GRADUATE SCHOOL AND RESEARCH CENTER AT THE HEART OF THE DIGITAL SOCIETY

EURECOM
Sophia Antipolis

Expert in the loop: Human-machine interaction to build safe AI-based systems

Maria A. Zuluaga

Reliable AI: Healthcare as a use case

06/2011
IBM Watson wins at Jeopardy[1]

12/2012
Watson to be trained for oncological patient treatment[2]

02/2013
Lung cancer treatment assistant: Watson's first commercial product[3]

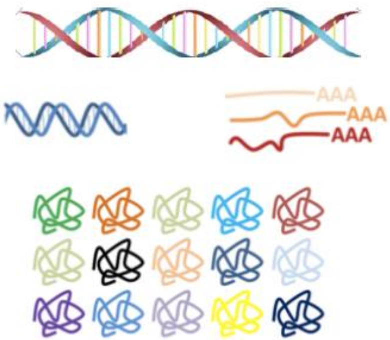
05/2015
Why Watson is taking so long?[4]

06/2018
Layoffs at Watson Health are linked to "problems with AI"[5]

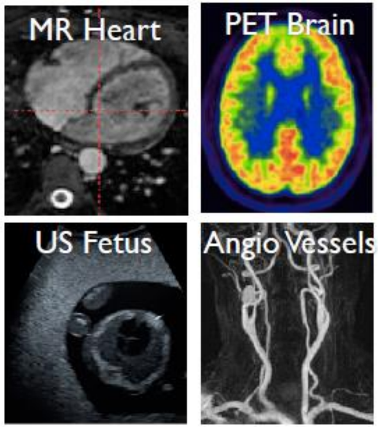
[1] IEEE Spectrum - http://disq.us/t/dyp9wt; [2] IEEE Spectrum - http://disq.us/t/gki3ri; [3] Forbes - https://bit.ly/2TBIYJi
[4] IEEE Spectrum - http://disq.us/t/1q91ji9; [5] IEEE Spectrum - http://disq.us/t/33p3080

The challenges

1. Data is inherently complex and expensive to collect



Molecular & -omics data



Biomedical images

Clinical records

The challenges

1. Data is inherently complex and expensive to collect
2. High-impact, high-risk decisions with little tolerance to errors

source: https://www.mskcc.org/locations/directory/memorial-hospital

> *"…the rate of agreement was about 33 percent — so the hospital decided not to buy the system "*
>
> STAT, Boston Globe Media - Sept 5, 2017

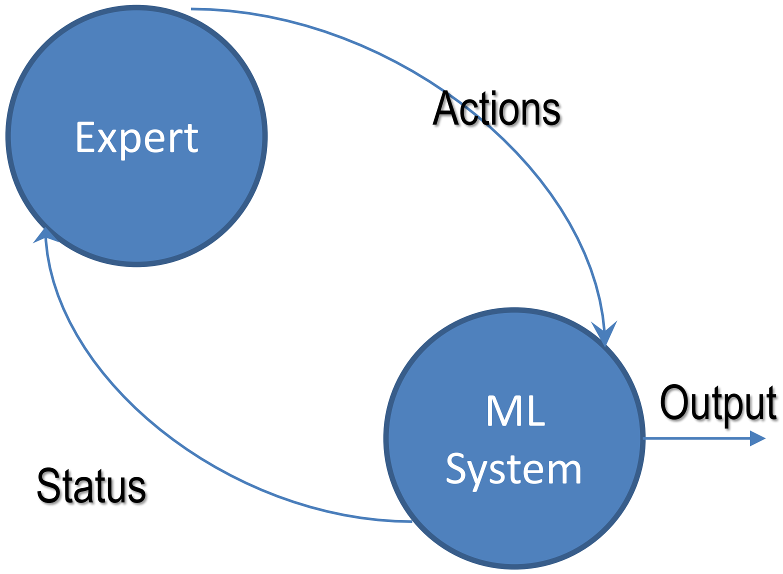
source: https://bit.ly/2xOq7QK

The challenges

1. Data is inherently complex and expensive to collect
2. High-impact, high-risk decisions with little tolerance to errors

There is a need for <u>robust and reliable</u> methods that can be safely deployed in practice

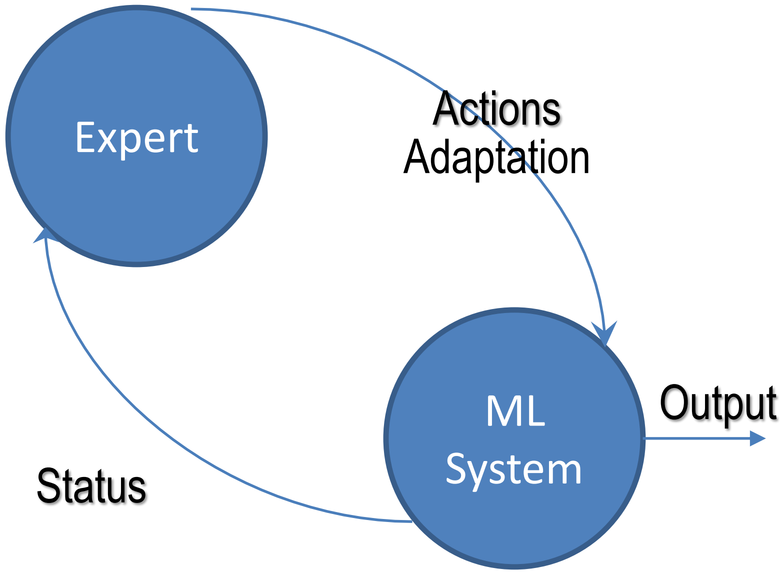
Expert in the loop



Expert in the loop

Challenge 1:

