A Story about Data

Avishek Sen Gupta
@avisheksengupta

Why are we here Anyway?

The aim of Data Analysis is:

- Insight
- Informed Decisions

The aim of Data Analysis is *not*:

- Pretty pictures
- Mysterious numbers
- Flailing around in N-space

Skills to Pay the Bills

Requires skills different from framework/API usage/knowledge, so don't get too cocky.

Skills to Pay the Bills

Being conversant in distributed systems is A Good Thing, but not Everything.

Skills to Pay the Bills

- These are Important Things to know:

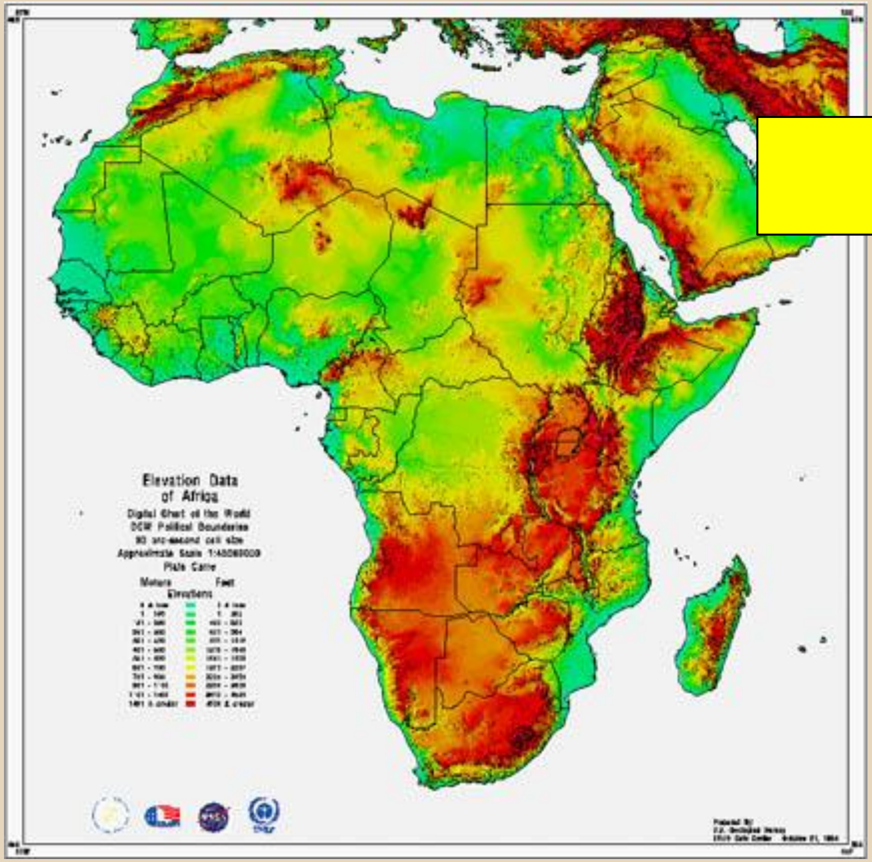
  - Basic (frequentist) statistics
  - Linear Algebra
  - R/Excel/any scripting language
  - Bayesian statistics
  - Basic acquaintance with numerical methods
  - Ability to read papers

Skills to Pay the Bills

Get into staring contests with equations: eventually, either you'll give up, or the equation will.

Meet the Cast

Continuous Data

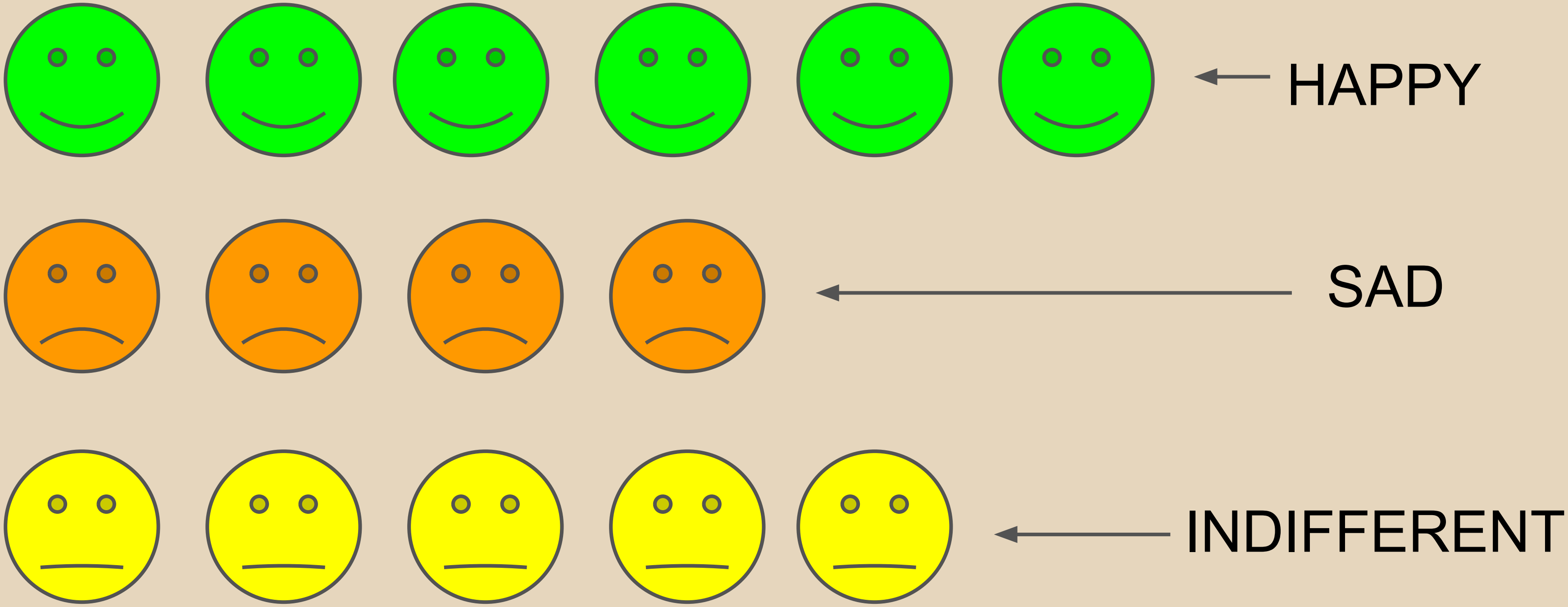


Why is this data continuous?

THINK!

Meet the Cast

Categorical Data



Meet the Cast

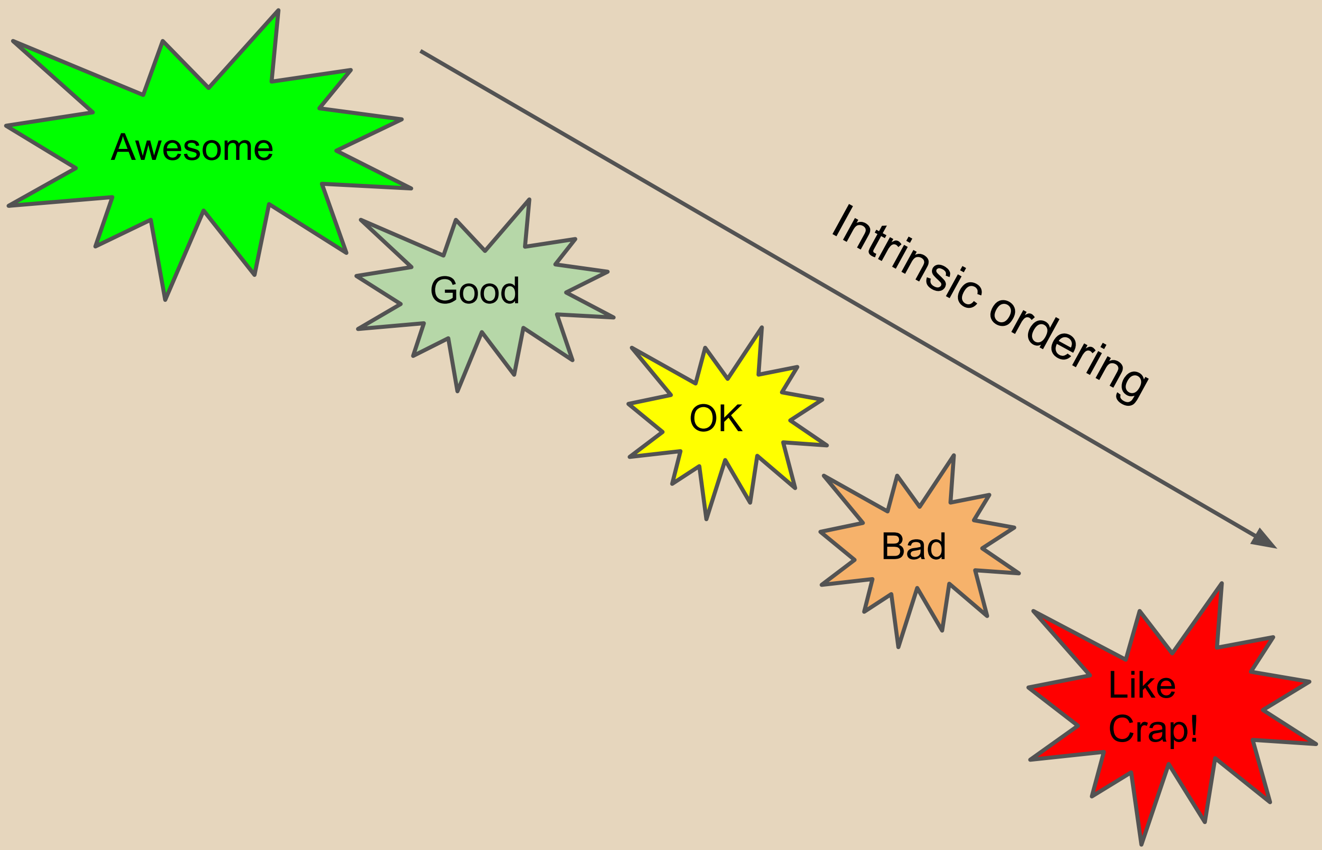
Whoa whoa whoa! Not so fast.

Two types of categorical data

- Ordinal
- Nominal

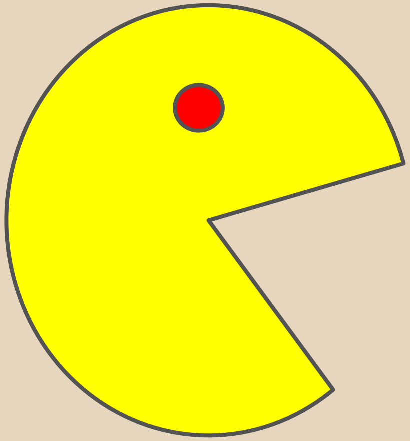
Meet the Cast

Ordinal Categorical



Meet the Cast

Nominal Categorical



Non-vegetarian

Vegetarian

No intrinsic ordering (unless you include politically incorrect jokes)

Frequentist Statistics

The Old Faithful

The Shape of the Data

We'd like to know if the squiggles representing our data can be represented more compactly.

Compact → Mathematical Representation

But Why??

Shorthand to indicate:

- Trend
- Future values (prediction)

The Shape of the Data

You have (broadly) two options.

- Plot the data itself.
- Plot the distribution of the data.

