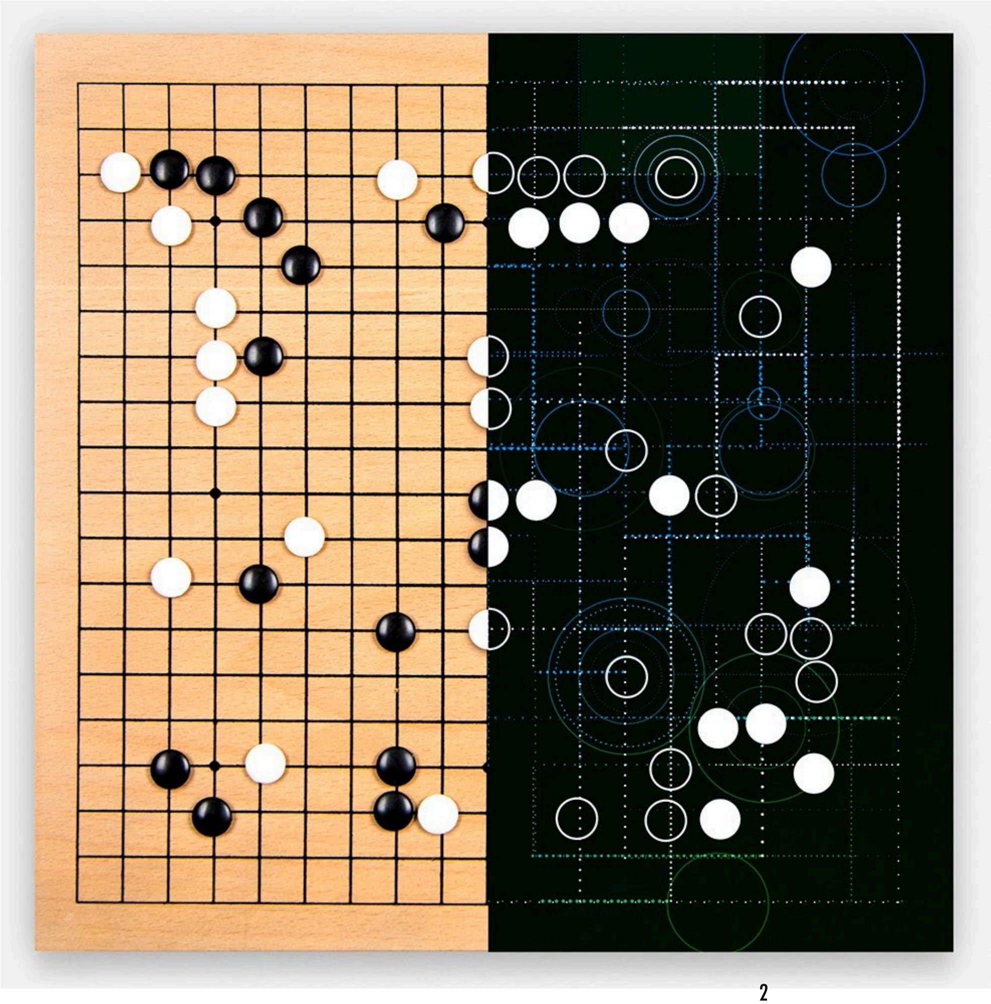
DEEP LEARNING IN NON-COGNITIVE DOMAINS

Truyen Tran
PRaDA, Deakin

DEEP LEARNING IN COGNITIVE DOMAINS

AlphaGo



DEEP LEARNING IN NON-COGNITIVE DOMAINS

Where humans need extensive training to do well

Domains with great diversity but small in size

Domains with great uncertainty, low-quality/missing data

Domains that demand transparency & interpretability.

… **healthcare, security, foods, water, manufacturing**

AGENDA

Introduction to PRaDA

Introduction to deep learning

Deep learning for [X], where X =

- Healthcare
- Software engineering
- Choice and ranking
- Anomaly detection
- Multi-relational databases
- Representation

The open room

CENTRE FOR PATTERN RECOGNITION AND DATA ANALYTICS

PRADA: MAKING DATA SPEAK

Domains

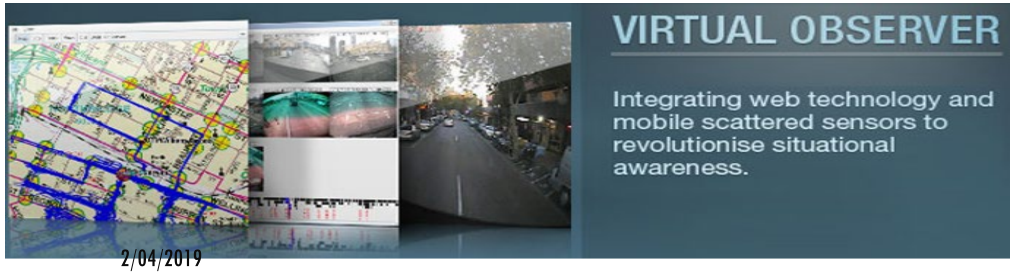
- Health
- Pervasive computing
- Social media
- Manufacturing
- Cybersecurity
- Now: in collaboration with UoW – software engineering and process mining

Methods

- Deep learning
- Bayesian nonparametrics (topic models included)
- Sparse methods (e.g., compress sensing)
- Probabilistic graphical models
- Distributed computing
- Optimization

PRADA: THE MAKING

4 Startups



Collaborators

IFM INSTITUTE FOR FRONTIER MATERIALS

DTI

autism west

Epworth

Barwon Health

KOLLING INSTITUTE

BLACK DOG INSTITUTE

AGENDA

Introduction to PRaDA

Introduction to deep learning

Deep learning for [X], where X =

- Healthcare
- Software engineering
- Choice and ranking
- Anomaly detection
- Multi-relational databases
- Representation

