Data Science as a function

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069: Applied Data Science
for Social Scientists

Spring 2023
Columbia University

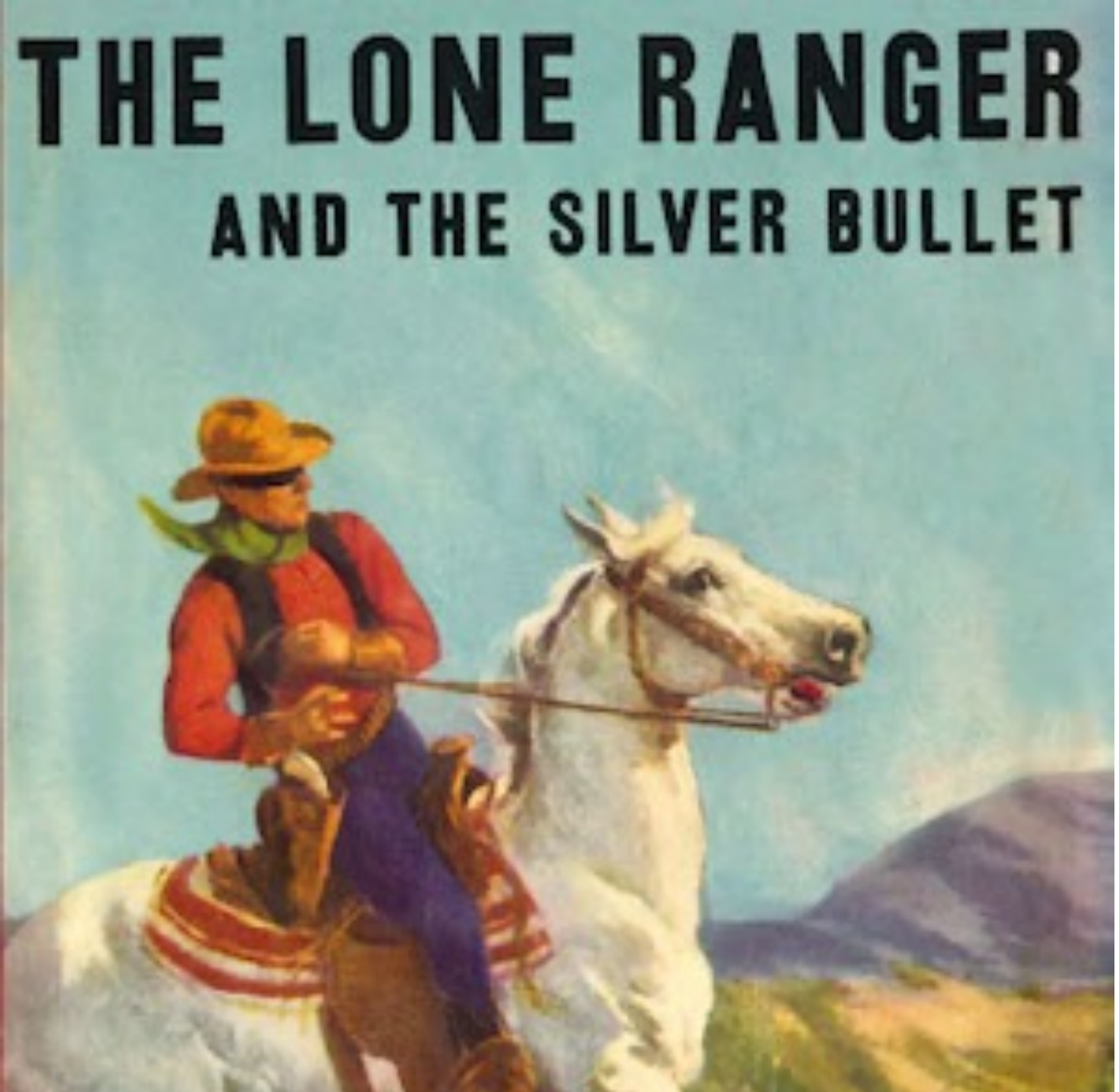

Two myths
about
Data Science

myth #1:

