

HUDK 4050: CORE METHODS IN EDM

Today

- Data sources
- Zotero
- Download some data

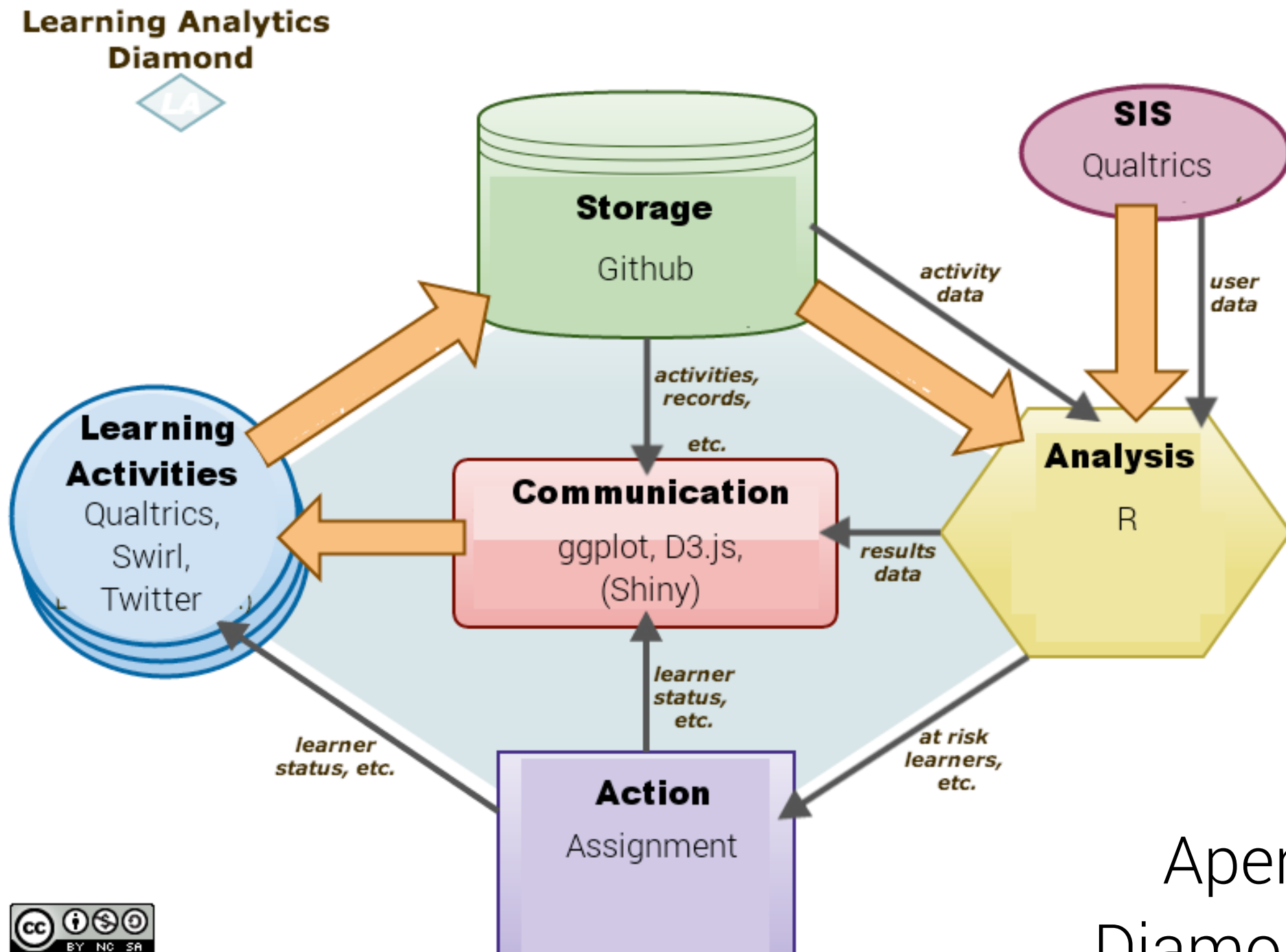
I Need Your Github Username