The open room

INTRODUCTION TO DEEP LEARNING

Google Brain

Google DeepMind

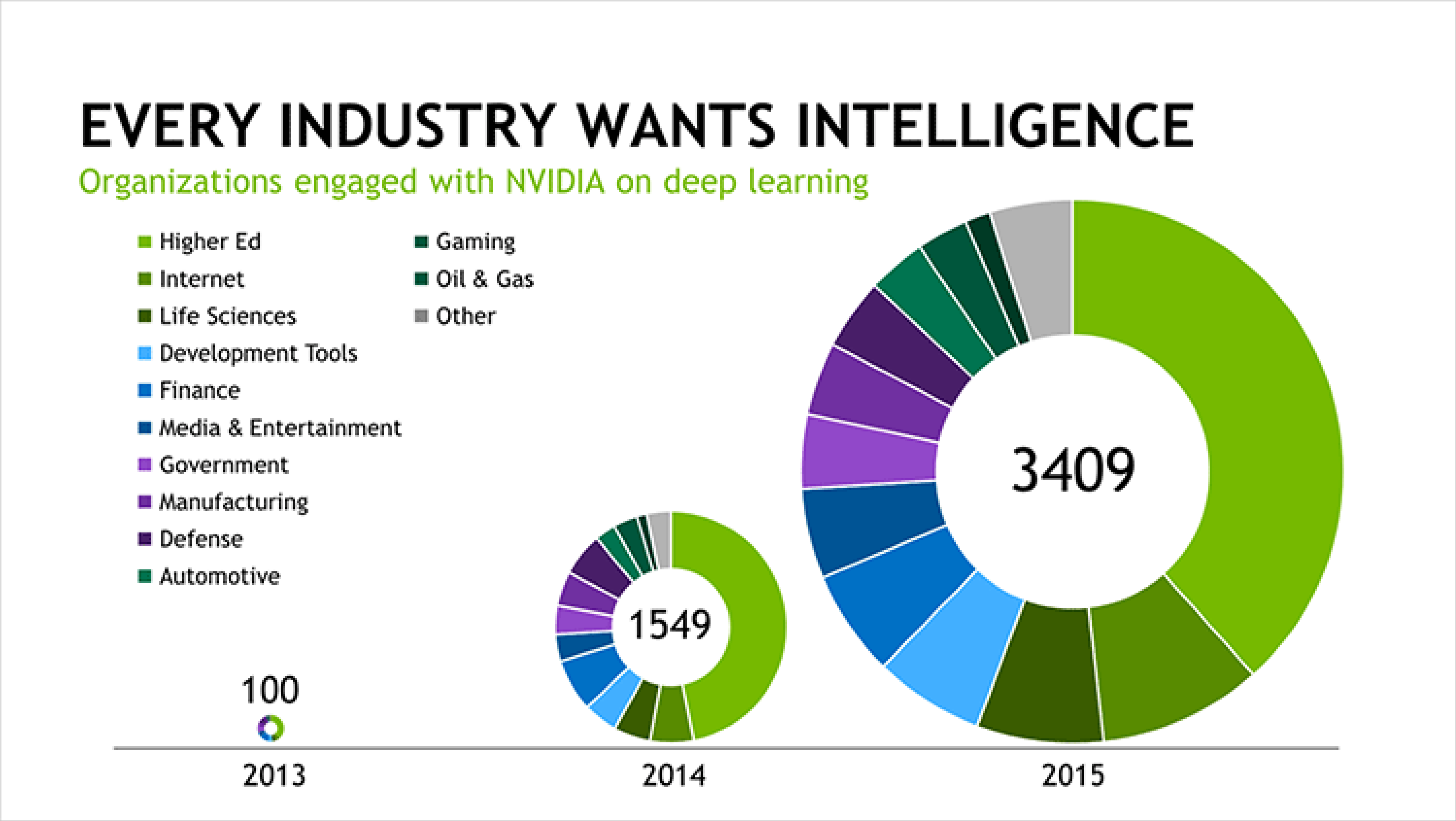
Facebook (FAIR)

Baidu

Microsoft

Twitter Cortex

IBM



DEEP LEARNING: MACHINE THAT LEARNS EVERYTHING

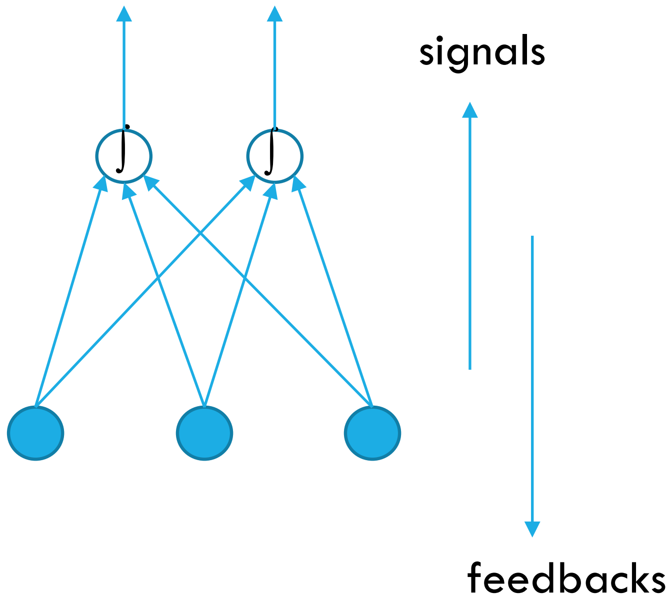
End-to-end machine learning – no human involved.

Models are **compositional**, e.g., object is composed of parts.

- → Models can be complex, but building block is simple and universal!
- → Learning is more efficient in multiple steps

Things can be learn: **Feature** | Selectivity | Invariance | Dynamics | Memory encoding and forgetting | Attention | Planning

THE BUILDING BLOCKS: FEATURE DETECTOR



WHY FEATURE LEARNING?

In typical machine learning projects, 80-90% effort is on feature engineering

- A right feature representation doesn't need much work. Simple linear methods often work well.

Vision: Gabor filter banks, SIFT, HOG, BLP, BOW, etc.

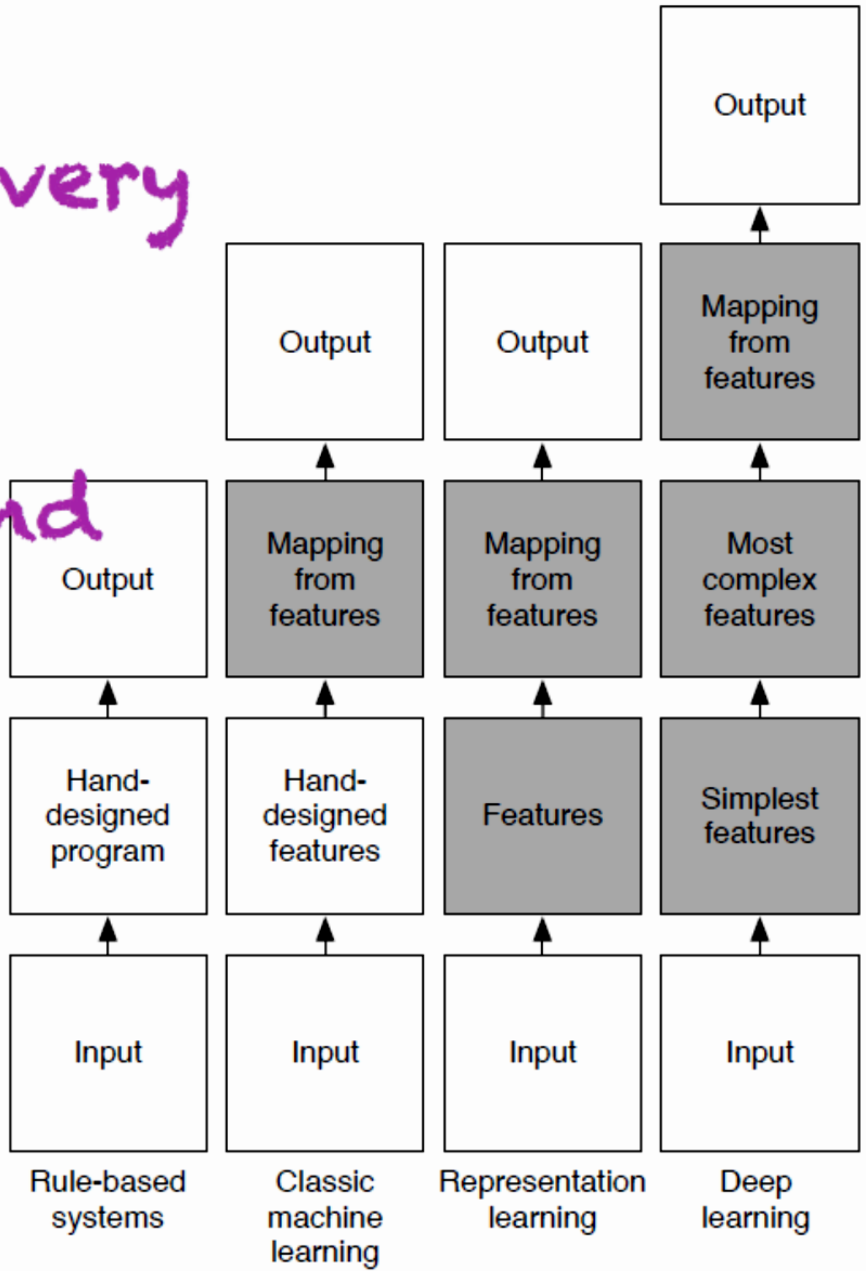
Text: BOW, n-gram, POS, topics, stemming, tf-idf, etc.