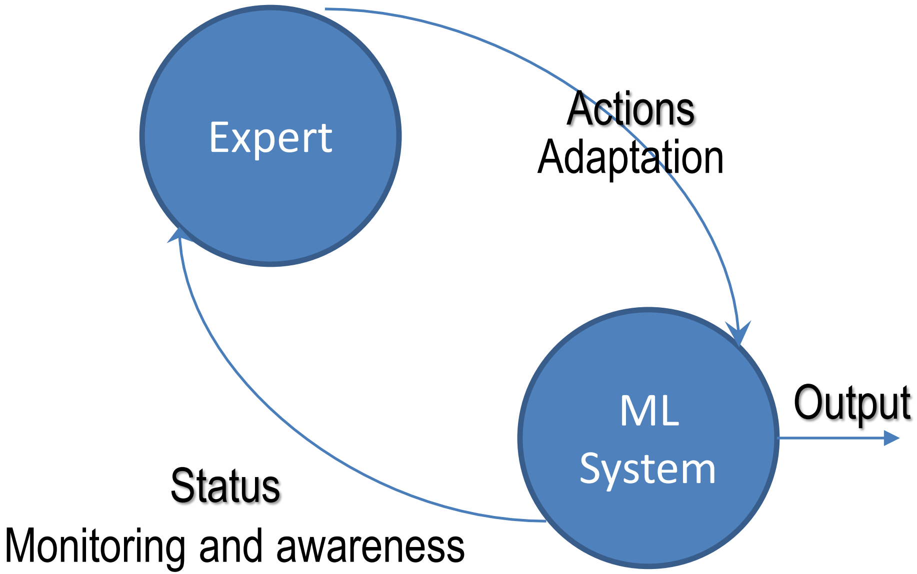
Interactivity



Expert in the loop

Challenge 1:
Interactivity

Challenge 2:
Feedback

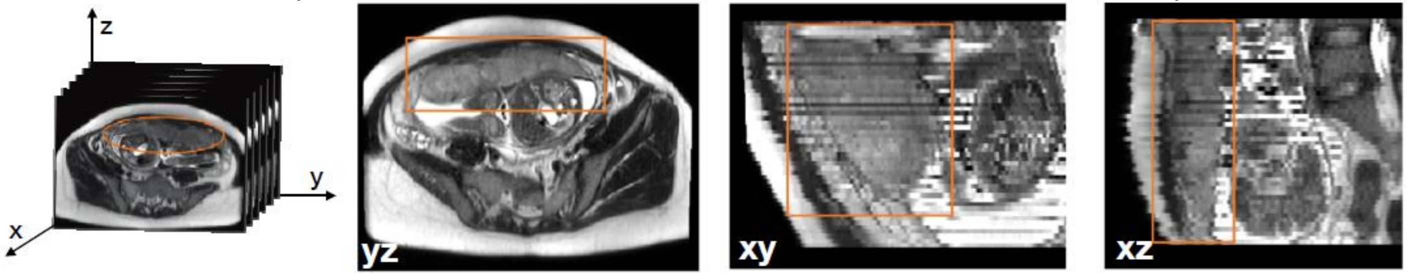


Challenge 1:

INTERACTIVITY TO DEAL WITH DATA

Low quality training data

Human placenta: 3D reference and v*iew from the 3 standard anatomical planes*



adapted: [Wang et al. MICCAI 2015]

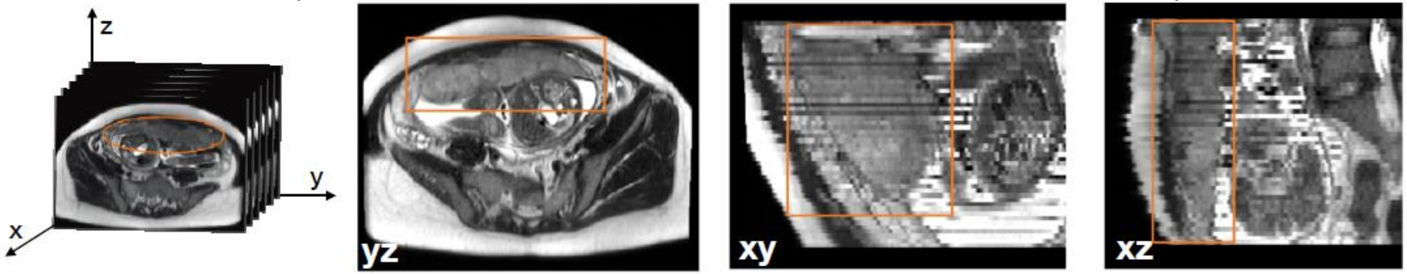
Challenges: Poor 3D image quality and varying locations

