Why ML-Based Perception Can’t be Trusted

And How to Fix It

Based on

Rick Salay, Krzysztof Czarnecki. A Safety Assurable Human-Inspired Perception Architecture. To appear in WAISE’22

Preprint arXiv:2205.07862

Assurance objective for Automated Driving

Status Quo Assurance Assumption

Society trusts that perception done by (mature, non-impaired) humans is sufficient for the safe operation of a car

CV ML Hope

Given enough training examples, the perception task can be assurably learned and done by CV ML, with (at least) human accuracy

Is CV ML hope realistic?
If so, what does it take?

Maybe we should look at how humans do perception…

Example: Classification task during driving

- Task: Object Classification

  Classify objects encountered in the world as being in class C or $\neg C$

  - e.g., Cyc is class of objects that are cyclists

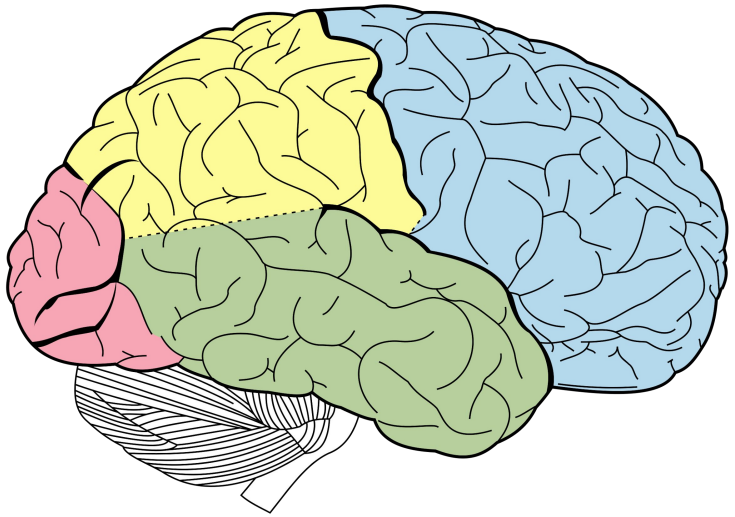
- Assumption
  - Information about objects in the world is obtained visually (as a main modality) as images
  - Effectively: object image classification

Cyc ¬Cyc

Dual process theory of human thinking

System/Type 1:

Fast, non-conscious
Wholistic, intuitive
Same across individuals
(even across species)



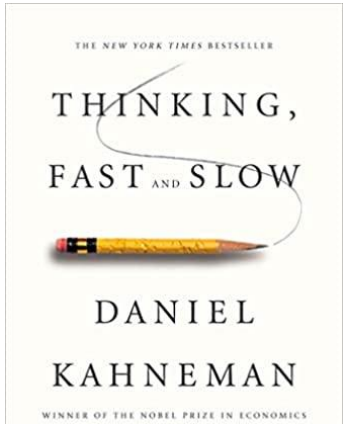
System/Type 2:

Slow, conscious
Sequential, reasoning
Varies across individuals,
correlated with intelligence
measures

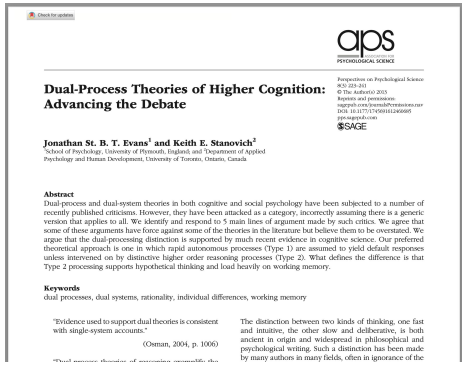
Relationship Between Type 1 and 2 Processes

- Prominent variant: *default-interventionism* [Kahneman11][Evans13]
  - Type 1 process *always* produces some default response, quickly
  - Type 2 process intervenes to produce a potentially different response only if "**difficulty**, **novelty**, and **motivation** combine to command the resources of working memory" [Evans13]
- Other hypotheses
  - (**Accuracy**) Default responses may be wrong [Evans13]
    - Humans often act as *cognitive misers* (in Type 1) by substituting a less accurate easy-to-evaluate characteristic for a harder one (leading to biases)
  - (**Uncertainty**) Confidence in default responses matter [Thompson11]
    - When people are confident, they are less likely to invoke Type 2 process

[Kahneman11]



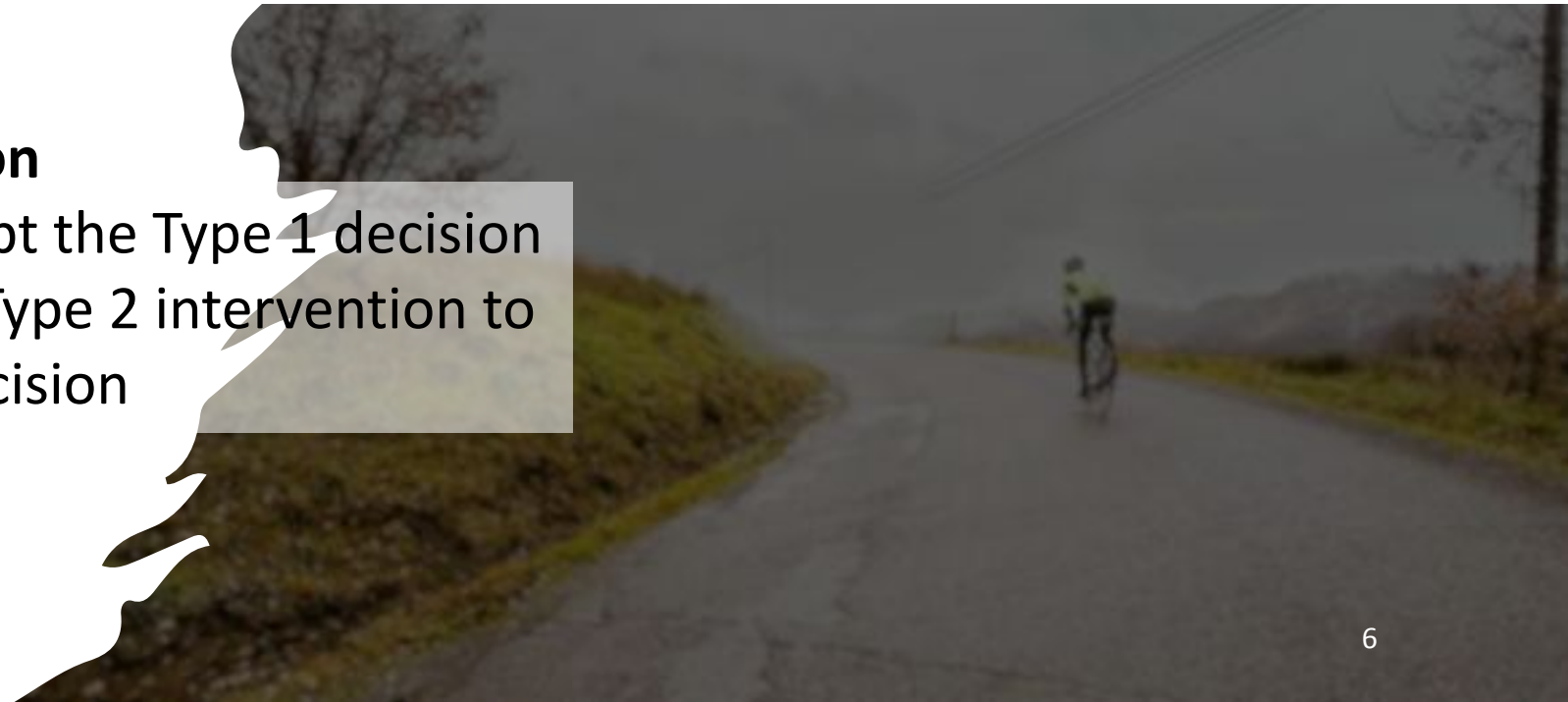
[Evans13]



Type 1 and 2 processes vs. urgency of high-risk decisions

Urgent high-risk decision

If Type 1 confident, accept the Type 1 decision
If Type 1 uncertain, take a cautious action (no time for Type 2 intervention)



Non-urgent high-risk decision

If Type 1 confident, accept the Type 1 decision
If Type 1 uncertain, use Type 2 intervention to improve upon Type 1 decision

Easy (Type 1 enough) vs. Hard (requires Type 2) Classification

- Easy – humans would *not* make type 1 classification errors on these
  - Typical instances: commonly experienced, definitively in class

  - Obvious non-instances: easy instances of other classes
  - Invariance Instances: only differ from typical instances by human vision invariance factors

    - (Bounded) variability in position, scale, pose, illumination, clutter, degradation, etc.

- Hard – humans *could* make type 1 classification errors on these

  - Novel instances (FN): look different from experienced instances

  - Subtle non-instances (FP): looks typical, requires observation of subtle atypical details

  - Aleatoric cases (FN/FP): not enough information to decide class (w[illegible]sing context)

Human Object Image Classification Performance Summary

- Type 1 (Fast default classification)
  - High confidence
    - Easy cases: optimal (accurate) classification
    - Hard subtle cases: FP classification, undetected and potentially unsafe

  - Low confidence
    - Remaining hard cases: apply conservative (i.e., safe) classification
    - Triggers Type 2 processing if need requires and time permits

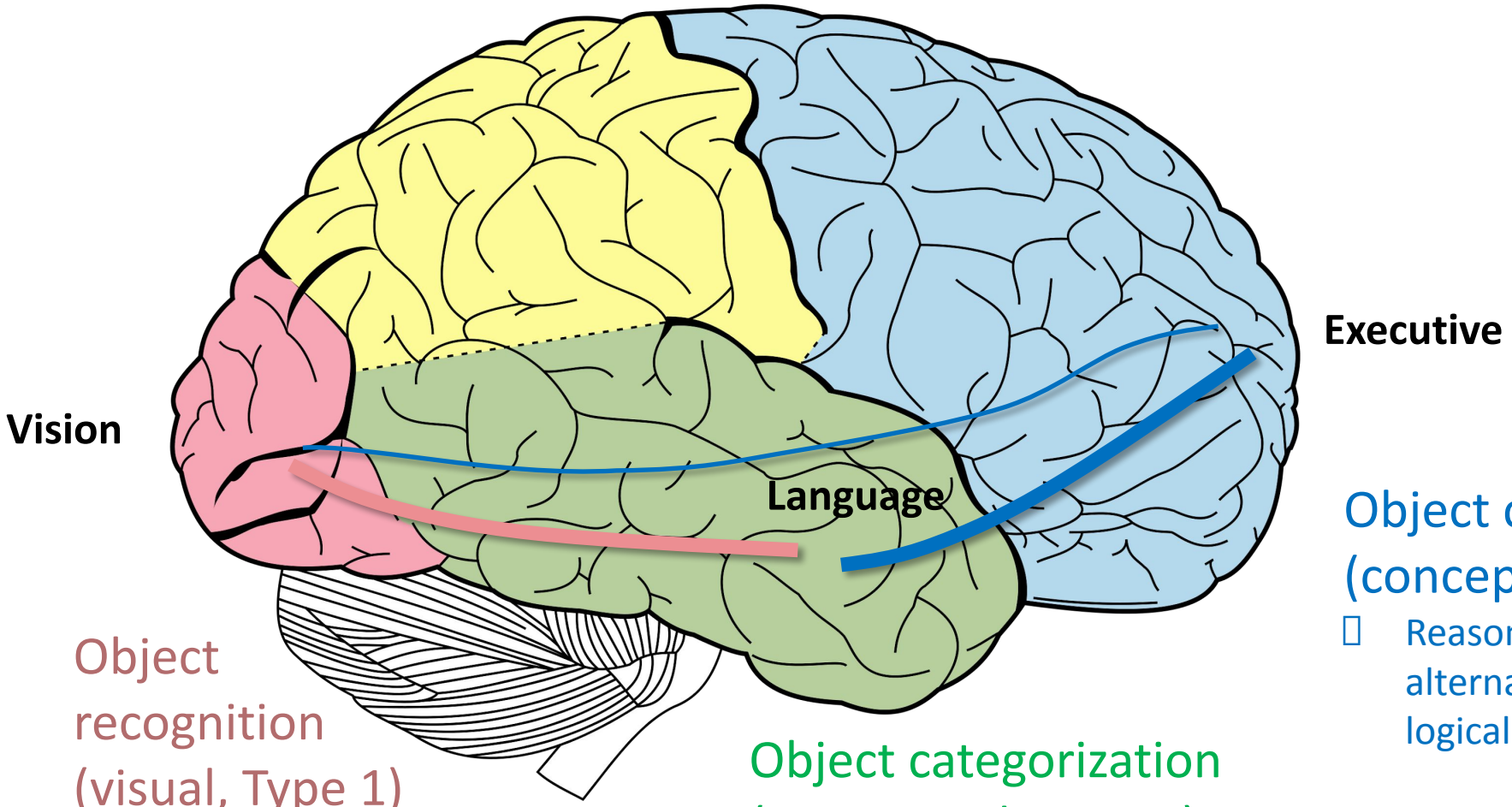
- Type 2 (Slow classification)
  - Use reasoning to improving classification accuracy as more time invested

ADS Assurance/Trust Implications

Claim: ADS trust by humans is affected by implicit expectations about performance on easy/hard cases

- Easy cases: Should not make fast classification errors since a human wouldn't
  - Errors are surprising and undermine trust
  - CV ML currently fails here:
    - e.g., adversarial errors are invariance instance errors (FNs)
- Hard cases: Fast classification errors acceptable but should be handled safely
  - Acting cautiously given high uncertainty
  - This depends on accurate uncertainty judgement to detect hard cases
  - CV ML currently fails here:
    - e.g., poor calibration/OOD detection

Brain Areas Involved in Object Image Classification



Object categorization (conceptual, Type 2)

- Reasoning, alternative hypotheses, logical arguments

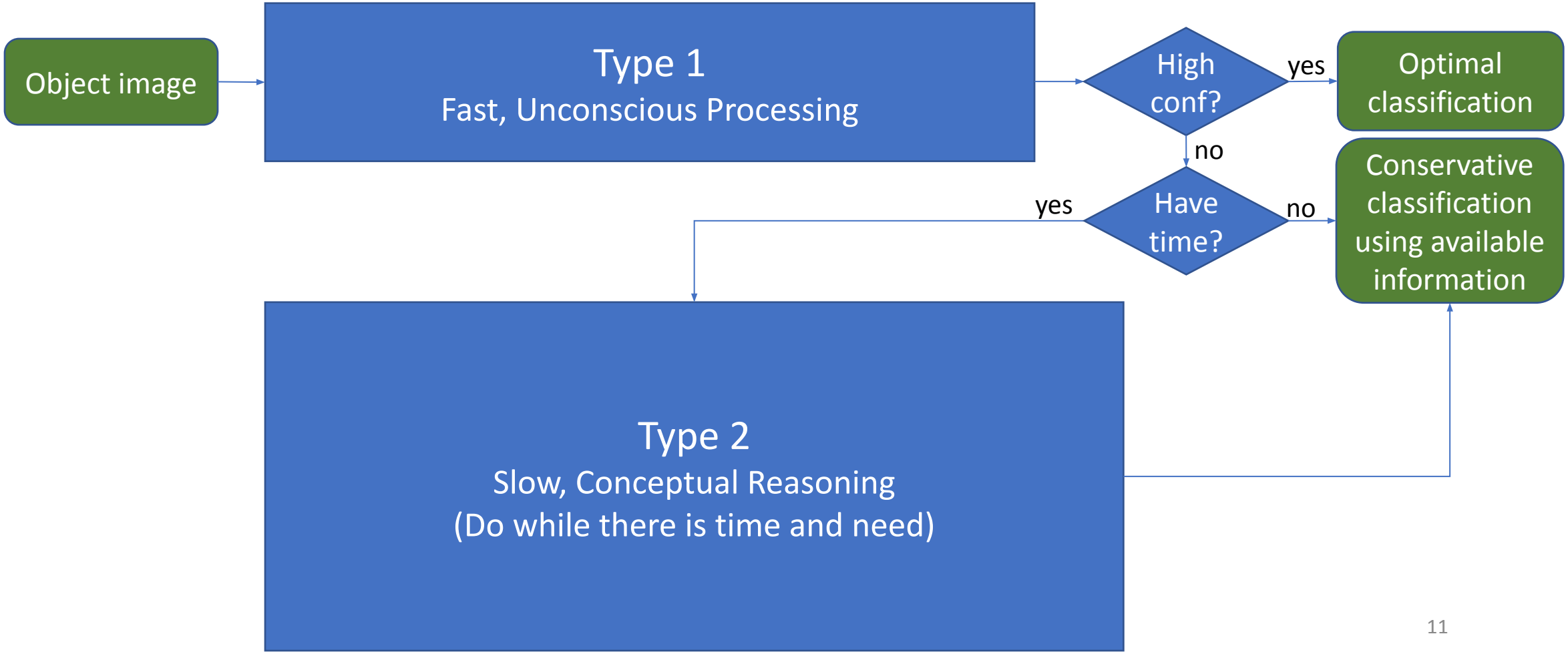
Object recognition (visual, Type 1)

