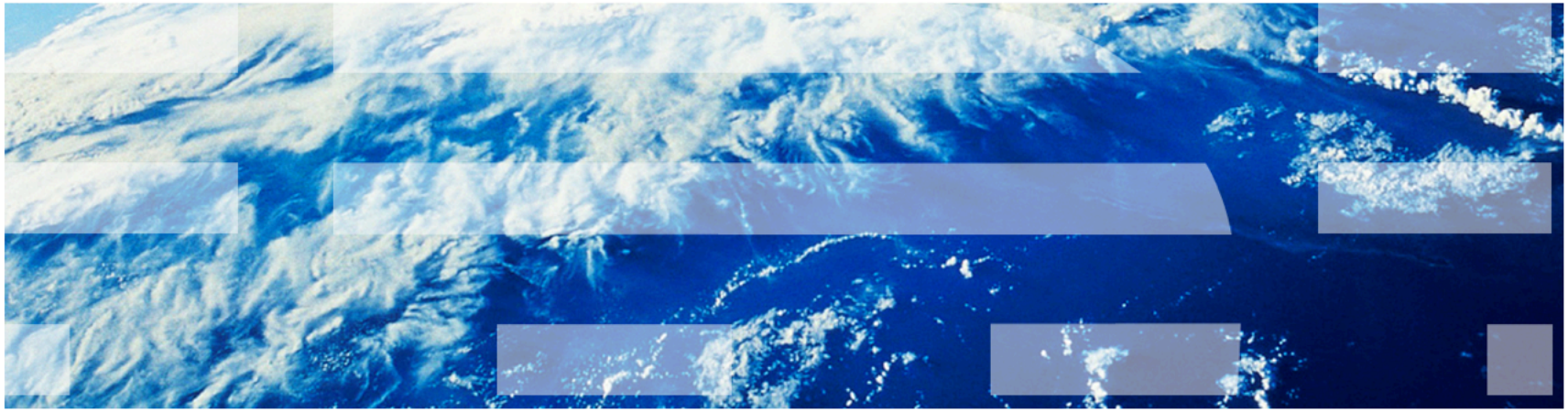
E6893 Big Data Analytics Lecture 13:

Big Data Analytics and Ai for Healthcare

Ching-Yung Lin, Ph.D.

Adjunct Professor, Dept. of Electrical Engineering and Computer Science



December 6th, 2019

Outline

- Mission
- What is Artificial Intelligence?
- AI Techniques for Artificial Empathy
- Homecare and Eldercare
- AI Robot
- Cancer Pathology Research
- Cancer Gene Research
- Medical Images
- Gaining Biomedical Knowledge
- Data Privacy Laws and Technologies
- Looking forward to the fast growth of Medical AI

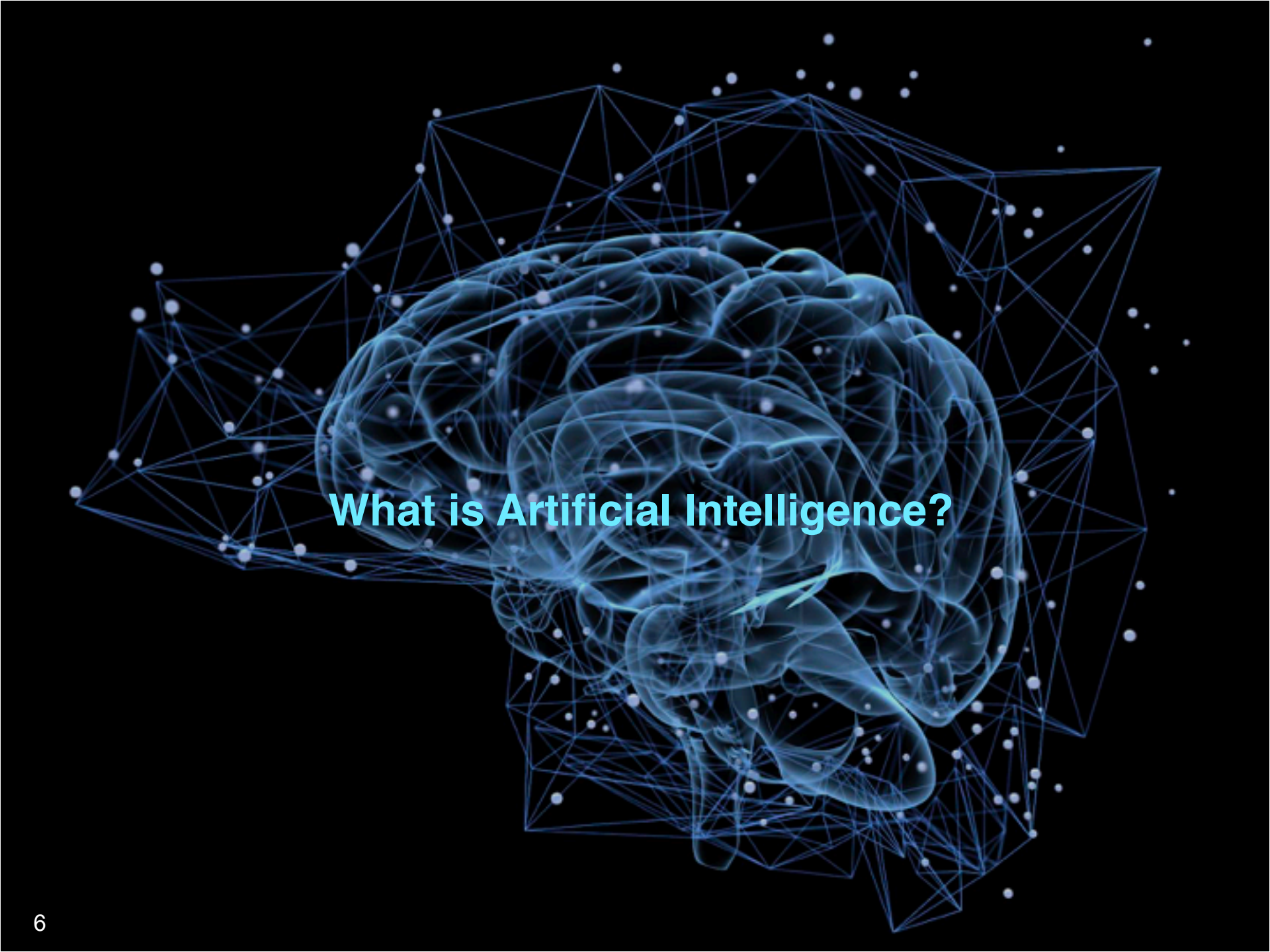
Example Areas

- AI Pathology Analytics System
  - AI pathology analysis system capable of interpreting pathology reports.
- Cancer Research
  - Collaborative research classifying the tumor type and predict the patient survival time after diagnosis by means of gene sequencing.
- AI Sensors for Eldercare and Patient Care
  - Utilizing multimodality sensors to automatically recognize abnormal events at home for eldercare or in hospital for patient care
- Artificial Empathy and Cognitive Robot
  - AI system that are cognitively smart and able to detect and react to our emotions because they have emotional intelligence, they are aware of our mental, social and emotional states, and familiar with our moods and preferences.

Humanoid robot, **Adam**:

- Biomedical Knowledge Graph
- NLP Systems for medical concept identification and semantic predicate extraction.
- Senses feelings visually and in speech.
- Interact through dialogs, touch and visuals.
- Face recognition.
- Object detection and segmentation.
- Face detection and tracking.
- Engagement status.
- Autonomous learning to act, react, and interact like humans.
- Gender and age detection.
- Question Answering.

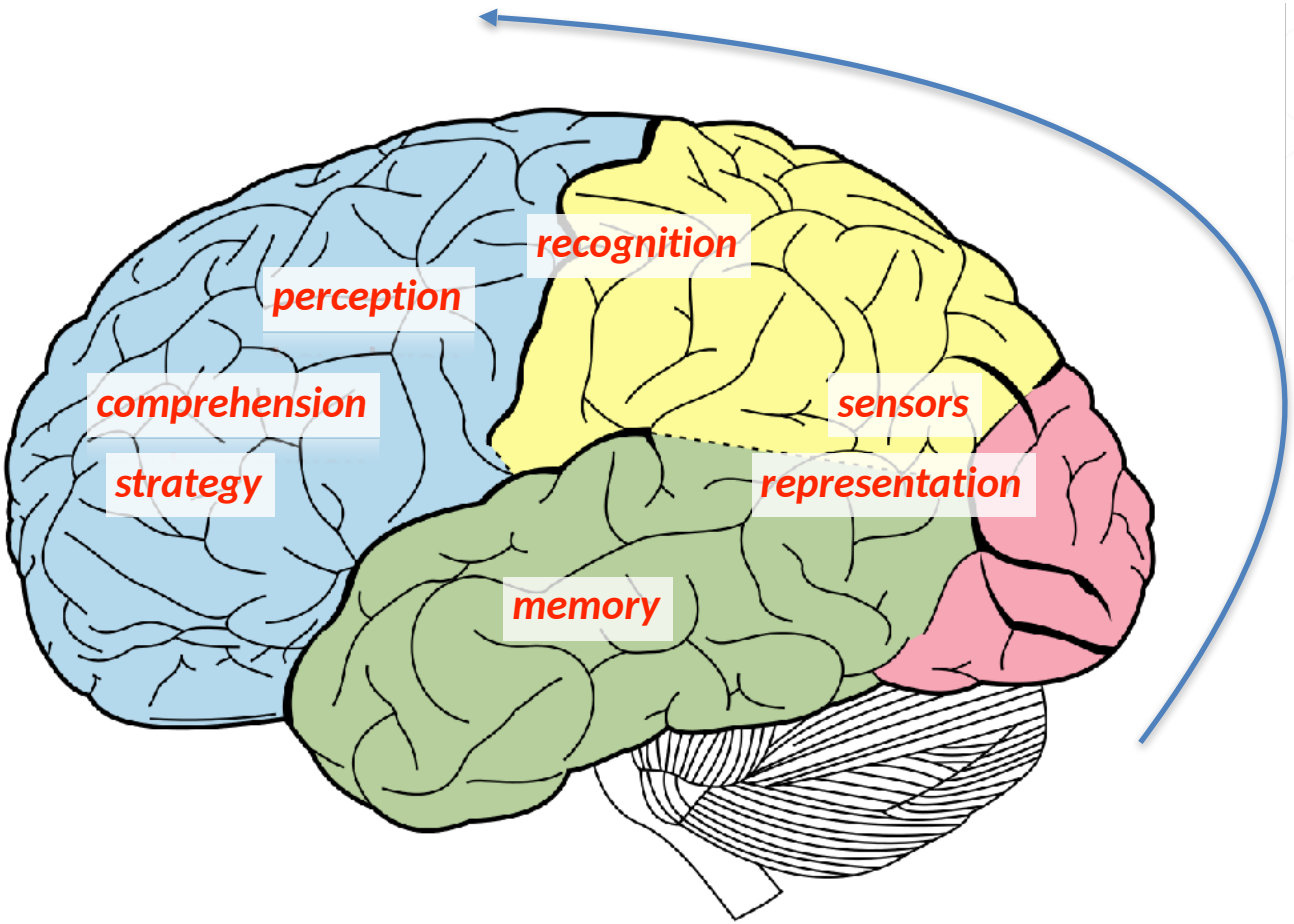
Dr. Adam

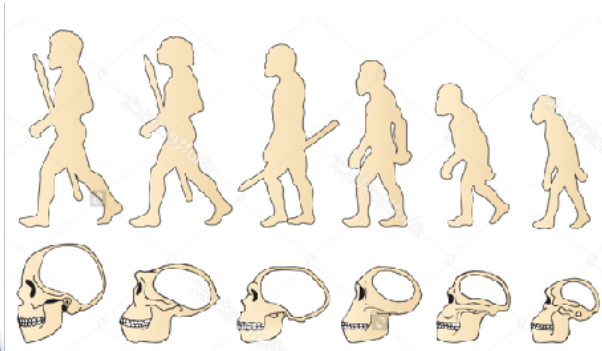

What is Artificial Intelligence?

Evolution of animal Intelligence

Columbia University

Direction of Evolution





Why is Graph Computing the core of AI?

How do kids grow intelligence?

"Airplane"

"Grandma"

"Grandma is in Taiwan"

"Auntie is also in Taiwan"

"I like grandma"

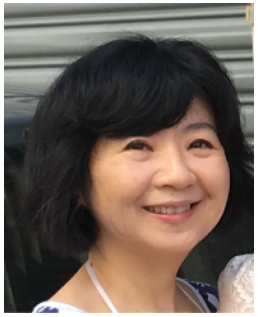
"I like grandpa"

The boy said:

"I like grandma and grandpa"

COLUMBIA UNIVERSITY

How does a Graph Computing mechanism reason and store data?



"see"

"Airplane"

"visit"

"Grandma"

"in"

"Taiwan"

"in"

"I"

"like"

"and"

"like"

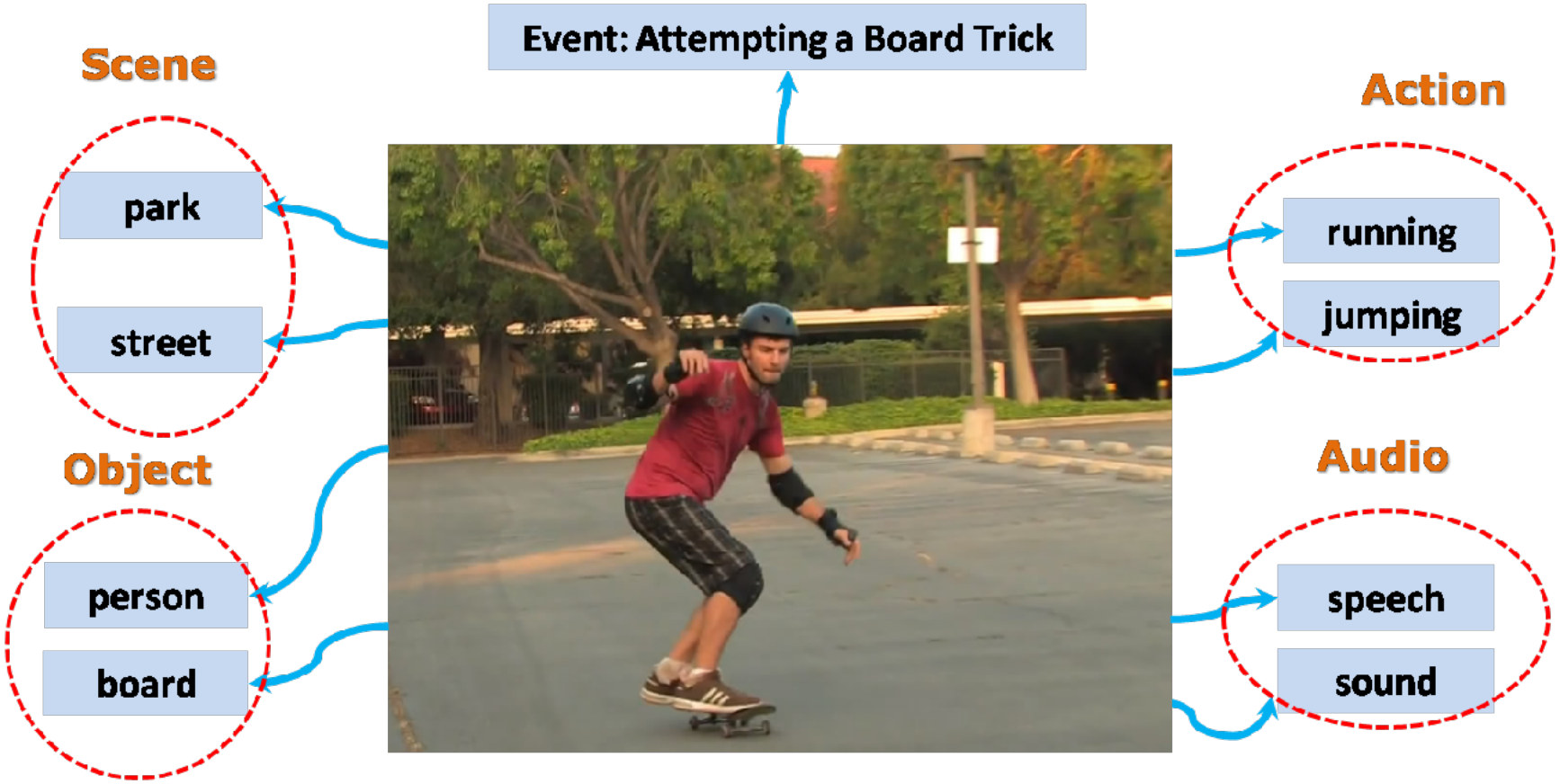
"Grandpa"

"Auntie"

Graph platform works like the human mind, connecting the dots when comprehending.

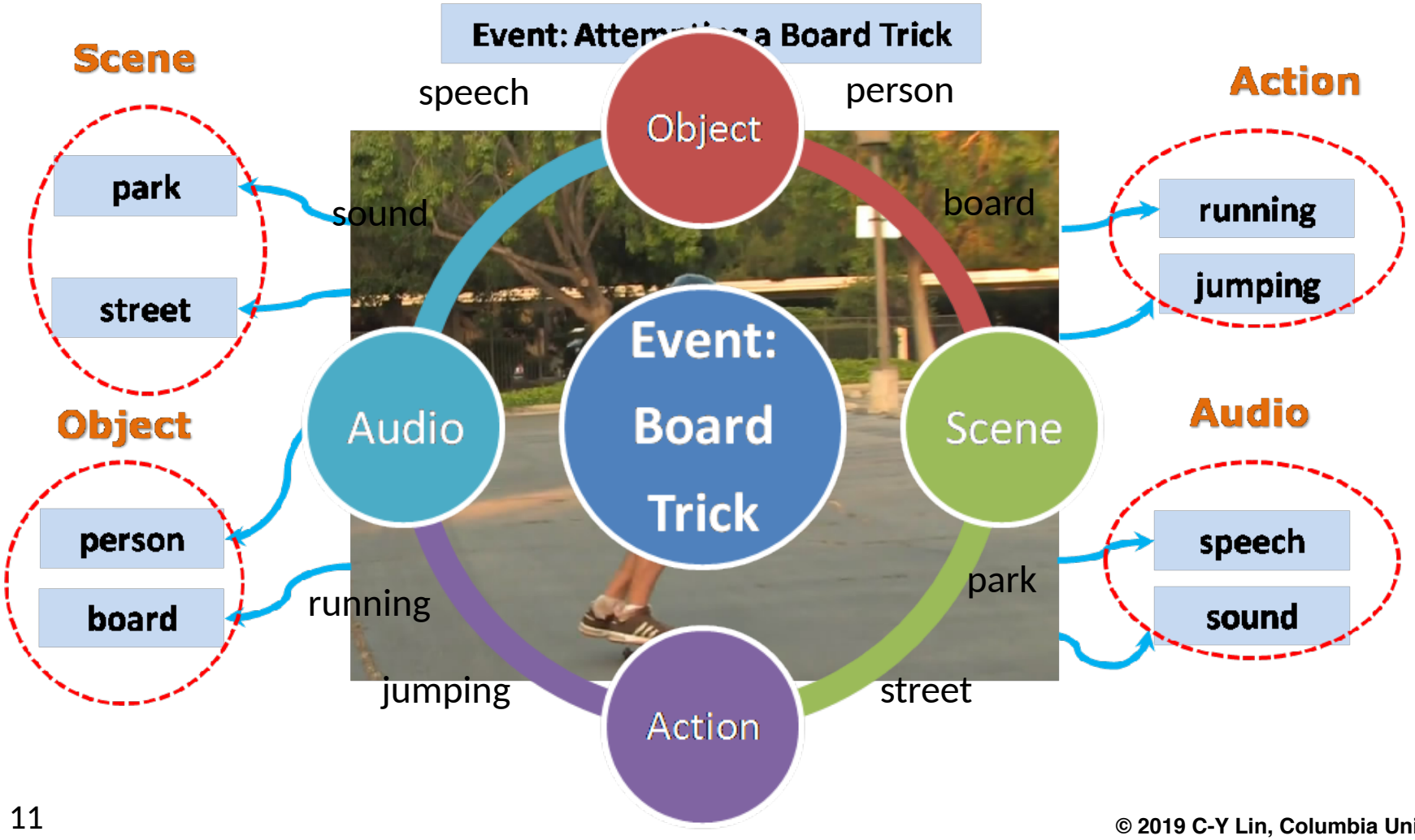
Example: Recognizing an Event

- Recognize an event from sensing -> feature extortion -> object/ scene/action detection -> semantics -> cognition inference



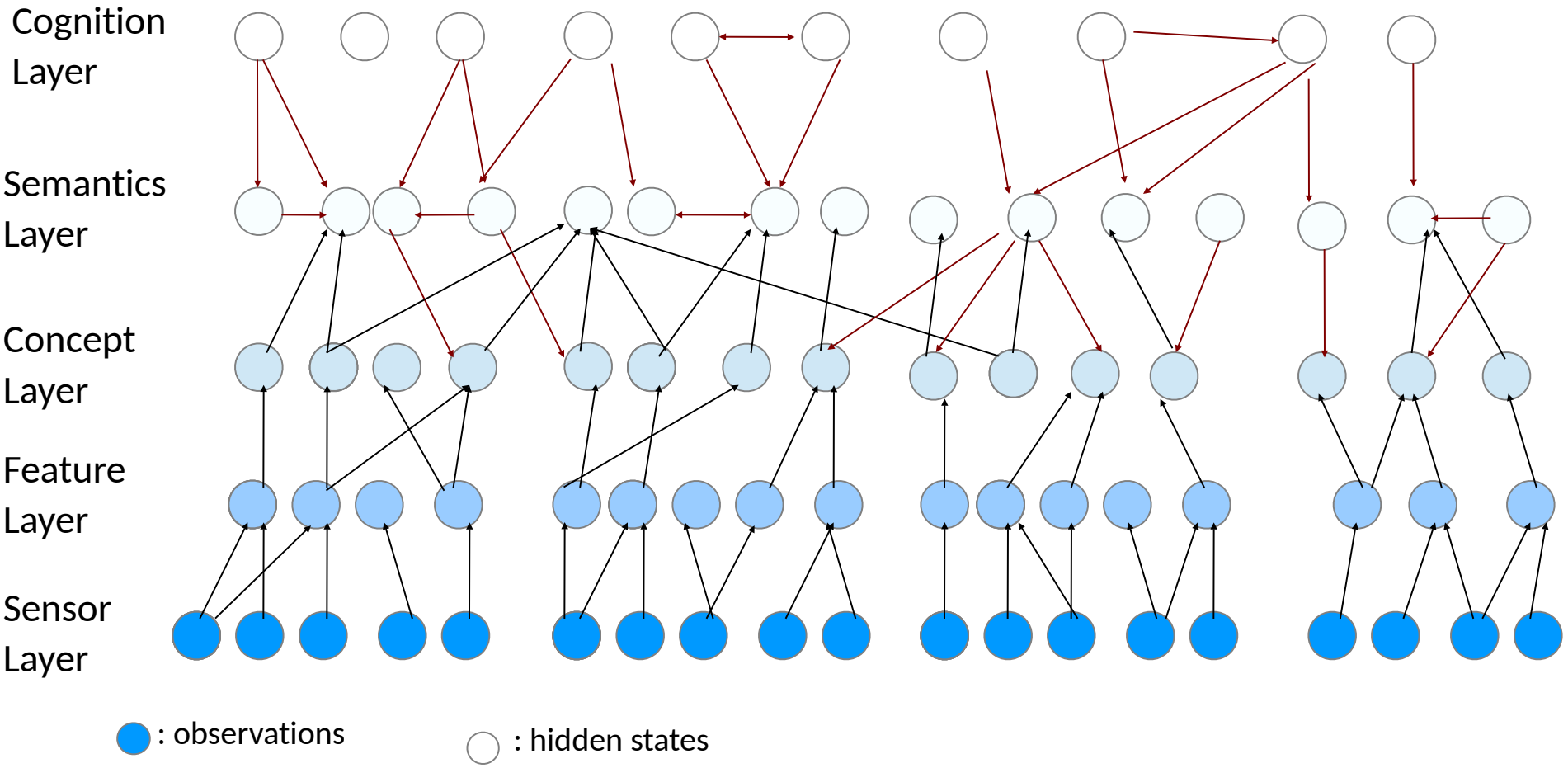
Example: Recognizing an Event

- Recognize an event from sensing -> feature extortion -> object/ scene/action detection -> semantics -> cognition inference