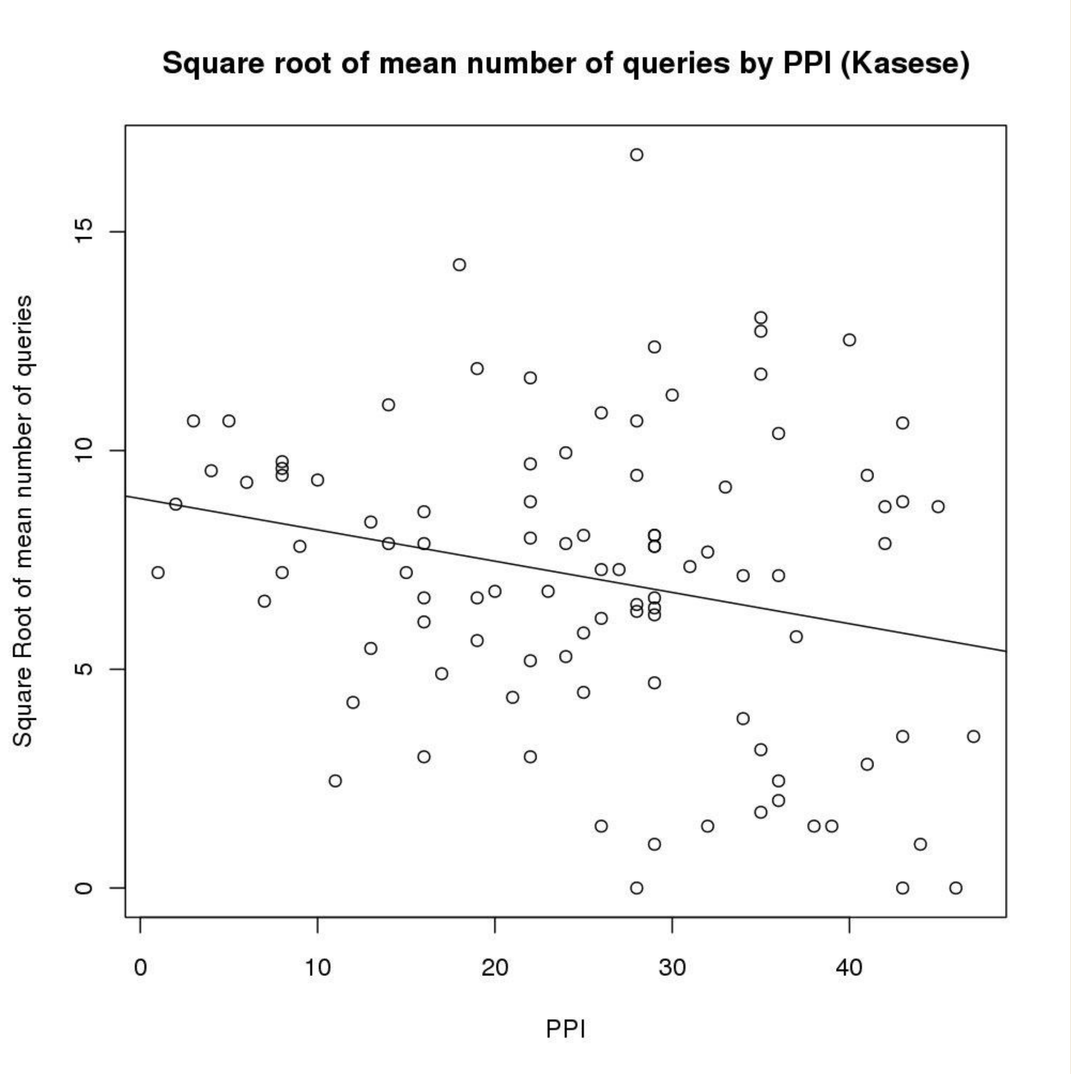
The Shape of the Data

There's tons of ways of visualising either the raw data, or it's summary.

- Box plots (distribution summary)
- Parallel coordinates (high-dimensional data)
- ...etc.

Visualisation is a massive field. Attend the visualisation session just after this one, for some more thoughts and discussion.

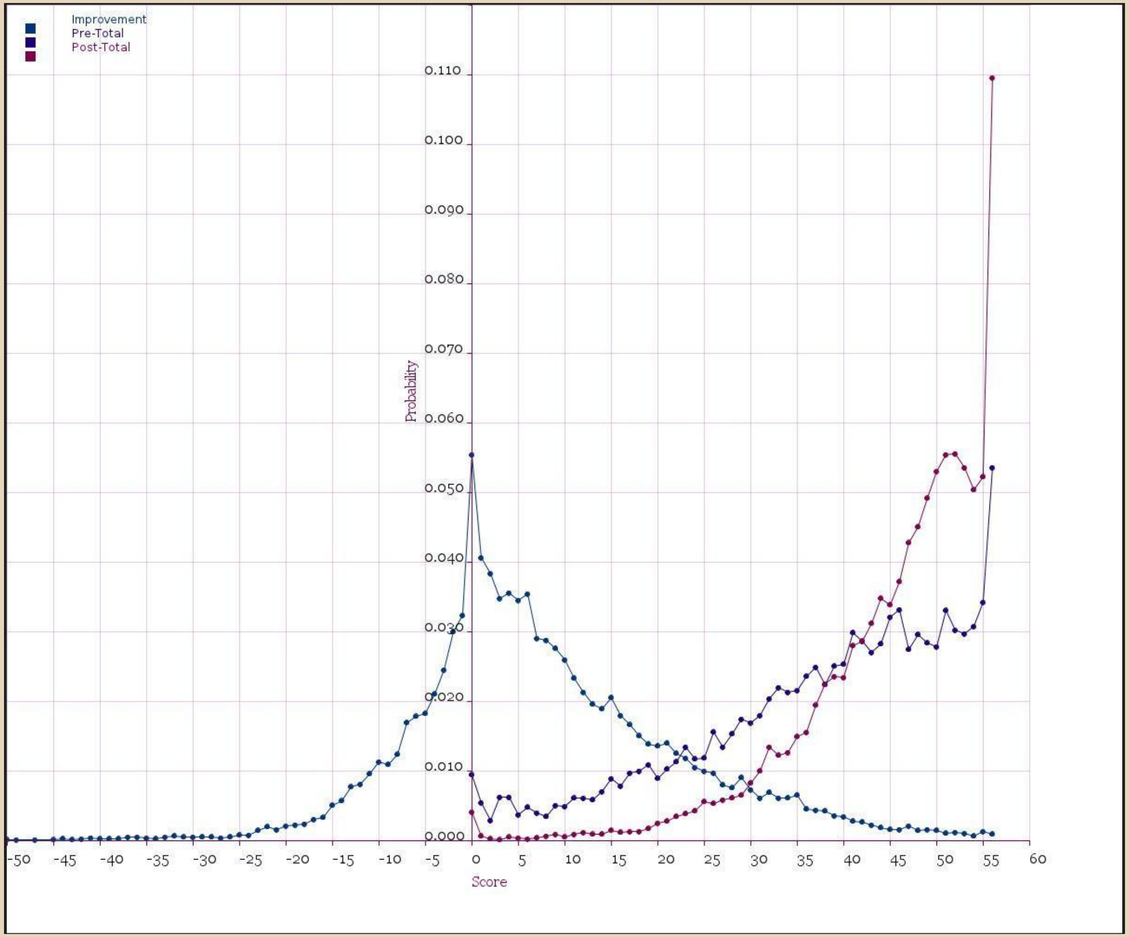
The Shape of the Data



Plotting the data itself

- Useful to get a quick intuitive sense of relationships (if any)
- Sometimes, variables may need transformation (more on this later).

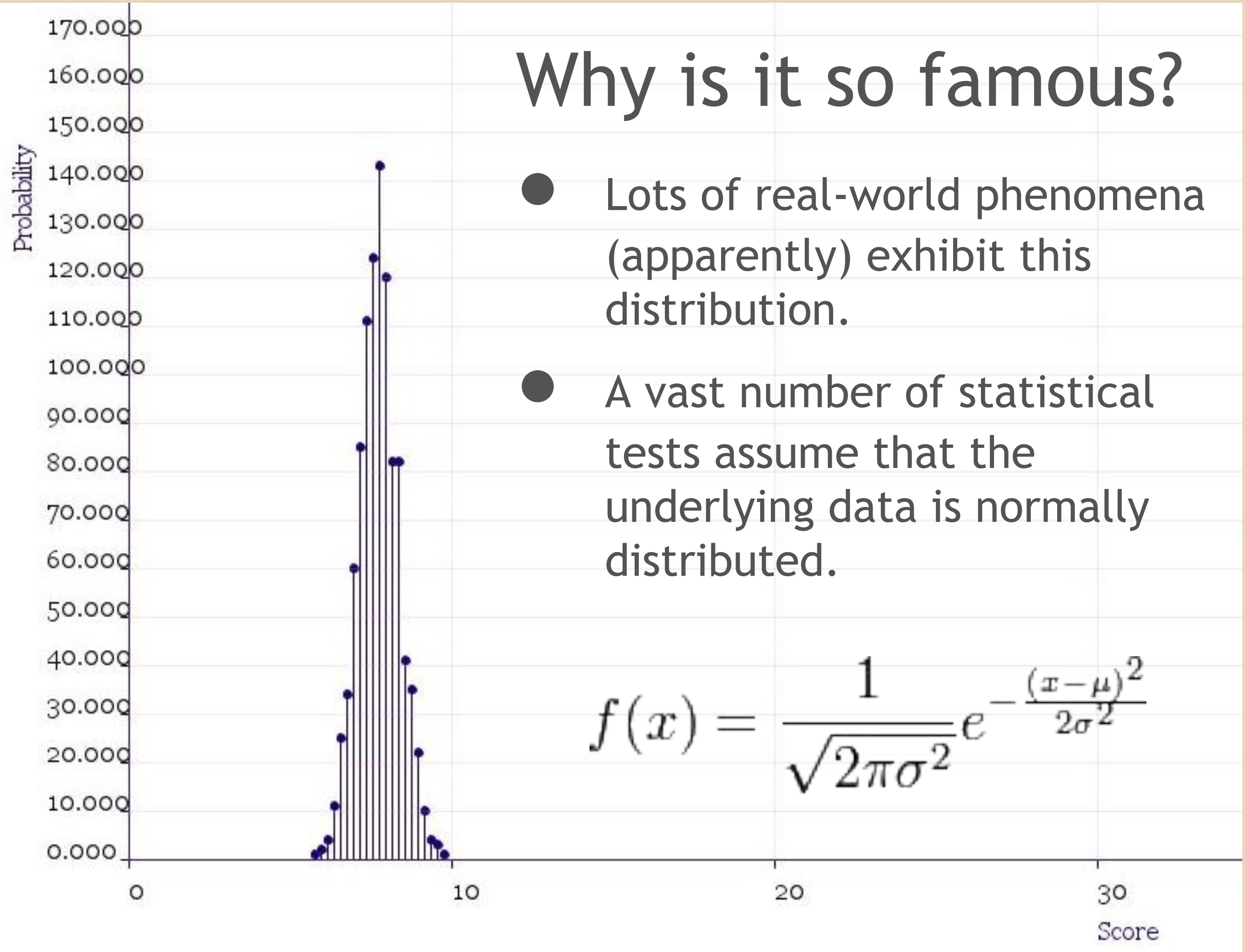
The Shape of the Data



Plotting the <u>distribution</u> of data

- Crisper representation of the data.
- Starting point for many tests and analyses.

The (in)famous Normal distribution



Why is it so famous?

