



INTRODUCTION TO THE DATA SCIENCE PIPELINE

ALEXANDRE GRAMFORT
CLAIRE BOYER

A word about me

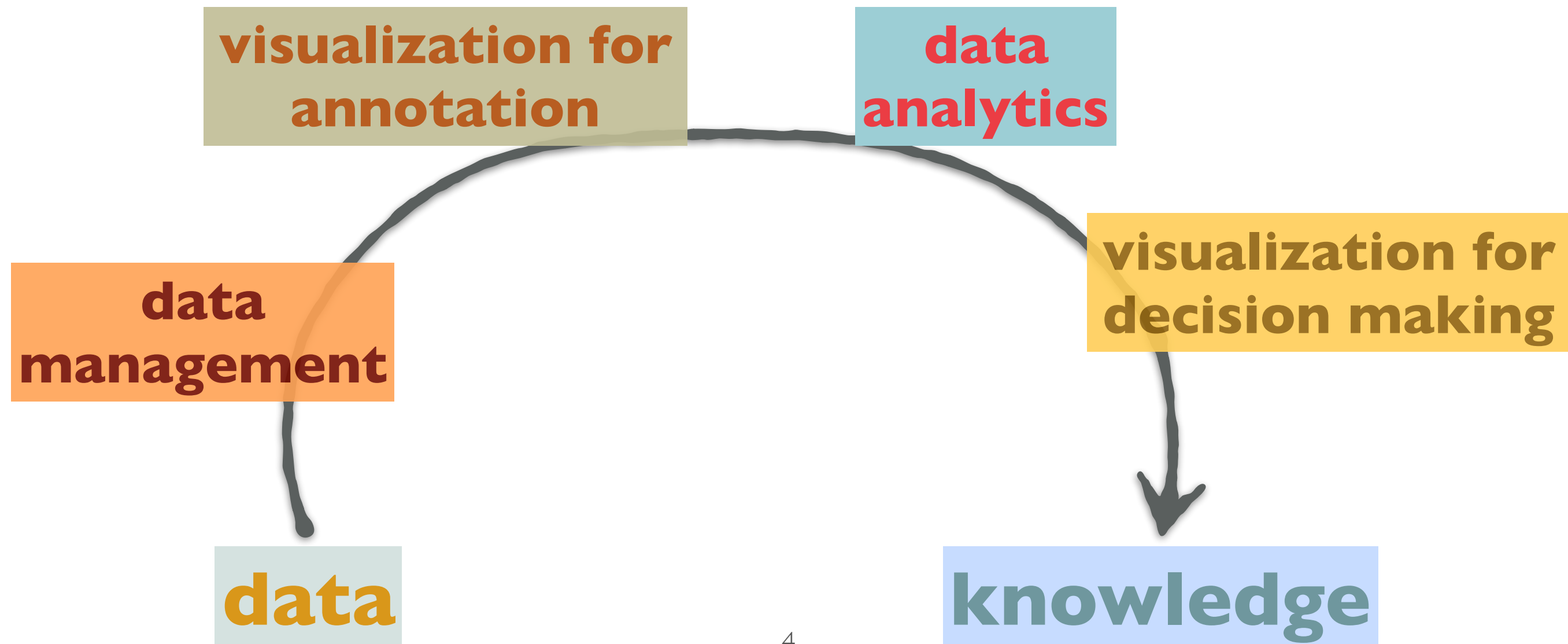
- Senior scientist (Directeur de Recherche) at **INRIA**
- <http://alexandre.gramfort.net>
- alexandre.gramfort@inria.fr
- co-author of scikit-learn  
- **Research topics:** machine learning, non-linear optimization, signal processing, applications in health care and particularly in neuroscience.

A word about Claire

- Associate Professor at **UPMC / ENS Ulm**
- <http://www.lpsm.paris/pageperso/boyer/>
- claire.boyer@upmc.fr
- **Research topics:** statistics, machine learning, inverse problems for imaging, compressed sensing

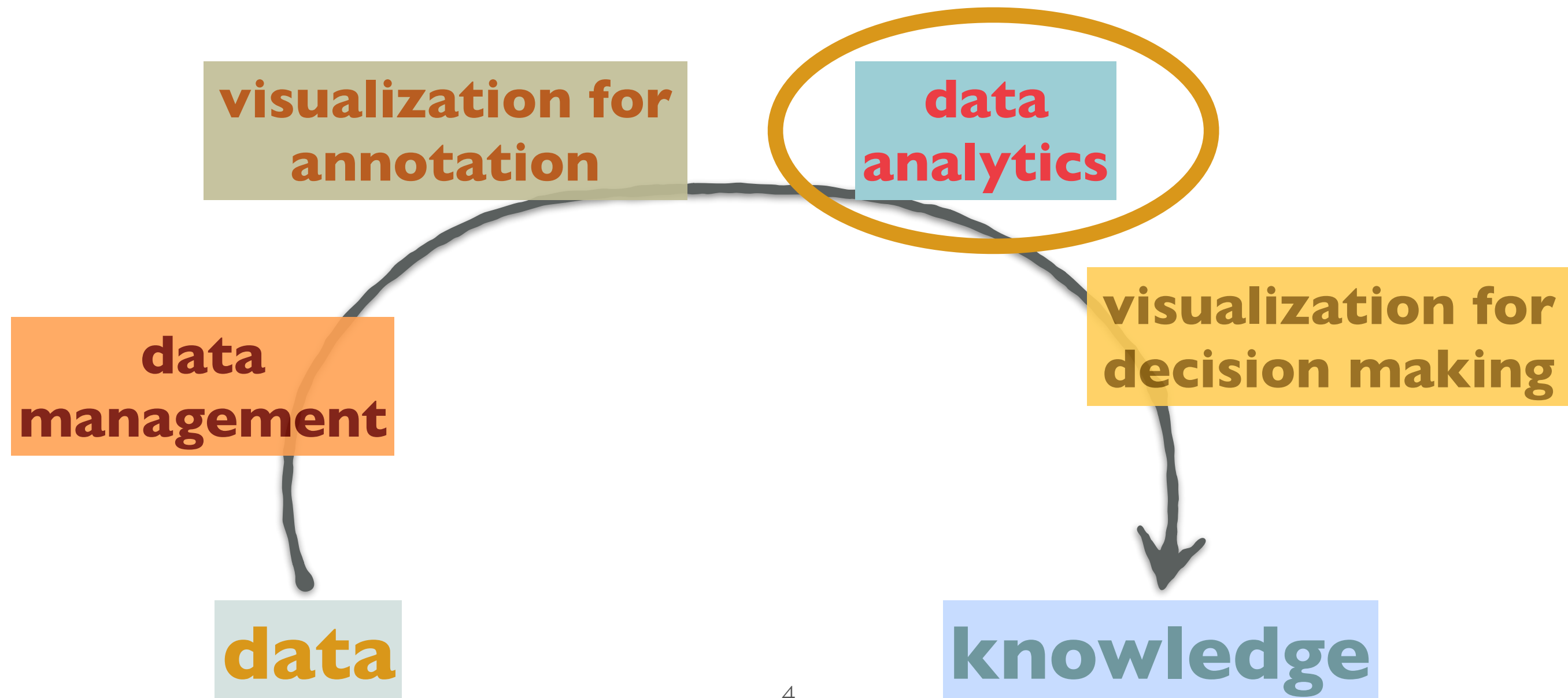
THE FULL DATA CHAIN

The products a data scientist/engineer can end up working on, **after the business case is identified**



THE FULL DATA CHAIN

The products a data scientist/engineer can end up working on, **after the business case is identified**



The data science ecosystem

THE DATA ANALYTICS PRODUCTION PIPELINE

THE DATA ANALYTICS PRODUCTION PIPELINE

web, database, sensors

THE DATA ANALYTICS PRODUCTION PIPELINE

web, database, sensors



raw data, stream

THE DATA ANALYTICS PRODUCTION PIPELINE

web, database, sensors



raw data, stream

**cleaning, formatting,
aligning, preprocessing**

THE DATA ANALYTICS PRODUCTION PIPELINE

web, database, sensors

raw data, stream

**cleaning, formatting,
aligning, preprocessing**

features **x**

THE DATA ANALYTICS PRODUCTION PIPELINE

web, database, sensors

raw data, stream

**cleaning, formatting,
aligning, preprocessing**

features **x**

prediction

THE DATA ANALYTICS PRODUCTION PIPELINE

web, database, sensors

raw data, stream

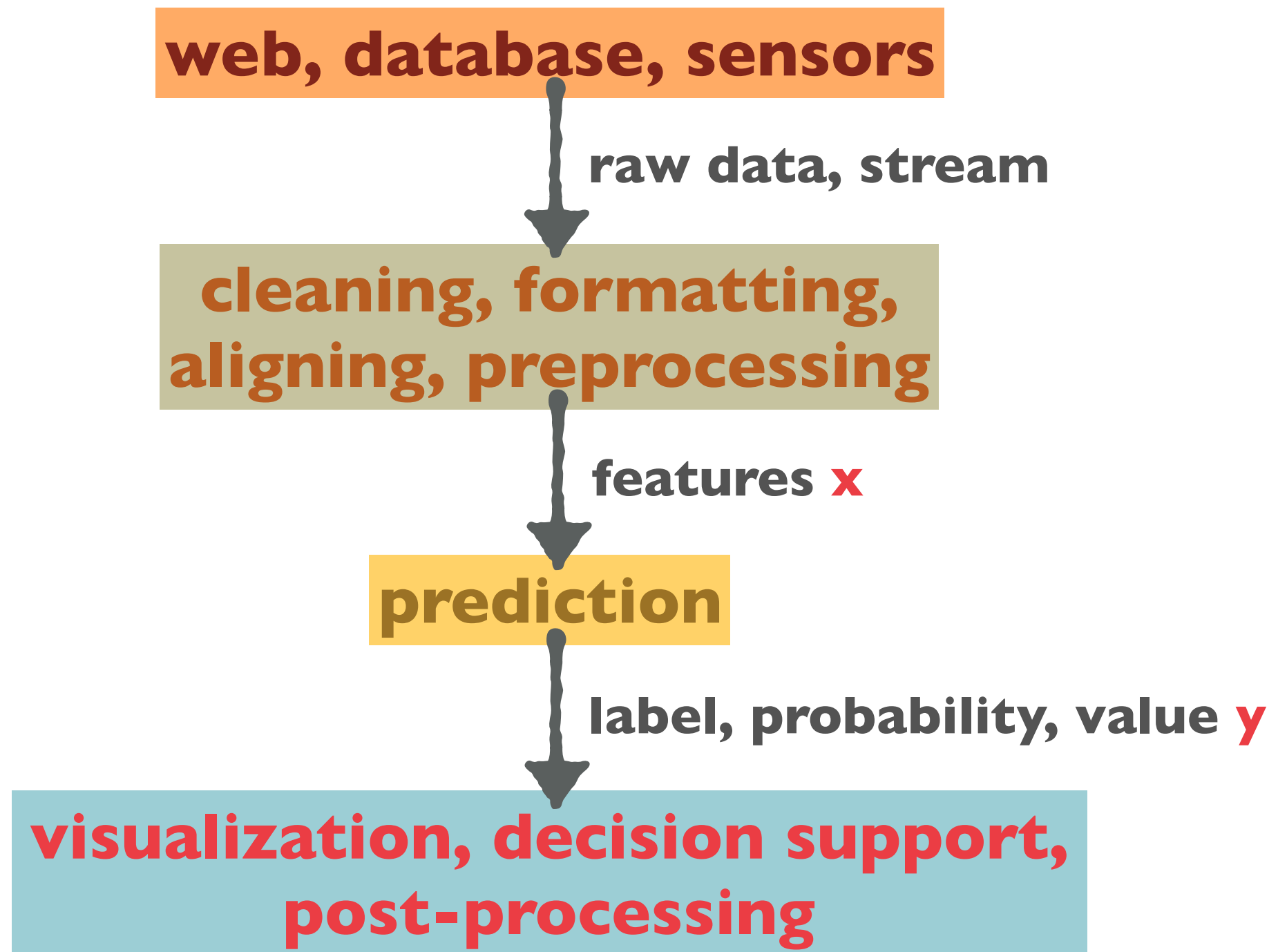
**cleaning, formatting,
aligning, preprocessing**

