Data Annotation & Evaluation

Dr. Tirthankar Dasgupta

Recap...

NLP and Text Processing Tasks

- Extracting Drug names, disease names, product mentions, email ids,
- addresses

Information Extraction

- Extracting Events / Sentiments Who did / said What, When, How, Why

Linguistic Analysis

- How many negative reviews – on which aspect of product / service

Statistical Analysis

- Risk assessment – new policy about environment / health care / education

Drawing Inference

- What are the causes a disease?
- What is the policy for Paternity leave?

Question Answering

- Summary of Financial obligations on signing a new contract with an alliance

Summarization

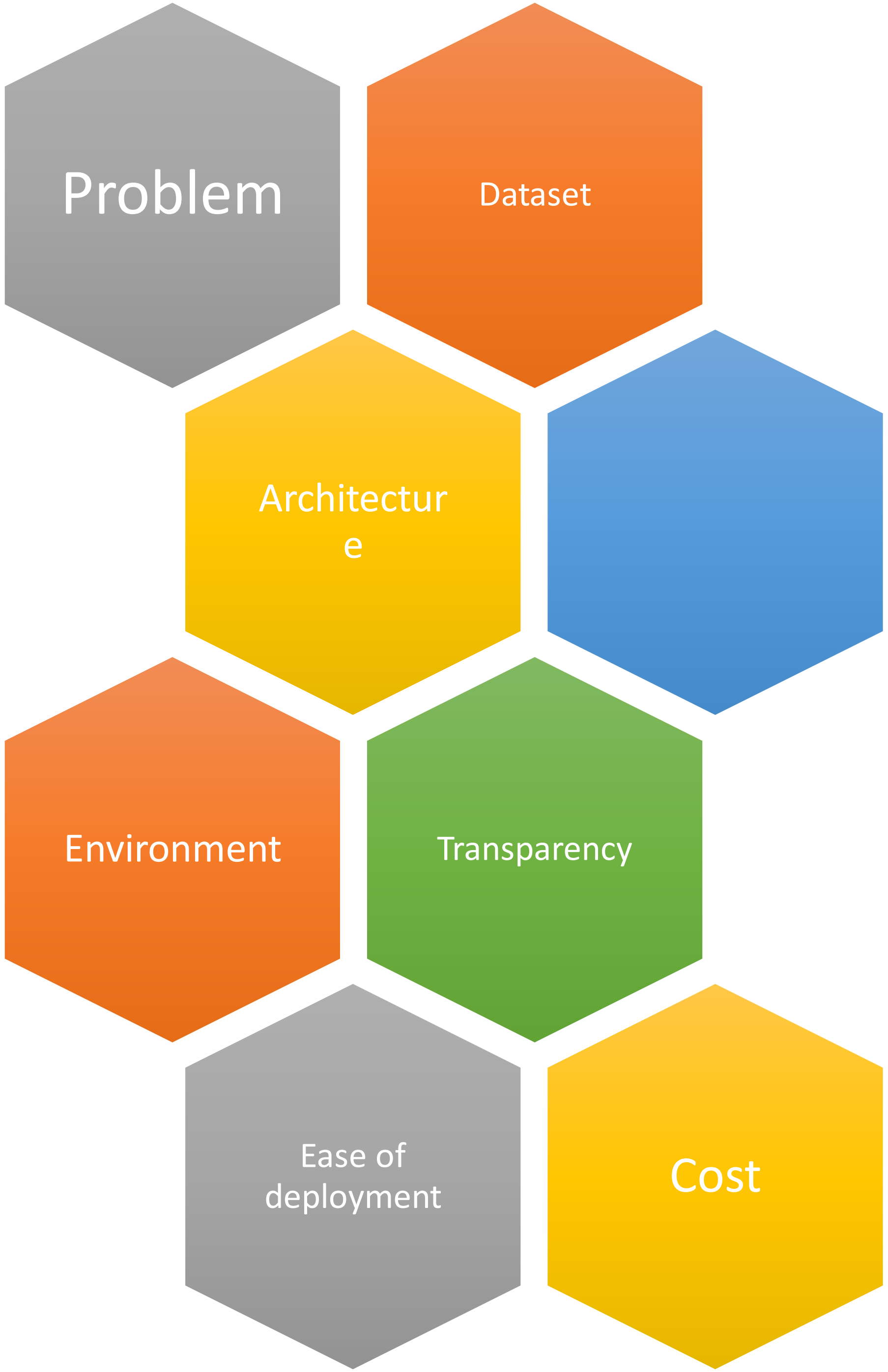
- General Purpose Conversation - Question
- Understand & Empathize
- Answer / Guide /Recommend

