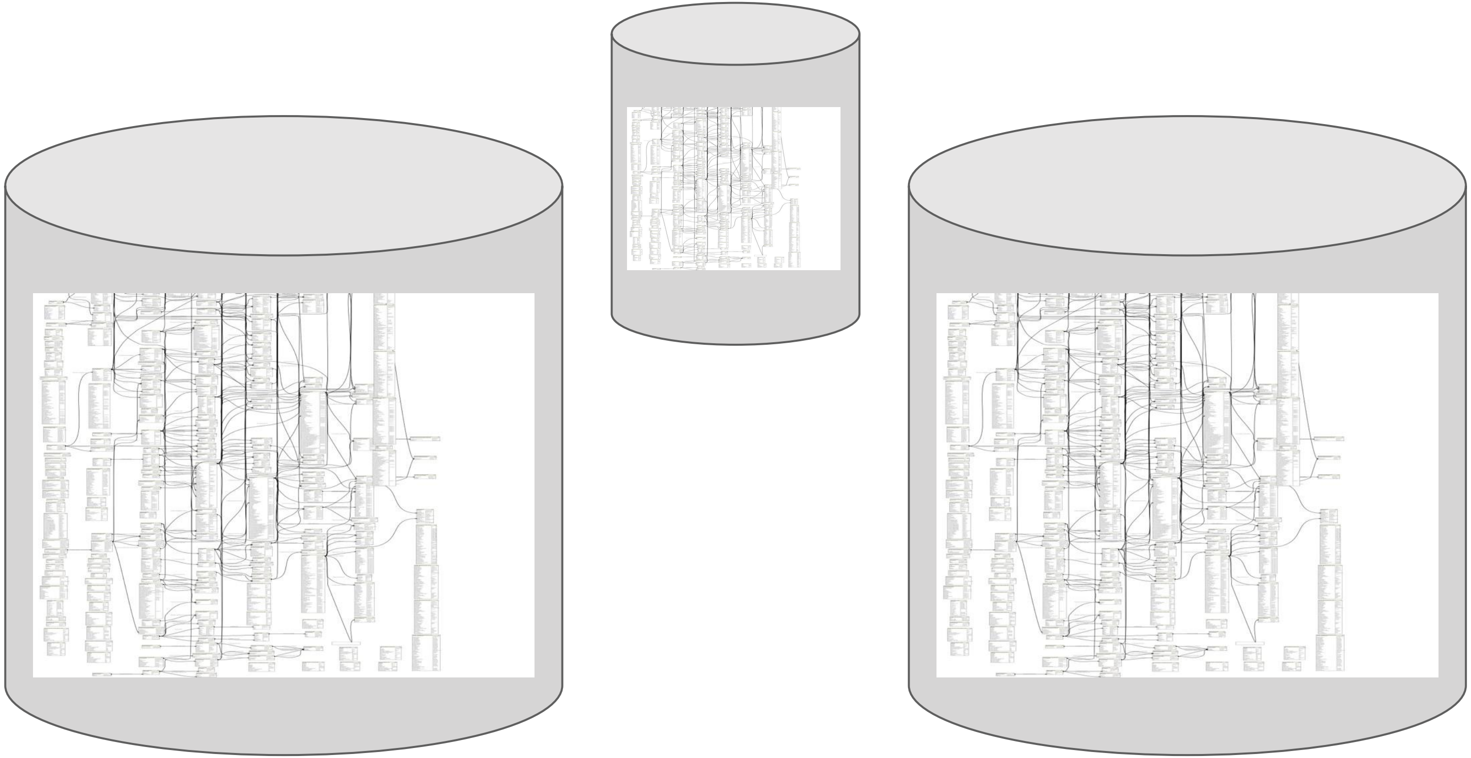
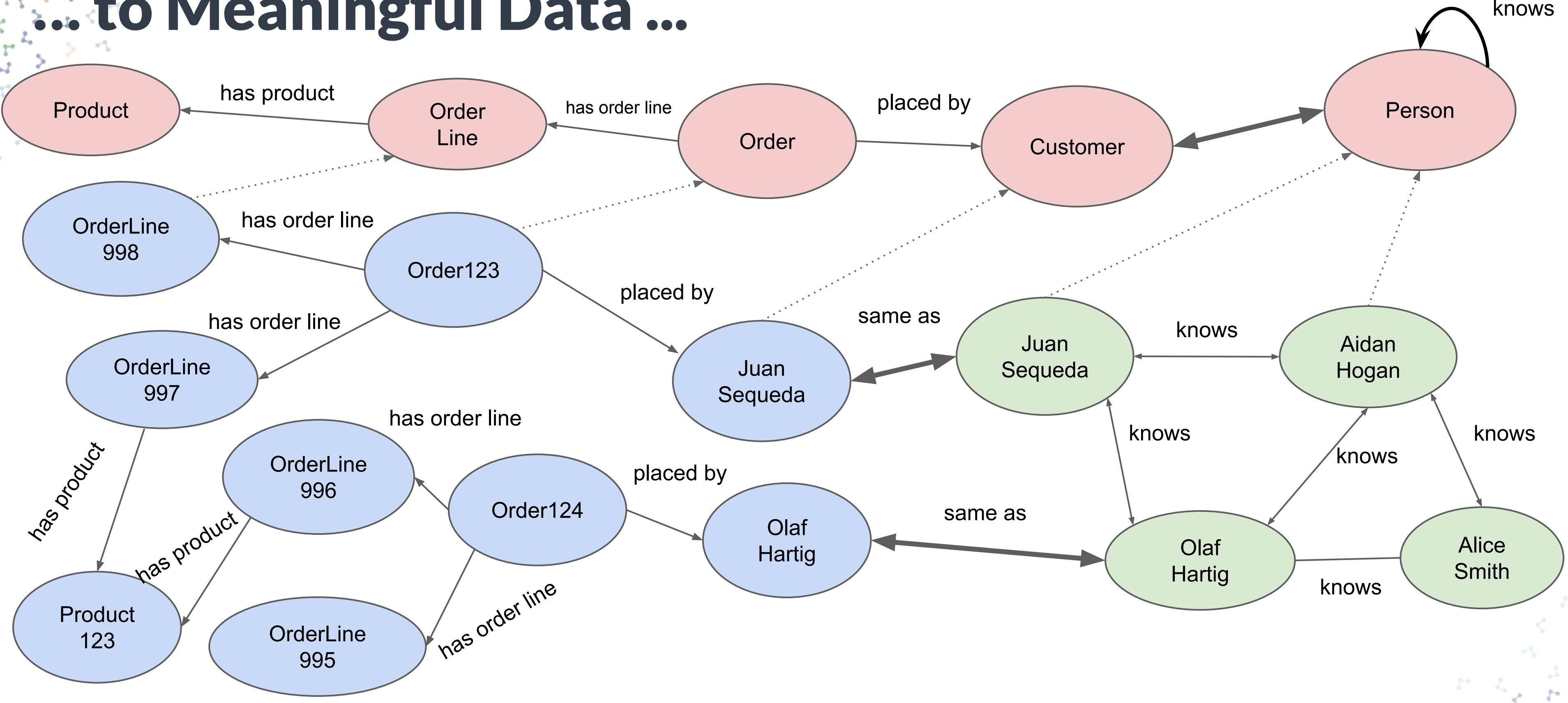
How do we get from Complex & Disparate Databases ...

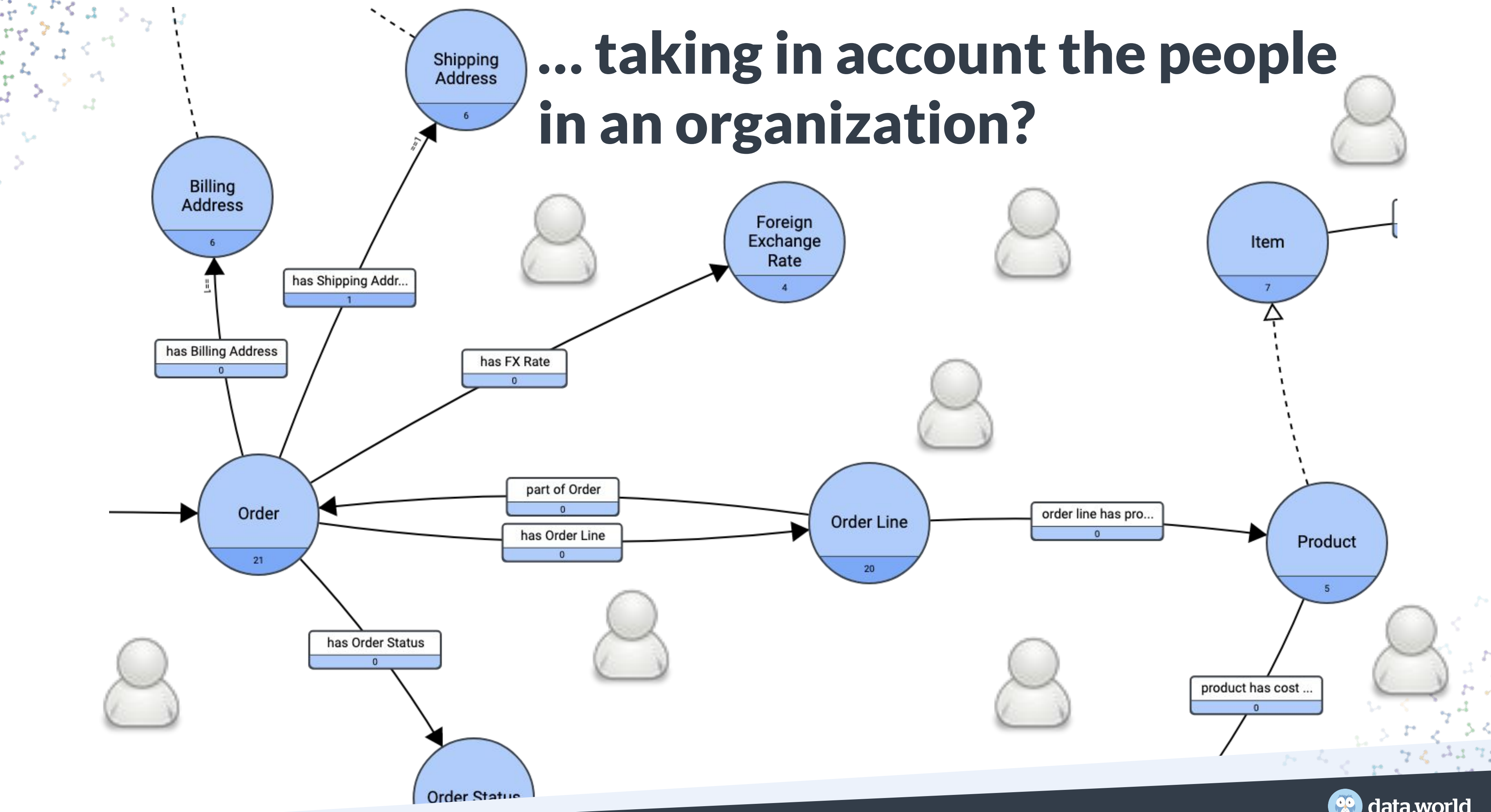


... to Meaningful Data ...



data.world

… taking in account the people in an organization?



data.world

The Socio-Technical Phenomena of Data Integration

data.world

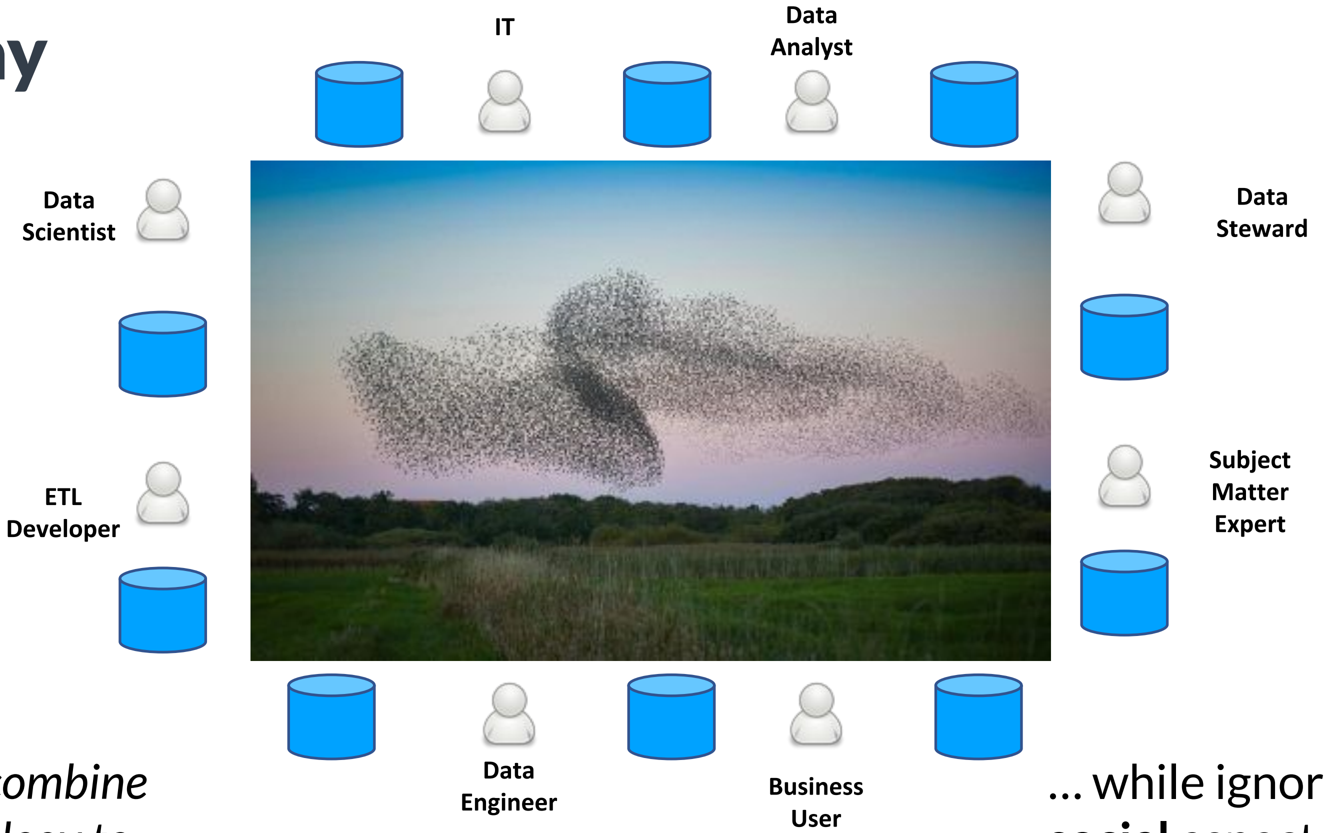
Juan Sequeda, PhD
Principal Scientist

@juansequeda

data.world

Takeaway

<u>My Thesis:</u>
We have been studying the phenomena of data integration from a technical perspective ...



