# 02 | Overview of The Data Science Process


Cynthia Rudin | MIT Sloan School of Management

# Module Overview

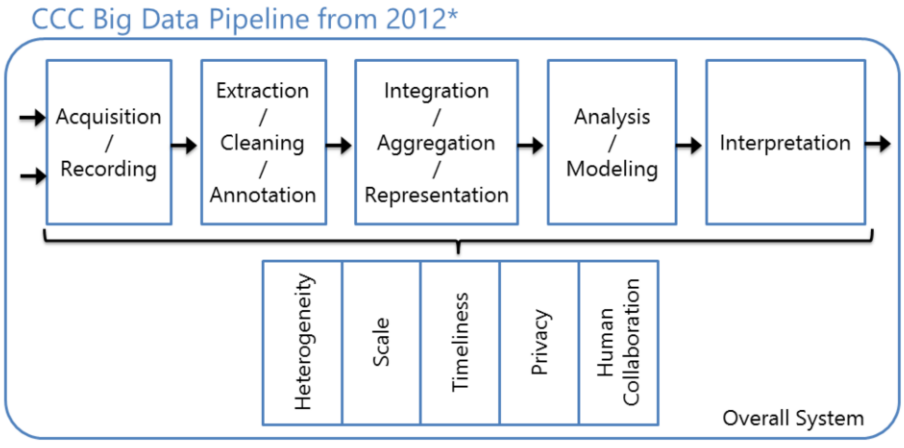
- Historical Notes on KDD, CRISP-DM, Big Data and Data Science and their relationship to Data Mining and Machine Learning
- Example of the knowledge discovery process

# Historical Notes on KDD, CRISP-DM, Big Data and Data Science and their relationship to Data Mining and Machine Learning


Cynthia Rudin | MIT Sloan School of Management

# Historical Notes

Term "Big Data" coined by astronomers Cox and Ellsworth in 1997




*From the Computing Community Consortium Big Data Whitepaper: http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

Let's start with historical notes. *Read* So it's actually kind of an old term that got a bit hyped up a few years ago after the CCC whitepapers on Big Data came out. I'm showing you the CCC big data pipeline from the whitepaper in 2012. It's nice, it's an effort to formalize the scientific process that goes into discovering knowledge from data. It's not just data mining, there's a lot more to it. In fact, data mining is the 4th step out of 5. Major steps in analysis of big data are shown in the flow at the top. Below it are big data needs that make these tasks challenging. Let me go through the 5 steps.

First you have to go through a process of actually obtaining the data. Perhaps that involves recording how users behave on different websites, so you have to write a script to collect the data properly.

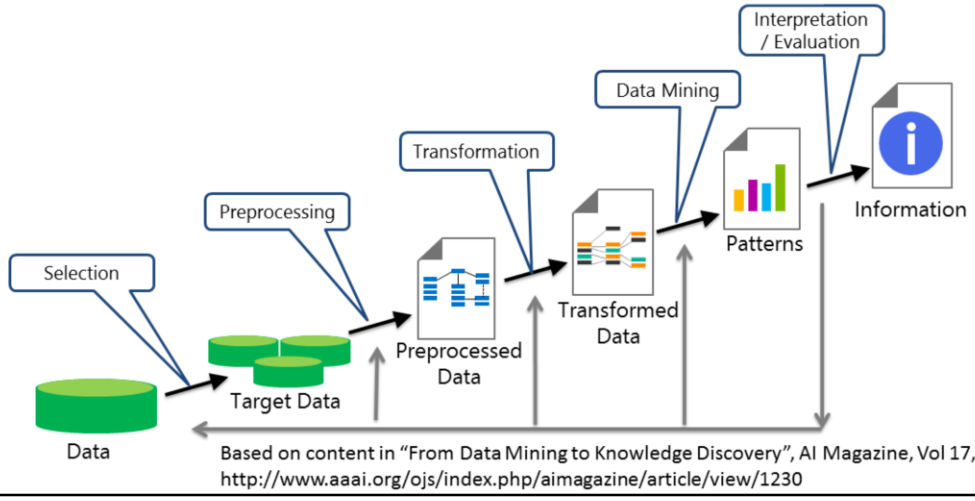
The second step is extraction/cleaning/annotation, where you try to reduce the noise a bit and get rid of the data you don't directly need. Or if you're working with free text that's easy to annotate, you might be able to turn it into a structured table. Basically at this step, you're turning the data from what is potentially a pile of rubbish into something you might actually be able to work with.

