Australian National University

Open-world Learning Tutorial COMP2550/4450/6445

Cheng Xue

Cheng.Xue@anu.edu.au

Icebreaking Time!

- Name?
- Bc/Honours/Masters? Which year?
- Where are you from?
- Ideally turn on your camera

About me:

About Me:

Cheng

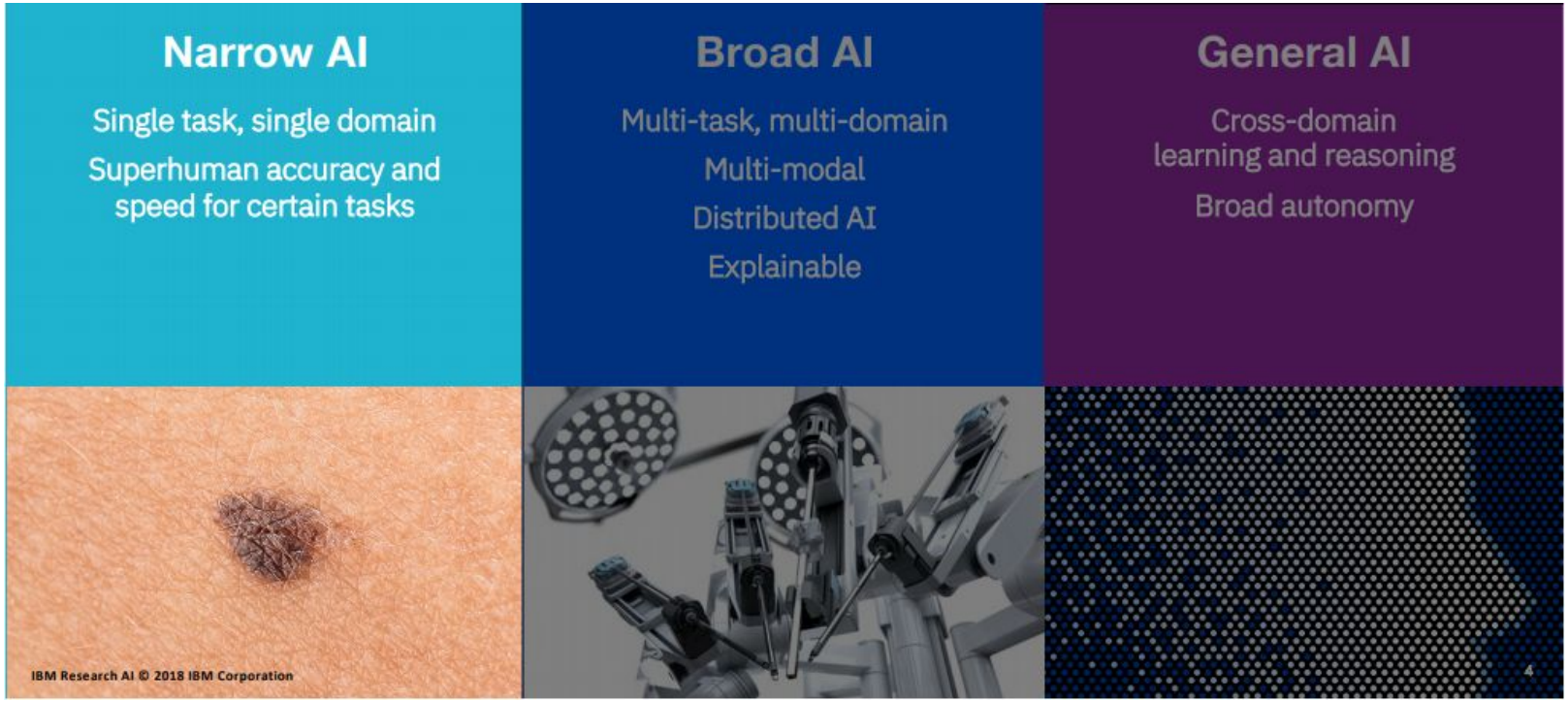
- PhD Candidate at School of Computing
- From China
- Master in Mathematics
- Currently work on the DARPA (Defense Advanced Research Projects Agency) SAILON project, which aims to build AI systems that can quickly detect novelty in an environment and adapt to it - open world learning.