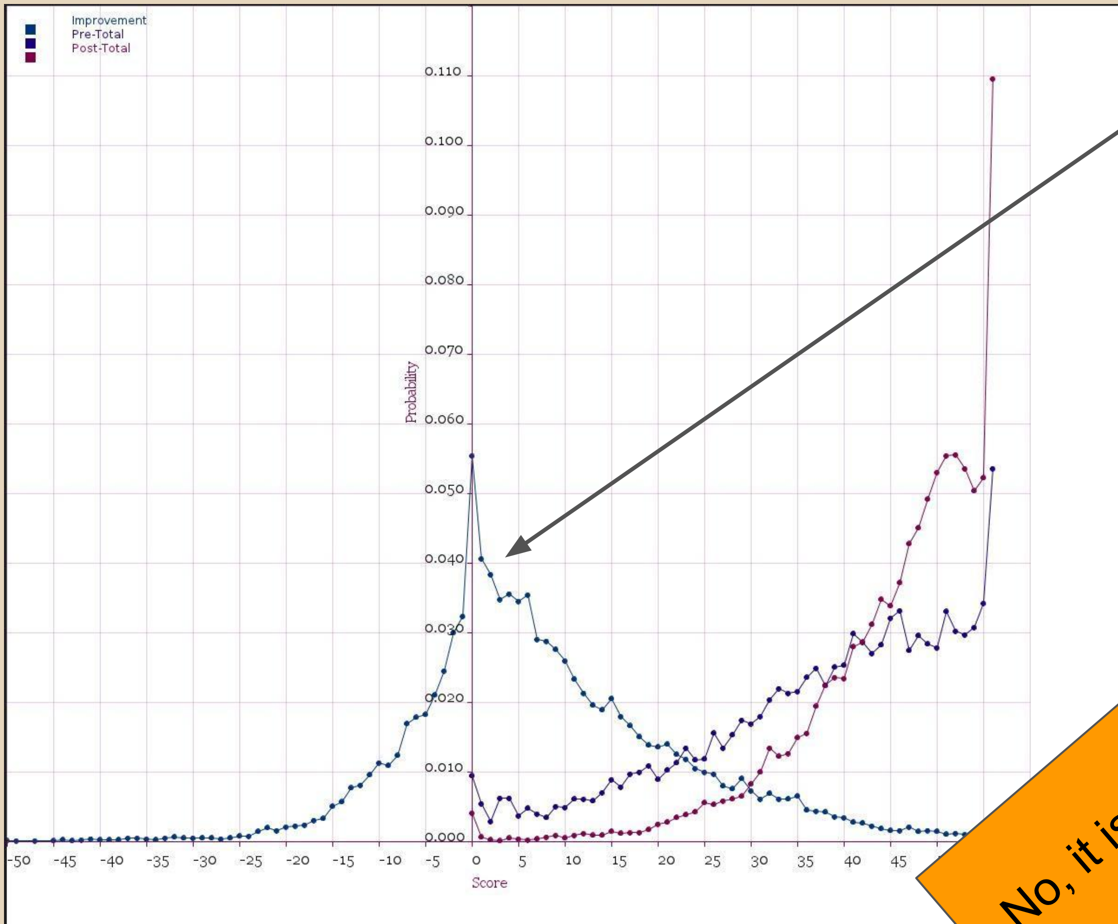
- Lots of real-world phenomena (apparently) exhibit this distribution.
- A vast number of statistical tests assume that the underlying data is normally distributed.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The sad truth about real-world distributions

- The Normal Distribution is only an approximation.
- Just because a distribution looks bell-shaped, does not imply that it is normally distributed.

The sad truth about real-world distributions



Is this distribution normal?

- It has only one peak.
- The values fall off on both sides.

No, it is *not* normal

Why?

How Normal?

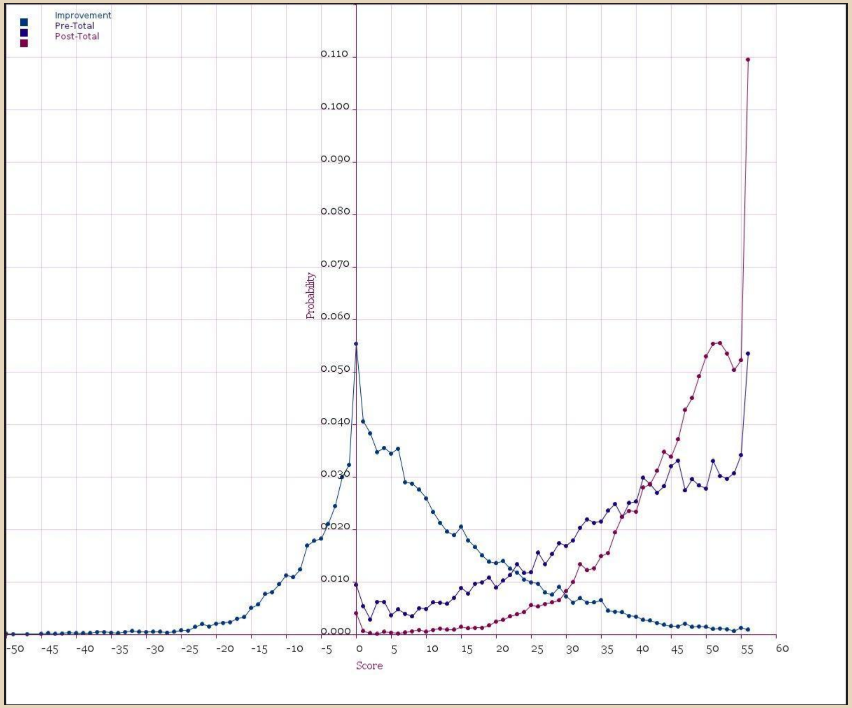
"How do I know of the Normal Distribution is a reasonable approximation of my data?"

3 Ways

How Normal?

"How do I know of the Normal Distribution is a reasonable approximation of my data?"

Visual inspection



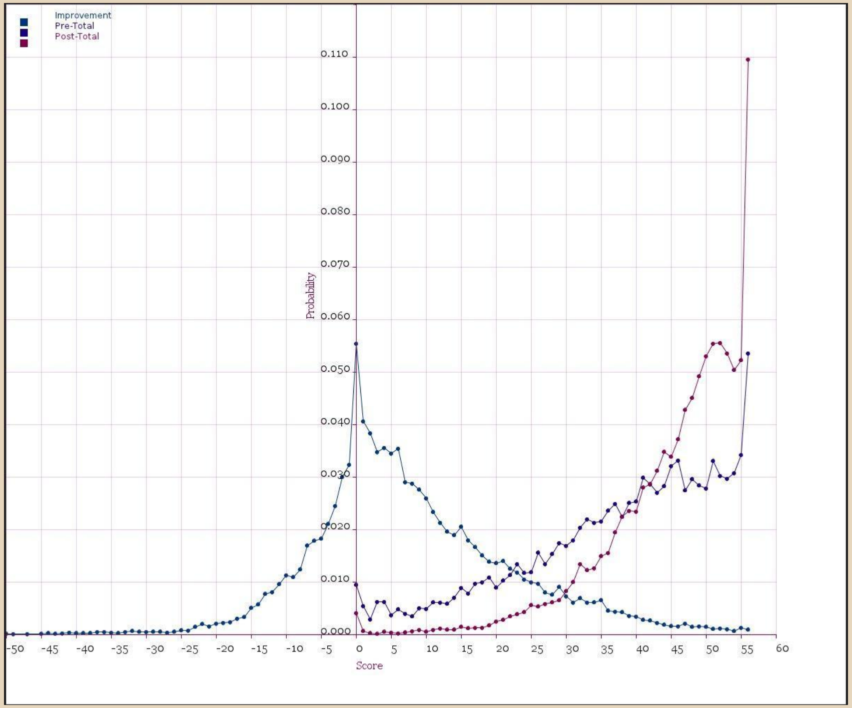
Just Look. And Judge.

Warning: your eyes may deceive you.

How Normal?

"How do I know of the Normal Distribution is a reasonable approximation of my data?"

Analytical tests

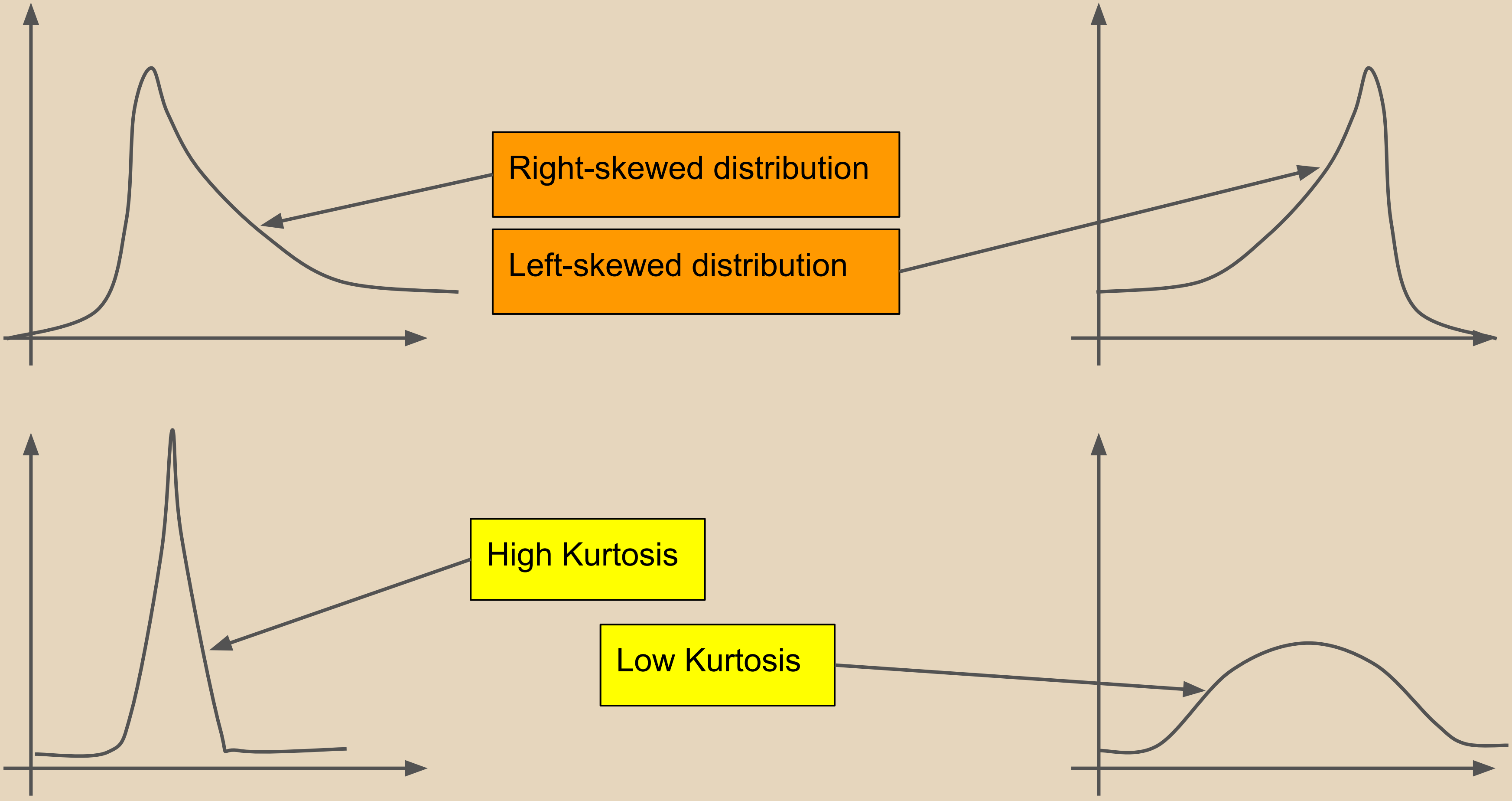


These tests use skew and kurtosis

Skew is a measure of how lopsided a distribution is.

Kurtosis is a measure of how peaked a curve is.

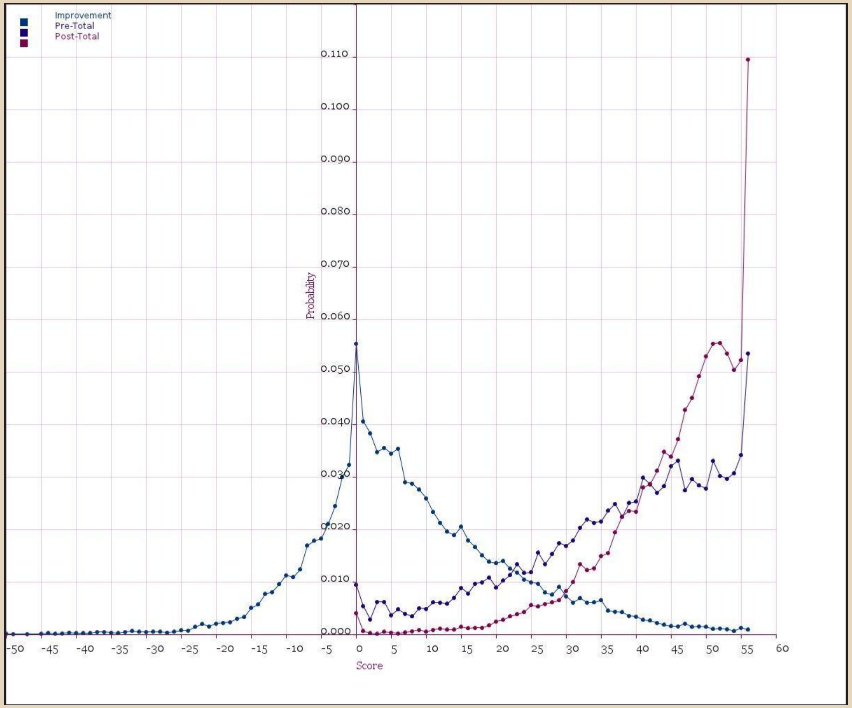
How Normal? (Skew and Kurtosis)



How Normal? (Analytical Tests contd.)