Data Science
is about
machine learning

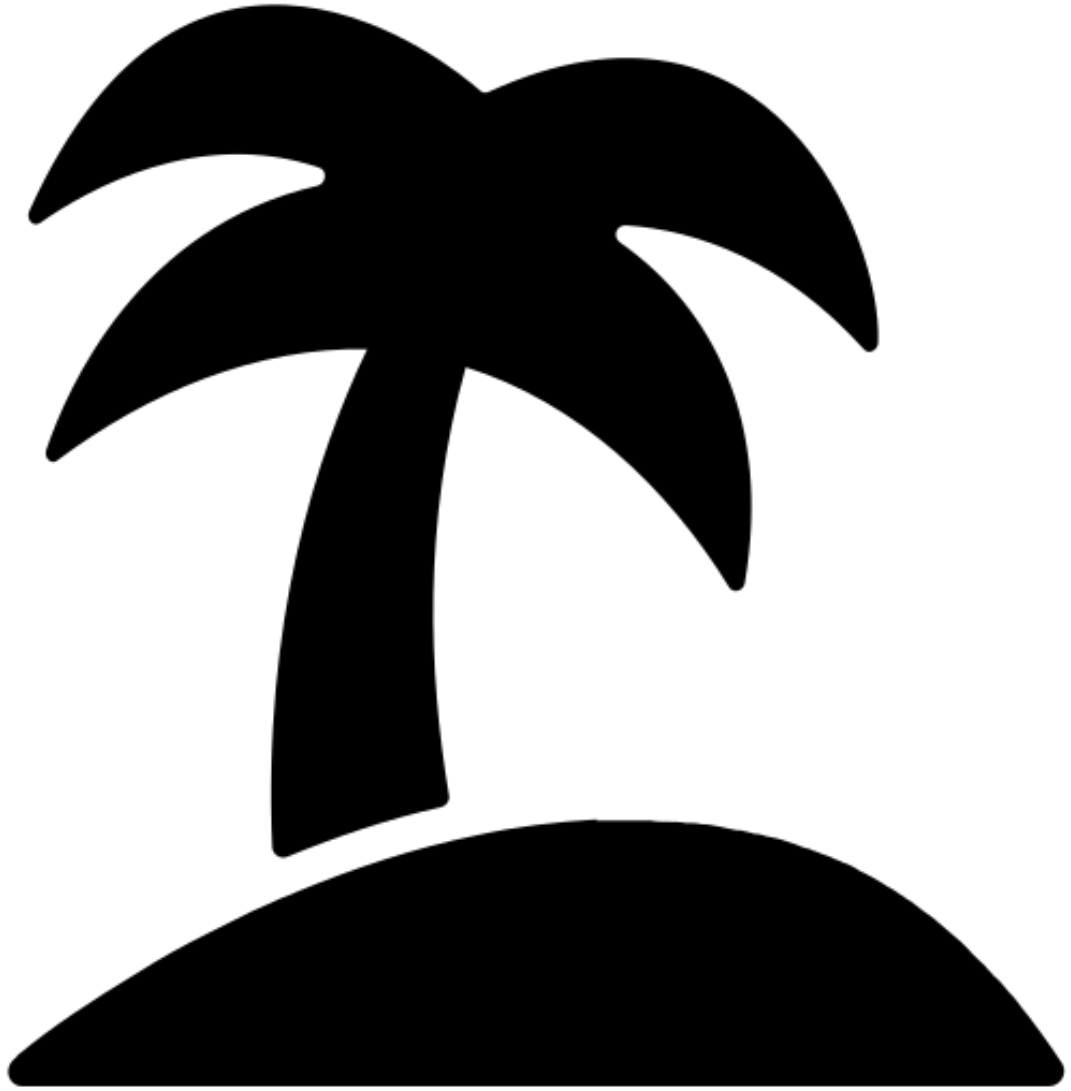
in practice:

Data Science is about building **Data Products** that solve a **business need or problem**



myth #2:

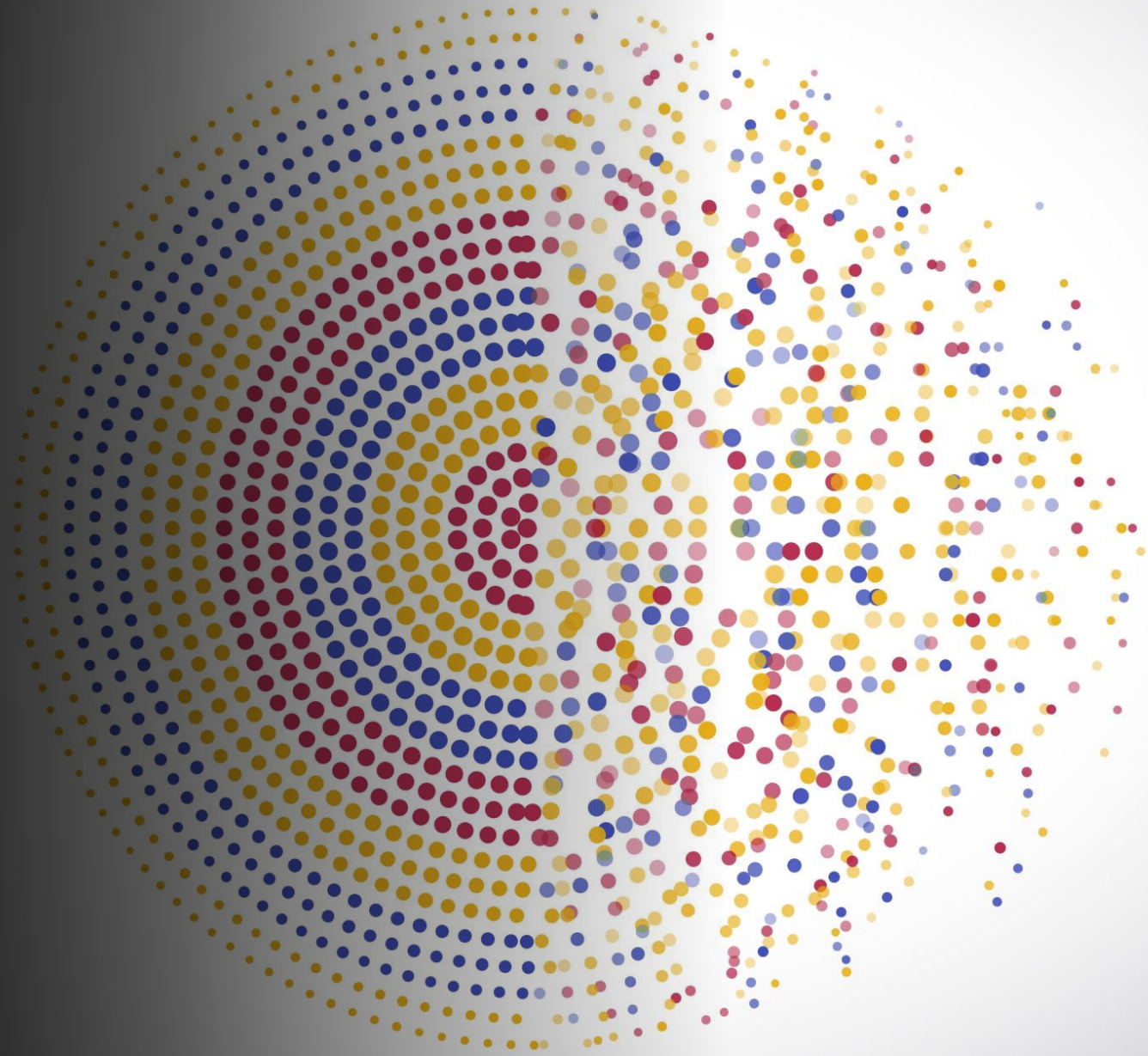
Data Scientists work alone



in practice:

no Data Scientist
is an island

Data Science in practice



starting point: a **conceptual framework**

Problem: a statement without (an appropriate) solution

Solution: a data product that (effectively) mitigates a problem

"[A] **data product** [...] facilitates an end goal through the use of data".

- DJ Patil, *Data Jujitsu* (2016)

what does the **Data Science Shop** do?