- Email: Cheng.Xue@anu.edu.au

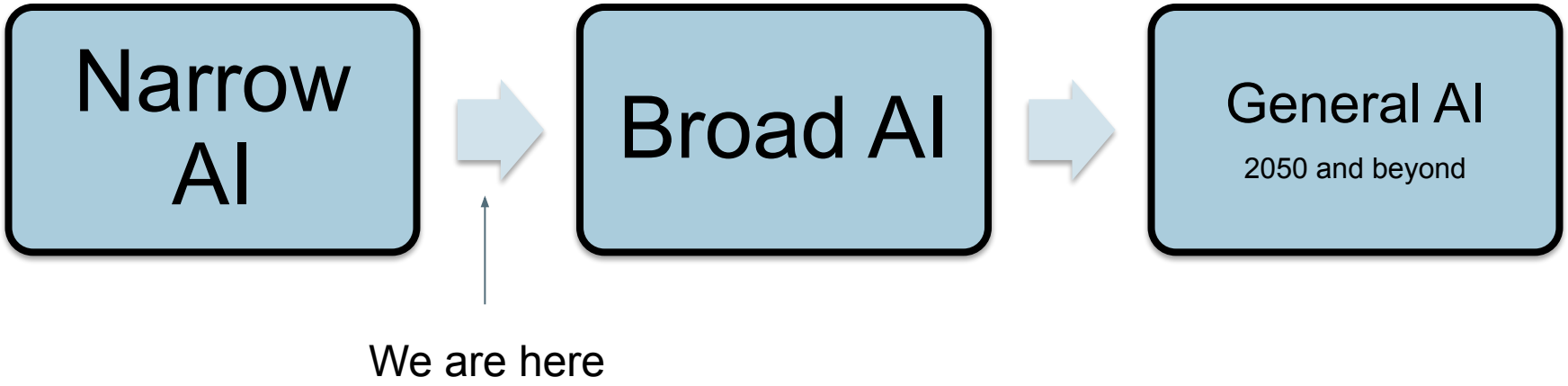
- Questions related to the topics, research, tutorial (are the papers too hard? Too easy?) or course in general are welcomed.

Let’s talk about Artificial Intelligence

The Evolution of AI



The Evolution of AI



Current AI Era

- What is “narrow” about today’s AI toolbox?

Caption Generation

man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

Karpathy and Li, 2015

Deep Reinforcement Learning

Generative Models

Gatys et al. 2015

Brock et al. 2018

Are we DOOMED ?

Elon Musk: artificial intelligence is our biggest existential threat

The AI investor says that humanity risks 'summoning a demon' and calls for more regulatory oversight

