DATA SCIENCE ESSENTIALS

UNIT-I

1. What is Data Science?

Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, mgmt & preservation of large collections of inf (both structured & unstrd)

Over the past few years, there's been a lot of hype in the media abt "data Science" & "Big data". A reasonable first reaction to all of this might be some combination of skepticism & confusion.

i) Big Data & Data Science Hype: attracting

What is eye-brow raising abt Big Data & data science? Let's count the ways:

1) There's a lack of def's around the most basic terminology. What does "data science" mean? What is "Big Data" anyway? What is the relationship b/w Big Data & data science? Is data science the science of Big Data? Is data science only the stuff going on in companies like Google & facebook & tech companies? Why do many people refer to Big Data as crossing disciplines (astronomy, finance, tech etc.) & data science as only taking place in tech? Just how big is big? & is it just a relative term?

2) There is distinct lack of respect for the researchers in academia & industry lab who have been working on this kind of stuff for yrs by statisticians, computer scientists, mathematicians, engineers & scientists of all types.

From the way the media describes it, machine learning alg's were invented & data was never "big" until Google came along.

3) The hype is crazy - people throw around tired phrases straight out of the height of the pre-financial crisis era like "Masters of the universe" to describe data scientists, & that doesn't bode well.

4) Statisticians already feel that they r studying & working on the "Science of Data". That's their bread & butter. May be, dear reader, r not a statistician & don't care, but imagine that for the statistician, this feels a little bit like how identity theft might feel for u. The media often describes data science

in a way that makes it sound like as if it's simply sh
& machine learning in the content of the tech industry.

ii) ~~Getting Past the Hype: (See 1b)~~

iii) Why Now:

We have massive amounts of data abt many aspects of our lives & simultaneously, an abundance of inexpensive computing power. Shopping, communicating, reading news, listening to music, searching for inf, expressing our opini - all this is being tracked online, as most people know.

What people might not know is that the "dataficatio of our offline behavior has started as well, mirroring th online data collection revolution. put the two together, & ther a lot to learn abt our behavior &, by extension, who we are a a species.

It's not just Internet data, though - it's finance, the medical industry, pharmaceuticals, bioinformatics, social welfare, govt, education, retail & the list goes on. There is a growing influence of data in most sectors & most industries. In some cases, the amt of data collected might be enough to be considered "big".

iv) Datafication:

Datafication is a modern technological trend turning many aspects of our life into computerised data & transforming this inf into new forms of value. Examples of datafication as applied to social & comm media r how Twitter datafies stray thoughts or datafication of HR by LinkedIn & others.

In May/June 2013 issue of Foreign Affairs, Kenneth Neil Cukier & Viktor Mayer-Schoenberger wrote an article called "The Rise of Big Data". In it they discuss the concept of datafication, & their eg of how we quantify friendship with "likes": it's the way everything we do, online or otherwise, ends up recorded for examination in someone's data storage units, or may be multiple storage units, & may be also for sale.

They define datafication as a process of "taking all aspects of life & turning them into data". As eg s, they mention that "Google's augmented-reality glasses datafy the gaze. Twitter tafies stray thoughts. LinkedIn datafies professional n/w's.

Datafication is an interesting Concept & led us to consider its importance with respect to people's intentions abt sharing their own data. we r being datafied, & rather our actions r, & when we "like" someone & something online, we r intending to be datafied, & at least we should expect to be. But when we merely browse the web, we r unintentionally, & at least passively, being datafied thru cookies that we might & might not be aware of. And when we walk around in a store, & even on the street, we r being datafied in a completely unintentional way, via sensors, cameras & Google glasses.

