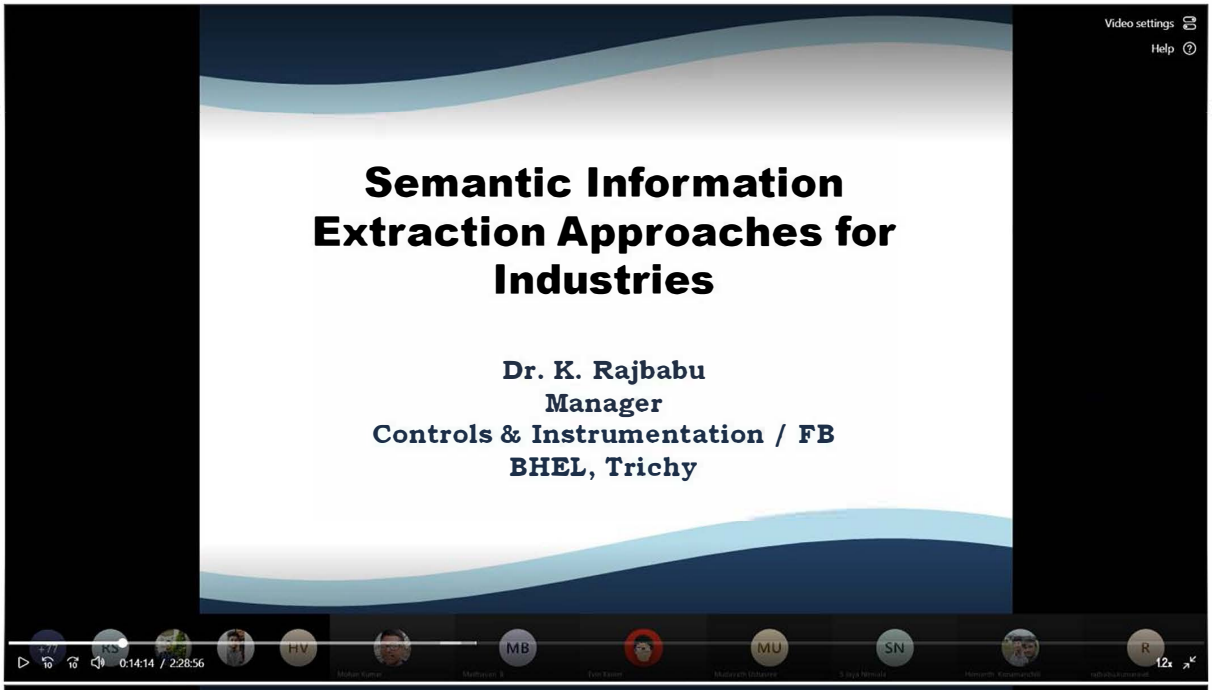
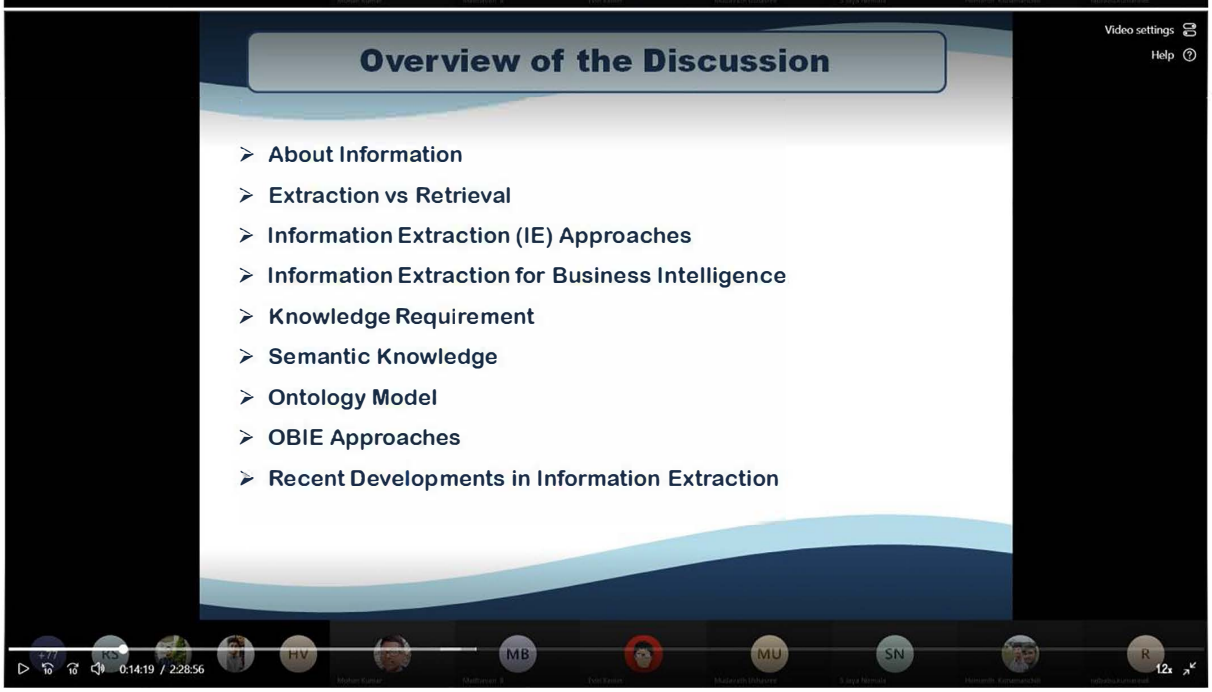
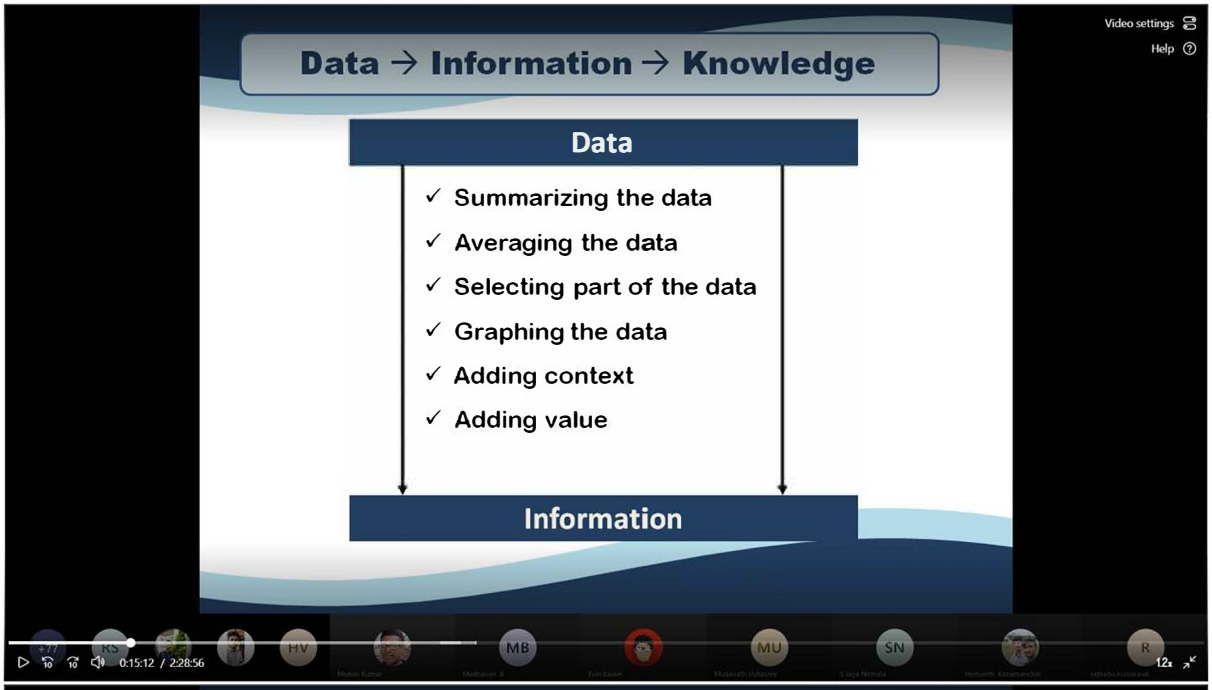
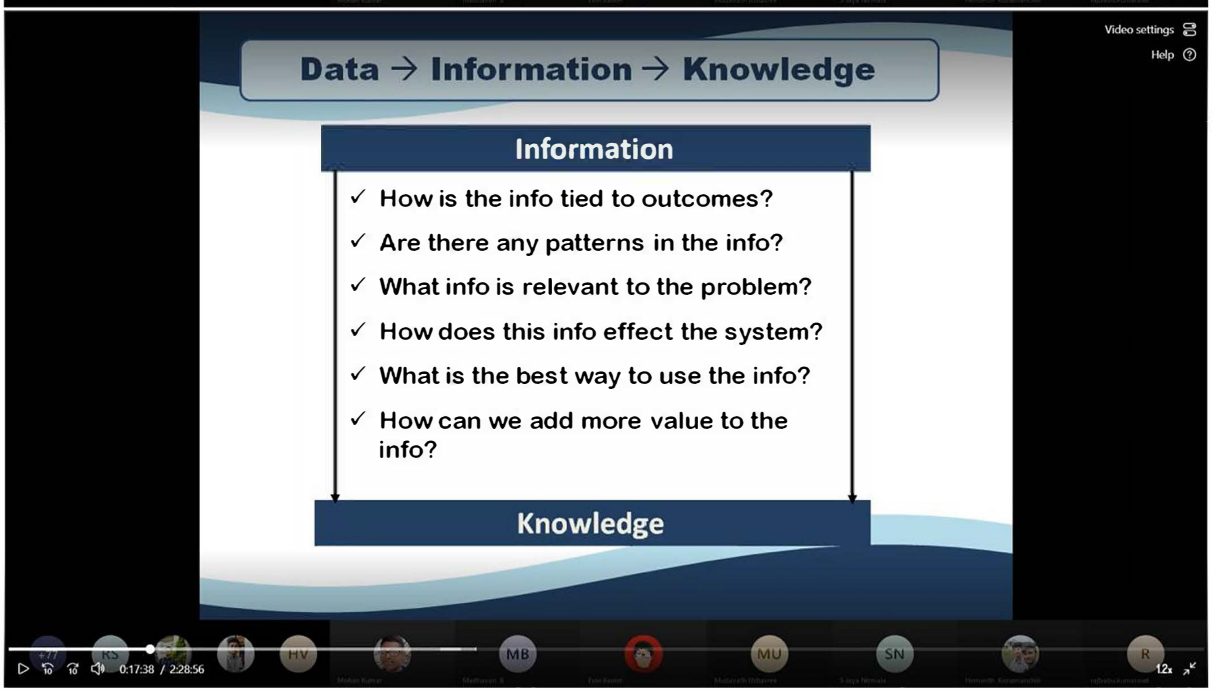
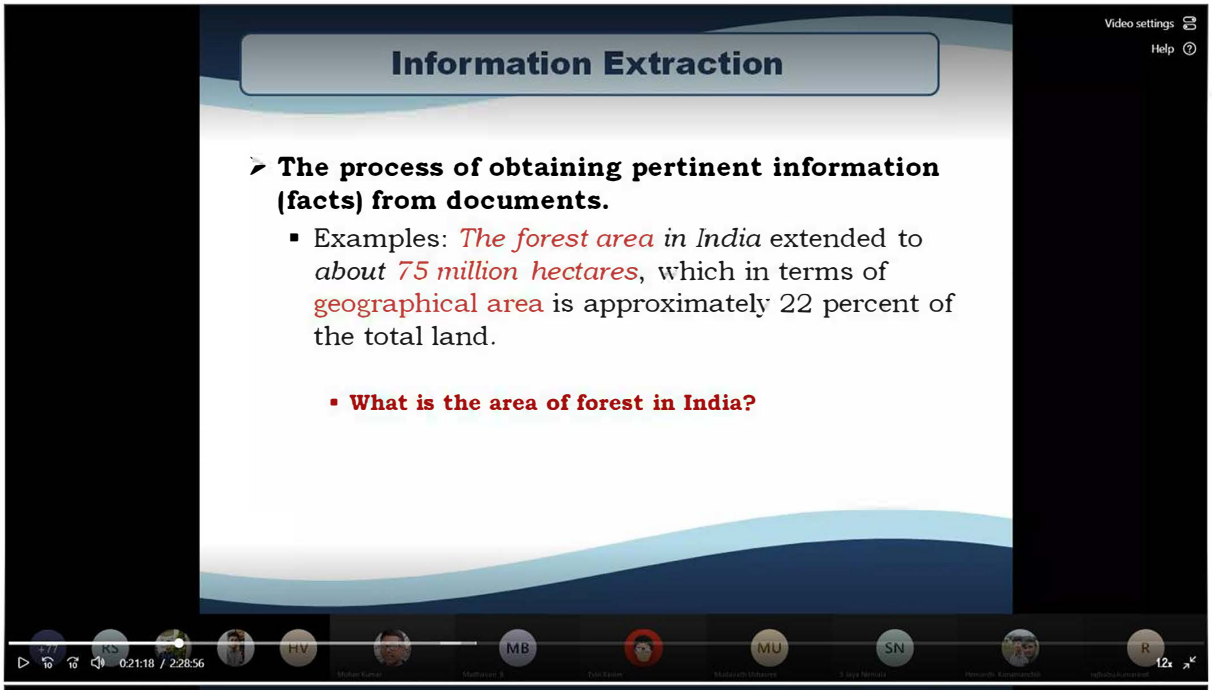
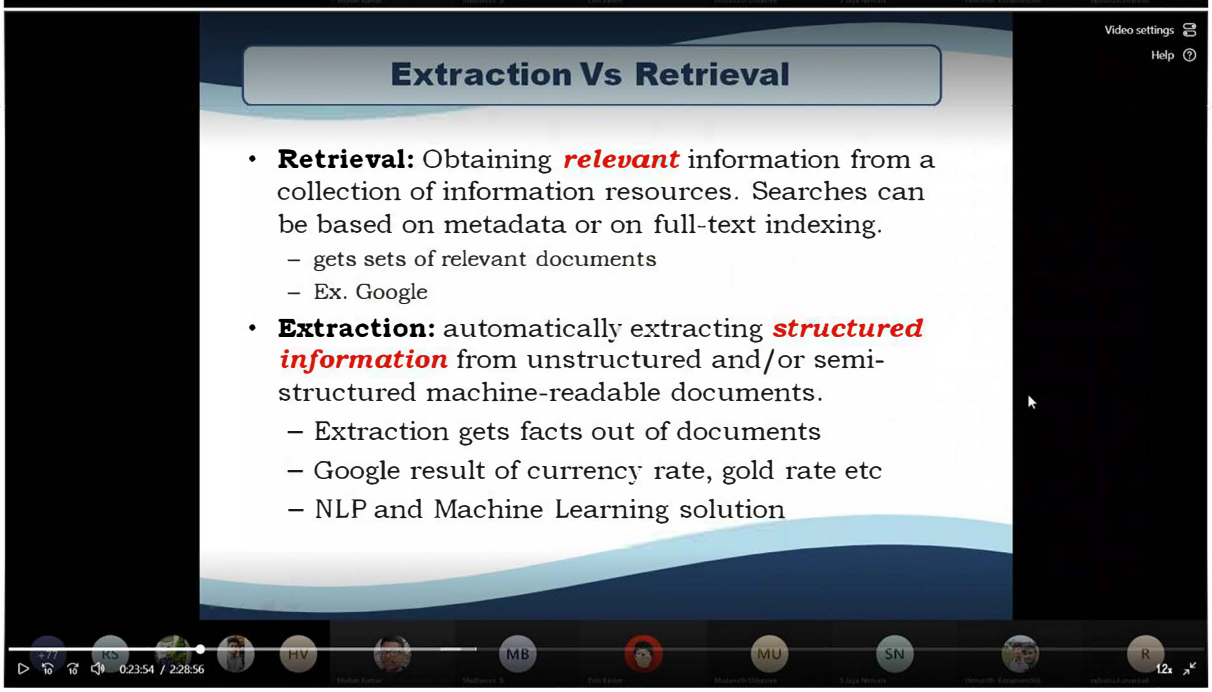

Video settings
Help
Semantic Information Extraction Approaches for Industries
Dr. K. Rajbabu
Manager
Controls & Instrumentation / FB
BHEL, Trichy
0:14:14 / 2:28:56
12x

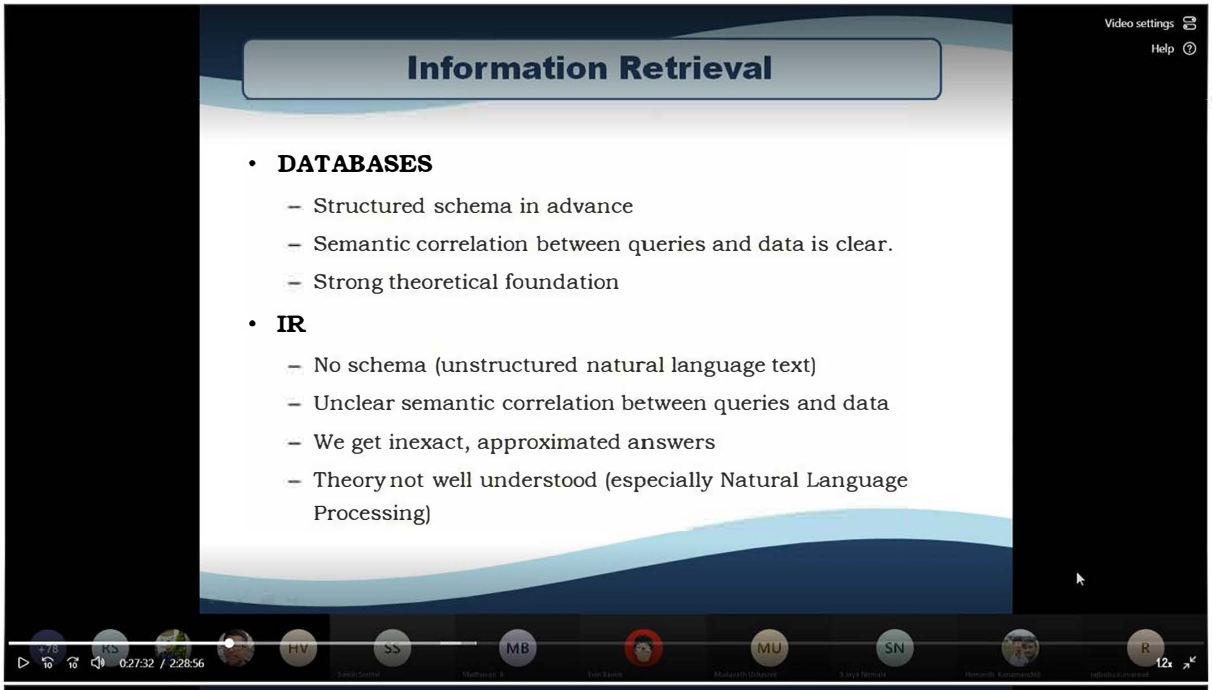

Video settings
Help
Overview of the Discussion
About Information
Extraction vs Retrieval
Information Extraction (IE) Approaches
Information Extraction for Business Intelligence
Knowledge Requirement
Semantic Knowledge
Ontology Model
OBIE Approaches
Recent Developments in Information Extraction
0:14:19 / 2:28:56
12x

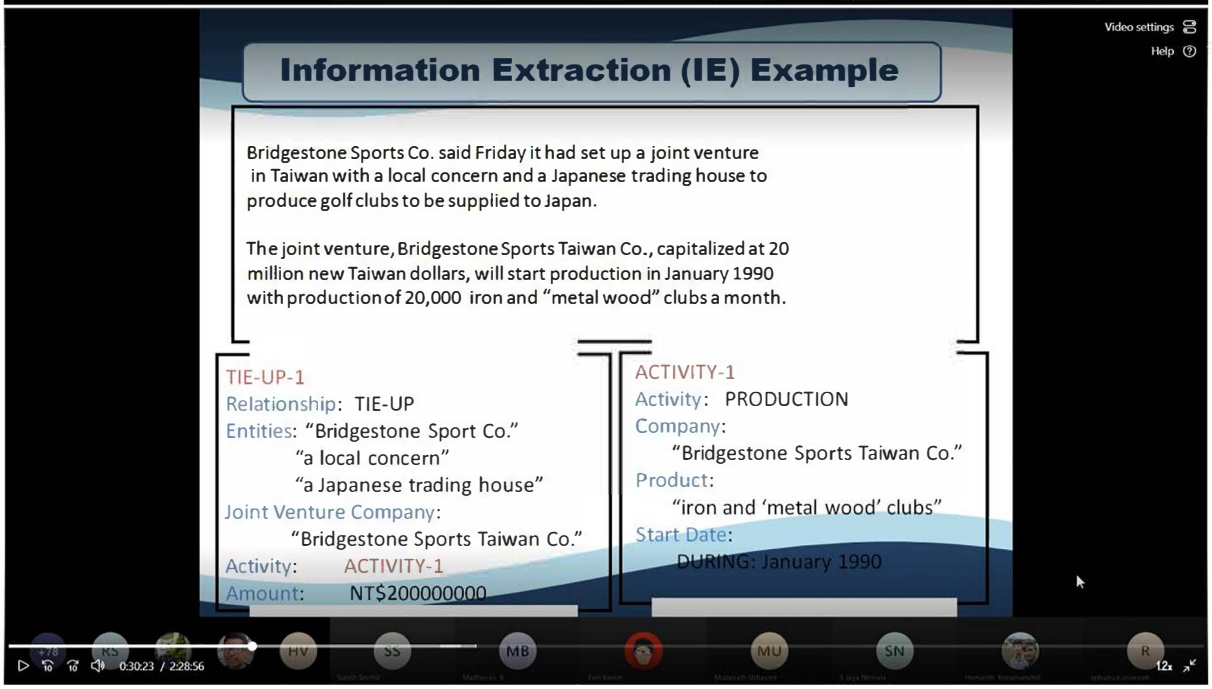

Video settings
Help
Data → Information → Knowledge
Data
✓ Summarizing the data
✓ Averaging the data
✓ Selecting part of the data
✓ Graphing the data
✓ Adding context
✓ Adding value
Information
0:15:12 / 2:28:56
12x

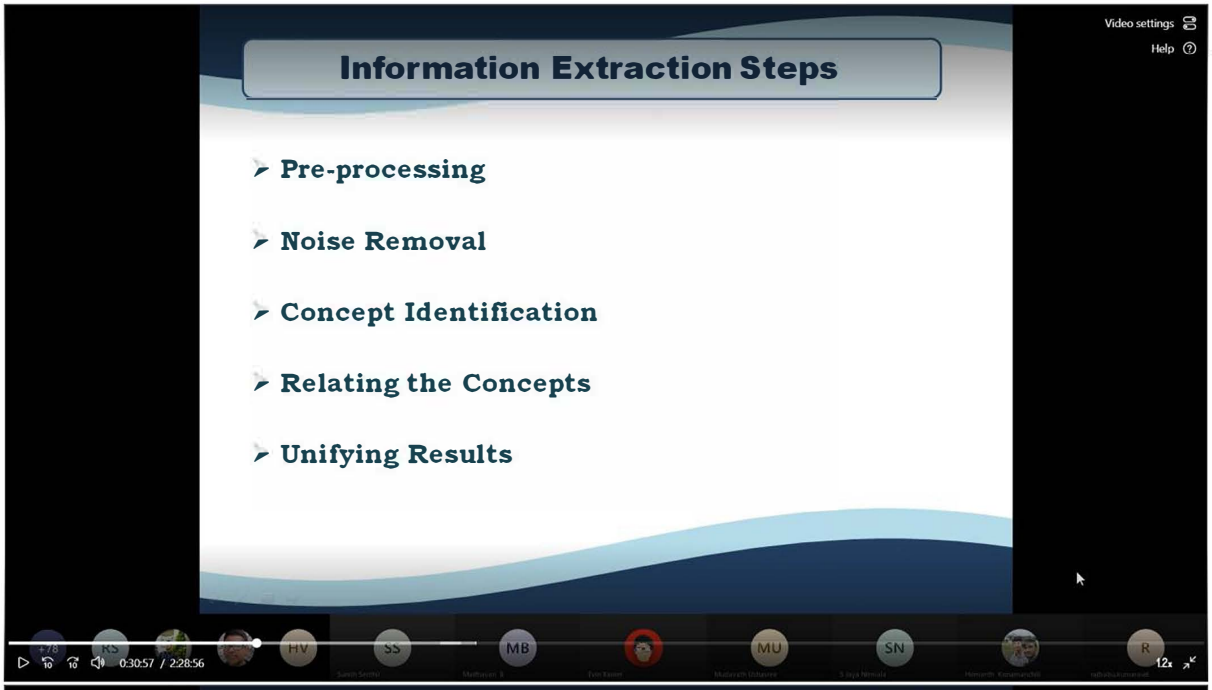

Video settings
Help
Data → Information → Knowledge
Information
✓ How is the info tied to outcomes?
✓ Are there any patterns in the info?
✓ What info is relevant to the problem?
✓ How does this info effect the system?
✓ What is the best way to use the info?
✓ How can we add more value to the info?
Knowledge
0:17:38 / 2:28:56
12x

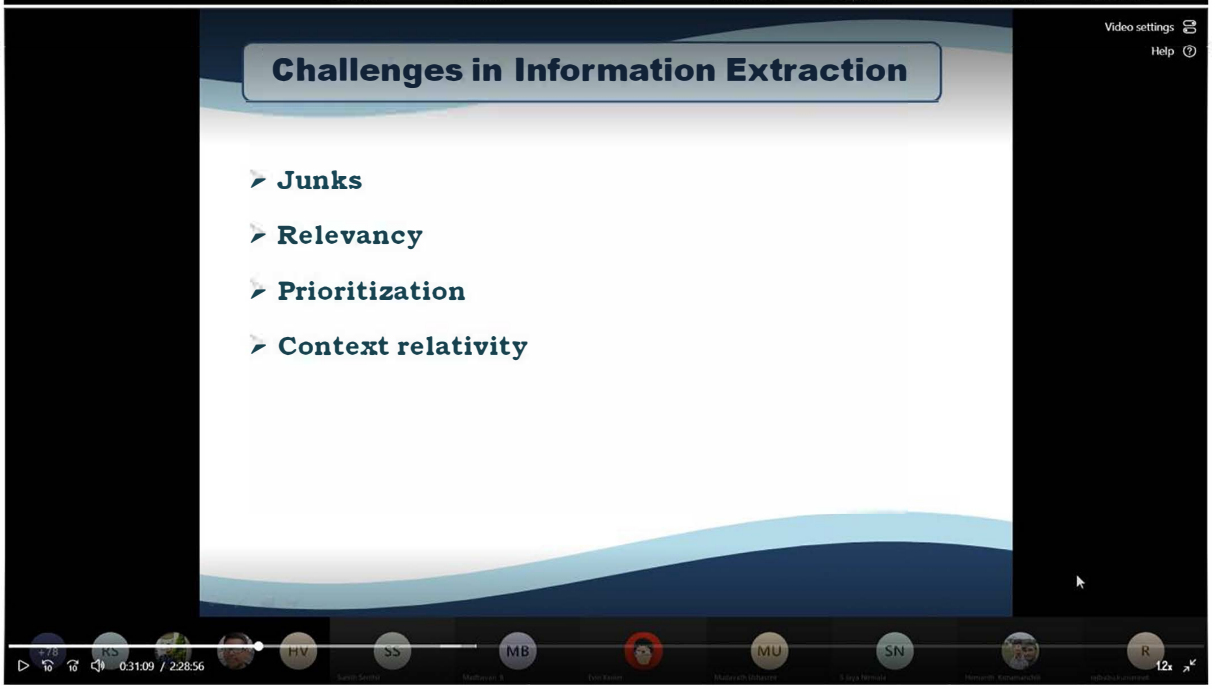

Video settings
Help
Information Extraction
The process of obtaining pertinent information (facts) from documents.
Examples: The forest area in India extended to about 75 million hectares, which in terms of geographical area is approximately 22 percent of the total land.
What is the area of forest in India?
0:21:18 / 2:28:56
12x

