Fundamentals of Data Science

Fundamentals of Data Science
23 September 2024
Prof. Fabio Galasso

SAPIENZA
UNIVERSITÀ DI ROMA

Briefly about myself

- Earlier work
  - Ph.D. and Pdoc in Cambridge
  - Research associate at MPI
  - Head of Computer Vision Dept at OSRAM
- Perception and Intelligence Lab (PINLab) at Sapienza
  - Research and Innovation in Perception (Computer Vision) and Machine Learning
  - Distributed and multi-agent intelligent systems
  - General intelligence (reasoning, meta-learning, domain adaptation)
  - Sustainable computing (low-power-consumption and constrained-computational-resource)
  - Interpretable AI

Lecture and Exercise

- Lectures:
  - Mondays 16:00-19:00 @ De Lollis
  - Fridays 10:00-12:00 @ De Lollis
  - Labs:
  - Thursdays 16:00-19:00 @ Lab 15 (via Tiburtina 205)
- Office hours
  - Thursdays 13:30-15:30 @Room 24, 2° floor, build. G, via Regina Elena 295
- Website: https://sites.google.com/di.uniroma1.it/fds-20242025/home
- Google Classroom Code: bc7iaui
  - For slides, course material, assignments and news
- **Subscribe to it now**

Exam

- Exam
  1. Theory: 50% (written)
  2. Practise: 50%, of which
     - 2/3 from assignments in Python, to be submitted by given deadlines during the course
     - 1/3 from a final project and presentation
- Assignments:
  - The assignments and the final projects **must be submitted in groups**
    - Groups must be **of size [3 – 5]**
    - Find a team today!
    - In order to take part in 1, it's needed a pass on part 2 of the exam
  - If you pass part 2, you may book the exam part 1 in the next calendar year
- Final project
  - Algorithms, objectives and topics for the final project may be freely chosen
  - Ideas for projects and resources for it would be discussed in class

Exam

- For the students of Data Science:
  1. (Course) Theory: 1/3 (written)
  2. (Course) Practise: 1/3, of which
     - 2/3 from assignments in Python, to be submitted by given deadlines during the course
     - 1/3 from a final project and presentation
  3. (Lab) Python programming lab: 1/3 (written)

- Same rules about the assignments and the final project

Assignments and final project

- Calendar
  - Assignment 1: 26 sept - 28 oct (4 weeks)
  - Assignment 2: 28 oct - 29 nov (4 weeks)
  - Final Project: 29 nov – 29 dec (4 weeks)
    - Project announcement on 11 Nov
    - First presentations of ideas on 25 Nov
    - Final project presentations on 16 Dec

Ethical Code of Conduct

- Plagiarism is an act of fraud
- Plagiarism is severely prohibited and, in any form, is regarded as a serious violation of the ethical code of conduct. Plagiarism includes the submission of an assignment or project whose source code or report bears strong resemblance to another persons's source code or report, including other AML projects and/or resources that can be found online. After submission, every project would be checked against plagiarism, including automatic detection tools
- Assignments and projects resulting incurring in plagiarism would be invalidated

Physical and Learning Disabilities

- Sapienza provides counseling and support
- You may reach out to:
  sportellodisabili@uniroma1.it and counselingdsa@uniroma1.it
- Or directly to:
  Prof. Tiziana Calamoneri
  Coordinator for Disabilities and DSA for the I3S Faculty
  http://wwwusers.di.uniroma1.it/~calamo/

Material

- Slides and coding scripts are distributed after lectures
- There is much material online
- Books (more at https://sites.google.com/di.uniroma1.it/fds-20242025/resources)
  - Data Science:
    - Bertsimas, O'Hair, Pulleyblank. The Analytics Edge.
    - Jure Leskove, Anand Rajaraman, Jeffrey D. Ullman, 2019.
      Mining of Massive Datasets. Cambridge University Press.
      (available at: http://infolab.stanford.edu/~ullman/mmdsn.html)
  - Machine Learning
    - Christopher M. Bishop, 2006. Pattern Recognition and Machine Learning
  - Deep learning
    - Ian Goofellow, Yoshua Bengio, Aaron Courville, 2017. Deep Learning
      (available at: https://www.deeplearningbook.org/)
  - Image Analysis and Recognition, Computer Vision
    - Richard Szeliski, 2010. Computer Vision: Algorithms and Applications
      (available at: http://szeliski.org/Book)

Coding References

- Coding examples and assignments would be in Python (3.x), leveraging the Pytorch (2.x) framework. The course provides an introduction to Pytorch
- Books for Python
  - Allen B. Downey, 2015. Think Python: How to Think Like a Computer Scientist (available at: https://www.greenteapress.com/thinkpython/thinkpython.html)
  - Jake VanderPlas, 2016. Python Data Science Handbook: Tools and Techniques for Developers: Essential Tools for working with Data (Book and notebooks available at: https://github.com/jakevdp/PythonDataScienceHandbook)
- Online tutorials for Python: https://docs.python.org/3/tutorial/
- Online tutorials for Pytorch: https://pytorch.org/tutorials/

Setup and computing

- A Linux OS is recommended
  - but Python, Pytorch and R also run on Windows
- Recommended Python distribution: anaconda (https://www.anaconda.com/distribution/)
- For running some exercises you may need a GPU
  - Use one on Google colab: https://colab.research.google.com
  - Refer to my prepared colab notebook (clone it) https://colab.research.google.com/drive/1e9FFE46ajCoXF-wjg7LMytkjlVlu4Zhr
- Refer to this tutorial on how to setup Pytorch in Google colab:
  - https://medium.com/dair-ai/pytorch-1-2-quickstart-with-google-colab-6690a30c38d

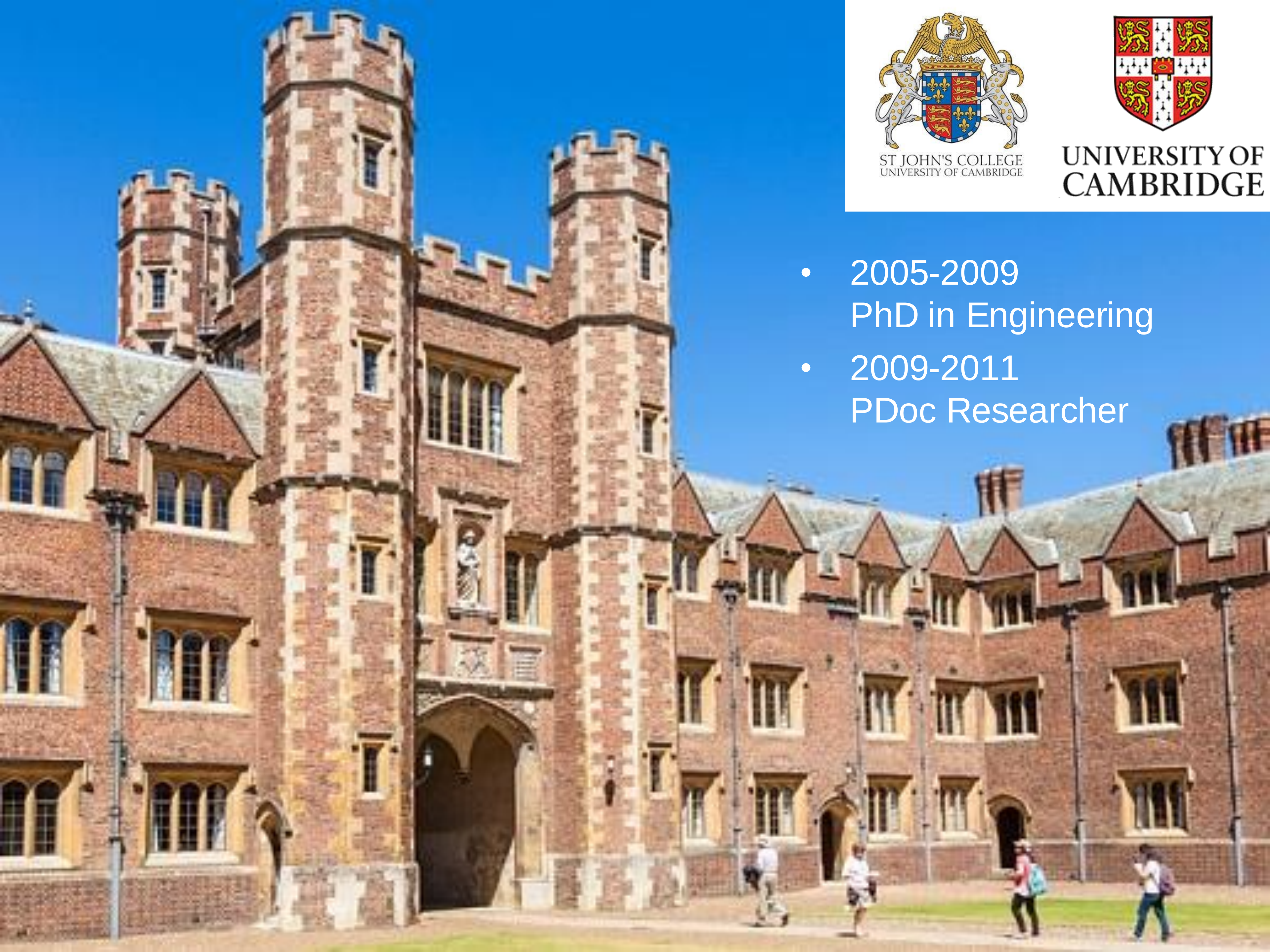
Pre-requisites

- Calculus and Linear Algebra
  - taking derivatives, understanding matrix vector operations and notation
  - Mathematics for Machine Learning (https://mml-book.github.io/book/mml-book.pdf) chapter 2, 3, 4 ,5
- Basic Probability and Statistics
  - basics of probabilities, gaussian distributions, mean, standard deviation, etc
  - Mathematics for Machine Learning (https://mml-book.github.io/book/mml-book.pdf) chapter 6

Syllabus

- Basics of digital image processing
- Discriminative models
  - Linear Regression
  - Logistic Regression
  - Multinomial Logistic Regression
- Optimization
  - Normal equation
  - Gradient Descent
  - Newton's Method
- Deep Neural Networks
  - Optimization and Back-propagation
  - Convolutional neural networks
- Bias/Variance
- Regularization
- Dimensionality Reduction
- Variational inference
- Advice on DS and ML

More about own research and the Perception and Intelligence Lab



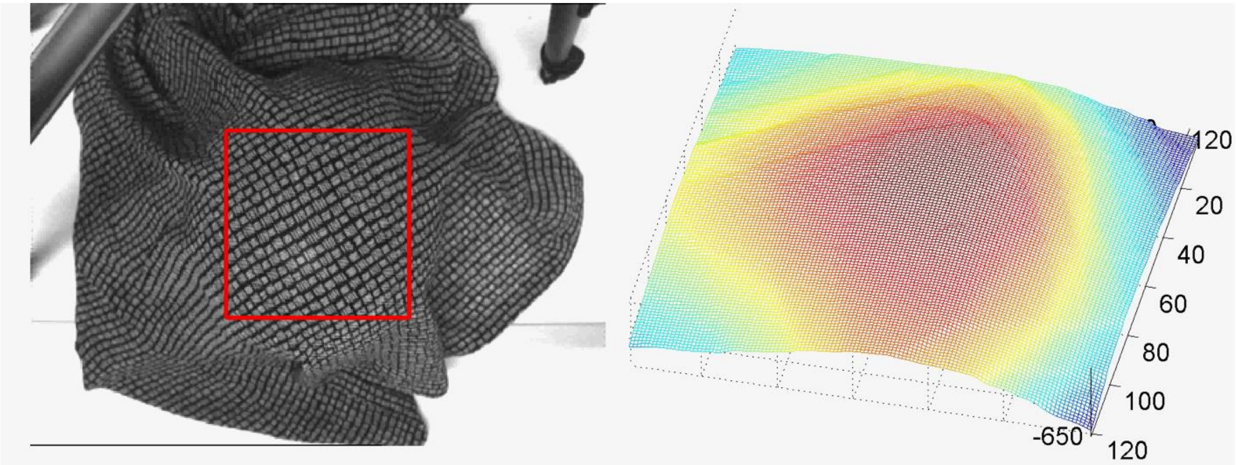
ST JOHN'S COLLEGE
UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF CAMBRIDGE

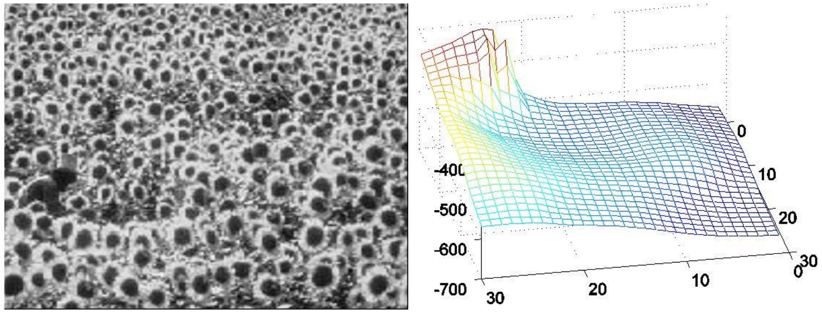
- 2005-2009
  PhD in Engineering
- 2009-2011
  PDoc Researcher

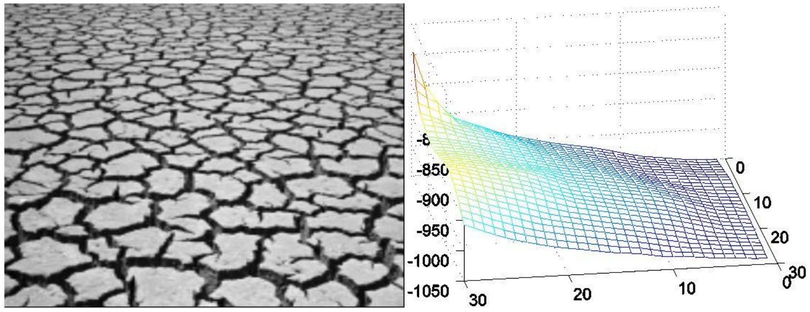
Textures and 3D reconstruction
BMVC’07, ISVC’07, ISVC’08, CVPR’09

- Leverage texture to recover the shape
  - In controlled environments



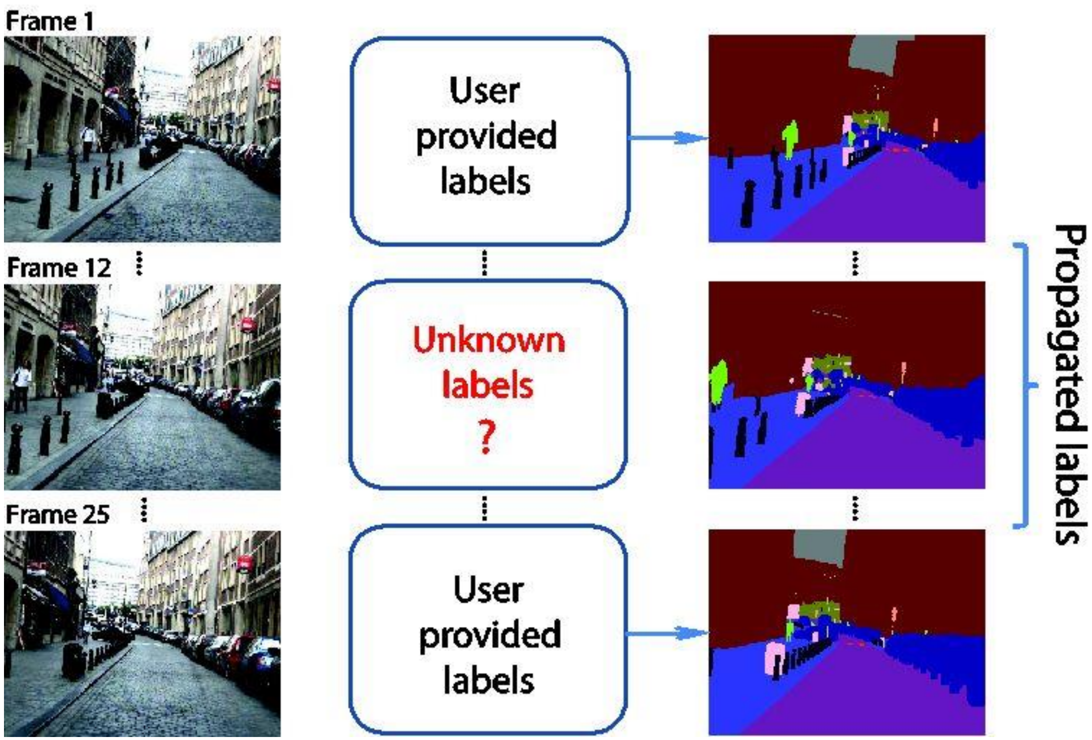
  - In the “wild”

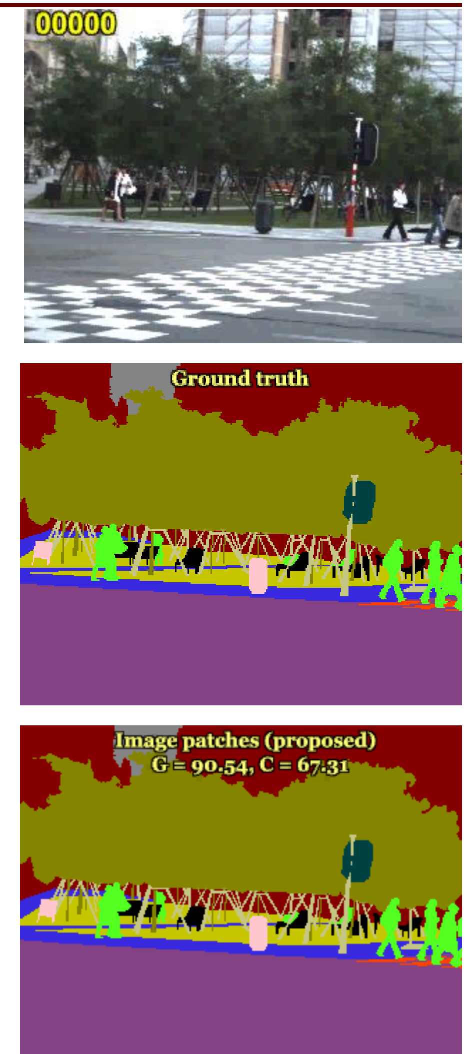




Label propagation and video analysis CVPR’10, ICCV’11

- Model videos with a graphical model
- First *label propagation*
  - Label non-annotated video frames automatically





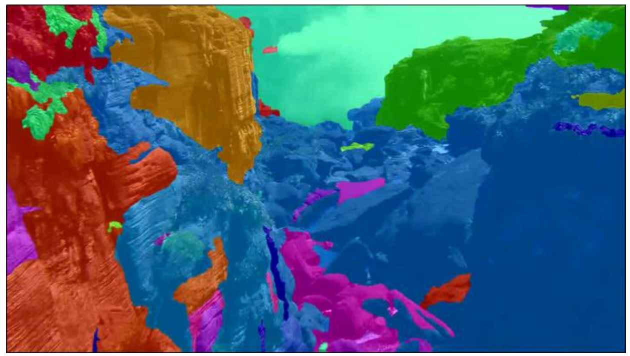
mpii max planck institut informatik

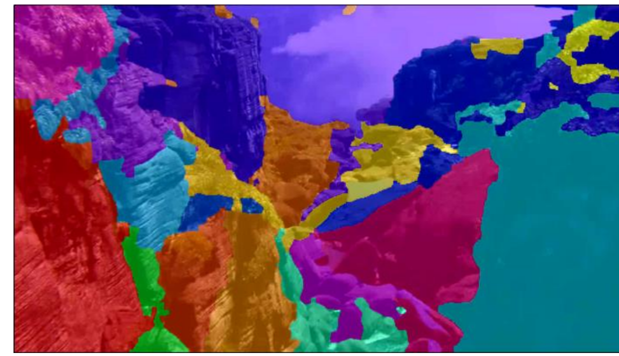
- 2011-2014
  PDoc Research Associate

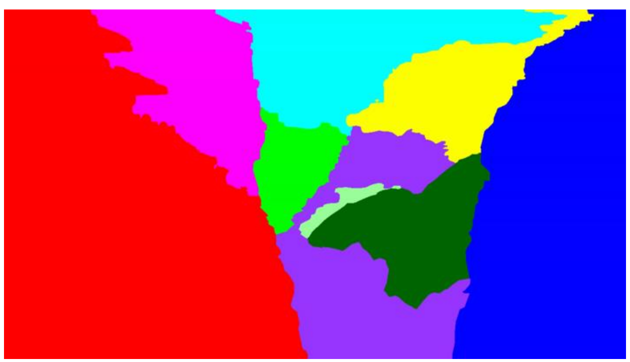
Video segmentation
ICCV’11, ACCV’12, ICCV’13, ACCV’16

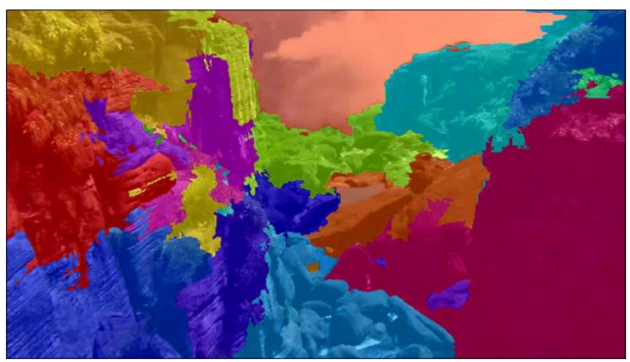
- *Learn from videos* to mimic the human perception


Video


Segm. propagation


Spectral graph reduction

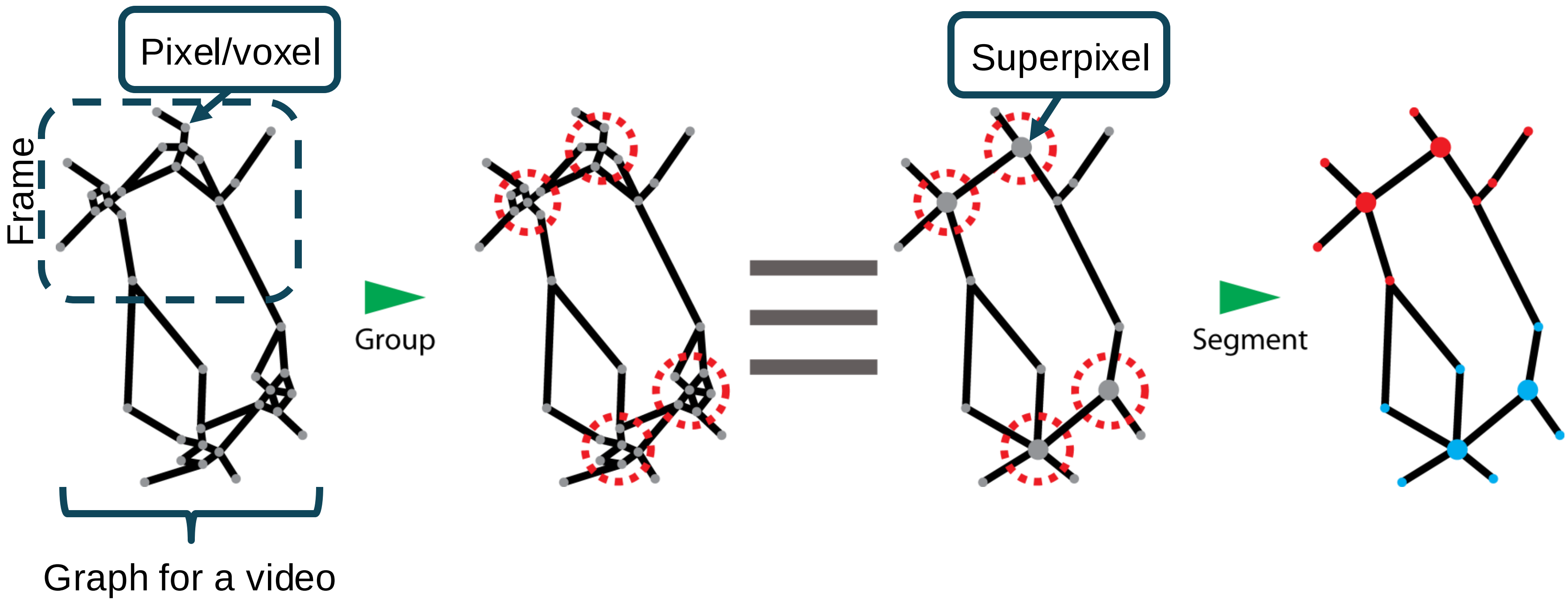

Ground truth


Grundmann et al. CVPR’10


Ours

Clustering with graphs
CVPR'14, GCPR'14, CVPR'15

- Learn graph representations for videos
- Learn grouping constraints
- Prove equivalence of graph reductions



- 2014-2019
  R&D Manager
  Head of Computer Vision Dept

Innovation transfer
Smart office

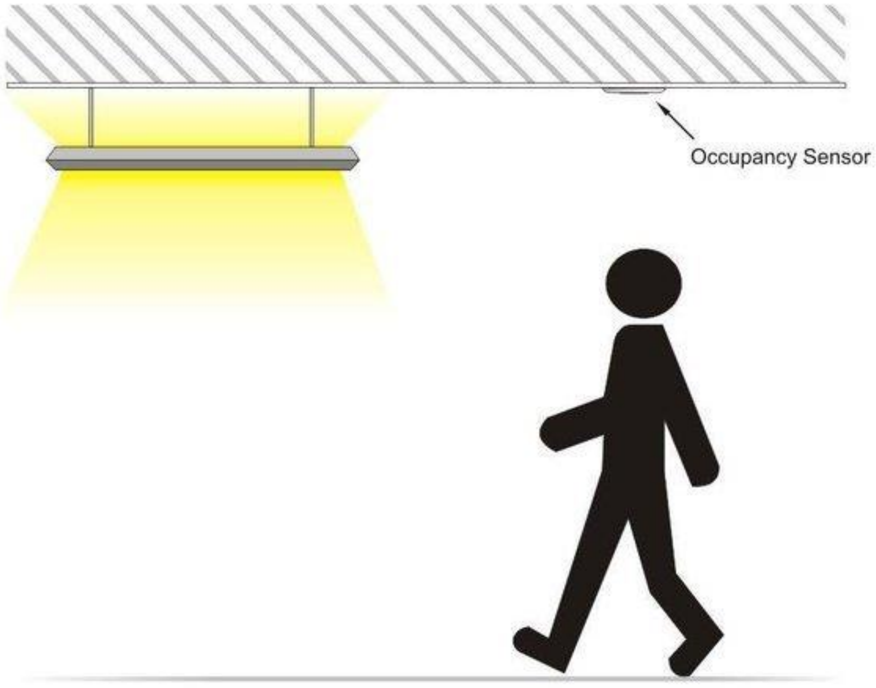
- Business potential: *light is everywhere people are*

Innovation transfer
Smart office

- Business potential: *light is everywhere people are*
- Objective:
  make light smarter by pairing it with a camera and some intelligence

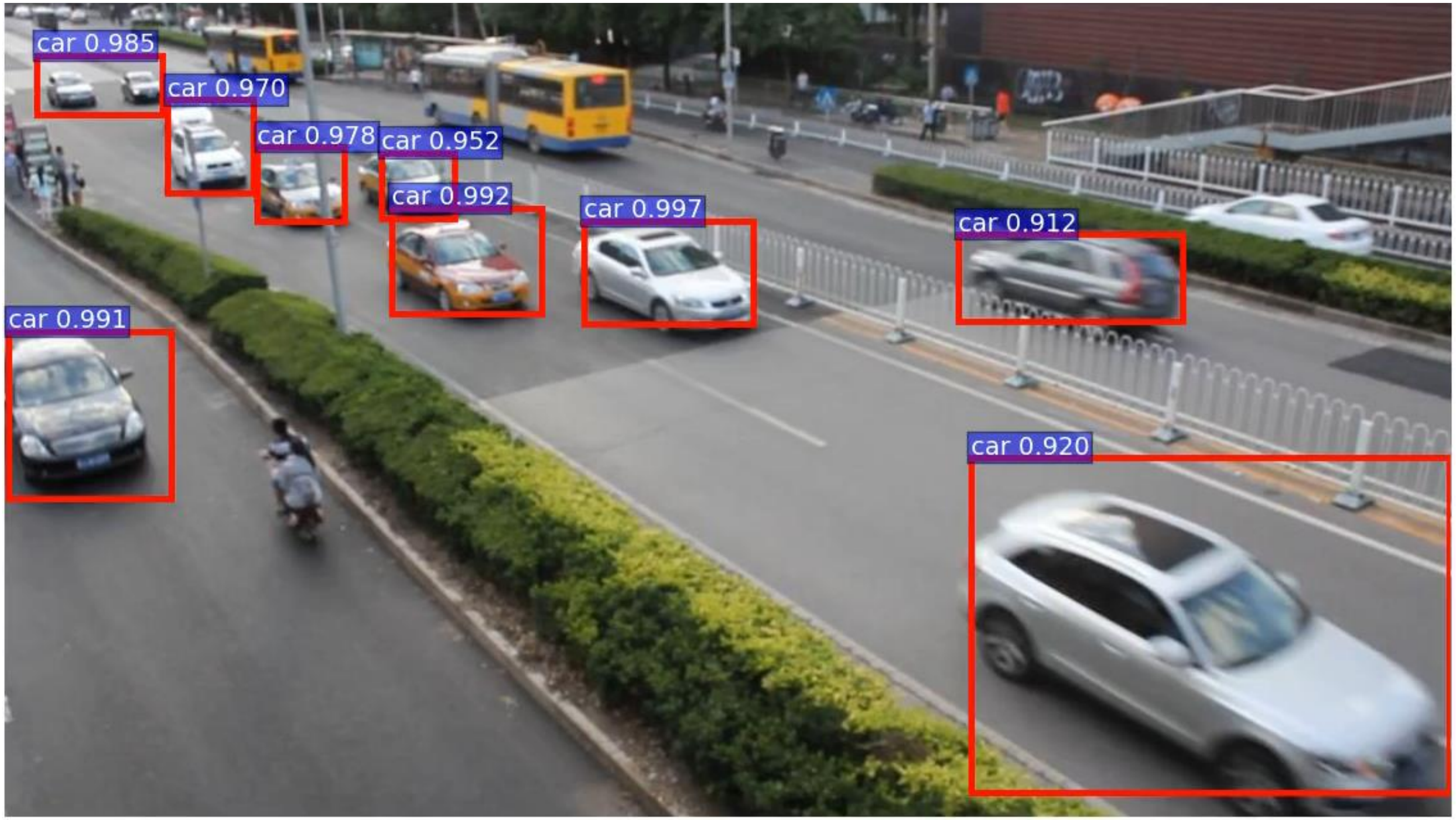
Innovation transfer
Smart office

- Business potential: *light is everywhere people are*
- Objective:
  make light smarter by pairing it with a camera and some intelligence
- Roadmap: *start by detecting the people*



