Data Science 101- Cheat Sheet

SPARKLINE
Igniting Insights

Summary of the 90-minute intro to DS with an emphasis on Image concepts. Join us for future talks as we dive into more detail and write code to do cooler stuff from scratch!

What we hope you takeaway from this:

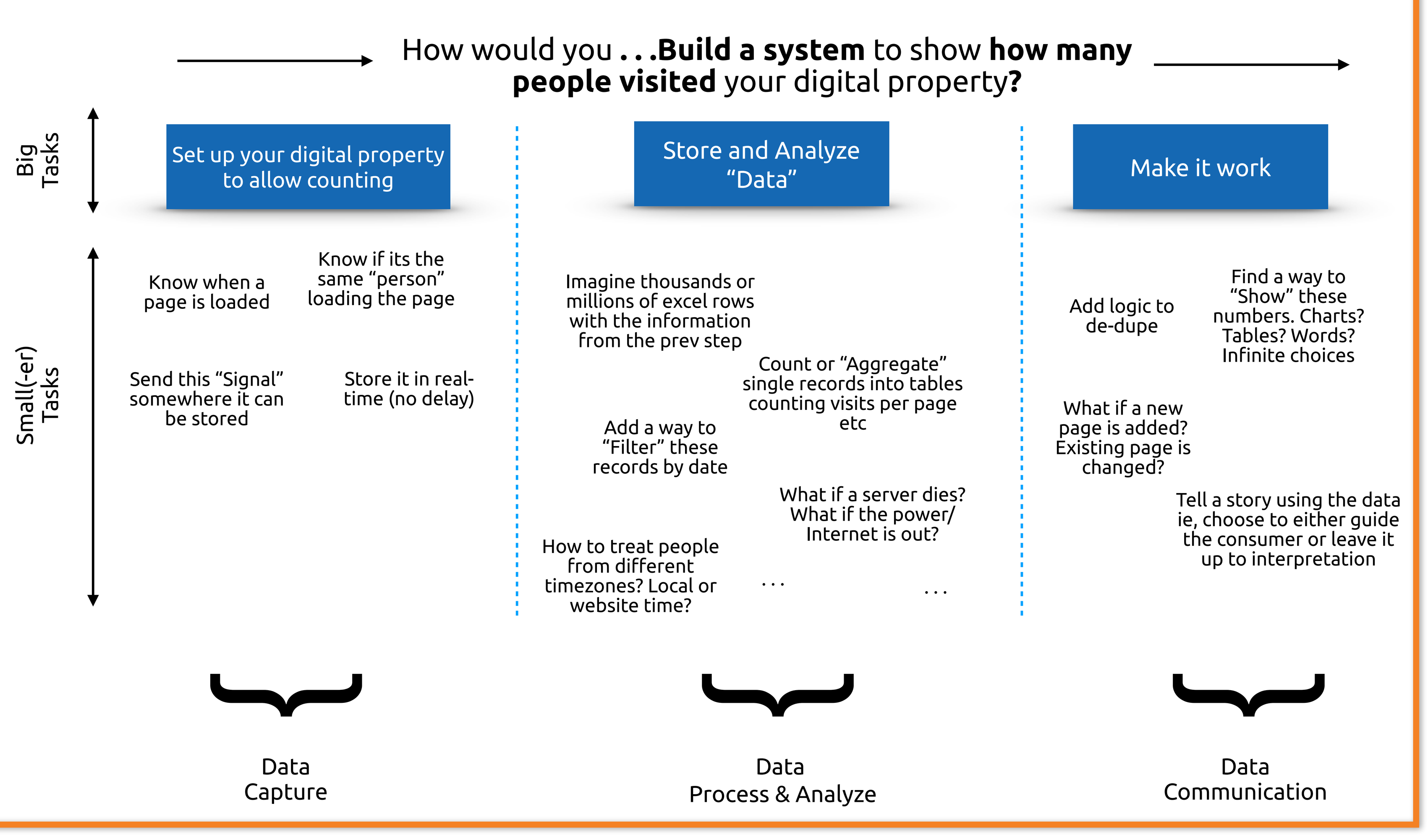
- Understand the ***Mechanics*** of the field
- Deep dive into one present hotness in the field: ***Deep Learning (CNNs)***
- What is data?
- High level overview of ***systems*** and ***techniques***
- Working with images