- Produces normalized object representations

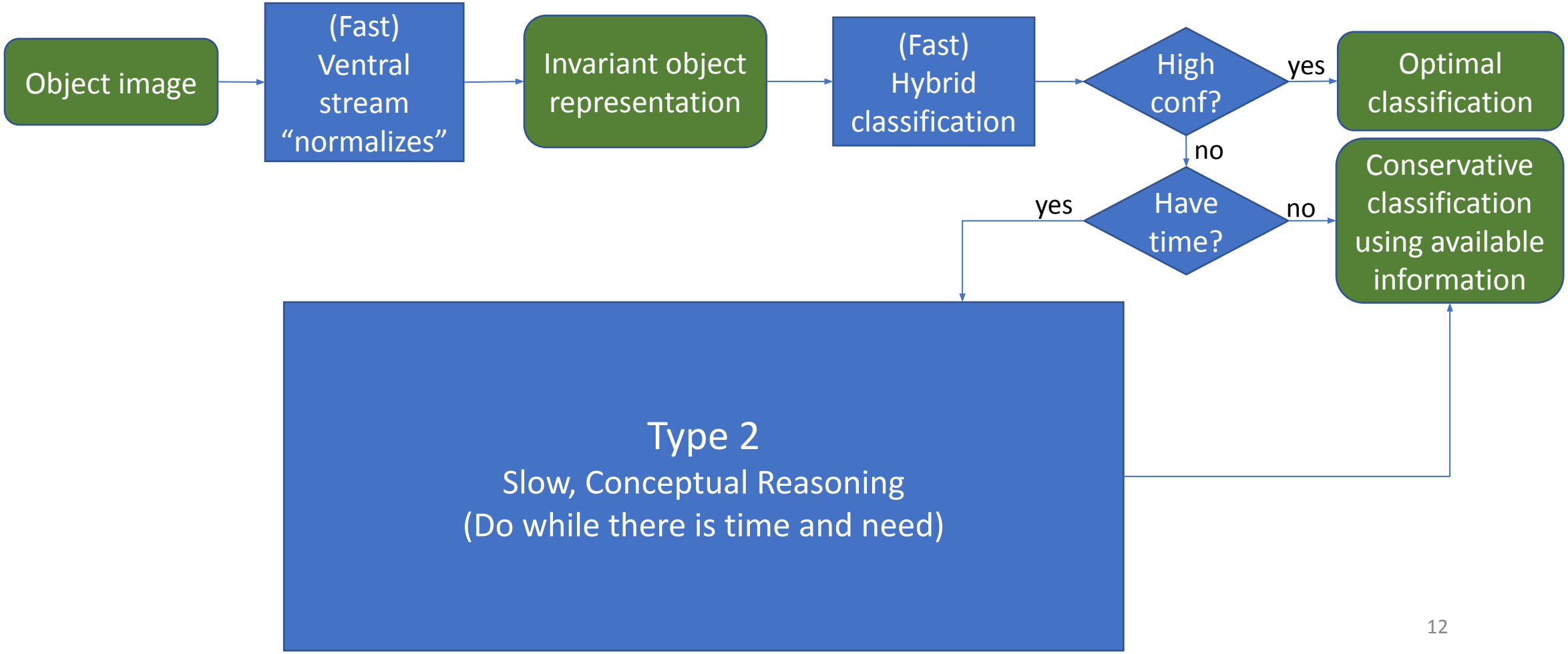
Object categorization (conceptual, Type 1)

- Rules, prototypes, exemplars

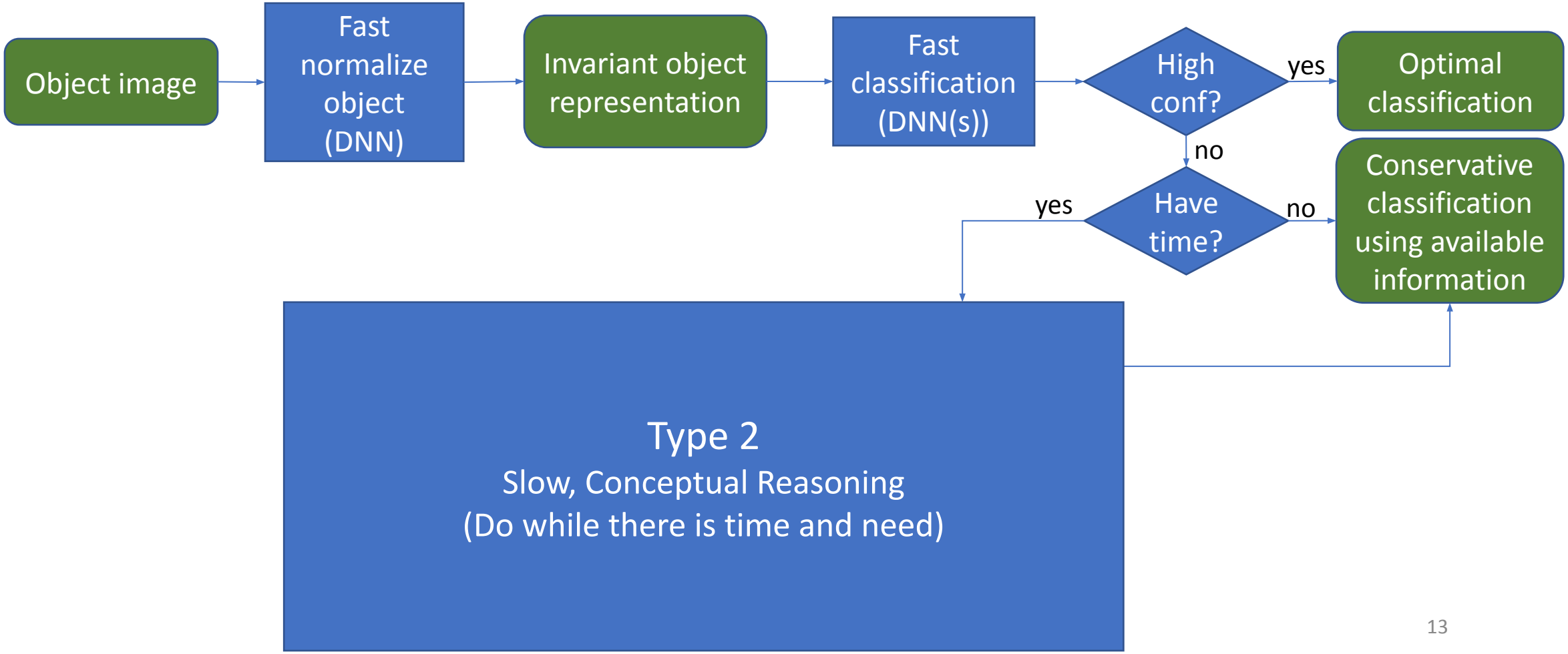
Classification Workflow in Humans – Summary



Classification Workflow in Humans – Summary

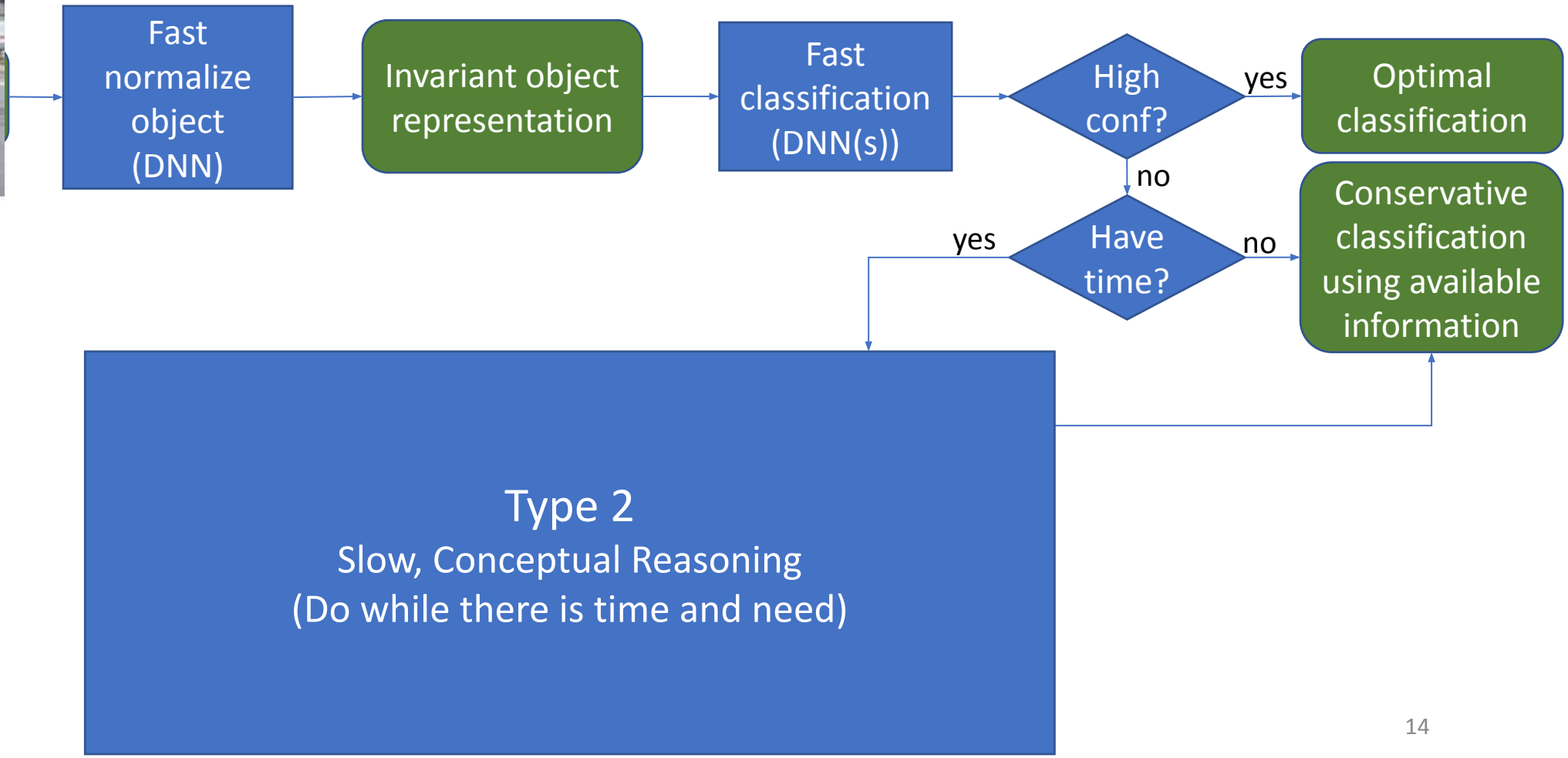


Safe human-inspired classification workflow (Type 1)



Safe human-inspired classification workflow (Type 1)

Easy case

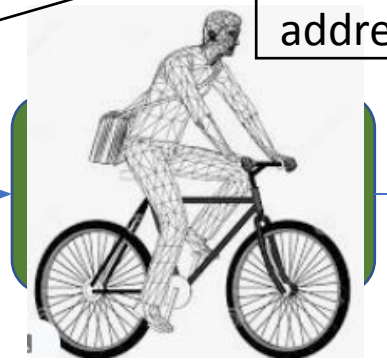


Safe human-inspired classification workflow (Type 1)

Easy case

Easy cases
generic "object backbone"
addresses invariance instances

Fast normalize object (DNN)



Fast classification (DNN(s))

High conf?

yes → Optimal classification

no → Have time?

yes → Type 2

no → Conservative classification using available information

Type 2
Slow, Conceptual Reasoning
(Do while there is time and need)

Safe human-inspired classification workflow (Type 1)

Easy case

Fast normalize object (DNN)

Invariant object representation

Fast classification (DNN(s))

Easy cases
Classifiers address typical instances (best-fit class)

High conf?

yes

Optimal classification

no

no

Conservative classification using available information

(Do w

Safe human-inspired classification workflow (Type 1)

Easy case

Fast normalize object (DNN)

Fast classification (DNN(s))

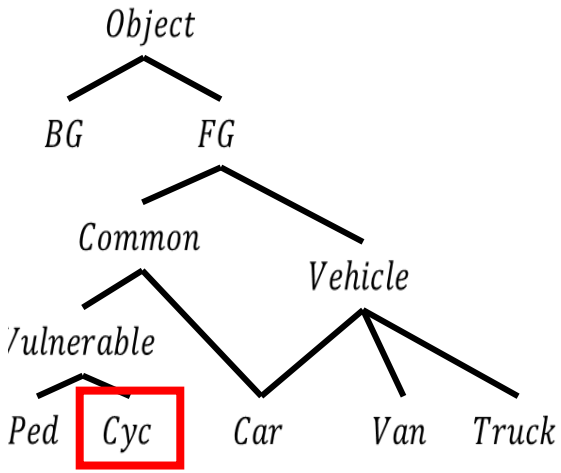
High conf?

yes → Cyc

no → Have time?

yes → Type 2

no → Conservative classification using available information

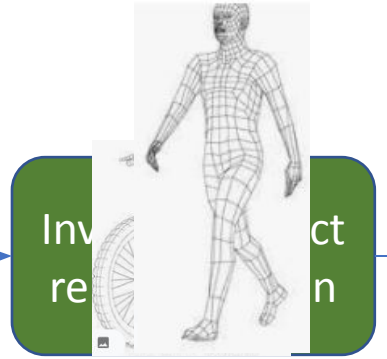


Type 2
Slow, Conceptual Reasoning
(Do while there is time and need)

Safe human-inspired classification workflow (Type 1)

Hard case

Fast normalize object (DNN)



Fast classification (DNN(s))

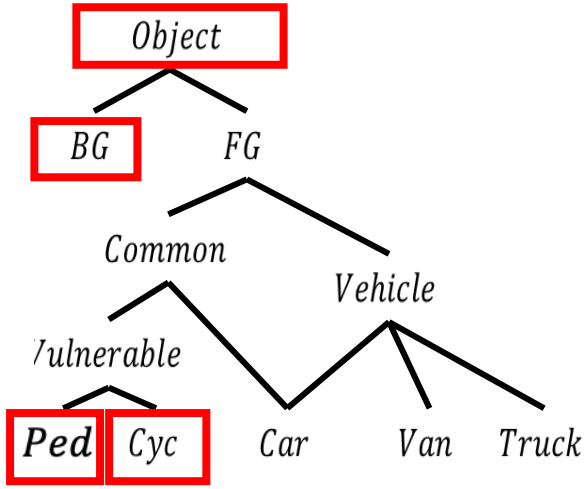
High conf?

yes → Optimal classification

no → Have time?

yes → Type 2

no → Most specific high confidence (under) classification



Type 2
Slow, Conceptual Reasoning
(Do while there is time and need)

Safe human-inspired classification workflow (Type 2)

Hard case

Fast normalize object (DNN)

Fast classification (DNN(s))

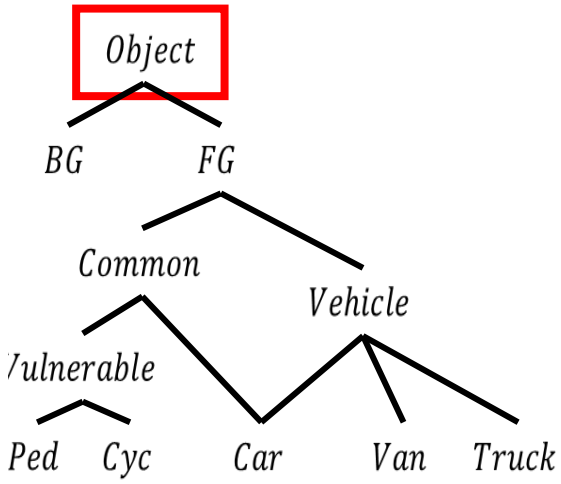
High conf?

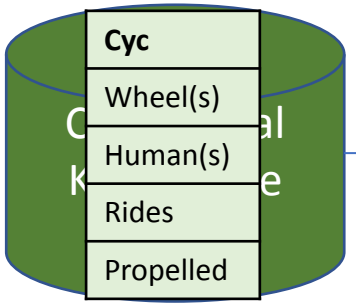
yes → Optimal classification

no → Have time?

yes → Slow Classification (Symbolic AI)

no → Object





? Wheel(s)
? Human(s)
? Rides
? Propelled

Slow Classification (Symbolic AI)

Answer

while time and need

Query

Fast query answering (DNN(s))

Safe human-inspired classification workflow (Type 2)

Hard case

Fast normalize object (DNN)

Fast classification (DNN(s))

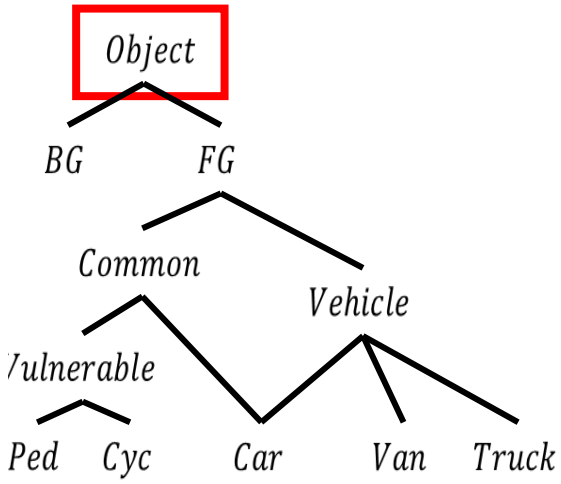
High conf?

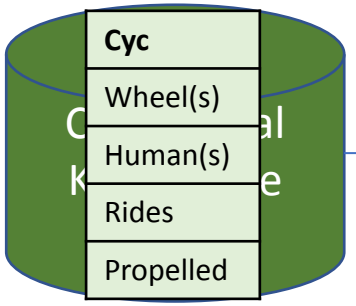
yes → Optimal classification

no → Have time?

yes → Slow Classification (Symbolic AI)

no → Object





? Wheel(s)
? Human(s)
? Rides
? Propelled

Slow Classification (Symbolic AI)

Answer

while time and need

Query

Fast query answering (DNN(s))

Safe human-inspired classification workflow (Type 2)

Hard case

Fast normalize object (DNN)

Fast classification (DNN(s))

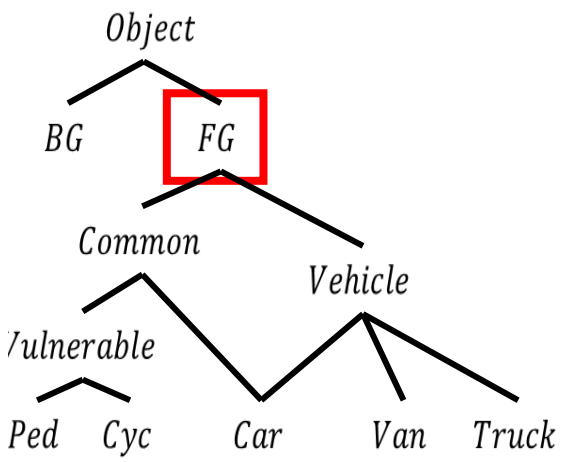
High conf?