features **x**

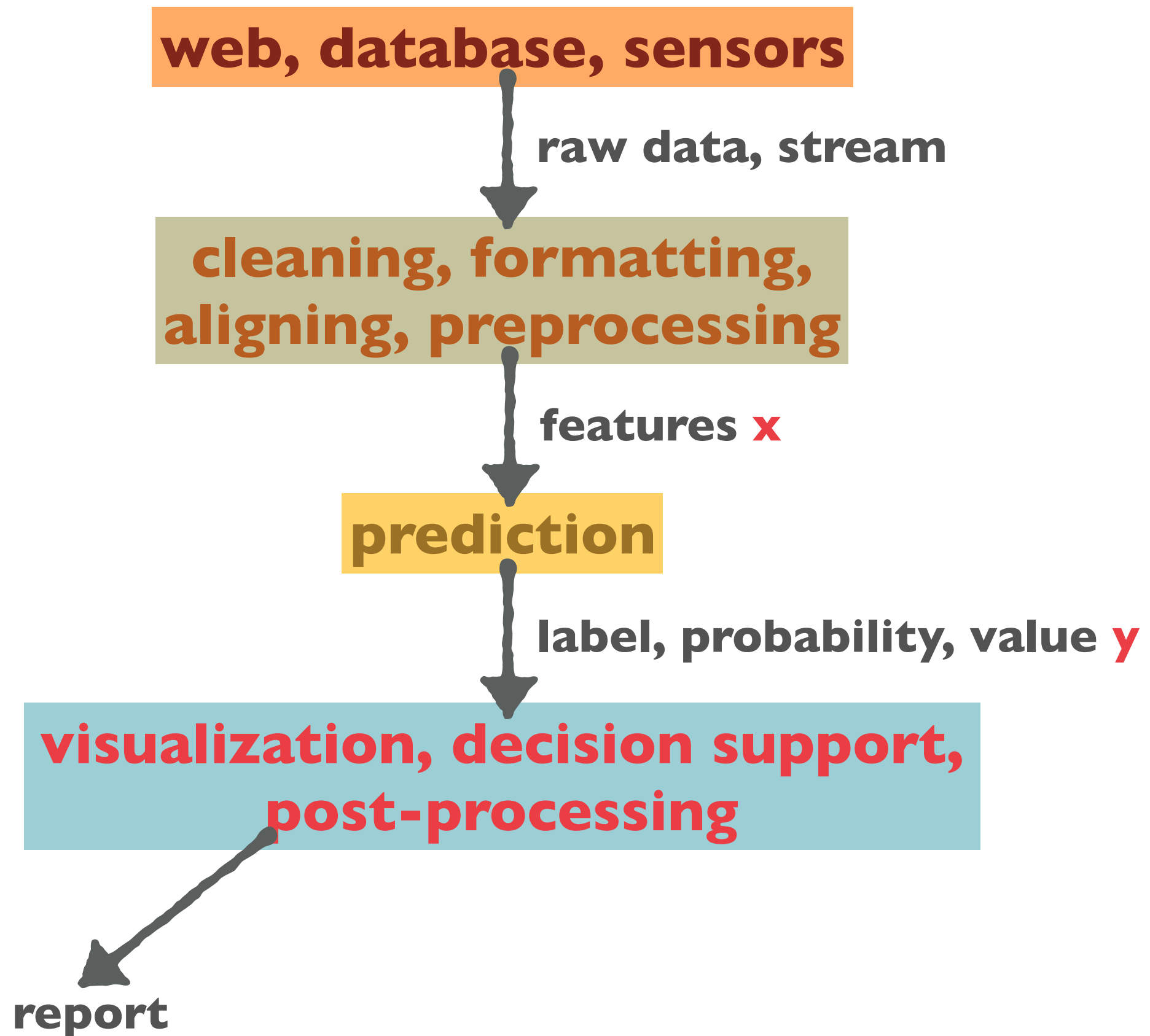
prediction

label, probability, value **y**

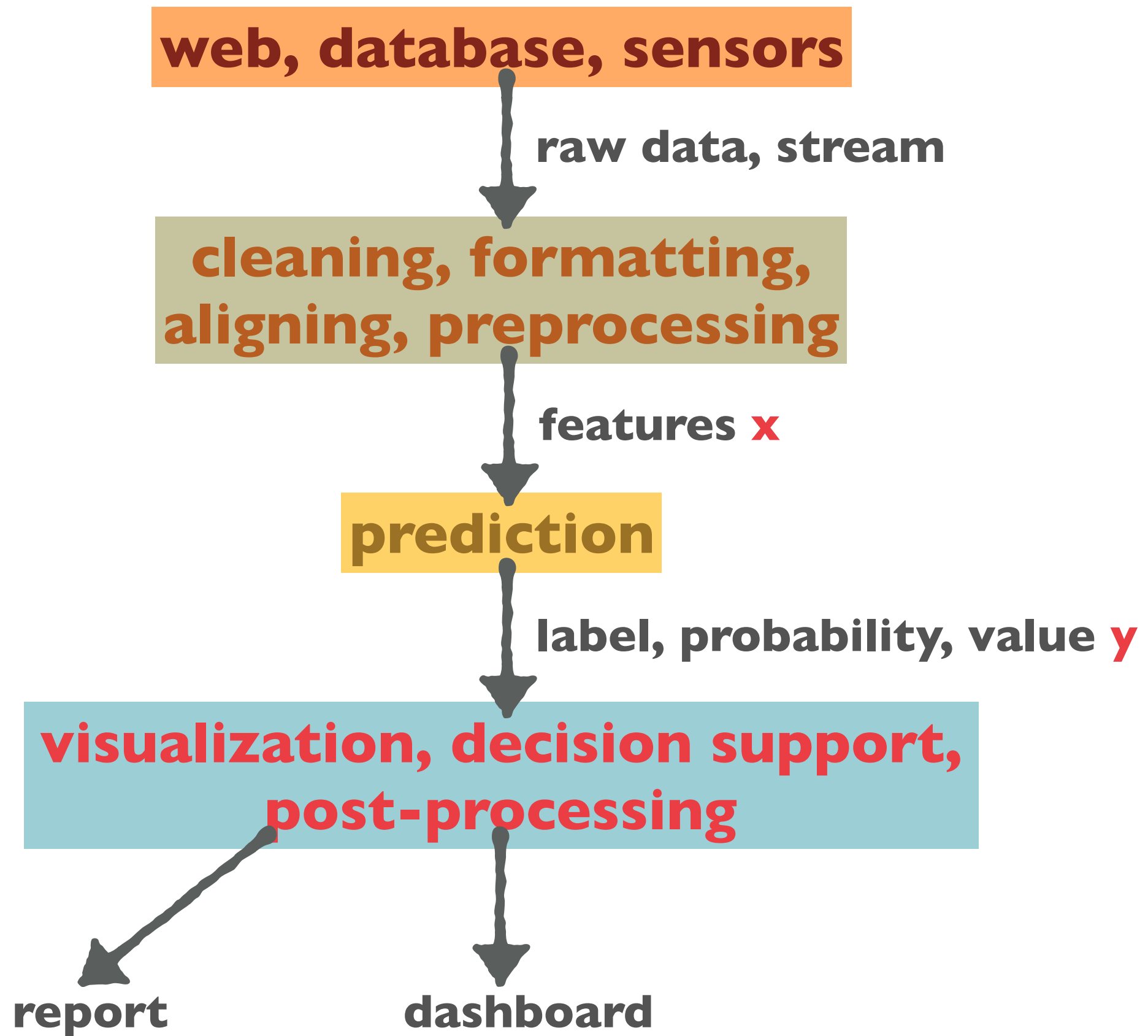
THE DATA ANALYTICS PRODUCTION PIPELINE



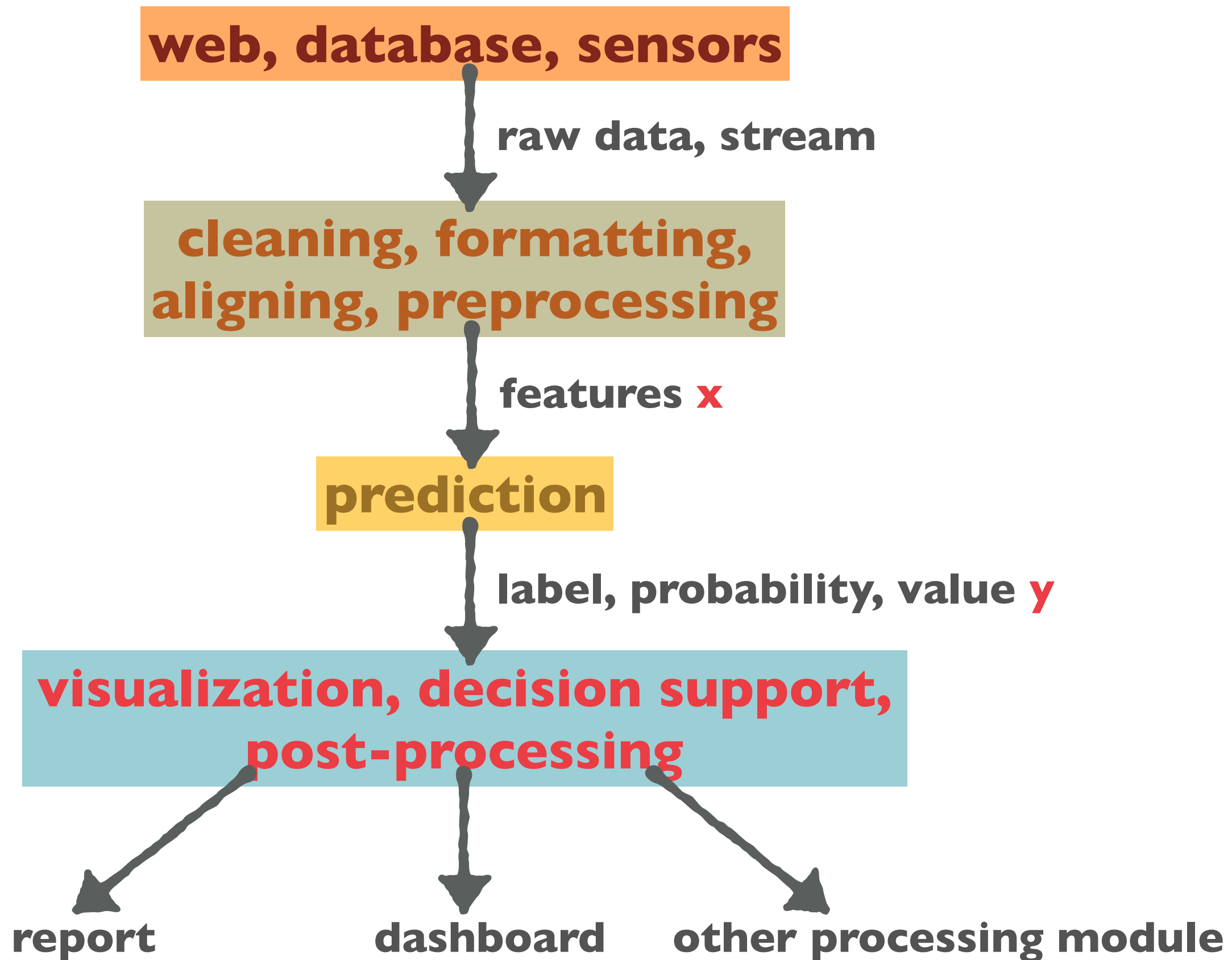
THE DATA ANALYTICS PRODUCTION PIPELINE



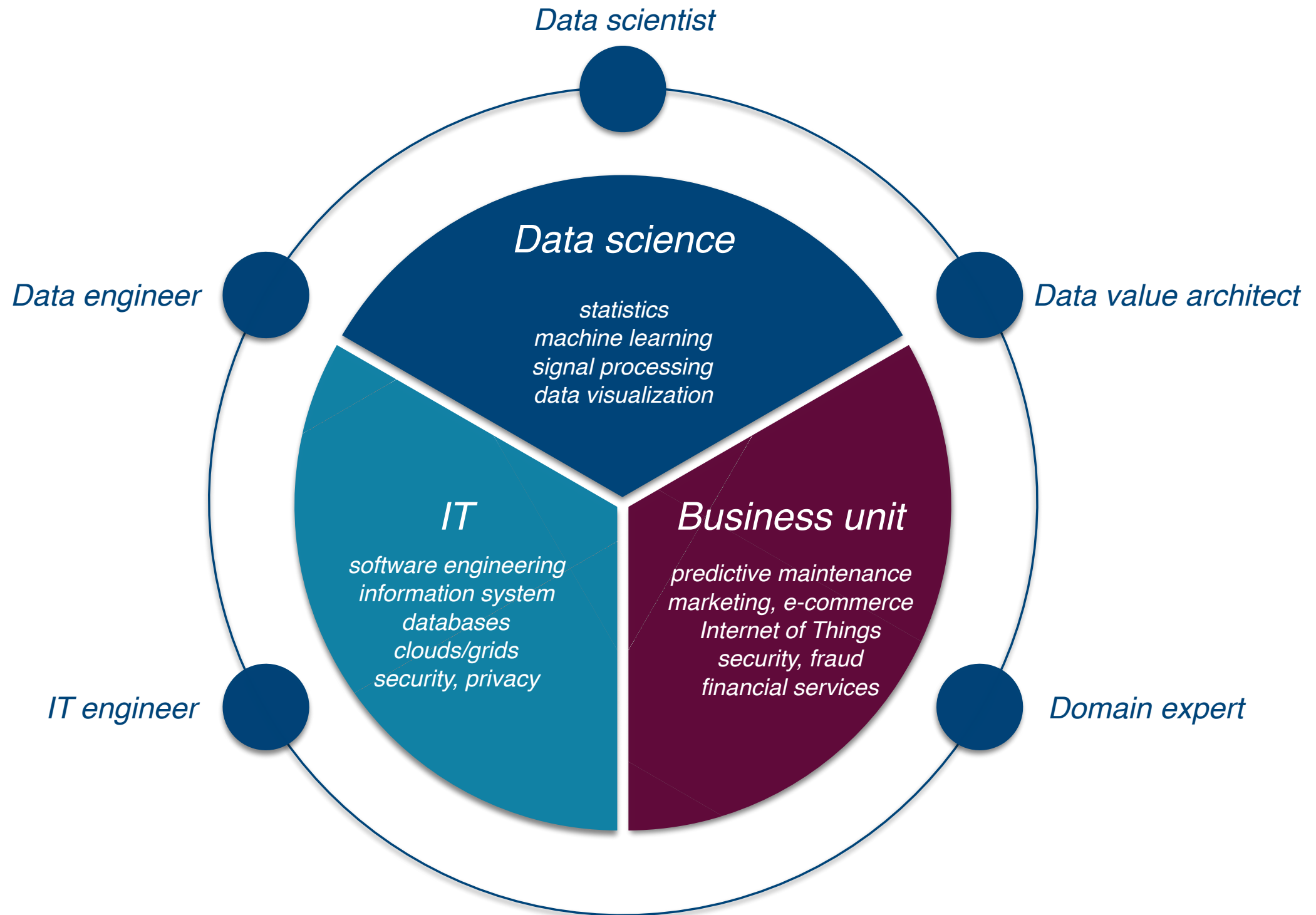
THE DATA ANALYTICS PRODUCTION PIPELINE



THE DATA ANALYTICS PRODUCTION PIPELINE



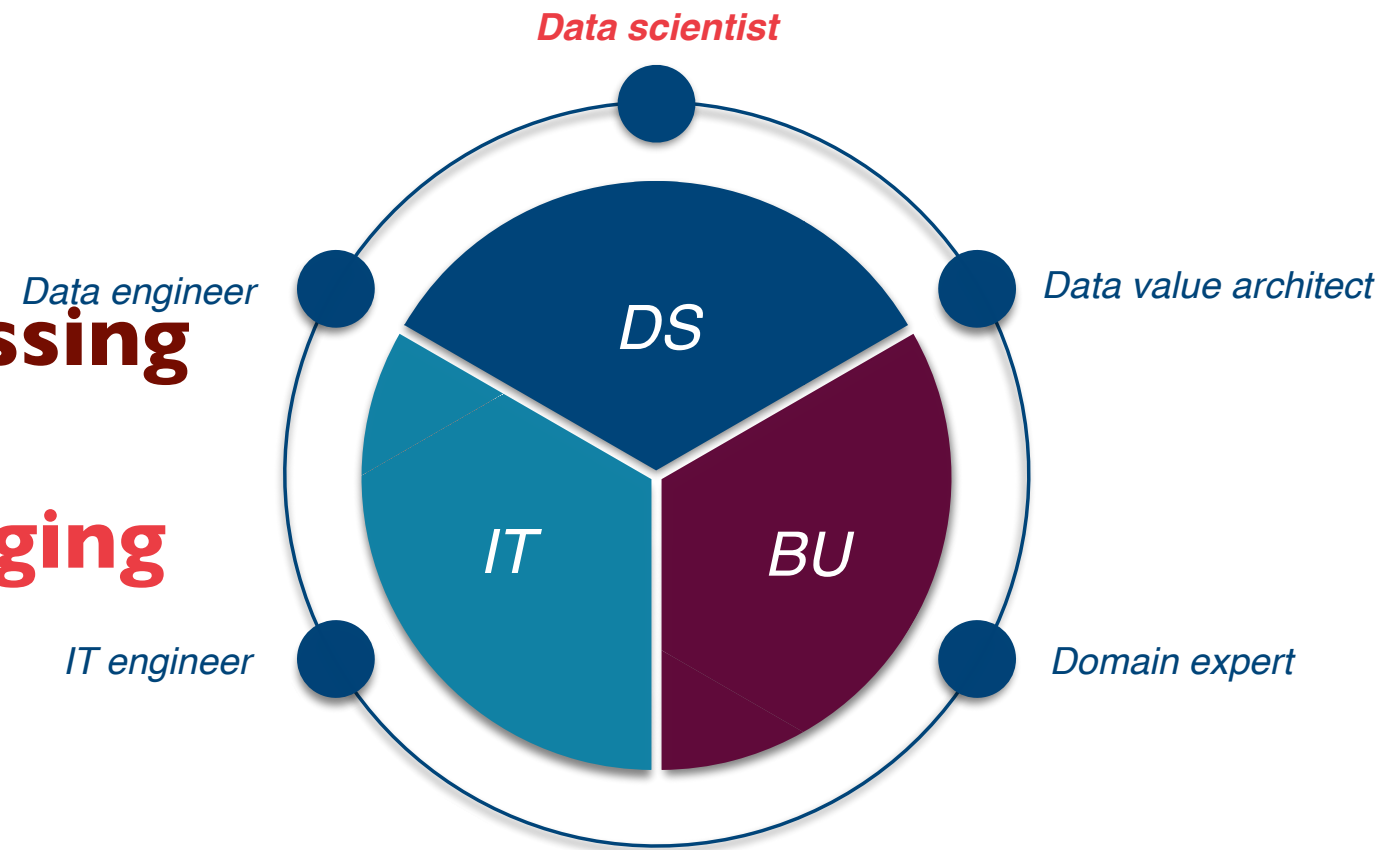
THE DATA SCIENCE ECOSYSTEM



DATA SCIENTIST

THE KAGGLER

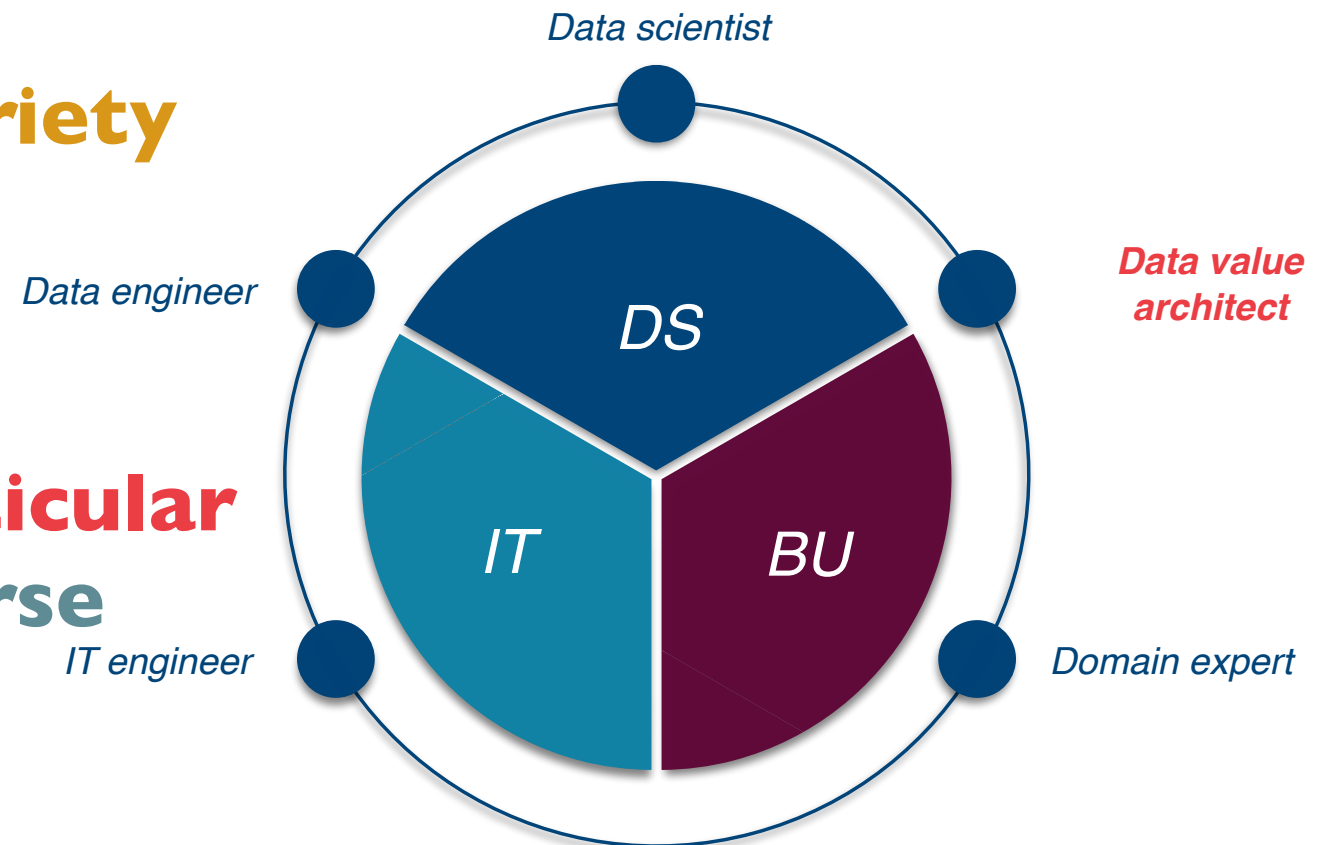
- Technical expert in **machine learning**, **statistics**, **visualization**, **signal processing**
- Efficient in **cleaning** and **munging**
- Knows the latest **techniques** and **tools**
- Can handle different **data types** and **loss metrics**
- Can build adequate **prototype workflows**
- Knows how to **tune** (optimize) and **blend** models



DATA VALUE ARCHITECT

THE *FORMALIZER*

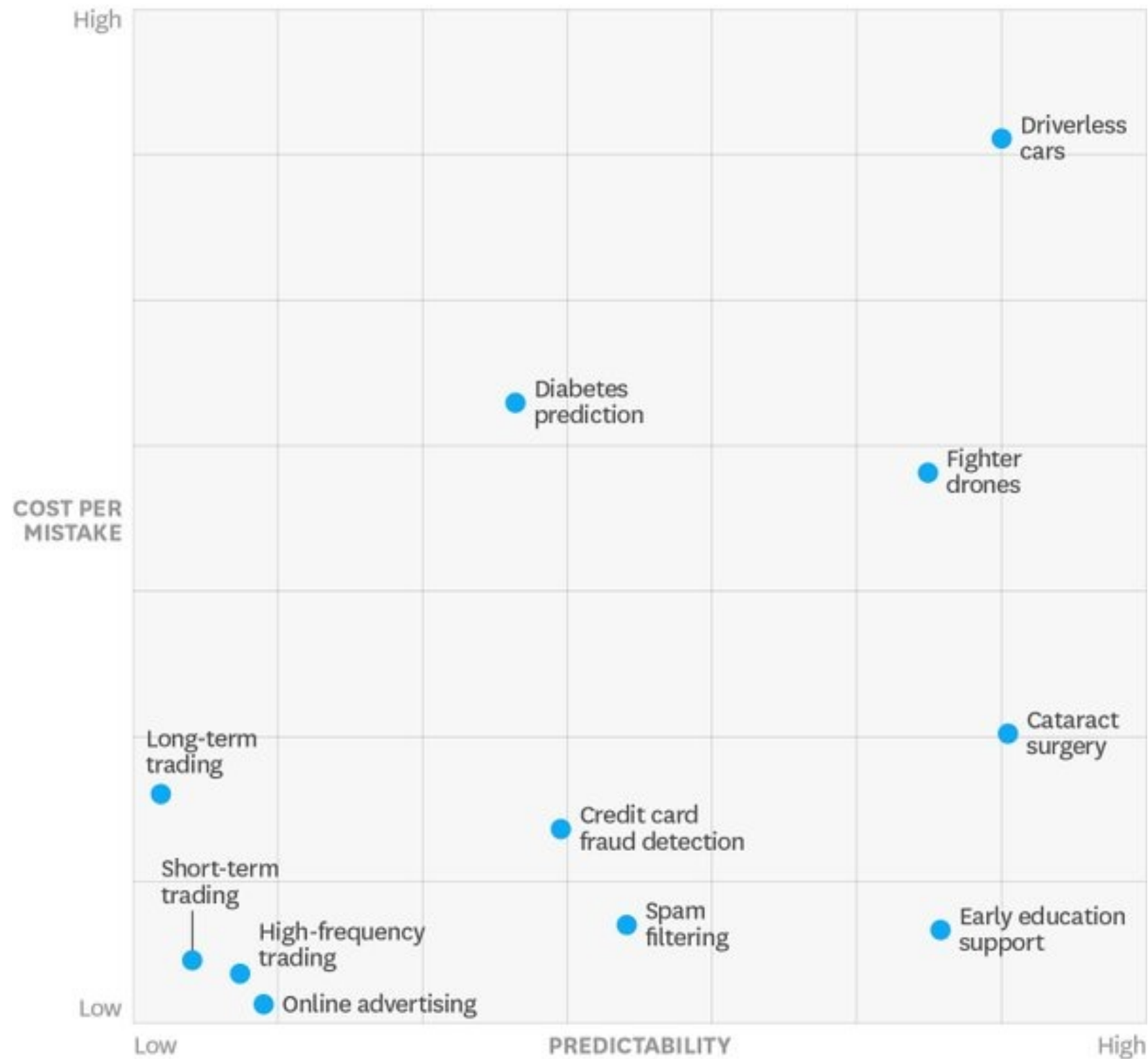
- Has experience with a **wide variety of business problems** and **technical solutions**
- Is possibly **expert in the particular domain**, or at least can **converse** with the domain expert
- Can translate **business goals into loss metrics**
- Can **formalize** adequate **prototype workflows**
- Can **estimate the costs** of building and running workflows
- Can define and dimension the **data collection** effort



DATA VALUE ARCHITECT MINDSET

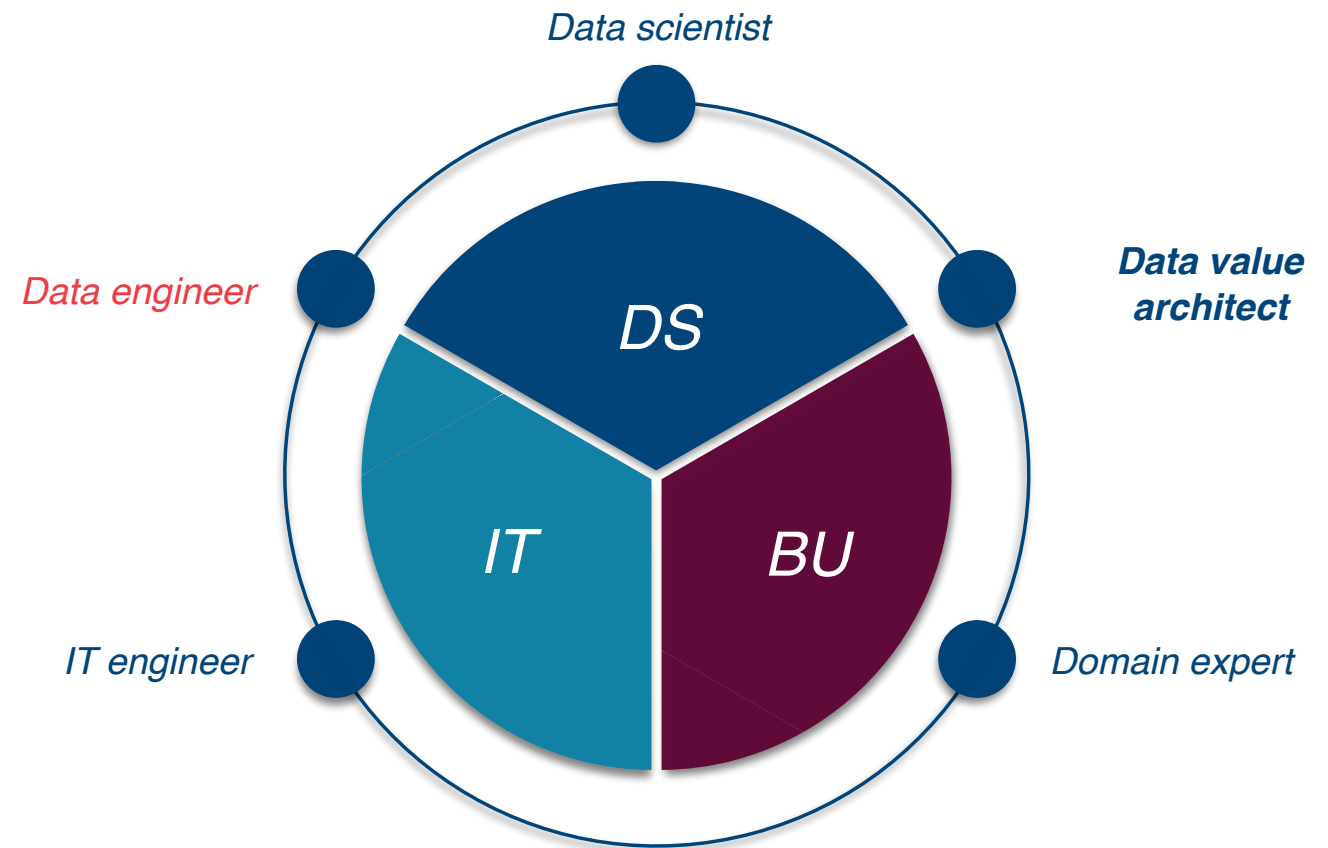
The Decision Automation Map

Plotting how well machines can do at making predictions against the costs of mistakes.

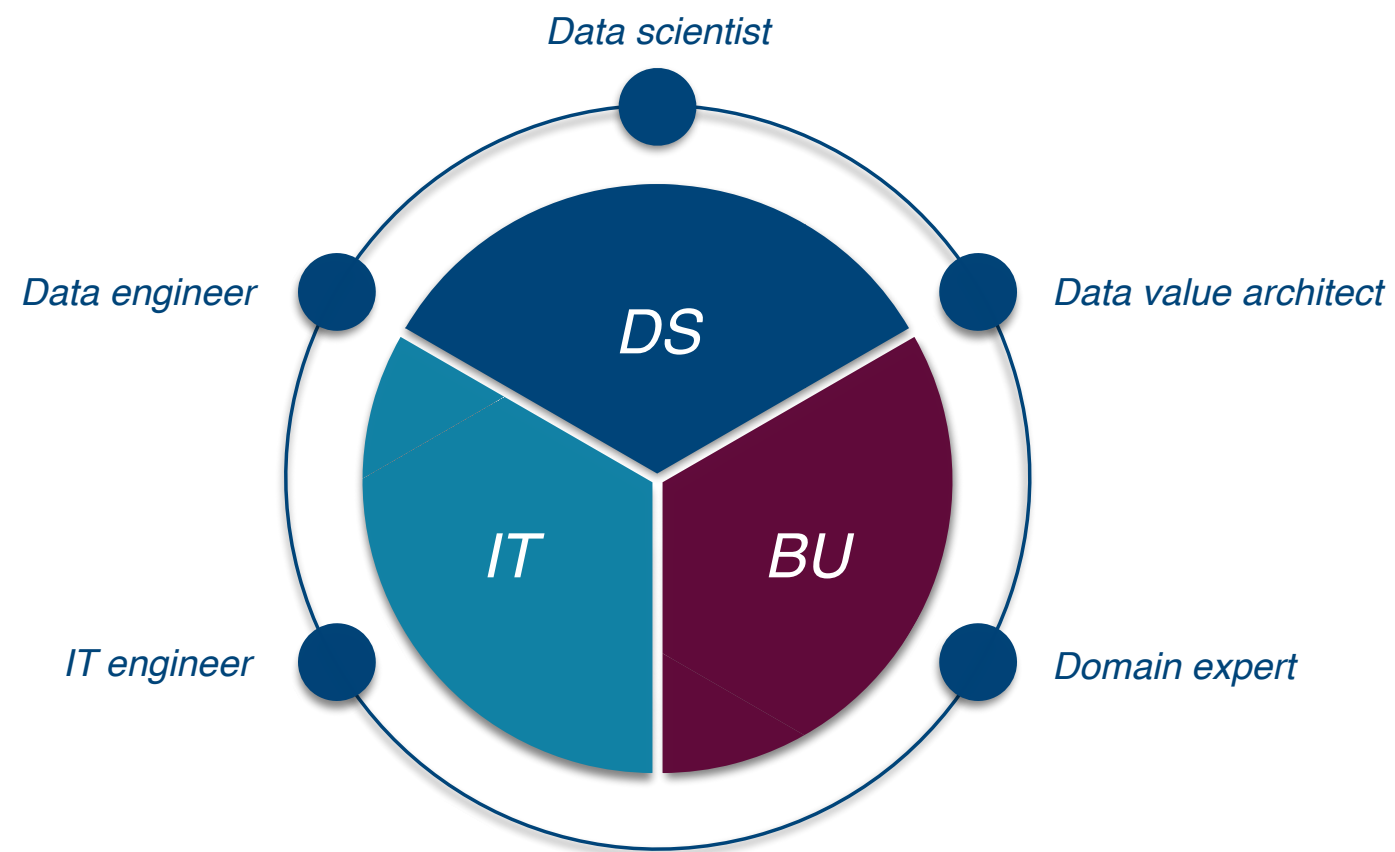


DATA ENGINEER

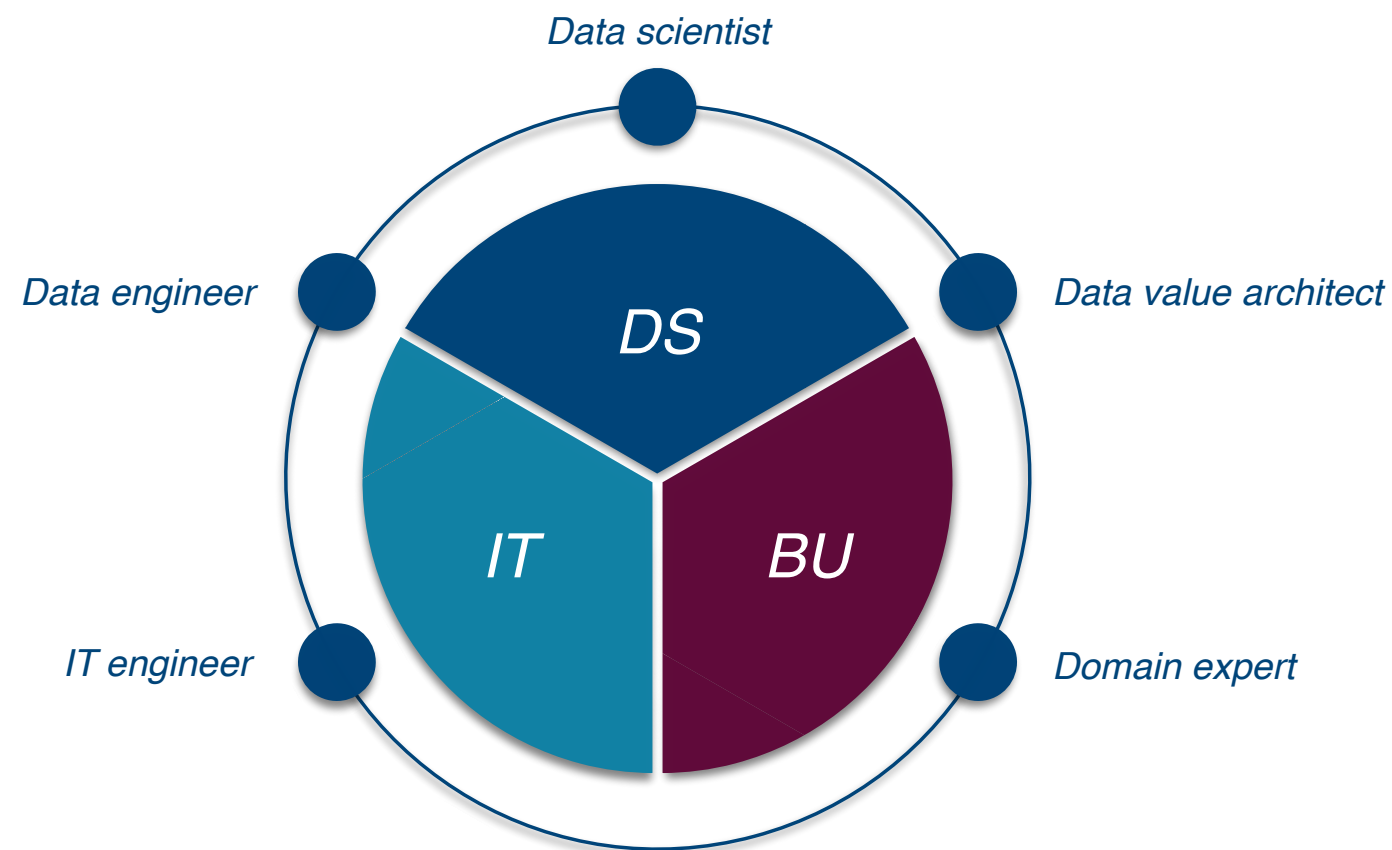
- Translates prototypes into **production workflows**, runs and maintains them
- Knows the latest **data engineering systems** and **architectures**
- Knows the **existing IT**
- Can **dimension the production workflows** and **estimate their costs**
- Knows the **basics of building** a data science workflow, and can feed the process by **extracting** and possibly **cleaning/munging** adequate data



BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY IT

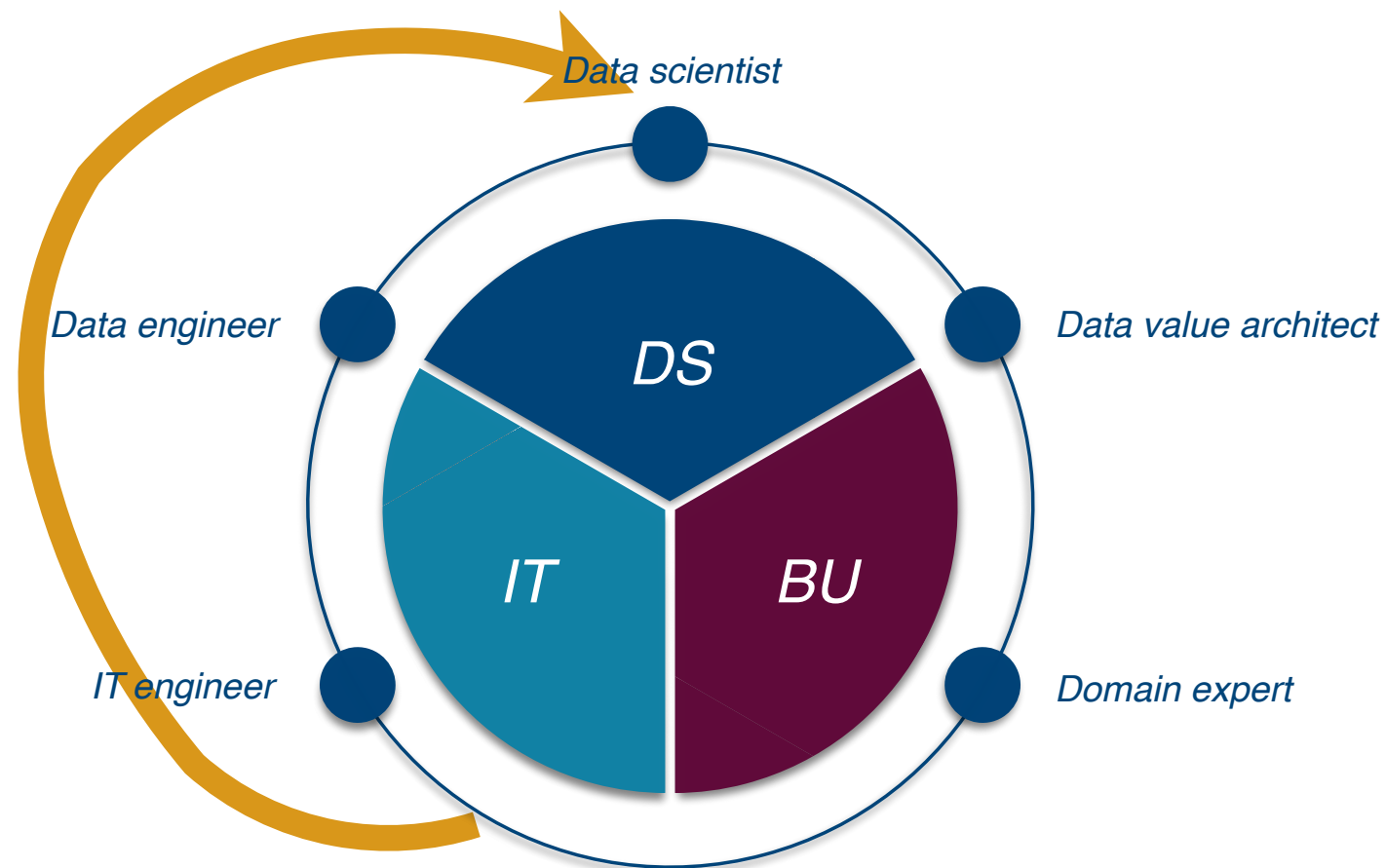


BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY IT



“Let’s install Hadoop”

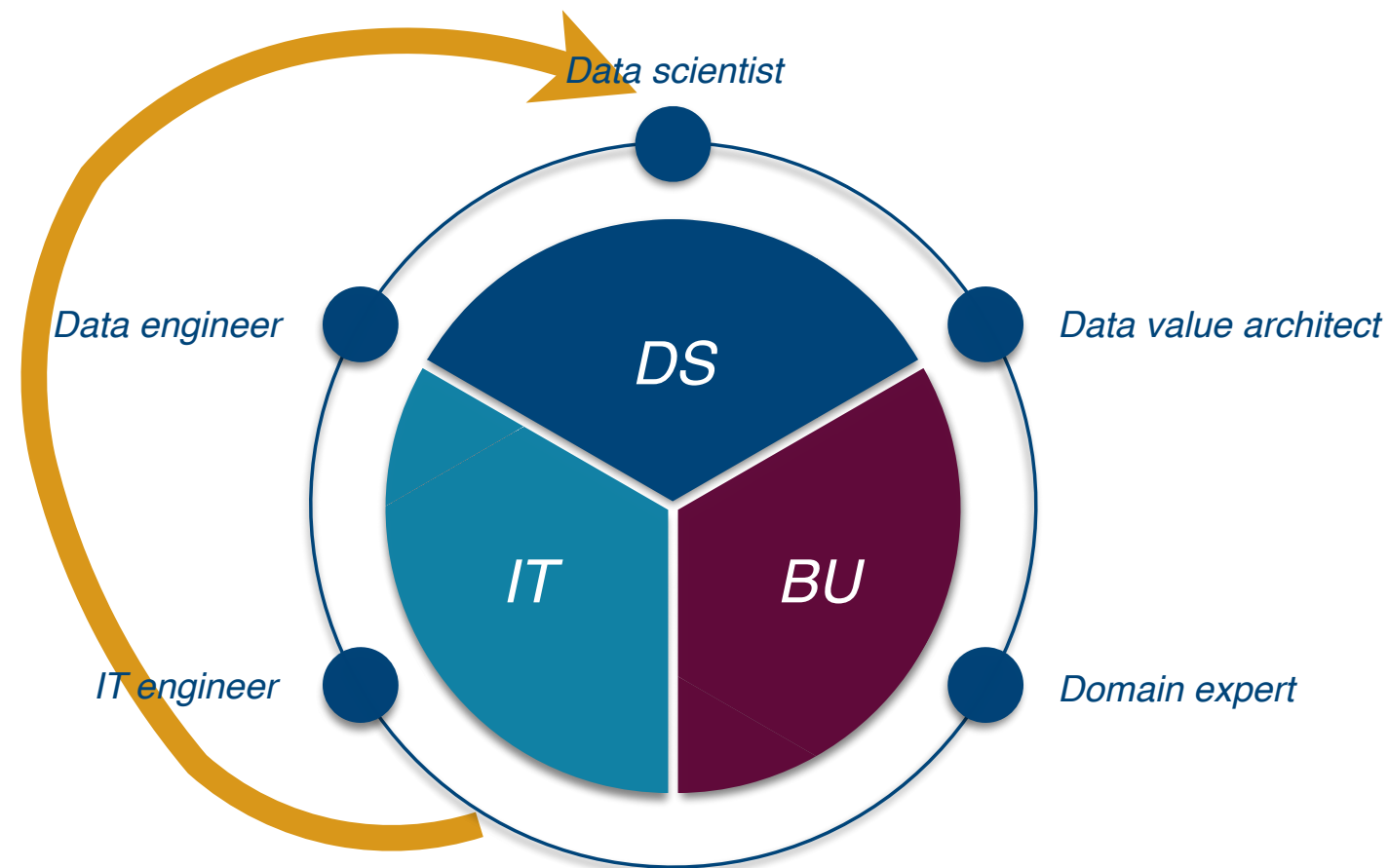
BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY IT



“Let’s install Hadoop”

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY IT

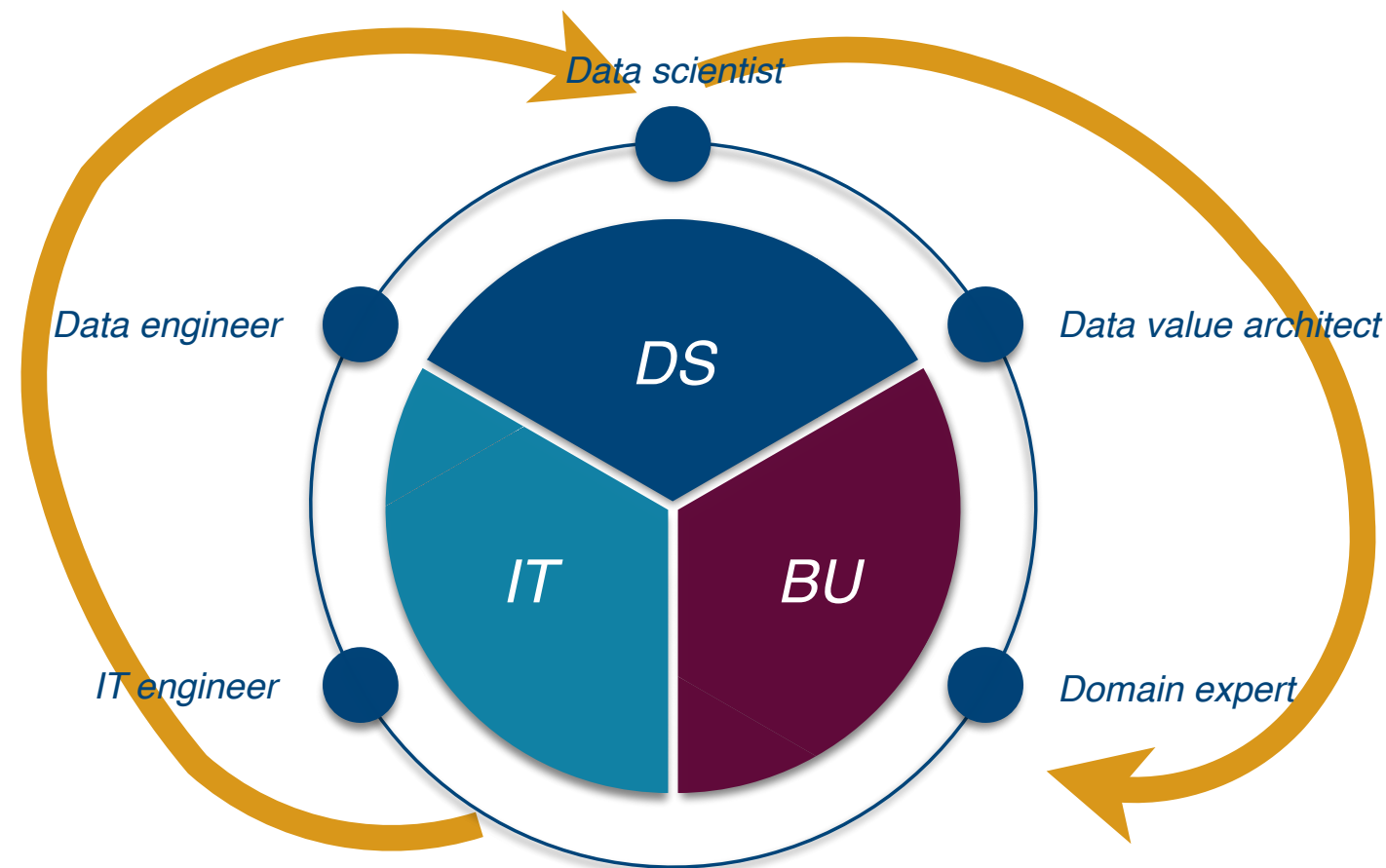
“Let’s hire data scientists”



“Let’s install Hadoop”

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY IT

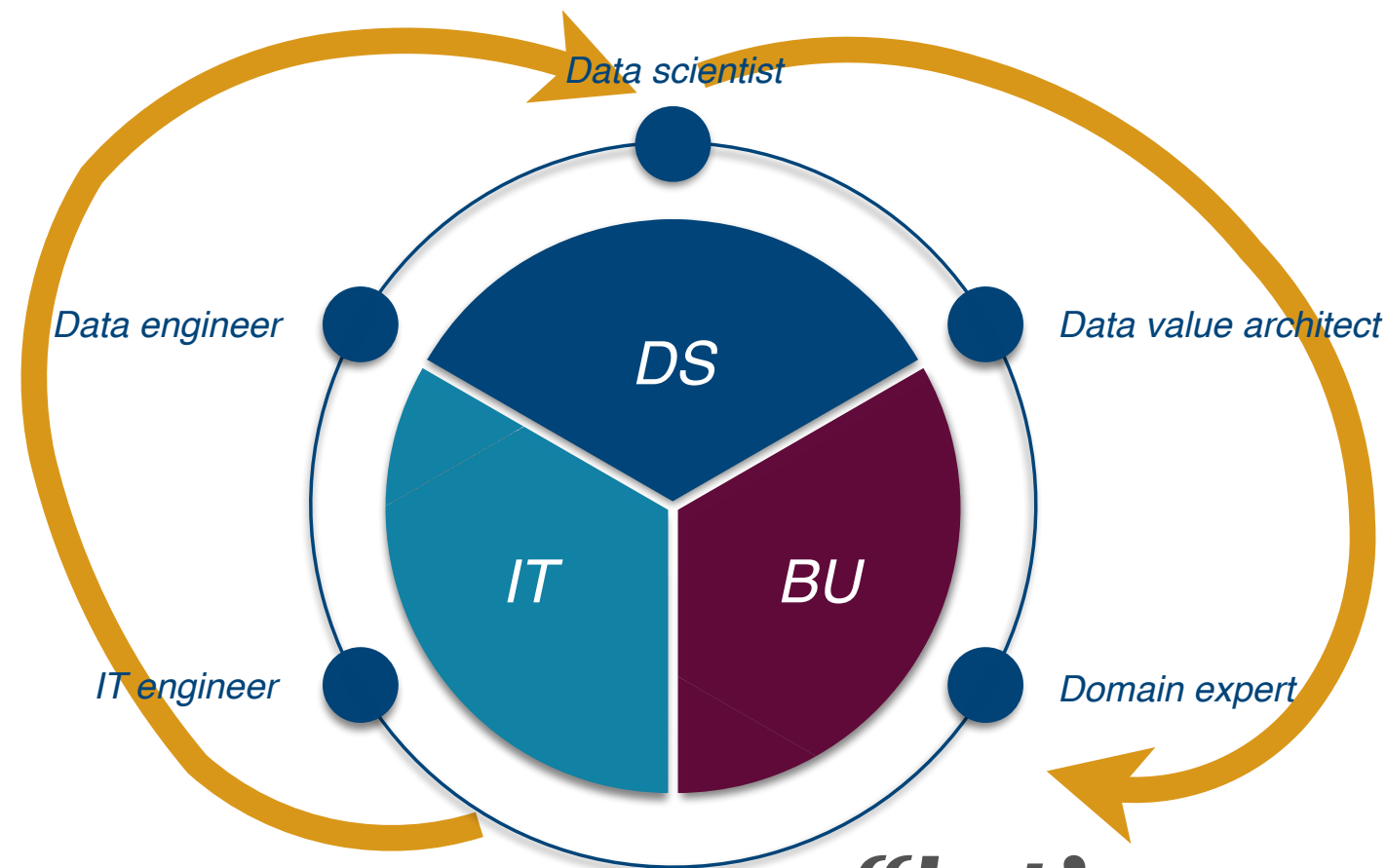
“Let’s hire data scientists”



“Let’s install Hadoop”

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY IT

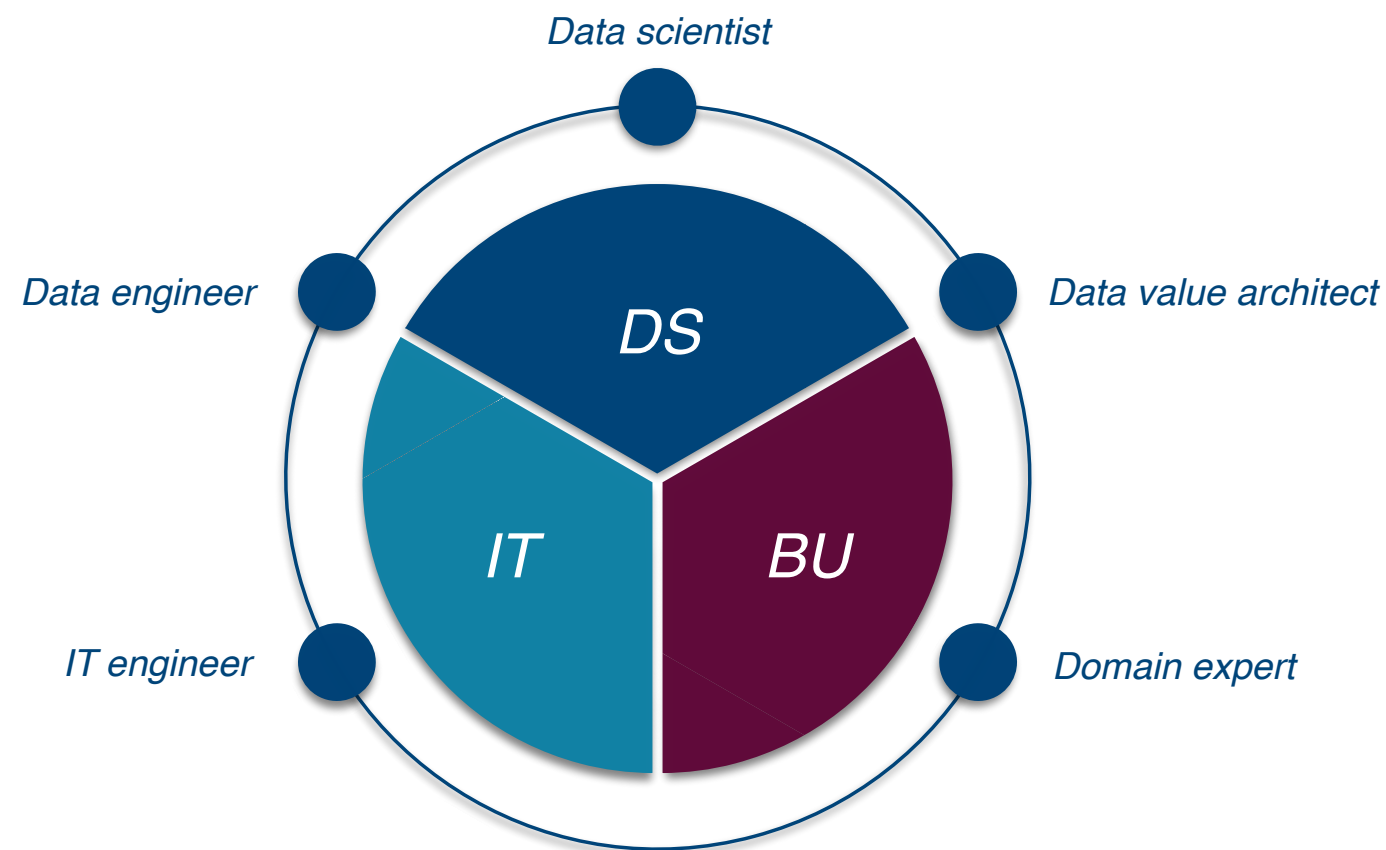
“Let’s hire data scientists”



“Let’s install Hadoop”

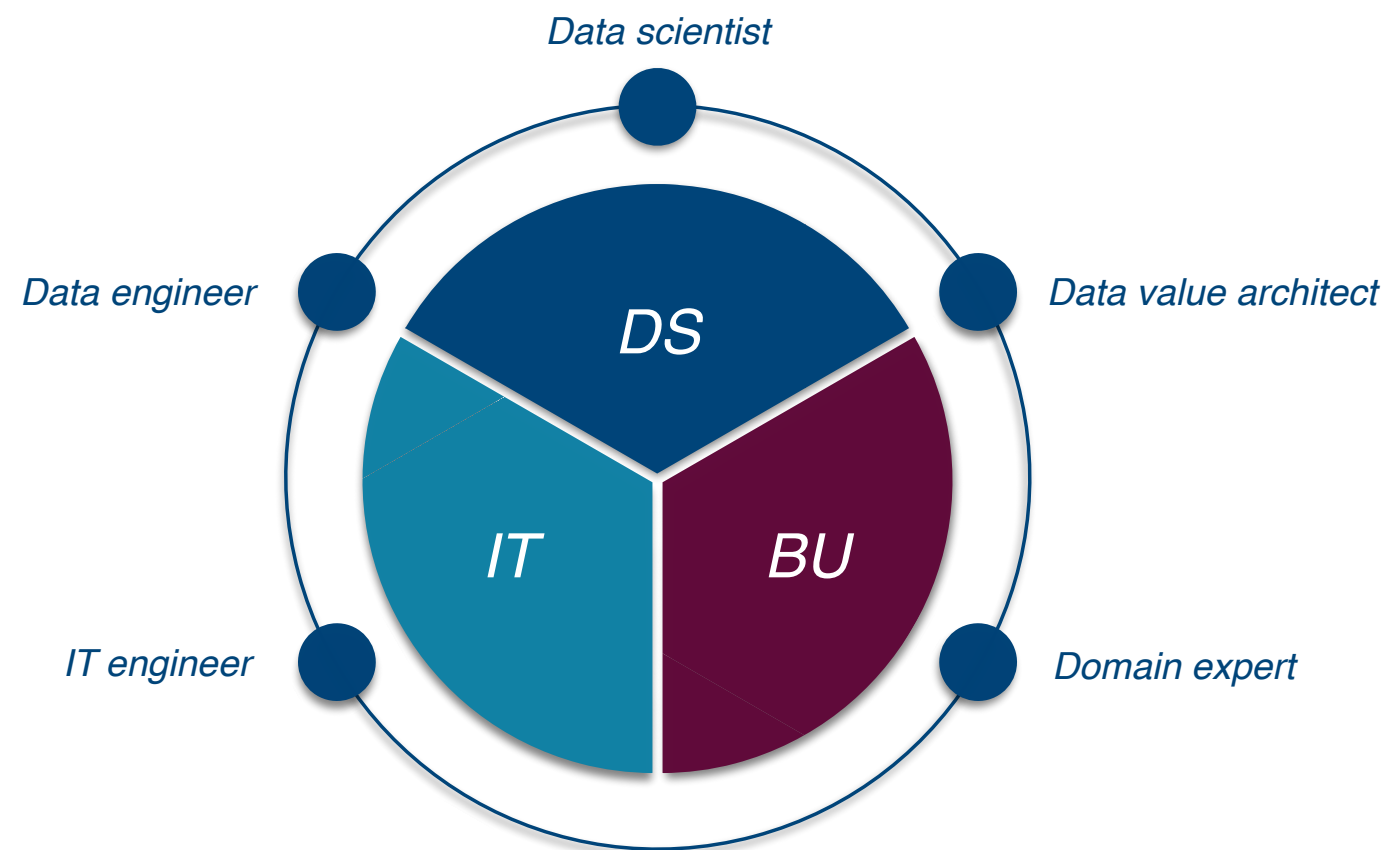
“Let’s see what business problems we can solve with the existing data science team and the infrastructure we bought”

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY DATA SCIENTISTS



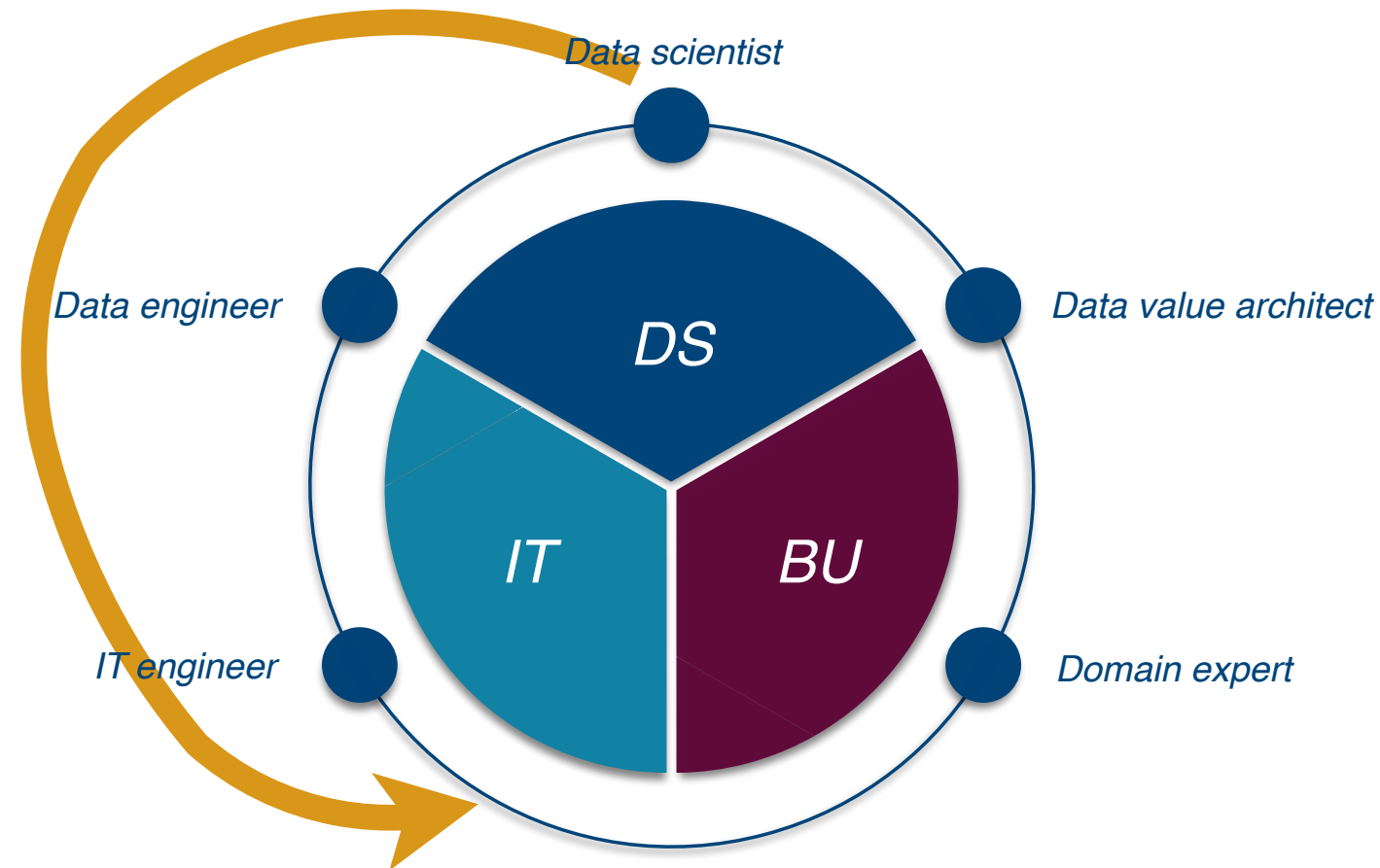
BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY DATA SCIENTISTS

“Let’s hire data scientists.”