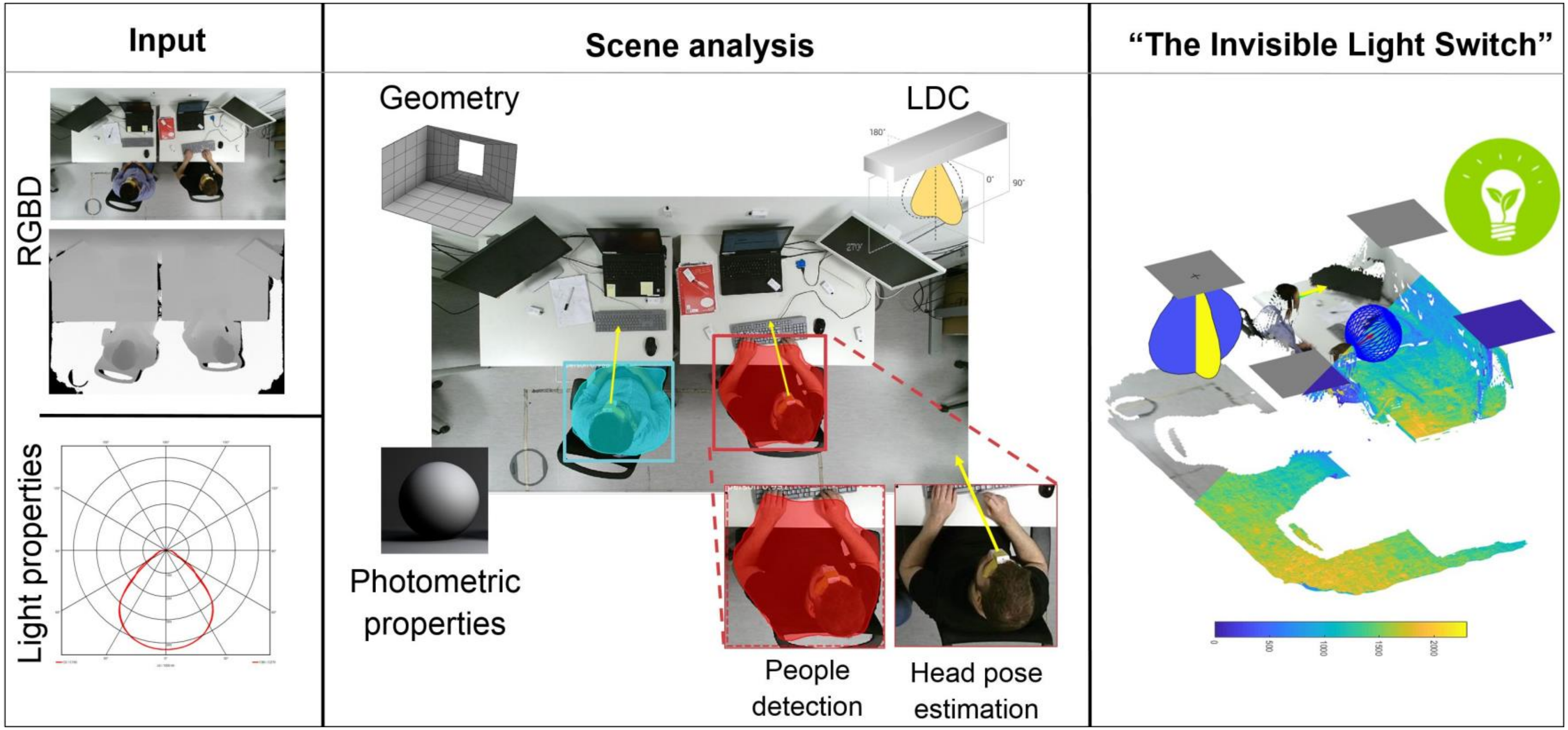
Detection, tracking and re-identification AVSS'17 (1st detector), AVSS'18 (3rd re-id tracker)

- Learn geometric (RPN) proposals
- Track across occlusion with re-identification



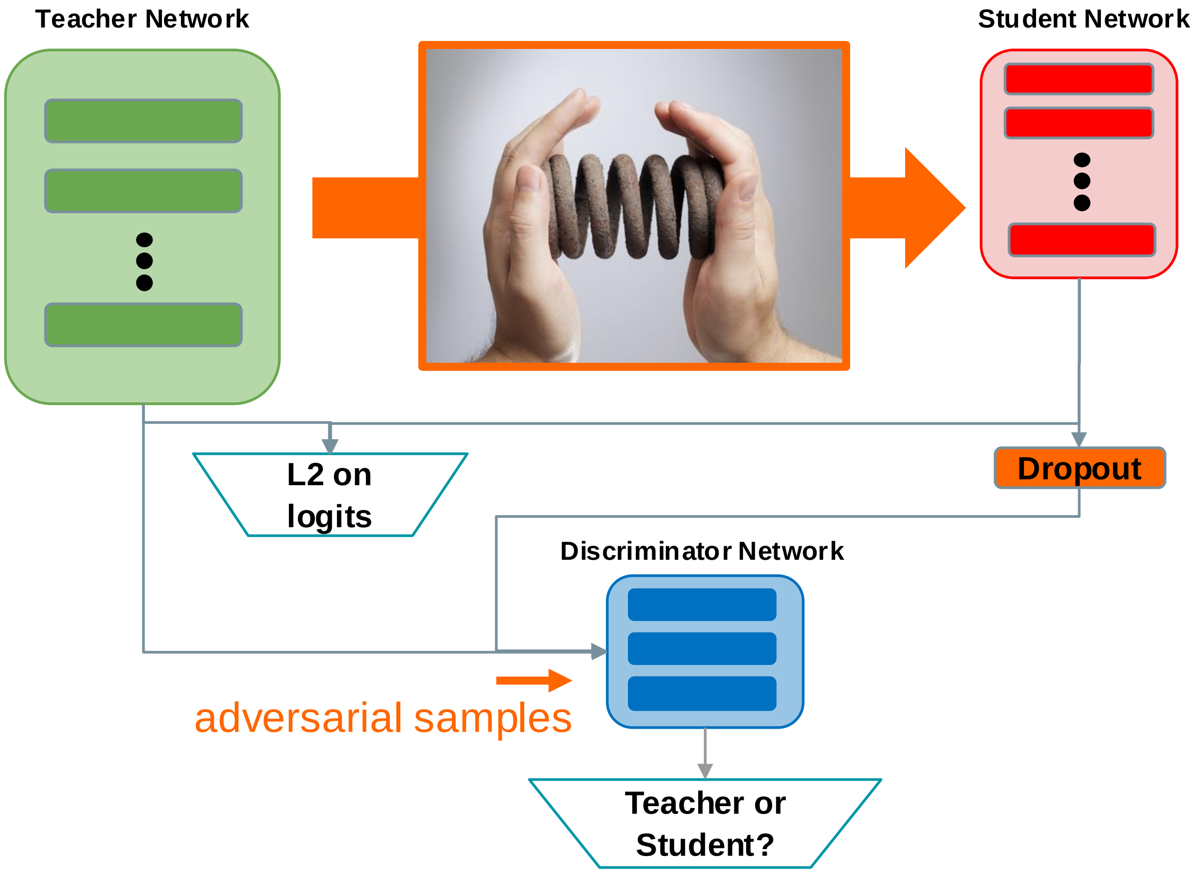
Detection, recognition, photometry
VISAPP'17, ICIP'18, *2x* WACV'19

- Detect people from top-views and estimate their gaze
- Model lighting and estimate the perceived illumination



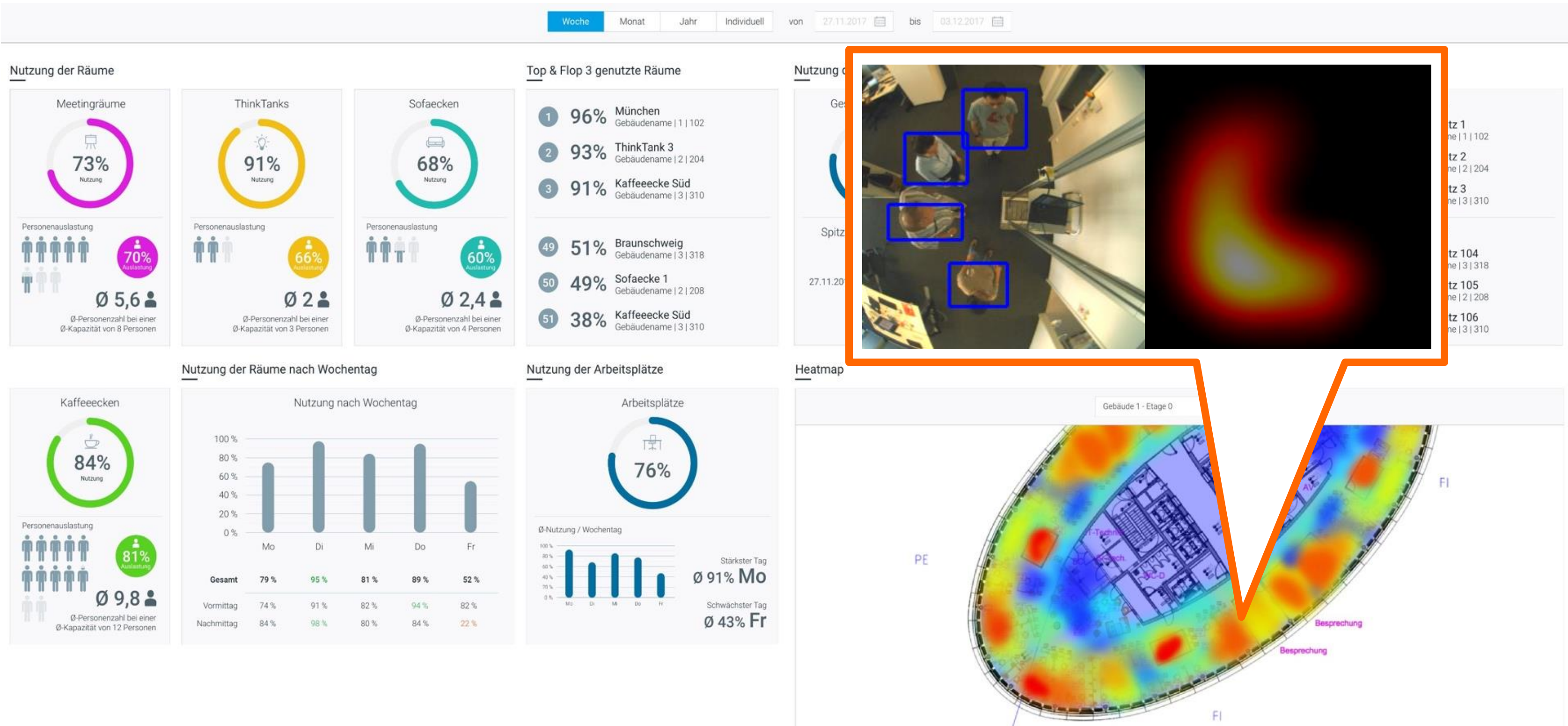
Model compression
CEFRL at ECCV'18

- First GAN for *adversarial network compression*



Innovation transfer
Smart office and retail

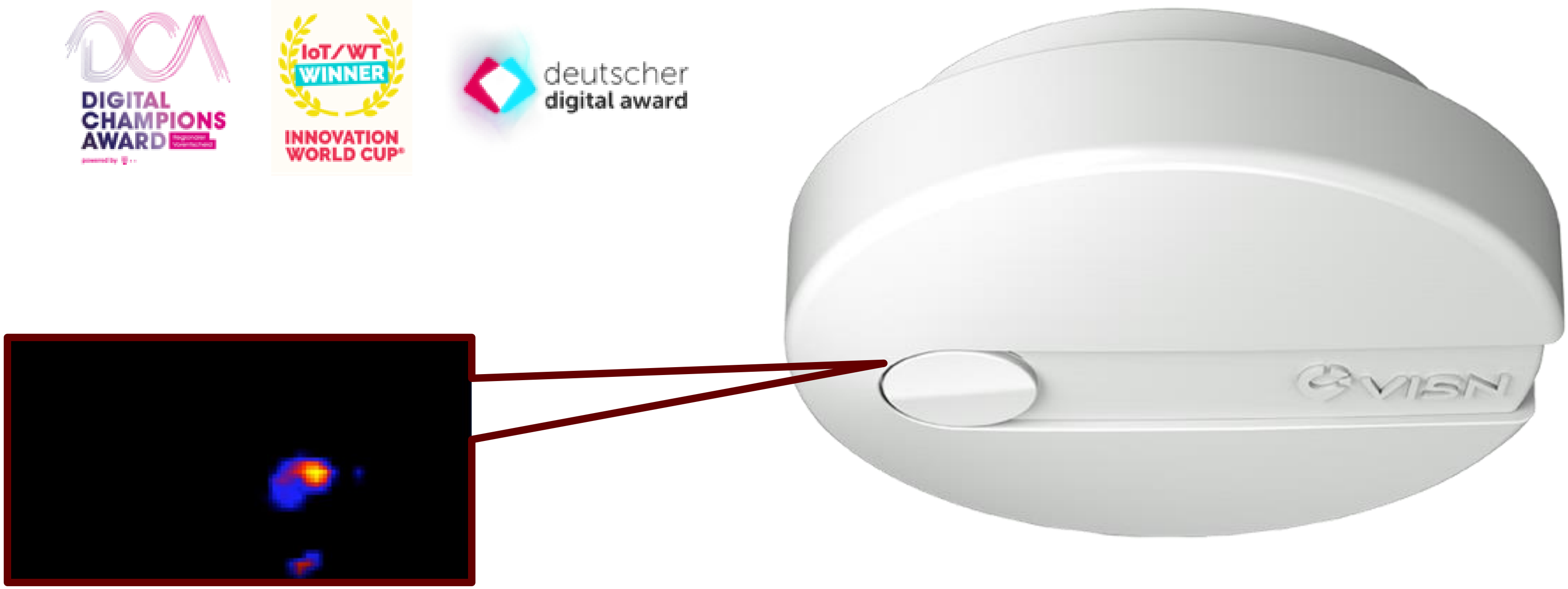
- Space utilization and people counting in offices (www.visn.io)
- Customer flow and conversion rate in retail (Edeka, *press release**)



* www.osram-group.de/de-DE/media/press-releases/pr-2018/17-12-2018b

Innovation transfer
Smart office and retail

- Space utilization and people counting in offices (www.visn.io)
- Customer flow and conversion rate in retail (Edeka, *press release**)
- Now on sale, also with thermal cameras



* www.osram-group.de/de-DE/media/press-releases/pr-2018/17-12-2018b

OSRAM

Bundesministerium für Wirtschaft und Energie

Innovation transfer
Smart city

- Installed in the city of Ulm
- Under evaluation for autonomous driving
- With partners:

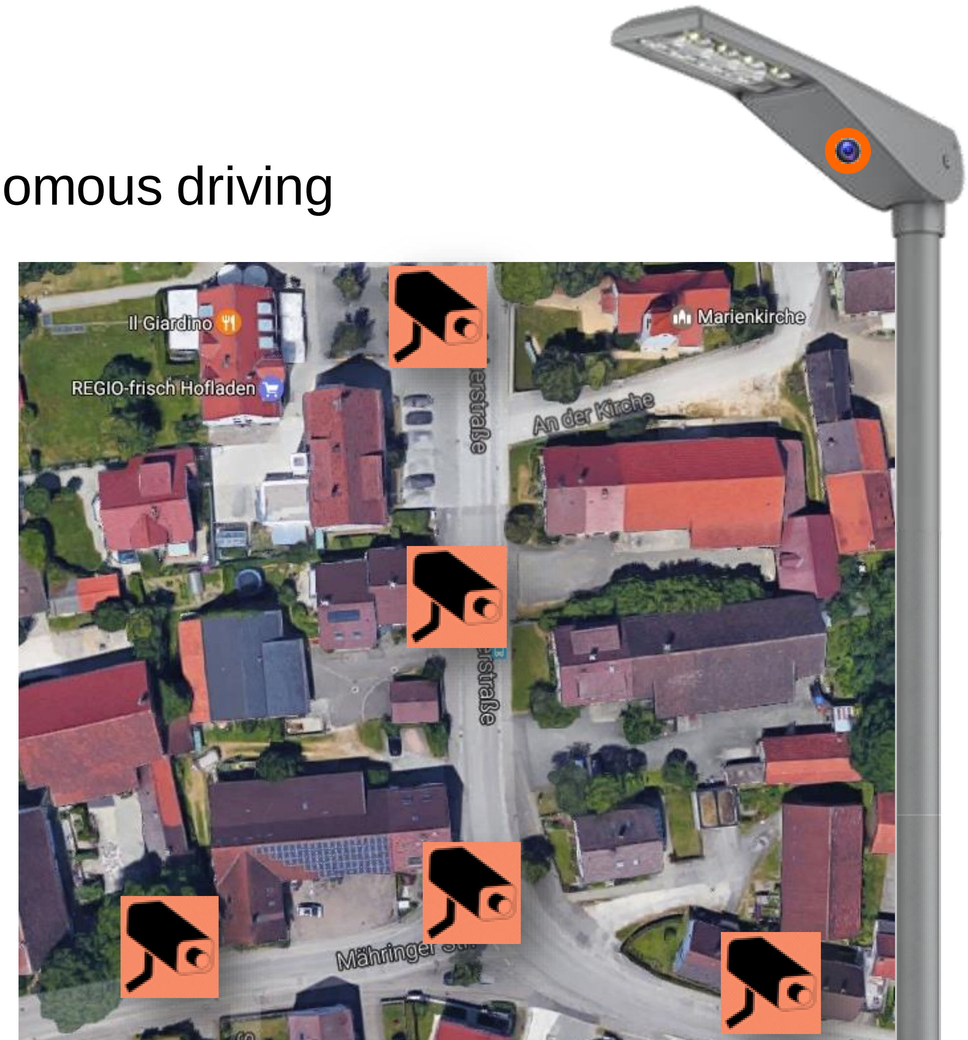
BOSCH

TOMTOM

NOKIA

universität ulm

DAIMLER



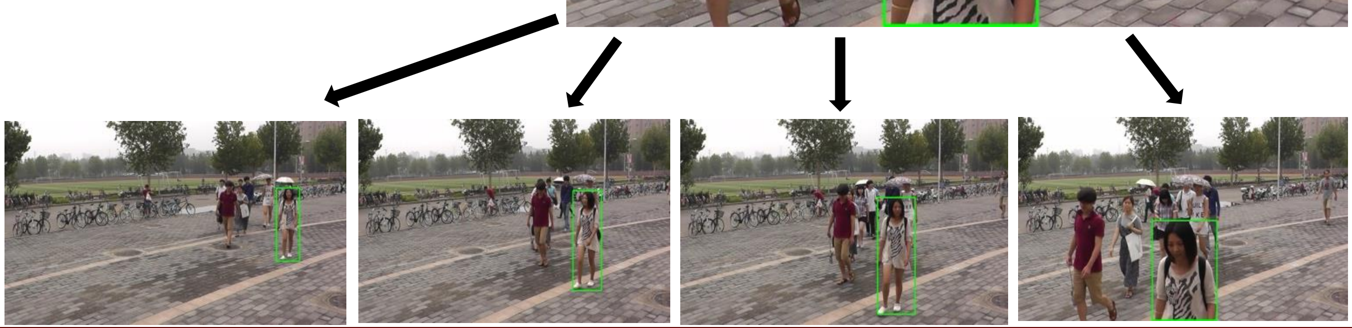
- Since 2019
  Perception and Intelligence Lab (PINLab)
  at Sapienza

Detection and re-identification of people
CVPR’19, BMVC’19-’20, IMAVIS’20, ACMSurveys’22

- Find queried people with a Siamese CNN model with Attention
- Unified re-id and few-shot learning, not just for people

Query

Gallery



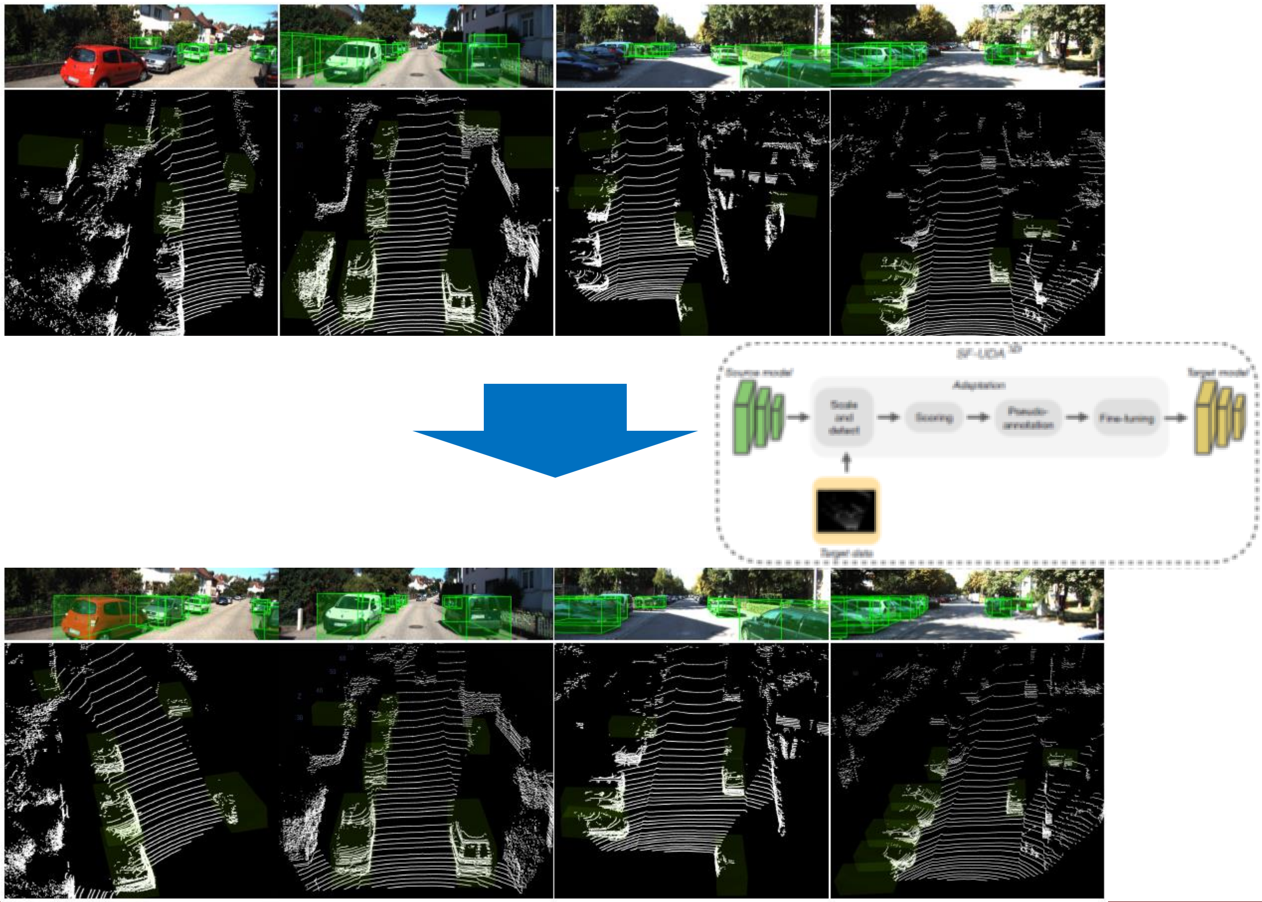
From Re-ID to Meta-Learning

- Surveillance
- Long-term tracking across-views
- One-shot understanding

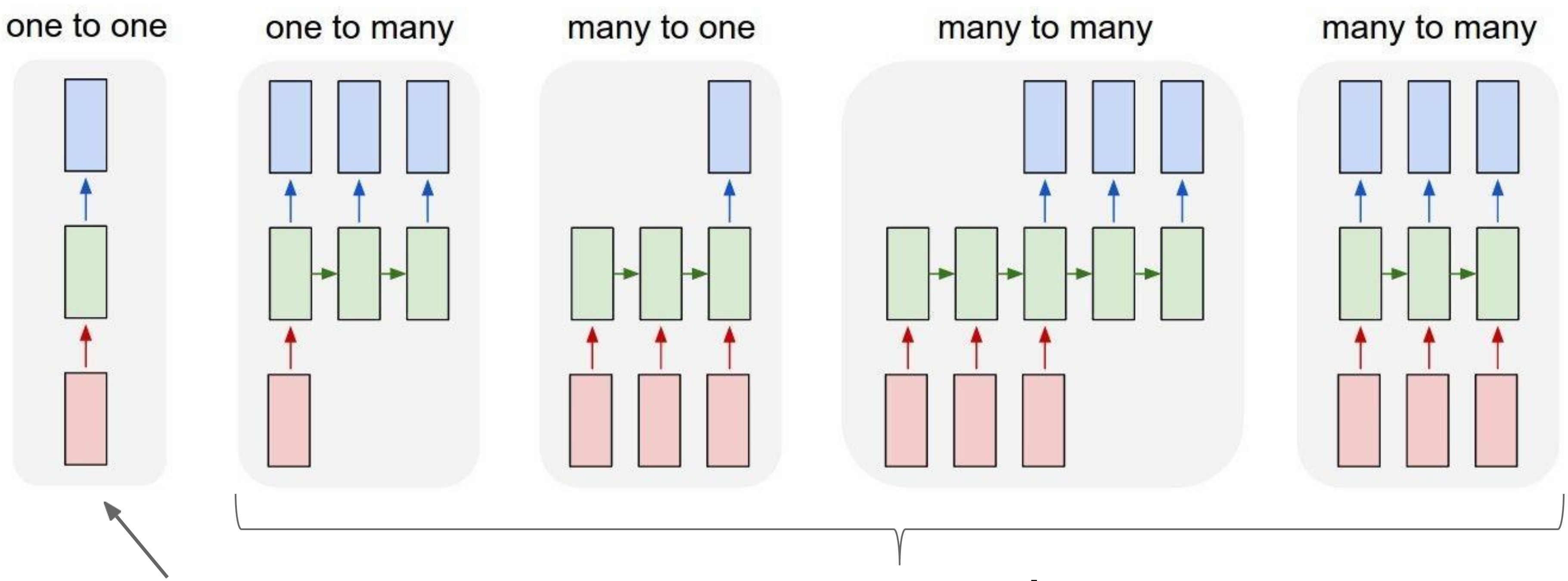
Image source: https://research.qut.edu.au/saivt/research-projects/person-re-identification-using-soft-biometrics/