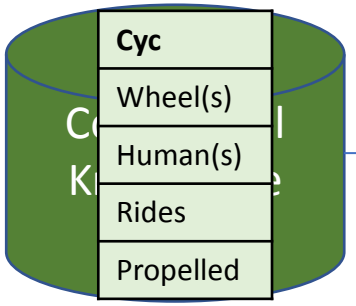
yes → Optimal classification

no → Have time?

yes → Slow Classification (Symbolic AI)

no → FG





✔ Wheel(s)
? Human(s)
? Rides
? Propelled

Slow Classification (Symbolic AI)

Query: is there a wheel?

while time and need

Answer: yes

Fast query answering (DNN(s))

Safe human-inspired classification workflow (Type 2)

Hard case

Fast normalize object (DNN)

Fast classification (DNN(s))

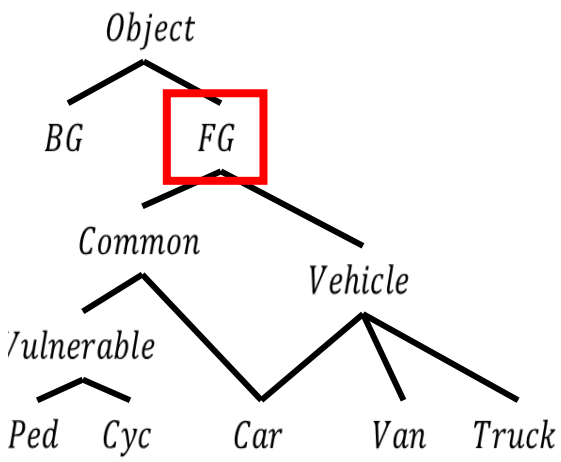
High conf?

yes → Optimal classification

no → Have time?

yes → Slow Classification (Symbolic AI)

no → FG



Cyc	Human
Wheel(s)	Head
Human(s)	Body
Rides	
Propelled	

✔Wheel(s)	? Head
? Human(s)	? Body
? Rides	
? Propelled	

Slow Classification (Symbolic AI)

Answer: yes

while time and need

Query: is there a human?

Fast query answering (DNN(s))

Safe human-inspired classification workflow

Hard case

Fast normalize object (DNN)

Fast classification (DNN(s))

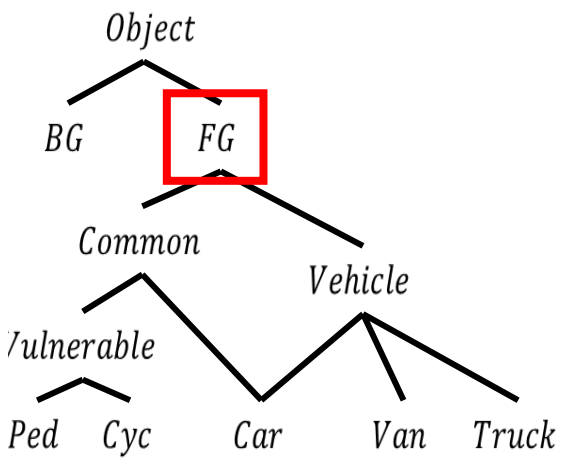
High conf?

yes → Optimal classification

no → Have time?

yes → Slow Classification (Symbolic AI)

no → FG



Cyc
Wheel(s)
Human(s)
Rides
Propelled

Human
Head
Body

✔ Wheel(s)
? Human(s)
? Rides
? Propelled

✔ Head
? Body

Slow Classification (Symbolic AI)

Answer: yes

while time and need

Query: is there a head?

Fast query answering (DNN(s))

Safe human-inspired classification workflow (Type 2)

Hard case

Fast normalize object (DNN)

Fast classification (DNN(s))

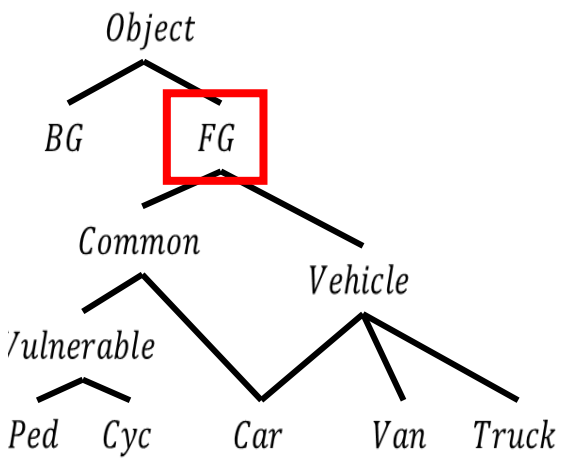
High conf?

yes → Optimal classification

no → Have time?

yes → Slow Classification (Symbolic AI)

no → FG



Cyc	Human
Wheel(s)	Head
Human(s)	Body
Rides	
Propelled	

✔ Wheel(s)	✔ Head
✔ Human(s)	✔ Body
? Rides	
? Propelled	

Slow Classification (Symbolic AI)

Answer: yes

while time and need

Query: is there a body?

Fast query answering (DNN(s))

Safe human-inspired classification workflow (Type 2)

Hard case

Fast normalize object (DNN)

Fast classification (DNN(s))

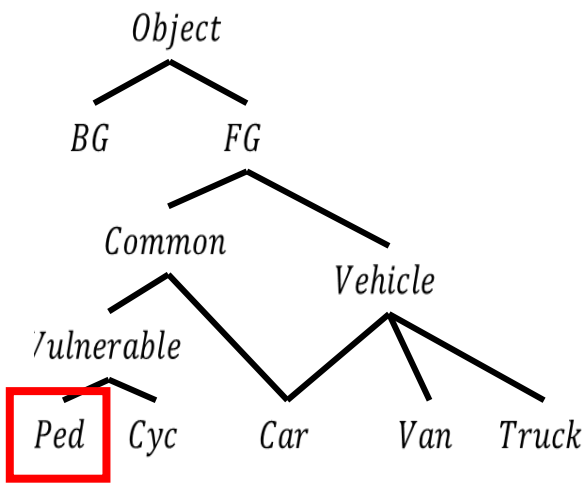
High conf?

yes → Optimal classification

no → Have time?

yes → Slow Classification (Symbolic AI)

no → Ped



Cyc	Ped
Wheel(s)	Human
Human(s)	Self-supported
Rides	
Propelled	

✔ Wheel(s)	✔ Human
✔ Human(s)	? Self-supported
? Rides	
? Propelled	

Slow Classification (Symbolic AI)

Answer: no

while time and need

Query: human rides?

Fast query answering (DNN(s))

Safe human-inspired classification workflow (Type 2)

Hard case

Fast normalize object (DNN)

Fast classification (DNN(s))

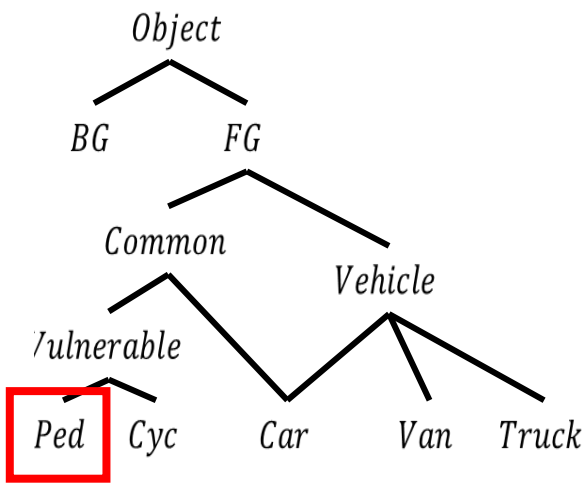
High conf?

yes → Optimal classification

no → Have time?

yes → Slow Classification (Symbolic AI)

no → Ped



Cyc	Ped
Wheel(s)	Human
Human(s)	Self-supported
Rides	
Propelled	

✔ Wheel(s)	✔ Human
✔ Human(s)	? Self-supported
☐ Rides	
? Propelled	

Slow Classification (Symbolic AI)

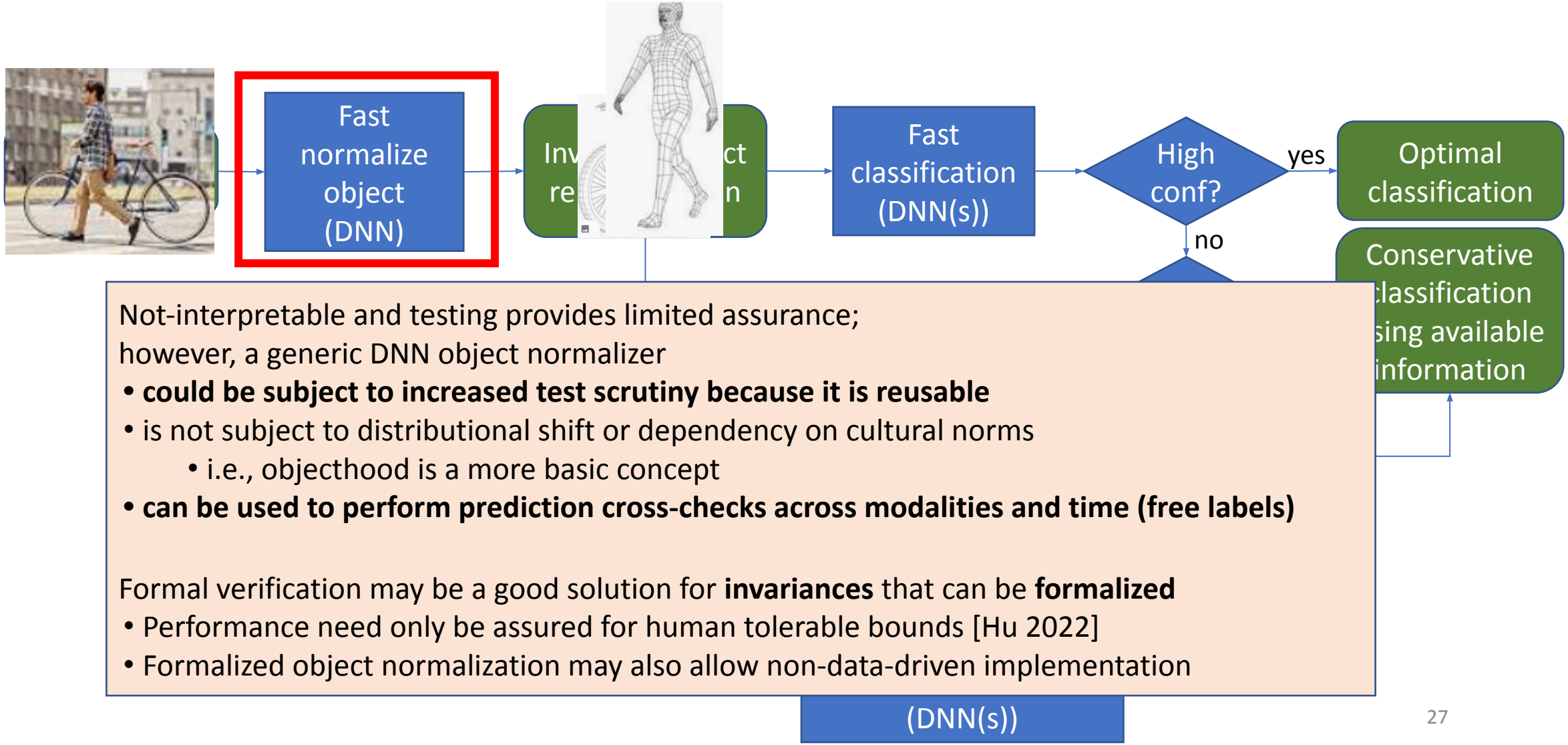
Answer: no

while time and need

Query: human rides?

Fast query answering (DNN(s))

Assurance



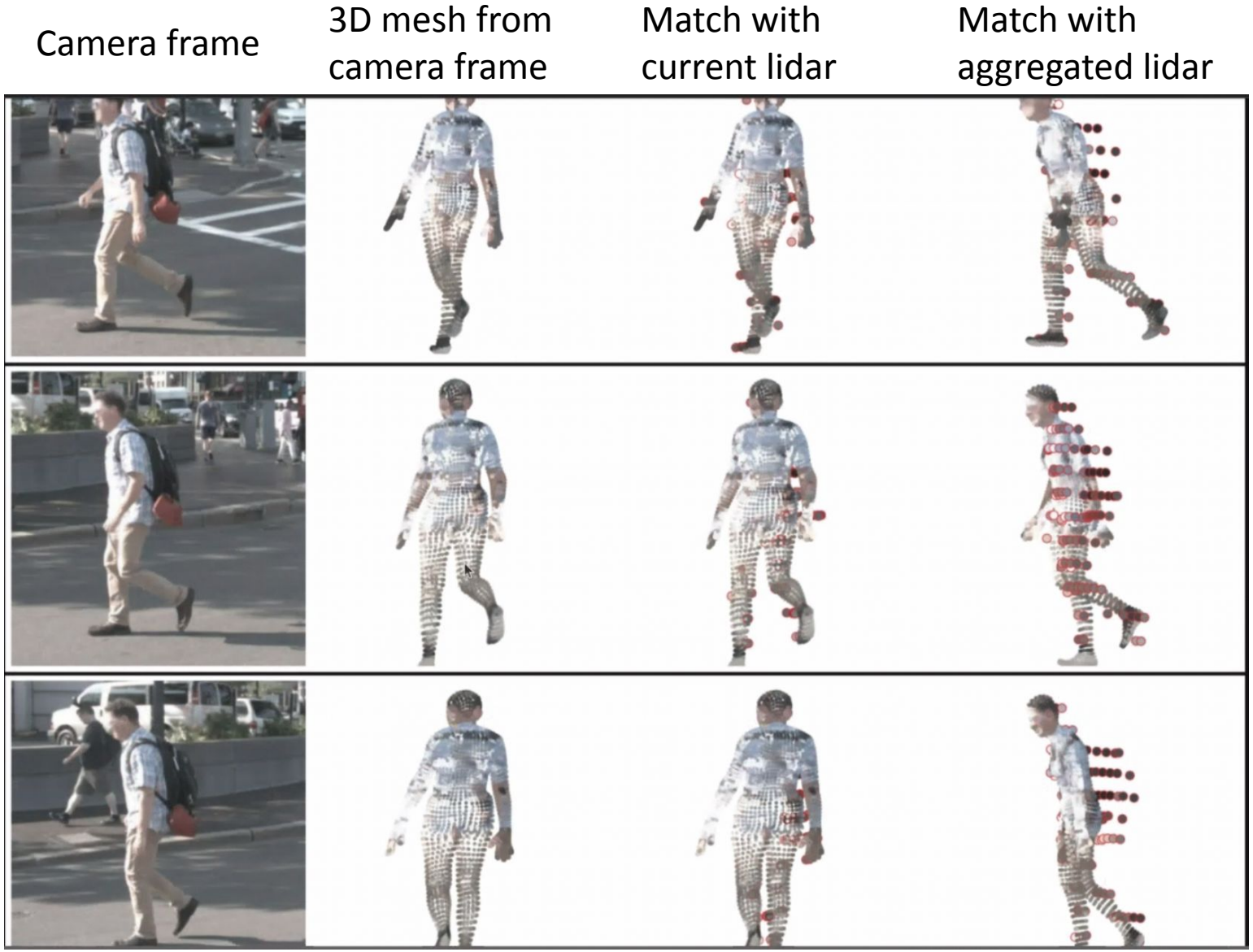
Not-interpretable and testing provides limited assurance;
however, a generic DNN object normalizer

- **could be subject to increased test scrutiny because it is reusable**
- is not subject to distributional shift or dependency on cultural norms
  - i.e., objecthood is a more basic concept
- **can be used to perform prediction cross-checks across modalities and time (free labels)**

Formal verification may be a good solution for **invariances** that can be **formalized**

- Performance need only be assured for human tolerable bounds [Hu 2022]
- Formalized object normalization may also allow non-data-driven implementation

Example of cross-modal prediction check



Assurance

Fast normalize object

Fast classification (DNN(s))

High conf?

yes → Optimal classification

no → Conservative classification using available information

Prototype/exemplar-based classifiers provide interpretability

- allows inspection by humans to provide evidence for correctness (global explanation)

Alignment with how humans represent concepts may be well-founded source of trust

- Normalized objects simplify classification problem

Issues

- **Calibrated uncertainty / out-of-distribution detection** (Type I / Type II decisions)
- Assurance of similarity judgement – can it be a separate generic reusable component?
- Bounding the occurrence of hard subtle cases

Fast query answering (DNN(s))

