

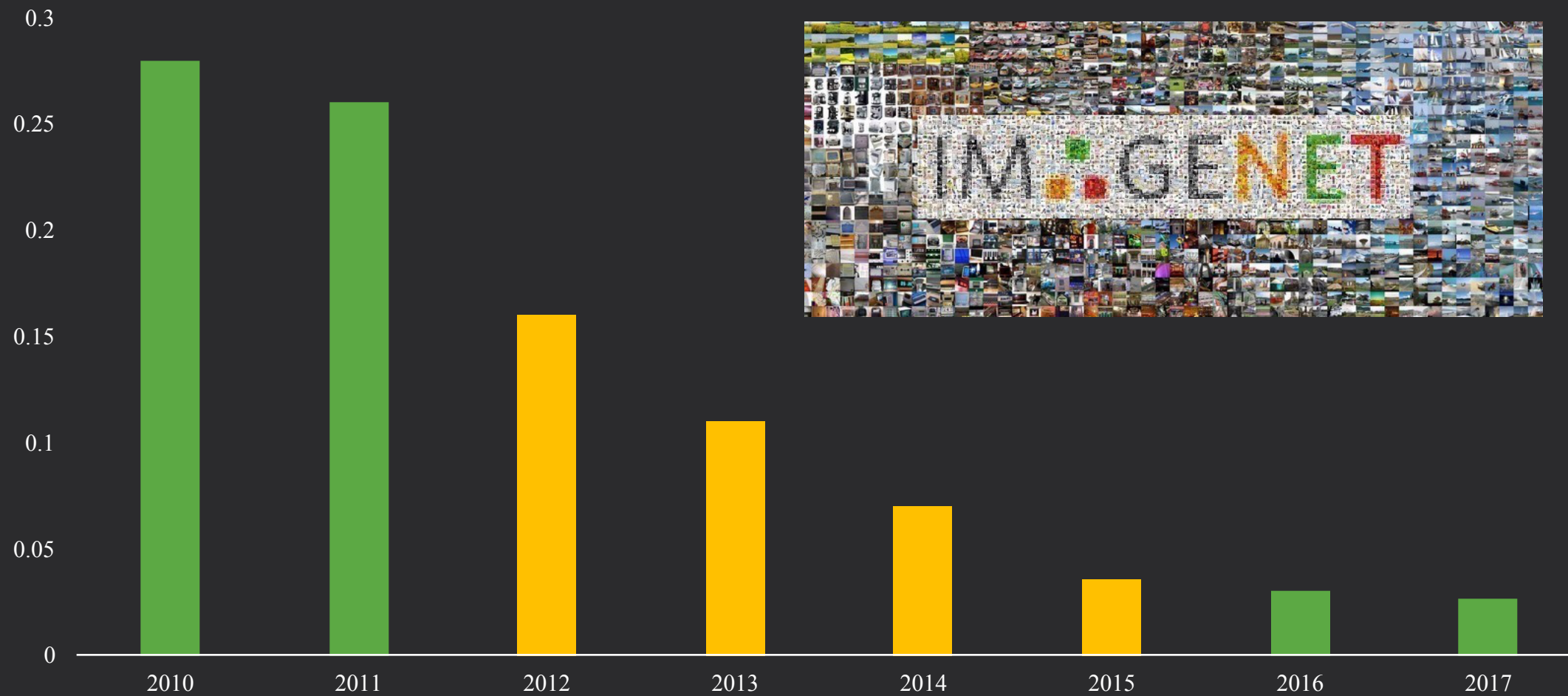
# New Exploration in Computer Vision

*Professor Dahua Lin*

*Co-founder, SenseTime*

*Director, CUHK-SenseTime Joint Lab*

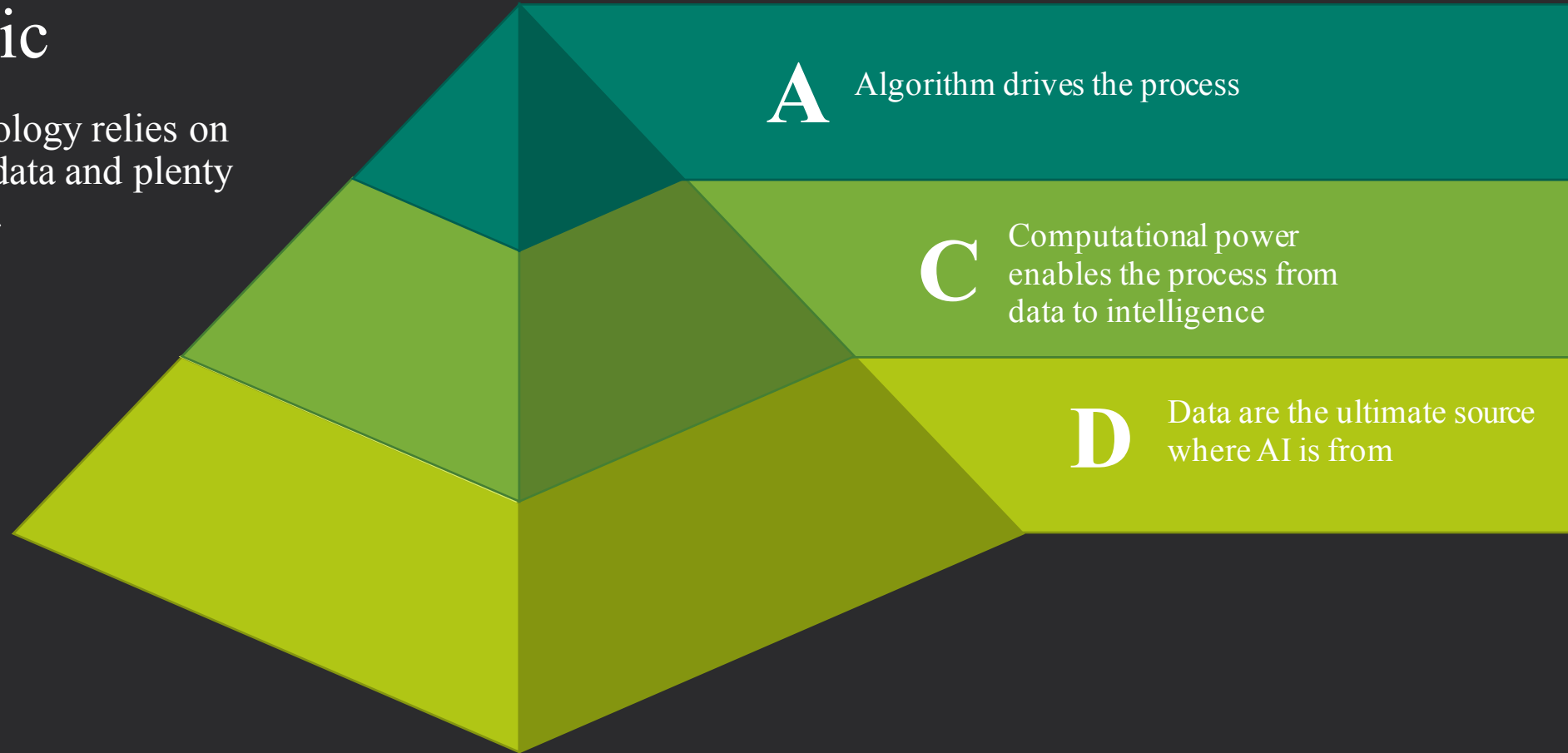
# Remarkable Progress in Eight Years



# The Technological Pyramid of AI

## AI is not a magic

The success of an AI technology relies on large amount of annotated data and plenty of computational resources.