On the right of this text box, an example problem is presented to illustrate an approach in certain cases also referred to as "***Decomposition***" ie, breaking down a large problem into smaller and smaller blocks that, separately are relatively simple, but taken together, solve a complex problem.

We also illustrate roughly, the "**Data Science Life Cycle**", showing the typical steps along the DS journey:

- ***Data Capture*** - Instrumenting your sources or "getting" data from "somewhere".
- ***Data Processing*** - Taking the "Raw" data from the step above and applying methods to change it to the use case at hand.
- ***Data Analysis*** - Extracting intelligence from the data using methods ranging from intuitive digging to statistical methods grounded in math.
- ***Data Communication*** - "Telling a story" by visualizing the data as pictures or charts or other transformed depictions of numeric information.
- ***Maintenance*** - Across the board

Layering on complexity to solve problems



An Intro to Statistical Methods: One story told in three different ways. . .

Name	Sq. Footage	Price
House 1	100	$175,000.00
House 2	500	$520,000.00
House 3	240	$340,000.00
House 4	80	$150,000.00
House 5	1500	$1,500,000.00

fig i: Imagine you spend some time looking at house prices in a particular location and manage to collect a bunch of house prices along with the size of the homes. A sample of this list could look like the table above

=

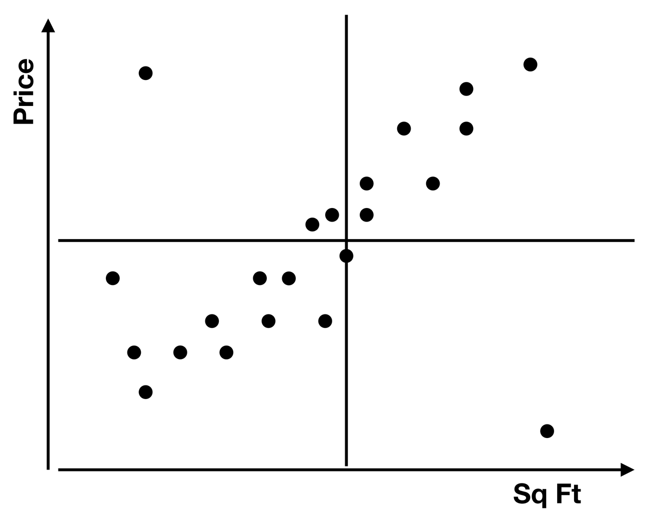


fig ii: Humans are really good at visual intuition. Simply looking at the chart above, one might see a pattern emerge where the dots (mostly) lie along a plane with a certain inclination.

*This is just an illustration of data from our imaginary housing market table to the left. We're simply laying the houses, depicted as dots here along an axis of price and another of the size of the house. These axes are referred to as "**Dimensions**"*