Layers of Machine Understanding

- *Structure Learning*
- *Evolutionary Behavioral Modeling & Prediction*



Layers of Machine Understanding

Columbia University

- *Structure Learning*
- *Evolutionary Behavioral Modeling & Prediction*

Future AI / Graphen's AI

Cognition Layer

Semantics Layer

Concept Layer

Feature Layer

Sensor Layer

: observations

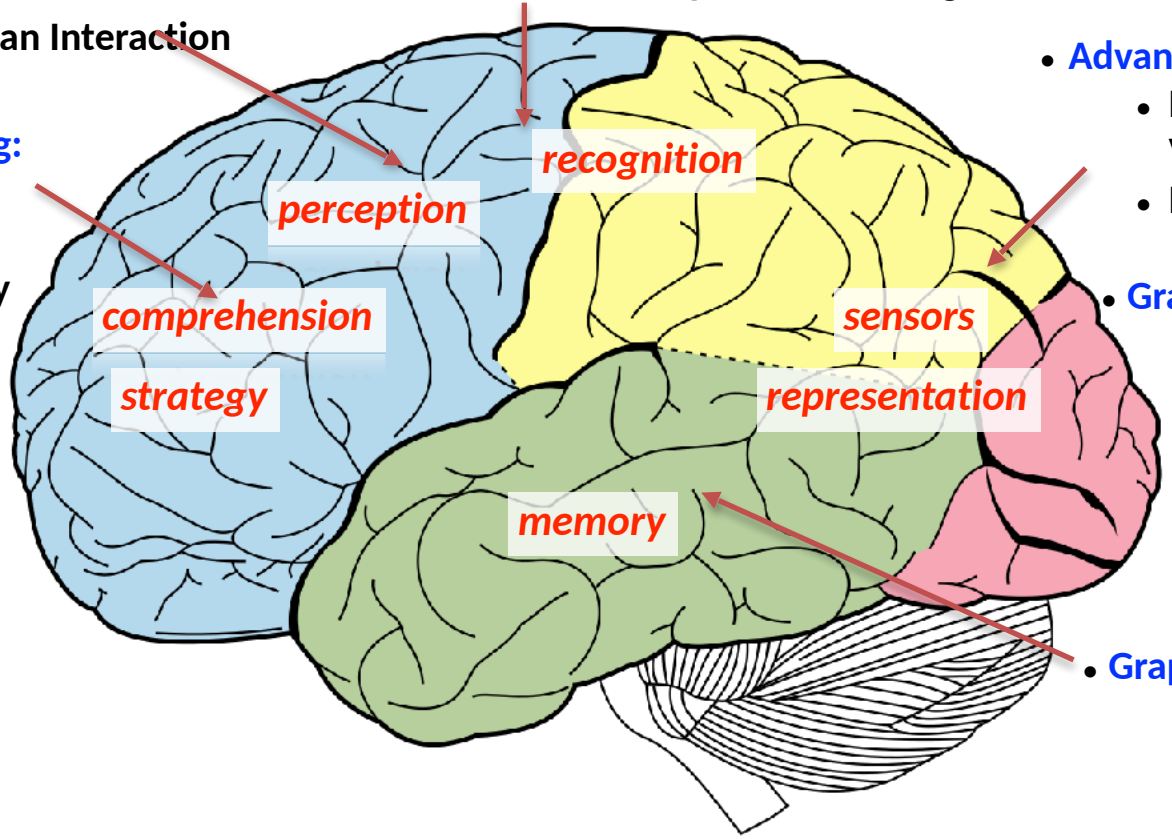
: hidden states

Most of Today's AI

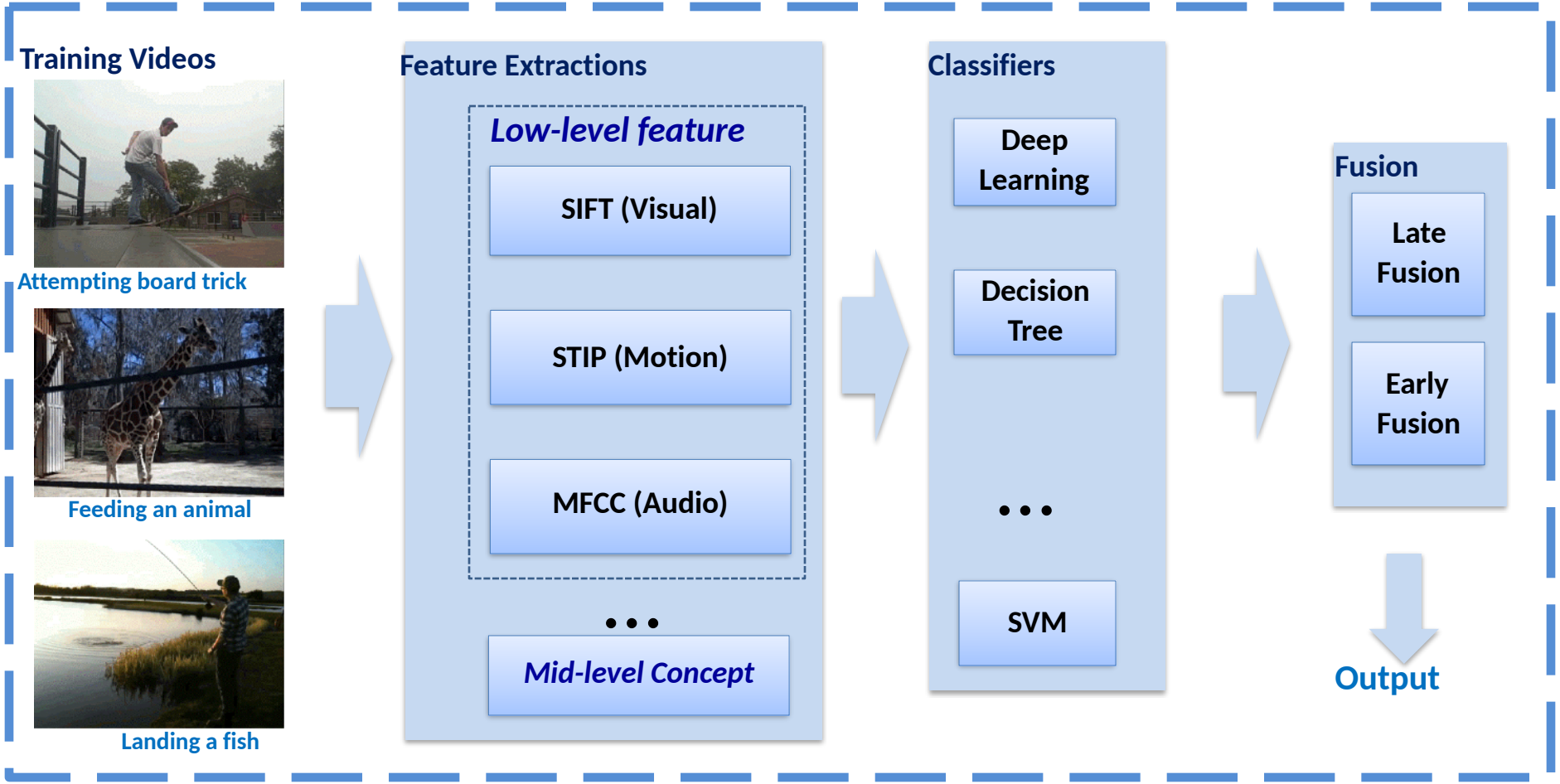
Human brain is a graph of 100B nodes and 700T edges. Graphen creates a next-generation AI platform that mimics ***full brain*** **functions**.

Graphen's Ardi AI platform makes machines remember, associate, reason, etc., like or beyond human.

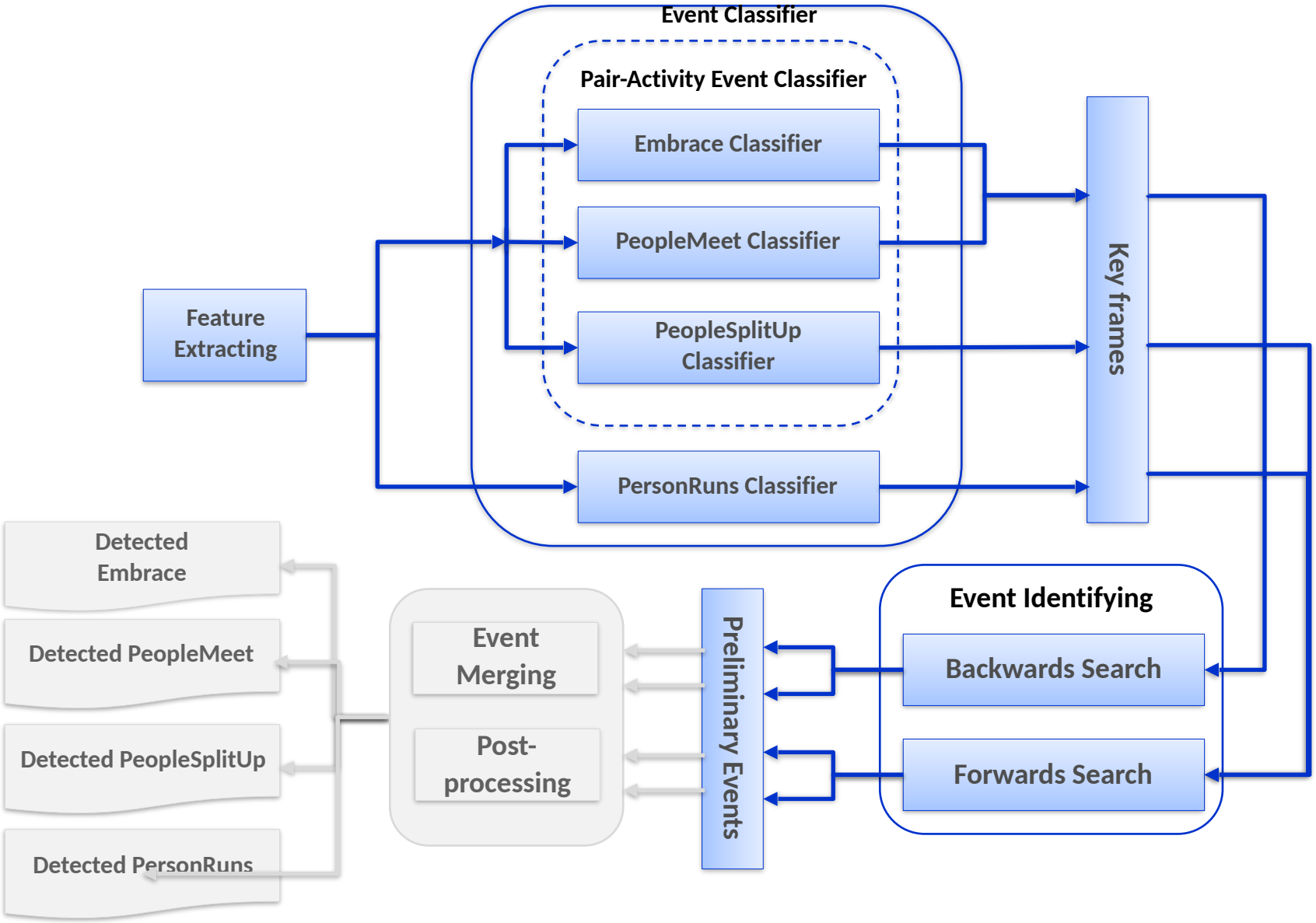
- **Machine Cognition:**
  - Robot Cognition Tools
  - Feeling
  - Robot-Human Interaction
- **Machine Learning:**
  - ML and Deep Learning
  - Autonomous Imperfect Learning
- **Machine Reasoning:**
  - Bayesian Networks
  - Game Theory Tools
- **Advanced Visualization:**
  - Dynamic and Interactiv Viz.
  - Big Data Viz.
- **Graph Analytics:**
  - Network Analysis
  - Flow Prediction
- **Graph Database:**
  - Distributed Native Database



Event Detection Baseline



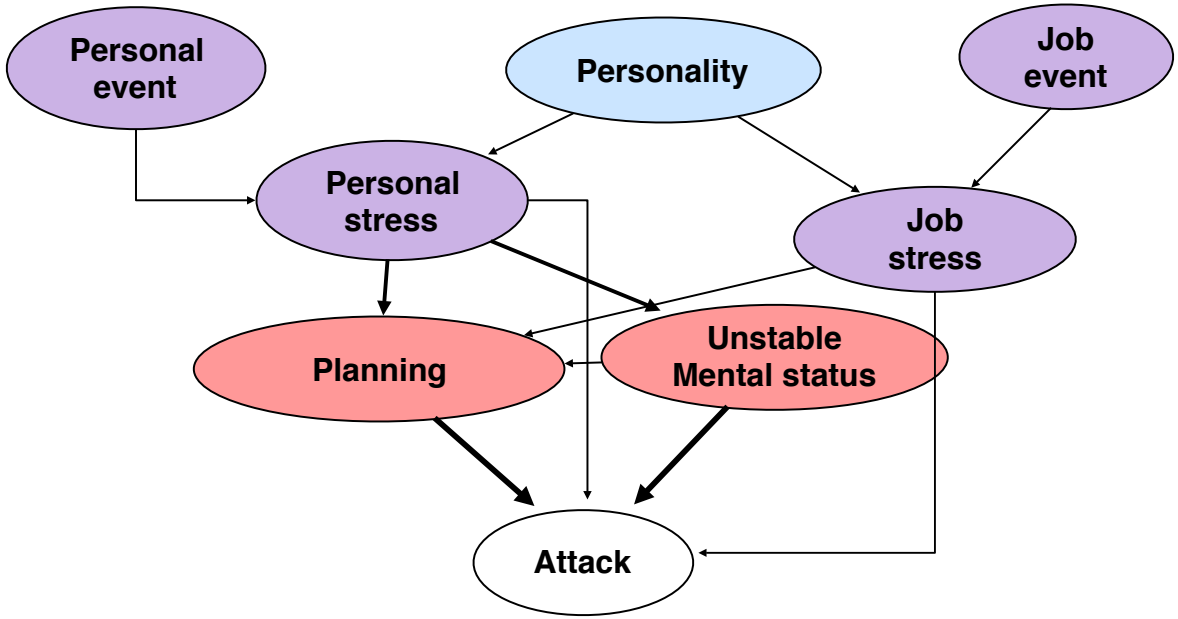
Event Classification Example



Example of AI-based Reasoning

- Personal stress:
  - Gender identity confusion
  - Family change (termination of a stable relationship)
- Job stress:
  - – Dissatisfaction with work
    - Job roles and location (sent to Iraq)
    - long work hours (14/7)

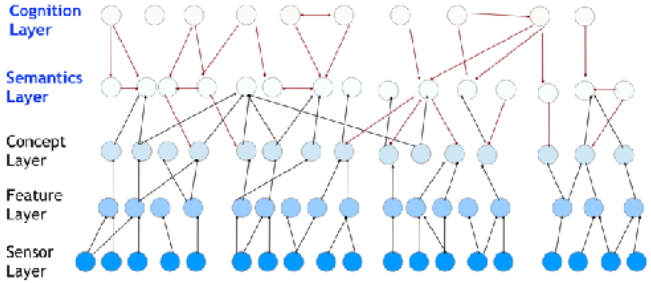
- Unstable Mental Status:
  - Fight with colleagues, write complaining emails to colleagues
  - Emotional collapse in workspace (crying, violence against objects)
  - Large number of unhappy Facebook posts (work-related and emotional)
- Planning:
  - Online chat with a hacker confiding his first attempt of leaking the information

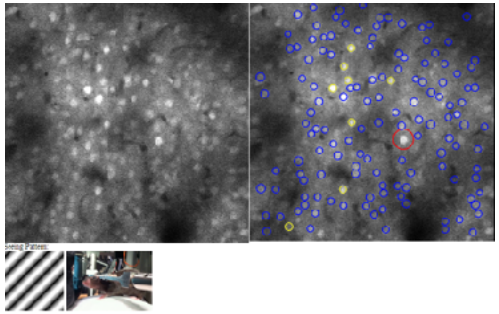


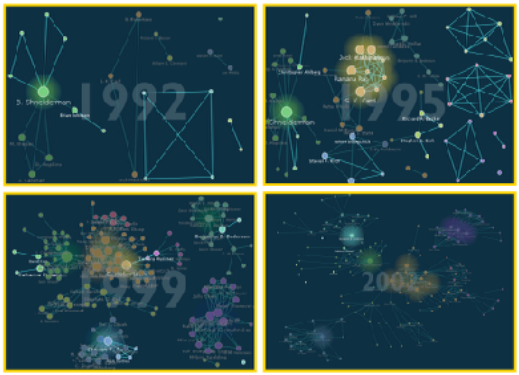
- Attack:

  – Brought music CD to work and downloaded/ copied documents onto it with his own account

Ardi Platform provides Key Differentiators

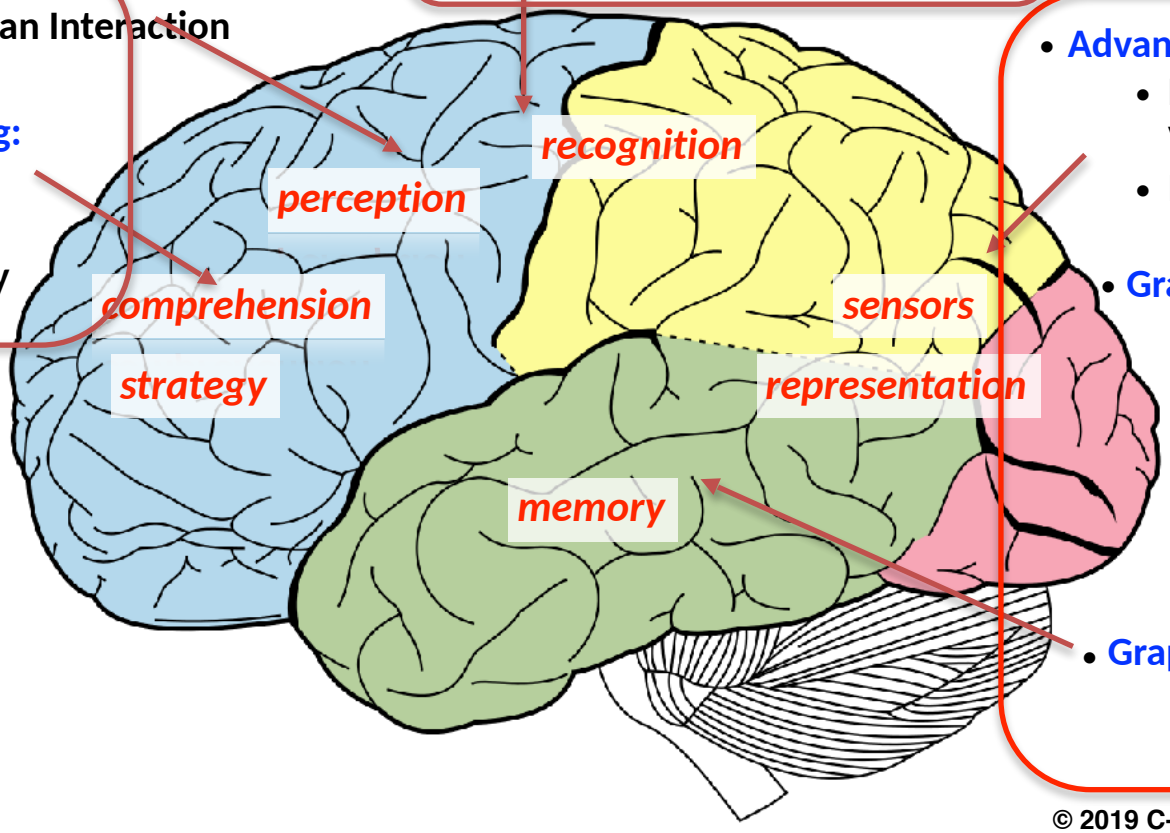




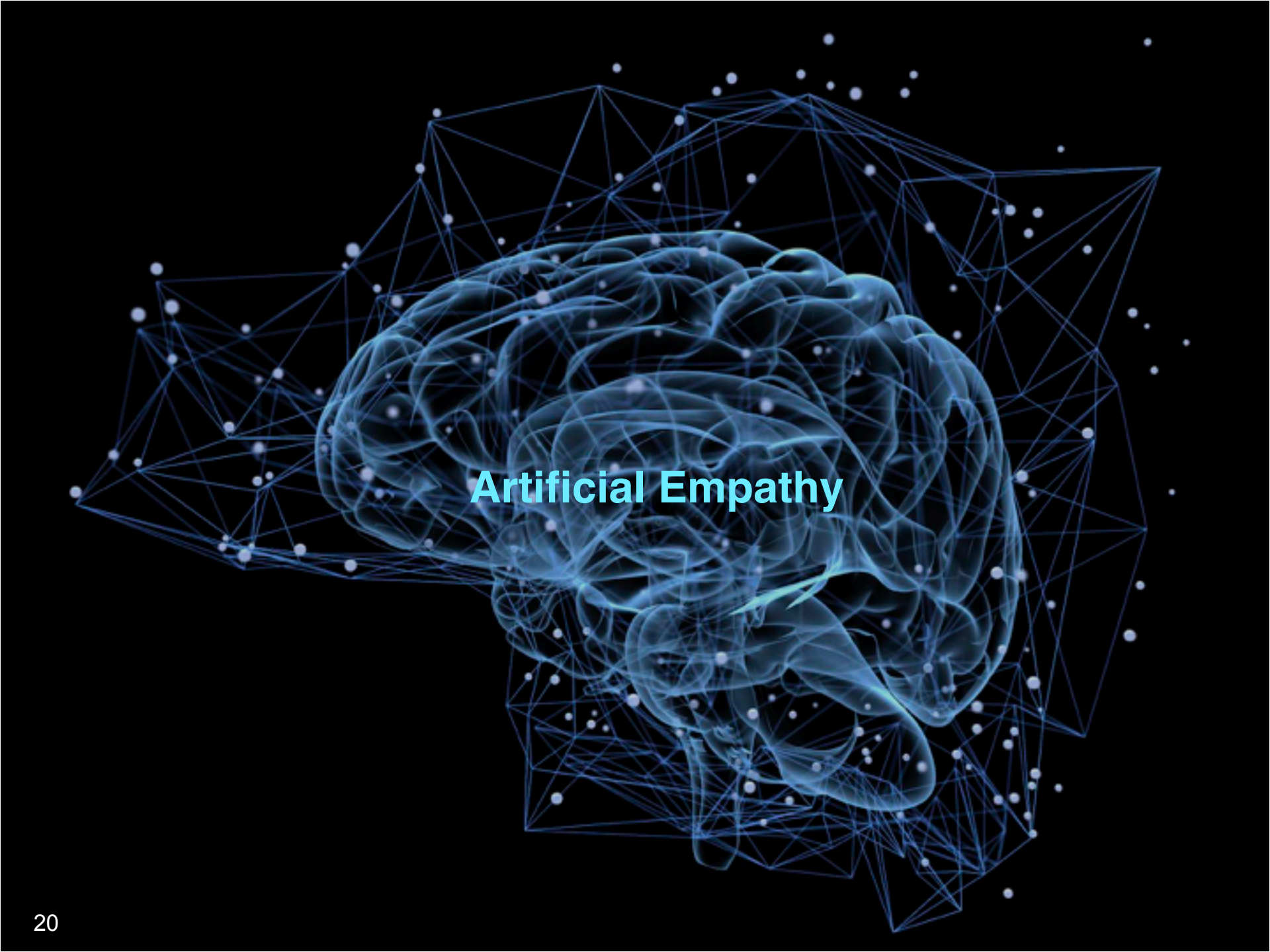


- **Machine Cognition:**
  - Robot Cognition Tools
  - Feeling
  - Robot-Human Interaction
- **Machine Reasoning:**
  - Bayesian Networks
  - Game Theory Tools

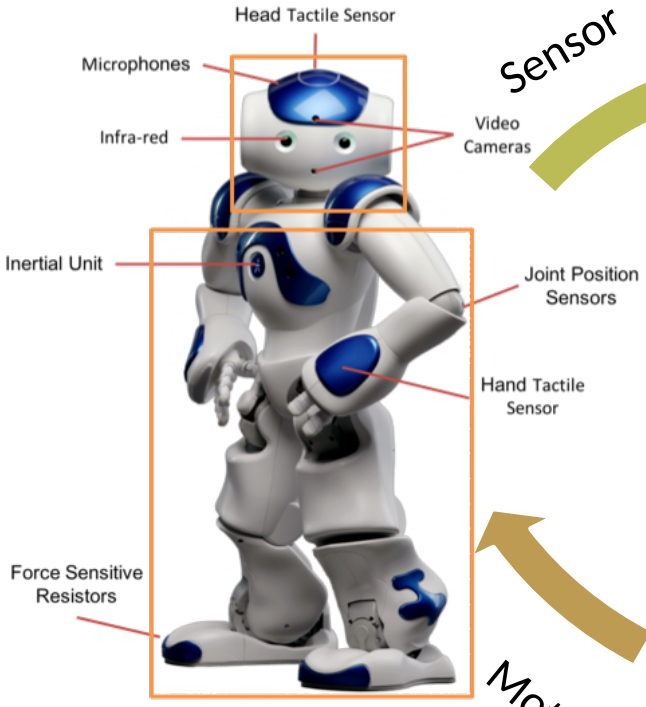
- **Machine Learning:**
  - ML and Deep Learning
  - Autonomous Imperfect Learning



- **Advanced Visualization:**
  - Dynamic and Interactive Viz.
  - Big Data Viz.
- **Graph Analytics:**
  - Network Analysis
  - Flow Prediction
- **Graph Database:**
  - Distributed Native Database


Artificial Empathy

Artificial Empathy



Sensor

- Audio
- Visual
- Infra-red

Multi-Modality Information

Understanding

Emotion Perception

Artificial Empathy

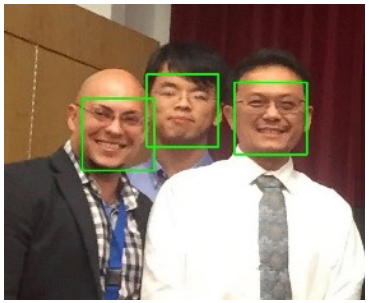
Expression

Emotion Contagion

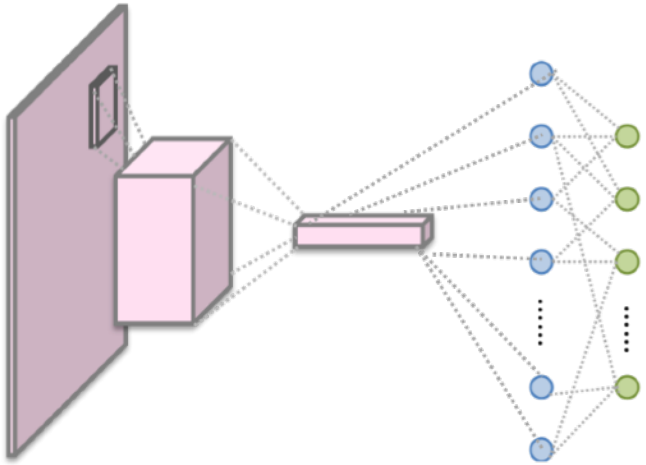
Motor Mimicry

Motion

Frame-Based Facial Expression Recognition



Face Detection



Convolutional Neural Network

Emotion Recognition

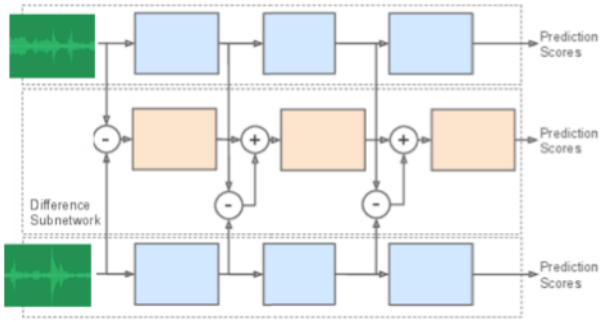
- Angry
- Disgust
- Fear
- **Happy**
- Sad
- Surprise
- Neutral