<u>My Question:</u>
How can we best combine people and technology to improve data integration?

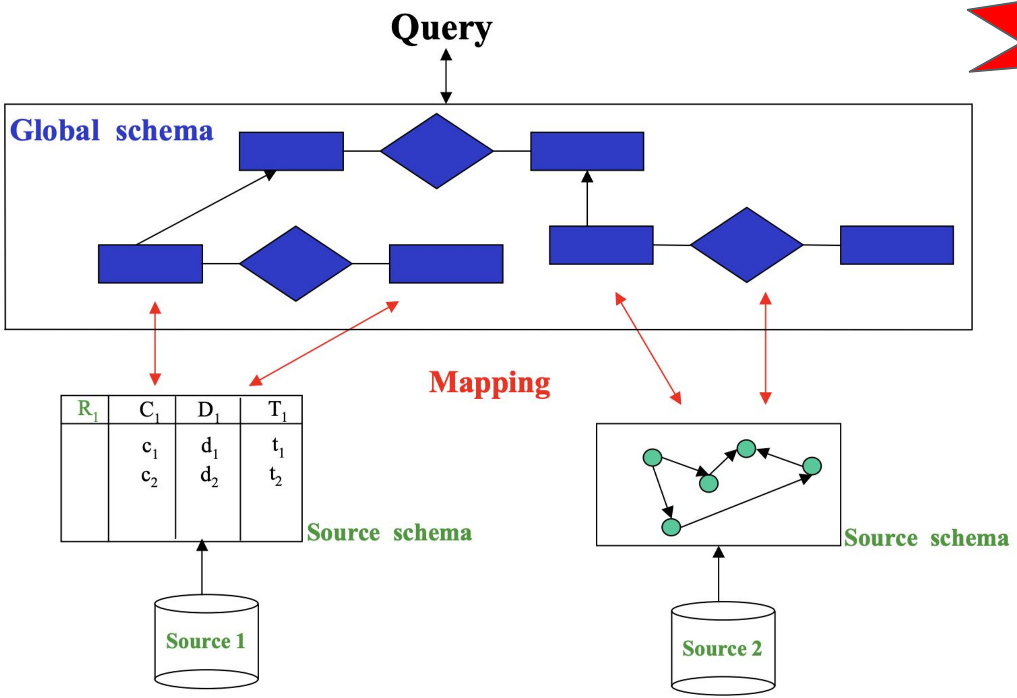
... while ignoring the **<u>social</u>** aspect

Data Integration

WHY?

“Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data.”

- Schema/Ontology Matching
- Record Linkage/Instance Matching
- Incomplete Data
- Data Quality
- Materialized vs Virtual
- ...



R_1	C_1	D_1	T_1
	c_1	d_1	t_1
	c_2	d_2	t_2

Maurizio Lenzerini (2002). "Data Integration: A Theoretical Perspective" . PODS 2002.

Real World Socio-**Technical** Problems

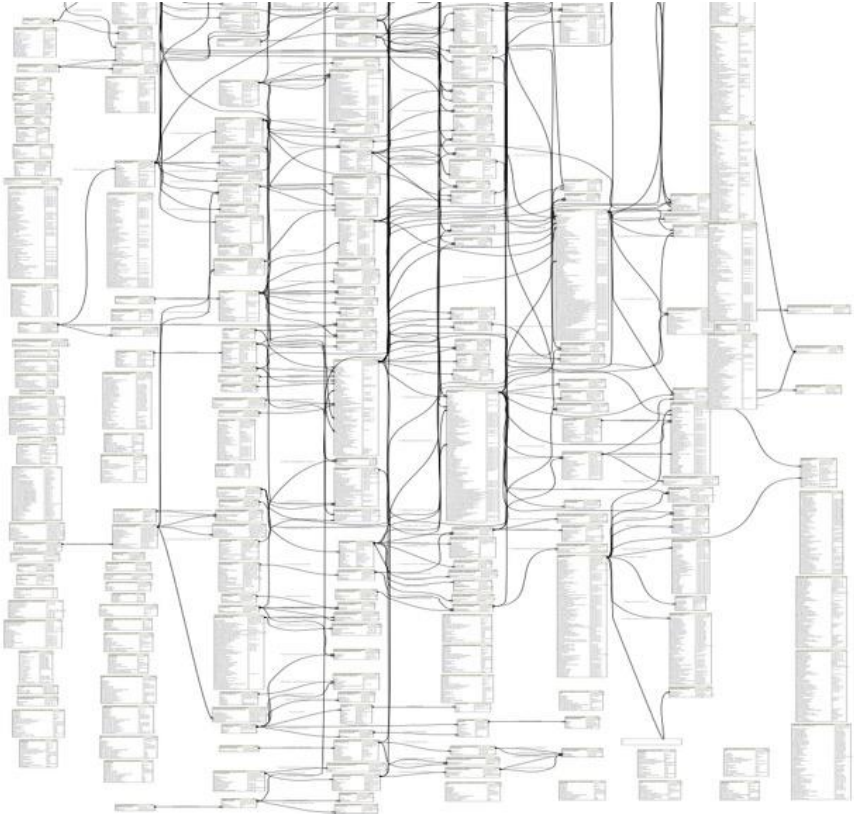
Technical 1: Understand the Source

Too many tables and attributes

Complex Relationships

Data experts unavailable

Master databases are off limits



Impossible to understand naming

Data is application centric

Documentation non-existent

Data quality unknown

Technical 2: Understand the Target

Sophisticated use of competency questions, test-driven development, ontology design patterns, reuse.

Populating ontology with data coming from a relational database is an afterthought and not part of methodologies.

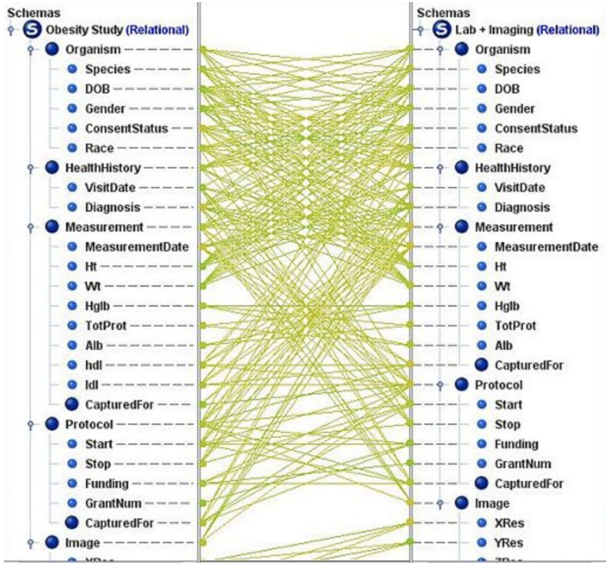
Numerous upper level ontologies for reuse: Good Relations, FIBO, Gist, Schema.org, etc.

Technical 3: Understand the Mappings

Ontology/Schema matching between relational schemas and ontologies works … in theory. But not in practice.

MasterOrder,
Order,
P_Order,
Order_P

segment1,
segment2,
…,
segment99

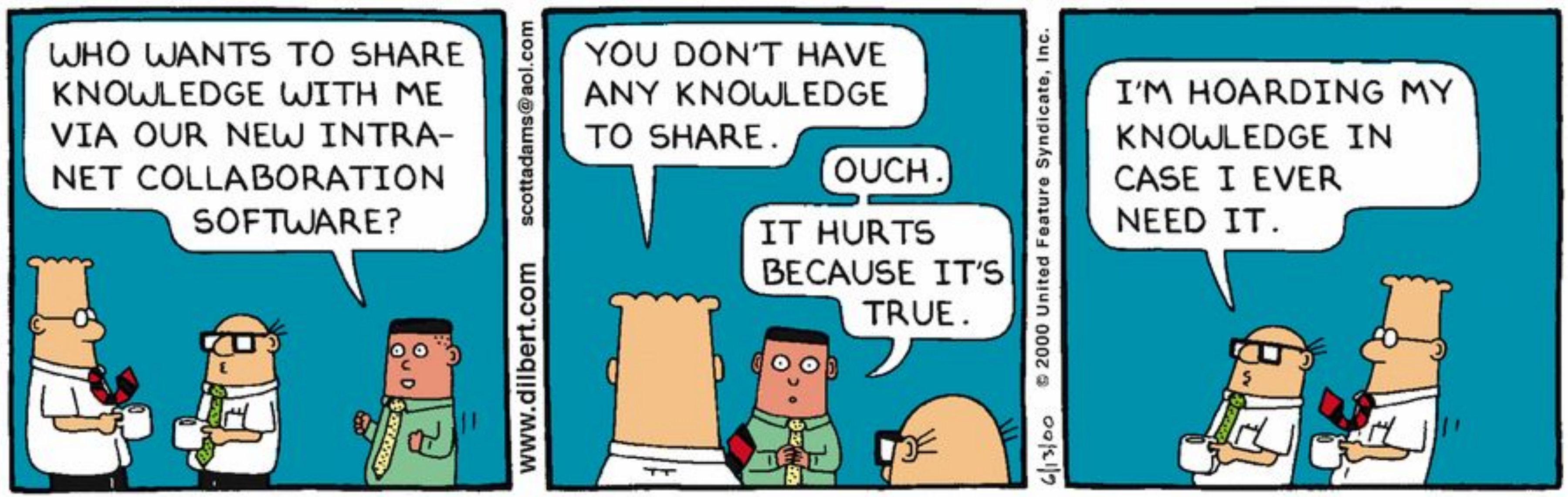


1-1 correspondence between table-classes and columns-properties are rare

Not plausible we will ever have large amounts of schemas & mappings to train ML models.

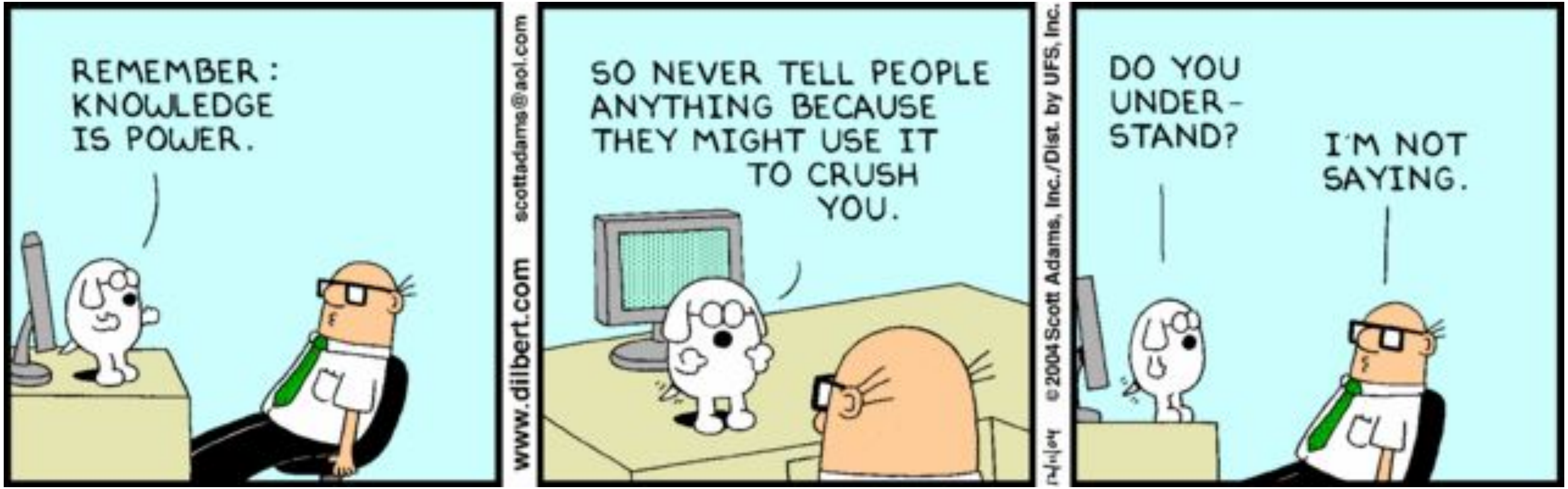
Real World <u>**Socio**</u>-Technical Problems