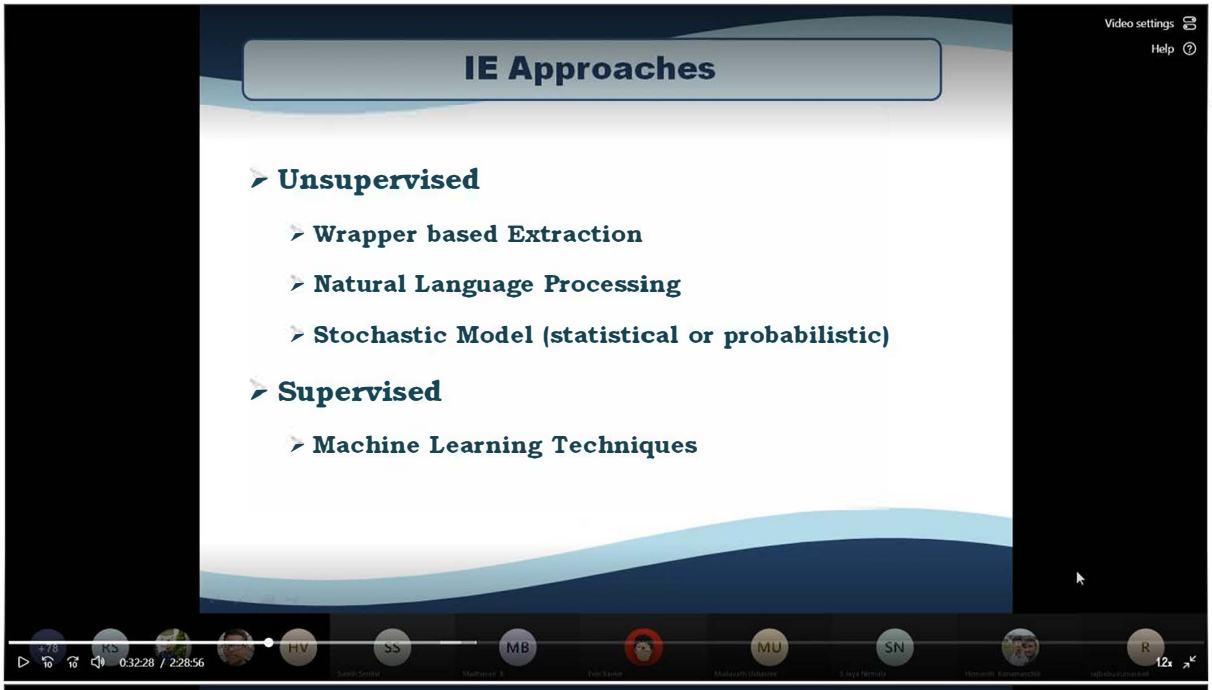

Video settings
Help
Extraction Vs Retrieval
Retrieval: Obtaining relevant information from a collection of information resources. Searches can be based on metadata or on full-text indexing.
gets sets of relevant documents
Ex. Google
Extraction: automatically extracting structured information from unstructured and/or semi-structured machine-readable documents.
Extraction gets facts out of documents
Google result of currency rate, gold rate etc
NLP and Machine Learning solution
0:23:54 / 2:28:56
12x

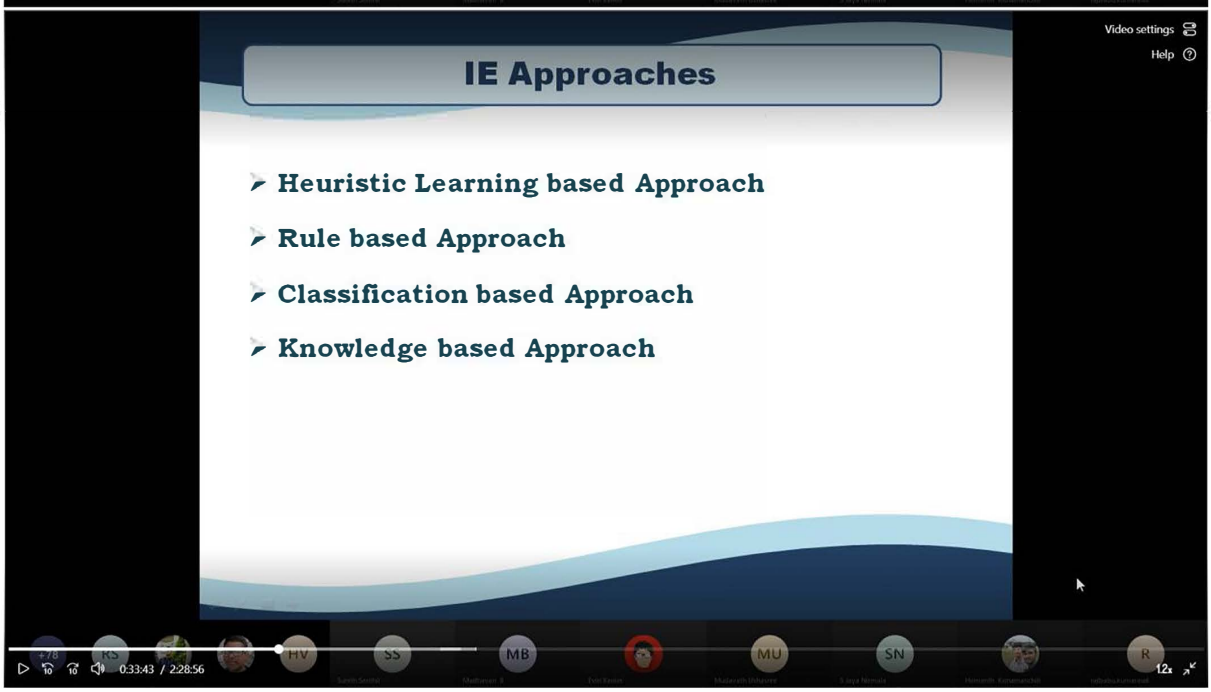

Video settings
Help
Information Retrieval
• DATABASES
– Structured schema in advance
– Semantic correlation between queries and data is clear.
– Strong theoretical foundation
• IR
– No schema (unstructured natural language text)
– Unclear semantic correlation between queries and data
– We get inexact, approximated answers
– Theory not well understood (especially Natural Language Processing)
0:27:32 / 2:28:56
12x


Video settings
Help
Information Extraction (IE) Example
Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.
The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.
TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000
ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990
0:30:23 / 2:28:56
12x


Video settings
Help
Information Extraction Steps
Pre-processing
Noise Removal
Concept Identification
Relating the Concepts
Unifying Results
0:30:57 / 2:28:56
12x

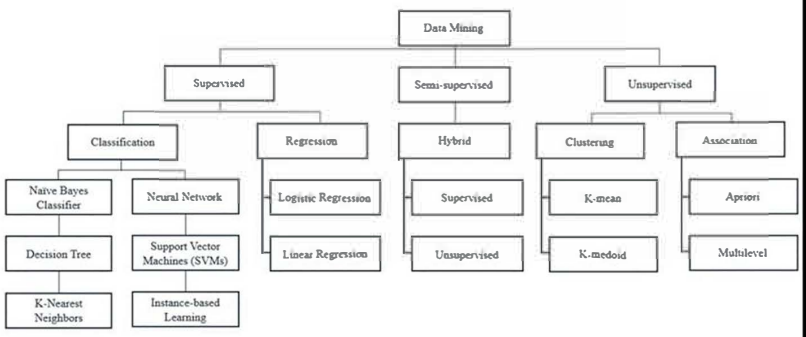

Video settings
Help
Challenges in Information Extraction
Junks
Relevancy
Prioritization
Context relativity
0:31:09 / 2:28:56
12x


Video settings
Help
IE Approaches
Unsupervised
Wrapper based Extraction
Natural Language Processing
Stochastic Model (statistical or probabilistic)
Supervised
Machine Learning Techniques
0:32:28 / 2:28:56
12x


Video settings
Help
IE Approaches
Heuristic Learning based Approach
Rule based Approach
Classification based Approach
Knowledge based Approach
0:33:43 / 2:28:56
12x

Video settings
Help

Taxonomy



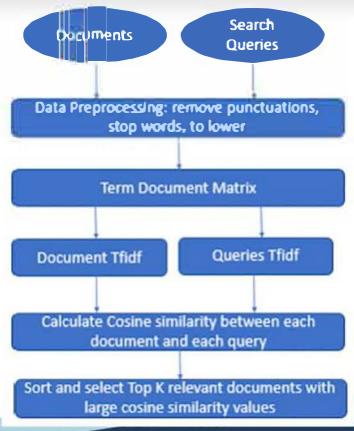
0:35:39 / 2:28:56

12x

Video settings
Help

Vector Space Model

- **Algebraic model of representing text documents**
- **Term- Document Matrix**
- **Similarity Measures: cosine similarity**



0:35:41 / 2:28:56

12x

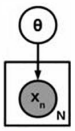
Video settings
Help

Probabilistic Model

➢ **Ex.: Mission to Mars, Playing Chess**

Assume data $\mathbf{X} = \{x_n\}_{n=1}^{N}$ generated from a probability distribution $p(x|\theta)$, in an i.i.d. (independent and identically distributed) fashion

$$x_1, \ldots, x_N \sim p(x|\theta)$$



The form of $p(x|\theta)$ (also called likelihood) depends on the type of the data

Assumptions about parameter θ can be encoded via a prior distribution $p(\theta)$

- Also corresponds to imposing a regularizer over θ (helps in generalization)

Goal: To estimate parameter θ, given data $\mathbf{X}$

Variations of this general view subsume most machine learning problems

- Regression, classification, clustering, dimensionality reduction, etc.

0:40:22 / 2:28:56

Video settings
Help

Probabilistic Model

Distribution	Domain	Picture	Parametric Form
Binomial	Binary		$Bin(x\|N,\theta) \propto \theta^{n}(1-\theta)^{N-n}$
Multinomial	K classes		$Mult(\bar{x}\|\bar{\theta}) \propto \prod \theta_k^{x_k}$
Beta	[0,1]		$Beta(\theta\|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
Gamma	$[0,\infty)$		$Gam(x\|a,b) \propto x^{a-1}\exp(-bx)$
Dirichlet	Simplex		$Dir(\bar{\theta}\|\bar{\alpha}) \propto \prod \theta_k^{\alpha_k-1}$
Gaussian	Reals		$Nor(x\|\mu,\sigma^2) \propto \exp\left[(x-\mu)^2/2\sigma^2\right]$

0:40:44 / 2:28:56

Probabilistic Model

- **Naive Bayes classifier is based on the Naive Bayes algorithm**

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

- **Logistic Regression classifier is based on the Logistic function**

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

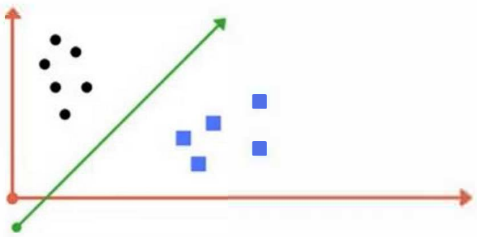
- **Conditional Random Fields (CRF)**

$$P(y, X, \lambda) = \frac{1}{Z(X)} exp\{\sum_{i=1}^{n} \sum_{j} \lambda_j f_i(X, i, y_{i-1}, y_i)\}$$

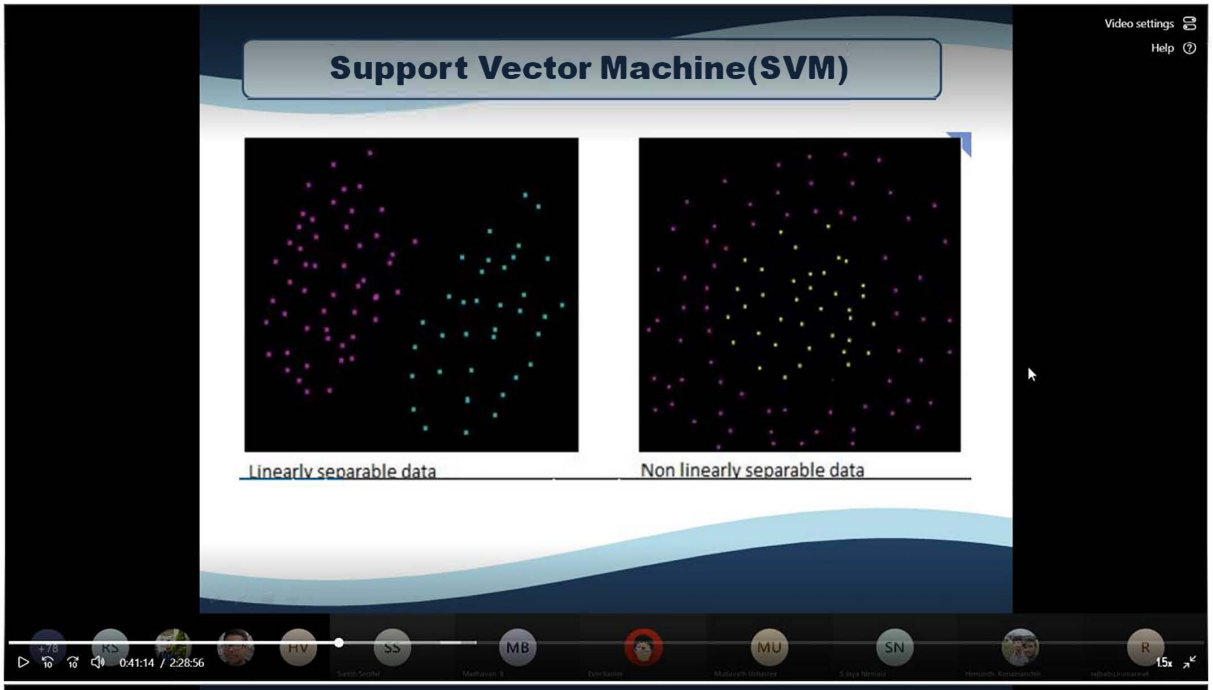
$$\text{Where: } Z(x) = \sum_{y' \in y} \sum_{i=1}^{n} \sum_{j} \lambda_j f_i(X, i, y'_{i-1}, y'_i)$$

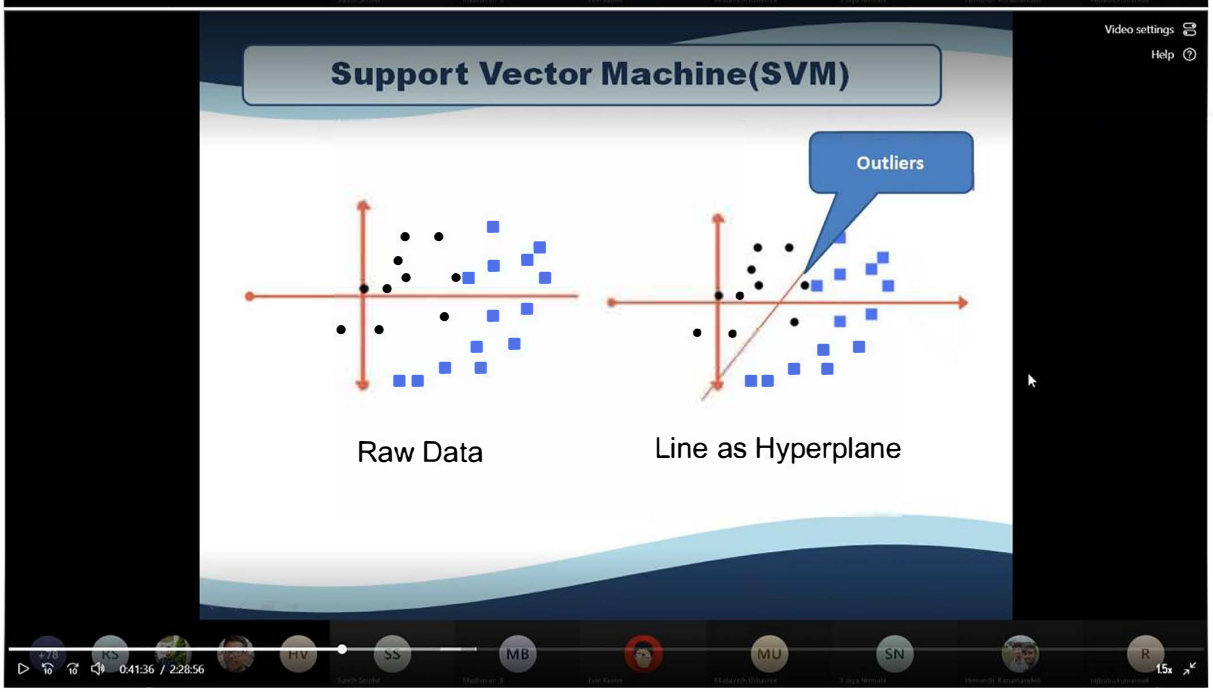
Support Vector Machine(SVM)

- **A Discriminative classifier which takes training data, algorithm and outputs a optimal hyerplane which categorizes new examples**



A line drawn between these data points classify the black dots and blue squares.

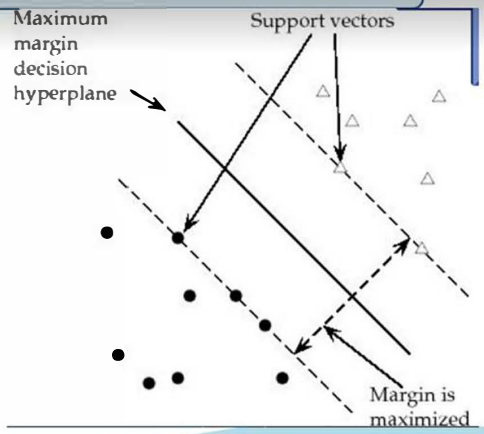

Video settings
Help
Support Vector Machine(SVM)
Linearly separable data
Non linearly separable data
0:41:14 / 2:28:56
15x


Video settings
Help
Support Vector Machine(SVM)
Outliers
Raw Data
Line as Hyperplane
0:41:36 / 2:28:56
15x

Video settings
Help

Support Vector Machine(SVM)

- **Tuning Parameters**
  - **Kernal**
  - **Regularization**
  - **Gamma**
  - **Margin**



0:41:44 / 2:28:56

Video settings
Help

Support Vector Machine(SVM)

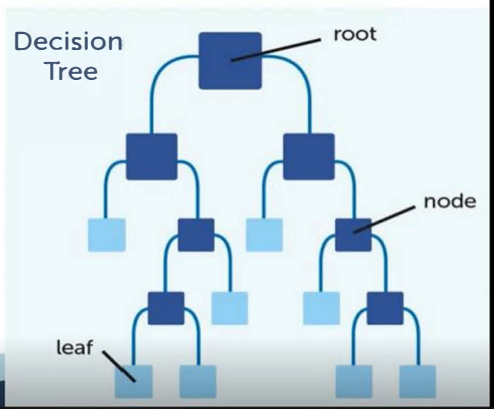
- **Kernel – Mathematical function for transforming data using linear algebra.**
- **Different SVM Algorithm use different kernel functions**
- **Linear**
- **NonLinear**
- **Radial Basis Functions (RBF)**
- **Sigmoid**
- **Polynomial**
- **Exponential**

0:41:58 / 2:28:56

Video settings
Help

Decision Tree

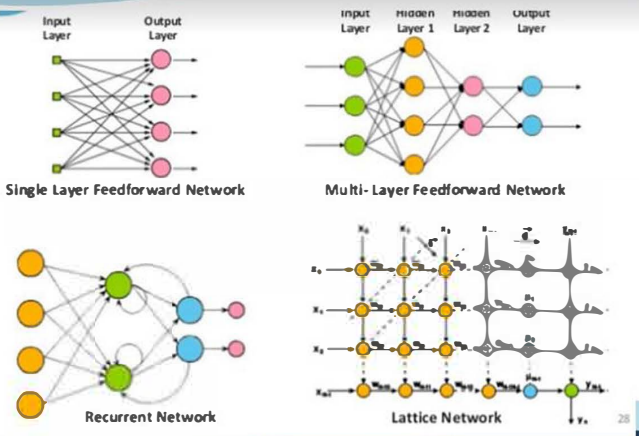
- Data is break down by making decisions using series of conditions.
- Classification and Regression algorithms
- Business Management
- CRM
- Fraudulent Statement
- Energy Consumption



0:42:32 / 2:28:56

Video settings
Help

Neural Networks

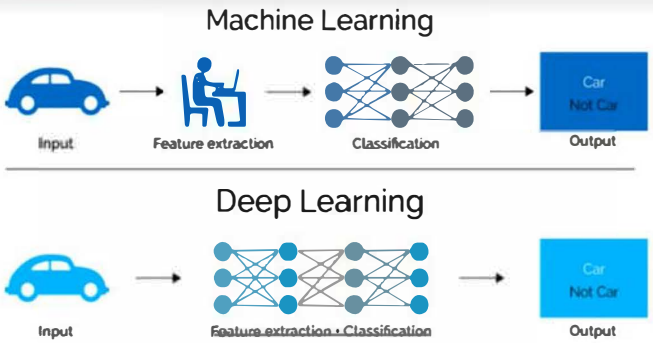


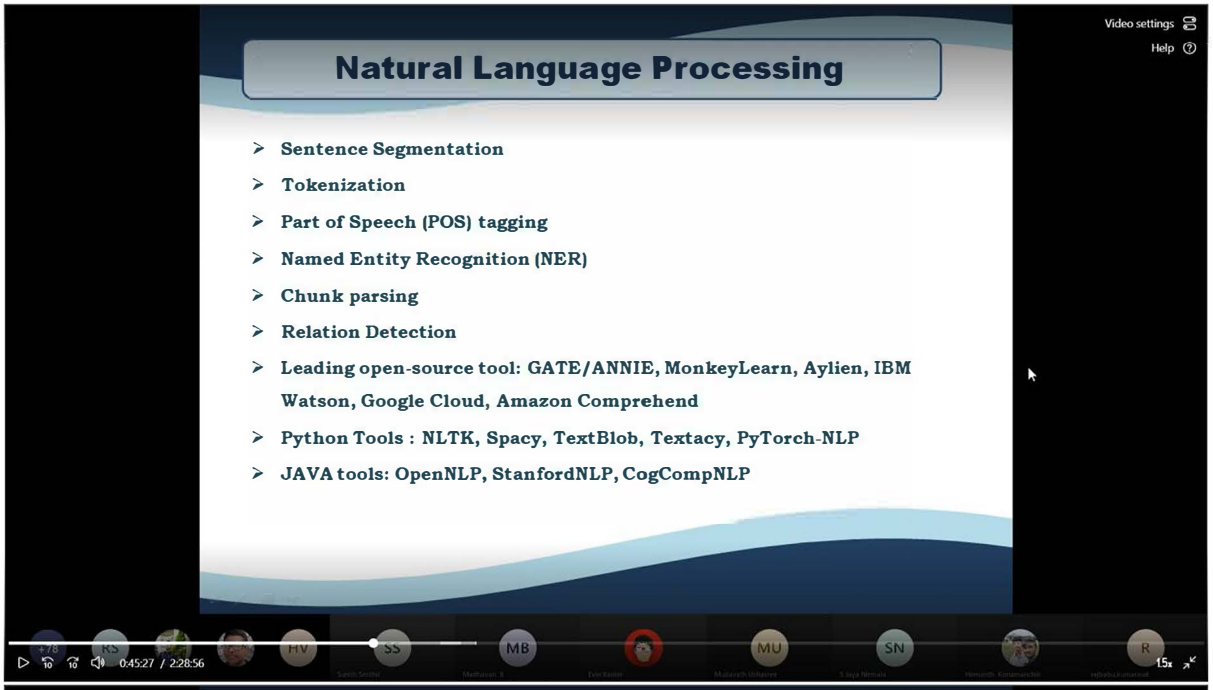
0:43:17 / 2:28:56

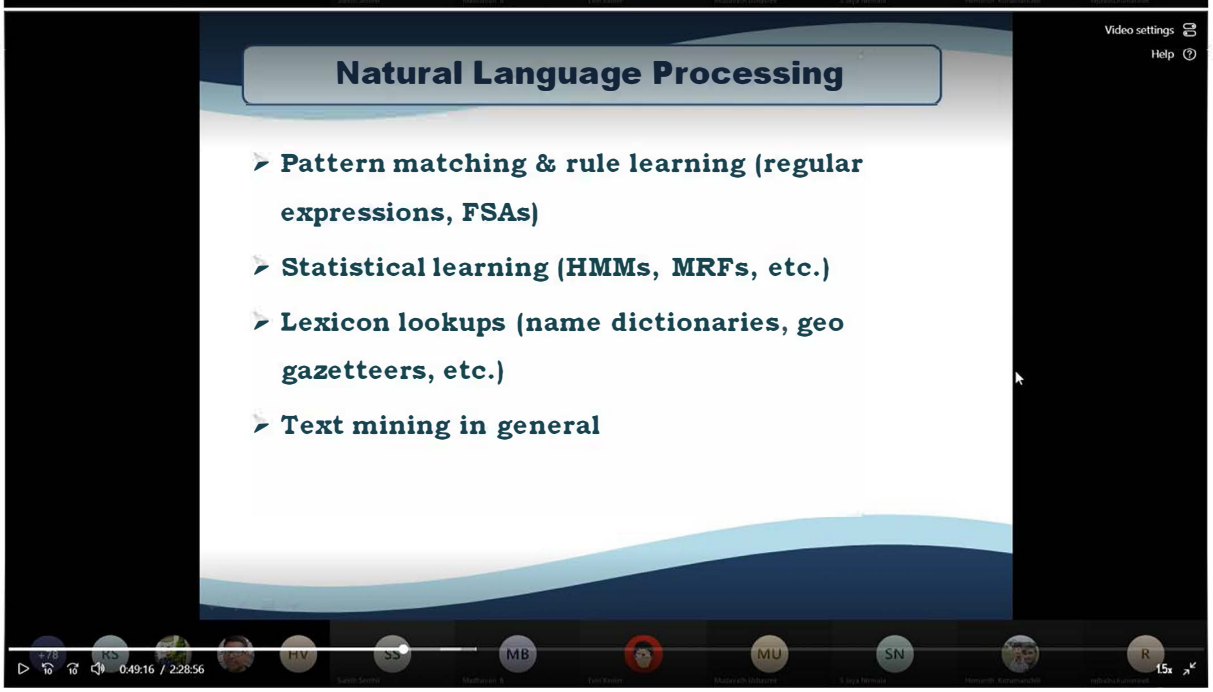
Neural Networks - Application

- **Classification**: Assigning each object to a known specific class
- **Clustering**: Grouping together objects similar to each other
- **Pattern Association**: Presenting of an input sample triggers the generation of specific output pattern
- **Function approximation**: Constructing a function generating almost the same outputs from input data as the modeled process
- **Optimization**: Optimizing function values subject to constraints
- **Forecasting**: Predicting future events on the basis of past history
- **Control**: Determining values for input variables to achieve desired values for output variables

Deep Learning




Video settings
Help
Natural Language Processing
Sentence Segmentation
Tokenization
Part of Speech (POS) tagging
Named Entity Recognition (NER)
Chunk parsing
Relation Detection
Leading open-source tool: GATE/ANNIE, MonkeyLearn, Aylien, IBM Watson, Google Cloud, Amazon Comprehend
Python Tools : NLTK, Spacy, TextBlob, Textacy, PyTorch-NLP
JAVA tools: OpenNLP, StanfordNLP, CogCompNLP
0:45:27 / 2:28:56
1.5x


Video settings
Help
Natural Language Processing
Pattern matching & rule learning (regular expressions, FSAs)
Statistical learning (HMMs, MRFs, etc.)
Lexicon lookups (name dictionaries, geo gazetteers, etc.)
Text mining in general
0:49:16 / 2:28:56
1.5x

Video settings
Help

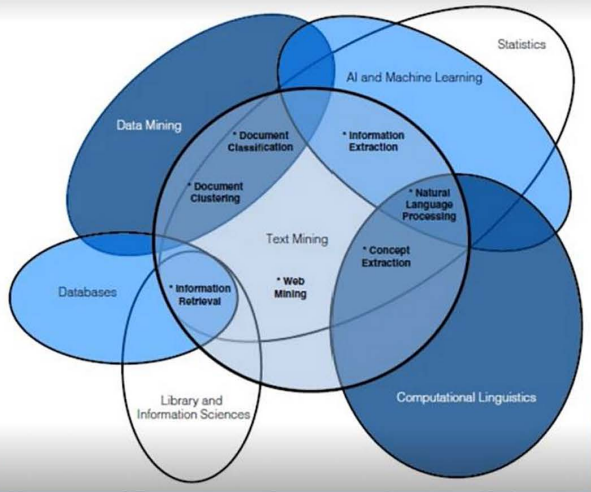
Discussion

- Supervised learning outperforms compared to unsupervised learning
- Domain specific approach
- Hybrid solutions
- Issue based approach based on inputs and expected outcome