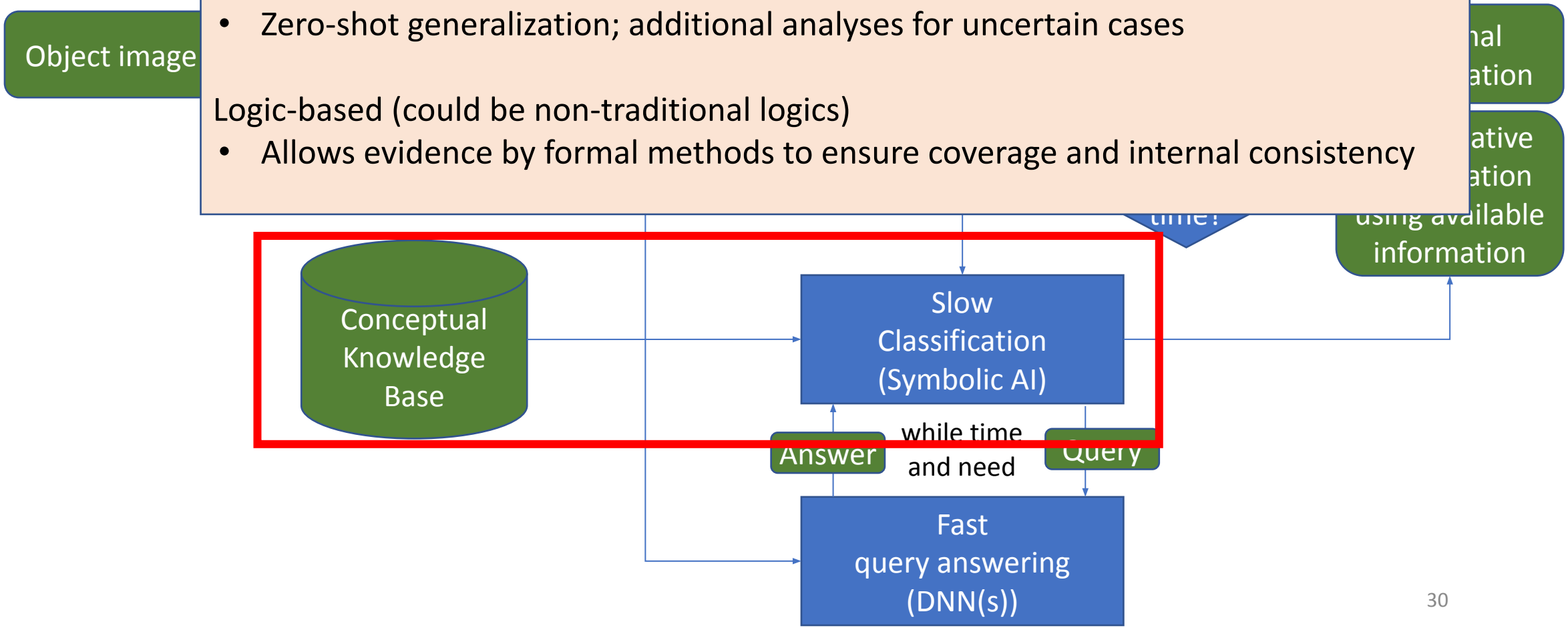
Assurance

Conceptual and linguistic, therefore interpretable

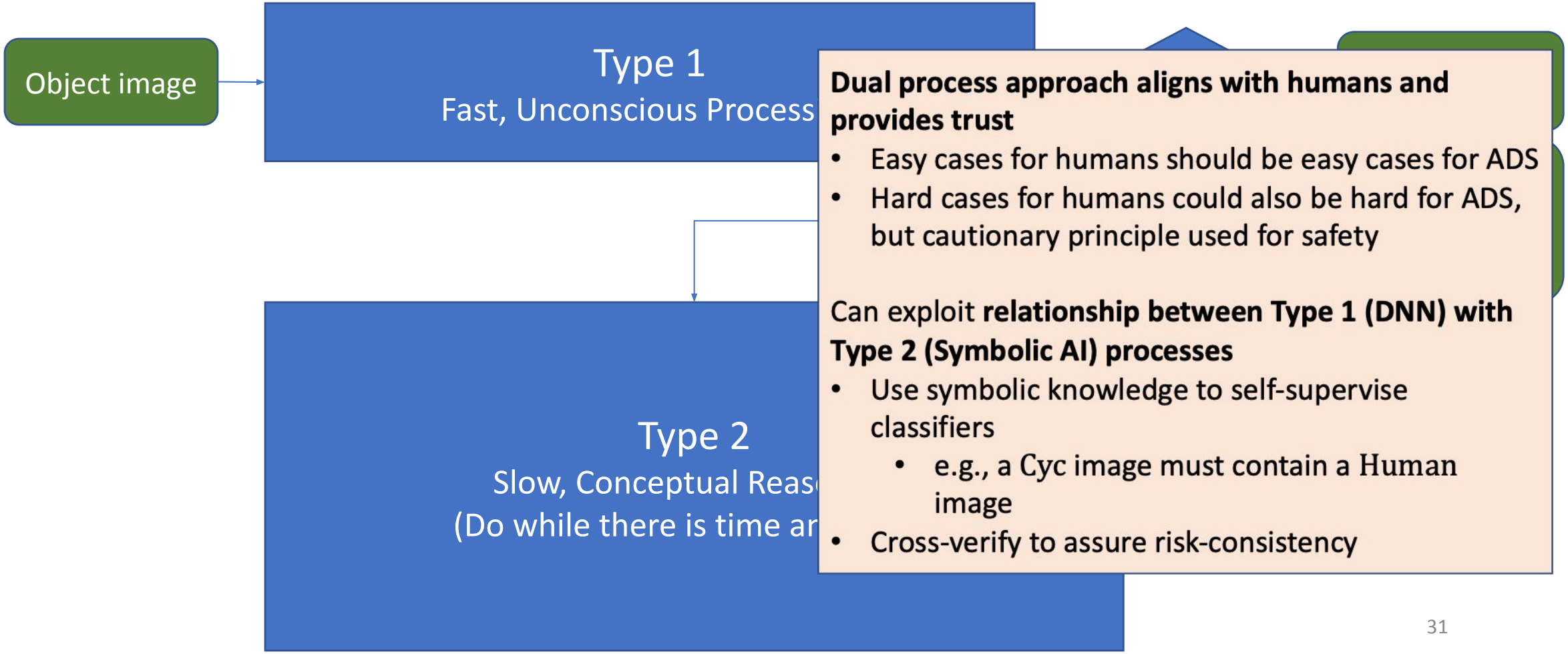
- Allows evidence via inspection by humans
- Can verify alignment with cultural-specific knowledge about object classes
- Zero-shot generalization; additional analyses for uncertain cases

Logic-based (could be non-traditional logics)

- Allows evidence by formal methods to ensure coverage and internal consistency



Assurance


Object image
Type 1
Fast, Unconscious Process
Type 2
Slow, Conceptual Reas
(Do while there is time ar
Dual process approach aligns with humans and provides trust
• Easy cases for humans should be easy cases for ADS
• Hard cases for humans could also be hard for ADS, but cautionary principle used for safety
Can exploit relationship between Type 1 (DNN) with Type 2 (Symbolic AI) processes
• Use symbolic knowledge to self-supervise classifiers
• e.g., a Cyc image must contain a Human image
• Cross-verify to assure risk-consistency

Summary

Understanding how humans do classification is useful both for ADS perception functionality and assurance

- Normalization, prototype/exemplar-based classification, need for Type 2 processing

Takeaway: human-inspired perception architectures are warranted

For more detail see

Rick Salay, Krzysztof Czarnecki. A Safety Assurable Human-Inspired Perception Architecture. Preprint arXiv:2205.07862

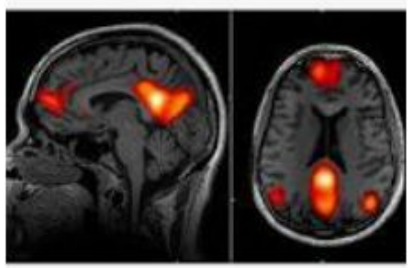
Appendix

Research on human vision tasks

Object recognition (Type 1)

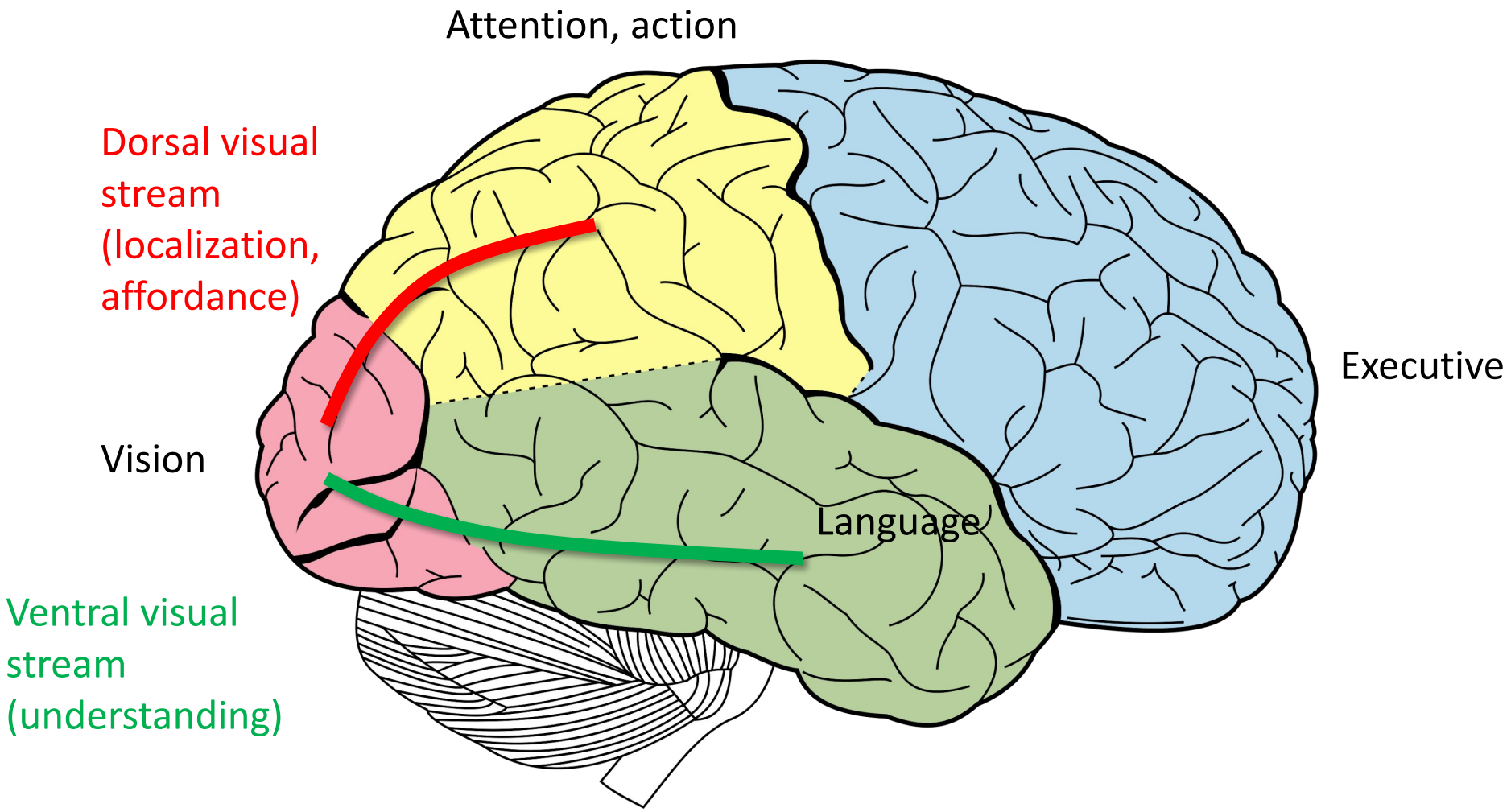
- Ability to assign labels to particular objects, from precise labels (“identification”) to coarse labels (“categorization”) [DiCarlo12]

- Studied in cognitive neuroscience of vision
  - fMRI studies



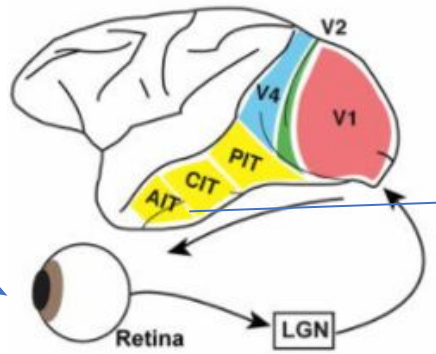
- Behavioural studies show robust 250–290 ms min RT for humans categorizing object classes [Thorpe11]

Major Brain Areas



Object Recognition

- Ventral stream is responsible for object recognition [DiCarlo12]
  - Dorsal stream deals with other tasks: object tracking, segmentation, obstacle avoidance, object grasping, etc.
- Key function: **map images into *invariant* object representations**
  - human vision is invariant to *identity-preserving image transformations*
    - specificity for some classes: faces, animate vs inanimate, tools, places
  - invariant object representations then used to access memory

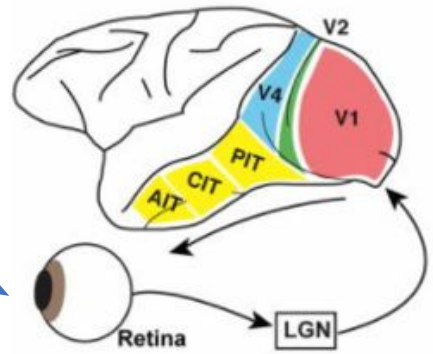


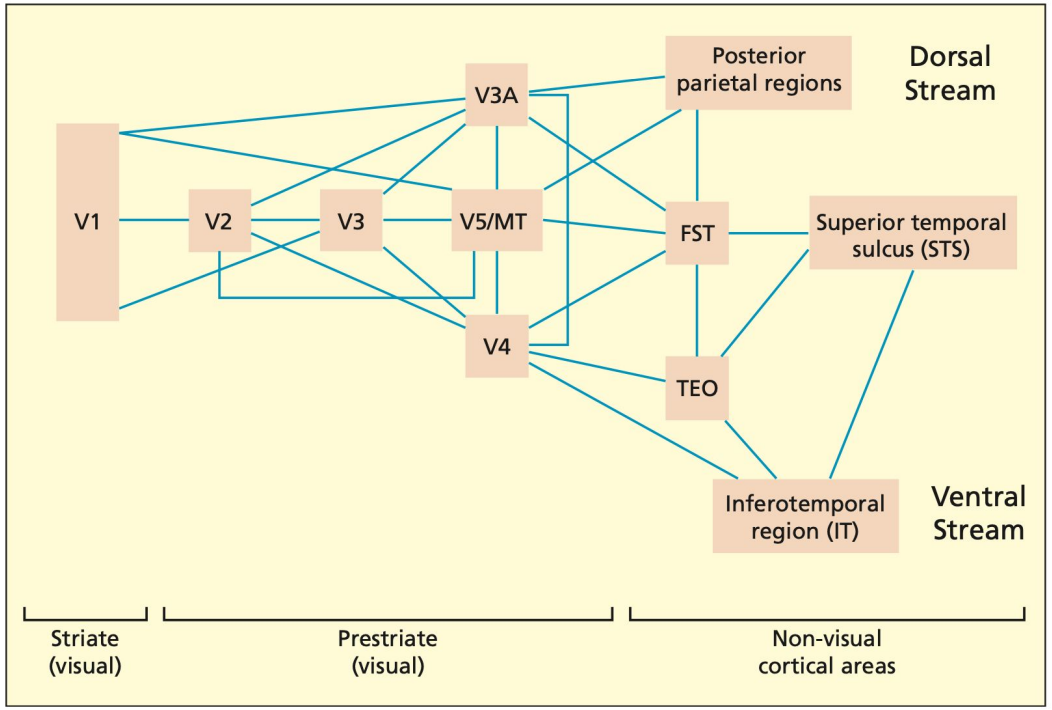
hierarchical architecture

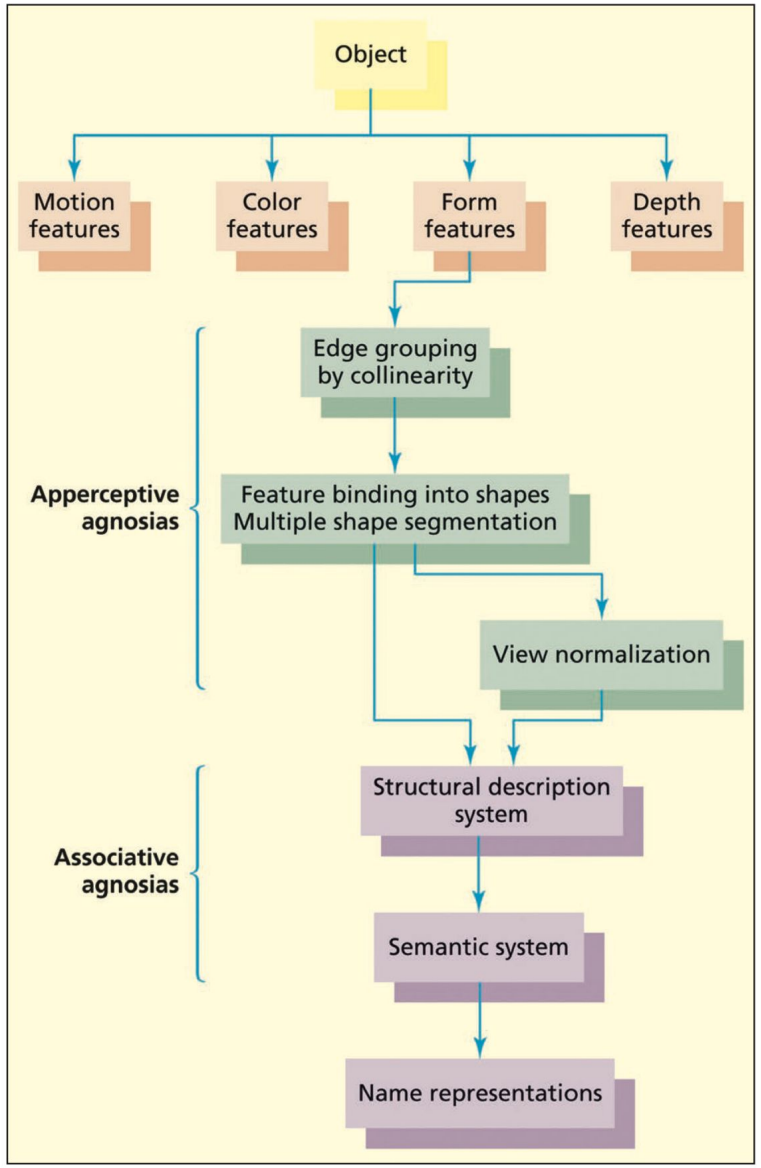
Invariant object representation theories [Hummel12]

- **View-based**: combinations/transformations of 2D views
  e.g., vector of 2D retinal coordinates [Edelman&Intrator03]
- **Structural description based**: 3D parts-based representation
  e.g., recognition-by-components [Biederman87]