SW: token, LOC, API calls, #loops, developer reputation, team complexity, report readability, discussion length, etc.

Try yourself on Kaggle.com!

Automating Feature Discovery

Discovering and representing higher-level abstractions

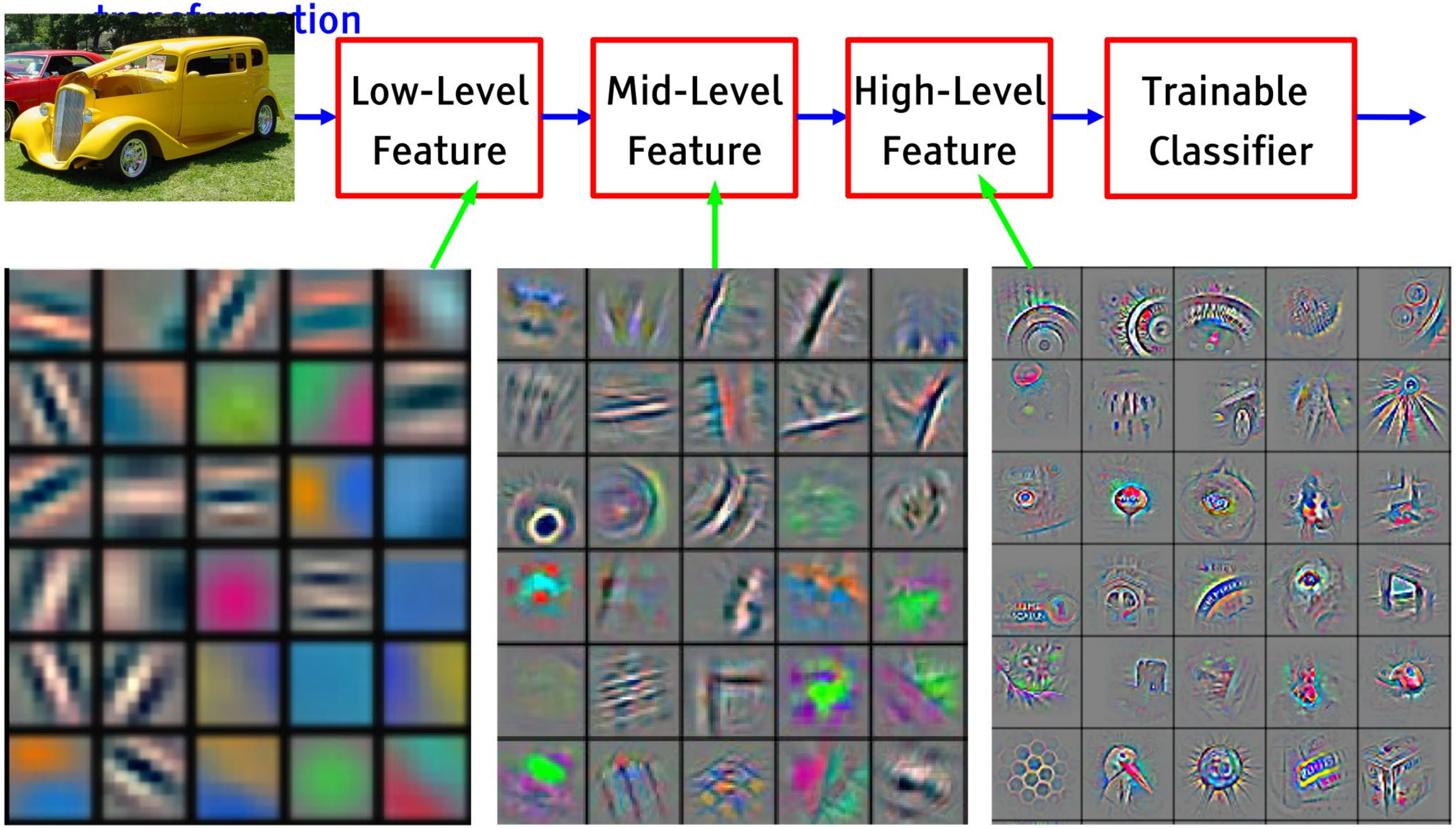


(Bengio, DLSS 2015)

Deep Learning = Learning Hierarchical Representations

Y LeCun
MA Ranzato

- It's deep if it has more than one stage of non-linear feature transformation



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

THREE MAIN ARCHITECTURES

Deep (DNN): Vector to vector

- Most existing ML/statistics fall into this category

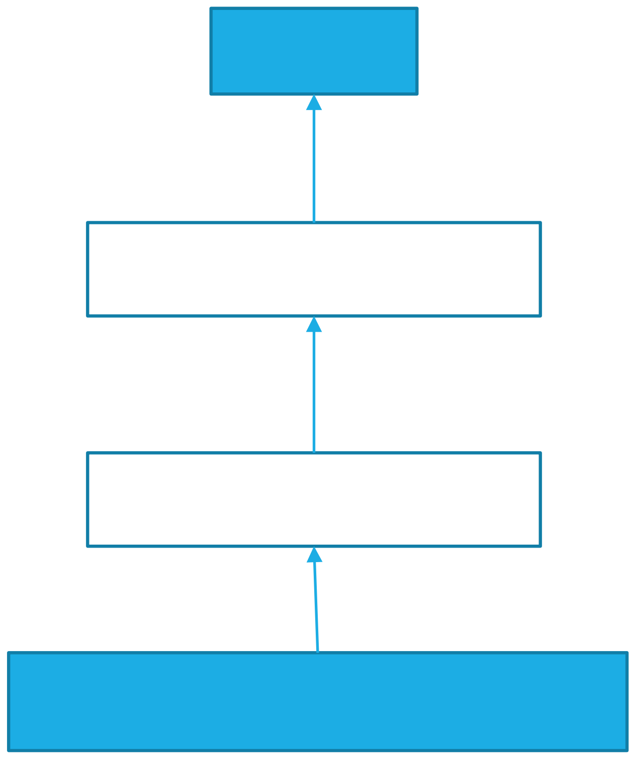
Recurrent (RNN): Sequence to sequence

- Temporal, sequential. E.g., sentence, actions, DNA, EMR
- Program evaluation/execution. E.g., sort, traveling salesman problem