State of the art

- Large number of interactions
- Constrained interactions
- Limited learning ability of the underpinning model

Low quality training data

Human placenta: 3D reference and *view from the 3 standard anatomical planes*



adapted: [Wang et al. MICCAI 2015]

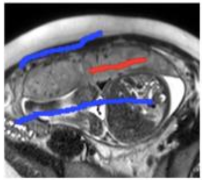
Challenges: Poor 3D image quality and varying locations

State of the art

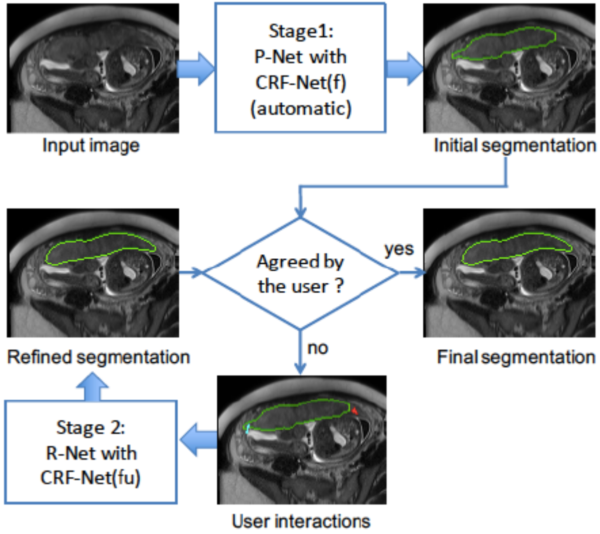
- Large number of interactions
- Constrained interactions
- Limited learning ability of the underpinning model

Minimally interactive
Flexible
Robust
Works in 2D & 3D

Human placenta:
yz plane



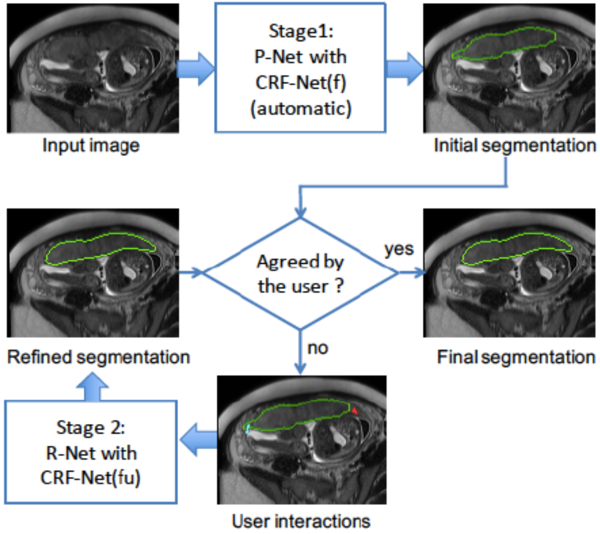
A Deep Interactive Framework for Image Segmentation



1. INTERACTION FOR DATA AUGMENTATION
2. INTERACTION TO DEFINE A PRIOR FOR SMOOTHER RESULTS

G. Wang, MA. Zuluaga, et al. IEEE TPAMI(41) 2019

A Deep Interactive Framework for Image Segmentation



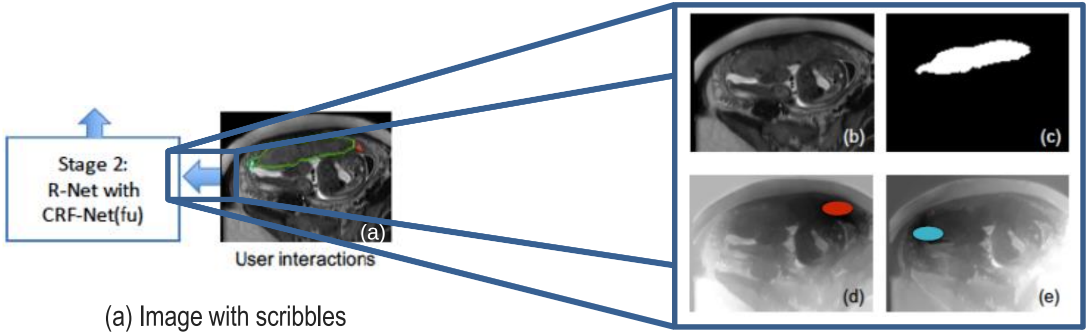
1. INTERACTION FOR DATA AUGMENTATION

The original image **I** and the initial segmentation are concatenated with the geodesic distance maps and used as input of the refinement net (R-Net).

A Deep Interactive Framework for Image Segmentation

1. Interaction for data augmentation

The original image **I** and the initial segmentation are concatenated with the geodesic distance maps and used as input of the refinement net (R-Net).