Feature Extraction

Features

Audio-Based Emotion Recognition

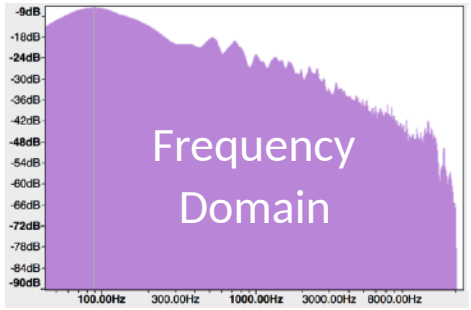
Audio Clips



Temporal Difference Network

Emotion Recognition

Angry
Disgust
Fear
Happy
Sad
Surprise
Neutral

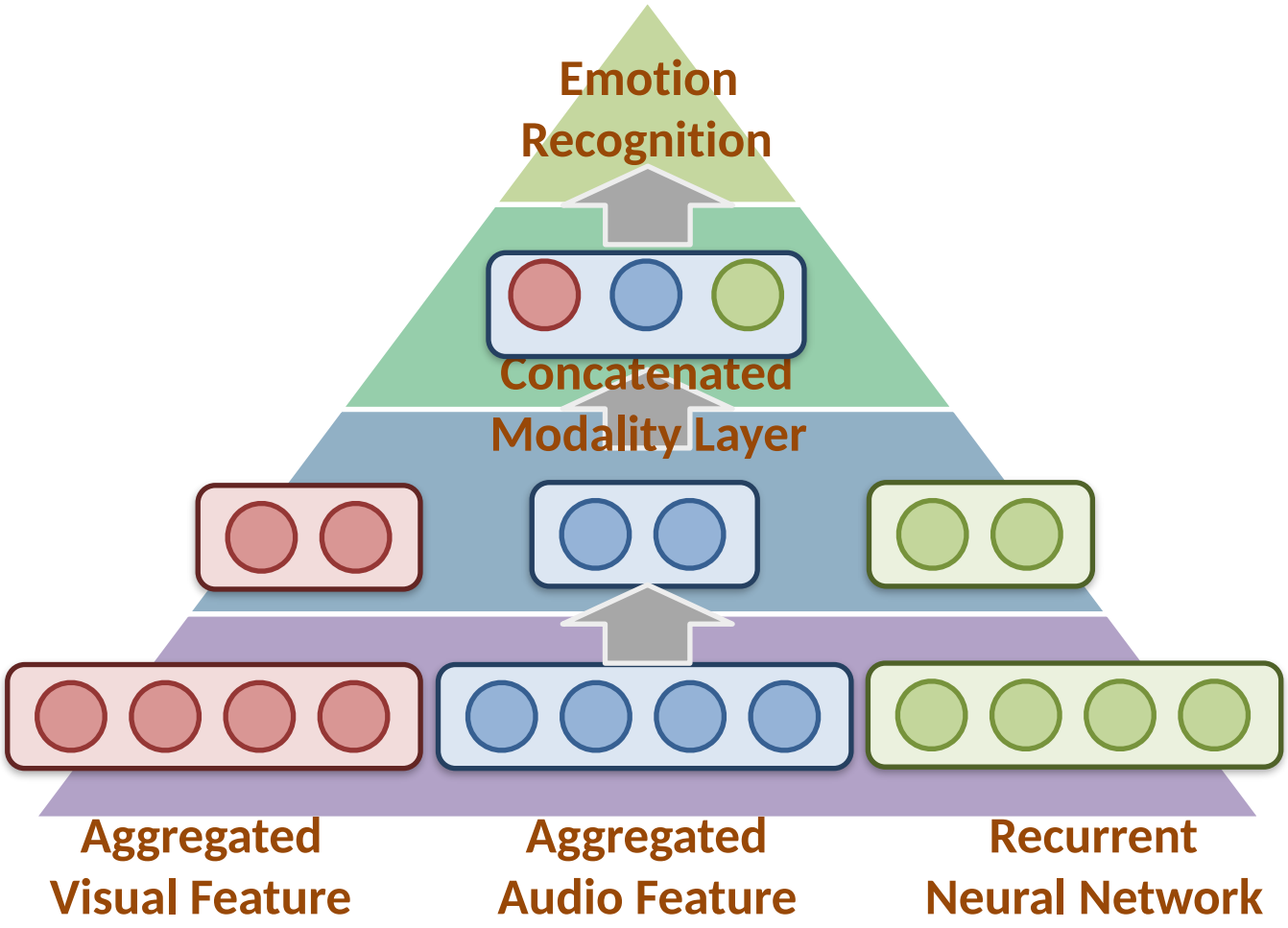


Low Level Feature Descriptors

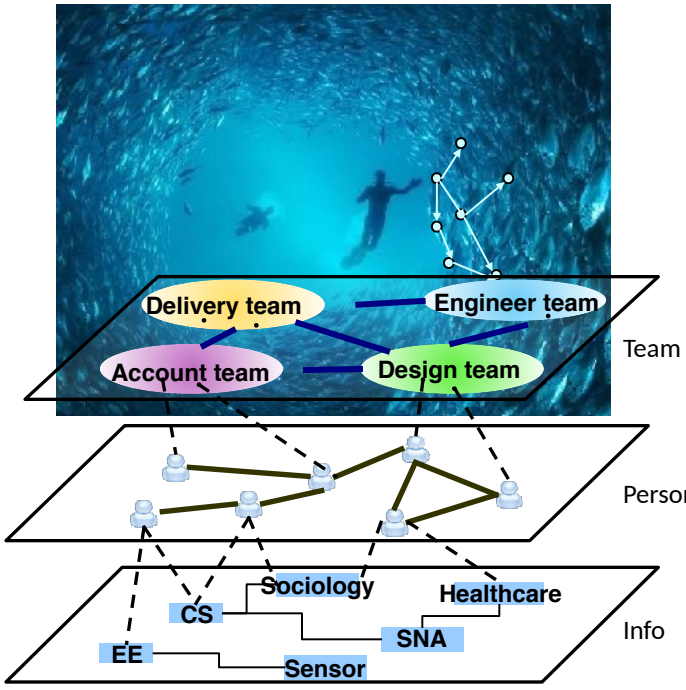
Feature Extraction

Features

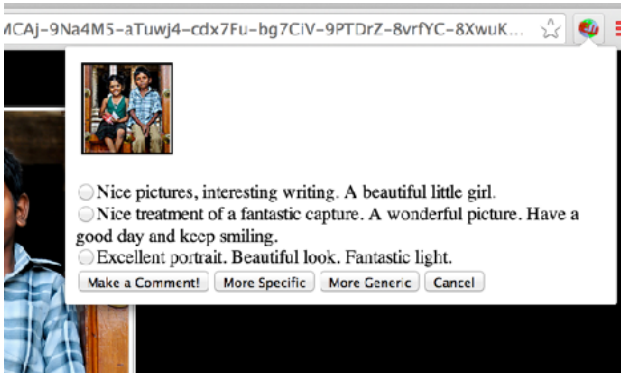
Emotion Recognition in Video



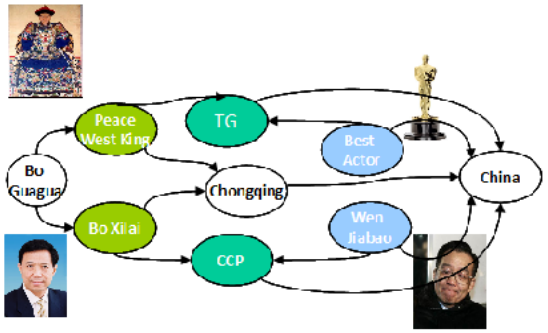
AI of Autonomous Large-Scale Human Understanding

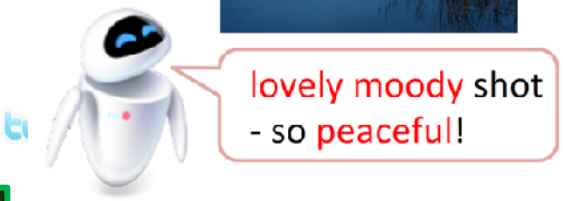


Information Evolution



Machine Interpretation





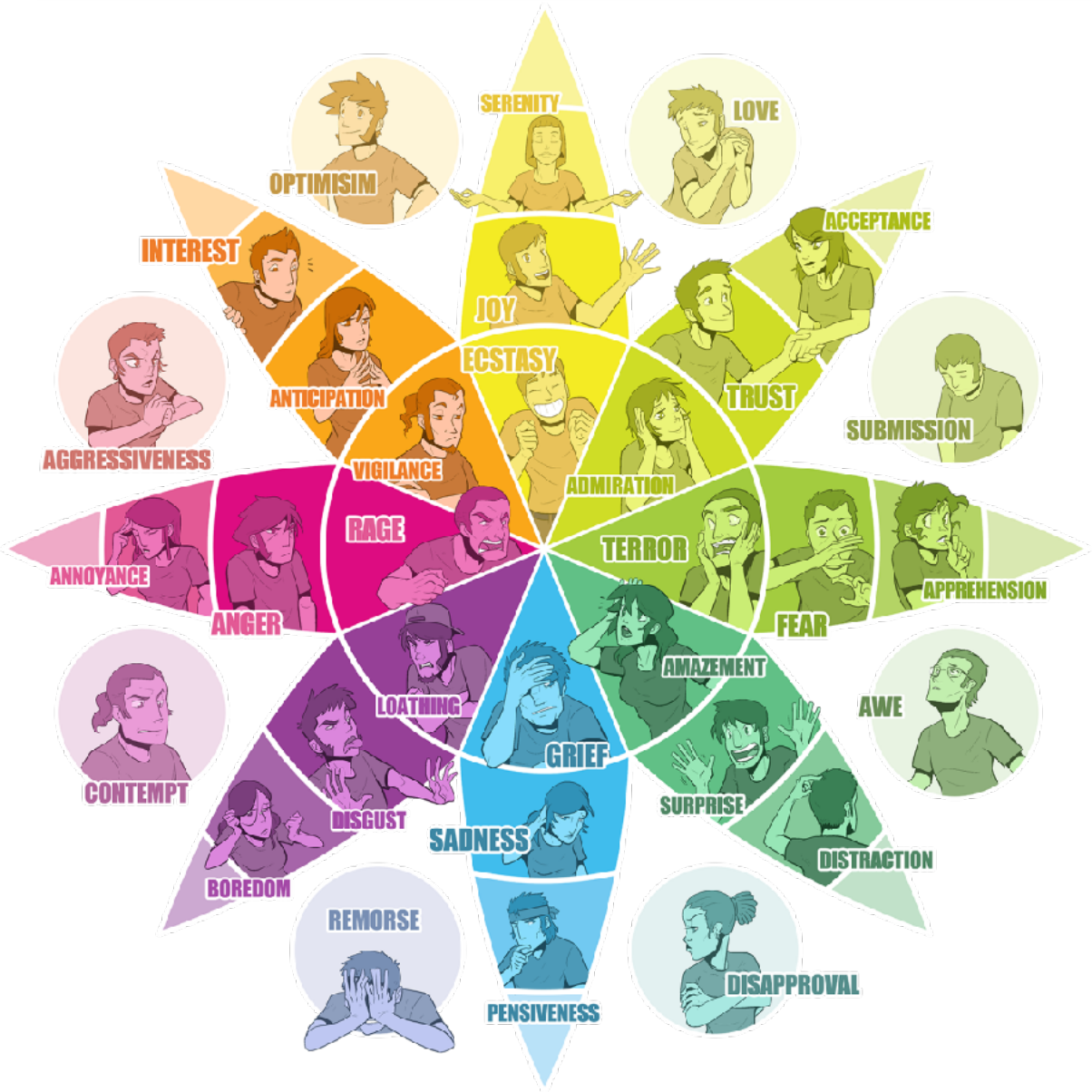
Machine Feeling — Detection results of “lonely dog”

Machine Feeling — Detection results of “crazy car”

Human Interpretation

Personality
Needs
Values
Trustworthiness
Trustingness
Influence

Making Robots Feel



Psychology emotion wheel (24 emotions, by Robert Plutchik)

Plenty on the Web: "*For content to go viral, it needs to be emotional,*" Dan Jones

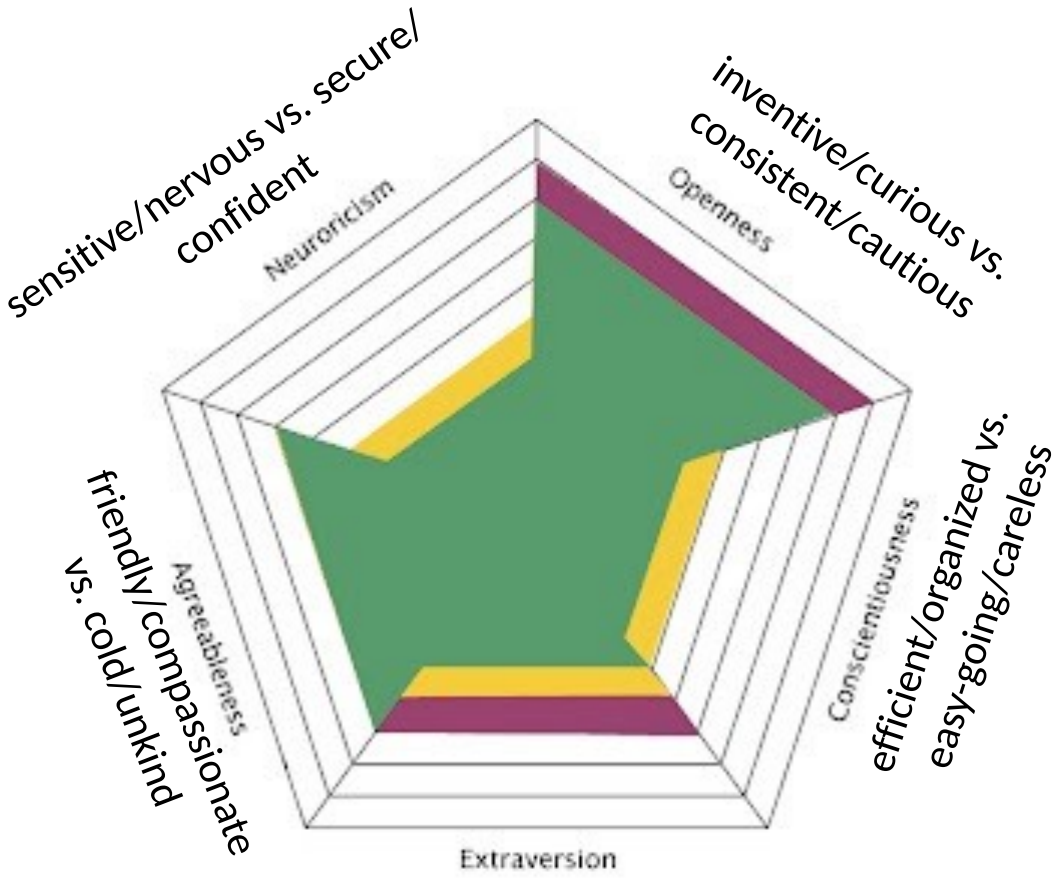
Visual Feeling Detector

'peaceful_country', 'incredible_house', 'wild_coast', 'peaceful_park'

'lonely_grave', 'tiny_spider', 'laughing_face', 'scary_skull',

Deriving Personality

Big5 Personality (OCEAN)



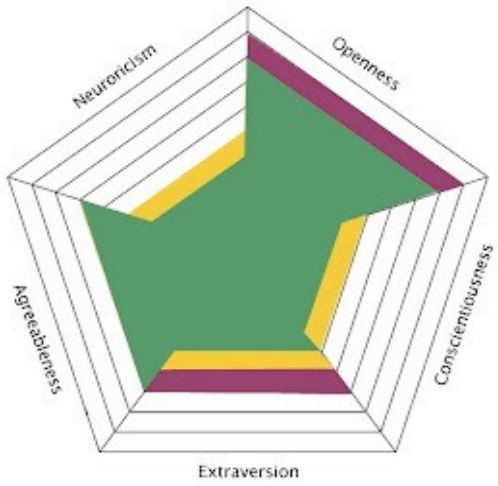
Deriving Personality

– **Mapping text to psycholinguistic category (LIWC) to BIG 5 Personality** [Yakoni '10] : 694 bloggers; 66 LIWC categories; ~2,500 words

– We extended the # of words to about 30,000 by combining with WordNet

LIWC Category	*N*	*E*	*O*	*A*	*C*
Total pronouns	0.06	0.06	-0.21***	0.11**	-0.02
First person sing.	0.12**	0.01	-0.16***	0.05	0
First person plural	-0.07	0.11**	-0.1*	0.18***	0.03
First person	0.1*	0.03	-0.19***	0.08*	0.02
Second person	-0.15***	0.16***	-0.12**	0.08	0
Third person	0.02	0.04	-0.06	0.08	-0.08
Negations	0.11**	-0.05	-0.13**	-0.03	-0.17***
Assent	0.05	0.07	-0.11**	0.02	-0.09*
Articles	-0.11**	-0.04	0.2***	0.03	0.09*
Prepositions	-0.04	-0.04	0.17***	0.07	0.06
			-0.08*	0.11*	0.04
			-0.12**	0.06	-0.06
			-0.15***	0.18***	0.04
			-0.11**	0.14**	-0.02
			0	0.15***	0.16***
			0	-0.15***	-0.18***

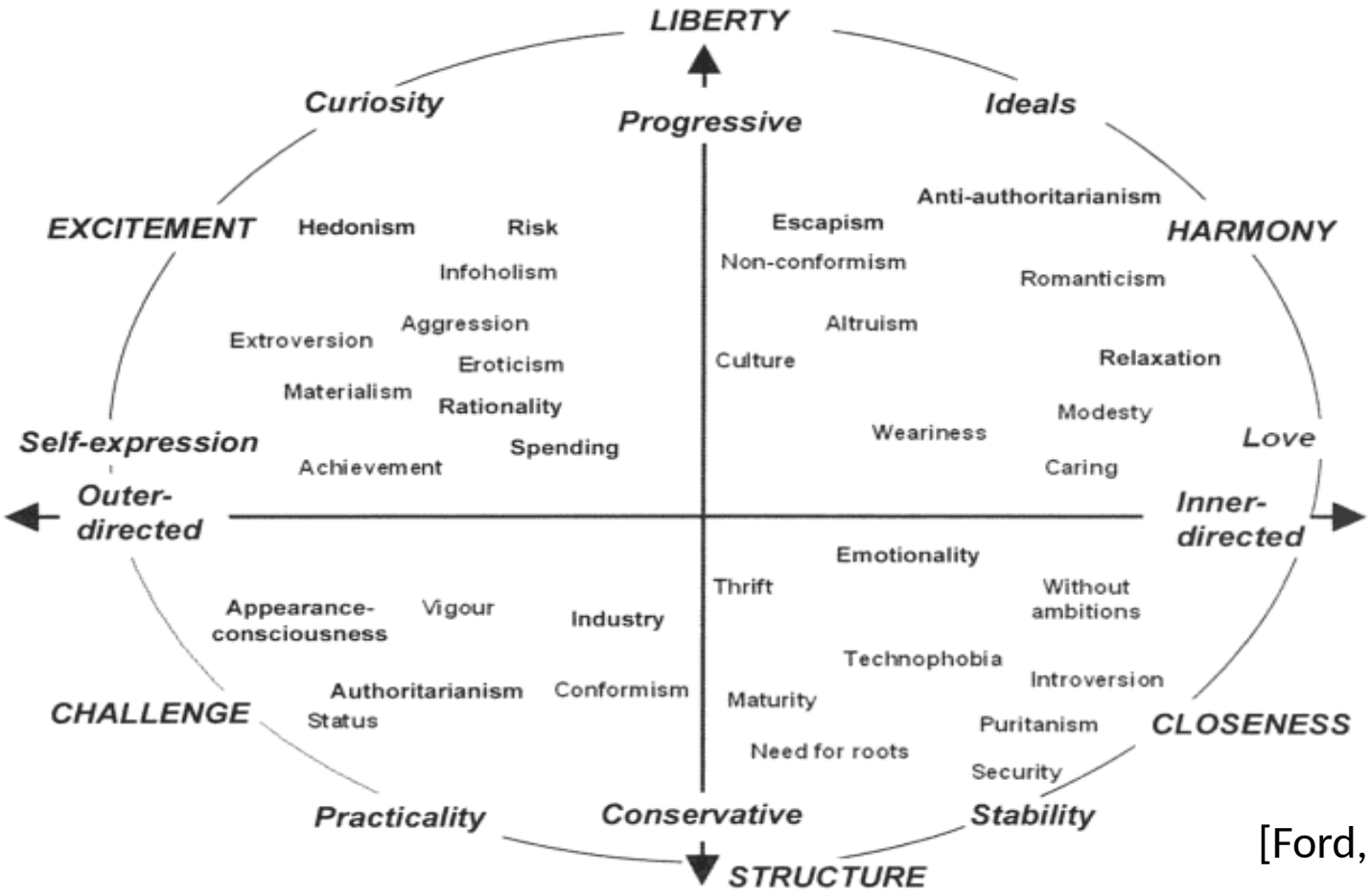
Trait	*No. of words sig. at p <.001*	*Top 20 words*
Neuroticism	24	awful (0.26), though (0.24), lazy (0.24), worse (0.21), depressing (0.21), irony (0.21), road (-0.2), terrible (0.2), Southern (-0.2), stressful (0.19), horrible (0.19), sort (0.19), visited (-0.19), annoying (0.19), ashamed (0.19), ground (-0.19), ban (0.18), oldest (-0.18), invited (-0.18), completed (-0.18)
Extraversion	20	bar (0.23), other (-0.22), drinks (0.21), restaurant (0.21), dancing (0.2), restaurants (0.2), cats (-0.2), grandfather (0.2), Miami (0.2), countless (0.2), drinking (0.19), shots (0.19), computer (-0.19), girls (0.19), gloriou (0.19), minor (-0.19), pool (0.18), crowd (0.18), sang (0.18), grilled (0.18
Openness	393	folk (0.32), humans (0.31), of (0.29), poet (0.29), art (0.29), by (0.28), universe (0.28), poetry (0.28), narrative (0.28), culture (0.28), giveaway 0.28), century (0.28), sexual (0.27), films (0.27), novel (0.27), decades (0.27), ink (0.27), passage (0.27), literature (0.27), blues (0.26)
Agreeableness	110	wonderful (0.28), together (0.26), visiting (0.26), morning (0.26), spring (0.25), porn (-0.25), walked (0.23), beautiful (0.23), staying (0.23), felt (0.23), cost (-0.23), share (0.23), gray (0.22), joy (0.22), afternoon (0.22) day (0.22), moments (0.22), hug (0.22), glad (0.22), fuck (-0.22)
Conscientiousness	13	completed (0.25), adventure (0.22), stupid (-0.22), boring (-0.22), adventures (0.2), desperate (-0.2), enjoying (0.2), saying (-0.2), Hawaii (0.19), utter (-0.19), it's (-0.19), extreme (-0.19), deck (0.18)



Deriving Needs

What do we model

- 12-dimension needs

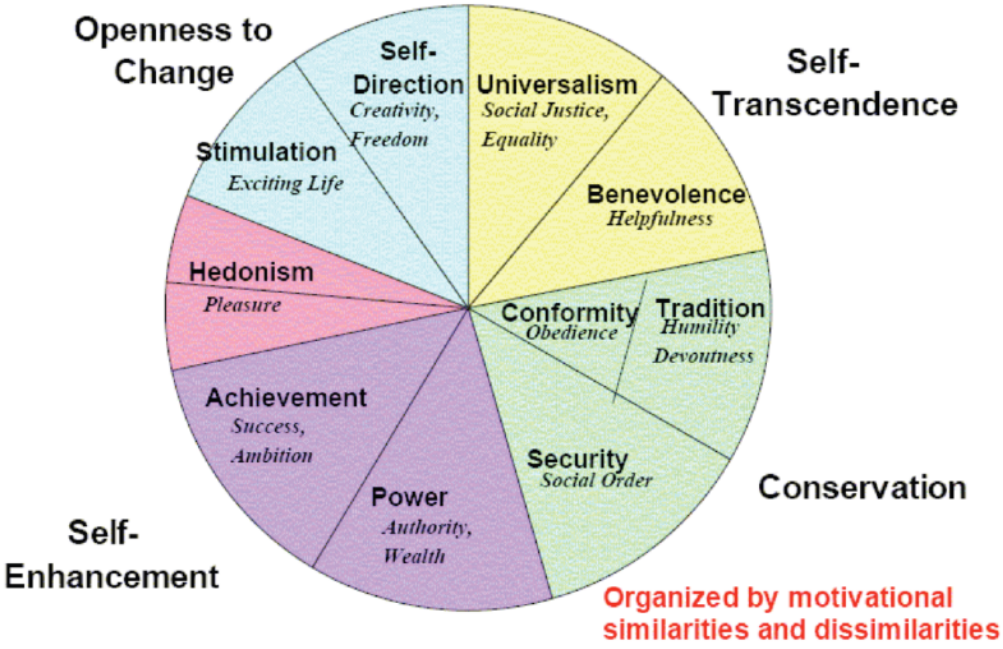


[Ford, 2005]

Deriving Value

- Why model value
  - Values motivate people and guide their actions
  - Values transcend specific actions and situations

[Schwartz 2006]




Event Detection for Hospital Care and Home Elder Care

Multimodality Intelligent Sensor for Homecare

Sensors in the home to detect or prevent accident

S : sensors

Fall down

Stroke

Tension

Sleep monitoring

Multimodality Intelligent Sensor for Homecare

Columbia University

Long-Term Monitoring and Logging of Personal Activities for Chronicle Disease and Health Maintenance

Watching TV

Talking on the phone

Sleep monitoring

S : sensors

Intelligent Multimodality Sensor

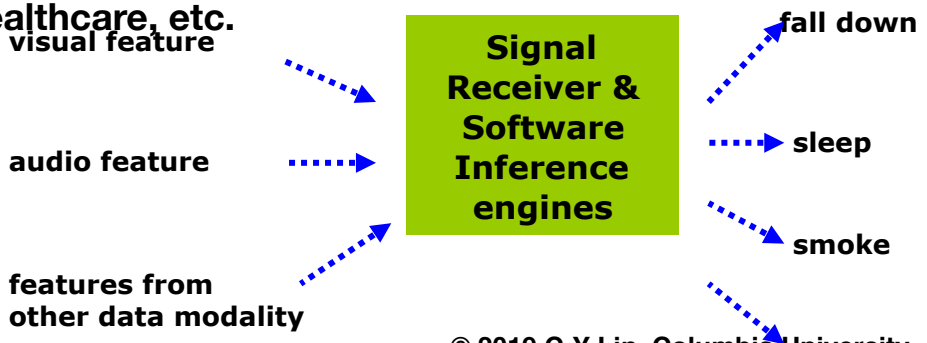
- Commodity:
  - **Fixed Low-Cost Multimodality Recognition Sensor installed under the ceiling or mounted on the wall.**
  - **A Replacement or extension of the smoke detector, intruder detector, baby monitor, etc.**