Convolutional (CNN): Image to vector/sequence/image

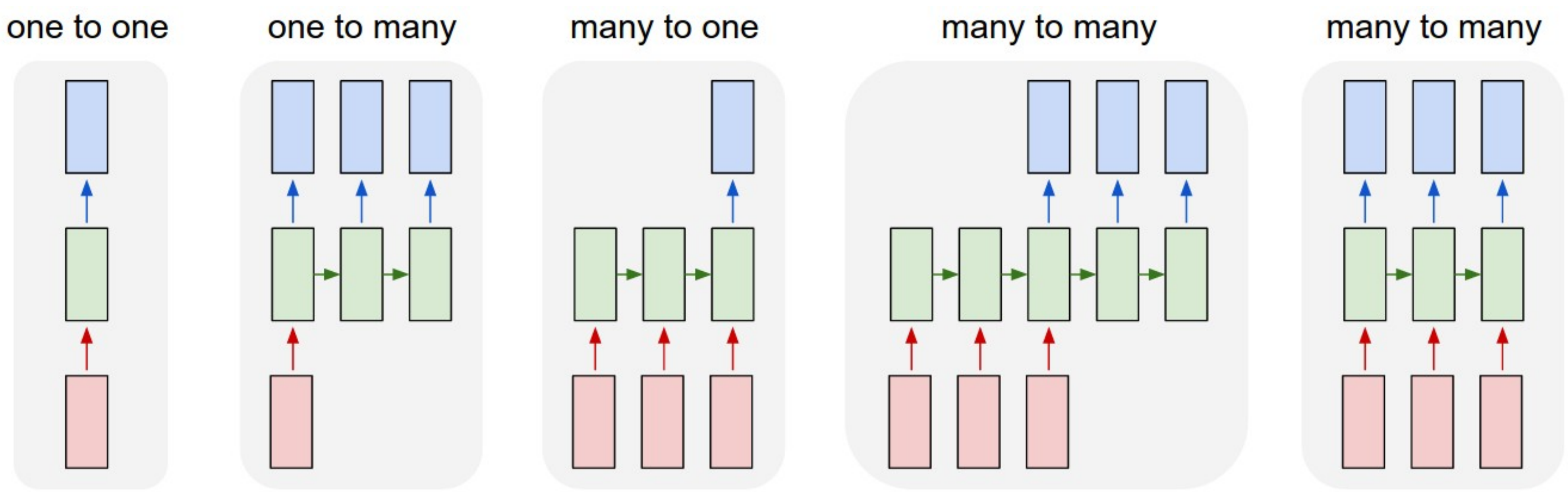
- In time: Speech, DNA, sentences
- In space: Image, video, relations

DNN FOR **VEC2VEC** MAPPING



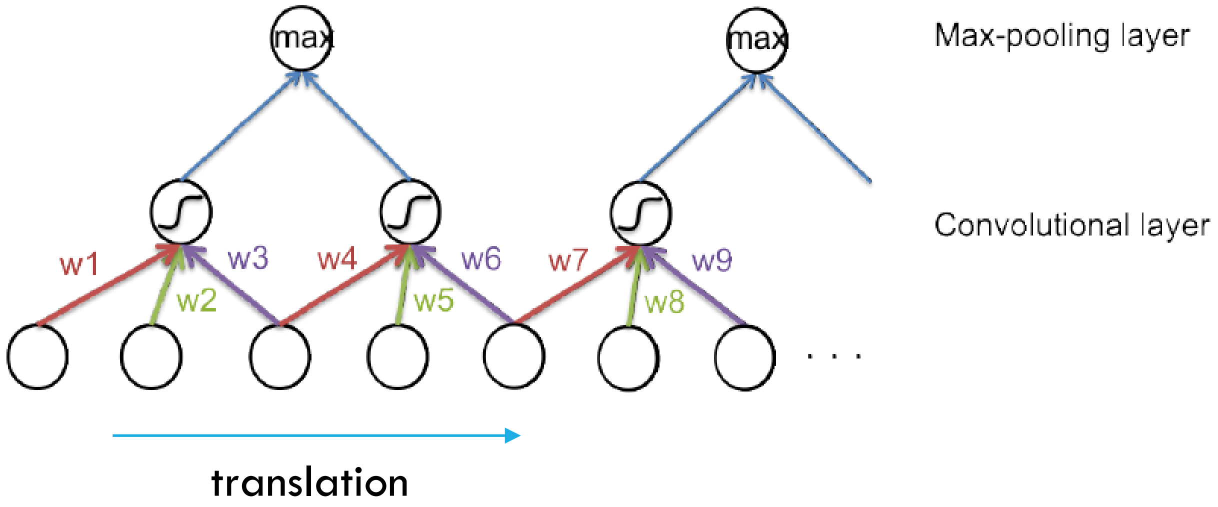
RNN FOR **SEQ2SEQ** MAPPING

deep neural nets with parameter tying

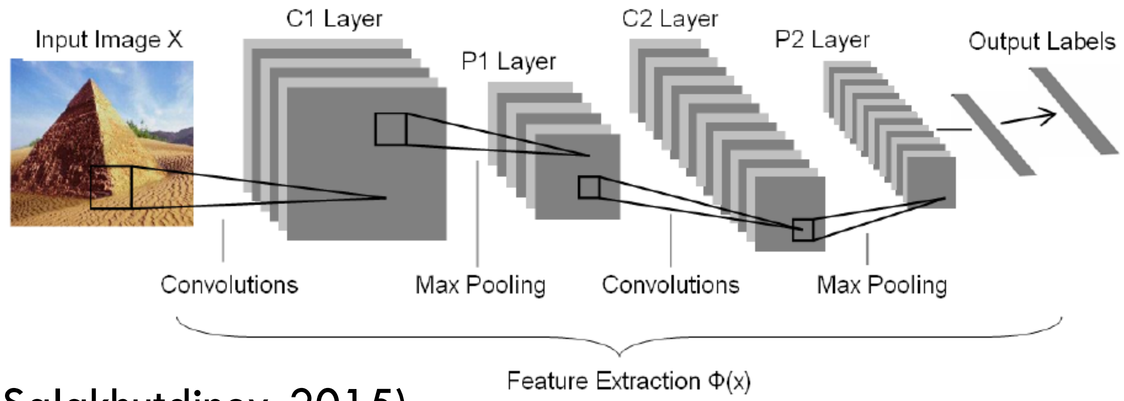


Source: http://karpathy.github.io/assets/rnn/diags.jpeg

CNN FOR TRANSLATION INVARIANCE



(Quoc Le, 2015)



(Russ Salakhutdinov, 2015)

WHEN DOES DEEP LEARNING WORK?

Lots of data (e.g., millions of images)

Strong, clean training signals (e.g., when human can provide correct labels – cognitive domains)

Data structures are well-defined (e.g., image, speech, NLP, video)

Data is compositional (luckily, most data is like this)

The more primitive (raw) the data, the more benefit of using deep learning.