Object Recognition

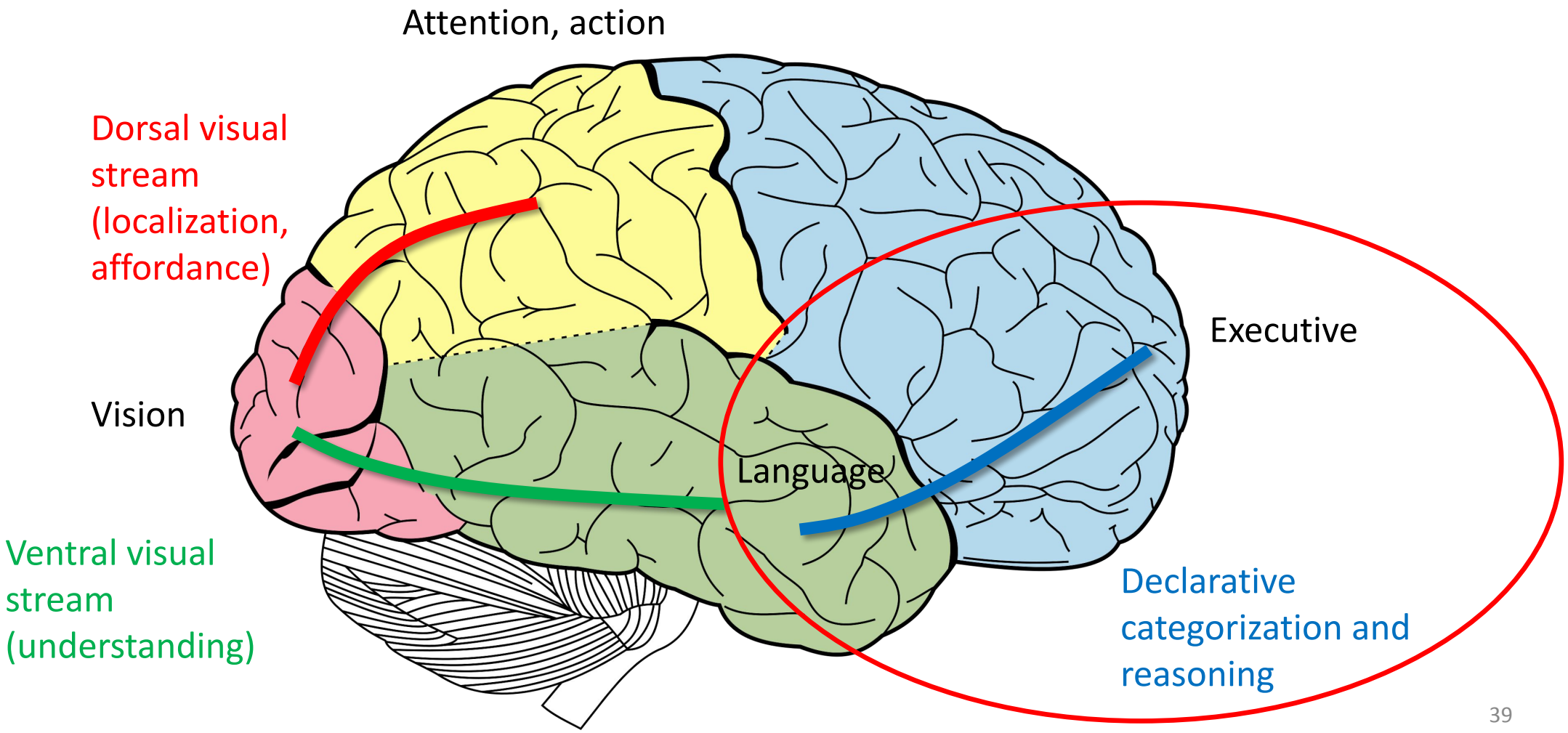






[Ward19]

Major Brain Areas – Beyond Vision



Object Categorization

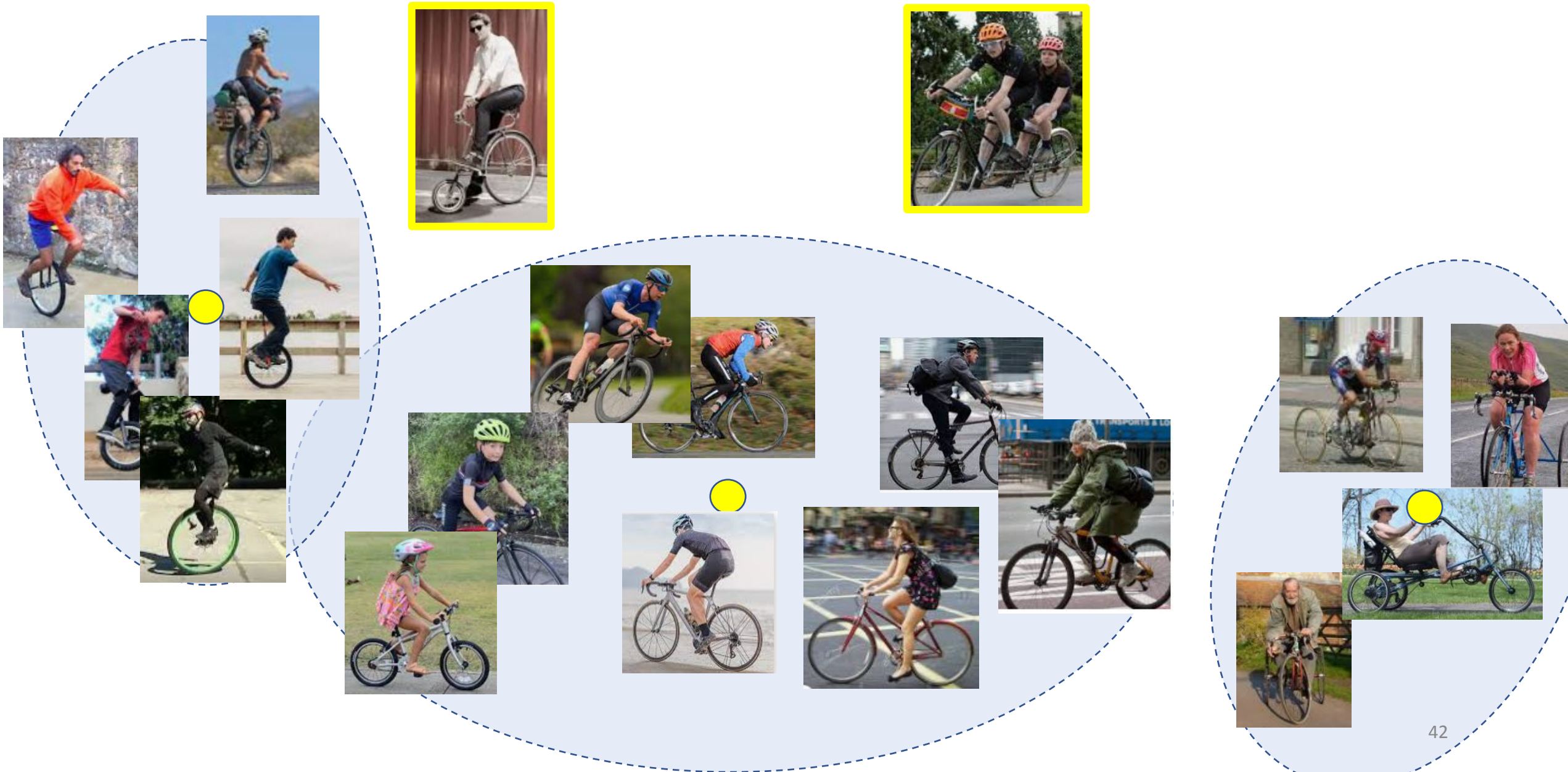
- Object categorization (Type 1 & 2)
  - the process of grouping objects based on similar/shared features [Goldstone17]
- Neurocognitive foundations of object categorization
  - Procedural categorization learning (Type 1) and declarative categorization learning (Type 2) [Ashby17]
- Focus is on concept learning and use
  - concept = mental representation of a category
  - studied in cognitive psychology, cognitive linguistics

Object Categorization Theories

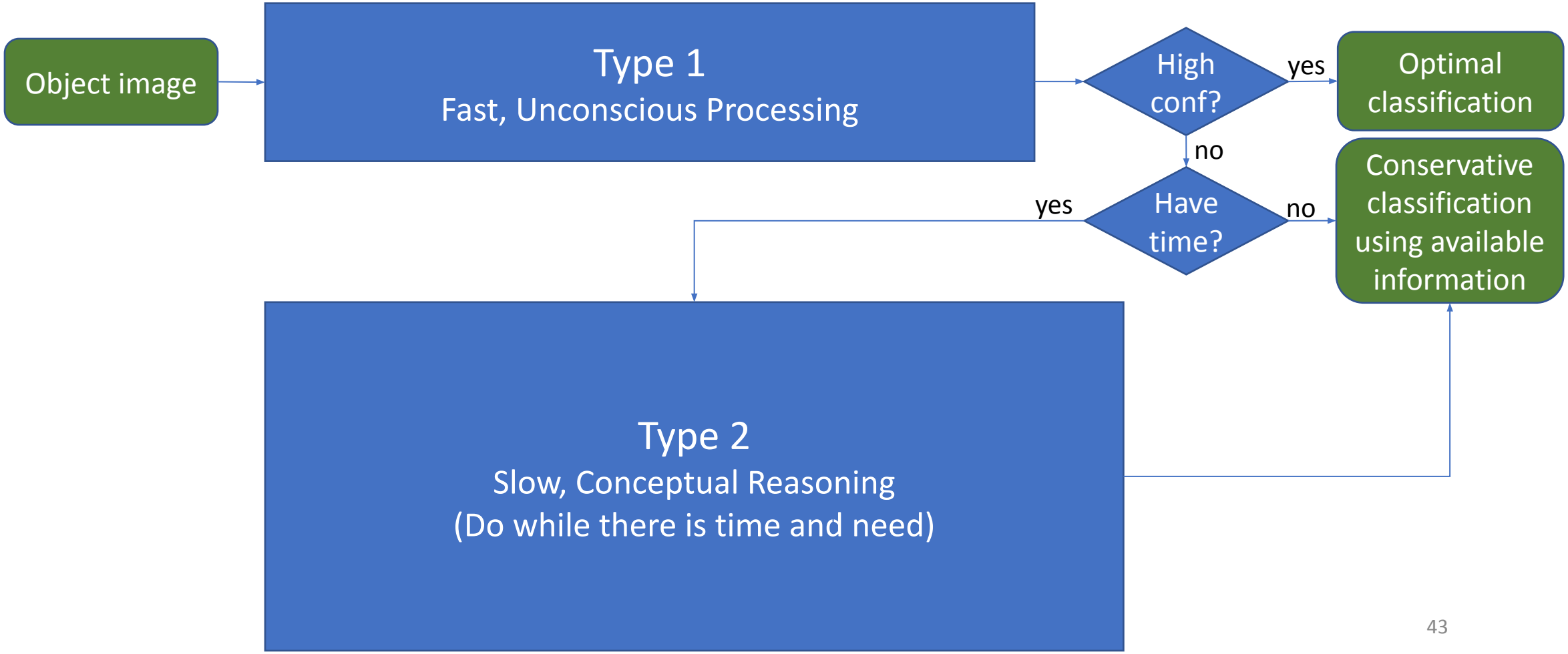
How are concepts represented in the human mind?

- **Rule** (classical): membership in terms of necessary and sufficient features
- **Prototype** [Rosch73]: membership by similarity to a prototype (summary)
  - instances have “family resemblances” (Wittengstein)
  - some instances are more central (prototypical) than others
- **Exemplar** [Medin&Schaffer78]: membership by collective similarity to exemplars
  - exemplars are individually memorized examples of category instances
- Each approach has limitations and recent opinion is a hybrid approach [Murphy16]
  - multi-categorization assigns instance to class with best fit

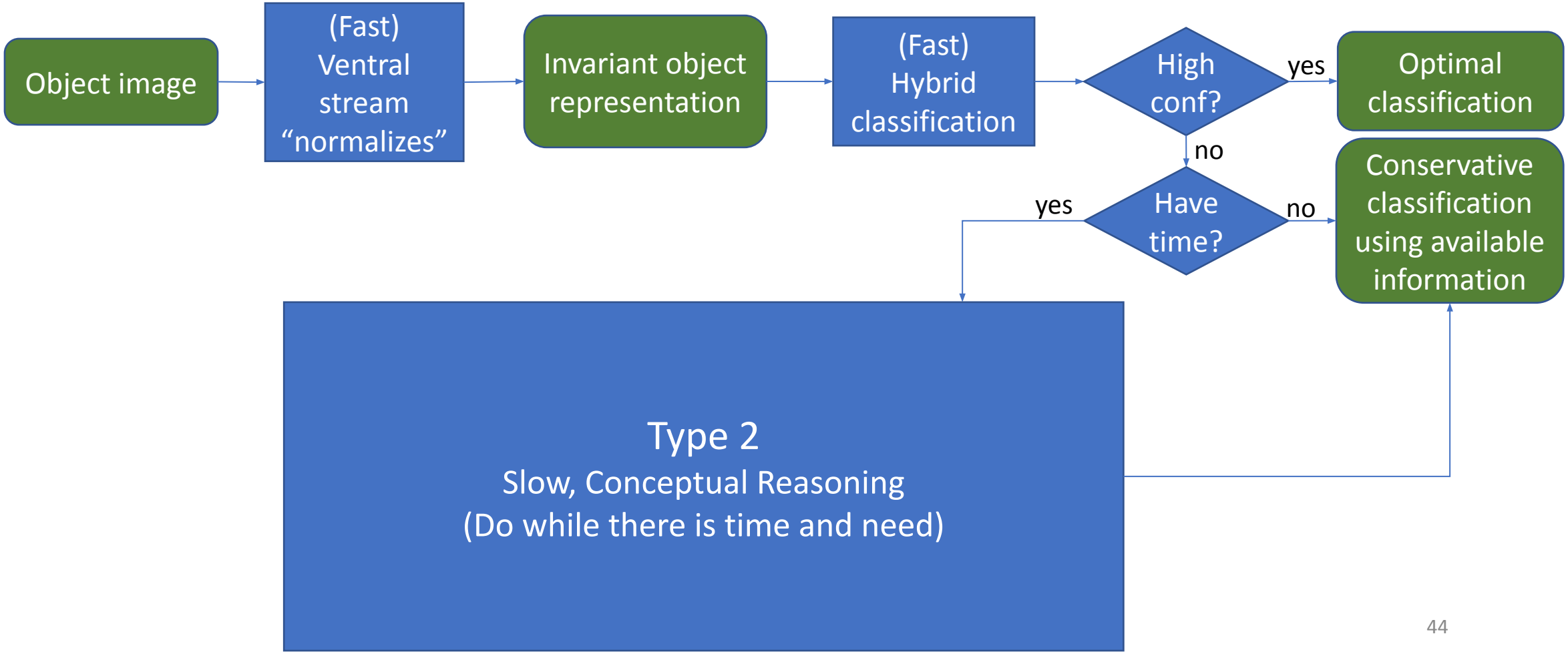
Hybrid prototype and exemplar Cyc classification



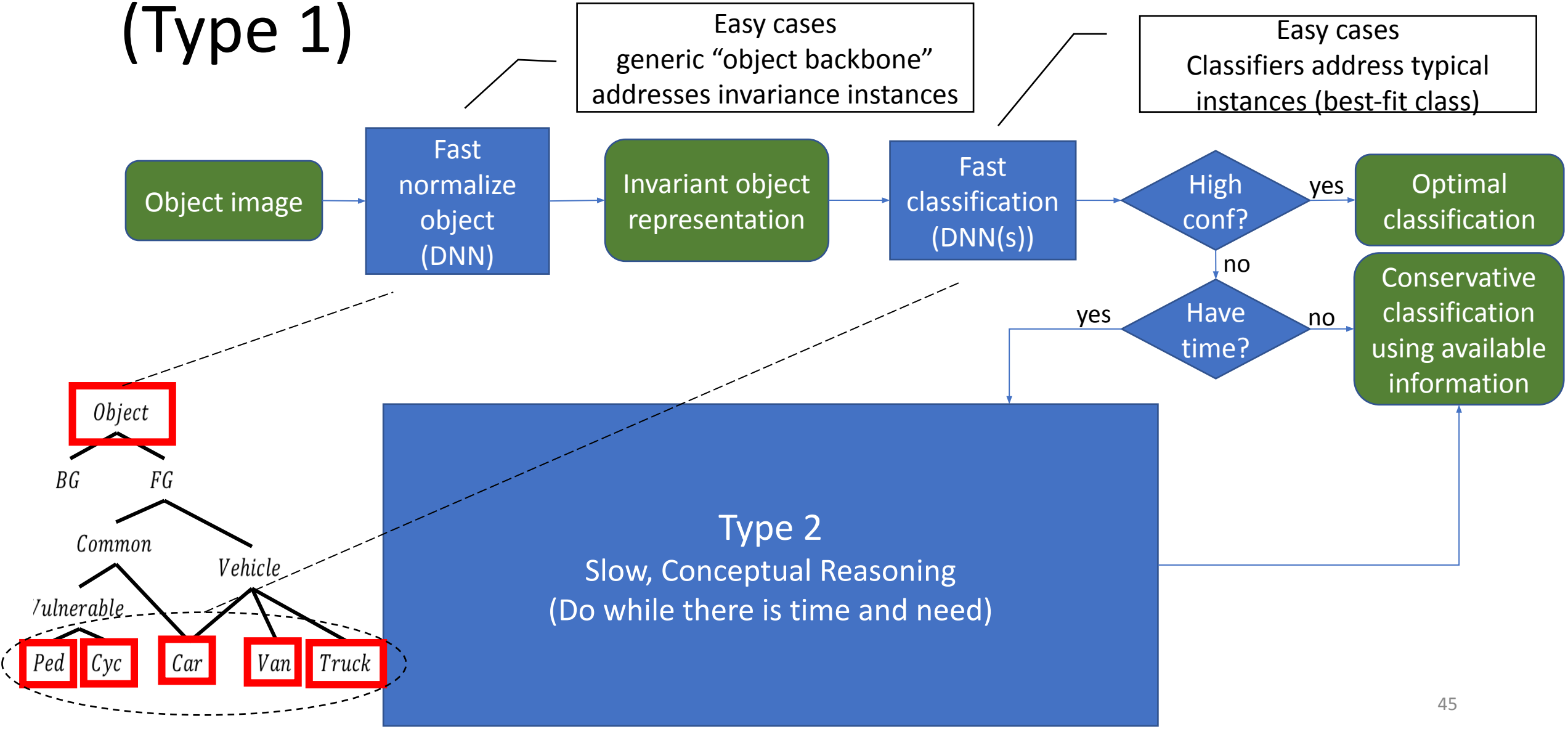
Safe human-inspired classification workflow