- Technical Innovation:
  - **Distributed Intelligence:**
    - **Integrated recognition-driven feature extraction module in the sensor**
    - **Only transmitting required features through wireless channel *(e.g., MVs, Color histogram, Sound MFCC coefficients, etc.)***
    - **Separate inference engines for behavior and event recognition.**
  - **Benefits:**
    - **Low-Energy Consumption**
    - **Low Data Transmission Bandwidth**
    - **High Privacy Protection**

Intelligent Multimodality Behavior-Recognition Software Engine

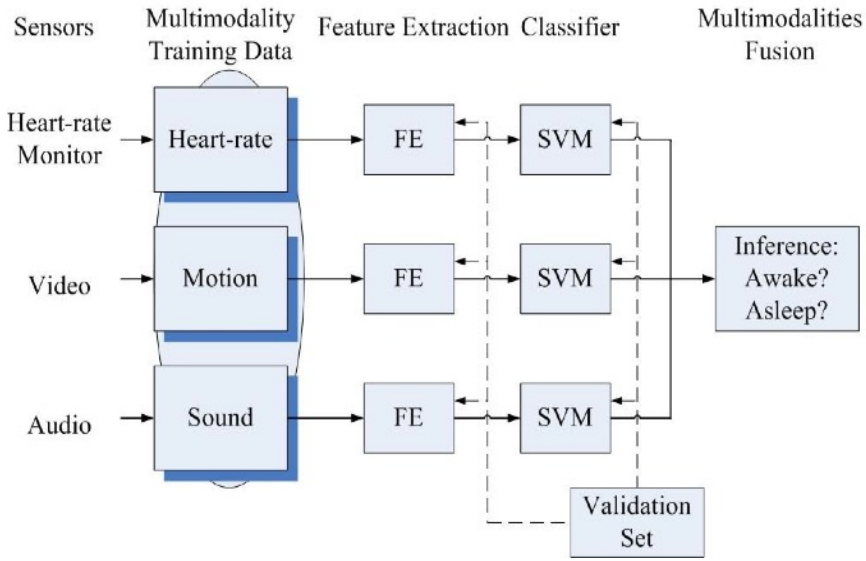
- **Commodity:**
  - **Extensible software applications residing in PC or other computing devices for behavior or event inference.**
  - **Standard-compliant wireless signal receiver for interfacing with the wireless sensor.**

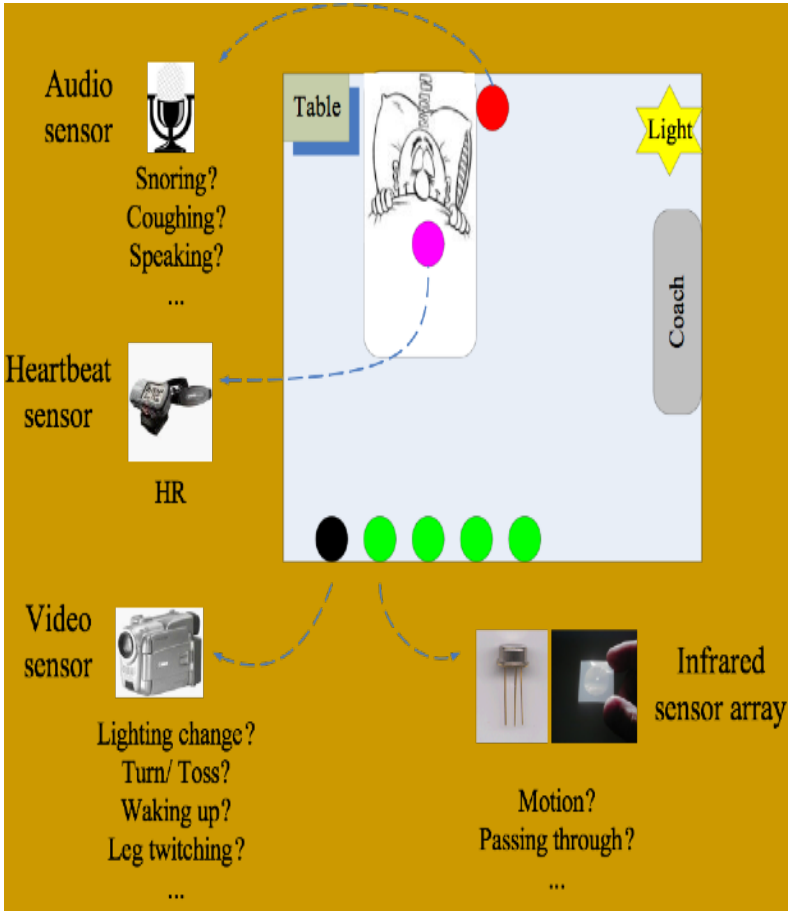
- **Technical Innovation:**
  - **Distributed Intelligence:**
    - **Developing Machine Learning and Data Mining Algorithms for Human Behavior or Environmental Context Recognition**
    - **Recognition is based on the received feature signals**
  - **Benefits:**
    - **Scalability: Consumers can buy any combination of software modules for different applications – surveillance, healthcare, etc.**
    - **Low Maintenance Cost**



Simple Multimodality Sensors for Sleep Situation Inference

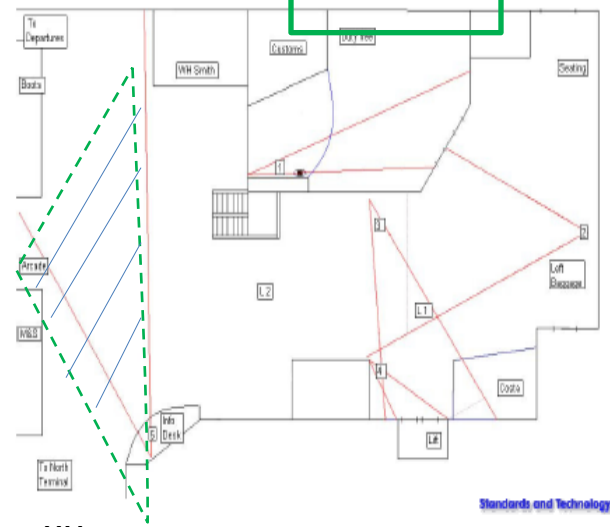
- Understand human night-time activity – *Sleep*
- What we have achieved:
  - Using visual-audio sensors to monitor a person's sleep patterns
  - Measurement of sleep quality

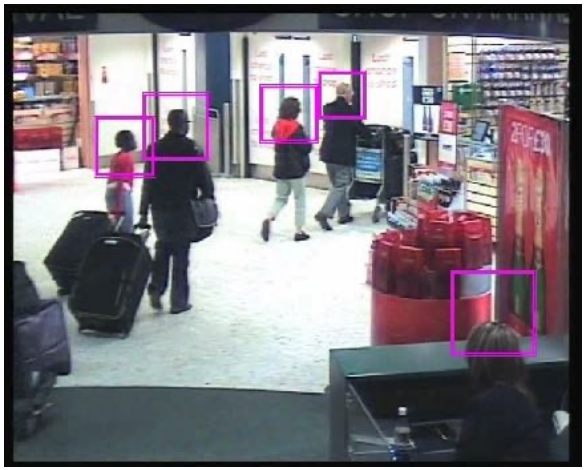


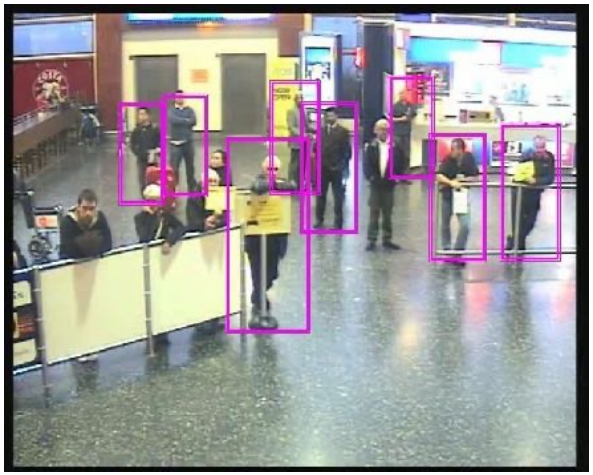


Detecting Events and Tracking People Flow

- Can be applied on Security or Commerce Applications

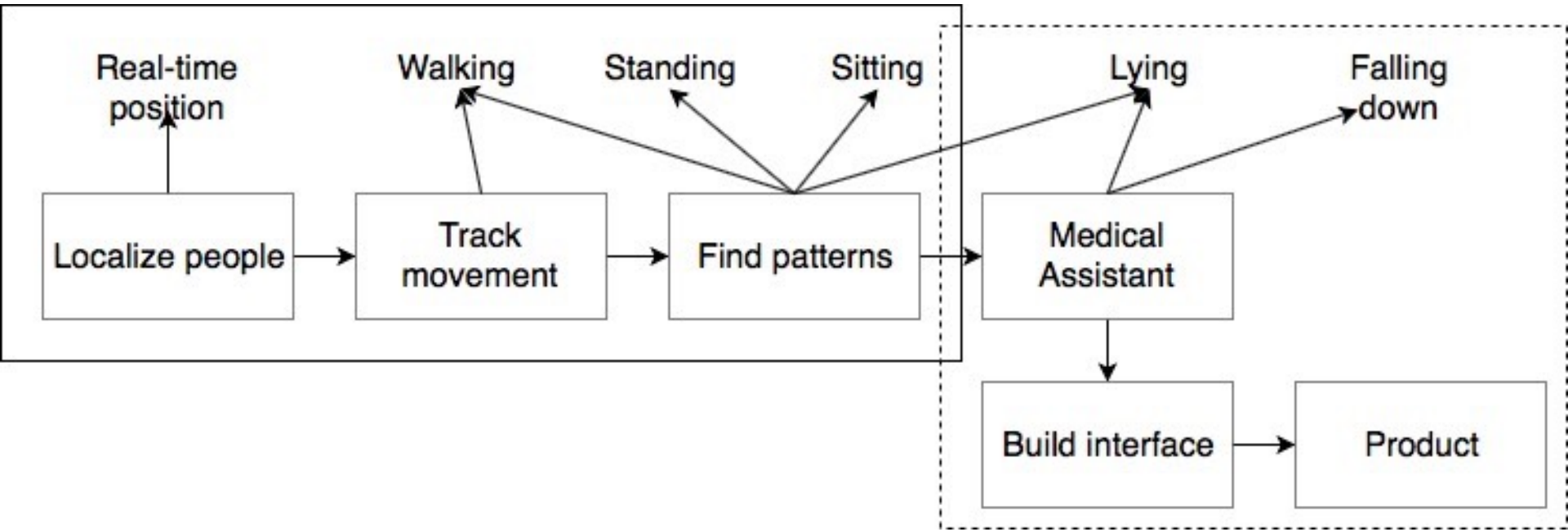




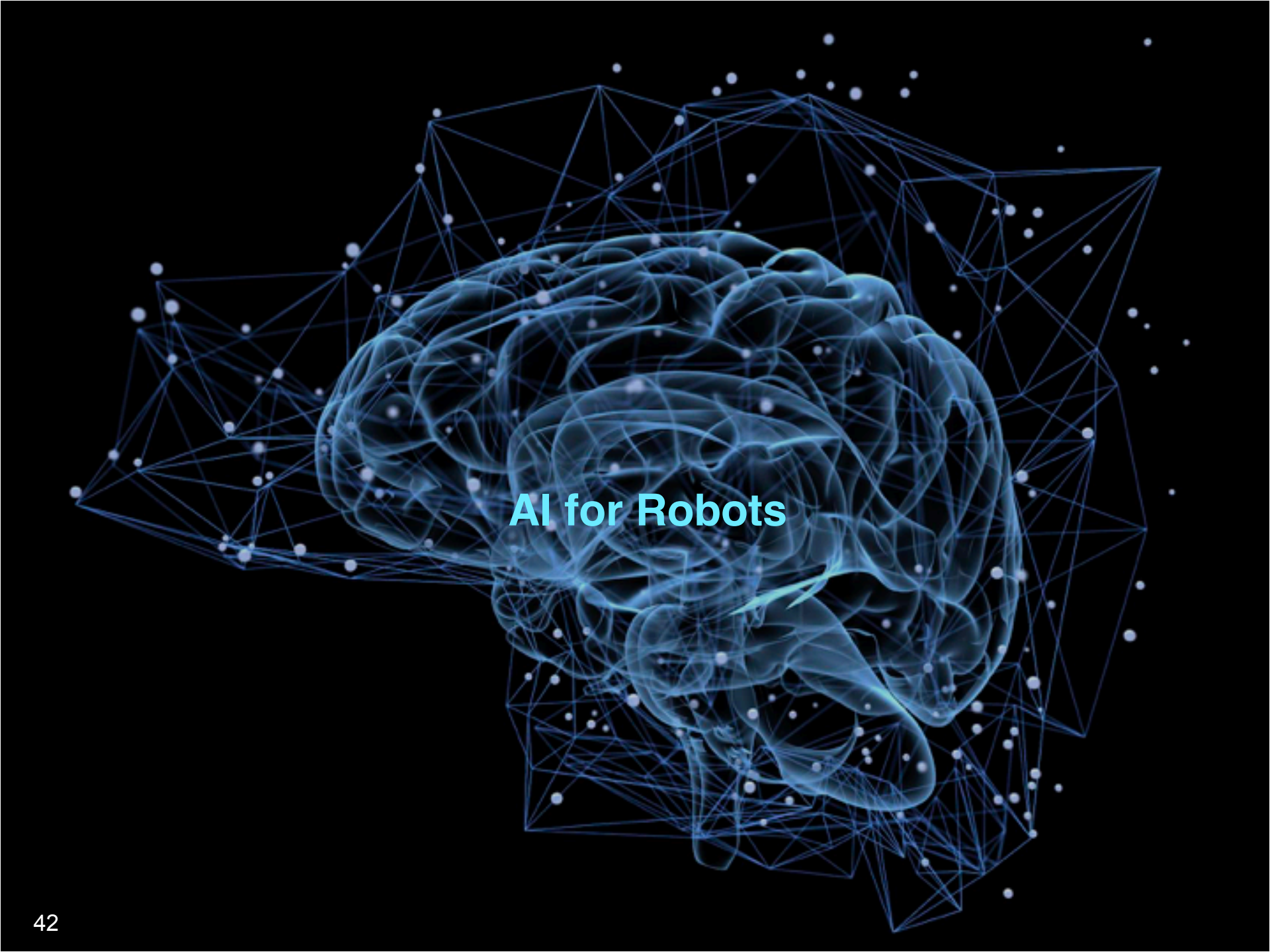


Anomaly Detection of Heart Signals

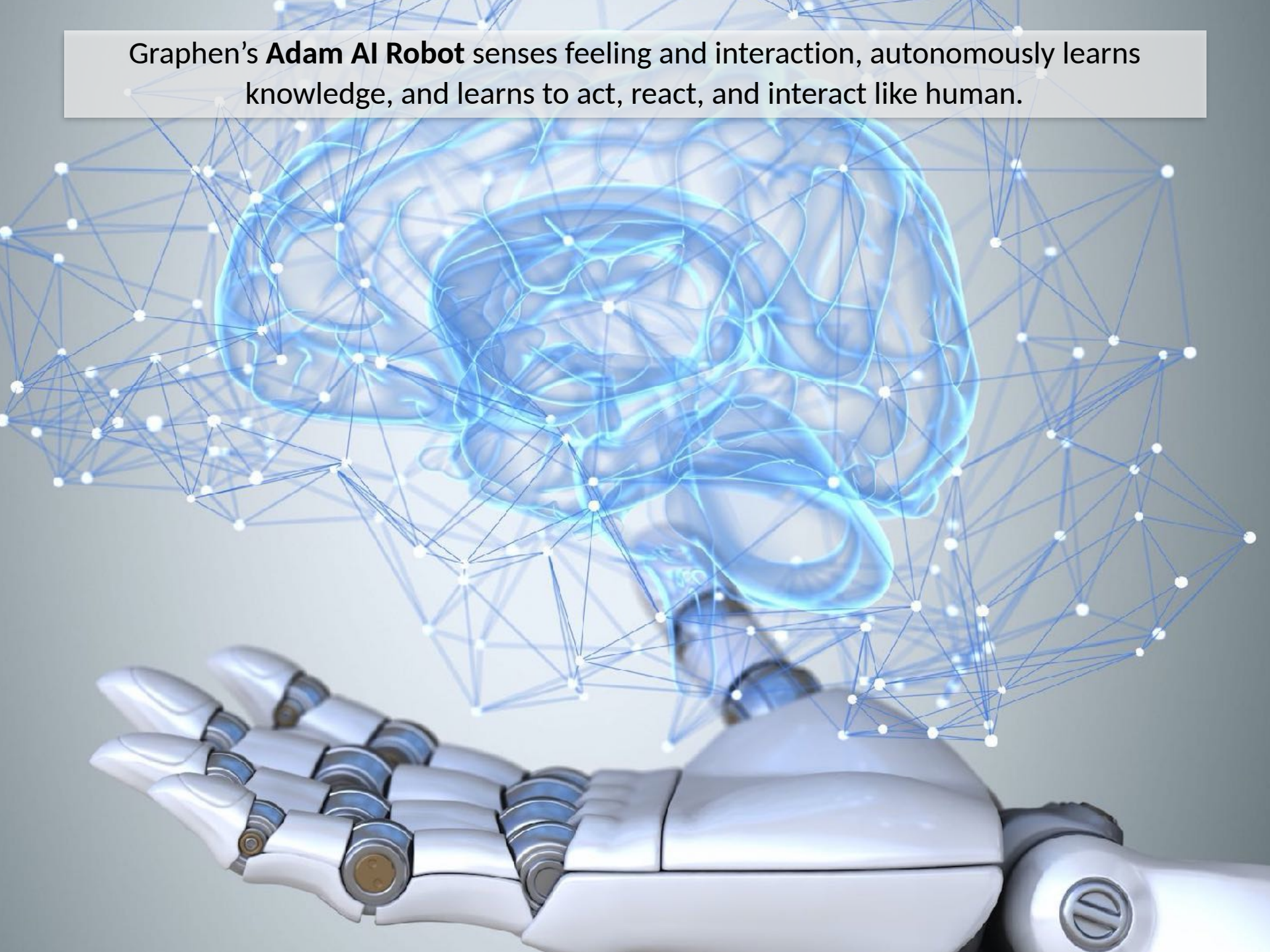
Fisheye Camera Project Pipeline



Real-time Movement Tracking


AI for Robots

Graphen's **Adam AI Robot** senses feeling and interaction, autonomously learns knowledge, and learns to act, react, and interact like human.



Graphen is Making Robots Learn Automatically

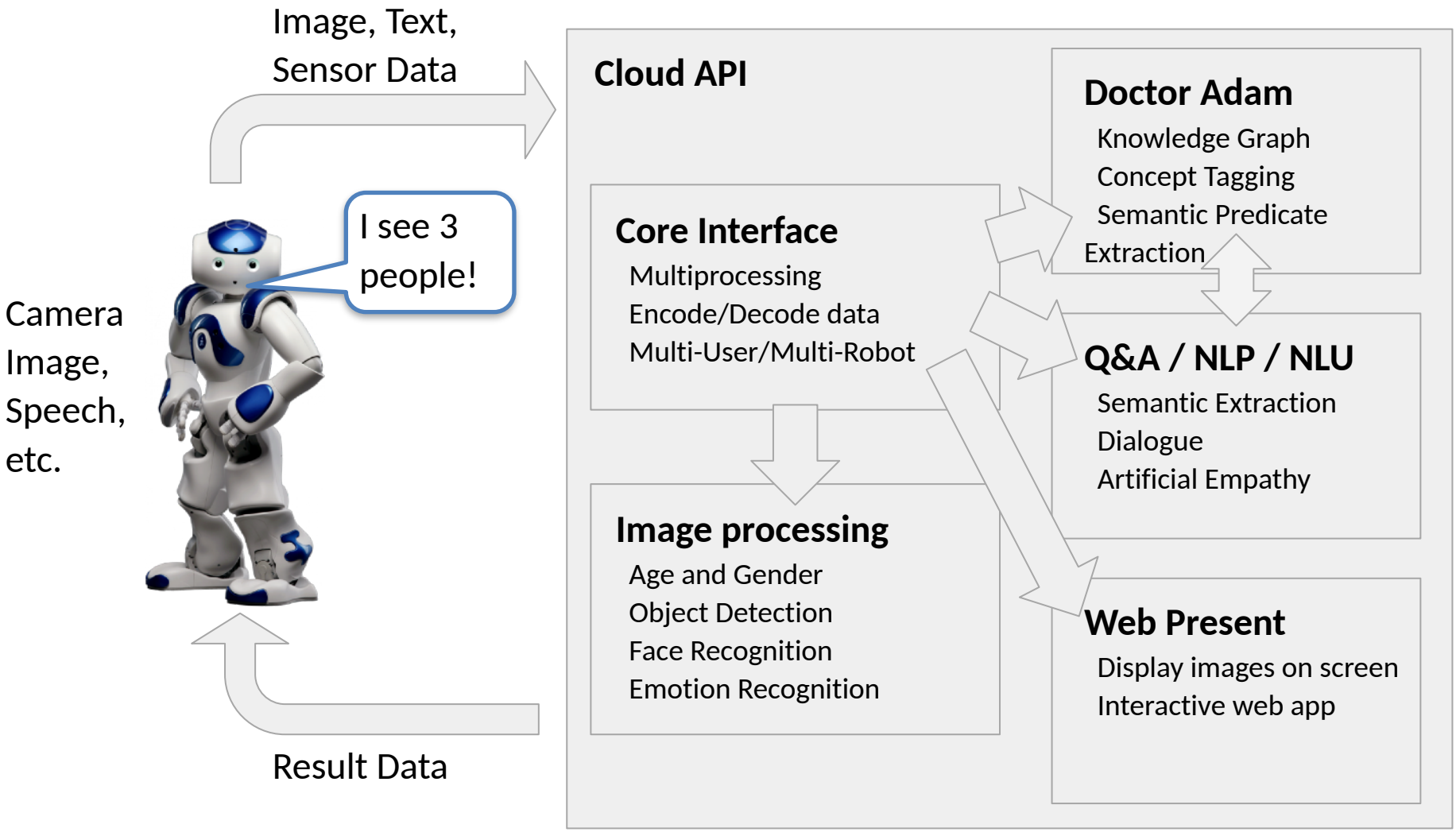
Image Source: http://wonderforgood.com/category/visual-storytelling/

https://www.youtube.com/watch?v=BV8qFeZxZF

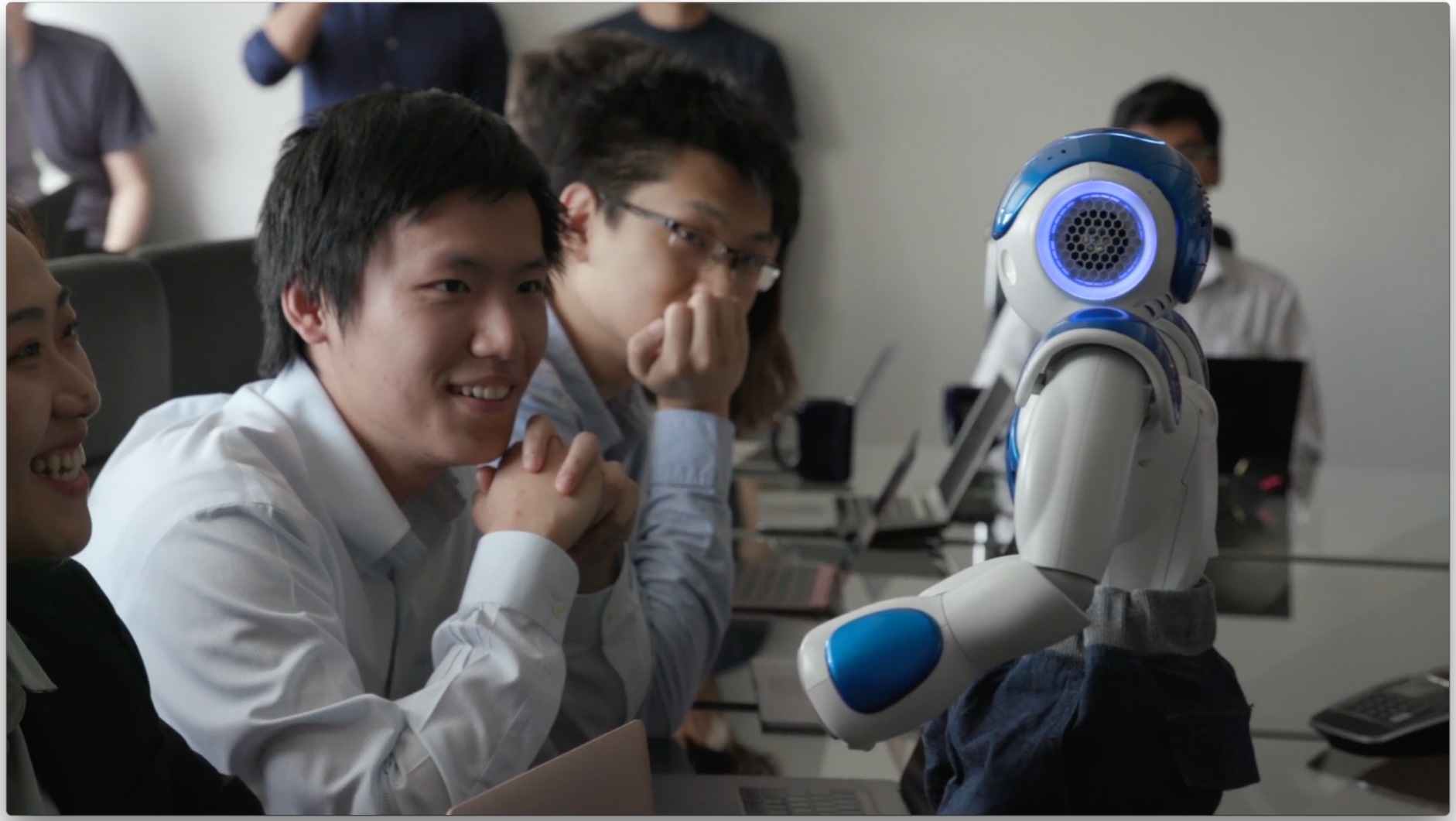
Why Humanoid Robots?