AGENDA

Introduction to PRaDA

Introduction to deep learning

Deep learning for [X], where X (non-cognitive) is:

- Healthcare
- Software engineering
- Choice and ranking
- Anomaly detection
- Multi-relational databases
- Representation

The open room

X = PROSPECTIVE HEALTHCARE

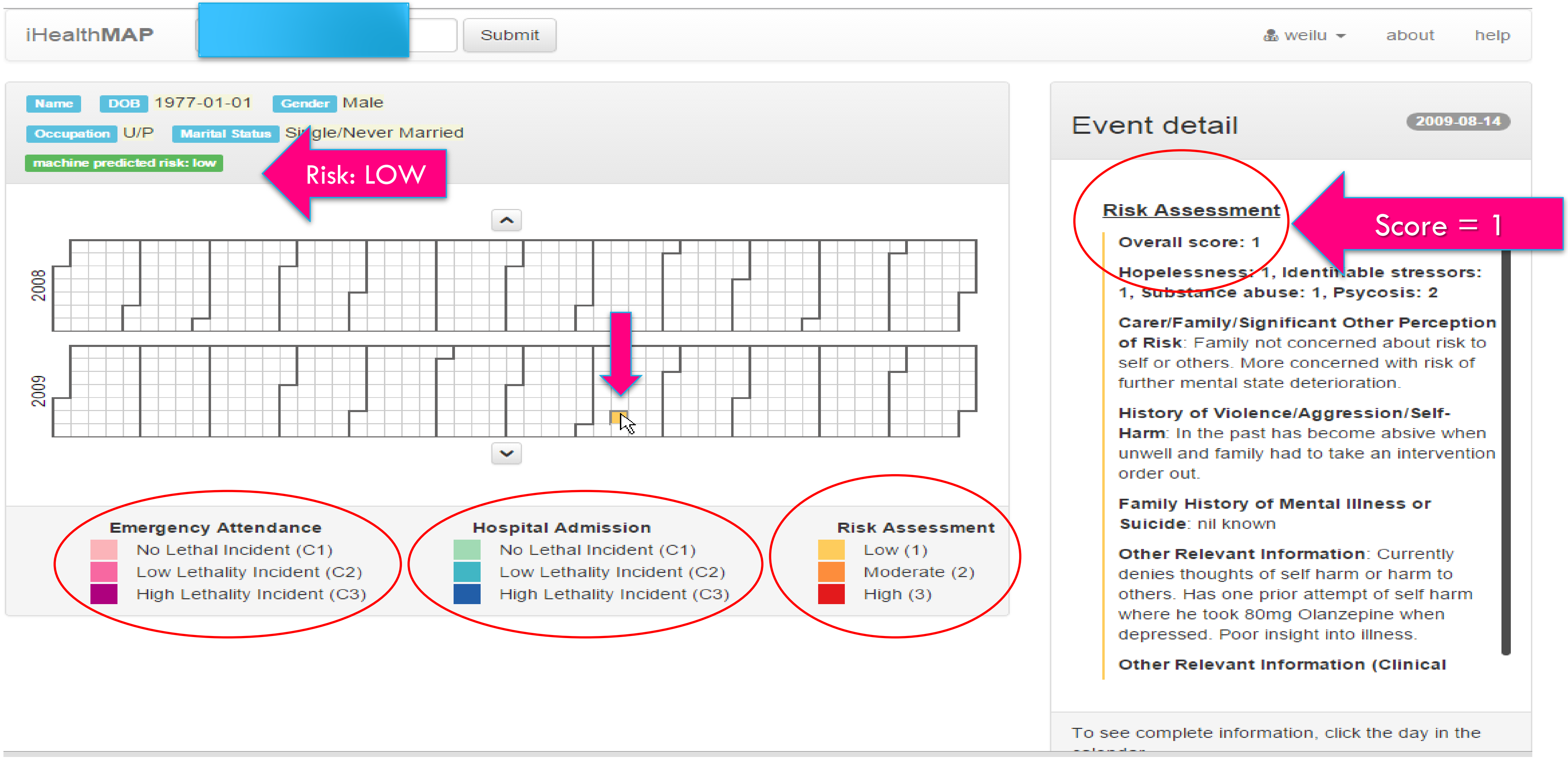
Promises yet to be delivered.

Bottleneck: data – providers not willing to share. Many issues – privacy, ethics, governance.

