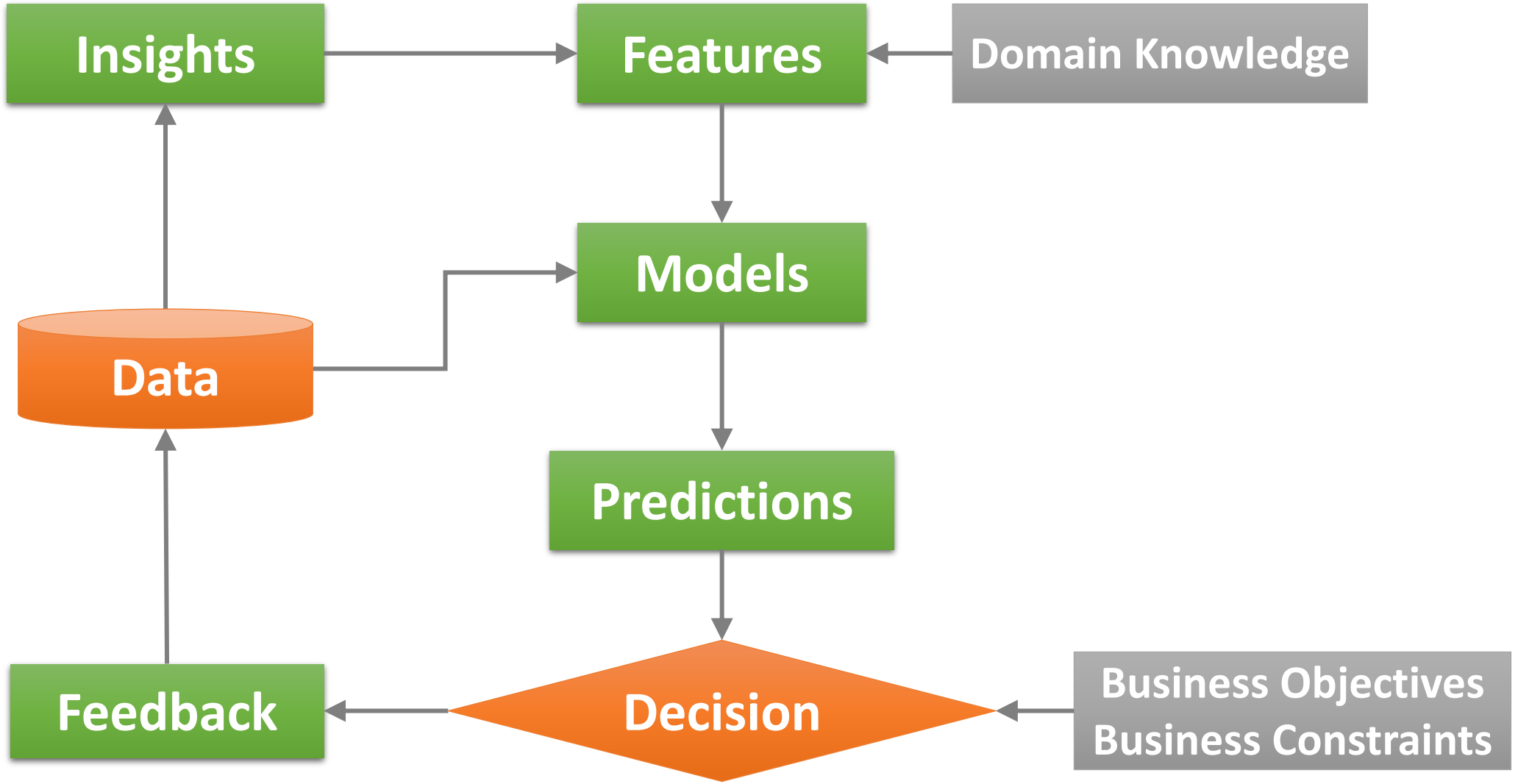
Machine Learning

DRIVING BETTER DECISIONS

Dr. Shailesh Kumar

From DATA to DECISIONS



The CLASSIFICATION Paradigm

Insights

Features

Agricultural / Botanical Knowledge

Classification

- Remote Sensing
- Text Classification
- Sentiment Analysis
- Medical diagnosis
- CRM systems

Normal vs. Stressed in a way?

Crop Yield & Total Cost

Water? Pesticide? Fertilizer?

Maximize Yield
Minimize Cost

The RECOMMENDATION Paradigm

Insights

Features

Seasonal vs. Causal purchase behavior

Recommendation

- Product (Amzaon)
- Movie (NetFlix)
- People (LinkedIn)
- News feeds
- Video (YouTube)

Probability of Conversion

Coupon converted
Repeat purchases
Customer Churns
Blocked Couponing

Product? Messaging?
Channel? Time?

Maximize Cross-sell
Maximize Up-sell
Increase Lifetime value
Minimize Coupon fatigue

The OUTLIER DETECTION Paradigm

Insights

Features

Seasonal usage, Multiple users,

Outlier Detection

- Fraud (Credit card)
- Intrusion (Security)
- Fault (Manufacturing)
- Anomaly (Sensors)
- Terrorism (behavior)

Probability of Fraud?

Customer Reports

Approve? Call customer? Block card? Block account?

Minimize Fraud
Maximize Convenience

The RETREIVAL Paradigm

hyderabad preschools Search

About 101,000 results (0.15 seconds) Advanced search

Profitable Ad

Summarize Information

Relevant Search Result

Local business results for **preschools** near **Hyderabad, Andhra Pradesh, India**

A Balarcade
www.balarcade.com - 040 65089937 - More

B Kinderkare Day Care
maps.google.com

C Tendril Preschools
www.kg.tendrildaycare.com - 040 40203350 - 2 reviews

D Mount Carmel Global School
www.mountcarmelhyd.com - 040 3252 3253 - More

E Sharada's kids
www.sharadaskids.com - 040 23703397 - More

F Euro Kids
www.eurokidsindia.com - 040 23114796 - More

G Gator Academy
maps.google.com - 040 2331 6517 - More

More results near Hyderabad, Andhra Pradesh, India »

Preschools in Hyderabad
Preschools indiaedu provides details on **Preschools** in **Hyderabad**.
preschools.indiaedu.com/hyderabad-preschools/ - Cached

KinderKare: **Preschool**, Day Care in **Hyderabad**, Gachibowli, Kondapur ...
KinderKare, **preschool**, creche and day care center is located in Jaibheri Enclave, Gachibowli, Kondapur near to hitech city, madhapur, Nanakram guda, .
www.kinderkare.in/ - Cached - Similar

Sponsored links

Apply for Franchisee
RootsToWings from Educomp
India's Premier **Preschool** Chain
www.RootsToWings.in
Andhra Pradesh

Roots to Wings-Preschool
A.S. Rao Nagar, Secunderabad.
Destination for NextGen Kids.
falconkidz.com/rootstowings.aspx
Hyderabad, Andhra Pradesh

Manthan Internatnl School
A Cambridge(CIPP) and CBSE School
Admissions Open at Madhapur, HYD
manthanschool.org
Andhra Pradesh