Social 1: Knowledge Hoarding

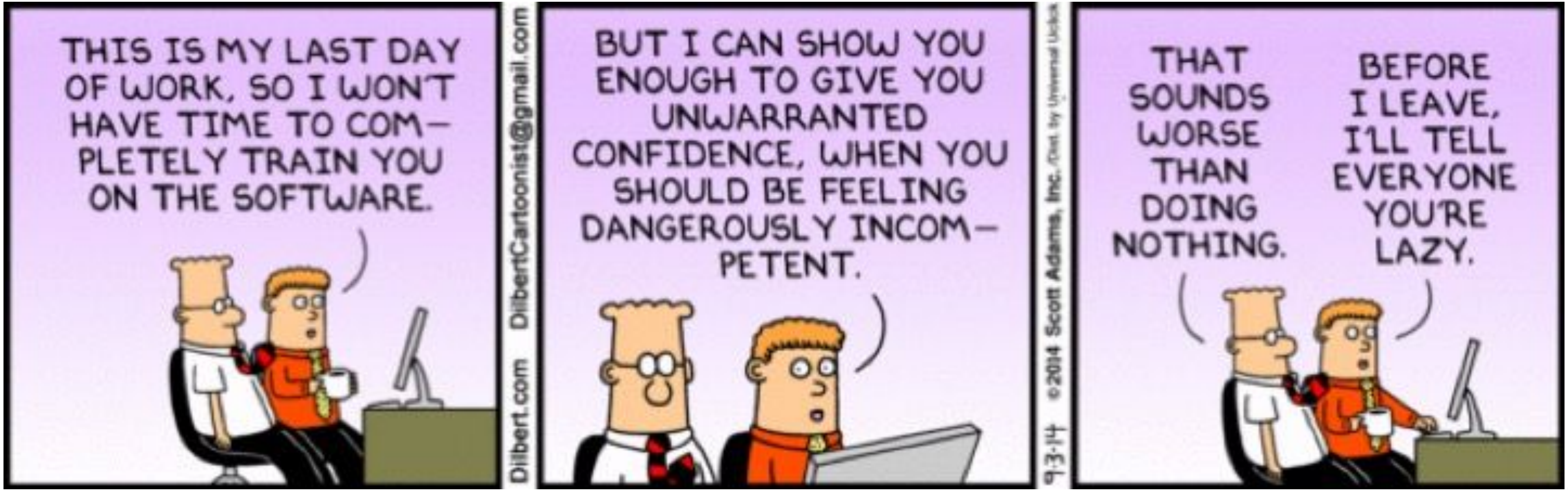

WHO WANTS TO SHARE KNOWLEDGE WITH ME VIA OUR NEW INTRA-NET COLLABORATION SOFTWARE?
scottadams@aol.com
www.dilbert.com
YOU DON'T HAVE ANY KNOWLEDGE TO SHARE.
OUCH.
IT HURTS BECAUSE IT'S TRUE.
© 2000 United Feature Syndicate, Inc.
6/13/00
I'M HOARDING MY KNOWLEDGE IN CASE I EVER NEED IT.

data.world

Social 2: Knowledge Aspiration


REMEMBER: KNOWLEDGE IS POWER.
scottadams@aol.com
www.dilbert.com
SO NEVER TELL PEOPLE ANYTHING BECAUSE THEY MIGHT USE IT TO CRUSH YOU.
© 2004 Scott Adams, Inc./Dist. by UFS, Inc.
12/11/04
DO YOU UNDER-STAND?
I'M NOT SAYING.

Social 3: Knowledge Retirement


THIS IS MY LAST DAY OF WORK, SO I WON'T HAVE TIME TO COMPLETELY TRAIN YOU ON THE SOFTWARE.
BUT I CAN SHOW YOU ENOUGH TO GIVE YOU UNWARRANTED CONFIDENCE, WHEN YOU SHOULD BE FEELING DANGEROUSLY INCOMPETENT.
THAT SOUNDS WORSE THAN DOING NOTHING.
BEFORE I LEAVE, I'LL TELL EVERYONE YOU'RE LAZY.

Dilbert.com DilbertCartoonist@gmail.com

9-3-14 ©2014 Scott Adams, Inc. /Dist. by Universal Uclick

Real World **Socio-Technical** Problems

data.world

Business Question

How many orders were placed in December 2019?

E-commerce: 118,595

Shipping: 114,324

Finance: 116,211

What do you mean by ...

What is an **Order**?

E-commerce: 118,595

When the user clicks "Order" on the website.

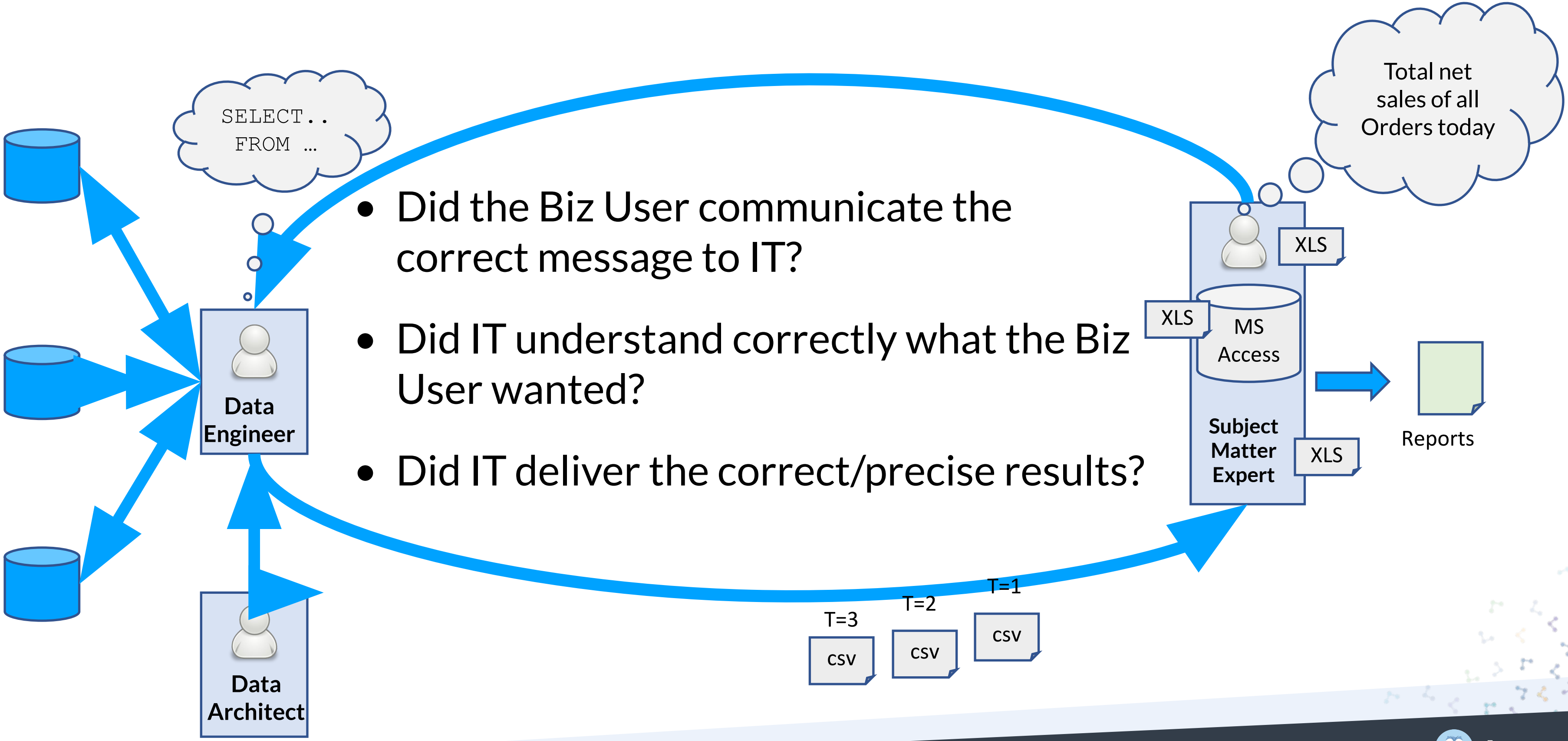
Shipping: 114,324

When the customer has received the product

Finance: 116,211

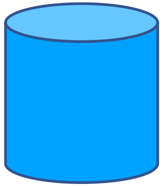
When it comes out of the billing system and the CC has been charged

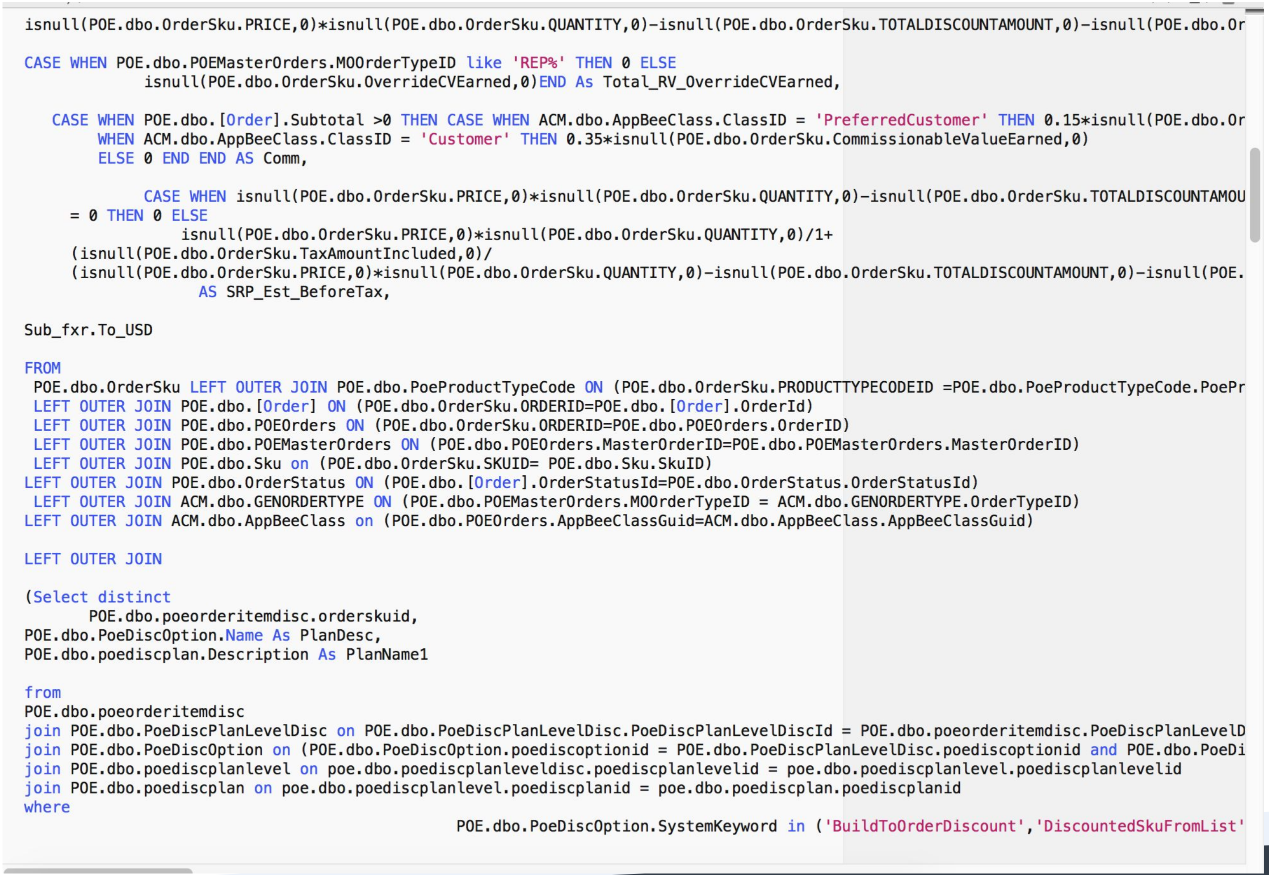
Spreadsheet Approach



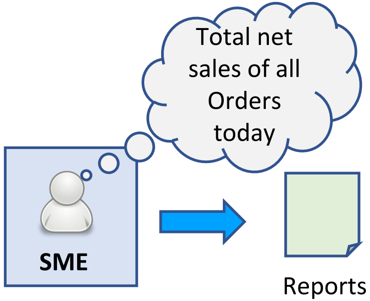
- Did the Biz User communicate the correct message to IT?
- Did IT understand correctly what the Biz User wanted?
- Did IT deliver the correct/precise results?