SkyNet

Stephen Hawking warns artificial intelligence could end mankind

By Rory Cellan-Jones
Technology correspondent

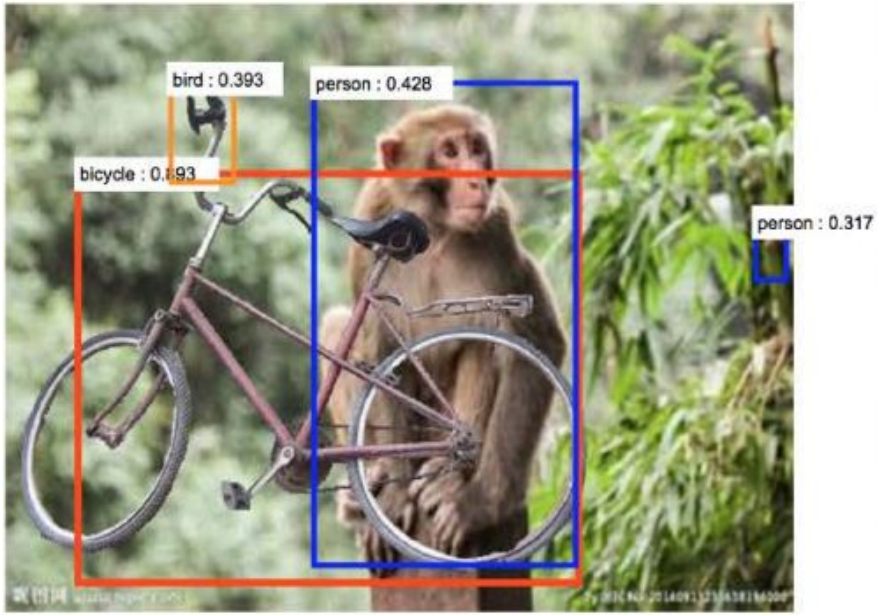
Why we still call it narrow AI

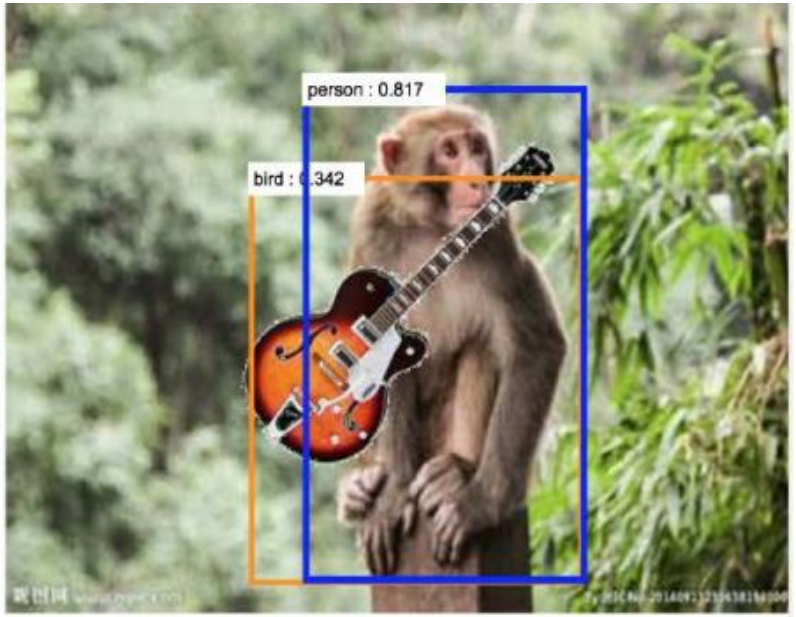
Why we still call it narrow AI

“Teddy Bear”

Meret Oppenheim, *Le Déjeuner en fourrure*

Object Recognition is solved?

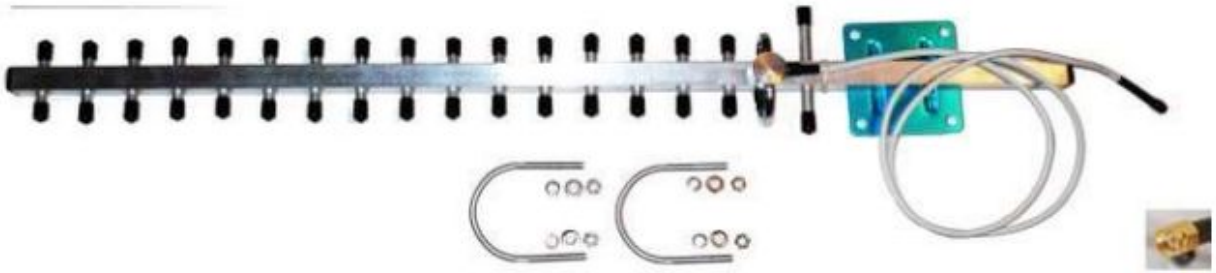




Wang et al. 2018

Need huge datasets to work

What's this?



