

# Data Science as a function

Marco Morales

[marco.morales@columbia.edu](mailto:marco.morales@columbia.edu)

Nana Yaw Essuman

[nanayawce@gmail.com](mailto:nanayawce@gmail.com)

GR5069: Applied Data Science  
for Social Scientists

Spring 2024  
Columbia University

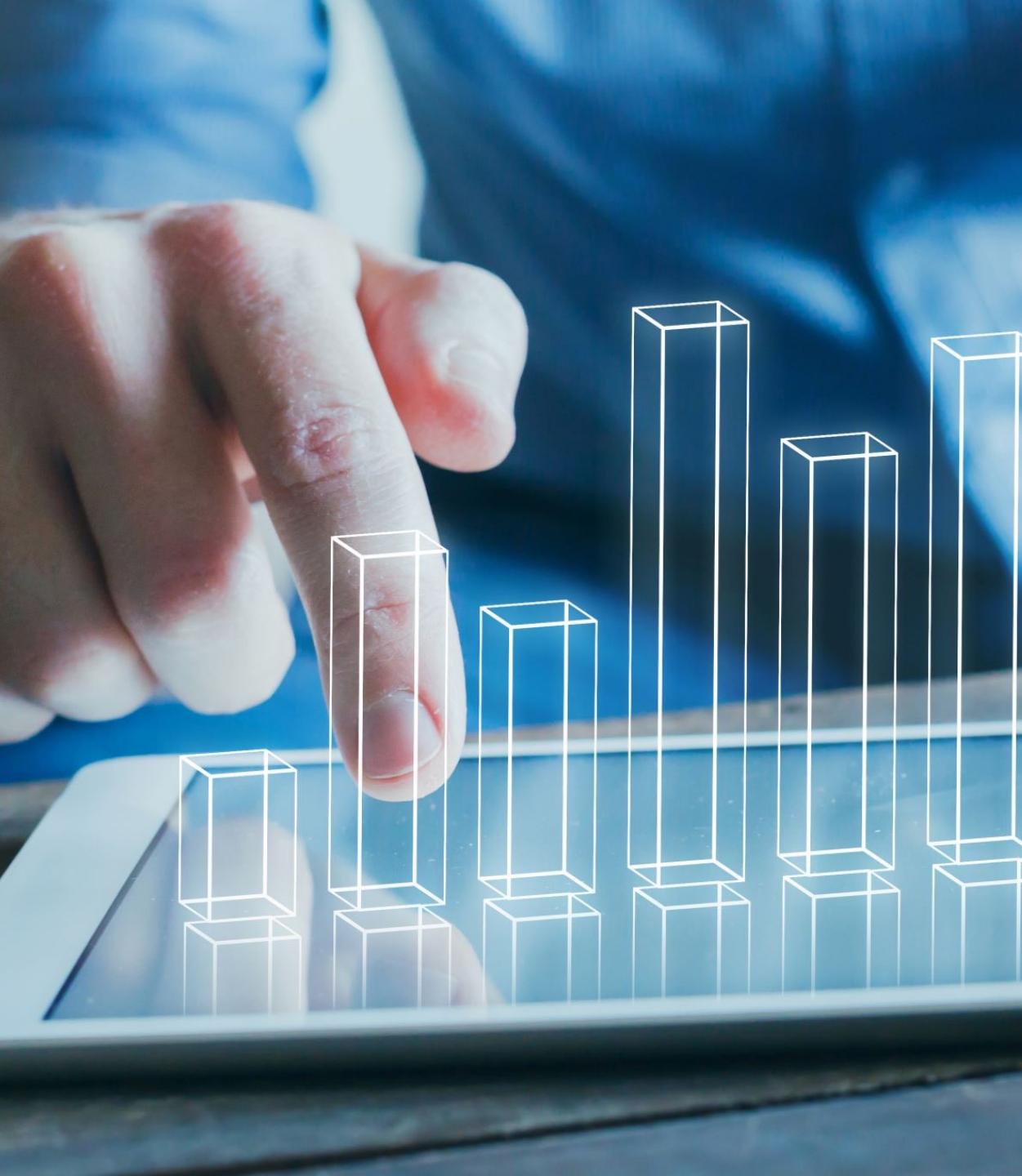


# Two myths about Data Science



# myth #1:

## Data Science is about machine learning



in practice:

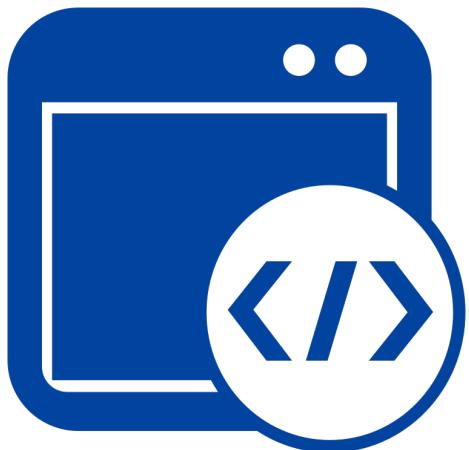
Data Science is  
about building  
**Data Products** that  
solve a business  
need or problem



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

- DJ Patil, *Data Jujitsu* (2016)