Blue Blocks Pre-School
(Birth to 6 years)Montessori system
Gachibowli, **Hyderabad**
BlueBlocks.in

LE Preschool Franchise
Master and Single unit franchise
Across India @LOWEST Franchise Fee
www.littleeinsteins.co.in

Discounts Toys, Furniture
Buy **Preschool** Equipments and Toys
Playschool books and furniture safe
www.MyKidsArena.com
Andhra Pradesh

Insight

Every ML Paradigm is an OPTIMIZATION Problem

- **Search** = Maximize **Click Through Rate (CTR)**
- **Ads** = Maximize Revenue = **CTR x Cost Per Click**

Key Questions in building an ML Solution

- **What DECISIONS does my business make?**
  - e.g. Which offer to send to which customer?
- **On what BASIS do I make those decisions?**
  - e.g. Past purchase behaviors of those/similar customers?
- **How do I quantify SUCCESS of my decisions?**
  - e.g. What fraction of offers get converted?
- **What data should I collect to EVALUATE my decisions?**
  - e.g. Did customers redeem the coupons – after how long, how often.
- **What data should I collect to IMPROVE my decisions?**
  - e.g. Point of sales data, Social data, Reviews, etc.
- **How do I improve my MODELS from the data I collect?**
  - More data, More features, Better Modeling, More Customization,…

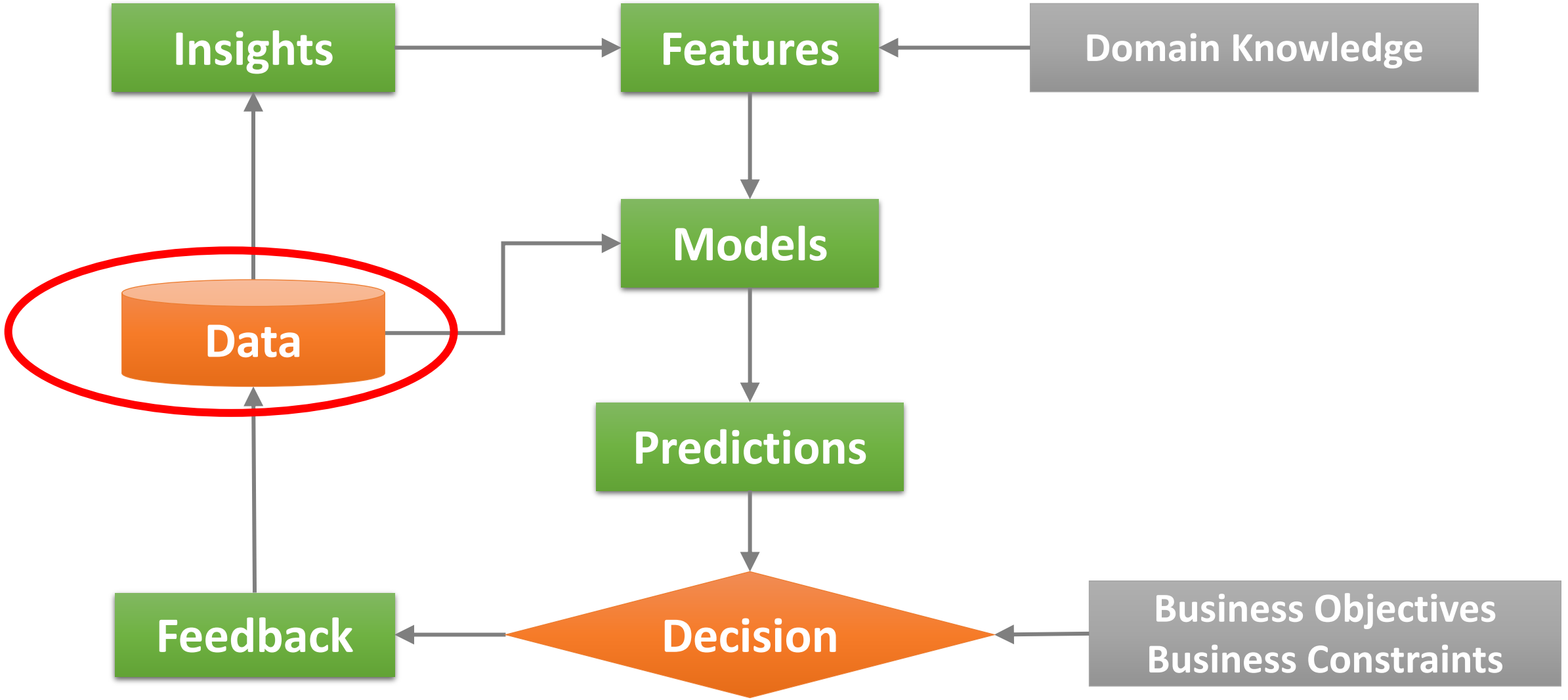
Machine Learning

ANATOMY OF DATA

Dr. Shailesh Kumar

Data comes in many shapes and forms.

Understanding DATA



Data TYPES

- **Numeric – a quantifiable** number
  - **Type** – integer (e.g. age), floating (e.g. price), time, date, …
  - **Stats** – min/max/median/mean/…
  - **Units** – (C/F), (KG/Lb), (Meter/Feet), (Sec/Min/Hrs),
  - **Distributions** – exponential/uniform/…
- **Ordinal – not quantifiable** but **ordered**
  - E.g. size = Small/Medium/Large/…
  - E.g. income bucket = Low/Medium/High/Wealthy/...
  - E.g. Relevance = Perfect/Excellent/Good/Fair/Bad/…
- **Symbolic – neither quantifiable, nor ordered**
  - E.g. color = red/green/blue/…
  - E.g. state/country/region/…
  - E.g. weather = rainy/cloudy/windy/…

Data ORGANIZATION

MULTIVARIATE

(rows (**examples**) of columns (**features**))

	feaure-1	feature-2	feature-3	feature-4	feature-5
example-1					
example-2					
example-3					
example-4					
example-5					
example-6					
example-7					

Low Dimensional and Dense

BASKET

(**sets** of things)

market basket, keyword list

	item-1	item-2	Item-3	Item-4 …	item-10M
example-1	**1**		**1**		
example-2				**1**	**1**
example-3					
example-4		**1**			
example-5				**1**	
example-6		**1**		**1**	
example-7			**1**		**1**

High Dimensional and Sparse

BAG

(**weighted sets** of things)

Bag-of-Words, Bag-of-Visual Words

	item-1	item-2	Item-3	Item-4	item-100M
example-1	**10**		**5**		
example-2				**13**	**11**
example-3					
example-4		**12**			
example-5				**21**	
example-6		**4**		**51**	
example-7			**1**		**32**

High Dimensional and Sparse

Data ORGANIZATION

1-D SEQUENCE

(list of “symbols”)