Yixiao	Li
Runkun	Han
Ruin	Wang
Wanruo	Zhang
Jingze	Dai
Mengjie	Xu

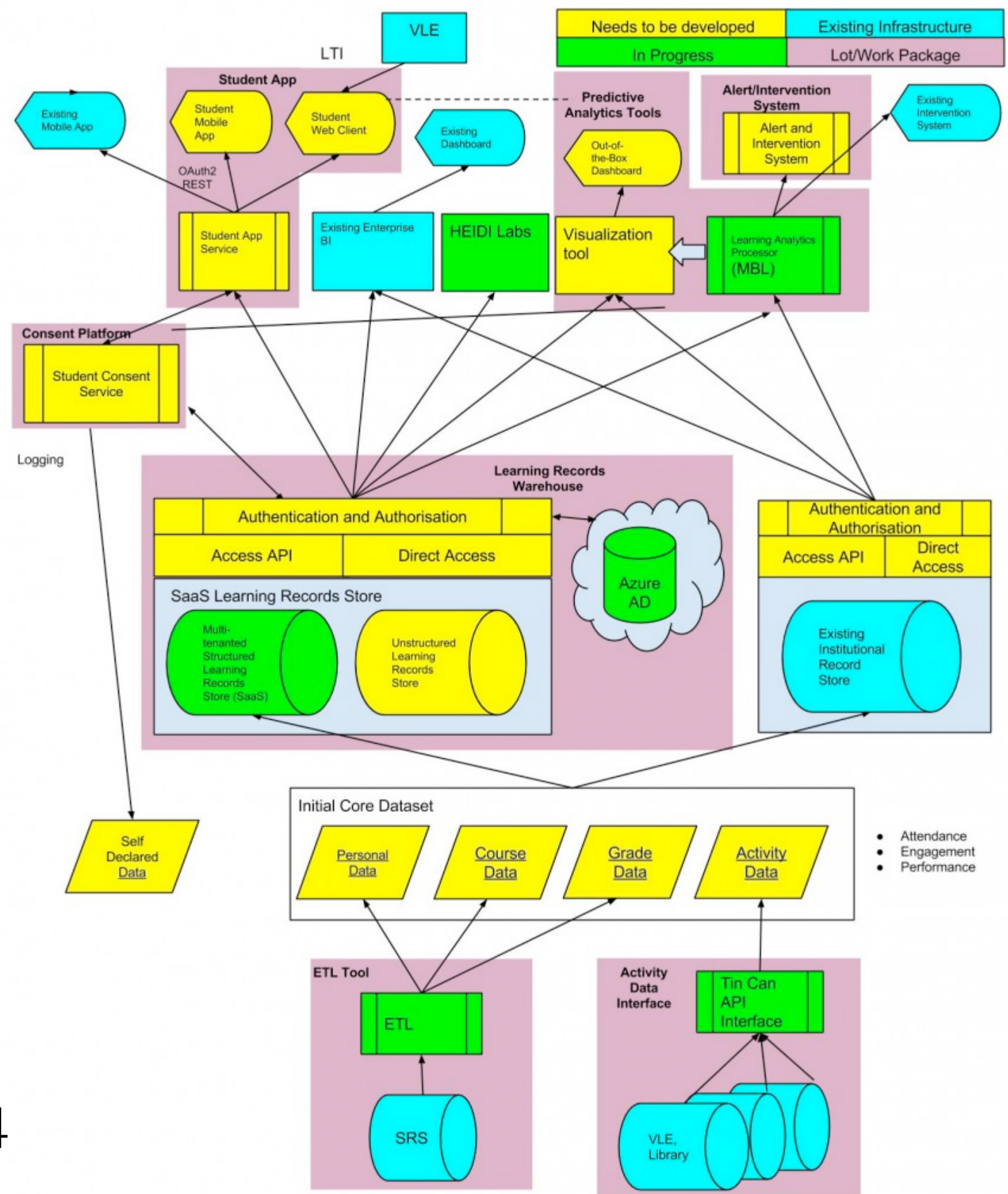
“We are what we measure.”