define the **problem** with crisp clarity

identify **data** available to tackle problem

evaluate whether data is appropriate to solve the problem

design a **solution**

select appropriate **tools** and **techniques**

decide how the solution will be **served**

build solution

identify **obstacles** to adoption

craft **strategy** to overcome obstacles

uncover unattended business needs

incorporate **new features** in roadmap

1 2 3 4 5 6 7 8 9 10 11

diagnosis

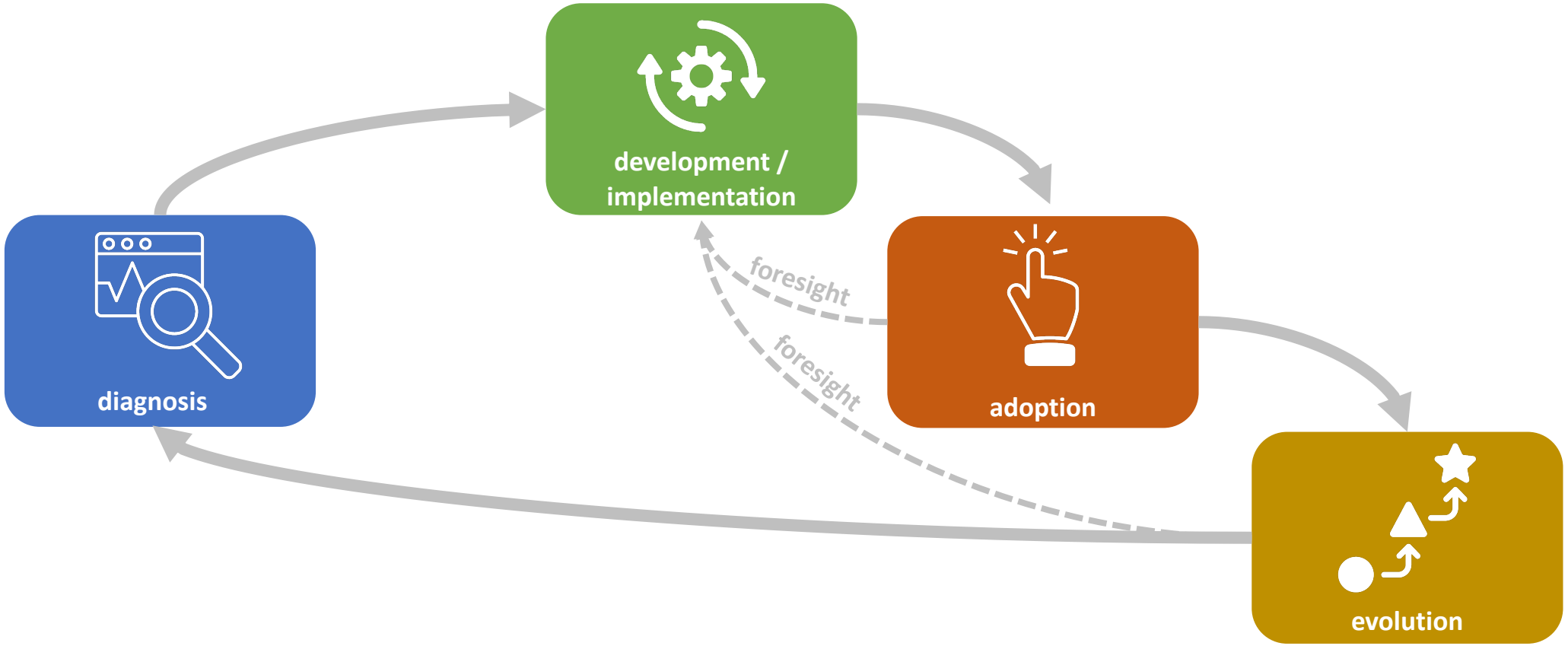
development / implementation

adoption

evolution

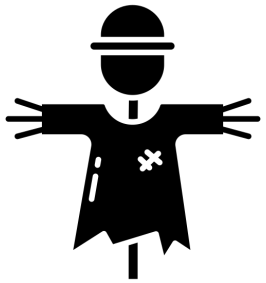
the Data Product cycle

what does the **Data Science Shop** do?

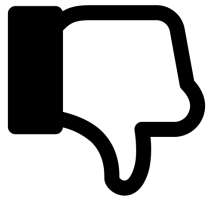


the Data Product cycle

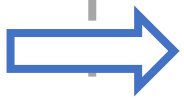
how does the **Data Science Shop** do it?

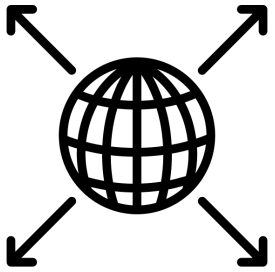


start small
(MVP)

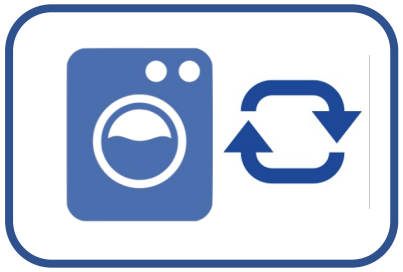


fail fast





scale up



iterate

who crews the **Data Science Shop**?

data scientist

- Define **correct questions**
- Prototype **ETL**
- **Model** data (apply **algorithms**)
- **Build** prototype solutions
- **Translate** solution outputs

data analyst

- **Query** data
- **Summarize and visualize** data
- Identify **trends**
- **Interpret** findings
- **Communicate** with business

data engineer

- Develop and maintain **data architecture**
  - data ingestion
  - data storage
  - data transformation
- Build **data pipelines**
- Productionize **ETL**
- Build processes to ensure **data quality**
- **Orchestrate** processes
- Build **working environments**

ML engineer

- **Productionize** algorithms
- **Scale** prototyped solutions
- **Optimize** computational performance
- Create **endpoints** for outputs
- **Orchestrate** processes
- Build **working environments**

project manager

- Develop **timelines**
- Task **planning**
- Resource **allocation**
- **Risk** monitoring