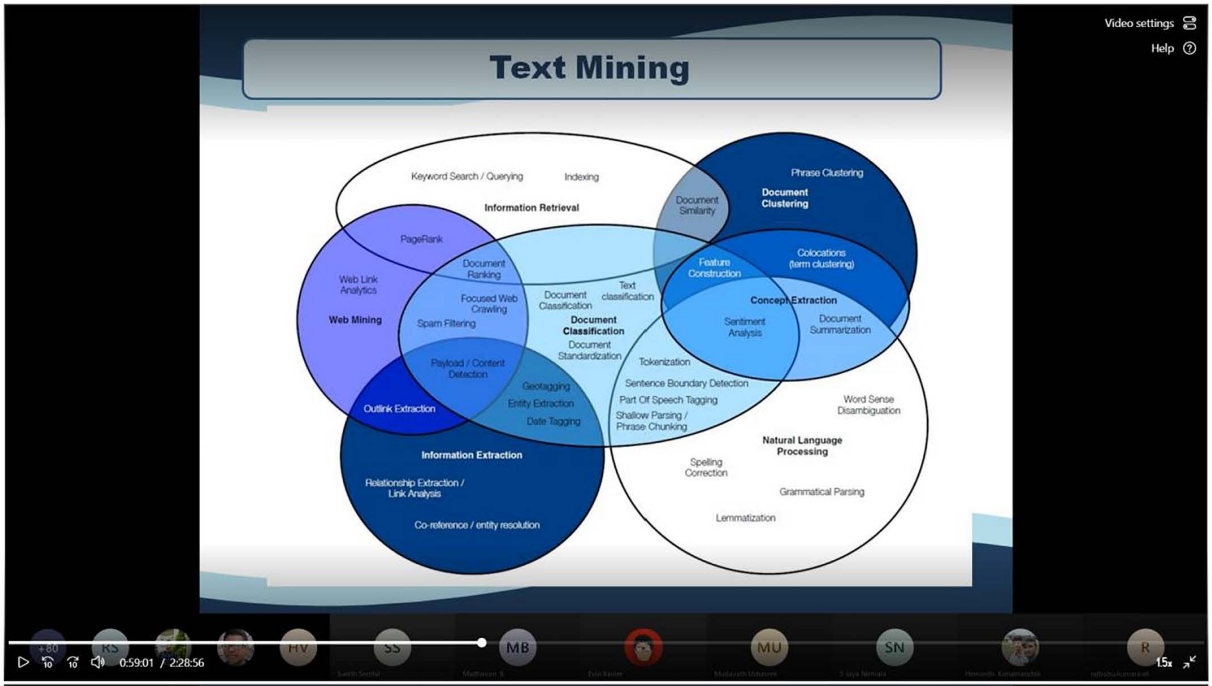
0:49:55 / 2:28:56

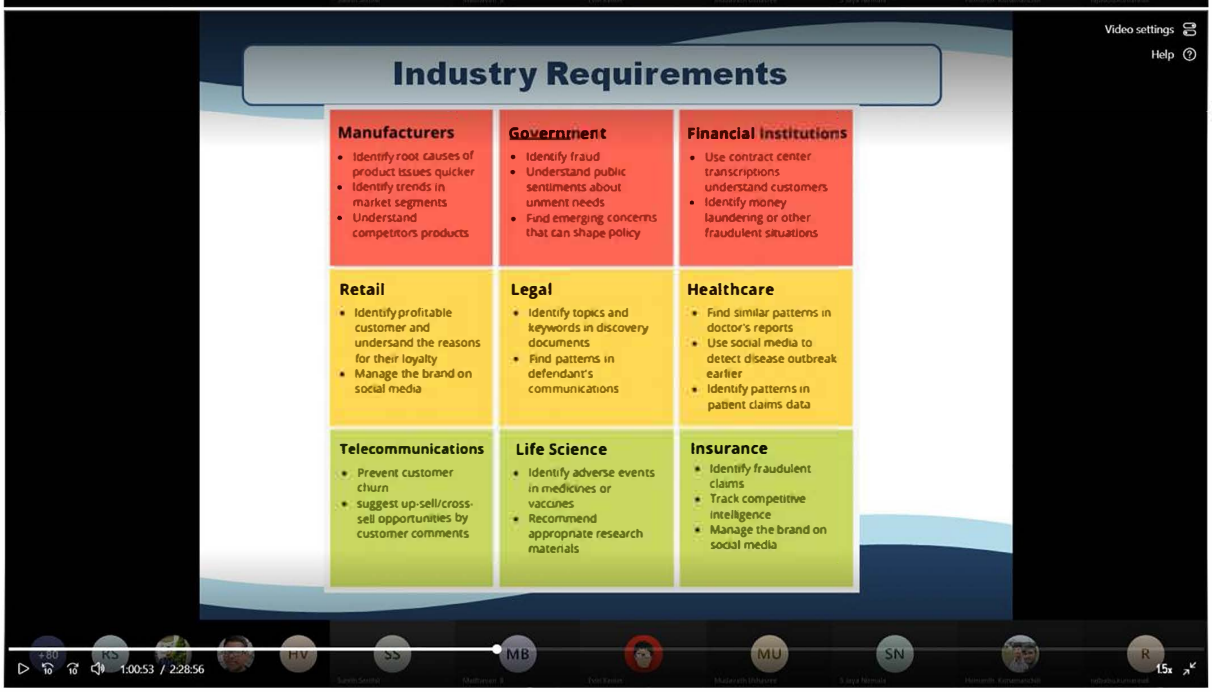
Video settings
Help

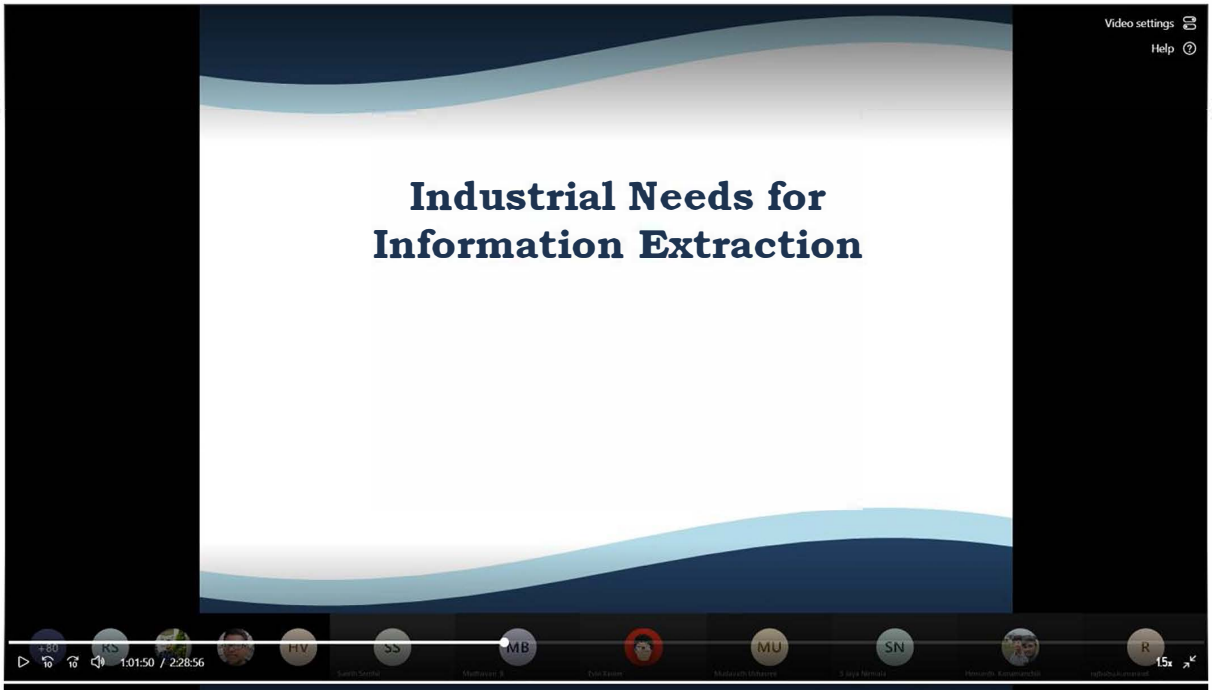
Text Mining

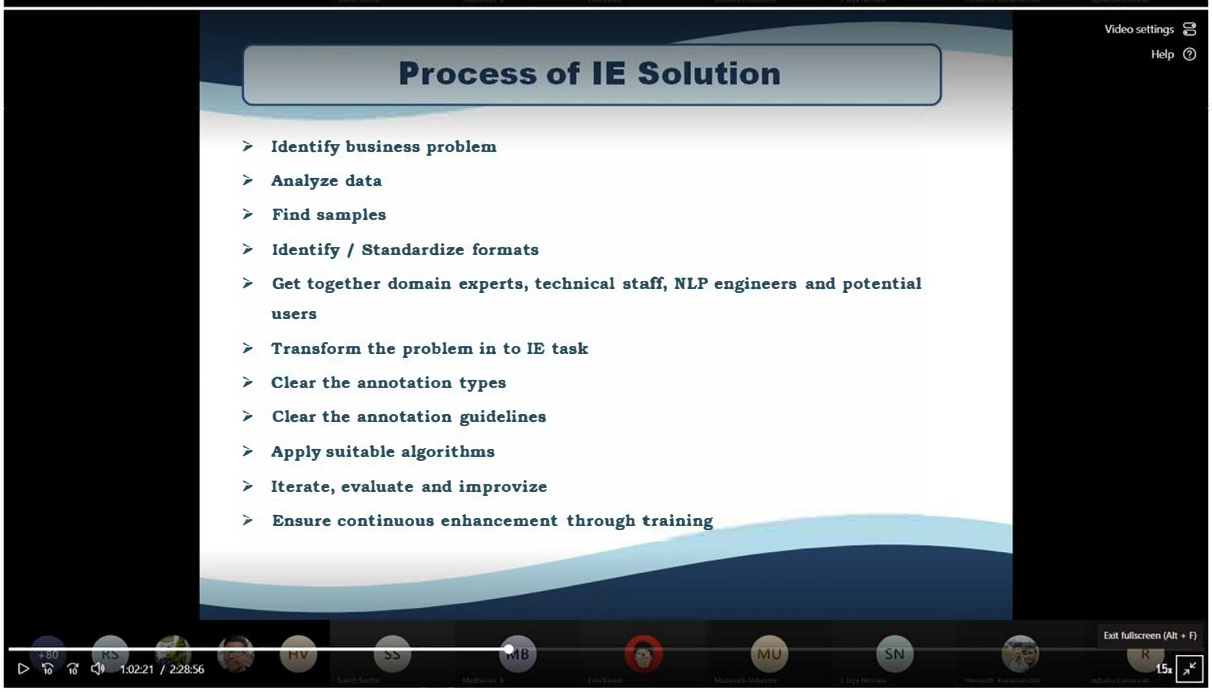


0:52:42 / 2:28:56

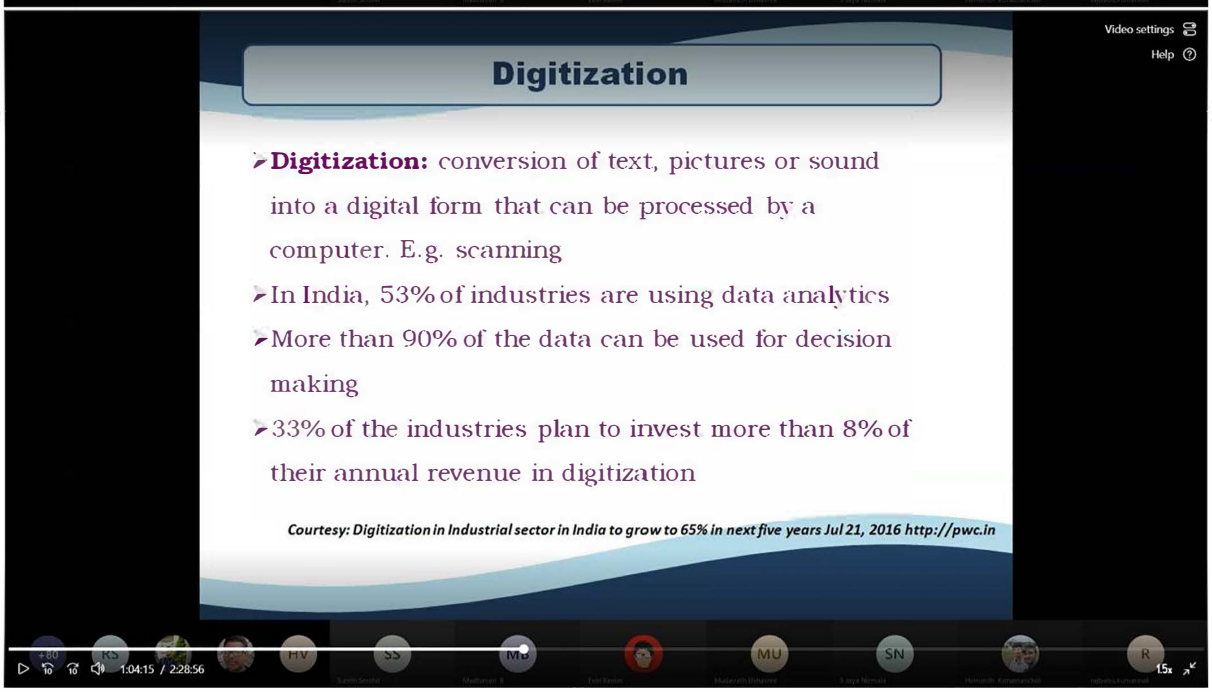

Video settings
Help
Text Mining
Keyword Search / Querying
Indexing
Information Retrieval
PageRank
Web Link Analytics
Web Mining
Focused Web Crawling
Spam Filtering
Document Ranking
Payload / Content Detection
Outlink Extraction
Information Extraction
Relationship Extraction / Link Analysis
Co-reference / entity resolution
Geotagging
Entity Extraction
Date Tagging
Document Classification
Text classification
Document Classification
Document Standardization
Document Similarity
Phrase Clustering
Document Clustering
Feature Construction
Colocations (term clustering)
Concept Extraction
Sentiment Analysis
Document Summarization
Tokenization
Sentence Boundary Detection
Part Of Speech Tagging
Shallow Parsing / Phrase Chunking
Word Sense Disambiguation
Natural Language Processing
Spelling Correction
Grammatical Parsing
Lemmatization
0:59:01 / 2:28:56
15x


Video settings
Help
Industry Requirements
Manufacturers
Identify root causes of product issues quicker
Identify trends in market segments
Understand competitors products
Government
Identify fraud
Understand public sentiments about unmet needs
Find emerging concerns that can shape policy
Financial Institutions
Use contract center transcriptions understand customers
Identify money laundering or other fraudulent situations
Retail
Identify profitable customer and undersand the reasons for their loyalty
Manage the brand on social media
Legal
Identify topics and keywords in discovery documents
Find patterns in defendant's communications
Healthcare
Find similar patterns in doctor's reports
Use social media to detect disease outbreak earlier
Identify patterns in patient claims data
Telecommunications
Prevent customer churn
suggest up-sell/cross-sell opportunities by customer comments
Life Science
Identify adverse events in medicines or vaccines
Recommend appropriate research materials
Insurance
Identify fraudulent claims
Track competitive intelligence
Manage the brand on social media
1:00:53 / 2:28:56
15x


Video settings
Help
Industrial Needs for Information Extraction
1:01:50 / 2:28:56
1.5x

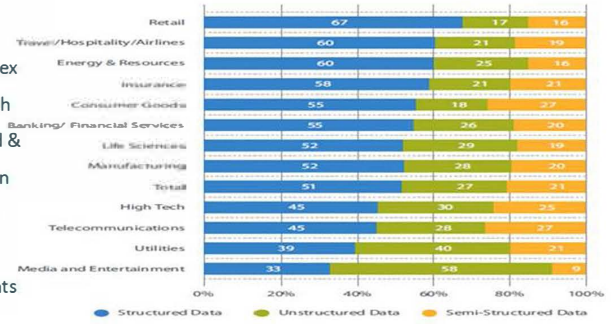

Video settings
Help
Process of IE Solution
Identify business problem
Analyze data
Find samples
Identify / Standardize formats
Get together domain experts, technical staff, NLP engineers and potential users
Transform the problem in to IE task
Clear the annotation types
Clear the annotation guidelines
Apply suitable algorithms
Iterate, evaluate and improvize
Ensure continuous enhancement through training
Exit fullscreen (Alt + F)
1:02:21 / 2:28:56
1.5x


Video settings
Help
Need of Information Extraction in Industry
Increasing thrust towards digitization
Cut-Throat Competitiveness
Showcase technical capability in Business
Industry 4.0
1:03:41 / 2:28:56
15x


Video settings
Help
Digitization
Digitization: conversion of text, pictures or sound into a digital form that can be processed by a computer. E.g. scanning
In India, 53% of industries are using data analytics
More than 90% of the data can be used for decision making
33% of the industries plan to invest more than 8% of their annual revenue in digitization
Courtesy: Digitization in Industrial sector in India to grow to 65% in next five years Jul 21, 2016 http://pwc.in
1:04:15 / 2:28:56
15x

Unstructured Text

Unstructured Text: Text-oriented; complex data structures which are hidden, collapsed & highly interpretable in nature e.g. e-mails, reports, business application documents etc.



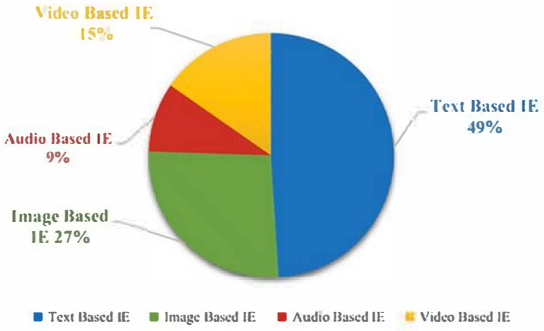
**80% of business-relevant information originates in unstructured form, primarily text*

Courtesy: Big Data Study http://sites.tcs.com/big-data-study/industries-unstructured-data/ 2013

Video settings
Help

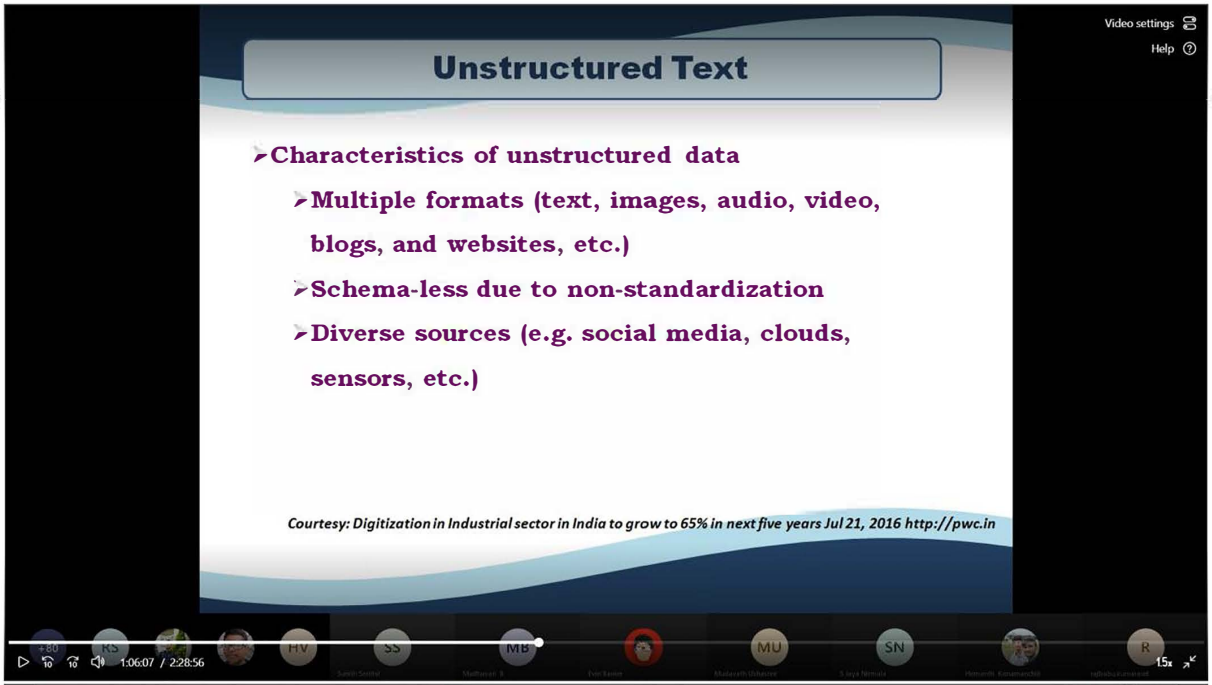
1:04:35 / 2:28:56

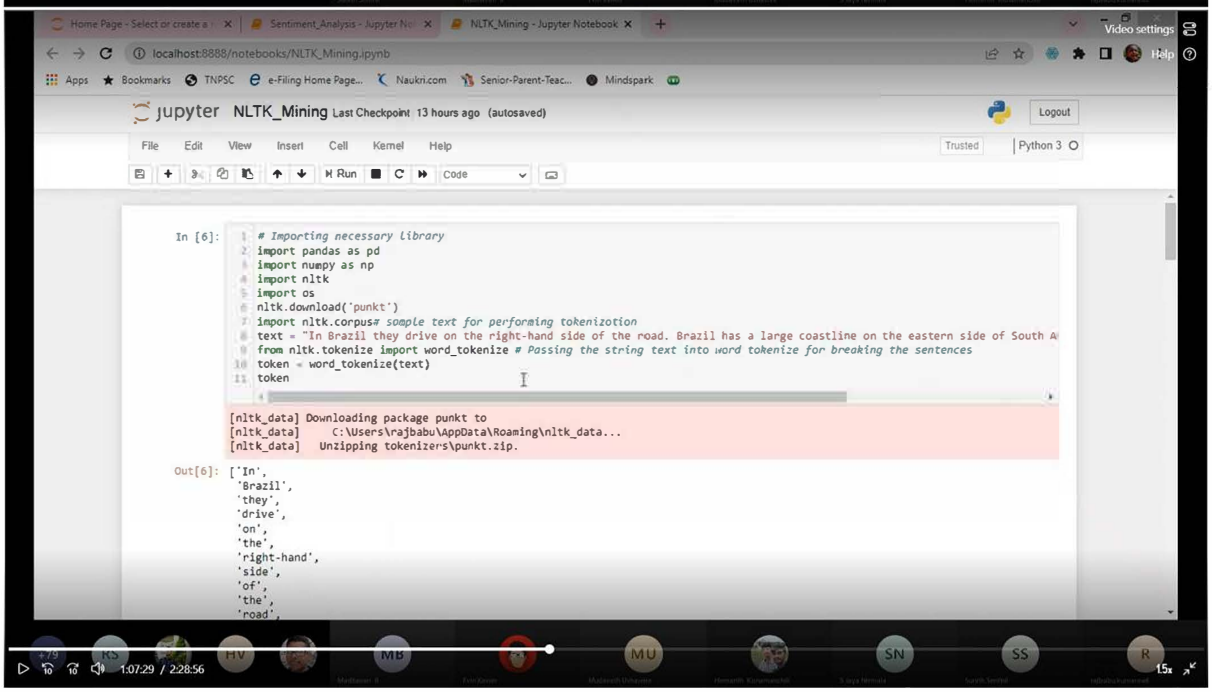
Information Extraction

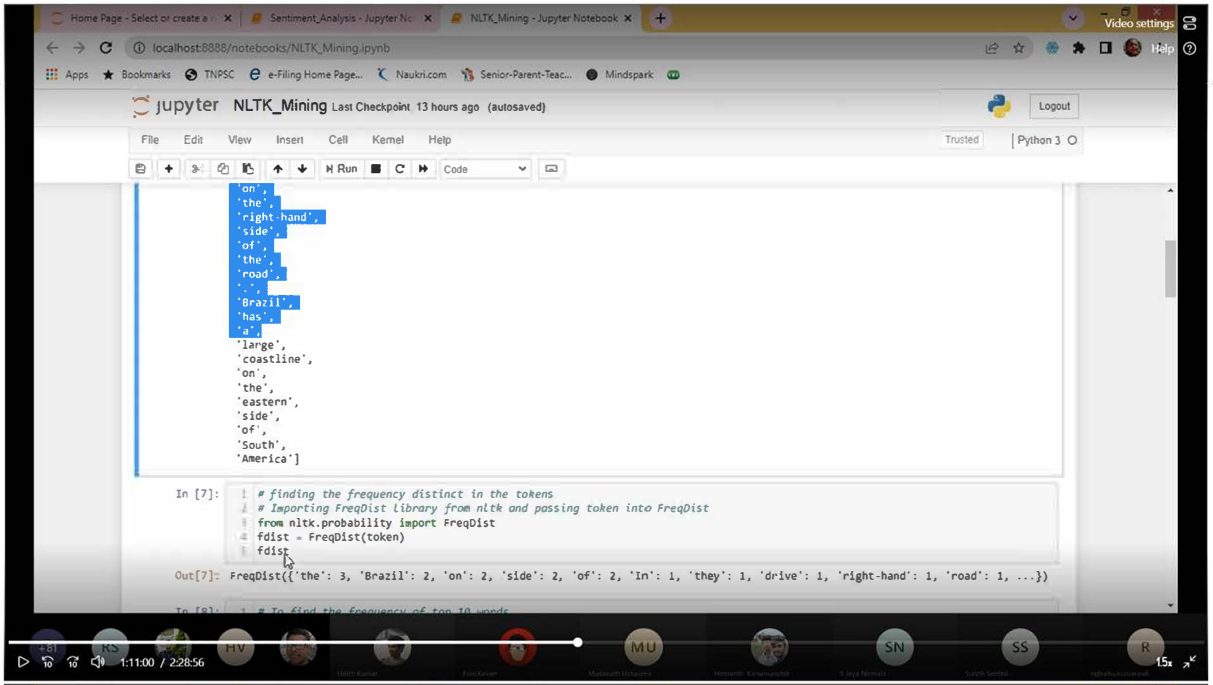


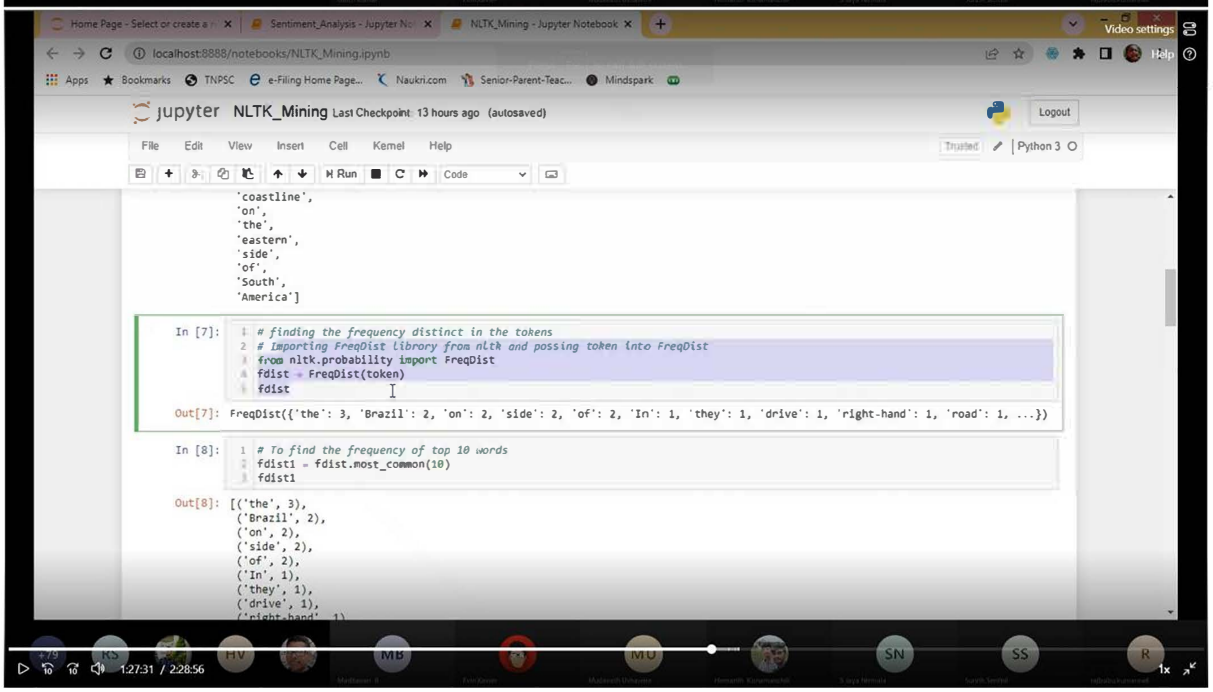
Video settings
Help

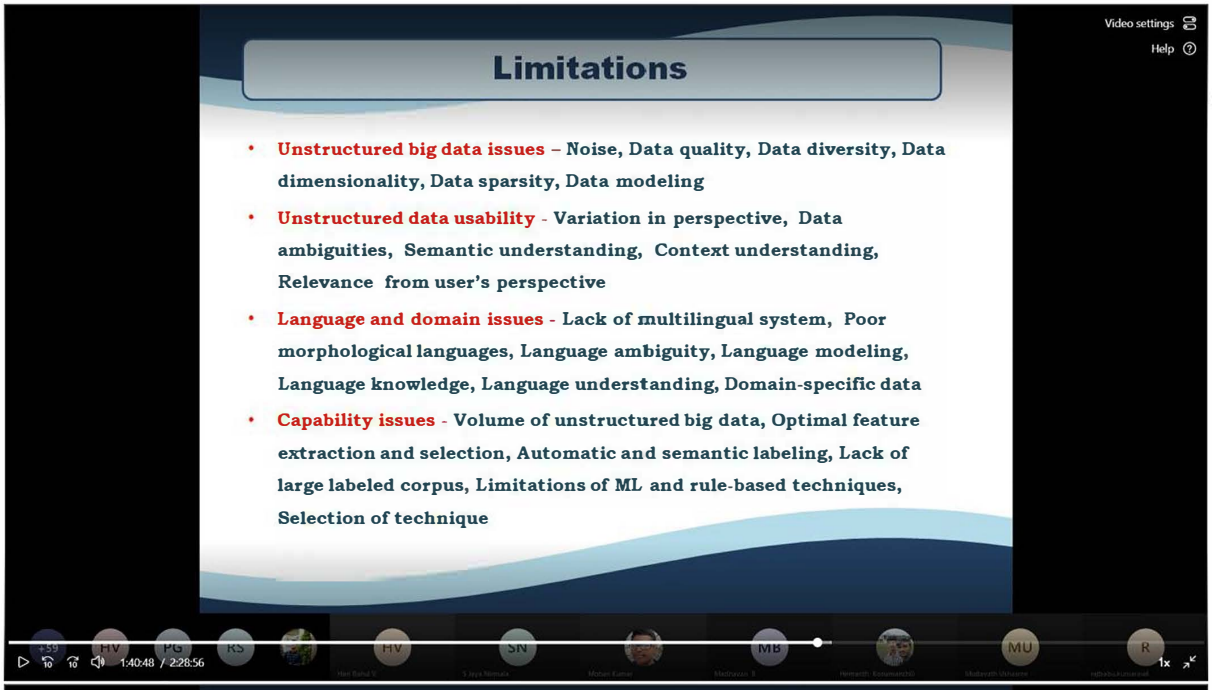
1:05:15 / 2:28:56

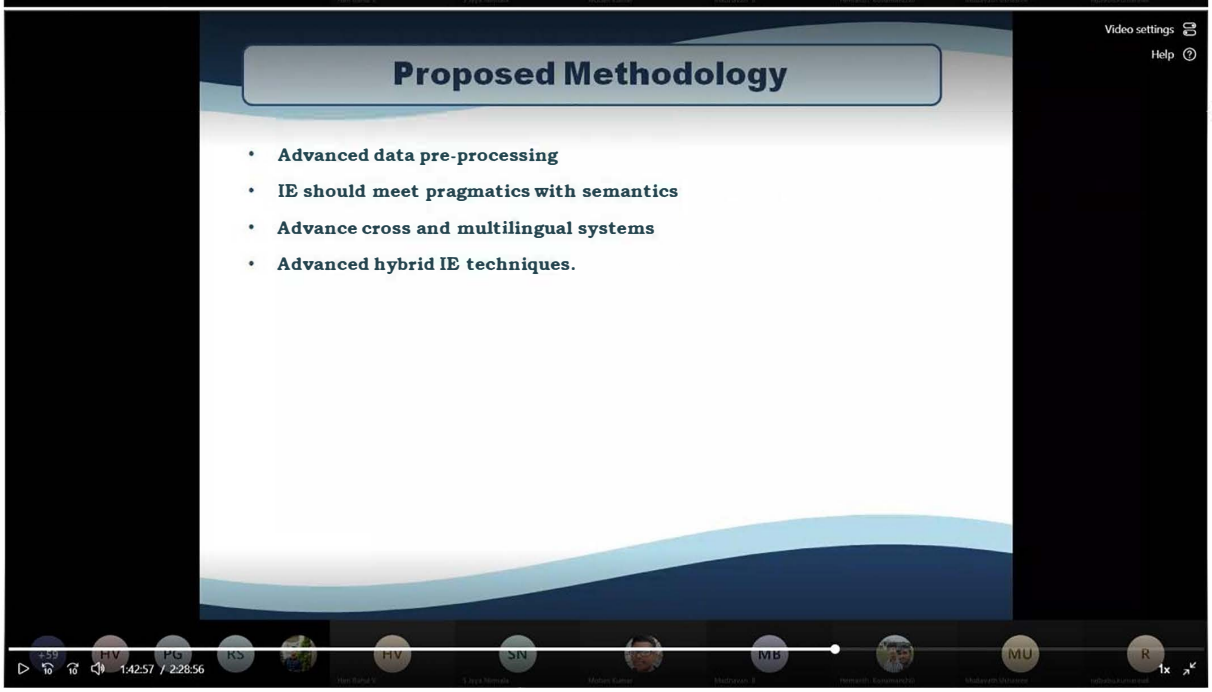

Video settings
Help
Unstructured Text
Characteristics of unstructured data
Multiple formats (text, images, audio, video, blogs, and websites, etc.)
Schema-less due to non-standardization
Diverse sources (e.g. social media, clouds, sensors, etc.)
Courtesy: Digitization in Industrial sector in India to grow to 65% in next five years Jul 21, 2016 http://pwc.in
1:06:07 / 2:28:56