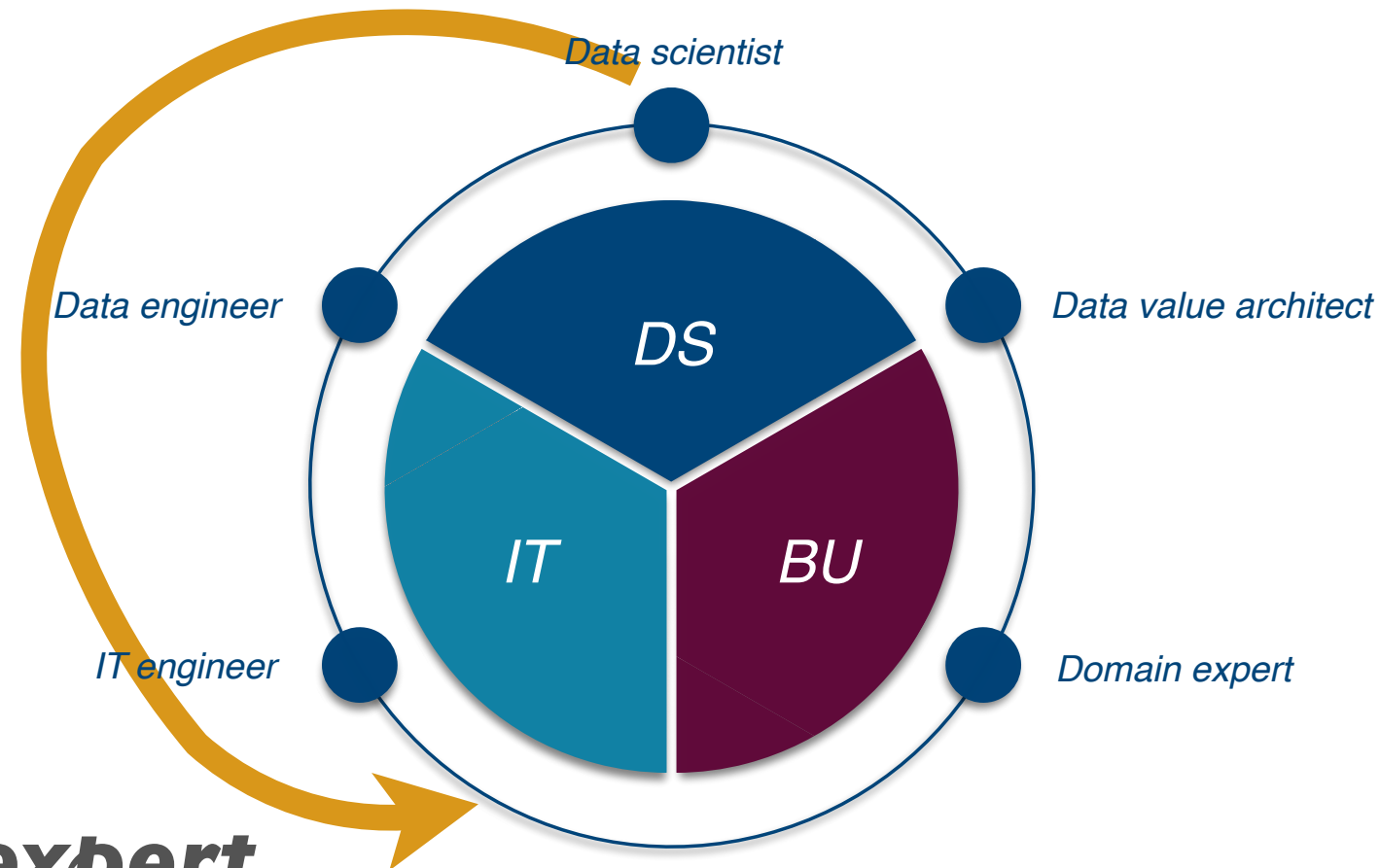
BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY DATA SCIENTISTS

“Let’s hire data scientists.”



BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY DATA SCIENTISTS

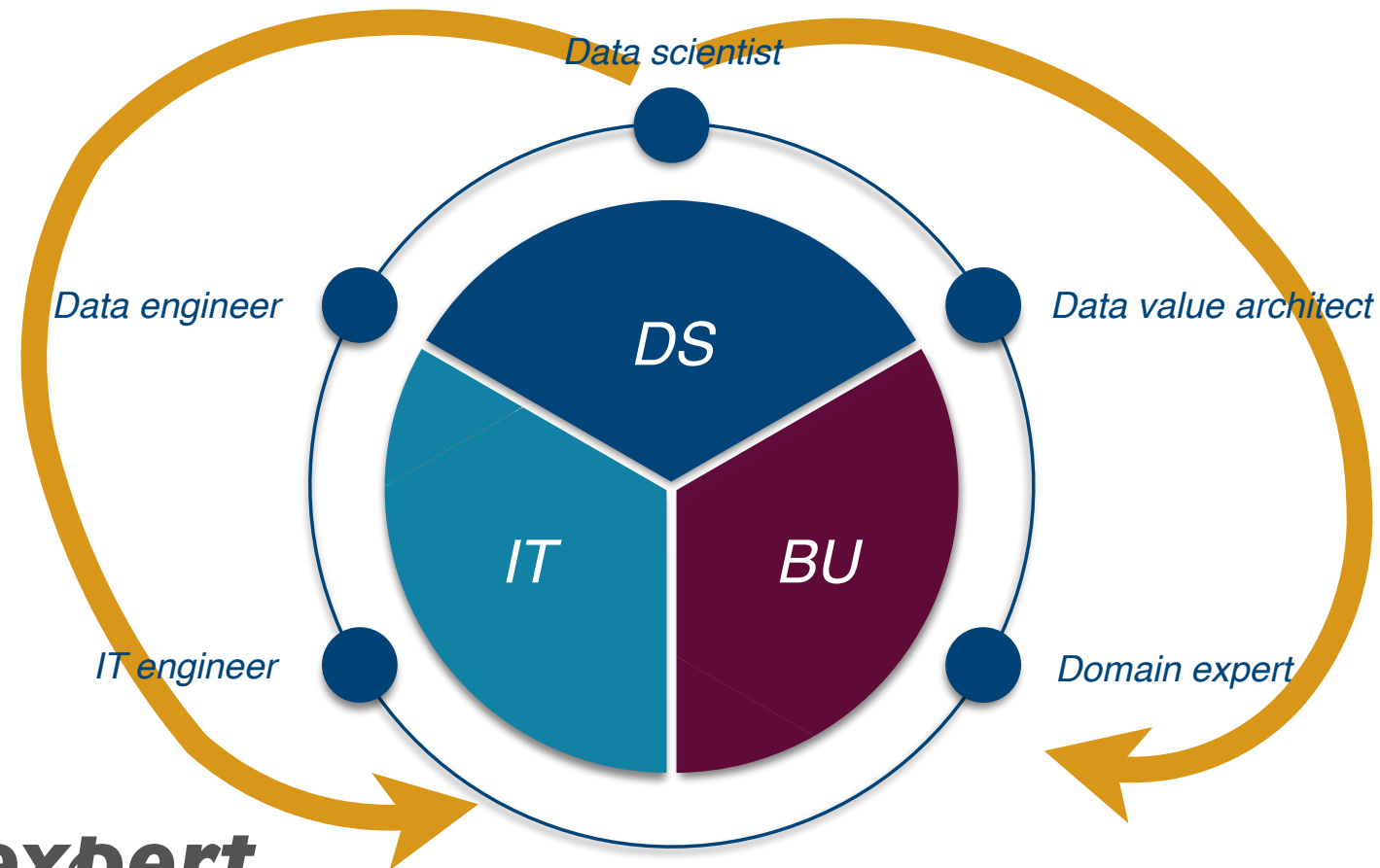
“Let’s hire data scientists.”



***“I’m an expert
of deep learning,
let’s buy a GPU cluster.”***

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY DATA SCIENTISTS

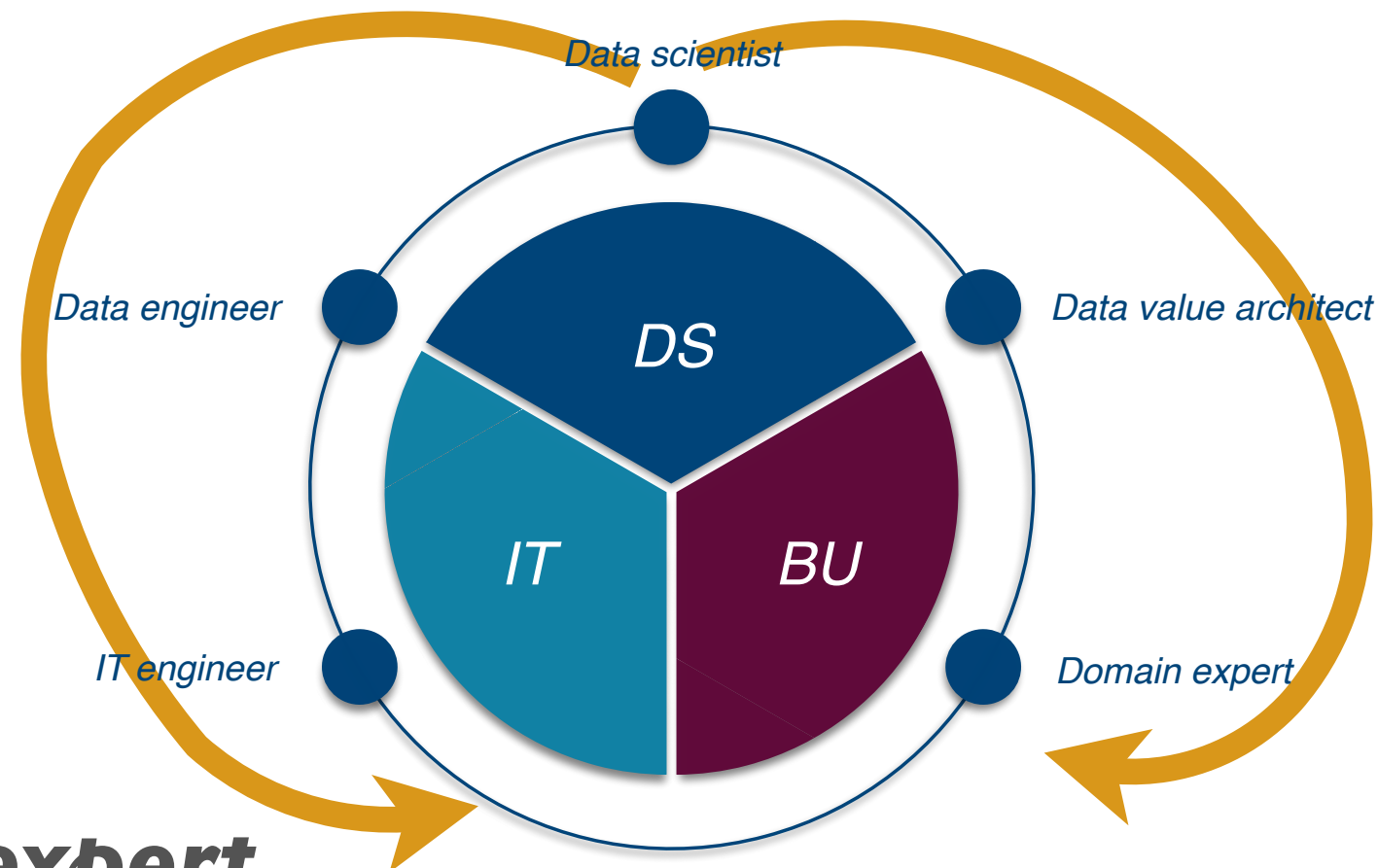
“Let’s hire data scientists.”



***“I’m an expert
of deep learning,
let’s buy a GPU cluster.”***

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY DATA SCIENTISTS

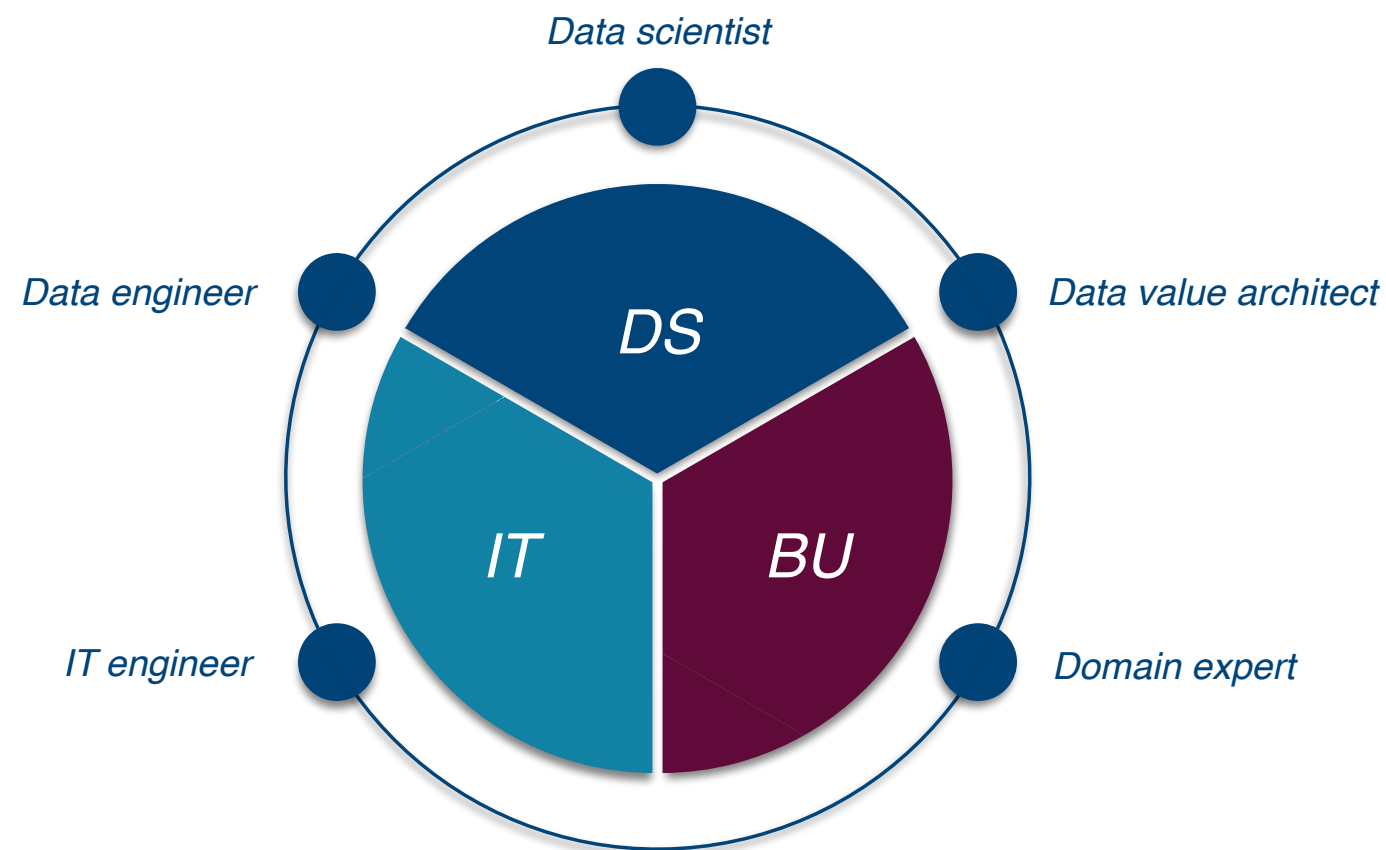
“Let’s hire data scientists.”



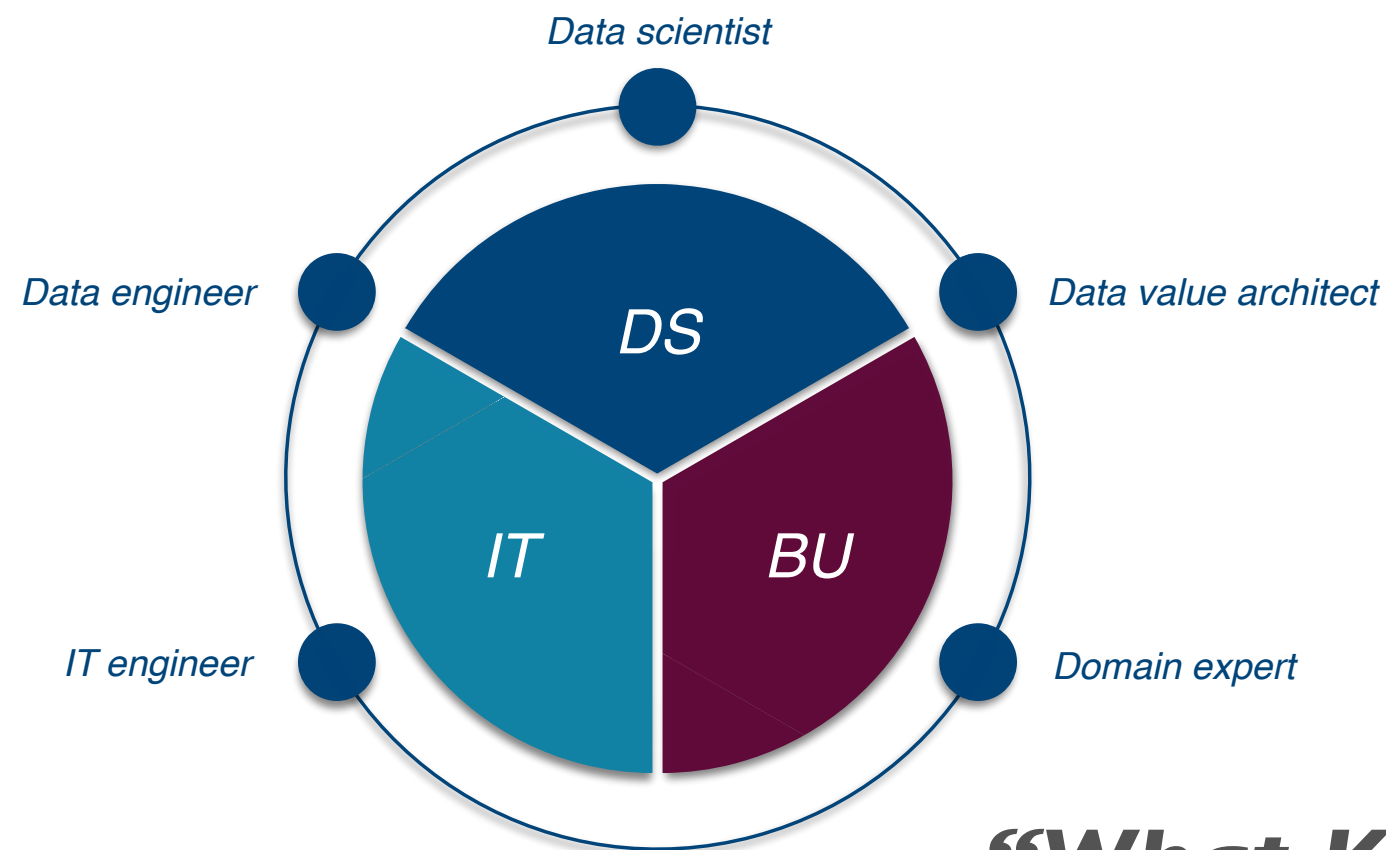
***“I’m an expert
of deep learning,
let’s buy a GPU cluster.”***