"How do I know of the Normal Distribution is a reasonable approximation of my data?"

Analytical tests



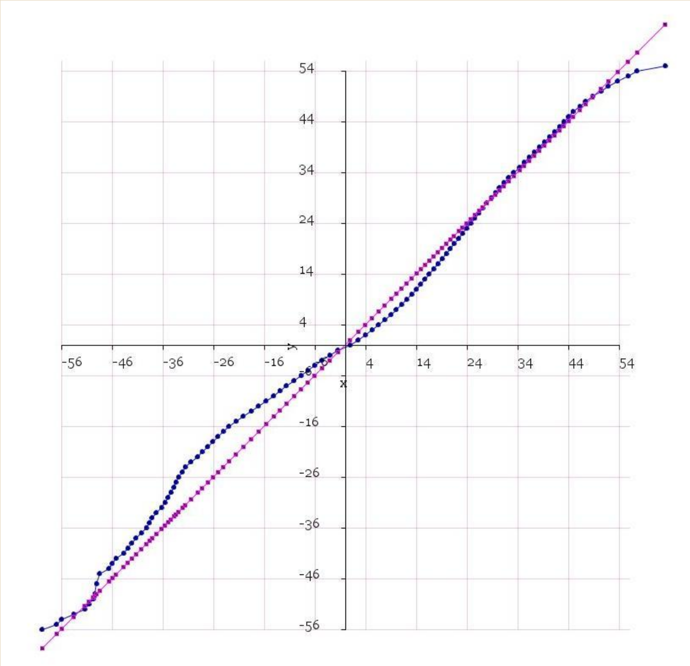
These tests combine skew and kurtosis to produce a normality metric.

eg: Jarque-Bera test, Anderson-Darling test

How Normal?

"How do I know of the Normal Distribution is a reasonable approximation of my data?"

Quantile-Quantile Plot



Maps the normal distribution to a straight, diagonal line

Maps your data using the same logic.

Check for deviations.

Not Normally distributed? Deal with it.

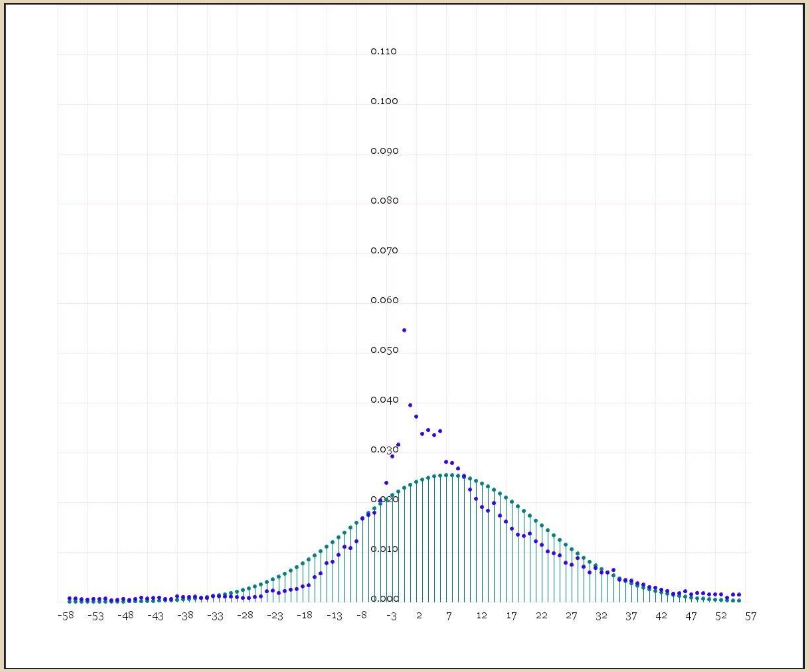
- Real-world data is not pretty.
- The Normal distribution is just an approximation.
- Don't try to force it; if it's not normal, it's not.

So, what can we do?

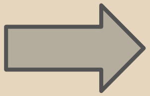
Not Normally distributed? Deal with it.

Idea: Transform the variable(s)

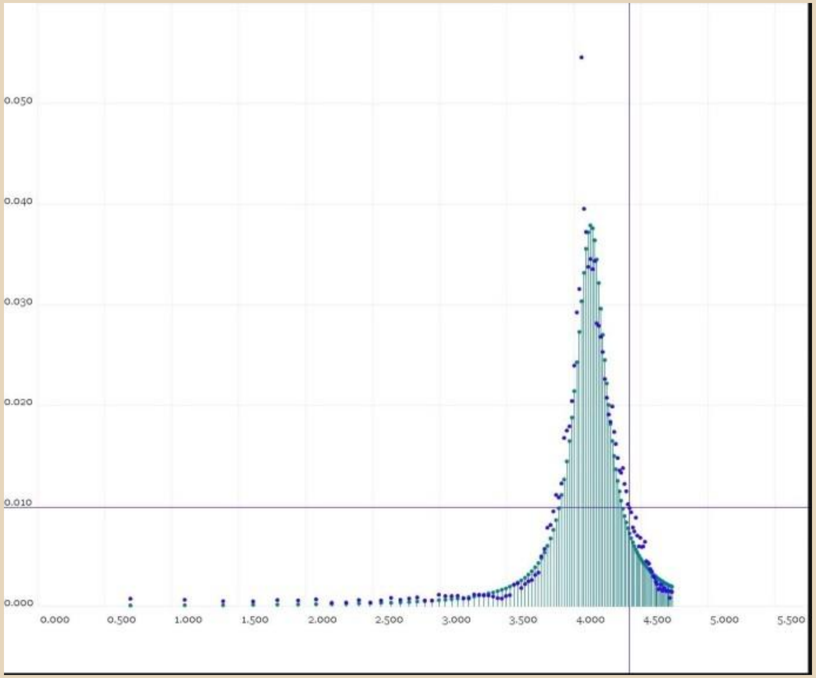
Best fit before



Log (x)



Best fit after



Not Normally distributed? Deal with it.

Idea: Transform the variable(s)

- You are not restricted to using a specific transformation.
- Use square-root, inverse, inverse cube, anything.

Box-Cox Transform

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ log(y_i), & \text{if } \lambda = 0 \end{cases}$$

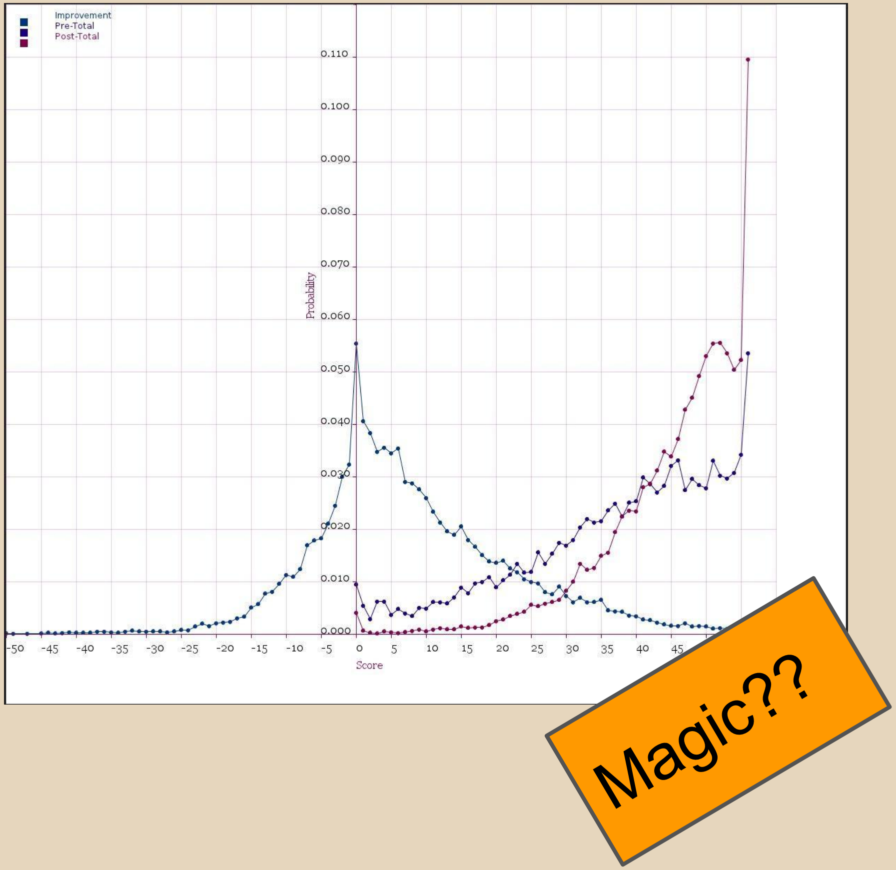
Not Normally distributed? Deal with it.

If none of this works...

- Screw the distribution. Who cares if it's Normal or not?
- Go for the jugular...that is, the mean.
- Bring out the heavy guns.