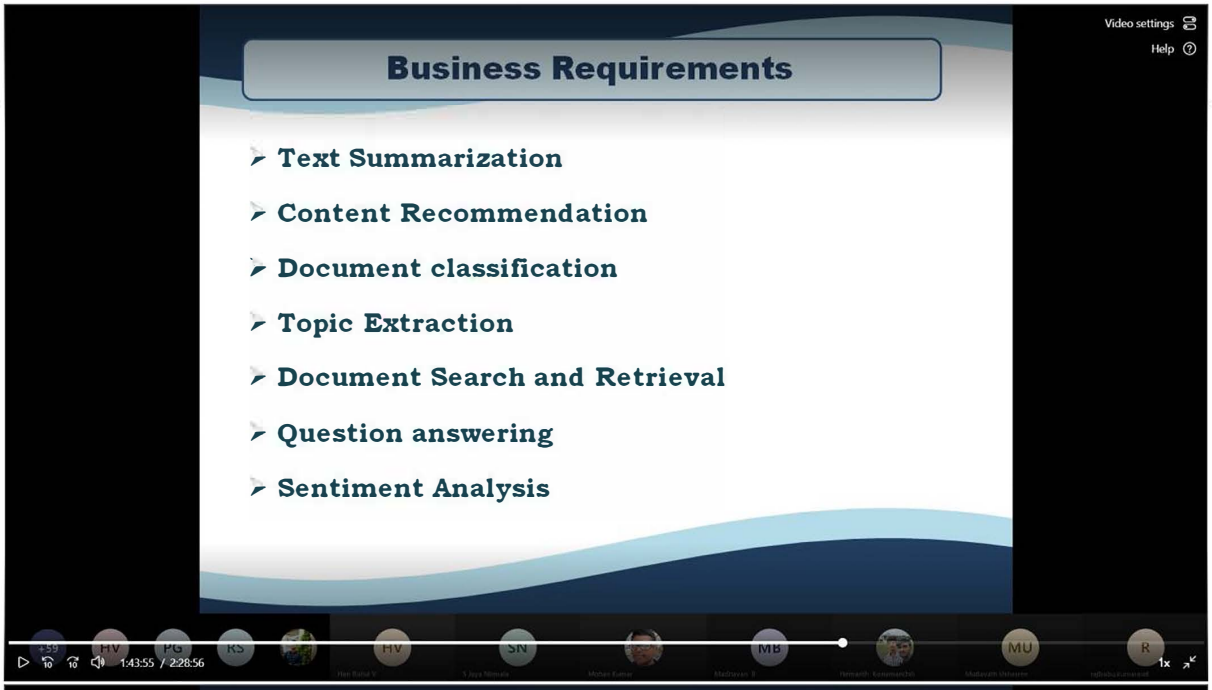

Video settings
Help
localhost:8888/notebooks/NLTK_Mining.ipynb
jupyter NLTK_Mining Last Checkpoint 13 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
In [6]:
# Importing necessary library
import pandas as pd
import numpy as np
import nltk
import os
nltk.download('punkt')
import nltk.corpus# sample text for performing tokenization
text = "In Brazil they drive on the right-hand side of the road. Brazil has a large coastline on the eastern side of South A
from nltk.tokenize import word_tokenize # Passing the string text into word tokenize for breaking the sentences
token = word_tokenize(text)
token
[nltk_data] Downloading package punkt to
[nltk_data] C:\Users\rajbabu\AppData\Roaming\nltk_data...
[nltk_data] Unzipping tokenizers\punkt.zip.
Out[6]:
['In',
'Brazil',
'they',
'drive',
'on',
'the',
'right-hand',
'side',
'of',
'the',
'road',
1:07:29 / 2:28:56

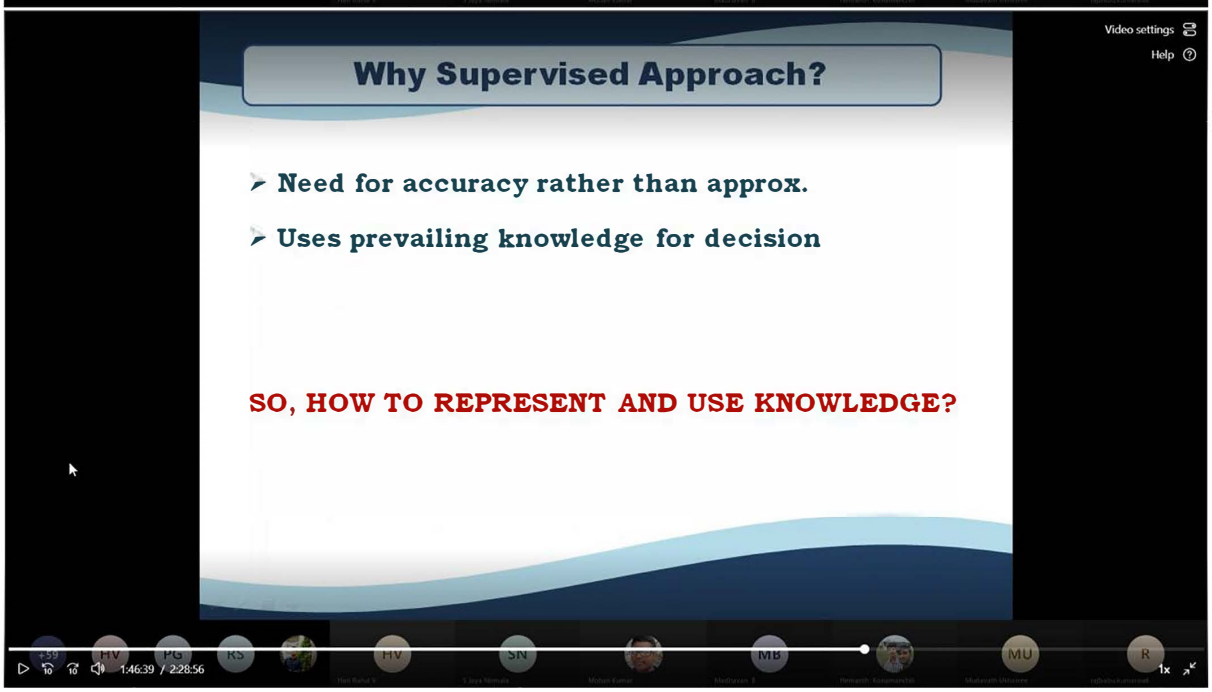

Home Page - Select or create a
Sentiment_Analysis - Jupyter No
NLTK_Mining - Jupyter Notebook
Video settings
localhost:8888/notebooks/NLTK_Mining.ipynb
Help
Apps
Bookmarks
TNPSC
e-Filing Home Page...
Naukri.com
Senior-Parent-Teac...
Mindspark
jupyter NLTK_Mining Last Checkpoint 13 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
'on',
'the',
'right-hand',
'side',
'of',
'the',
'road',
'.',
'Brazil',
'has',
'a',
'large',
'coastline',
'on',
'the',
'eastern',
'side',
'of',
'South',
'America']
In [7]:
# finding the frequency distinct in the tokens
# Importing FreqDist library from nltk and passing token into FreqDist
from nltk.probability import FreqDist
fdist = FreqDist(token)
fdist
Out[7]: FreqDist({'the': 3, 'Brazil': 2, 'on': 2, 'side': 2, 'of': 2, 'In': 1, 'they': 1, 'drive': 1, 'right-hand': 1, 'road': 1, ...})
1:11:00 / 2:28:56
1.5x

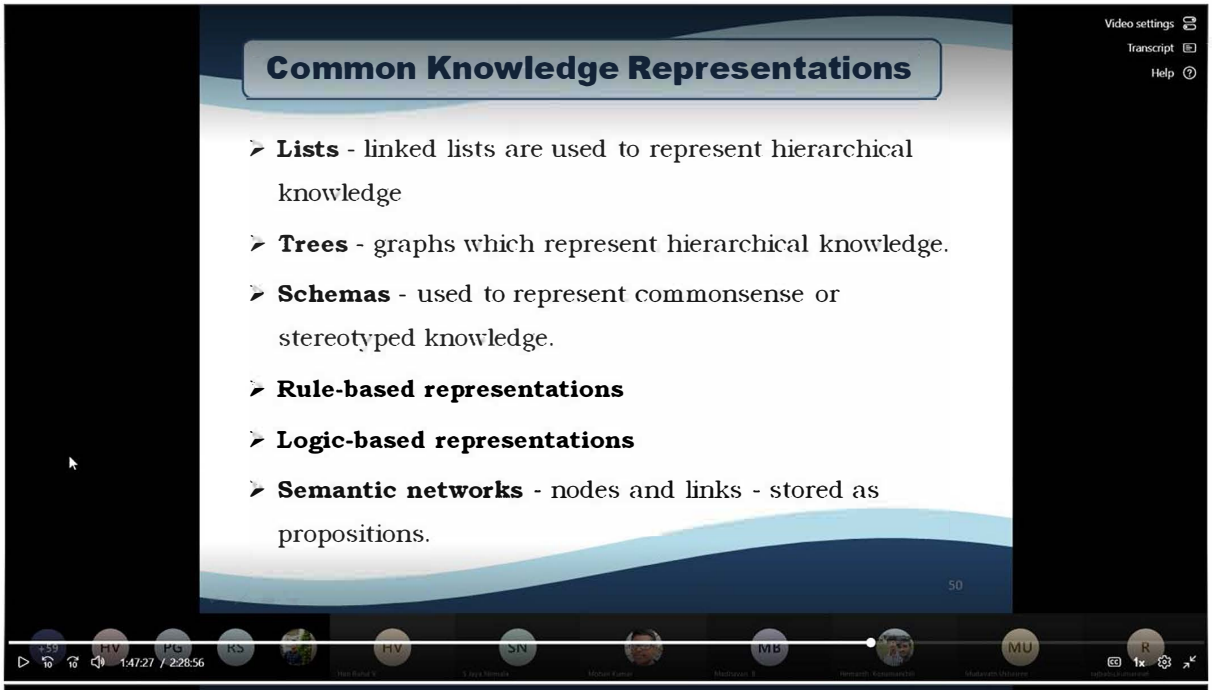

Home Page - Select or create a
Sentiment_Analysis - Jupyter No
NLTK_Mining - Jupyter Notebook
Video settings
localhost:8888/notebooks/NLTK_Mining.ipynb
Help
Apps
Bookmarks
TNPSC
e-Filing Home Page...
Naukri.com
Senior-Parent-Teac...
Mindspark
jupyter NLTK_Mining Last Checkpoint 13 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Code
'coastline',
'on',
'the',
'eastern',
'side',
'of',
'South',
'America']
In [7]:
# finding the frequency distinct in the tokens
# Importing FreqDist library from nltk and passing token into FreqDist
from nltk.probability import FreqDist
fdist = FreqDist(token)
fdist
Out[7]: FreqDist({'the': 3, 'Brazil': 2, 'on': 2, 'side': 2, 'of': 2, 'In': 1, 'they': 1, 'drive': 1, 'right-hand': 1, 'road': 1, ...})
In [8]:
# To find the frequency of top 10 words
fdist1 = fdist.most_common(10)
fdist1
Out[8]: [('the', 3),
('Brazil', 2),
('on', 2),
('side', 2),
('of', 2),
('In', 1),
('they', 1),
('drive', 1),
('right-hand', 1)
1:27:31 / 2:28:56
1x

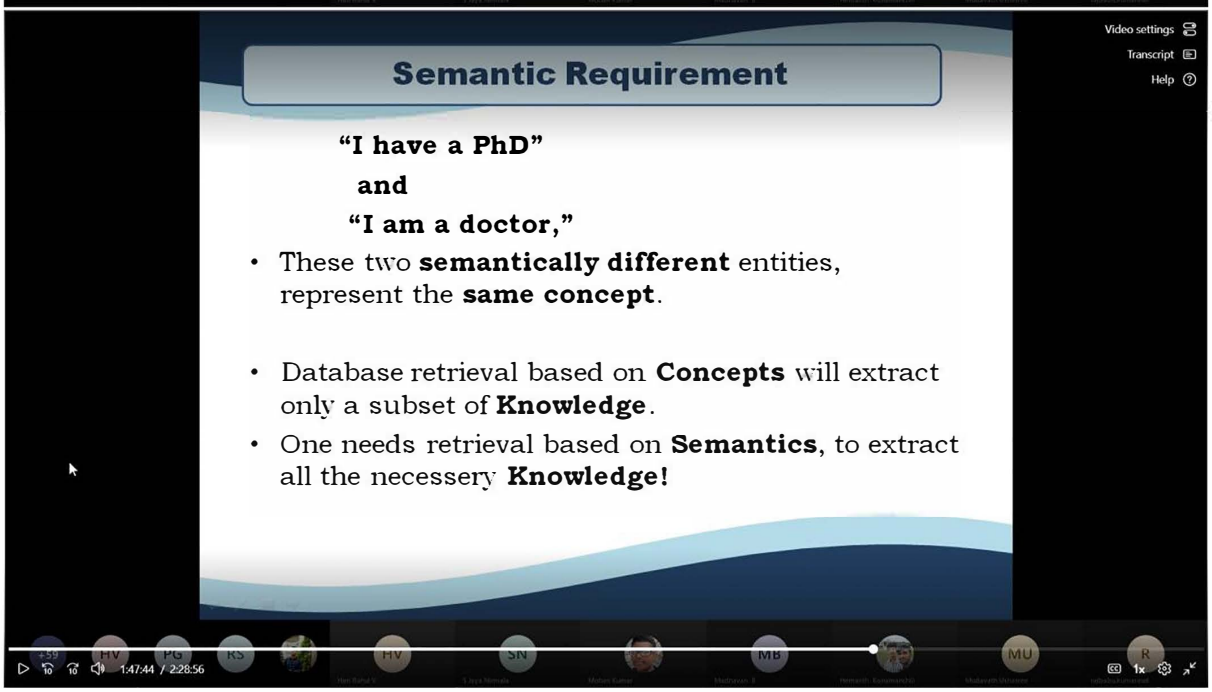

Video settings
Help
Limitations
• Unstructured big data issues – Noise, Data quality, Data diversity, Data dimensionality, Data sparsity, Data modeling
• Unstructured data usability - Variation in perspective, Data ambiguities, Semantic understanding, Context understanding, Relevance from user's perspective
• Language and domain issues - Lack of multilingual system, Poor morphological languages, Language ambiguity, Language modeling, Language knowledge, Language understanding, Domain-specific data
• Capability issues - Volume of unstructured big data, Optimal feature extraction and selection, Automatic and semantic labeling, Lack of large labeled corpus, Limitations of ML and rule-based techniques, Selection of technique
1:40:48 / 2:28:56
1x


Video settings
Help
Proposed Methodology
• Advanced data pre-processing
• IE should meet pragmatics with semantics
• Advance cross and multilingual systems
• Advanced hybrid IE techniques.
1:42:57 / 2:28:56
1x


Video settings
Help
Business Requirements
Text Summarization
Content Recommendation
Document classification
Topic Extraction
Document Search and Retrieval
Question answering
Sentiment Analysis
1:43:55 / 2:28:56
1x

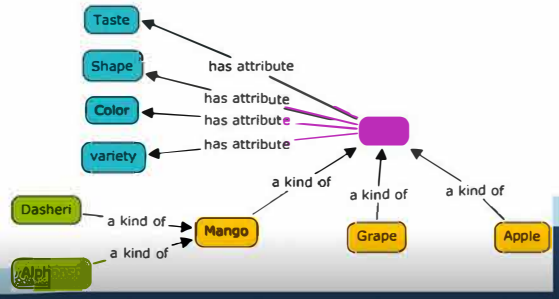

Video settings
Help
Why Supervised Approach?
Need for accuracy rather than approx.
Uses prevailing knowledge for decision
SO, HOW TO REPRESENT AND USE KNOWLEDGE?
1:46:39 / 2:28:56
1x


Video settings
Transcript
Help
Common Knowledge Representations
Lists - linked lists are used to represent hierarchical knowledge
Trees - graphs which represent hierarchical knowledge.
Schemas - used to represent commonsense or stereotyped knowledge.
Rule-based representations
Logic-based representations
Semantic networks - nodes and links - stored as propositions.
50
1:47:27 / 2:28:56

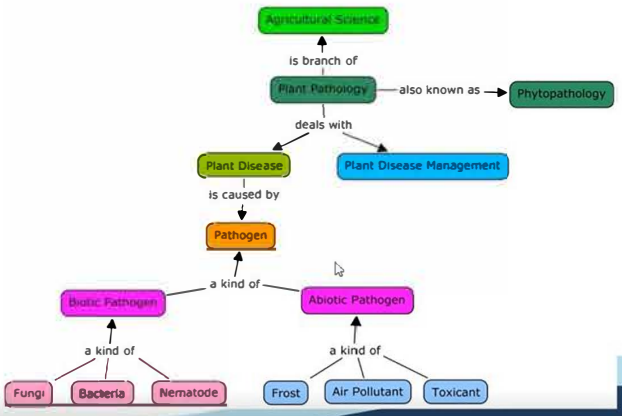

Video settings
Transcript
Help
Semantic Requirement
"I have a PhD"
and
"I am a doctor,"
These two semantically different entities, represent the same concept.
Database retrieval based on Concepts will extract only a subset of Knowledge.
One needs retrieval based on Semantics, to extract all the necessery Knowledge!
1:47:44 / 2:28:56

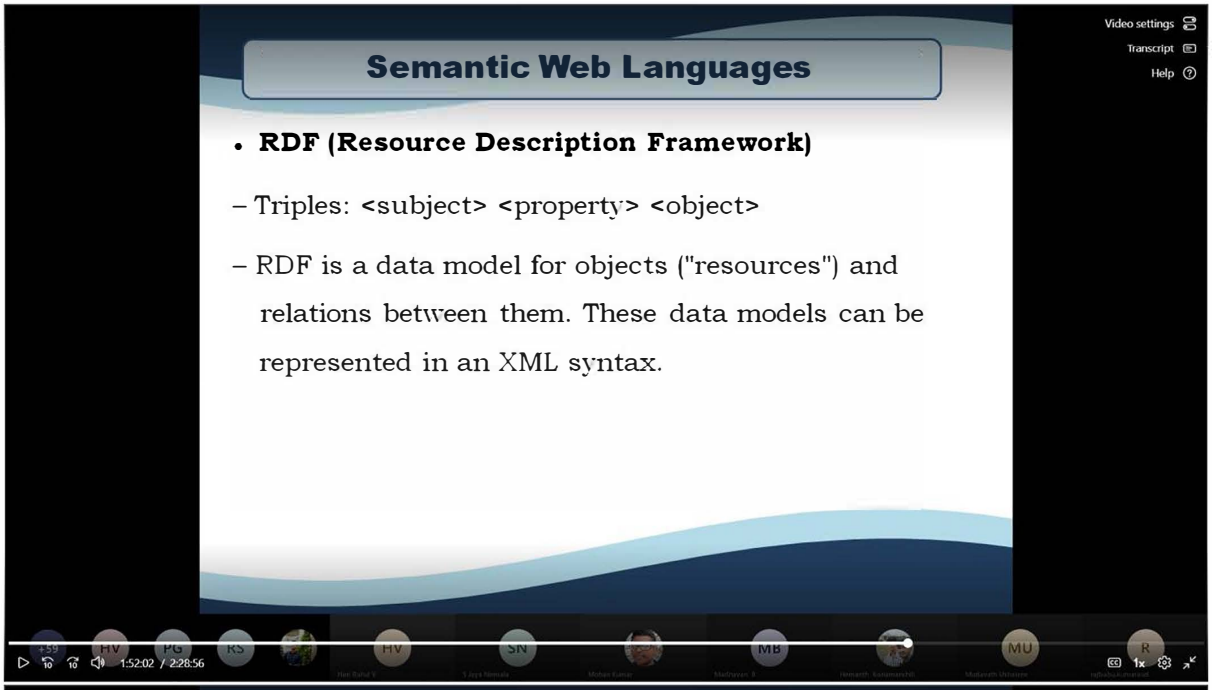
Semantic Network

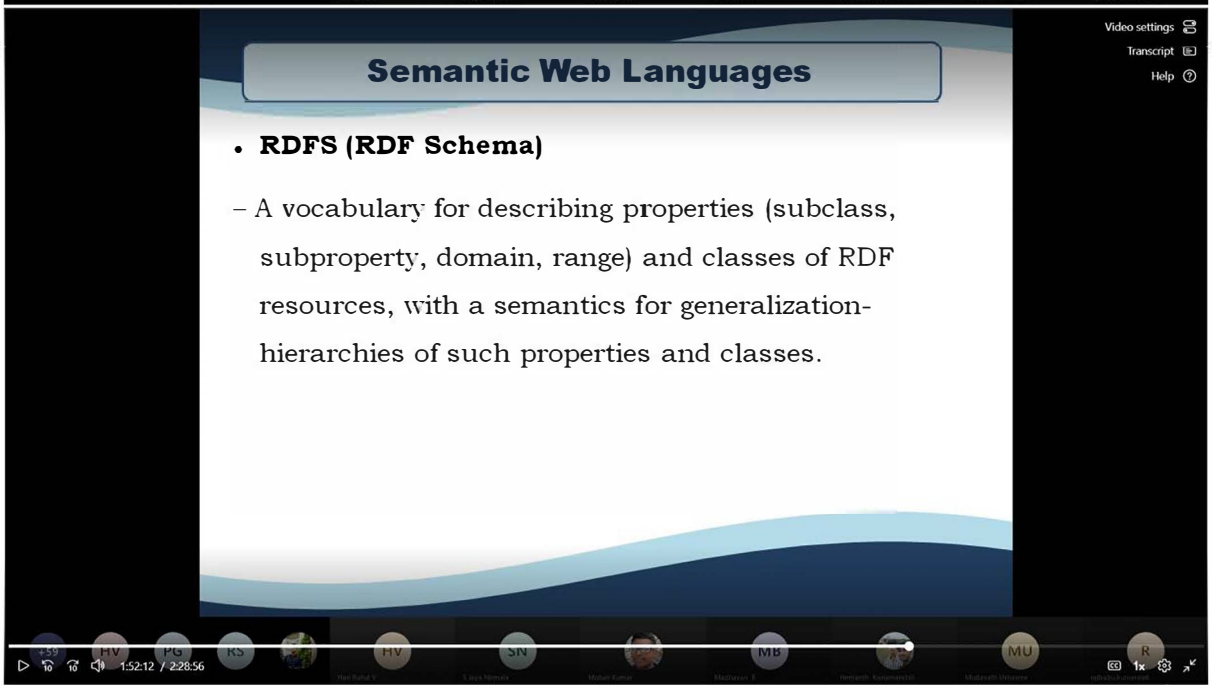
- Semantic networks or concept maps reflect cognition.
- Visual models of concepts of some specific domain connected by some type of relationship (link/arc).
- Can be easily extended and are easy to learn.