***“I’m an expert of deep
learning, let’s see what it
can do for your business.”***

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY BUSINESS

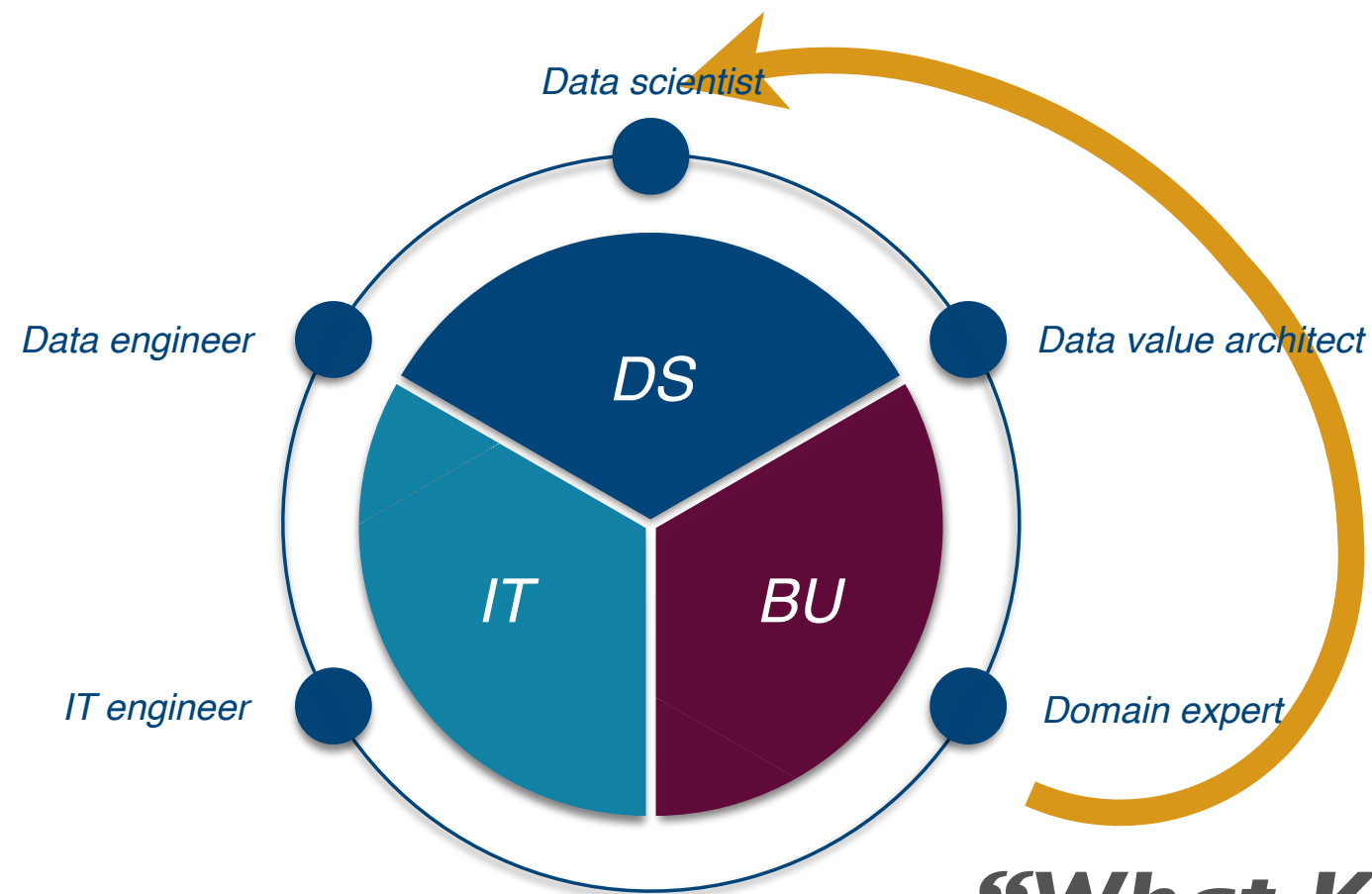


BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY BUSINESS



***“What KPI can we
improve with data?
What data should we
collect?”***

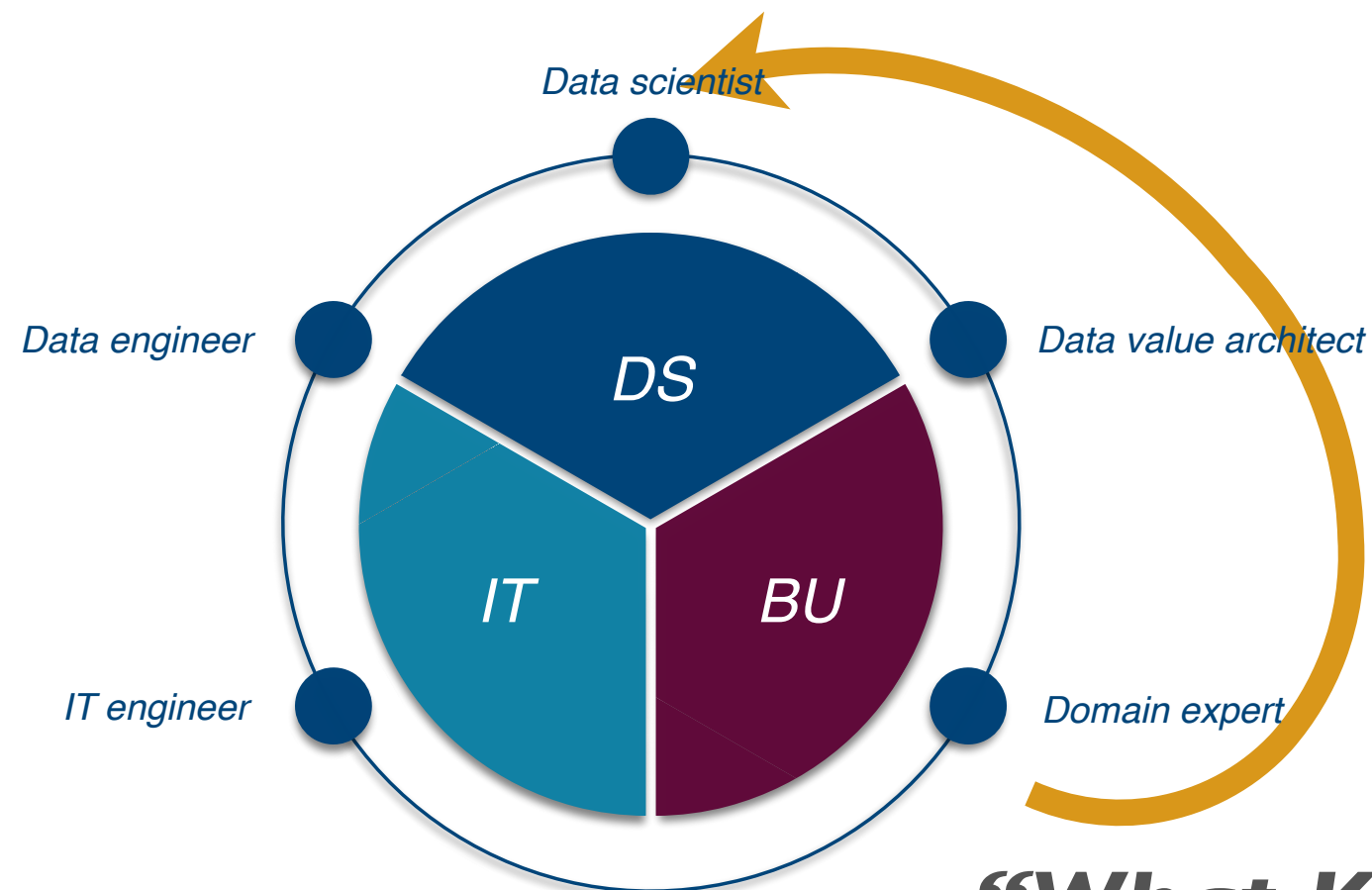
BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY BUSINESS



***“What KPI can we
improve with data?
What data should we
collect?”***

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY BUSINESS

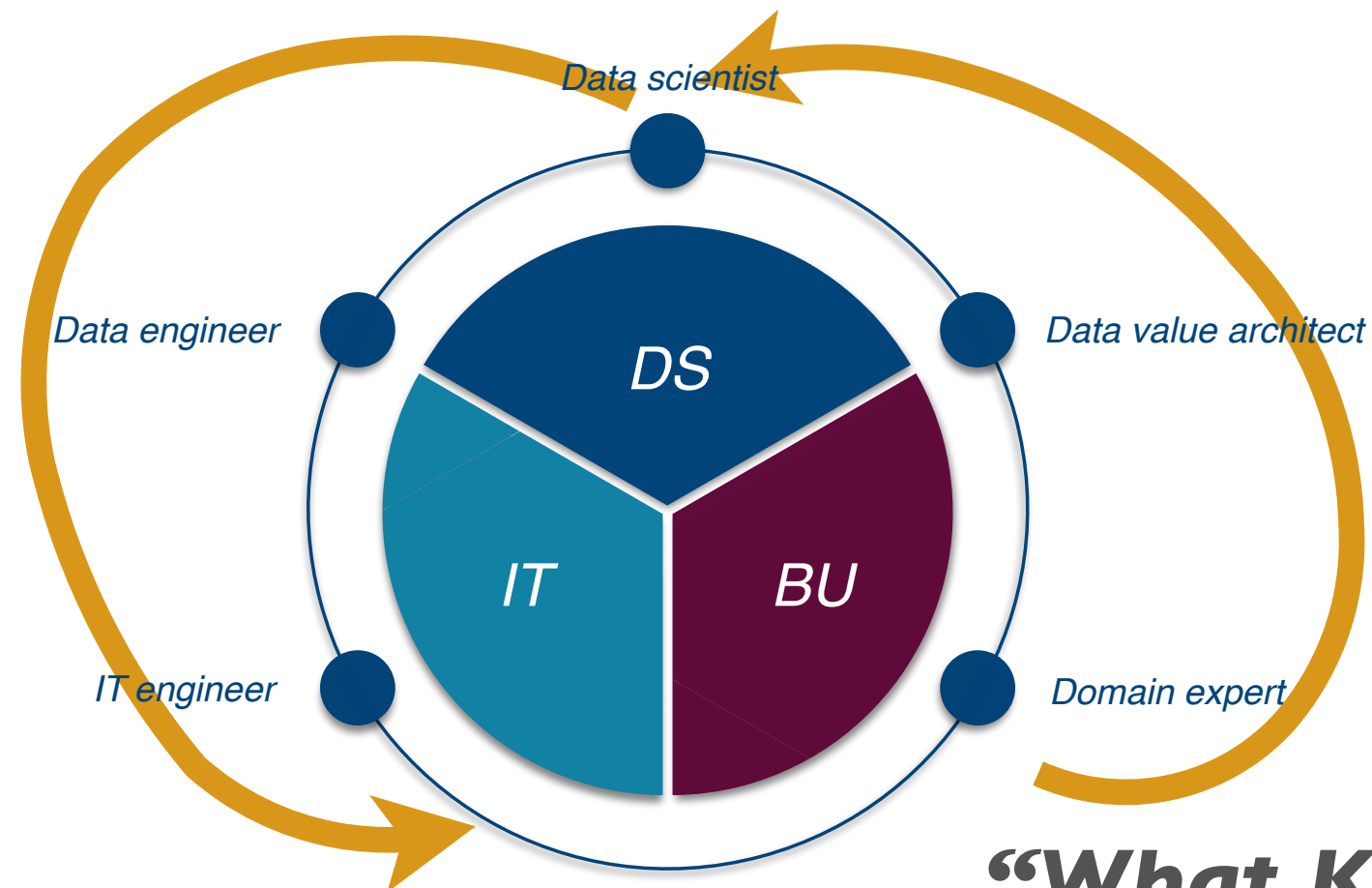
“Let’s hire data scientists for prototyping the business case.”



***“What KPI can we
improve with data?
What data should we
collect?”***

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY BUSINESS

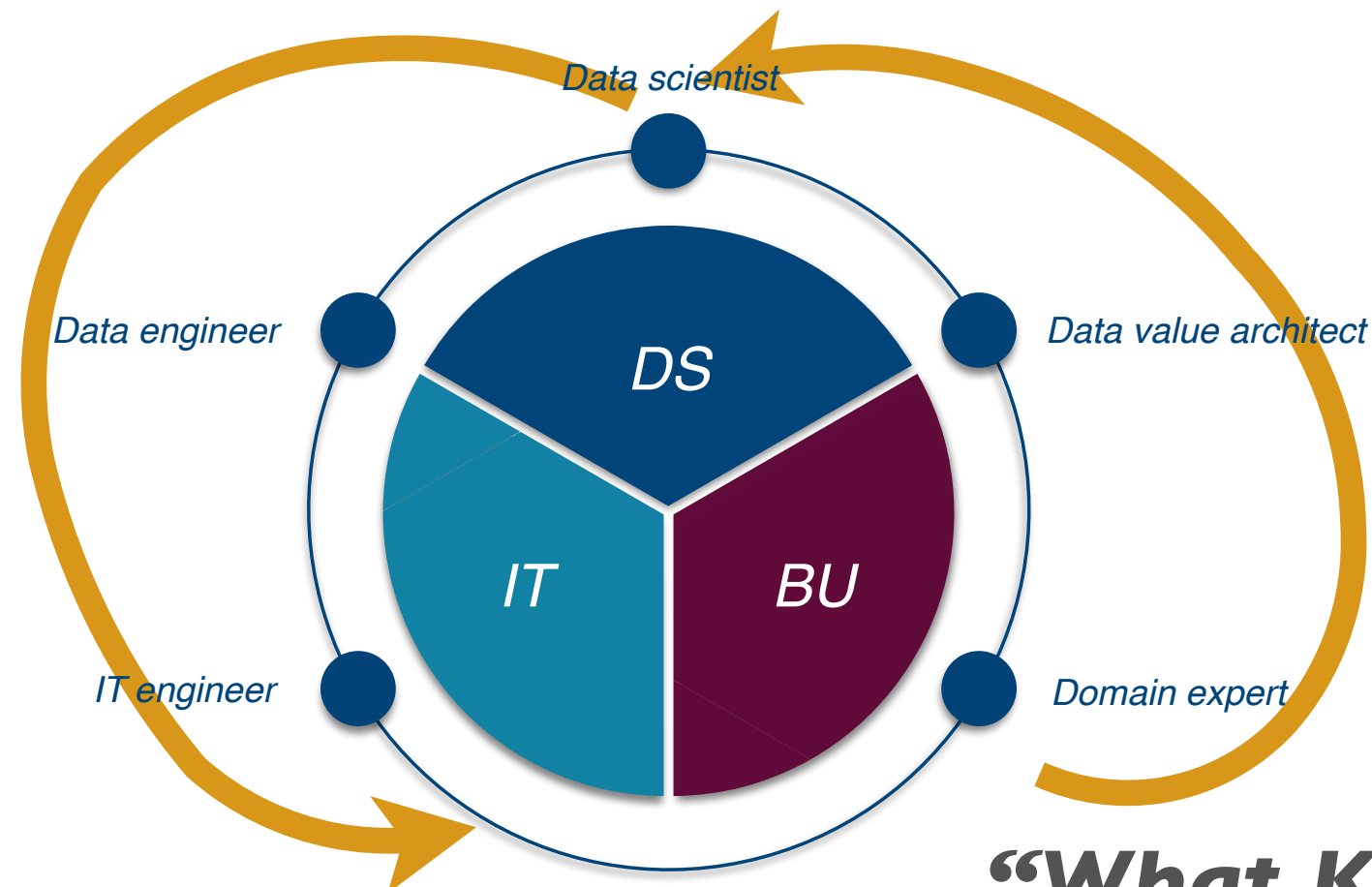
“Let’s hire data scientists for prototyping the business case.”



***“What KPI can we improve with data?
What data should we collect?”***

BUILDING A DATA SCIENCE ECOSYSTEM DRIVEN BY BUSINESS

“Let’s hire data scientists for prototyping the business case.”



“Let’s build a system for putting the prototype into production.”

***“What KPI can we improve with data?
What data should we collect?”***

A case study

A CASE: CAR RENTAL PREDICTIVE MAINTENANCE

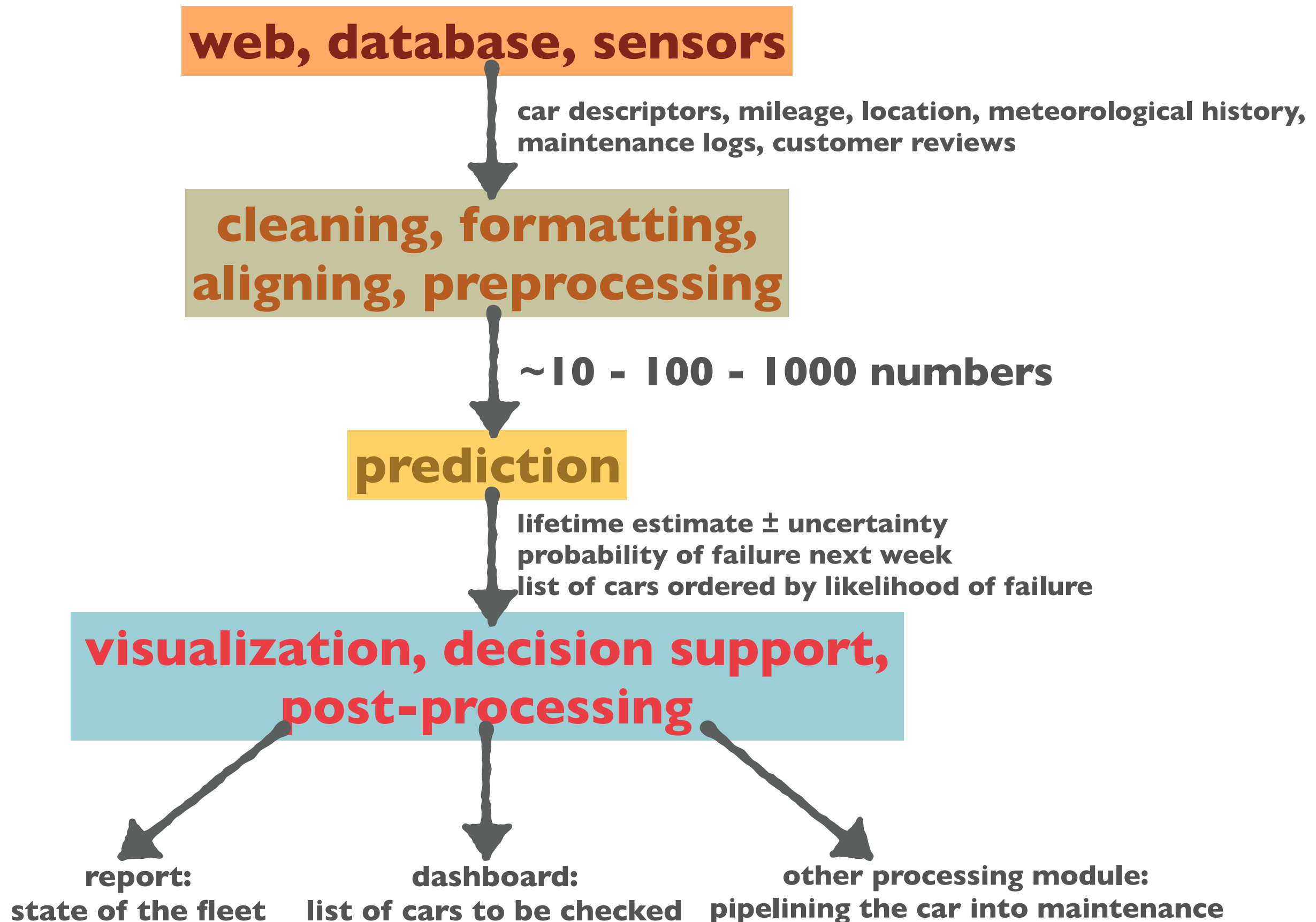
AB-Rent is a multinational car rental company. They have a fleet of 500K vehicles, ranging from sub-economy cars to 12 foot trucks. Annual sales are about 6B\$ per year with a profit of 100M\$. AB-Rent follows an extensive maintenance schedule to avoid the rental of defective cars. There are check-ups (pre-determined by the car manufacturer) scheduled by mileage and age of the car. After every rental, mechanics visually inspect and drive the cars to check for defects. A defective car, if it were rented, would generate both direct costs (replacement, towing, refunds) and indirect costs (reputation, customer churn). Maintenance costs total about 10% of annual sales (~600M\$).

To decrease maintenance and defect costs, AB-Rent decided to launch a Big Data project. The goal is to predict more accurately when a car needs a check-up and repair.

1. **KPI.** What do we want to predict and how do we measure the quality of the prediction? How will a better prediction improve the bottomline?
2. **Business process.** Do you want to have decision support, a fully automated system, or to know just the factors which are important? How will the agent use the system? Why are these important questions to ask?
3. **Data science metrics.** What should be the quantitative prediction? How do we measure success?
4. **Data hunt.** What data do we need to develop a predictor?

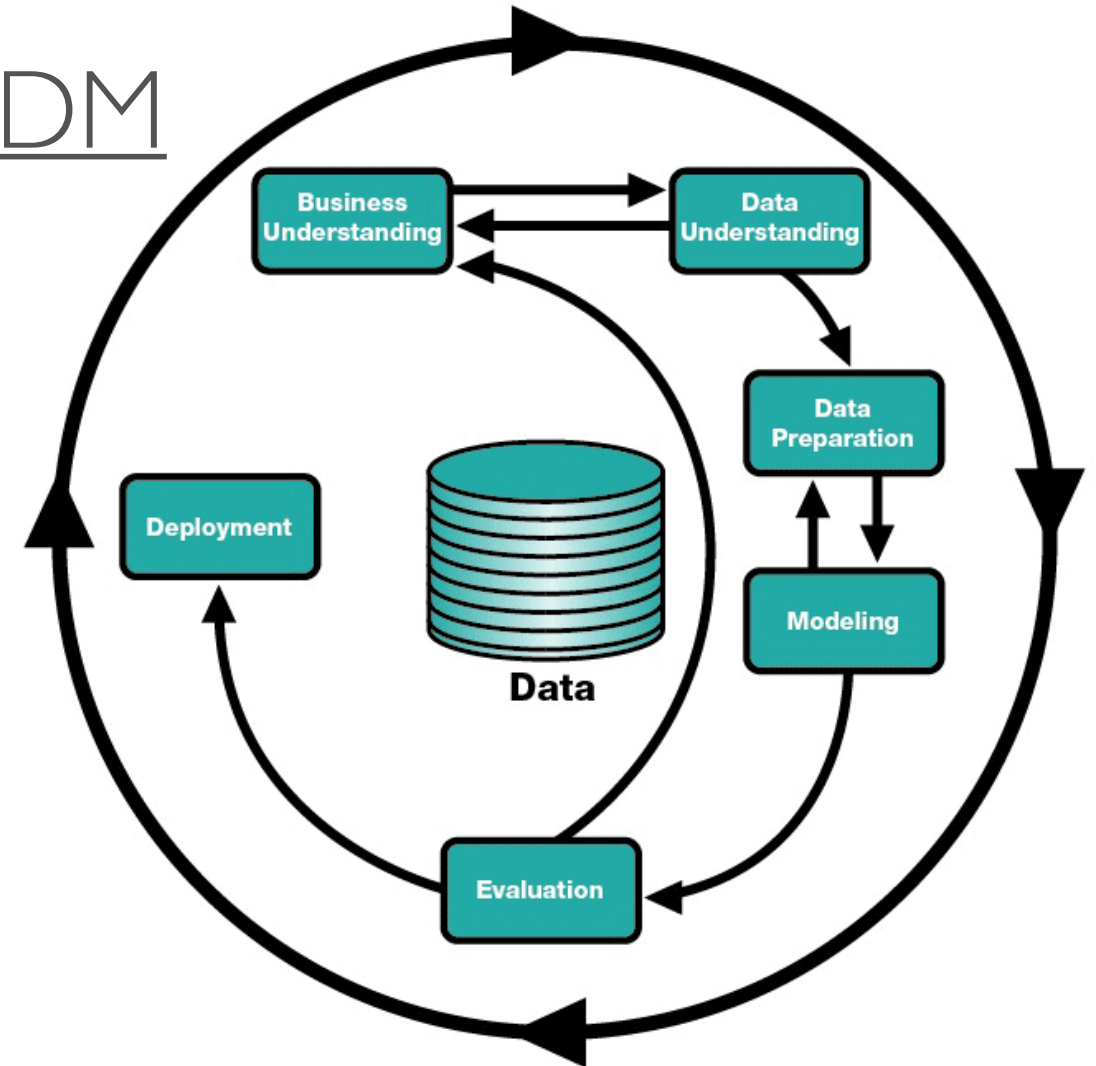
The data analytics production pipeline and development loop

THE DATA ANALYTICS PRODUCTION PIPELINE



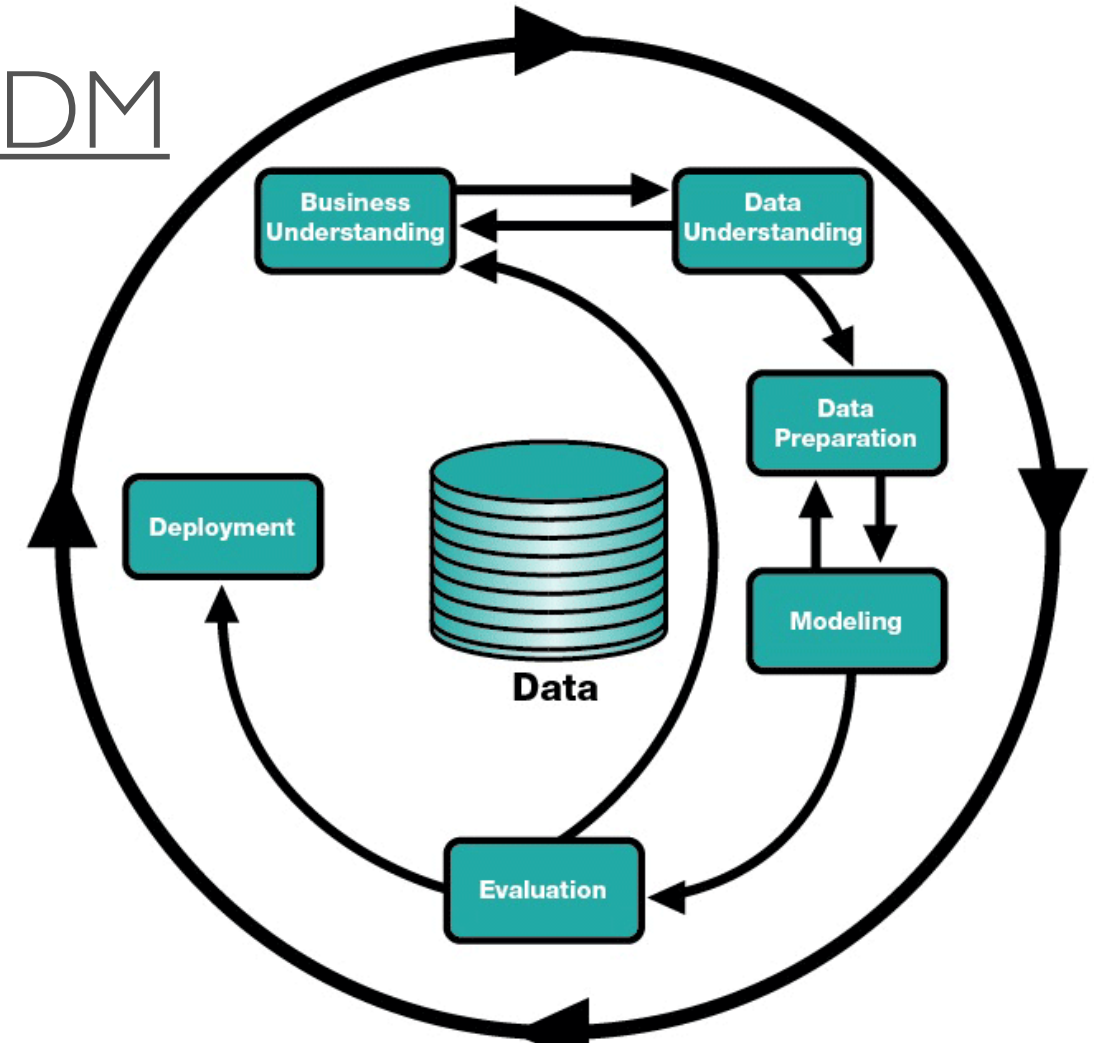
THE DATA ANALYTICS BUILDING PIPELINE

IBM CRISP-DM



THE DATA ANALYTICS BUILDING PIPELINE

IBM CRISP-DM



Dataiku

Define Your Final Goal

I want to predict which users are most likely to churn

Collect Your Data

Gather your nicest logs, your transaction and CRM data

Explore Your Data

See what you got, try to detect patterns and anomalies, etc.