At the third step you are going to set up the data in a way that is directly conducive to data mining. All the text documents are linke properly to the other structured tables, and you have one central database where the information is stored neatly.

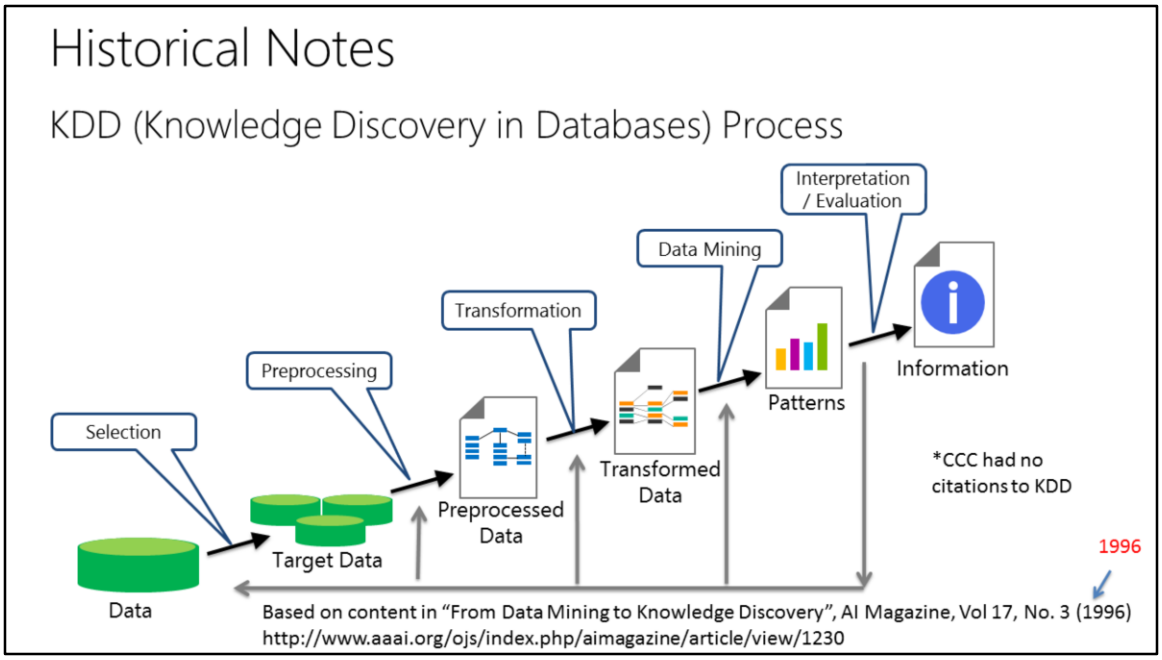
:
Ok so that’s the big data pipeline. Now I’m going to show you another, separate big data pipeline, constructed by a separate group of people, at another time. So this is a second group of people trying to formalize the discovery of knowledge from data. Ready?