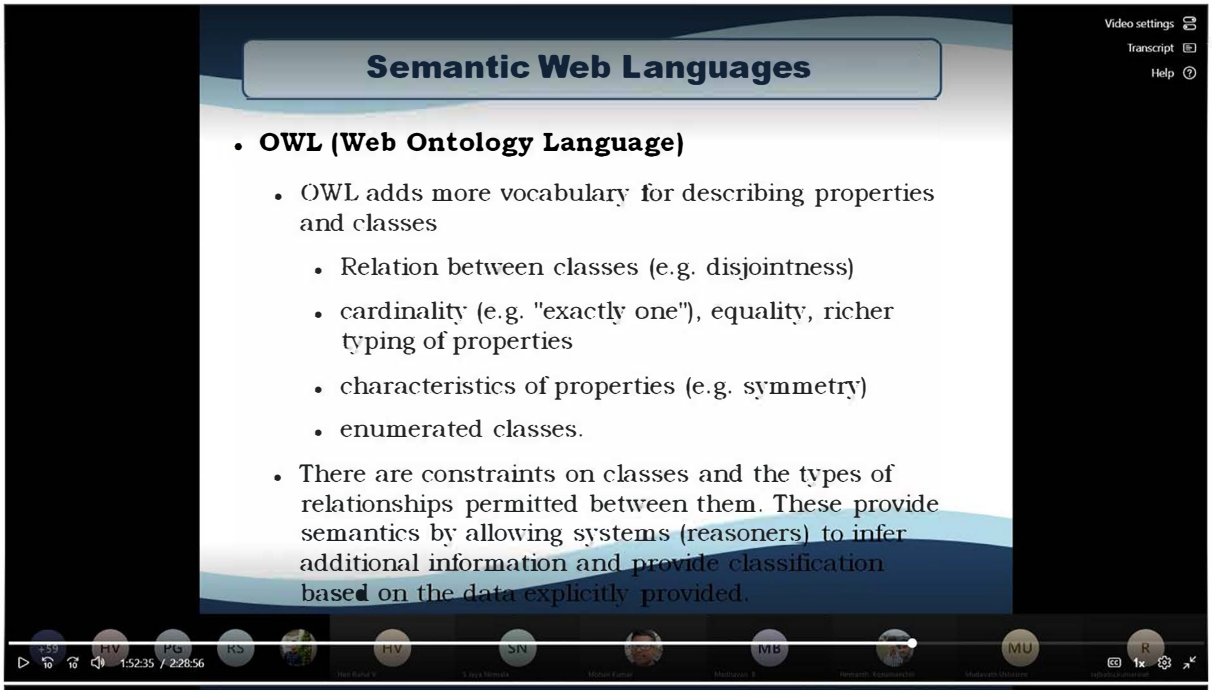


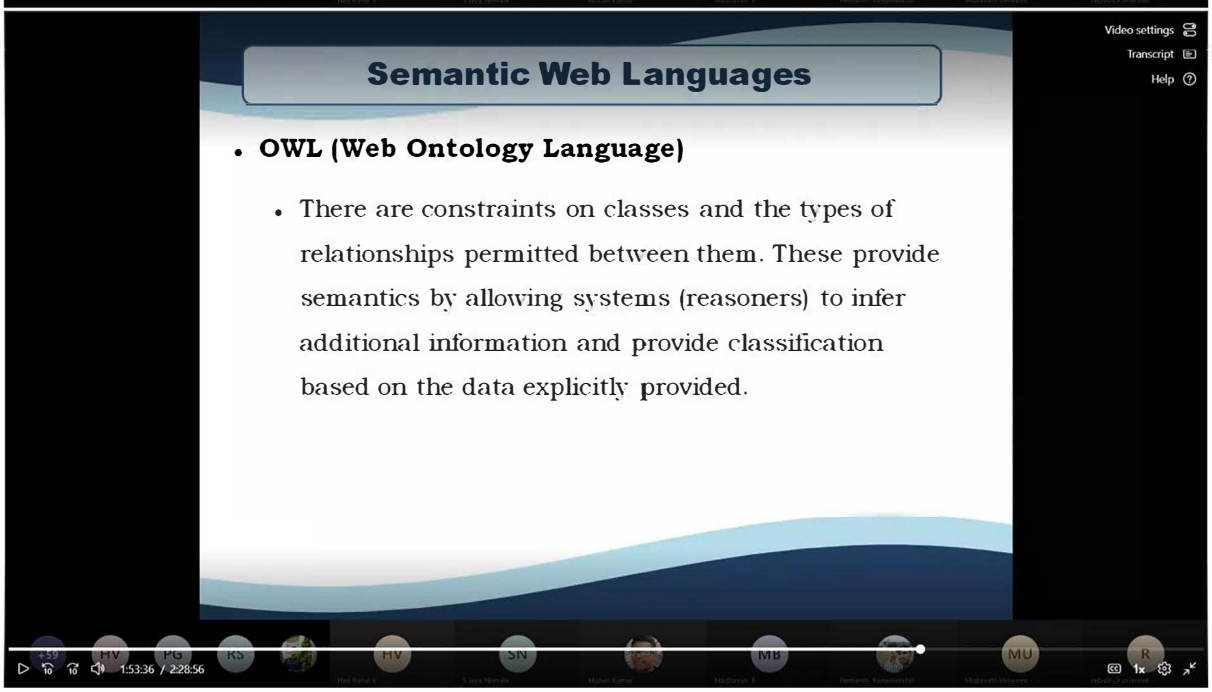
Semantic Network

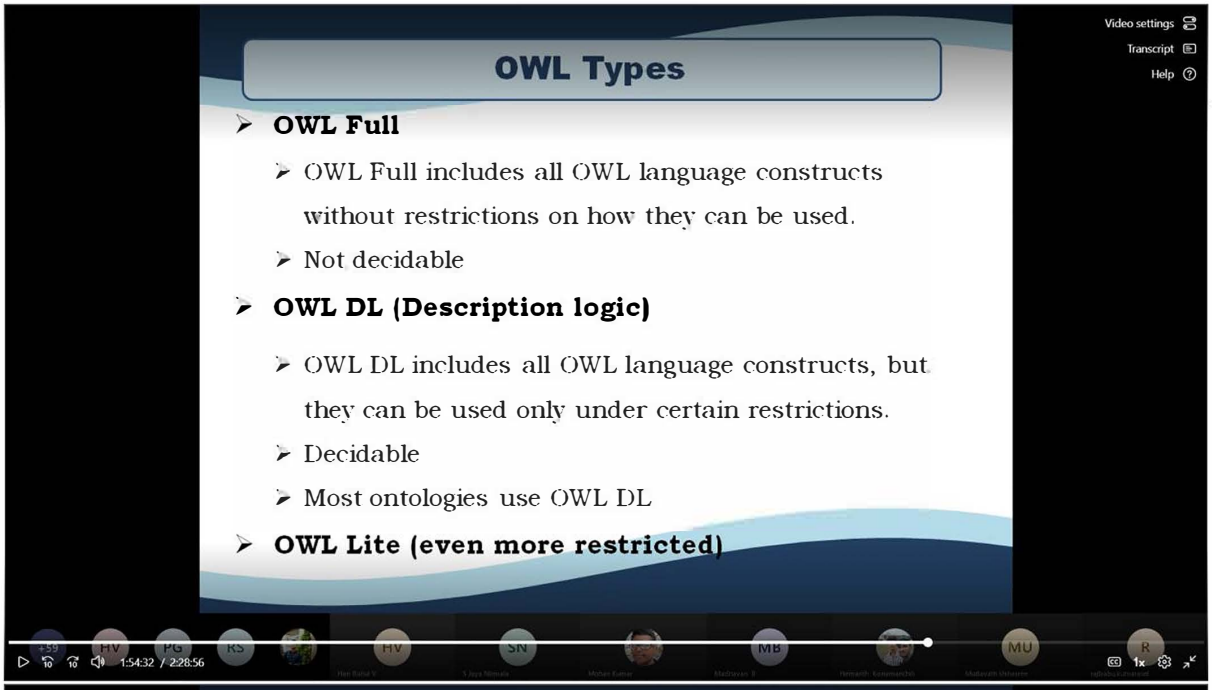


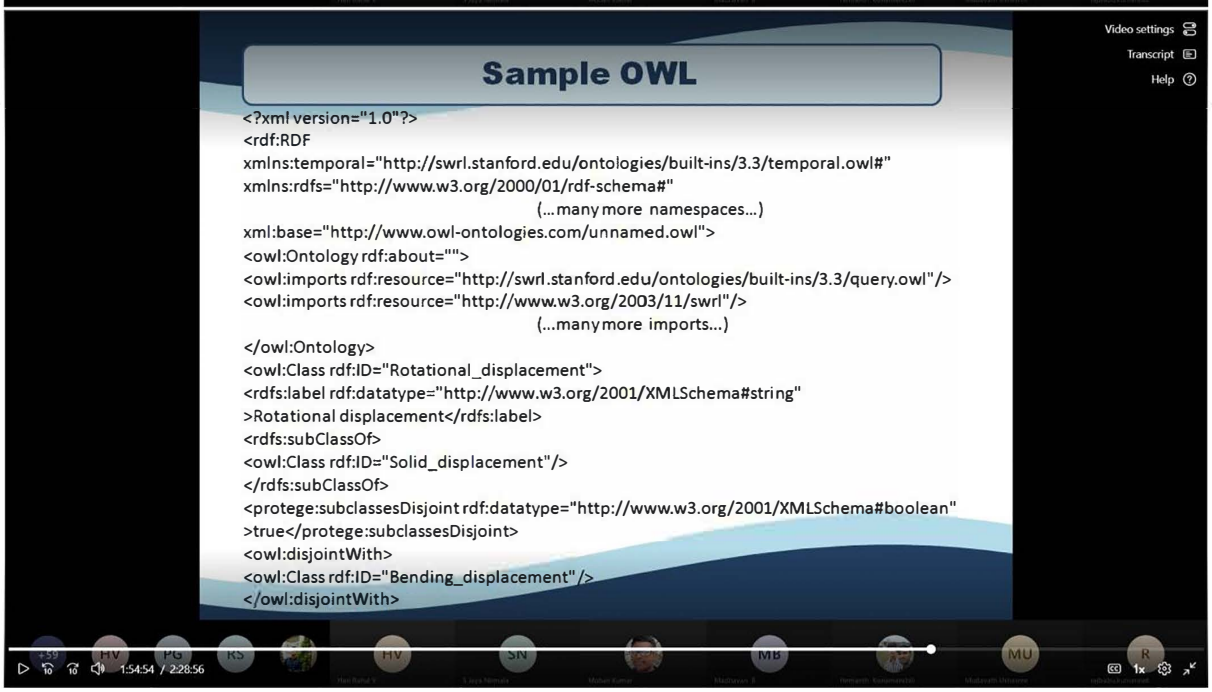

Video settings
Transcript
Help
Semantic Web Languages
• RDF (Resource Description Framework)
– Triples: <subject> <property> <object>
– RDF is a data model for objects ("resources") and relations between them. These data models can be represented in an XML syntax.
1:52:02 / 2:28:56

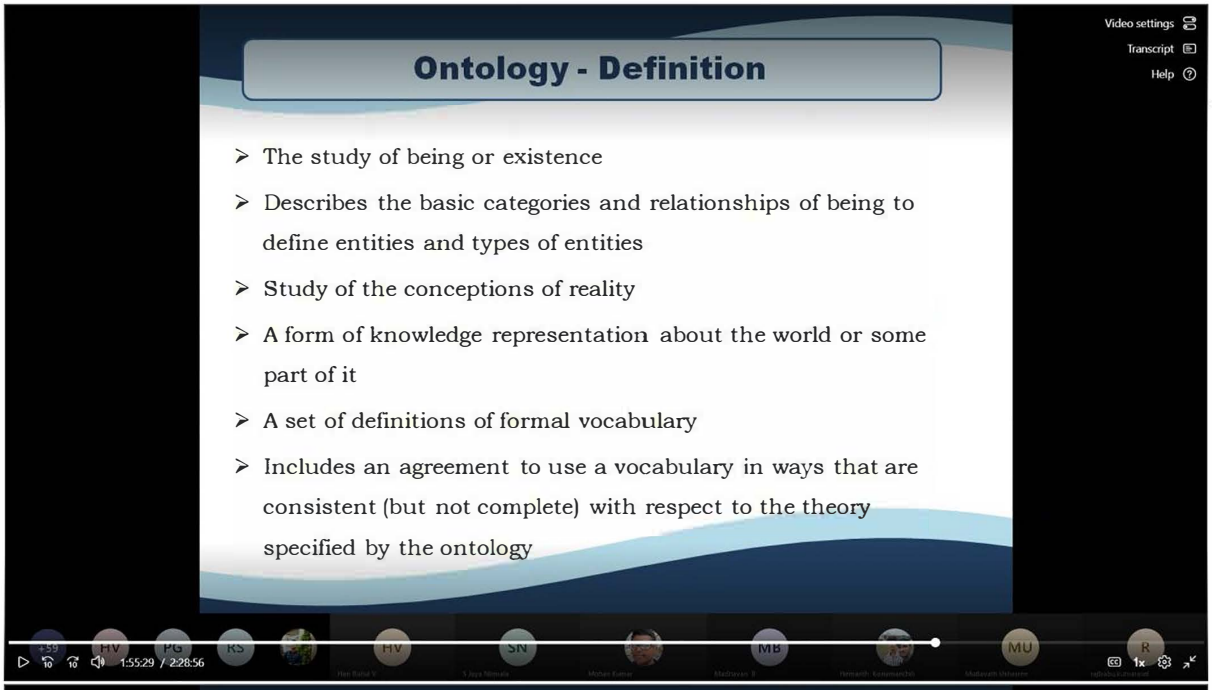

Video settings
Transcript
Help
Semantic Web Languages
• RDFS (RDF Schema)
– A vocabulary for describing properties (subclass, subproperty, domain, range) and classes of RDF resources, with a semantics for generalization-hierarchies of such properties and classes.
1:52:12 / 2:28:56

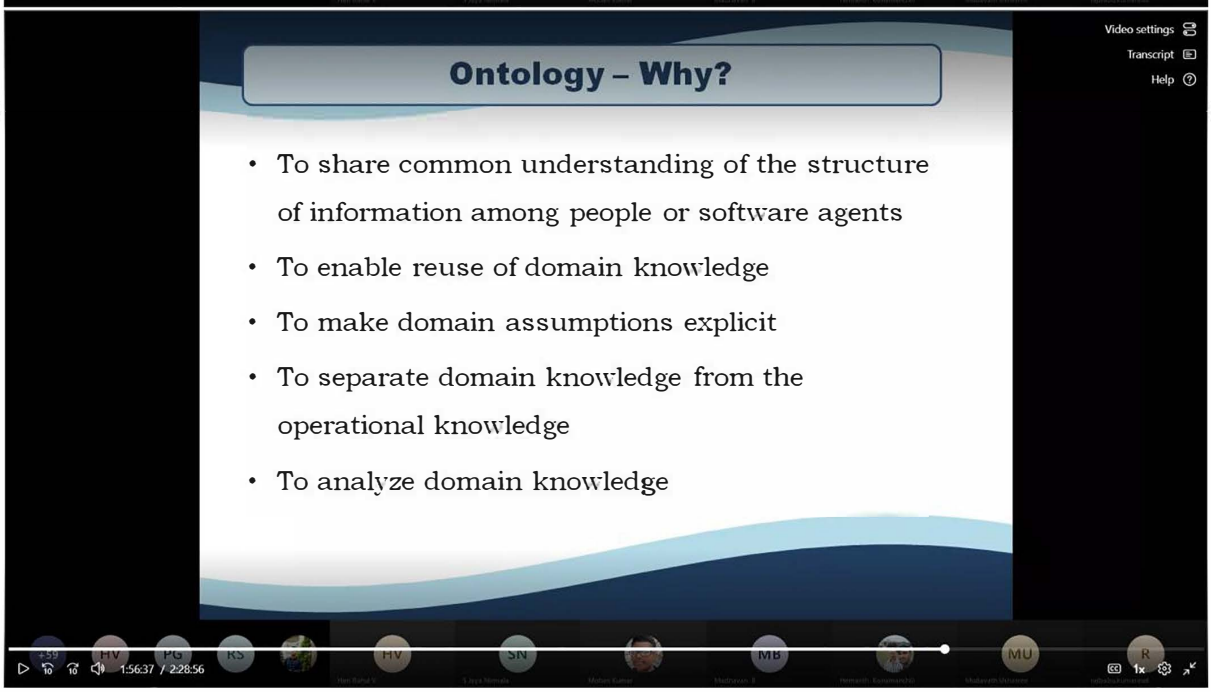

Video settings
Transcript
Help
Semantic Web Languages
• OWL (Web Ontology Language)
• OWL adds more vocabulary for describing properties and classes
• Relation between classes (e.g. disjointness)
• cardinality (e.g. "exactly one"), equality, richer typing of properties
• characteristics of properties (e.g. symmetry)
• enumerated classes.
• There are constraints on classes and the types of relationships permitted between them. These provide semantics by allowing systems (reasoners) to infer additional information and provide classification based on the data explicitly provided.
1:52:35 / 2:28:56

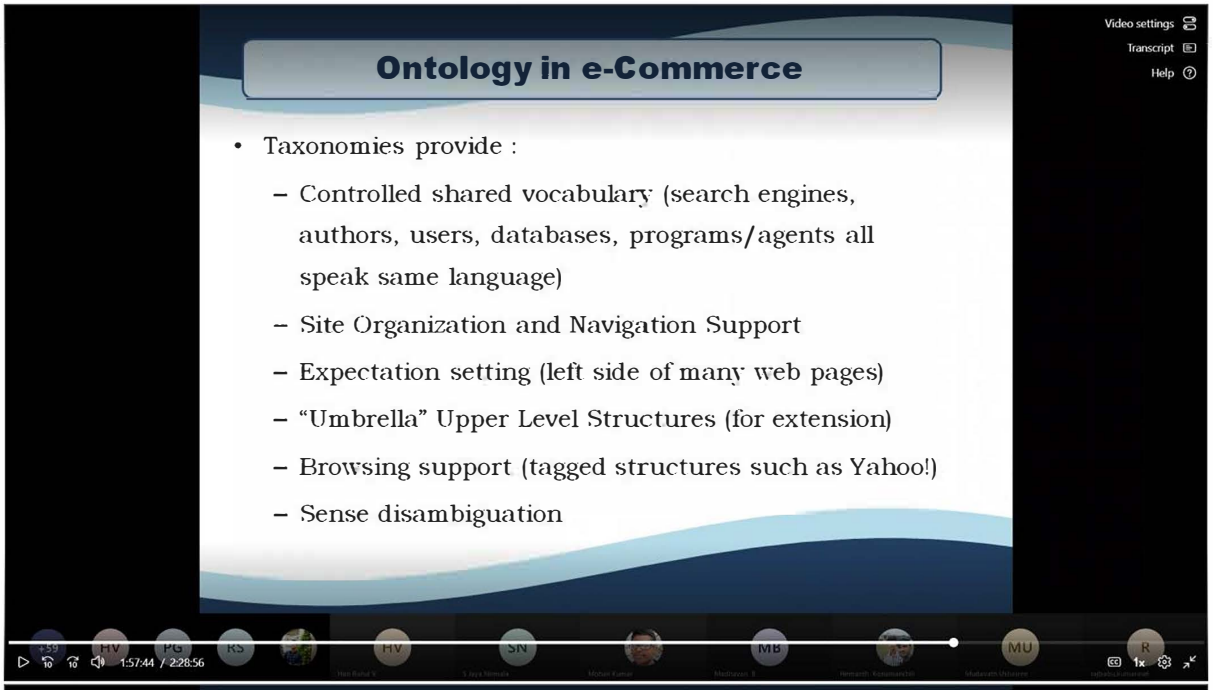

Video settings
Transcript
Help
Semantic Web Languages
• OWL (Web Ontology Language)
• There are constraints on classes and the types of relationships permitted between them. These provide semantics by allowing systems (reasoners) to infer additional information and provide classification based on the data explicitly provided.
1:53:36 / 2:28:56

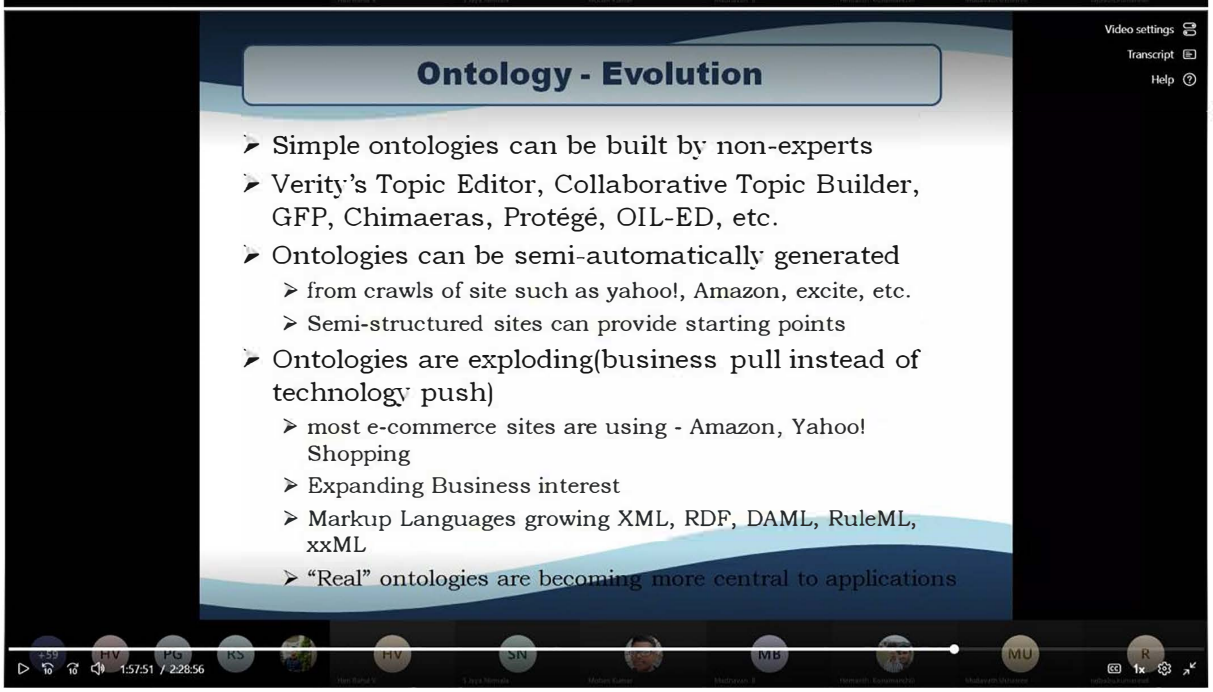

Video settings
Transcript
Help
OWL Types
OWL Full
OWL Full includes all OWL language constructs without restrictions on how they can be used.
Not decidable
OWL DL (Description logic)
OWL DL includes all OWL language constructs, but they can be used only under certain restrictions.
Decidable
Most ontologies use OWL DL
OWL Lite (even more restricted)
1:54:32 / 2:28:56

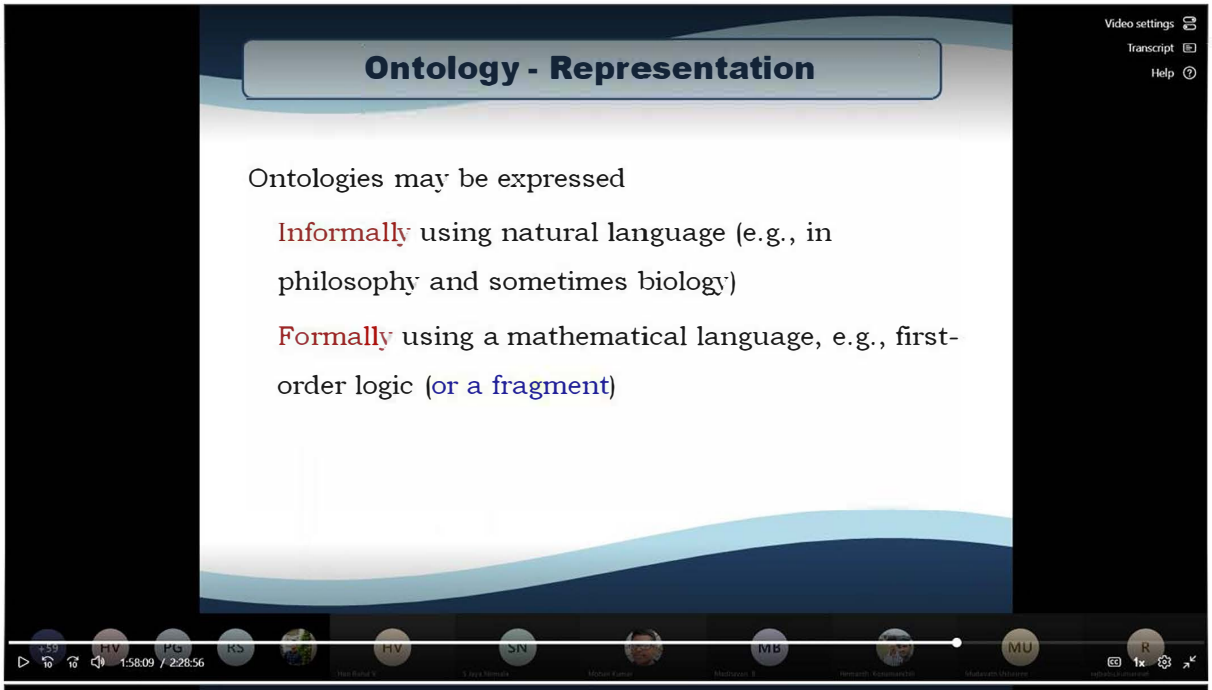

Video settings
Transcript
Help
Sample OWL
<?xml version="1.0"?>
<rdf:RDF
xmlns:temporal="http://swrl.stanford.edu/ontologies/built-ins/3.3/temporal.owl#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
(...many more namespaces...)
xml:base="http://www.owl-ontologies.com/unnamed.owl">
<owl:Ontology rdf:about="">
<owl:imports rdf:resource="http://swrl.stanford.edu/ontologies/built-ins/3.3/query.owl"/>
<owl:imports rdf:resource="http://www.w3.org/2003/11/swrl"/>
(...many more imports...)
</owl:Ontology>
<owl:Class rdf:ID="Rotational_displacement">
<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>Rotational displacement</rdfs:label>
<rdfs:subClassOf>
<owl:Class rdf:ID="Solid_displacement"/>
</rdfs:subClassOf>
<protege:subclassesDisjoint rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
>true</protege:subclassesDisjoint>
<owl:disjointWith>
<owl:Class rdf:ID="Bending_displacement"/>
</owl:disjointWith>
1:54:54 / 2:28:56

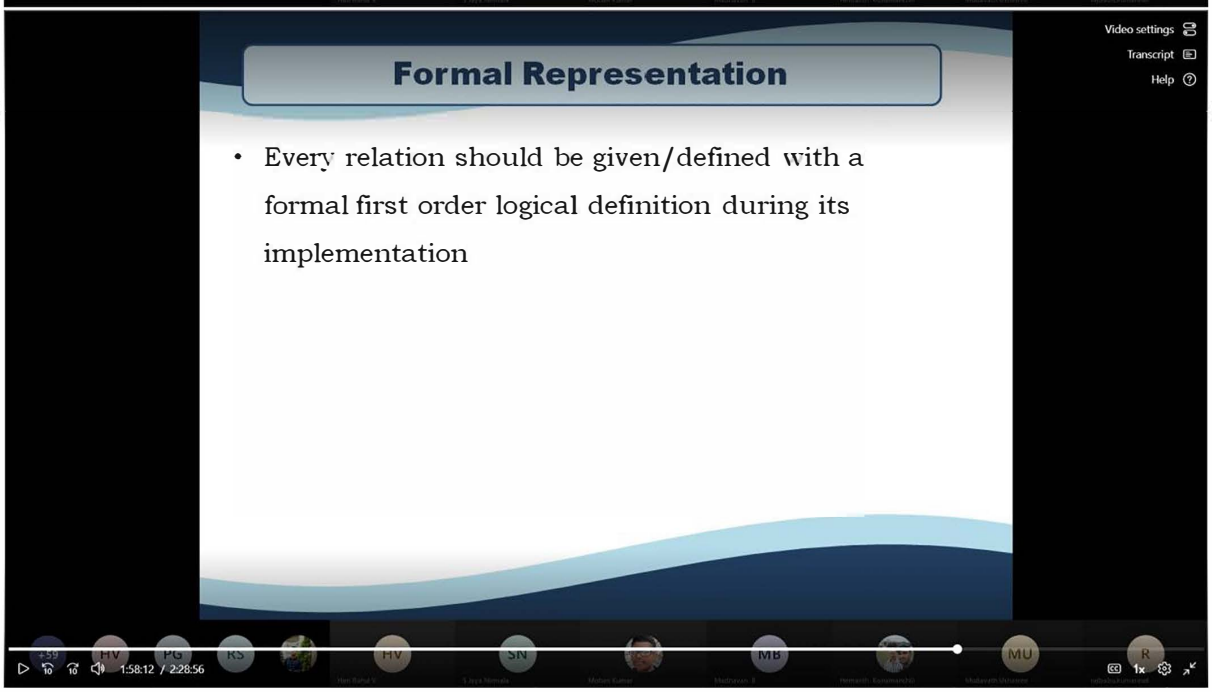

Video settings
Transcript
Help
Ontology - Definition
The study of being or existence
Describes the basic categories and relationships of being to define entities and types of entities
Study of the conceptions of reality
A form of knowledge representation about the world or some part of it
A set of definitions of formal vocabulary
Includes an agreement to use a vocabulary in ways that are consistent (but not complete) with respect to the theory specified by the ontology
1:55:29 / 2:28:56

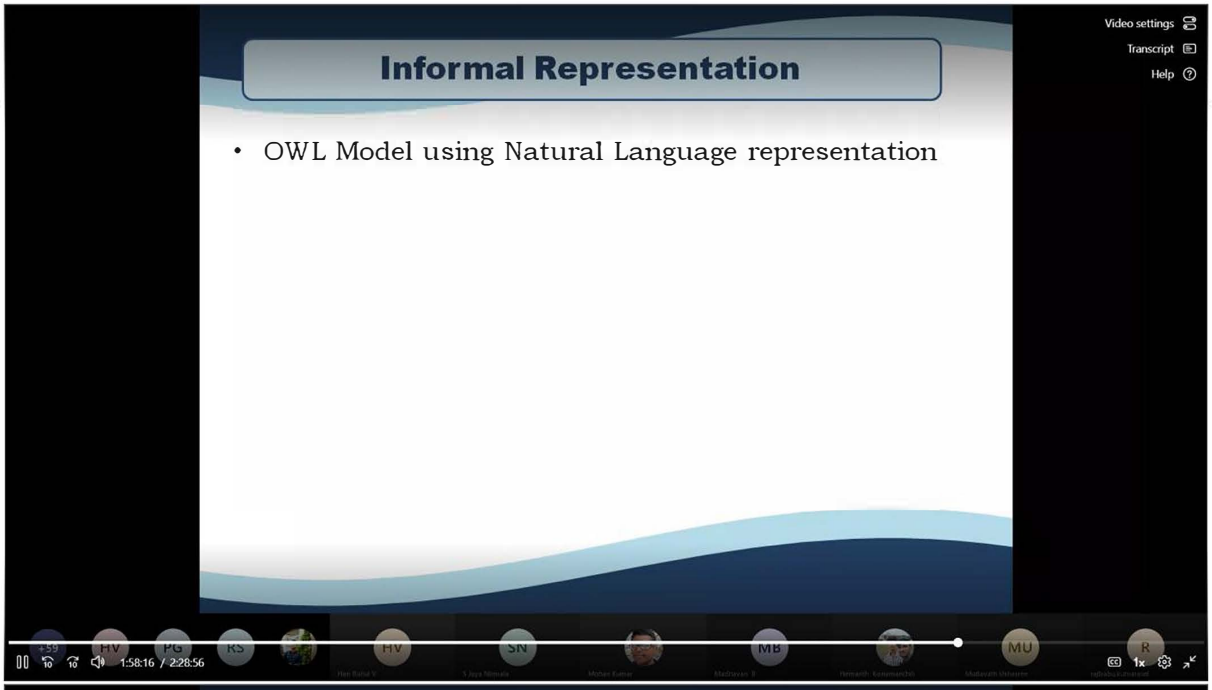

Video settings
Transcript
Help
Ontology – Why?
To share common understanding of the structure of information among people or software agents
To enable reuse of domain knowledge
To make domain assumptions explicit
To separate domain knowledge from the operational knowledge
To analyze domain knowledge
1:56:37 / 2:28:56