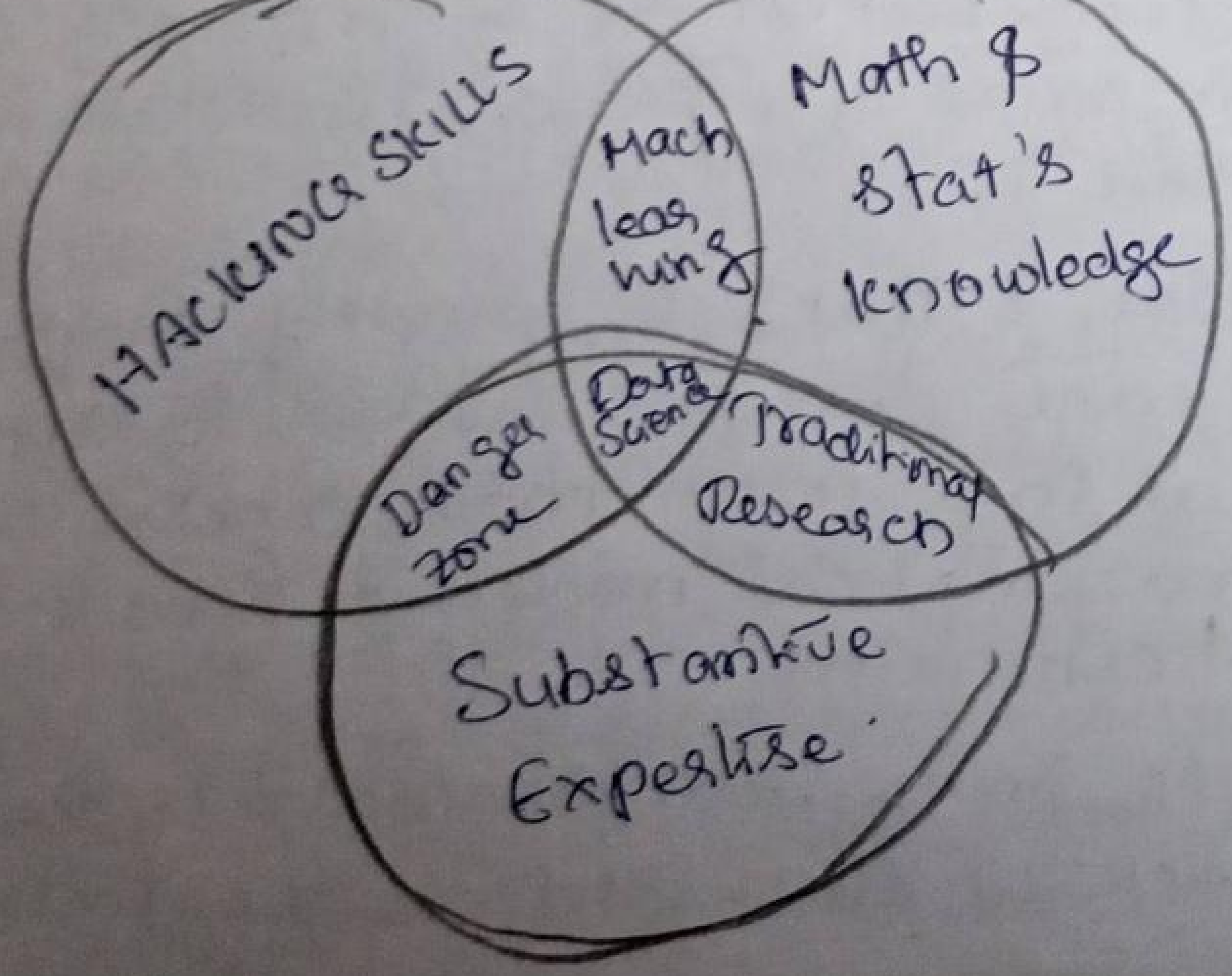
If we want to think bigger, if we want our "we" to refer to people in general, we'll be swimming against the tide.

iv) The Current Landscape:

So, what is data science? Is it new, & is it just stat's & analytics rebranded? Is it real, & is it pure hype? And if it's new & if it's real, what does that mean?

This is an ongoing discussion, but one way to understand what's going on in this industry is to look online & see what current discussions r taking place. This doesn't necessarily tell us what data science is, but it atleast tells us what other people think it is, & how they're perceiving it.