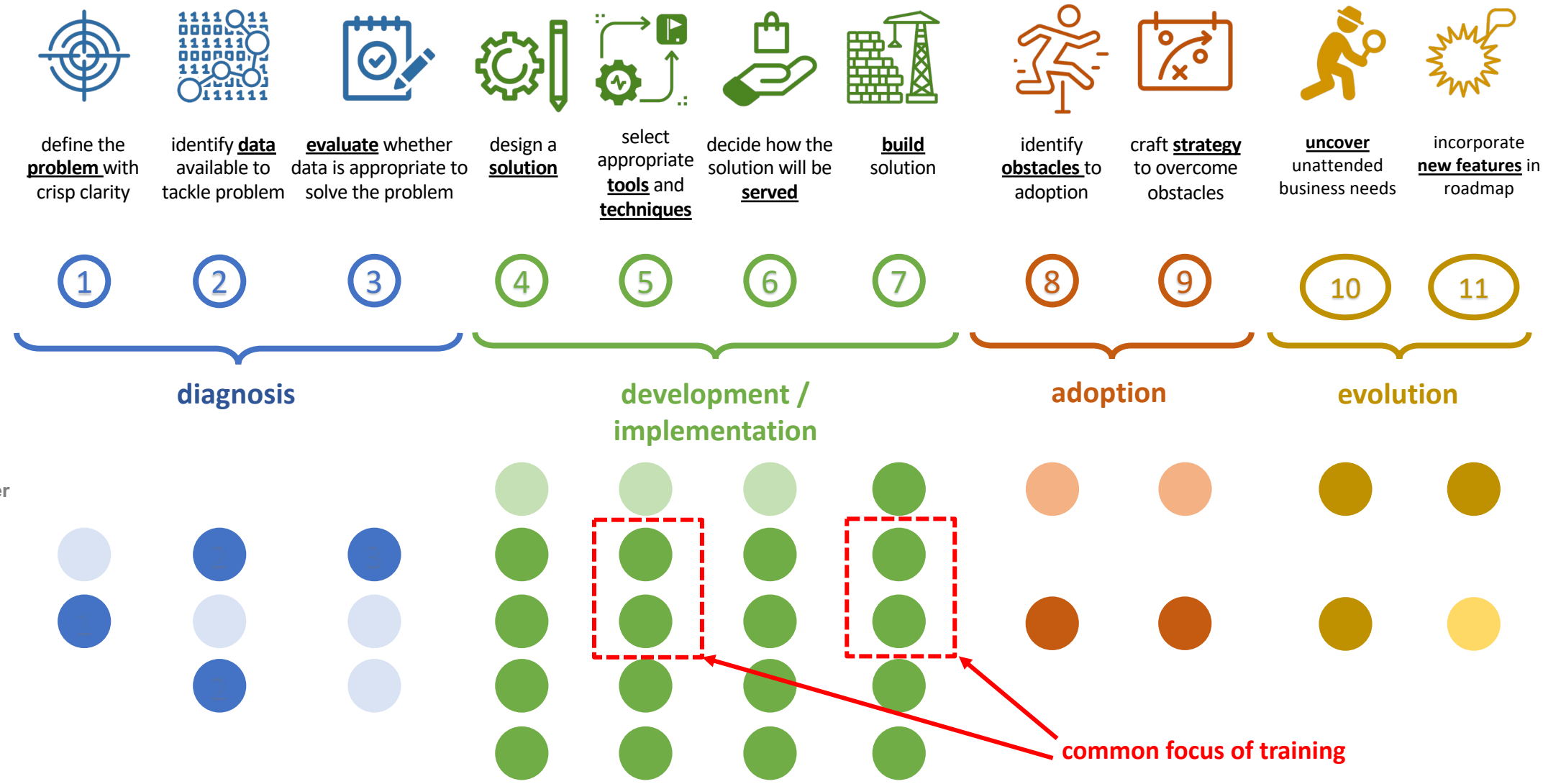
the **Data Science shop** crew in detail

	 data scientist	 data analyst	 data engineer	 ML engineer	 project manager

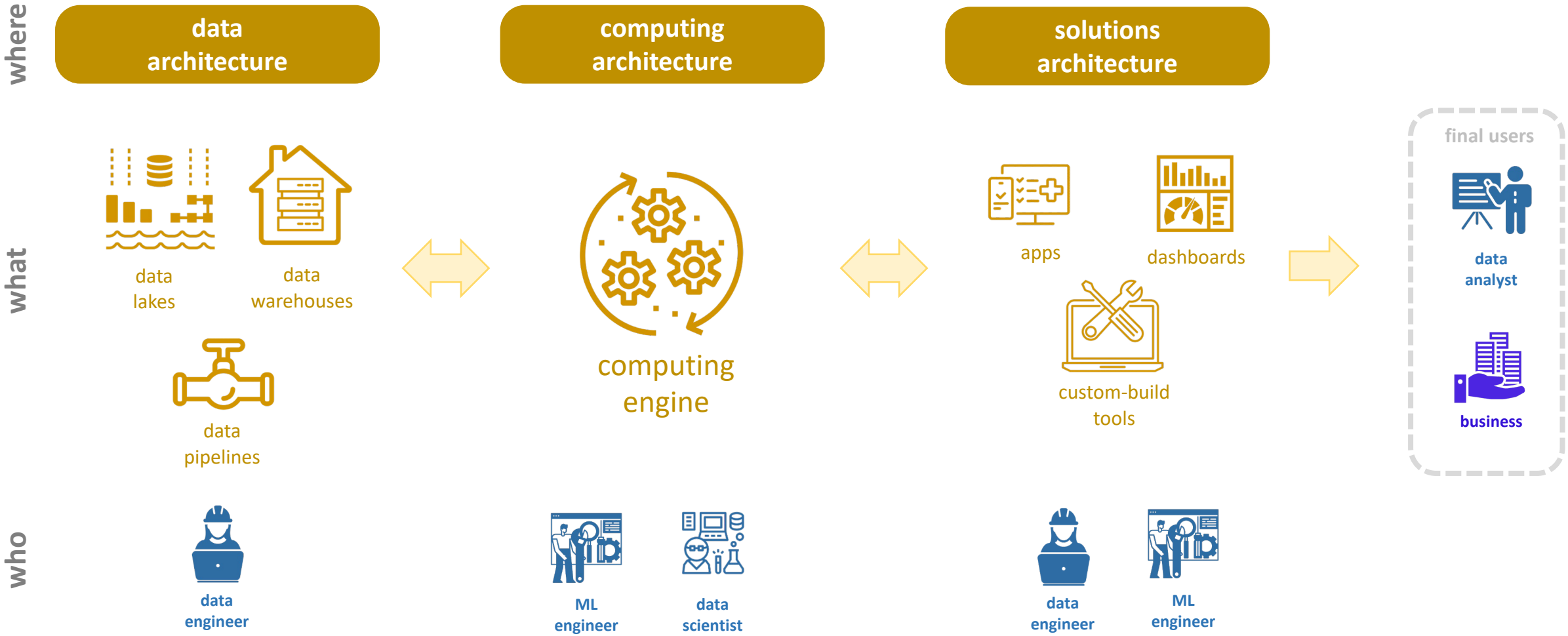
tasks	• Define **correct questions**
	• Prototype **ETL**
	• **Model data** (apply algorithms)
	• **Build** prototype solutions
	• **Translate** solution outputs
		• **Query** data(bases)
		• **Summarize** and **visualize** data
		• Identify **trends**
		• **Interpret** findings
		• **Communicate** with business
			• Develop and maintain **data architecture**
			  ○ data ingestion
			  ○ data storage
			  ○ data security
			  ○ data transformation
			• Build **data pipelines**
			• Productionize **ETL** (prototypes)
			• Build **data quality processes**
			• Orchestrate **processes**
			• Build **working environments**
				• **Productionize** algorithms
				• **Scale** prototyped solutions
				• **Optimize** computational performance
				• Create **endpoints** for outputs
				• **Orchestrate** processes
				• Build **working environments**
					• Develop **timelines**
					• Task **planning**
					• Resource **allocation**
					• **Risk** monitoring