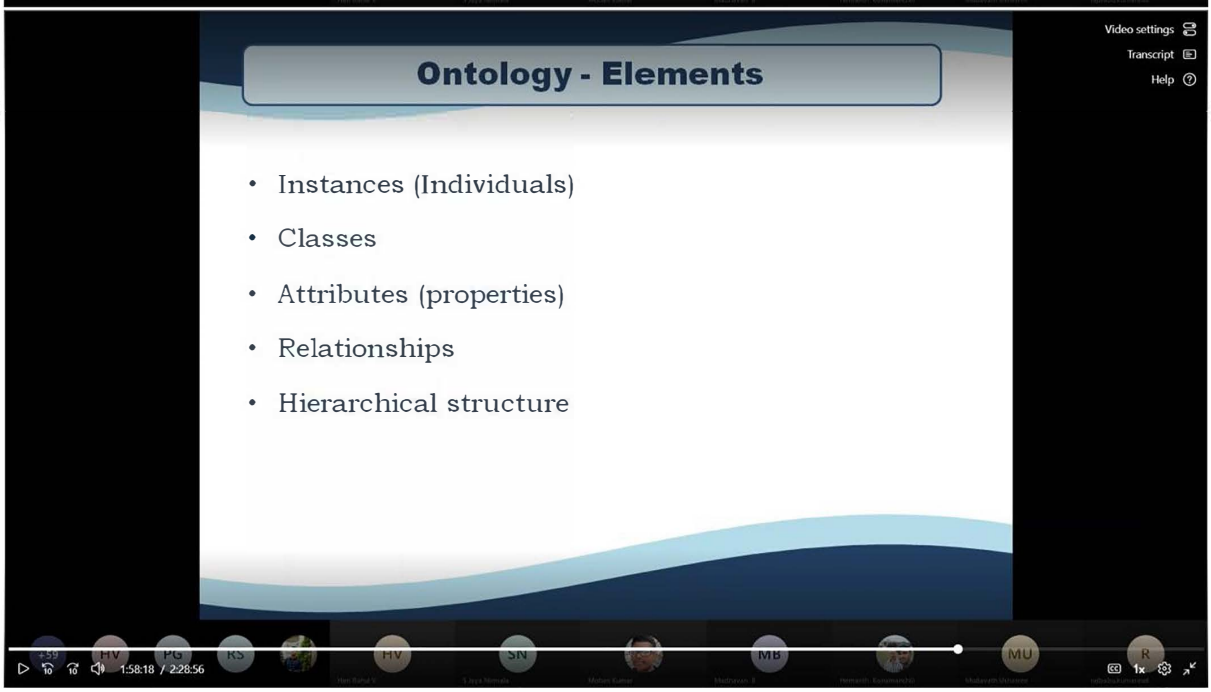

Video settings
Transcript
Help
Ontology in e-Commerce
• Taxonomies provide :
– Controlled shared vocabulary (search engines, authors, users, databases, programs/agents all speak same language)
– Site Organization and Navigation Support
– Expectation setting (left side of many web pages)
– "Umbrella" Upper Level Structures (for extension)
– Browsing support (tagged structures such as Yahoo!)
– Sense disambiguation
1:57:44 / 2:28:56

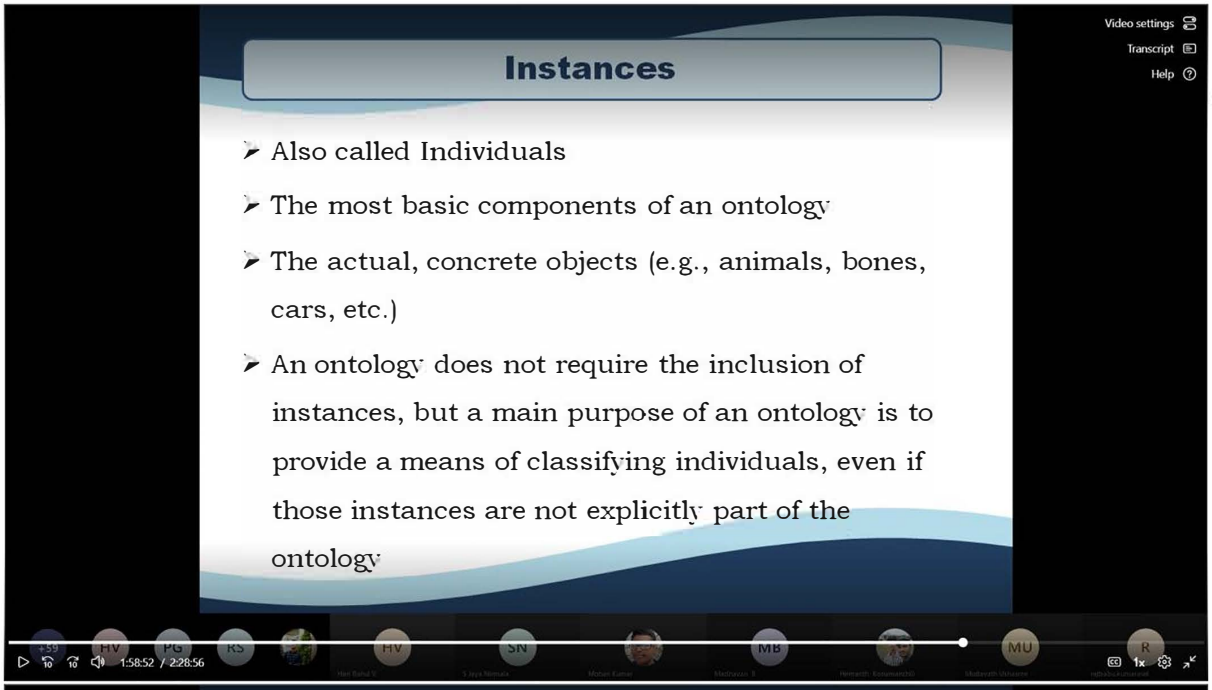

Video settings
Transcript
Help
Ontology - Evolution
➢ Simple ontologies can be built by non-experts
➢ Verity's Topic Editor, Collaborative Topic Builder, GFP, Chimaeras, Protégé, OIL-ED, etc.
➢ Ontologies can be semi-automatically generated
➢ from crawls of site such as yahoo!, Amazon, excite, etc.
➢ Semi-structured sites can provide starting points
➢ Ontologies are exploding(business pull instead of technology push)
➢ most e-commerce sites are using - Amazon, Yahoo! Shopping
➢ Expanding Business interest
➢ Markup Languages growing XML, RDF, DAML, RuleML, xxML
➢ "Real" ontologies are becoming more central to applications
1:57:51 / 2:28:56

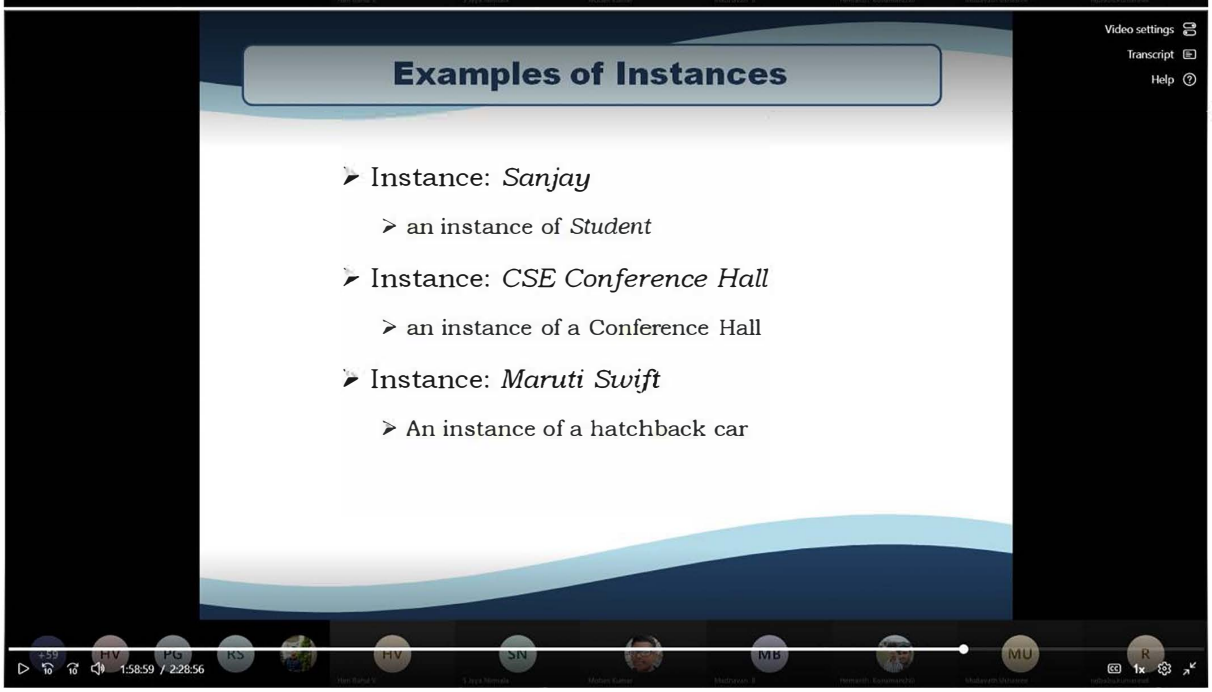

Video settings
Transcript
Help
Ontology - Representation
Ontologies may be expressed
Informally using natural language (e.g., in philosophy and sometimes biology)
Formally using a mathematical language, e.g., first-order logic (or a fragment)
1:58:09 / 2:28:56

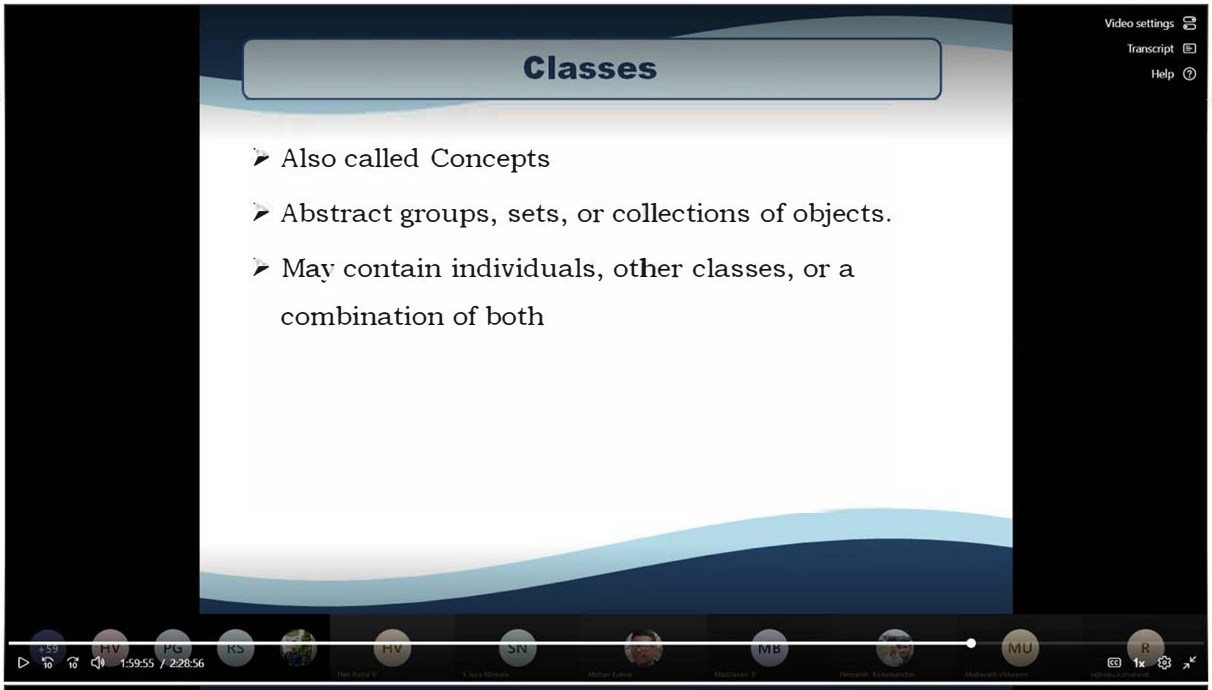

Video settings
Transcript
Help
Formal Representation
Every relation should be given/defined with a formal first order logical definition during its implementation
1:58:12 / 2:28:56

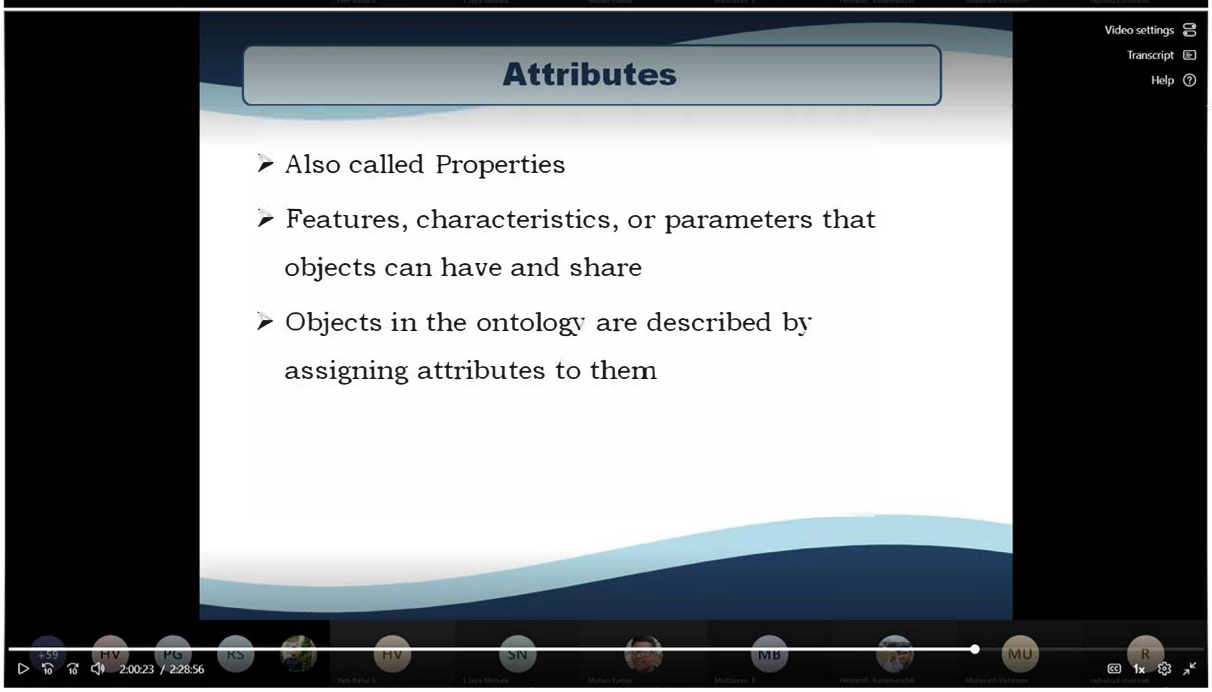

Video settings
Transcript
Help
Informal Representation
• OWL Model using Natural Language representation
1:58:16 / 2:28:56

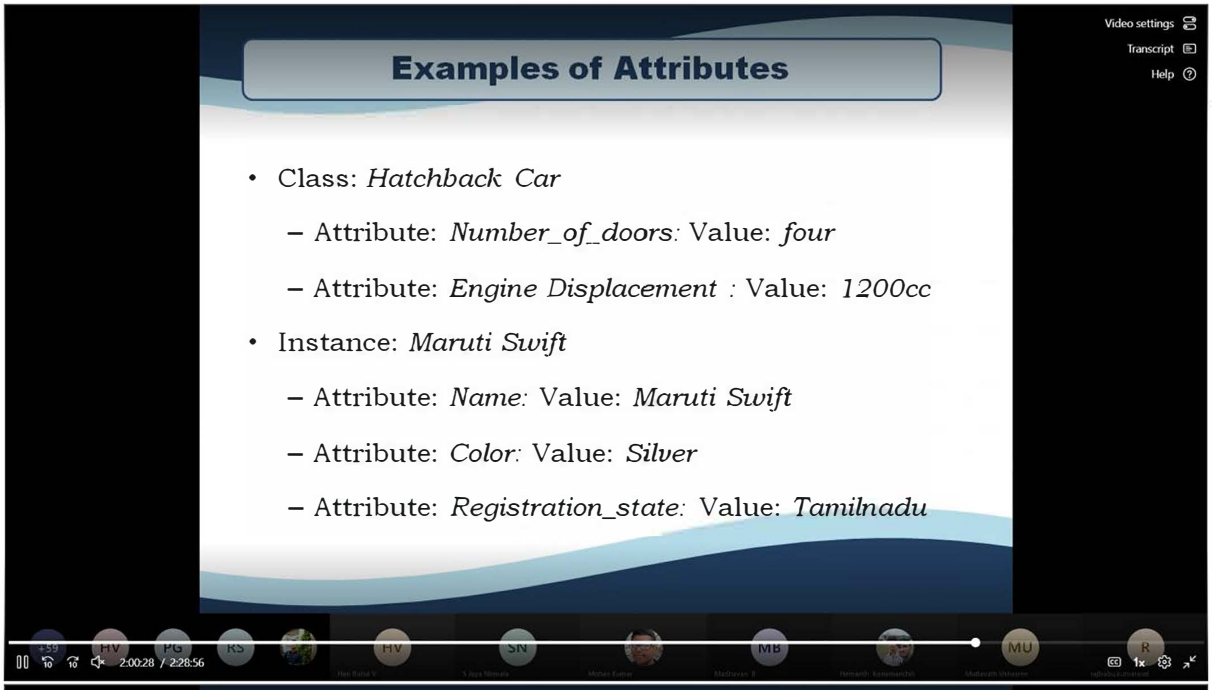

Video settings
Transcript
Help
Ontology - Elements
• Instances (Individuals)
• Classes
• Attributes (properties)
• Relationships
• Hierarchical structure
1:58:18 / 2:28:56

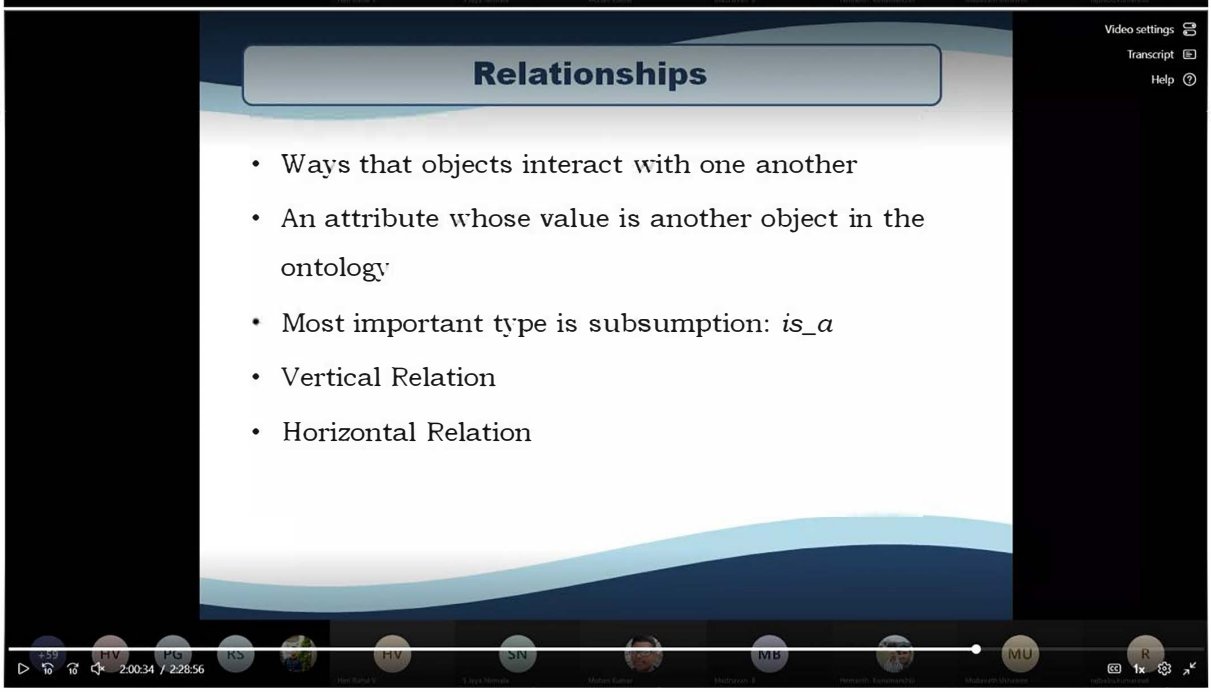

Video settings
Transcript
Help
Instances
Also called Individuals
The most basic components of an ontology
The actual, concrete objects (e.g., animals, bones, cars, etc.)
An ontology does not require the inclusion of instances, but a main purpose of an ontology is to provide a means of classifying individuals, even if those instances are not explicitly part of the ontology
1:58:52 / 2:28:56

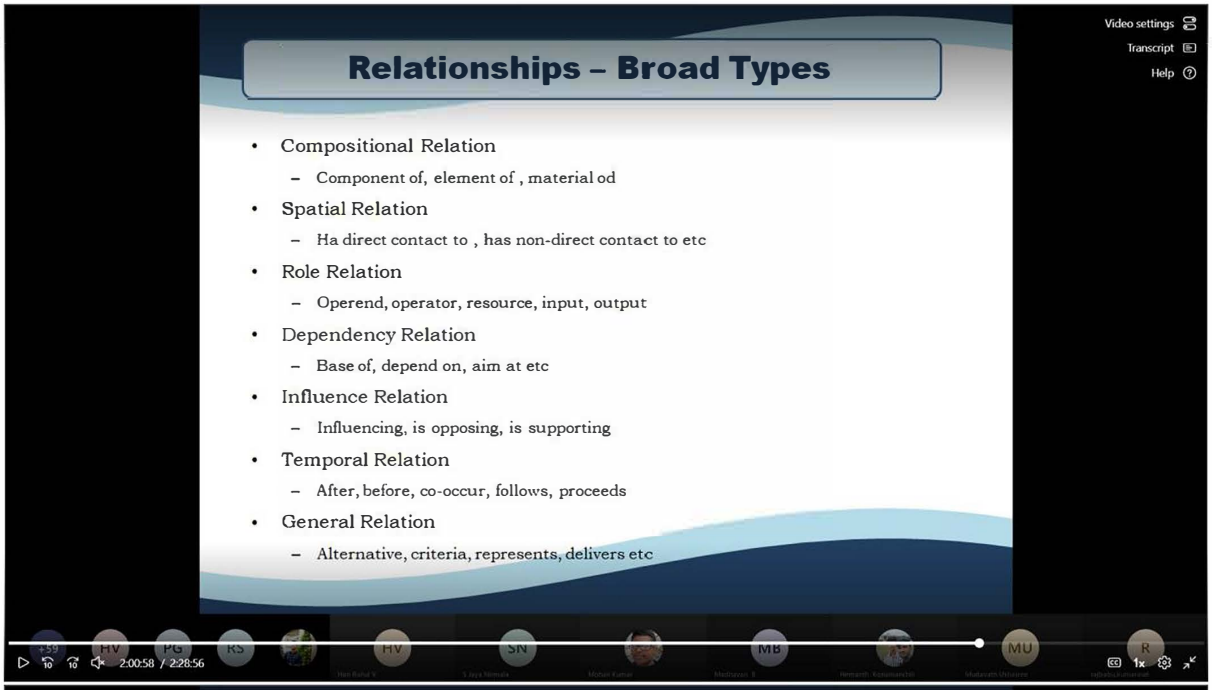

Video settings
Transcript
Help
Examples of Instances
Instance: Sanjay
an instance of Student
Instance: CSE Conference Hall
an instance of a Conference Hall
Instance: Maruti Swift
An instance of a hatchback car
1:58:59 / 2:28:56

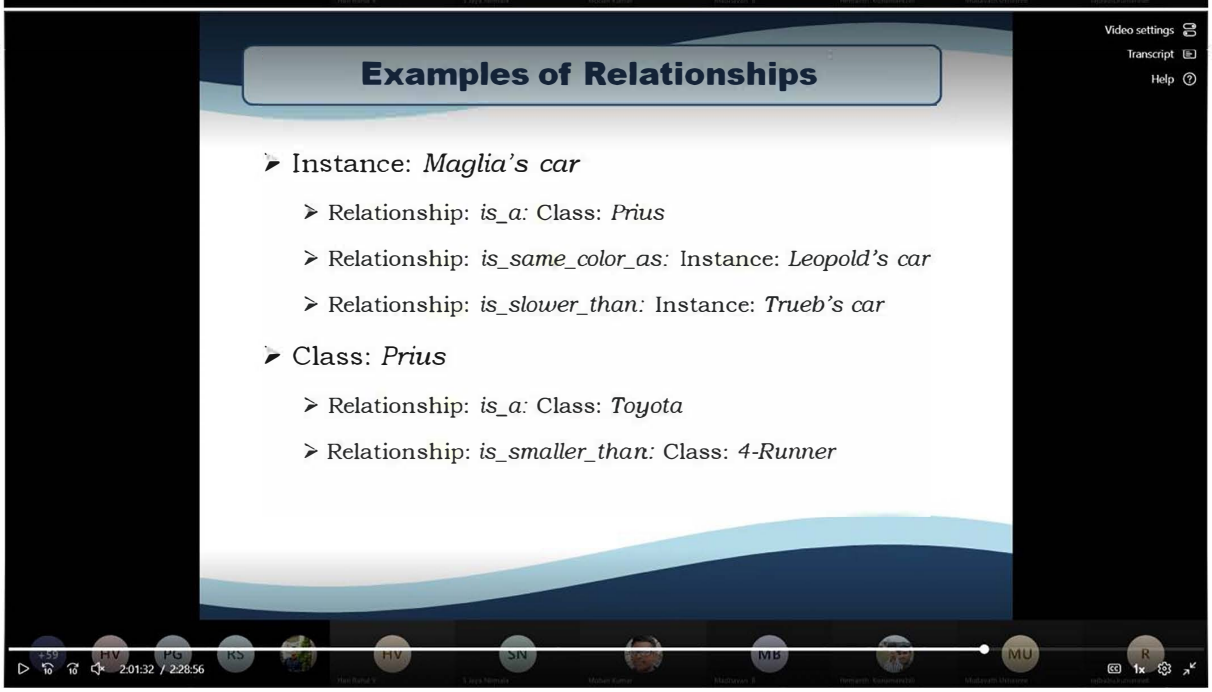

Video settings
Transcript
Help
Classes
Also called Concepts
Abstract groups, sets, or collections of objects.
May contain individuals, other classes, or a combination of both
1:59:55 / 2:28:56