Domain adaptation for 3D car detection
3DV’20, ECCV’22, TPAMI’23

- Adapt the detector to changes in the LiDAR sensor



Process Sequences

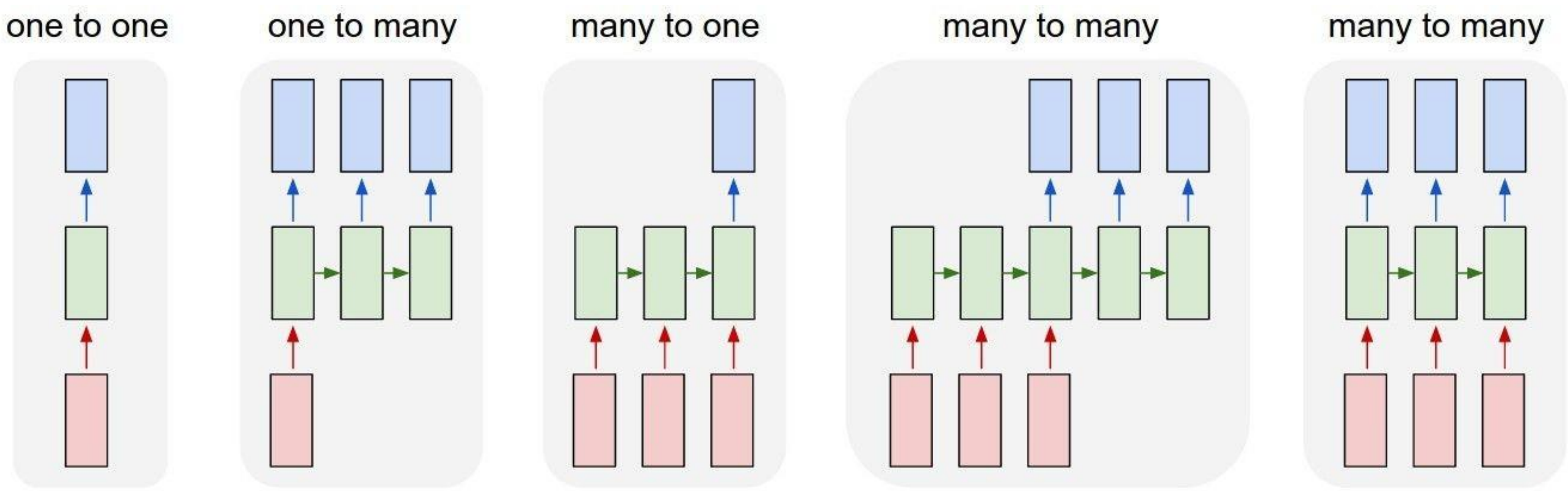


Vanilla Neural Networks
classification, detection, segmentation

Sequence tasks
captioning, video classification, forecasting, …

Image credits: Fei-Fei Li, Justin Johnson, Serena Yeung

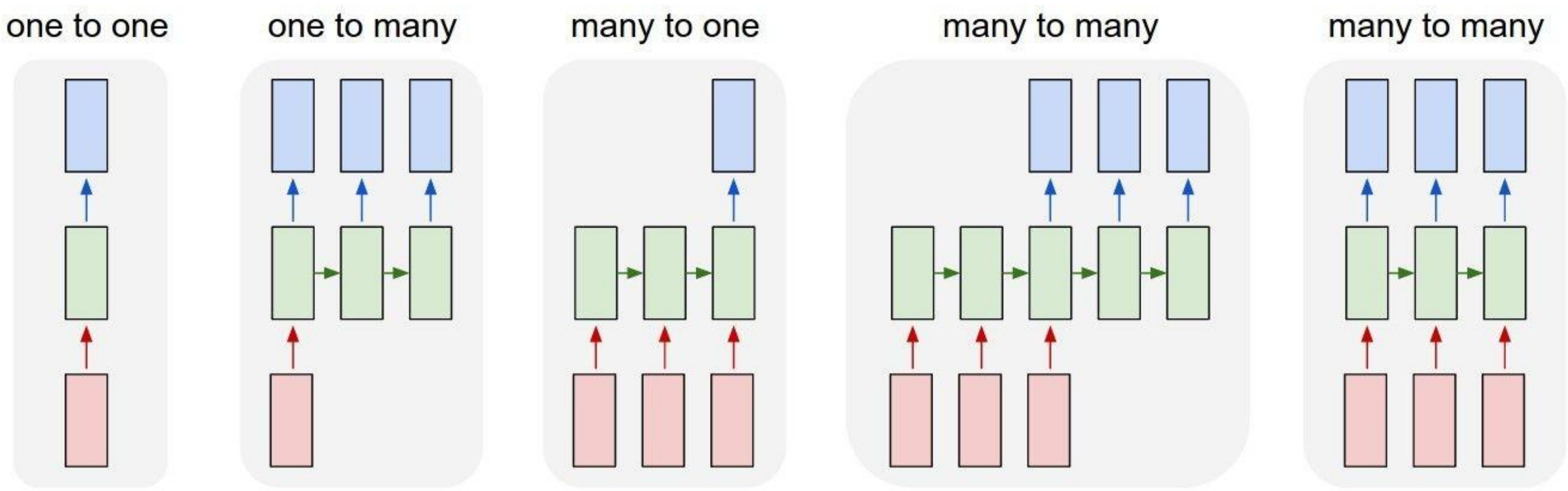
Process Sequences



e.g. **Image Captioning**
image -> sequence of words

Image credits: Fei-Fei Li, Justin Johnson, Serena Yeung

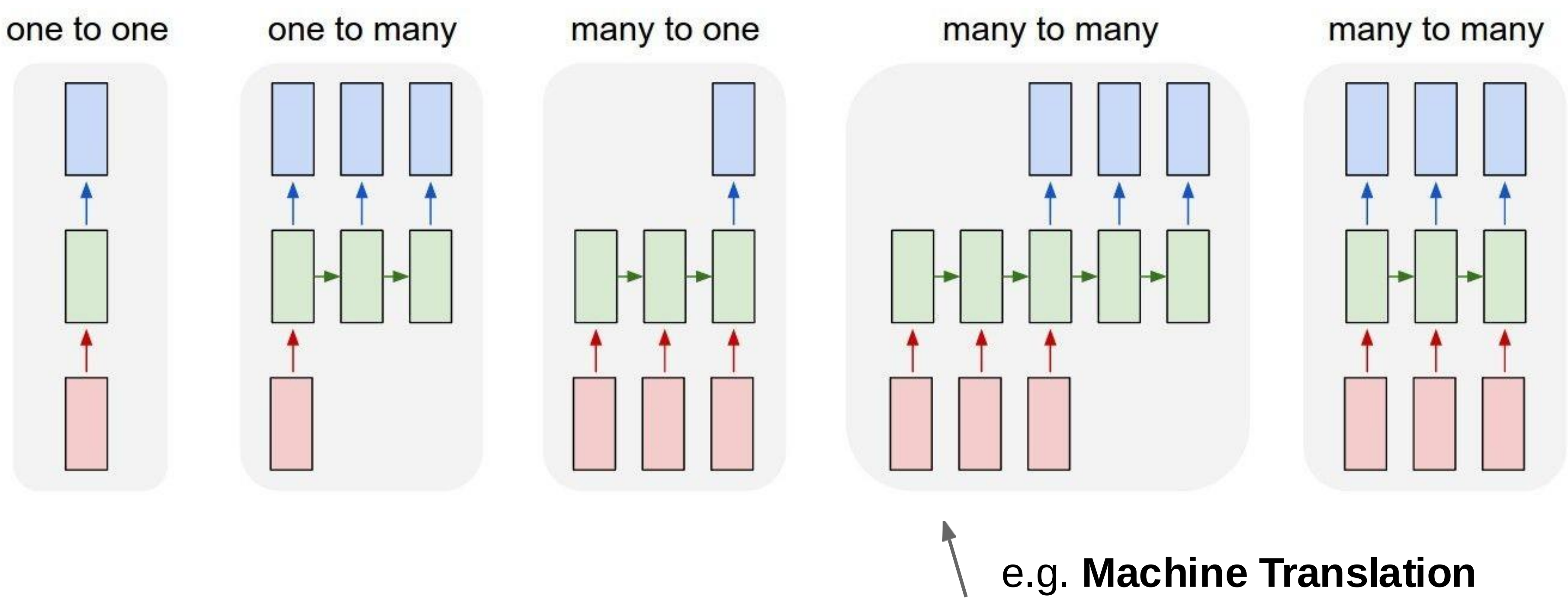
Process Sequences



e.g. **Sentiment Classification**
sequence of words -> sentiment

Image credits: Fei-Fei Li, Justin Johnson, Serena Yeung

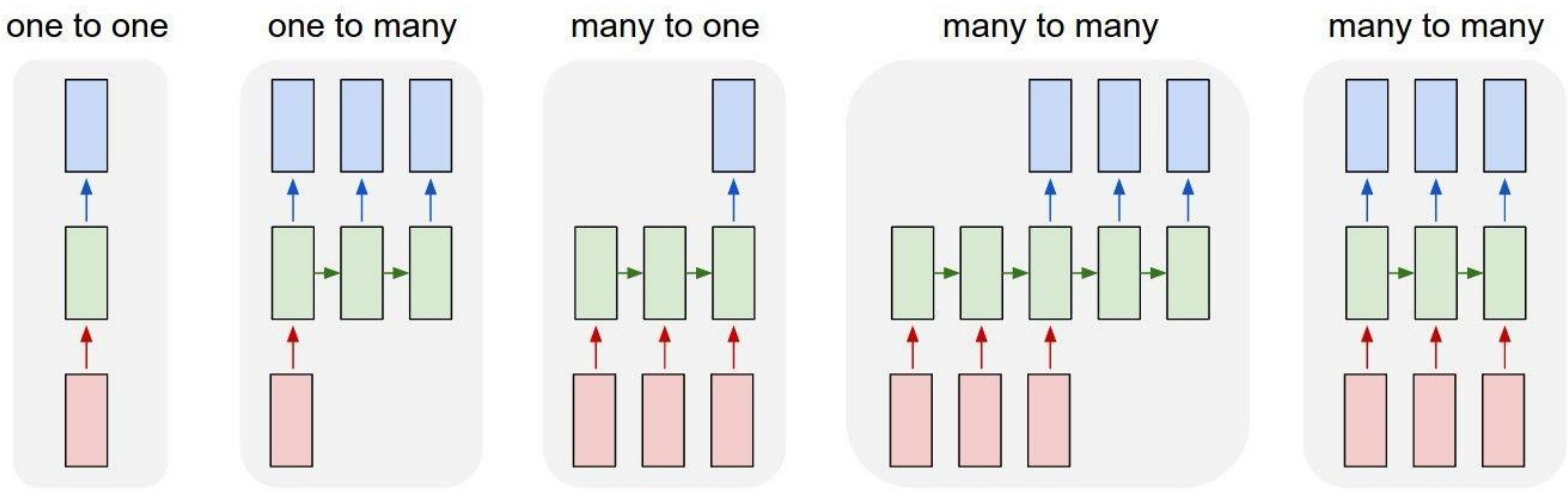
Process Sequences



e.g. **Machine Translation**
seq of words -> seq of words

Image credits: Fei-Fei Li, Justin Johnson, Serena Yeung

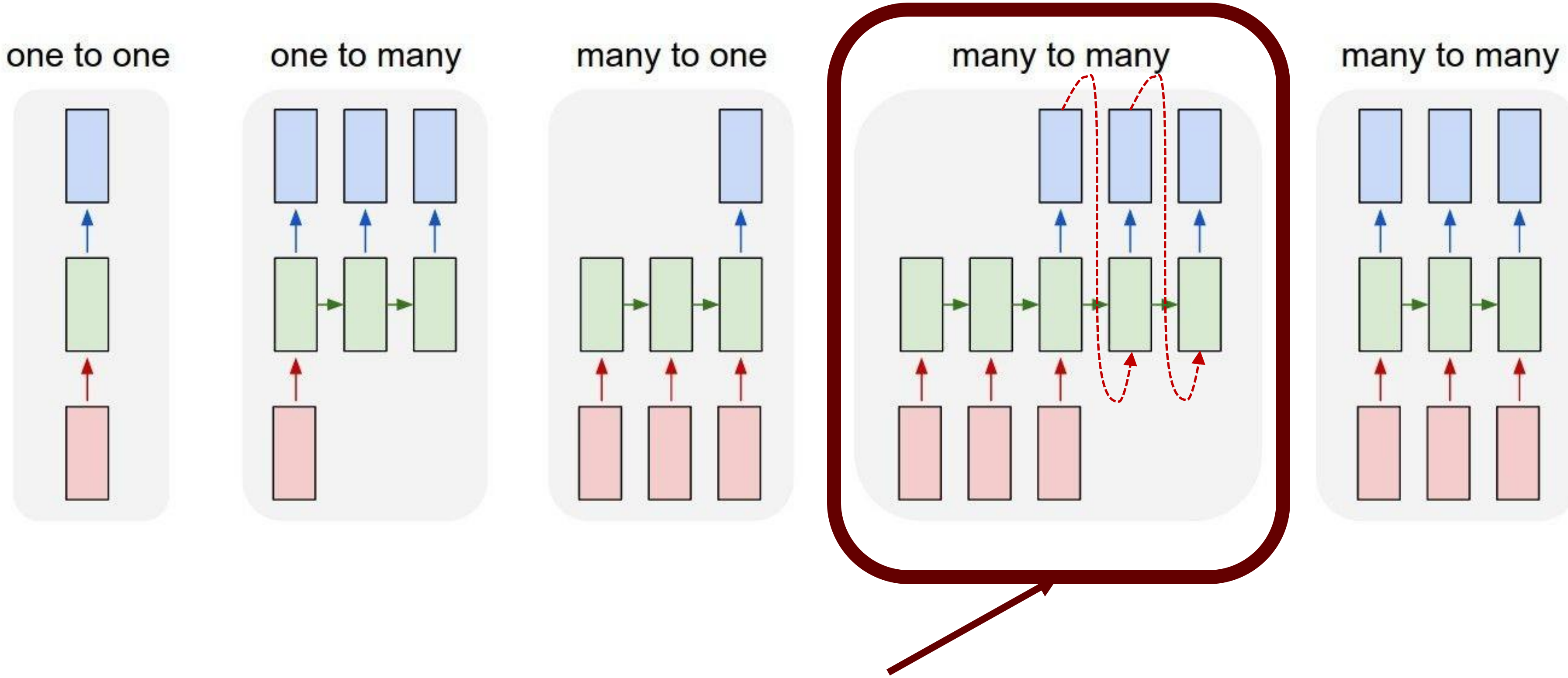
Process Sequences



e.g. **Video classification on frame level**

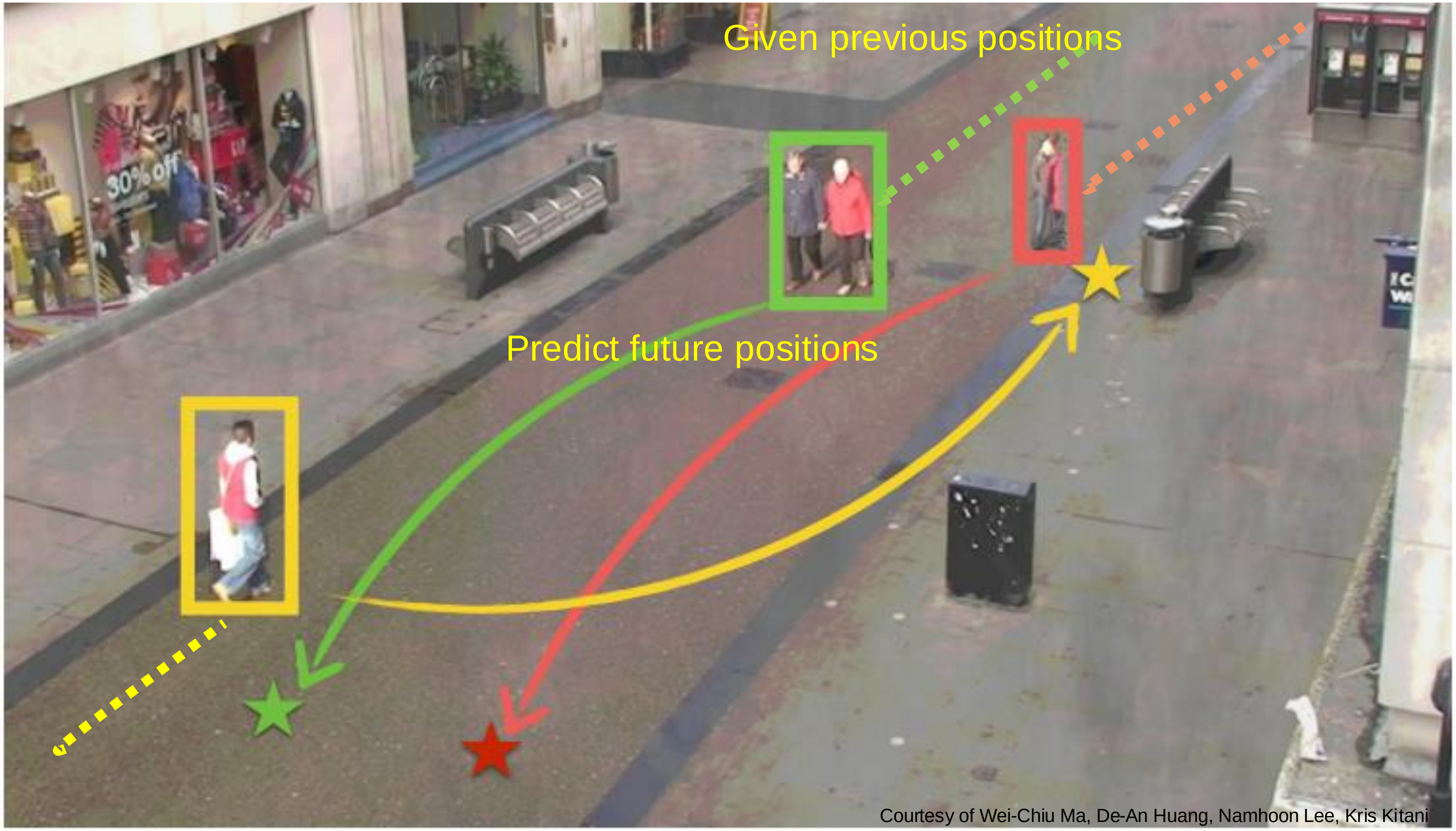
Image credits: Fei-Fei Li, Justin Johnson, Serena Yeung

Forecasting



Given some observations (history)
Predict the future auto-regressively

People Trajectory Forecasting



Courtesy of Wei-Chiu Ma, De-An Huang, Namhoon Lee, Kris Kitani

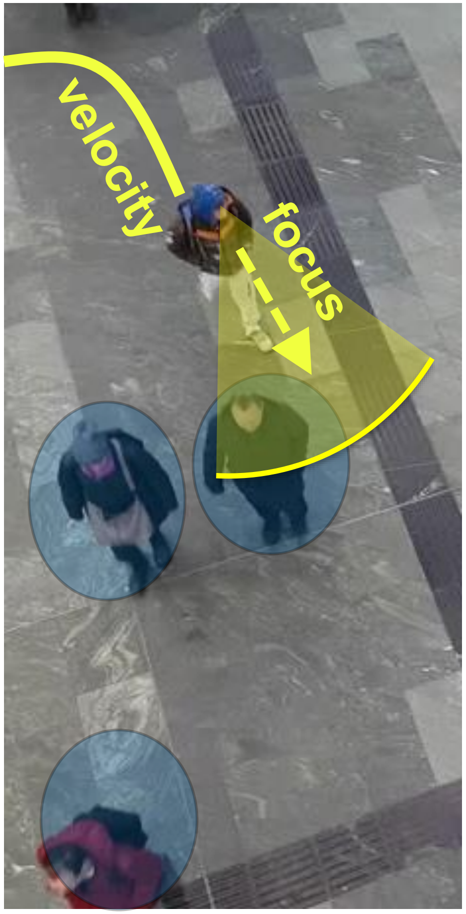
People Trajectory Forecasting

- For safety of autonomous vehicles

http://www.woostercollective.com/post/3d-optical-illusion-painted-on-street-to-make-drivers-slow-down

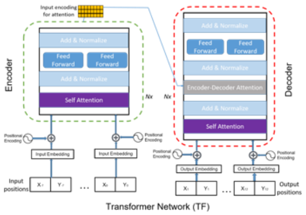
MX-LSTM: trajectory and head pose forecasting WACV'18, CVPR'18, TPAMI'19

- Predict jointly future motion and visual attention
- Condition social pooling on focus



Trajectory Forecasting with Transformers
ICPR'20, Pattern Recognition'23

- A better temporal model counts more than a social model
  - #1 on ETH+UCY, #2 on TrajNet

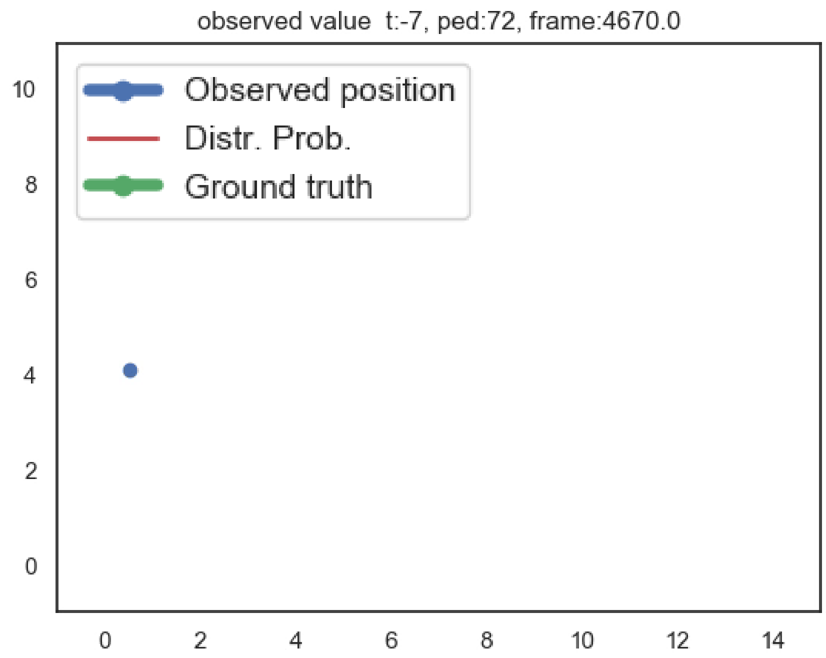


TrajNet

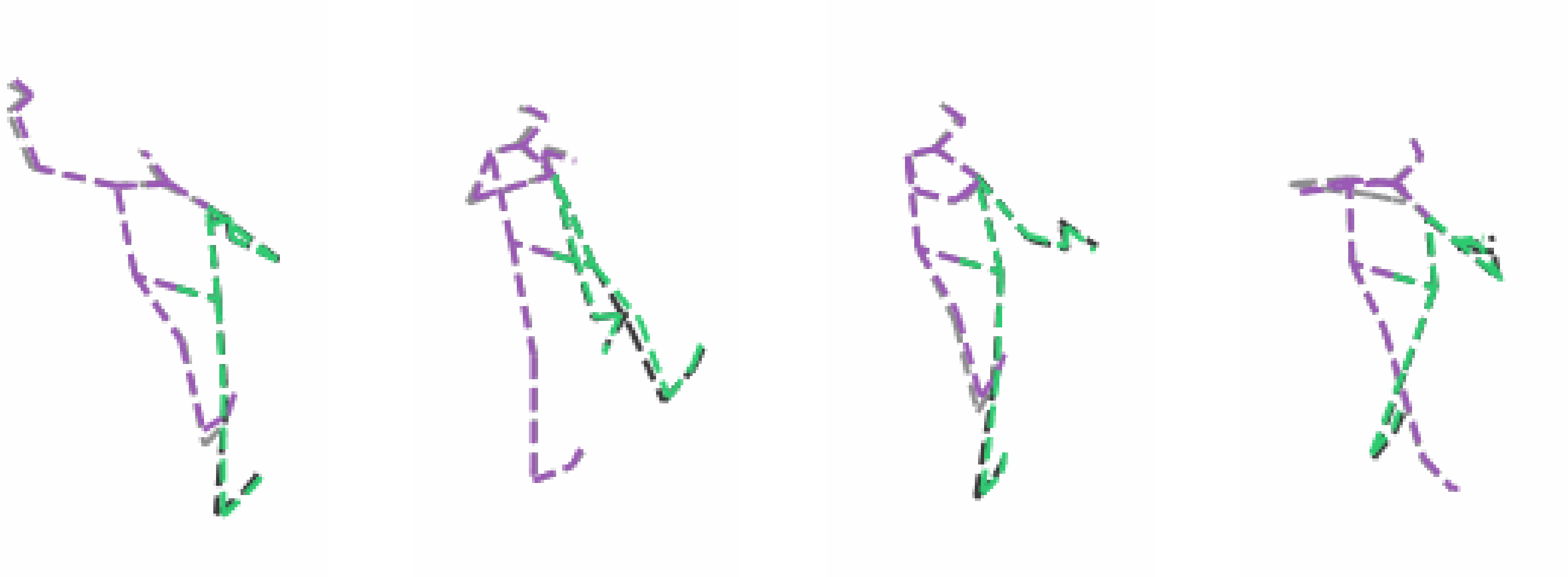
Rank	Method	Avg	FAD	MAD	Context	Cit.	Year
1	Ikg-TnT	0.769	1.183	0.356	s	[6]	2020
2	*TF* ***(ours)***	*0.776*	*1.197*	*0.356*	/		*2020*
3	REDv3	0.781	1.201	0.360	/	[4]	2019
7	SR-LSTM	0.816	1.261	0.37	s	[25]	2019
9	S.Forces (EWAP)	0.819	1.266	0.371	s	[13]	1995
15	Temp. ConvNet (TCN)	0.841	1.301	0.381	/	[3]	2018
16	TF_q	*0.858*	*1.300*	*0.416*	/		*2020*
17	N-Linear Seq2Seq	0.860	1.331	0.390	/	[4]	2018
18	MX-LSTM	0.887	1.374	0.399	s	[12]	2018
34	LSTM	1.140	1.793	0.491	/	[1]	2018
36	S-GAN	1.334	2.107	0.561	s	[10]	2018

ETH+UCY

	LSTM-based				TF-based ***(ours)***
	Individual	Social		Soc.+ map	Ind.
	S-GAN-ind [10]	S-GAN [10]	Trajectron++ [22]	Soc-BIGAT [14]	TF_q (ours)
ETH	0.81/1.52	*0.87/1.62*	*0.43/0.86*	*0.69/1.29*	**0.61 / 1.12**
Hotel	0.72/1.61	*0.67/1.37*	*0.12/0.19*	*0.49/1.01*	**0.18 / 0.30**
UCY	0.60/1.26	*0.76/1.52*	*0.22/0.43*	*0.55/1.32*	**0.35 / 0.65**
Zara1	0.34/0.69	*0.35/0.68*	*0.17/0.32*	*0.30/0.62*	**0.22 / 0.38**
Zara2	0.42/0.84	*0.42/0.84*	*0.12/0.25*	*0.36/0.75*	**0.17 / 0.32**
Avg	0.58/1.18	*0.61/1.21*	*0.20/0.39*	*0.48/1.00*	**0.31 / 0.55**

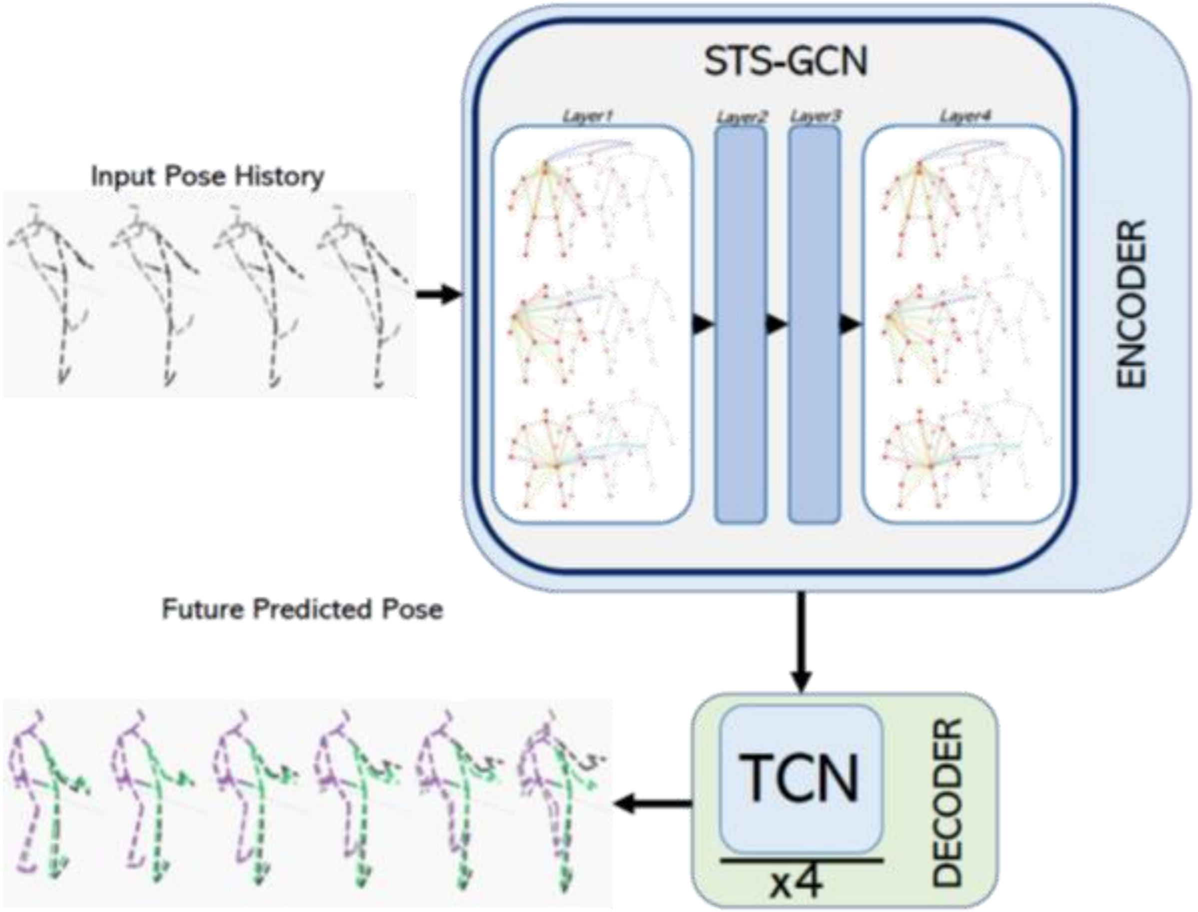


Human Pose Forecasting



Space-Time-Separable Graph Convolutional Network
STS-GCN, ICCV'21

- Encode body kinematics, decode future poses



Human Pose Forecasting for Human-Robot Cooperation

- For human-robot cooperation in shared workspaces
  - E.g. [Matthias et al. ISR'16]

http://www.woostercollective.com/post/3d-optical-illusion-painted-on-street-to-make-drivers-slow-down

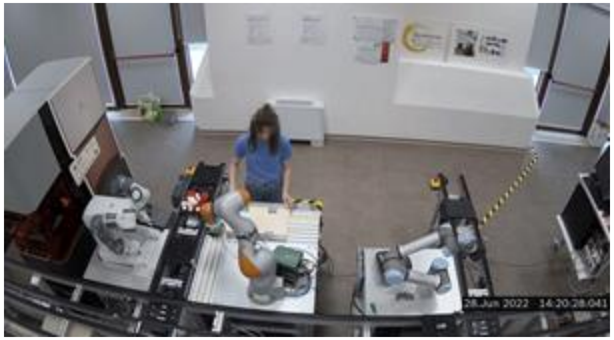
Human Pose Forecasting for Human-Robot Cooperation

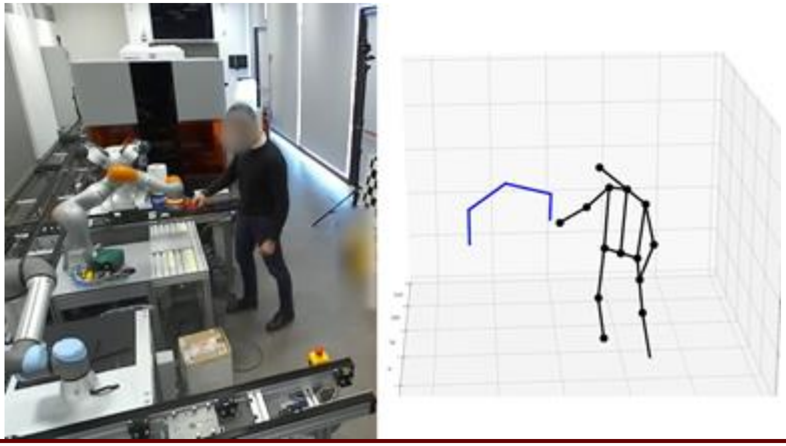
- For human-robot cooperation in shared workspaces
  - E.g. [Matthias et al. ISR’16]
- Teammates consider the consequences of their actions on others
  - E.g. [Shah et al. ACM-HRI’11]

CHICO: Cobots and Humans in Industrial COllaboration ECCV'22

- HRC in 7 industrial actions (reproduced assembly line, KUKA cobot)
  - Markerless, 3 RGBD camera views
  - 20 actors, 15 annotated joints, ~230 cobot-person collisions







CHICO: Cobots and Humans in Industrial COllaboration ECCV'22

- HRC in 7 industrial actions (reproduced assembly line, KUKA cobot)
  - Markerless, 3 RGBD camera views
  - 20 actors, 15 annotated joints, ~230 cobot-person collisions
- Tasks
  - Predict the human motion

Pose Forecasting	Average		Inference time	Parameters
msec	*400*	*1000*	*1000*	*1000*
HisRep	54.6	91.6	91	3.4 M
MSR-GCN	54.1	90.7	252	6.29 M
STS-GCN	53.0	87.4	**23**	57.6 k
SeS-GCN	**48.8**	**85.3**	**23**	58.6 k

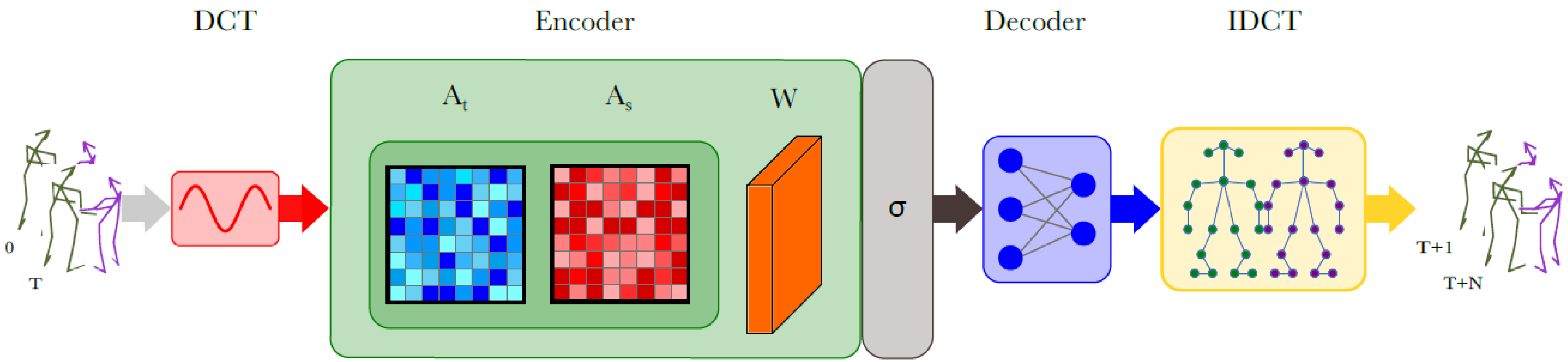
  - Detect collision

Collision Detection	1000 msec		
Metrics	*Prec*	*Rec*	*F1*
HisRep	0.63	0.58	0.56
MSR-GCN	0.63	0.30	0.31
STS-GCN	0.68	**0.61**	0.63
SeS-GCN	**0.84**	0.54	**0.64**

HisRep: W. Mao, et al., History repeats itself: Human motion prediction via motion attention. ECCV, 2020.
MSR-GCN: Dang et al,, Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction, ICCV, 2021.