Query Approach

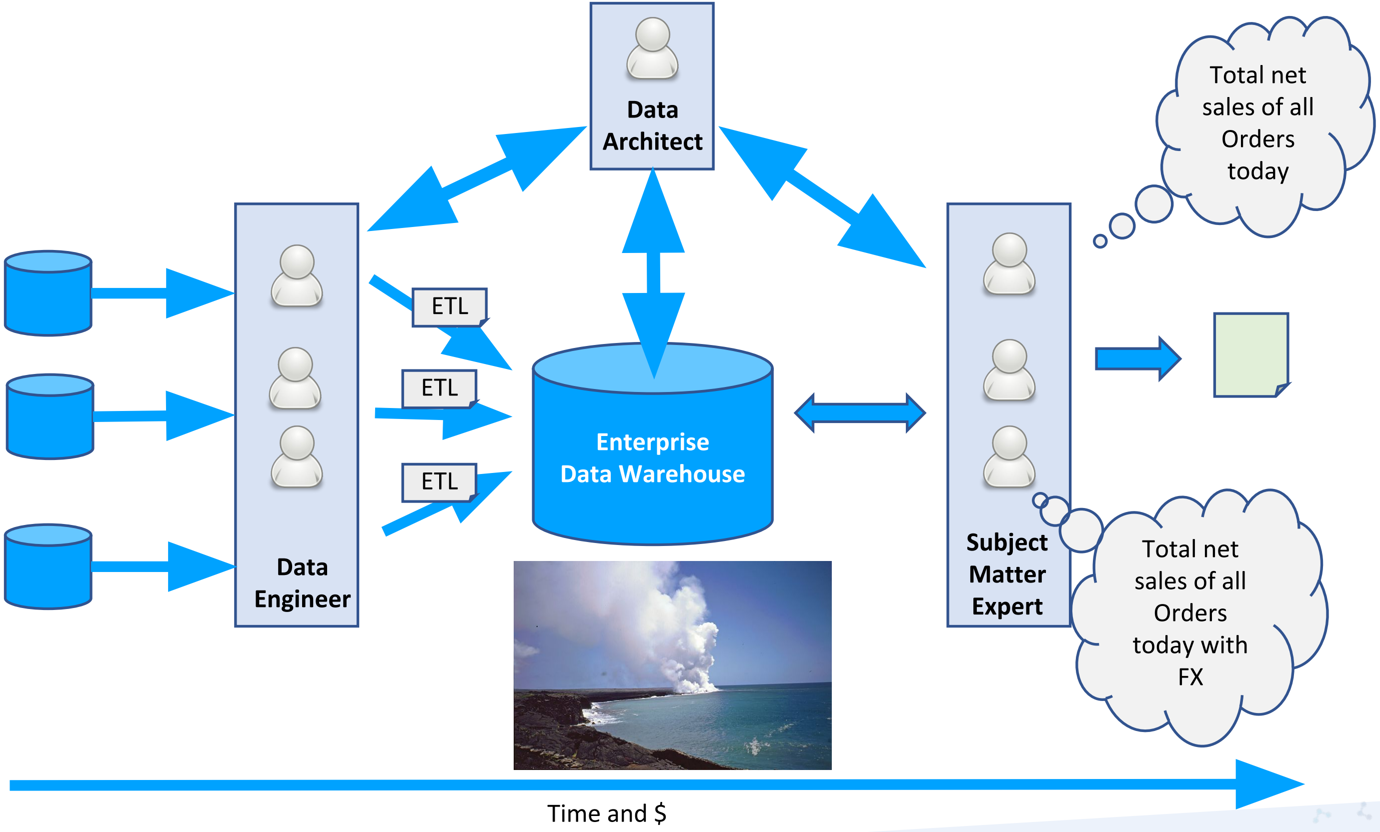




```
isnull(POE.dbo.OrderSku.PRICE,0)*isnull(POE.dbo.OrderSku.QUANTITY,0)-isnull(POE.dbo.OrderSku.TOTALDISCOUNTAMOUNT,0)-isnull(POE.dbo.Or

CASE WHEN POE.dbo.POEMasterOrders.MOOrderTypeID like 'REP%' THEN 0 ELSE
             isnull(POE.dbo.OrderSku.OverrideCVEarned,0)END As Total_RV_OverrideCVEarned,

   CASE WHEN POE.dbo.[Order].Subtotal >0 THEN CASE WHEN ACM.dbo.AppBeeClass.ClassID = 'PreferredCustomer' THEN 0.15*isnull(POE.dbo.Or
        WHEN ACM.dbo.AppBeeClass.ClassID = 'Customer' THEN 0.35*isnull(POE.dbo.OrderSku.CommissionableValueEarned,0)
        ELSE 0 END END AS Comm,

             CASE WHEN isnull(POE.dbo.OrderSku.PRICE,0)*isnull(POE.dbo.OrderSku.QUANTITY,0)-isnull(POE.dbo.OrderSku.TOTALDISCOUNTAMOU
     = 0 THEN 0 ELSE
                 isnull(POE.dbo.OrderSku.PRICE,0)*isnull(POE.dbo.OrderSku.QUANTITY,0)/1+
     (isnull(POE.dbo.OrderSku.TaxAmountIncluded,0)/
     (isnull(POE.dbo.OrderSku.PRICE,0)*isnull(POE.dbo.OrderSku.QUANTITY,0)-isnull(POE.dbo.OrderSku.TOTALDISCOUNTAMOUNT,0)-isnull(POE.
                   AS SRP_Est_BeforeTax,

Sub_fxr.To_USD

FROM
 POE.dbo.OrderSku LEFT OUTER JOIN POE.dbo.PoeProductTypeCode ON (POE.dbo.OrderSku.PRODUCTTYPECODEID =POE.dbo.PoeProductTypeCode.PoePr
 LEFT OUTER JOIN POE.dbo.[Order] ON (POE.dbo.OrderSku.ORDERID=POE.dbo.[Order].OrderId)
 LEFT OUTER JOIN POE.dbo.POEOrders ON (POE.dbo.OrderSku.ORDERID=POE.dbo.POEOrders.OrderID)
 LEFT OUTER JOIN POE.dbo.POEMasterOrders ON (POE.dbo.POEOrders.MasterOrderID=POE.dbo.POEMasterOrders.MasterOrderID)
 LEFT OUTER JOIN POE.dbo.Sku on (POE.dbo.OrderSku.SKUID= POE.dbo.Sku.SkuID)
LEFT OUTER JOIN POE.dbo.OrderStatus ON (POE.dbo.[Order].OrderStatusId=POE.dbo.OrderStatus.OrderStatusId)
 LEFT OUTER JOIN ACM.dbo.GENORDERTYPE ON (POE.dbo.POEMasterOrders.MOOrderTypeID = ACM.dbo.GENORDERTYPE.OrderTypeID)
LEFT OUTER JOIN ACM.dbo.AppBeeClass on (POE.dbo.POEOrders.AppBeeClassGuid=ACM.dbo.AppBeeClass.AppBeeClassGuid)

LEFT OUTER JOIN

(Select distinct
       POE.dbo.poeorderitemdisc.orderskuid,
POE.dbo.PoeDiscOption.Name As PlanDesc,
POE.dbo.poediscplan.Description As PlanName1

from
POE.dbo.poeorderitemdisc
join POE.dbo.PoeDiscPlanLevelDisc on POE.dbo.PoeDiscPlanLevelDisc.PoeDiscPlanLevelDiscId = POE.dbo.poeorderitemdisc.PoeDiscPlanLevelD
join POE.dbo.PoeDiscOption on (POE.dbo.PoeDiscOption.poediscoptionid = POE.dbo.PoeDiscPlanLevelDisc.poediscoptionid and POE.dbo.PoeDi
join POE.dbo.poediscplanlevel on poe.dbo.poediscplanleveldisc.poediscplanlevelid = poe.dbo.poediscplanlevel.poediscplanlevelid
join POE.dbo.poediscplan on poe.dbo.poediscplanlevel.poediscplanid = poe.dbo.poediscplan.poediscplanid
where
                                            POE.dbo.PoeDiscOption.SystemKeyword in ('BuildToOrderDiscount','DiscountedSkuFromList'
```



Data Warehouse Approach



Data Architect

Data Engineer

ETL

ETL

ETL

Enterprise Data Warehouse

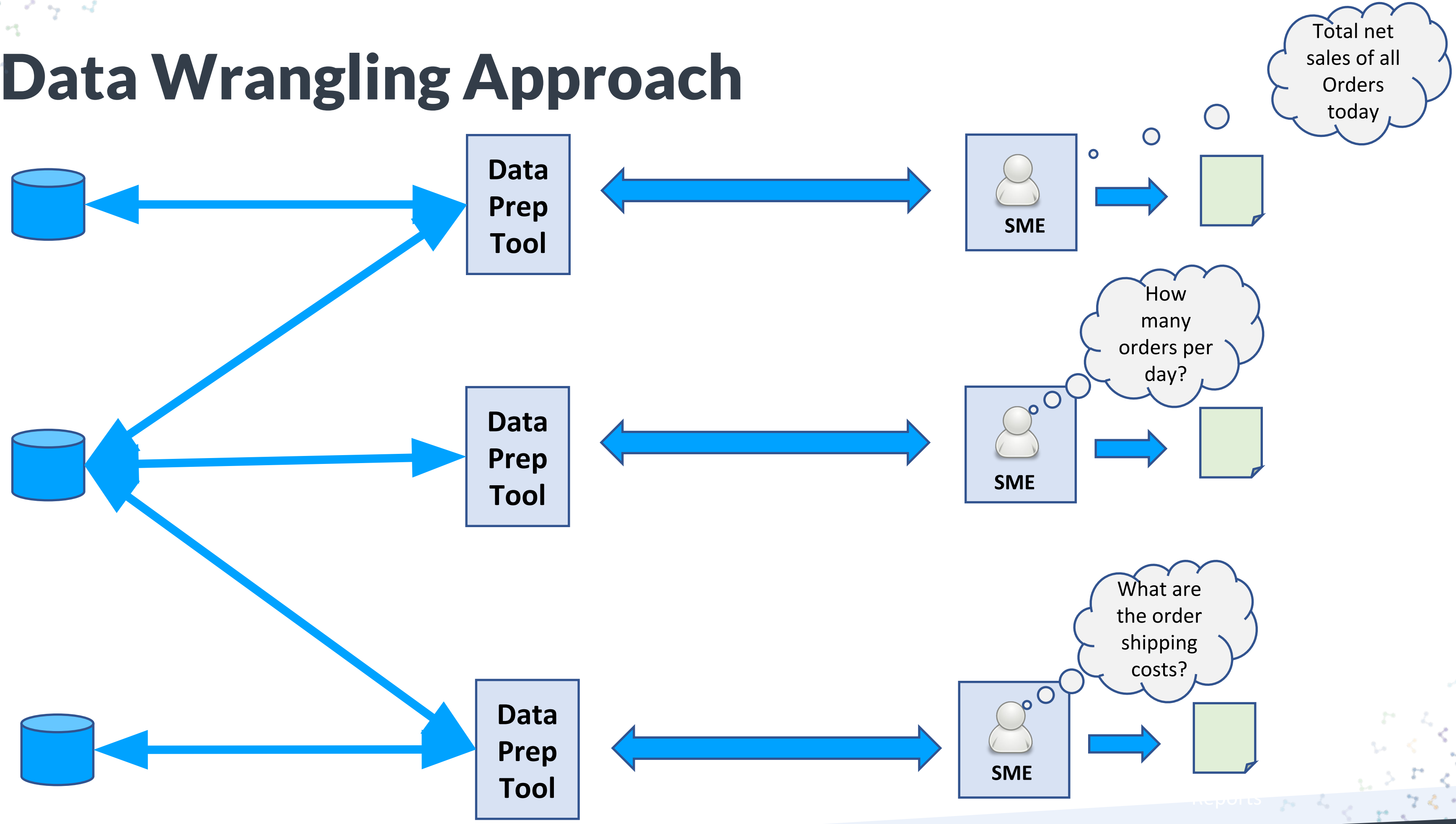
Subject Matter Expert

Total net sales of all Orders today

Total net sales of all Orders today with FX

Time and $

Data Wrangling Approach



Why is this hard?

Strings to Things

What is an Order?

An order is if it had shipped or the accounts receivable had been received.

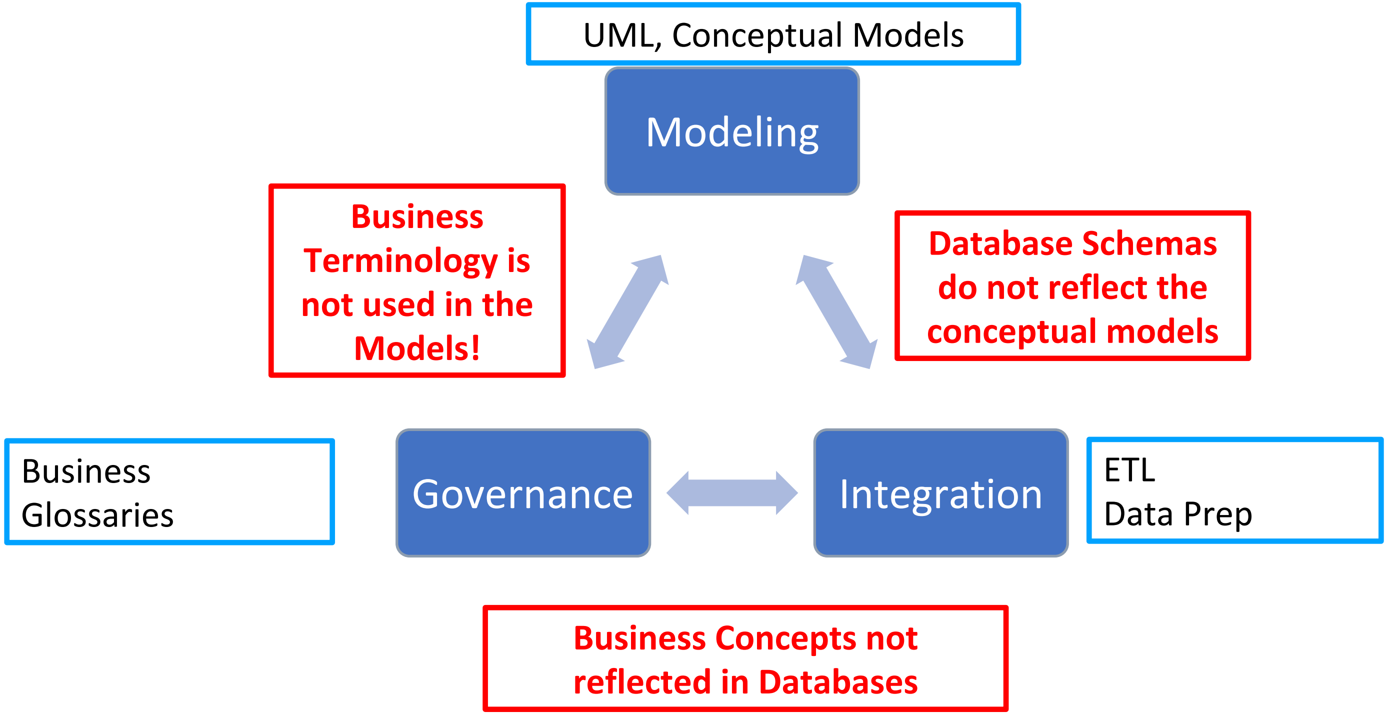
```
SELECT moid
FROM masterorder m
JOIN order o on m.oid = o.oid
WHERE ordertype IN (2,3)
```

What is the net sales of an order?