=

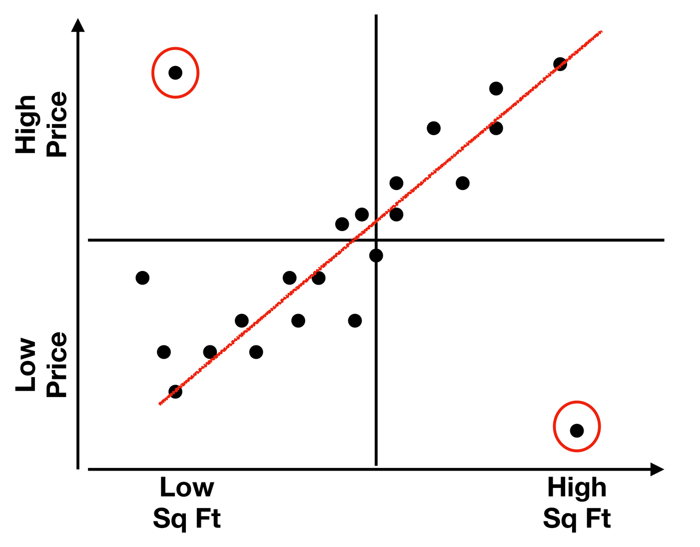


fig iii: Further adding detail to our chart, we see the imaginary line that approximates the positions of the dots along with "outliers" that don't lie on the line or reasonably near it.

*One likely knows from high school that this line can be described by a formula or an "**Equation**" that describes the "**Relationship**" between the two dimensions shown here.*

If we figure out what this relationship for the line is, we may guess-timate the price of a house if we know how big it is or vice-versa. *You may be surprised to know this is actually a statistical method to predict stuff.*

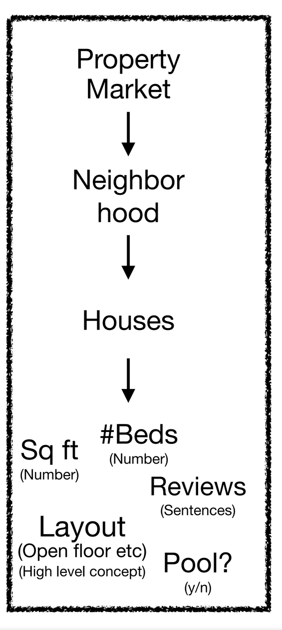
Shifting Gears: Lets talk fundamentals

So. . .What *is* "Data"

Well, the short answer might be "*It depends*". In the modules above, we've been implying indirectly that "Data" may be loosely described as "**Information**".

At a high level, when humans describe something like a ***Property Market***, we're talking about a ***high-level abstract Concept***.

Using the approach of layering complexity we saw earlier, let's break the concept of a 'Property Market" into simpler and simpler descriptions:

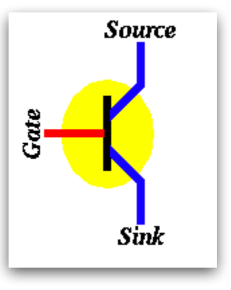


As shown in the figure to the left, we can describe a market as being loosely connected to a geographic neighborhood with houses.

Each house is characteristically different, were one to describe each of these houses, one would likely talk about the ***size*** which is a numeric value but also use other descriptors such as what the ***layout*** is, now this is likely a sentence and not representable at first glance using a number, but it's still ***information*** about the house. So now, we see that information or data also comes in many flavors or ***types*** ranging from *numbers*, true or false values (*boolean*) and even sentences describing a ***category*** of architectural choices etc

Data as a physical 'thing'

So far, we've limited ourselves to imagination, but if we were to do something useful with these concepts we just saw, we'll need a way to translate information to something we can operate on.



The image on the right is one of the fundamental building blocks of the "Computing" industry that has helped Humanity make such enormous advances over the past half century or so, the humble Transistor.

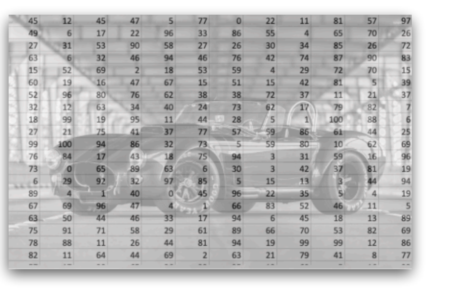
It's likely that ancient humans used the presence or absence of an object (lets say, a rock) to denote the presence or absence of an item (lets say, a cow) to convey meaning and help with calculations. (Look up how an Abacus works)

To do this electronically, the ingenious device in the image uses voltage applied at the gate to either allow current to flow from the source to the sink or block the flow if no voltage is applied.