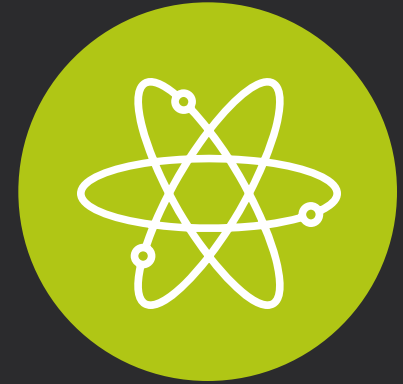
# Computer Vision: Beyond Performance



**Efficiency**



**Data Cost**



**Quality**

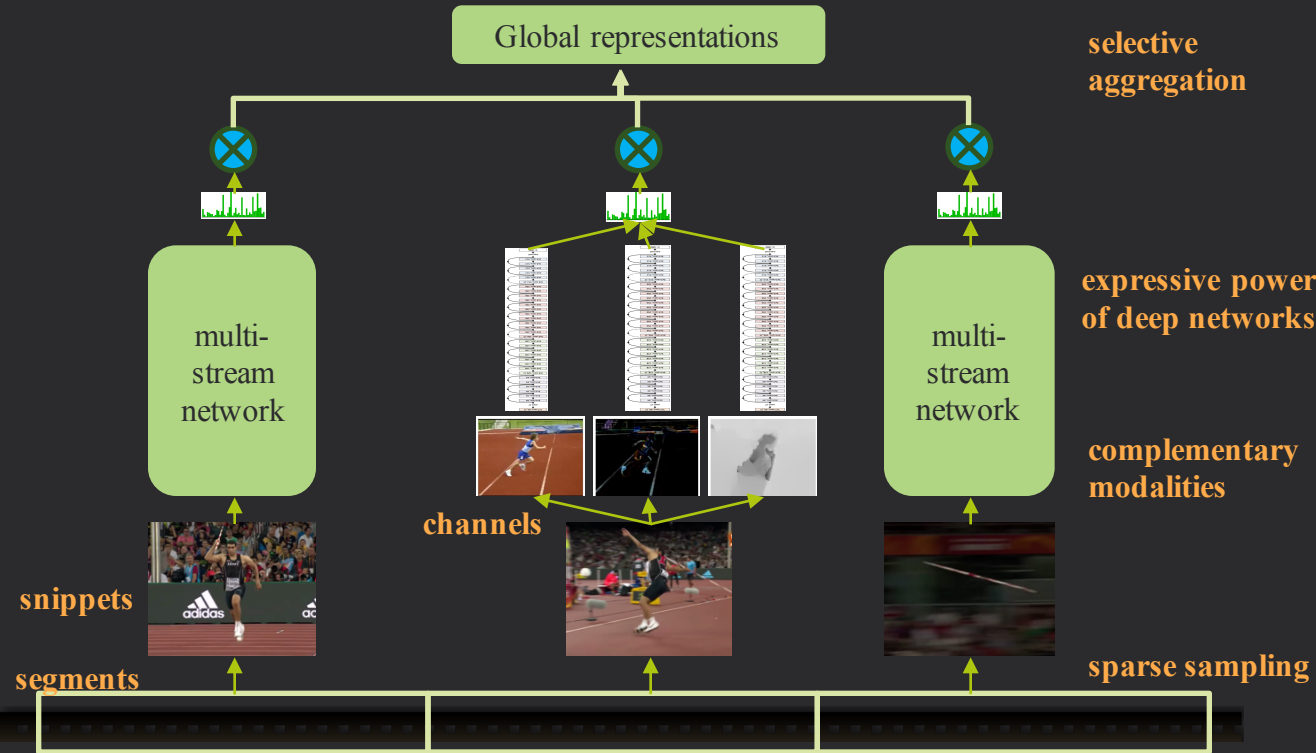


## **The key to efficiency**

Allocate resources to where they are truly needed



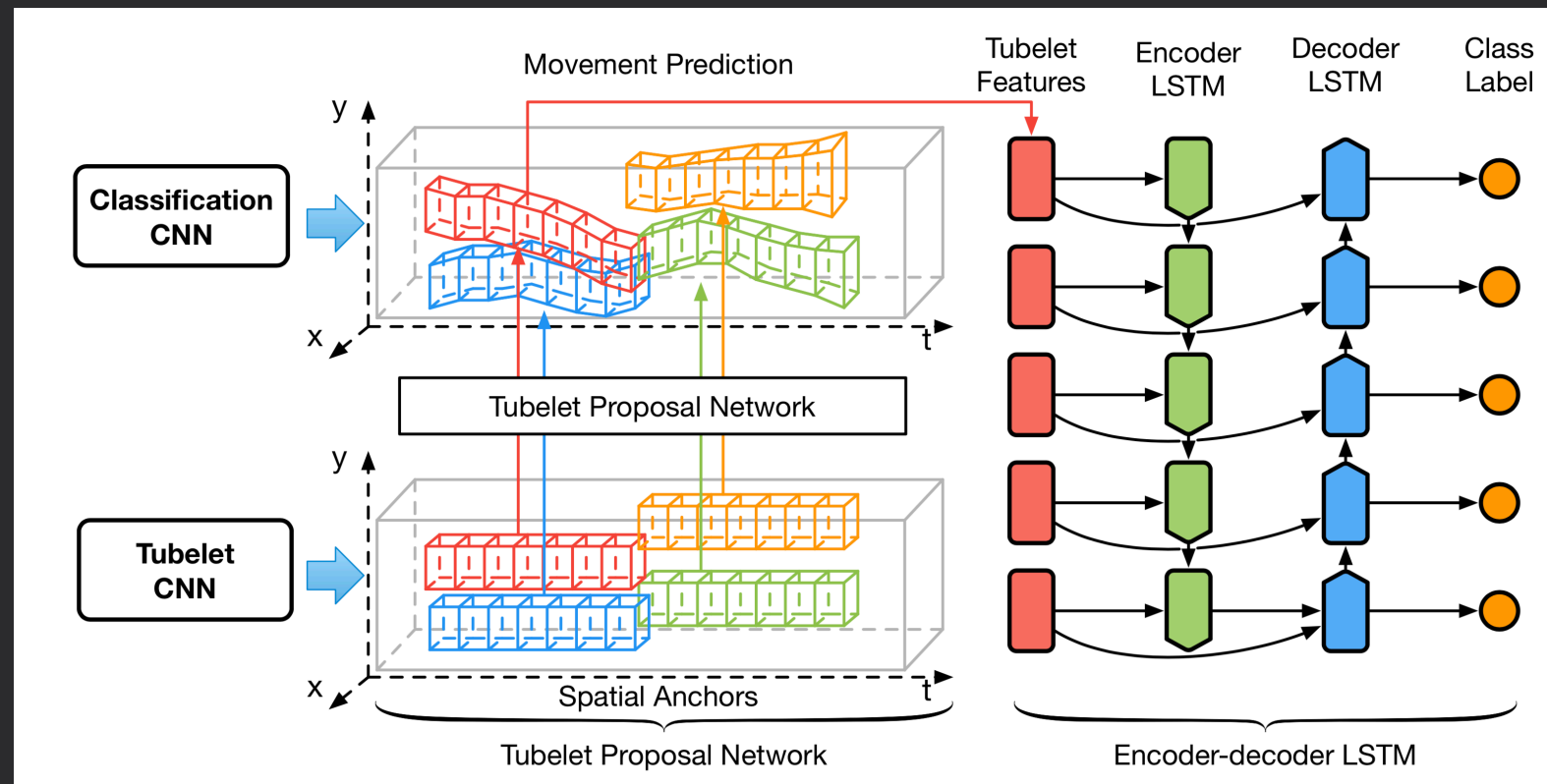
# Temporal Segmental Networks



- ✓ With randomized sparse sampling, it can handle long videos with limited memory capacity.