Best Practices for Two-Body Pose Forecasting CVPR'23 wks

- What single-person best practice transfers to forecasting 2
  - ☑ Frequency input representation
  - ☑ Space-time separable GCN encoders
  - ☑ Learned graph connectivity and weights
  - Attention
  - Hierarchical body parts
  - ☑ CNN Vs. MLP decoders



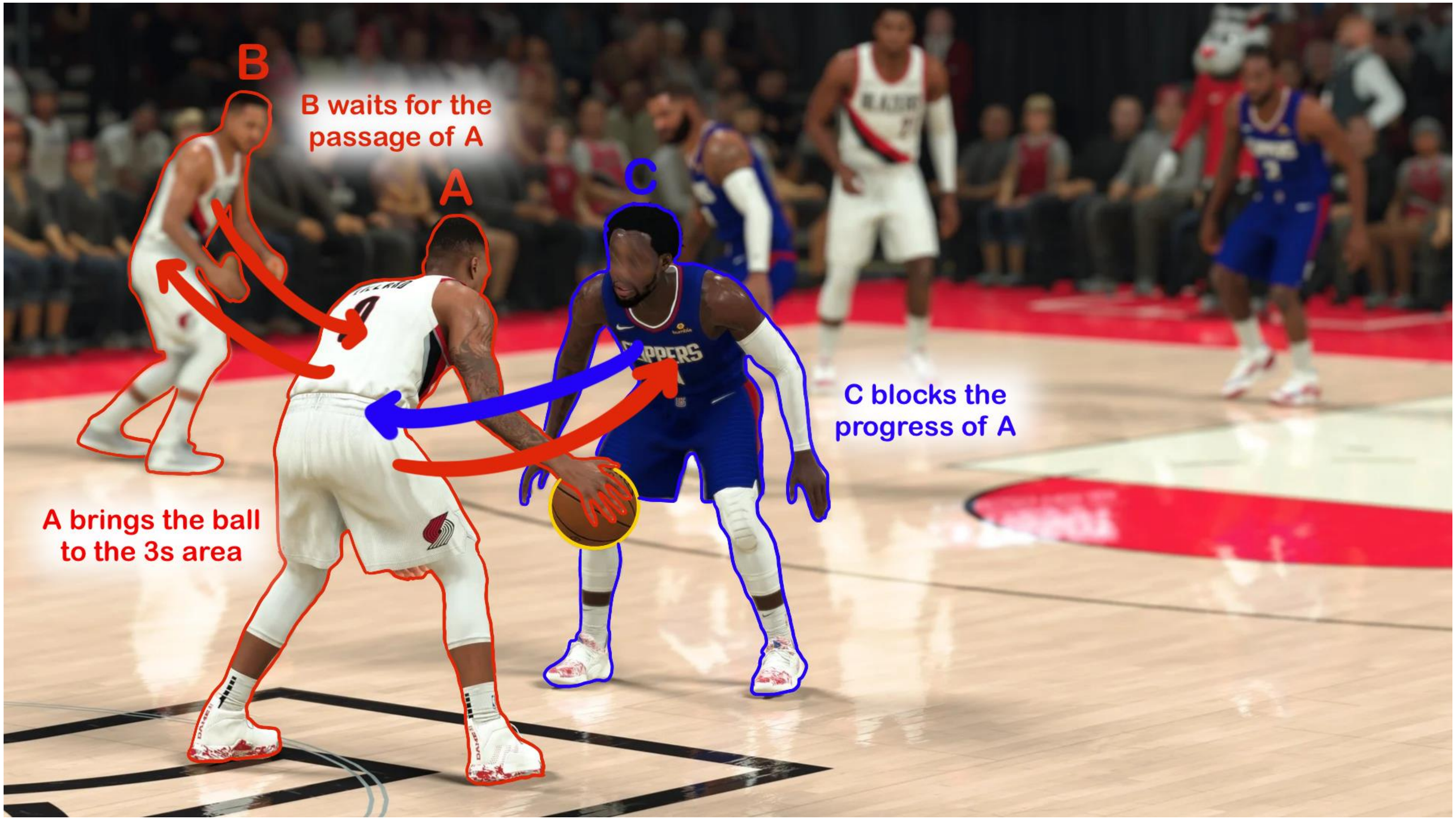
Best Practices for Two-Body Pose Forecasting CVPR’23 wks

- What single-person best practice transfers to forecasting 2
  - ☑ Frequency input representation
  - ☑ Space-time separable GCN encoders
  - ☑ Learned graph connectivity and weights
  - Attention
  - Hierarchical body parts
  - ☑ CNN Vs. MLP decoders

	Model	Input Repr.	Encoding					Decoding	MPJPE ↓				Param. ↓
		Freq. Enc. ✓	Learn. ✓	Sep. ✓	Init. ✓	Att.	Hier.	FC ✓	*200*	*400*	*600*	*1000*	(M)
1	[17]	✓	✓			✓			55	112	162	238	8.5
2	Space-time GCN		✓						108	152	255	379	1.08
3	*(kin. tree)*			✓					81	129	183	260	0.18
4			✓	✓					55	112	156	224	0.18
5	Input repr. practice	✓	✓	✓					41	88	135	219	0.18
6			✓	✓	✓				53	106	148	216	0.18
7	Encoder practices		✓	✓		✓†			55	112	157	228	9.9
8			✓	✓			✓		51	104	148	223	0.18
9	Decoder practices		✓	✓				✓	51	104	145	212	**0.17**
10		✓	✓	✓				✓	41	89	133	208	**0.17**
11		✓	✓	✓			✓	✓	51	104	146	217	**0.17**
12	**Best model**	✓	✓	✓	✓			✓	**39**	**86**	**129**	**202**	**0.17**

About latent roles in forecasting players in team sports ICLR'23 wks

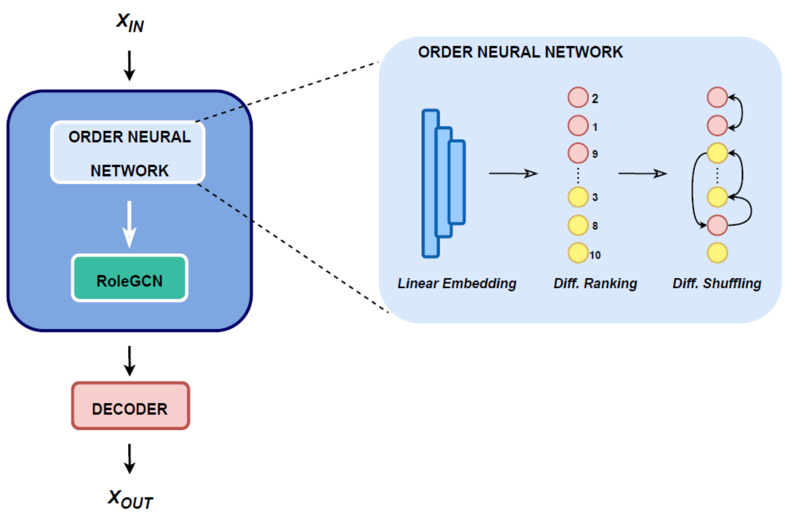
- Learn **role-based** interaction between basketball players

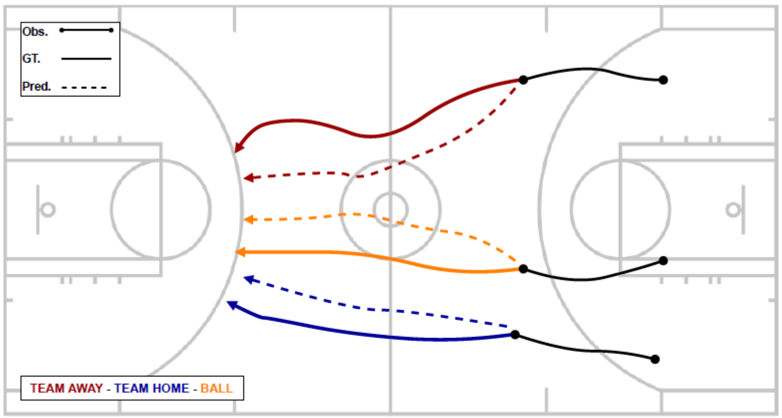


SAPIENZA UNIVERSITÀ DI ROMA

About latent roles in forecasting players in team sports ICLR'23 wks

- Learn **role-based** interaction between basketball players
  - Sort players (Order-NN)
  - Model role-based interaction with learned affinity terms (RoleGCN)
  - Decode future player positions

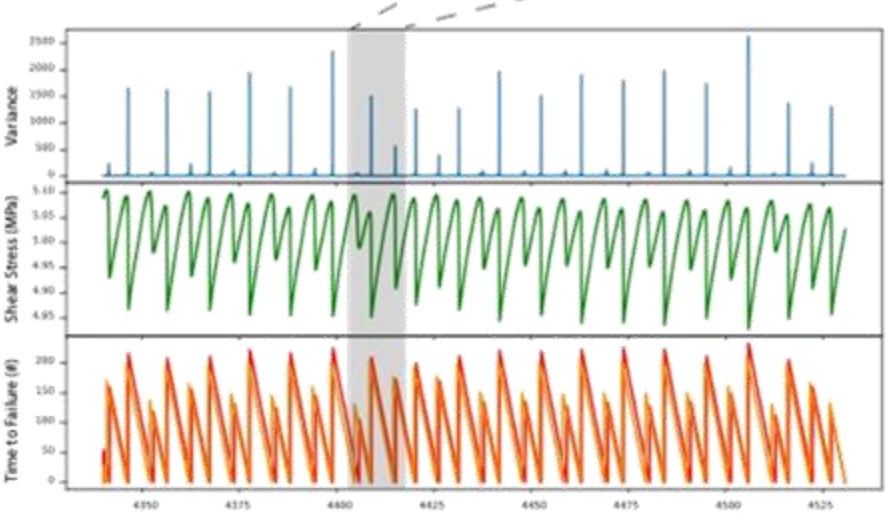




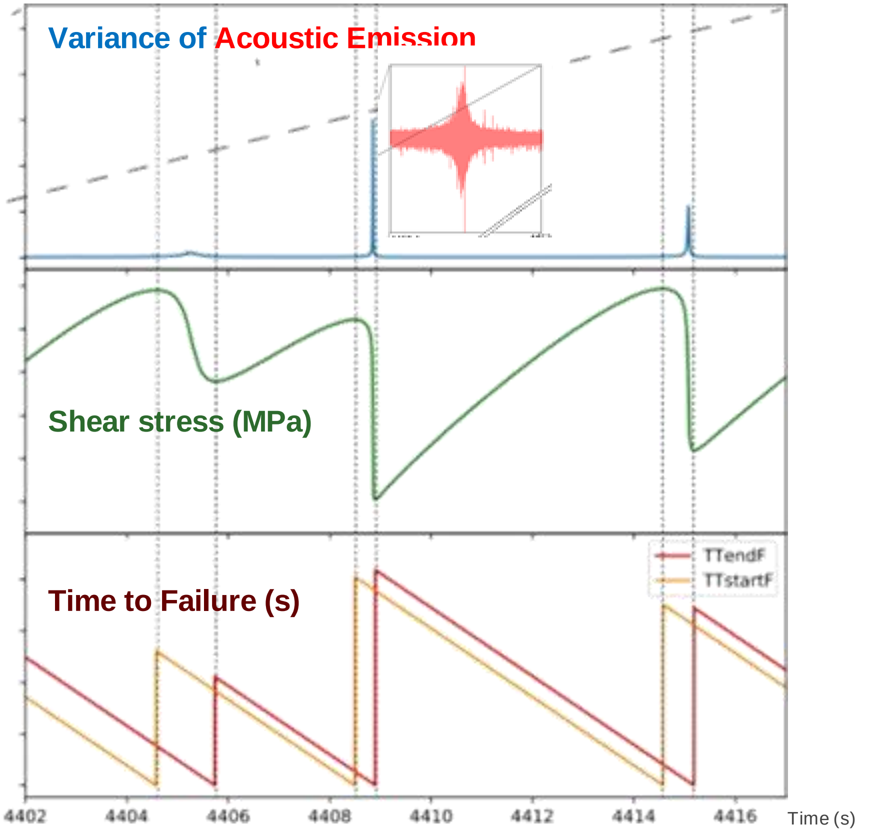
Also Forecasting ESC’21, EPSL’22

VIRTUAL 37th General Assembly of the European Seismological Commission ESC 2021 19-24 September 2021

- Earthquake forecasting
  - Observed: acoustic emissions
  - Latent: fault zone stress
  - Output: time to earthquake



Experiment p4679: fault gouge quartz; aperiodic slow and fast events

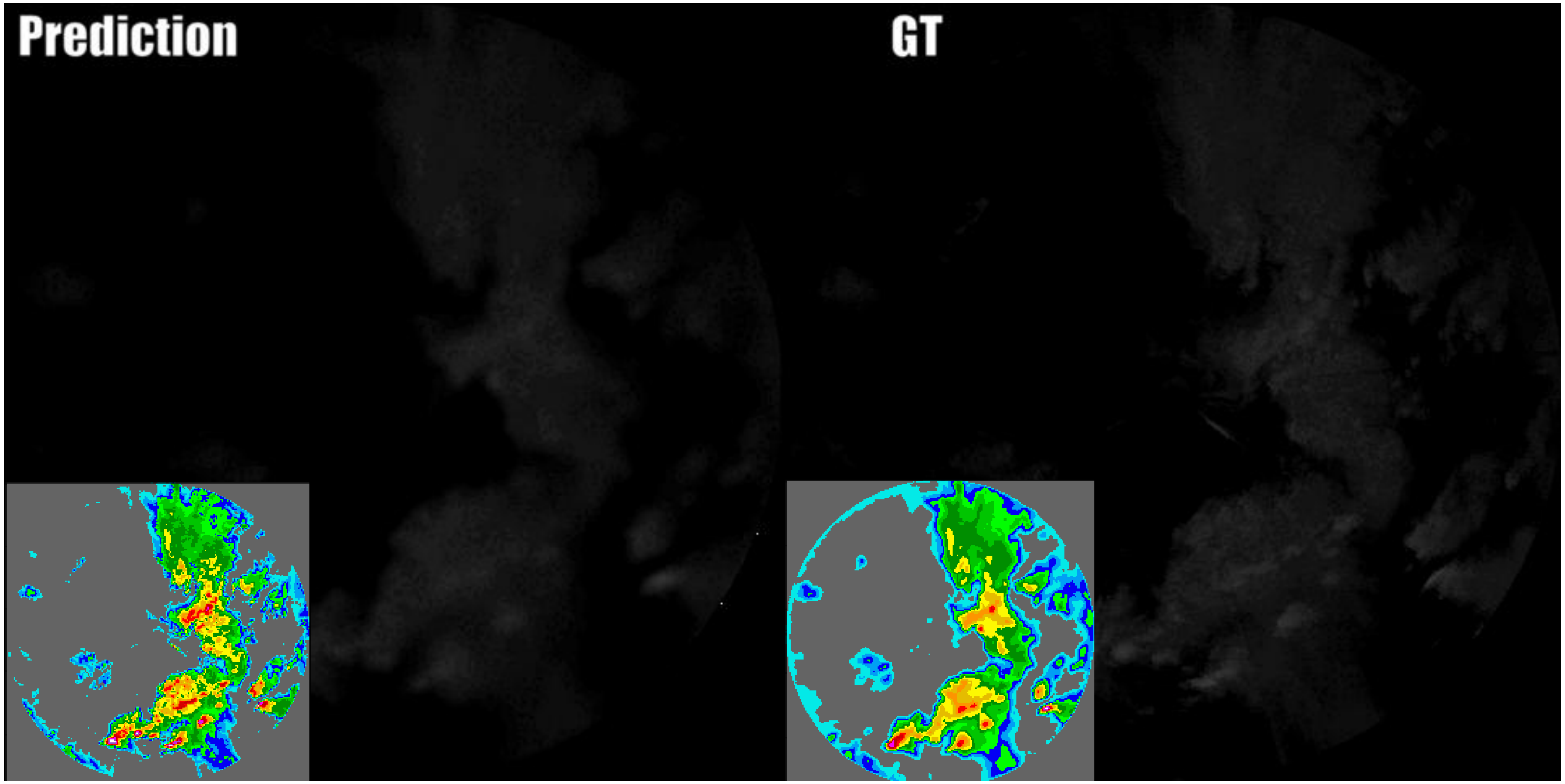


Laurenti, Tinti, Galasso, Franco, Marone (2021). ESC’21

Also Forecasting EGU’22

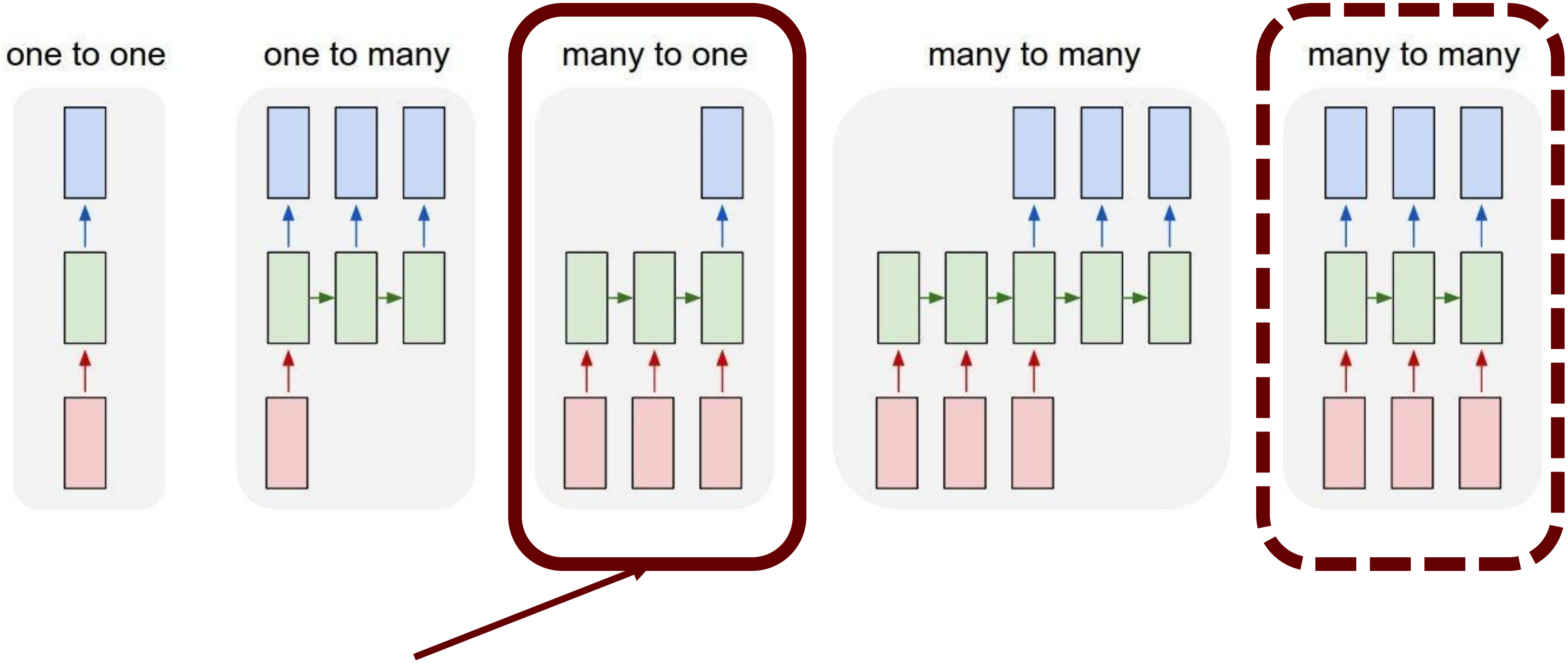
EGU General Assembly 2022

- Precipitation forecasting
  - Unet3D + STS-GCN for space-time predictions



Trappolini, Scofano, Sampieri, Messina, Galasso, Di Fabio, Marzano (2022). EGU’22

(Video) Anomaly Detection

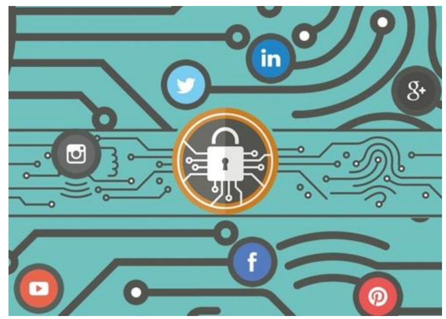


Given a sequence of observations
Predict anomalies (with *uncertainty)*

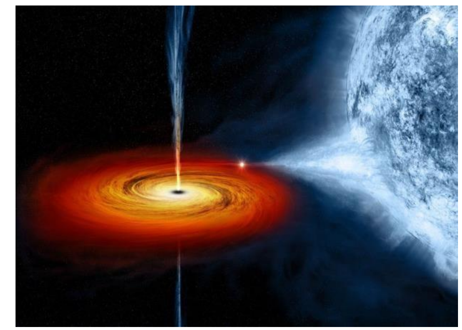
Anomaly Detection Applications

Cybersecurity:
attacks, malware, malicious apps/URLs, biometric spoofing

Social Network and Web Security:
false/malicious accounts, false/hate/toxic information

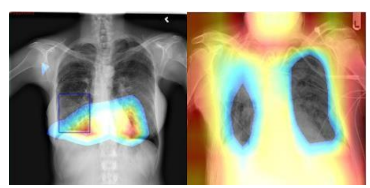


Astronomy:
Anomalous events

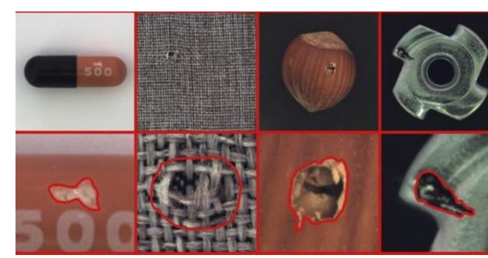


Finance:
credit card/insurance frauds, market manipulation, money laundering, etc.

Healthcare:
lesions, tumours, events in IoT/ICU monitoring, etc.



Industrial Inspection:
Defects, micro-cracks



Slide credit: Guansong Pang, Longbing Cao, Charu Aggarwal

Anomaly Detection Applications

Rover-Based Space Exploration:
unknown textures

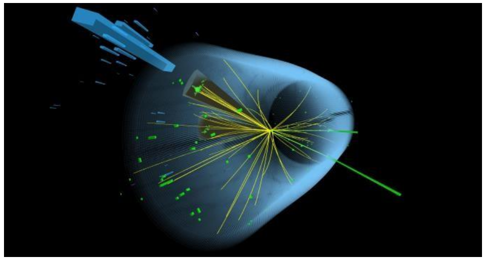
Bedrock
(Sol 1032)

Drill hole and tailings
(Sol 1496)

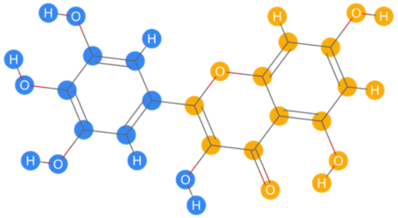
Video surveillance:
anomalous behavior, accidents, fights..

Normal Frame
Anomalous Frame

High-Energy Physics:
Higgs boson particles



Material Science:
exceptional molecule graphs



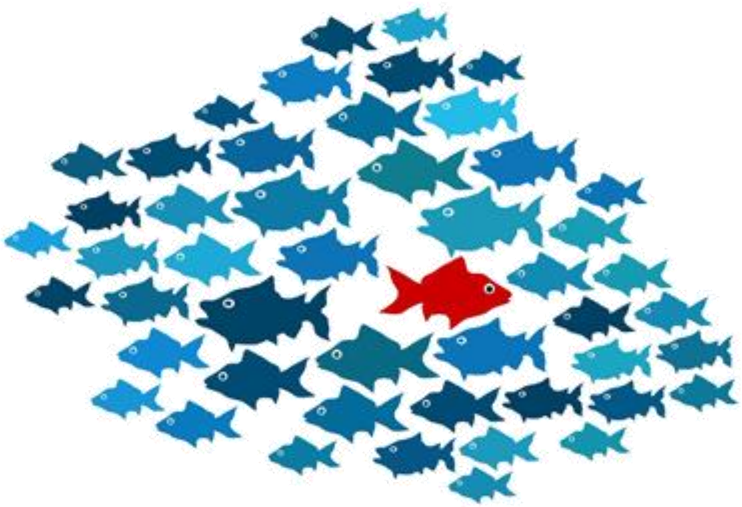
Drug Discovery:
rare active substances

Slide credit: Guansong Pang, Longbing Cao, Charu Aggarwal

Anomaly Detection
AIM’23, CVPR-wks’23, Pattern Recognition’23 (u. rev.)

- Target data
  - Financial series (NAB)
  - IT systems (YAHOO)
  - Mars aerospace measurements (NASA)
  - Medical data on elderly from sensor data (CASA)
  - Industrial water treatment (SWaT)
  - Anomalous human behavior (UBnormal)

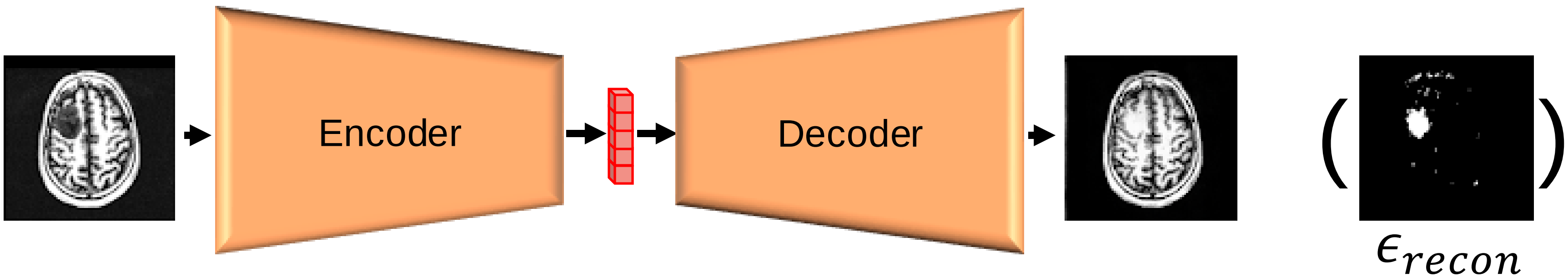
- **Real-world problem formulation**
  - Train on normalcy just (*aka* **OCC**)
  - Novel classes of *test* anomaly (**open set**)



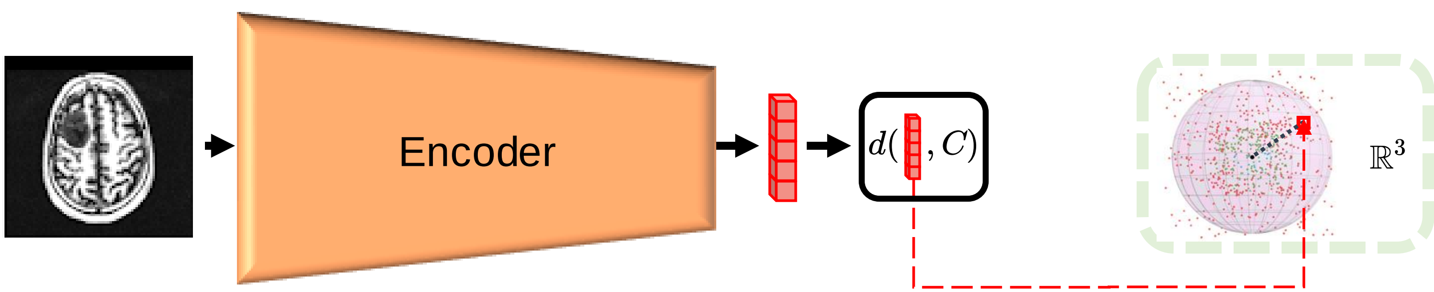
https://static.tildacdn.com/tild3131-3237-4364-b662-663731666262/anomaly_detection.png

Anomaly Detection

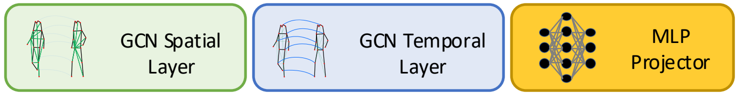
- Learn to reconstruct normalcy, compare input Vs. reconstructed



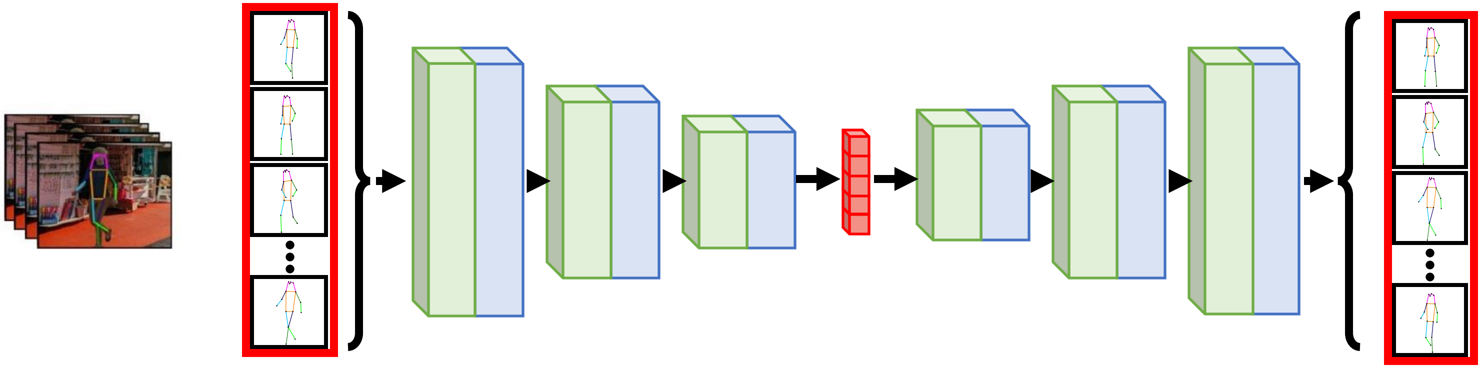
- Constrain normalcy into a hypersphere, measure dist. from center



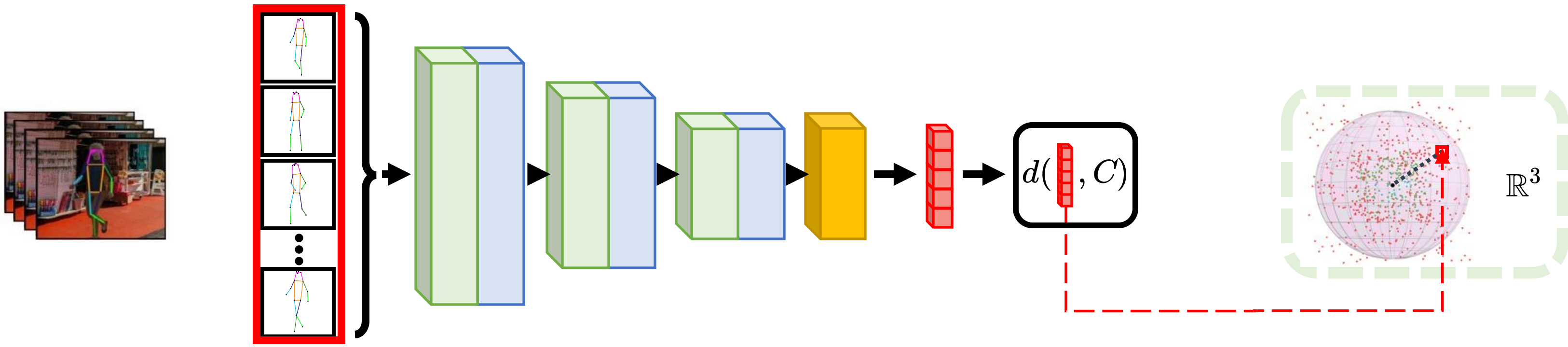
Skeletal Motion-based Anomaly Detection



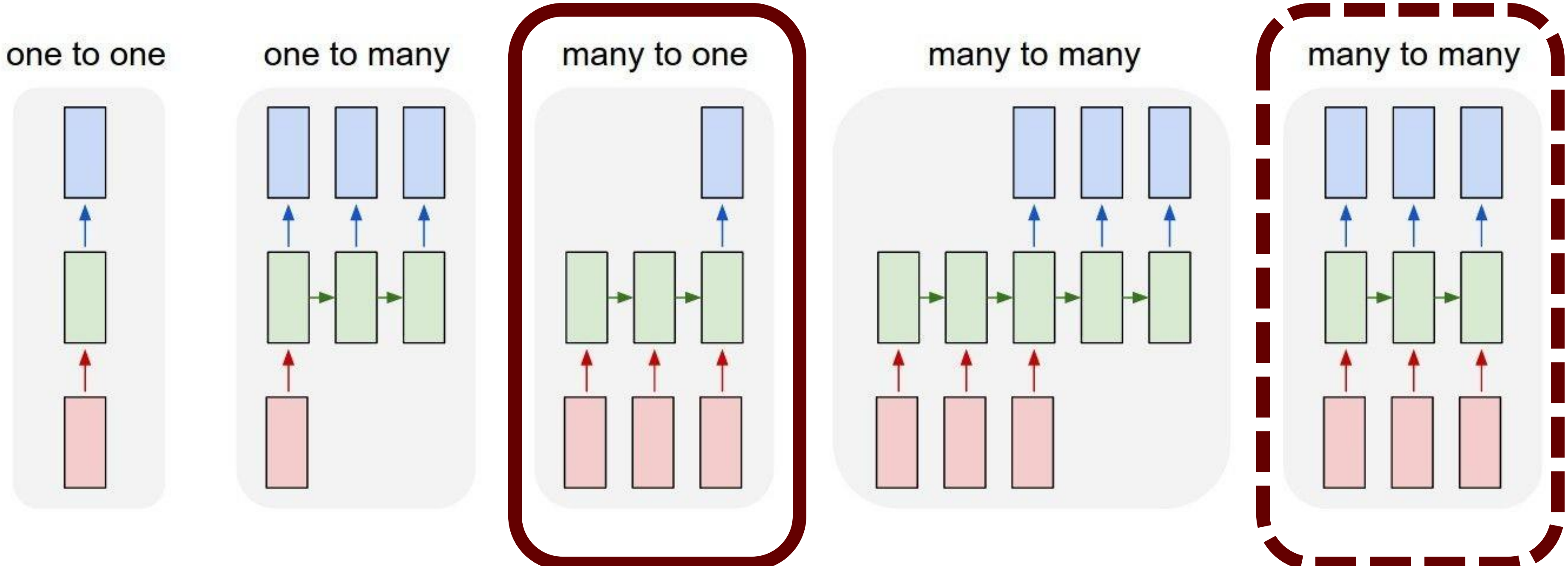
- Learn to reconstruct normalcy, compare input Vs. reconstructed



- Constrain normalcy into a hypersphere, measure dist. from center



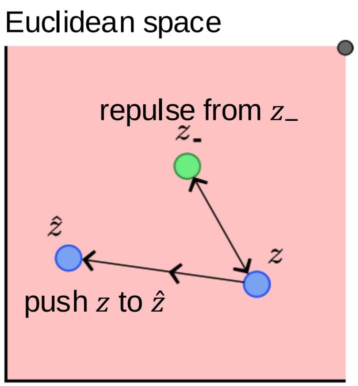
Activity recognition

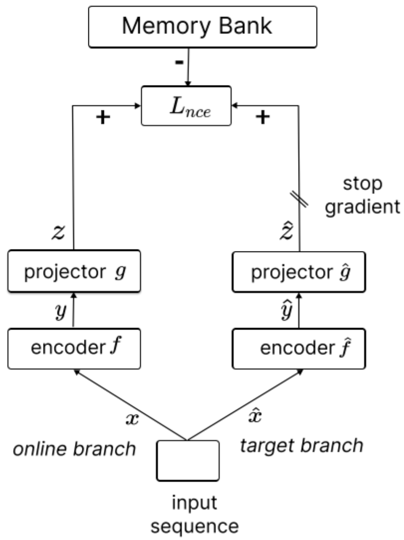


Recognize actions
from skeletal motions (with *uncertainty*)