For the scope of this module, lets take it for granted that solving this problem allowed one to electronically manipulate these depictions of information and thus do calculations much quicker than with large real-world objects like a rock.

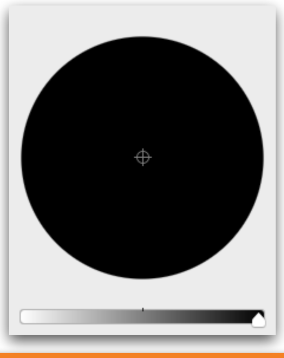
All modern computing devices operating on data work in part thanks to this principle.

Working with Images



A black and white image can be represented as a collection of points or ***pixels*** of grayscale intensities. Similarly, a color image may be represented as a collection of pixels where each pixel now is represented by three colors, **R**ed **B**lue and **G**reen - a mix of these colors in the right proportion leads to a new color and most colors visible to humans can be represented as a mix of these three ***primary colors***.

You may have seen controls similar to the ones below in color pickers in your favorite tools:



Neural Nets for absolute beginners:

So images are numbers, how do you modify them?

Passing a filter of numbers, called a ***kernel*** over an image and performing the ***convolution*** operation as described to the right over the elements of the image matrix and the kernel modifies the input image.

Different kernels modify the image differently.

For example, to the right, we see the effect of using a filter with large numbers in the center and values decreasing proportionally with distance from the center to give what's called a "Gaussian blur" to the image.

We also see the effect of performing the convolution operation using a different kernel to detect horizontal lines.

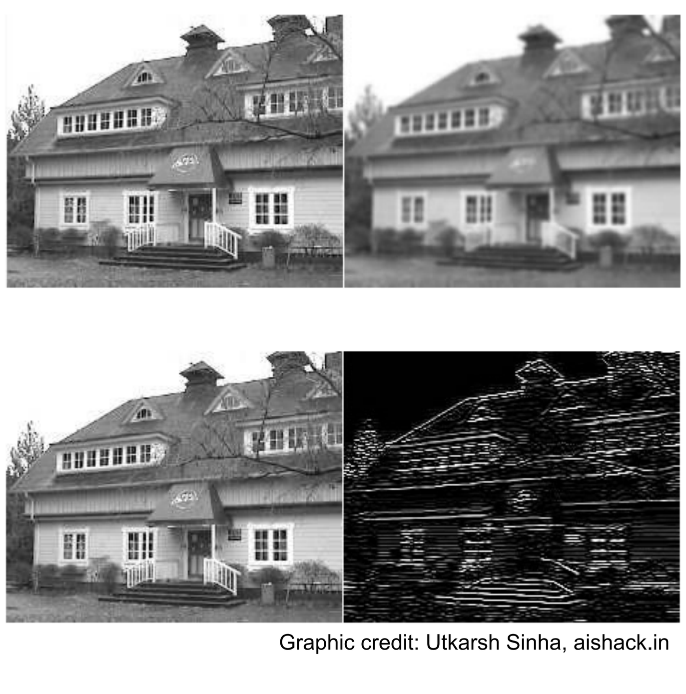
Notice how the values in the second kernel are high in the central horizontal row and negative everywhere else? Also the sum of all numbers in the second kernel cancel out to zero.