Net sales of an order is calculated by subtracting the tax and the shipping cost from the final price and also adjusting based upon the discount given. However, if the currency of the order is not in USD or CAD, then the shipping tax must be subtracted.

```
SELECT
    moid,
    o.ordertotal - ot.finaltax -
    CASE WHEN o.currencyid in (“USD”, “CAD”)
        THEN o.shippingcost
        ELSE o.shippingcost = ot.shippingtax
    END as ordernetsales
FROM masterorder m
JOIN order o on m.oid = o.id
JOIN ordertax ot on o.oid = ot.oids
```

People and Tool Fragmentation


UML, Conceptual Models
Modeling
Business Terminology is not used in the Models!
Database Schemas do not reflect the conceptual models
Business Glossaries
Governance
Integration
ETL
Data Prep
Business Concepts not reflected in Databases

Real World Example

An E-Commerce Case Study

Background

- Customer invested in expensive EDW
- Business Users were skeptical of the data feeding the BI reports
- EDW was not being used.
- BU generating reports directly from sources.

Challenge

- IT quickly became the bottleneck
- Friction between BU and IT due to lack of agreed terminology.
- Tribal knowledge in Excel, MS Access.
- Different answers for the same question.
- How is Tableau going to be successful?

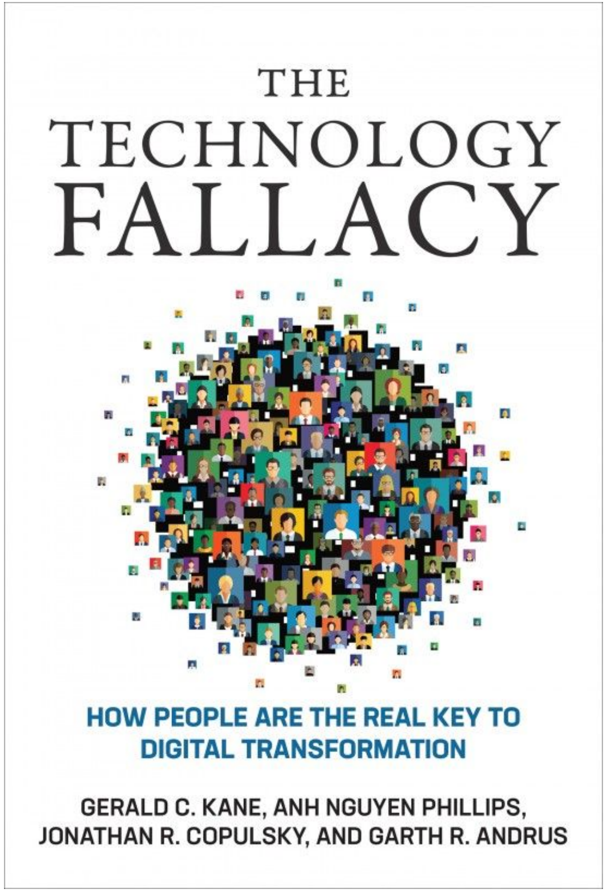
Customer Need

- Consistent, understandable and trusted data view across the multiple relational databases.

- Tableau needs to consume the trusted data.

- Large number of business users should generate reports using the same trusted data.

- Agile approach in order to start showing value quickly.

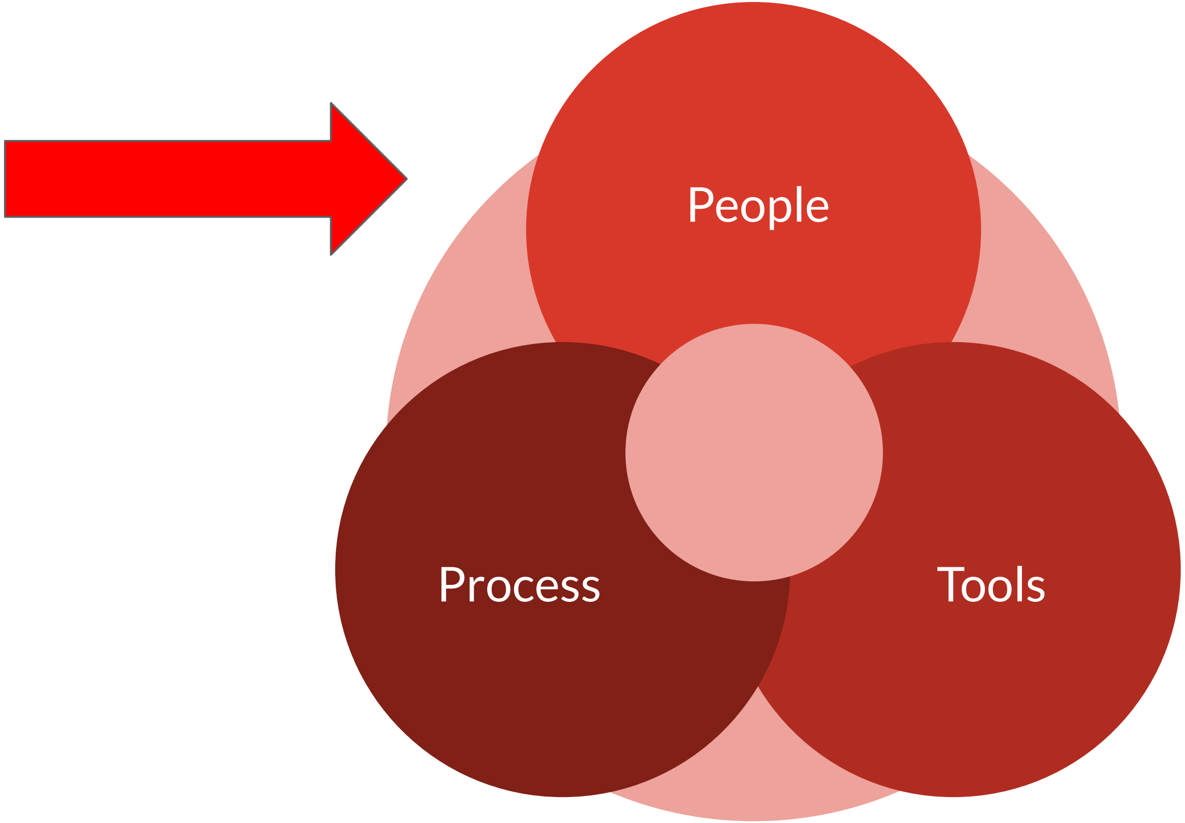


THE TECHNOLOGY FALLACY

HOW PEOPLE ARE THE REAL KEY TO DIGITAL TRANSFORMATION

GERALD C. KANE, ANH NGUYEN PHILLIPS, JONATHAN R. COPULSKY, AND GARTH R. ANDRUS

"The mistaken belief that just because business challenges are caused by digital technology, that they also need to be solved by digital technology."


People
Process
Tools

People

Data Engineer

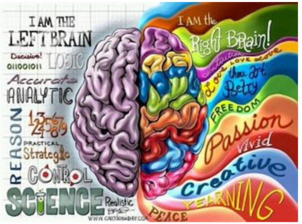
Understand database schemas, including how the data are interconnected.

Knowledge Scientist

Serves as the communication and developer bridge between Data Engineers and Business Users

Data Access

"Geeky Person"



Business Modeling

"People Person"

Business User

SME who understand the business

Knowledge Scientist vs Data Scientist

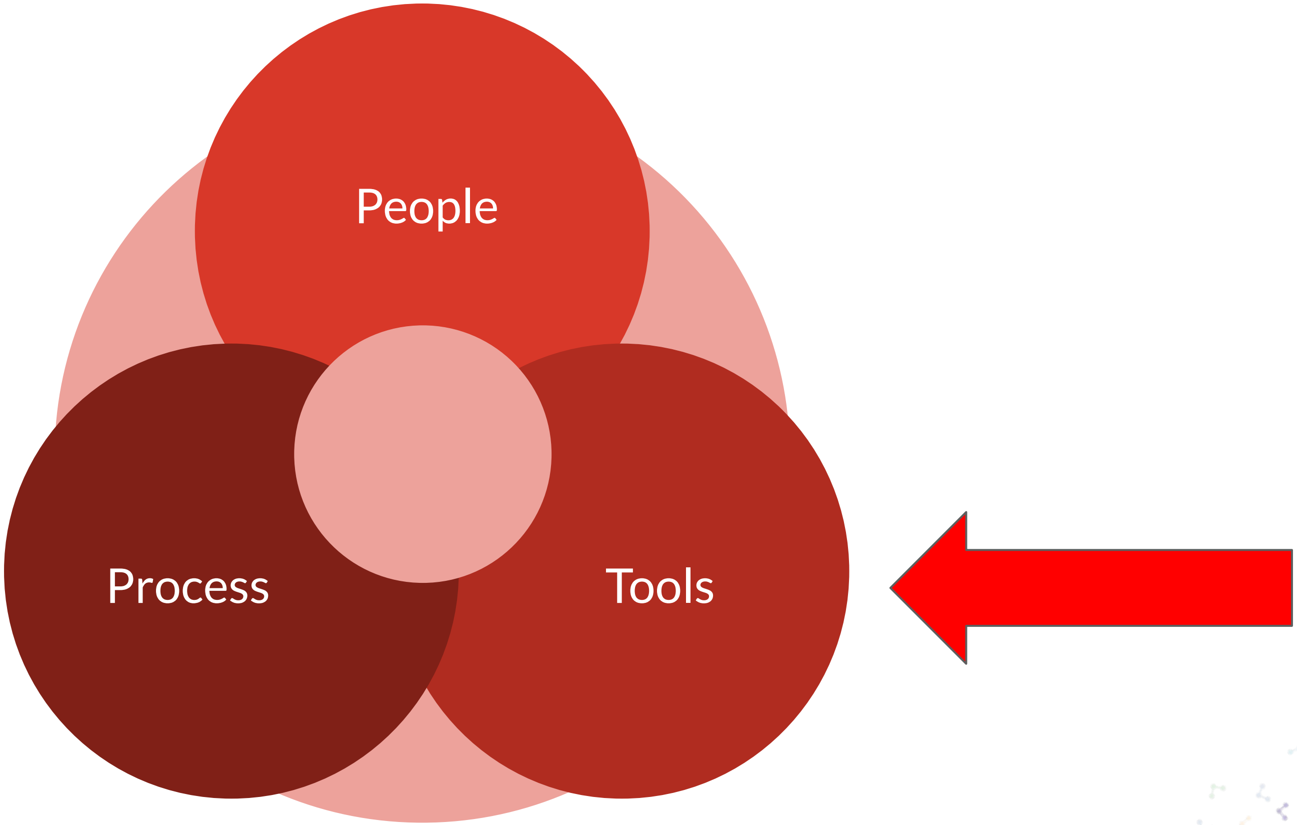
"Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data, which is an inefficient data strategy"

→ Knowledge Scientist

https://www.infoworld.com/article/3228245/data-science/the-80-20-data-science-dilemma.html

Knowledge Scientist

Data Scientist



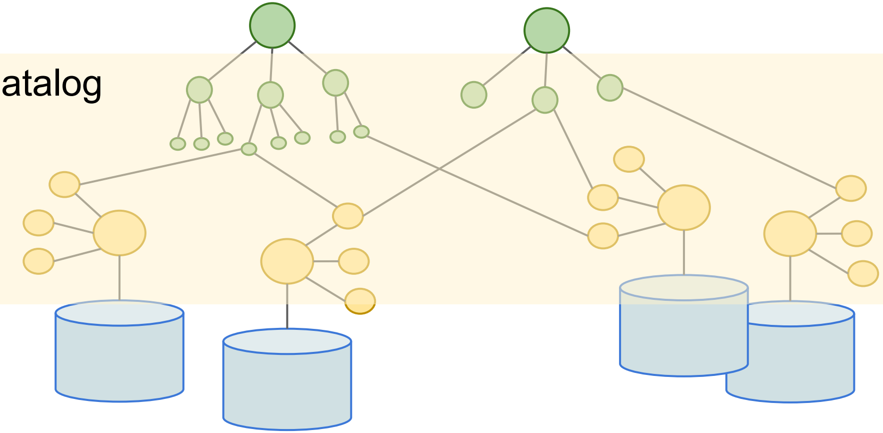
People

Process

Tools

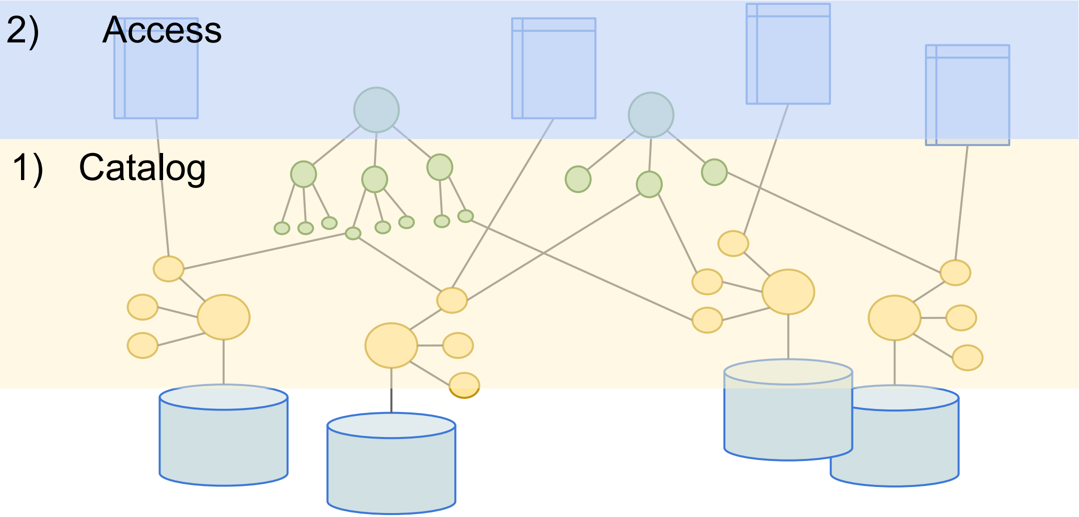
What data do we have?