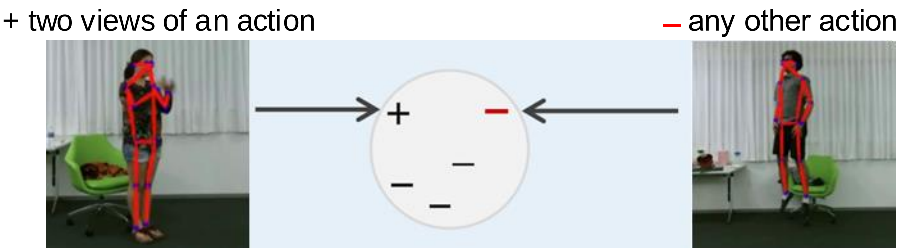
Skeleton-based SSL for action representations

- SoA builds on SkeletonCLR [Li et al. CVPR'21]



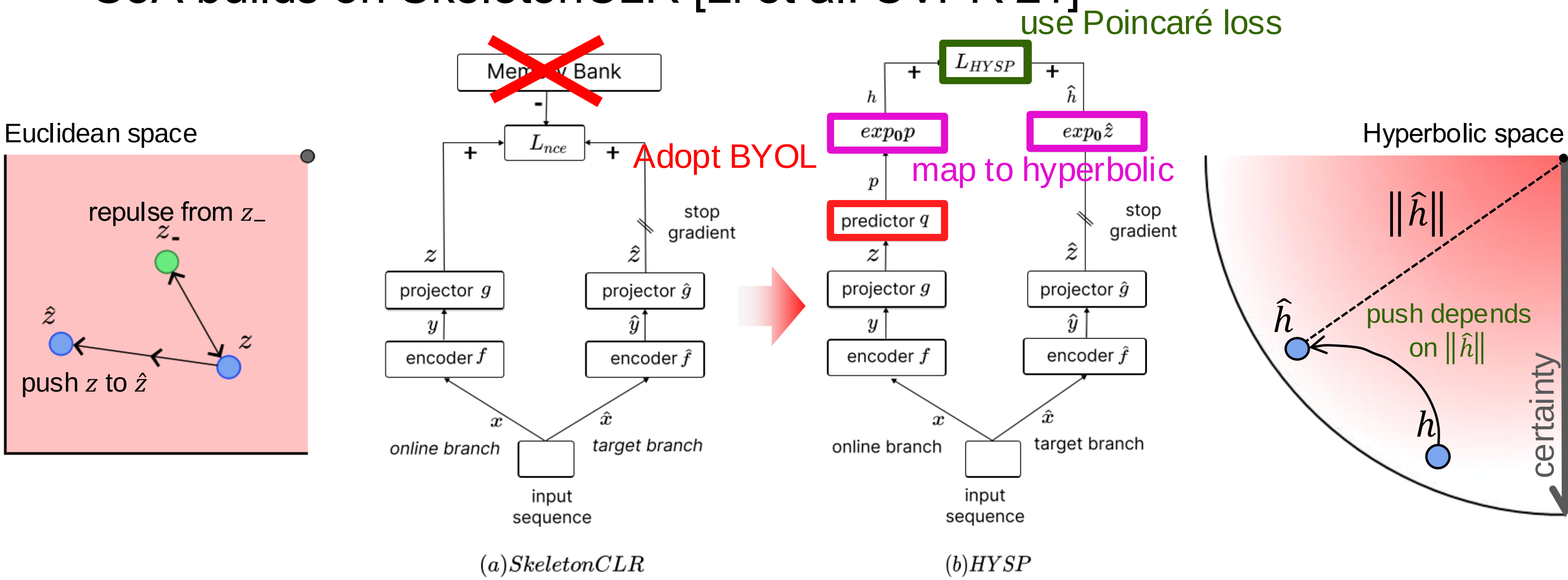


$(a) SkeletonCLR$



Hyperbolic Self-paced SSL (HYSP) ICLR'23

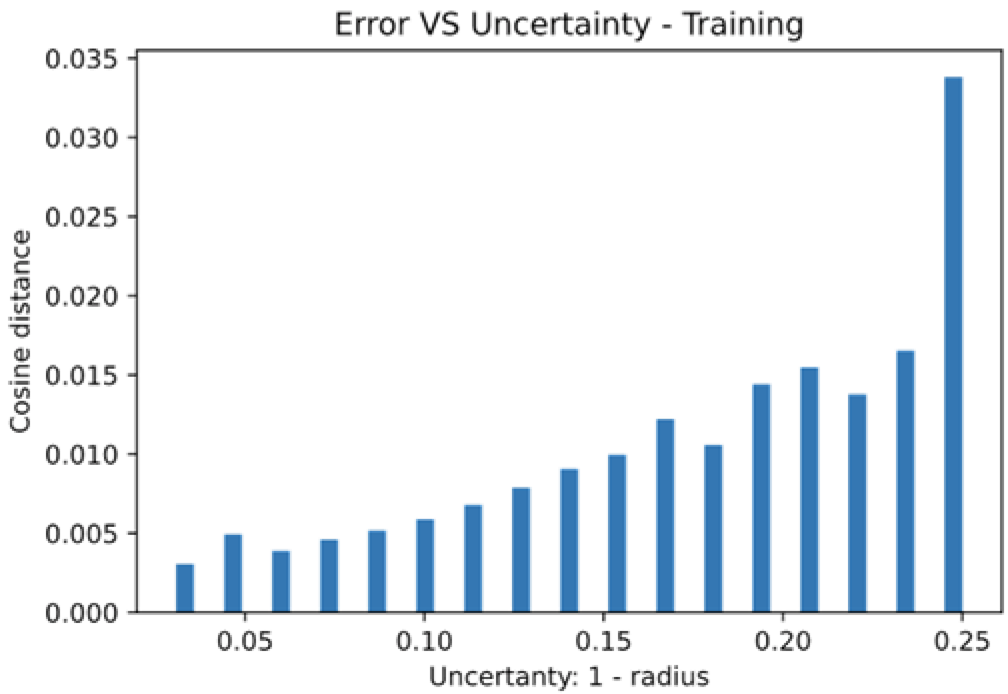
- SoA builds on SkeletonCLR [Li et al. CVPR'21]

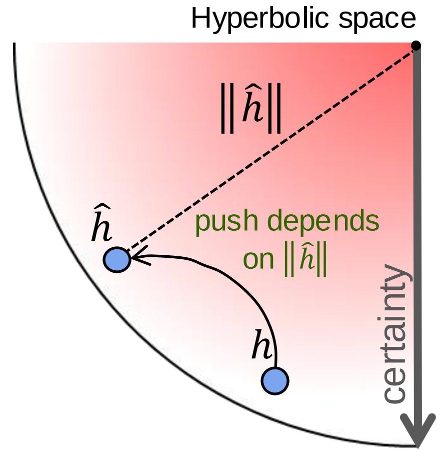


- Proposition: use hyperbolic uncertainty to self-pace SSL
  - More certain samples should drive learning more predominantly
- Hyperbolic Self-paced Self-Supervised Learning (HYSP)

HYSP after training

- End-to-end trained uncertainty matches the intuition
  - Large sample uncertainty corresponds to larger prediction errors (larger cosine distance)
  - Learn larger uncertainty for more ambiguous actions





Fabio Galasso

What is Data Science

Data is Everywhere

- Explosion in data-driven scientific discovery, business practices, medicine, education, politics, societal interventions, …
- And it's just the beginning
  - Ability to collect data across many domains will continue to accelerate
  - Data analysis techniques will continue to improve

"Data is the oil of the 21st century"

The Two Steps of Working with Data

(1) Collect data

Via computers, sensors, people, events …

(2) Do something with it

Make decisions, confirm hypotheses, gain insights, predict future …

"Data Science" = Going from (1) to (2)

This introduction

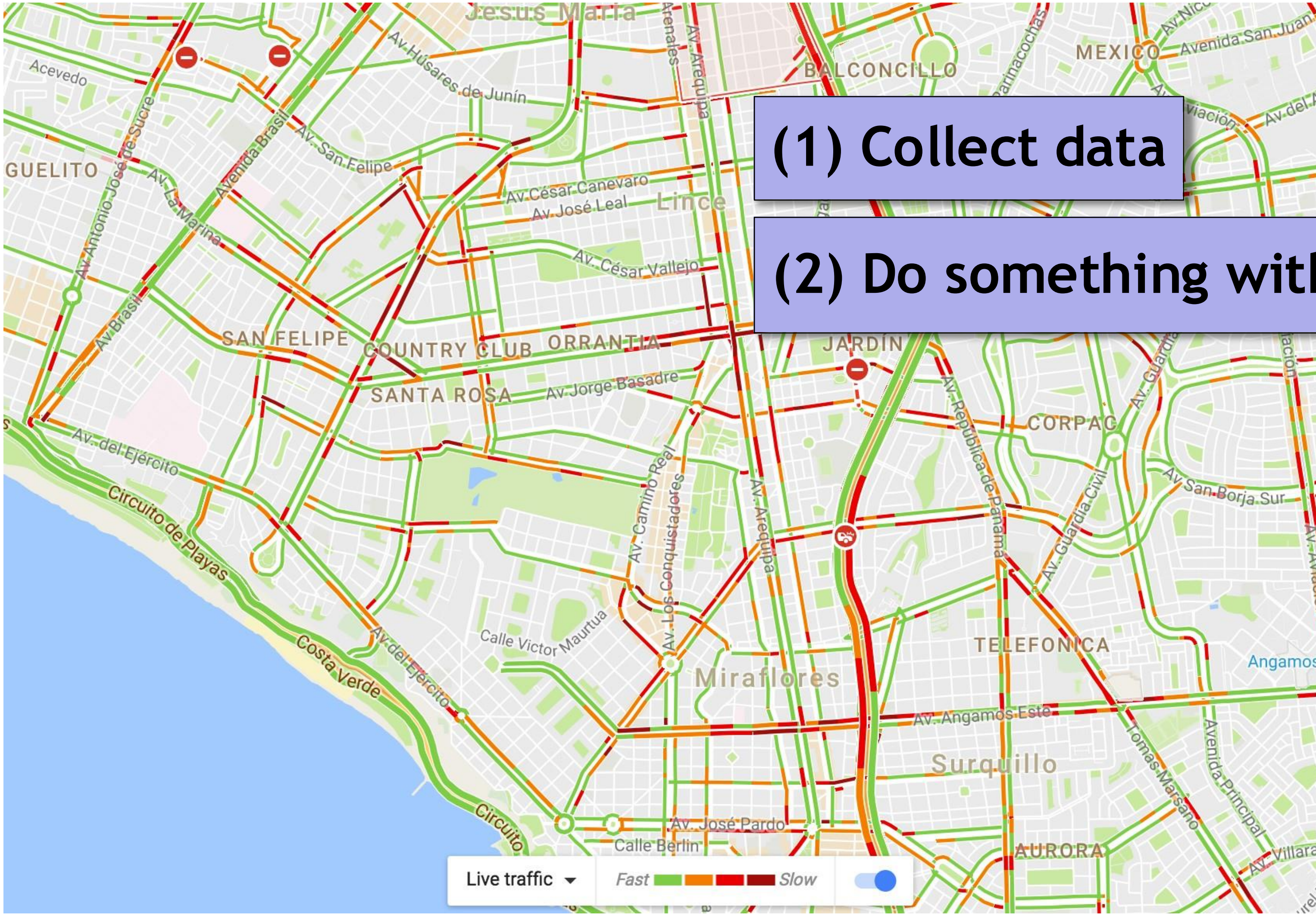
- Promises of data science
  - Applications and services
- Data tools and techniques
  - Database management systems
  - Data mining and machine learning
- Pitfalls in data science
  - Correlation and causation
  - Underfitting and overfitting
  - Privacy and a few others
- Data systems and platforms

Promises of Data Science

(1) Collect data

(2) Do something with it

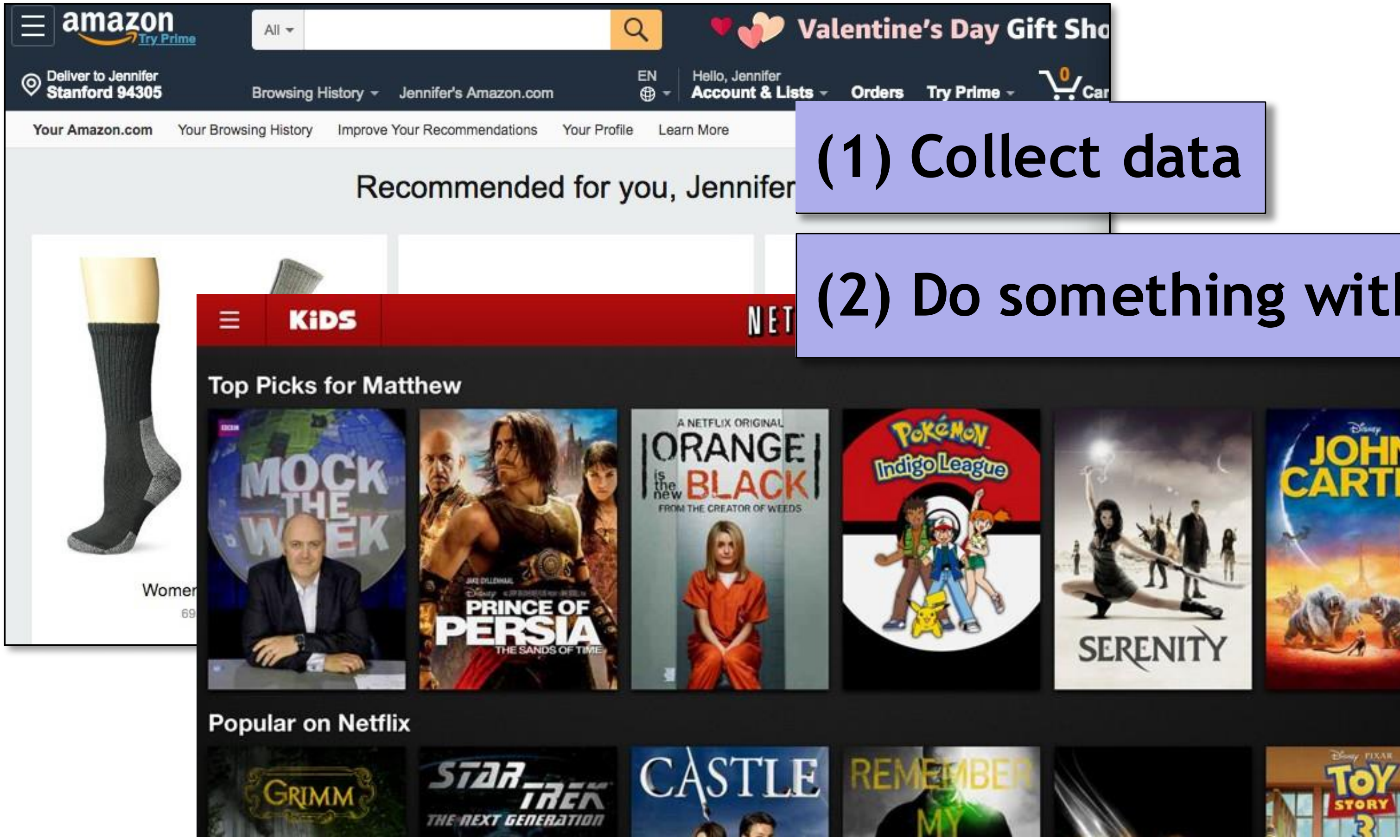
Traffic



(1) Collect data

(2) Do something with it

Recommender Systems

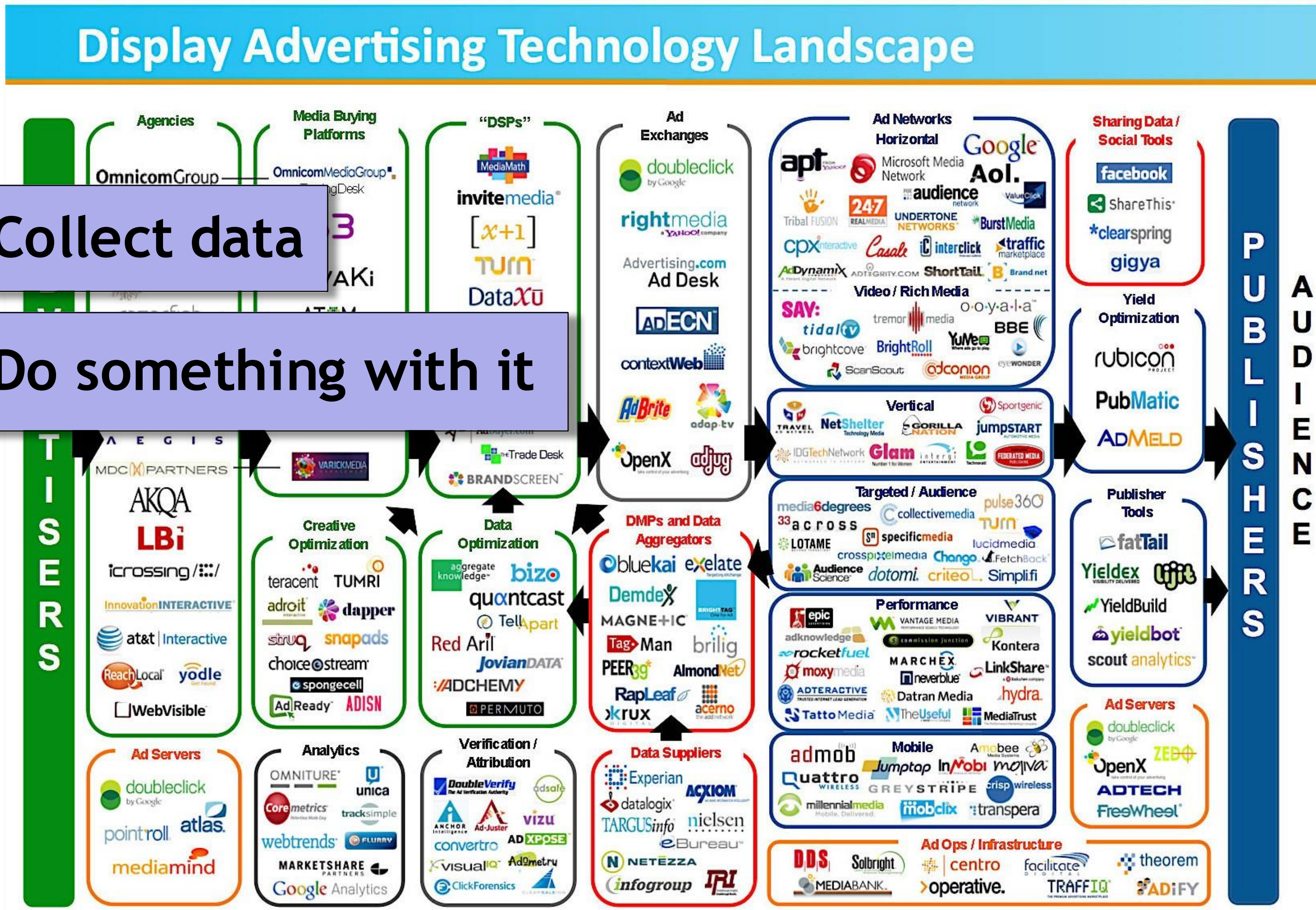


(1) Collect data

(2) Do something with it

+ music, news, friends, romantic partners, and many more!

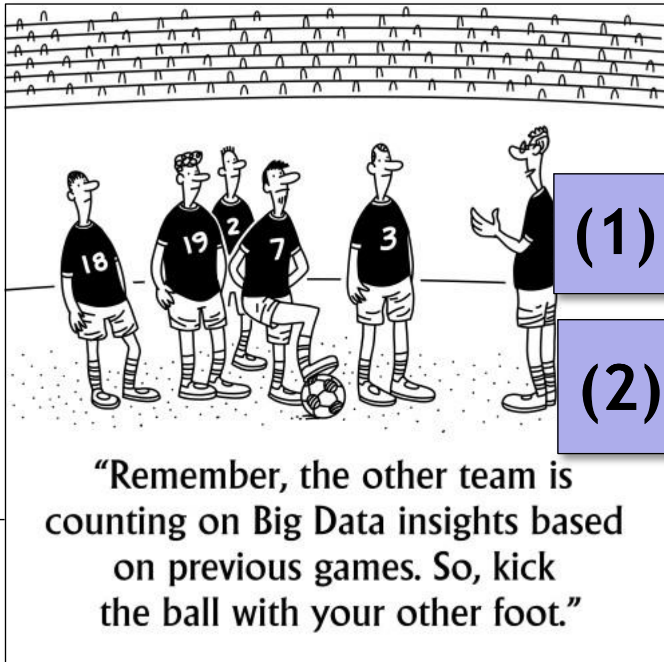
Online Advertising



(1) Collect data

(2) Do something with it

Sports



(1) Collect data

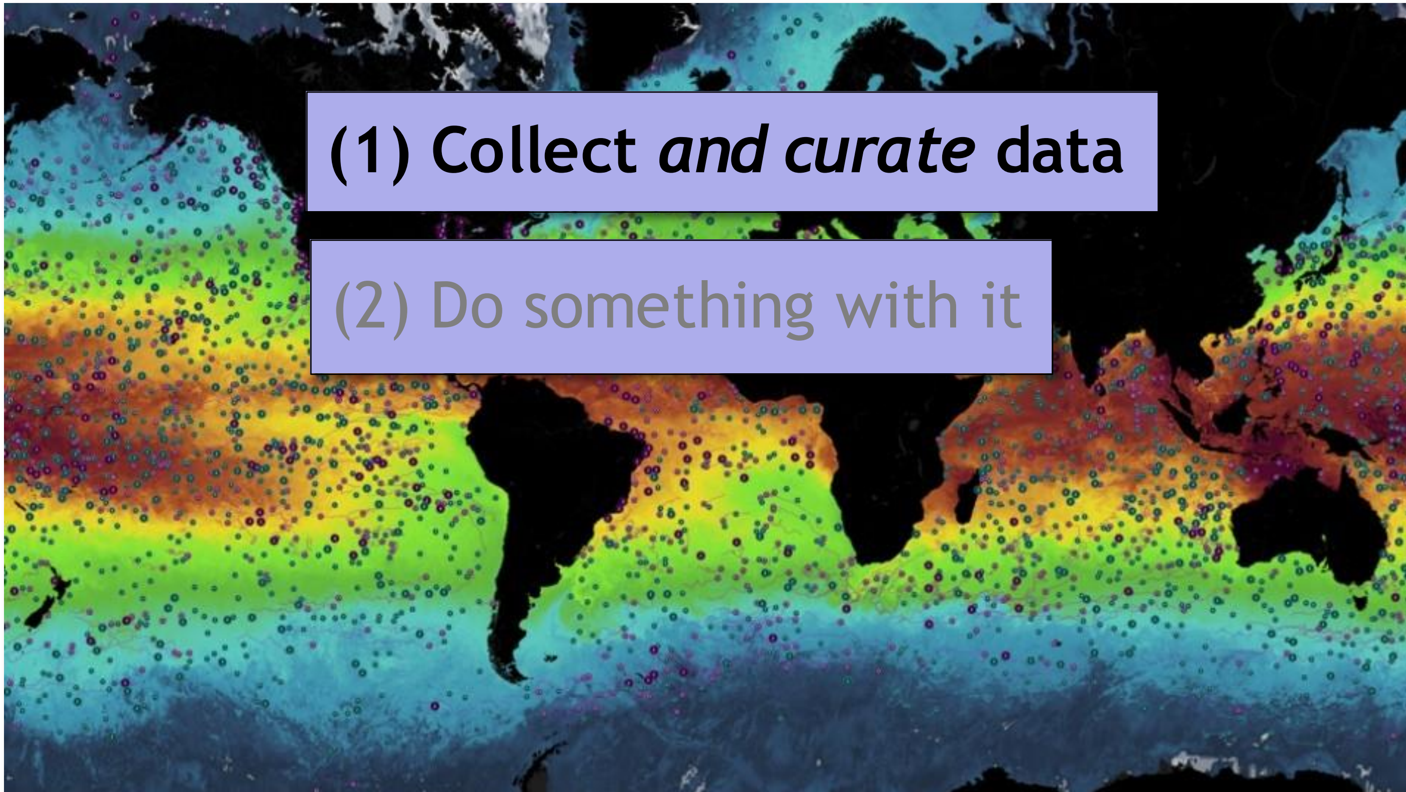
(2) Do something with it

How Big Data is Changing the World of Football

How big data gave the German football team a leg up

Saheli Roy Choudhury | @sahelirc
Thursday, 7 Jul 2016 | 12:39 AM ET
CNBC

Ocean Health



44,000 sensors, over 2 billion measurements

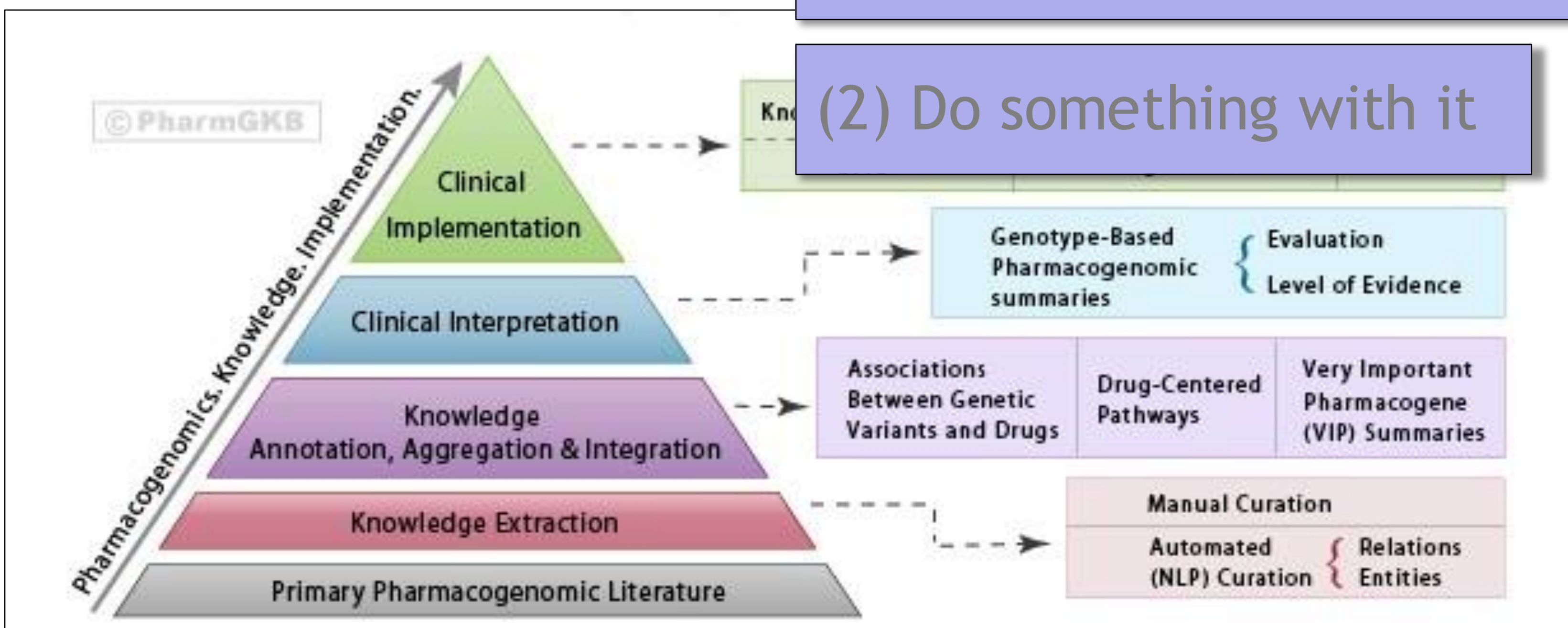
Physical, chemical, biological …

Genetics-Medicine Relationships

PharmGKB collects, curates, and disseminates knowledge about how human genetics affects response to medicines

(1) Collect *and curate* data

(2) Do something with it



And Many More

- Weather prediction
- Medical diagnosis
- Financial markets
- Resource management
- Computational social science
- Smart buildings and cities
-

The list goes on and on,
and it's still early days

Data Tools and Techniques

- Basic Data Manipulation and Analysis
  Performing well-defined computations or asking well-defined questions (“queries”)
- Data Mining
  Looking for patterns in data
- Machine Learning
  Using data to build models and make predictions
- Data Visualization
  Graphical depiction of data
- Data Collection and Preparation

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions ("queries")

- Average January low temperature for each country over last 20 years
- Number of items over $100 bought by females between ages 20 and 30
- Frequency of specific medicine relieving specific symptoms
- The ten stocks whose price varied the most over the past year

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions (“queries”)

- Aver cour
- Num fema
- Freq specific symptoms
- The ten stocks whose price varied the most over the past year

- Spreadsheets
- Relational (SQL) database systems
- “NoSQL” / scalable systems
- Programming languages with data support (e.g., Python, R)

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People who like movie X also like movie Y
- Patients who respond well to medicines X and Y also respond well to medicine Z
- Students going to the same university are frequently online friends
- Wealthier people are moving from cities to suburbs

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People … ovie Y
- Patien … nes X and Y … Z
- Studen … y are frequently online friends
- Wealthier people are moving from cities to suburbs

> - Frequent item-sets
> - Association rules
> - Specialized techniques for graphs, text, multimedia

Machine Learning

Using data to build models and make predictions

- Customers who are women over age 20 are likely to respond to an advertisement
- Students with good grades are predicted to do well on entrance exams
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

Machine Learning

Using data to build models and make predictions

- Customers ... over age 20 are likely to res... rtisement
- Students wi... re predicted to do well on ent...
- Roughly: Basic data analysis and data mining give answers from the available data, while machine learning uses the available data to make predictions about missing or future data

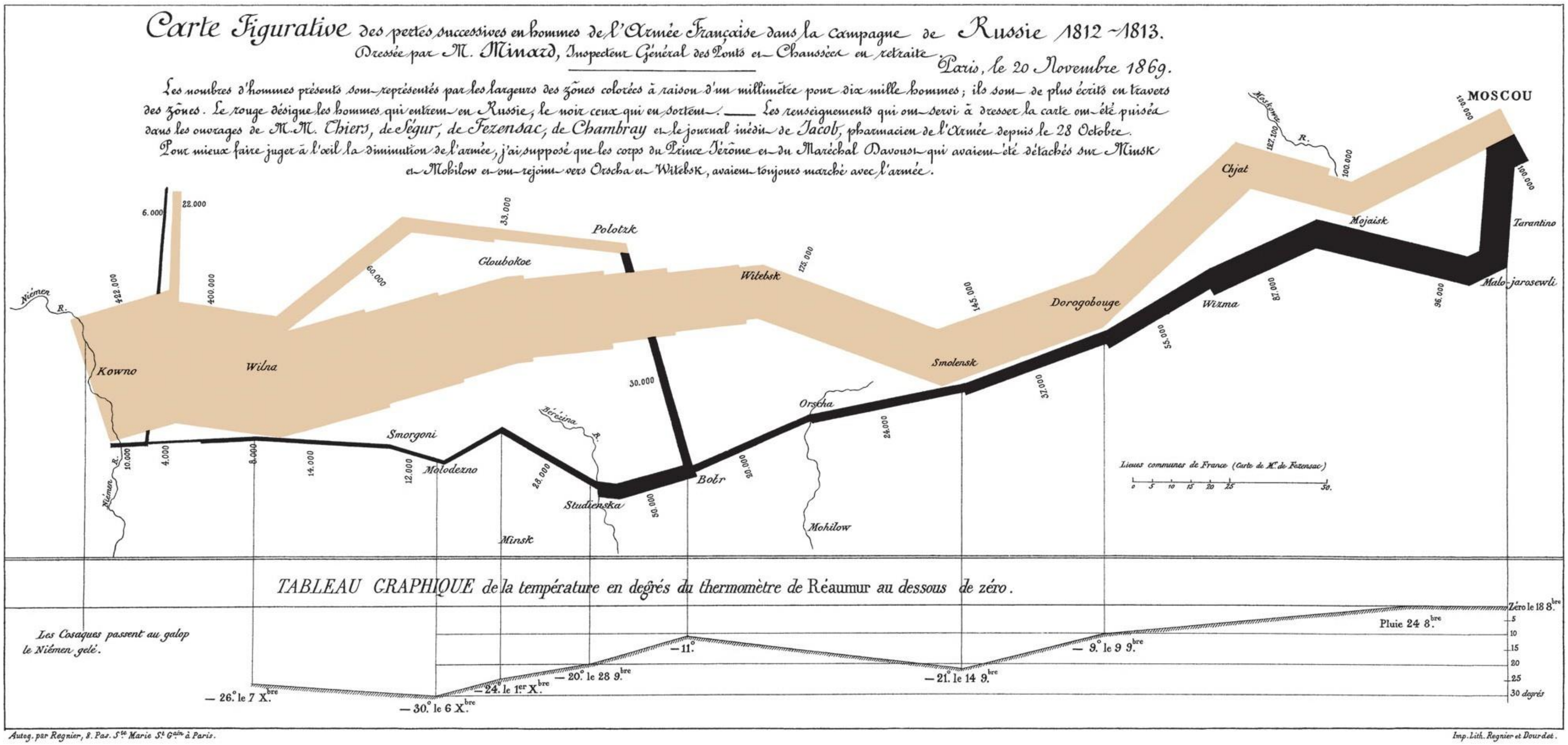
- Regression
- Classification
- Clustering