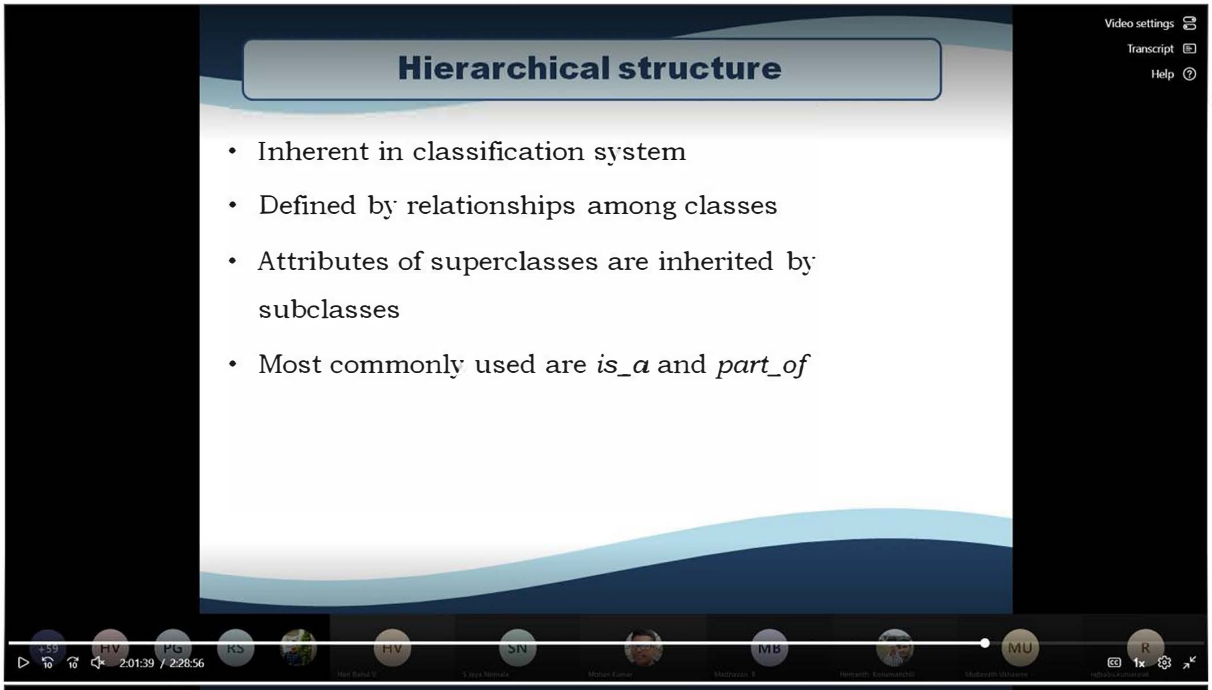

Video settings
Transcript
Help
Attributes
Also called Properties
Features, characteristics, or parameters that objects can have and share
Objects in the ontology are described by assigning attributes to them
2:00:23 / 2:28:56

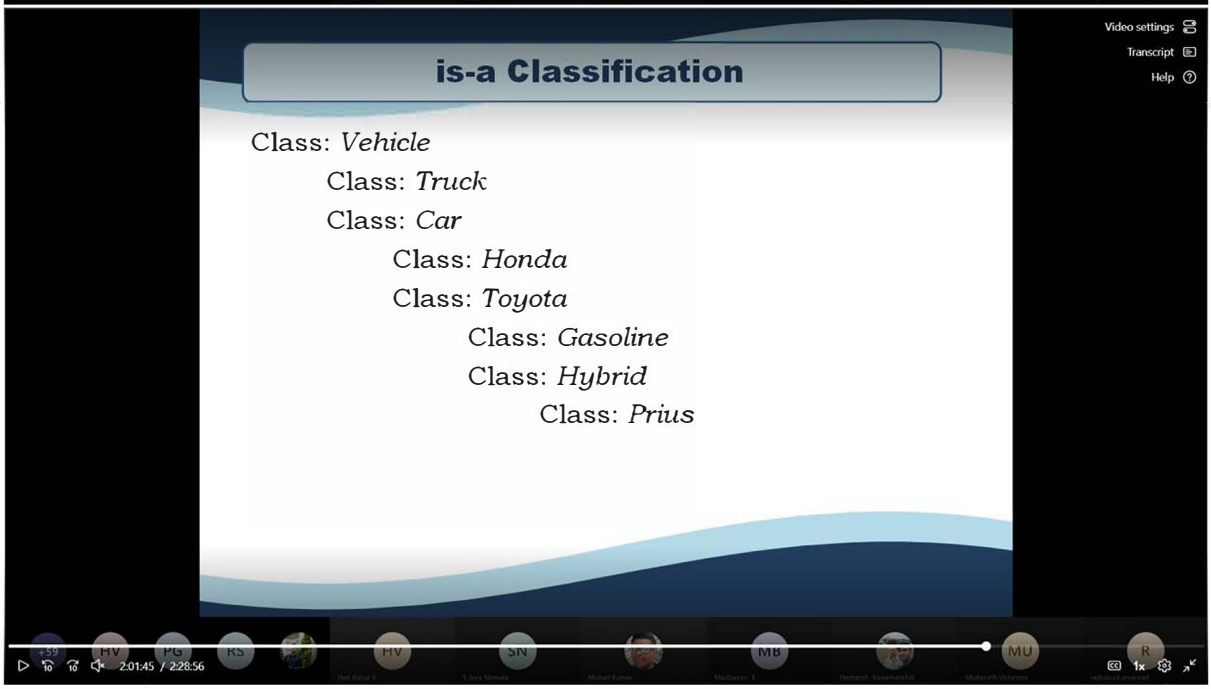

Video settings
Transcript
Help
Examples of Attributes
• Class: Hatchback Car
– Attribute: Number_of_doors: Value: four
– Attribute: Engine Displacement : Value: 1200cc
• Instance: Maruti Swift
– Attribute: Name: Value: Maruti Swift
– Attribute: Color: Value: Silver
– Attribute: Registration_state: Value: Tamilnadu
2:00:28 / 2:28:56

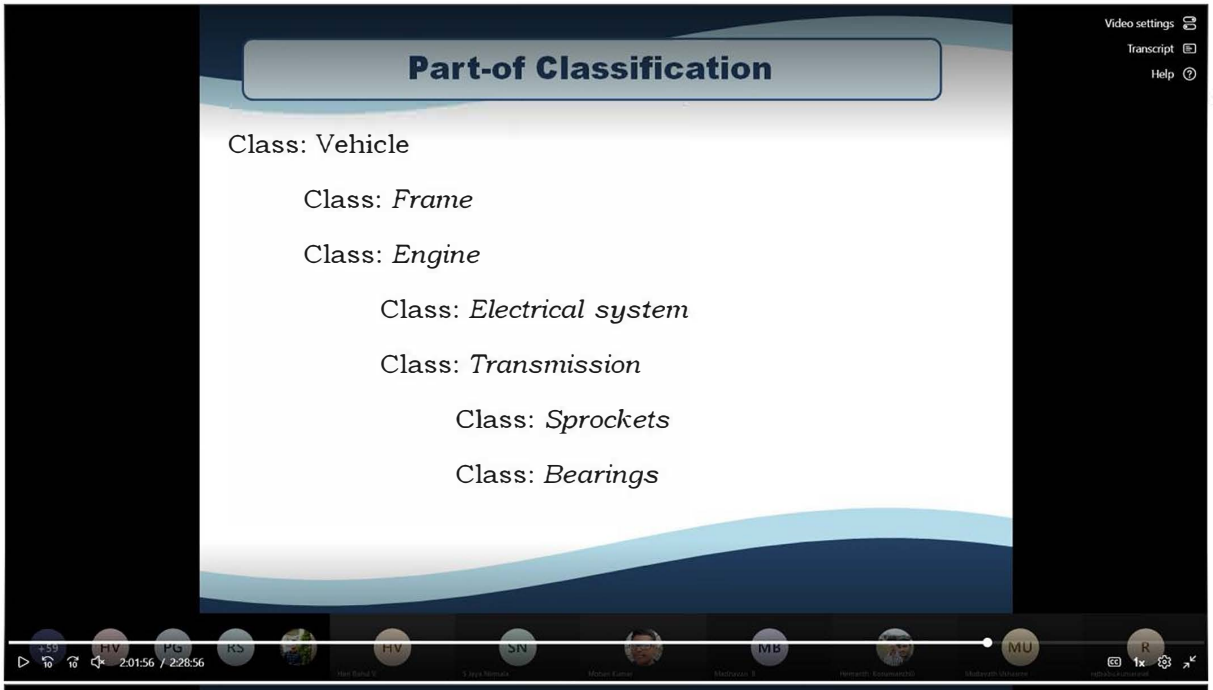

Video settings
Transcript
Help
Relationships
• Ways that objects interact with one another
• An attribute whose value is another object in the ontology
• Most important type is subsumption: is_a
• Vertical Relation
• Horizontal Relation
2:00:34 / 2:28:56

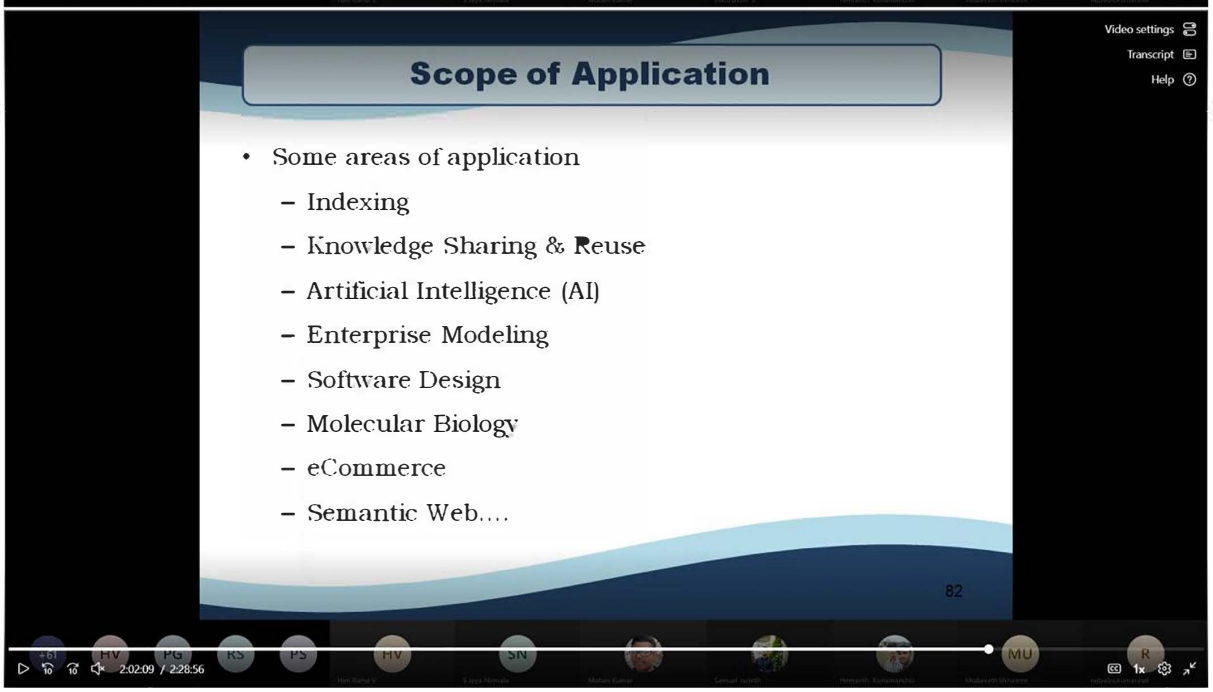

Video settings
Transcript
Help
Relationships – Broad Types
• Compositional Relation
– Component of, element of , material od
• Spatial Relation
– Ha direct contact to , has non-direct contact to etc
• Role Relation
– Operend, operator, resource, input, output
• Dependency Relation
– Base of, depend on, aim at etc
• Influence Relation
– Influencing, is opposing, is supporting
• Temporal Relation
– After, before, co-occur, follows, proceeds
• General Relation
– Alternative, criteria, represents, delivers etc
2:00:58 / 2:28:56

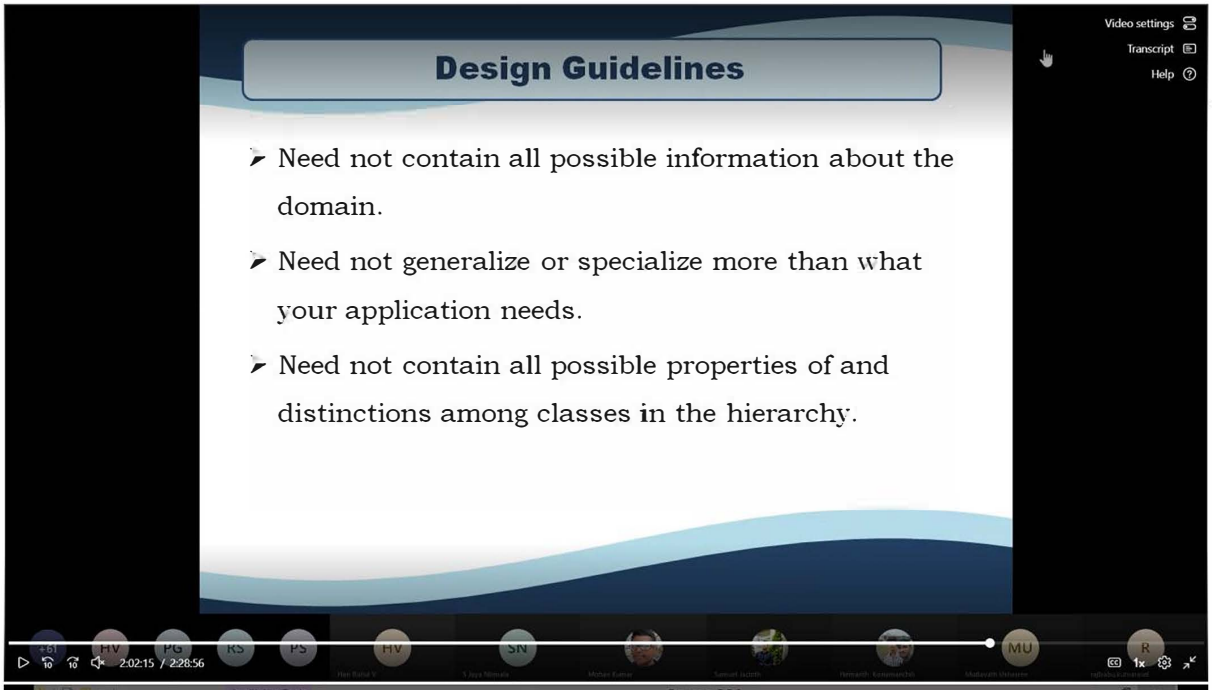

Video settings
Transcript
Help
Examples of Relationships
➢ Instance: Maglia's car
➢ Relationship: is_a: Class: Prius
➢ Relationship: is_same_color_as: Instance: Leopold's car
➢ Relationship: is_slower_than: Instance: Trueb's car
➢ Class: Prius
➢ Relationship: is_a: Class: Toyota
➢ Relationship: is_smaller_than: Class: 4-Runner
2:01:32 / 2:28:56

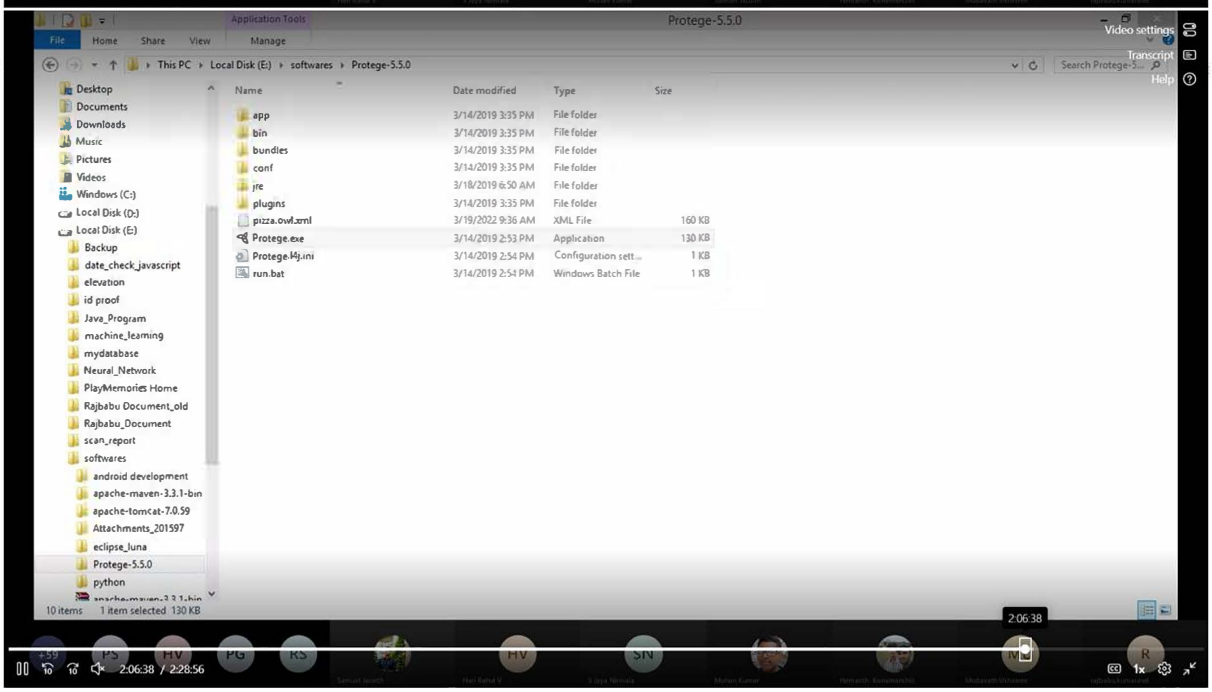

Video settings
Transcript
Help
Hierarchical structure
• Inherent in classification system
• Defined by relationships among classes
• Attributes of superclasses are inherited by subclasses
• Most commonly used are *is_a* and *part_of*
2:01:39 / 2:28:56

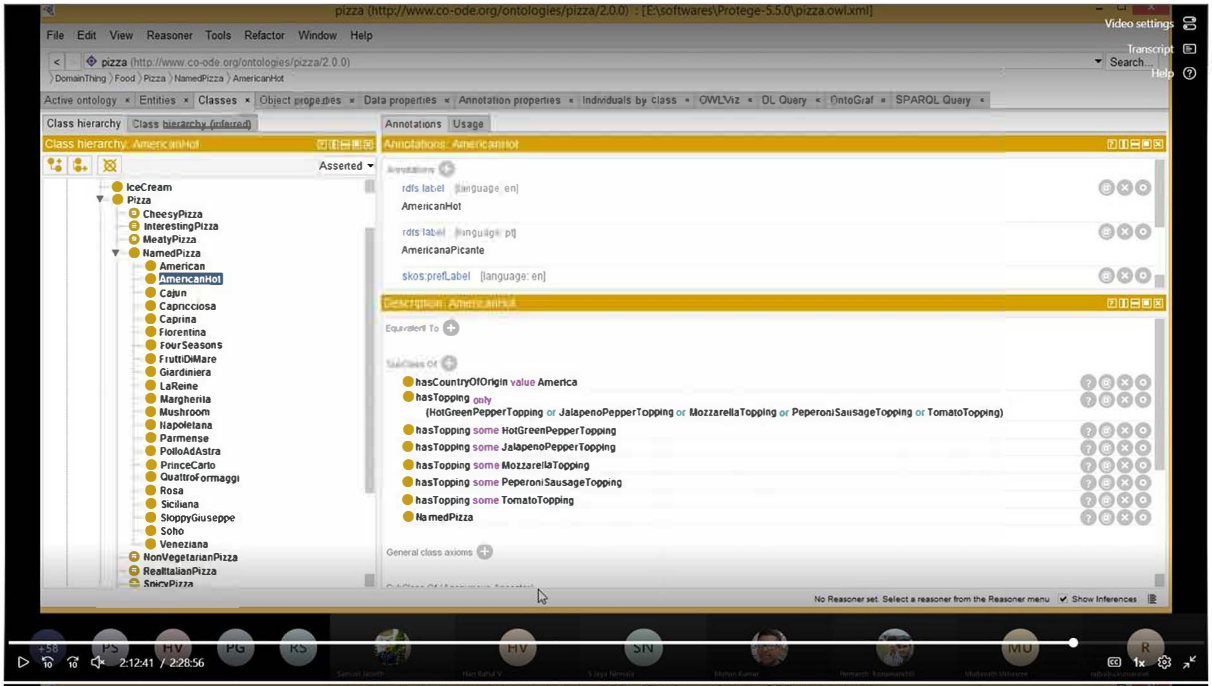

Video settings
Transcript
Help
is-a Classification
Class: *Vehicle*
Class: *Truck*
Class: *Car*
Class: *Honda*
Class: *Toyota*
Class: *Gasoline*
Class: *Hybrid*
Class: *Prius*
2:01:45 / 2:28:56

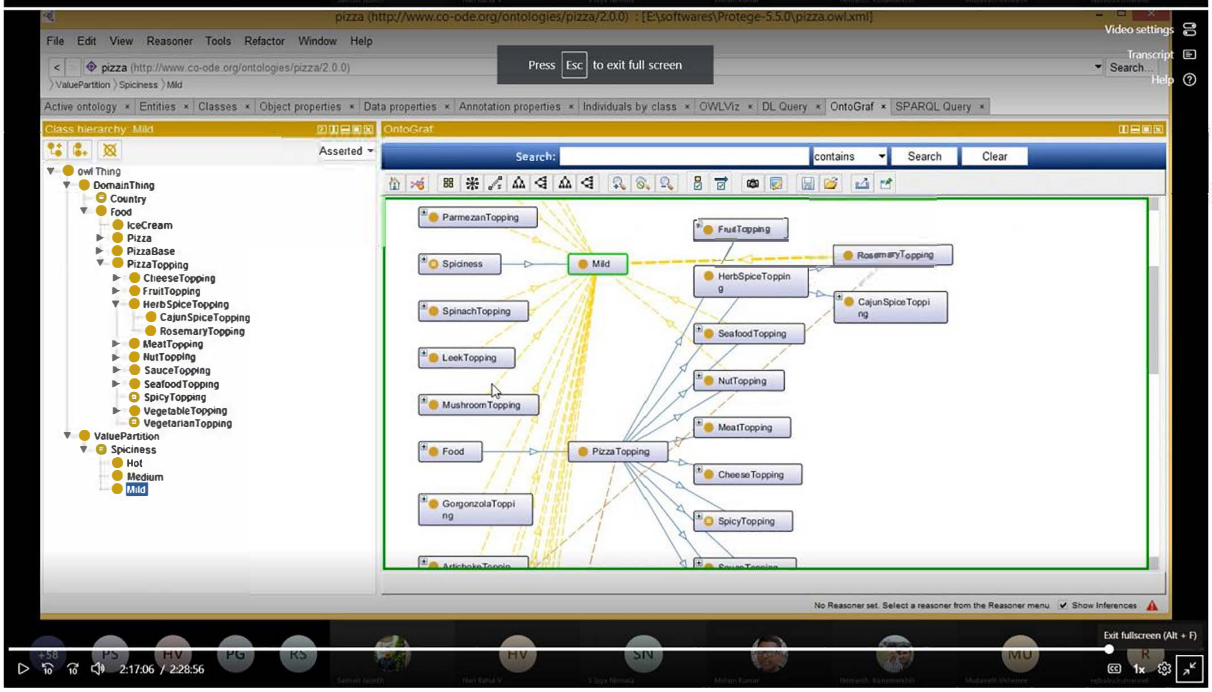

Video settings
Transcript
Help
Part-of Classification
Class: Vehicle
Class: *Frame*
Class: *Engine*
Class: *Electrical system*
Class: *Transmission*
Class: *Sprockets*
Class: *Bearings*
2:01:56 / 2:28:56

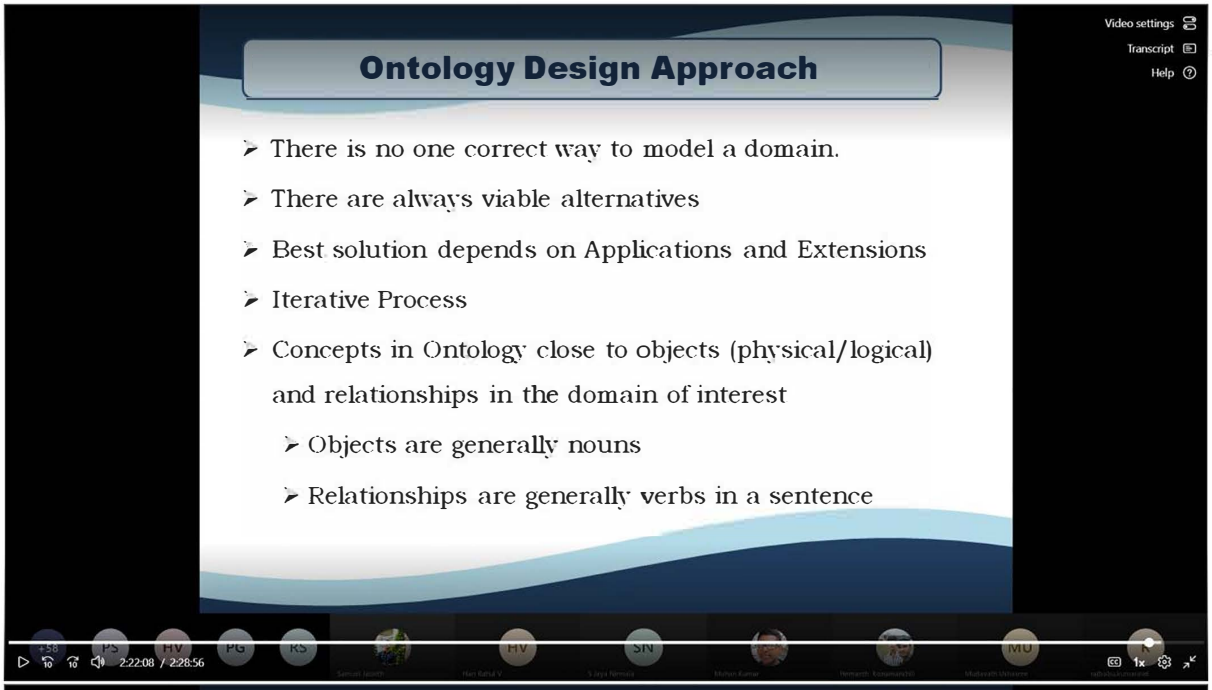

Video settings
Transcript
Help
Scope of Application
• Some areas of application
– Indexing
– Knowledge Sharing & Reuse
– Artificial Intelligence (AI)
– Enterprise Modeling
– Software Design
– Molecular Biology
– eCommerce
– Semantic Web....
82
2:02:09 / 2:28:56

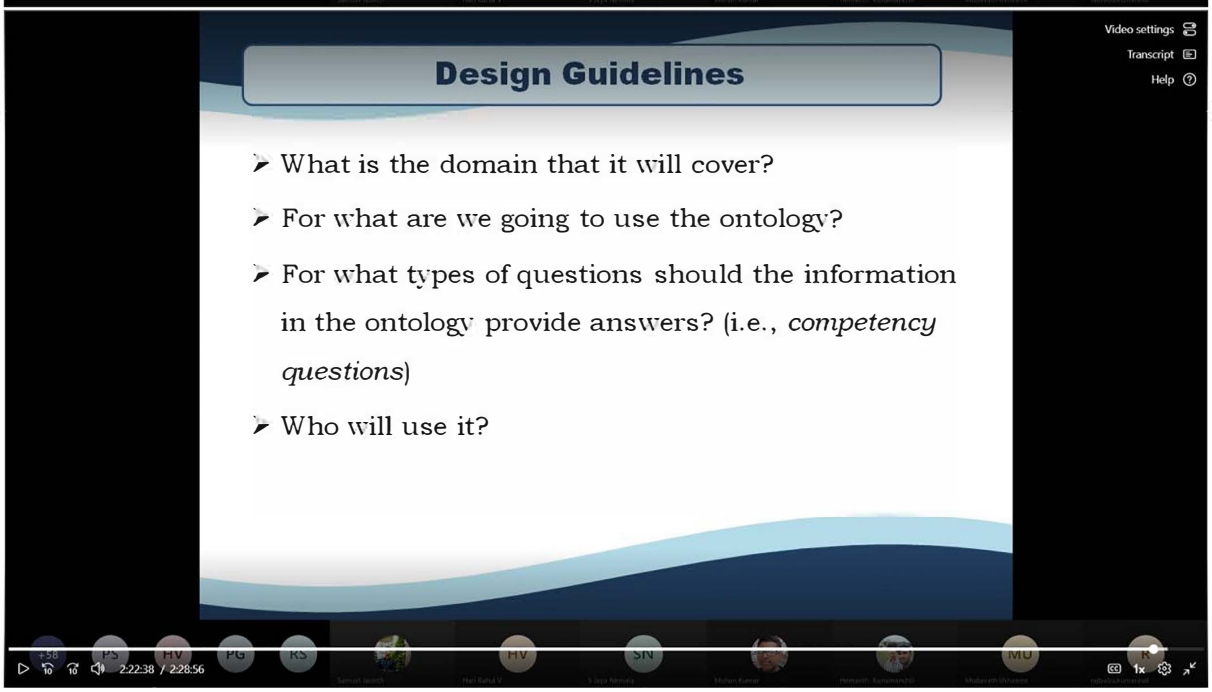

Design Guidelines
Need not contain all possible information about the domain.
Need not generalize or specialize more than what your application needs.
Need not contain all possible properties of and distinctions among classes in the hierarchy.
Video settings
Transcript
Help
2:02:15 / 2:28:56