1) Catalog

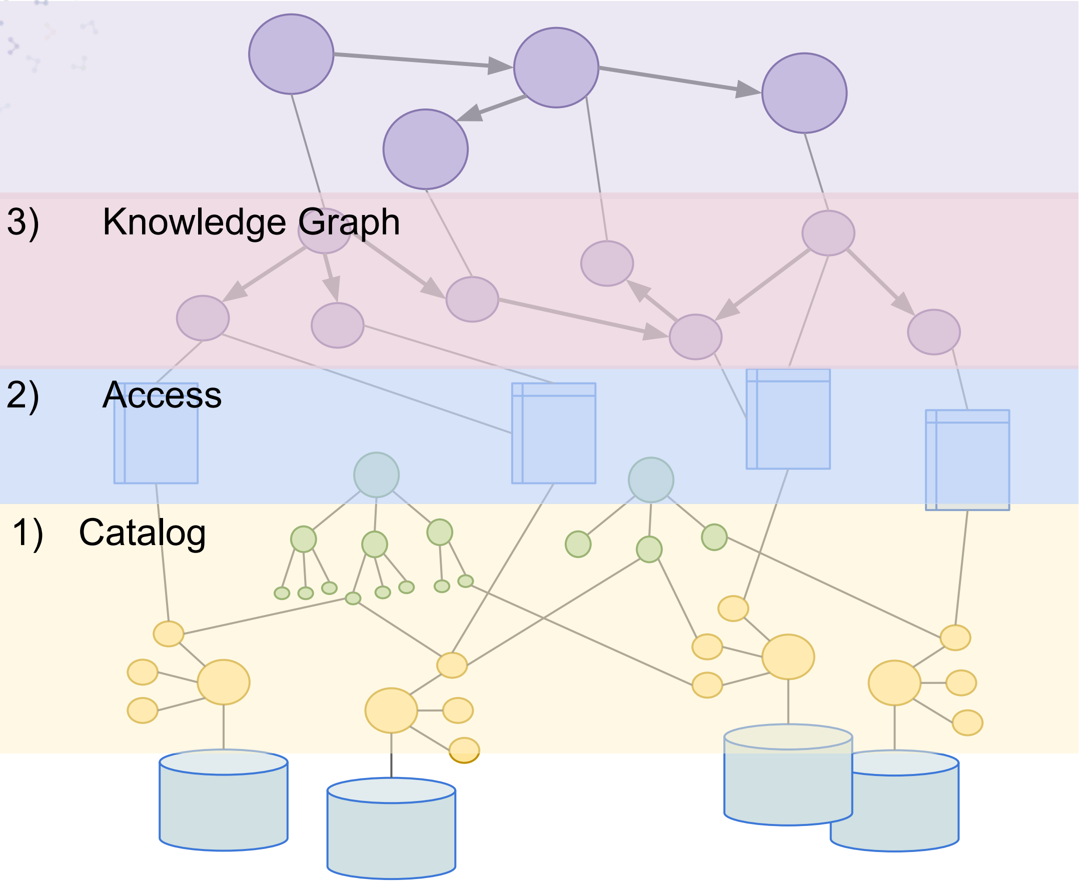


Let’s catalog the data!

How can I access the data?



Let’s access the data!

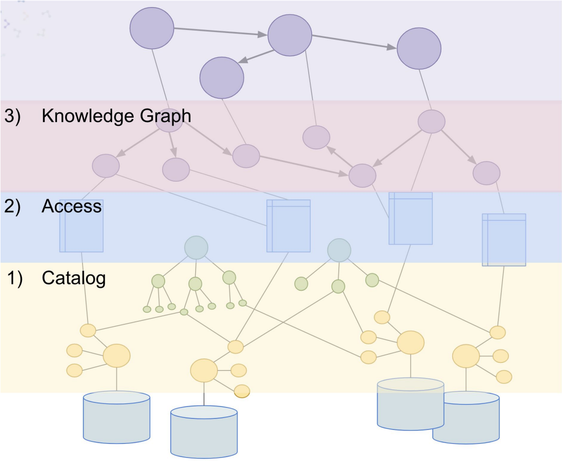


I don’t understand my data!

Let’s add knowledge to the data!

Data Cloud

A data cloud is where your data is available to your people and your machines – it's where your data assets are leveraged.



+

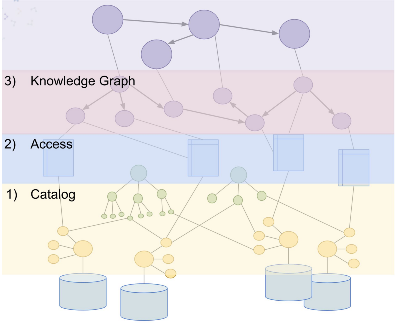
= **Data Cloud**

The 2019 LinkedIn Top Startups Are Growing Fast

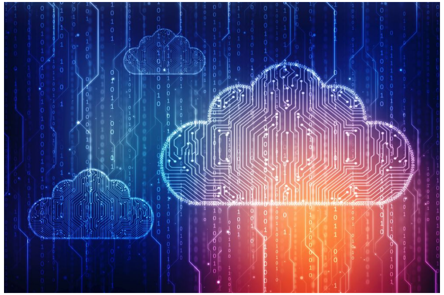
1) Snowflake
2) …

Hybrid Data Cloud

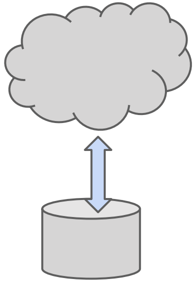
The data cloud will evolve to be hybrid, combining elements of your on-prem systems and systems native to public clouds.



+

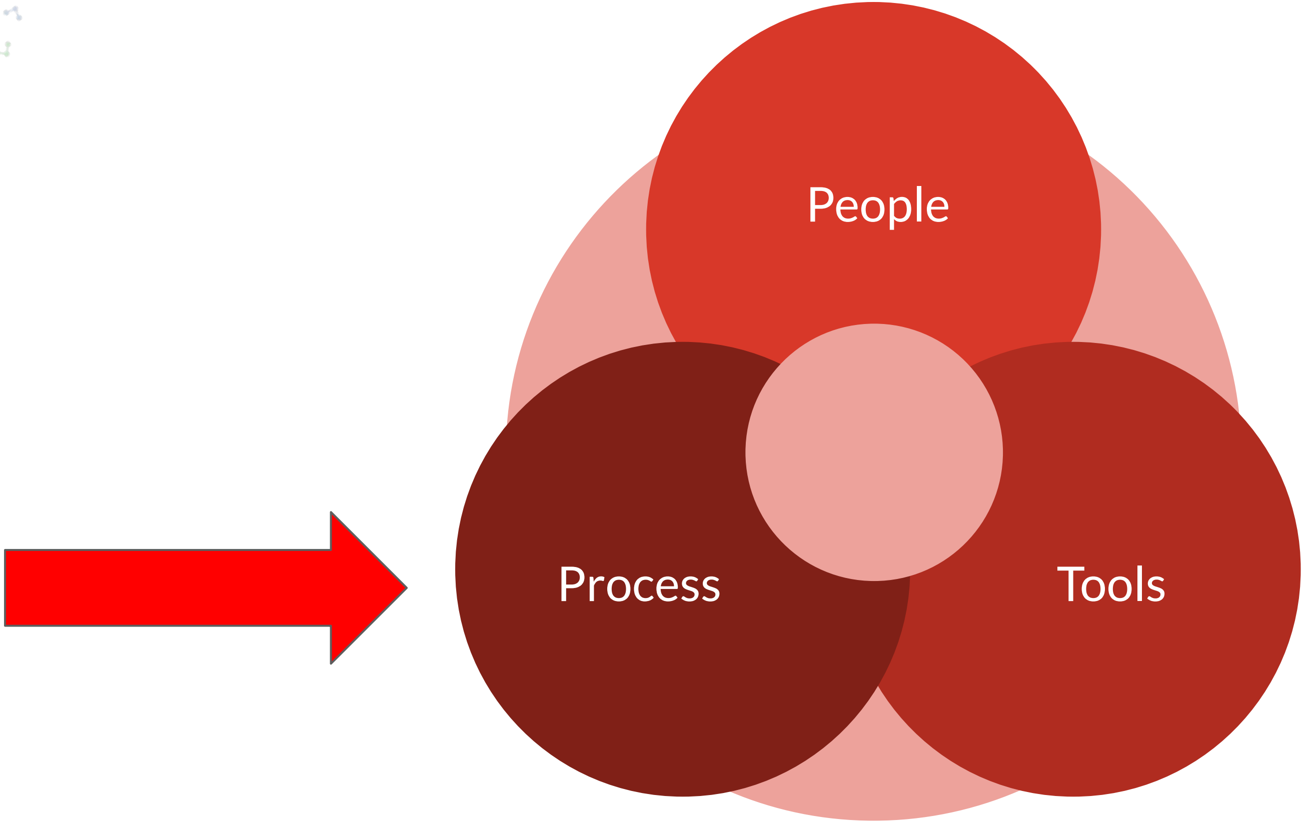


+



Knowledge Graph Virtualization

= **Hybrid Data Cloud**

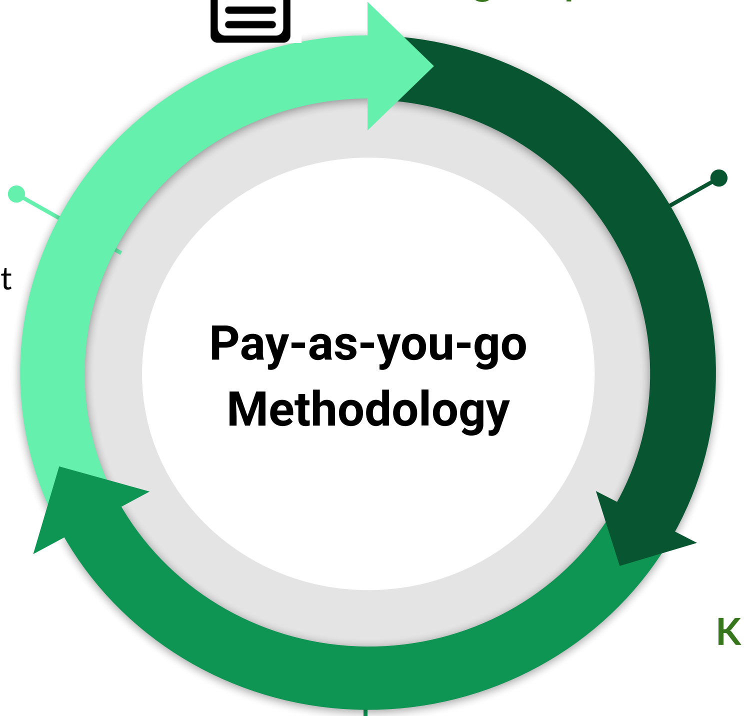

People
Process
Tools

Knowledge Report

Knowledge Capture

1. Analyze as-is process
2. Collect Documentation
3. Develop Knowledge Report

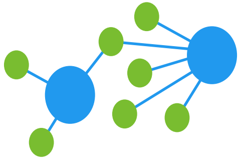
Business Question



Pay-as-you-go Methodology

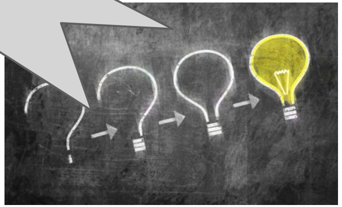
Knowledge Implementation

4. Create/Extend Ontology
5. Implement Mapping
6. Create Extract Queries
7. Validate Data



Enterprise Knowledge Graph

THIS IS HOW WE MEASURE SUCCESS!