- Interactivity
  - Human-Robot Interaction
  - Conversation
- Semantic/Contextual Understanding
  - Artificial Empathy
  - Environment understanding
- Gesture and Body Language
  - Eye contact
  - Expression through physical body movement
  - Dance and sports
- Playfulness
  - Fun personality and activities
  - Teaching
- Integration to Human-Centric World

Robot Cloud API Services

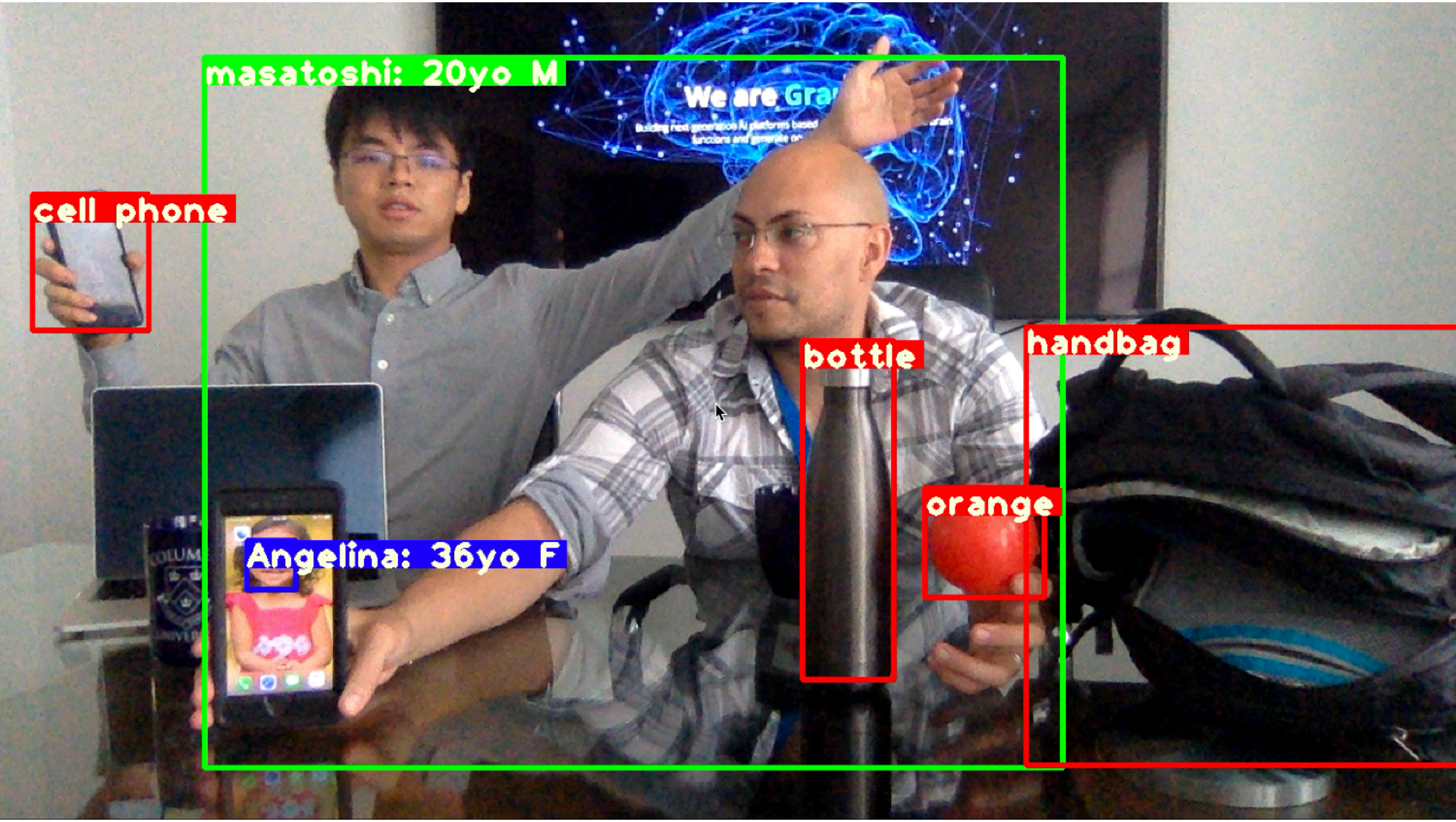


Human-Robot Interaction

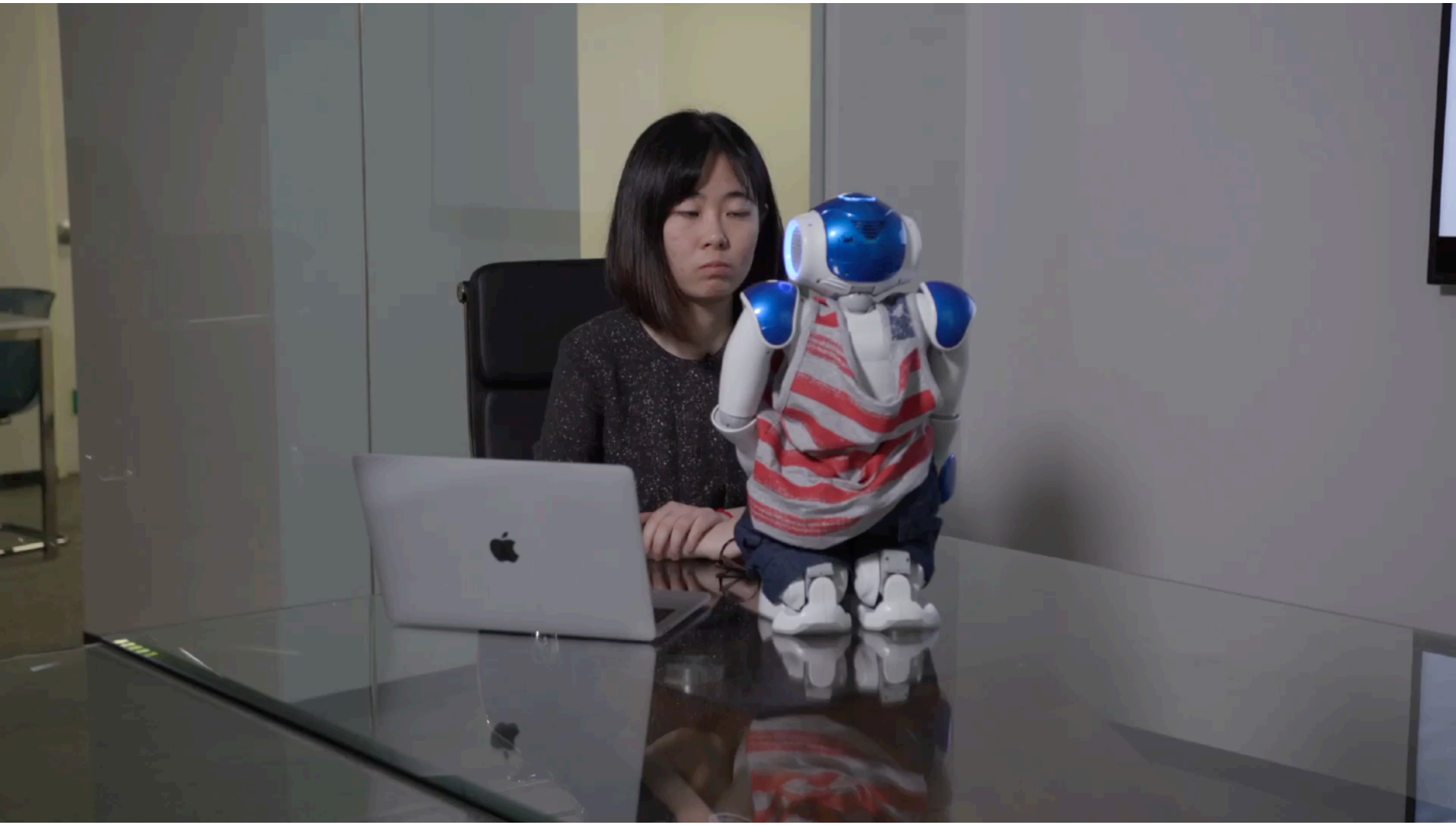


Human-Robot Interaction

What the Robot Sees



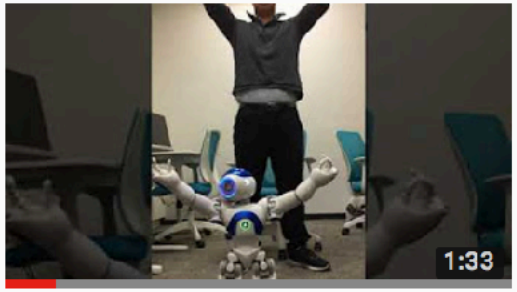
Emotion and Cheers



How Robot cheers you up

Adam Demos on YouTube

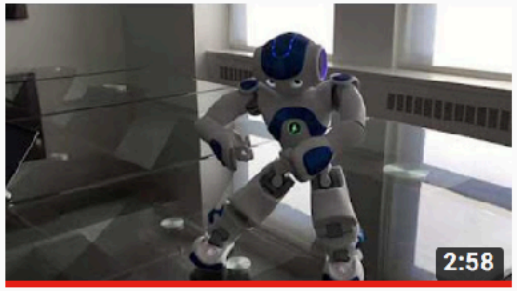
COLUMBIA UNIVERSITY



Graphen Adam Demo – Mimicking Your Actions

Graphen • 89 views • 3 months ago

Check out how **Graphen** cognitive robot **Adam** is mimicking the actions of a user! The robot is learning how to behave based on his



Graphen Adam Demo - Facial Recognition, Response to Commends and Dancing

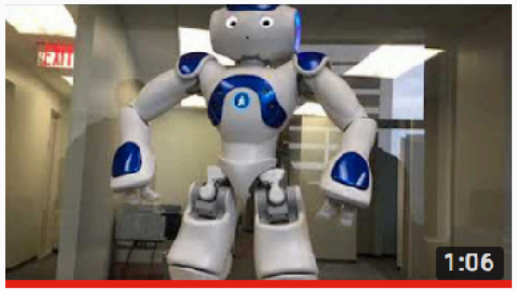
Graphen • 214 views • 3 months ago

AI robot **Adam** is a cognitive robot that is currently under development at **Graphen**. He is built on the NAO robot hardware.

Graphen Adam Demo - Emotion Interactions

Graphen • 122 views • 2 months ago

Graphen develops novel cognitive robot technologies that aim to make robot be more human like. In this demo, we showcase the



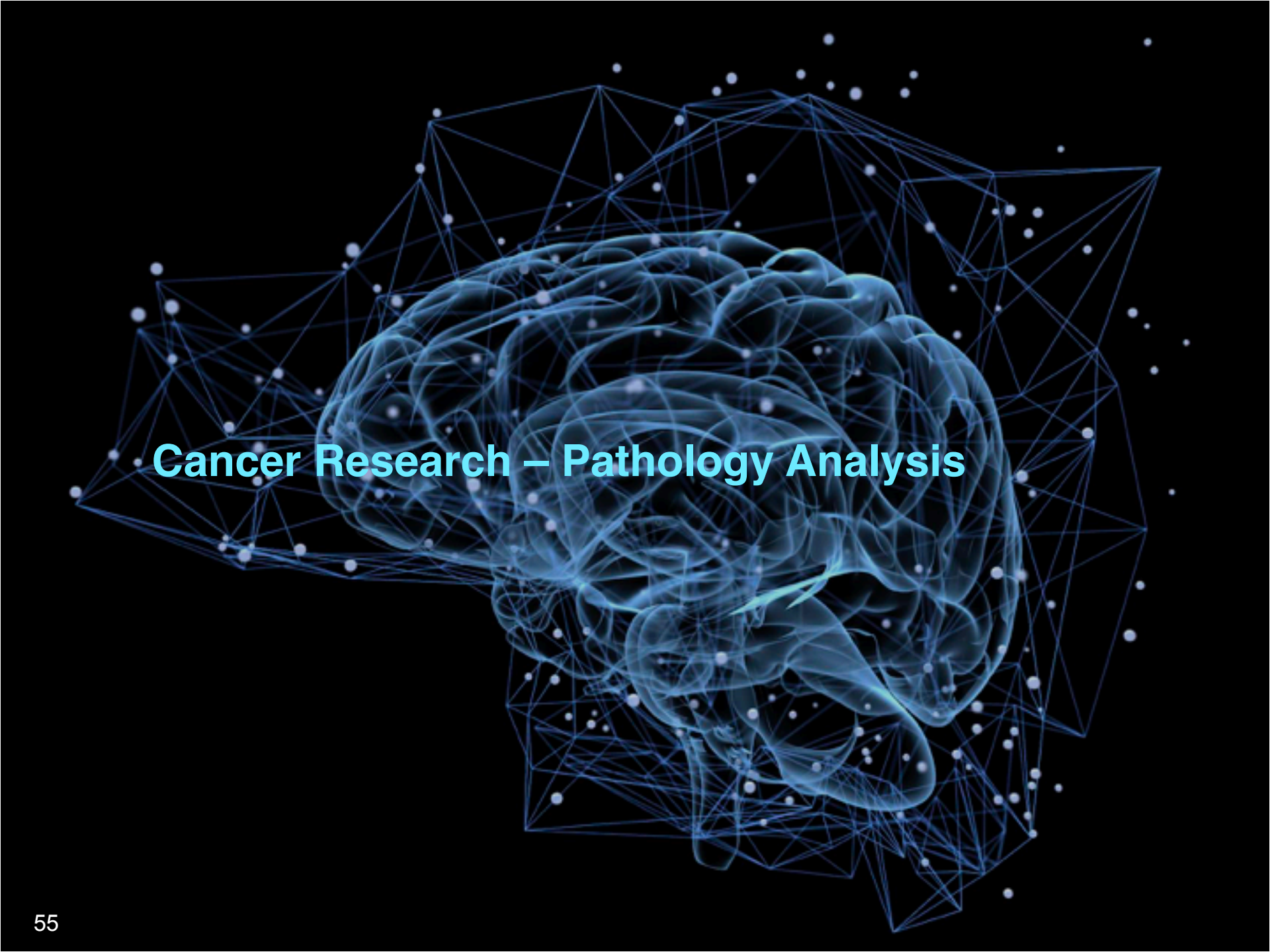
Graphen Robot Adam - Doing Maths

Graphen • 63 views • 2 months ago

Graphen's AI robot **Adam** is doing math calculations. **Adam** is a project that creates human-like intelligence. **Graphen** is focused on

Sample functions of existing "Adam for Healthcare"

1. **Healthcare and Education for Kids**
   - Emotion Recognition & Imitation
   - Mimic Me – Robot Remote Control
   - Face Tracking & Engagement Status
   - Sports
   - Story Telling
2. **Healthcare Provider Intelligence Augmentation**
   - Semantic Predicate Extraction
   - Optical Character Recognition (OCR)
   - Big Data Analytics
3. **Elderly Care**
   - Face Recognition & Object Detection
   - Voice and Visual Emotion Detection
   - Health Information
   - Medication & Appointment Reminder
   - Fall Assessment


Cancer Research – Pathology Analysis

Pathology Result AI System

- Description
  - Collaborative project with the largest and oldest private cancer center in the world in designing an AI system capable of interpreting pathology reports for cancer diagnosis.
- Contribution
  - Besides automating a time consuming and manual process performed by Pathologist, it transforms a highly unstructured pathology report into a structured source of data for further analysis and research.

Definitions

- Pathology
  - Is a branch of medical science that involves the study and diagnosis of disease through the examination of surgically removed organs, tissues, bodily fluids, and in some cases the whole body (autopsy).
- Stain
  - A dye or other coloring material that is used in microscopy to make structures visible.
- Pathology Report
  - A pathology report is a document that contains the diagnosis determined by examining cells and tissues under a microscope.
- Pathologist
  - The medical doctor who conducts the examination of the specimen and writes the pathology report.
- Oncology
  - Is a branch of medicine that deals with the prevention, diagnosis, and treatment of cancer.

Tissue Processing

- The following steps should be completed on the tissue removed from patient before it can be examined under a microscope :
  1. Cut into thin sections, the following 2 methods are used to make the tissue firm enough to cut into thin sections:
     1. frozen sections.
     2. paraffin-embedded (permanent) sections.
  2. Placed on slides.
  3. Stained with dyes.
- This process can take up to 10 days from the moment the biopsy was performed to the Oncologist finally getting the pathology report.

Pathology Report

Columbia University

Hospital Name
Address

Surgical Pathology Report

Patient: Last Name, First Name
MRN: Medical Record Number
DOB: Date of Birth (Age: #)
Gender: M/F

Accession Number: Specimen Identification
Procedure: Date
Attending: Doctor's Name

Clinical History: Large Gastric Mass

Specimen: Gastric Mucosa

Diagnosis

Stomach, Partial Gastrectomy:

- Malignant Epithelioid Gastrointestinal Stromal Tumor
- Tumor Size 10 x 9 x 8 cm
- Cell Type: Epithelioid and Spindled
- High cellularity; present
- Mucosal Invasion: Focally present adjacent to ulceration
- Mucosal ulceration present
- Mitotic Count: 10/50 HPF
- Myxoid background: Focally present
- Foci of necrosis present
- CD117, vimentin, and CD34: uniformly positive

Gross Description

The specimen consists of an approximately 5 x 7 cm portion of gastric mucosa that is surrounded and underlying by a lobulated mass which is 10 x 9 x 8 cm. The central portion of the mass appears to have an approximately 1.5-cm ulcer. The mucosa away from the area of ulceration is partially removed from the underlying tumor. The underlying mass appears encapsulated and lobular. Gross sections show the lesion to consist of several different patterns. A single area has a gray to gray-tan pattern with an area of central necrosis showing a fairly uniform appearance whereas; other regions of the tumor are gray white- and somewhat lobular in appearance. Areas of yellow necrosis are scattered through the tumor. Representative portions submitted.

Microscopic Description

Sections through the neoplasm show it to be primarily a high cellular neoplasm. The cells are in part arranged in fascicles and clusters with enlarged elongate nuclei having relatively find nucleoli. In some areas, the fascicles have an interwoven appearance. Mitotic figure up to 10:50 HPF. A few areas show foci of necrosis with the cells appearing to be surrounded by somewhat myxoid stroma. Foci of displayed necrosis are present. The lesions appear circumscribed, although not specifically encapsulated. It focally involved the mucosa and shows full thickness ulceration. The tumor immediately beneath the mucosal area of ulceration has a nearly lobular somewhat spindled growth pattern. Some areas of the tumor have a slightly more rounded nuclei and somewhat epithelioid appearance. The cells appear to be arranged in groups and clusters. Some of the cells have cyptoplasmic vacuoles. These areas also show a prominent mitotic activity. Some mitotic figured are abnormal and atypical. The tumor contains numerous relatively open vascular channels which appear to be part of the neoplasm. The tumor has a pseudo capsule and in some areas appear to be nearly covered.

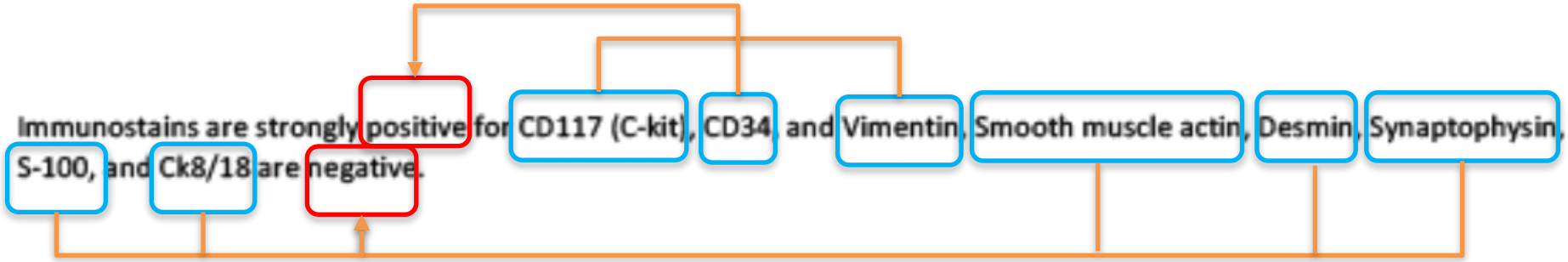
Immunostains are strongly positive for CD117 (C-kit), CD34, and Vimentin, Smooth muscle actin, Desmin, Synaptophysin, S-100, and Ck8/18 are negative.

Comment

Immunostains were performed on the core biopsy and demonstrate that the tumor cells are positive for CD117. The findings are consistent with the above diagnosis.

Objectives

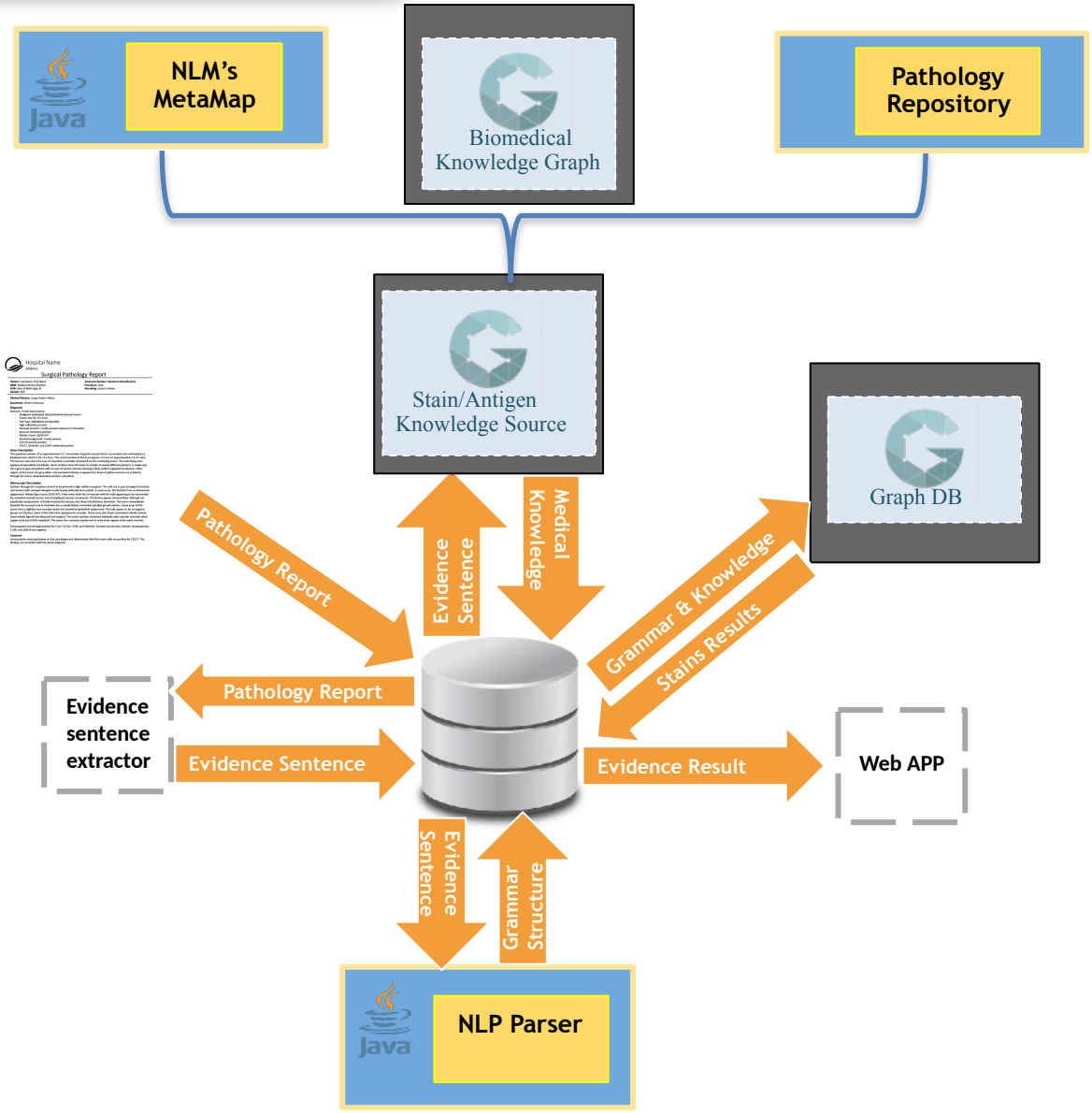
1. Extract evidence sentence from Pathology Result.
2. Identify, within the evidence sentence, the stains or antigens and their respective result, strength, percentage of cells and location.



Immunohistochemistry for PD-L1 (Cell Signaling, Clone E1L3N) was performed with the following results: 30% of tumor cells show weak membranous labeling

Output: (PD-L1, Positive, Weak, 30%, Membranous)

System Infrastructure



Preliminary Results

Immunohistochemical stain for TTF1 is positive, consistent with a pulmonary adenocarcinoma. The tumor cells are negative for synaptophysin, chromogranin, CD56 and ASCL1, and show Ki-67 proliferative index of approximately 70%.

	Phrase	Original Phrase	CUI	Preferred Name	TUI	Semantic Type	Result
0	TTF-1	TTF1	C1610981	TTF1 protein, human	T116	Amino Acid, Peptide, or Protein	positive
1	ASCL1	ASCL1	C0083344	ASCL1 protein, human	T116	Amino Acid, Peptide, or Protein	negative
2	CHROMOGRANIN-A	chromogranin	C0055633	Chromogranin A	T116	Amino Acid, Peptide, or Protein	negative
3	CD56	CD56	C0282587	Neural Cell Adhesion Molecules	T116	Amino Acid, Peptide, or Protein	negative
4	SYNAPTOPHYSIN	synaptophysin	C0085255	Synaptophysin	T116	Amino Acid, Peptide, or Protein	negative
5	Ki-67	Ki-67	C0208804	Ki-67 Antigen	T116	Amino Acid, Peptide, or Protein	negative

Immunohistochemical stains reveal that the tumor cells are positive for TTF-1 while negative for CDX-2 and CK20.