Which one is the object?

How many objects are there?

Even ImageNet (with 14 million images) has problems

IMAGENET




MIT-IBM Watson AI Lab

ObjectNet

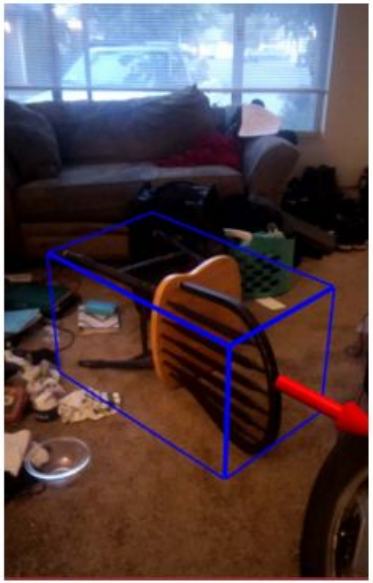
Boris Katz
MIT

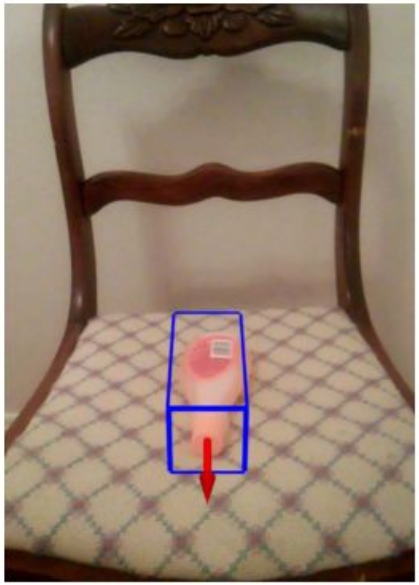
Andrei Barbu
MIT

Dan Gutfreund
IBM

#MITIBM #AI

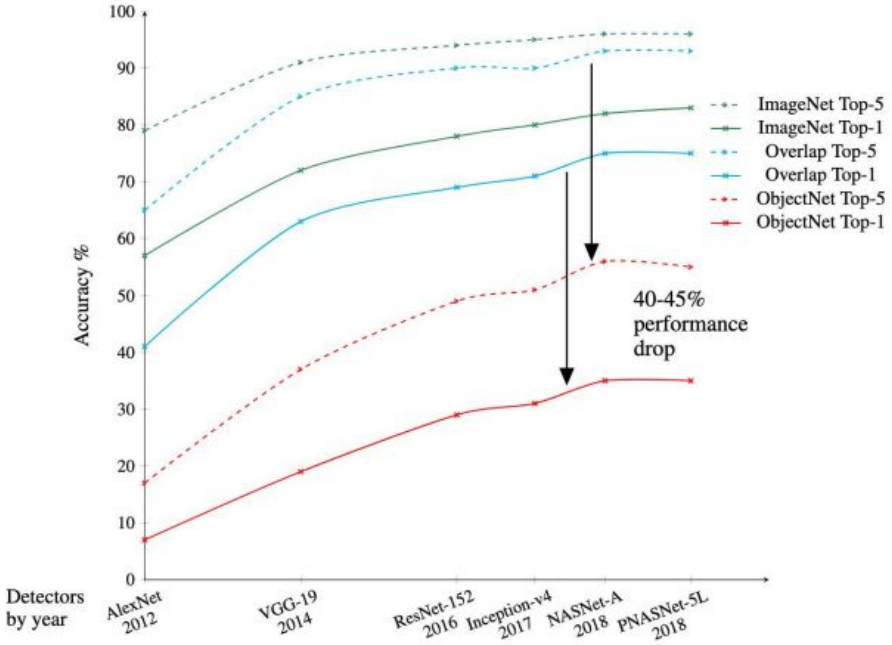
ObjectNet

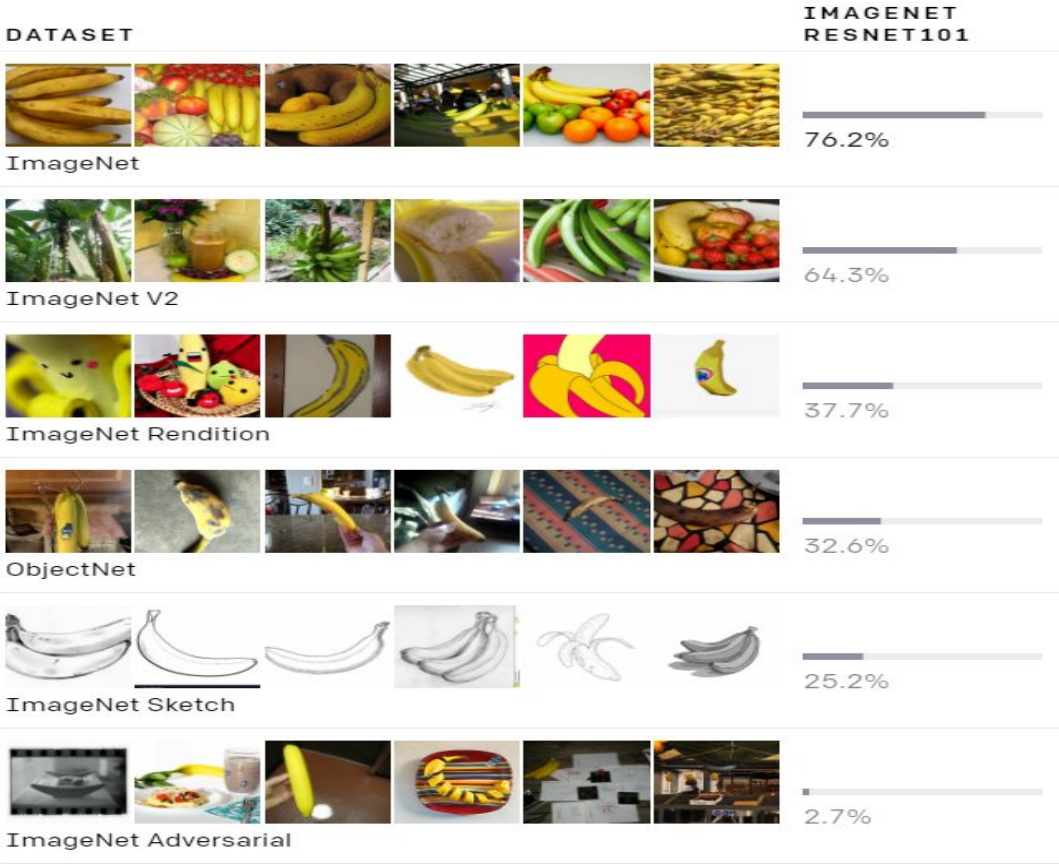




- ~50K images
- ~300 object classes
- 4 different room types

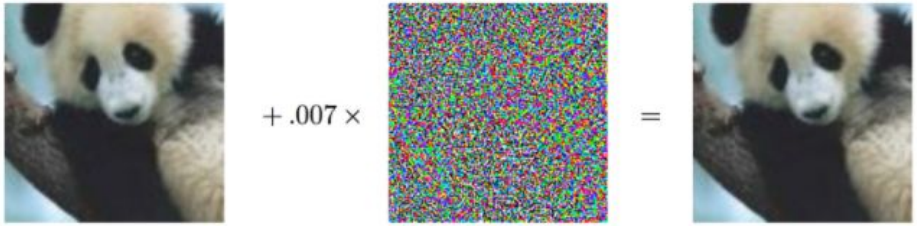
Testing ImageNet-trained models on ObjectNet





OpenAI Clip 2019

Deep Learning can be hacked



Original Top-3 inferred captions:

1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

Adversarial Top-3 captions:

1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

Chen et al. 2018

Generating adversarial patches against YOLOv2

- https://www.youtube.com/watch?v=MIbFvK2S9g8

Even with standard datasets, best training condition, no adversarial attack, well-placed objects, Deep learning models still struggle.