Gene sequence, Text

```
TGGAAGAGGCCTCAGCAGGCCCAGGCCACCTGGAGGGAGAGCAGACCTGCGGCTGAGGATGCAGGGCTCC
CGGGCACGGTGCTAGCCCTGCCTTGAGACACCCCGAGAGCTGTGGGAAGAGCTGTGGGATCCCCTATTGC
ATCACAAAGCGGCCCTGGAGGGCTGGTCTTTATTTTGATGAGGCTGAGAAGGGAAGGCTGCGGGCATGTT
TAATCCGCACGCTTTAGACTCCCCGGCTGTGATTTTTGACAATGGCTCGGGGTTCTGCAAAGCGGGCCTG
TCTGGGGAGTTTGGACCCCGGCACATGGTCAGCTCCATCGTGGGGCACCTGAAATTCCAGGCTCCCTCAG
CAGAGGCCAACCAGAAGAAGTACTTTGTGGGGGAGGAGGCCCTGTACAAGCAGGAGGCCCTGCAGCTGCA
```

1-D SERIES

(list of “numbers”)

Stock market, IoT data

2-D SERIES

(mage)

3-D SERIES

(videos)

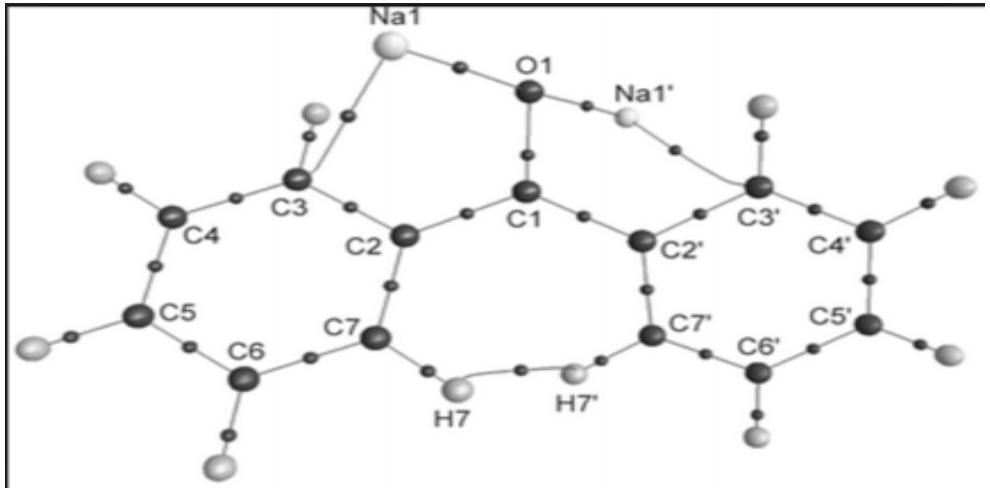
4-D SERIES

(3-D videos)

Data ORGANIZATION

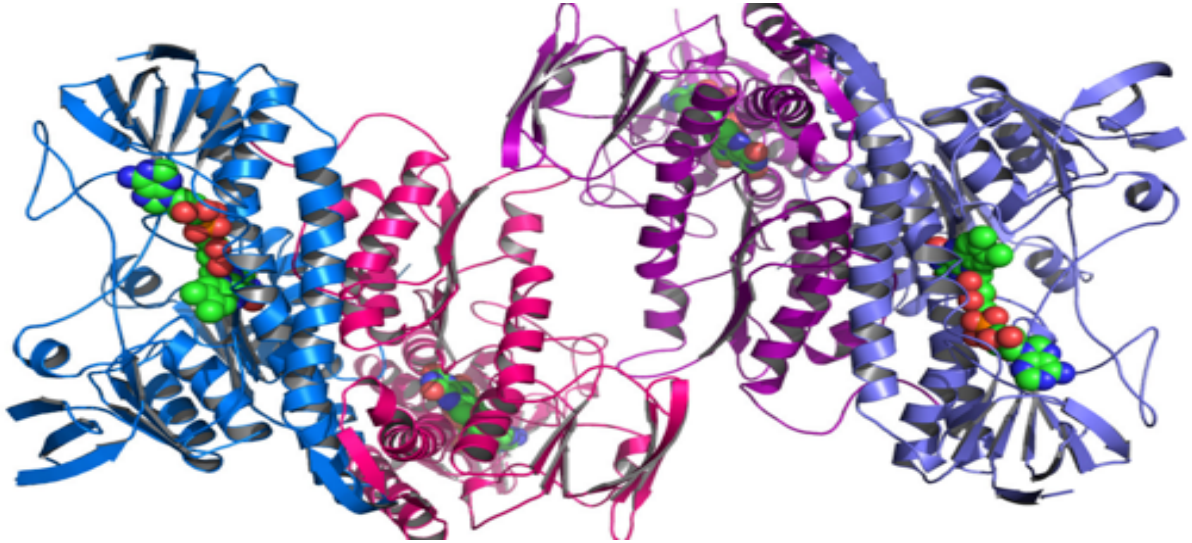
GRAPHS

Chemical compounds
Social Networks
Knowledge Graphs
Telecom/Traffic Networks



3-D SHAPES

(list of (x,y,z) coordinates)
Motion Sensors, Protein Structures



Data MODALITIES

- **STRUCTURED** – fixed columns in a table
  - Multivariate data
  - Mix of numeric and symbolic features

- **UNSTRUCTURED** -- Arbitrary size data points
  - **SEQUENCE** : biological, speech, …
  - **SERIES** : stock market, etc.
  - **TEXT** : pages, queries, tweets, ads, blogs, news, …
  - **IMAGE** : regular, medical, remote sensing,…
  - **VIDEO** : regular, movies, security, surveillance,…

Machine Learning

NUANCES IN DATA

Dr. Shailesh Kumar

Data is not always **clean** or **simple** in nature

Data Nuance 1 | DATA NOISE (Text Data)

NOISE IN TEXT DATA | Grammar, Spelling Mistakes, LOL's, Mixed Languages, etc.

garlic use on pizza wa tasty

- sicilian sauce wa tasty
- pizza though oi so yummy
- even marinara chicken sandwich deliciou
- eggplant parmesan wa deliciou
- crust on frie chicken wa real good though

probab one of worst meal ive ever had

- thi wa most disgust place at which ive ever eaten
- most unsatisfy place i have ever eaten
- absolute worst service ive ever receive
- would never ever go back
- never ever again