Clean Your Data

Group your information by customer, join datasets, etc. to get a clear view of each customer's activity

Visualize Your Data

Build Your Model

Create models to predict which users are most likely to churn

Put Your Model In Production

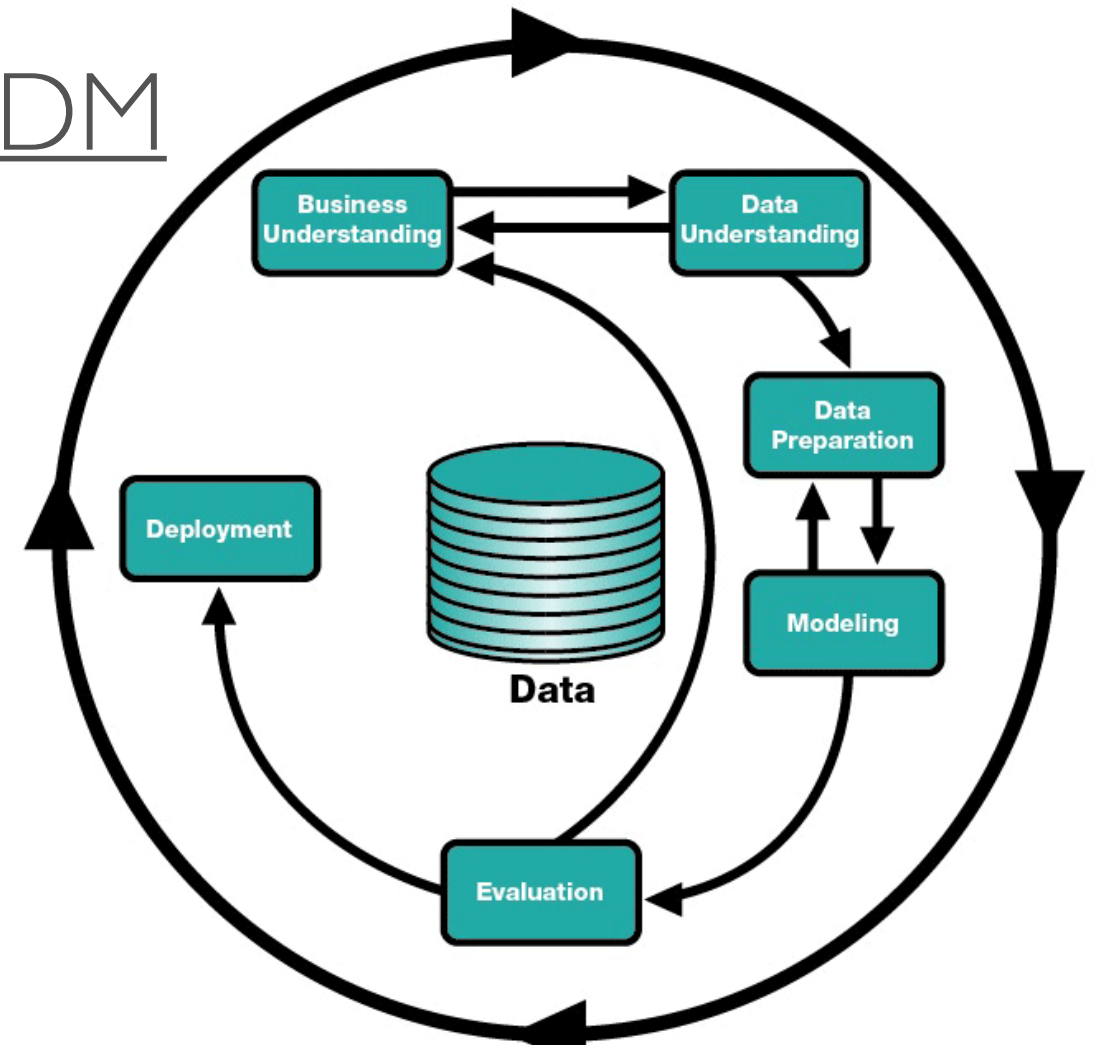
Use your model to send targeted emails to potential future churners

Visualize Your Data

THE DATA ANALYTICS BUILDING PIPELINE

IBM CRISP-DM

1996



2016

Dataiku

Define Your Final Goal

I want to predict which users are most likely to churn

Collect Your Data

Gather your nicest logs, your transaction and CRM data

Explore Your Data

See what you got, try to detect patterns and anomalies, etc.

Clean Your Data

Group your information by customer, join datasets, etc. to get a clear view of each customer's activity

Visualize Your Data

Build Your Model

Create models to predict which users are most likely to churn

Put Your Model In Production

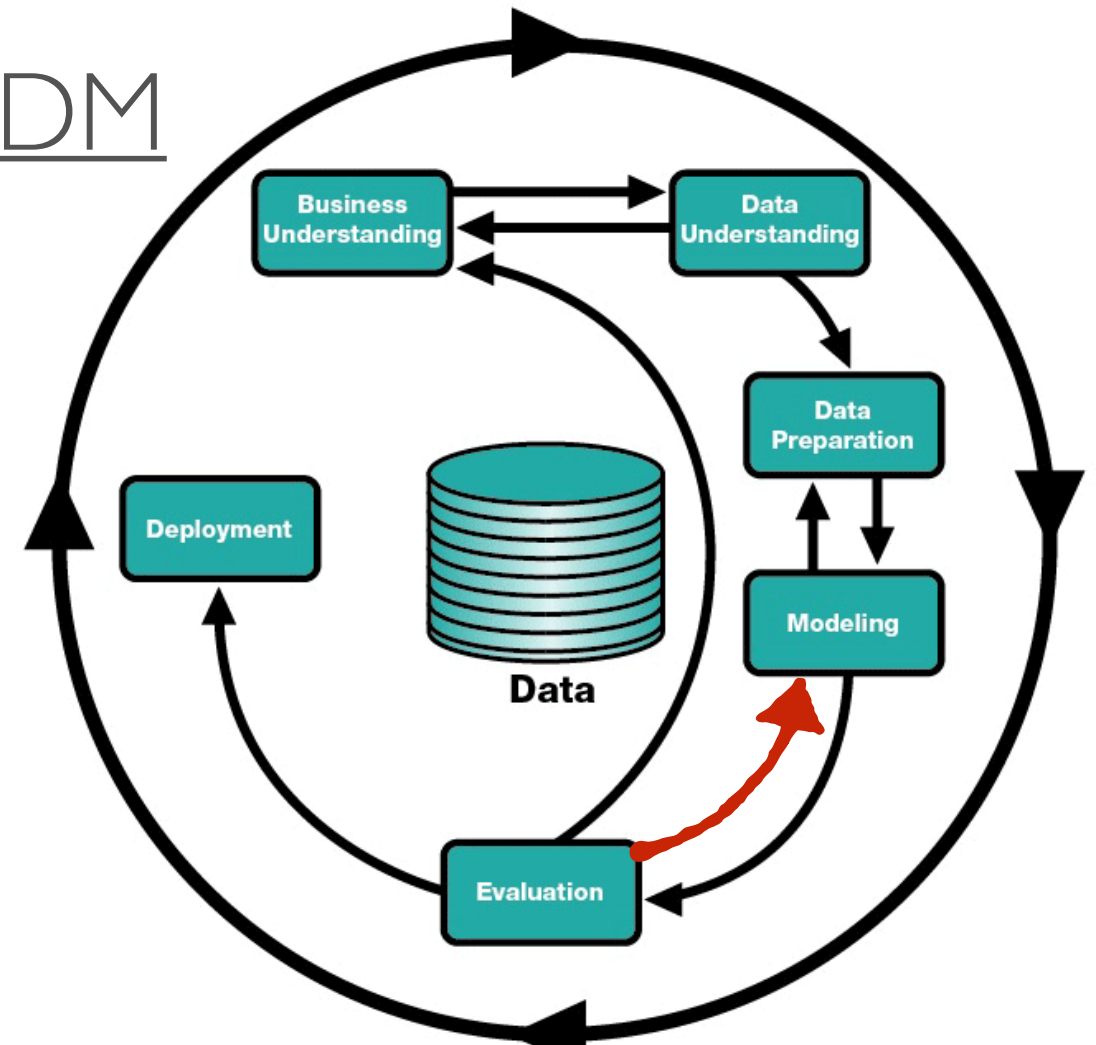
Use your model to send targeted emails to potential future churners

Visualize Your Data

THE DATA ANALYTICS BUILDING PIPELINE

IBM CRISP-DM

1996



2016

Dataiku

Define Your Final Goal

I want to predict which users are most likely to churn

Collect Your Data

Gather your nicest logs, your transaction and CRM data

Explore Your Data

See what you got, try to detect patterns and anomalies, etc.

Clean Your Data

Group your information by customer, join datasets, etc. to get a clear view of each customer's activity

Visualize Your Data

Build Your Model

Create models to predict which users are most likely to churn

Put Your Model In Production

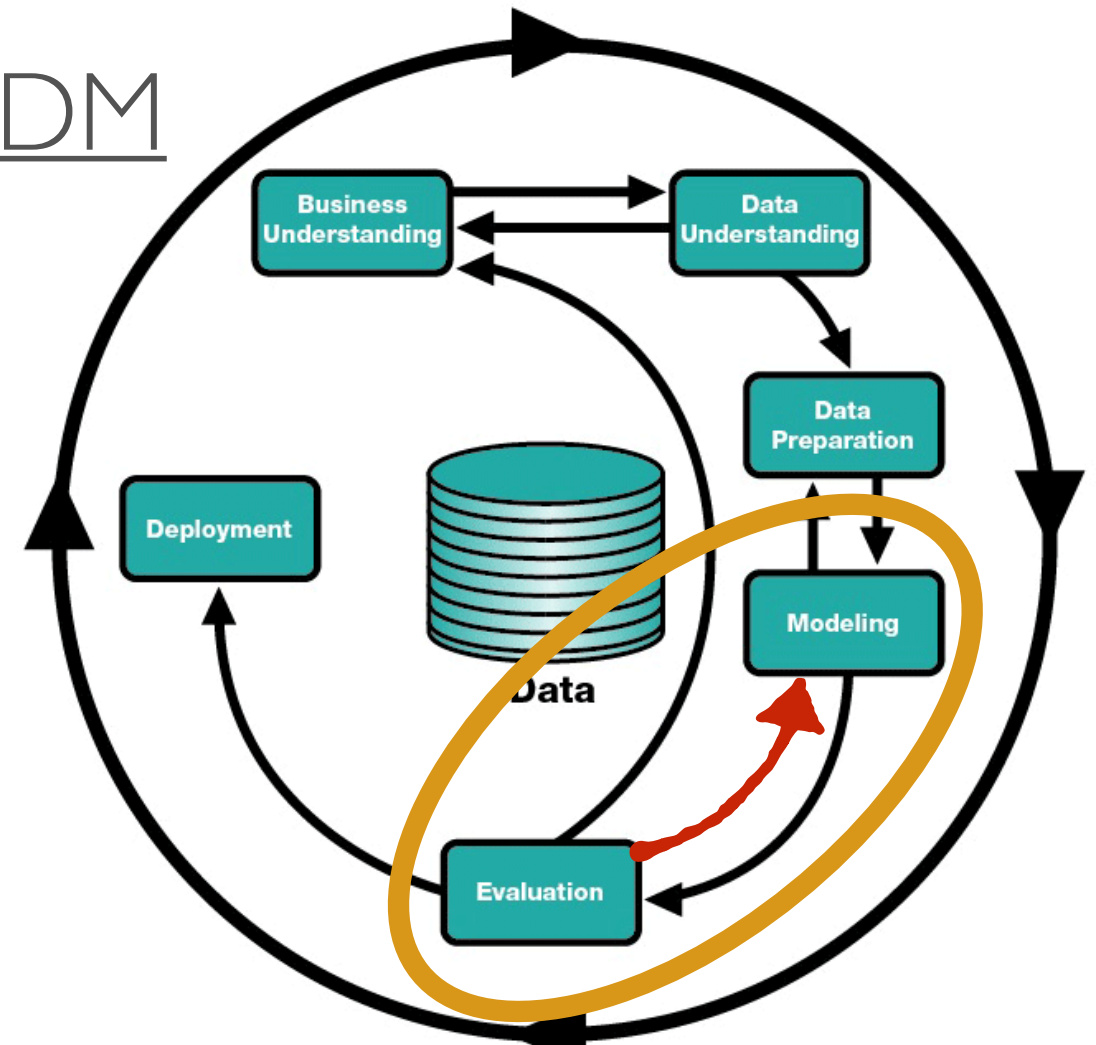
Use your model to send targeted emails to potential future churners

Visualize Your Data

THE DATA ANALYTICS BUILDING PIPELINE

IBM CRISP-DM

1996



2016

Dataiku

Define Your Final Goal

I want to predict which users are most likely to churn

Collect Your Data

Gather your nicest logs, your transaction and CRM data

Explore Your Data

See what you got, try to detect patterns and anomalies, etc.

Clean Your Data

Group your information by customer, join datasets, etc. to get a clear view of each customer's activity

Visualize Your Data

Build Your Model

Create models to predict which users are most likely to churn

Put Your Model In Production

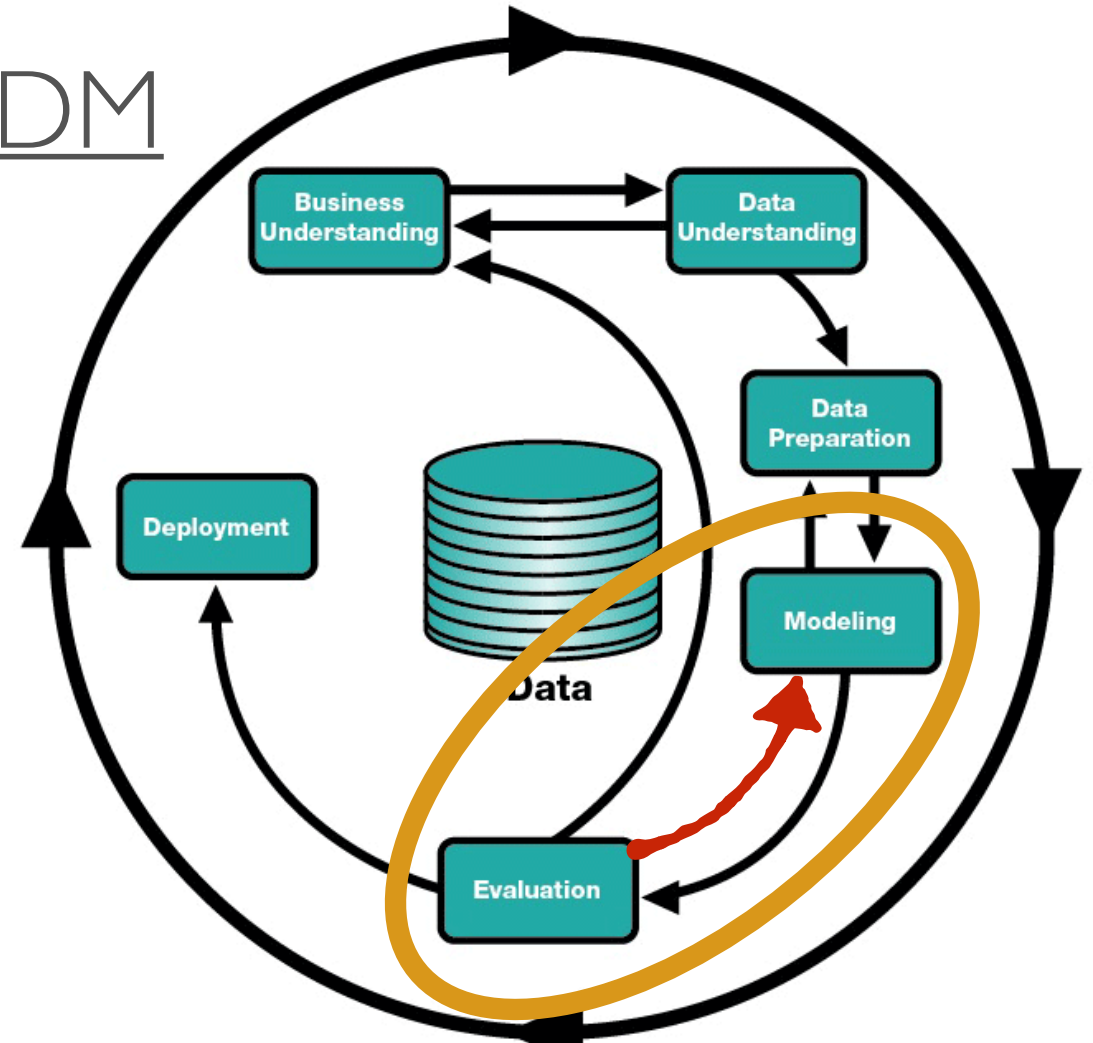
Use your model to send targeted emails to potential future churners

Visualize Your Data

THE DATA ANALYTICS BUILDING PIPELINE

IBM CRISP-DM

1996



2016

Dataiku

Define Your Final Goal

I want to predict which users are most likely to churn

Collect Your Data

Gather your nicest logs, your transaction and CRM data

Explore Your Data

See what you got, try to detect patterns and anomalies, etc.

Clean Your Data

Group your information by customer, join datasets, etc. to get a clear view of each customer's activity