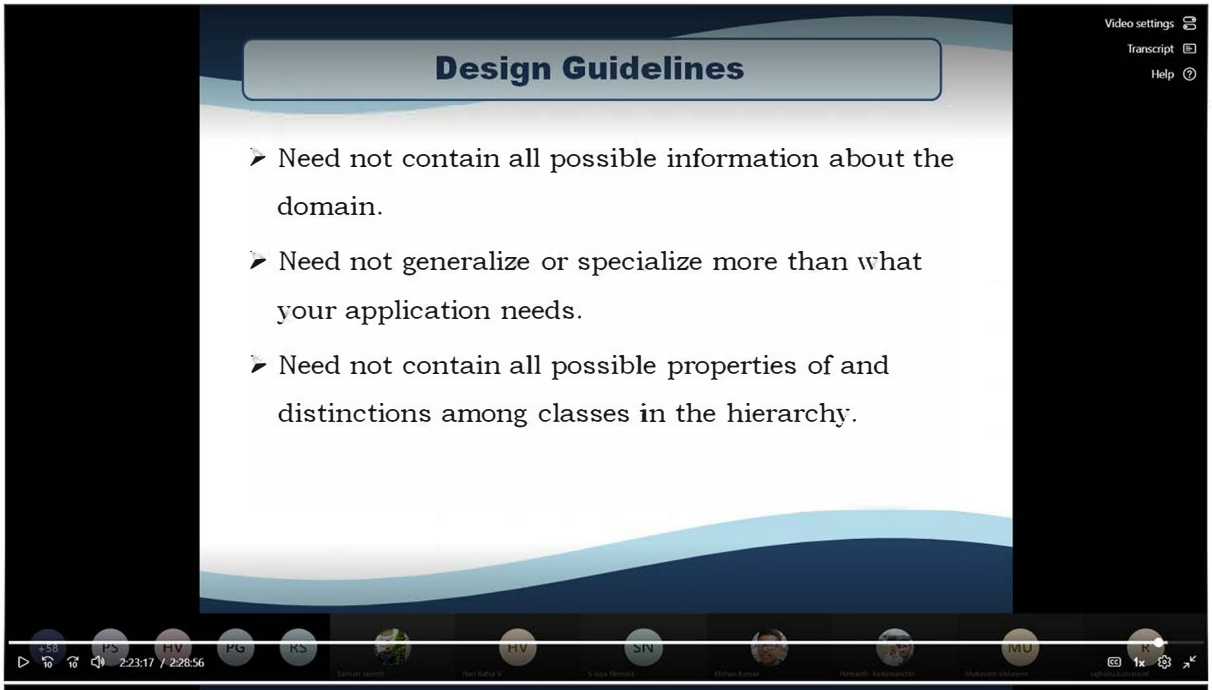

Protege-5.5.0
This PC › Local Disk (E:) › softwares › Protege-5.5.0
Name
Date modified
Type
Size
app
bin
bundles
conf
jre
plugins
pizza.owl.xml
Protege.exe
Protege.l4j.ini
run.bat
Video settings
Transcript
Help
2:06:38
2:06:38 / 2:28:56

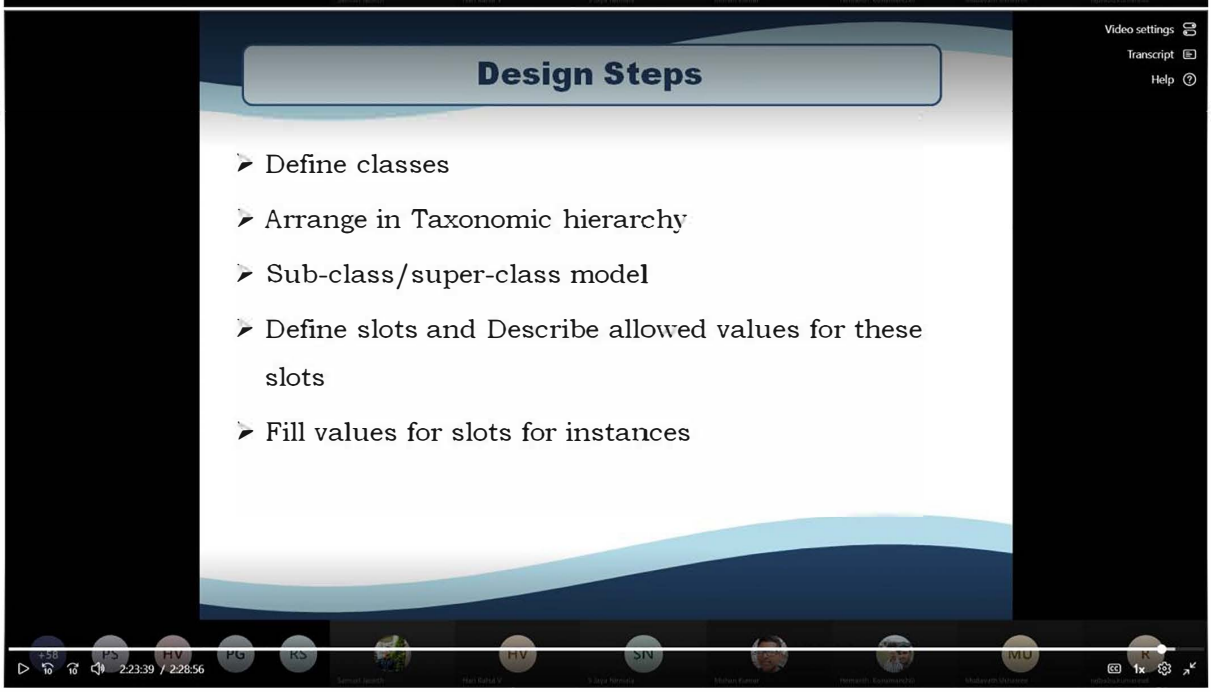

pizza (http://www.co-ode.org/ontologies/pizza/2.0.0) : [E:\softwares\Protege-5.5.0\pizza.owl.xml]
File Edit View Reasoner Tools Refactor Window Help
Video settings
Transcript
Help
pizza (http://www.co-ode.org/ontologies/pizza/2.0.0)
Search...
DomainThing) Food) Pizza) NamedPizza) AmericanHot
Active ontology × Entities × Classes × Object properties × Data properties × Annotation properties × Individuals by class × OWLViz × DL Query × OntoGraf × SPARQL Query ×
Class hierarchy
Class hierarchy (inferred)
Class hierarchy: AmericanHot
Asserted
IceCream
Pizza
CheesyPizza
InterestingPizza
MeatyPizza
NamedPizza
American
AmericanHot
Cajun
Capricciosa
Caprina
Fiorentina
FourSeasons
FruttiDiMare
Giardiniera
LaReine
Margherita
Mushroom
Napoletana
Parmense
PolloAdAstra
PrinceCarlo
QuattroFormaggi
Rosa
Siciliana
SloppyGiuseppe
Soho
Veneziana
NonVegetarianPizza
RealItalianPizza
SpicyPizza
Annotations
Usage
Annotations: AmericanHot
rdfs:label [language: en]
AmericanHot
rdfs:label [language: pt]
AmericanaPicante
skos:prefLabel [language: en]
Description: AmericanHot
Equivalent To
SubClass Of
hasCountryOfOrigin value America
hasTopping only
(HotGreenPepperTopping or JalapenoPepperTopping or MozzarellaTopping or PeperoniSausageTopping or TomatoTopping)
hasTopping some HotGreenPepperTopping
hasTopping some JalapenoPepperTopping
hasTopping some MozzarellaTopping
hasTopping some PeperoniSausageTopping
hasTopping some TomatoTopping
NamedPizza
General class axioms
No Reasoner set. Select a reasoner from the Reasoner menu
Show Inferences
2:12:41 / 2:28:56
1x

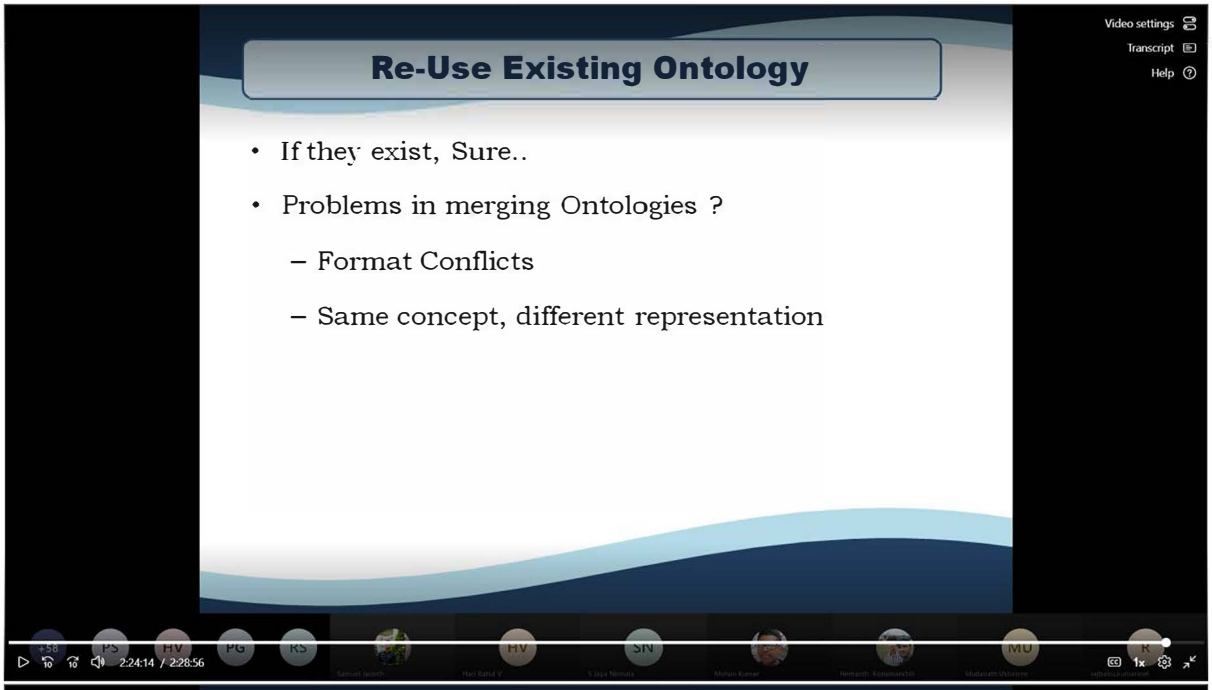

pizza (http://www.co-ode.org/ontologies/pizza/2.0.0) : [E:\softwares\Protege-5.5.0\pizza.owl.xml]
File Edit View Reasoner Tools Refactor Window Help
Video settings
Transcript
Help
Press Esc to exit full screen
pizza (http://www.co-ode.org/ontologies/pizza/2.0.0)
Search...
ValuePartition) Spiciness) Mild
Active ontology × Entities × Classes × Object properties × Data properties × Annotation properties × Individuals by class × OWLViz × DL Query × OntoGraf × SPARQL Query ×
Class hierarchy: Mild
Asserted
owl:Thing
DomainThing
Country
Food
IceCream
Pizza
PizzaBase
PizzaTopping
CheeseTopping
FruitTopping
HerbSpiceTopping
CajunSpiceTopping
RosemaryTopping
MeatTopping
NutTopping
SauceTopping
SeafoodTopping
SpicyTopping
VegetableTopping
VegetarianTopping
ValuePartition
Spiciness
Hot
Medium
Mild
OntoGraf
Search:
contains
Search
Clear
ParmezanTopping
Spiciness
SpinachTopping
LeekTopping
MushroomTopping
Food
GorgonzolaTopping
Mild
PizzaTopping
FruitTopping
HerbSpiceTopping
SeafoodTopping
NutTopping
MeatTopping
CheeseTopping
SpicyTopping
RosemaryTopping
CajunSpiceTopping
No Reasoner set. Select a reasoner from the Reasoner menu
Show Inferences
Exit fullscreen (Alt + F)
2:17:06 / 2:28:56
1x

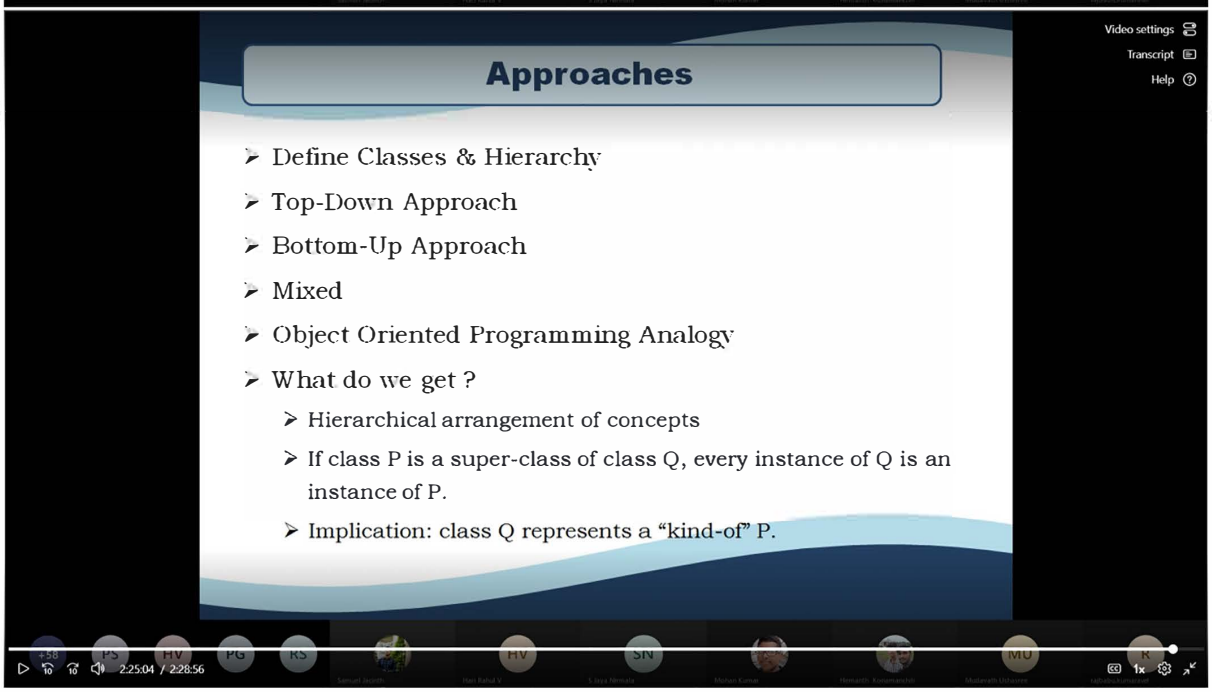

Video settings
Transcript
Help
Ontology Design Approach
There is no one correct way to model a domain.
There are always viable alternatives
Best solution depends on Applications and Extensions
Iterative Process
Concepts in Ontology close to objects (physical/logical) and relationships in the domain of interest
Objects are generally nouns
Relationships are generally verbs in a sentence
2:22:08 / 2:28:56

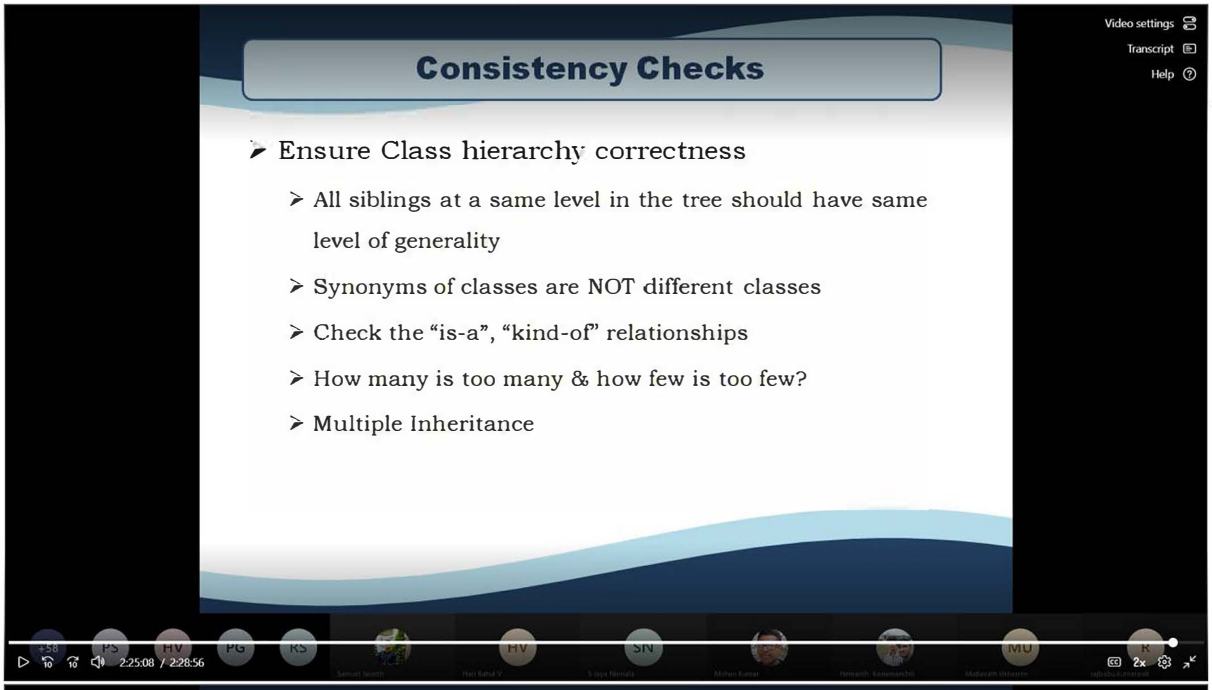

Video settings
Transcript
Help
Design Guidelines
What is the domain that it will cover?
For what are we going to use the ontology?
For what types of questions should the information in the ontology provide answers? (i.e., *competency questions*)
Who will use it?
2:22:38 / 2:28:56

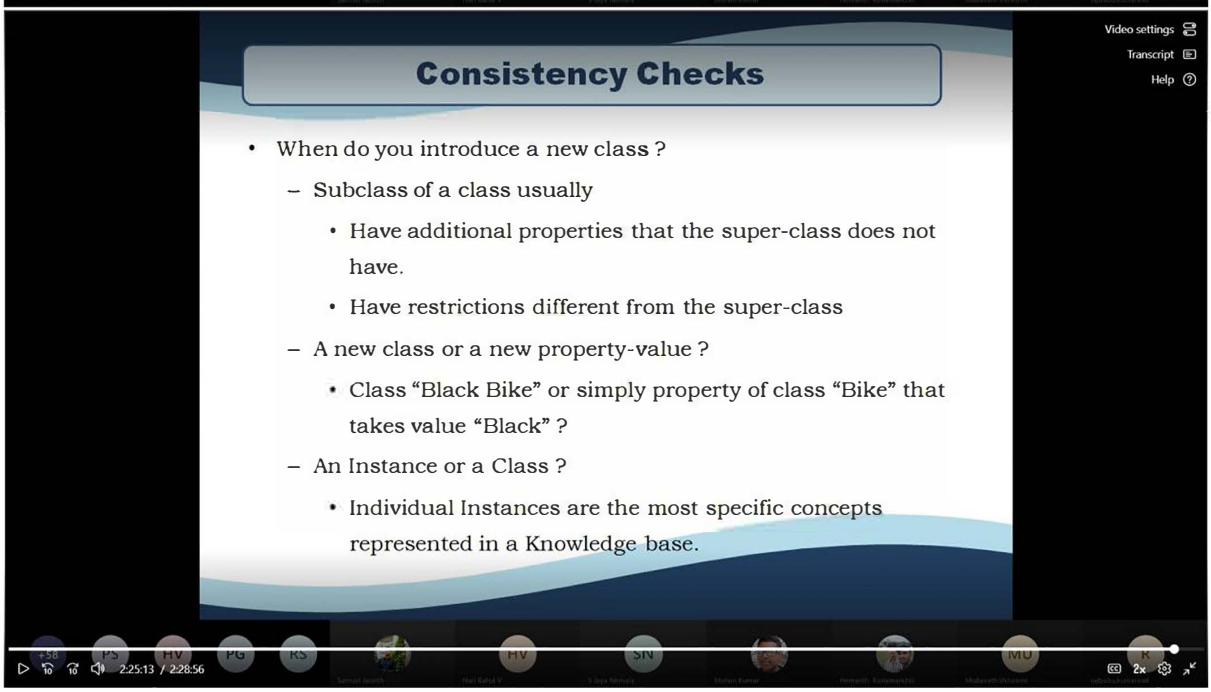

Video settings
Transcript
Help
Design Guidelines
Need not contain all possible information about the domain.
Need not generalize or specialize more than what your application needs.
Need not contain all possible properties of and distinctions among classes in the hierarchy.
2:23:17 / 2:28:56

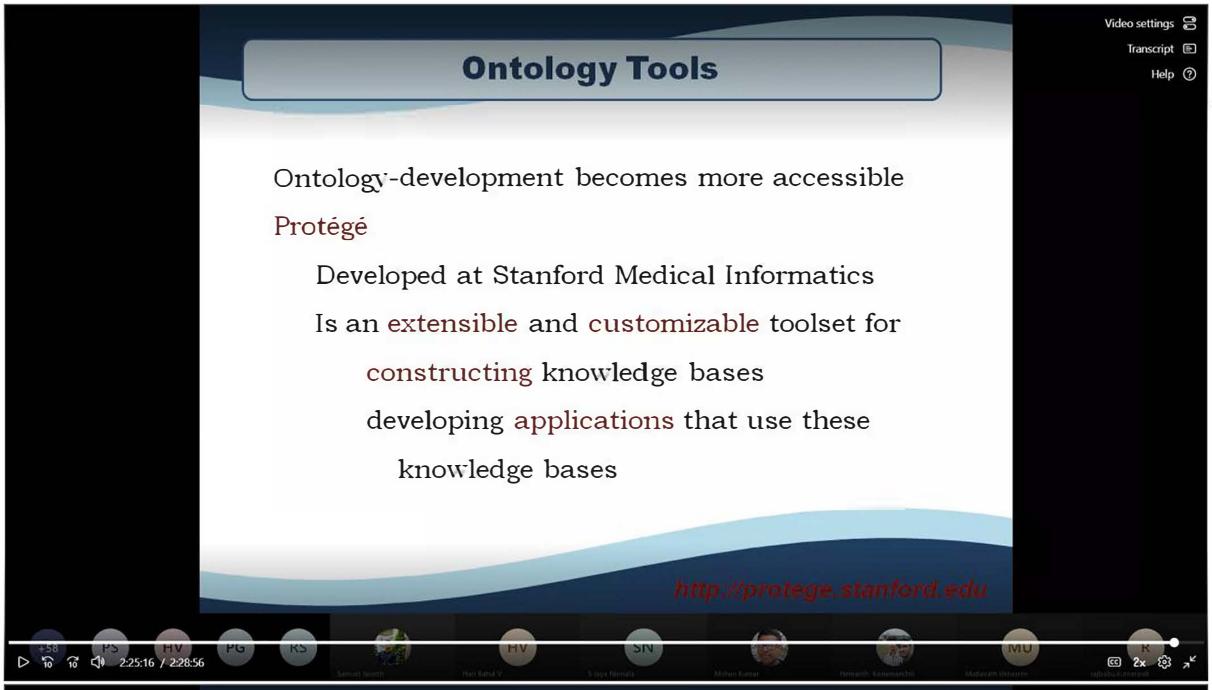

Video settings
Transcript
Help
Design Steps
Define classes
Arrange in Taxonomic hierarchy
Sub-class/super-class model
Define slots and Describe allowed values for these slots
Fill values for slots for instances
2:23:39 / 2:28:56

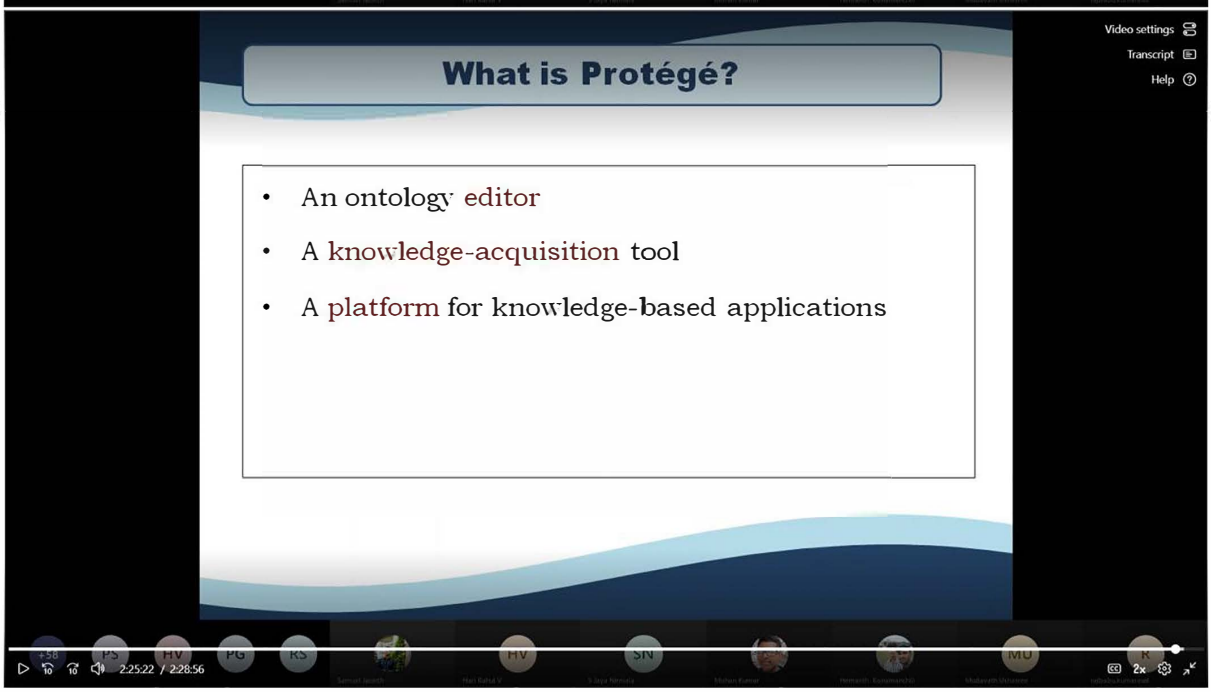

Video settings
Transcript
Help
Re-Use Existing Ontology
• If they exist, Sure..
• Problems in merging Ontologies ?
– Format Conflicts
– Same concept, different representation
2:24:14 / 2:28:56


Video settings
Transcript
Help
Approaches
➢ Define Classes & Hierarchy
➢ Top-Down Approach
➢ Bottom-Up Approach
➢ Mixed
➢ Object Oriented Programming Analogy
➢ What do we get ?
➢ Hierarchical arrangement of concepts
➢ If class P is a super-class of class Q, every instance of Q is an instance of P.
➢ Implication: class Q represents a "kind-of" P.
2:25:04 / 2:28:56


Video settings
Transcript
Help
Consistency Checks
Ensure Class hierarchy correctness
All siblings at a same level in the tree should have same level of generality
Synonyms of classes are NOT different classes
Check the "is-a", "kind-of" relationships
How many is too many & how few is too few?
Multiple Inheritance
2:25:08 / 2:28:56

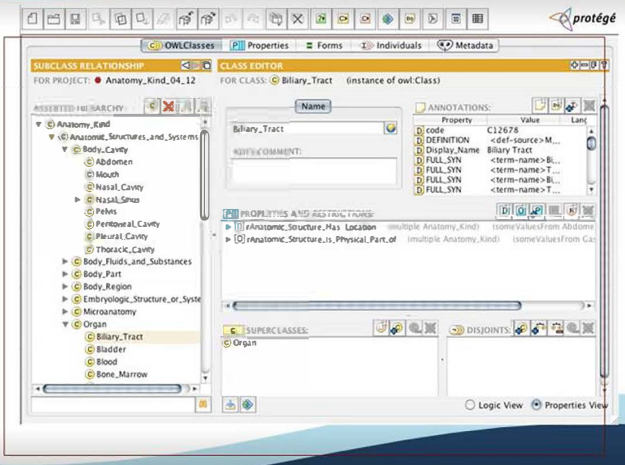

Video settings
Transcript
Help
Consistency Checks
When do you introduce a new class ?
Subclass of a class usually
Have additional properties that the super-class does not have.
Have restrictions different from the super-class
A new class or a new property-value ?
Class "Black Bike" or simply property of class "Bike" that takes value "Black" ?
An Instance or a Class ?
Individual Instances are the most specific concepts represented in a Knowledge base.
2:25:13 / 2:28:56


Video settings
Transcript
Help
Ontology Tools
Ontology-development becomes more accessible
Protégé
Developed at Stanford Medical Informatics
Is an extensible and customizable toolset for
constructing knowledge bases
developing applications that use these
knowledge bases
http://protege.stanford.edu
2:25:16 / 2:28:56


Video settings
Transcript
Help
What is Protégé?
An ontology editor
A knowledge-acquisition tool
A platform for knowledge-based applications
2:25:22 / 2:28:56

Video settings
Transcript
Help

Protégé – Ontology Editor



Exit fullscreen (Alt + F)
2:25:24 / 2:28:56
2x

Video settings
Transcript
Help

What makes Protégé Different?

- Easy-to-use graphical interface
- Scalability
  - currently can handle up to 5 million concepts
- Plugin architecture
  - active international community of plugin developers
- It's a platform for other applications
  - Integration with Eclipse (Mayo Clinic)
  - A server and a client for (Semantic) Web Services
- Open source

Exit fullscreen (Alt + F)
2:25:25 / 2:28:56
2x


Video settings
Transcript
Help
THANK YOU
THANK YOU VERY MUCH
For Queries & Contact
rajbabu@bhel.in
+91 9489202899
2:27:30 / 2:28:56
2x