For eg, on Quora there's a discussion from 2010 abt "what is Data Science?", ~~Discoll~~ Driscoll then refers to Drew Conway's Venn diagram of data science from 2010 as shown,



2) Statistical Inference :

diff things unknown not known

The world we live ~~in~~ is complex, random & uncertain. At the same time, it's one big data generating machine.

travel

As we commute to work on subways ~~in~~ & in cars, as our blood moves thru our bodies, as we're shopping, emailing, procrastinating at work by browsing the Internet & watching the stock market, as we're building things, eating things, talking to our friends & family abt things, while factories r producing products, this all atleast potentially produces data.

data/gap in future

Imagine spending 24 hrs looking out the window, & for every minute, counting & recording the no of people who pass by. Or gathering up everyone who lives within a mile of ur house & making them tell u how many email msg's they receive everyday for the next year. The point here is that the processes in our lives r actually data generating processes.

We'd like ways to describe, understand, & make sense of these processes, in part bcoz as scientists we just want to understand the world better, but many times, understanding these processes is part of the soln to problems we're trying to solve.

After seperating the process from the data collection, we can see clearly that there r 2 src's of randomness & uncertainity. Namely, the randomness & uncertainity underlying the process itself, & the uncertainity associated with ur underlying data collection methods.

Once u have all this data, u have somehow captured the world, or certain traces of the world. But u can't go walking around with a huge excel spreadsheet or db of millions of transactions & look at it and, with a snap of a finger, understand the world & process that generated it.

So, u need a new idea, & that's to simplify those captured traces into something more comprehensible, to something that somehow captures it all in a much more concise way, & that something could be mathematical models & func's of the data, known as Statistical estimators.

This overall process of going from the world to the data, & then from the data back to the world, is the field of statistical Inference.

More precisely, statistical inference. is the discipline that concerns itself with the development of procedures, methods & theorems that allow us to extract meaning & inf from data that has been generated by stochastic (random) processes.

3) Populations & Samples:

In classical statistical literature, a distinction is made b/w the population & the sample. The word population immediately makes us think of the entire world population 7 billion people. But put that image out of ur head, bcoz in statistical inference population isn't used to simply describe only people. It could be any set of objects or units, such as tweets / photographs / stars.

If we could measure the characteristics & extract characteristics of all those obj's, we'd have a complete set of observations, & the convention is to use N to represent the total no of observations in the population.

Suppose ur population was all emails sent last year by employees at a huge corporation, BigCorp. Then a single observation could be a list of things: the sender's name, the list of recipients, data sent, text of e-mail, no of chars in the email, no of sentences in the email, no of verbs in the email, & the length of time until first reply.

When we take a sample, we take a subset of the units of size n in order to examine the observations to draw conclusions & make inferences abt the population. There r different ways u might go abt getting this subset of data, & u want to be aware of this sampling mechanisms bcoz it can introduce biases into the data, & distort it, so that the subset is not a "mini-me" shrunk-down version of the population. Once that happens, any conclusions u draw will simply be wrong & distorted.

In the BigCorp email eg, u could make a list of all the employees & select 1/10th of those people at random & take all the email they ever sent, & that would be ur sample. Alternatively, u could sample 1/10th of all email sent each day at random, & that would be ur sample. Both these methods r reasonable, & both methods yield the same sample size. But if u look them & counted how many email msg's each person sent, & used that to estimate the underlying distribution of emails sent by all individuals at BigCorp, u might get entirely different answers.

So if even getting a basic tally get distorted when you're using a reasonable - sounding sampling method, imagine what can happen to more Compl alg's & models if u have'nt taken into account the proce that got data into ur hands.

4) Statistical Modeling:

Before u get too involved with the data & start Coding, it's useful to draw a picture of what u think the underlying Process might be with ur model. what influences what? what causes what? what's a test of what?

But different people think in different ways. Some prefer to express these kinds of relationships in terms of math. The mathematical expressions will be general enough that they have to include parameters, but the values of these parameters r not yet known.