Business Answer

Self Service Analytics

8. Build Report
9. Answer Business Question
10. Move to Production

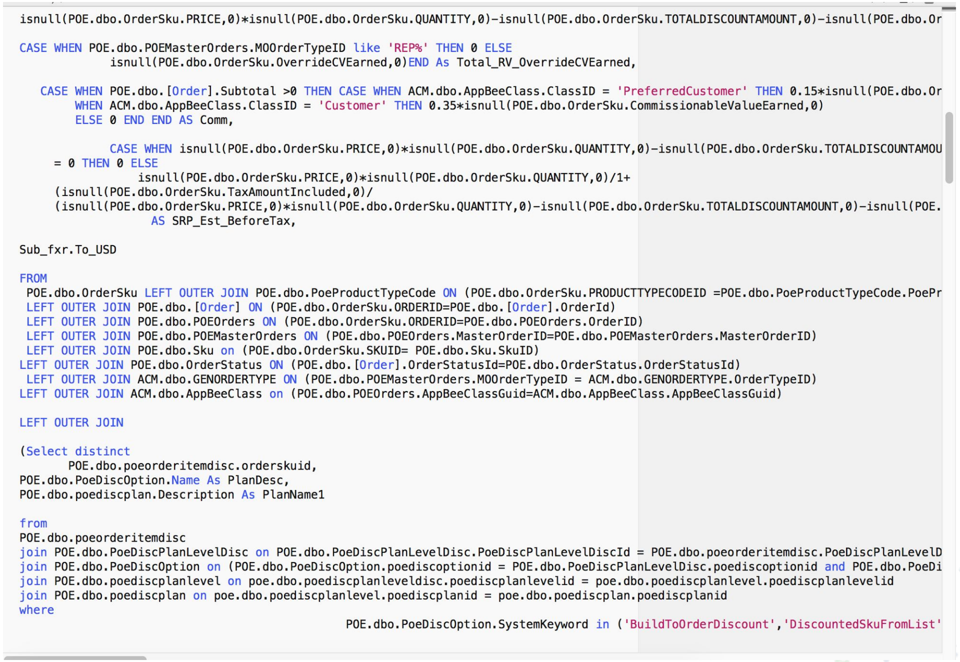
A Pay-as-you-go Methodology to Design and Build Enterprise Knowledge Graphs from Relational Databases. ISWC 2019

1) Analyze as-is Process and Workflow

WHAT is the business problem/question?	How many orders were placed in a given time period per their status?
WHY do we need to answer these questions?	Depending on whom is asked, different answers can be provided. Unaware of the source of the problem, the executives are vexed by inconsistencies across established business reports.
WHO produces/consumes the data?	The Finance department, specifically the CFO
HOW is this the business question answered today?	A business analyst asks the IT developer for this information every morning.
WHERE is the data?	There is a proprietary Order Management System and Oracle E-Business Suites.
WHEN will it be consumed?	Every morning they want to know this number

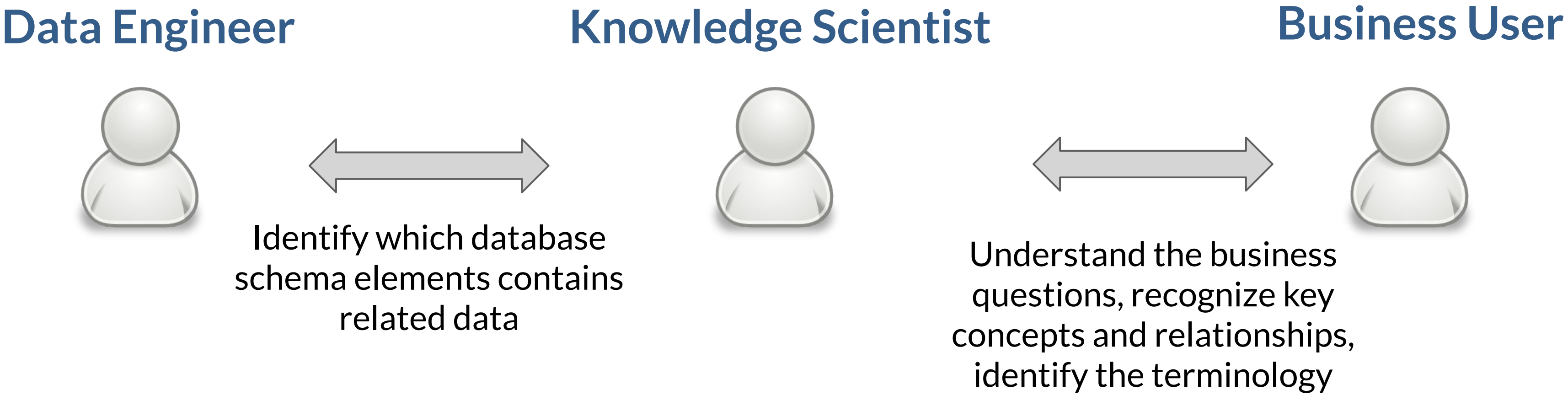
2) Collect Documentation

- Focus on HOW and WHAT
- Documentation, Wiki, SQL queries, Excel, ETL scripts, MS Access
- Interviews to understand the people and tech workflow
  - Who talks to who?
  - What is being shipped around?
  - Reverse engineer reports, queries, datasets



```
isnull(POE.dbo.OrderSku.PRICE,0)*isnull(POE.dbo.OrderSku.QUANTITY,0)-isnull(POE.dbo.OrderSku.TOTALDISCOUNTAMOUNT,0)-isnull(POE.dbo.Or

CASE WHEN POE.dbo.POEMasterOrders.MOOrderTypeID like 'REP%' THEN 0 ELSE
             isnull(POE.dbo.OrderSku.OverrideCVEarned,0)END As Total_RV_OverrideCVEarned,

   CASE WHEN POE.dbo.[Order].Subtotal >0 THEN CASE WHEN ACM.dbo.AppBeeClass.ClassID = 'PreferredCustomer' THEN 0.15*isnull(POE.dbo.Or
        WHEN ACM.dbo.AppBeeClass.ClassID = 'Customer' THEN 0.35*isnull(POE.dbo.OrderSku.CommissionableValueEarned,0)
        ELSE 0 END END AS Comm,

             CASE WHEN isnull(POE.dbo.OrderSku.PRICE,0)*isnull(POE.dbo.OrderSku.QUANTITY,0)-isnull(POE.dbo.OrderSku.TOTALDISCOUNTAMOU
     = 0 THEN 0 ELSE
                 isnull(POE.dbo.OrderSku.PRICE,0)*isnull(POE.dbo.OrderSku.QUANTITY,0)/1+
     (isnull(POE.dbo.OrderSku.TaxAmountIncluded,0)/
     (isnull(POE.dbo.OrderSku.PRICE,0)*isnull(POE.dbo.OrderSku.QUANTITY,0)-isnull(POE.dbo.OrderSku.TOTALDISCOUNTAMOUNT,0)-isnull(POE.
                  AS SRP_Est_BeforeTax,

Sub_fxr.To_USD

FROM
 POE.dbo.OrderSku LEFT OUTER JOIN POE.dbo.PoeProductTypeCode ON (POE.dbo.OrderSku.PRODUCTTYPECODEID =POE.dbo.PoeProductTypeCode.PoePr
 LEFT OUTER JOIN POE.dbo.[Order] ON (POE.dbo.OrderSku.ORDERID=POE.dbo.[Order].OrderId)
 LEFT OUTER JOIN POE.dbo.POEOrders ON (POE.dbo.OrderSku.ORDERID=POE.dbo.POEOrders.OrderID)
 LEFT OUTER JOIN POE.dbo.POEMasterOrders ON (POE.dbo.POEOrders.MasterOrderID=POE.dbo.POEMasterOrders.MasterOrderID)
 LEFT OUTER JOIN POE.dbo.Sku on (POE.dbo.OrderSku.SKUID= POE.dbo.Sku.SkuID)
LEFT OUTER JOIN POE.dbo.OrderStatus ON (POE.dbo.[Order].OrderStatusId=POE.dbo.OrderStatus.OrderStatusId)
 LEFT OUTER JOIN ACM.dbo.GENORDERTYPE ON (POE.dbo.POEMasterOrders.MOOrderTypeID = ACM.dbo.GENORDERTYPE.OrderTypeID)
LEFT OUTER JOIN ACM.dbo.AppBeeClass on (POE.dbo.POEOrders.AppBeeClassGuid=ACM.dbo.AppBeeClass.AppBeeClassGuid)

LEFT OUTER JOIN

(Select distinct
       POE.dbo.poeorderitemdisc.orderskuid,
POE.dbo.PoeDiscOption.Name As PlanDesc,
POE.dbo.poediscplan.Description As PlanName1

from
POE.dbo.poeorderitemdisc
join POE.dbo.PoeDiscPlanLevelDisc on POE.dbo.PoeDiscPlanLevelDisc.PoeDiscPlanLevelDiscId = POE.dbo.poeorderitemdisc.PoeDiscPlanLevelD
join POE.dbo.PoeDiscOption on (POE.dbo.PoeDiscOption.poediscoptionid = POE.dbo.PoeDiscPlanLevelDisc.poediscoptionid and POE.dbo.PoeDi
join POE.dbo.poediscplanlevel on poe.dbo.poediscplanleveldisc.poediscplanlevelid = poe.dbo.poediscplanlevel.poediscplanlevelid
join POE.dbo.poediscplan on poe.dbo.poediscplanlevel.poediscplanid = poe.dbo.poediscplan.poediscplanid
where
                                               POE.dbo.PoeDiscOption.SystemKeyword in ('BuildToOrderDiscount','DiscountedSkuFromList'
```

3) Knowledge Report



Data Engineer

Knowledge Scientist

Business User

Identify which database schema elements contains related data

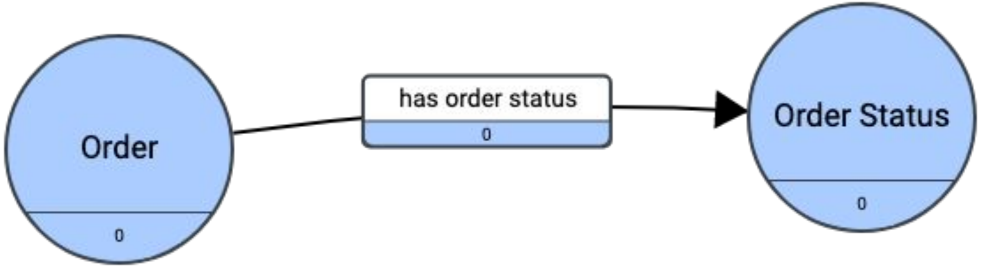
Understand the business questions, recognize key concepts and relationships, identify the terminology

Disagreements?
Focus on the business questions to drive consensus

Knowledge Report can be understood by everyone!

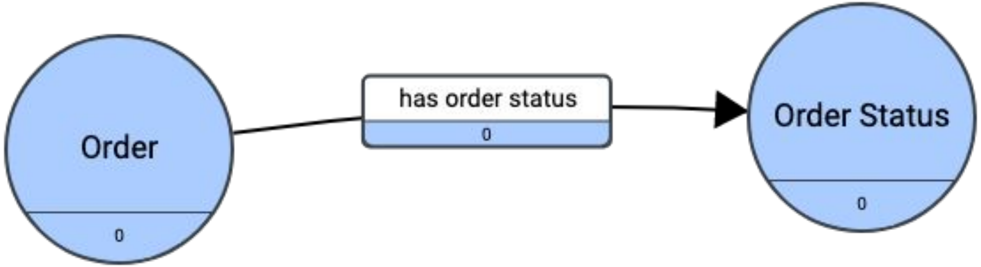
Knowledge Reports mimics the Intermediate Representations (IRs) from METHONTOLOGY

3) Knowledge Report: Concept



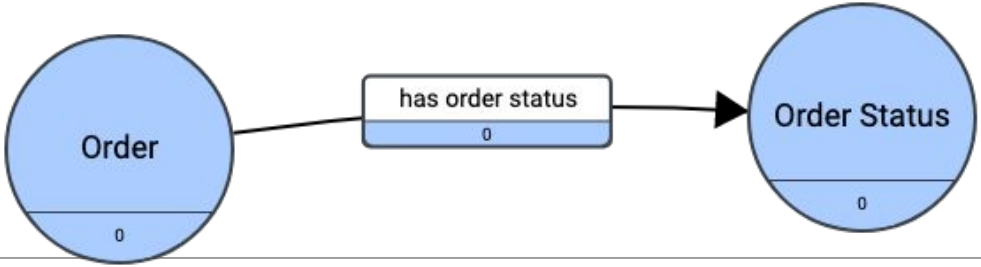
Concept Name	Order
Concept ID	Order
Concept Instance ID	`moid`
Table Name/Query	`SELECT moid FROM masterorder m JOIN order o on m.oid = o.oid WHERE ordertype in (2,3)`

3) Knowledge Report: Attribute



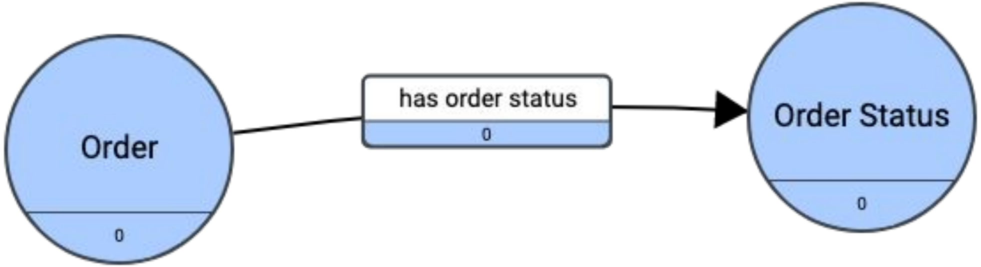
Attribute Name	Order Date
Attribute ID	orderDate
Applied to Concept	Order
Table Name/Query	`select moid, orderdate from masterorder m join order o where m.oid = o.id`
Column Name	`orderdate`
Datatype	`date`
Is NULL possible?	No
Cardinality	1:13) Knowledge Report: Concept

3) Knowledge Report: Relationship



Relationship Name	has order status
Relationship ID	hasOrderStatus
From Concept	Order
To Concept	Order Status
Table Name/Query	`select moid, ostid, max(orderstatusdate) from OrderStatus group by orderstatusdate`
Cardinality	1:1

3) Knowledge Report: Tabular Extract



```
SELECT ?Order_Number ?Order_Date ?Order_Status
WHERE {
 ?o a :Order;
     :orderNumber ?Order_Number;
     :orderDate ?Order_Date;
     :hasOrderStatus [
          :orderStatus ?Order_Status
     ].
}
```