–Paolo Blikstein, 2013

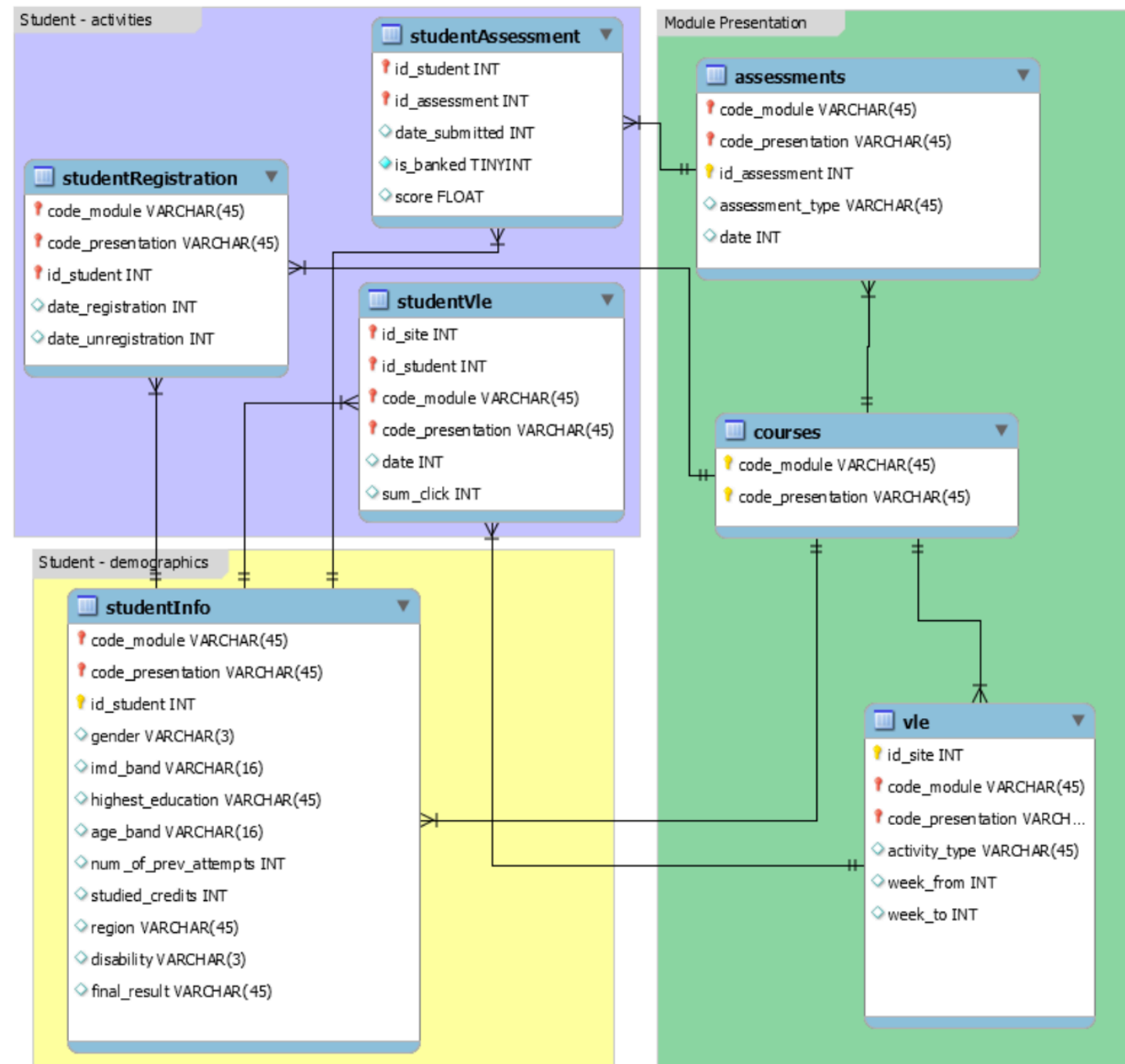
Where does data come from?



JISC Learning Architecture, 2014



Real World Example



https://analyse.kmi.open.ac.uk/open_dataset

Standardizing Data Communication

- Limited vocabulary for describing learning
- So that LMS can communicate with tools with SIS, etc
- Scorm = limit information
- Tin Can = limit syntax
- *All this comes from the US Defense Dept?*



I call it **Tin Can** | **xAPI**.