Natural Language generation

- Any language to any language

Machine Translation

Things to consider



Problem

Dataset

Architectur
e

Environment

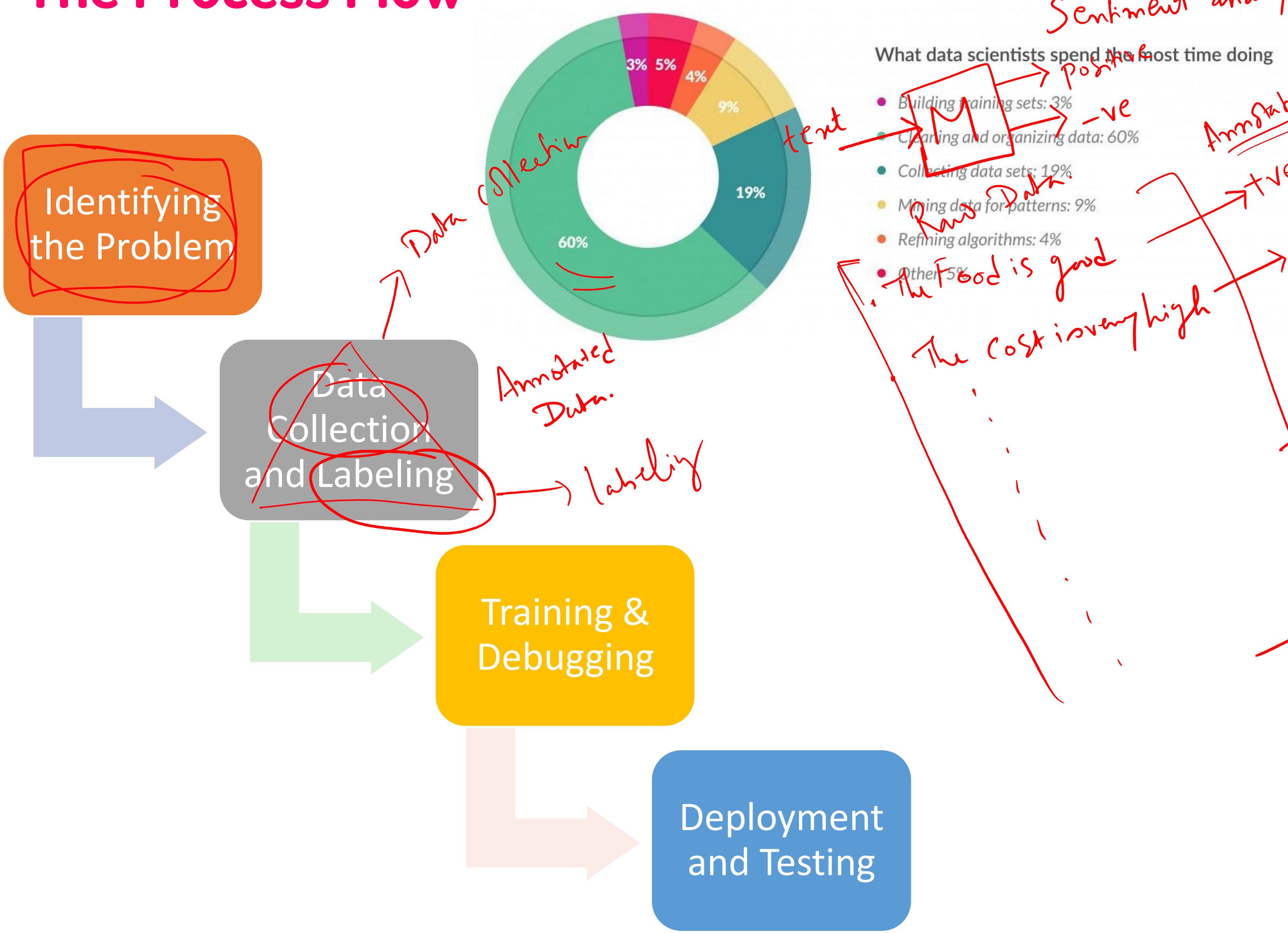
Transparency

Ease of
deployment

Cost

It is so cool to apply deep learning!

The Process Flow


Identifying the Problem
Data Collection and Labeling
Training & Debugging
Deployment and Testing
What data scientists spend the most time doing
3%
5%
4%
9%
19%
60%
● Building training sets: 3%
● Cleaning and organizing data: 60%
● Collecting data sets: 19%
● Mining data for patterns: 9%
● Refining algorithms: 4%
● Other: 5%

Sentiment analysis Tool/model

text → M → positive

→ -ve

Annotation

Raw Data

The Food is good → +ve

The cost is very high → -ve

→ neutral

Data collection

Annotated Data.

labeling

Types of NLP Problem

definition of "Word"

- Phoneme boundary identification (Speech)
- Morphological analyzer
- Information Retrieval
- Information Extraction
  - NER
  - Keyphrase extraction
  - Relation extraction
- Topic Identification
- POS
- Role labeling
- Word Sense Disambiguation
- Textual similarity
- Summarization
- Machine Translation
- QA
- Conversation