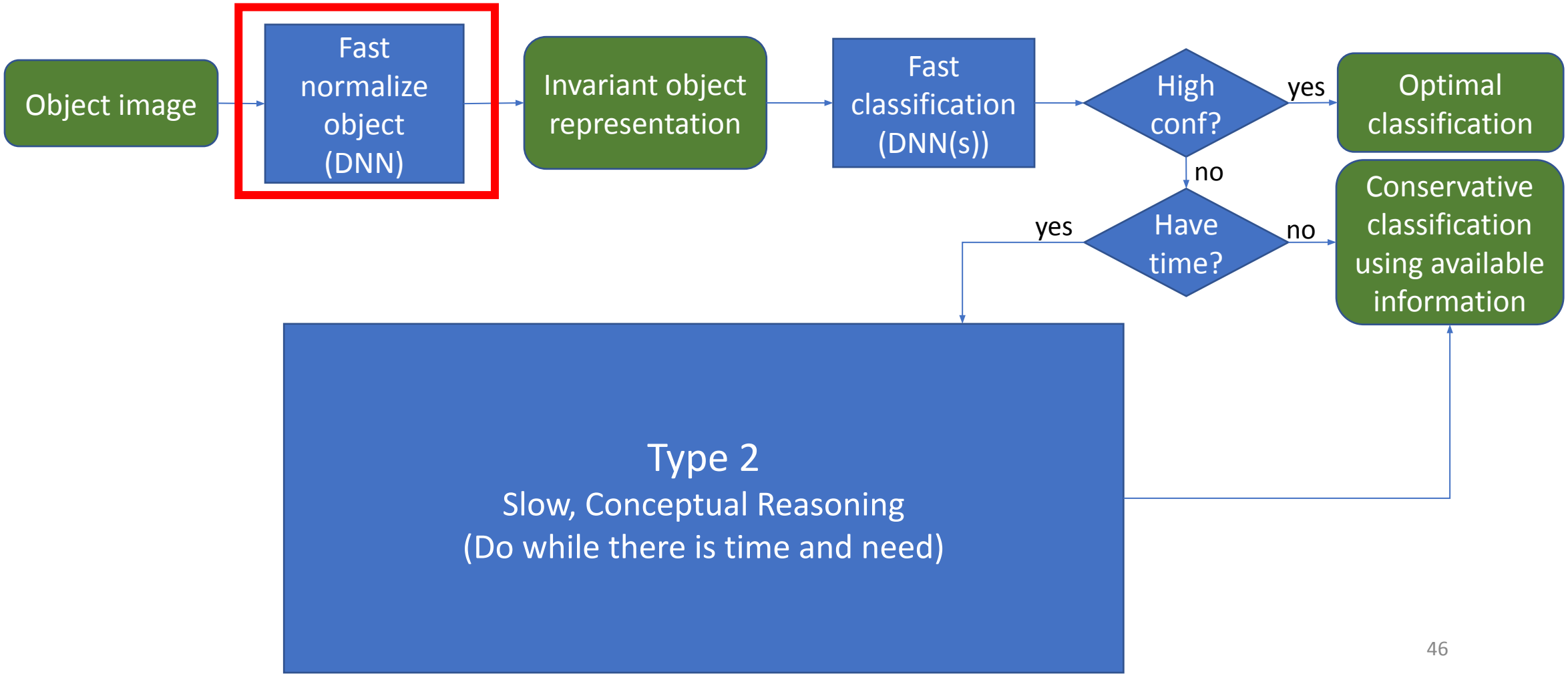
Safe human-inspired classification workflow (Type 1)



Safe human-inspired classification workflow (Type 1)



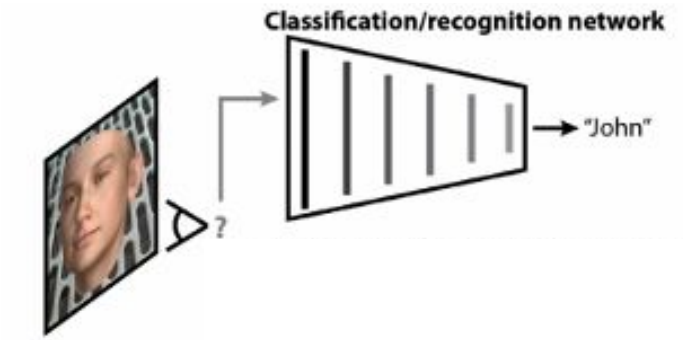
Safe human-inspired classification workflow (Type 1)



Object Normalization: Inverse Graphics

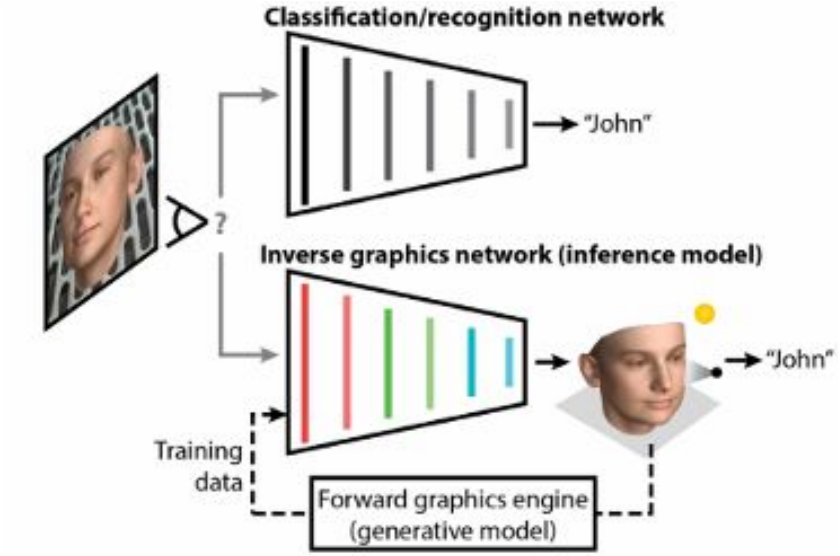
- Idea: invert the graphics computing rendering process of producing an image from a specification of 3D mesh, lighting, texture, etc.
- 3d-aware scene manipulation via inverse graphics
  - Yao, Shunyu, Tzu Ming Harry Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, William T. Freeman, and Joshua B. Tenenbaum. "." *arXiv preprint arXiv:1808.09351* (2018).
- Taking inverse graphics seriously
  - Hinton, Geoffrey (2013), capsule networks
- Cerberus: A multi-headed derenderer.
  - Deng, Boyang, Simon Kornblith, and Geoffrey Hinton. *arXiv preprint arXiv:1905.11940* (2019).
- Efficient inverse graphics in biological face processing.
  - Yildirim, Ilker, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. *Science advances* 6, no. 10 (2020)
  - Uses a DCNN/GAN architecture to simulate ventral stream processing and shows it compares with human processing

Object Normalization using Inverse Graphics



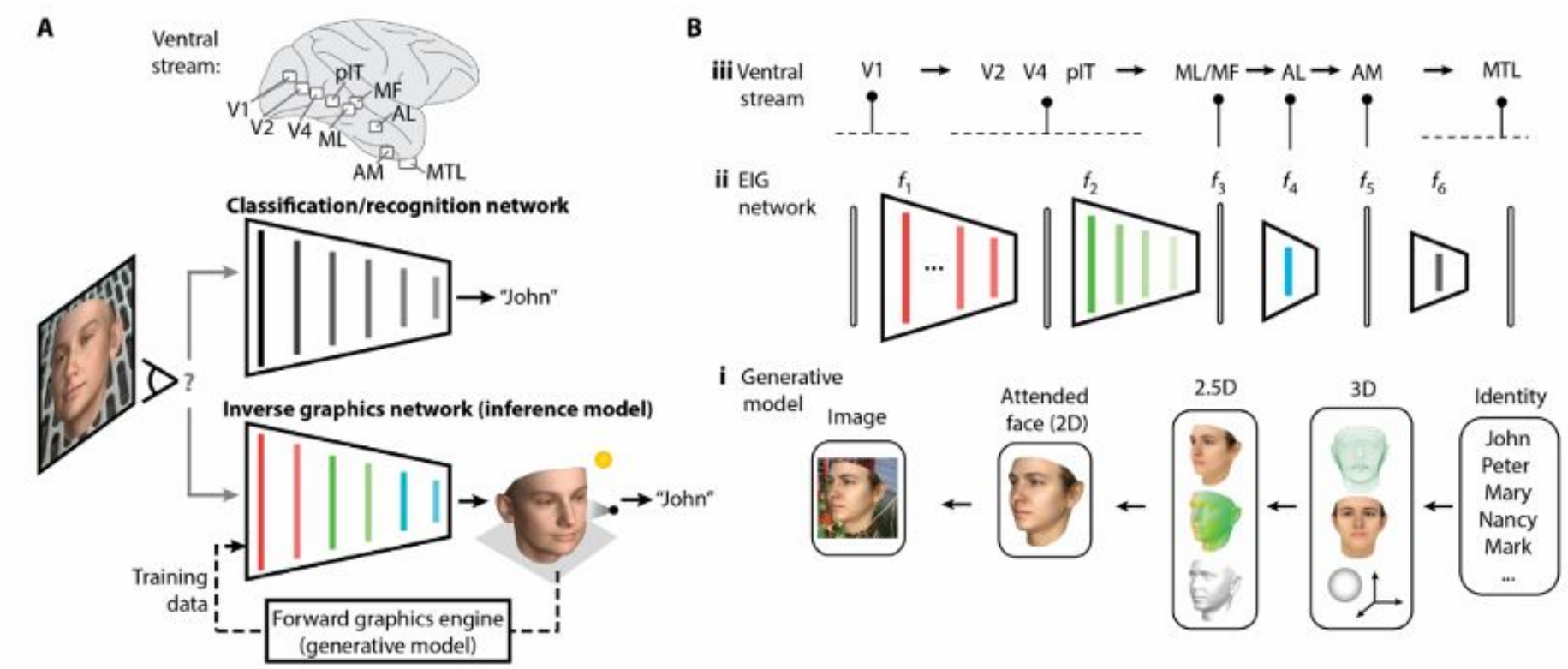
- Efficient inverse graphics in biological face processing.
  - Yildirim, Ilker, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. *Science advances* 6, no. 10 (2020)
  - Uses a DCNN/GAN architecture to simulate ventral stream processing and shows it compares with human processing

Object Normalization using Inverse Graphics



- Efficient inverse graphics in biological face processing.
  - Yildirim, Ilker, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. *Science advances* 6, no. 10 (2020)
  - Uses a DCNN/GAN architecture to simulate ventral stream processing and shows it compares with human processing

Object Normalization using Inverse Graphics



- Efficient inverse graphics in biological face processing.
  - Yildirim, Ilker, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. *Science advances* 6, no. 10 (2020)
  - Uses a DCNN/GAN architecture to simulate ventral stream processing and shows it compares with human processing

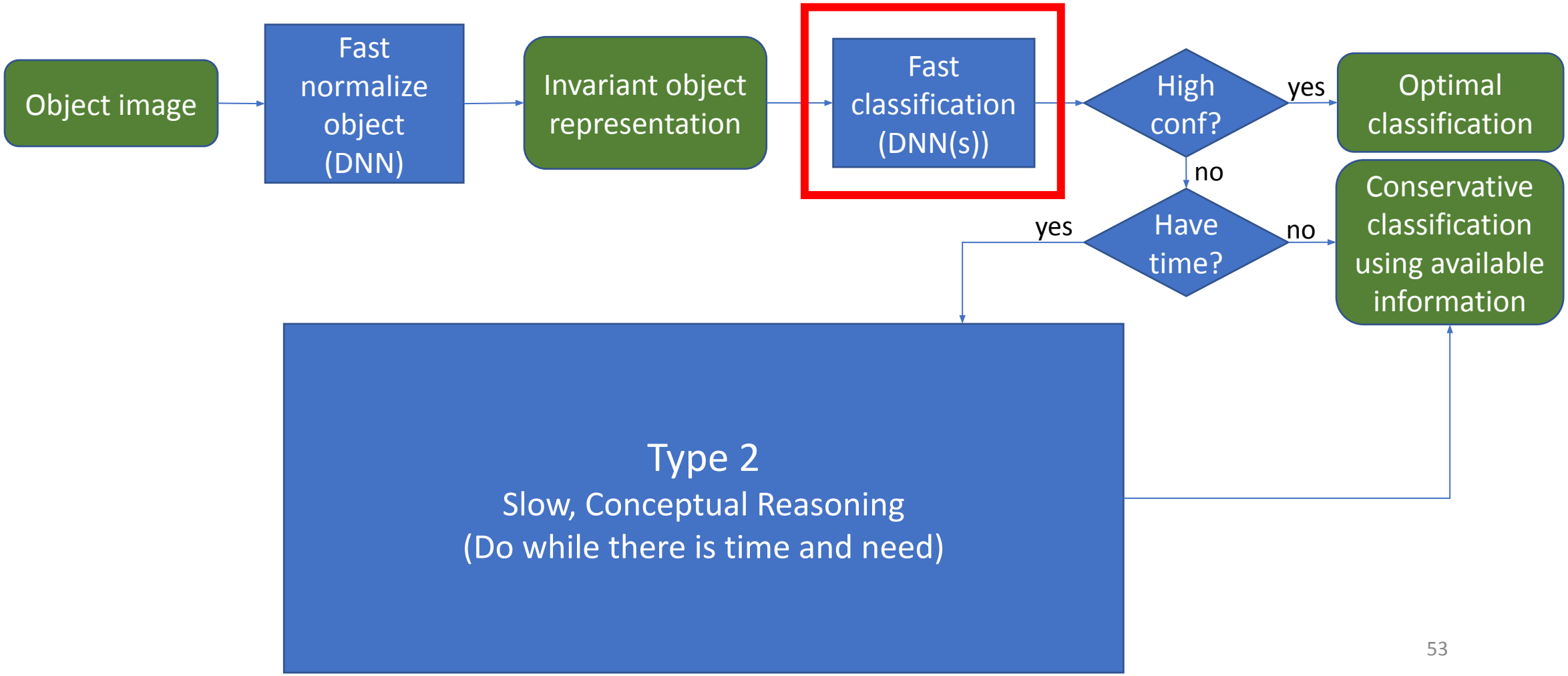
Why is human object recognition invariant?

- Generic knowledge about object classes is learned in an *embodied* way by interacting with objects [Smith05]
  - not by looking at a dataset of images!
- Embodied learning includes
  - image invariants – image-taking configuration change doesn't affect the object class
    - variability in position, scale, pose, illumination, clutter, etc.
  - object invariants – object manipulations (generally) do not affect object class
    - adding/removing (non-essential) parts
    - putting object inside, on top, beneath another object, etc.
  - beyond invariants
    - occluding an object doesn't mean it is gone (object constancy)
    - objects are composed of parts which are themselves objects
    - commonsense physics of causal physical relationships between objects [Fischer16]
- Knowledge obtained this way represents an "Embodiment Prior"
  - Assurance: although not interpretable, knowledge is obtained similarly to humans

Object Normalization using Embodied AI

- Idea: Train a DNN for object invariances by allowing it to manipulate objects in a simulated environment rather than giving it a training dataset of (augmented) images
  - really, a version of active learning
- A Survey of Embodied AI: From Simulators to Research Tasks
  - Duan, Jiafei, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan, arXiv preprint arXiv:2103.04918 (2021)
  - Surveys nine recent simulators developed for embodied AI
- Retrospective learning of spatial invariants during object classification by embodied autonomous neural agents
  - Caudell, Thomas P., Cheri T. Burch, Mustafa Zengin, Nathan Gauntt, and Michael J. Healy. In The 2011 International Joint Conference on Neural Networks
  - learns invariants by approaching objects in different ways in simulation

Safe human-inspired classification workflow (Type 1)



Hybrid prototype and exemplar Cyc classification

Prototype/Exemplar based Classifiers

- Prototype
  - Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions
    - Li, Oscar, Hao Liu, Chaofan Chen, and Cynthia Rudin.." *arXiv preprint arXiv:1710.04806* (2017).
  - Interpretable Image Recognition with Hierarchical Prototypes
    - Hase, Peter, Chaofan Chen, Oscar Li, and Cynthia Rudin. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, pp. 32-40. 2019.
- Exemplar
  - Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning
    - Papernot, Nicolas, and Patrick McDaniel. *arXiv preprint arXiv:1803.04765* (2018).
  - This looks like that: deep learning for interpretable image recognition
    - Chen, Chaofan, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. In *Advances in neural information processing systems*, pp. 8930-8941. 2019.
- Why do classification this way? Interpretability, few-shot learning!

This looks like that: deep learning for interpretable image recognition.