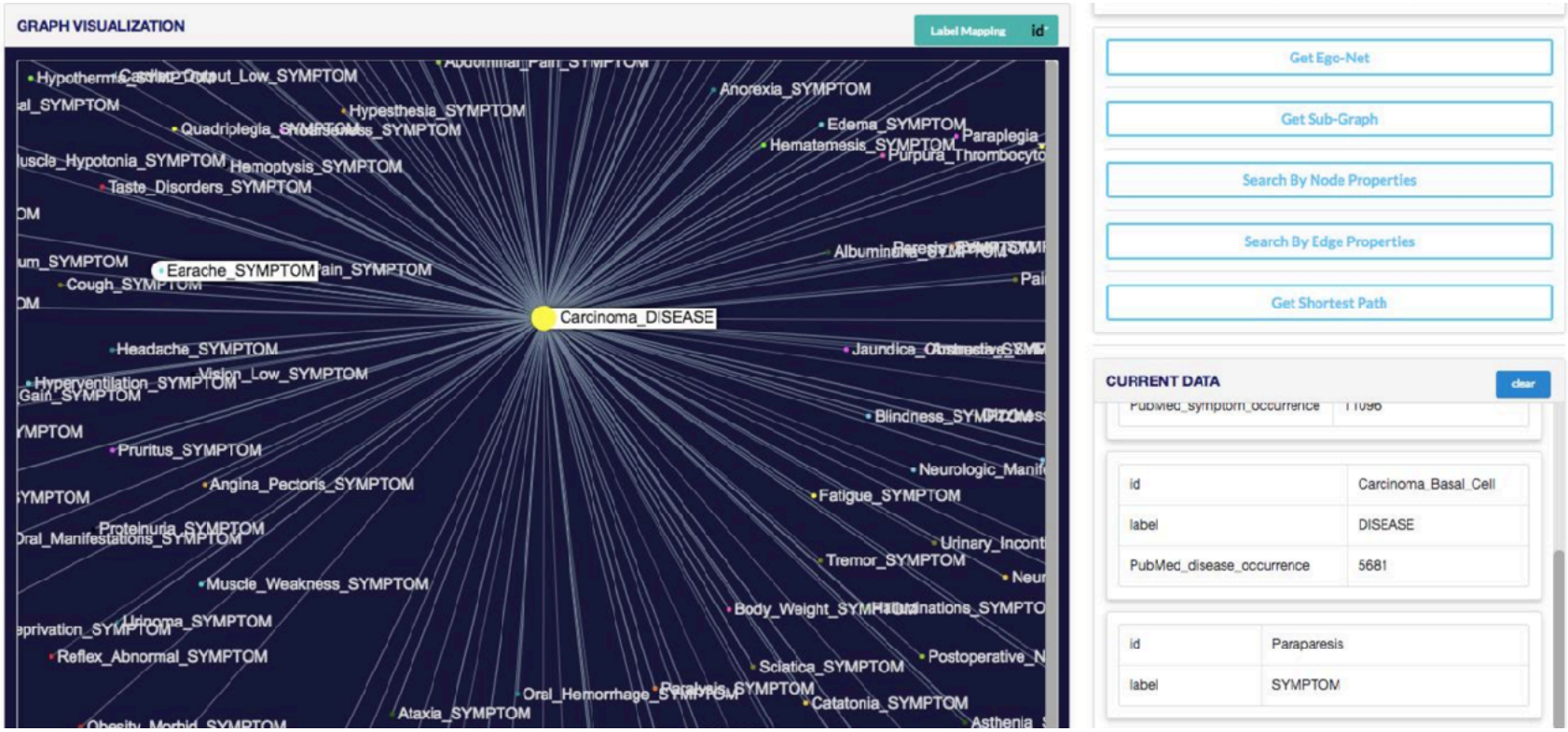
	Phrase	Original Phrase	CUI	Preferred Name	TUI	Semantic Type	Result
0	CDX2	CDX-2	C1505661	CDX2 Protein, human	T116	Amino Acid, Peptide, or Protein	negative
1	CK20	CK20	C0219510	cytokeratin 20	T116	Amino Acid, Peptide, or Protein	negative
2	TTF-1	TTF-1	C1610981	TTF1 protein, human	T116	Amino Acid, Peptide, or Protein	positive

By immunostains, the tumor was positive for CK7 and negative for TTF1, CK20, p40, PAX8, SF1 and Melan A

	Phrase	Original Phrase	CUI	Preferred Name	TUI	Semantic Type	Result
0	CK20	CK20	C0219510	cytokeratin 20	T116	Amino Acid, Peptide, or Protein	negative
1	TTF-1	TTF1	C1610981	TTF1 protein, human	T116	Amino Acid, Peptide, or Protein	negative
2	CK7	CK7	C3272783	Keratin, Type II Cytoskeletal 7	T116	Amino Acid, Peptide, or Protein	positive
3	PAX-8	PAX8	C3273568	Paired Box Protein Pax-8	T116	Amino Acid, Peptide, or Protein	negative
4	MELAN-A	Melan	C1522481	MART-1 Antigen	T116	Amino Acid, Peptide, or Protein	negative
5	SF-1	SF1	C0171961	Steroidogenic Factor 1	T116	Amino Acid, Peptide, or Protein	negative
6	P40	p40	C4079844	deltaNp63-p40 protein	T116	Amino Acid, Peptide, or Protein	negative

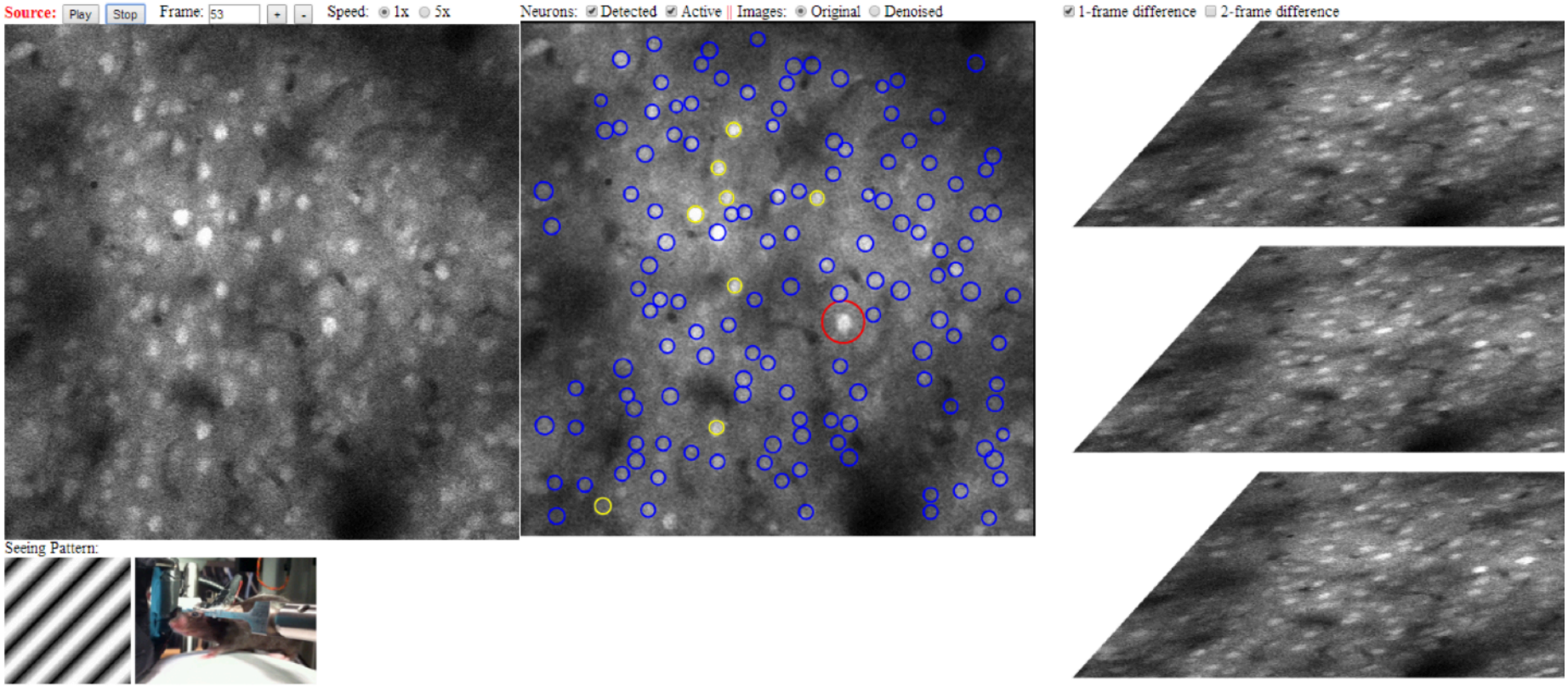
Cancer Pathology Knowledge Mining

Disease-Symptom network

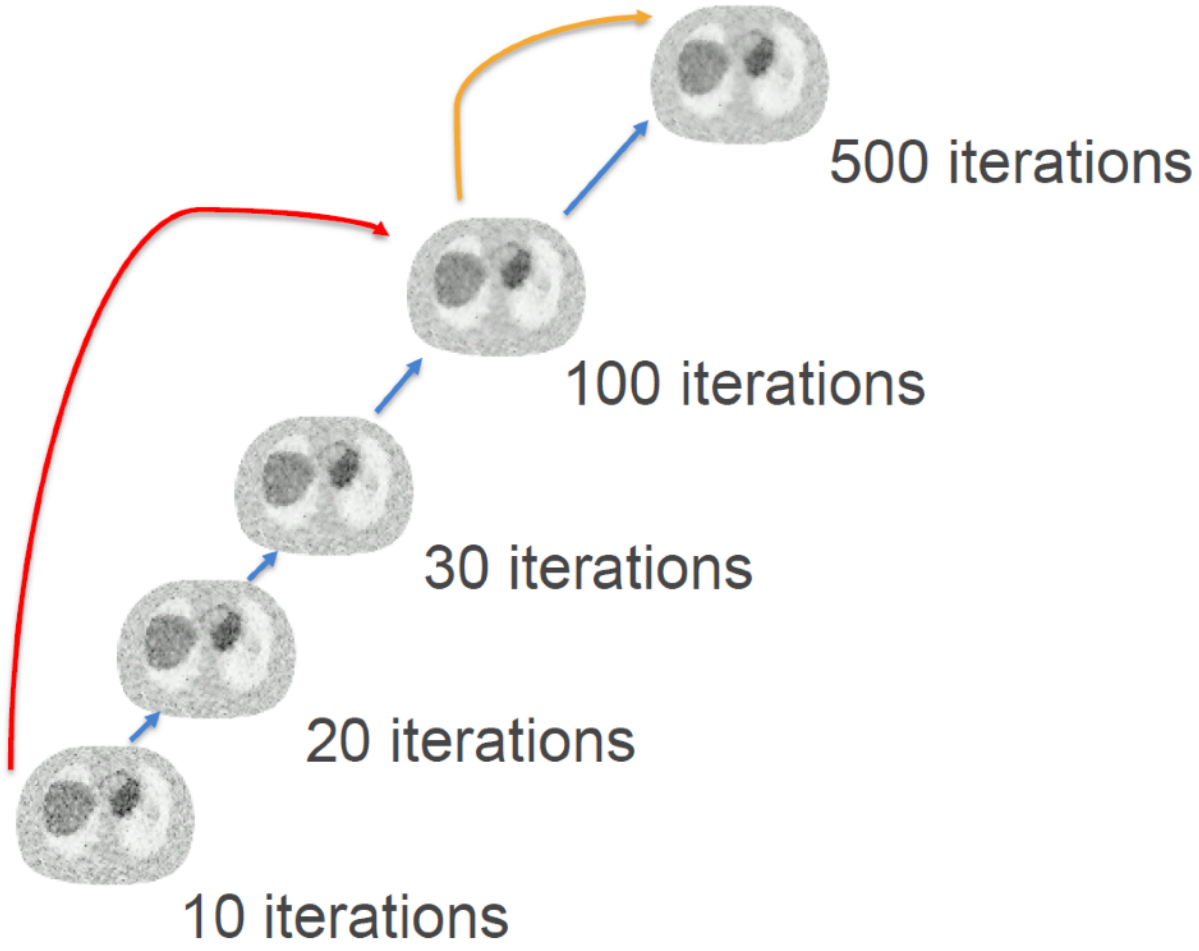



Deep Learning and other AI
techniques for Medical Imaging

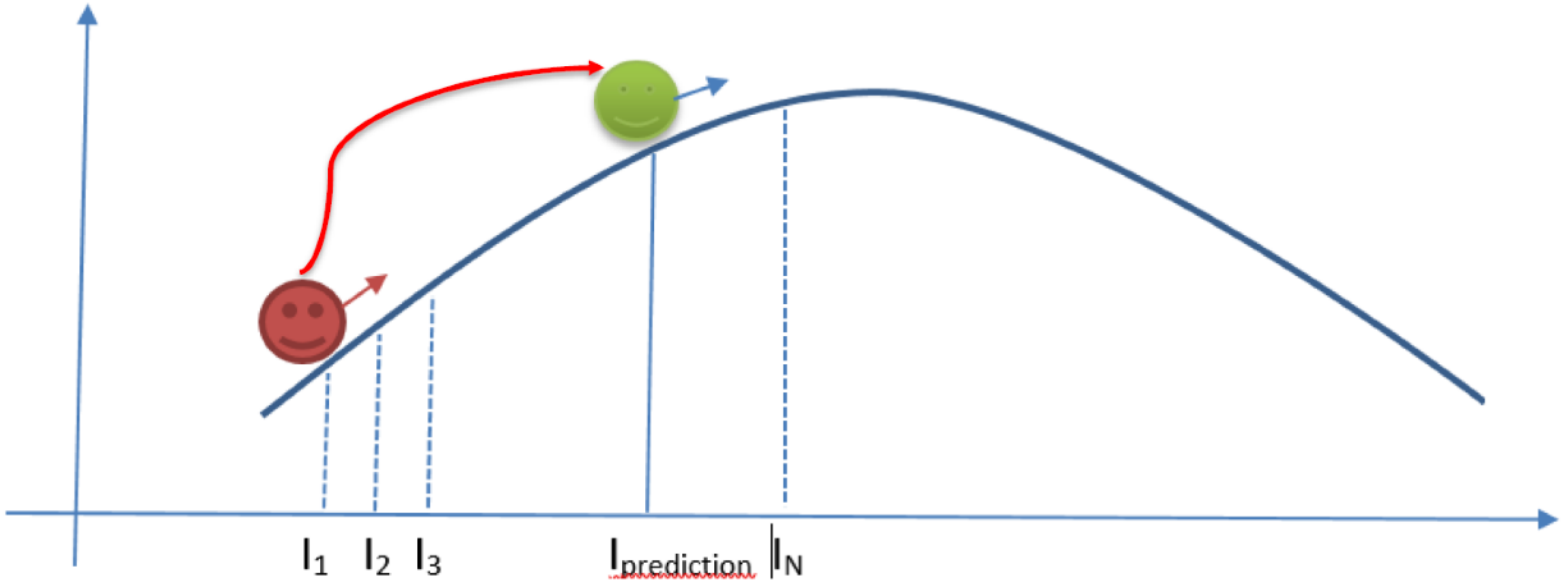
Understand and Reconstruct Brain Network

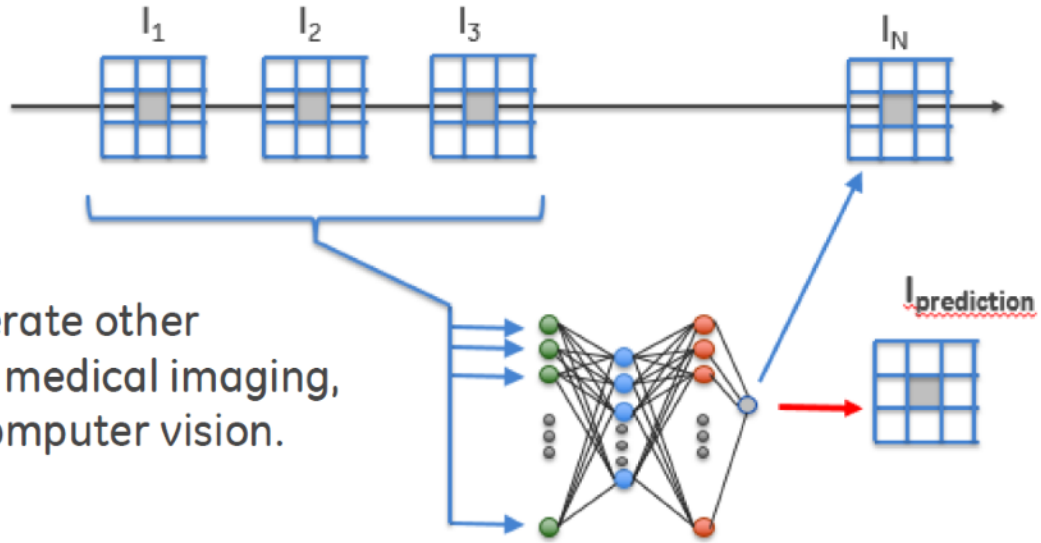


Deep Learning for Optimization Problems in Imaging



Methods: Predicting convergence curve



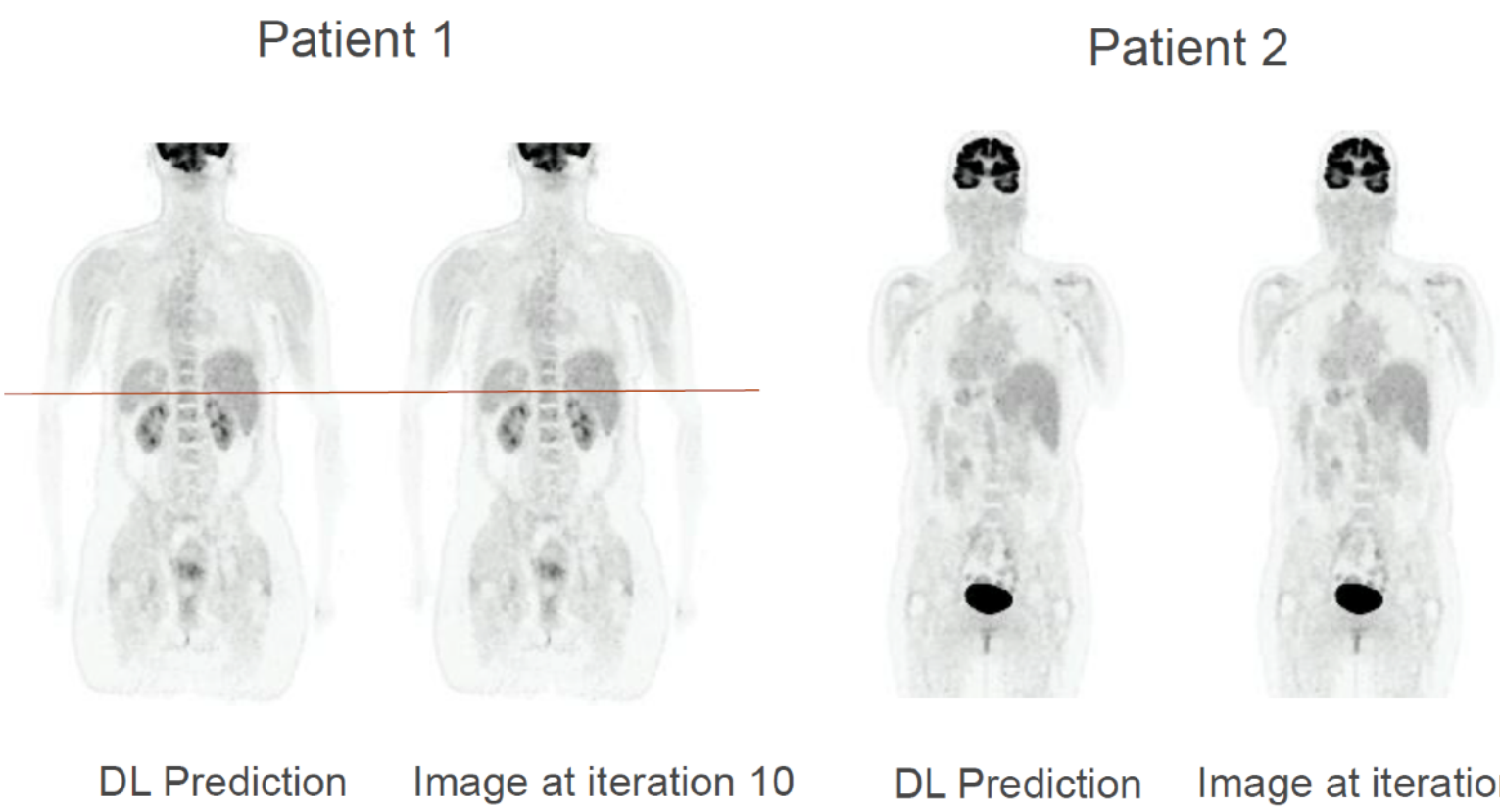


Idea applicable to accelerate other optimization problem in medical imaging, image processing and computer vision.