In mathematical expressions, the Convention is to use Greek letters for Parameters & Latin letters for data. So, for eg, if u have two columns of data, x and y, & u think there's a linear relationship, you'd write down $y = \beta_0 + \beta_1 x$. You don't know what β_0 & β_1 are in terms of actual no's yet, so, they're the parameters.

Other people prefer pictures & will first draw a diagram of data flow, possibly with arrows, showing how things affect other things & what happens over time. This gives them an abstract picture of the relationships before choosing eqn's to express them.

Remember, it's always good to start simply. There is a trade-off in modeling b/w simple & accurate. Simple models may be easier to interpret & understand. Oftentimes the crude, simple model gets u 90% of the way there & only takes a few hrs to build & fit, whereas getting a more complex model might take months & only get u to 92%..

Some of the building blocks of these models are Probability distributions.

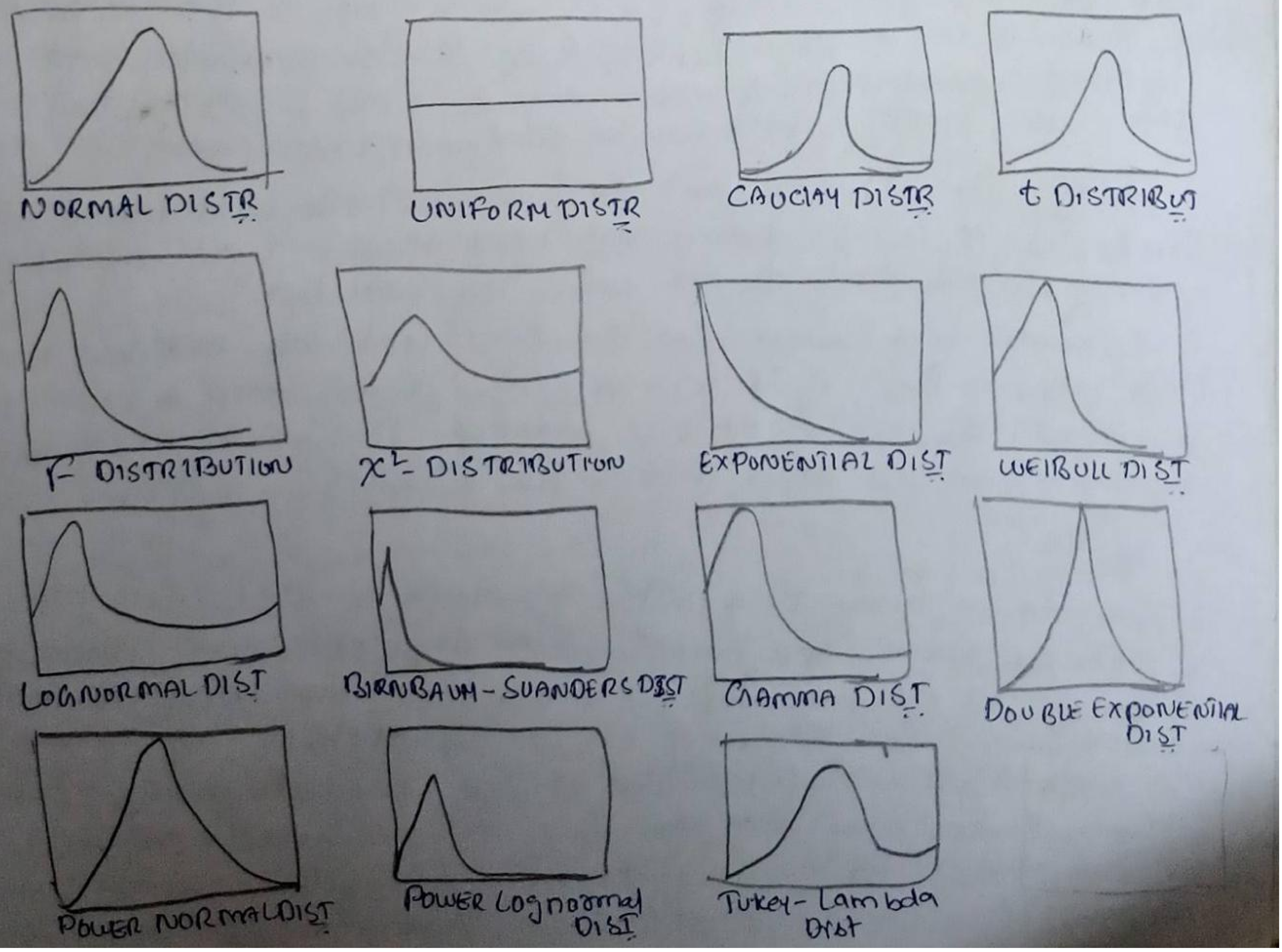
Probability Distributions:

Prob distributions r the foundation of Statistical models. When we get to linear regression & Naive Bayes, u will see how this happens in practice. One can take multiple semesters of courses on prob theory, & so it's a tall challenge to condense it down for u in a small section.

Back in the day, before computers, scientists observed real-world phenomenon, took measurements, & noticed that certain mathematical shapes kept reappearing. The classical eg is the height of humans, following a normal distr - a bell shaped curve, also called a Gaussian distr, named after Gauss.

Natural processes tend to generate measurements whose emphorical shape could be approximated by mathematical func's with a few param's that could be estimated from the data.

Not all processes generate data that looks like a named distr, but many do. We can use these func's as building blocks of our models. The following fig is an illustration of various common shapes. There is actually an infinite no of possible distr's (based on data patterns the prob will be distributed)



NORMAL DISTR

UNIFORM DISTR

CAUCHY DISTR

t DISTRIBUT

F DISTRIBUTION

χ^2- DISTRIBUTION

EXPONENTIAL DIST

WEIBULL DIST

LOGNORMAL DIST

BIRNBAUM-SUANDERS DIST

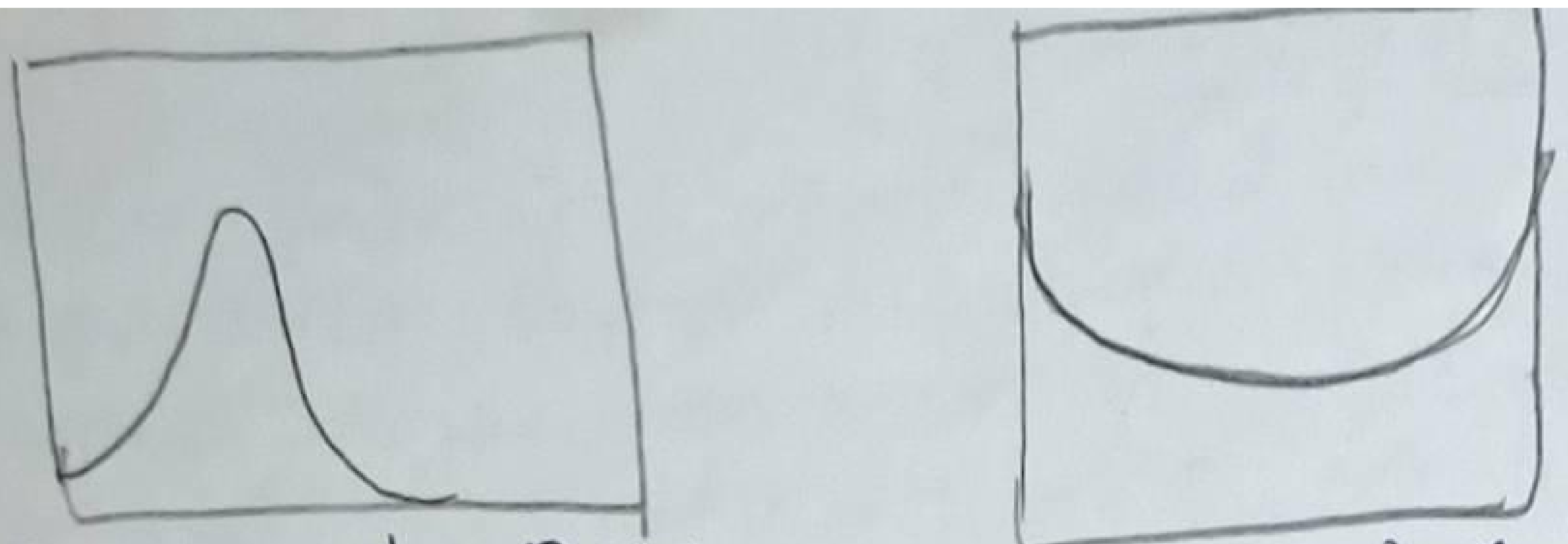
GAMMA DIST

DOUBLE EXPONENTIAL DIST

POWER NORMAL DIST

POWER Lognormal DIST

Tukey-Lambda Dist



Extreme value Distri   Beta Distribution

Fig: A bunch of Continuous density func's (aka prob distri's)

They r to be interpreted as assigning a prob to a sub, of possible outcomes, & have corresponding func's. For eg → the normal distribution is written as:

$$N(x|\mu,\sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The parameters μ is the mean & median & controls where distribution is centered, & the parameter 'σ' controls how spread out the distribution is. This is the general functional form, but for specific real-world phenomenon, these parameters have actual no's as values, which we can estimate from the data.

A random var denoted by x & y can be assumed to have a corresponding prob distri, P(x), which maps x to a +ve real no. In order to be a prob density func, we're restricted to the set of func's such that if we integrate P(x) to get the area under the curve, it is 1, so it can be interpreted as prob.

For eg, let x be the amt of time until the next bus arrives (in sec's). x is a random var bcoz there is variation & uncertainity in the amt of time until the next bus.