- ✓ Effective combination of appearance and motion with multi-stream architecture.
- ✓ Selective aggregation can effectively capture long-term temporal structures.
- ✓ **No.1 in ActivityNet 2016**
- ✓ **Widely adopted by practical systems for video classification & tagging.**

# Video Object Detection

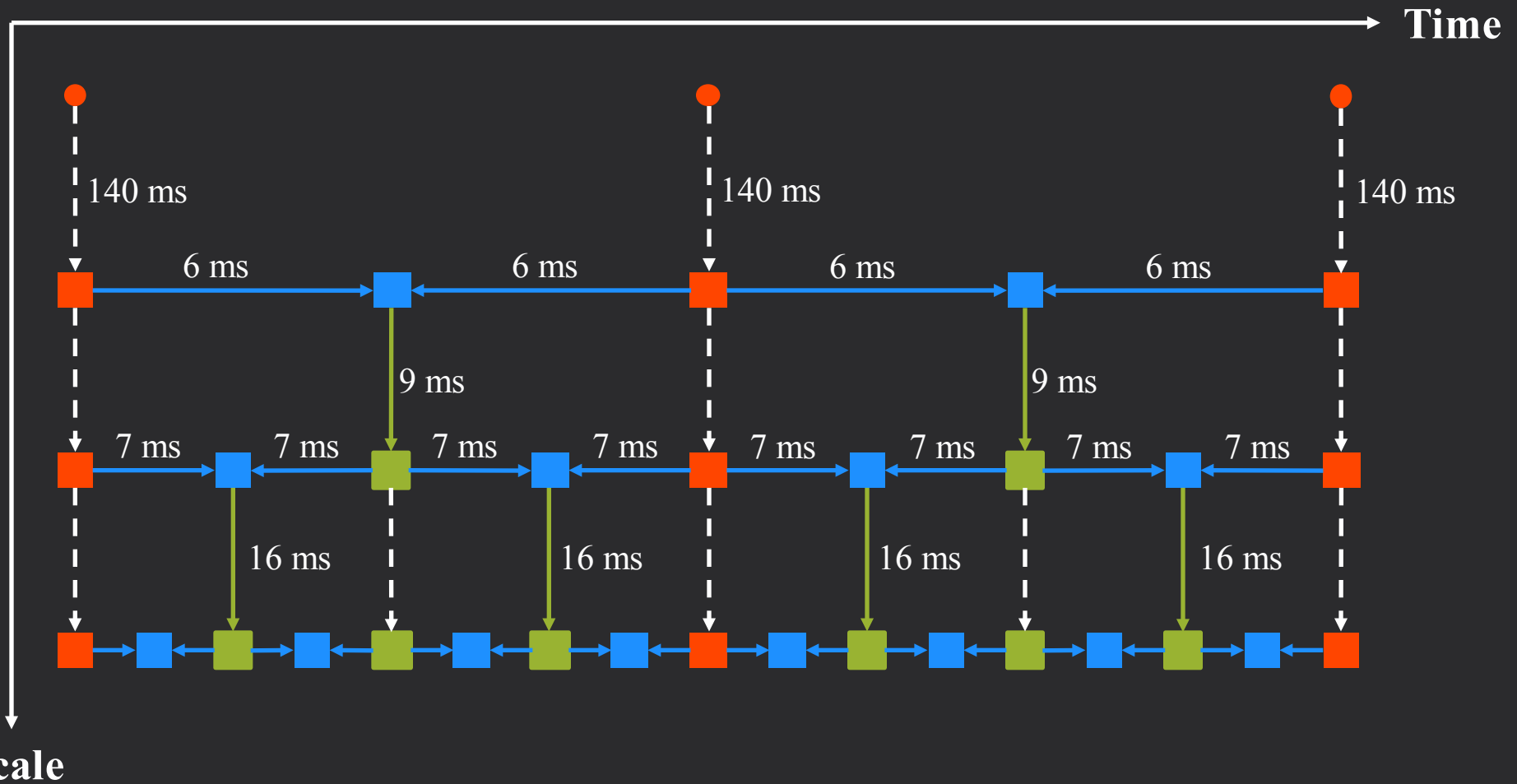
The winner of the VID track of ImageNet 2016