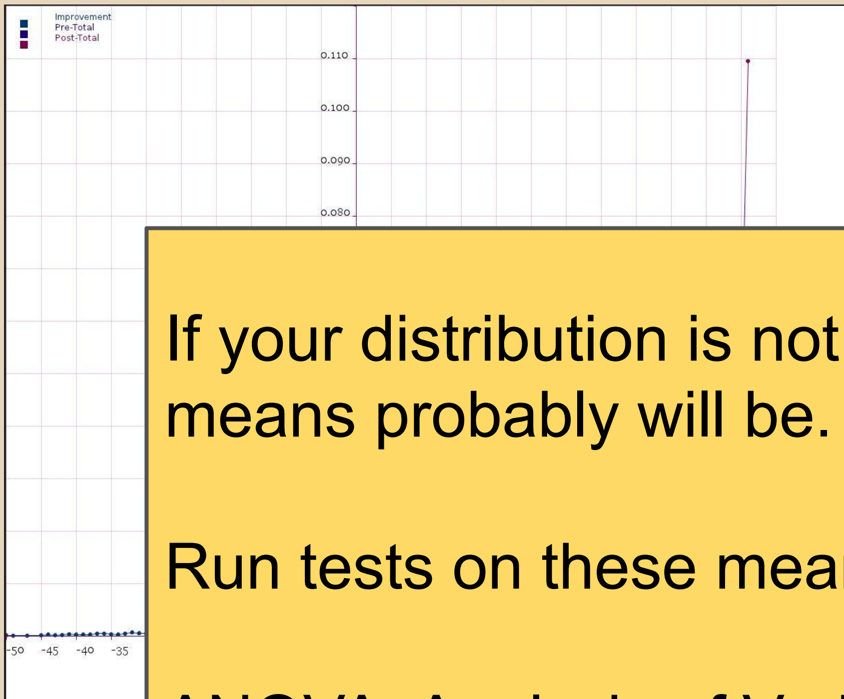
The Central Limit Theorem

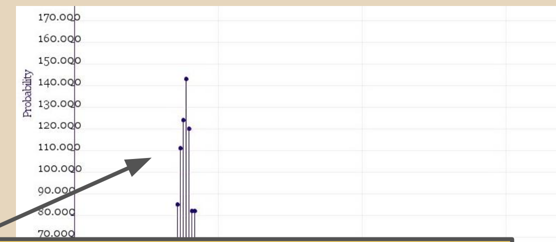
Central Limit Theorem

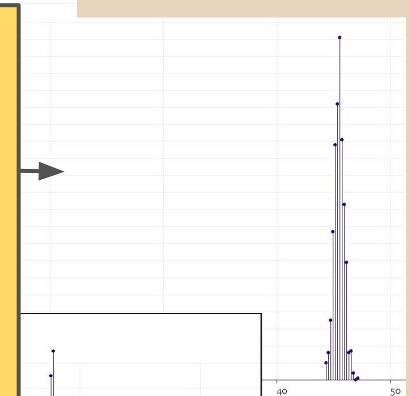


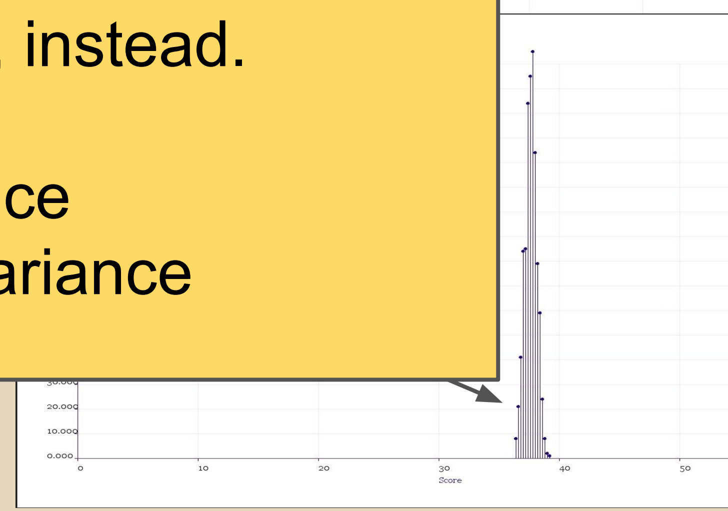
- Take any data set.
- Grab 20 random data points from it.
- Find the mean value of this subset. Call it M1.
- Repeat the above to get M1, M2, M3,..., Mn.
- These means will be distributed normally, irrespective of how your original dataset was distributed.

Central Limit Theorem









If your distribution is not normal, your means probably will be.

Run tests on these means, instead.

ANOVA: Analysis of Variance
ANCOVA: Analysis of Covariance

Answering questions : Hypothesis Tests

- Was this drug effective?
- Did this session help?
- Did the shock therapy make you less psychotic?

What if we want to answer such questions?

Formulate the question properly, sir!

Answering questions : Hypothesis Tests

- Was this drug effective?
- Did this session help?
- Did the shock therapy make you less psychotic?

Formulate the question properly, sir!

How do we restate this problem mathematically?

Answering questions : Hypothesis Tests

Formulate the question properly, sir!

How do we restate this problem mathematically?

Example: A drug is being field tested. A group of volunteers (some have the disease, some don't) take the drug. After the trial period, they are re-tested for the condition.

Answering questions : Hypothesis Tests

Formulate the question properly, sir!

How do we restate this problem mathematically?

Both groups have the same people.

There are two 'groups'

- The first group represents t[illegible]e of the volunteers before the drug t[illegible].
- The second group represents the state of the volunteers after the drug trial.

Answering questions : Hypothesis Tests

There are two 'groups'

- The first group represents the state of the volunteers before the drug trial.
- The second group represents the state of the volunteers after the drug trial.

So, how are the people within these groups distributed?

Key insight: If the drug had no discernible effect, the populations within these two groups will follow a Chi-Square distribution.

Answering questions : Hypothesis Tests

Contingency Table for Drug Test example

When/Condition	After:Present	After:Absent
Before: Present	A	B
Before: Absent	C	

- Chi-Square Test
- McNemar's Test for Matched Pairs
- Fisher's Exact Test

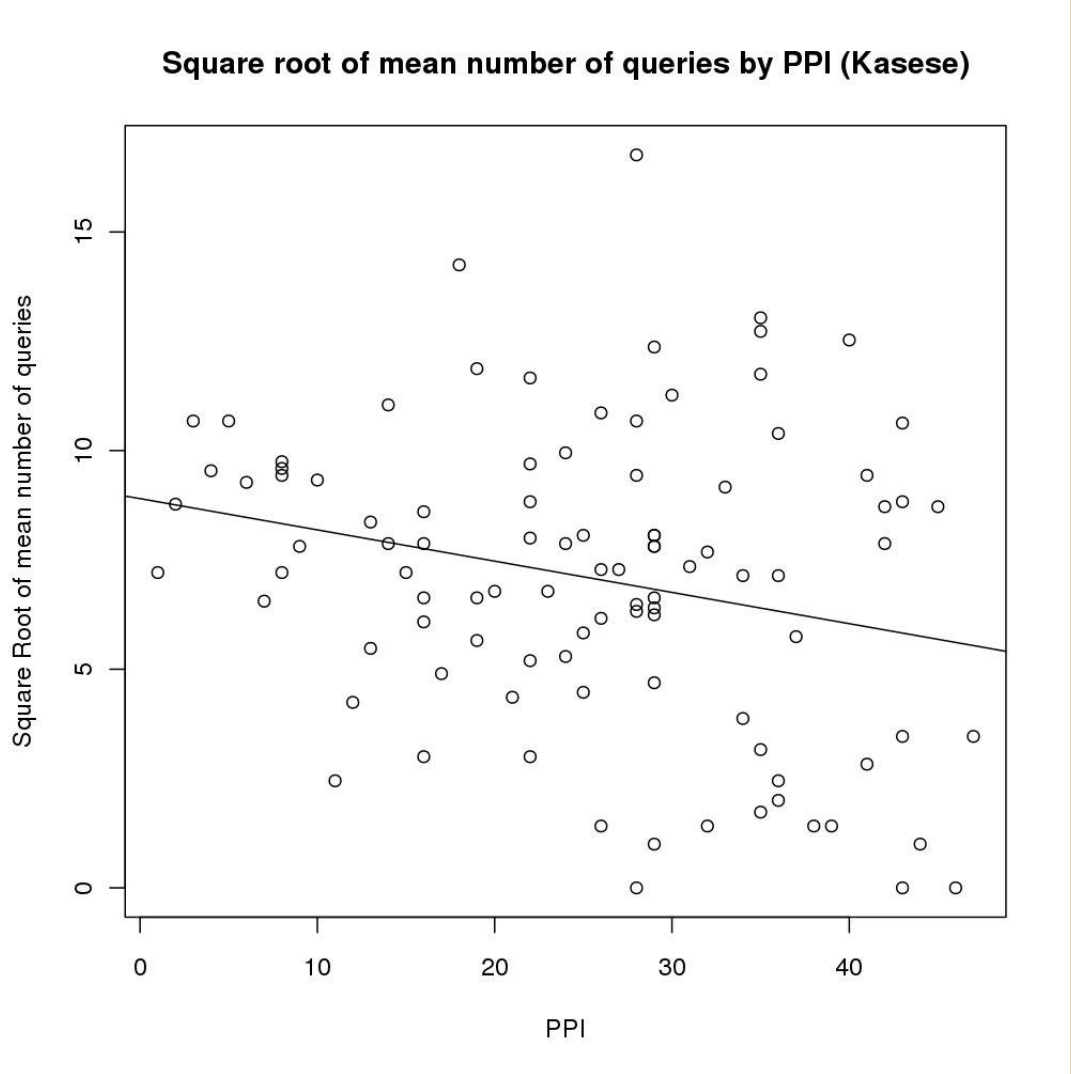
Hypothesis tests start out with the hyp are independently distributed, using tables
This hypothesis is then either rejected, or not.

'Classical' Predictors and Models: The Minefield starts Right Here

Use these tools. Use them wisely.

- Linear Regression
- Decision Trees
- Logistic Regression
- Loglinear Analysis
- ...more

Linear Regression



Find the best predictor, for:

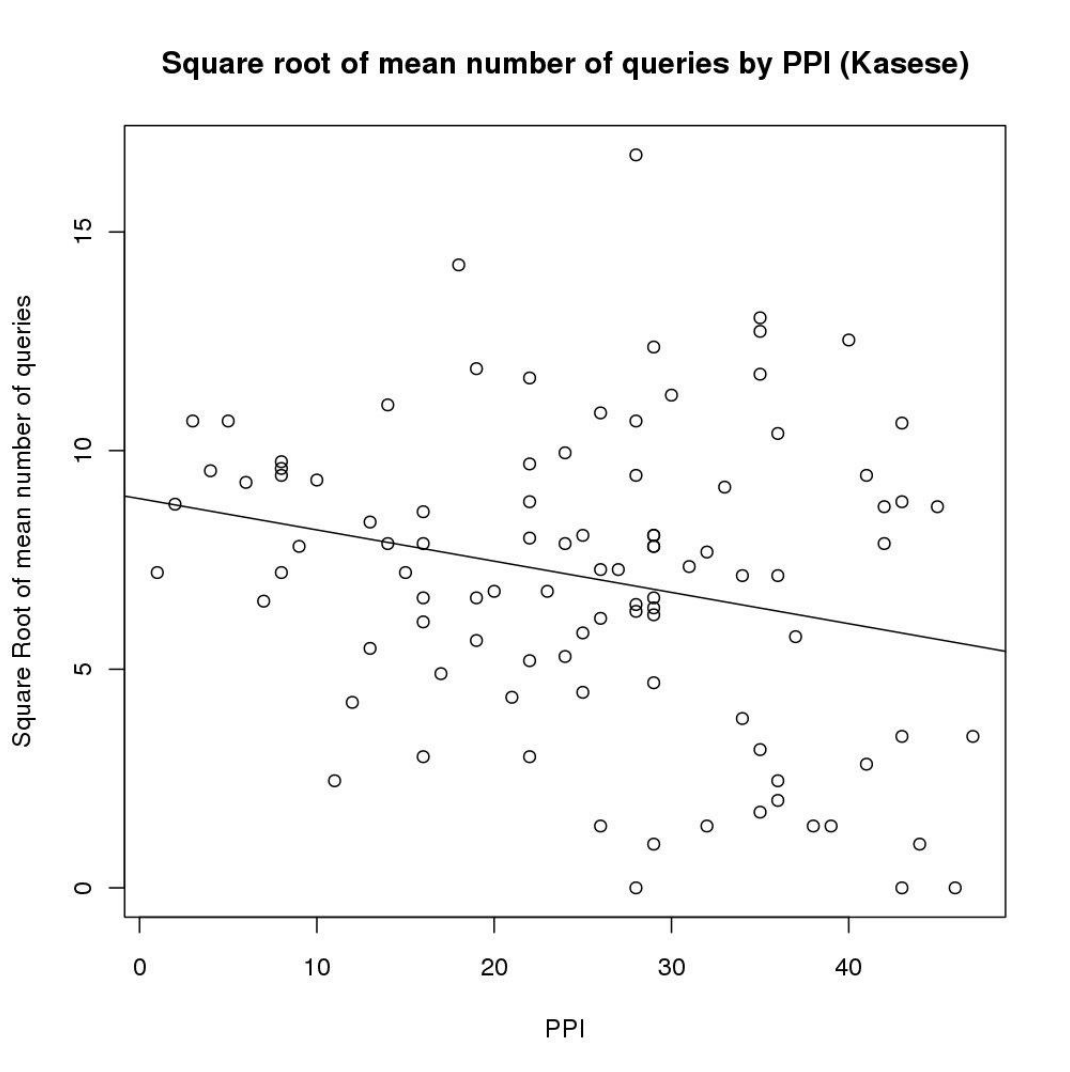
y=mx + c

Seems simple, intuitive, has a closed-form solution.

Yes?

Maybe...