# Historical Notes

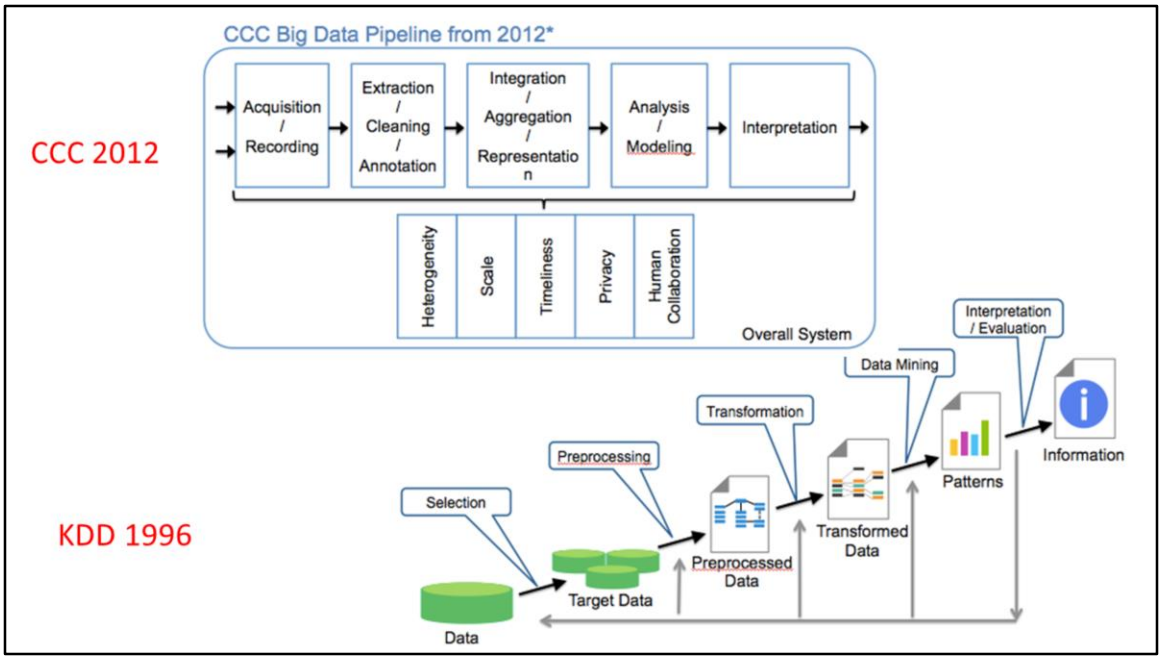
## KDD (Knowledge Discovery in Databases) Process




selection, preprocessing, transformation.
Does it look familiar? It should, because it's the same as on the previous slide! Same exact steps. So you're probably not surprised, so what, two groups came up with the same diagram. But look. This diagram is from 1996.



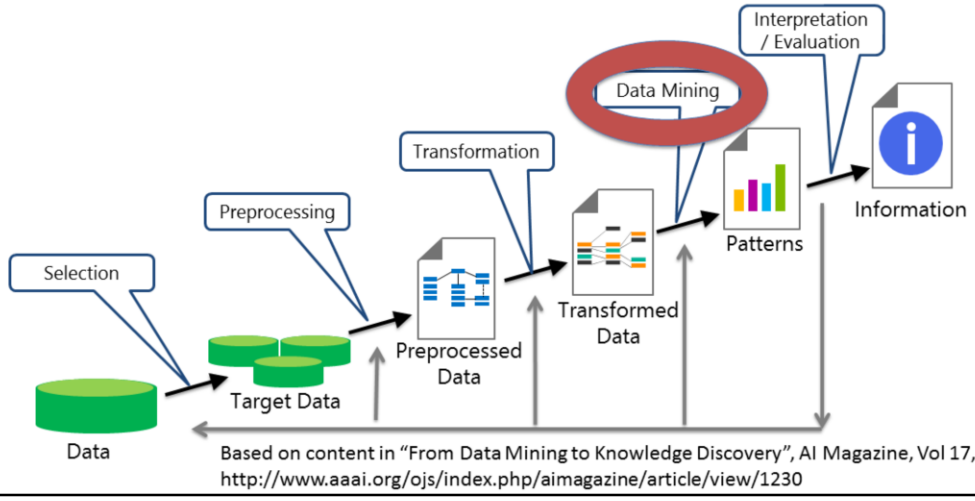

Way before the CCC had "big data" as a twinkling in their eyes. And the CCC people did not cite the original KDD paper – they didn't know about each other. Perhaps there is something fundmental here about this process. Maybe it's like an important number, like pi or the golden ratio or something. This sort of universal process that people separately discover. Anyway, so this is the kdd process.




It seems so abstract but there may be something fundamental about this.
BTW reading both of these articles is really inspiring. One thing I want to point out again is that data mining is a part of the process.