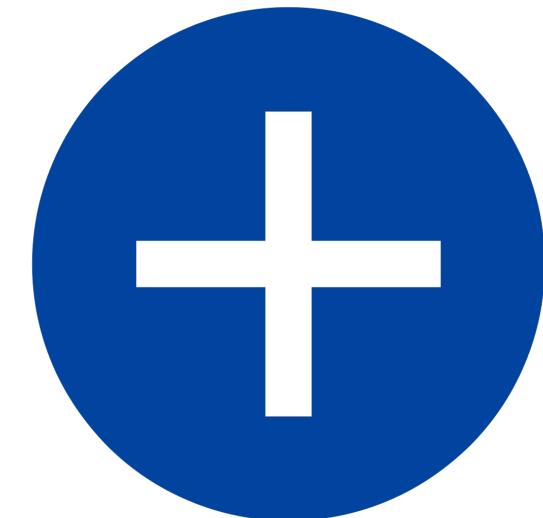
data products are a special kind of digital solution



software products



data products



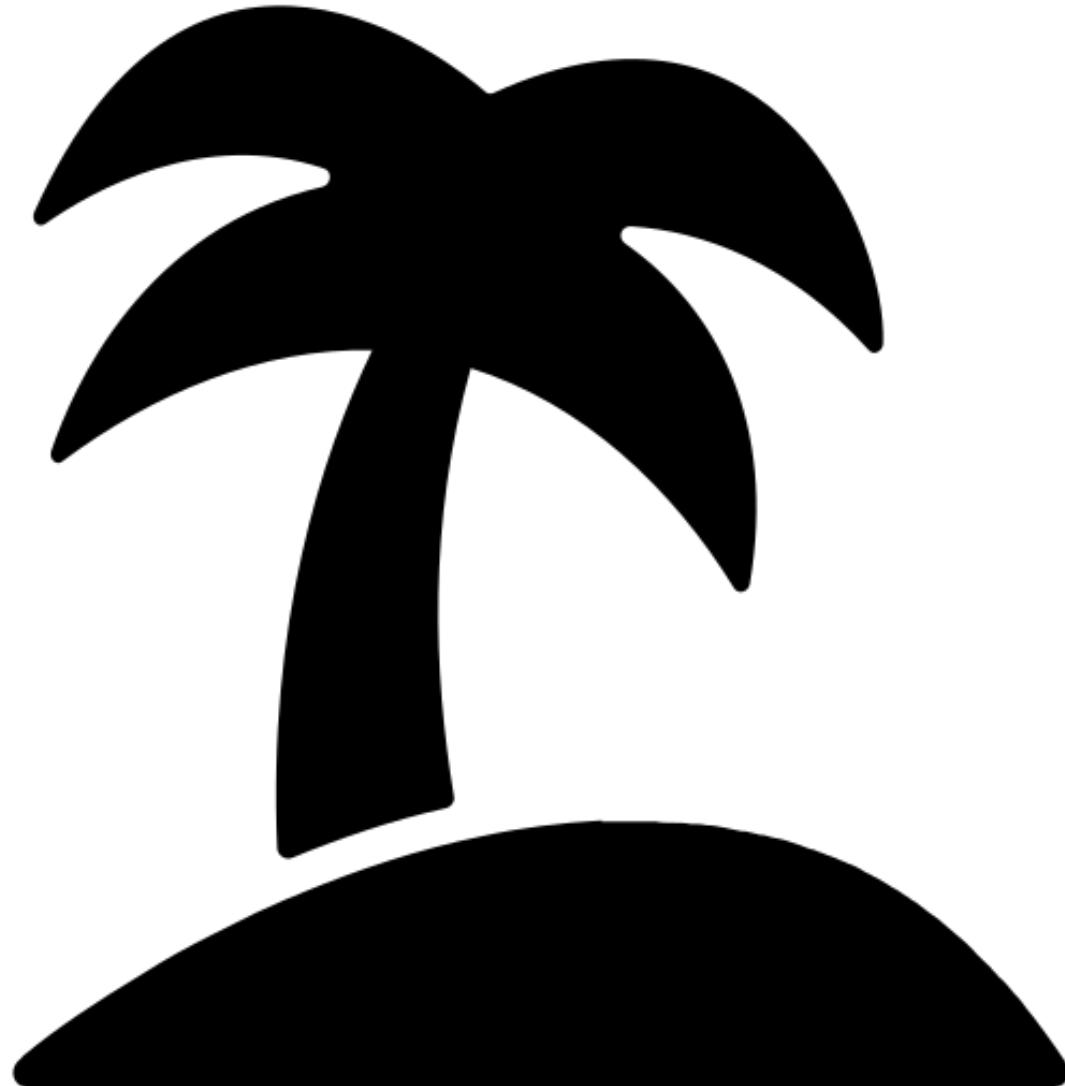
other digital solutions

# THE LONE RANGER AND THE SILVER BULLET



myth #2:

Data Scientists  
work alone

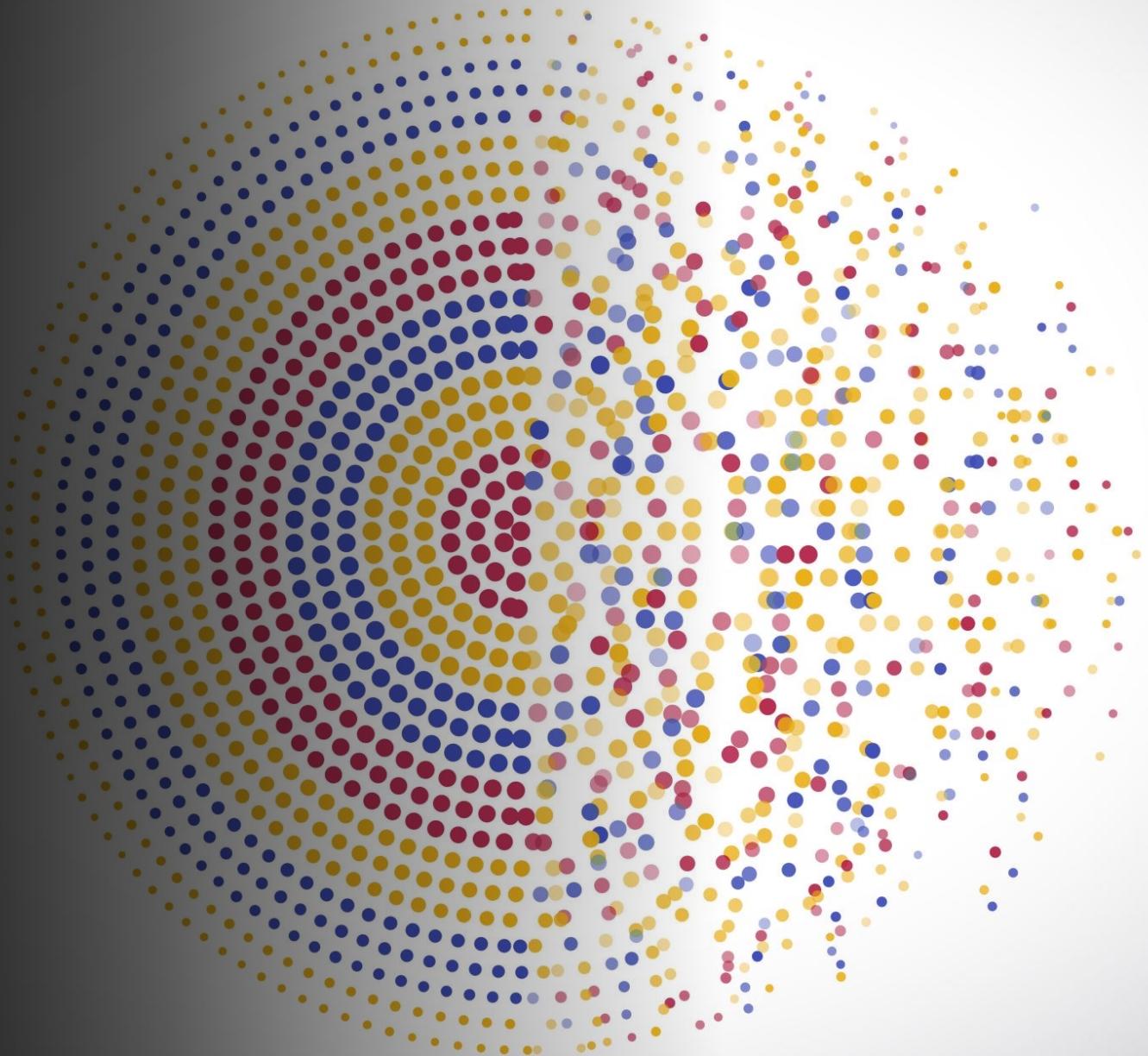


in practice:  
no Data Scientist  
is an island



# Data Science in practice

---





DATA  
SCIENCE  
SHOP

by Marco Morales, PhD

UNOFFICIAL  
TRAILER

think of Data Science as a play on Broadway!



# there is a cast...

Zazu



confidant

Scar



antagonist

Simba



protagonist

Rafiki



tertiary character

# the cast performs on a scenery...



lights

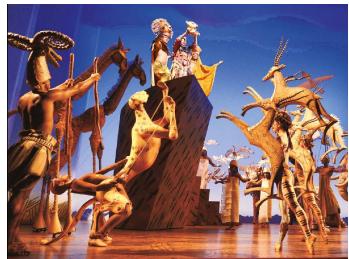
costumes

sound

structures

# the plot is built from a sequence of scenes...

ACT 1



[...]



ACT 2



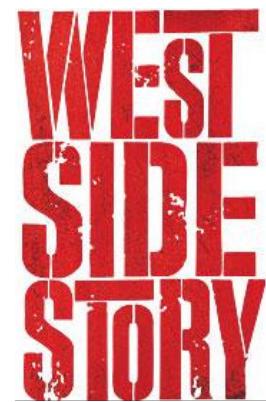
[...]



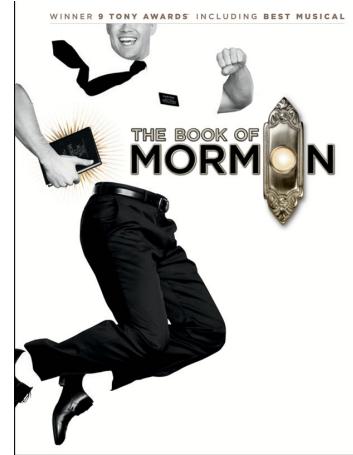
scene 14  
“Hakuna Matata”

scene 27  
“Pride Rock”

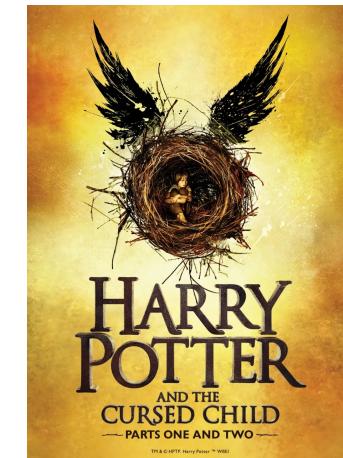
# the play could be of any genre...