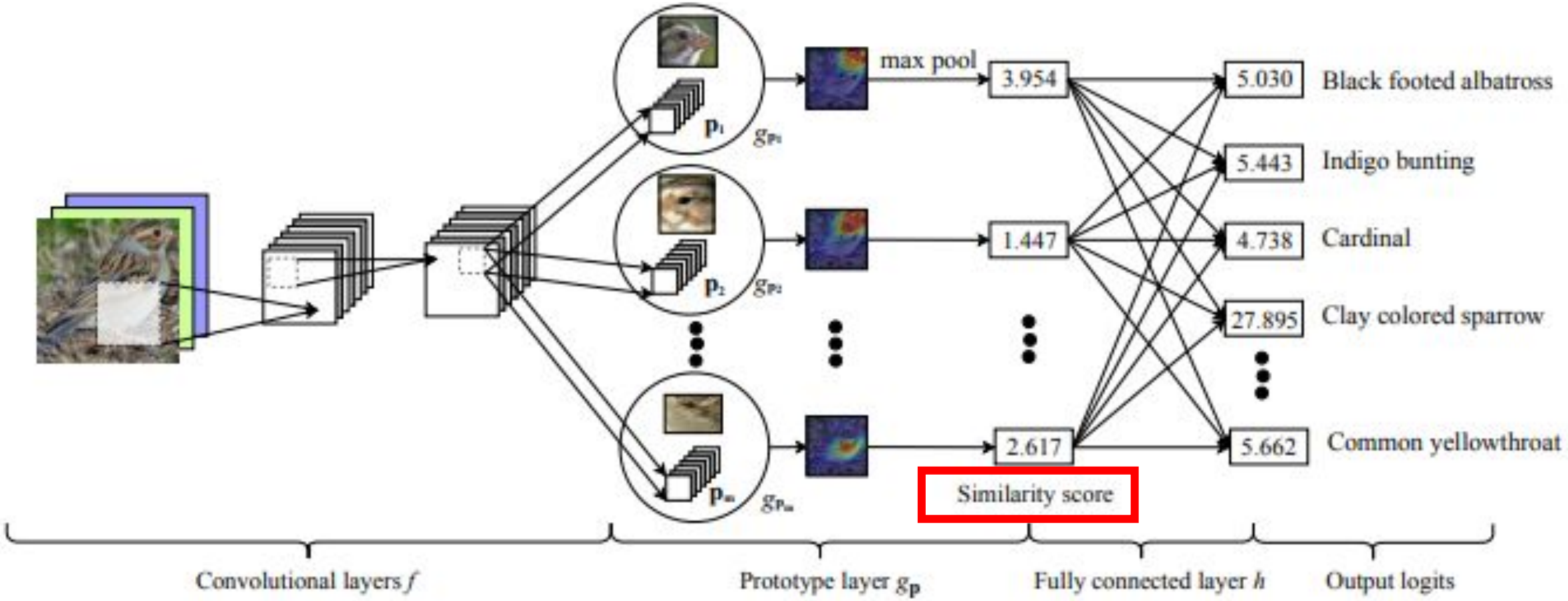
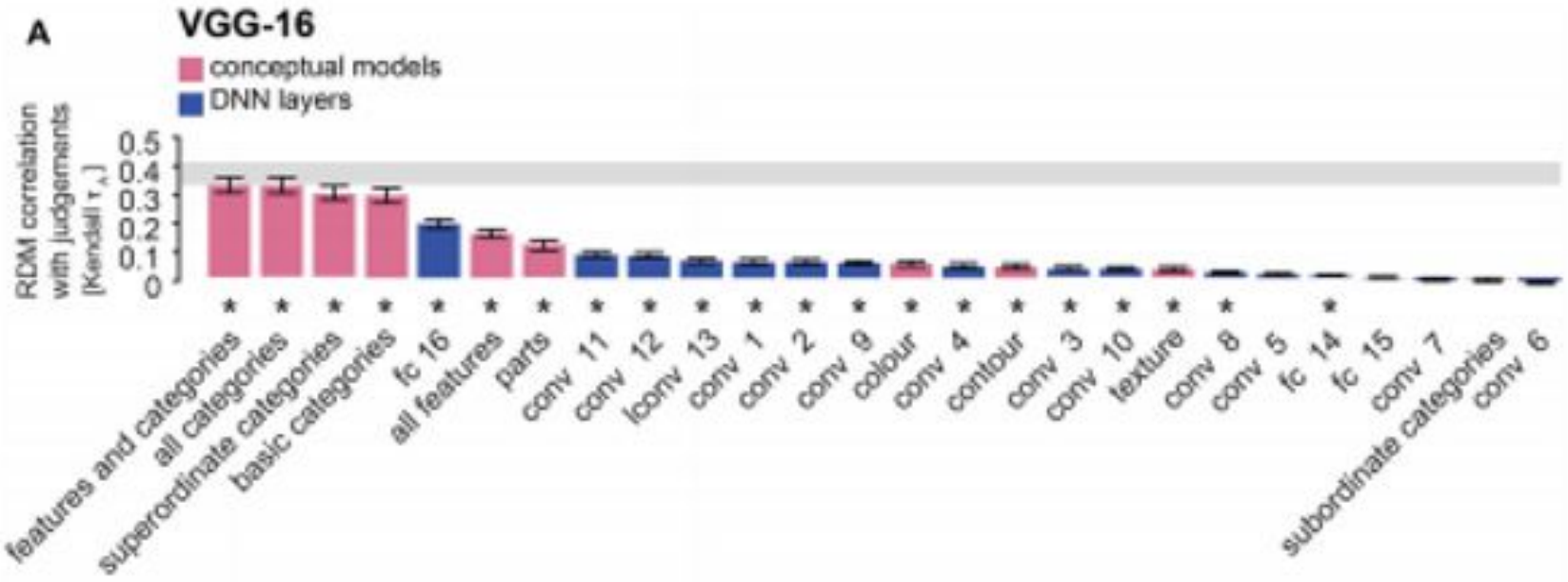


Figure 2: ProtoPNet architecture.

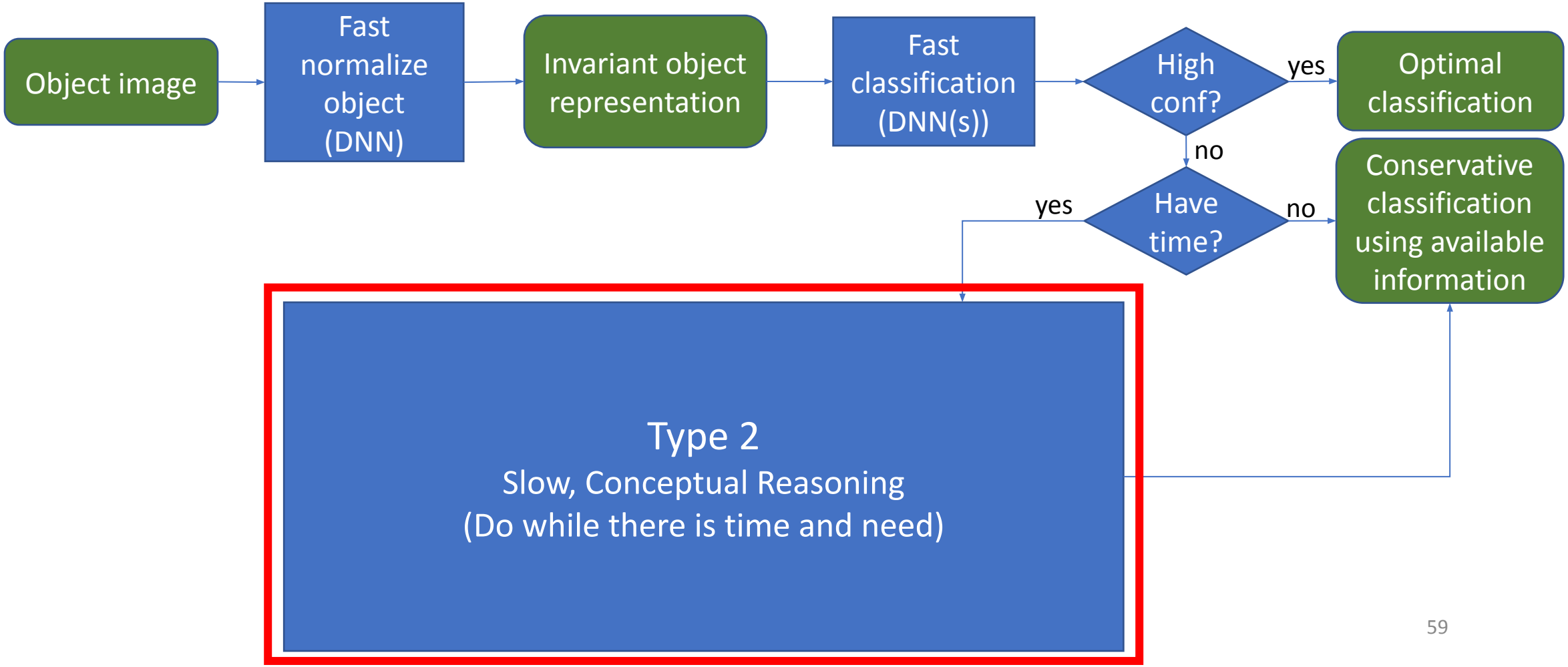
Similarity Judgement

- Prototype and Exemplar approaches depend on *similarity judgement*
- Various theories about human similarity judgement [Goldstone12]
  - Geometric, Feature-based, Alignment-based, Transformation-based
- Can DNN's do similarity judgement like humans? Yes and No. [Jozwik17]
  - Compared human image similarity judgements to image-image distance metrics using DNN feature representations and using human-understandable conceptual feature representations
  - DNNs correlated well but were outperformed by human-understandable categorical features



parts	colour	contour	texture	all categories
arm	black	arched	brick	animal
back	blonde	arrow	furry	armadillo
beard	blue	coiled	glass	artificial
branches	brown	curved	hairy	aubergine
brick	green	cylindrical	leafy	baboon
bristles	grey	domed	long	banana
building	red	pear-shaped	metallic	body part
cheeks	white	rectangular	plastic	bottle
chest	wooden	rounded	shiny	building
collar	woolly	spiky	spiky	camel
core	yellow	straight	stubbly	carnivore
dimples	socks/pink	symmetrical	sharp/scaly	carrots
dress	wet/water	cloves/bulbous	spire/ steeple/tall	chimpanzee
ear	purple/seat/wheels	round/circular	beak/feathers/feathery/wings	cold-blooded
eye		cubic/hob/square		courgette
eyelashes				crocodile
feet				dancer
fur				door

Safe human-inspired classification workflow (Type 1)



Question: Why is Type 2 really necessary for classification?

- CV ML assumption: object image class $\mathrm{I}C$ can be fully characterized using only *visual features*
  - visual feature = mathematical property of an image
- With enough training examples, can't a CV ML achieve arbitrarily good accuracy on $\mathrm{I}C$?
- Problem: classes used by humans and ADS correspond to human concepts
  - classes are conceptual rather than visual
  - also: socially constructed, subject to change across culture/time

E.g., Can Cyc be represented using only visual features?

- Observation:

  A cyclist is not defined by how it looks (visual features) but by what it consists of, how it works, (conceptual features) etc.!

- Cyc: has wheel(s), carries human rider(s), propelled by rider effort, etc.
- Precisely deciding Cyc (or ICyc) membership requires *reasoning about conceptual features*
  - parts, relationships, causality, commonsense physics [Fischer16], etc.

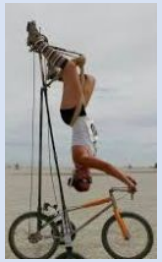
E.g., Can Cyc be represented using only visual features?

Visual features alone work for subsets of instances but performance is necessarily limited

FN: can always find unusual cyclists that fit conceptual description but not visual

✔ Wheel(s)	
✔ Human(s)	
✔ Rides	
✔ Propelled	

✔ Wheel(s)	
✔ Human(s)	
✔ Rides	
✔ Propelled	

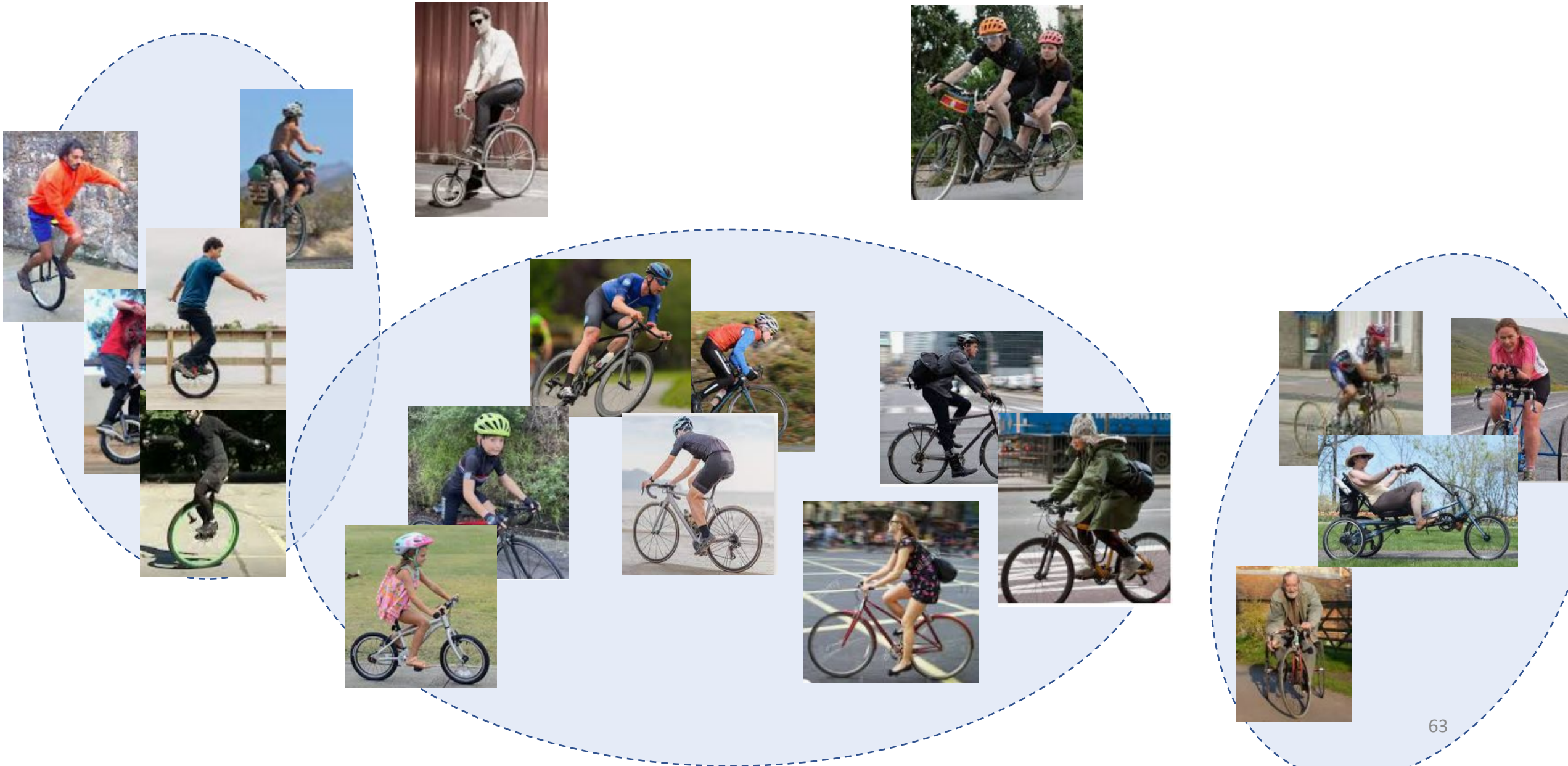


FP: can always find images that look like cyclists, but on careful inspection, aren't

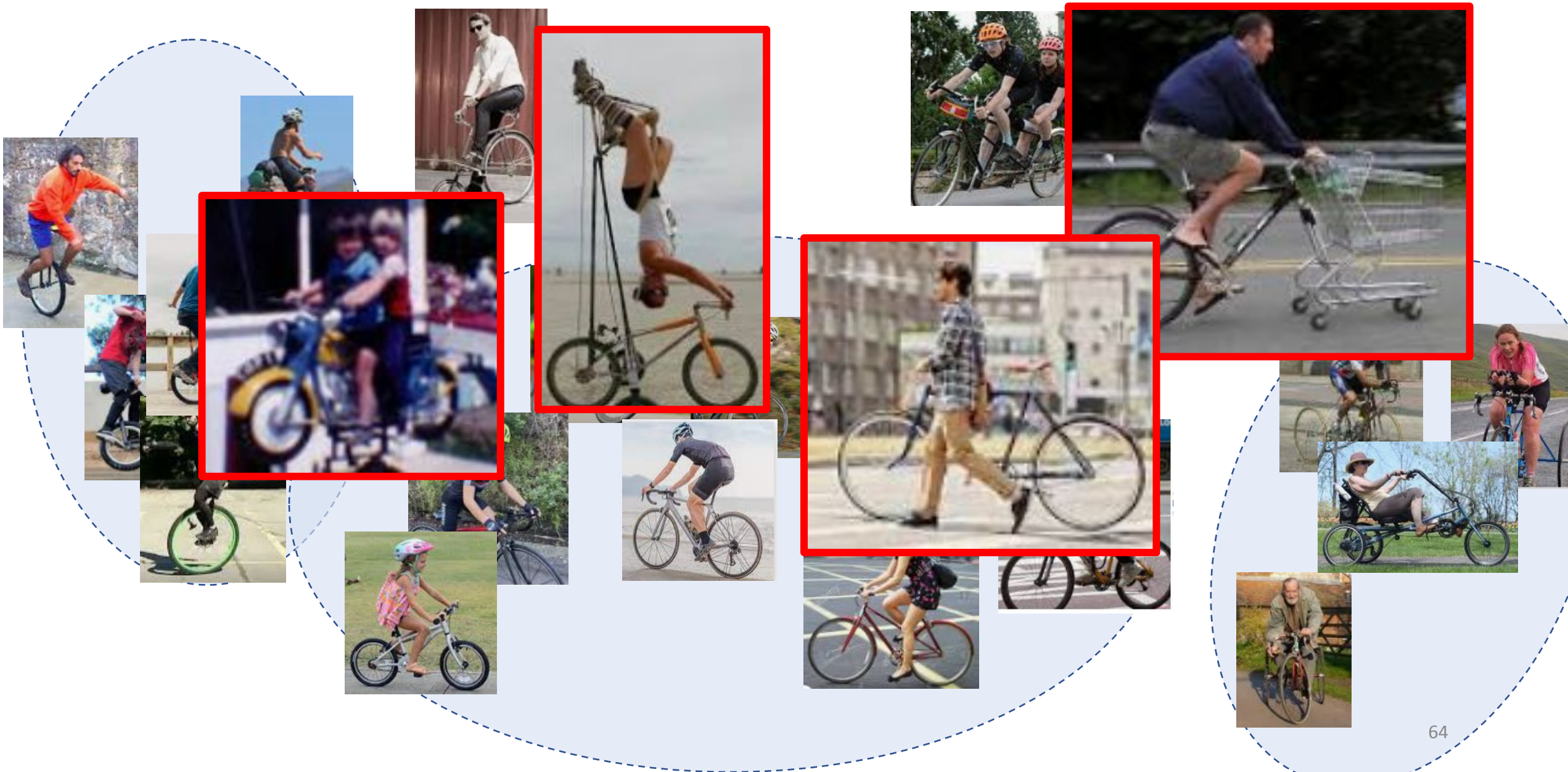
☐ Wheel(s)	
✔ Human(s)	
✔ Rides	
☐ Propelled	