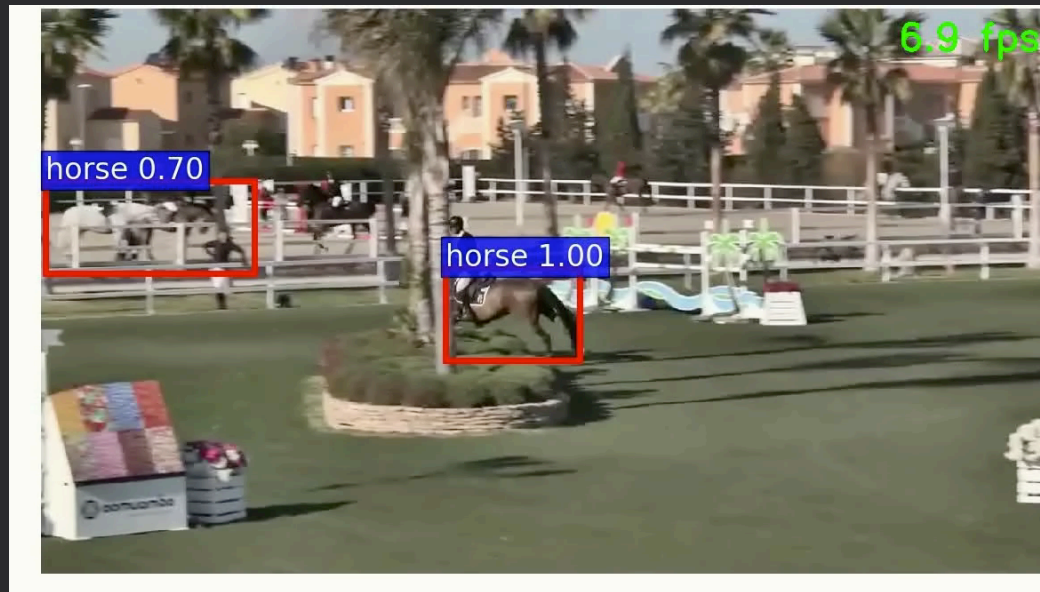
# Video Object Detection in Real-Time

Scale-Time Lattice

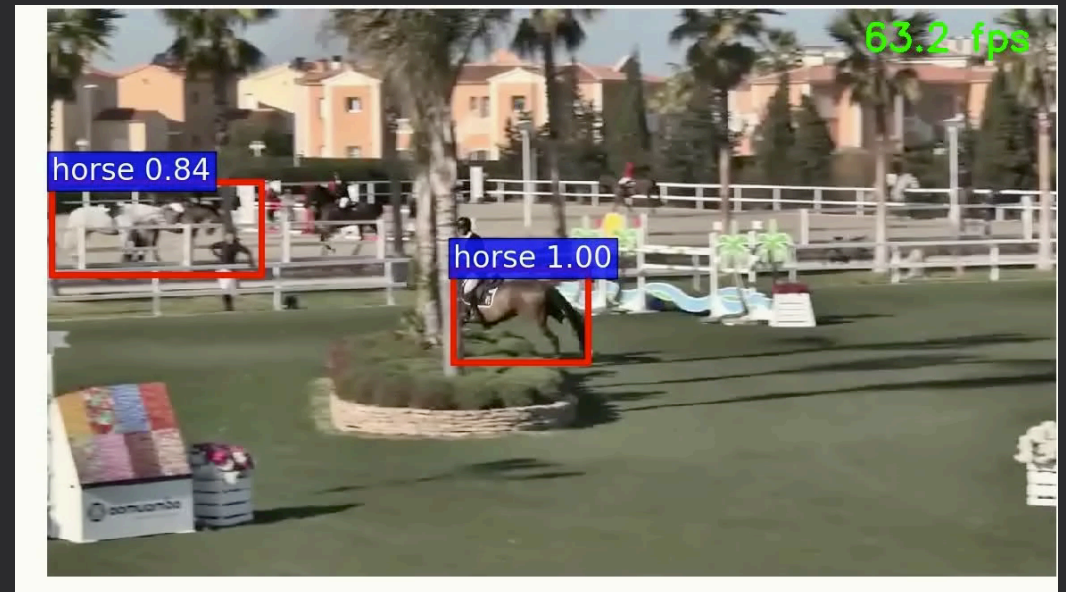




# Video Object Detection in Real-Time

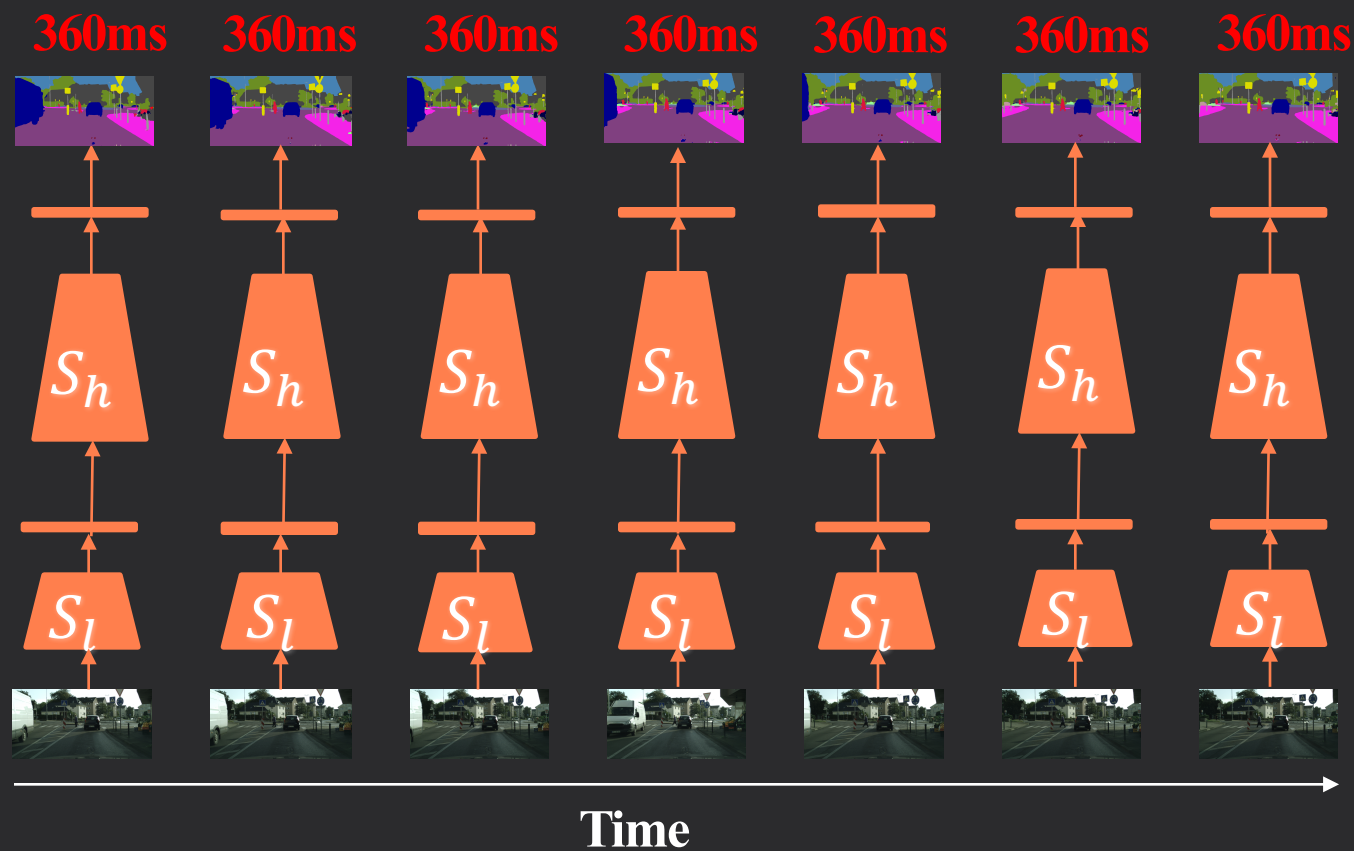


7 fps



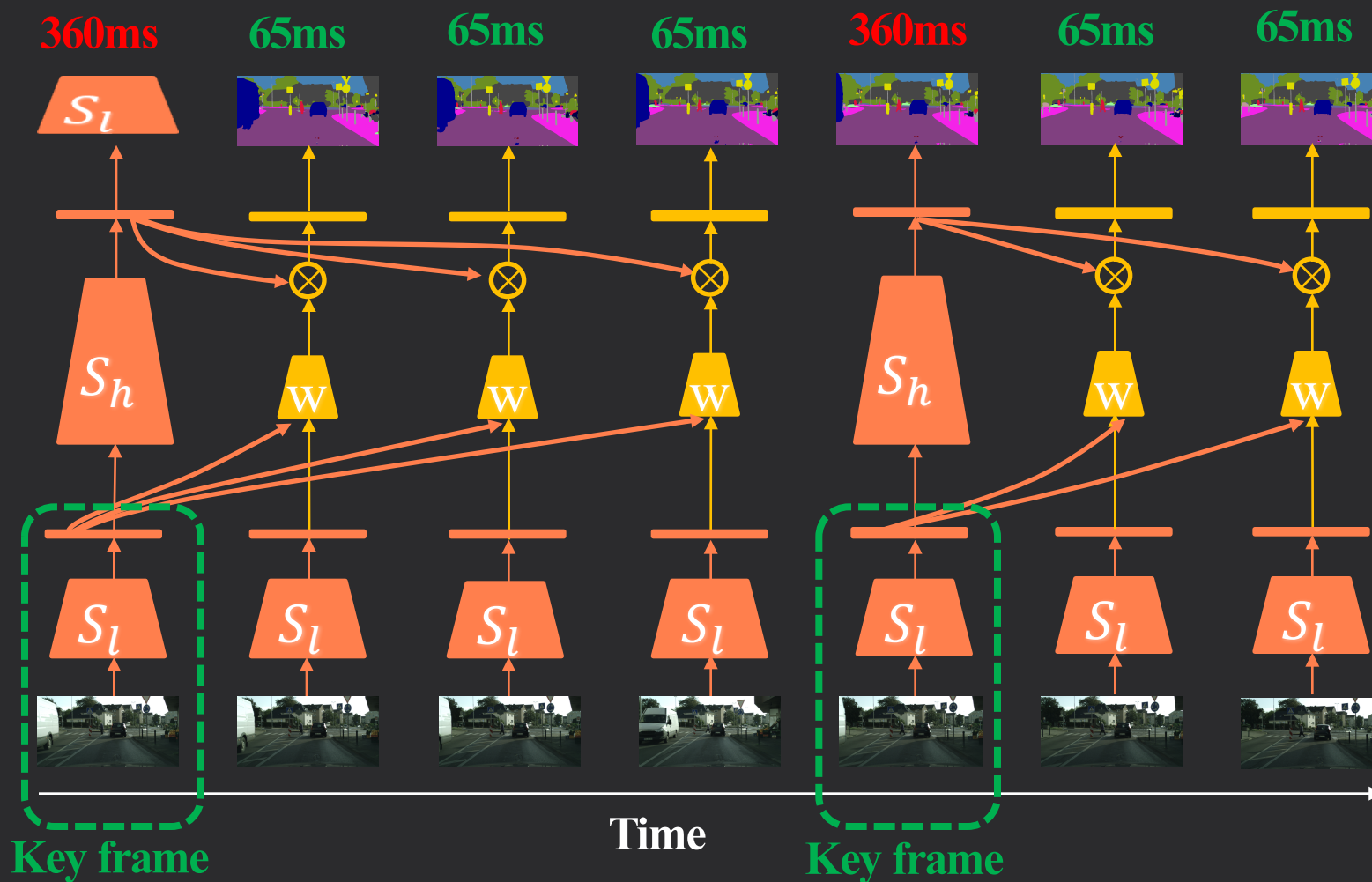
62 fps + more reliable

# Video Semantic Segmentation





# Semantic Segmentation in Real-Time



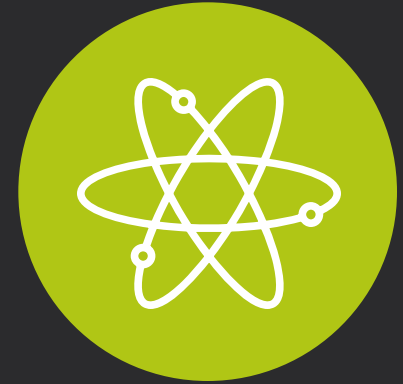