tragedy



comedy



drama

# in a nutshell...

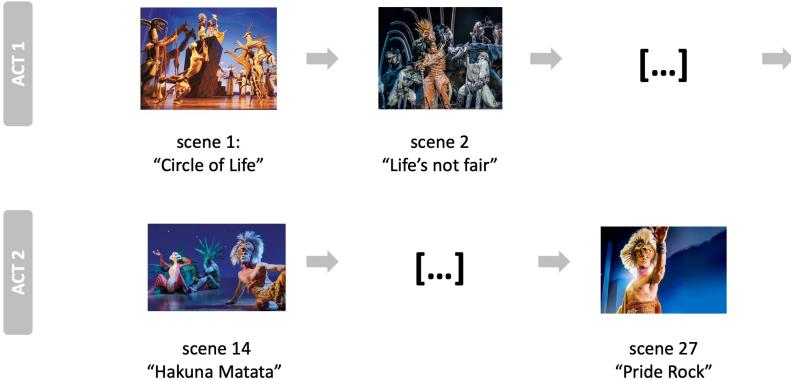
who



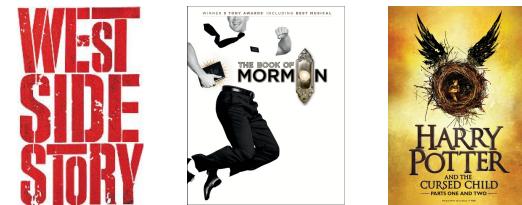
where

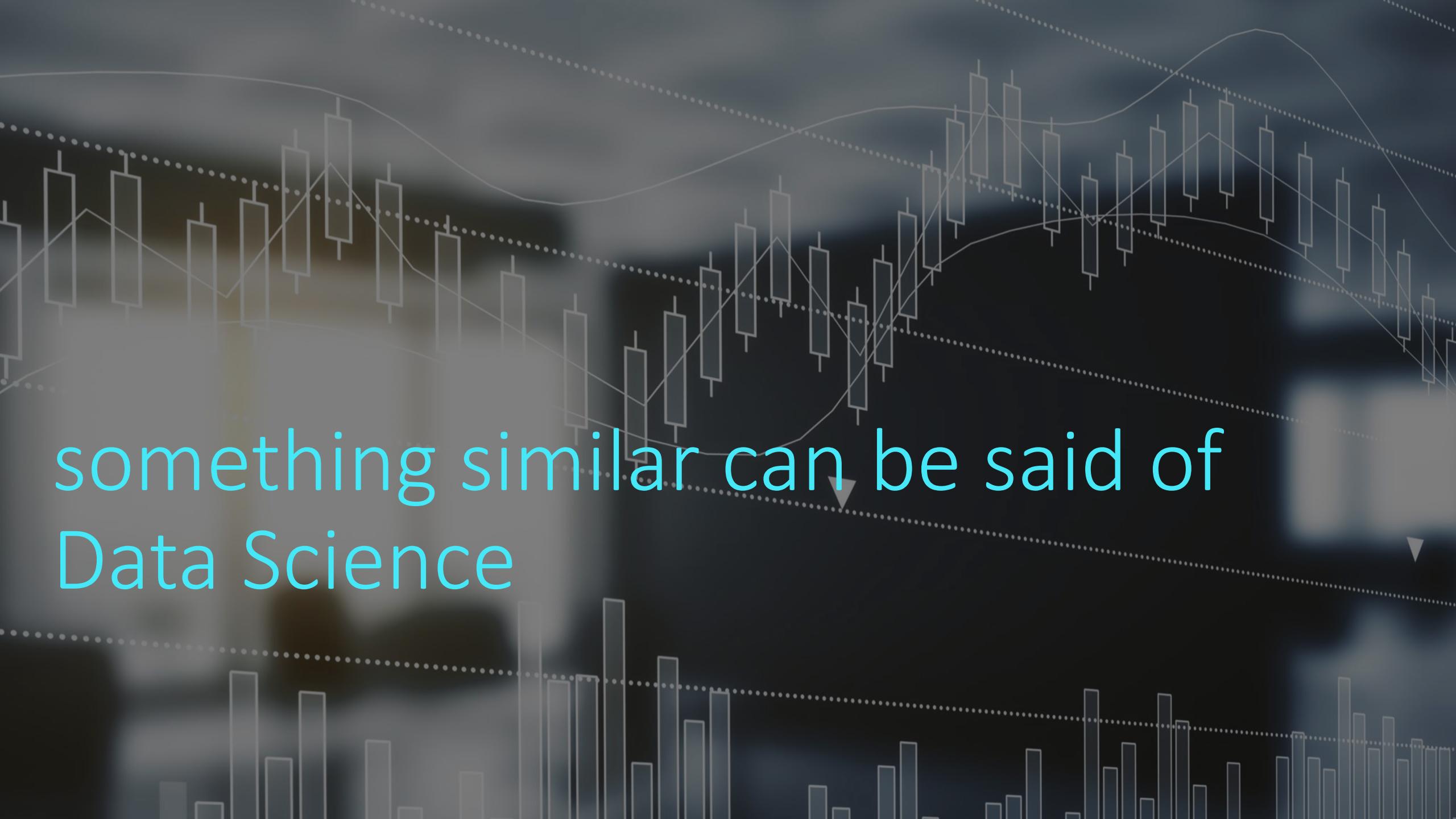


how



what





something similar can be said of  
Data Science

```
>>> re.sub(pattern, repl)
```

```
pattern = [  
    'CAST',  
    'SCENERY',  
    'PLOT',  
    'GENRES'  
]
```

```
repl = [  
    'DATA SCIENCE SHOP CREW',  
    'TECHNOLOGY ENVIRONMENTS',  
    'DATA PRODUCT CYCLE',  
    'DATA PRODUCTS'  
]
```

# 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(bases)
- **summarize** and **visualize** data
- **identify trends**
- **interpret** findings
- **communicate** with business



## data engineer

- develop and maintain **data architecture**
  - data ingestion
  - data storage
  - data security
  - data transformation
- build **data pipelines**
- productionize **ETL**
- build **data quality processes**
- 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