Visualize Your Data

Build Your Model

Create models to predict which users are most likely to churn

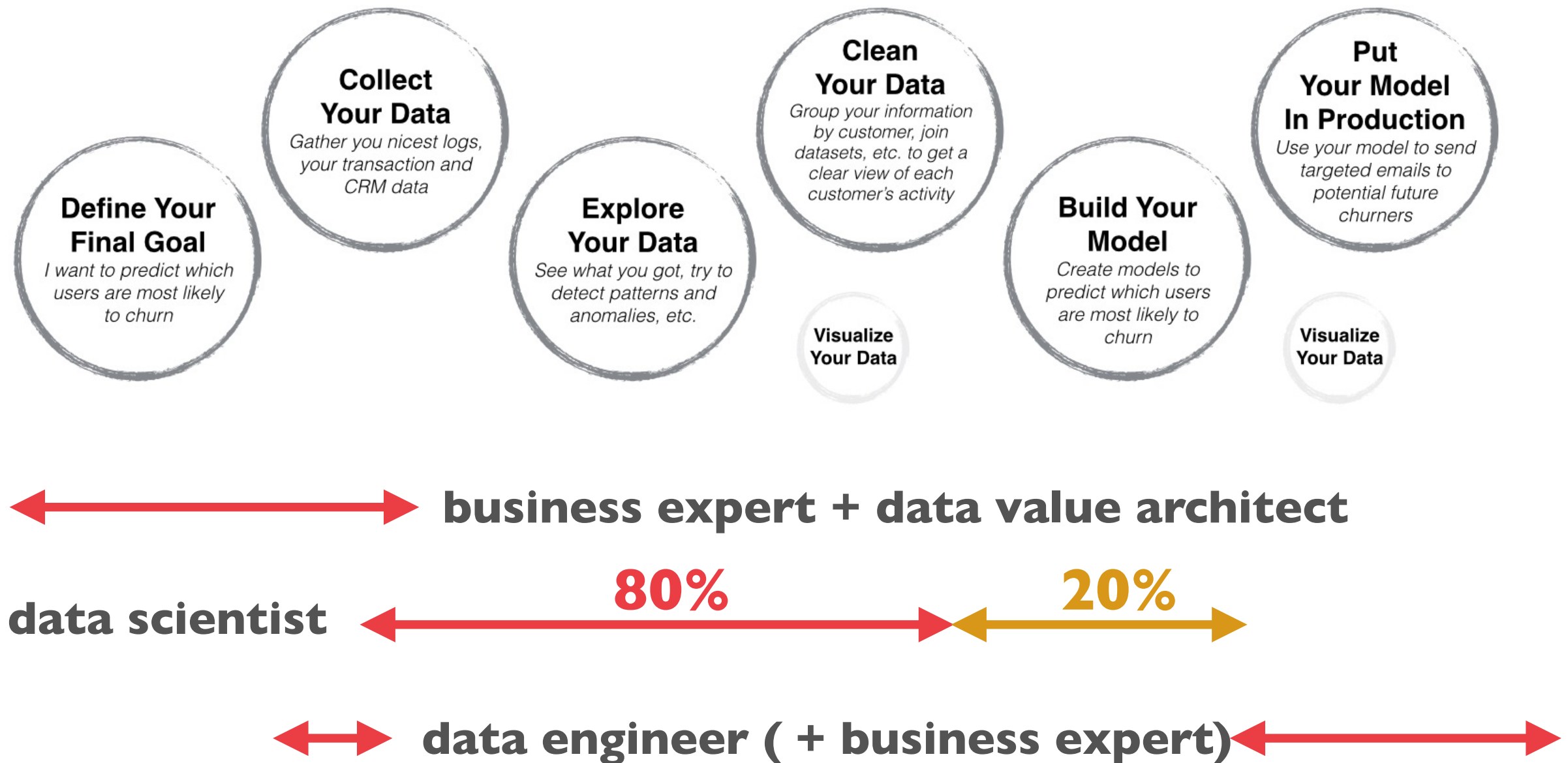
Put Your Model In Production

Use your model to send targeted emails to potential future churners

Visualize Your Data

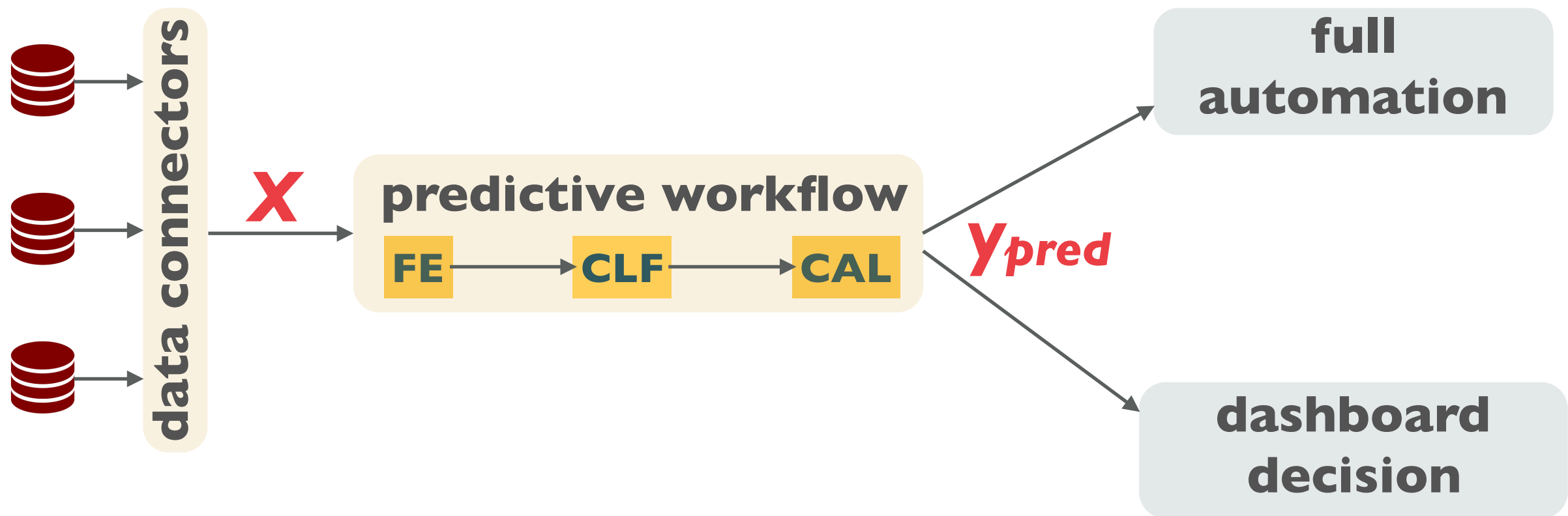
THE DATA ANALYTICS BUILDING PIPELINE

WHO DOES WHAT AND WHEN



THE DATA FLOW

data flow →



THE IDEAL SEQUENCE

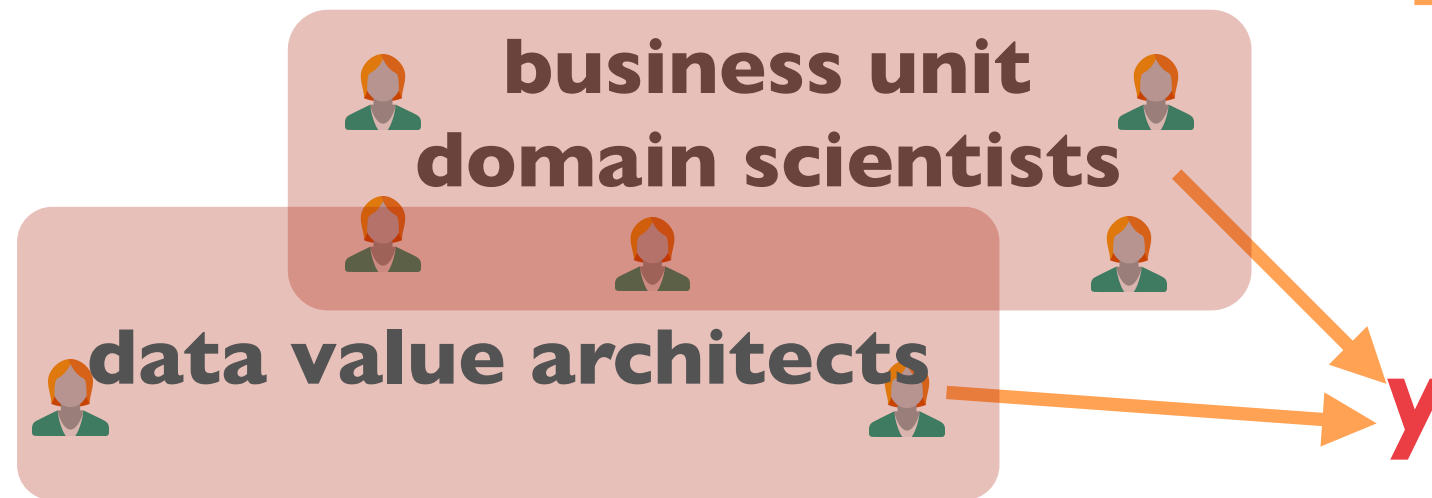


THE IDEAL SEQUENCE

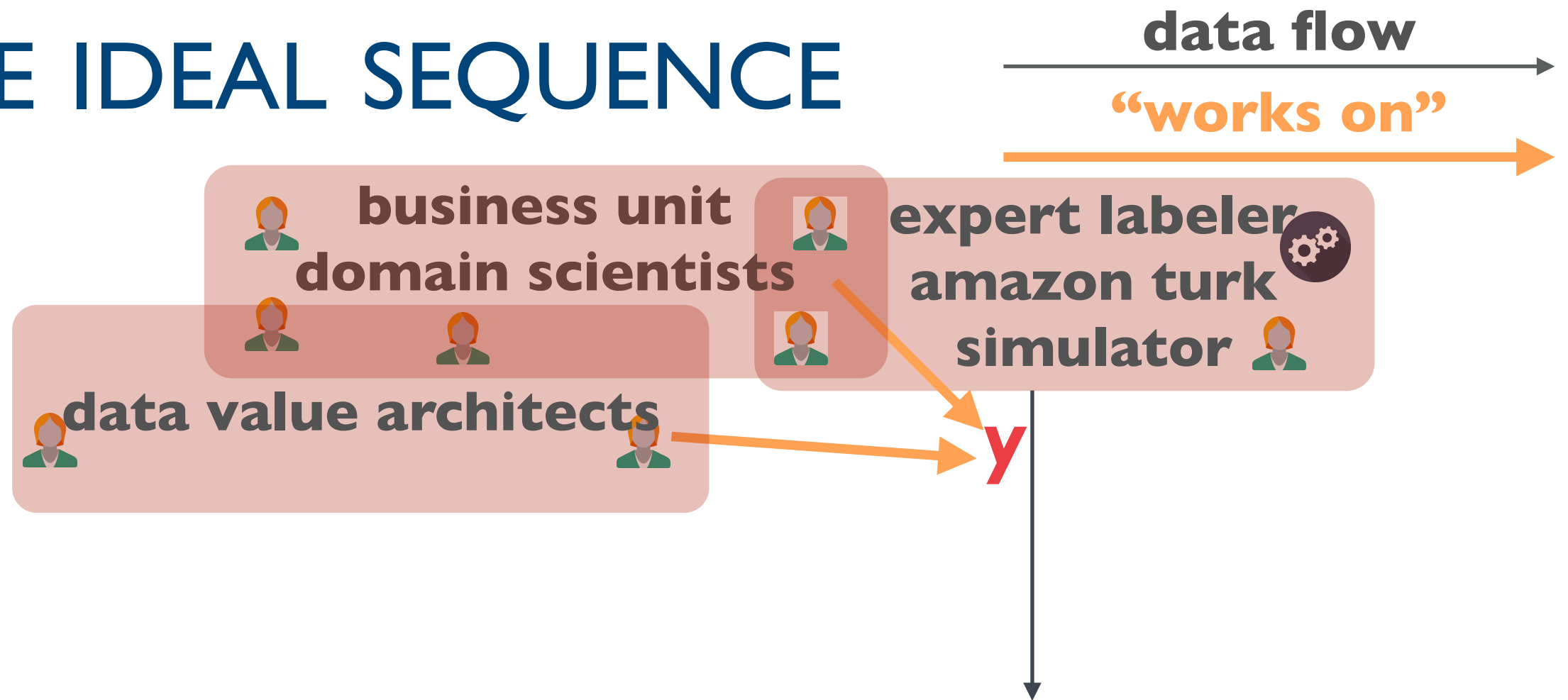


y

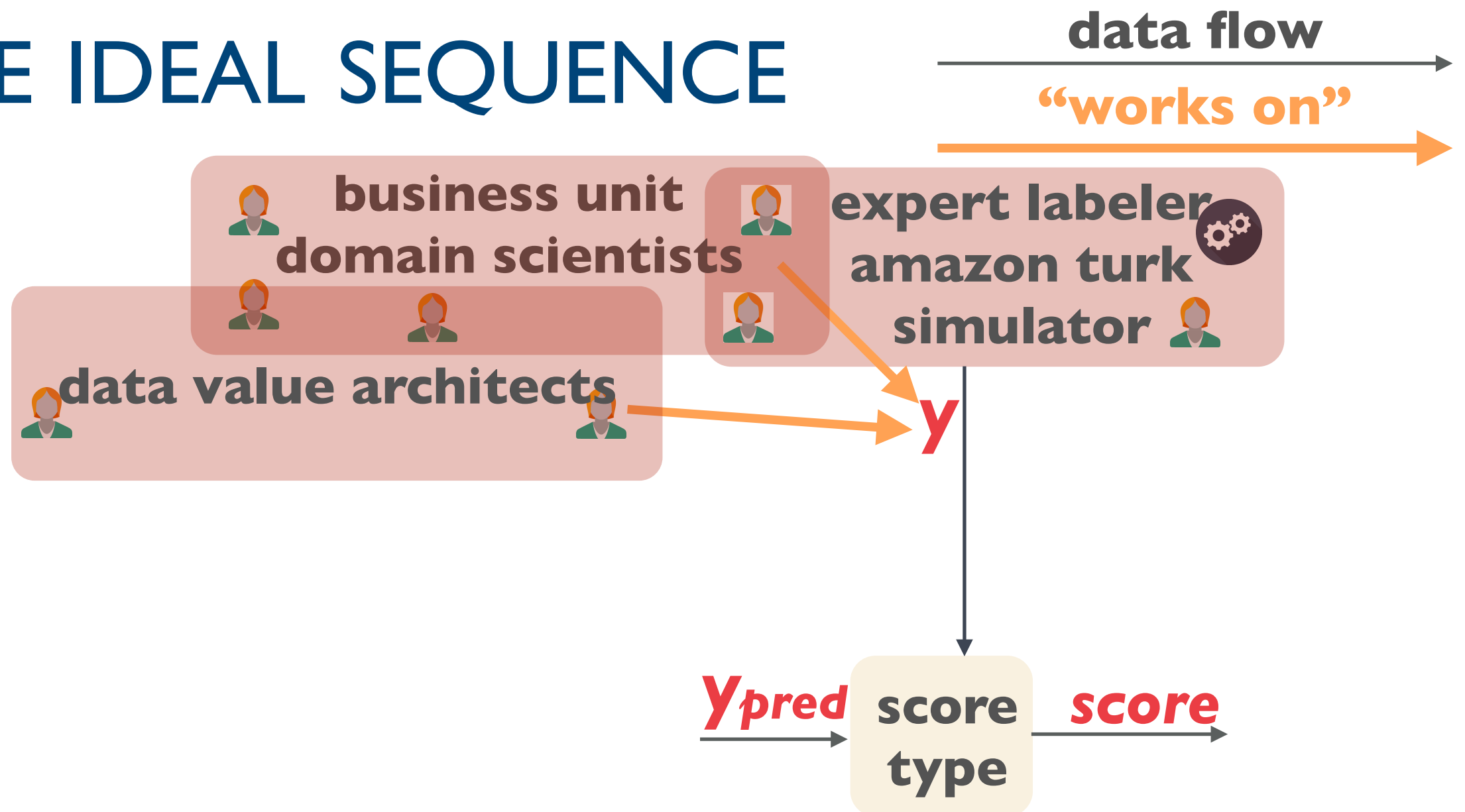
THE IDEAL SEQUENCE



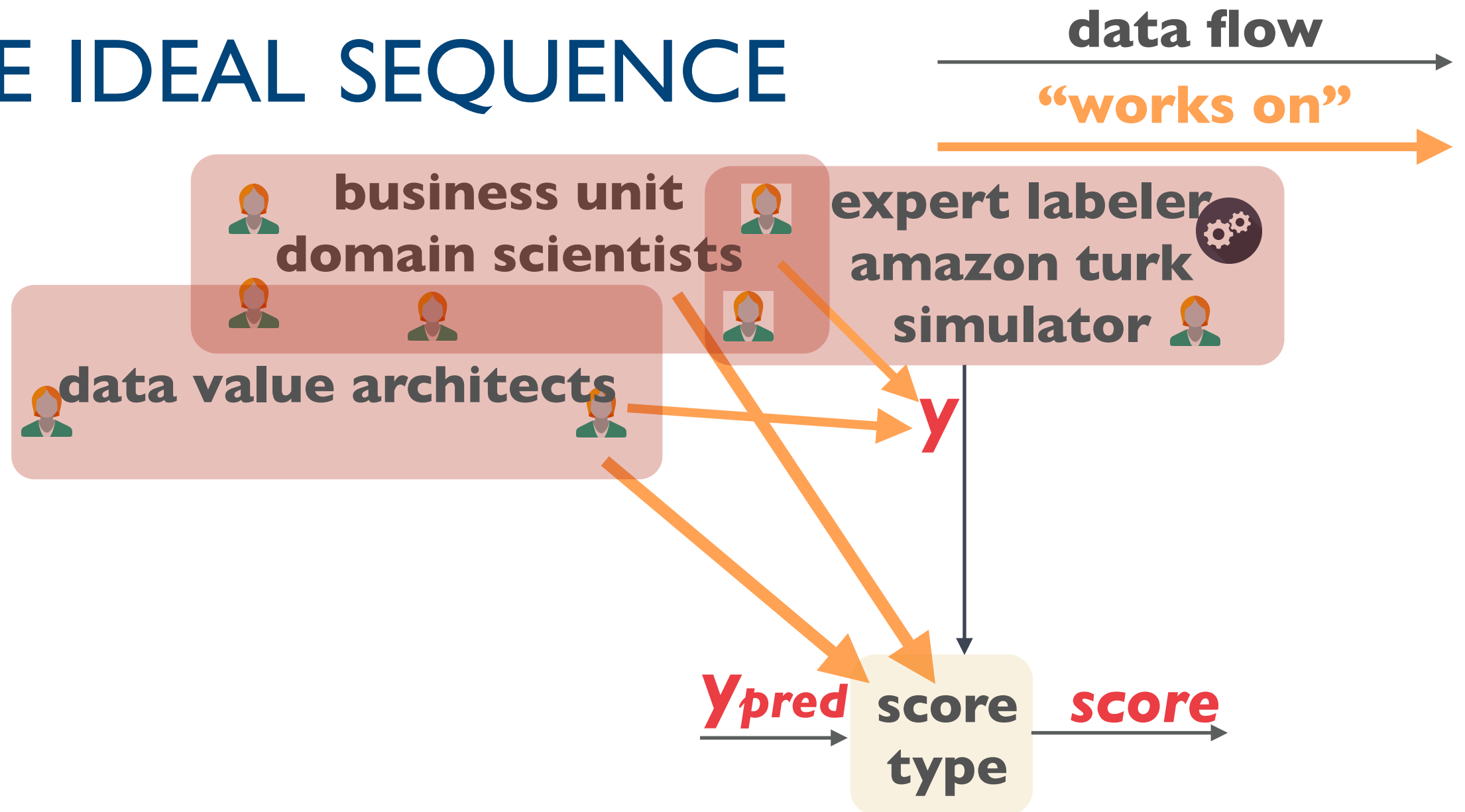
THE IDEAL SEQUENCE



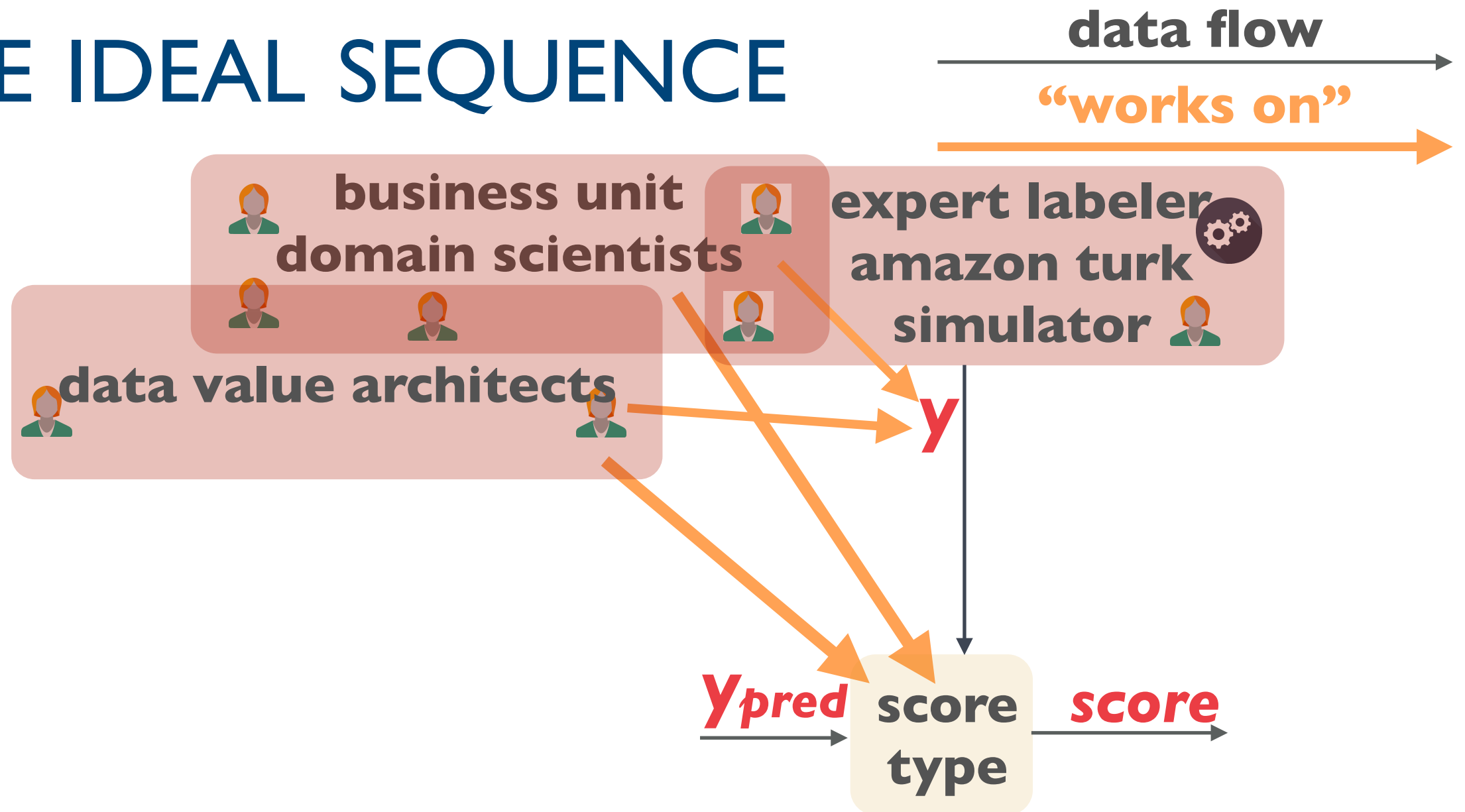
THE IDEAL SEQUENCE



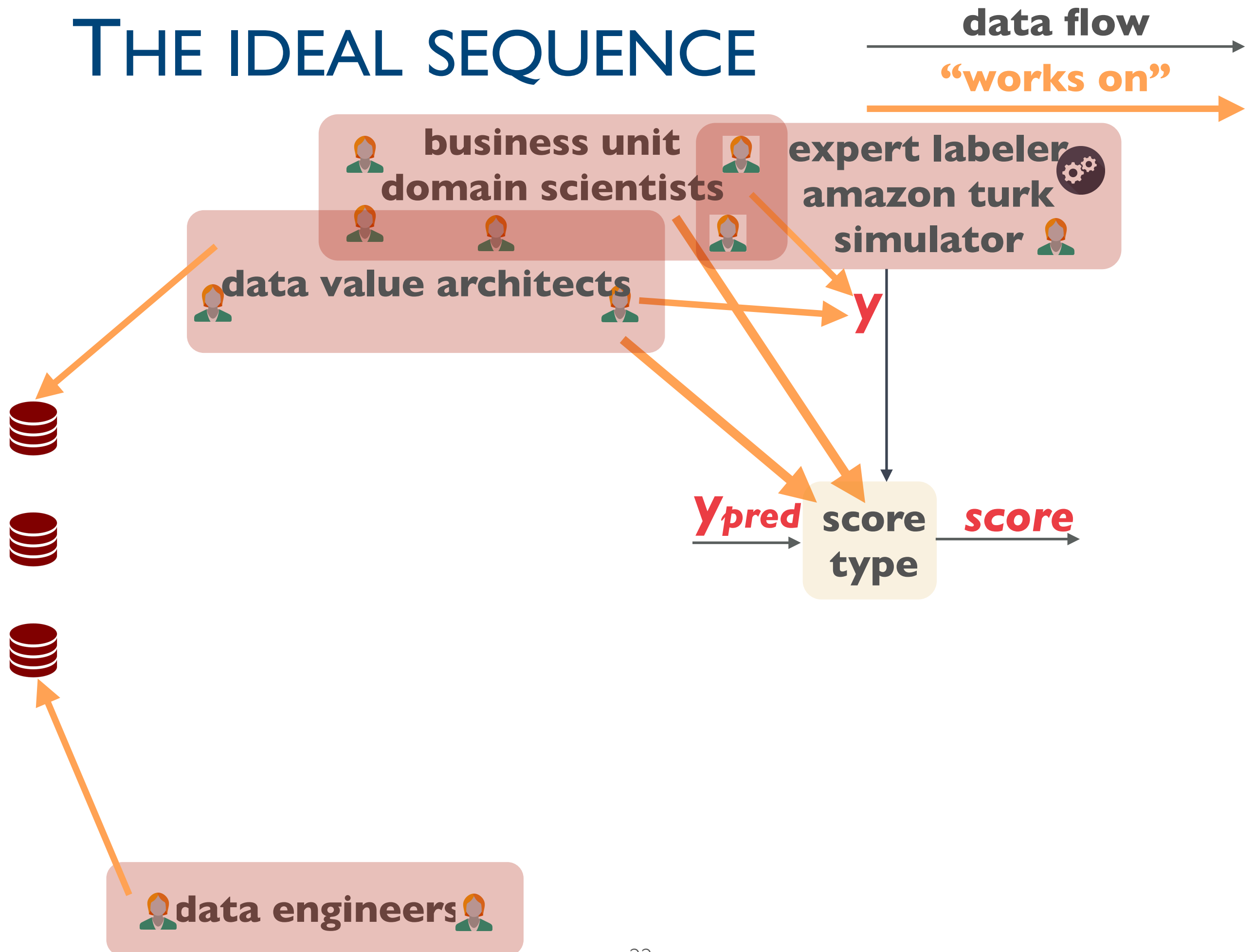
THE IDEAL SEQUENCE



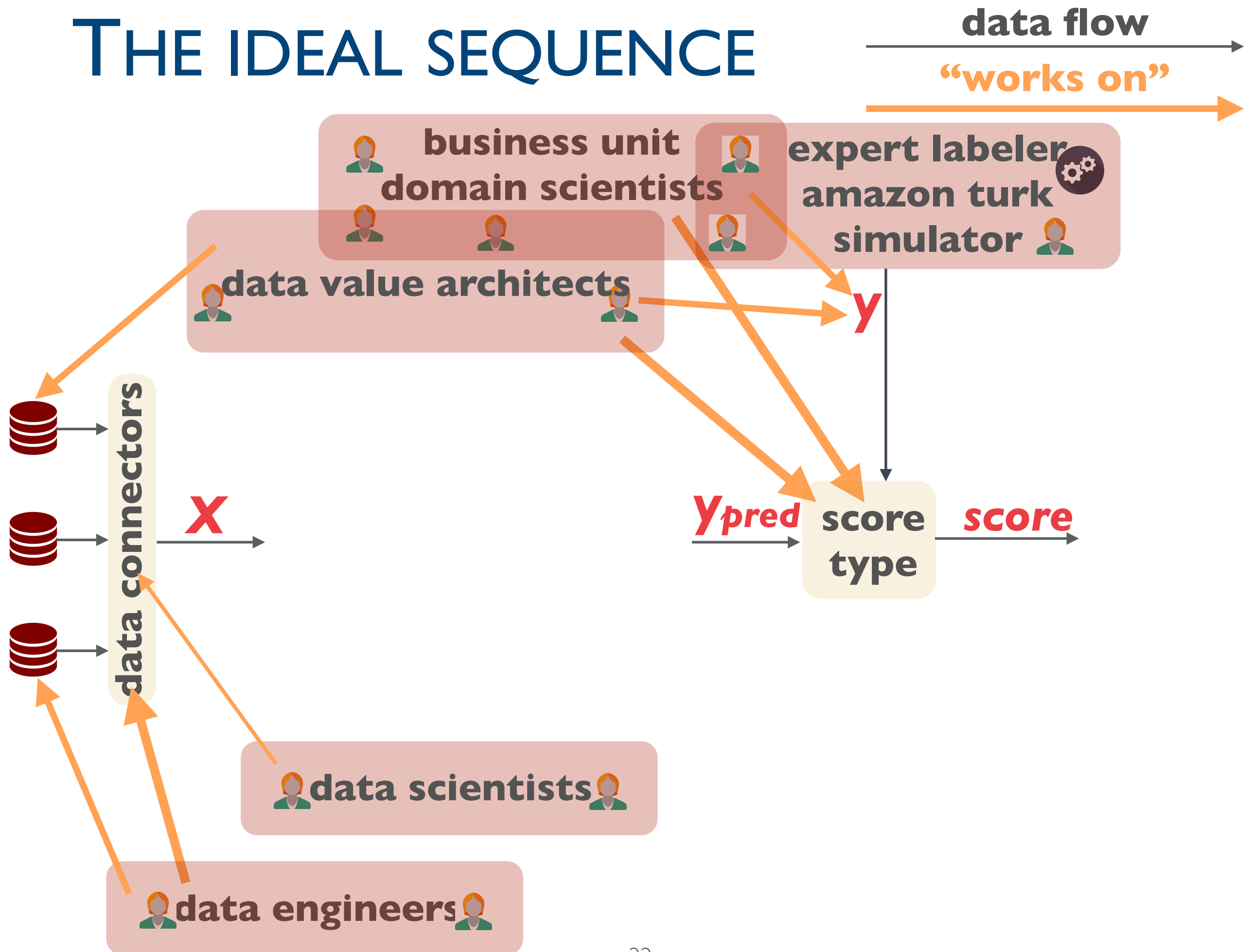
THE IDEAL SEQUENCE



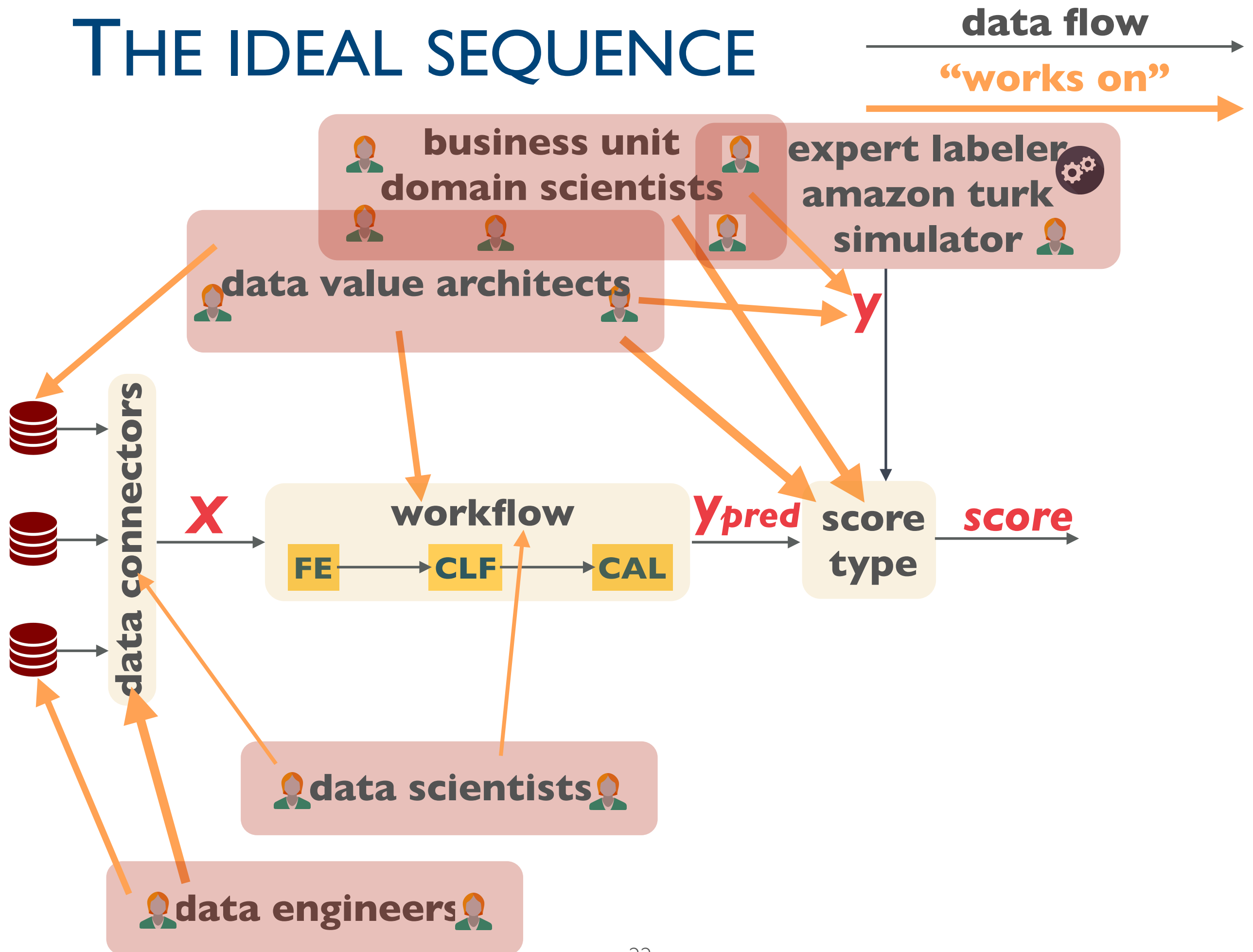
THE IDEAL SEQUENCE



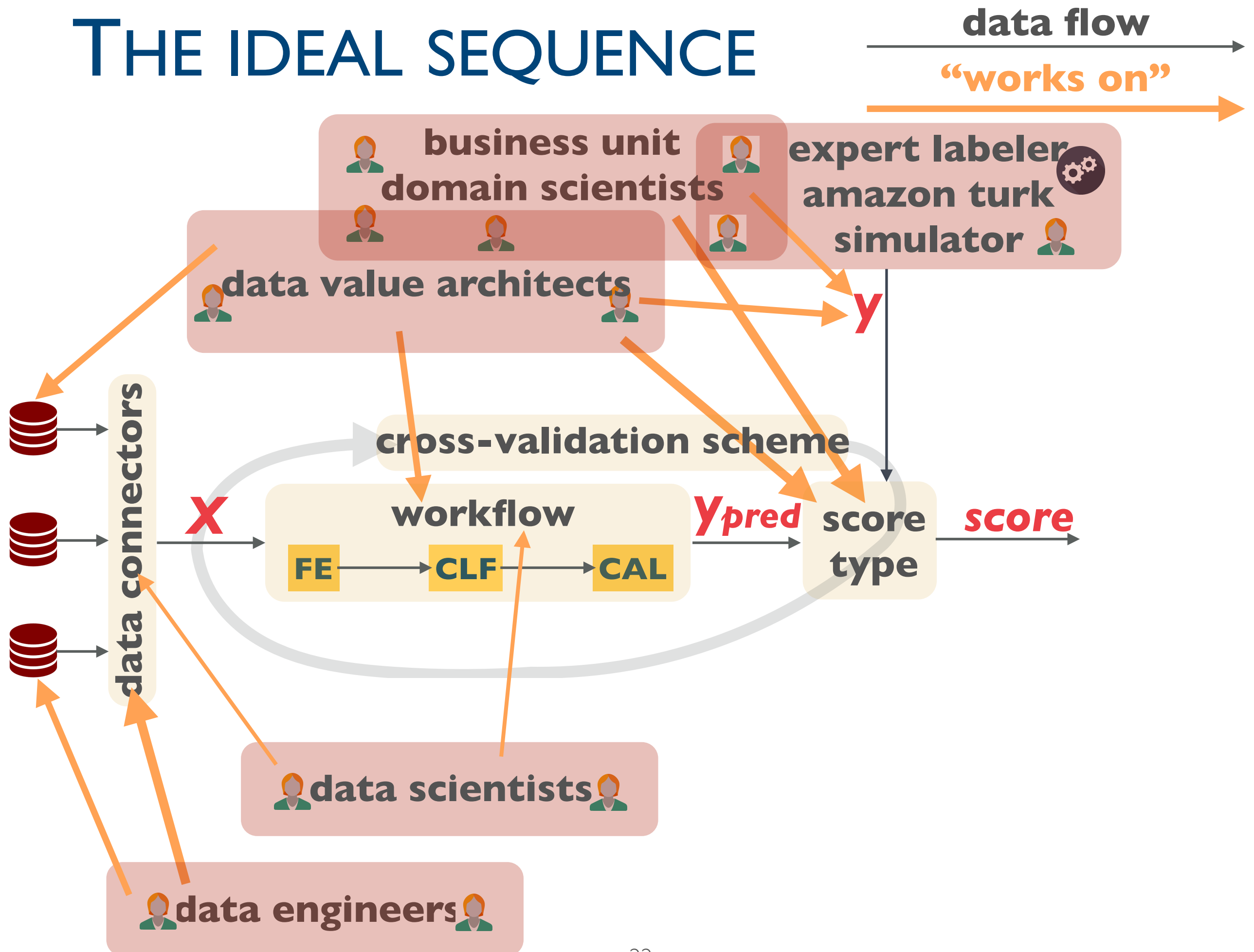
THE IDEAL SEQUENCE



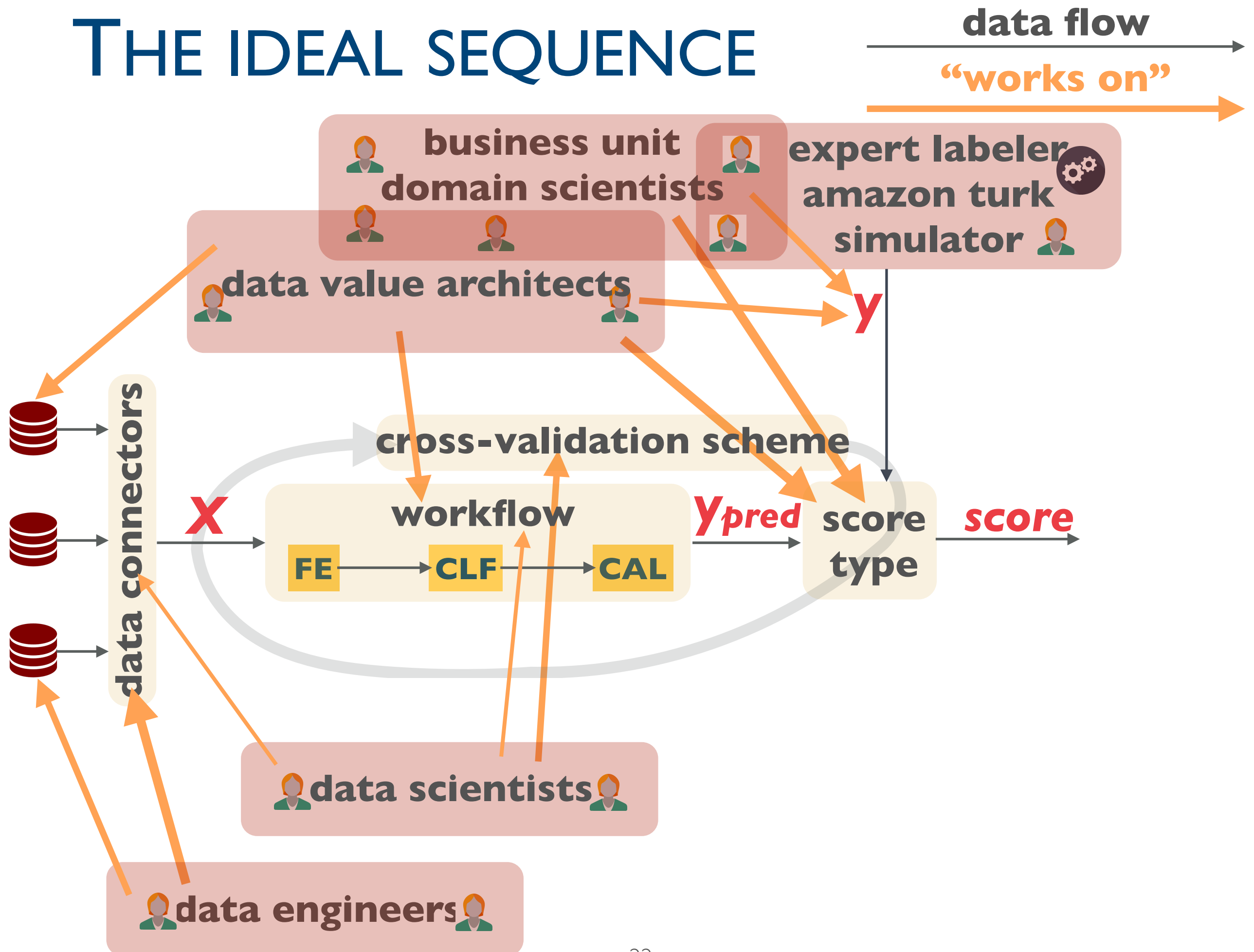
THE IDEAL SEQUENCE



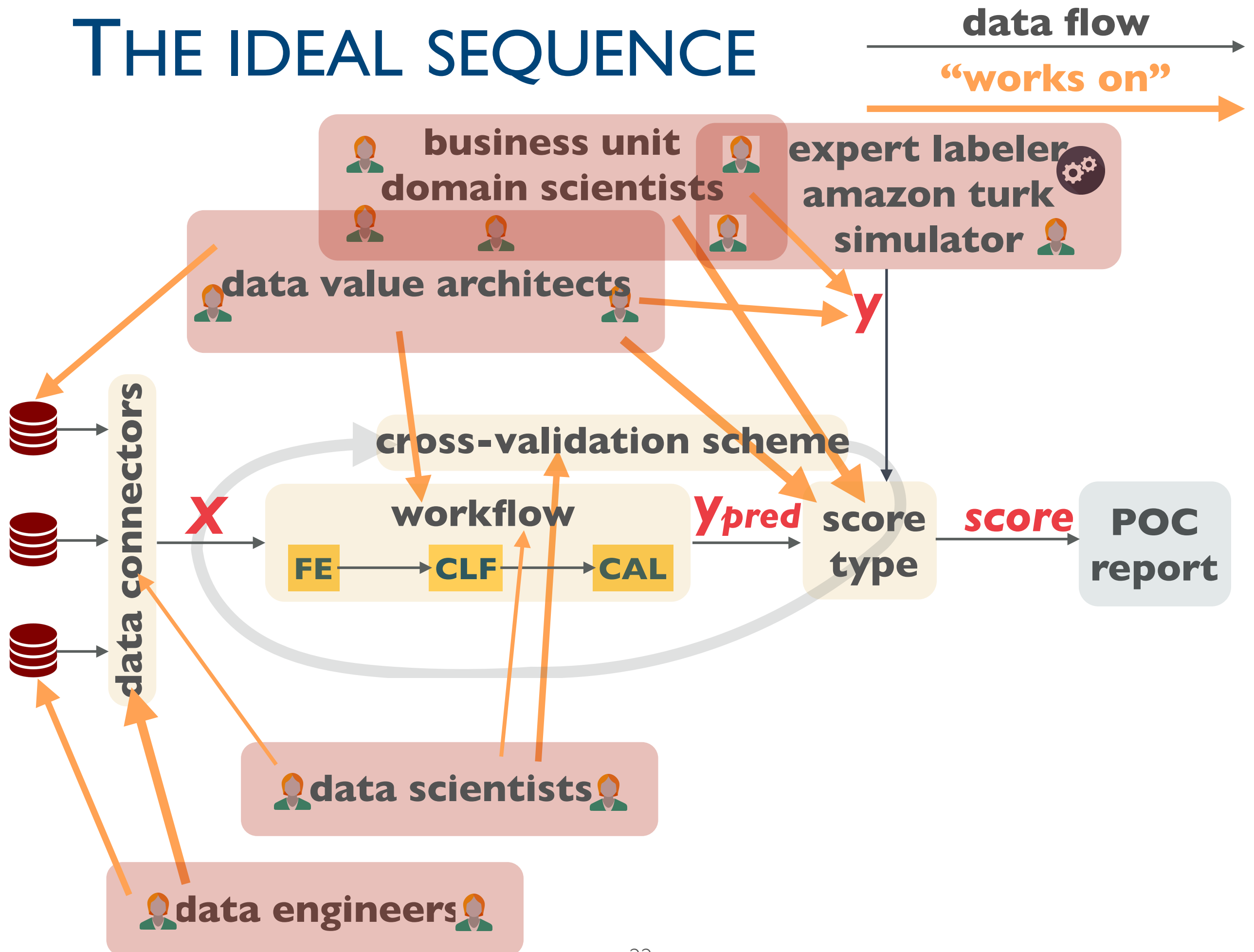
THE IDEAL SEQUENCE



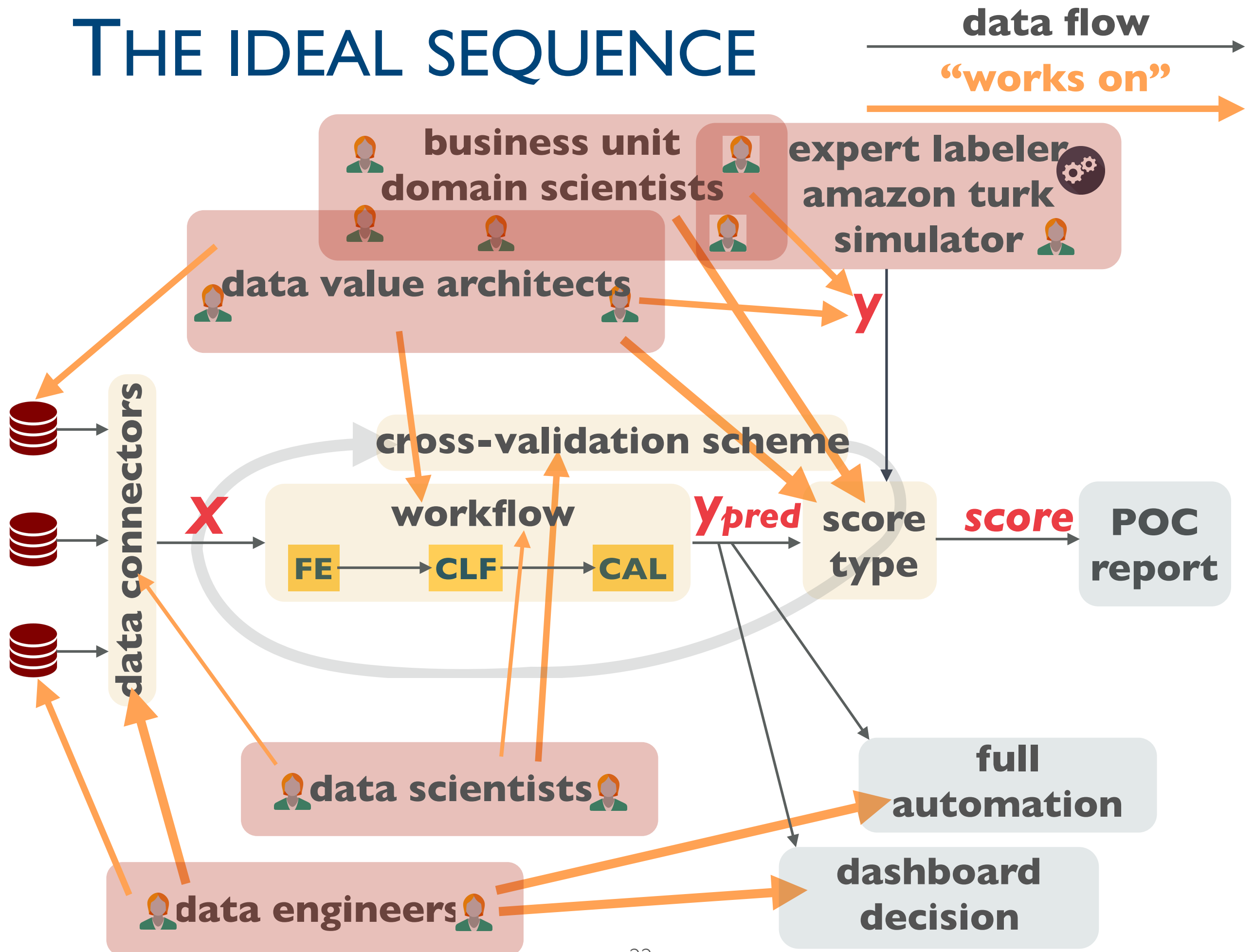
THE IDEAL SEQUENCE



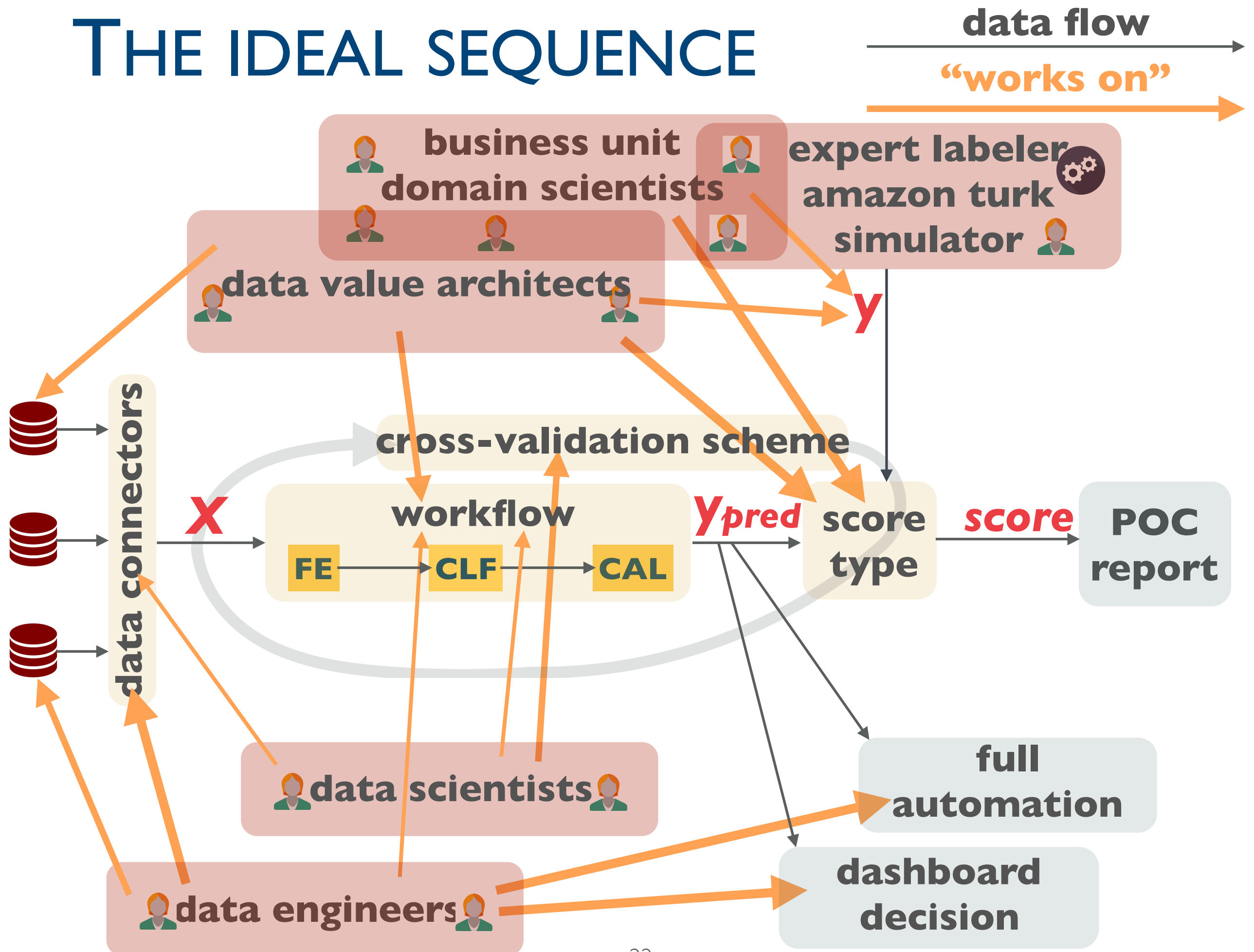
THE IDEAL SEQUENCE



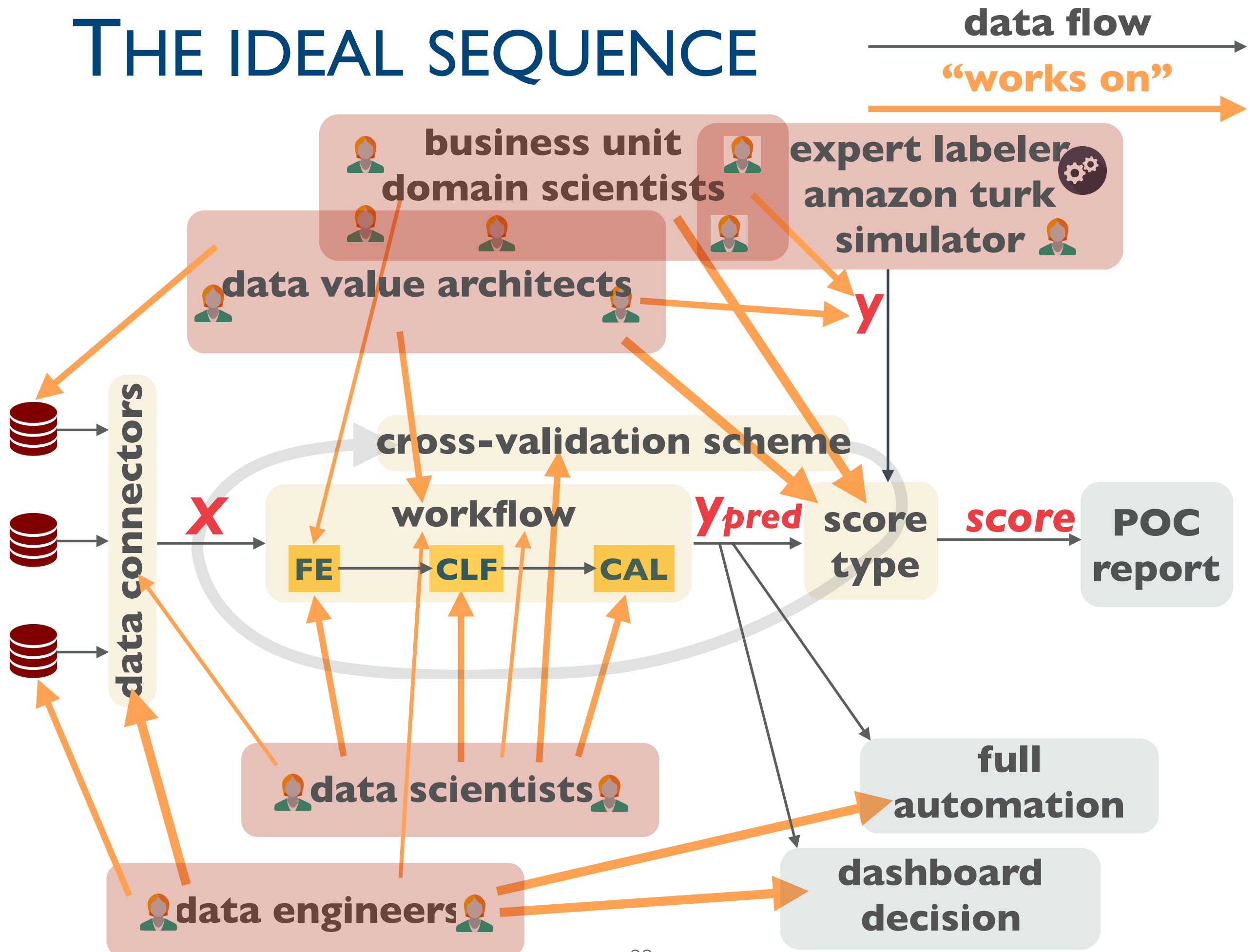
THE IDEAL SEQUENCE



THE IDEAL SEQUENCE



THE IDEAL SEQUENCE



THE DATA ANALYTICS BUILDING PIPELINE

- It is **trial and error**
 - little if any theory-based, model-based design
 - even research (development of new algorithms) is (mostly) trial and error