Suppose we know that the time until the next bus has a prob density func of $P(x) = 2e^{-2x}$. If we want to know the likelihood of the next bus arriving in b/w 12 & 13 minutes, then we find the area under the curve b/w 12 & 13 by $\int_{12}^{13} 2e^{-2x}$.

How do we know this is the rt distri to use? Well, there r 2 possible ways: we can conduct an experiment where we show up at the bus stop at a random time, measure how much time until the next bus, & repeat this experiment over & over again. Then we look at the measurements, plot them, & approximate the func as discussed. Or, bcoz we r familiar with the fact that "waiting time" is a common enough real-world phenomenon that a distri called exponential distrⁿ has been invented to describe it, know that it takes the form $p(x) = \lambda e^{-\lambda x}$

In addition to denoting distrⁿ's of single random var's with func's of one var, we use multivariate func's called joint distrⁿ's to do the same thing for more than one random var.

we also have what is called a Conditional distrⁿ, $p(x|y)$, which is to be interpreted as the density funcⁿ of x given a particular value of Y.

when we're working with data, conditioning corresponds to subsetting. So for egⁿ, suppose we have a set of user-level data for Amazon.com that lists for each user the amt of money spent last month on Amazon, whether the user is male/female, & how many items they looked at before adding the first item to shopping cart.

If we consider X to be the random var that represents the amount of money spent, we can look at the distrⁿ of money spent across all users, & represent it as $p(x)$.

we can then take the subset of users who looked at more than 5 items b4re buying anything & look at the distrⁿ of money spent among these users. Let Y be the random var that represents noⁿ of items looked at, then $p(x|y>5)$ would be the corresponding Conditional distrⁿ.

when we observe data points, i.e, $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, we r observing realizations of a pair of random var's. when we have an entire dataset with n rows & K col's, we r observing 'n' realizations of the joint distrⁿ of those K random var's.

15 45 50 min ≈ optimal

min, max

6) Fitting a Model: (House constrn)

factor | measure

fitting a model means that u estimate the parameters of the model using the observed data. u r using ur data as evidence to help approximate the real world mathematical process that generated the data. Fitting the model often involves optimization methods & alg's such as maximum likelihood estimation, to help get the parameters.