What is BIG Data?

- It is relative
- Process vs Study
- Depends on the domain of study: ed (MB-GB) vs ed tech (TB) vs astrophysics (PB) vs business (EB)

Common File Formats

DBF:

- Database format
- Microsoft Access, some freeware
- Table

Common File Formats

XML:

- Semantic Web
- Extensible Markup Language
- Export web page data
- Hierarchy like HTML with tags to delimit

```
<row>
```

```
  <Year>2016</Year>
```

```
  <Course>EDCTGE2550</Course>
```

```
  <Price>Priceless</Price>
```

```
</row>
```

Common File Formats

JSON:

- JavaScript Object Notation
- Similar to XML, most common server-browser format

```
{  
  {10, 12, 15, 100},  
  {100, 200, 150, 500},  
  {9, 8, 8, 7},  
};
```

Common File Formats

Fixed Width:

- Create a grid with text using spaces

```
Year.....Course.....Price.....  
2016.....EDCTGE2550....Priceless.....
```

Common File Formats

CSV (TSV):

- Comma Separated Value (Tab Separated Value)
- Most common data format
- Lightweight, easy to interpret - but you can run into trouble with text

```
Year,Course,Price  
2016,HUDK4050,Priceless
```

Yeah, but where do WE get data?

Open Data Sets

- *Government*: NYC (<https://nycopendata.socrata.com/>), Whitehouse (<https://open.whitehouse.gov/>), UK (<https://data.gov.uk/>)
- *Research Labs*: ASSISTments (<https://sites.google.com/site/assistmentsdata/>), PSLC DataShop (<https://pslcdatashop.web.cmu.edu/>)
- *Private release*: Harvard/MIT MOOC Data (<https://dataverse.harvard.edu/dataverse/mxhx>)
- LearnSphere (<https://learnsphere.org>)

Yeah, but where do WE get data?