 - the data scientist's best friend is a **well-designed experimental studio** for facilitating fast iterations of
 - **what data to use**: car descriptors, mileage, location, meteorological history, maintenance logs, customer reviews
 - **what features to select or engineer**: which descriptors to feed into the predictor, how to digest meteorological history into a small set of numbers, predictive about car failure
 - **what predictors to use**: linear regression (classical statistics), random forests (scikit learn), neural networks (deep learning)
 - **how to parametrize the predictors**: how many trees in the forests, how many neurons

THE DATA ANALYTICS BUILDING PIPELINE

- Data-driven predictors should **work well on future (unseen) data**. This simple fact drives everything, the mindset and the design.
 - this is the **same paradigm as in classical statistics**
 - use **historical data to select and fit a model**, then use the model to **make predictions on new data**
 - but **we only have historical data**: we need to **“simulate” past and future** on existing data
 - the **train-test loop**
 - use (eg) 80% of the data for selecting the features, selecting the models, optimizing the models
 - use 20% of the data to test the model, to measure the predictive quality
 - we will need a **quantitative measure of quality/performance** because the data scientist will have to **test a lot of different choices**
 - qualitative, human-in-the-loop measures slow down the development loop
 - designing the quantitative measure is crucial: it should be tightly coupled to the KPI that we want to improve