In machine learning we use the complete/whole dataset that we have.

So why visualize it?

① Know what you have $\xrightarrow{\text{so you can}}$ know what to do with it

② Machine learning algorithms are sensitive to missing data and data attributes that are of very different scales -

e.g. 0, 1, 2 for years of education while \$100,000 → \$1,000,000 for salary.

(3) Non-representative data or skewed data can affect the quality of what you can infer from the data.

How to systematically explore data?

There is no recipe to do it - but there are some guidelines.

A) START by getting the data in a form that you can manipulate easily.

- Excel sheet
- Dataframe / database table
- File that can be read by Orange.

B) Get a handle on the <u>structure</u> of the data.

rows, # columns

types of information

Numerical versus categorical

C) For numerical data attributes

- Summary statistics
  Mean, Median, Distribution
  ↓ outliers

Display as box plots
histograms
(representativeness)

D) For categorical data

- Unique values of each category
- Balance / Distribution of these values within a category
  (representativeness)

Display as bar graphs
facet plots

E) For both numerical and categorical data, find and handle missing values.

Empty cells ' '

Truly empty cells – e.g. NaN

Figure out how to handle missing instances

- remove column(s)
- remove row(s)
- impute values
  - mean
  - median
  - mode

F) Relationships between Categorical attributes

Pivots, Facet plots, Bar graphs

G) Relationships between numerical attributes

Scatter plots
Correlation tables
Overlapping histograms
Correlation + Distribution diagrams