Cheng et.al., Fully 3D 2017

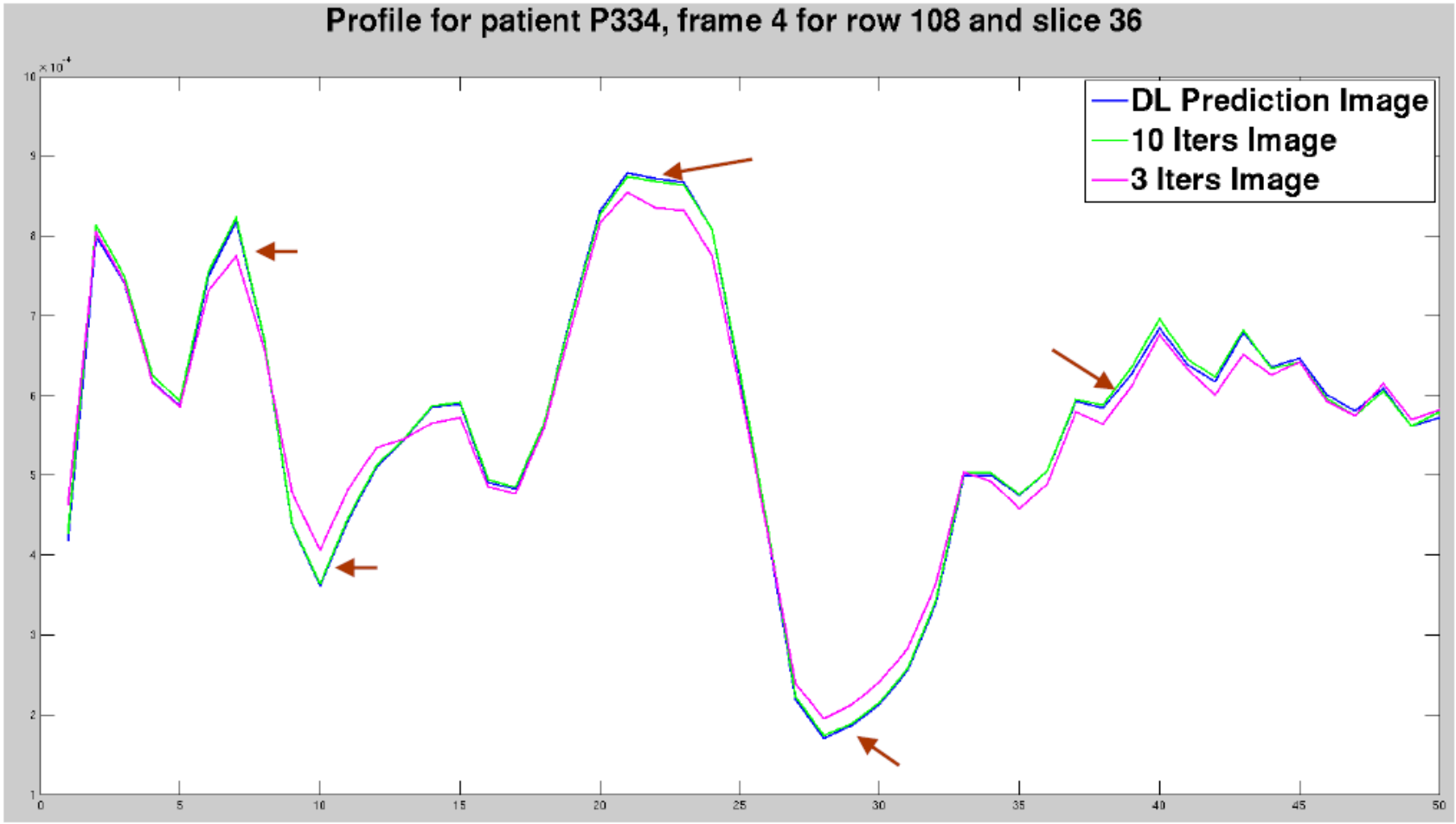
Deep Learning Results

Visual Image Results



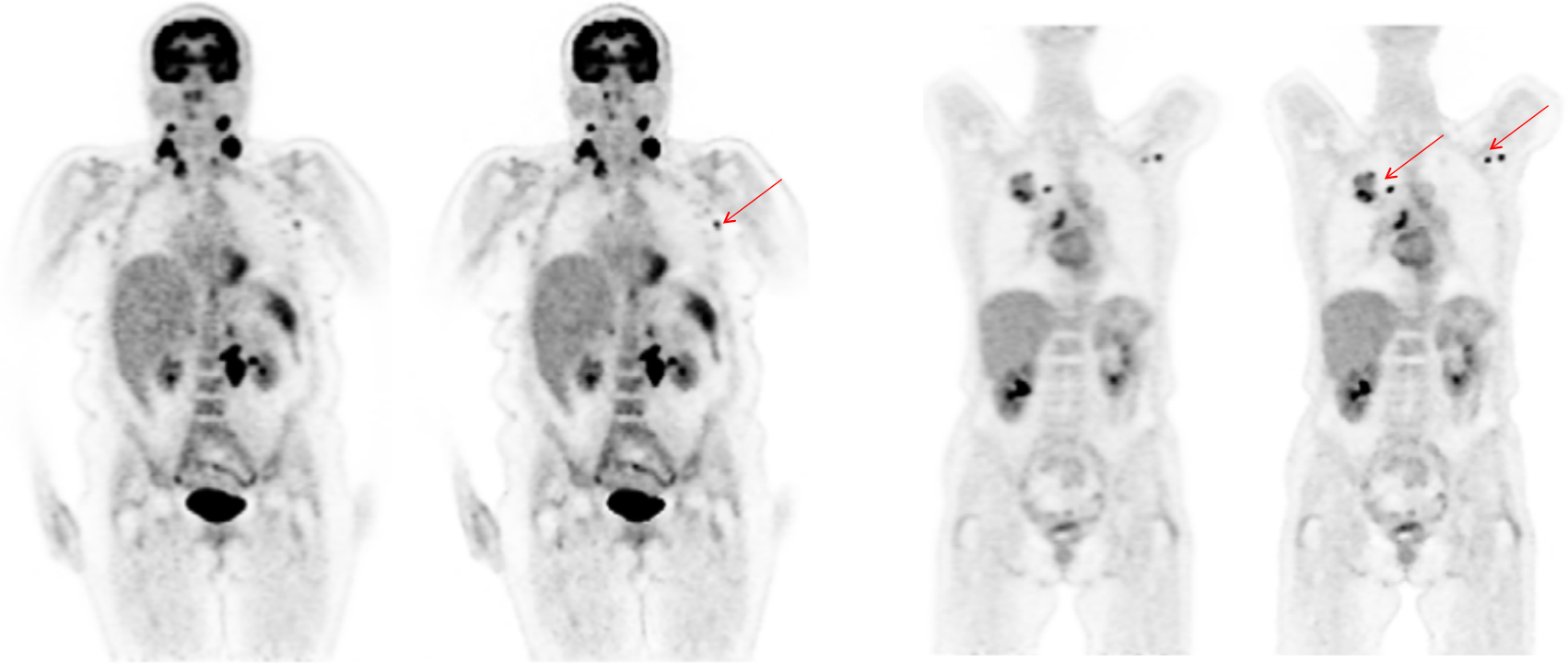
Cheng et.al., Fully 3D 2017

Deep Learning Results: Quantitative Comparison



Cheng et.al., Fully 3D 2017

Better tumor detection: PET/CT



Cheng, et.al., PMB 2015

Image Registration

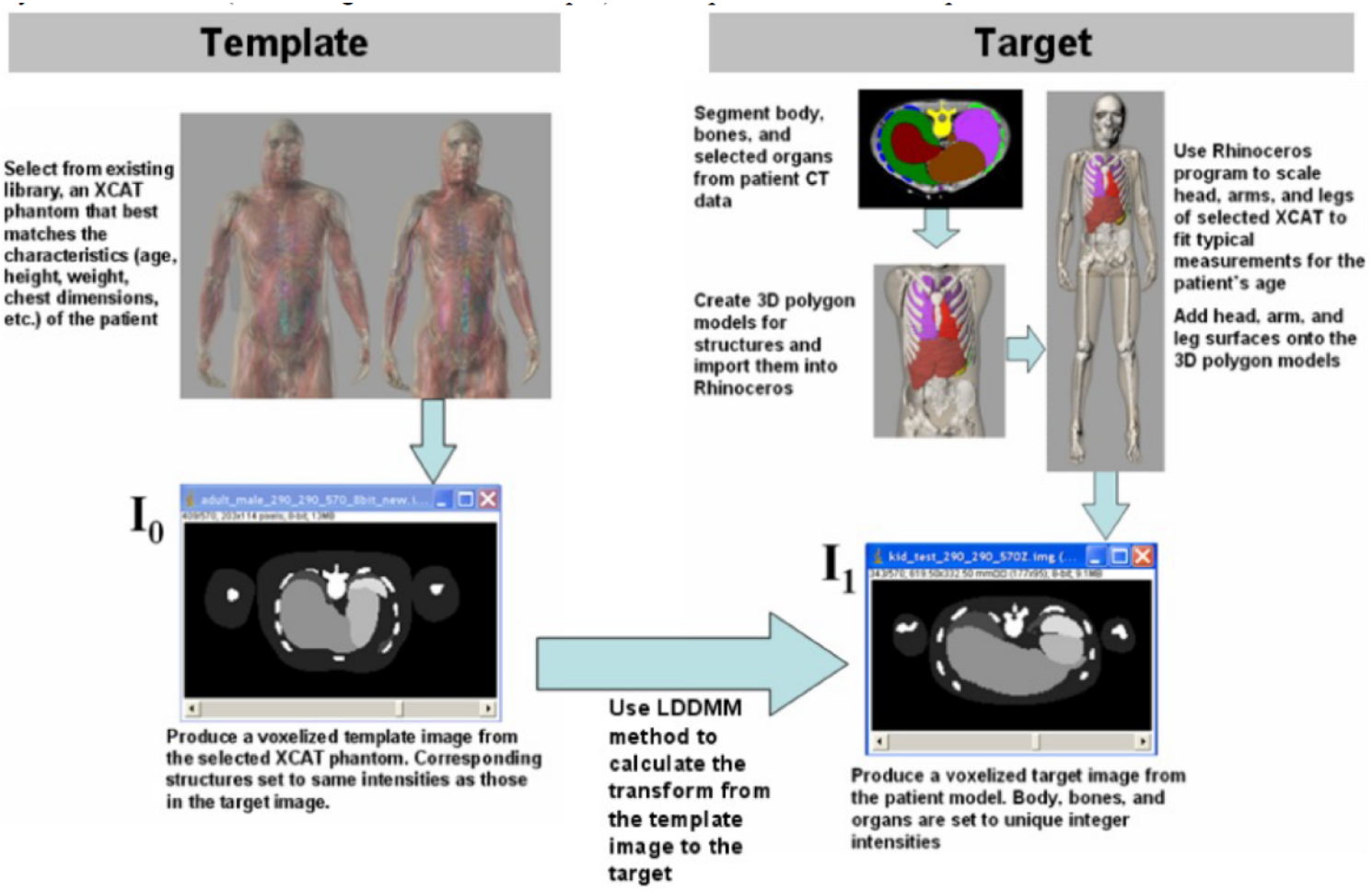


Fig. 3. Procedure for calculating the LDDMM transform. To calculate the LDDMM transform, template and target images are required. The above steps are performed to create these images.

Cheng, et.al., SPIE 2009

Image Registration(cont'd)

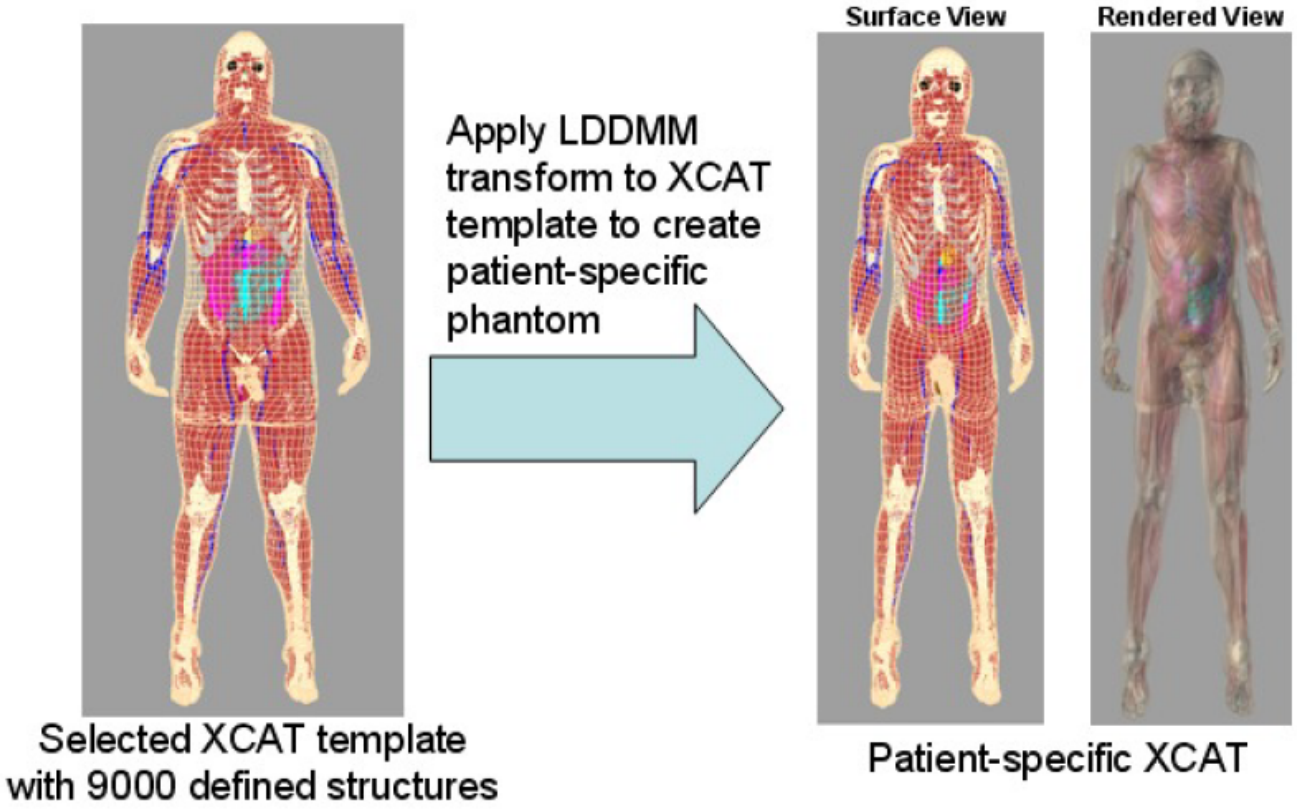
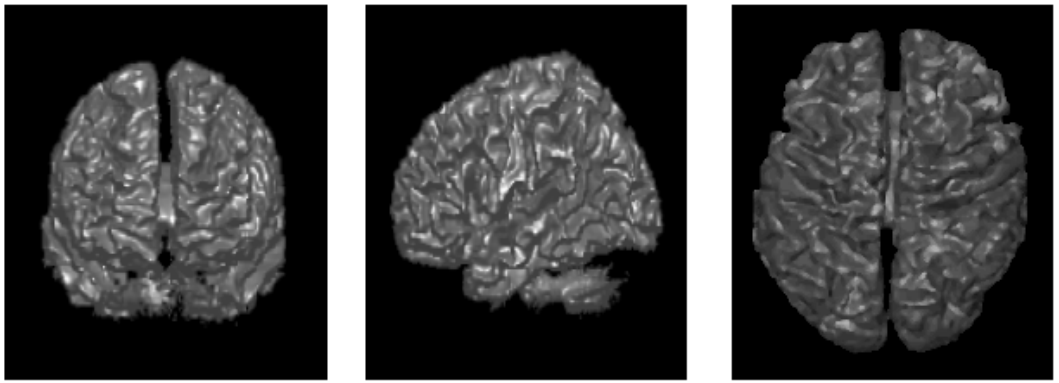


Fig. 4. Once the LDDMM transform is calculated, it can be used to transform the template XCAT phantom to define the detailed anatomy of the patient. Organs not segmented from the patient CT are interpolated using the LDDMM transform.

Cheng, et.al., SPIE 2009

Image Segmentation: brain MRI



(a) (b) (c)

Fig. 3. Three views of the interface between GM/WM for case 7. (a) Coronal view. (b) Sagittal view. (c) Axial view

Table 1. Comparison of our method with manual segmentation using DSC (%)

Tissue	Case1	2	3	4	5	6	7	8	9	10	11	12
WM	87	86	84	86	84	85	89	84	80	87	88	79
GM	77	76	70	74	72	78	80	74	72	79	79	71
Average	82	81	77	80	78	82	85	79	76	83	84	75

Cheng, et.al., IPMI, 2005

IPMI: Information Processing in Medical Imaging is the premier International conference for medical image analysis.

The first paper from China in nearly 40 years published in IPMI

Image Segmentation: Heart MRI

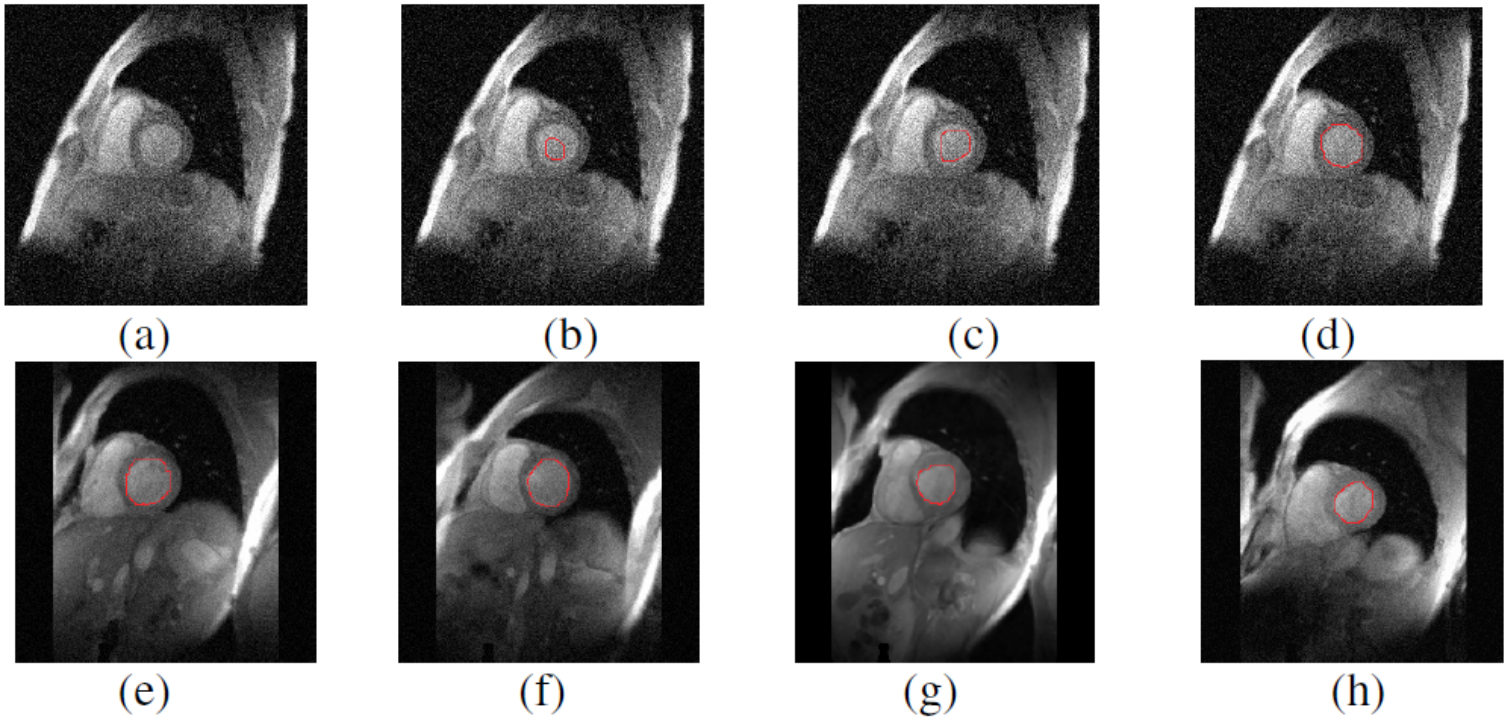
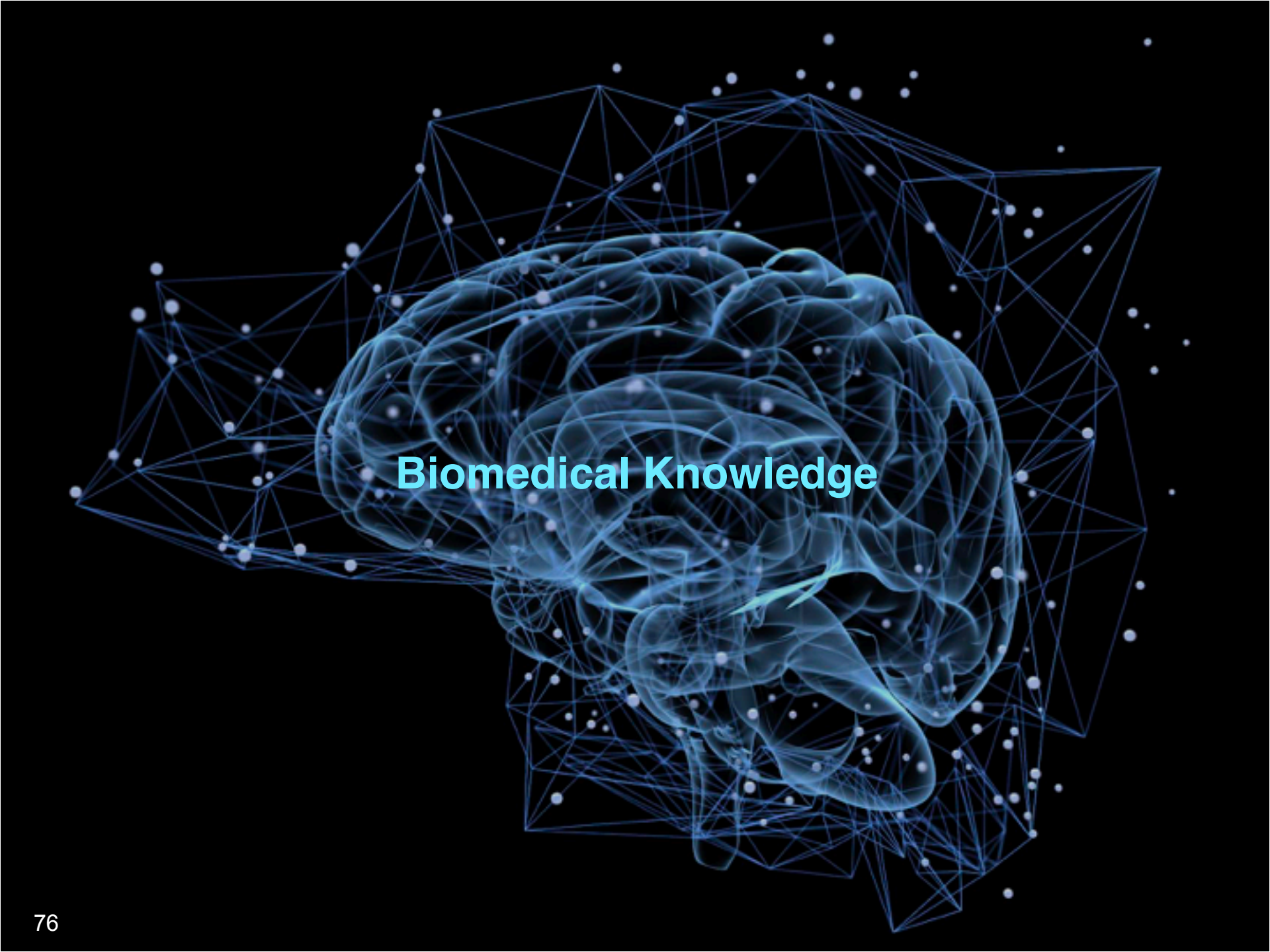


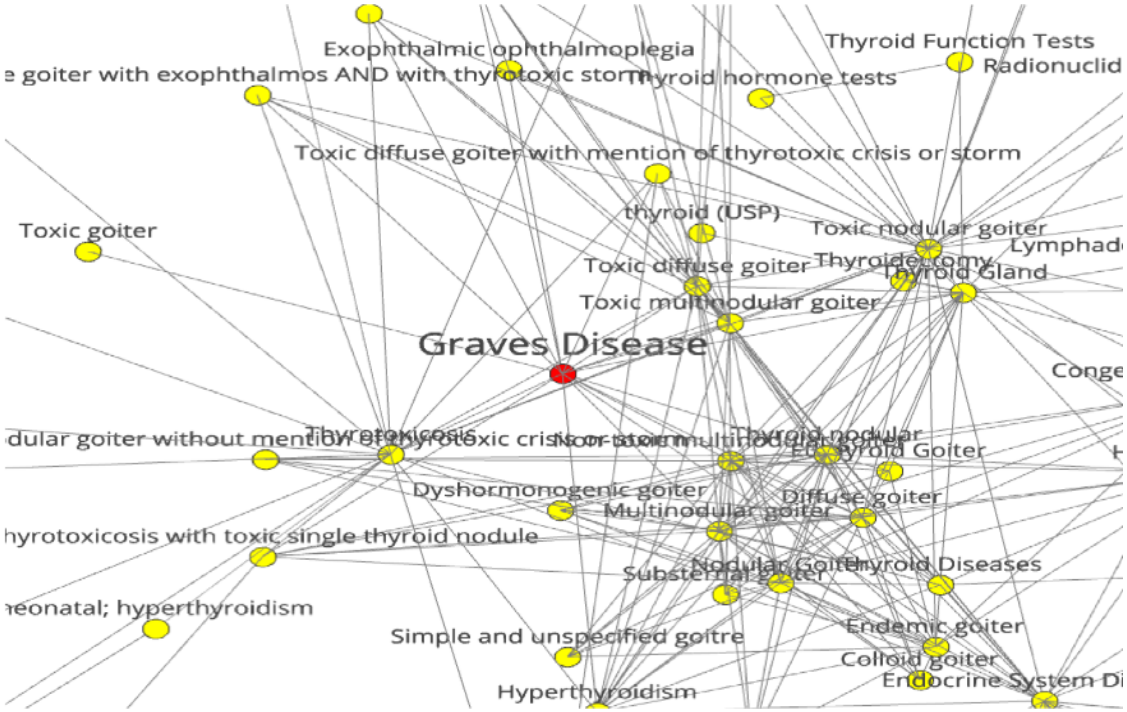
Fig. 4. Segmentation results for MR cardiac images: (a) is the original image for case 2, (b)~(d) are the three steps for the segmentation process; (e)~(h) are the final results for case 3,7,1 and 8, separately. The active contour is in red.

Cheng, et.al., International Conference on Computer Vision Workshop CVBIA, 2005


Biomedical Knowledge

Biomedical Knowledge Graph

- 3.5 million biomedical concepts with 7.5 million relationships among them.



What is Knowledge?

- Knowledge is a familiarity, awareness, or understanding of **information** about real things or abstract concepts acquired through experience or education by perceiving, discovering, or learning.
- What is information?
  - Information is any entity or form that resolves uncertainty or provides the answer to a question of some kind.
  - It is thus related to **data** and **knowledge**, as data represents values attributed to parameters, and knowledge signifies understanding of real things or abstract concepts.
- Knowledge $_{\text{how we retain}}$ Information $_{\text{obtained by}}$ Data $_{\text{represent}}$ Real or Abstract Concept.

Knowledge Organization System (KOS)

- What is KOS?
  - KOS is a general term that refers to, among other things, a set of elements, often structured and controlled, that can be used for describing objects, indexing objects, browsing collections, etc.
  - Well-understood knowledge organization schema (i.e. Taxonomy, Thesauri) that allow the organization of concepts into concept schemes where is possible to indicate relationships between the terms or concepts contained on the schema.
- Commonly found in cultural heritage institutions such as libraries and museums. They are also used in other scientific areas, such as Biology and Chemistry, where naming and classifying are important.
- Why is KOS important?
  - They can make search more robust (instead of simple keywords matching, related words, for example, can also be considered).
  - They can help to build more intelligent browsing interfaces (following the hierarchical structure, and explore broader/ narrower terms, etc.).
  - They can help us to formally organize our knowledge for a given domain, therefore promote reuse of the knowledge and also facilitate data interoperability.

Taxonomy, Thesaurus and Ontologies

- What is Taxonomy?
  - Taxonomy is the science of classification.
  - Often used in a more general way, referring to the classification of things or concepts, as well as the schema underlying the classification.
  - Only provide the narrower and broader (parent, child) relationship.
- What is Thesaurus?
  - Thesaurus can be understood as a extension to taxonomy. It add relationship types to the schema so that the concepts can be related to each other not only in a hierarchy way.
  - Some examples of these additional relationships are:
  - Broader term, narrower term, top term, related term, preferred term, synonym term, etc.
- What is Ontology?
  - An Ontology formally defines a common set of terms, and relationships among them, that are used to describe and represent a domain or an area of knowledge.
  - It offers the terms one can use when creating RDF documents in the domain.
  - It Provides a way to reuse domain knowledge.
  - Together with ontology description languages (such as RDFS and OWL), it provides a way to encode knowledge and semantics such that machines can understand (**Semantic Web**).

Thesauri vs. Ontology

- KOSs are used for knowledge organization, while ontologies are used for knowledge representation.
- Compared to ontologies, KOSs' descriptive capability is simply far too weak, which is also the reason why KOSs cannot be used to represent knowledge.
- KOSs are semantically much less rigorous than ontologies, and no formal reasoning can be conducted by just having KOSs.
- Why we need to port KOSs to the Semantic Web:
  - The KOSs schemes would become machine-readable and can be exploited in a much more effective and intelligent way.
  - Promotes reuse of KOSs schemes, and further promotes interoperability.
  - Allows KOSs to leverage all the new ideas and technologies originating from the Semantic Web.
- How do we port KOS into the Semantic Web?
  - Simple Knowledge Organization Systems, or SKOS , is an RDF vocabulary for representing KOSs, such as taxonomies, thesauri, classification schemes and subject heading lists. It is used to port existing KOSs into the shared space of the Semantic Web.

What is the Semantic Web?

Columbia University

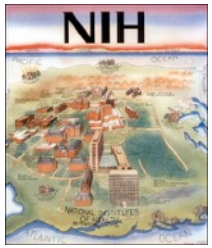
- Tim Berners-Lee the inventor of the Semantic Web defines it as follow:
  - "The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. . . . a web of data that can be processed directly and indirectly by machines."
- As the inventor of the World Wide Web, Berners-Lee hopes that eventually computers will be able to use the information on the Web, not just present the information.
- Semantic Web about having data as well as documents on the Web so that machines can process, transform, assemble, and even act on the data in useful ways.
- The purpose of this extra information is to enable the computers to understand the meaning or semantic of these documents.
- The Semantic Web provides the technologies and standards to make the current Web machine-understandable

Columbia University

Technology/Tool	Description/Purpose
Resource Description Framework (RDF)	To model or represent knowledge on the Web. Is the building block of the Semantic Web.
RDF/XML,Notation-3, Turtle and N-Triples	Serialization syntaxes for creating and reading concrete RDF models.
RDF Schema (RDFS), Web Ontology Language (OWL)	To create the common data model schema. Use to define Classes (terms), Sub Classes, Properties (terms relationships), Sub Properties and Constraints.
Semantic markup, RDFa, Microformats	To add metadata to a page by using common ontologies.
SPARQL Query Language	Language used to query RDF knowledge graph.

- The **Linked Data** is realized by using the Semantic Web standards and technologies, the result is a **Web of Data**.

