UNIT-II

1) Exploratory Data Analysis:

Exploratory data analysis (EDA) is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there.

EDA is the first step towards building a model. It's traditionally presented as a bunch of histograms & stem & leaf plots. But EDA is a critical part of the data science process, & also represents a philosophy & way of doing stat's practiced by a strain of statisticians coming from the Bell Labs tradition.

John Tukey, a mathematician at Bell labs, developed exploratory data analysis in contrast to confirmatory data analysis, which concerns itself with modeling & hypotheses as described in the previous section. In EDA, there is no hypothesis & there is no model. The "exploratory" aspect means that u understanding of the problem u r solving, & might solve, is changing as u go.

The basic tools of EDA r plots, graphs & summary stat's. Generally speaking, it's a method of systematically going thru the data, plotting distributions of all var's (using box plots), plotting time series of data, transforming var's, looking at all pairwise relationships blw var's using scatterplot matrices & generating summary stat's for all of the them. At the very least that would mean computing their mean, min, max, the upper & lower quartiles, & identifying outliers.

But as much as EDA is a set of tools, it's also a mindset. And that mindset is abt ur relationship with the data. u want to understand the data - gain intuition, understand the shape of it, & try to connect ur understanding of the process that generated the data to data itself. EDA happens blw u & the data & isn't abt providing anything to anyone else.

In stat's, EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used & not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling / hypothesis testing task.

EDA is different from Initial data analysis, which focuses more narrowly on checking assumptions required for model fitting & hypothesis testing.

2) Philosophy of Exploratory Data Analysis:

"Long before worrying abt how to convince others, u first have to understand what's happening ur'self"
- Andrew Gelman.

While at Google, Rachel was fortunate to work alongside 2 former Bell labs | AT&T statisticians - Daryl pregibon & Diane Lambert, who also work in this vein of applied stat's - and learned from them to make EDA a part of her best practices.

Yes, even with very large scale data, they did EDA. In the world of data in an Internet-Legg Company, EDA is done for some the same reasons it's done with smaller datasets, but there are additional reasons to do it with data that has been generated from logs.

There r imp reasons anyone working with data should do EDA. Namely, to gain intuition abt the data; to compare sons bt distributions; for sanity checking (making sure the data is of the scale u expect, in the format u thought it should be); to find out where data is missing | if there r outliers; & to summarize the data.

In the context of data generated from logs, EDA also helps with debugging the logging process. For eg, "Patterns" u find in the data could actually be something wrong in the logging process that needs to be fixed. If u never go to the trouble of debugging, u'll continue to think ur Patterns r real. The ~~eggs~~ engineers we've worked with r always grateful for help in this area.

In the end, EDA helps u make sure the product is performing as intended.

Although there's lots of visualization involved in EDA, we distinguished bt EDA & data visualization in that EDA is done toward the beginning of analysis, & data visualization, as it's used in our vernacular, is done towards the end to communicate one's findings. With EDA, the graphics are solely done for u to understand what's going on.

With EDA, u can also use the understanding u get to inform & improve the development of alg's. For eg, Suppose u r trying to develop a ranking alg that ranks content that u r showing to users. To do this u might want to develop a

Before u decide how to measure Quantify Popularity (which could be, for eg, highest frequency of clicks, or the post with most no of comments, or comments above some threshold, or some weighted avg of many metrics), u need to understand how the data is behaving, & the best way to do that is looking at it & getting ur hands dirty.

Plotting data & making comparisons can get u extremely far, & is far better to do than getting dataset & immediately running a regression just bcoz u know how. It's been observed to analysts & data scientists that EDA has not been enforced as a critical part of the process of working with data. Take this opportunity to make it part of your process.

2. The Data Science process:-

Let's put it all together into what we define as the data science process. The more eg's u see of people doing data science, the more you'll find that they fit into the general framework shown below:

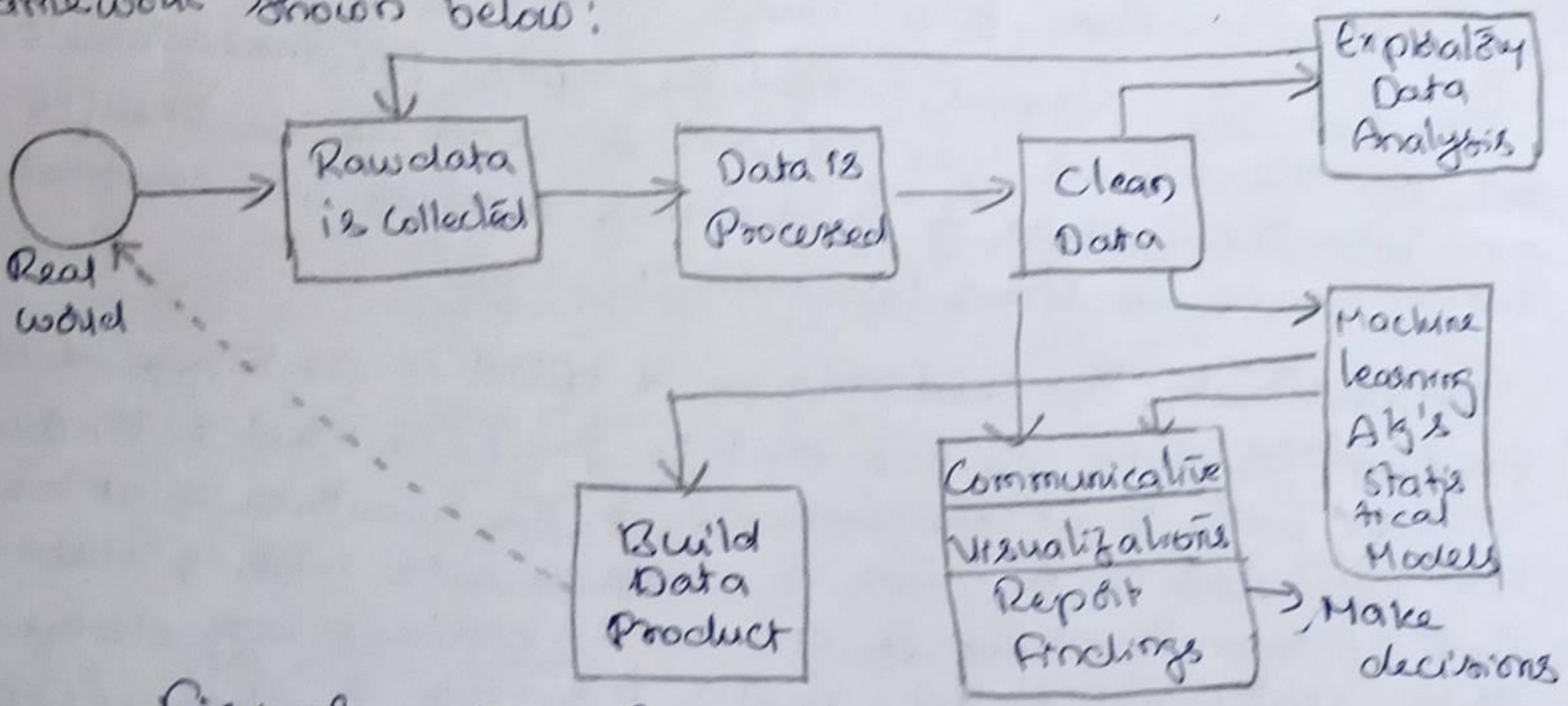


Fig: The data Science process.

First we have the real world. Inside the real world r lots of people busy at various activities. Some people r using Google+, others r competing in the olympics, there r spammers sending spam, & there r people getting their blood drawn. Say we have data on one of these things.

Specifically, we'll start with raw data - logs, olympics records, Enron emp emails, & recorded genetic material. We want to process this to make it clean for analysis. So

we build & use pipelines of data munging: joining, scr
wrangling, & whatever u want to call it. To do this we'll
tools such as python, shell scripts, R & SQL & all of the
above.

Eventually we get the data down to a nice format, like some
-ng with columns:

name | event | year | gender | event time

Once we have this clean dataset, we should be doing some
kind of EDA. In the course of doing EDA, we may realize
that it isn't actually clean bcoz of duplicates, missing value
absurd outliers & data that wasn't actually logged & incor
-rectly logged. If that's the case, we may have to go back to
collect more data, & spend more time cleaning the dataset

Next, we design our model to use some alg's like KNN,
linear regression, Naive Bayes, & ~~Bayes~~ something else. The
model we choose depends on the type of problem we're trying
to solve, of course, which could be a classification problem,
prediction problem, & a basic description problem.