outputs	• **prototyped solutions** • **science-backed solutions**	• **insights**	• **data architectures** • **quality-checked data pipelines**	• **computation-optimized solutions** • **production-ready solutions**	• **roadmaps** • **execution**
skills	• critical thinking (about data) • statistics • data visualization • hacking • algorithms • explanation / prediction • communication • translation	• dense business knowledge • data querying • data visualization • communication	• advanced programming skills • advanced software engineering • cloud computing • database design • data architecture design • distributed systems • communication	• advanced programming skills • advanced software engineering • advanced cloud computing • advanced optimization math • algorithms (intermediate) • distributed systems • communication	• leadership • negotiation • team building • planning • basic technical acumen • communication • translation

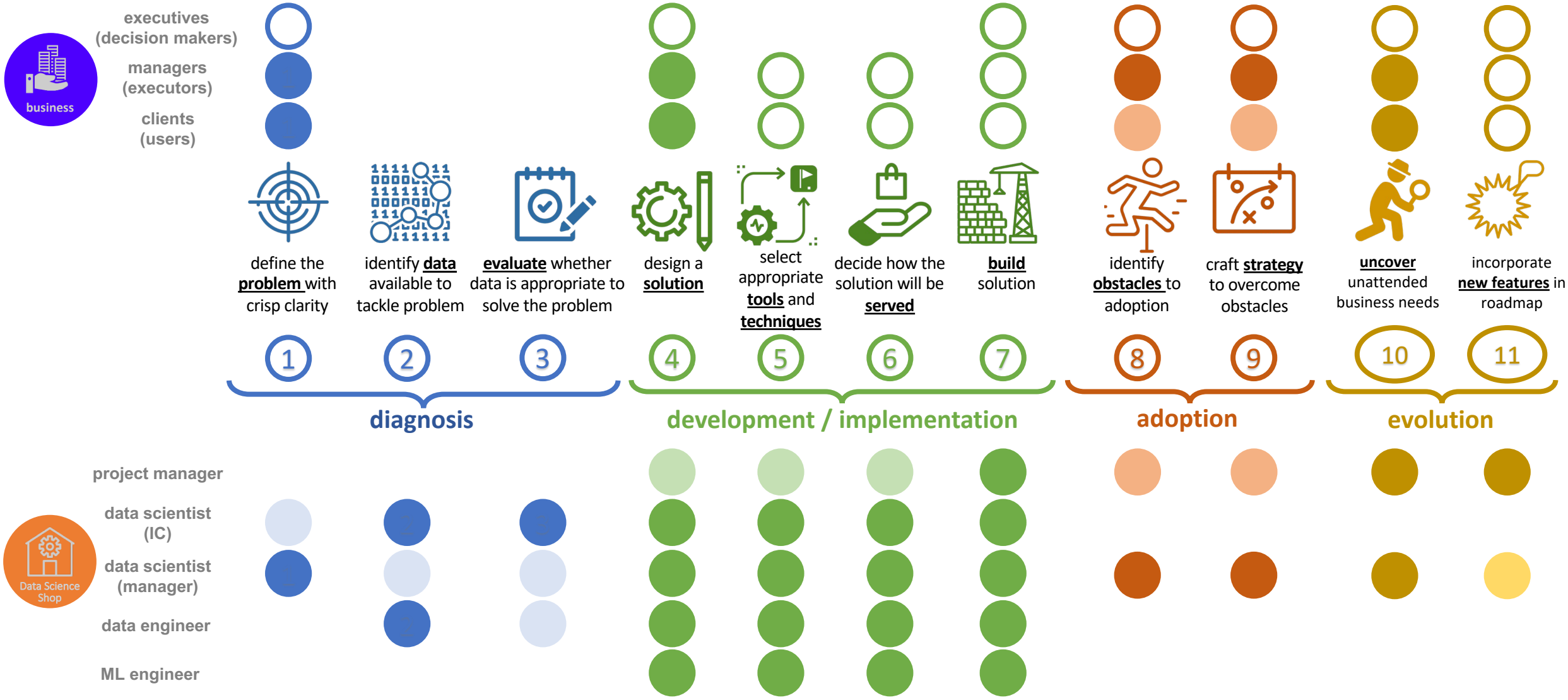
how does a **Data Science Shop** operate?