P
Org
loc
Currency

Syllable boundary identification

- Classification
  - Word level
  - Sentence level
  - Document level
- Clustering
- Generation

Play + ing = Playing → inflection form of "Play"

un+done = undone

unputdownable

un + put + down + able
prefix
suffix

Ram is playing cricket

Ram is play with Siva

Cat vs Cats

Play, playing, Player

blackboard

"cloud nine"

"grand father"

Applications

- POS tagging
- Sentiment Analysis
- Aspect Extraction

Parts-of-Speech

Saw → V

IDENTIFY PARTS OF SPEECH

John saw the Girl

PN V DT N

"saw" is more likely to be a verb V
rather than a noun N

↑PN V DT V

John saw the saw

The second "saw" is a noun N because a noun N is more likely to follow a determiner.

			NN		
			RB		
	VBN		JJ		VB
PRP	VBD	TO	VB	DT	NN
She	promised	to	back	the	bill

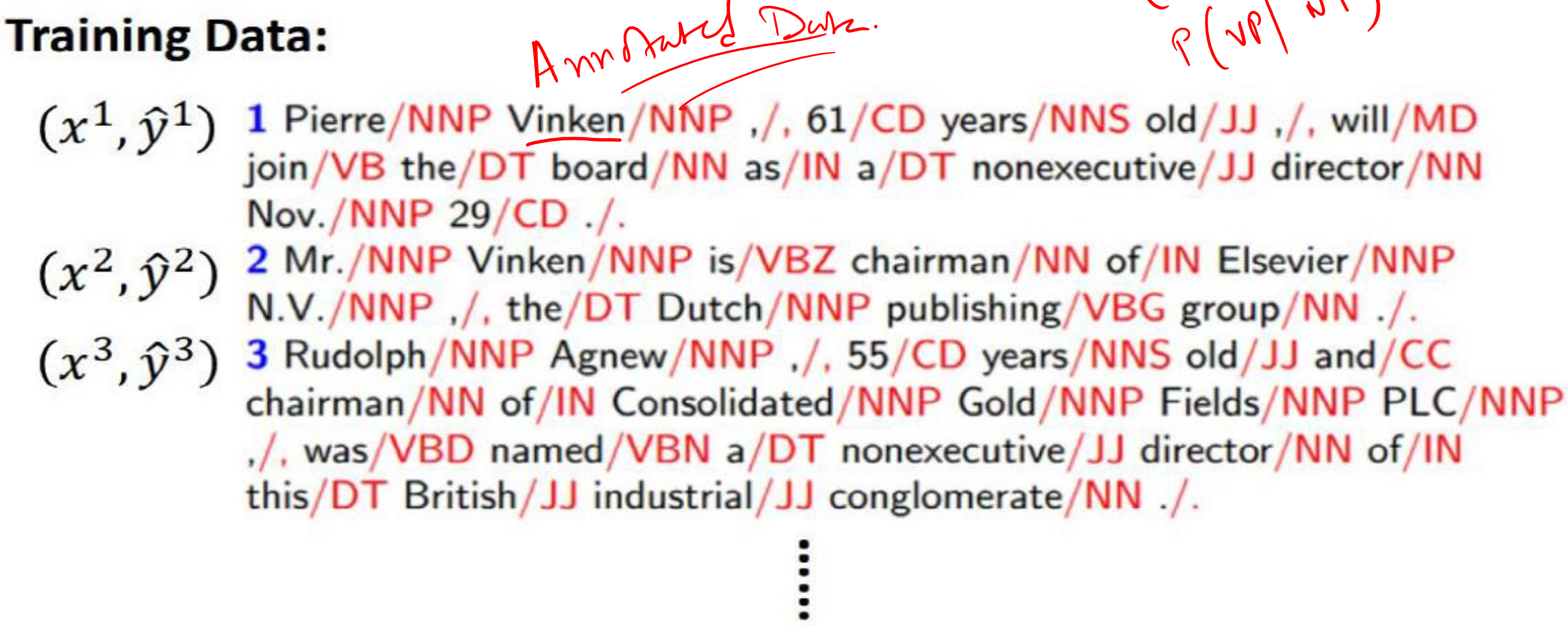
ESTIMATING THE PROBABILITIES

~100 POS
~1,000 unique words
100,00

- How can I know P(V|PN), P(saw|V) ...... ?
- Obtaining from training data

P(NNP | NNP, CD, NP) P(NN | V) . . . .
P(VP | NP)

Annotated Data.



Training Data:

$(x^1, \hat{y}^1)$ 1 Pierre/NNP Vinken/NNP ,/, 61/CD years/NNS old/JJ ,/, will/MD join/VB the/DT board/NN as/IN a/DT nonexecutive/JJ director/NN Nov./NNP 29/CD ./.