Kernel

0	0	0	5	0	0	0
0	5	18	32	18	5	0
0	18	64	100	64	18	0
5	32	100	100	100	32	5
0	18	64	100	64	18	0
0	5	18	32	18	5	0
0	0	0	5	0	0	0

-1	-1	-1
2	2	2
-1	-1	-1

Input Image **Output Image**



Graphic credit: Utkarsh Sinha, aishack.in

The Convolution Operation

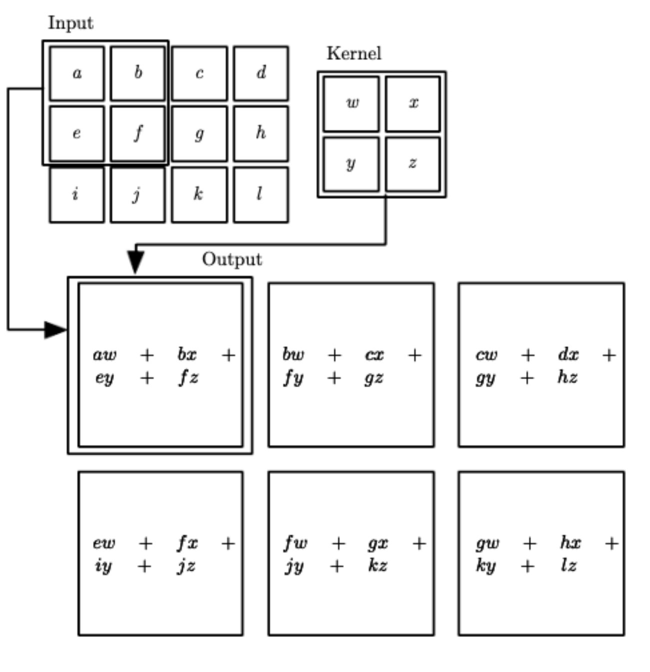
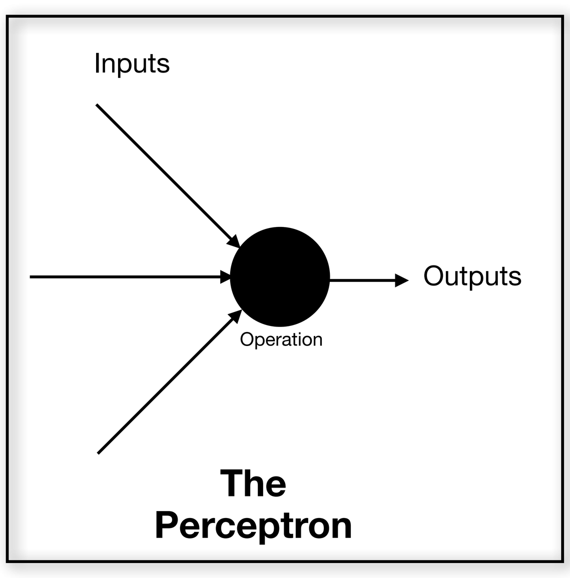


fig: The 2D convolution operation without kernel flipping.

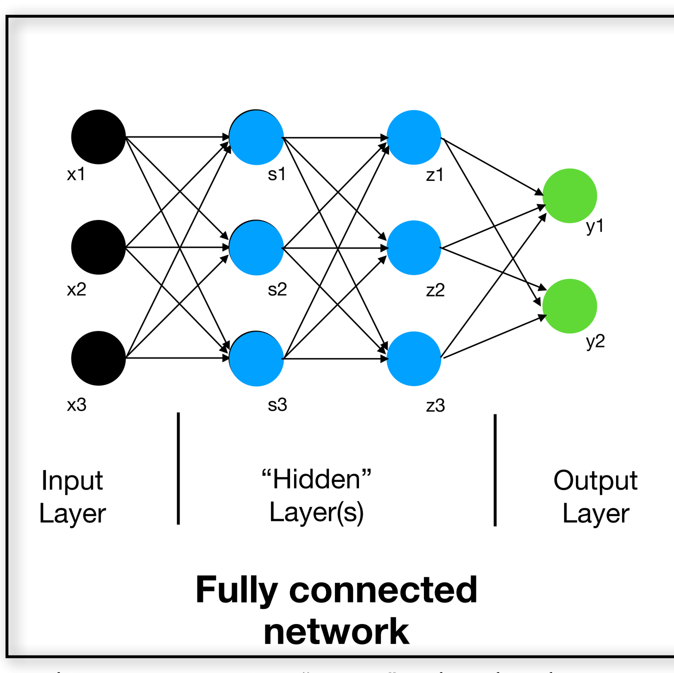
Credit: Deep Learning, Goodfellow et al, 2016, Ian Goodfellow and Yoshua Bengio and Aaron Courville, MIT Press