where does the **Data Science Shop** operate?



when does the business intervene?



what **Data Products** can the **Shop** build?

	dashboards	answers	deployments	custom-built tools
what	frontends for automated data summarization and visualization	science-backed answers to business questions (explanatory / predictive)	stand-alone algorithmic outputs that integrate to business processes	end-to-end proprietary applications developed to fulfill a business objective
who	 data engineer, data analyst	 data engineer, data scientist	 data engineer, ML engineer, data scientist	 data engineer, ML engineer, data scientist, data analyst, project manager
where	data architecture solutions architecture	data architecture	data architecture computing architecture	data architecture computing architecture solutions architecture

the **problem** defines the **type of shop**

Problem: a statement without (an appropriate) solution

Solution: a data product that (effectively) mitigates a problem

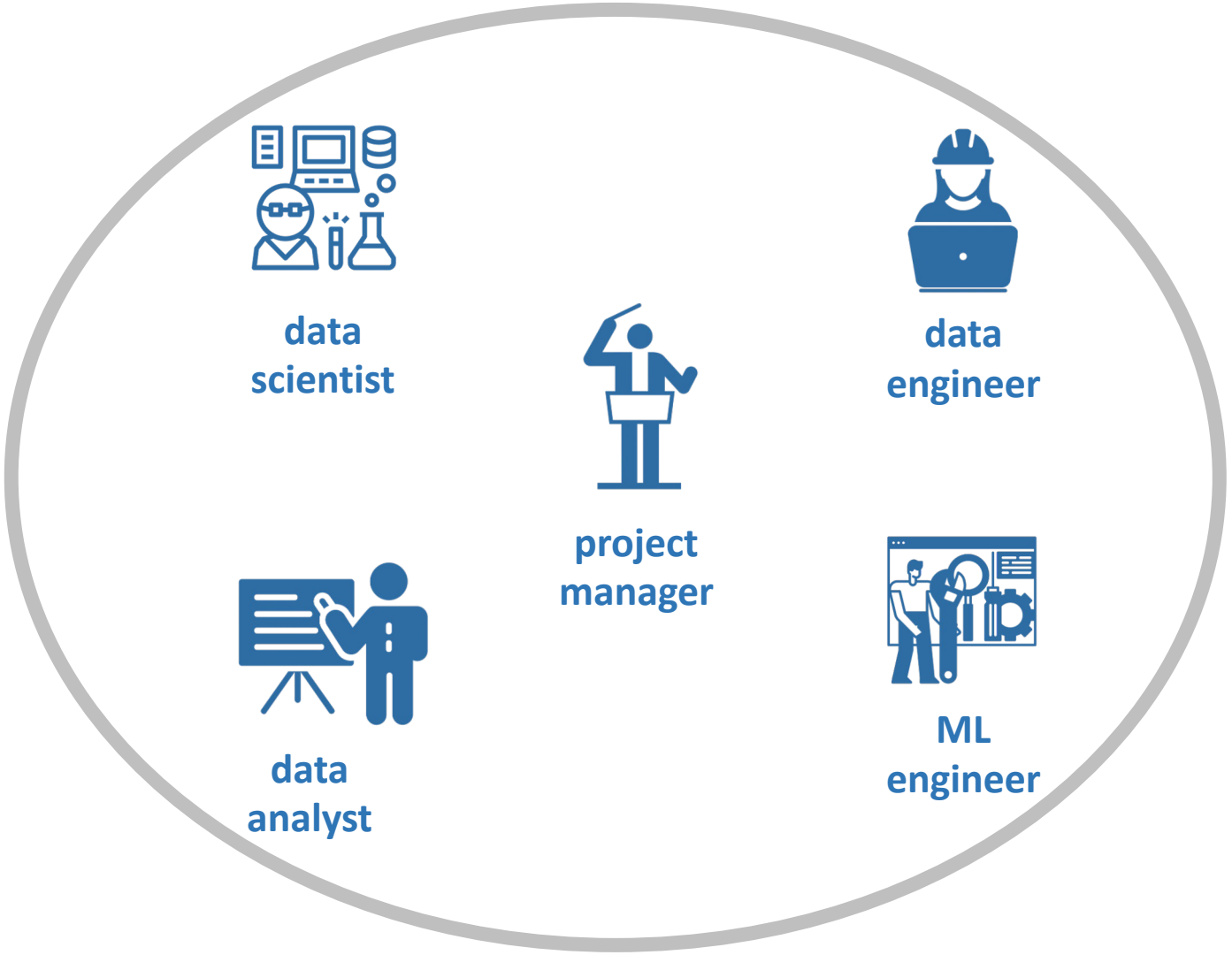
the **Data Science Shop**: *mutatis mutandis*