Cut a Deal

- Happens often but will lose autonomy/\$\$\$/control of results

Automated acquisition

- APIs, web scraping, beacons

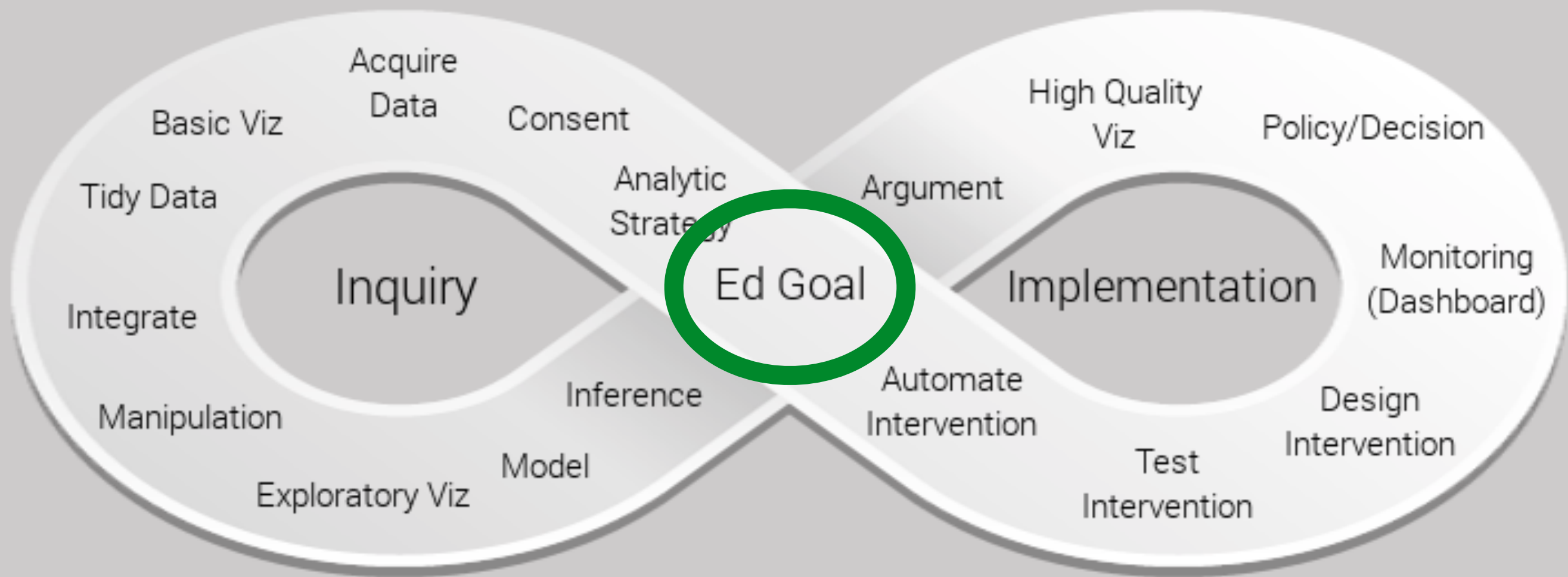
Generate

- Make your own!

Let's Practice!

- Go to: https://analyse.kmi.open.ac.uk/open_dataset
- Explore the codebook
- Download the dataset
- Open a new project in RStudio
- Uncompress the file into the new folder for the RStudio Project
- File -> New File -> RMarkdown
- Write the code to load your data into R and hit run
- How many students are there in the file?
- What is the average for all assessments
- Can you visualize one of the variables?

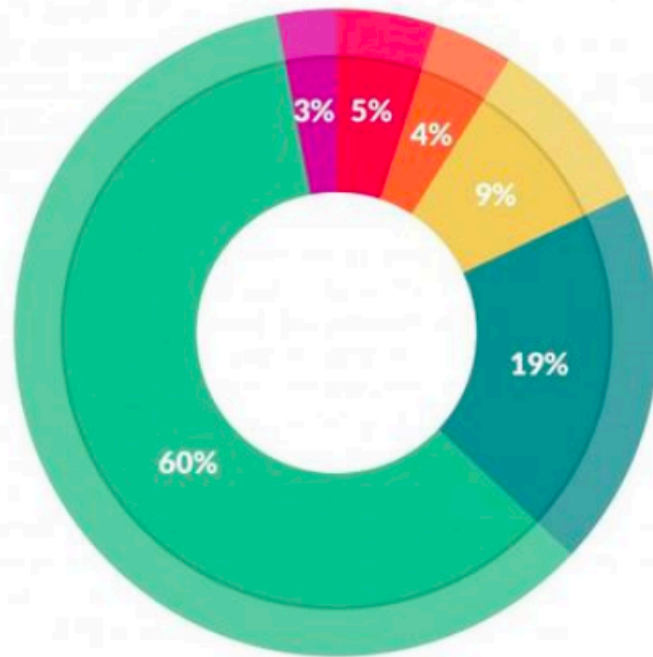
Ed Data Science Cycle



The Ideal Model

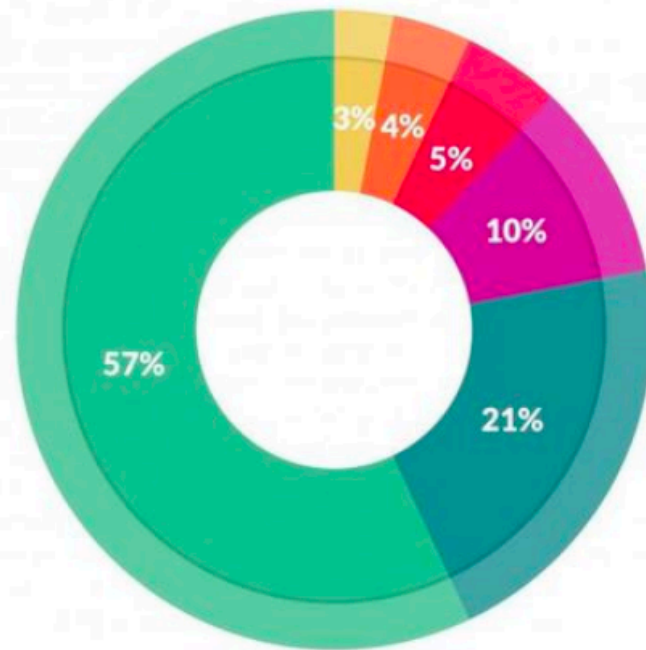
Data → Insight → Automate

Reality 1



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Reality 2

The ideal model:

Data → Insight → Automate

VERY RARELY HAPPENS



Workflow

Workflow

Data → Process → Knowledge → Action

What am I
counting?