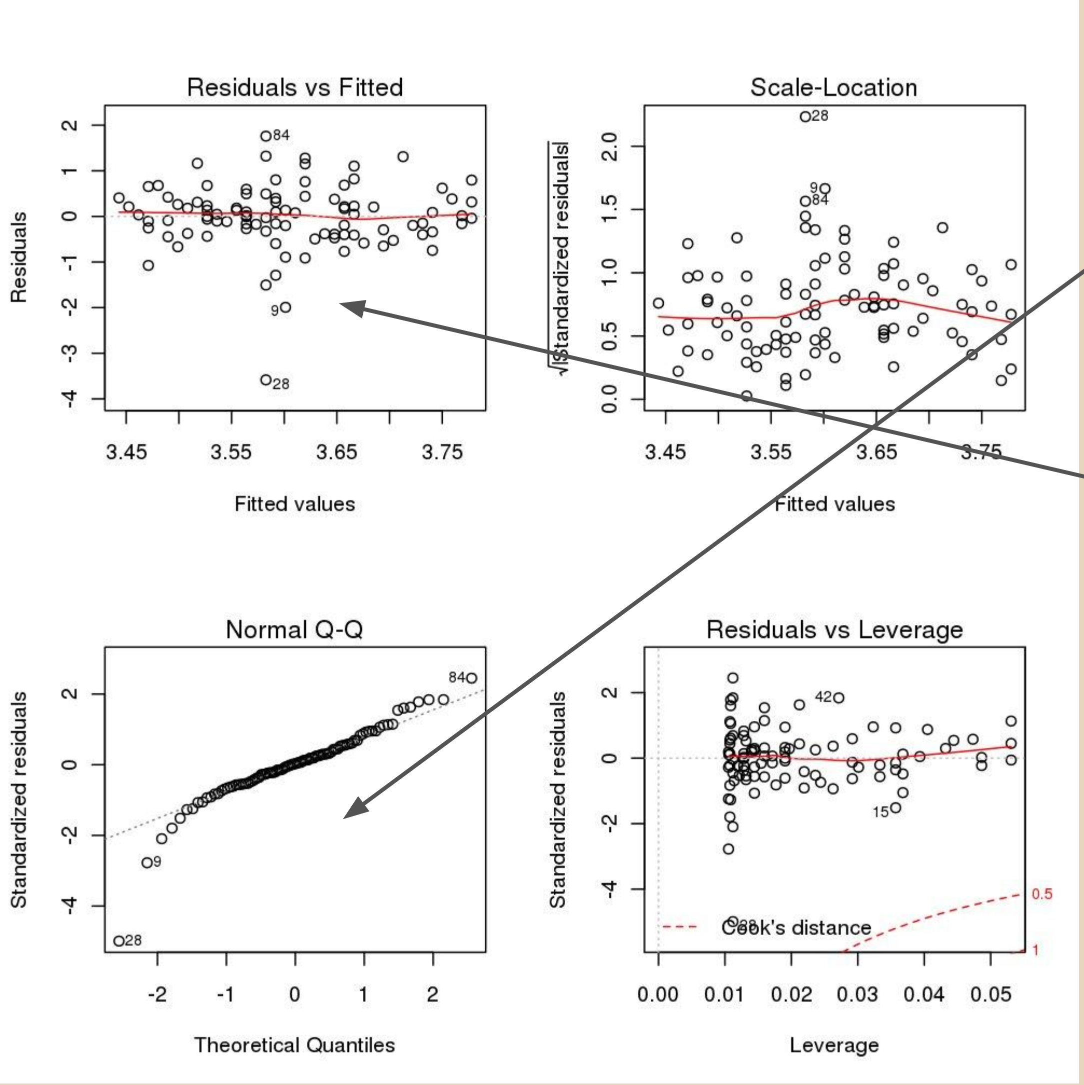
Linear Regression



Use Linear Regression carelessly, and your results will be:

- At best, misleading.
- At worst, plain wrong!

Linear Regression

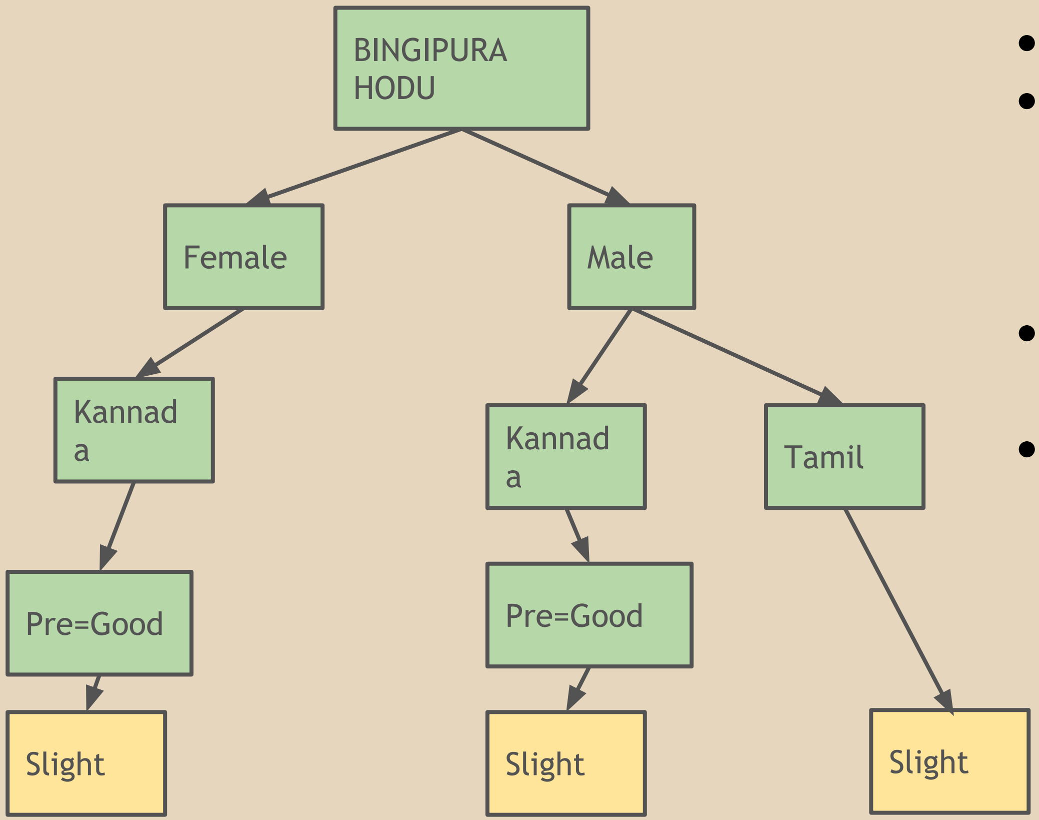


Assumptions

- Residuals must be normally distributed.
- Residuals show no discernible trend.

Decision Trees

"Given that your height is 5'2", and you are cross-eyed, are you an orangutan?"



- Deterministic
- Uses a metric called Information Gain to decide the hierarchy of attributes to branch on.
- Must be reconstructed every time new data is added.
- May not be able to resolve all scenarios.

Binomial/Logistic Regression

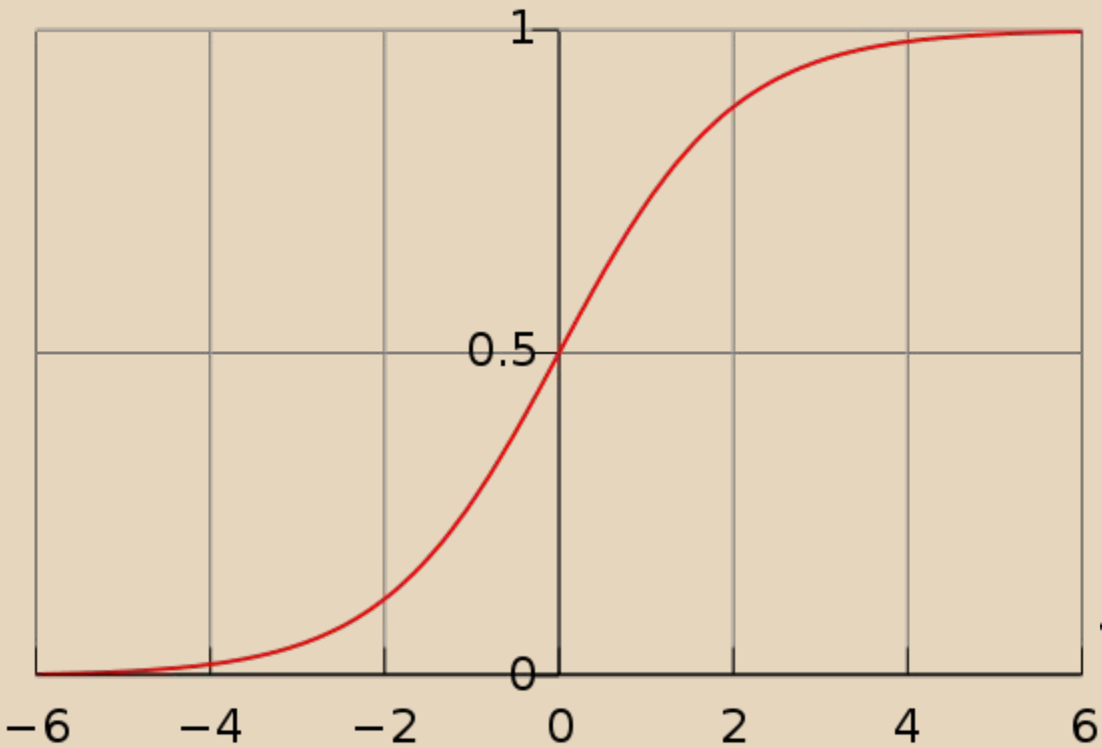
Scenario: Calculating the risk of heart disease

Assume that we want to calculate the risk of heart disease, given:

- Age
- Sex
- Cholesterol level

→ Factors

Binomial/Logistic Regression



Use the logistic curve as a model for risk.
Why?

- It is bounded within [0,1].
- It can support multiple variables.