Building up complexity. Dense (& Sparsely) connected "Deep" networks



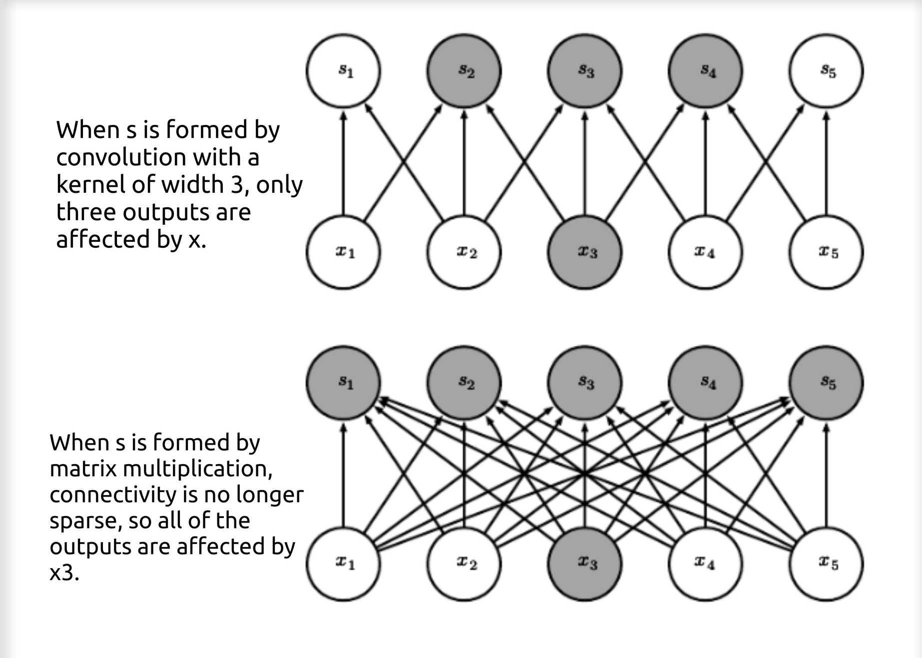
For the purpose of a 101, consider this to be a simple object that accepts inputs, performs an action on them and spits something out as an output.

The lines also have ***weights*** ie, a multiplier for the inputs



Stacking perceptrons into "***Layers***" and stacking layers on each other sequentially gives us the architecture for a "***Network***".

The nodes may either be all connected to all nodes in the next layer making it a "Densely connected" or "***Fully connected***" network or otherwise connected sparsely using an example method involving convolution as described to the right



(Top) : Sparsely connected network

(Bottom): Densely connected network

Credit: Deep Learning, Goodfellow et al, 2016, Ian Goodfellow and Yoshua Bengio and Aaron Courville, MIT Press

Did you like this module?

Check out courses on our Analytics Academy for more:

Get $900 off on our "21st Century Analyst - Boosting Careers in Data Analysis & Data Science" course until 31st July with the Discount Code: **SCL_399**

Our next module with Shopee is on Jul 14, watch the code league space for RSVPs opening and join us if you can! https://careers.shopee.sg/codeleague/#training

Questions about the Sparkline Training Tool or related? Reach out to Sarah at: https://www.linkedin.com/in/sarahrebello/

Any questions about the content of this module? Or suggestions to make it better? Please drop AK a note at: https://www.linkedin.com/in/a-k-hemanth-kumar-b8287974/

SPARKLINE
Igniting Insights