H) Relationships between categorical and numerical attributes

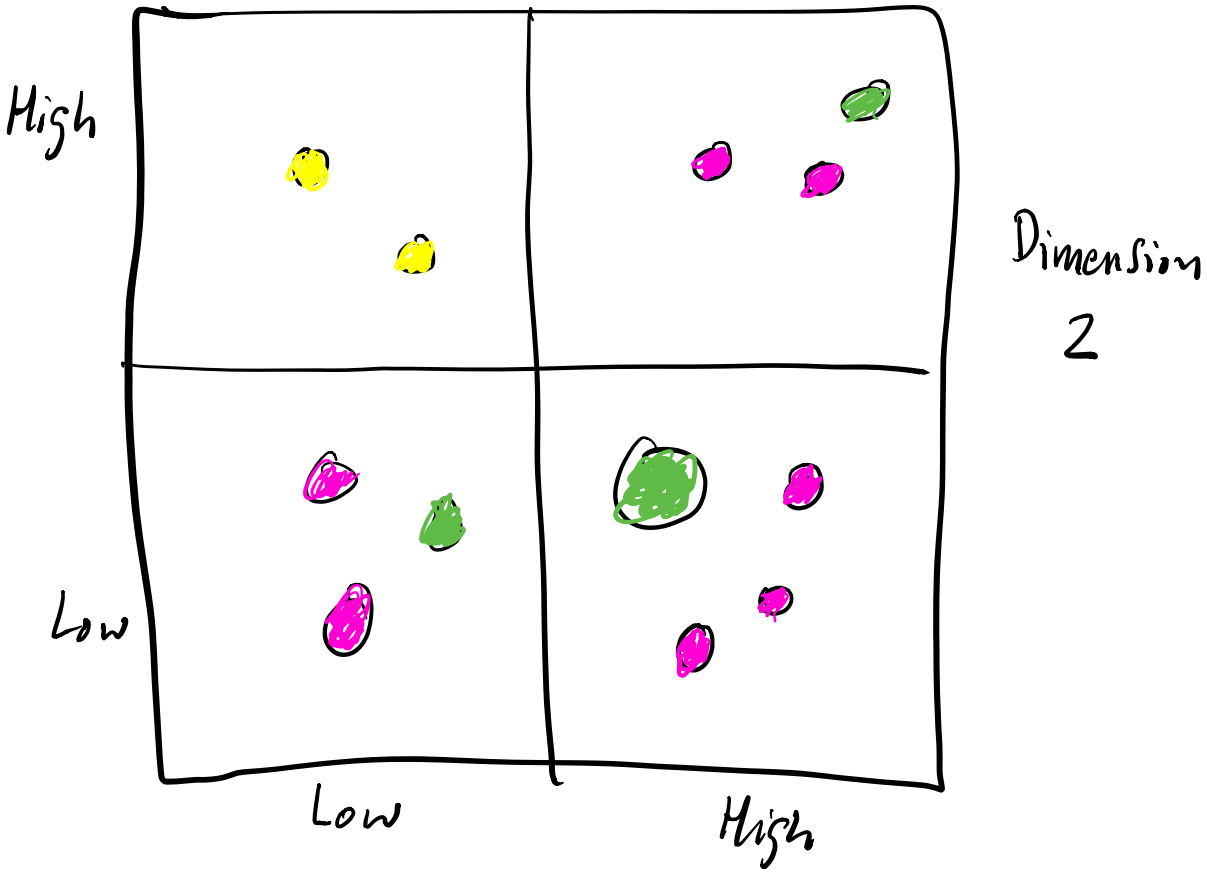
Facet plots
Density plots
Jitter plots

What happens when there are more than 3 features to visualize?

There are a few tricks we can use for displaying as many as 7 dimensions comfortably on a piece of 2-D paper.

Let's have a look...