# Computer Vision: Beyond Performance



**Efficiency**



**Data Cost**



**Quality**

# People Behind the Scene



## **How to reduce the annotation cost?**

Exploit the links available in the data and environment



# Learning Object Detectors from Documentaries

Frames



Timeline



Subtitles

She needs to be good. Her **cubs** have already got huge appetites

Most animals fear **sloth bears** as but not apparently **wild boar**, at least not in this food around.

**langurs** are the friends of **spotted deer**.



appearance



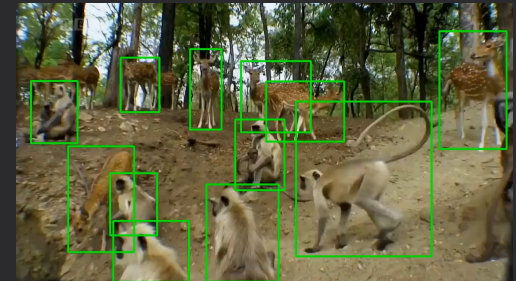
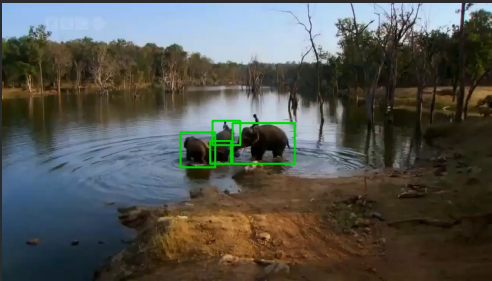
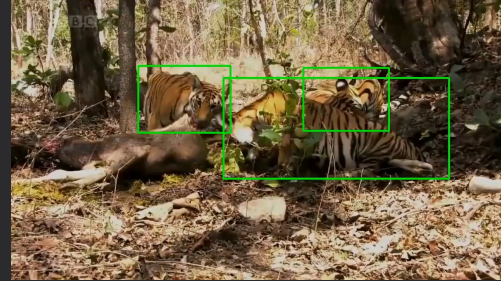
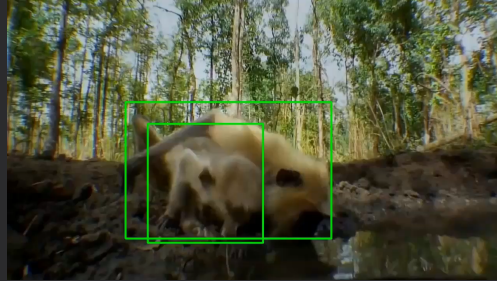
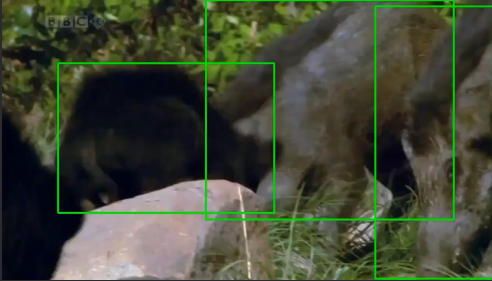
geometry