(a) Image with scribbles

(b) Image **I** (c) Segmentation
(d) Foreground GM (e) Background GM

$$G(i, \mathcal{S}, \mathbf{I}) = \min_{j \in \mathcal{S}} D_{geo}(i, j, \mathbf{I})$$

$$D_{geo}(i, j, \mathbf{I}) = \min_{p \in \mathcal{P}_{i,j}} \int_0^1 ||\nabla \mathbf{I}(p(s)) \cdot \mathbf{u}(s)|| ds$$

A Deep Interactive Framework for Image Segmentation

Stage1:
P-Net with
CRF-Net(f)
(automatic)

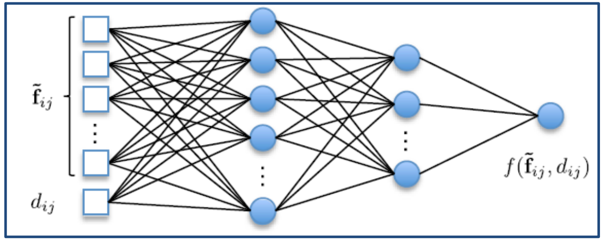
Stage 2:
R-Net with
CRF-Net(fu)

2. INTERACTION TO DEFINE A PRIOR FOR SMOOTHER RESULTS

Trainable conditional random field (CRF) using freeform functions as pair-wise potentials

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{(i,j)\in\mathcal{N}} \psi_p(x_i, x_j)$$

Scores from P- or R-Net

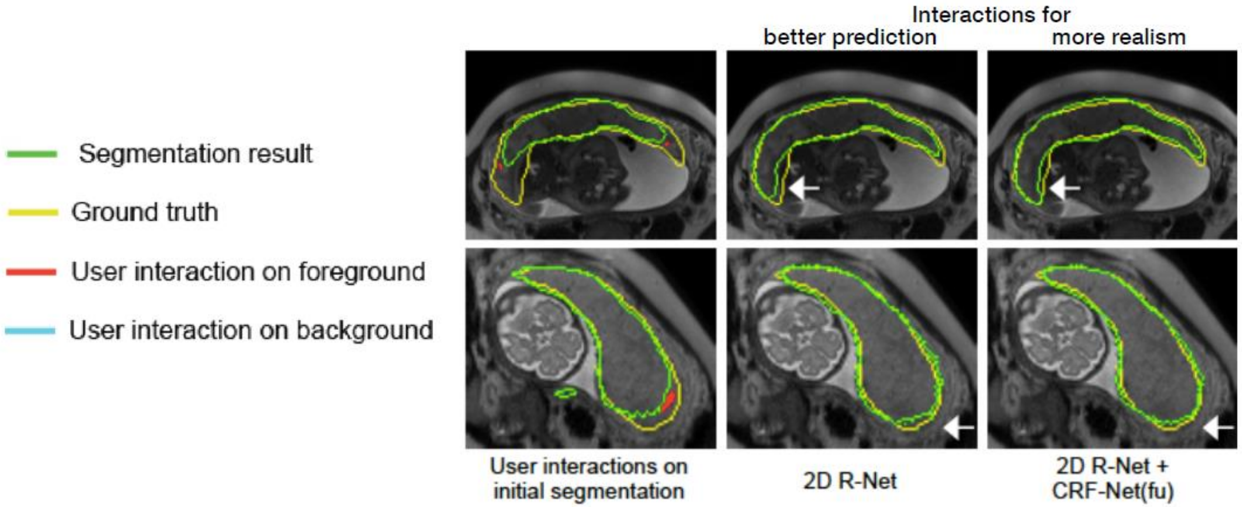


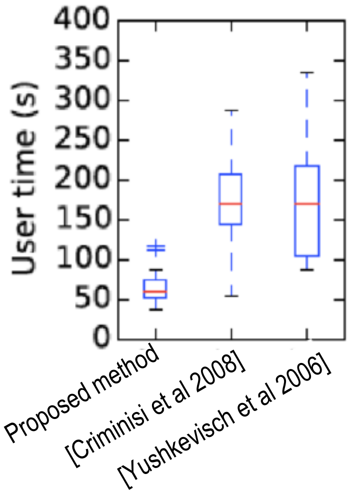
Mean-field approximation Q to infer the CRF ψ_P

$$Q_i(x_i|\mathbf{I}) = \begin{cases} 1 & \text{if } i \in \mathcal{S}_{fb} \text{ and } x_i = s_i \\ 0 & \text{if } i \in \mathcal{S}_{fb} \text{ and } x_i \neq s_i \\ \frac{1}{Z_i} e^{-E(x_i)} & \text{otherwise,} \end{cases}$$