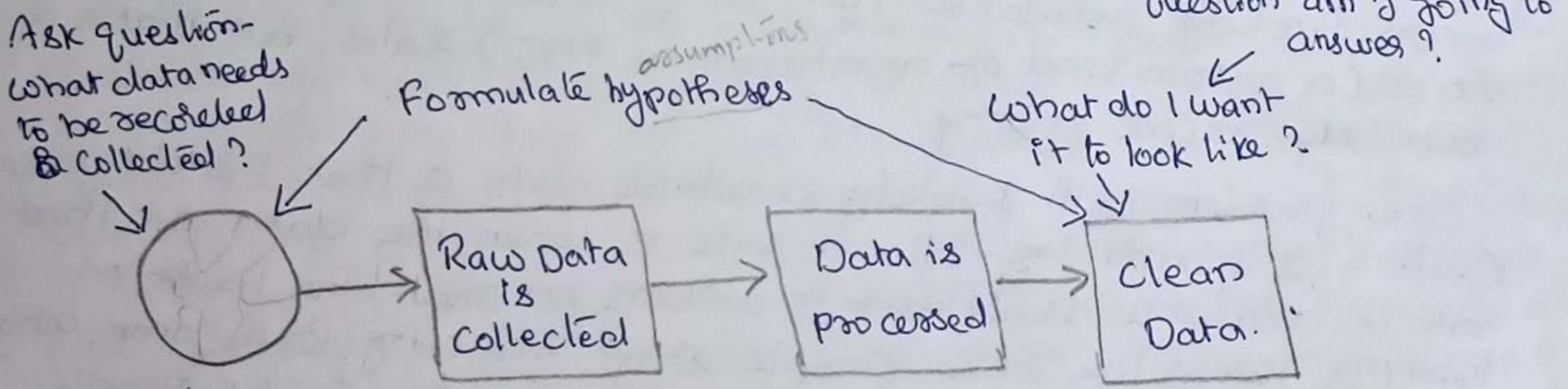
we then can interpret, visualize, report, & communicate our
results. This could take the form of repeating the results up to
our boss & coworkers, & publishing a paper in a journal & going
out & giving academic talks abt it.

Alternatively, our goal may be to build & prototype a "data
Product". A data product that is productionized & that users
interact with is at one extreme & the weather is at the other,
but regardless of the type of data u work with & "data product"
that gets built on top of it - be it public policy determined
by a statistical model, health insurance, & election polls that
gets built on top of it - be it public policy determined widely
reported & perhaps influence viewer opinions - u should
consider the extent to which ur model is influencing the
very phenomenon that u r trying to observe & understand.

3) A data scientist role in this process: (3)

This model so far seems to suggest this will all magically happen without human intervention. By "human" here, we mean "data scientist". Someone has to make the decisions abt what data to collect, & why. That person needs to be formulating questions & hypotheses & making a plan for how the problem will be attacked. And that someone is the data scientist or our beloved data science team.

Let's revise & at least add an overlay to make clear that the data scientist to be involved in this process throughout, meaning they r involved in the actual coding as well as in the higher level process as shown below:



Human behaving	- Email	- Pipelines	- Clean
Biology	- logs	- web-scraping	- outliers
finance	- Medical records	- cleaning	- Missing values
Internet	- Surveys	- Munging	- Debugging
Medicine	- Blood drawn	- Joining	- Table.
Sociology	- Olympic records	- wrangling	
olympics	- NYT web pages (articles)		

Fig: The data-scientist is involved in every part of this process.

4) Case Study: RealDirect:

Doug Perlson, the CEO of RealDirect, has a background in real estate law, startups & online advertising. His goal with RealDirect is to use all the data he can access abt real estate to improve the way people sell & buy houses.

Normally, people sell their homes abt once every 7 yrs, & they do so with the help of professional brokers & current data. But there's a problem both with the broker system & the data quality. RealDirect addresses both of them.

First, the brokers. They r typically "free agents" operating on own - think of them as home sales consultants. This means that guard their data aggressively, & the really good ones have lots experience. But in the grand schemes of things, that really means they have only slightly more data than the interpersonal brokers.

RealDirect is addressing this problem by hiring a team of licensed real estate agents who work together & pool their knowle -dge. To accomplish this, it built an interface for sellers, giving them useful data-driven tips on how to sell their house. It also uses interaction data to give real-time recommendations on what to do next.

The team of brokers also become data experts, learning to use inf-collecting tools to keep tabs on new & relevant data & to access publicly available inf. For eg, u can now get data on Co-op (a certain kind of apartment in NYC) sales, but that's a relatively recent change.

One problem with publicly available data is that it's old news - there's a 3 months lag b/w a sale & when the data abt that sale is available. RealDirect is working on real-time feeds on things like when people start searching for a home, what the initial offer is, the time b/w offer & close, & how people search for a home online.

Ultimately, good inf helps both the buyer & the seller. At least if they're honest.

(ii) How does RealDirect make Money?