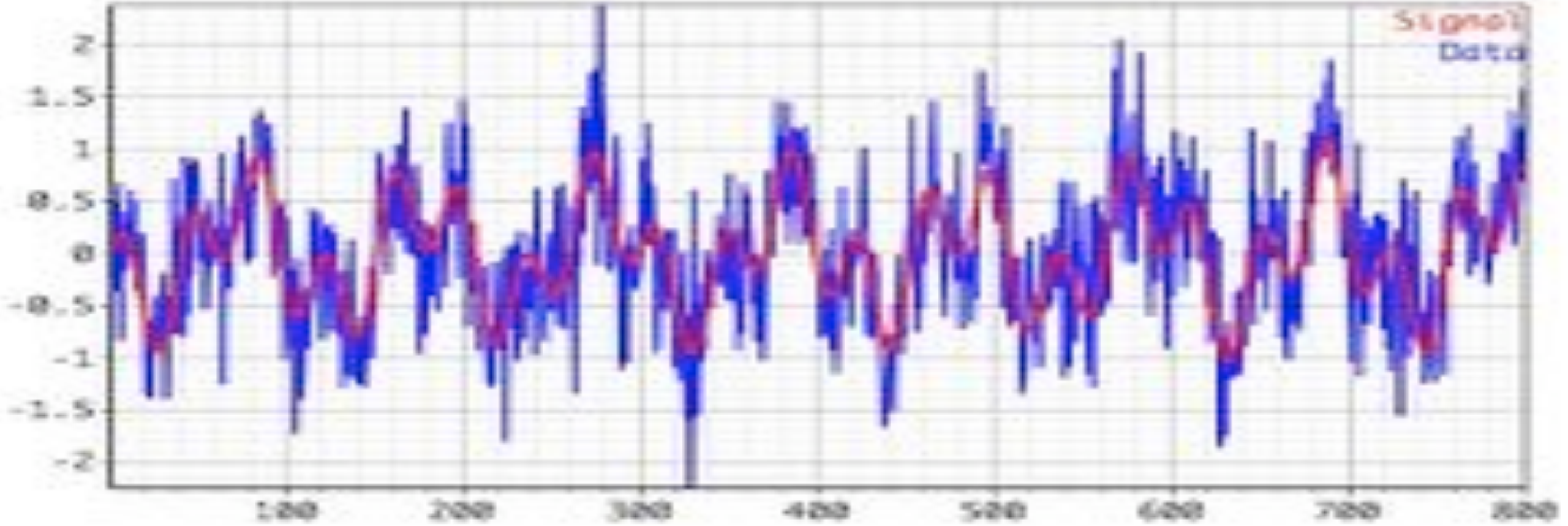
Comments on YouTube and News Articles → Twitter Data → Call Centre Notes → Product / Movie Reviews → Blogs, Quora → News Articles → Journal Papers

- **Its essential to know the "noise profile" of your data.**
- The ML algorithms we choose must be robust to the noise in the data
- E.g. NLP techniques cannot be used on User Generated Content or Twitter data
- E.g. NLP techniques can easily used on News articles

Data Nuance 1 | DATA NOISE (Time Series Data)

- High Frequencies are typically noise.
- Low Frequencies are typically signa
- Where do draw the boundary?

$$g(t) = a_0 + \sum_{m=1}^{\infty} a_m \cos\left(\frac{2\pi mt}{T}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2\pi nt}{T}\right)$$



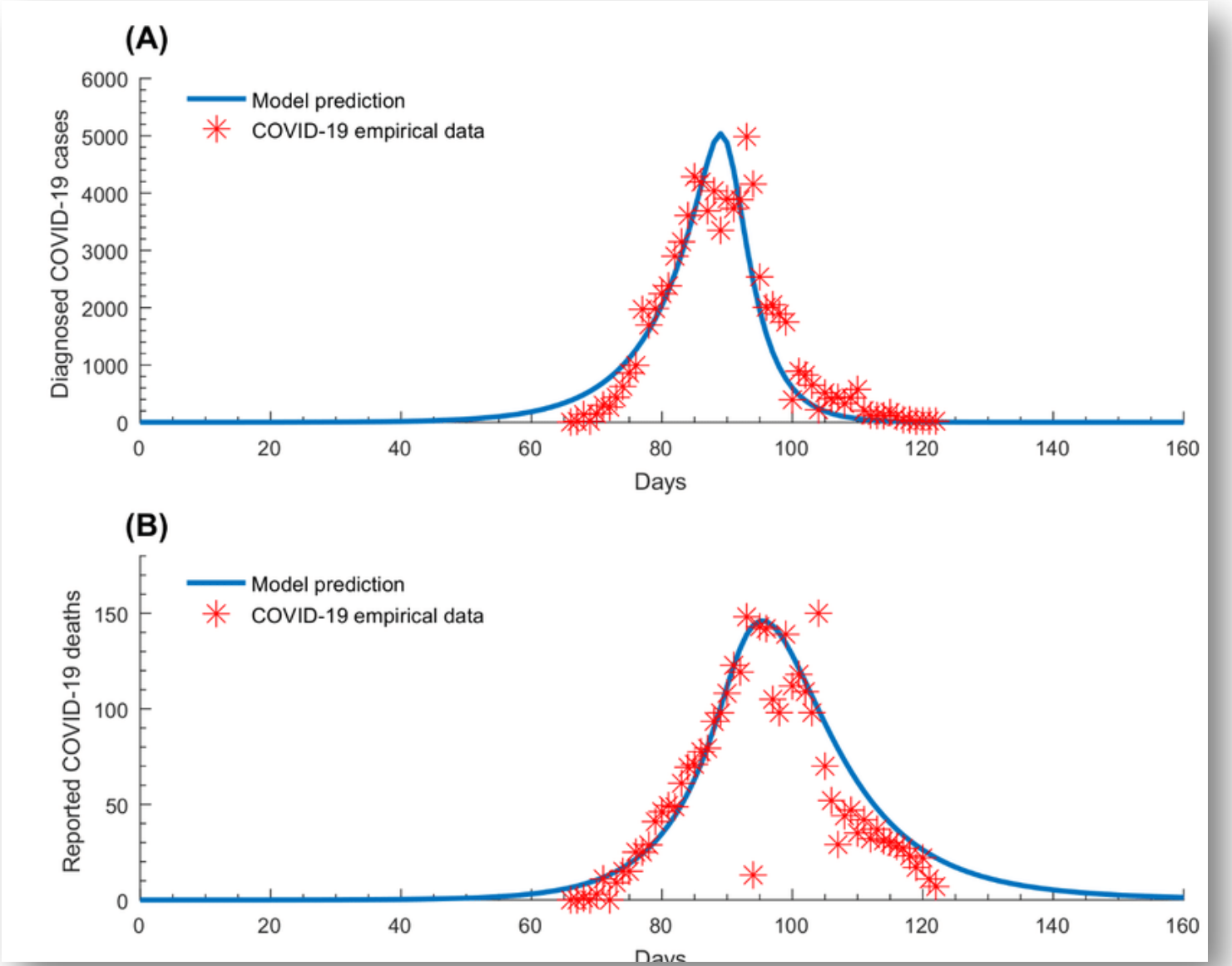
Data Nuance 1 | DATA NOISE (Count Statistics)

Sources of Noise?

- Not enough testing
- Not testing in the right places
- Testing kit noise
- Reporting Errors
- Testing gaps

MODELS help "smooth" this noise out

- Models = Grammar of data
- Fit the curve with the model
- Get a smooth shape that best "fits"
- Everything else is considered noise



Insight

Machine Learning = Finding Signal in Noise

Data Nuance 2 | GRAMMAR (Market Basket Data)

Few buy a complete “logical” item-set in same basket

- already have other products
- buy them from another retailer
- buy them at a different time
- got them as gifts
- ….

It’s a **Projections** of **latent customer** intentions

Data Nuance 2 | GRAMMAR (Market Basket Data)

It's a **Mixture** of **Projections** of **latent** intentions

Insight

You cannot find the needle unless you understand the nature of the haystack!