G. Wang, MA. Zuluaga, et al. IEEE TPAMI(41) 2019

A Deep Interactive Framework for Image Segmentation

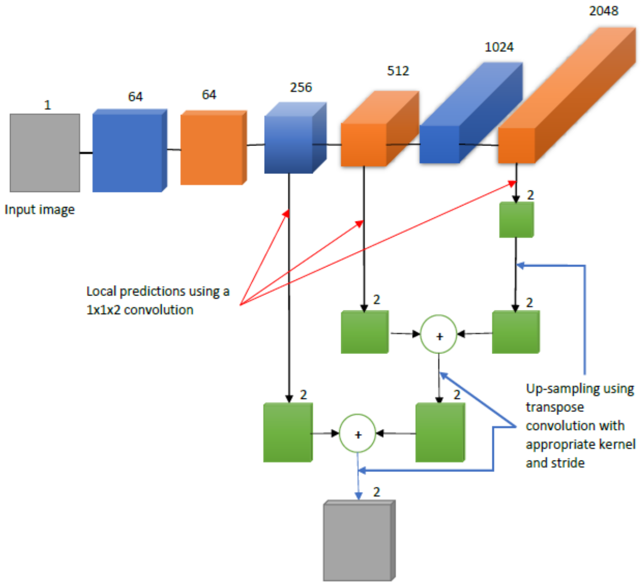


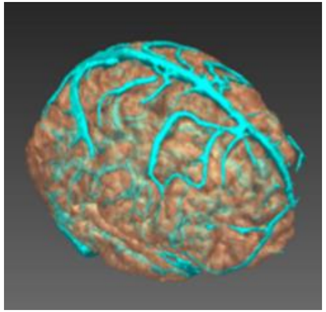


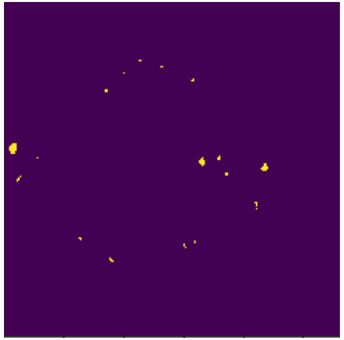
G. Wang, MA. Zuluaga, et al. IEEE TPAMI(41) 2019

Currently: Deal with small objects

P. Mathur







- Understand the effects of “bad” interactions

Challenge 2:

FEEDBACK

Hidden Technical Debt of ML Systems

Sculley et al. NeurIPS 2015

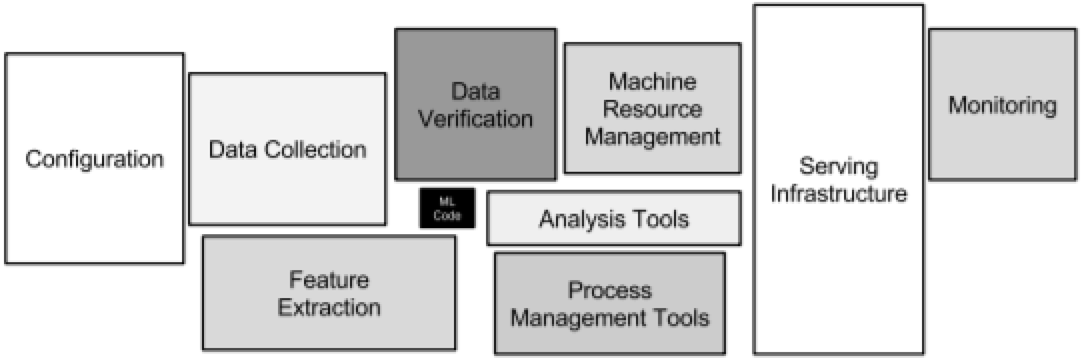


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Hidden Technical Debt of ML Systems

Sculley et al. NeurIPS 2015

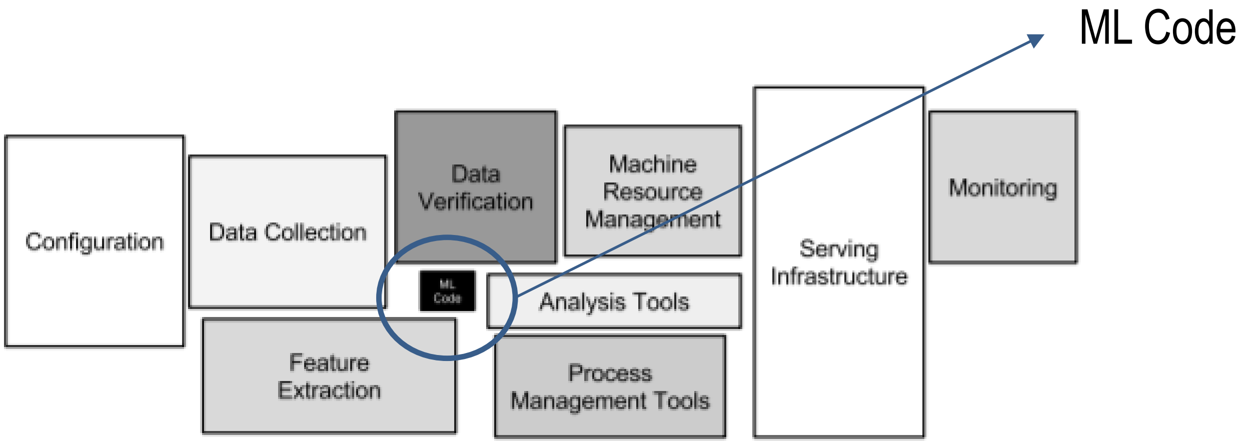


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Hidden Technical Debt of ML Systems