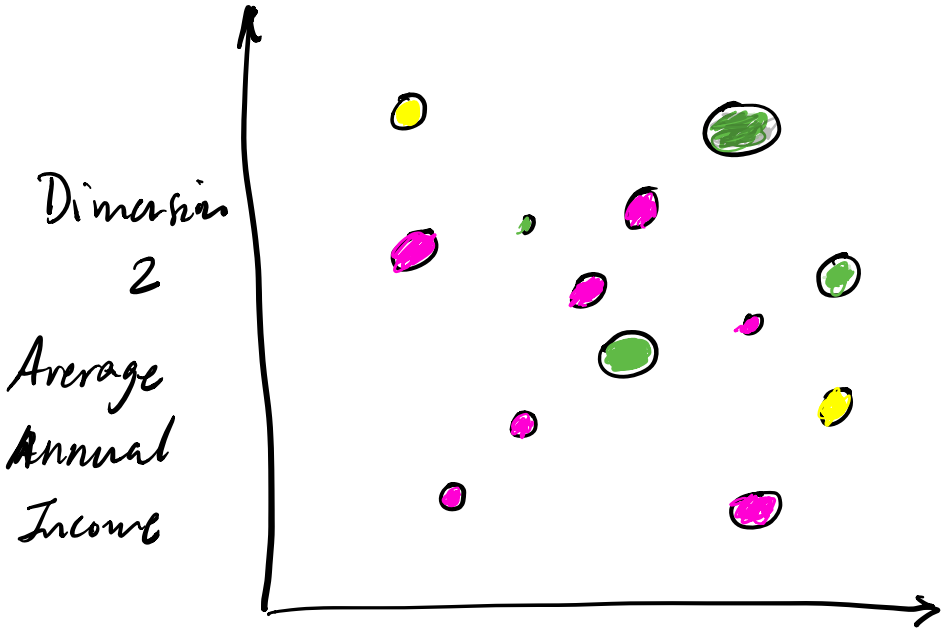
2 x 2 Chart — 4 Dimensions



Dimension 3 - Bubble Size ○ ○ ○

Dimension 4 - Bubble Color

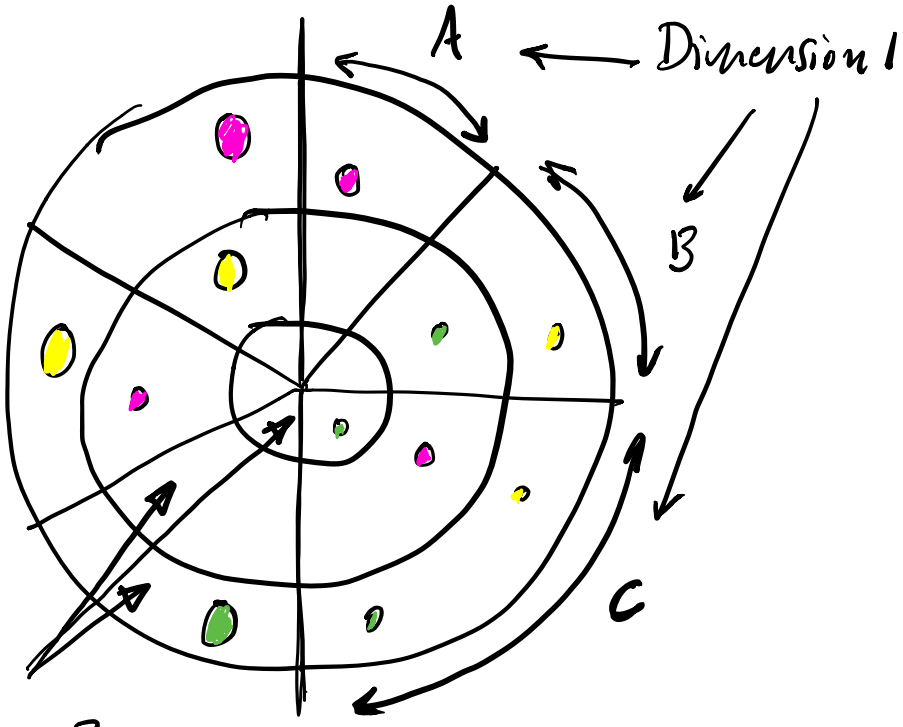