# where does the Data Science Shop operate?

where

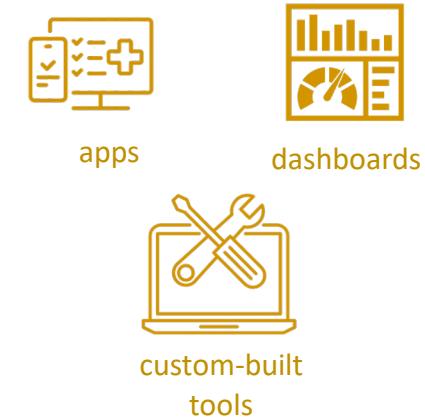
data  
architecture



computing  
architecture



solutions  
architecture



what

who



data  
engineer



ML  
engineer



data  
scientist

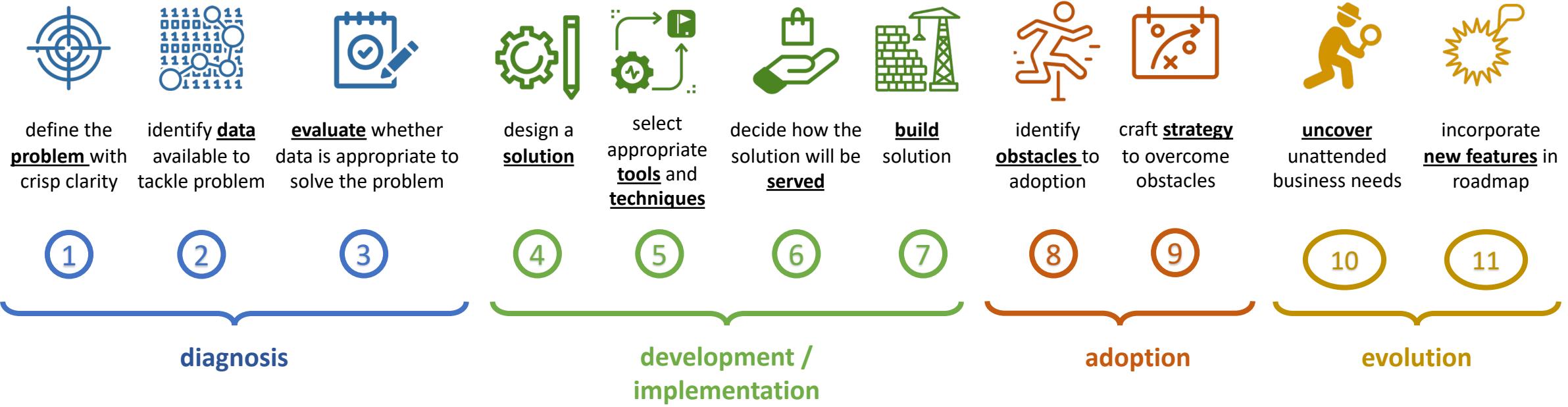


data  
engineer



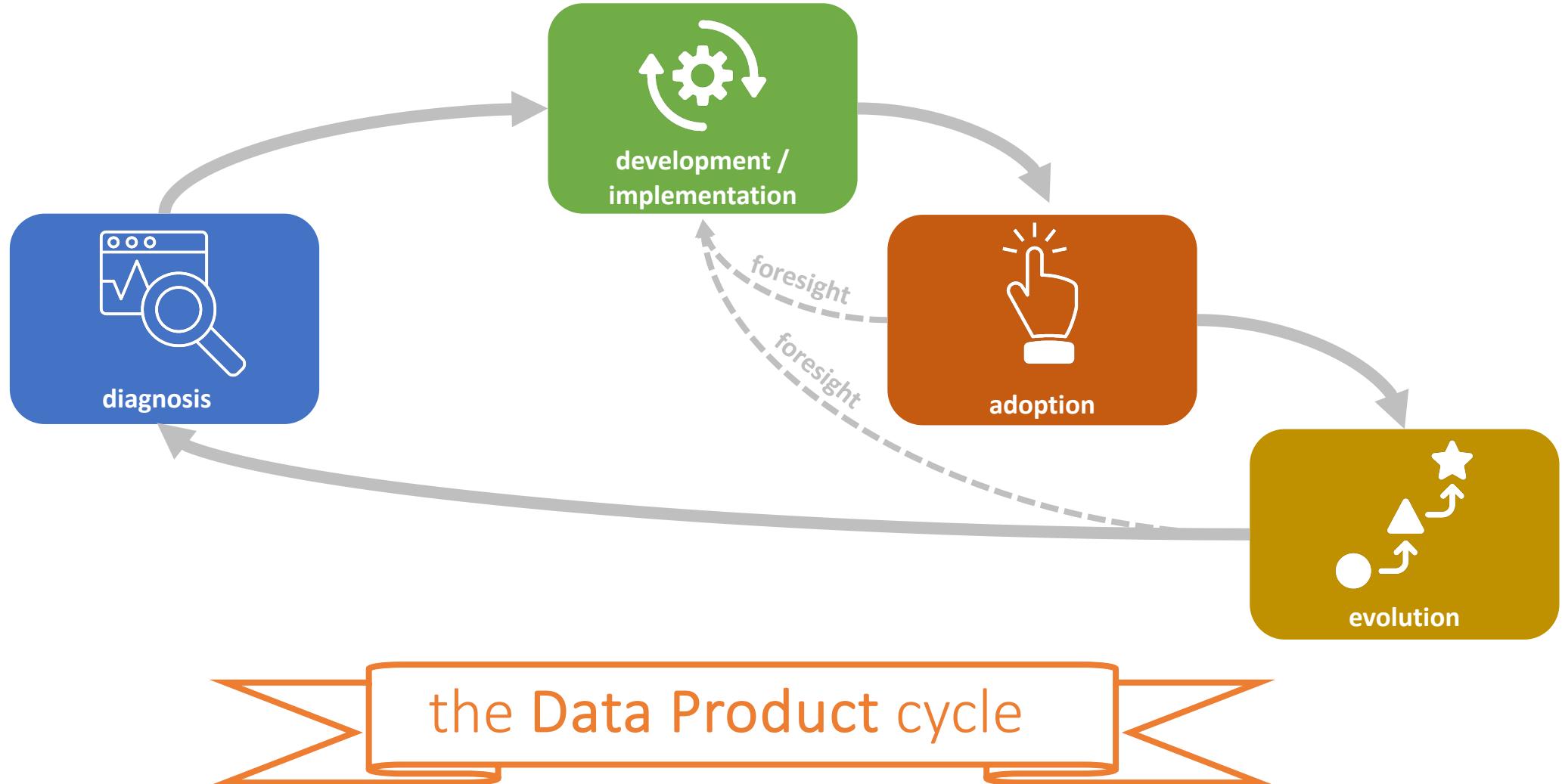
ML  
engineer

# how is a Data Product built?



the Data Product cycle

# how is a Data Product built?

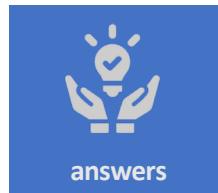


# what Data Products can the Shop build?

what



frontends for  
automated data  
summarization and  
visualization



science-backed  
answers to business  
questions  
(explanations, scenarios,  
projections, causes)