Sculley et al. NeurIPS 2015

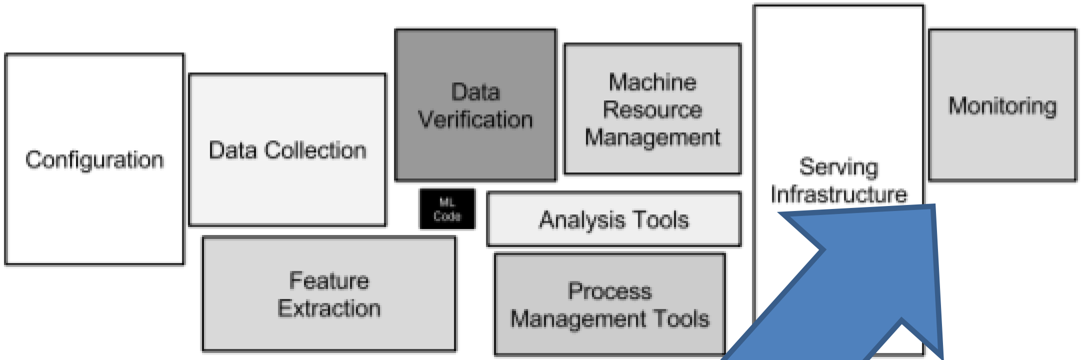


Figure 1: Only a small fraction of real-world ML systems is … of the ML code, as shown by the small black box in the middle. The required surroun… …ructure is vast and complex.

Our idea:

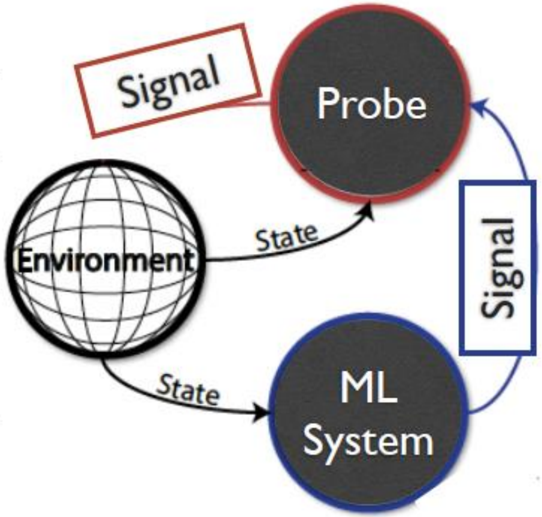
ML Systems to monitor ML Systems

INSTITUT CARNOT
Télécom & Société numérique

EURECOM
Sophia Antipolis

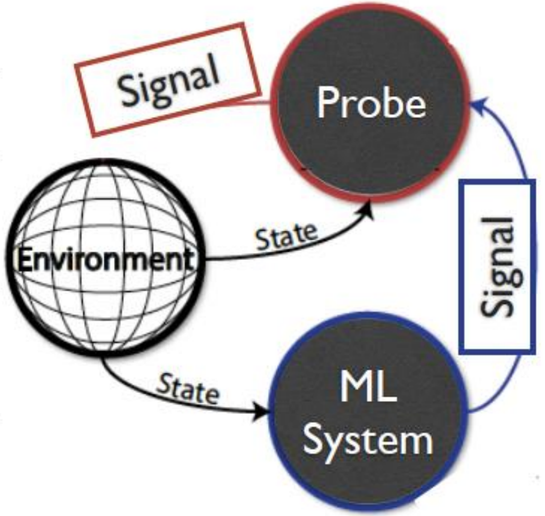
Model monitoring: Probe principle

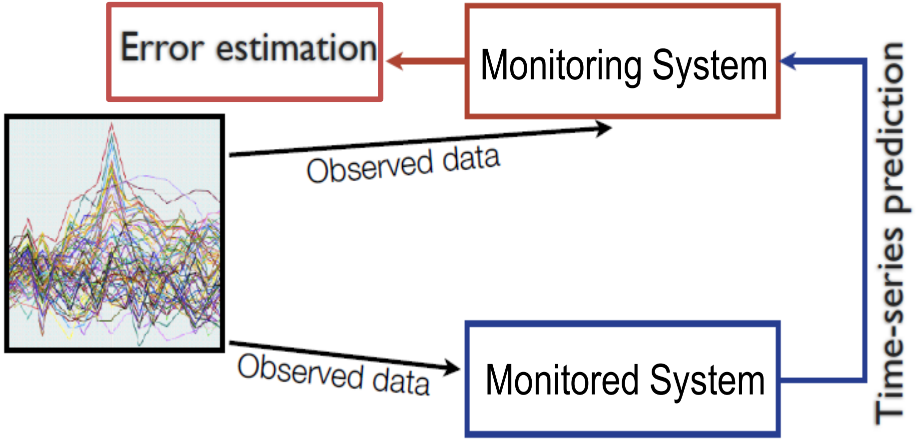
Adapted from [Haber et al, 2018]



Model monitoring: Probe principle

Adapted from [Haber et al, 2018]





Hypothesis: We can train a model to monitor the performance of other models over time

INSTITUT CARNOT Télécom & Société numérique

R. Candela, MA. Zuluaga, P. Michiardi, M. Filippone (under review)

EURECOM

Model monitoring: Empirical study

Materials

- Data: 23.4K time series from the M4 benchmark
- Monitoring models: LSTM, CNN, Bayesian CNN, GP
- Monitored models: 10 different methods

Model monitoring: Empirical study

Materials

- Data: 23.4K time series from the M4 benchmark
- Monitoring models: LSTM, CNN, Bayesian CNN, GP
- Monitored models: 10 different methods

Method

- Measure the capacity of the monitoring models to estimate the forecast prediction error of the monitored models

$$sMAPE = \frac{1}{h}\sum_{t=1}^{h} 2\frac{|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|}$$