First it offers a subscription to sellers - abt $395 a month - to access the selling tools. Second, it allows sellers to use RealDirect's agents at a reduced commission, typically 2% of the sale instead of the usual 2.5% or 3%. This is where the magic of data pooling comes in: it allows RealDirect to take a smaller commission b/c it's more optimized, & therefore gets more volume.

The site itself is best thought of as a platform for buyers & sellers to manage their sale & purchase process. There r statuses for each person on site: active, offer rejected, showing, in contract etc., Based on ur status, different actions r suggested by the software.

There's some challenges they have to deal with as well, of course. First off, there's a law in New York that says u can't show all the current housing listings unless those listings reside behind a registration wall, so RealDirect requires registration. On the one hand, this is an obstacle for buyers, but seriously buyers r likely willing to do it. Moreover, places that don't require registration, like Zillow, aren't true competitors to RealDirect because they r merely showing listings without providing additional service. Doug pointed out that u also need to register to use pinterest, & it has lots of users inspite of this.

RealDirect comprises licensed brokers in various established realtor associations, but even so it has had its share of hate mail from realtors who don't appreciate its approach to cutting commission costs. In this sense, RealDirect is breaking directly into a guild. On the other hand, if a realtor refused to show houses bcoz they r being sold on RealDirect, the potential buyers would see those listings elsewhere & complain. So the traditional brokers have little choice but to deal with RealDirect even if they don't like it. In other words, the listings themselves r sufficiently transparent so that the traditional brokers can't get away with keeping their buyers away from these houses.

Doug talked abt key issues that a buyer might care about - nearby parks, subway, & schools, as well as the comparison of prices per ~~the~~ square foot of apartments sold in the same building & block. This is the kind of data they want to increasingly collect as part of the service of RealDirect.

Exercise: Real Direct Data Strategy:

u have been hired as Chief data scientist at realdirect.com, & report directly to CEO. The company doesn't yet have its data plan in place. It's looking to u to come up with a data strategy.

→ u can use any or all of the datasets.
→ First challenge: load in & clean up data. Next, conduct EDA to find out missing values.
→ once the data is in good shape, conduct EDA to visualize & make comparisons. If u have time, start looking for meaningful patterns in this dataset.
→ Summarize ur findings in a brief report aimed at the CEO.

Data scientist role (continue):

We can think of the data science process as an extension of variation of the scientific method:

→ Ask a question → Do background research
→ Construct a hypothesis
→ Test ur hypothesis by doing an experiment
→ Analyze ur data & draw a conclusion
→ Communicate ur results.

In both the data science process & the scientific method, not every problem requires one to go thru all steps, but almost all problems can be solved with some combination of stages.

Exploratory Data Analysis Tools:

Now a days, ample of tools r available in the market which r free & quite interesting to work with. These tools doesn't require u to code explicitly but simple drag-drop clicks does the job.

1) Weka: Data mining s/w in java:

Weka is a collection of machine learning alg's for data mining tasks. The alg's can either be applied directly to a dataset or called from ur own java code. Weka contains tools for data pre-processing, classification, regression, clustering and visualization. It is an open src s/w issued under GNU General Public license.

2) R: R is a lang & environment for statistical computing & graphics. R provides a wide variety of statistics & graphical techniques & is highly extensible.

3) Gephi: The Open Graph Viz platform:

Gephi is the leading visualization & exploration s/w for all kinds of graphs & nw's. Gephi is open src & free. It runs on windows, Mac OS & Linux. Gephi is a tool for data analysts & scientists keen to explore & understand graphs.

the goal is to help data analysts to make hypotheses & discover patterns.

4) OpenRefine:

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it, transforming it from one format into another, & extending it with web services & external data.

5) Orange:

It is an open src machine learning & data visualization tool for novice & expert. It is an interactive data analysis tool with a large toolbox. It performs simple data analysis with clever data visualization.

6) Trifacta:

Trifacta's wrangling tool is challenging the traditional methods of data cleaning & manipulation. Since excel posses limitations on data size, this tool has no such boundaries & u can securely work on big data sets.

7) Rapid Miner:

It is more than just a data cleaning tool. It extends its expertise in building ML models. Not just a GUI, it also extends support to people using Python & R for model building.

8) Qlikview:

It is one of the most popular tool in business intelligence industry around the world. With its state of art visualization capabilities, you'd be amazed by the amt of clrs u get while working on data. It has an inbuilt recommendation engine to update u from time to time about best visualization methods while working on datasets.

9) Data Cracker:

It's a data analysis tool which specializes on survey data. Many companies do survey but they struggle to analyze it statistically. Survey data is never clean. It comprises of

lot of noise & inappropriate value. This tool reduces our
agony & enhances our experience of working on messy data!