linguistic



# Without any human intervention, we learn







# Recognize People by Exploiting Contextual Links

Who are they? With Face + Visual Context + Social Context



Person-Event



Person-Person



Framework



Rose

Jack

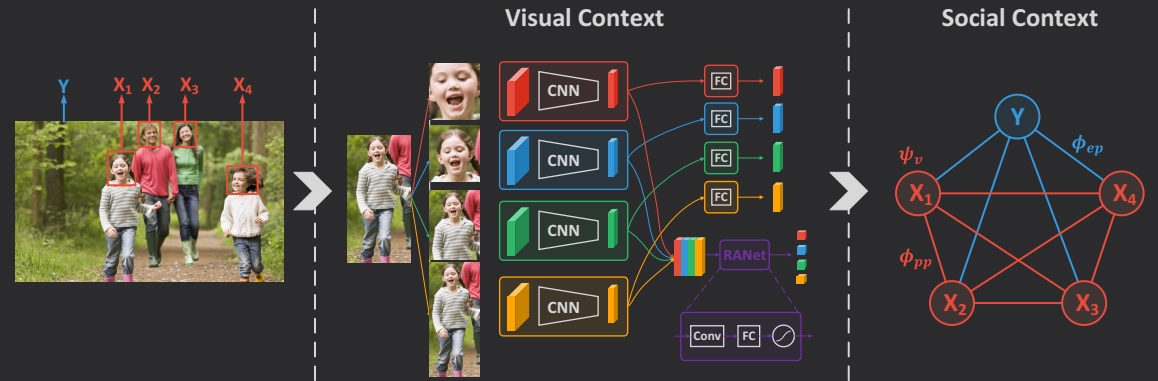
Caledon

Molly

Ruth

Edward

Brock



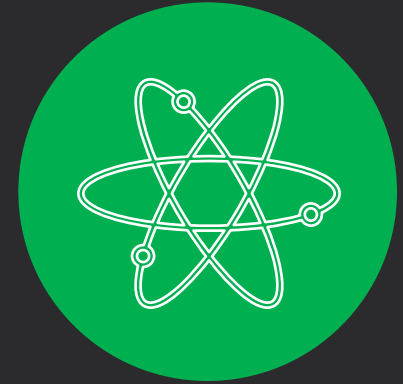
# Computer Vision: Beyond Performance



**Efficiency**



**Data Cost**



**Quality**



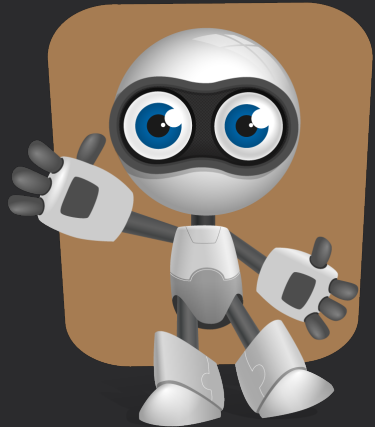
**We wish AI can improve  
the quality of our life,  
but quality is not just  
about accuracy**

# Captioning: How far is AI from a human being?

He is flying through the air while riding a snowboard

He is flying through the air while riding a snowboard

He is flying through the air while riding a snowboard



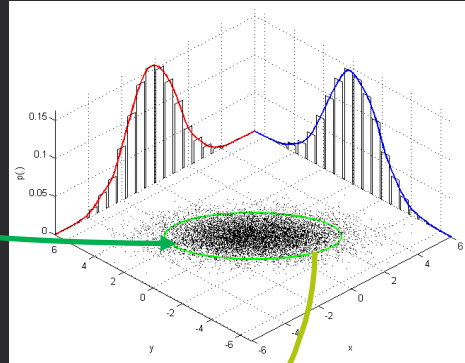




# A More Lively Captioning Framework



G



E



a man with stunts on his skis in the snow



a man on a skateboard in a snowy park



a man skiing down a mountain

# A bit more fun

Sad



Exciting



Shock



Clapping



Bowing



Cheer up






## A bit more fun

纯属虚构，如有雷同实属巧合





**AI will make our world better,  
but we still have a long way ahead ...**