Results: Monitoring

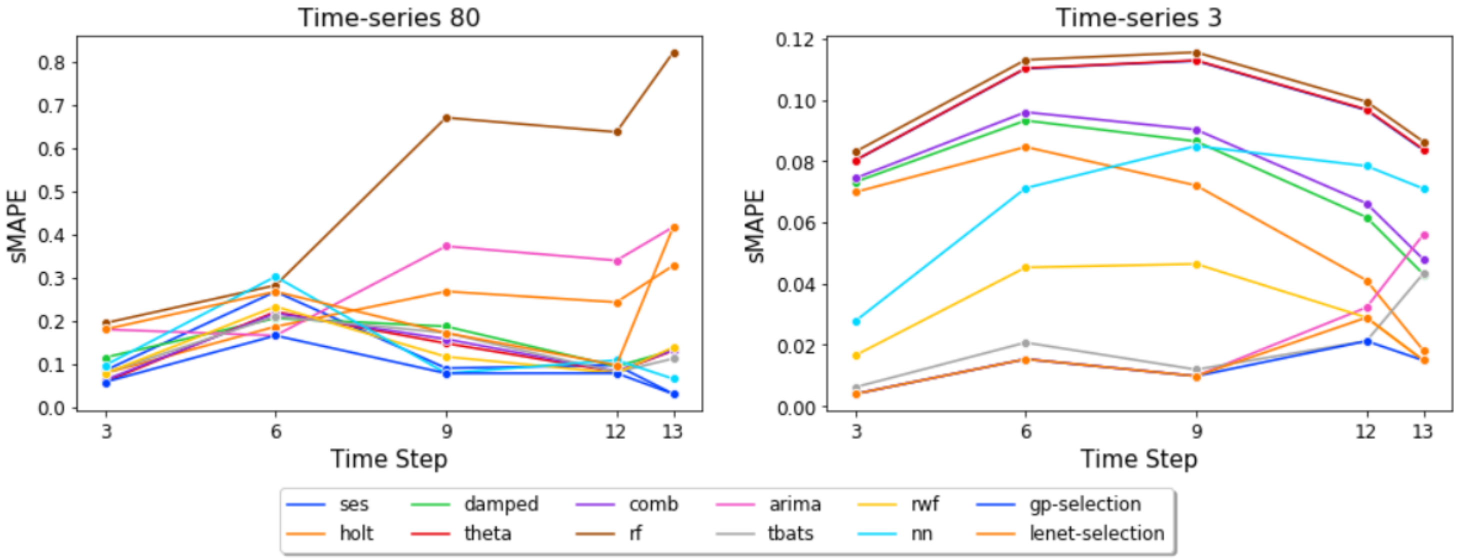
	Ground Truth		Monitoring models						Baseline	
			LSTM		Bayes-LeNet		GPs			
	YEARLY - $h = 6$									
	model	*sMAPE*	*monitored model*	*sMAPE*	*monitored model*	*sMAPE*	*monitored model*	*sMAPE*	*model*	*sMAPE*
1	comb	0.166 (0.196)	damped	0.132 (0.076)	comb	<u>0.144</u> (0.056)	comb	0.153 (0.218)	rf	0.225 (0.269)
2	damped	0.167 (0.31)	comb	0.133 (0.094)	theta	<u>0.146</u> (0.058)	damped	0.155 (0.237)	damped	0.368 (0.310)
3	theta	0.169 (0.183)	theta	0.136 (0.072)	damped	<u>0.149</u> (0.054)	theta	<u>0.159</u> (0.230)	holt	0.370 (0.322)
4	holt	0.176 (0.225)	holt	0.139 (0.125)	holt	0.156 (0.126)	holt	<u>0.163</u> (0.273)	comb	0.385 (0.313)
5	ses	0.191 (0.182)	ses	0.167 (0.076)	ses	0.175 (0.067)	ses	0.184 (0.214)	theta	0.406 (0.320)
6	rf	0.221 (0.241)	rf	0.208 (0.156)	rf	0.212 (0.098)	rf	0.210 (0.266)	ses	0.437 (0.324)

Results: Monitoring

	Ground Truth		*Monitoring models*						Baseline	
			LSTM		Bayes-LeNet		GPs			
	YEARLY - $h = 6$									
	model	*sMAPE*	*monitored model*	*sMAPE*	*monitored model*	*sMAPE*	*monitored model*	*sMAPE*	*model*	*sMAPE*
1	comb	0.166 (0.196)	damped	0.132 (0.076)	comb	<u>0.144</u> (0.056)	comb	0.153 (0.218)	rf	0.225 (0.269)
2	damped	0.167 (0.31)	comb	0.133 (0.094)	theta	<u>0.146</u> (0.058)	damped	0.155 (0.237)	damped	0.368 (0.310)
3	theta	0.169 (0.183)	theta	0.136 (0.072)	damped	<u>0.149</u> (0.054)	theta	<u>0.159</u> (0.230)	holt	0.370 (0.322)
4	holt	0.176 (0.225)	holt	0.139 (0.125)	holt	0.156 (0.126)	holt	<u>0.163</u> (0.273)	comb	0.385 (0.313)
5	ses	0.191 (0.182)	ses	0.167 (0.076)	ses	0.175 (0.067)	ses	0.184 (0.214)	theta	0.406 (0.320)
6	rf	0.221 (0.241)	rf	0.208 (0.156)	rf	0.212 (0.098)	rf	0.210 (0.266)	ses	0.437 (0.324)