Bubble Chart - 4 Dimensions



Dimension 3 - Bubble Size

Dimension 4 - Bubble Color

Bullseye Chart - 4 Dimensions



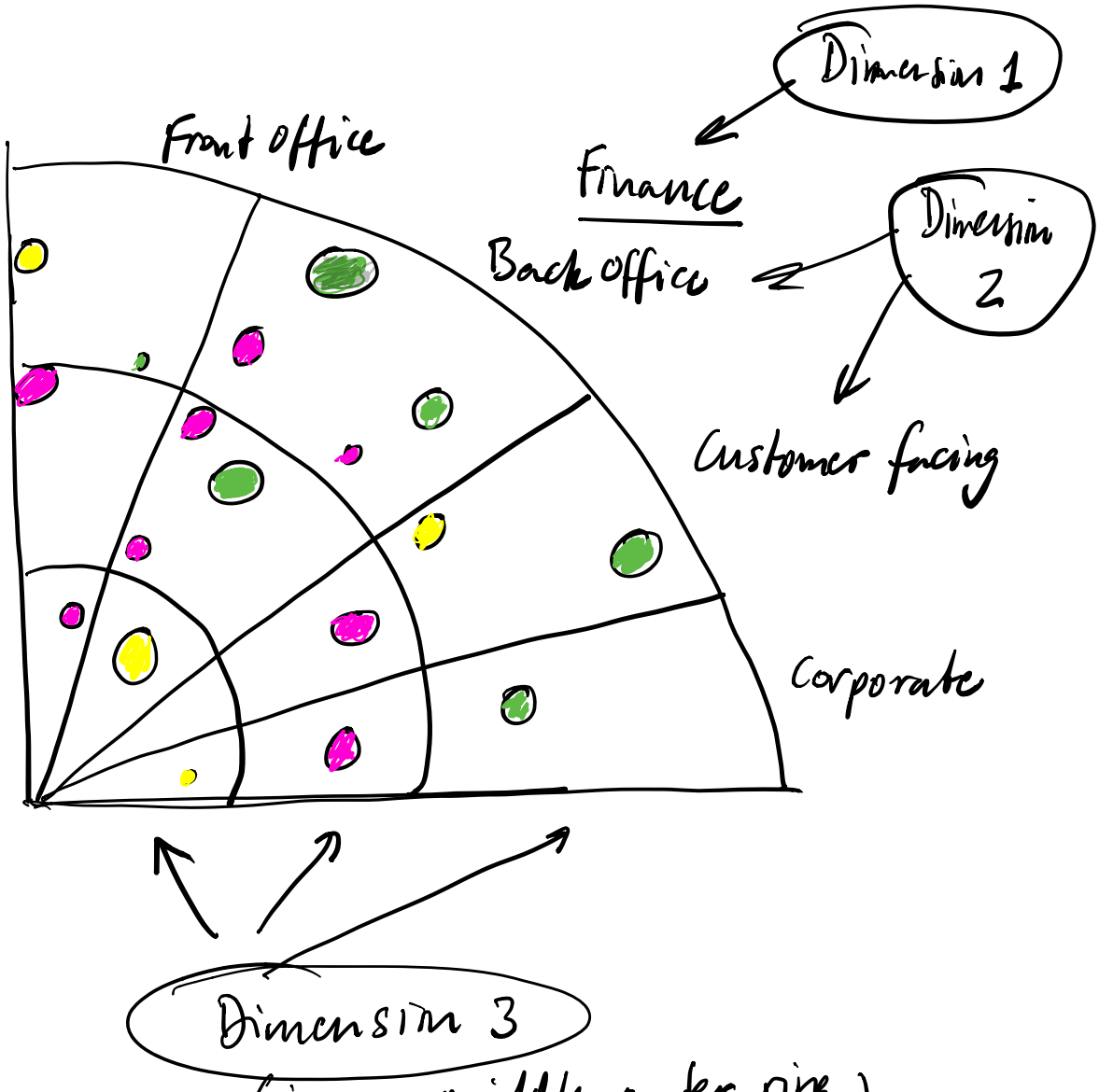
Dimension 2

(Outer, middle, inner ring)