[developed]

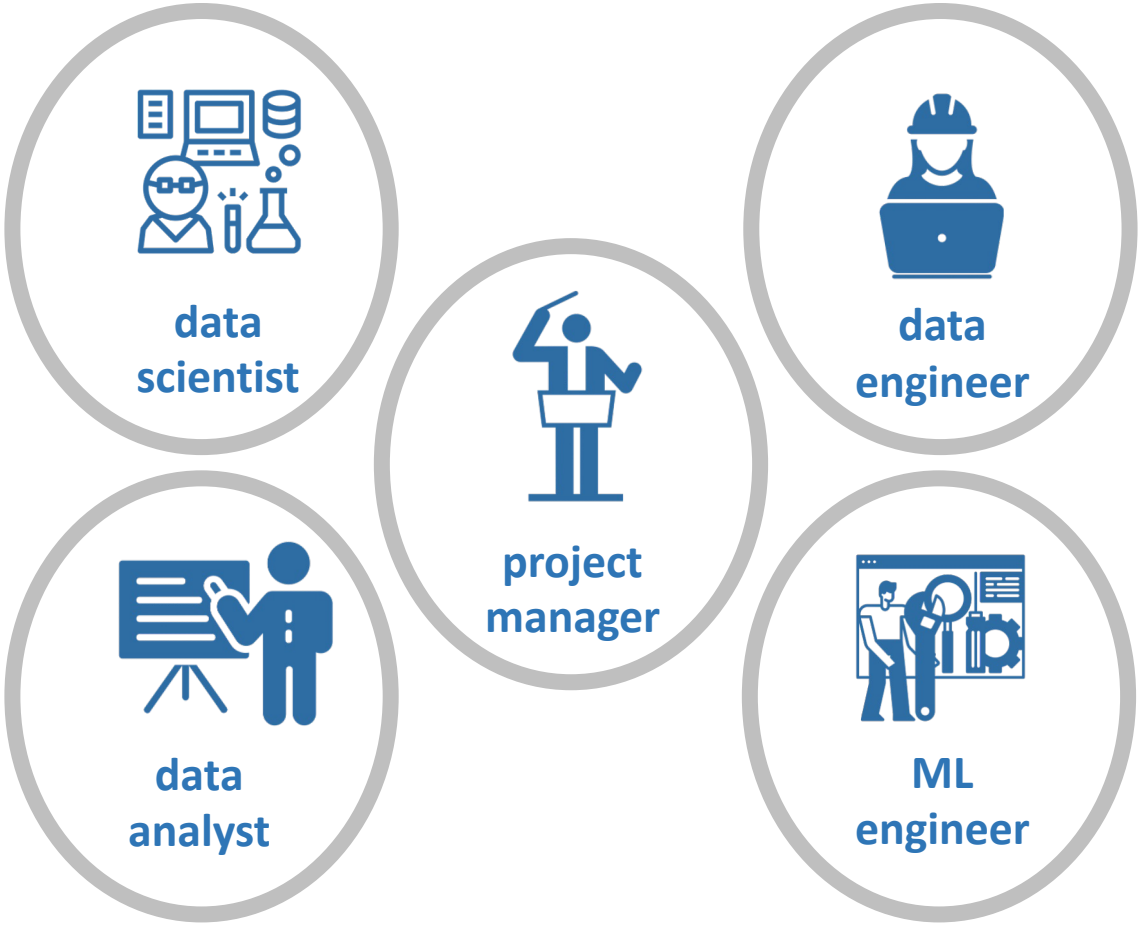
[embryonic]

circa 2010: the **unicorn** approach

⇔



today: (toward) specialization



Data Science as a function

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069: Applied Data Science
for Social Scientists

Spring 2023
Columbia University