How many blocks are on the right of the three-level tower?

Will the block tower fall if the top block is removed?

Are there more trees than animals?

What is the shape of the object closest to the large cylinder?

Deep learning is spreadsheet on steroids and we are far from declaring victory.

-DARPA

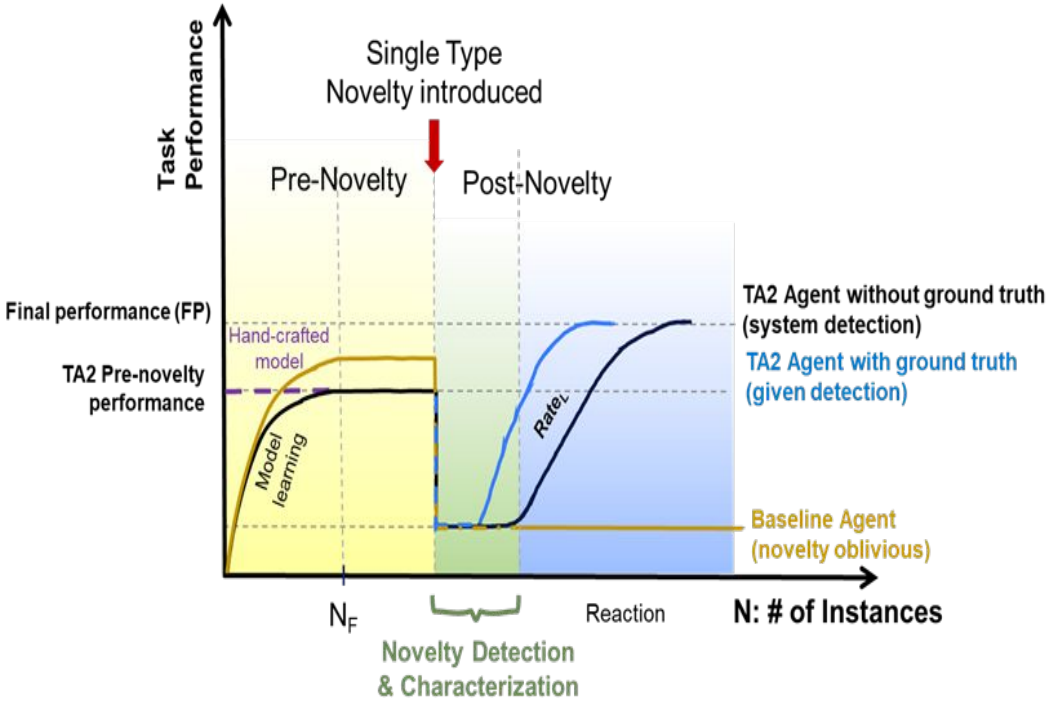
SAILON

Goal:

Develop the underlying scientific principles and general engineering techniques and algorithms needed to create AI systems that **act appropriately and effectively in novel situations which occur in open worlds**

Objectives:

- Develop scientific principles to **quantify** and **characterize** novelty in open world domains
- Create AI systems that can **detect** and **accommodate** novelty in open world domains.



Examples

- rules for chess were changed mid-game
  - How would an AI system know if the board had become larger, or
  - if the object of the game was no longer to checkmate your opponent's king but to capture all his knights?
  - Or what if rooks could now move like bishops?
- conditions changed for autonomous vehicle
  - off-road environment?
  - unseen weather?
  - hostile environment with enemy troops, roadblocks and bunkers?
  - broken bridges?
  - a hole on the roads?