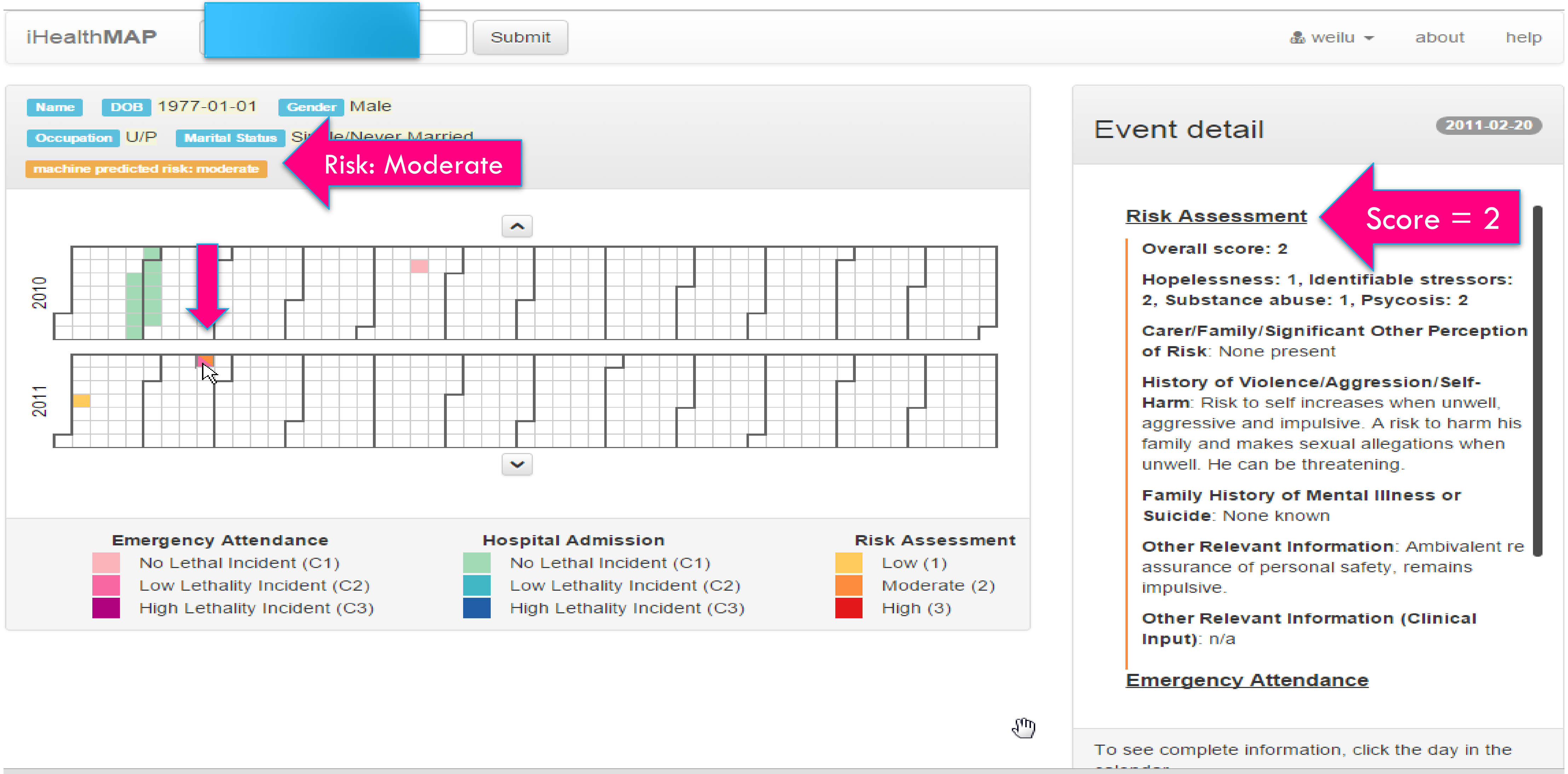
Requirements

- Transparency & interpretability
- Correctly model characteristics of healthcare data (e.g., Irregular timing, Interventions, Regular motifs)

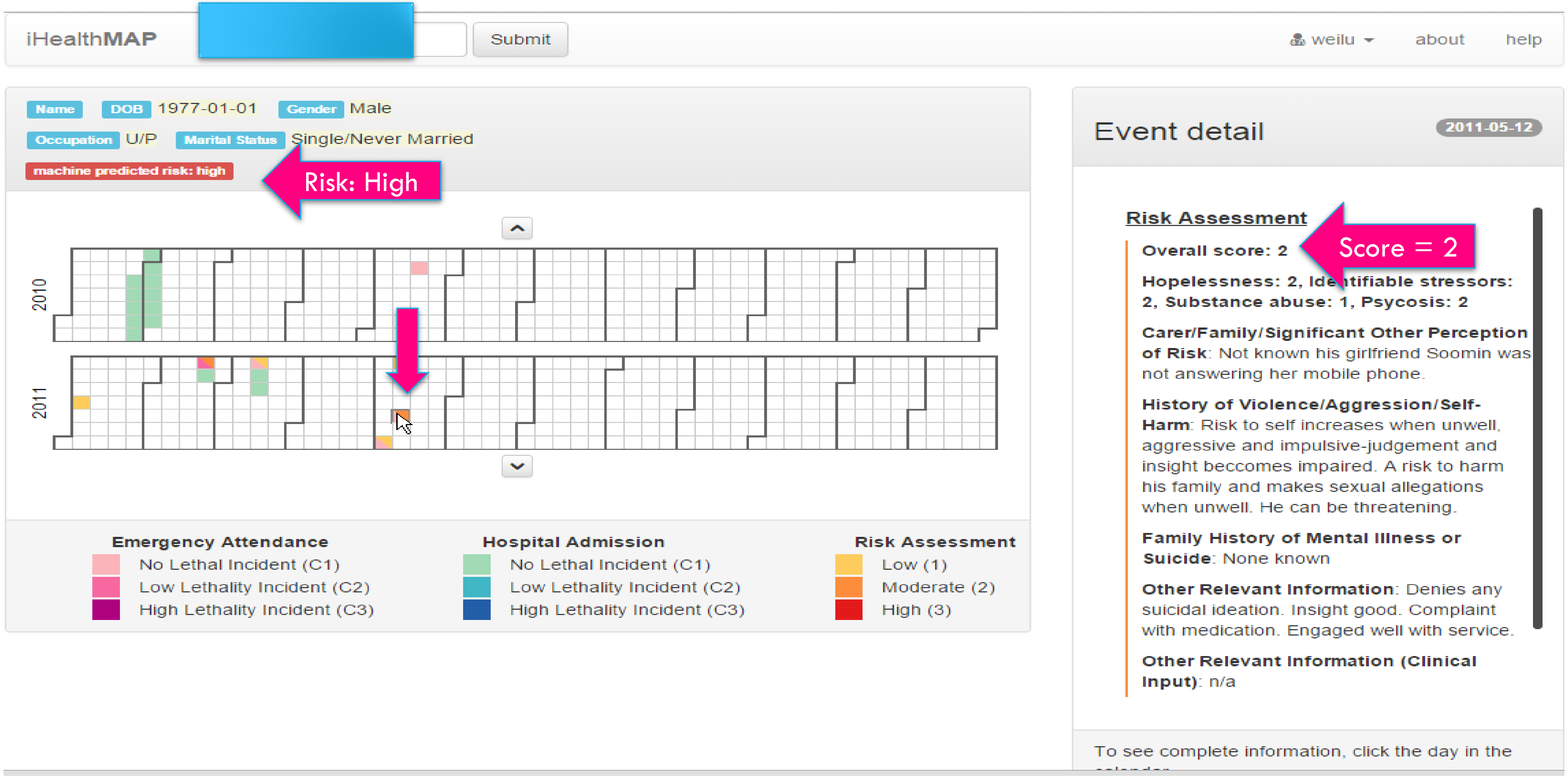
A wide range of modalities: *health processes, EMR, questionaires, imaging, biomarkers, NLP (clinical notes, pubmed, social media), wearable devices, genomics.*