Data Nuance 3 | VARIABILITY (IMAGEES)







Variability in IMAGES:

- Pose | Scale | Object | Illumination

Data Nuance 3 | INVARIANCE (SPEECH, TEXT, OCR)

SPEECH INVARIANCES

- AGE GROUP
- GENDER
- ACCENT
- EMOTION
- BACKGROUND NOISE
- VOCAL TRACK HEALTH

TEXT

- PARAPHRASING
- TENSE
- ACTIVE/PASSIVE
- SEMANTIC
- LANGUAGE

OCR

- FONT
- BOLD/NOT
- ITALICS/NOT
- COLOR
- HAND WRITING

Insight

Machine Learning = the art of INVARIANCE LEARNING

Data Nuance 4 | MISSING DATA

SOURCES OF MISSING DATA

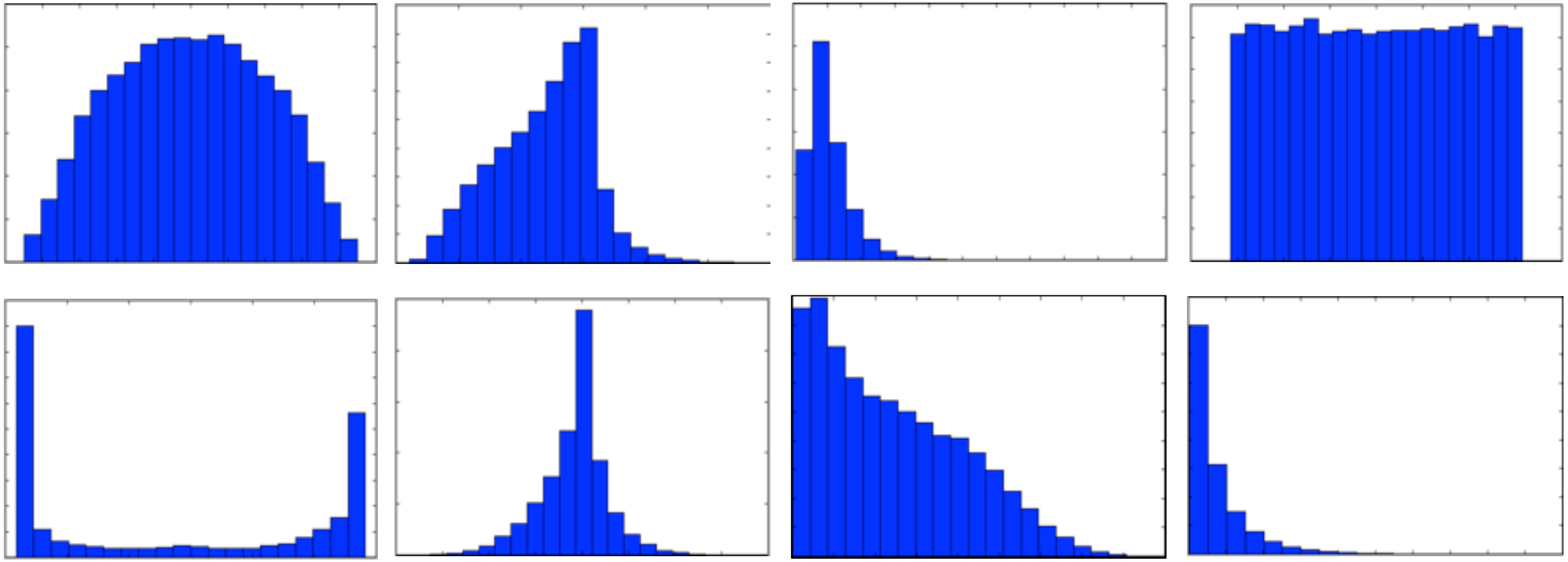
- Sensor Malfunction
- Logging Malfunction
- Human Error
- Optional Fields
- Downtimes
- Not Digitized

DATA IMPUTATION METHODS

- Statistical Imputation
  - Most common value (categorical feature)
  - Average value (Numeric feature)
- Density based
  - Most likely value given others
- Model based
  - Predicted from features

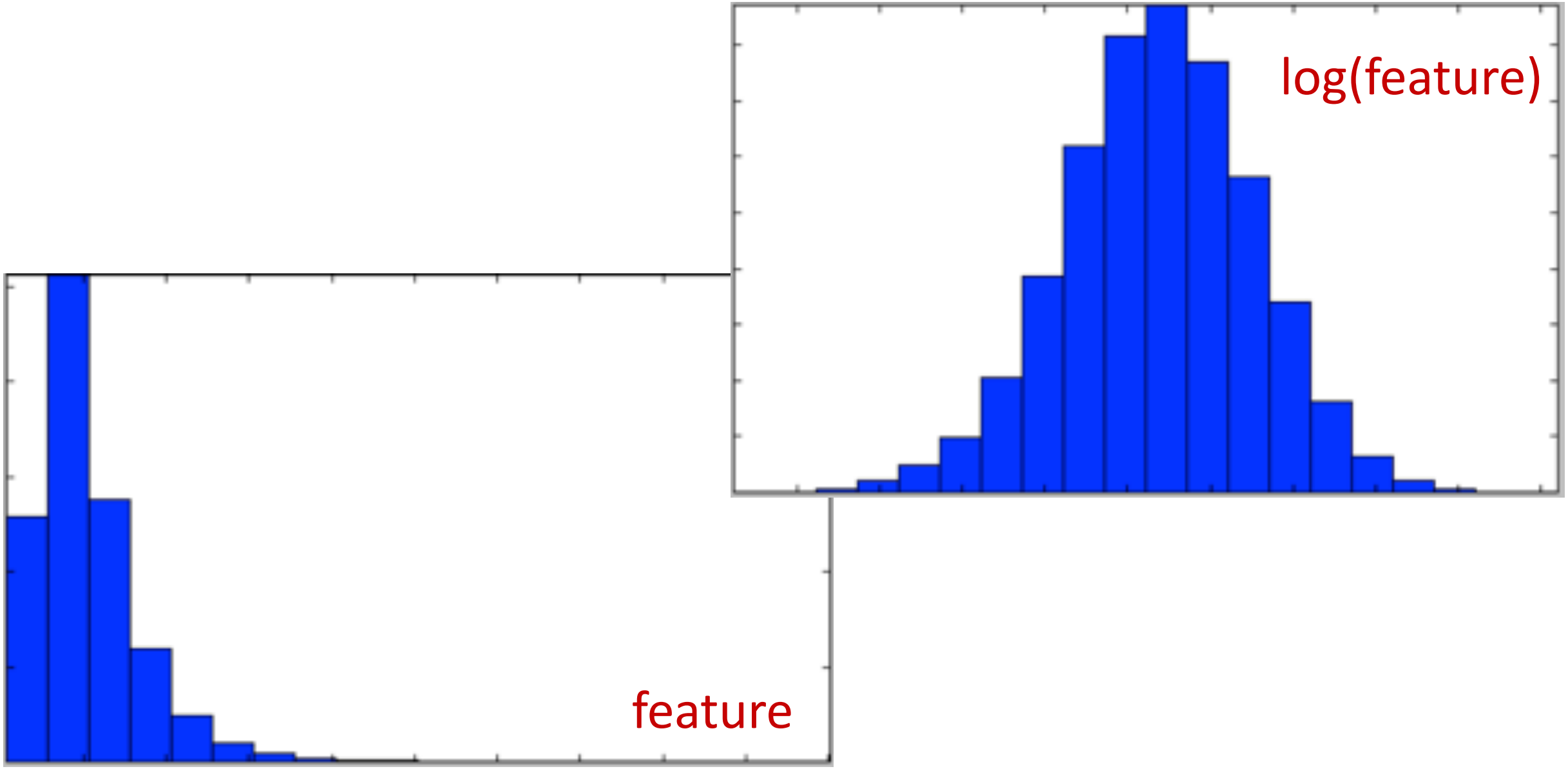
Data Nuance 5 | NOT-SO-NORMAL DISTRIBUTION