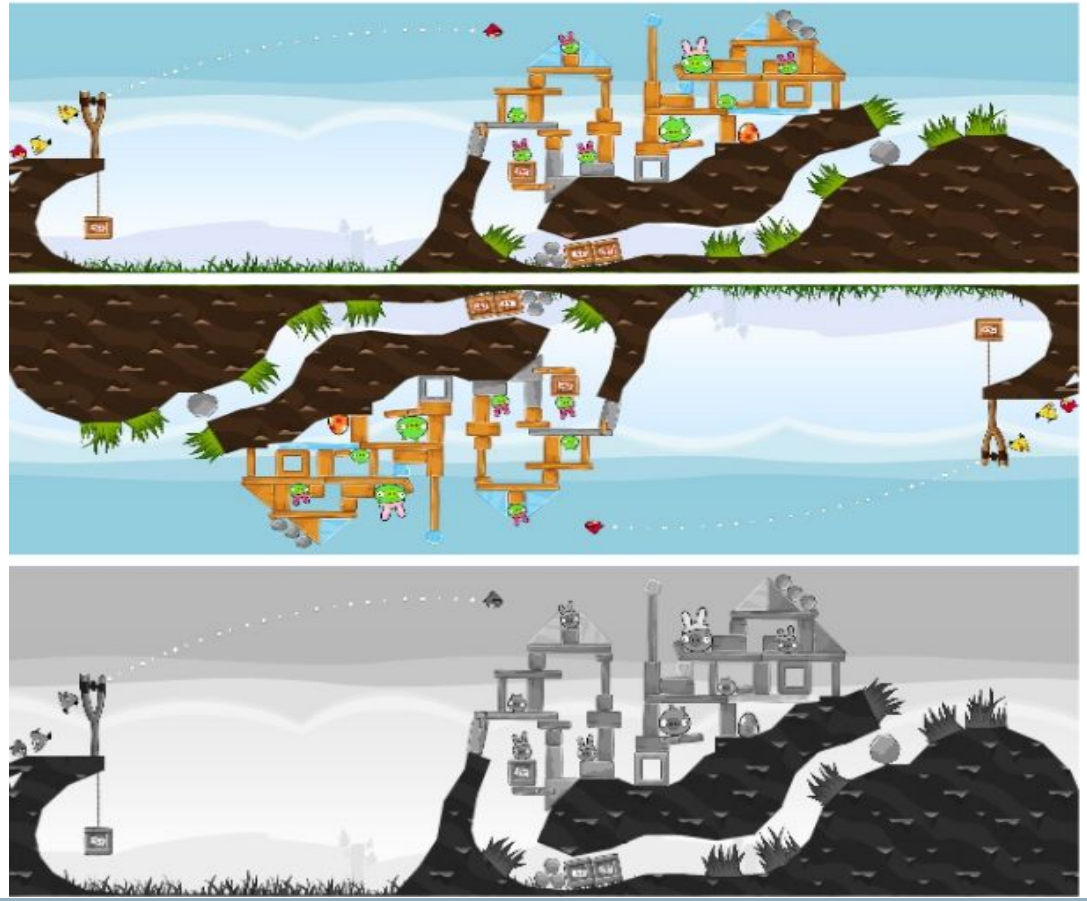
Angry Birds

Angry Birds is a popular physics puzzle game we use to develop

and test AI capabilities for the physical world. AI Agents need

to find physical actions that solve a given task, but:

- Only partial knowledge about the environment is available.
- Consequences of actions are unknown in advance.
- There are (infinitely) many actions to choose from.
- Unpredictable novelty occurs frequently and in many different ways.