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            data  
engineer    scientist



data  
engineer    ML  
engineer    data  
scientist



data  
engineer    ML  
engineer    data  
scientist    data  
analyst    project  
manager

where



solutions  
architecture



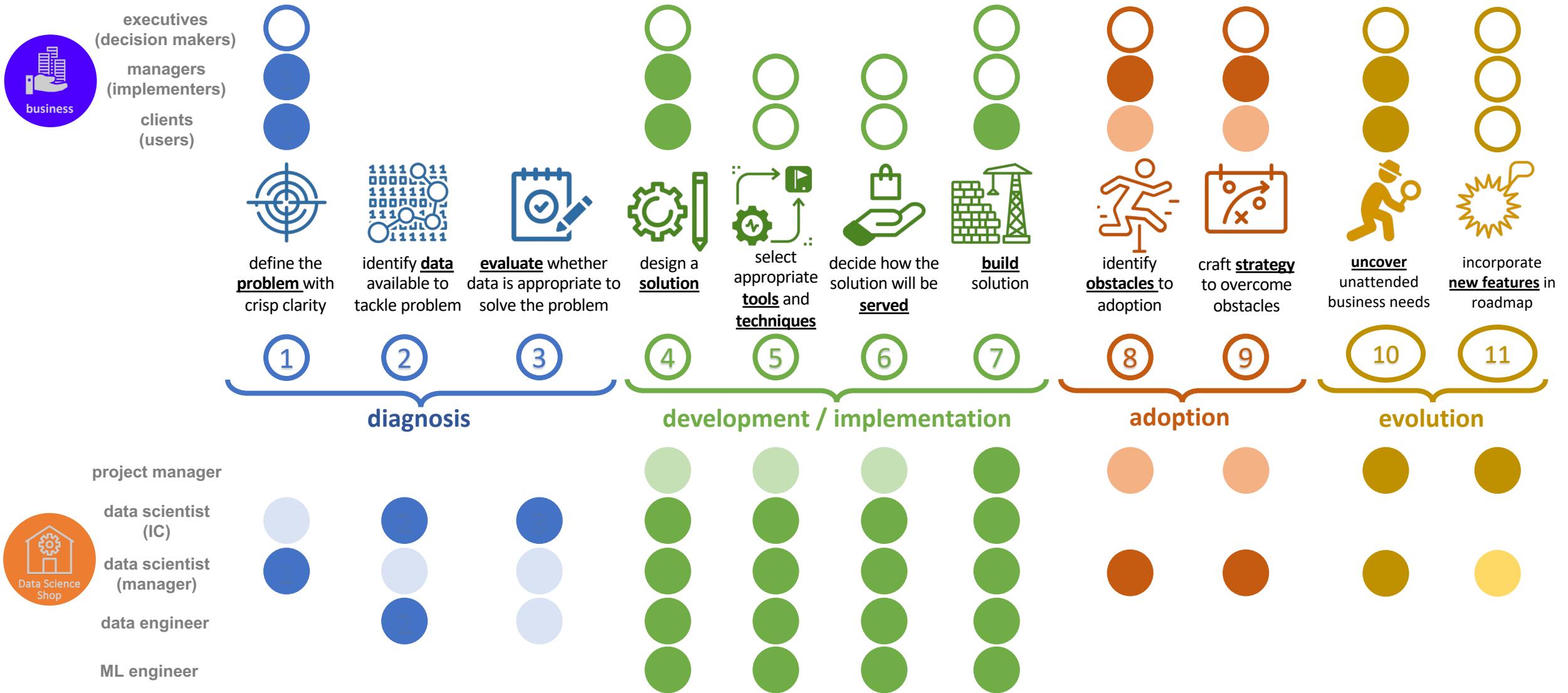
computing  
architecture



computing  
architecture



# how does the Data Science Shop operate?



the problem defines the type of shop