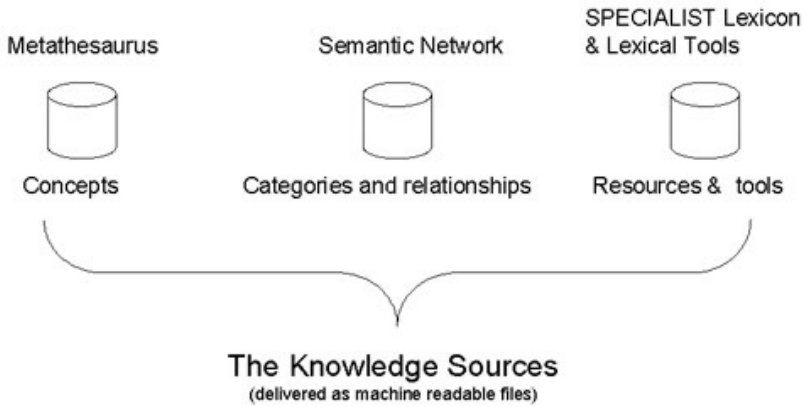
Biomedical Datasets: Thesaurus



- National Institutes of Health (NIH)
  - National Medical Research Agency.
  - Made of 27 different components called Institutes, including the National Library of Medicine.
- National Library of Medicine (NLM)
  - Founded in 1836 as a small collection of medical books and journals in the office of the United States Army Surgeon General.
  - Became part of NHI in 1962.
  - Is the world's largest biomedical library.
  - Conducts research, development, and training in biomedical informatics and health information technology.
- Unified Medical Language System (UMLS)
  - Is part of NLM.
  - The UMLS integrates and distributes biomedical key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.

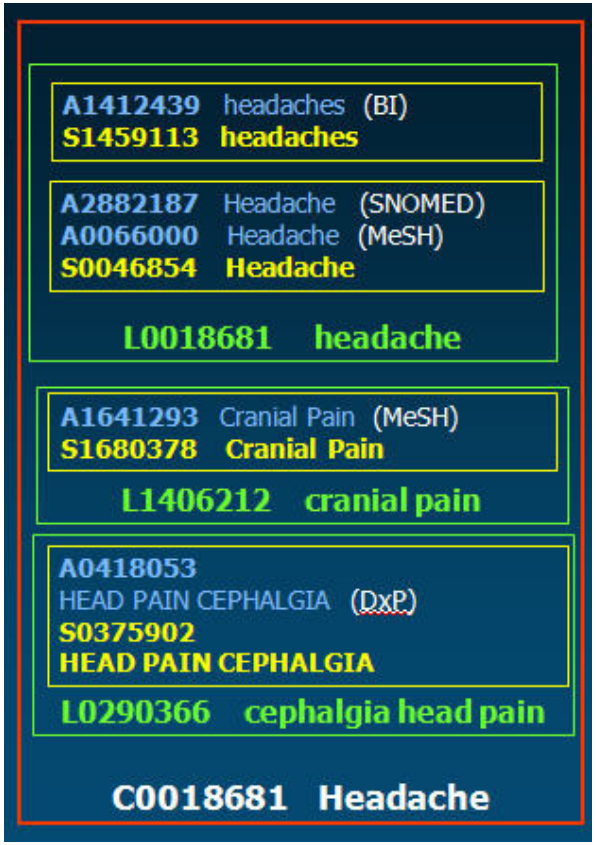
UMLS Knowledge Source

- The purpose of the UMLS Knowledge Source is to facilitate the development of computers systems.
- The UMLS distributes its data through the following 3 knowledge sources:
  - The Metathesaurus: contains over one million biomedical concepts from over 100 source vocabularies.
  - The Semantic Network: defines 133 broad categories and fifty-four relationship between categories for labeling the biomedical domain.
  - The SPECIALIST Lexicon & Lexical Tools: provides lexical information and scripts for language processing.



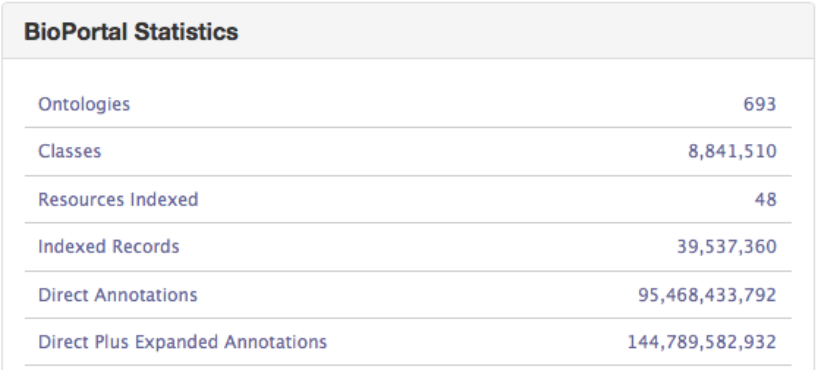
UMLS Metathesaurus

- Organized by concepts. A concept represent the abstract meaning on a conceptual level without referring to any concrete implementation.
- Can be conceptualized as a very big graph with hundreds of millions of vertices representing the CUIs, LUIs, SUIs, Normalized Words, Normalized Strings (group of words), semantic types and group of semantics.
- It connects alternative names and interpretations of the same concept, and then highlights relationships between different concepts.
- All concepts in the Metathesaurus are assigned at least one Semantic Type to provide categorization to the large vocabulary.
- **Concept Unique Identifiers (CUI)**
  A concept is a meaning. A meaning can have many different names. A key goal of Metathesaurus construction is to understand the intended meaning of each name in each source vocabulary and to link all the names from all of the source vocabularies that mean the same thing (the synonyms).
- **Lexical (term) Unique Identifiers (LUI)**
  LUI link strings that are lexical variants.
- **String Unique Identifiers (SUI)**
  Each unique concept name or string in each language in the Metathesaurus has a unique and permanent string identifier (SUI). Any variation in character set, upper-lower case, or punctuation difference is a separate string, with a separate SUI.
- **Atom Unique Identifiers (AUI)**
  The basic building blocks or "atoms" from which the Metathesaurus is constructed are the concept names or strings from each of the source vocabularies.



Biomedical Datasets: Ontologies

- BioPortal
  - The word's largest biomedical ontology repository.
  - Contains more than 600 different ontologies and more than 8 million entities that define a wide range of concepts.
  - Even though it provides lexical information (labels, definitions, etc.), comprehensive mappings between semantic types or properties are frequently missing

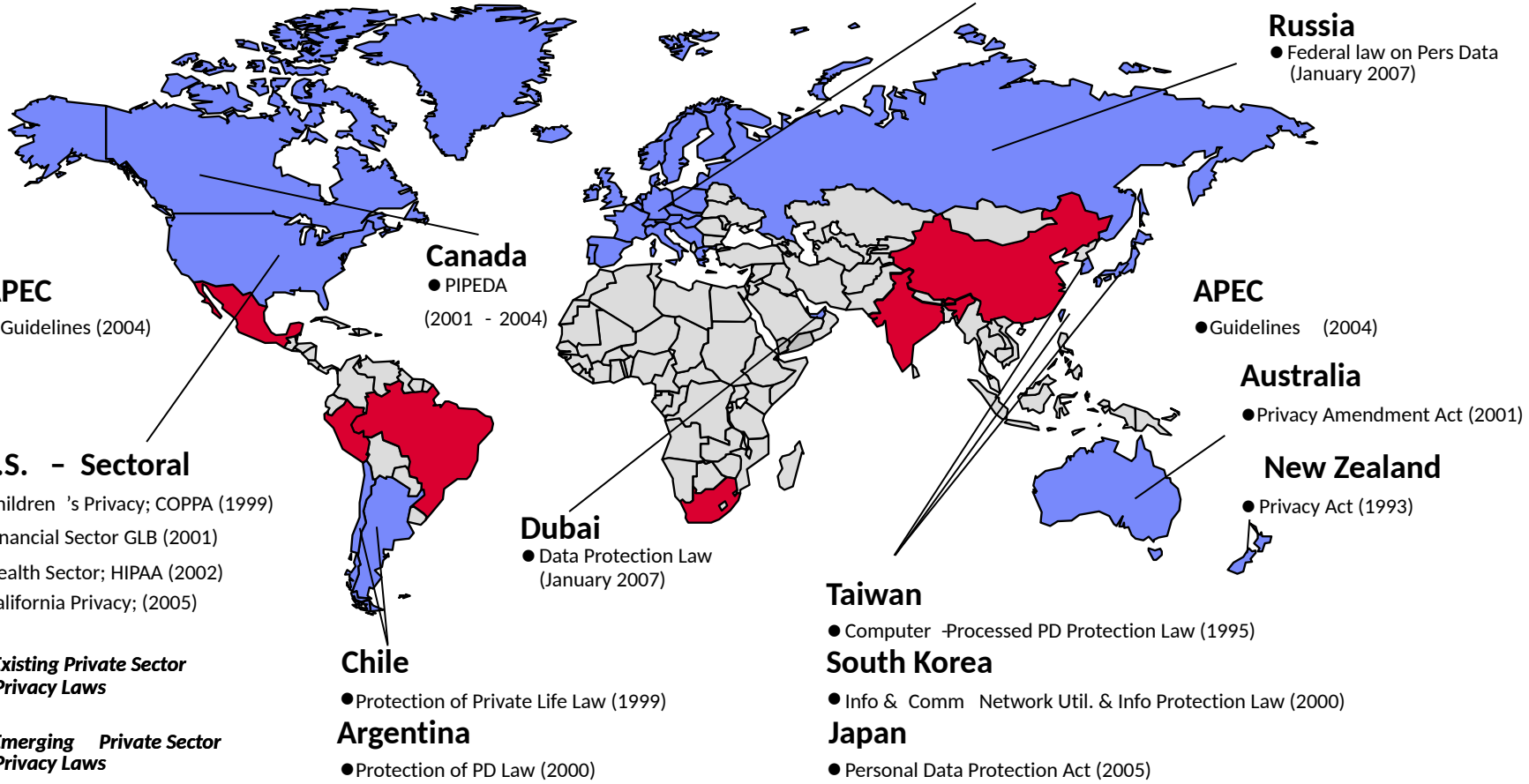


BioPortal Statistics

Ontologies	693
Classes	8,841,510
Resources Indexed	48
Indexed Records	39,537,360
Direct Annotations	95,468,433,792
Direct Plus Expanded Annotations	144,789,582,932


Data Privacy Law and Technologies

Worldwide Privacy Law Landscape (2009)



What was 'solved' via GDPR?

- **In what cases 'Personal Data' will no longer be considered protected by the law?**
  - EU GDPR Introduction, Item 26:
    - The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. **The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.** This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

Why the above sentence is important? The major purpose of Data Protection Laws is to realize the Article 12 of the Universal Declaration of Human Rights. Once a data is no longer associated to a person, it will not dishonor, attack, or interfere his life or family.

What was 'solved' via GDPR?

- **Is it my right to opt out my Health Record in a government database?**
  - **No**

- **Is it my right to opt out my pseudomized Health Record in a government database further used for scientific research or statistics?**
  - **Yes or No**
  - **Yes. That is my personal data.**
  - **No. Missing data causes incorrect statistical results that may contradict public interest.**

- **Is it my right to opt out my anonymized Health Record release by a government database?**
  - **No. The anonymized data are no longer a person's data.**

What was ‘solved’ via GDPR?

- **Can someone try to access my data without my consent?**
  - A controller goes into a system to check my data? ➔ Violation of Item 83.

Latest Privacy Regulation in US

California's New Landmark Data Privacy Regulation and What Companies Need to Do to Comply

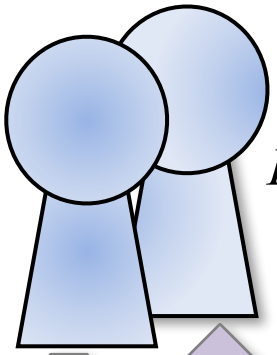
By Tyler Newby, Jim Koenig, Hanley Chew, Avery Brown and Chieh Tung

What You Need to Know Now

- The new law takes effect January 1, 2020, but there's a lot to do so you need to start work now.
- The new law expands the definition of personal information and gives California consumers increased privacy rights and protections for that information.
- California consumers can stop a business from selling their personal information. But, a business can offer financial incentives to get consumers to agree to the sale of their personal information.
- Businesses cannot discriminate against California consumers for exercising their rights under the new law, but can still charge different prices for goods and services in limited situations.
- Both the California Attorney General and individuals can bring a claim for violations, and businesses may be on the hook for up to $7,500 per violation.

Signed by Governor:
June 28, 2018

Encrypted Domain Data Mining



Surgeons

Physicians

Columbia University

Doctors

Alice

Bob

Carol

I. Doctor submits encrypted query to system
(Private Information Retrieval)

Encrypted results returned to doctor

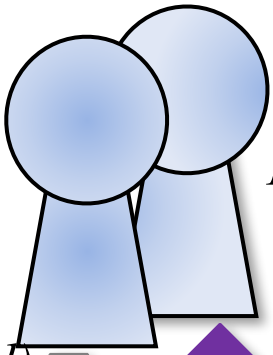
III. Privacy preserving ranking

2. Sends to server via secure channels

II. Server measures encryption-domain similarity at encrypted database

1. Patients' diagnosis results encrypted before doctor submits to database.

Encrypted Domain Data Mining



Surgeons

Physicians

Doctors

I. Doctor submits encrypted query to system
(Private Information Retrieval)

Alice

Bob

Carol

2. Sends to server via secure channels

Encrypted results returned to doctor

III. Privacy preserving ranking

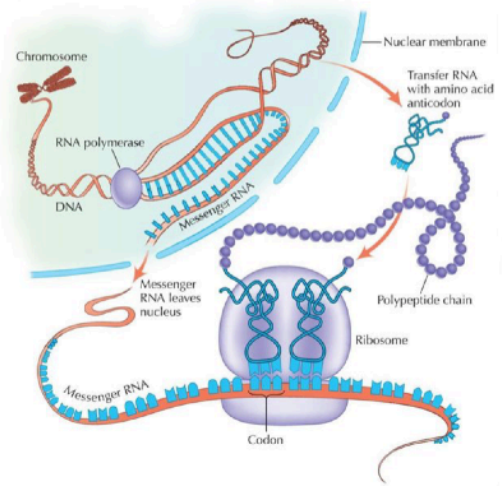
II. Server measures encryption-domain similarity at encrypted database

1. Patients' diagnosis results encrypted before doctor submits to database.


Medical Intelligence

Integral View of Complex Diseases – Omics and KGs

COLUMBIA UNIVERSITY

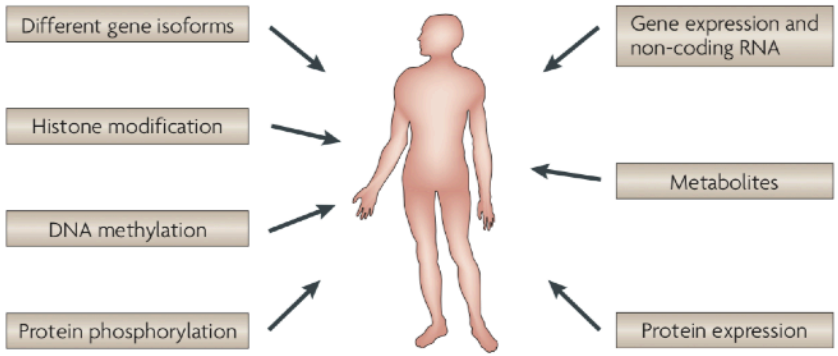


DNA

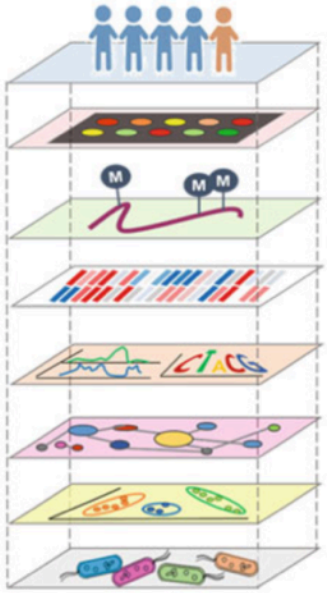
RNA

PROTEIN

a Many different types of data can be systematically scored



Healthy



Genomics

Epigenomics

Transcriptomics

Regulomics

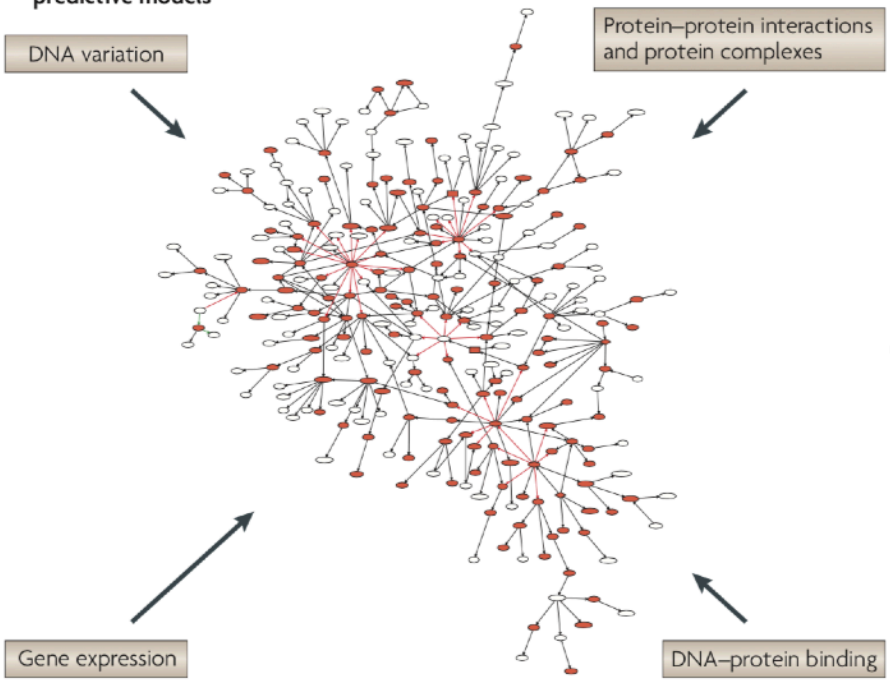
Proteomics

Metabolomics

Metagenomics

Diseased

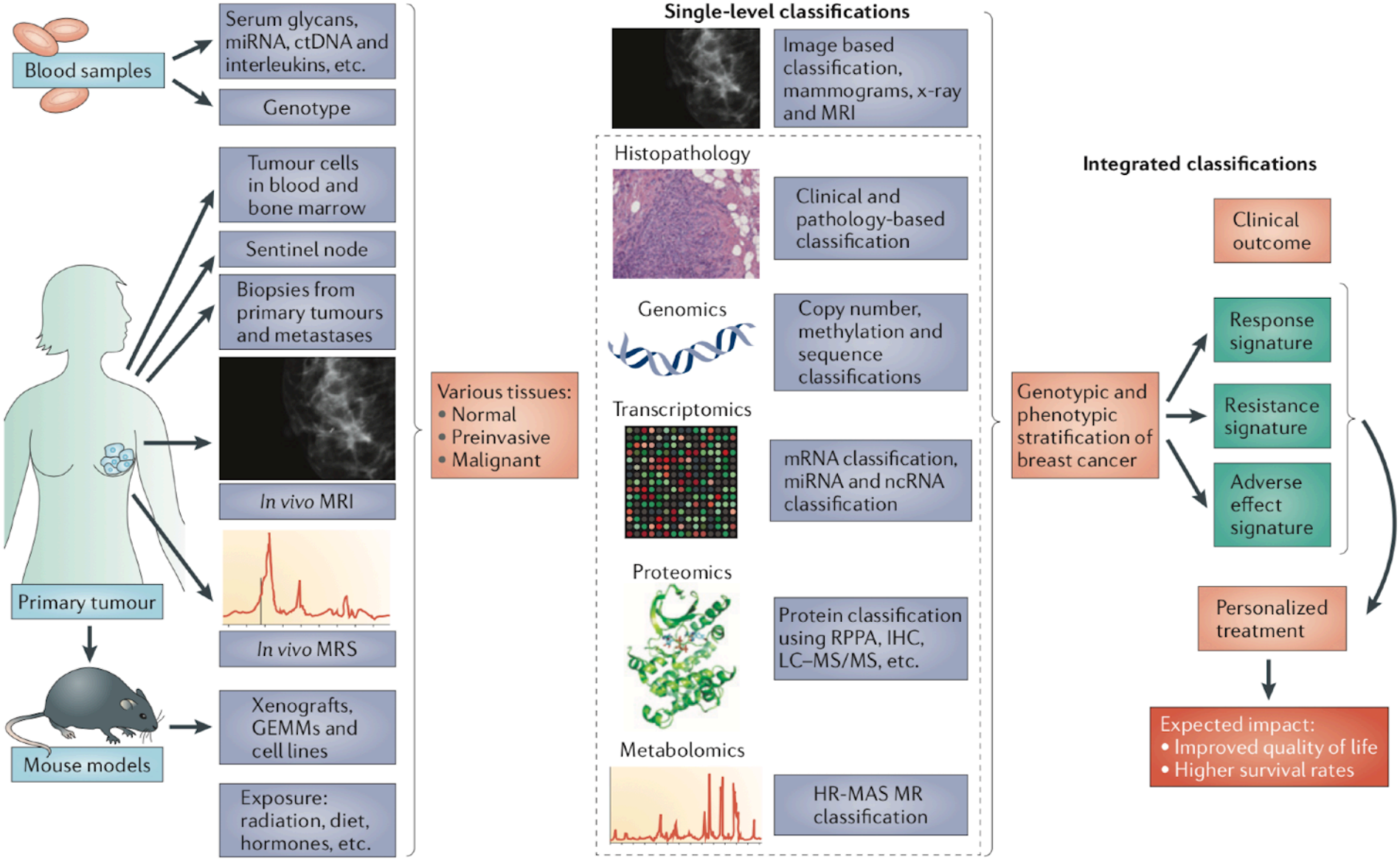
b These data can be integrated to build predictive models



Graphen Precision Medicine

Making foundation towards Precision Medicine -- Breast Cancer as an example

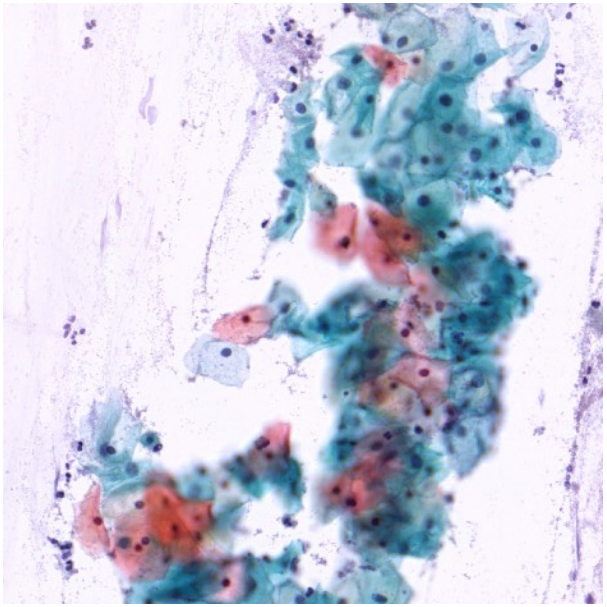
Columbia University



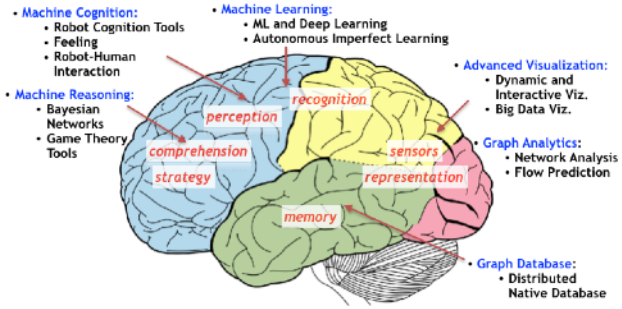
Autonomous Machine Learning

COLUMBIA UNIVERSITY

- Our method shows 2% miss rate on cervical cancer (pap smear) image detection. Human detection accuracy: 10%-40% miss rate.

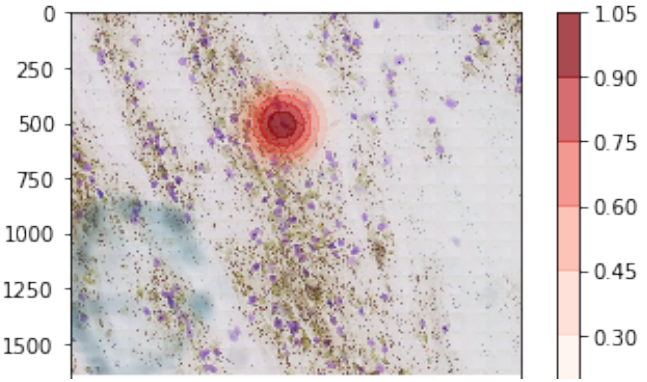


- No need for annotation

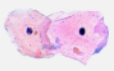
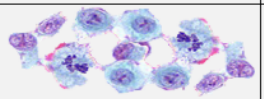
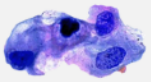


Graphen Images

- Which Cells are suspicious?



- Why suspicious?

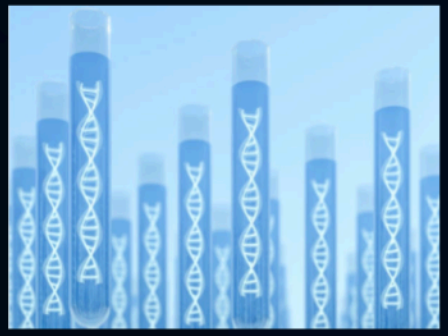
Normal	Cancer	
		Large, variably shaped nuclei
		Many dividing cells; Disorganized arrangement
		Variation in size and shape
		Loss of normal features

GRAPHEN

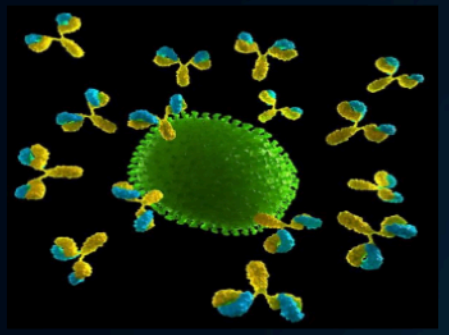
Building Next-Generation AI Solutions

Graphen Proteogenomics Pathway Analyzer

Update Data



Look at Pathways



Network Analysis

Login

Advancing healthcare related AI technologies for human life