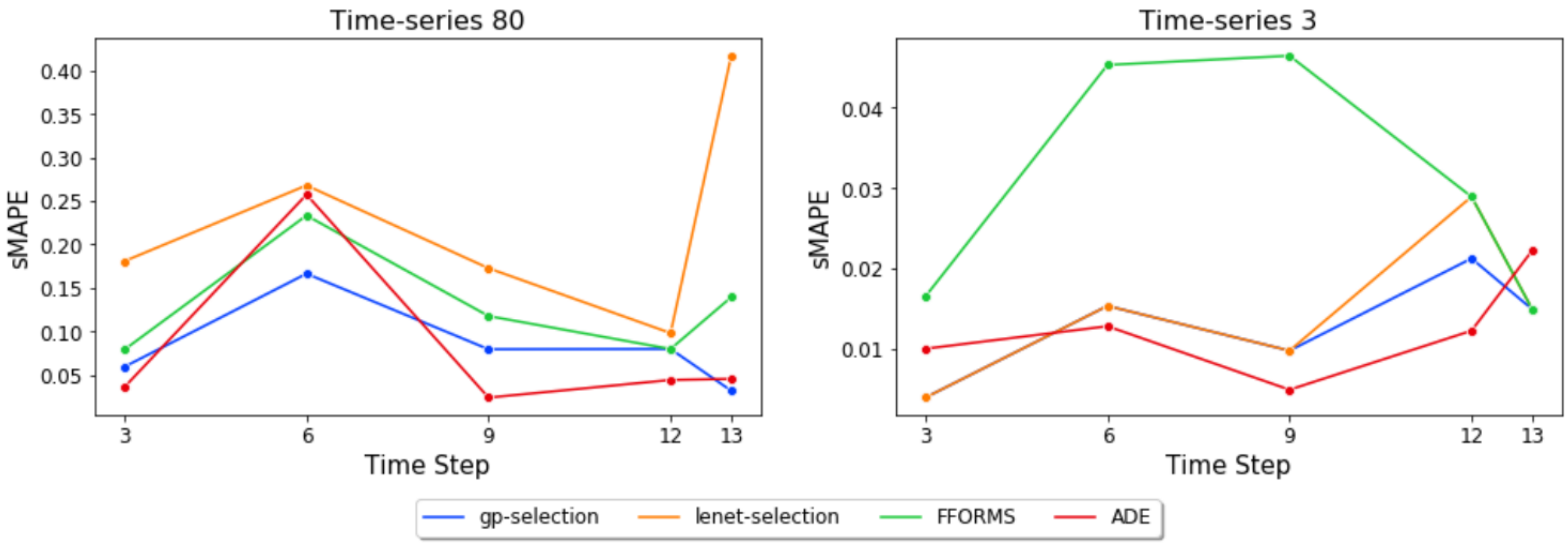
- Further direction: Can we use it for model selection?

Results: Model Selection



R. Candela, MA. Zuluaga, P. Michiardi, M. Filippone (under review)

Error monitoring vs meta-learning



- Competitive results using simpler “base” learners

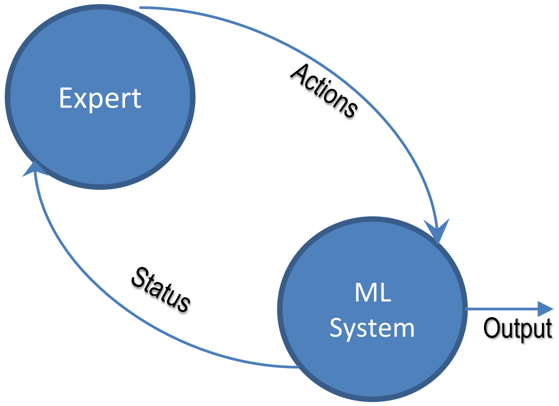
R. Candela, MA. Zuluaga, P. Michiardi, M. Filippone (under review)

INSTITUT CARNOT Télécom & Société numérique

EURECOM

Summary

- Interactive ML is a promising way to achieve <u>robustness and reliability</u> by keeping experts involved and informed
- Demonstrations in two aspects:
  - Efficient user input to deal with training data
  - System feedback to alert the expert



Acknowledgments

EURECOM

Prof. P. Michiardi
Prof. M. Filippone
R. Candela
P. Mathur

KCL

Prof. S. Ourselin
Prof. T. Vercauteren

UCL

Prof. A. David, MD
Dr. R. Pratt
Dr. T. Doel
Dr. R. Rodionov

KU Leuven

Prof. J. Deprest, MD
Dr. M. Aertsen

University of Electronic Science and Technology of China

Dr. G. Wang

GOSH

Dr. P.A. Patel, MD

Great Ormond Street Hospital

Amadeus

INSTITUT CARNOT
Télécom & Société numérique

EURECOM

Expert in the loop: Human-machine interaction to build safe AI-based systems

THANK YOU

INSTITUT CARNOT
Télécom & Société numérique

EURECOM
Sophia Antipolis