**Problem:** a statement without (an appropriate) solution

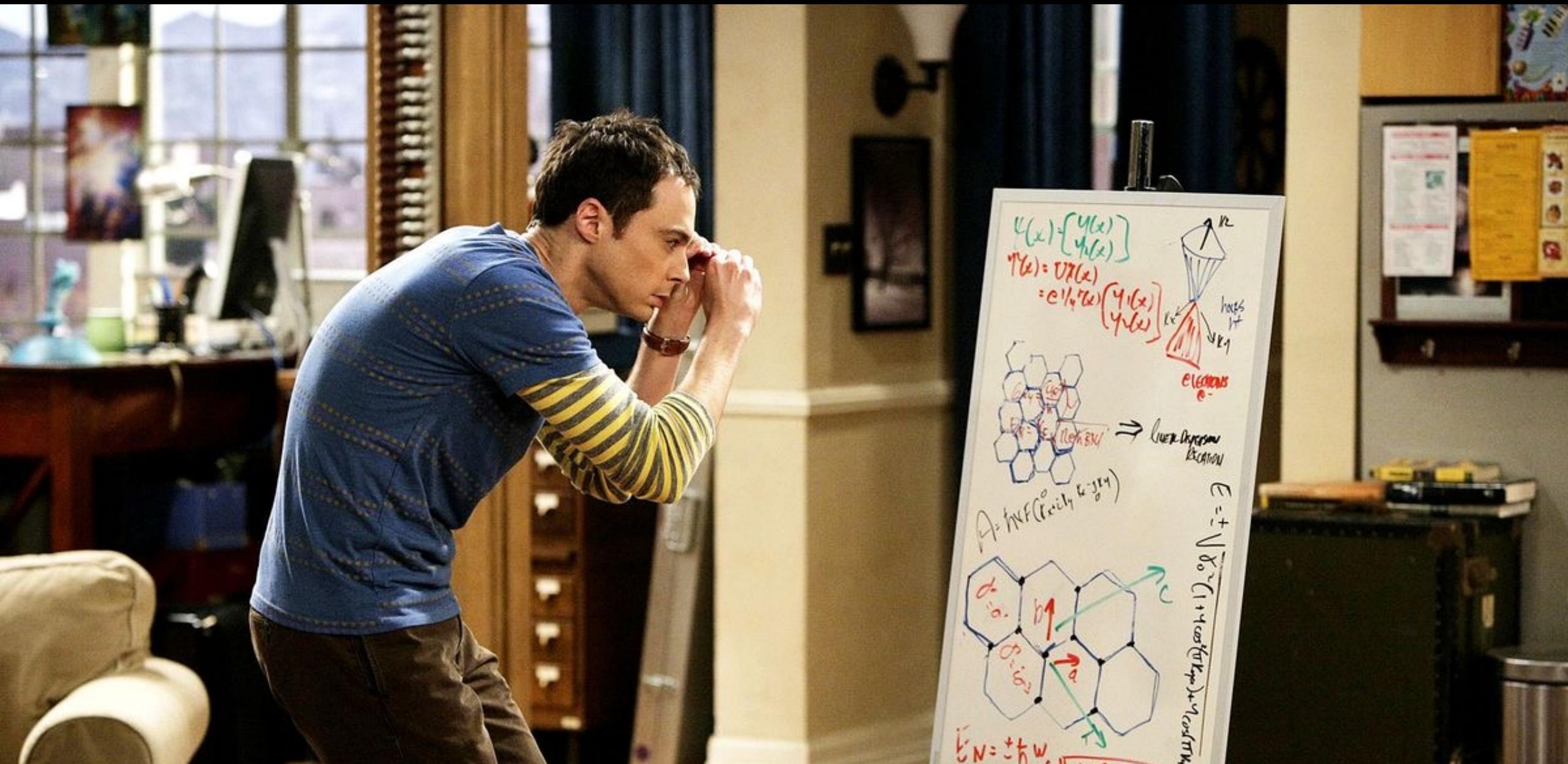
**Solution:** a data product that (effectively) mitigates a problem



DATA  
SCIENCE  
SHOP

**COMING SOON**

# well, there's a little more to it than that....



1

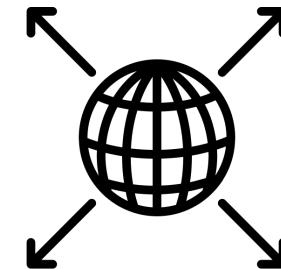
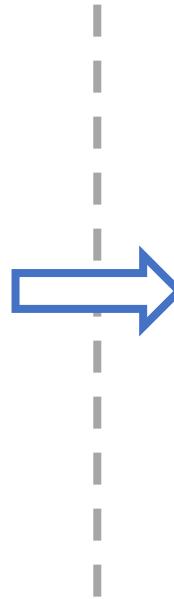
# how does the Data Science Shop do it?



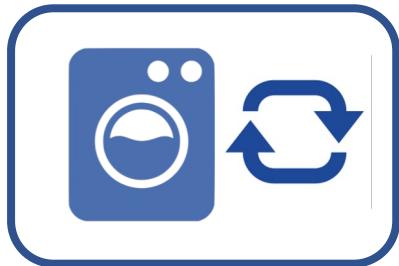
start small  
(MVP)