# Historical Notes

## KDD (Knowledge Discovery in Databases) Process

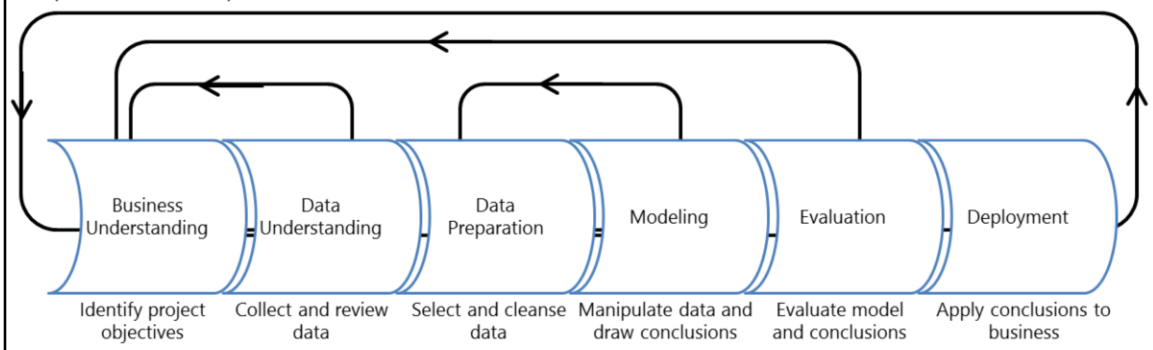



One part, not all of it. Granted, it's like the climax of a story, but like any story, the set up of the story really makes the whole thing work. *pause*

# Historical Notes

From 2000, 77 pages

CRoss Industry Standard Process for Data Mining (CRISP-DM)




Now, KDD wasn't the only game in town before the CCC whitepapers. This is CRISP-DM – Cross.... These guys were really serious about formalizing the process. They wrote a 77 page manual on how to do this stuff. I actually like CRISP-DMs process the best, because it includes business understanding and data understanding. I also love all the backwards arrows showing how many iterations you do within this process on a regular basis. Anyway, this is the process I tend to think of.

## Historical Notes

- The stages are basically the same no matter who invents or reinvents the (knowledge discovery / data mining / big data / data science) process. You may not always need all the stages.
- Data science is an iterative process.
  - Backwards arrows on most process diagrams.

Just to summarize

# Example of the knowledge discovery process

Cynthia Rudin | MIT Sloan School of Management

## Knowledge Discovery Process Example

- I'll walk you through the knowledge discovery process with an example – the process of predicting power failures in Manhattan.

Just to summarize

# Motivation for Example

- In NYC the peak demand for electricity is rising.
- The infrastructure dates back to the 1880's from the time of Thomas Edison.
- Power failures occur fairly often (enough to do statistics) and are expensive to repair
- We want to determine how to prioritize manhole inspections in order to reduce the number of manhole events (fires, explosions, outages) in the future.
- This is a real problem.

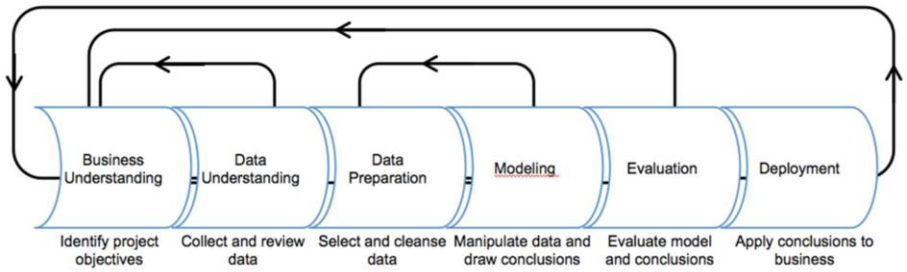
## Stages in the knowledge discovery process

- Opportunity Assessment & Business Understanding
- Data Understanding & Data Acquisition
- Data Cleaning and Transformation
- Model Building
- Policy Construction
- Evaluation, Residuals and Metrics
- Model Deployment, Monitoring, Model Updates

these are essentially taken from CRISP-DM.

# Stages in the knowledge discovery process

- Opportunity Assessment & Business Understanding
- Data Understanding & Data Acquisition
- Data Preparation, including Cleaning and Transformation
- Model Building
- Policy Construction
- Evaluation, Residuals and Metrics
- Model Deployment, Monitoring, Model Updates




See CRISP-DM for the 77 page description

## Opportunity Assessment & Business Understanding

What do you really want to accomplish and what are the constraints? What are the risks? How will you evaluate the quality of the results?

- For manhole events the general goal was to "predict manhole fires and explosions before they occur." We made it more precise:
  - Goal 1: Assess predictive accuracy for predicting manhole events in the year before they happen.
  - Goal 2: Create a cost-benefit analysis for inspection policies that takes into account the cost of inspections and manhole fires. Determine how often manholes need to be inspected.

*read* These goals are concrete – we can form concrete assessments of how well we did with these tasks.