https://www.kaggle.com/c/higgs-boson

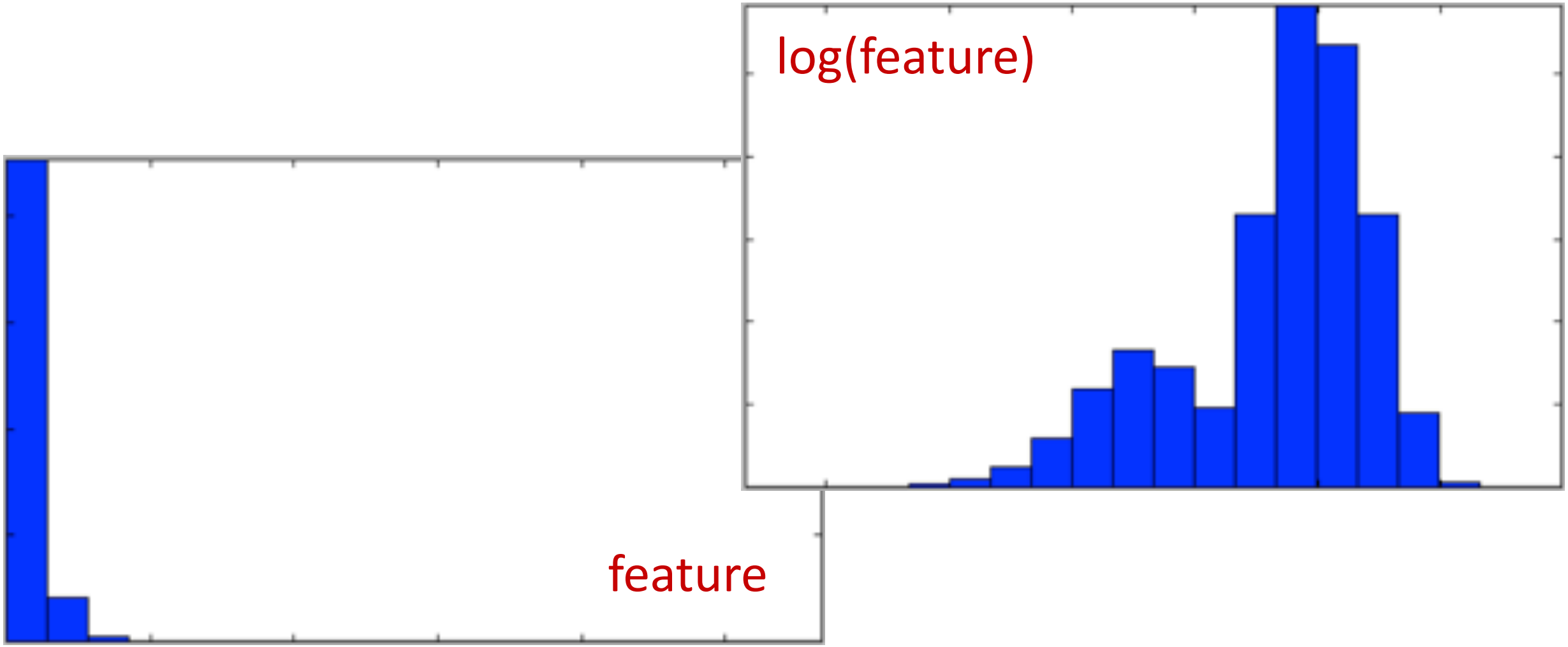


None of the features in this dataset have a normal distribution!!

Data Nuance 5 | NOT-SO-NORMAL DISTRIBUTION



Data Nuance 5 | NOT-SO-NORMAL DISTRIBUTION



Data Nuance 6 | HETEROGENEOUS FEATURES

Variable	Reference Range
C-reactive protein (mg/liter)	<3.0
Troponin I (ng/ml)	0–0.4
D-dimer (ng/ml)	<400
Sodium (mmol/liter)	136–145
Potassium (mmol/liter)	3.5–5.1
Chloride (mmol/liter)	98–107
Carbon dioxide (mmol/liter)	20–31
Blood urea nitrogen (mg/dl)	9–23
Creatinine (mg/dl)	0.7–1.3
Glucose (mg/dl)	80–140
Calcium (mg/dl)	8.7–10.4
Total protein (g/dl)	5.7–8.2
Globulin (g/dl)	2.0–3.0
Aspartate aminotransferase (U/liter)	10–40
Alanine aminotransferase (U/liter)	10–49
Anion gap (mmol/liter)	5–16
Albumin (g/dl)	3.2–4.8
Total bilirubin (mg/dl)	0.3–1.2
Alkaline phosphatase (U/liter)	46–116
Lactate dehydrogenase (U/liter)	120–246
Lactic acid (mmol/liter)	0.4–2.0
White-cell count (per mm^3)	3800–11,000
Hemoglobin (g/dl)	13.2–17.0
Hematocrit (%)	39.0–50.0
Platelet count (per mm^3)	150,000–400,000
Absolute neutrophil count (per mm^3)	1900–7400
Absolute lymphocyte count (per mm^3)	1100–3900
Erythrocyte sedimentation rate (mm/hr)	1–20

- **Features might be in different scales & distributions**
  - Incomes: 1,00,000 – 100,00,000
  - Age: 20 – 75
  - Rent / Buy: 0 – 1
- **Min-Max Normalization** to **range [0,1]**

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

When would this not work well?

What should we do to fix it?