Novelty

- Novelty is not anomaly.
  - Novelty: samples that share some common space with the trained distribution, which are usually concepts or classes which **the model could include when expanding its knowledge.**
    - If you train a network specialized in different dog breeds, an example would be a new dog breed that was not in the training set. Furthermore, if the classes are more complex, some novelty out-of-distribution could be new viewpoints or modifications of an existing learned class.
  - Anomaly: samples that are not related with the trained model
- Not all novelty are adaptable
  - Given you can only speak English, an adaptable novelty would be switch each “I” you would say to “You” and each “You” to “I”. You need to think about it, but you can do it. Speaking french, although it shares some common space, is not adaptable. You need to learn.
- Different Novelty Types:
  - novelty doesn’t affect the performance of the agent. e.g. chess board changed colour
  - novelty can not be accommodated. e.g. the opponent’s king can’t be captured by all your pieces
  - novelty once used correctly benefits performance, and doesn’t affect performance if not detected. e.g. knight captures all opponent’s pieces on the way.
  - novelty affects performance and require accommodation to restore performance. e.g. goal of the game changed from capture the king to queen.
  - You name it!

What are the open research problems?

- novelty theory
  - do all novelties in different domains share same characteristics? what are the same ones and what are the different ones?
  - How can we tell if the novelty is adaptable or not?
  - How can we define the difficulty to detect and/or react to a novelty?
  - Can novelty be generated automatically? How can we make sure the generated novelties are meaningful?
  - How can we systematically evaluate agent's performance in open worlds? what are the metrics should be used?
  - what are the important characteristics a testing domain should have?
- react to novelty
  - how to reliability detect novelties?
  - how to quickly accommodate novelties?
  - Can there be a general method that works for all novelties?
- You name it!

Recap of the Assignments

The Five Assignments

- #1. Tutorial topic (15%)
- #2. Paper review (15%)
- #3. Team project report (30%)
- #4. Team project presentation (15%)
- #5. Individual project proposal (15%)
- In addition: Tutorial mark for project updates and discussions (10%)

The Five Assignments

- #1. Tutorial topic (15%)
- #2. Paper review (15%)
- #3. Team project report (30%)
- #4. Team project presentation (15%)
- #5. Individual project proposal (15%)
- **In addition: Tutorial mark for project updates and discussions (10%)**

Three Phases

- Selection phase
  - 5 weeks from Mar 1 – Apr 1
  - Discuss the papers you read.
  - Brainstorming: discuss the research questions and research problems that would be interesting to solve/answer and how.
  - You need to form a group (of 3) and agree on a topic by Apr 1. If you have agreed on the group and topic before Apr 1, let me know.
- Project phase
  - 4 weeks from Apr 19 – May 13
  - You start working on your project, study the literature and do some experiments, implementations, analysis etc.
  - 15-20 minutes reporting and discussion per project. You can contribute ideas to others project or asking questions.
- Presentation phase
  - 2 weeks from May 19 to May 27
  - Each team gives 15 min talk in the tutorial and 4 minute lightning talk during one the the lecture on May 17 and May 24.
  - Submit the report

What will happen during the tutorials?

Phase 1:

1. You will be randomly assigned to a channel with another 2 students until you have formed a team and then you will be assigned to the channel with your teammates.
2. You will discuss with the students about the papers you read and think about possible project topics.

Phase 2:

1. Everyone will be in the same channel.
2. You will present (15-20 minutes) the progress you made over the week.
3. Share ideas and criticism with others.

Phase 3:

1. 4 minutes lightning talk **during one lecture.**
2. 15 minutes talk during the tutorial.
3. students from other tutorials will also join.

How you get participation mark in the Selection Phase

- Read the papers before the tutorial.
- Discuss the papers you read. I will call each of you individually and you will need to report to me what you have read and discuss your thoughts.
- Brainstorming: discuss the research questions and research problems that would be interesting to solve/answer and how.
  - Active participation of the tutorials is encouraged (1.0 for excellent contribution, 0.5 for standard, 0.0 for no or very little)

discussion question for today

Do you think to be able to accommodate novelty, the agent need to understand what is novel or not? why?