## Data Understanding & Data Acquisition

Data were:

- Trouble tickets – free text documents typed by dispatchers documenting problems on the electrical grid.
- Records of information about manholes
- Records of information about underground cables
- Electrical shock information tables
- Extra information about serious events
- Inspection reports
- Vented cover data

Cables were not matched to manholes

## Data Understanding & Data Acquisition

Data were:

- Trouble tickets – free text documents typed by dispatchers documenting problems on the electrical grid.
- Records of information about manholes
- Records of information about underground cables
- Electrical shock information tables
- Extra information about serious events
- Inspection reports
- Vented cover data

V's of Big Data include "Variety" and "Veracity"

Cables were not matched to manholes

## Data Understanding & Data Acquisition

Data were:

- Trouble tickets – free text documents typed by dispatchers documenting problems on the electrical grid.
- Records of information about manholes
- Records of information about underground cables
- Electrical shock information tables
- Extra information about serious events
- Inspection reports
- Vented cover data

How do you know what data to trust?

You don't. Over half of our information was useless. We didn't know which half. This ties into the evaluation – if you don't know which data to trust, how are you going to evaluate the quality of the results? Those are the kinds of problems we were stuck on. It wasn't the ML that was the hard part here. Most of the time the ML isn't the hard part. If you're taking this class, probably ML is maybe a little intimidating, but that's not the part you should be intimidated by. Hopefully this class will help you in that respect.

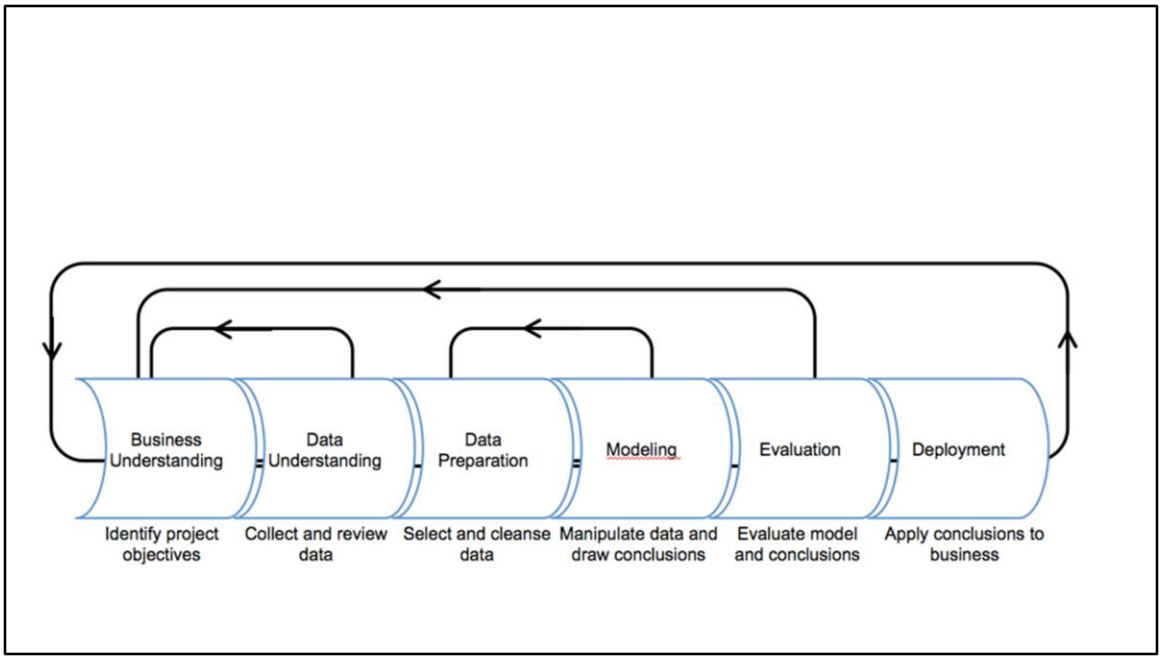
# Data Cleaning and Transformation

- Sometimes 99% of the work

# Data Cleaning and Transformation

- Turn free text into structured information:
  - Trouble tickets turned into a vector like:
    - Serious / Less Serious / Not an Event
    - Year
    - Month
    - Day
    - Manholes involved
    - ...
- Try to integrate tables (create unique identifiers):
  - If you join manholes to cables, half of the cable records disappear

Before this step you might have a giant pile of garbage.