Data Visualization

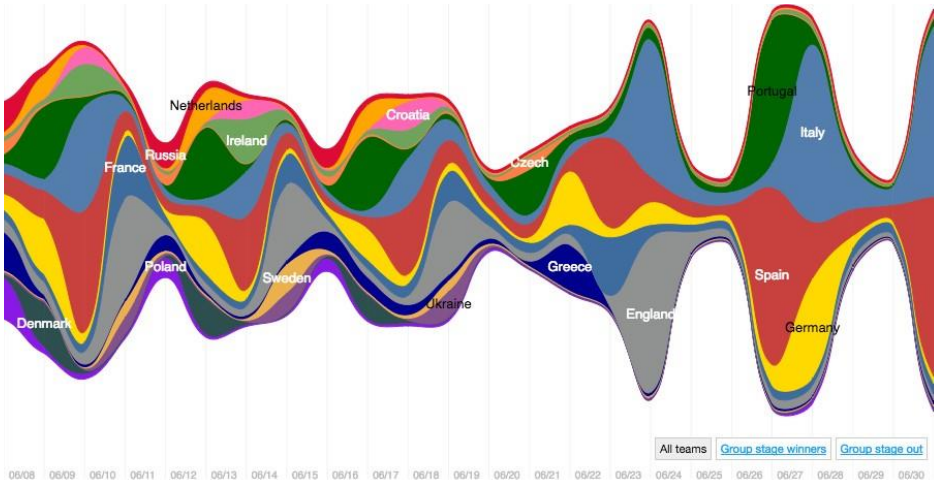
"A picture is worth a thousand words"

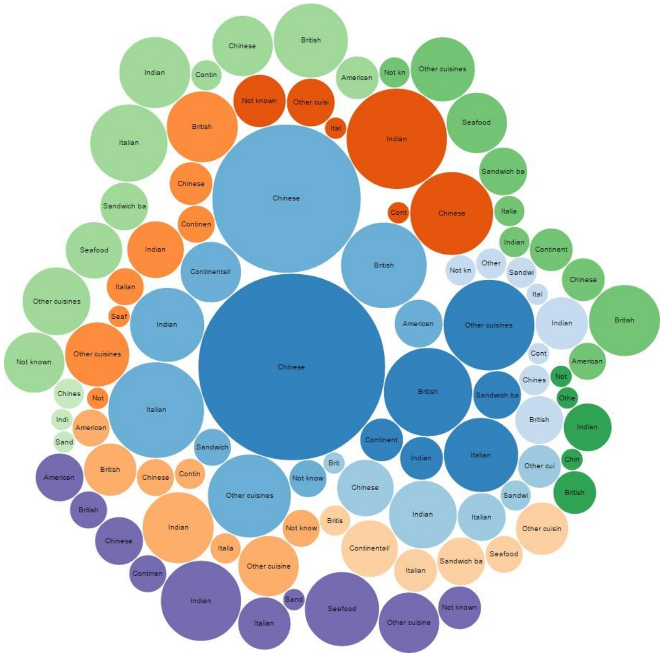
Early Data Visualization

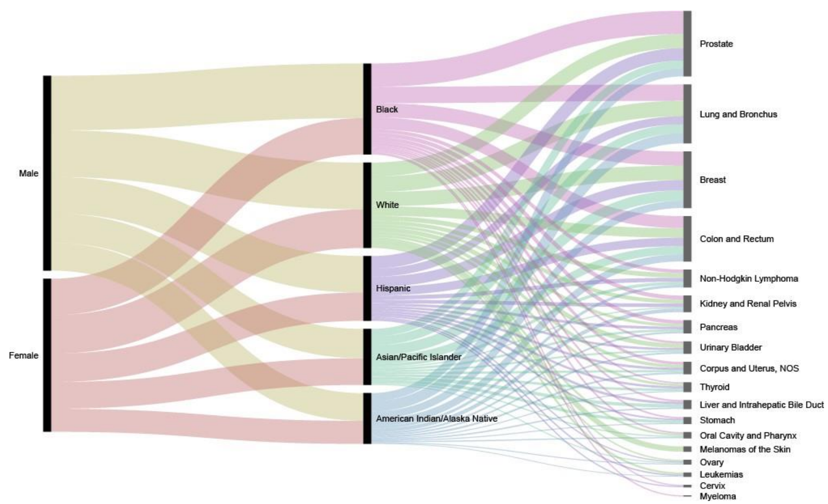
Napoleon's Army



Fancy Data Visualization



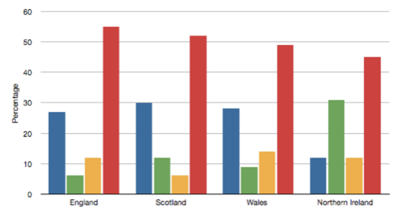


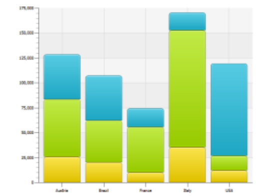


Basic Data Visualization

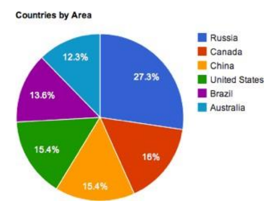
Don’t underestimate the power of basic visualizations

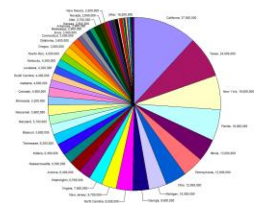
- Bar charts



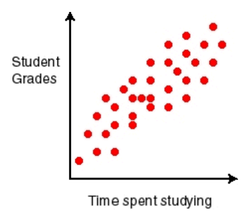


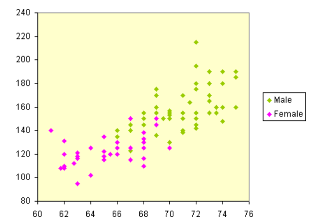
- Pie charts



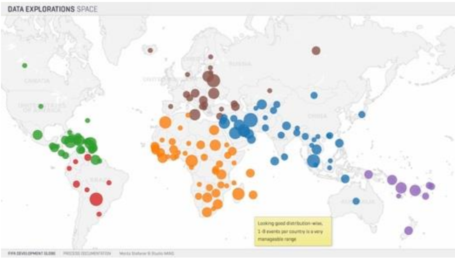


- Scatterplots

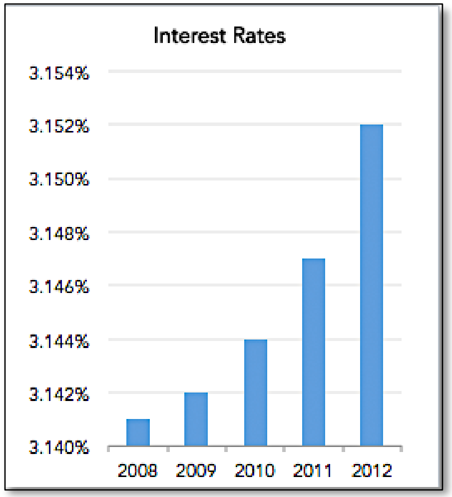


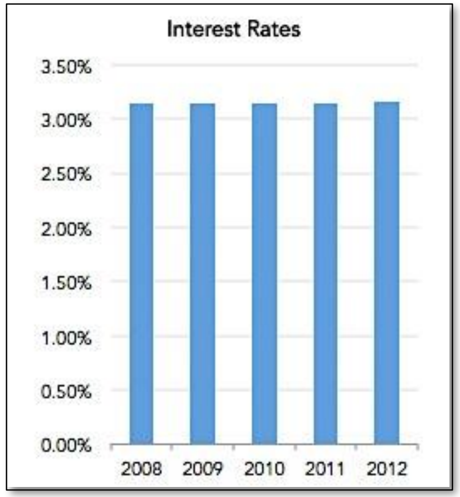


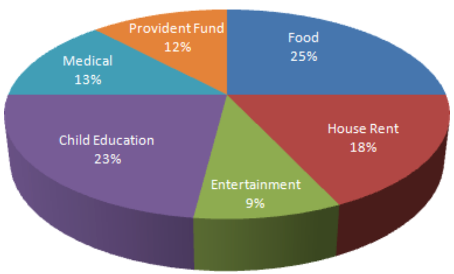
- Maps

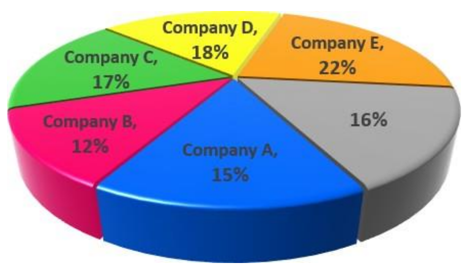


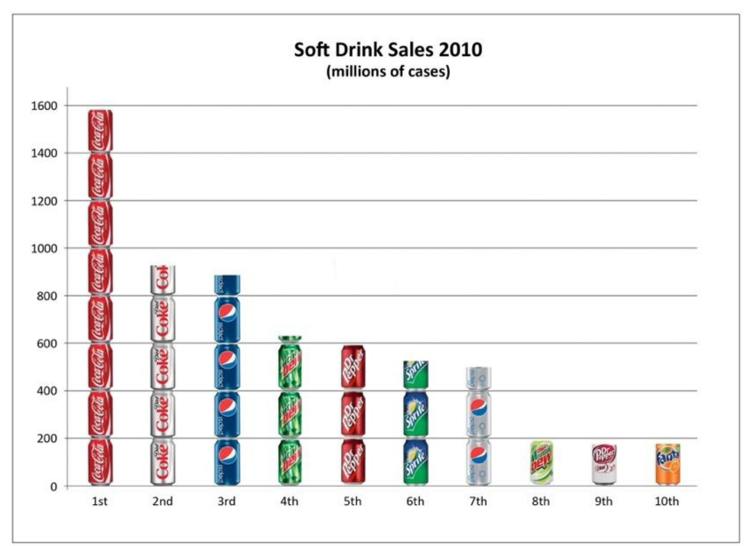
Misleading Data Visualization











Data Collection and Preparation

The "dirty" secret of working with data

- Extracting data from difficult sources
- Filling in missing values
- Removing suspicious data
- Making formats, encoding, and units consistent
- De-duplicating and matching

Data preparation often consumes 80% or more of the effort in a data-driven project

Pitfalls of Data Science

(1) Collect data

(2) Do something with it

correct

SAPIENZA
UNIVERSITÀ DI ROMA

Correlation and Causation

Data analysis, data mining, and machine learning can reveal relationships between data values

Correlation – Values track each other

- Height and Shoe Size
- Grades and Entrance Exam Scores

Causation – One value directly influences another

- Education Level → Starting Salary
- Temperature → Cold Drink Sales

Correlation and Causation

“Correlation does not imply causation”

Correlation – Values track each other

- Height and Shoe Size
- Grades and Entrance Exam Scores

Causation – One value directly influences another

- Education Level → Starting Salary
- Temperature → Cold Drink Sales

Correlation and Causation

"Correlation does not imply causation"

Correlation –
- Height anc
- Grades anc

Causation – C
- Education
- Temperatu

Correlation and Causation

> “Correlation does not imply causation”

- Correlation can be result of causation from a hidden “confounding variable”
- A and B are correlated because there’s a hidden C such that C → A and C → B
  - ❖Homeless population and crime rate
    Confounding variable: unemployment
  - ❖Forgetfulness and poor eyesight
    Confounding variable: age
  - ❖ Height and shoe size
  - ❖ Grades and entrance exam scores

Correlation and Causation

"Correlation does not imply causation"

- Correlation can be result of causation from a hidden "confounding variable"
- A and B are correlated because there's a hidden C such that C → A and C → B

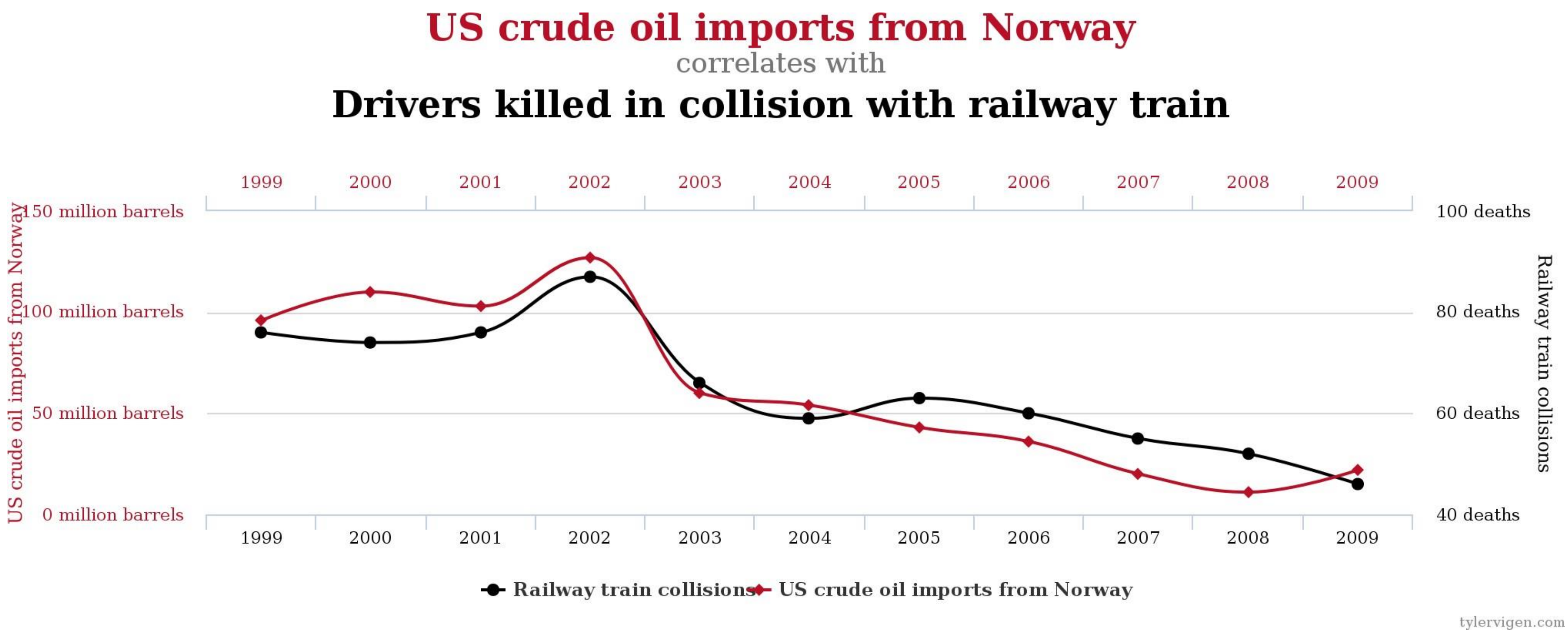
- Correlation is usually "easy" to test
- Causation is typically impossible to test

Correlation and Causation

"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

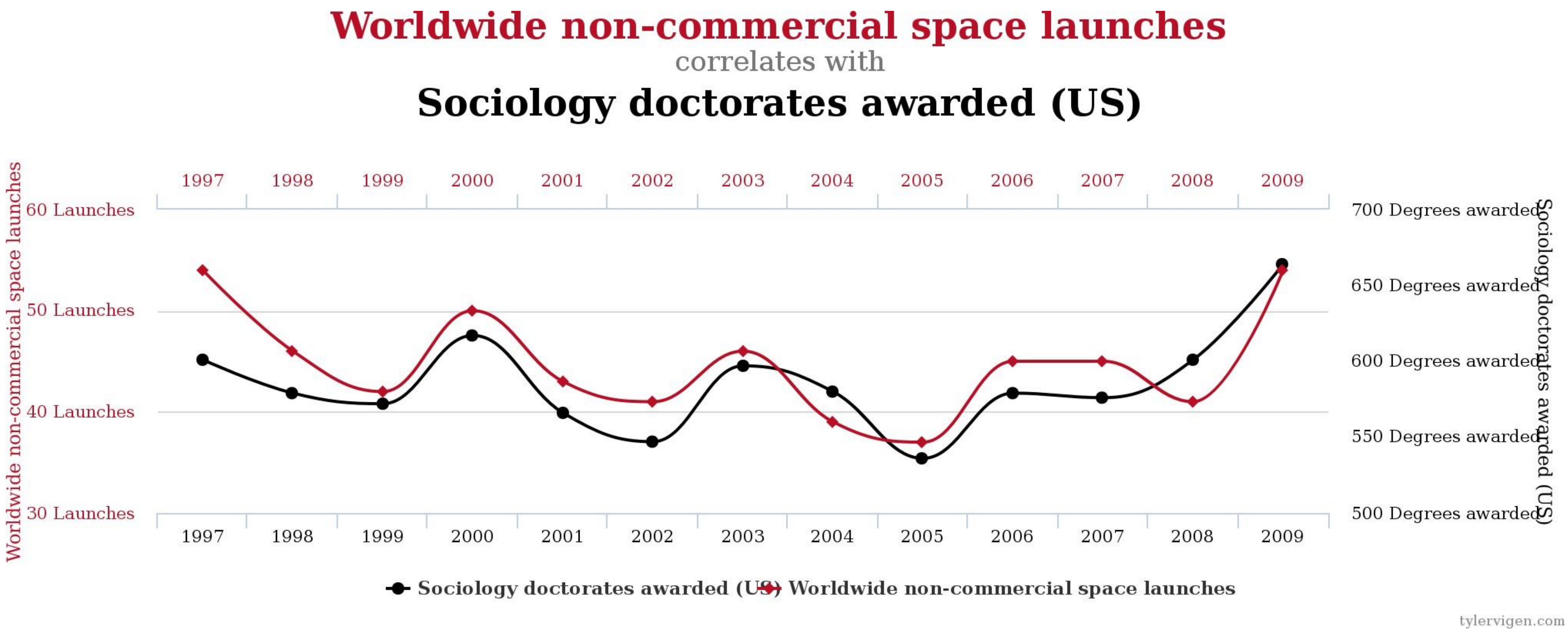
Excellent health statistics - smokers are less likely to die of age related illnesses.'

Surprising Correlation #1

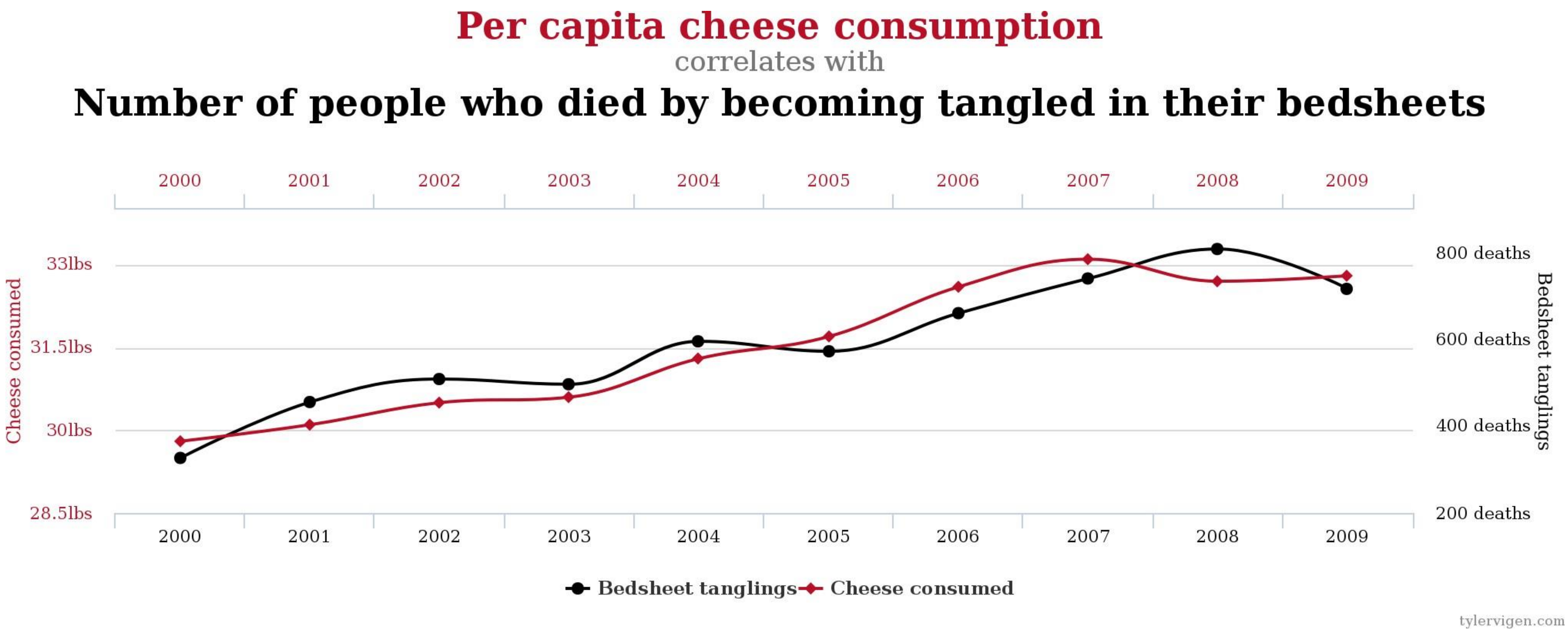


SAPIENZA
UNIVERSITÀ DI ROMA

Surprising Correlation #2



Surprising Correlation #3



“Spurious Correlations” Website

http://www.tylervigen.com/

Underfitting and Overfitting

Machine learning uses data to create a "model" and uses model to make predictions

- Customers who are women over age 20 are likely to respond to an advertisement
- Students with good grades are predicted to do well on entrance exams

Underfitting

Model used for predictions is **too simplistic**

- 60% of men and 70% of women responded to an advertisement, therefore all future ads should go to women
- If a furniture item has four legs and a flat top it is a dining room table

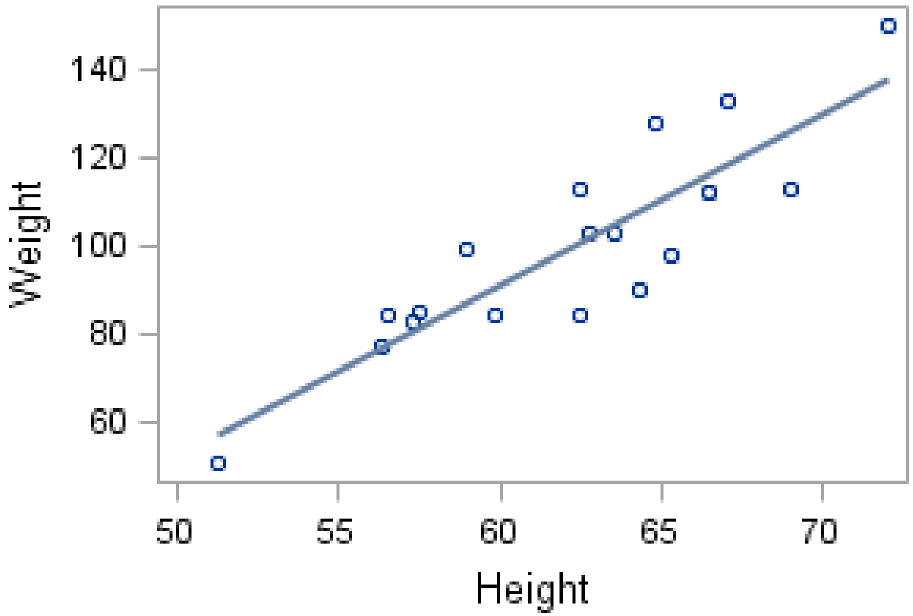
Overfitting

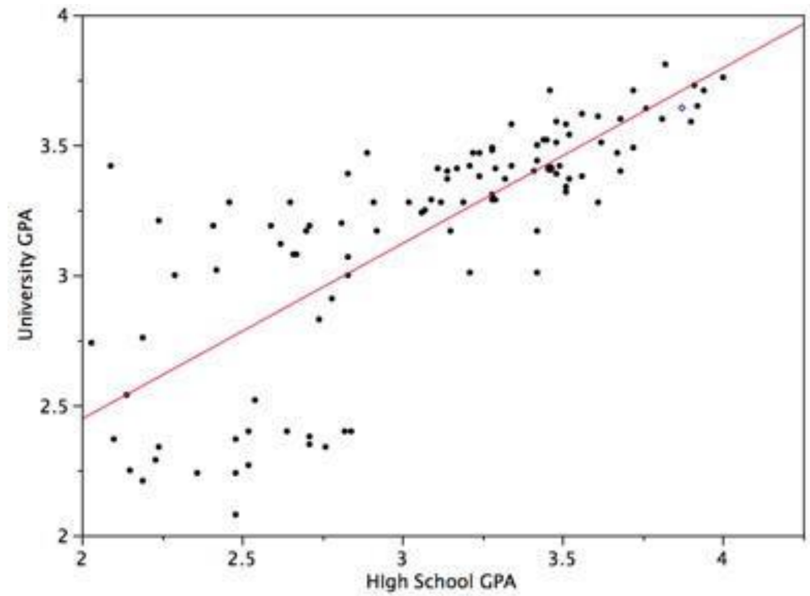
Model used for predictions is **too specific**

- The best targets for an advertisement are married women between 25 and 27 years with short black hair, one child, and one pet dog
- If a furniture item has four 100 cm legs with decoration and a flat polished wooden top with rounded edges then it is a dining room table

Regression

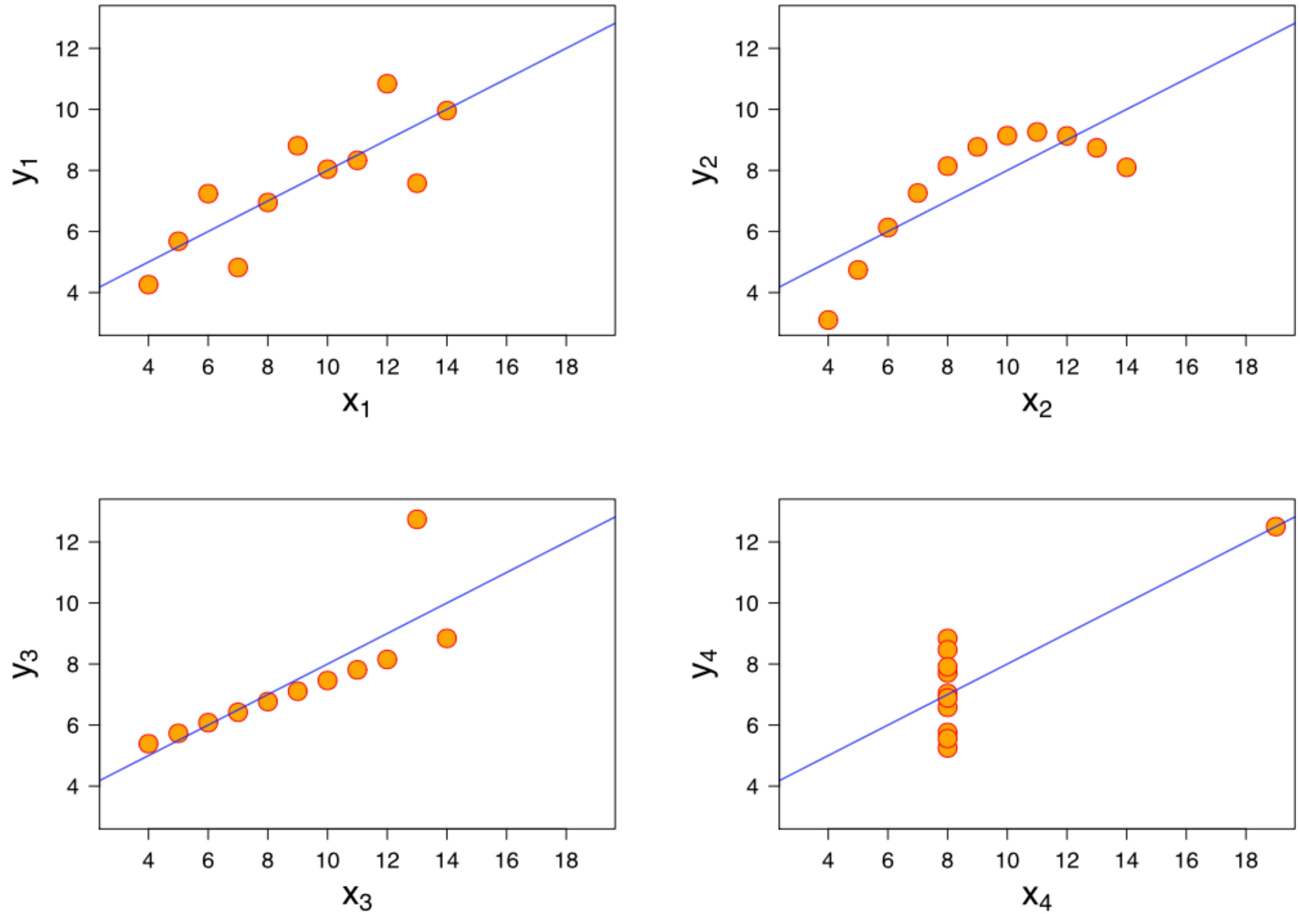
- Fit a line or curve to a set of points (model)
- Use model to predict values for new points





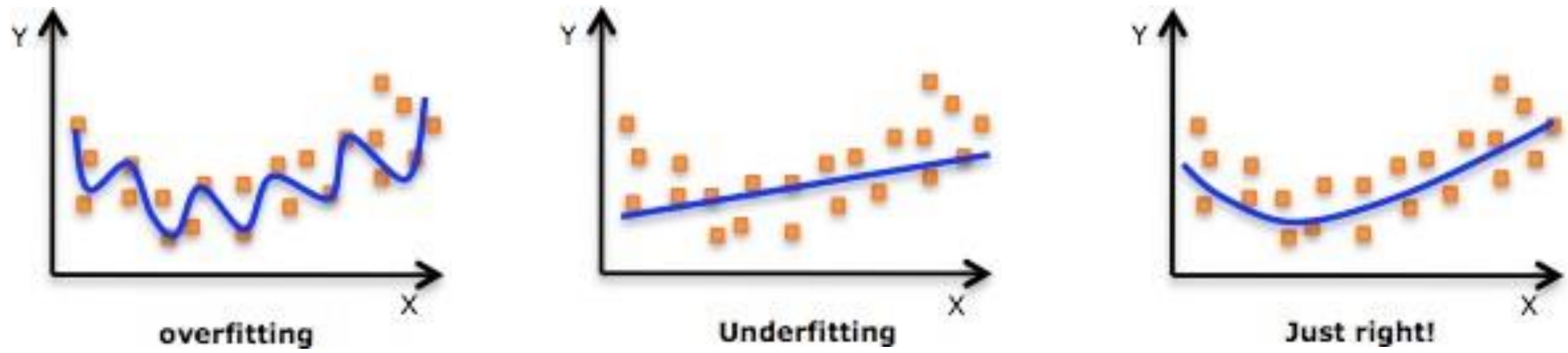
Underfitting

Model is **too simplistic**



Overfitting

Model is **too specific**



Soccer Match Prediction Scam

- Friday: receive email from "Psychic Sally" predicting which teams will be the winners in the weekend's five soccer matches. She's right about all of them!
- Same thing the following weekend: five games, all winners predicted correctly
- And the following one: five more correct
- Fourth Friday: Sally offers to give you her predictions for the coming weekend's games, for a fee

Should you do it?

Soccer Match Prediction Scam

How many contacts must Sally start with on week one to ensure she has 100 potential buyers by week four, i.e., 100 people who received 15 correct predicted winners?