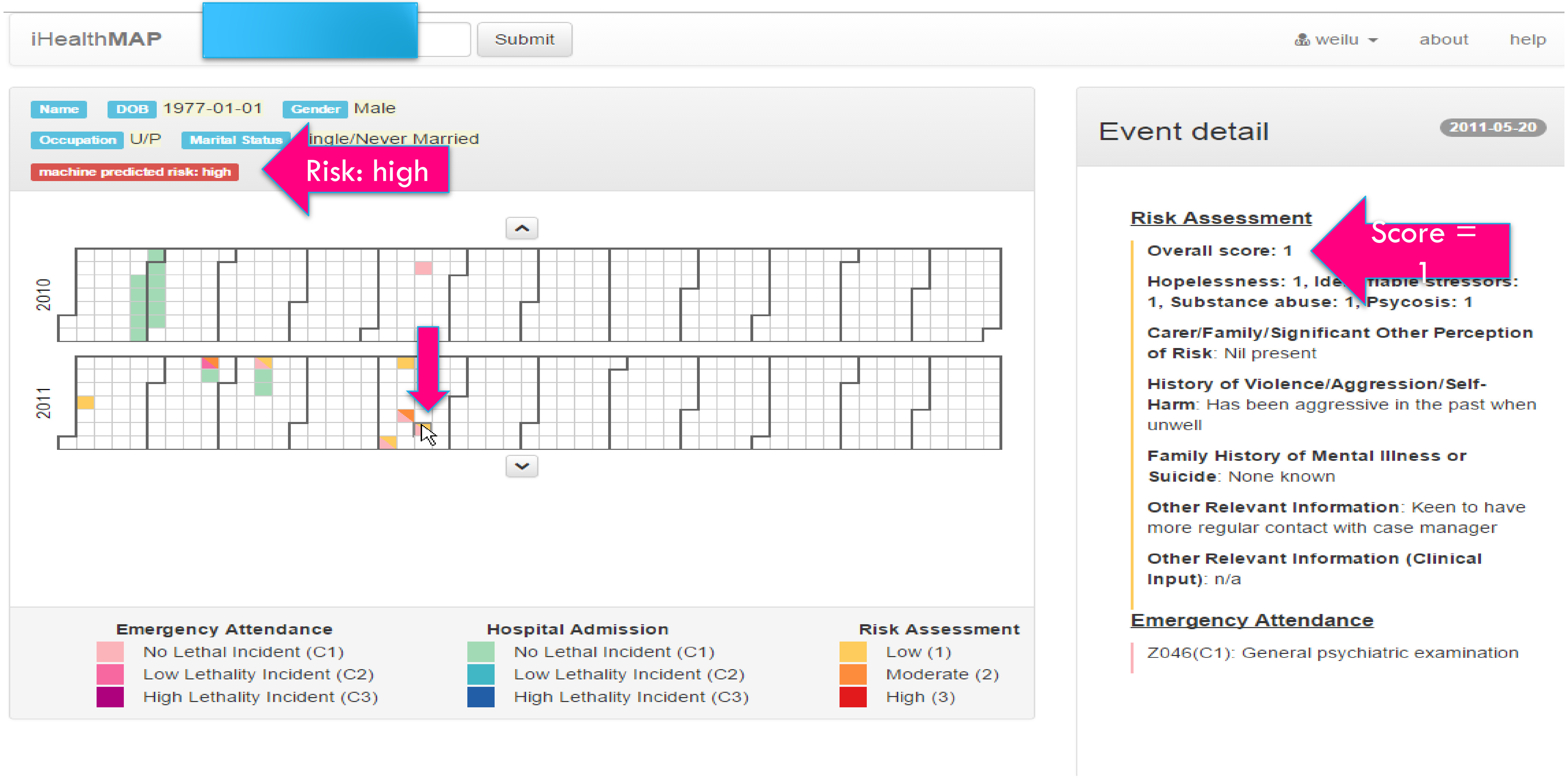


iHealthMAP
Submit
weilu
about
help
Name
DOB 1977-01-01
Gender Male
Occupation U/P
Marital Status Si gle/Never Married
machine predicted risk: low
Risk: LOW
2008
2009
Emergency Attendance
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Hospital Admission
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Risk Assessment
Low (1)
Moderate (2)
High (3)
Event detail
2009-08-14
Risk Assessment
Score = 1
Overall score: 1
Hopelessness: 1, Identifiable stressors: 1, Substance abuse: 1, Psycosis: 2
Carer/Family/Significant Other Perception of Risk: Family not concerned about risk to self or others. More concerned with risk of further mental state deterioration.
History of Violence/Aggression/Self-Harm: In the past has become absive when unwell and family had to take an intervention order out.
Family History of Mental Illness or Suicide: nil known
Other Relevant Information: Currently denies thoughts of self harm or harm to others. Has one prior attempt of self harm where he took 80mg Olanzepine when depressed. Poor insight into illness.
Other Relevant Information (Clinical
To see complete information, click the day in the

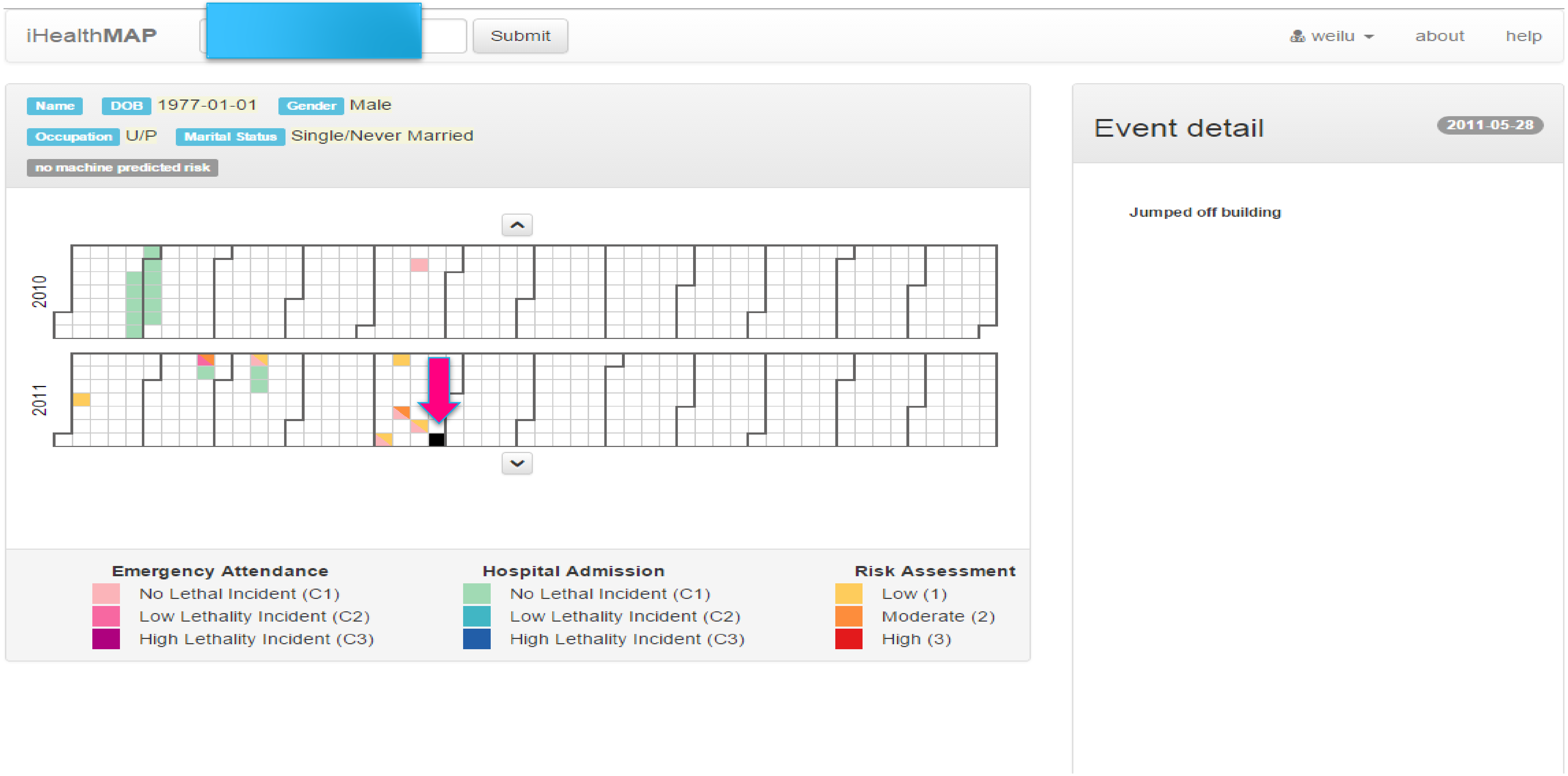


iHealthMAP
Submit
weilu
about
help
Name
DOB 1977-01-01
Gender Male
Occupation U/P
Marital Status Si le/Never Married
machine predicted risk: moderate
Risk: Moderate
2010
2011
Emergency Attendance
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Hospital Admission
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Risk Assessment
Low (1)
Moderate (2)
High (3)
Event detail
2011-02-20
Risk Assessment
Score = 2
Overall score: 2
Hopelessness: 1, Identifiable stressors: 2, Substance abuse: 1, Psycosis: 2
Carer/Family/Significant Other Perception of Risk: None present
History of Violence/Aggression/Self-Harm: Risk to self increases when unwell, aggressive and impulsive. A risk to harm his family and makes sexual allegations when unwell. He can be threatening.
Family History of Mental Illness or Suicide: None known
Other Relevant Information: Ambivalent re assurance of personal safety, remains impulsive.
Other Relevant Information (Clinical Input): n/a
Emergency Attendance
To see complete information, click the day in the