I like to think of the knowledge discovery process sort of like alchemy. Alchemy is where you try to turn lead into gold. Except that alchemy doesn't work, but data often does!

Business understanding is like – am I really going to get gold out of this? What constitutes gold here

Data understanding is like – what are my raw materials that I'm going to collect

Data prep is an important step – that's where you purify all the raw ingredients and get rid of all the junk and contaminants.

Modeling is where you actually do the transformation of the raw ingredients into gold – that's why people are obsessed with it, and why other people think it's magic, but really, as long as you did the processing right, this step is generally pretty easy. And it does work like magic. But you'll see it's definitely not magic!

Evaluation is where the appraiser comes and tells you how much the gold is worth, and of course deployment is where the whole village comes to buy your gold. Back to the modeling step.

# Model Building

- Often predictive modeling, meaning machine learning or statistical modeling
- If you want to answer a yes/no question, this is classification.
  - For manholes, will the manhole explode next year? Y/N
- If you want to predict a numerical value, this is regression.
- If you want to group observations into similar-looking groups, this is clustering.
- If you want to recommend someone an item (e.g., book/movie/product) based on ratings data from customers, this is a recommender system.
- Note: There are many other machine learning problems.

Predictive doesn't necessarily mean in the future time-wise. It could mean prediction of a new circumstance you haven't seen before.

## Policy Construction

How will your model be used to change policy?

- E.g., for manholes, how should we recommend changing the inspection policy based on our model?
- E.g., consider using social media and customer purchase data to determine customer participation if Starbucks moves into New City. Once the model is created, how to optimize where the shops are located, how big they are, and where the warehouses are located.

Model building is predictive, Policy Construction is prescriptive.

## Evaluation

How do you measure the quality of the result?
Evaluation can be difficult if the data do not provide ground truth.

- For manhole events, we had engineers at Con Edison withhold high quality recent data and conduct a blind test.

That's how they knew they could trust us.

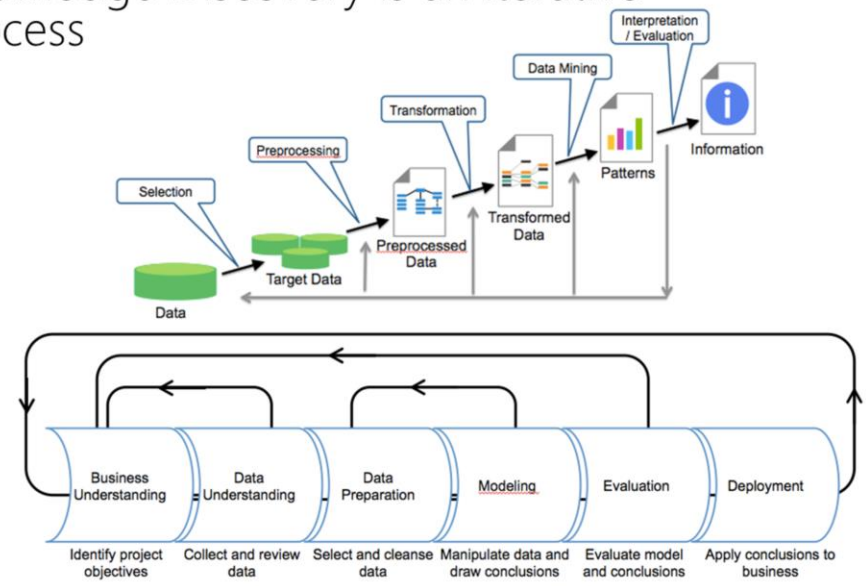
## Deployment

- Getting a working proof of concept deployed stops 95% percent of projects.
- Don't bother doing the project in the first place if no one plans to deploy it.*
- Keep a realistic timeline in mind. Then add several months.
- While the model is deployed it will need to be updated and improved.

* Unless it's fun.

It's like "Hey I turned lead into gold" and the reaction is "well, that's nice"
Goals might change, data itself might change, data might change as a reaction to the new policy, you observe something that you didn't expect and need to put it in, etc.

# Knowledge Discovery is an Iterative Process



## Summarize

- Several attempts to make the process of discovering knowledge scientific
  - KDD, CRISP-DM,CCC Big Data Pipeline
- All have very similar steps
  - Data Mining is only one of those steps (but an important one)