(Assume no draws)

Data Privacy

Of significant concern in some sectors

- Individual data collected covertly
  - Edward Snowden, “metadata” argument
- Individual data collected legally but used questionably
  - Individual “information trails” are enormous
  - Target stores pregnancy mailing
- Individual data deduced from “anonymous” public data
  - Governor of Massachusetts health record

Languages, Systems, Platforms

- Spreadsheets

  Surprisingly versatile and powerful for data analysis tasks, provided data is not *too* large

- Programming languages with data support
  - R Language – powerful statistical features
  - Python – general-purpose language with R-like add-ons (Pandas, SciPy, scikit-learn)

Languages, Systems, Platforms

- Relational Database Management Systems
  - Also called RDBMS, SQL Systems
  - Long-standing solution for reliability, efficiency, powerful query processing
  - Works for all but truly extreme data sizes, or highly unstructured data
- "NoSQL" Systems
  - Distributed/scalable processing
  - Some specifically target unstructured data (documents, graphs)

Languages, Systems, Platforms

- Specialized languages on scalable systems
  - MapReduce / Hadoop
  - Spark generalized data flow
- Systems for data preparation
- Systems for data visualization

Languages, Systems, Platforms

- Data processing in the cloud
  - Amazon Web Services, Google Cloud, Microsoft Azure
  - Data storage
  - Data processing: SQL, Hadoop, Spark
  - Machine learning libraries
  - Integration with visualization systems

How Much Data is There?

Complete works of William Shakespeare
- 5 megabytes

Average individual
- 50 gigabytes (10,000 Shakespeares)

USA Library of Congress
- 10 terabytes (2 million Shakespeares)

Uploaded to Facebook daily
- 1 petabyte (200 million Shakespeares)

Produced by humanity daily
- 2.5 exabytes (500 trillion Shakespeares)

“Big Data”

Some domains produce vast quantities of data, and some analyses require “big data” to be effective

- Most tools and techniques apply to data of all sizes
- Big insights can come from small/medium data

> Sometimes twenty Spark servers in the cloud are required.
> More often a laptop with SQL, Python, or simple spreadsheets does the job.

Some Key Principles

- *use many data sources* (the plural of anecdote is not data)
- *understand how the data were collected* (sampling is essential)
- *weight the data thoughtfully* (not all polls are equally good)
- *use statistical models* (not just hacking around in Excel)
- *understand correlations* (e.g., states that trend similarly)
- *think like a Bayesian, check like a frequentist* (reconciliation)
- *have good communication skills* (What does a 60% probability even mean? How can we visualize, validate,and understand the conclusions?)

Some Challenges

- *massive data* (500k users, 20k movies, 100m ratings)
- *curse of dimensionality* (very high-dimensional problem)
- *missing data* (99% of data missing; *not* missing at random)
- *extremely complicated set of factors that affect people's ratings of movies* (actors, directors, genre, ...)
- *need to avoid overfitting* (test data vs. training data)

What is Machine Learning

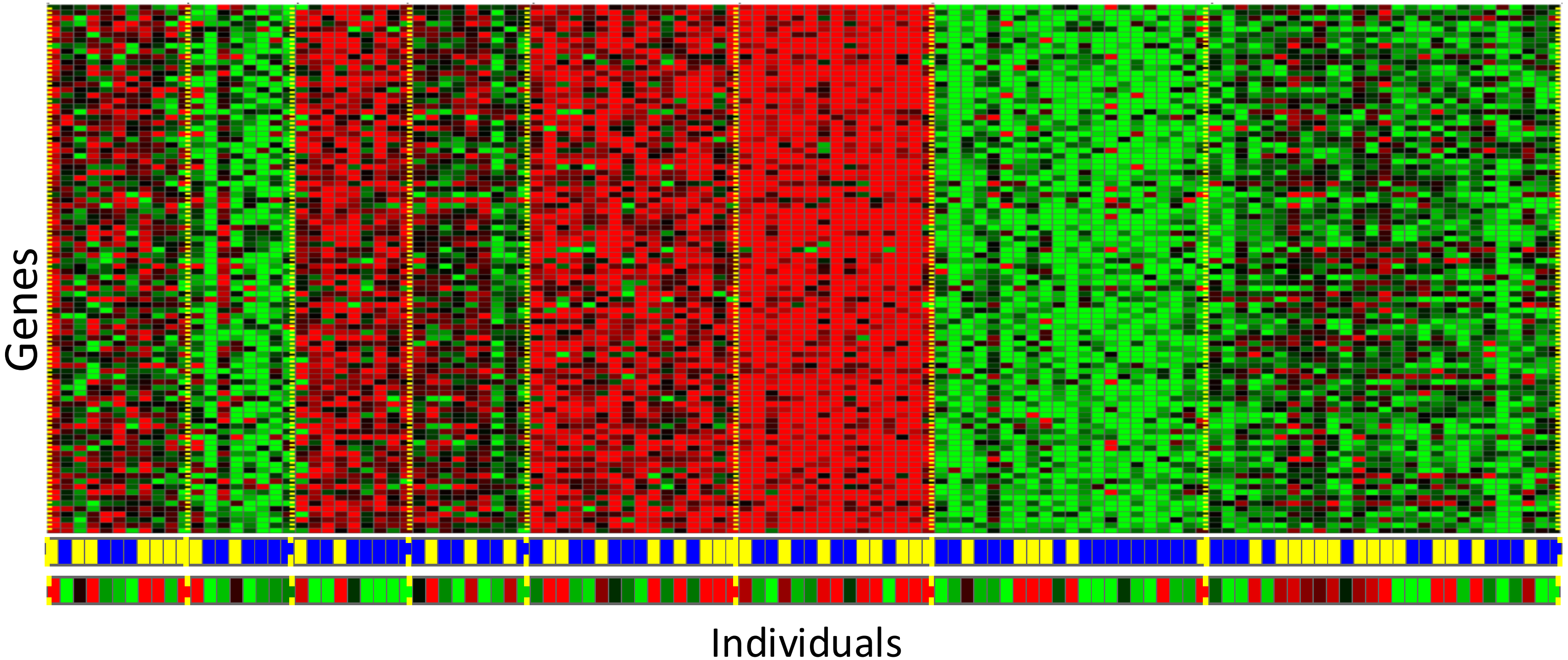
Machine Learning Applications

- Cluster news on the web for news.google.com

SAPIENZA
UNIVERSITÀ DI ROMA

Machine Learning Applications

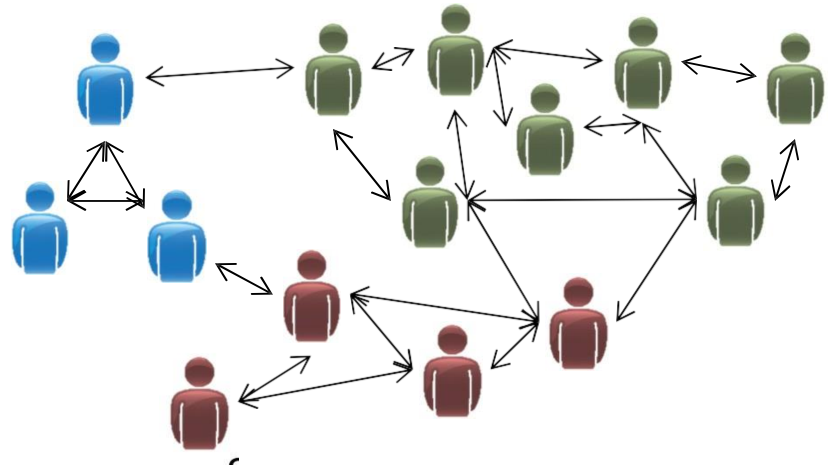
- Group individuals according to their genes



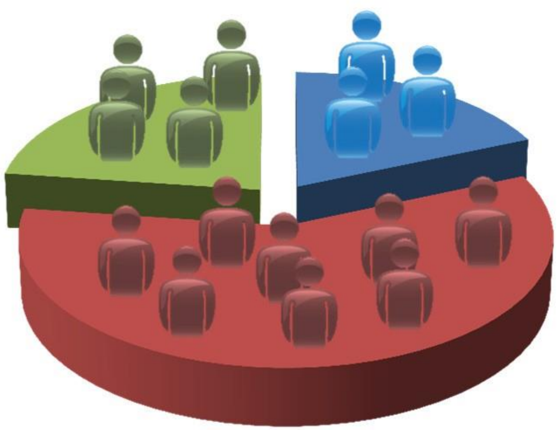
[Source: Daphne Koller]

Machine Learning Applications

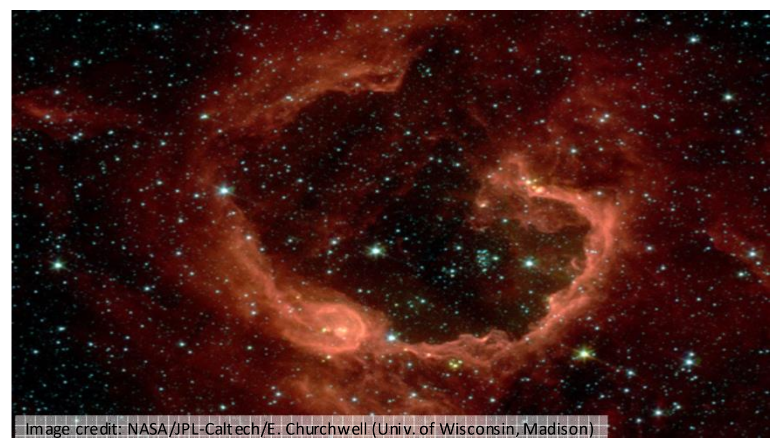
Organize computing clusters



Social network analysis



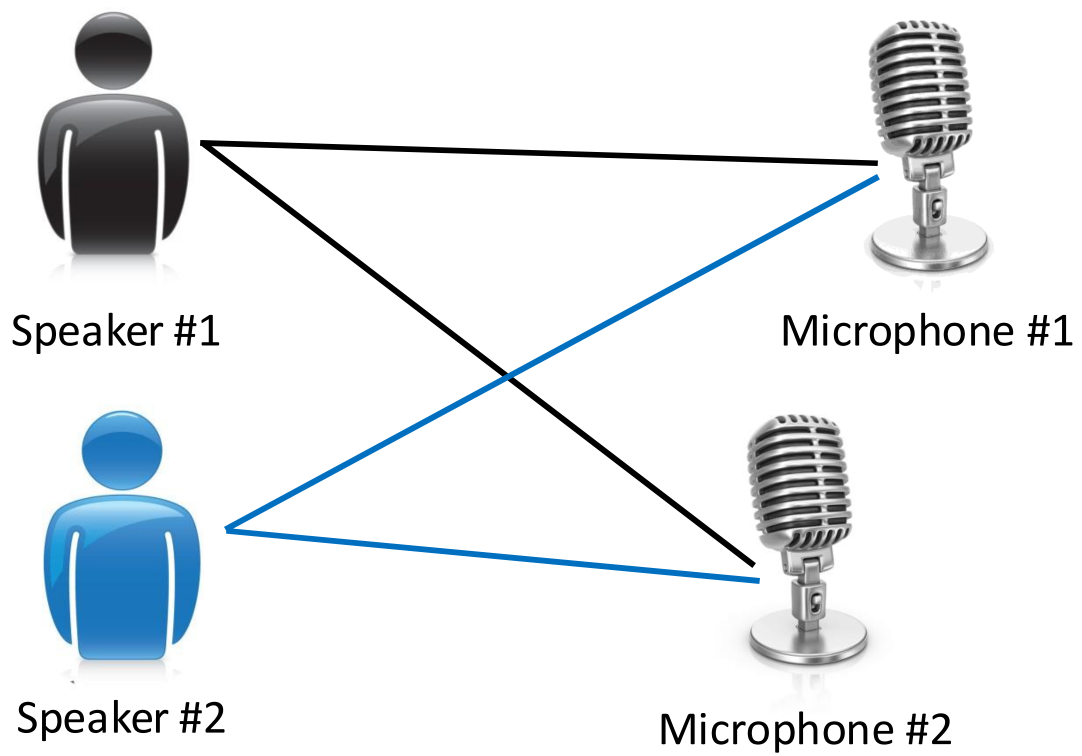
Market segmentation



Astronomical data analysis

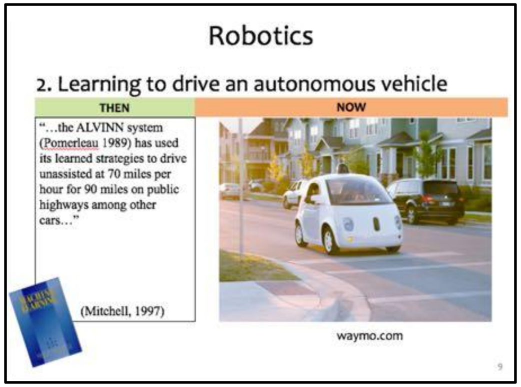
Machine Learning Applications

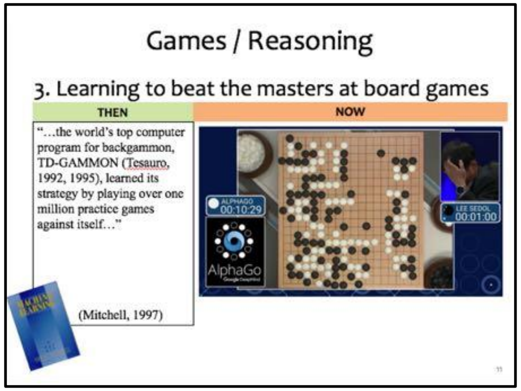
- Cocktail party problem

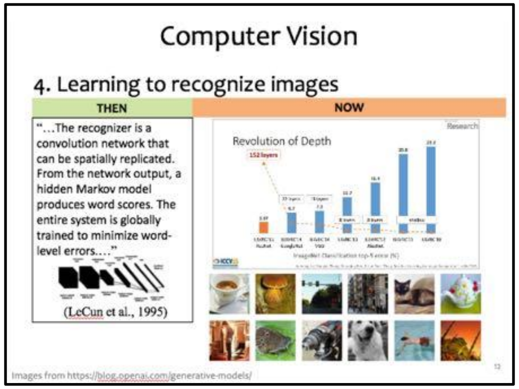


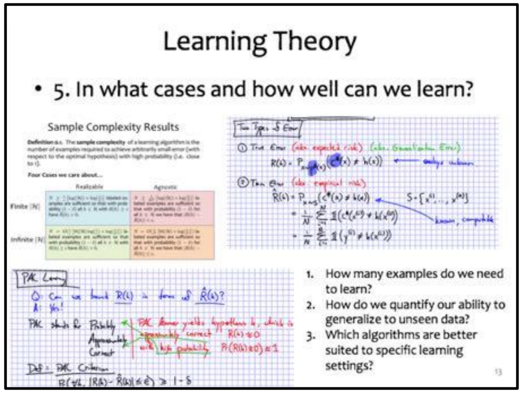
What is ML?


Speech Recognition
1. Learning to recognize spoken words
THEN
NOW
"…the SPHINX system (e.g. Lee 1989) learns speaker-specific strategies for recognizing the primitive sounds (phonemes) and words from the observed speech signal…neural network methods…hidden Markov models…"
(Mitchell, 1997)
Source: https://www.stonetemple.com/great-knowledge-box-showdown/#VoiceStudyResults


Robotics
2. Learning to drive an autonomous vehicle
THEN
NOW
"…the ALVINN system (Pomerleau 1989) has used its learned strategies to drive unassisted at 70 miles per hour for 90 miles on public highways among other cars…"
(Mitchell, 1997)
waymo.com

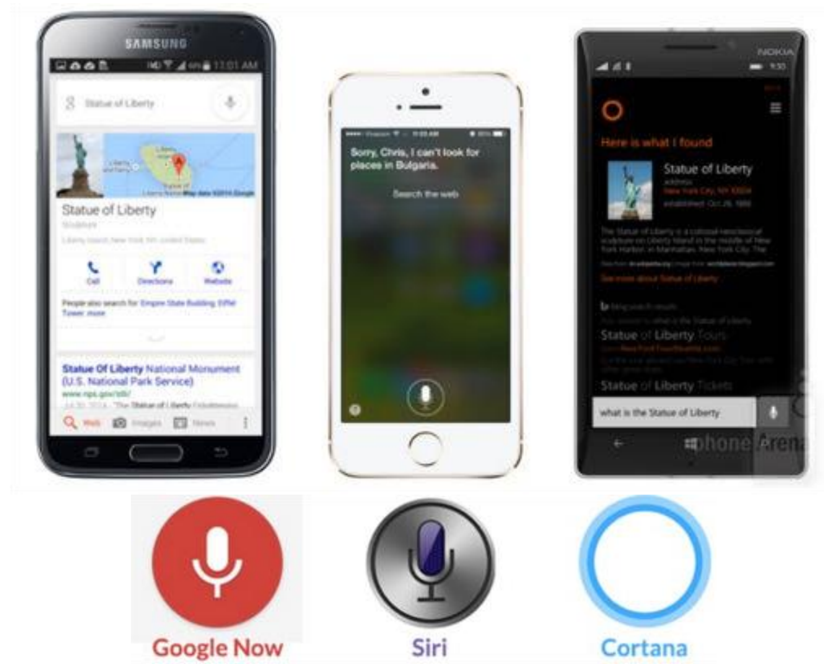

Games / Reasoning
3. Learning to beat the masters at board games
THEN
NOW
"…the world's top computer program for backgammon, TD-GAMMON (Tesauro, 1992, 1995), learned its strategy by playing over one million practice games against itself…"
(Mitchell, 1997)
AlphaGo


Computer Vision
4. Learning to recognize images
THEN
NOW
"…The recognizer is a convolution network that can be spatially replicated. From the network output, a hidden Markov model produces word scores. The entire system is globally trained to minimize word-level errors.…"
(LeCun et al., 1995)
Revolution of Depth
152 layers
Images from https://blog.openai.com/generative-models/


Learning Theory
• 5. In what cases and how well can we learn?
Sample Complexity Results
1. How many examples do we need to learn?
2. How do we quantify our ability to generalize to unseen data?
3. Which algorithms are better suited to specific learning settings?

Speech Recognition

1. Learning to recognize spoken words

THEN	NOW
"…the SPHINX system (e.g. Lee 1989) learns speaker-specific strategies for recognizing the primitive sounds (phonemes) and words from the observed speech signal…neural network methods…hidden Markov models…" (Mitchell, 1997)	

Source: https://www.stonetemple.com/great-knowledge-box-showdown/#VoiceStudyResults

Robotics

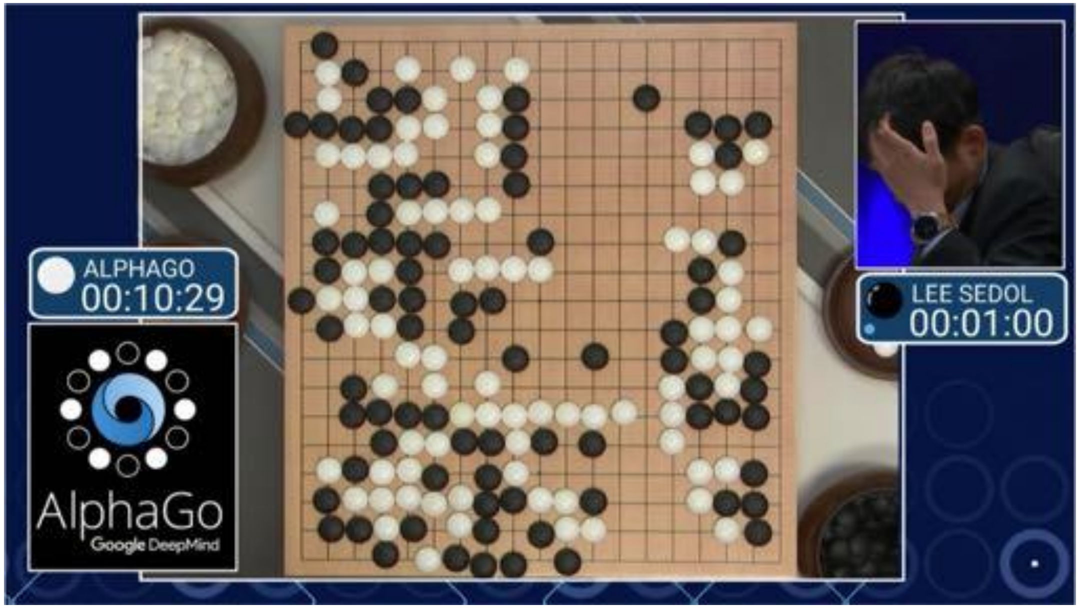
2. Learning to drive an autonomous vehicle

THEN	NOW
"…the ALVINN system (Pomerleau 1989) has used its learned strategies to drive unassisted at 70 miles per hour for 90 miles on public highways among other cars…" (Mitchell, 1997)	 https://www.geek.com/wp-content/uploads/2016/03/uber.jpg

SAPIENZA
UNIVERSITÀ DI ROMA

Games / Reasoning

3. Learning to beat the masters at board games

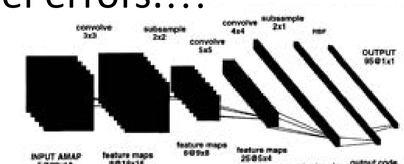
THEN	NOW
"…the world's top computer program for backgammon, TD-GAMMON (Tesauro, 1992, 1995), learned its strategy by playing over one million practice games against itself…" (Mitchell, 1997)	

Computer Vision

4. Learning to recognize images

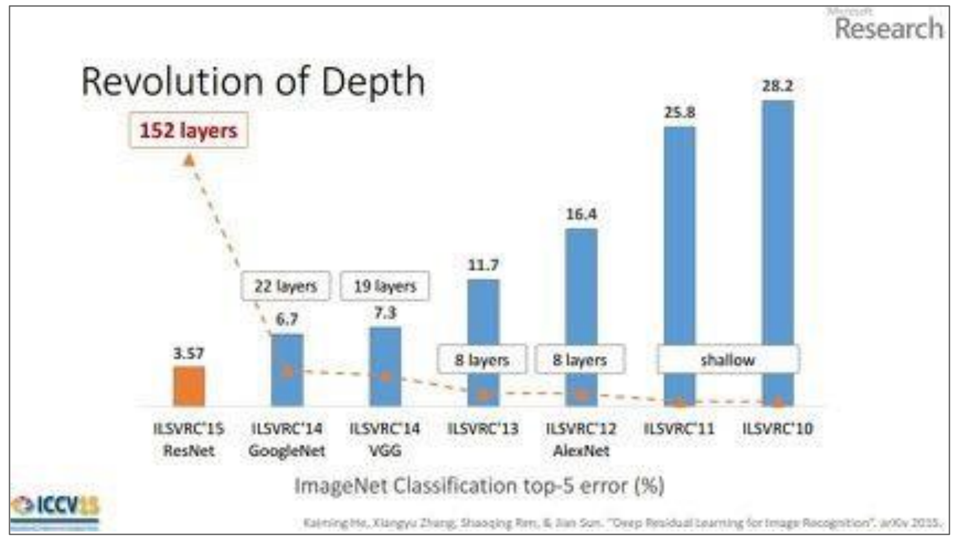
THEN

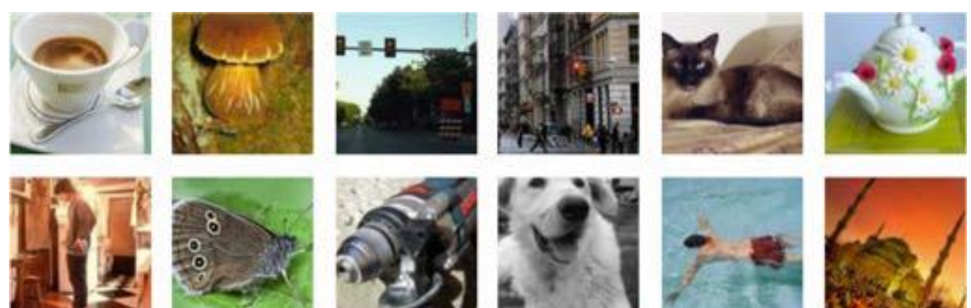
"…The recognizer is a convolution network that can be spatially replicated. From the network output, a hidden Markov model produces word scores. The entire system is globally trained to minimize word-level errors.…"



(LeCun et al., 1995)

NOW





Images from https://blog.openai.com/generative-models/

Learning Theory

- 5. In what cases and how well can we learn?

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

	Realizable	Agnostic
Finite $\|\mathcal{H}\|$	$N \geq \frac{1}{\epsilon}\left[\log(\|\mathcal{H}\|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	$N \geq \frac{1}{2\epsilon^2}\left[\log(\|\mathcal{H}\|) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| < \epsilon$.
Infinite $\|\mathcal{H}\|$	$N = O(\frac{1}{\epsilon}\left[\text{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	$N = O(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.

Two Types of Error

① True Error (aka. expected risk) (aka. Generalization Error)

$$R(h) = P_{x \sim p^*(x)}\left(c^*(x) \neq h(x)\right)$$

always unknown.

② Train Error (aka. empirical risk)

$$\hat{R}(h) = P_{x \sim S}\left(c^*(x) \neq h(x)\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(c^*(x^{(i)}) \neq h(x^{(i)})\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(y^{(i)} \neq h(x^{(i)})\right)$$

$S = \{x^{(1)}, \ldots, x^{(N)}\}$

known, computable

PAC Learning

Q: Can we bound $R(h)$ in terms of $\hat{R}(h)$?
A: Yes!

PAC stands for Probably Approximately Correct

PAC learner yields hypothesis h, which is approximately correct $R(h) \approx 0$ with high probability $Pr(R(h) \approx 0) \approx 1$

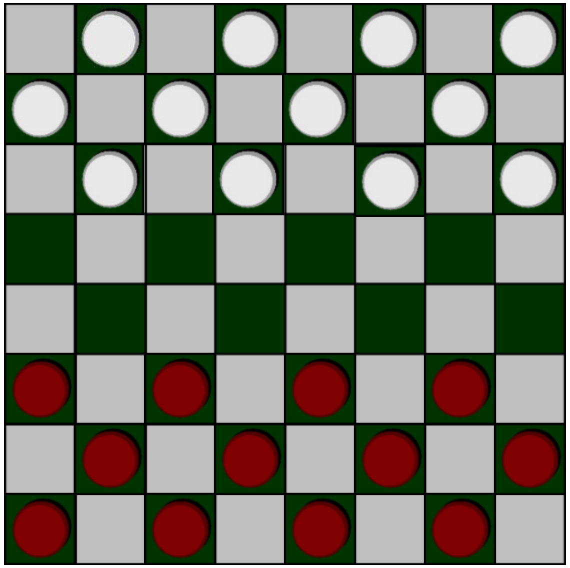
Def: PAC Criterion

$$Pr\left(\forall h, |R(h) - \hat{R}(h)| \leq \epsilon\right) \geq 1 - \delta$$

1. How many examples do we need to learn?
2. How do we quantify our ability to generalize to unseen data?
3. Which algorithms are better suited to specific learning settings?

What is Machine Learning

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed



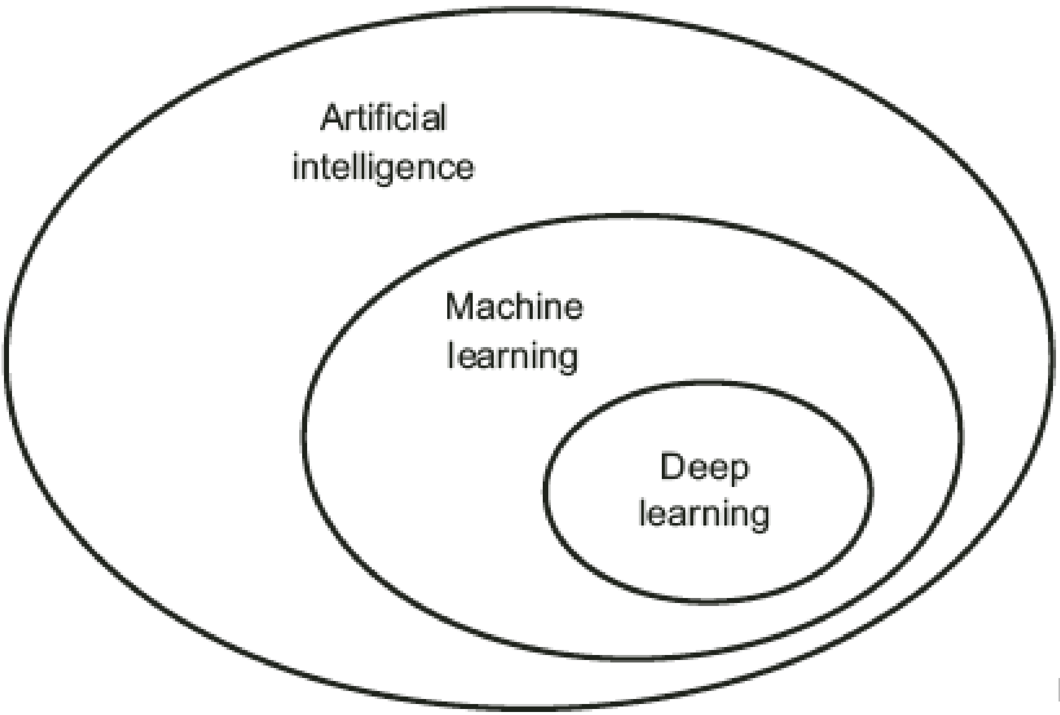
What is Machine Learning

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed

- Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E

What is Machine Learning

- Grew out of work in AI
- New capability for computers



What is Machine Learning

- Grew out of work in AI
- New capability for computers
- Examples:
  - Database mining
    - Large datasets from growth of automation/web
    - E.g. web click data, medical records, biology, engineering
  - Applications which cannot be programmed by hand
    - E.g. autonomous driving, handwriting recognition, most of Natural Language Processing (NLP), Computer Vision
  - Self-customizing programs
    - E.g. Amazon, Netflix product recommendations
  - Understanding human learning (brain, real AI)

What is Machine Learning

- "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

- Suppose your email program watches which emails you do or do not mark as spam and, based on that, learns how to better filter spam. What is the task T in this setting?
  - Classifying emails as spam or not spam.
  - Watching you label emails as spam or not spam.
  - The number (or fraction) of emails correctly classified as spam/not spam.
  - None of the above—this is not a machine learning problem.

Capturing the Knowledge of Experts

Solution #1: Expert Systems

- Over 20 years ago, we had rule based systems
- Ask the expert to
  1. Obtain a PhD in Linguistics
  2. Introspect about the structure of their native language
  3. Write down the rules they devise

Give me directions to Starbucks

```
If: "give me directions to X"
Then: directions(here, nearest(X))
```

How do I get to Starbucks?

```
If: "how do i get to X"
Then: directions(here, nearest(X))
```

Where is the nearest Starbucks?

```
If: "where is the nearest X"
Then: directions(here, nearest(X))
```

Capturing the Knowledge of Experts

Solution #2: Annotate Data and Learn

- Experts:
  - **Very good at** answering questions about specific cases
  - **Not very good at** telling **HOW** they do it
- 1990s: So why not just have them tell you what they do on **SPECIFIC CASES** and then let **MACHINE LEARNING** tell you how to come to the same decisions that they did

Capturing the Knowledge of Experts

Solution #2: Annotate Data and Learn

1. Collect raw sentences $\{x_1, \ldots, x_n\}$
2. Experts annotate their meaning $\{y_1, \ldots, y_n\}$

x_1: How do I get to Starbucks?

y_1: `directions(here, nearest(Starbucks))`

x_2: Show me the closest Starbucks

y_2: `map(nearest(Starbucks))`

x_3: Send a text to John that I'll be late

y_3: `txtmsg(John, I'll be late)`

x_4: Set an alarm for seven in the morning

y_4: `setalarm(7:00AM)`

Data Science Vs Machine Learning

A Data Scientist Is...

"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."