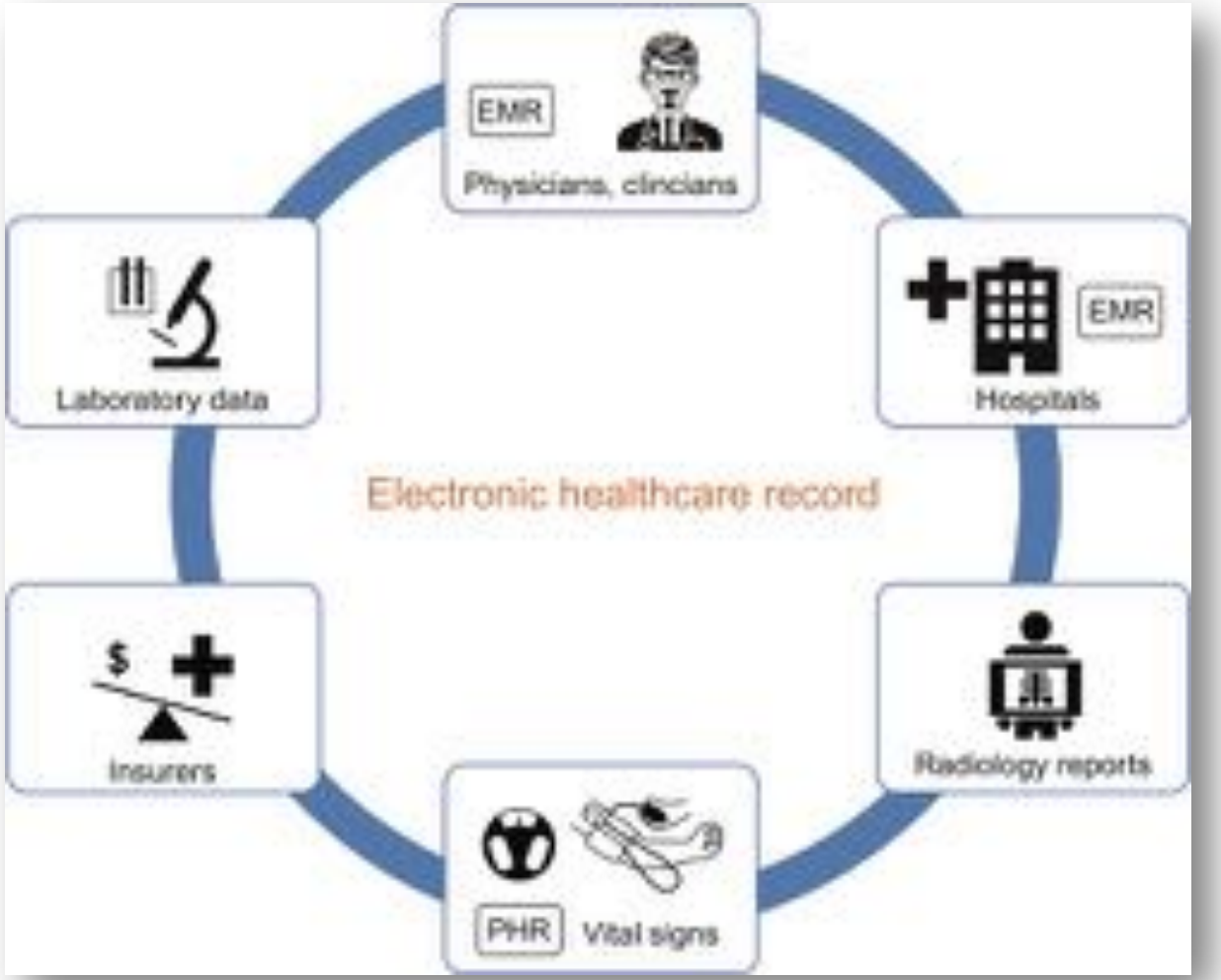
- **Z-Scoring Normalization** – makes data **zero mean**, **unit norm**

$$x' = \frac{x - \mu}{\sigma}$$

When would this not work well?

What should we do to fix it?

Data Nuance 7 | MULTI-MODAL FEATURES



- Series of Transactions
- Of different types
- ECG – Time Series
- X-Ray/MRI – Image data
- Symptoms – Basket Data
- Medical Report – Text Data
- Genomics - Sequence Data
- Pharmacy – Market Basket
- Blood report – Multi-variate
- Fitbit – Time series data

NUANCES IN OUTPUT DATA

Dr. Shailesh Kumar

Collecting/Managing Output data is an art in itself

Importance of LABELLED DATA

CLOUD INFRASTRUCTURE
AWS, GCE, BlueMix, Azure

AI COMPUTE PARADIGMS
Tensorflow, MxNet, SparkML…

BIG DATA PARADIGMS
Spark, Hadoop, Kafka, Pregel…

LABELED DATA
(problem dependent)
Unique to every organization

What are we PREDICTING?

- Defining CHURN
- Defining PART FAILURE
- Defining ENGAGEMENT
- Defining VALID CLICK
- Defining CLASS HIERARCHY
- Formulating OPTIMIZATION

IMPLICIT FEEDBACK | The Silent Innovation in ML

The quality of the feedback determines the quality of your models

- Which search result for which query by whom in what context?
- How long did they stay on that page? Is there a position bias?

Feedback need not just be binary

- Did they **click** the top recommended video?
- What fraction of that video they **watched**?

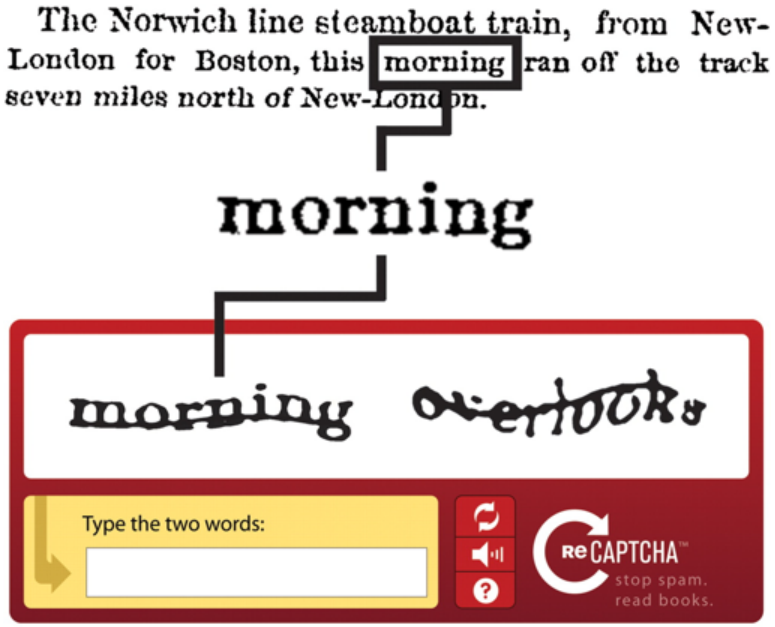
Same Entity Multiple degrees of Implicit Feedback

- What did you search, browse, read/write-review, Wish-list, Buy?
- What did you buy in same visit, in recent visits, for same address?