Dimension 3 - Bubble Size o O O

Dimension 4 - Bubble Color

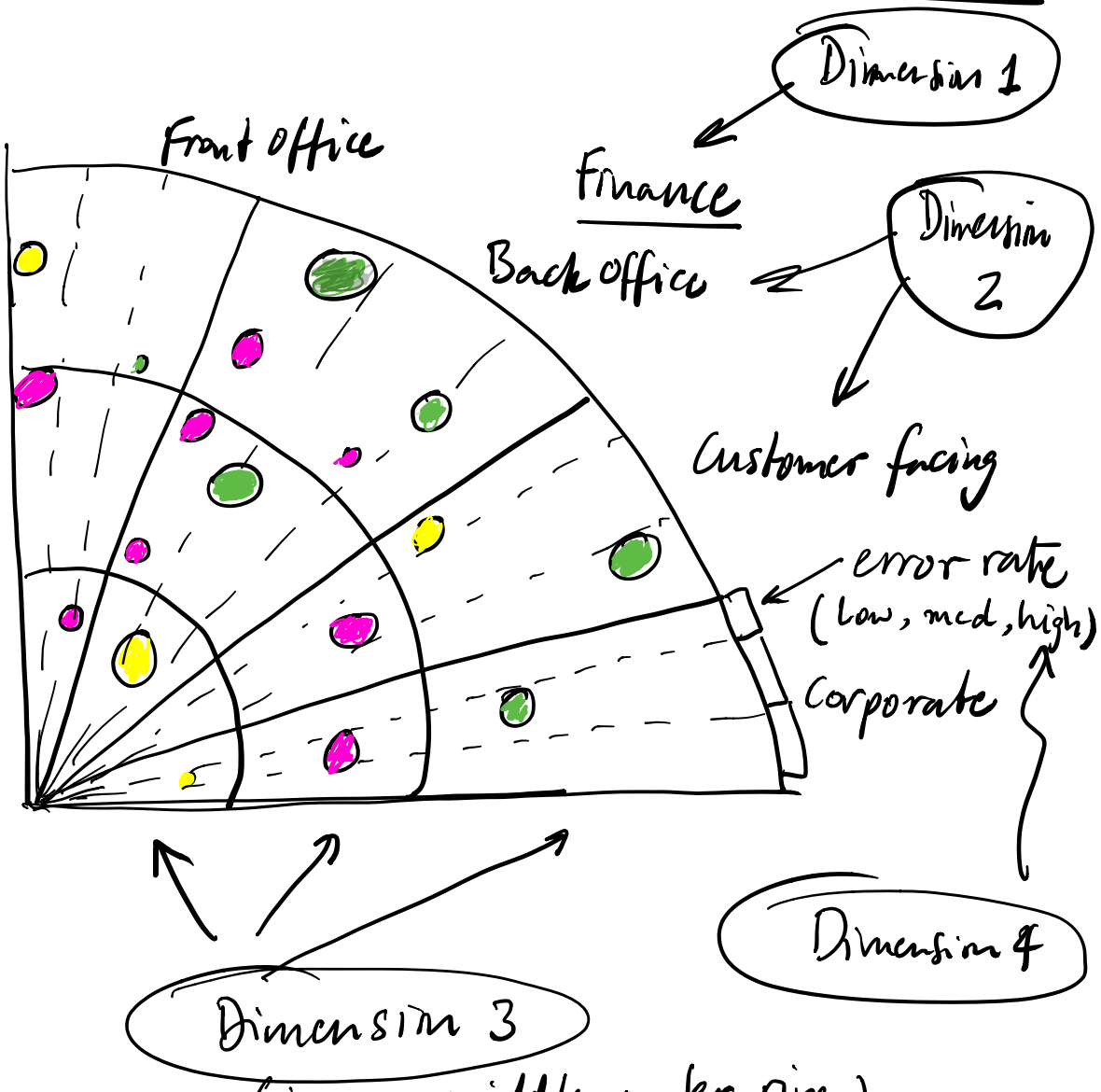
Bullseye Chart - 5 Dimensions



Dimension 4 - Bubble Size ○ ○ ○

Dimension 5 - Bubble Color ● ● ●

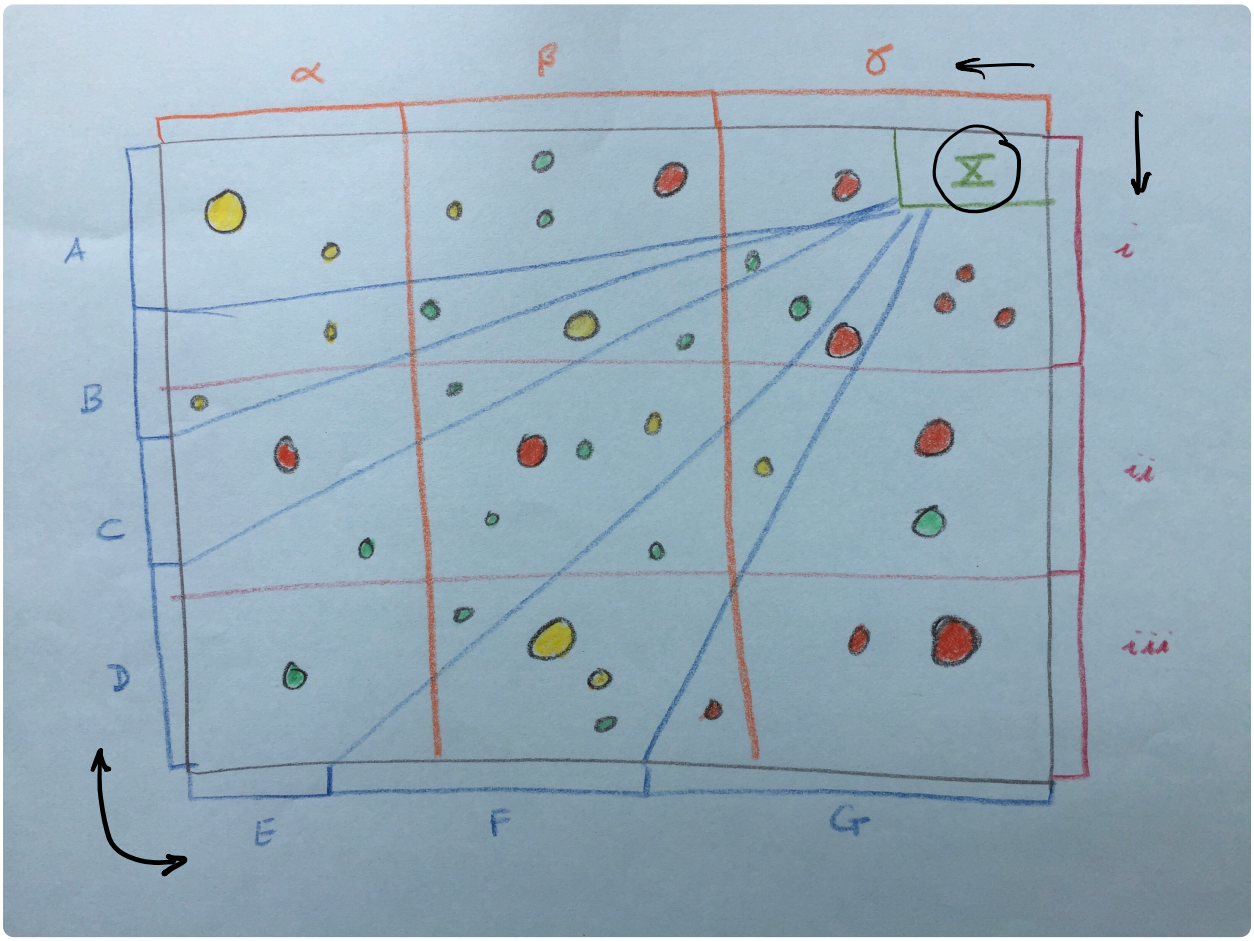
Bullseye Chart – 6 Dimensions



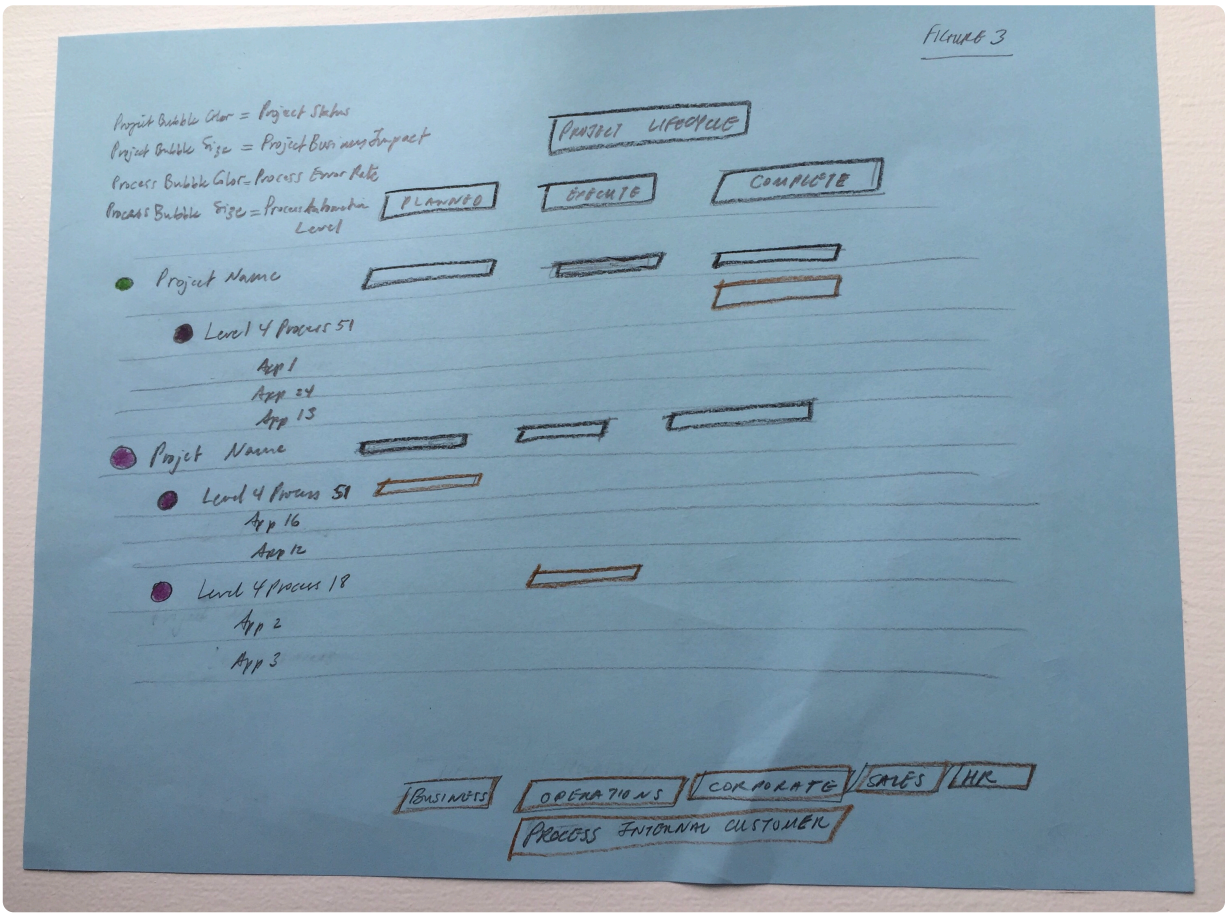
Dimension 5 – Bubble Size ○ ○ ○

Dimension 6 – Bubble Color

Roadmap Diagrams 6 + Dimensions



Gantt Chart – 7 Dimensions



But what if you have 10, or 20, or 100 dimensions?

(In machine learning problems it's not unusual to have 100,000 dimensions!)

Answer: Find out if a select handful of features matter more than the rest and use these features to visualize the dataset.

Some common techniques for finding relevant features (if they exist).

- Domain knowledge (useful but use with care)
- Convert / transform certain features and drop the rest
- Combine / transform one or more features into a single feature and drop the rest.
- Rank the relevance of features and choose the top n.
- Reduce the dimensionality of the dataset

Using compression algorithms.