Order Number	Order Date	Order Status

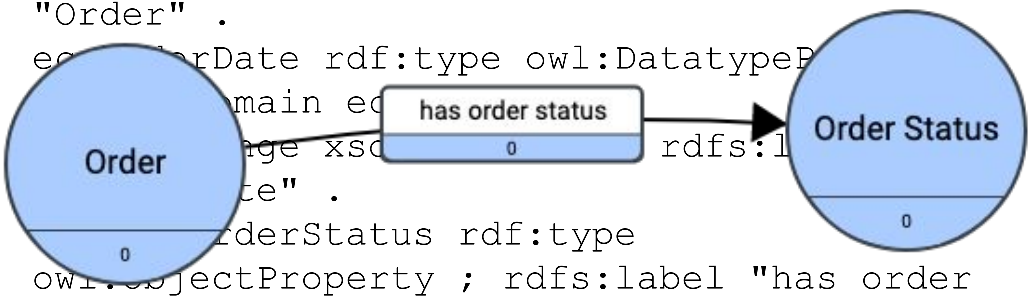
Knowledge Capture

Knowledge Implementation

4) Create/Extend Ontology

Knowledge
Report

```
ec:Order rdf:type owl:Class ; rdfs:label
"Order" .
ec      rderDate rdf:type owl:DatatypeP
             main ec
             nge xsd           rdfs:l
             te" .
            rderStatus rdf:type
ow    jectProperty ; rdfs:label "has order
status"
  rdfs:domain ec:Order ; rdfs:range
ec:OrderStatus .
```

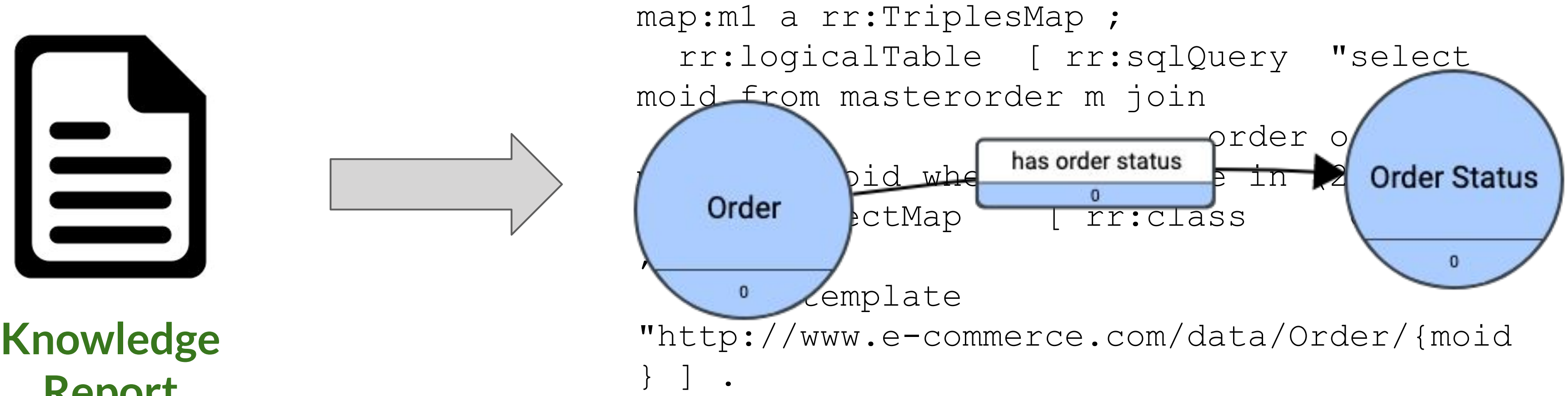


OWL Ontology *

Create the ontology using Gra.fo

* Property Graph Schema too

5) Implement Mapping



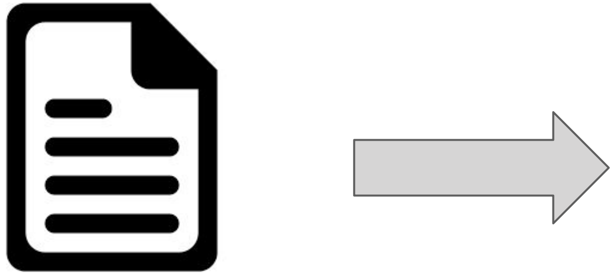
```
map:m1 a rr:TriplesMap ;
  rr:logicalTable  [ rr:sqlQuery  "select
moid from masterorder m join
                              order o
                oid whe           e in (2
                ectMap      [ rr:class
,
               template
"http://www.e-commerce.com/data/Order/{moid
} ] .
```

Knowledge Report

R2RML Mapping

Implement the mappings using Gra.fo

6) Extract/Tabular Queries



Knowledge Report

```
SELECT ?Order_Number ?Order_Date ?Order_Status
WHERE {
?x a :Order;
   :orderNumber ?Order_Number;
   :orderDate ?Order_Date;
   :hasOrderStatus [
      :orderStatusName ?Order_Status;
   ]
}
```

SPARQL Query

Execute the SPARQL query on Ultrawrap ETL/NoETL, which uses the R2RML mapping from the previous step.

7) Validate Data

- Counts: compare the number of results between extract and source
- Null: checking the validity of NULL values
- Duplicates: Check that expected Cardinality holds
- Sharing sample data to business users in a spreadsheet
- Creating sample visualizations in a BI tool

Knowledge Implementation

Self Service Analytics

8) Build Business Report

- **The ontology enables the simplified view but at the end, it is the tabular extract that the business user wants to access.**

Order Number	Order Date	Order Status

😊

```
SELECT * FROM Orders
```

☹️

```
SELECT m.moid as OrderNumber,
      o.orderdate as OrderDate,
      ost.statustype as OrderStatusName
FROM masterorder m JOIN order o ON m.oid = o.oid
JOIN (SELECT moid, ostid, max(orderstatusdate)
     FROM OrderStatus GROUP BY orderstatusdate) os
      ON m.moid = os.moid
JOIN OrderStatusType ost ON os.ostid = ostid.ostid
WHERE m.ordertype in (2,3)
```

9) Answer the Business Question

- The BI report should answer the original business question
    - The <u>What</u> in Step 1

- This report is shared with the stakeholders who asked the original business question
    - The <u>Who</u> in Step 1

- If they accept the BI report as an answer to their question, then this is ready to move to production. → SUCCESS

10) Move to Production

- Is the data live (virtual) or ETLed (materialized)?

- Common refresh schedules are daily, weekly, monthly or on demand.

- Determine extract time window:
    - Is the cache going to update the entire extract?
    - or is only yesterday’s data going to be pushed to the cache?
    - or last weeks?

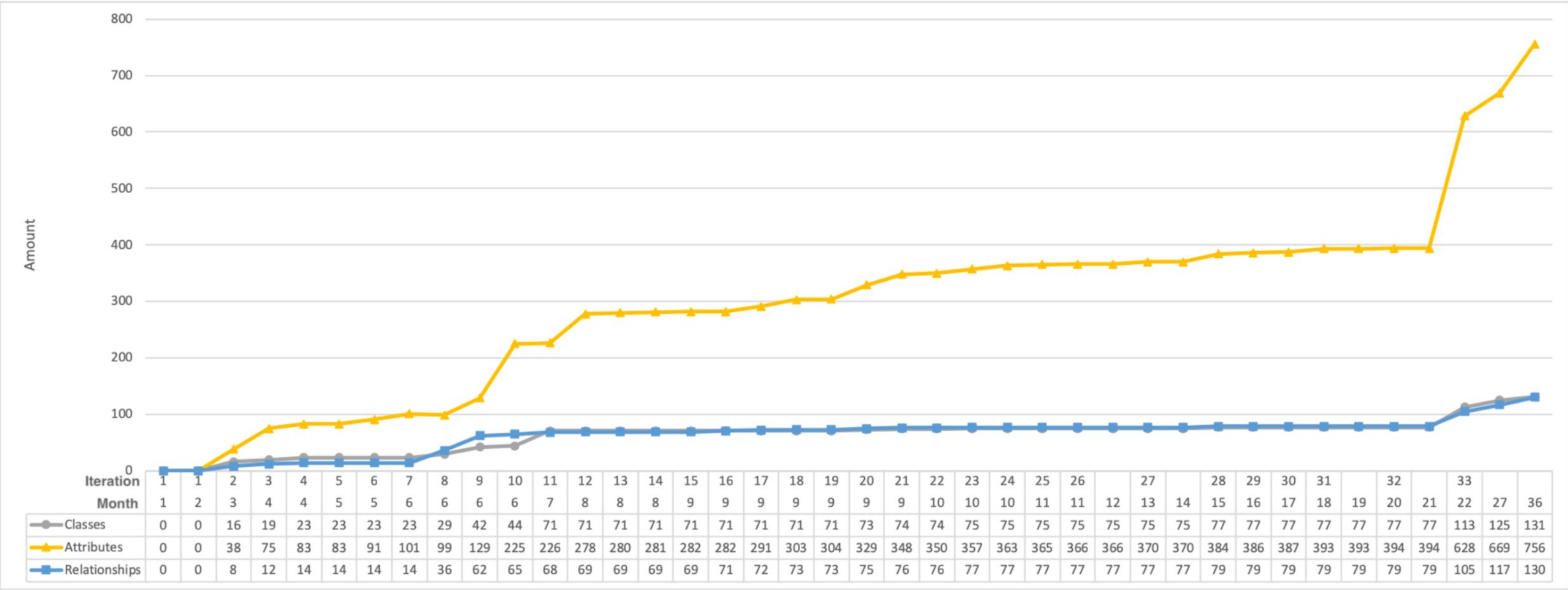
The E-Commerce Case Study Results

- Goal of First iteration: replicate most trusted BI report, the daily sales report that all C-level executives viewed every morning.

- 3 Business Users, 2 IT users and 1 Knowledge Scientist involved in 2 months.

- Current daily sales report was being generated by one SQL query that a business user would execute every morning.

- The Knowledge Capture phase revealed that the daily sales report encompassed 16 concepts, 38 attributes and 8 relationships. The customer was surprised to see how much knowledge was “hidden” within just one report.



Iteration	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26		27		28	29	30	31		32		33		
Month	1	2	3	4	4	5	5	6	6	6	6	7	8	8	8	9	9	9	9	9	9	9	10	10	10	11	11	12	13	14	15	16	17	18	19	20	21	22	27	36
Classes	0	0	16	19	23	23	23	23	29	42	44	71	71	71	71	71	71	71	71	71	73	74	74	75	75	75	75	75	75	75	77	77	77	77	77	77	77	113	125	131
Attributes	0	0	38	75	83	83	91	101	99	129	225	226	278	280	281	282	282	291	303	304	329	348	350	357	363	365	366	366	370	370	384	386	387	393	393	394	394	628	669	756
Relationships	0	0	8	12	14	14	14	14	36	62	65	68	69	69	69	69	71	72	73	73	75	76	76	77	77	77	77	77	77	77	79	79	79	79	79	79	79	105	117	130

- **Rapid Growth Phase (Month 1 - 7)**
- **Consistent Growth Phase (Month 8-22)**
- **Independent Growth Phase (Month 23-present)**

How can we best combine people and technology to improve data integration?

Interested? Hiring! Looking for research partners!

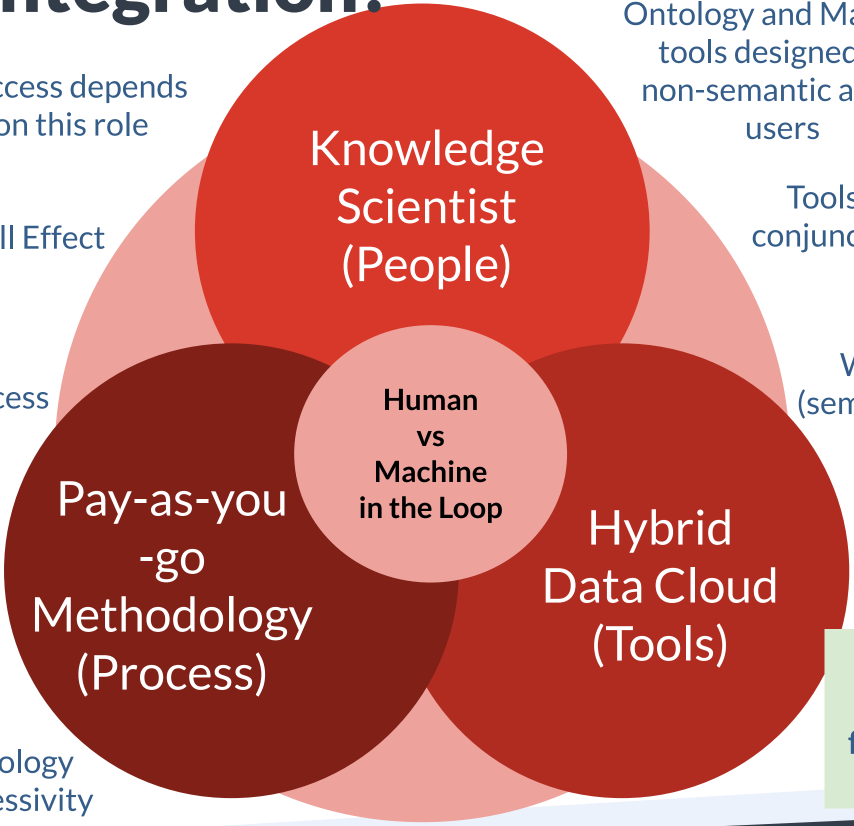
Success depends on this role

Snowball Effect

ual Process

Maintenance & Evolution

Ontology Expressivity


Knowledge Scientist (People)
Pay-as-you-go Methodology (Process)
Hybrid Data Cloud (Tools)
Human vs Machine in the Loop

Ontology and Mapping tools designed for non-semantic aware users

THANK YOU

Tools should be designed in conjunction with a methodology

What can be (semi)-automated?

- Dataset Search
- Nulls… so what?
- Data Quality & Mappings
- SQL Query Log Analysis
- Learn Mappings from Source Queries

juan@data.world
@juansequeda

data.world