fail fast



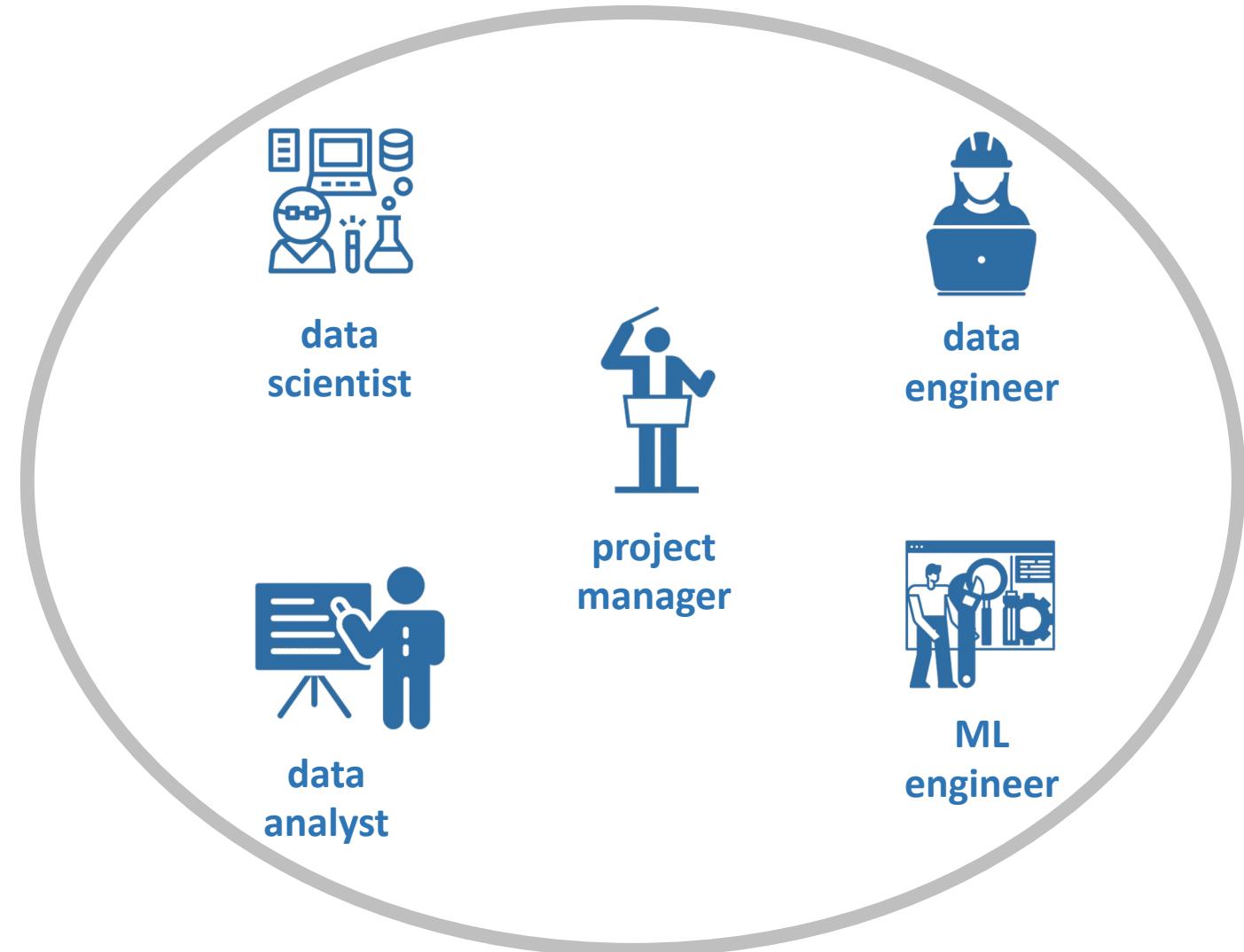
scale up



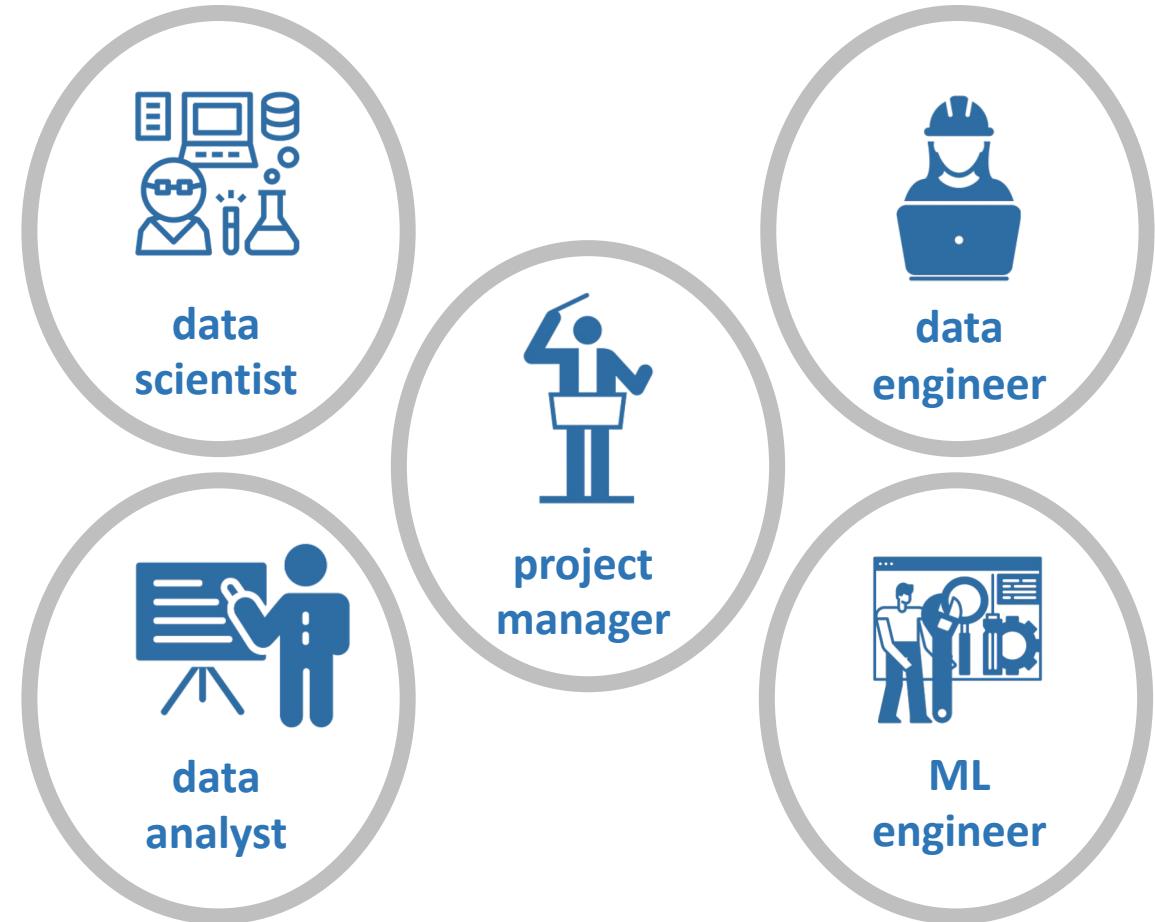
iterate

2

# circa 2010: the unicorn approach



# today: specialization & collaboration!



# the Data Science shop crew in detail

tasks      outputs      skills



data scientist

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



data analyst

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



data engineer

- 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**



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

- **prototyped solutions**
- science-backed solutions

- **insights**

- **data architectures**
- **quality-checked data pipelines**

- **computation-optimized solutions**
- **production-ready solutions**

- **roadmaps**
- **execution**

- 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

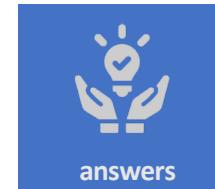
3

# the Data Science Shop: *mutatis mutandis*



Data Science  
Shop

[developed]



answers



deployments

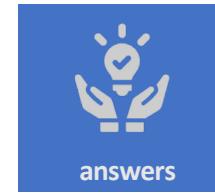


custom-built tools



Data Science  
Shop

[embryonic]



answers



deployments



custom-built tools



# Data Science as a function

Marco Morales

[marco.morales@columbia.edu](mailto:marco.morales@columbia.edu)

Nana Yaw Essuman

[nanayawce@gmail.com](mailto:nanayawce@gmail.com)

GR5069: Applied Data Science  
for Social Scientists

Spring 2024  
Columbia University