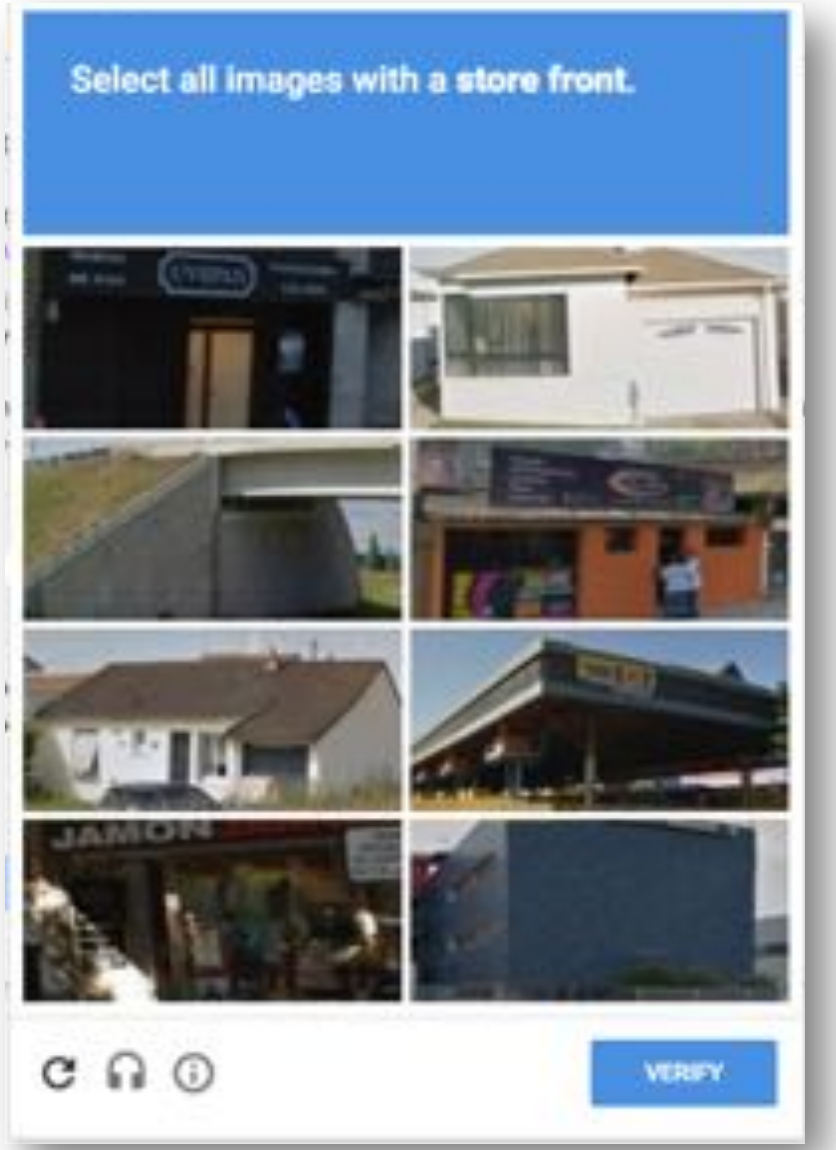
LABELD DATA | Whatever it takes!

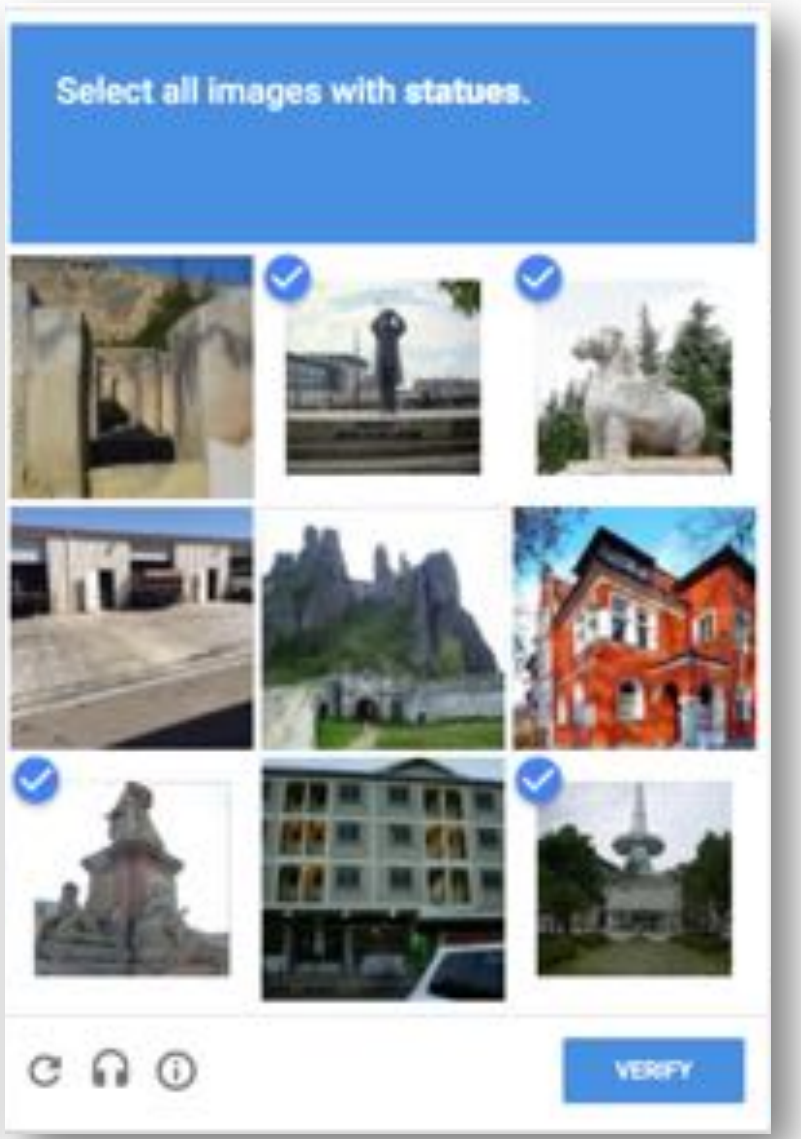
- **Label data is sometimes costly**
  - Remote Sensing – travel to the site
  - Astronomy – line of sight is rare
  - Samples from MARS – costly to obtain
  - Biological Experiments might be costly to do
  - Legal e-discovery – expensive lawyers

- **Crowd Sourcing labeled data**
  - Click logs – in web search
  - LabelMe – annotating images
  - ReCaptcha – labels for OCR data

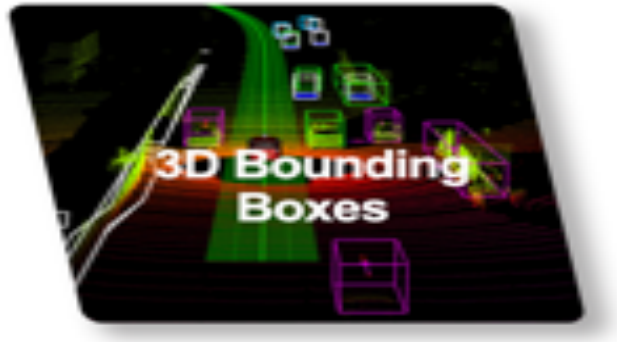


LABELD DATA COLLECTION





Labelling all kinds of data is possible now



Labelling is a whole new segment in “AI Economy”

LABELD DATA COLLECTION Platforms

https://aws.amazon.com/sagemaker/groundtruth