$(x^2, \hat{y}^2)$ 2 Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsevier/NNP N.V./NNP ,/, the/DT Dutch/NNP publishing/VBG group/NN ./.

$(x^3, \hat{y}^3)$ 3 Rudolph/NNP Agnew/NNP ,/, 55/CD years/NNS old/JJ and/CC chairman/NN of/IN Consolidated/NNP Gold/NNP Fields/NNP PLC/NNP ,/, was/VBD named/VBN a/DT nonexecutive/JJ director/NN of/IN this/DT British/JJ industrial/JJ conglomerate/NN ./.

⋮

Structured prediction (sequence tagging) – label for a word depends on other labels

- Rather than classifying each word independently

Applications

- POS tagging
- Sentiment Analysis
- Aspect Extraction

Sentiment Analysis

This is a good book. → Positive

This book is simply unputdownable! → More Positive

This is a bad book. → Negative

The first chapter of the book is great, but the rest is a junk! → Positive & Negative

+ve

The food is good but the prices are (very high). +ve

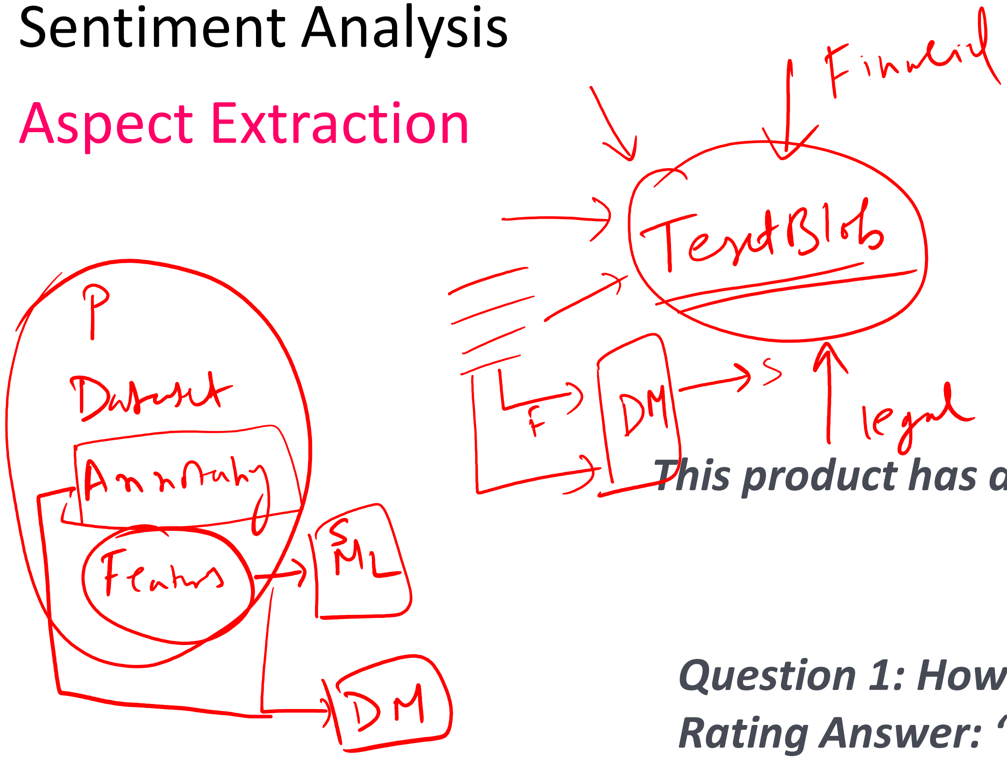
-ve

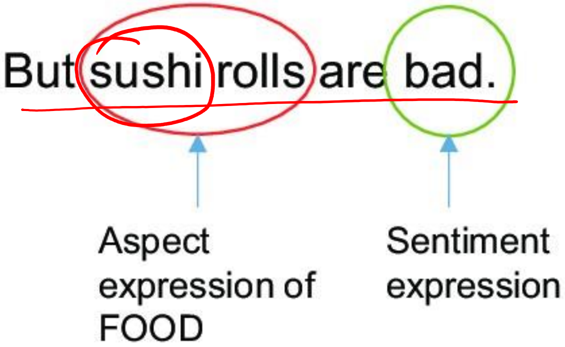
Applications

Aspect Analysis

- POS tagging
- Sentiment Analysis
- Aspect Extraction

I was tempted to buy this product as I really like its design, but its price is not very good.





But sushi rolls are bad.

Aspect expression of FOOD

Sentiment expression

This product has a good price; the one my brother purchased has a good design.

Question 1: How would you rate our service?
Rating Answer: “10”

Question 2: Motivate your answer.
Textual Answer: “Customer care”