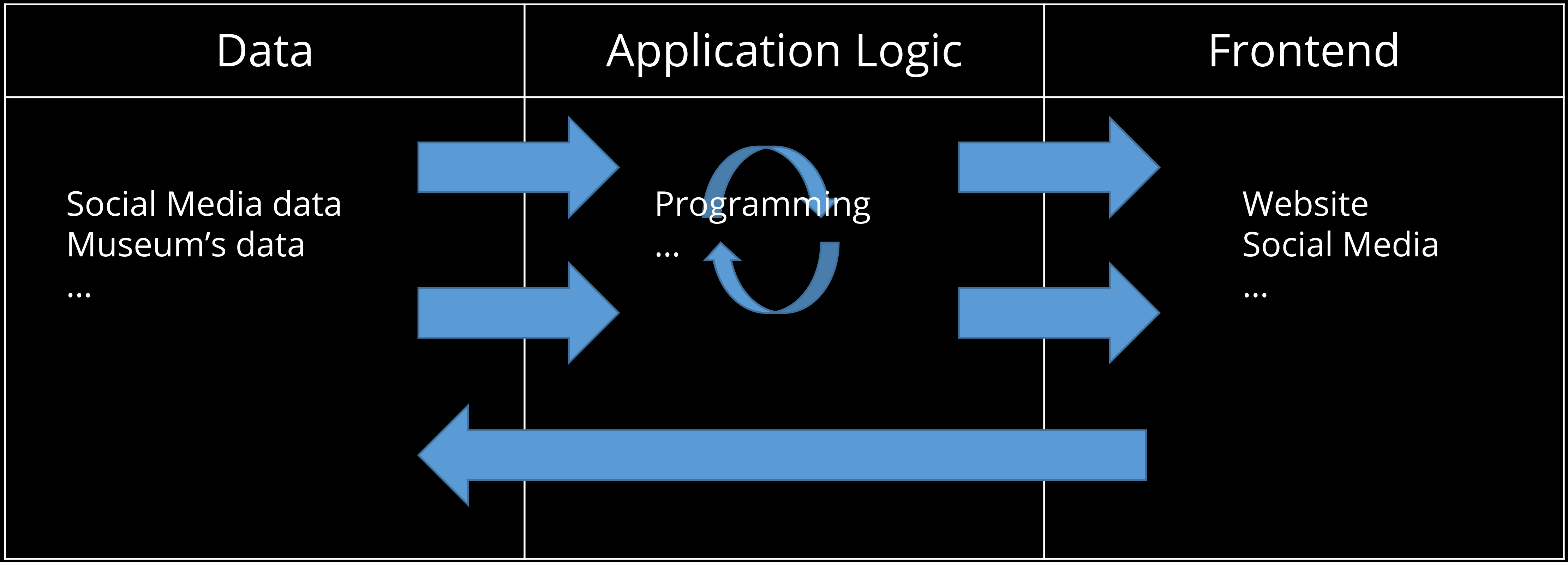
- Function as a service (FaaS): "serverless" architecture



Monolithic application vs. Microservices



Application planning



Data	Application Logic	Frontend
Social Media data Museum’s data …	Programming …	Website Social Media …

Next

- Look at our sample data
- Do something with our sample data
- Try out for yourself