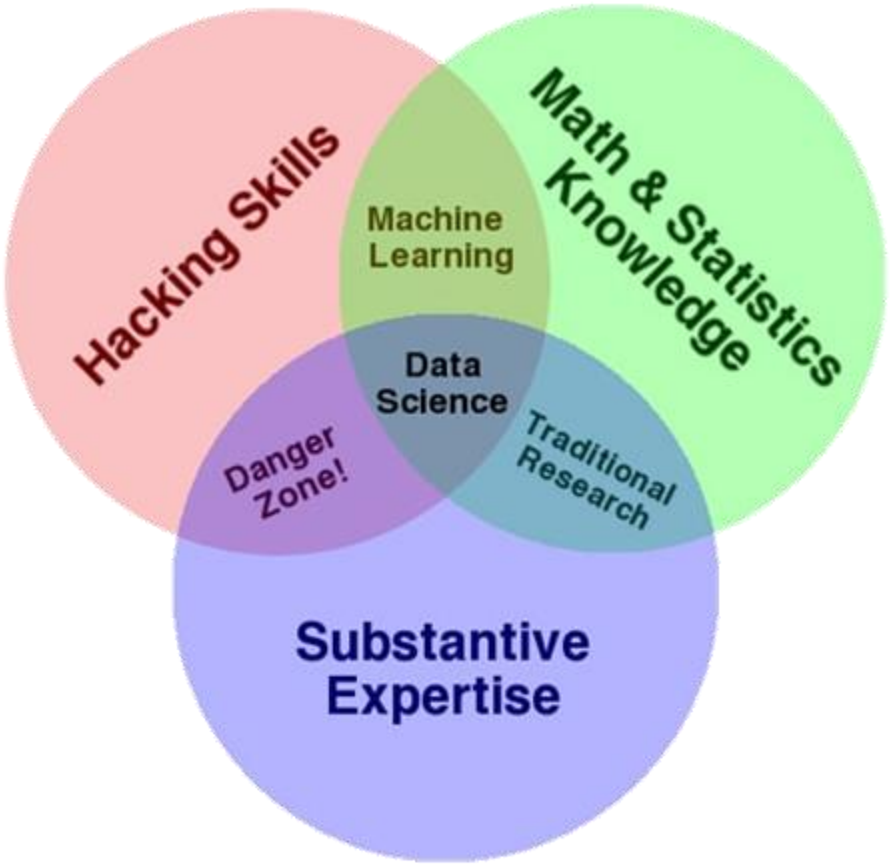
- Josh Blumenstock

"Data Scientist = statistician + programmer + coach + storyteller + artist"

- Shlomo Aragmon

Data Science Vs. Machine Learning

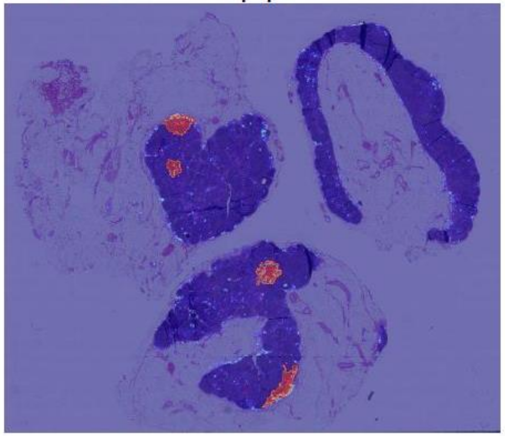
- Hugh Conway, 2010



What is Perception (Computer Vision)

Machine Learning and Perception (Computer Vision)

Medical applications



Robotics

Security

Gaming

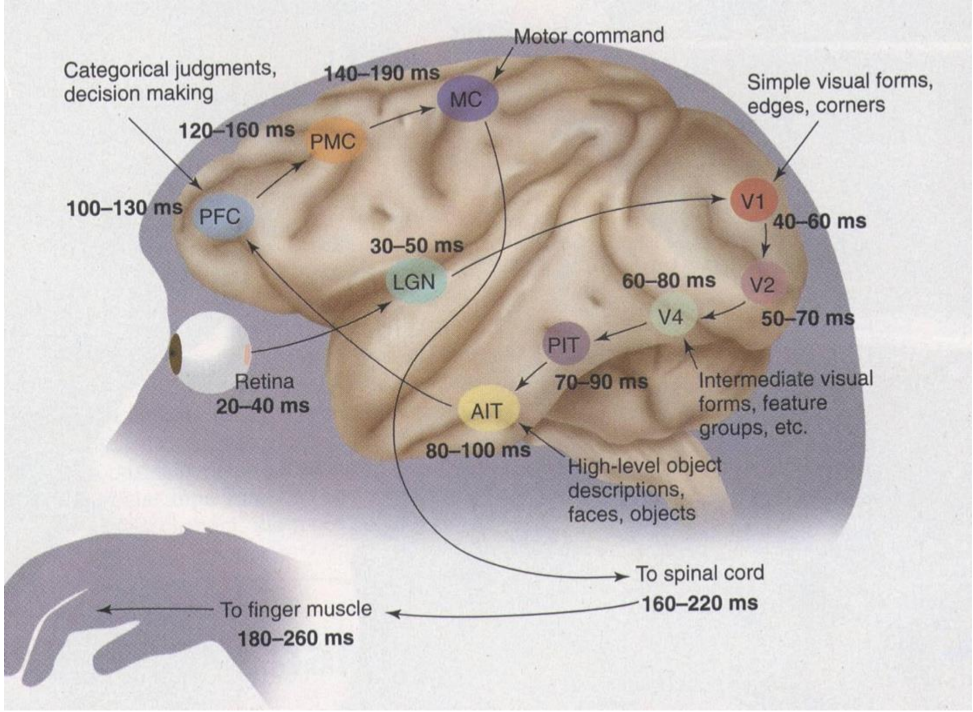
Mobile devices

Driving

Courtesy of A. Torralba, @ICVSS'18

Computer Vision

- Science
  - Foundations of perception. How do *WE* see?
  - computer vision to explore “computational model of human vision”



SAPIENZA UNIVERSITÀ DI ROMA

Computer Vision

- Science
  - Foundations of perception. How do *WE* see?
  - computer vision to explore “computational model of human vision”

- Engineering
  - How do we build systems that perceive the world
  - computer vision to solve real-world problems: cars to detect pedestrians

Computer Vision

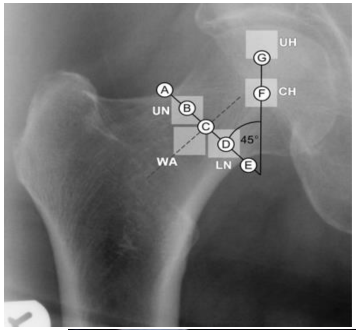
- Science
  - Foundations of perception. How do *WE* see?
  - computer vision to explore “computational model of human vision”
- Engineering
  - How do we build systems that perceive the world
  - computer vision to solve real-world problems: cars to detect pedestrians
- Applications
  - medical imaging (computer vision to support medical diagnosis, visualization)
  - surveillance (to follow/track people at the airport, train-station, ...)
  - entertainment (vision-based interfaces for games)
  - graphics (image-based rendering, vision to support realistic graphics)
  - car-industry (lane-keeping, pre-crash intervention, …)
  - …

Some Applications

- US Post office
  - At the mail processing plant, ***machines*** separate mail by shape and size, and orient them so their addresses are right-side up and facing the same direction
  - ***An optical scanner*** scans the address, and then a bar code representing the specific address is sprayed on the front of the envelope
  - If the scanner can't read the address, the letter is manually sorted

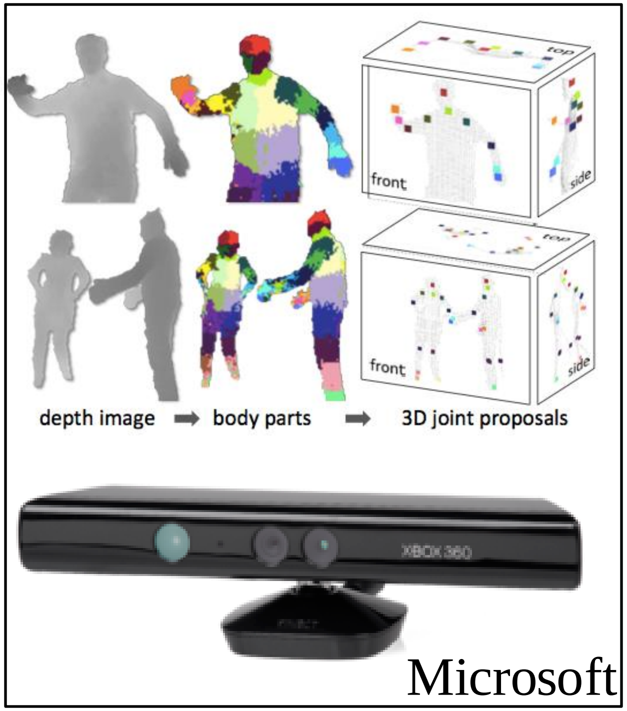
Some Applications

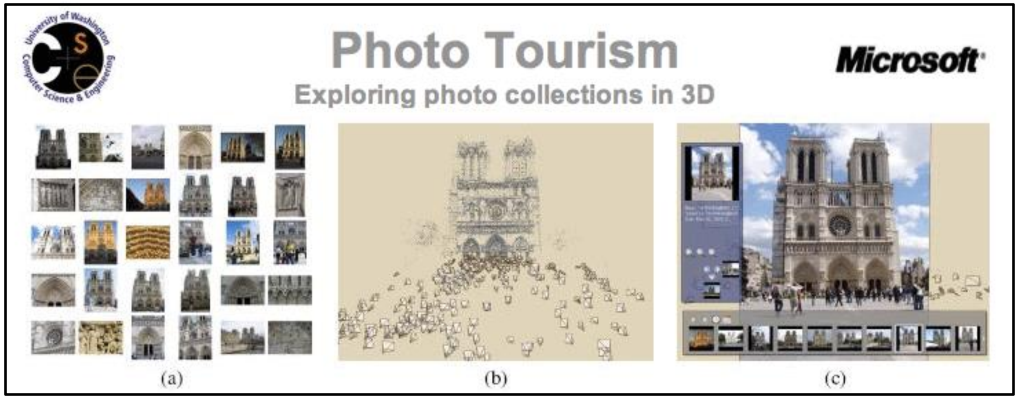
- License Plate Recognition
  - London Congestion Charge https://tfl.gov.uk/modes/driving/congestion-charge
- Security/Surveillance
  - Face Recognition
    - Apple’s Face ID: chance of 1-in-1-million that a random person could unlock your phone
  - Biometric passport (*aka* e-passport) has an embedded electronic chip which contains biometric information
    - Currently standardized biometrics are facial recognition, fingerprint recognition, and iris recognition
  - Airport Security (People Tracking)
- Medical Imaging
  - (Semi-)automatic segmentation and measurements
- Robotics
- Autonomous driving



More Applications

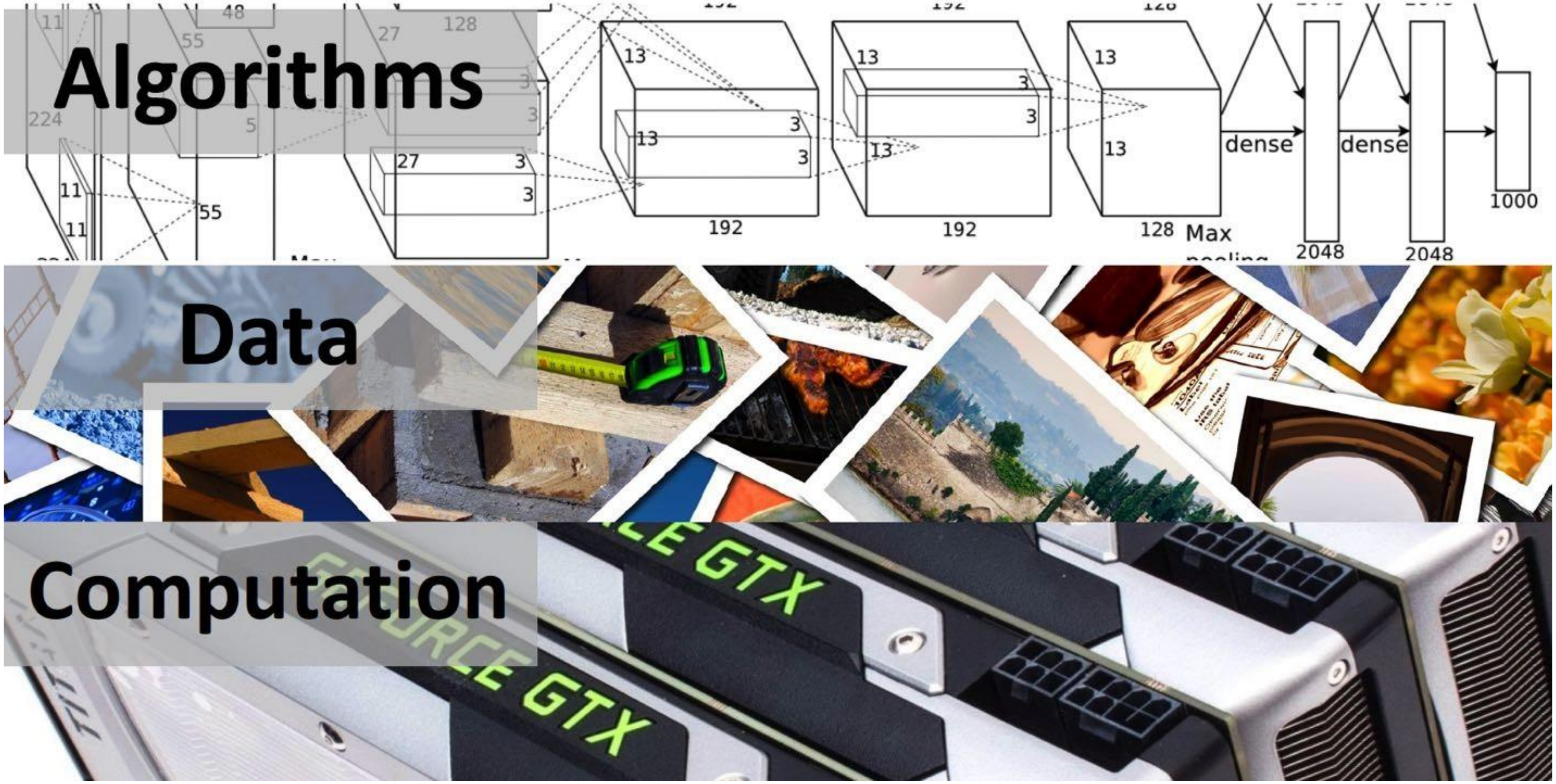
- Vision on Cellphones:
  - e.g. Google Goggles
- Vision for Interfaces:
  - e.g. Microsoft Kinect
- Reconstruction



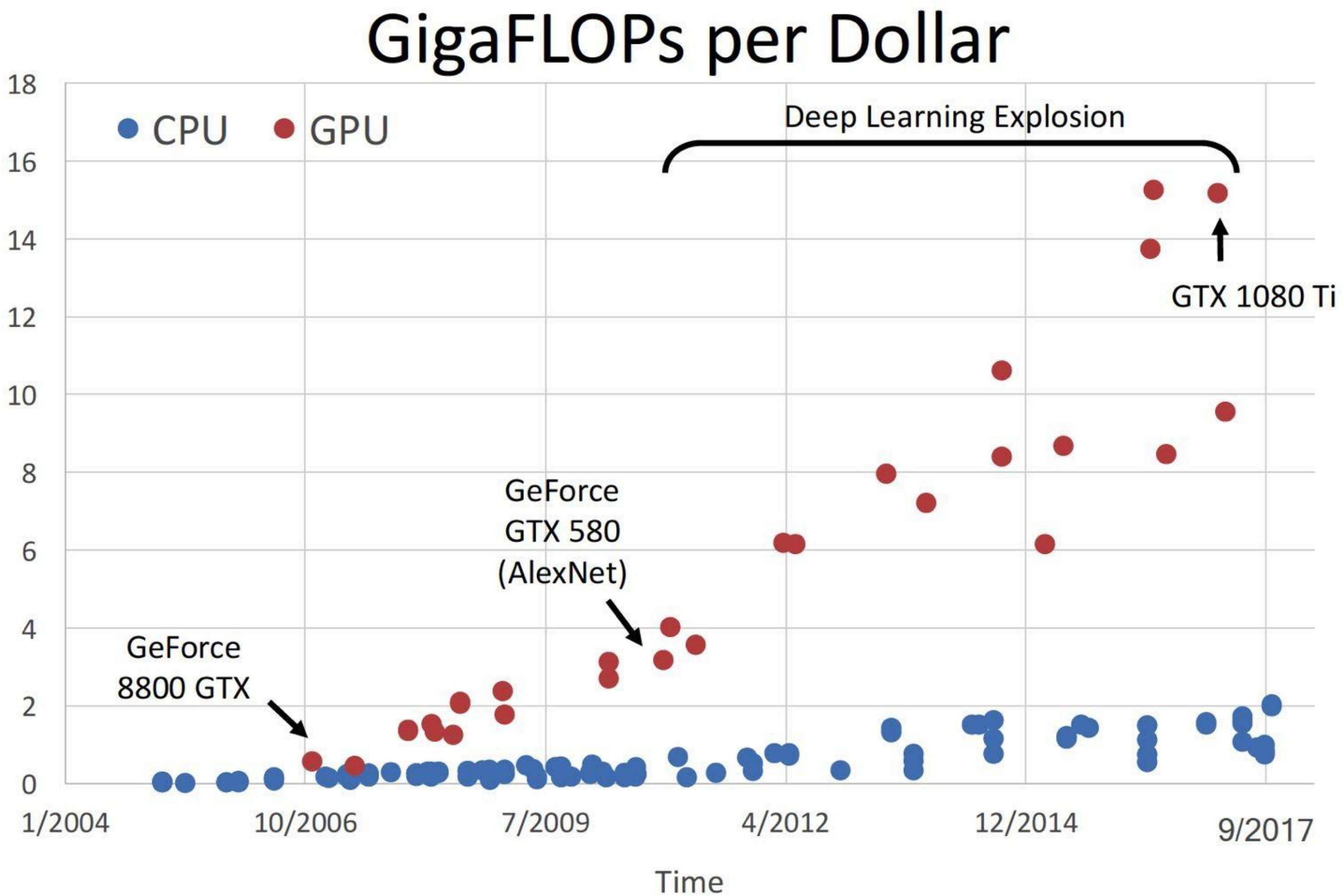


Preamble on Deep Learning

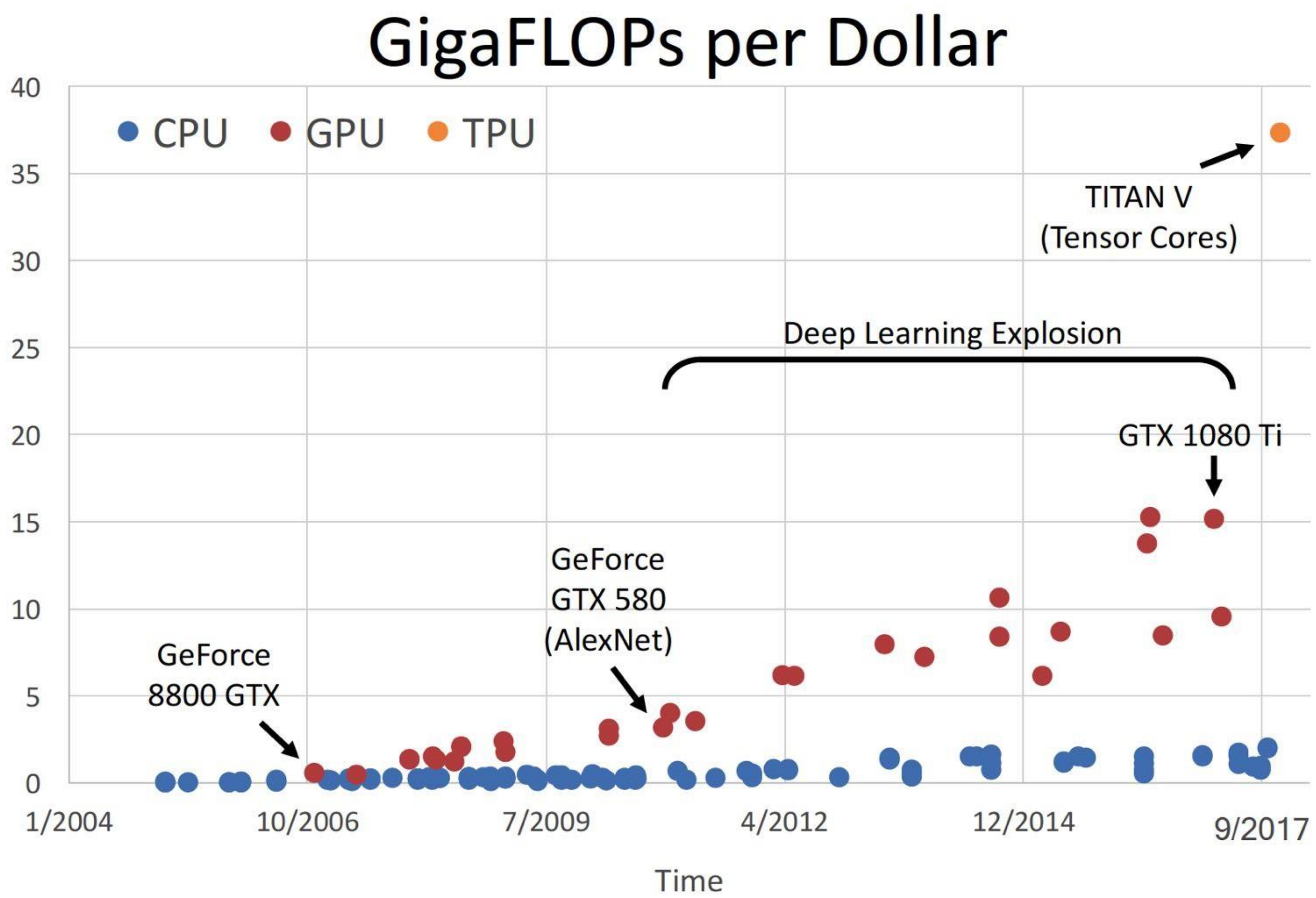
Keys to successes



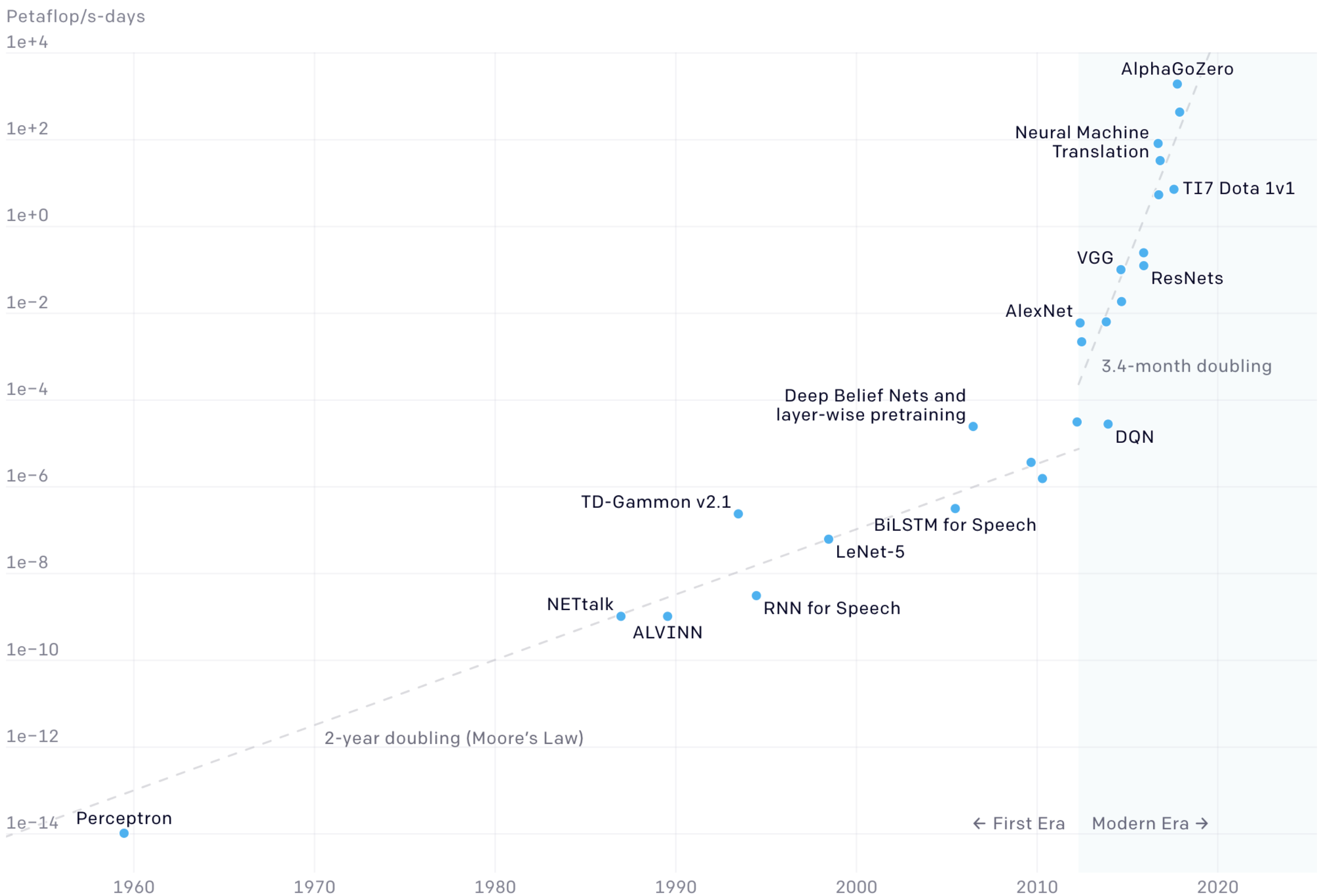
Keys to successes
Computation



Keys to successes
Computation

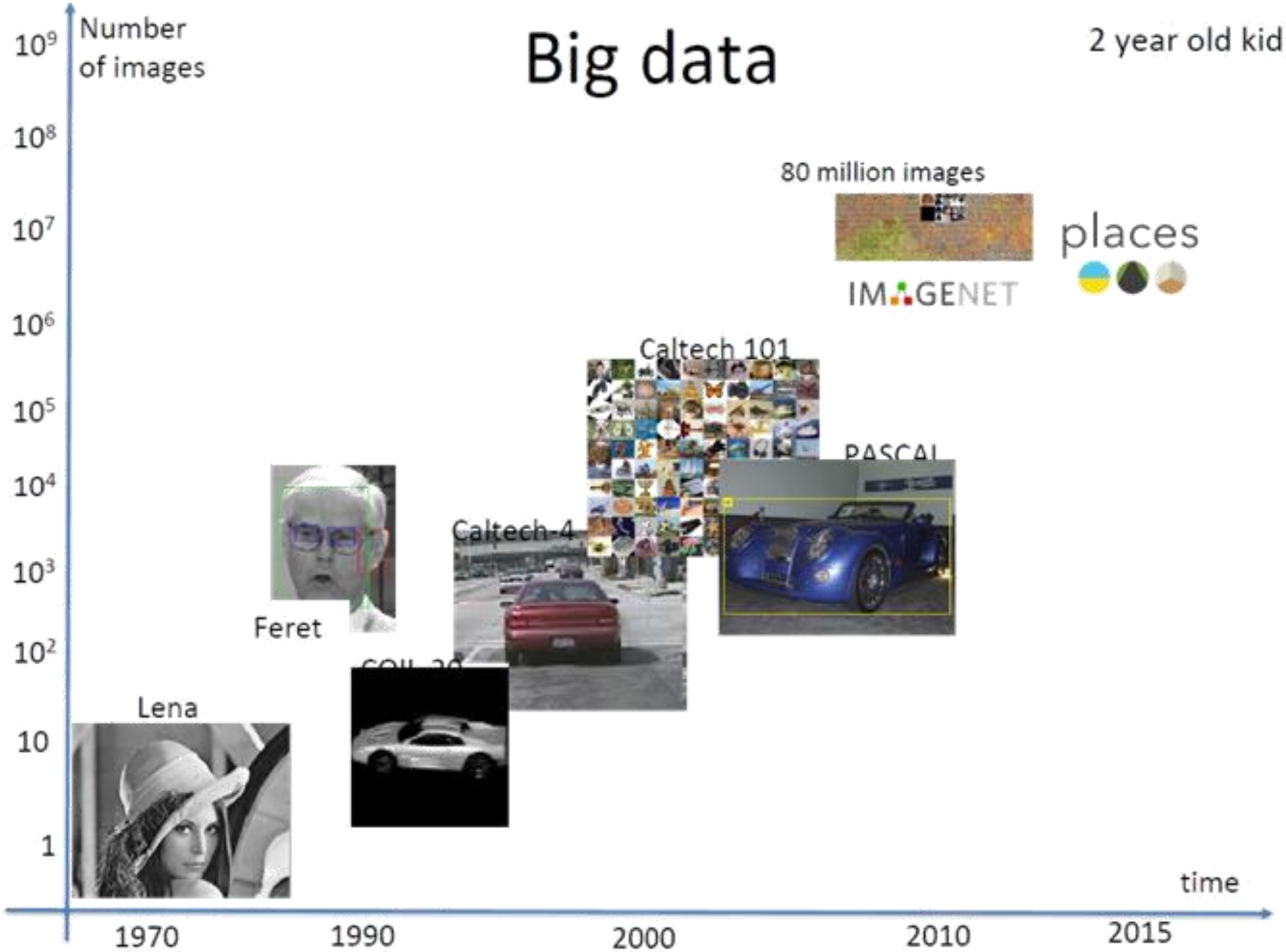


Keys to successes
Computation



Courtesy of OpenAI

Keys to successes
Data



Courtesy of A. Torralba @ICVSS'18

IMAGENET

www.image-net.org

22K categories and **14M** images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
- Food
- Materials
- Structures
- Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
- Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009

IMAGENET Large Scale Visual Recognition Challenge

The Image Classification Challenge:

1,000 object classes

1,431,167 images

Output:
Scale
T-shirt
<u>Steel drum</u>
Drumstick
Mud turtle

✔

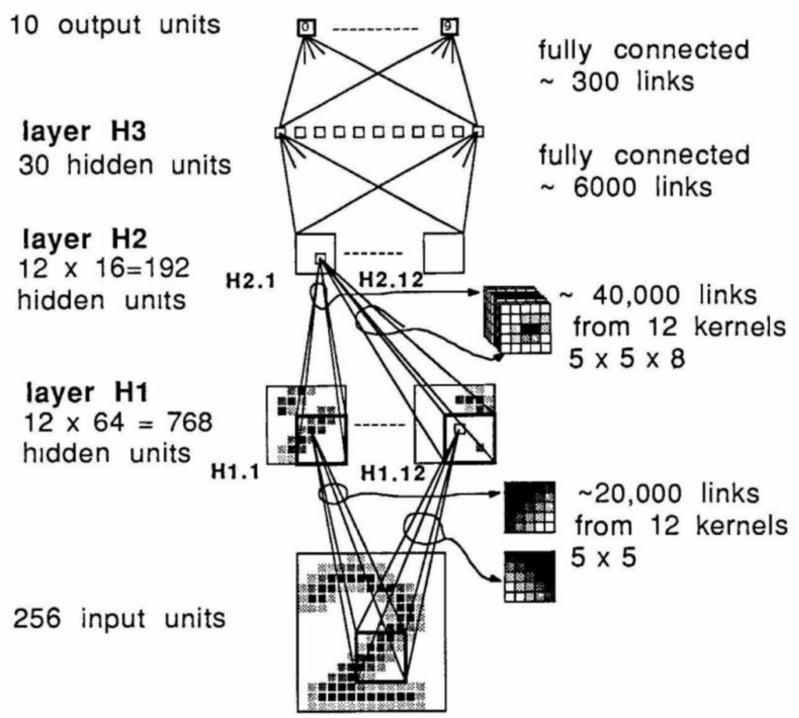
Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

Russakovsky et al. IJCV 2015

Keys to successes
Algorithms

- Progress in modelling
  - Cognitron/Neocognitron [Fukushima 1971-1982]
  - Pooling [Riesenhuber and Poggio 1999]
  - Convnet's [LeCun et al. 1989]
  - Non-linearities [Nair, Hinton 2010]
  - DropOut [Krizhevsky et al. 2012]
  - Batch Normalization [Ioffe Szegedy 2015]
  - Identity mapping [He et al. 2015]
  - Attention [Bengio et al. 2015]
  - …



SAPIENZA UNIVERSITÀ DI ROMA

Thank you

Acknowledges: some slides and material from Jennifer Widom, Matthew R. Gormley, Bernt Schiele, Mario Fritz, Michael Black, Bill Freeman, Fei-Fei, Justin Johnson, Serena Yeung

SAPIENZA
UNIVERSITÀ DI ROMA