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k,$$

Factors

Bayesian Statistics

The Snarling Dog on the other side of the Ring

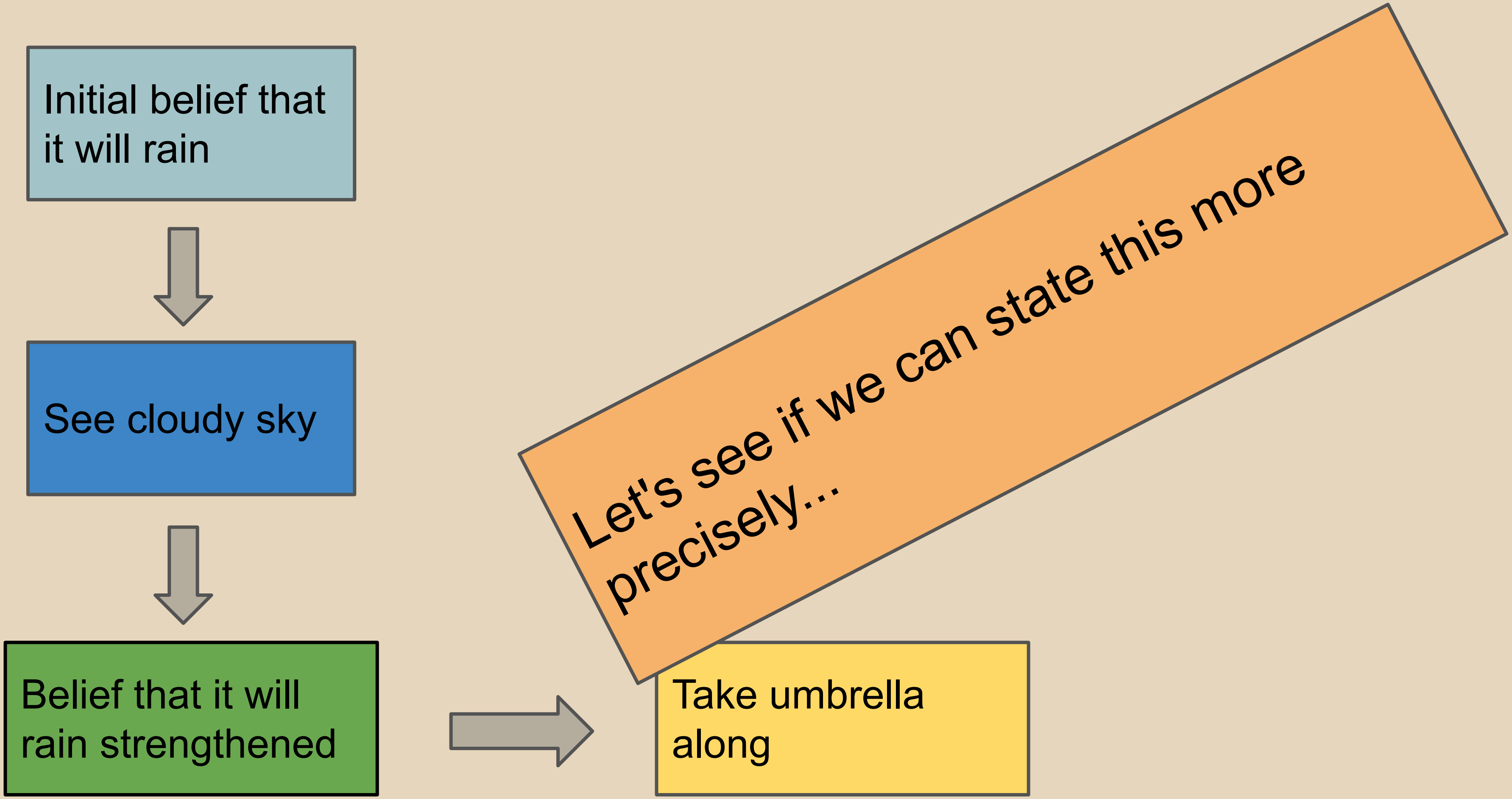
Bayesian Statistics

Person 1: "Why did you bring an umbrella if it's not raining?"
Person 2: "The sky is overcast, you idiot."
Person 1: "So you *believe* that i[illegible]ht rain?"
Person 2: "...maybe."

What is evident in the above conversation?

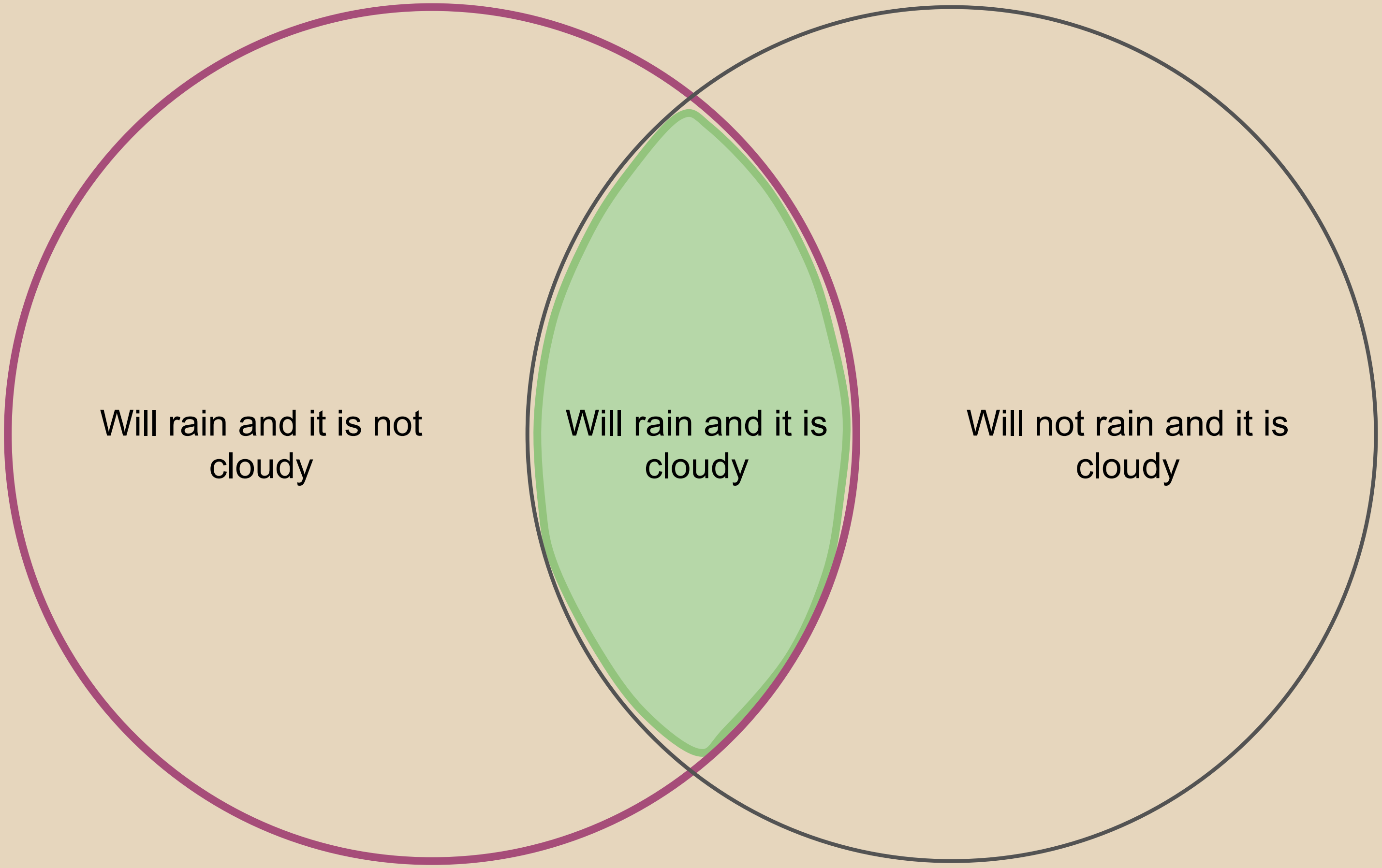
- It is not raining right now.
- It *is* cloudy.
- The overcast day increases Person 2's belief that it will rain.

Bayesian Statistics



Bayesian Statistics

Back to school, folks!



Bayesian Statistics

P(A) = Belief that it will rain, sans any evidence.
P(B) = Belief that it will be cloudy.
P(A.B) = Belief that it is cloudy and it will rain
P(A|B) = Belief that it will rain, given that *is* cloudy.

$P(A|B) = P(A.B)/P(B)$

Bayes Theorem

Using the same logic:

$P(B|A) = P(A.B)/P(A)$

This, gentlemen, right here is the foundation of Bayesian Statistics.

Therefore:
$P(A|B).P(B) = P(B|A).P(A)$

Why Bayesian Statistics?

...or, more informally, why the f**k did you bring an umbrella when it's not raining?

- Allows us to incorporate our "hunch".
- Can integrate conflicting opinions for analysis.

Uses:

- Classification (spam filters, etc.)
- Prediction (Stochastic models for stock price modelling)
- etc.

Bayesian Classifiers

Answers questions of the form:

"Given that I have some information about a potential event E, how probable is it that E will occur?"

Examples:

"Given that this student has gotten a First Class in CS and OOPs, what is the probability that she will score well in Databases too?"

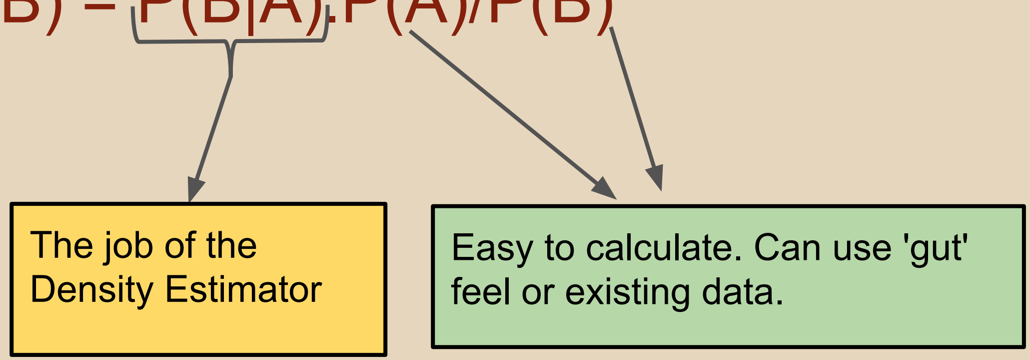
"Given that this patient has high cholesterol and is aged 65, what is the probability that he will suffer from a heart attack in the next 2 years?"

Density Estimators

Remember this? This is your standard Bayes Classifier.

$$P(A|B).P(B) = P(B|A).P(A)$$

$$\Rightarrow P(A|B) = P(B|A).P(A)/P(B)$$



Density Estimators

Example

A1	B1
A2	B2
A3	B3
A1	B3
A2	B1

P(A=A1|B=B1) = ?
How many cases of B=B1 are there? 2.
They are (A1, B1) and (A2, B1).

Now within this reduced set, how many times does A=A1 occur? Once, for (A1, B1).

Therefore P(A=A1|B=B1) = 1/2 = 0.5

In cases with more number of variables, or continuous data, such a calculation will be either tedious or downright impossible.

Density Estimators

Example : The simplest discrete density estimator

A1	B1	0.5
A1	B2	0
A1	B3	0.5
A2	B1	0.5
A2	B2	0.5
A2	B3	0
A3	B1	0
A3	B2	0
A3	B3	1

P(A=A1|B=B1) = ?
However,
Another way is:
P(A=A1|B=B1) = P(B=B1|A=A1).P(A=A1)/P(B=B1)
P(A=A1)= 2/5 = 0.4
P(B=B1)= 2/5 = 0.4
From the table to the left, we get:
P(B=B1|A=A1) = 0.5

P(A=A1|B=B1) = 0.5 * 0.4 / 0.4 = 0.5

Density Estimators

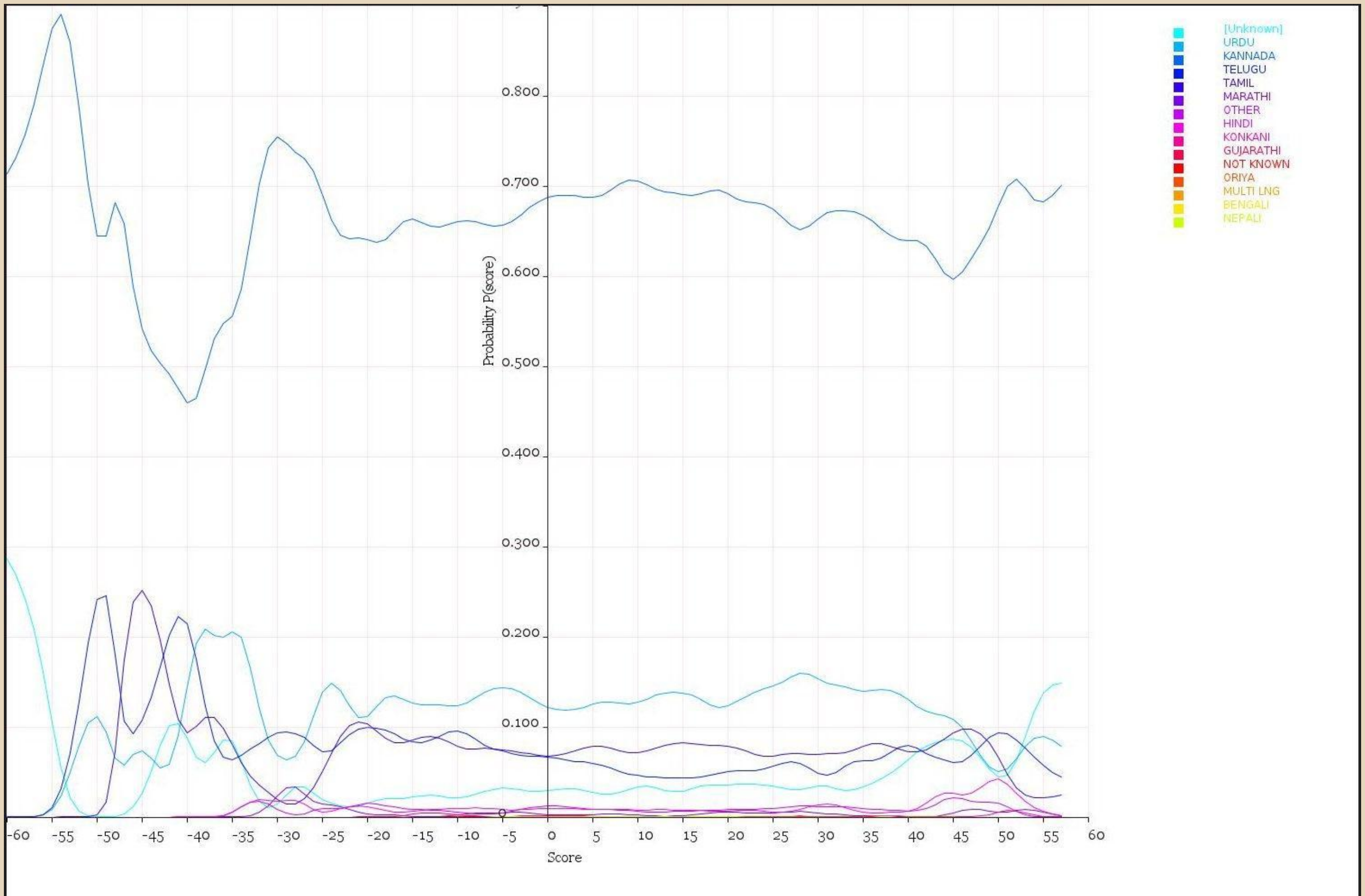
Different types of Density Estimators

- Joint DE
- Naive DE (suitable for massive number of attributes)
- Kernel DE
- Gaussian DE
- Bayesian Belief Networks

Plug in the Density Estimator of your choice into the Bayes Classifier to perform classification tasks.