✔ Wheel(s)	
✔ Human(s)	
☐ Rides	
✔ Propelled	

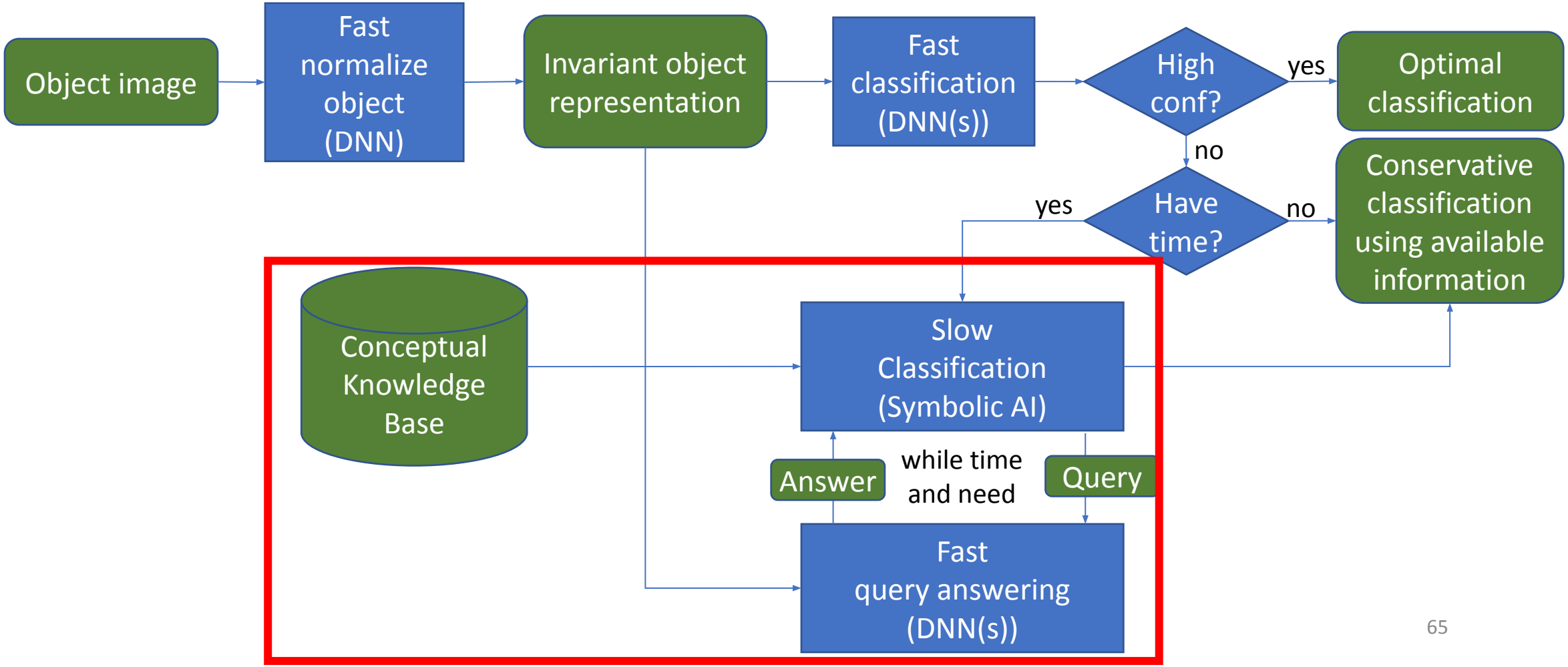
Visual feature clustering + exemplars only approximate a conceptual class



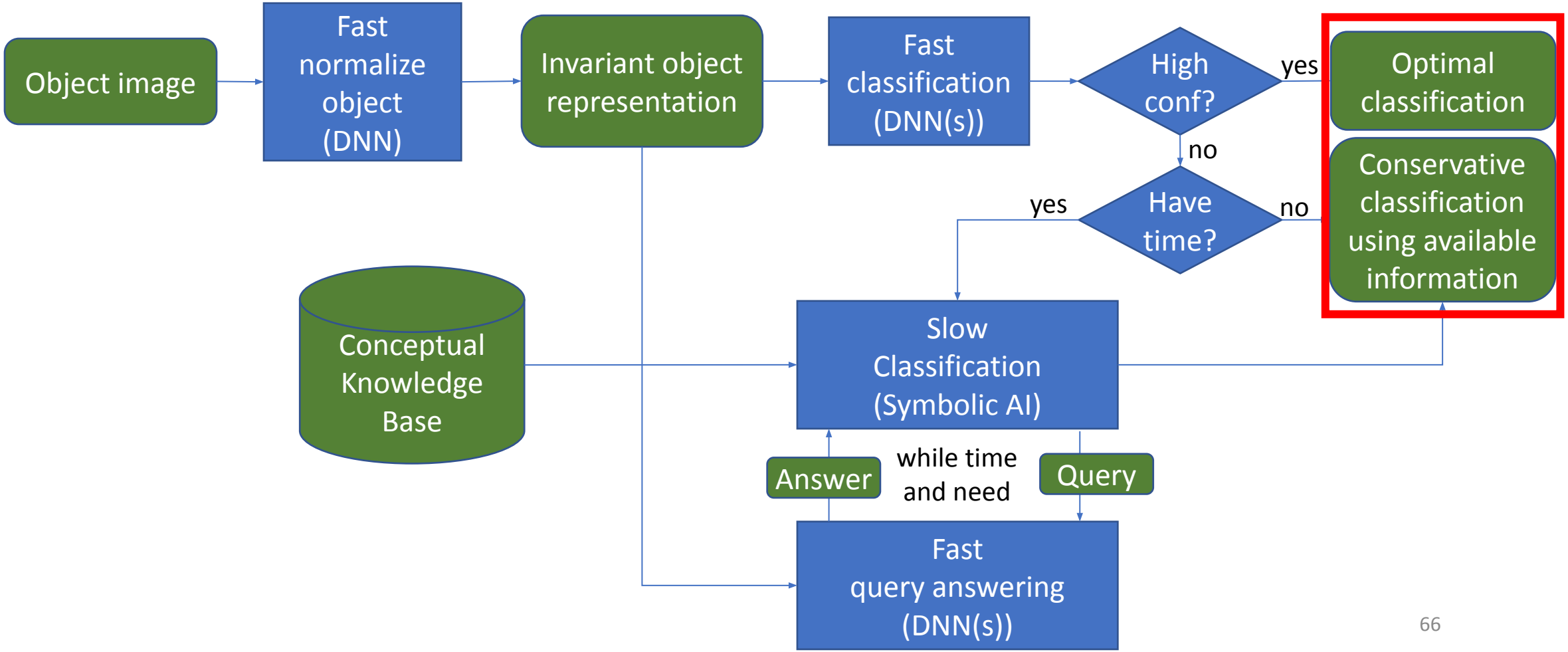
But why should these necessarily generalize correctly to unseen cases?



Safe human-inspired classification workflow



Safe human-inspired classification workflow



Human Object Image Classification

- Type 1 (Fast default classification)
  - High confidence
    - Easy cases: optimal (accurate) classification
    - Hard subtle cases: FP classification, undetected and potentially unsafe

  - Low confidence
    - Remaining hard cases: conservative (i.e., safe) classification
      - follows cautionary principle
    - Triggers Type 2 processing if need requires and time permits

- Type 2 (Slow classification)
  - Use reasoning to improving classification accuracy as more time invested

Guaranteeing safety

- Fast (type 1) classifier must have following properties
  1. Well-calibrated at high end: high-confidence => optimal (accurate) classification
  2. “Risk-consistent” with slow (type 2) classifier
     - conservative classification must at least as safe as classification by slow classifier
- How to get risk-consistency?
  - One way: ensure fast classifier is monotonic with slow classifier
    - fast classification must be same or superclass of slow classification
  - Risk-consistent since safe action for class is safe for all subclasses

References

- [Ashby17] FG Ashby, VV Valentin. Multiple systems of perceptual category learning: Theory and cognitive tests. Handbook of categorization in cognitive science, 2017: 157-188
- [Aven18] Aven, Terje. "How the integration of System 1-System 2 thinking and recent risk perspectives can improve risk assessment and management." Reliability Engineering & System Safety 180 (2018): 237-244.
- [Aven19] Aven, Terje. "The cautionary principle in risk management: Foundation and practical use." Reliability Engineering & System Safety 191 (2019): 106585.
- [Chen18] Chen, Chaofan, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. "This looks like that: deep learning for interpretable image recognition." arXiv preprint arXiv:1806.10574 (2018).
- [DiCarlo12]James J. DiCarlo, Davide Zoccolan, Nicole C. Rust. How Does the Brain Solve Visual Object Recognition?, Neuron, Volume 73, Issue 3, 2012: 415-434,
- [Epstein94] Epstein, Seymour. "Integration of the cognitive and the psychodynamic unconscious." American psychologist 49, no. 8 (1994): 709.
- [Evans13] Evans, Jonathan St BT, and Keith E. Stanovich. "Dual-process theories of higher cognition: Advancing the debate." Perspectives on psychological science 8, no. 3 (2013): 223-241.
- [Fabre11] Fabre-Thorpe, Michèle. "The characteristics and limits of rapid visual categorization." Frontiers in psychology 2 (2011): 243
- [Fischer18] Fischer, Jason, John G. Mikhael, Joshua B. Tenenbaum, and Nancy Kanwisher. "Functional neuroanatomy of intuitive physical inference." Proceedings of the national academy of sciences 113, no. 34 (2016): E5072-E5081
- [Hummel13] Hummel, John E. "Object recognition." Oxford handbook of cognitive psychology (2013): 32-46.
- [Jozwik17] Jozwik, Kamila M., Nikolaus Kriegeskorte, Katherine R. Storrs, and Marieke Mur. "Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments." Frontiers in psychology 8 (2017): 1726.
- [Kahneman11] Kahneman, Daniel. Thinking, fast and slow. Macmillan, 2011.
- [Murphy16] Murphy, Gregory L. "Is there an exemplar theory of concepts?." Psychonomic Bulletin & Review 23, no. 4 (2016): 1035-1042
- [Smith05] Smith, Linda, and Michael Gasser. "The development of embodied cognition: Six lessons from babies." Artificial life 11, no. 1-2 (2005): 13-29.
- [Thompson11] Thompson, Valerie A., Jamie A. Prowse Turner, and Gordon Pennycook. "Intuition, reason, and metacognition." Cognitive psychology 63, no. 3 (2011): 107-140.
- [Ward19] Jamie Ward. The Student's Guide to Cognitive Neuroscience. 4th Edition. Routledge, 2019

- [DiCarlo12] James J. DiCarlo, Davide Zoccolan, Nicole C. Rust. How Does the Brain Solve Visual Object Recognition?, Neuron, Volume 73, Issue 3, 2012, Pages 415-434,
- [Ward19] Jamie Ward. The Student's Guide to Cognitive Neuroscience. 4th Edition. Routledge, 2019
- [Ashby17] FG Ashby, VV Valentin. Multiple systems of perceptual category learning: Theory and cognitive tests. Handbook of categorization in cognitive science, 2017, Pages 157-188

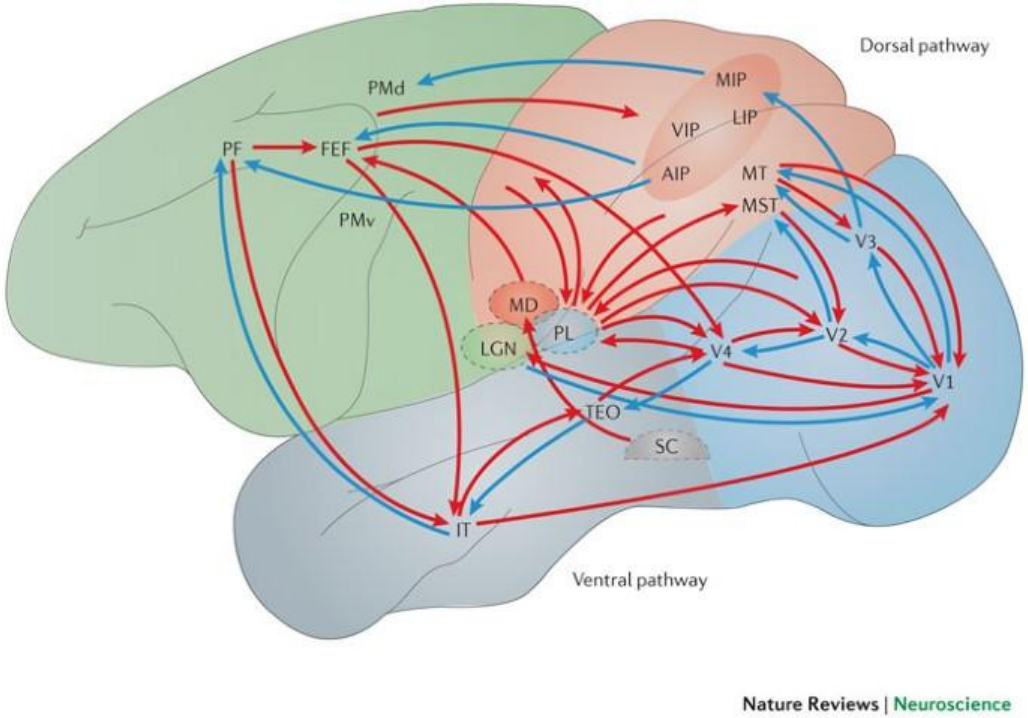
Apparent Visual Feature Paradox

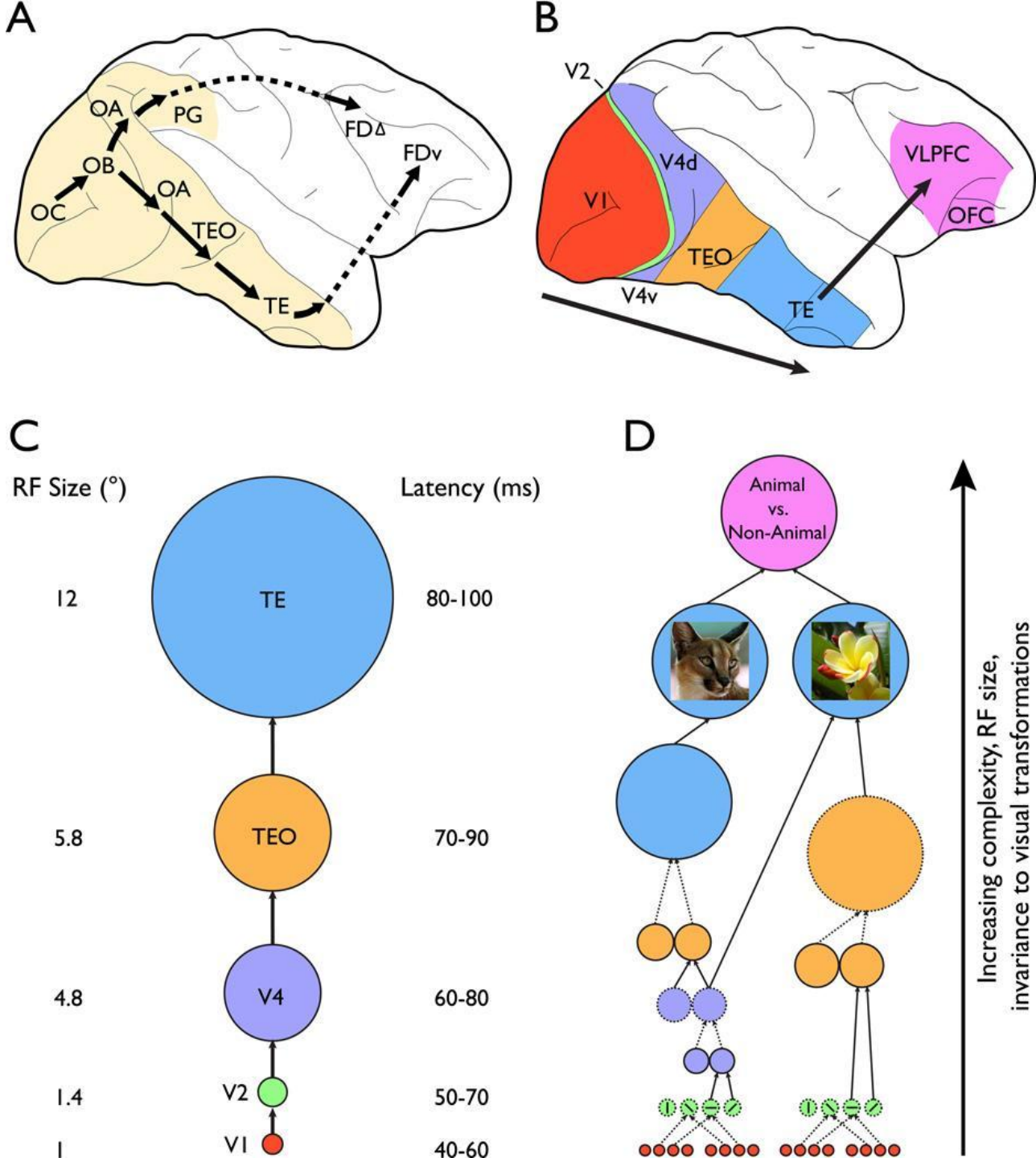
- Apparent Paradox
  - Visual features are not enough
    - Conceptual features are required for precise classification
  - We must check for conceptual features using visual feature queries
    - Since we can only get information about objects through senses (symbol grounding problem)
  - ∴ Visual features must be enough!
- Why not a paradox
  - A visual-feature approximation of ICyc only uses features local to typical cyclists
  - But the queries reachable using conceptual reasoning encompasses the entire knowledge base
    - Each query relies on local visual-feature based approximation + further conceptual reasoning

Risk Management by Humans

- Determining Risk
  - *Risk perception* is Type 1 processing (much Psych literature on this)
    - sensing/feeling uncertainties and the potentials for hazards
  - *Risk assessment* is Type 2 processing
    - a reasoned assessment of consequences, severities and occurrence probabilities
- Risk management in engineering (e.g., where to build a nuclear plant)
  - Determining risk is done using risk assessment only [Aven18]
    - argues that risk perception should be used as well
  - Risk assessment can be narrow and ignores uncertainties in assumptions that risk perception could detect
    - *Cautionary principle* is a norm [Aven19]: If the consequences of an activity could be serious and subject to uncertainties, then cautionary measures should be taken, or the activity should not be carried out
  - Risk perception without assessment can be biased and distorted
    - e.g., difference between 0 and 1 life lost feels different that between 500 and 501 lives lost

In Primates


Dorsal pathway
PMd
MIP
LIP
VIP
PF
FEF
AIP
MT
MST
PMv
V3
MD
PL
LGN
V4
V2
V1
TEO
SC
IT
Ventral pathway
Nature Reviews | Neuroscience


A
OA
PG
FDΔ
FDv
OB
OA
OC
TEO
TE
B
V2
V4d
VLPFC
V1
OFC
TEO
V4v
TE
C
RF Size (°)
Latency (ms)
12
TE
80-100
5.8
TEO
70-90
4.8
V4
60-80
1.4
V2
50-70
1
V1
40-60
D
Animal
vs.
Non-Animal
Increasing complexity, RF size,
invariance to visual transformations