In fact, when u estimate the param's, they r actually estimators, meaning they themselves r func's of the data.

who can easily

Once u fit the model, u actually can write it as $y = 7.2x + 4$ for eg, which means that ur guess is that this eqn & func'nal form expresses the relationship b/w ur 2 var's, based on ur assumption that the data followed a linear pattern.

Fitting the model is when u start actually coding: ur code will read in the data, & you'll specify the func'nal form that u wrote down on the piece of paper. Then R & Python will use built-in optimization methods to give u the most likely values of the parameters given the data.

As u gain sophistication, or if this is one of ur areas of expertise, u'll dig around in the optimization methods used. Initially u should have an understanding that optimization is taking place & how it works, but u don't have to code this part urself - it underlies the R & Python func's.

7) Introduction to R :

R is a programming lang & s/w environment for statistical analysis, graphics representation & reporting. R was created by Ross Ihaka & Robert Gentleman at University of Auckland, New Zealand, & is currently developed by the R Development core team in 1993.

R is freely available under GNU General public license. This prg'mming lang was named R, based on first letter of first name of 2 authors.

The core of R is an interpreted computer lang which allows branching & looping as well as modular prg'mming using func's. R allows integration with the procedures written in the C, C++, .NET, Python & FORTRAN lang's for efficiency.

Features of R :

→ R is a well-developed, simple & effective prg'ming lang which includes conditions, loops, user defined recursive func's & i/p & o/p facilities

→ R has an effective data handling & storage facility

→ R provides a suite of operators for calculations on arrays, lists, vectors & matrices

→ R provides a large, coherent & integrated collection of tools for data analytics.

→ R provides graphical facilities for data analysis & display either directly at the computer / printing at the paper

→ R is world's most widely used statistics prg'mming lang. It is the no 1 choice of data scientists & supported by a vibrant & talented community of contributors.

Datafication:

Business Examples:

→ Social media is a great example of datafying aspects of users daily lives

→ Facebook datafies our friendship & posts

→ Twitter datafies our followers, Tweets & interactions.

→ LinkedIn datafies our professional contacts, locations, likes, posts.

Personal Examples:

→ Going for a jog - one can monitor distance, speed, pulse, heart rate

→ Sleep schedule - quality of sleep, duration, sleeping without interruption.

→ Shopping - how much food to purchase, finding lowest prices, monitoring quantities consumed in a household.

Datafication refers to the collective tools, technologies & processes used to transform an organization to a data driven enterprise.

The Current Landscape:

It is difficult to narrowly define the skills of a data scientist bcoz they r naturally interdisciplinary, yet they exist at intersections of disciplines that do not often merge. In a general sense, there r three primary areas of expertise needed to be successful data scientist.

First one must have Hacking Skills. It doesn't mean malicious computer hacking & unauthorized disclosure of inf. Rather hacking skills in this context mean expertise proficiency working with large, unstructured chunks of electronic data. Simply, a hacker is who can easily

.. D. ю

navigate the required set of tools needed to be a data scientist.

Second, one needs a basic understanding of mathematics & stat's. as these fundamentals will inform all of the analysis.

Finally, & perhaps most importantly, a data scientist must have some real & independent existence substantive expertise in the data being analyzed.

The cautionary word on the combination of hacking skills & substantive expertise, which is identified as danger zone. This is where people who "know sufficient enough to be dangerous" & the most problematic area of diagram.

Those in this category may be perfectly capable of extracting & structuring data, likely related to a field they know quite a bit about. They may even have sufficient technological make quick decisions acumen to run a linear regression & report the coefficients; but they lack any understanding of what those coefficients measures a particular property mean & how to interpret them.

Given the customers of intelligence products, & the support stakes at play in the community, even a brief lapse into the danger zone can have causing great damage catastrophic results.

Data Science Defi

It is the field of applying advanced analytics techniques & scientific principles to extract valuable inf from data for business decision making, strategic planning & others uses.