Bayesian Classifier

Bayesian Classifier curve P(Language | Score)



Dimensional Reduction

...because only an Idiot assumes that Every Detail is Relevant.

Say you have:

- One million data p
- Each data poi iables.

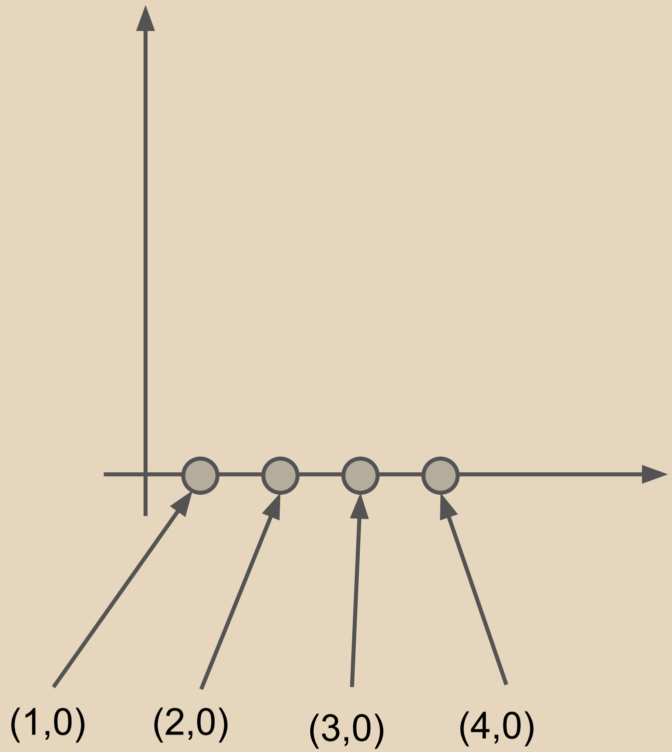
...and you analysis / classificatio cl g, etc.

There's no way to put this delicately:
You. Are. Screwed.

Not everything is parallelisable.
Brute force won't solve everything.

Dimensional Reduction

A Simple Example



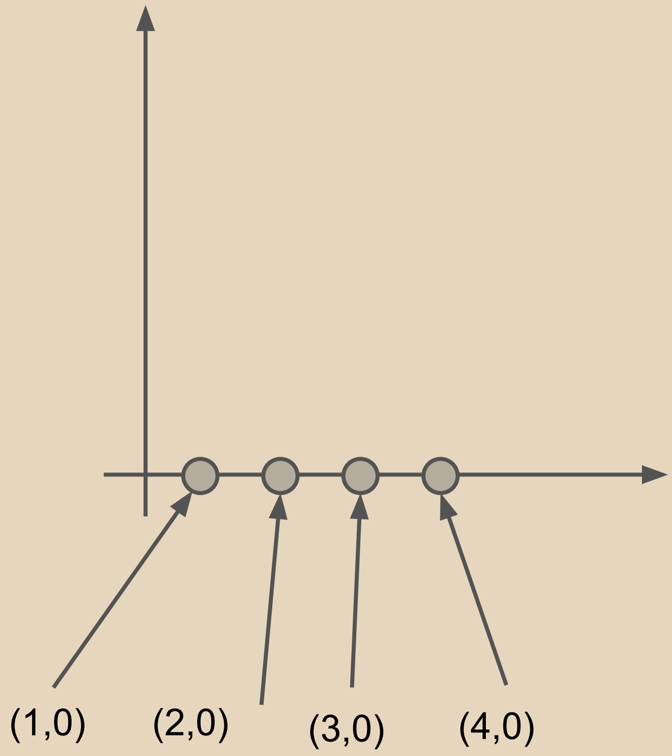
So, we have these 4 points.
They are on the 2D plane.
Each of them is represented by a pair of numbers.

Any way we c... de them to rep... sser

Think of this as a compression problem.

Dimensional Reduction

A Simple Example

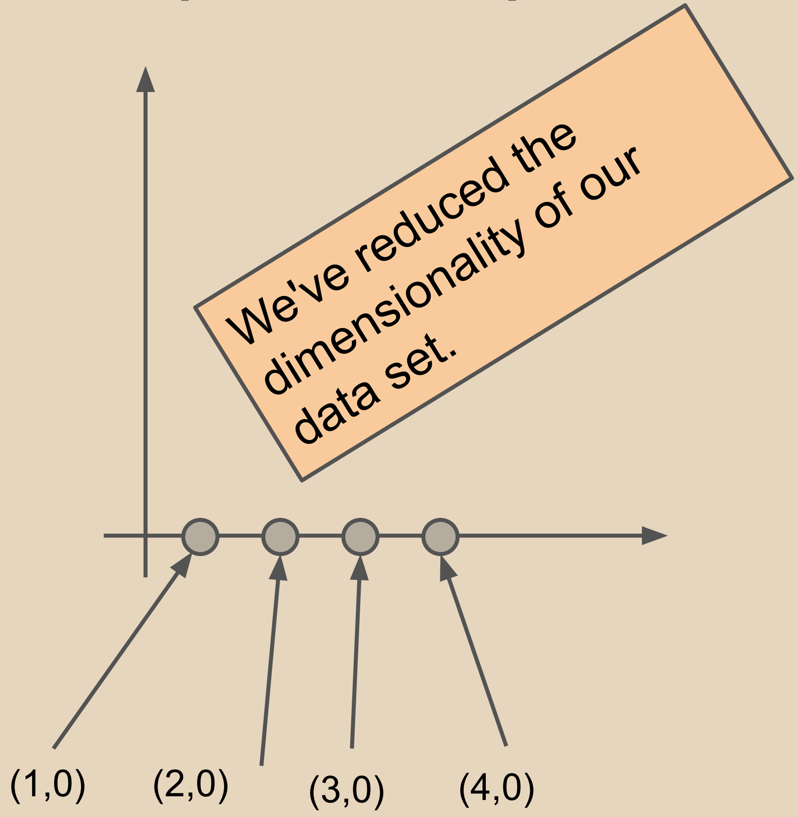


Any way we can recode them to represent them in lesser space?

Idea: Since we know that they are on the X-axis, store just the x-components, and specify the vector that they lie on, i.e., the X-axis, y=0.

Dimensional Reduction

A Simple Example

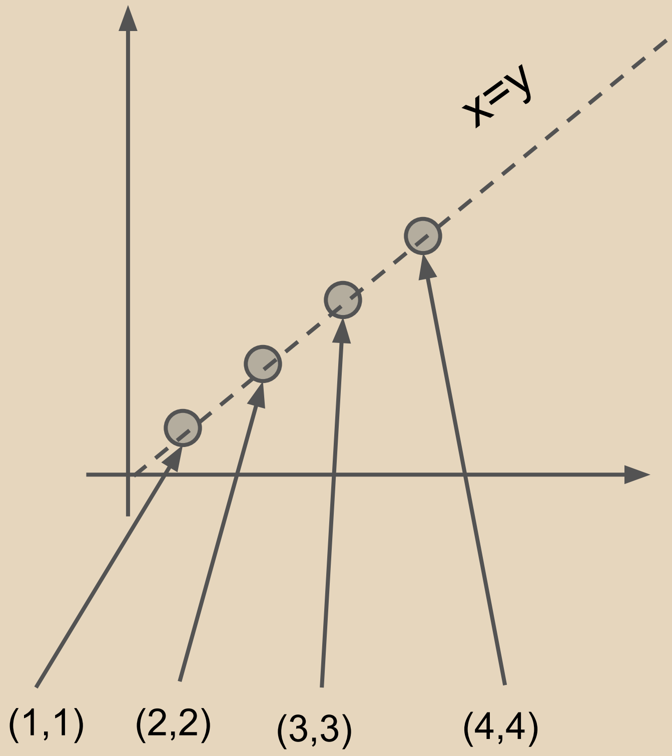


Idea: Since we know that they are on the X-axis, store just the x-components, and specify the vector that they lie on, i.e., the X-axis, y=0.

New representation of this data:
[1,2,3,4], y=0

Dimensional Reduction

Extending the Example

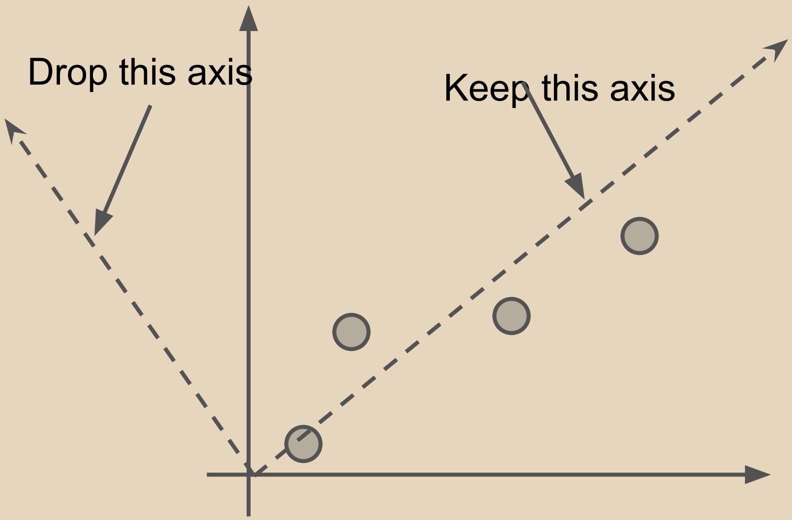


How do we apply our idea to this data set?
If we know that they all lie on x=y, we just need to change the vector, so our representation becomes:

[1,2,3,4], x=y

Dimensional Reduction

Extending the Example



This is lossy compression. Some variation will be lost when you discard an axis.

Basic Idea: Find a new set of coordinates which maximises variation of data along the first axis, and minimises the variation of data along the other one.

Disregard the axis with the least variation of data, then you've reduced the dimensionality of your data.

Dimensional Reduction

Extending the Example

- PCA is used extensively in dealing with massively-dimensional data sets.
- PCA implementations are non-trivial.
- Go learn Linear Algebra if you want to know how it works.

Basic Idea: Find a new set of
...hich maximises
...a along the
...minimises the
...a along the

... is called
Principal Component
Analysis.

More stuff I wish I'd time for...

- Neural Networks
- Genetic Algorithms

...maybe another time :-)

Questions?

http://avishek.net