How do I make
meaning from count?

What information does
the process give me?

How do I respond
to the information?



Count raised
hands



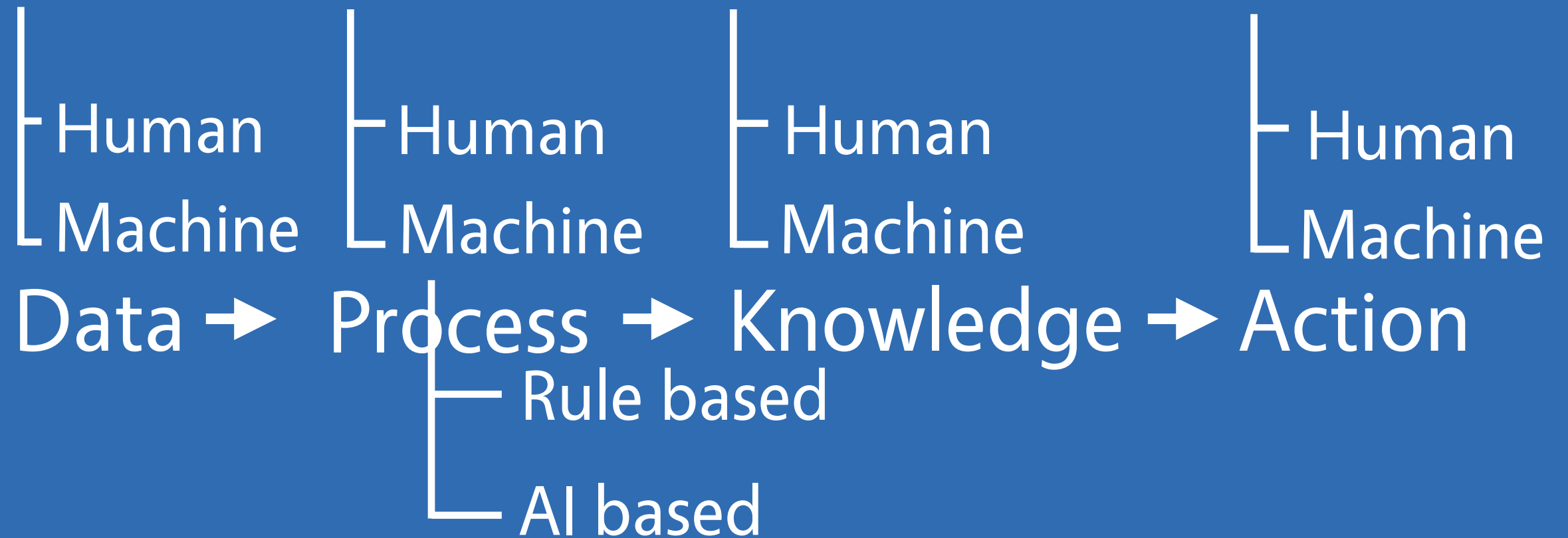
Is that enough?



Yes

No





What does this look
like in education?

Workflow

Data → Process → Knowledge → Action



Zotero

- Install Firefox & Zotero (Assignment 1)
- Clone the assignment1 repository to a new project in RStudio
- Open Zotero
- File -> Import -> Navigate to your assignment1 project folder and import the hudk4050-references.rdf file
- Click on “Siemens & Baker” in the list of refs
- Add a note within under the notes tab on the top right
- Right (cmd) click on the folder
- Choose “Export Collection...”
- Choose CSV from the drop down menu and choose your assignment1 project as the location
- Open R and type the code:

```
DF <- read.csv("hudk4050-references.csv", header = TRUE)
```
- Behold! Your bibliography including your notes (under the “notes” column)