iHealthMAP
Submit
weilu
about
help
Name
DOB 1977-01-01
Gender Male
Occupation U/P
Marital Status Single/Never Married
machine predicted risk: high
Risk: High
2010
2011
Emergency Attendance
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Hospital Admission
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Risk Assessment
Low (1)
Moderate (2)
High (3)
Event detail
2011-05-12
Risk Assessment
Score = 2
Overall score: 2
Hopelessness: 2, Identifiable stressors: 2, Substance abuse: 1, Psycosis: 2
Carer/Family/Significant Other Perception of Risk: Not known his girlfriend Soomin was not answering her mobile phone.
History of Violence/Aggression/Self-Harm: Risk to self increases when unwell, aggressive and impulsive-judgement and insight beccomes impaired. A risk to harm his family and makes sexual allegations when unwell. He can be threatening.
Family History of Mental Illness or Suicide: None known
Other Relevant Information: Denies any suicidal ideation. Insight good. Complaint with medication. Engaged well with service.
Other Relevant Information (Clinical Input): n/a
To see complete information, click the day in the



iHealthMAP
Submit
weilu
about
help
Name
DOB 1977-01-01
Gender Male
Occupation U/P
Marital Status
ingle/Never Married
machine predicted risk: high
Risk: high
2010
2011
Emergency Attendance
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Hospital Admission
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Risk Assessment
Low (1)
Moderate (2)
High (3)
Event detail
2011-05-20
Risk Assessment
Score = 1
Overall score: 1
Hopelessness: 1, Ide tiable stressors: 1, Substance abuse: 1, Psycosis: 1
Carer/Family/Significant Other Perception of Risk: Nil present
History of Violence/Aggression/Self-Harm: Has been aggressive in the past when unwell
Family History of Mental Illness or Suicide: None known
Other Relevant Information: Keen to have more regular contact with case manager
Other Relevant Information (Clinical Input): n/a
Emergency Attendance
Z046(C1): General psychiatric examination


iHealthMAP
Submit
weilu
about
help
Name
DOB 1977-01-01
Gender Male
Occupation U/P
Marital Status Single/Never Married
no machine predicted risk
2010
2011
Event detail
2011-05-28
Jumped off building
Emergency Attendance
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Hospital Admission
No Lethal Incident (C1)
Low Lethality Incident (C2)
High Lethality Incident (C3)
Risk Assessment
Low (1)
Moderate (2)
High (3)

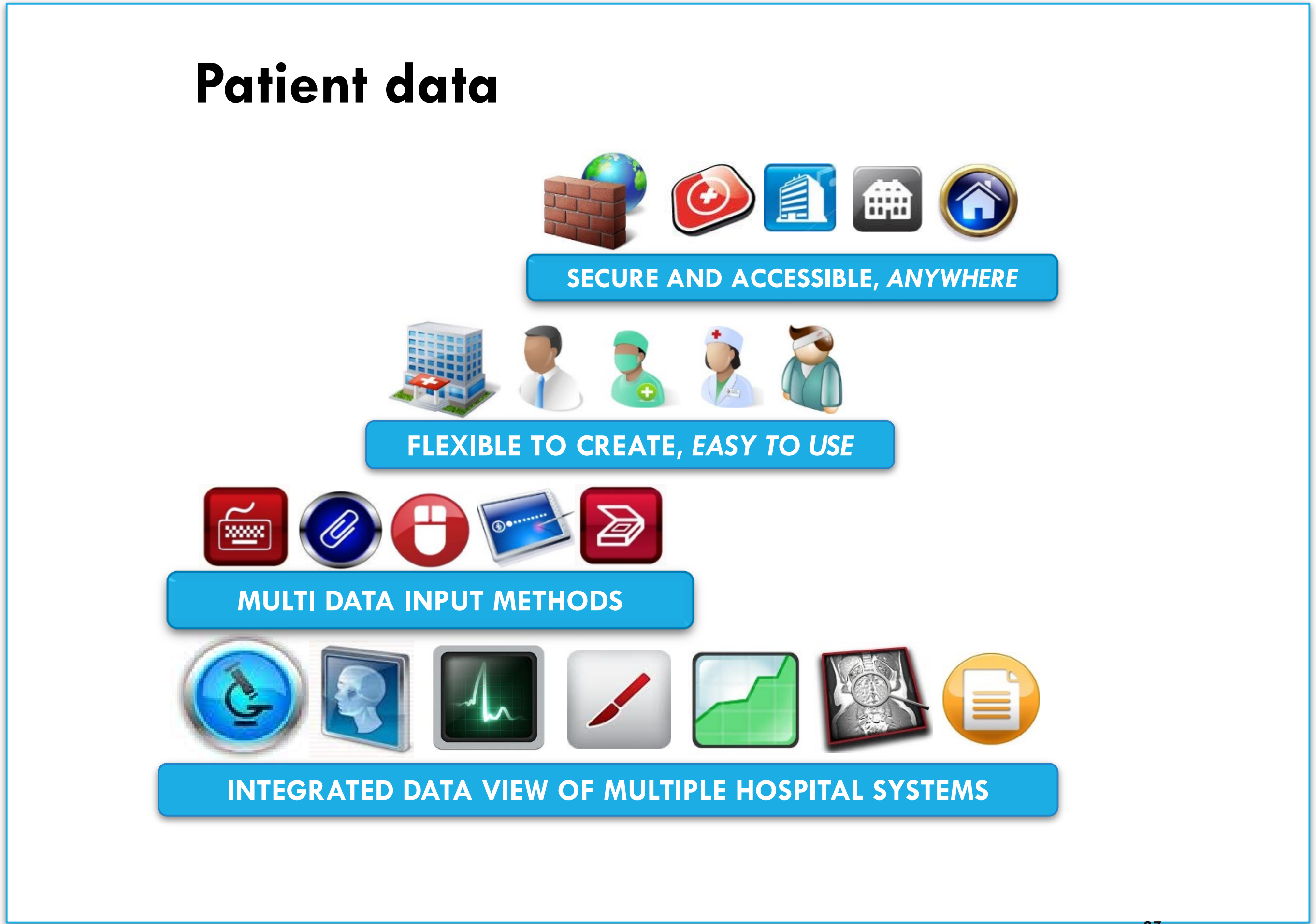
DEEP ARCHITECTURES FOR HEALTHCARE

Our primary goal: predicting future risk!

DNND – vector input & vector output, no sequences

DEEPR (CNN) – repeated motifs, short sequences

DEEPCARE (RNN) – long-term dependencies, long sequences

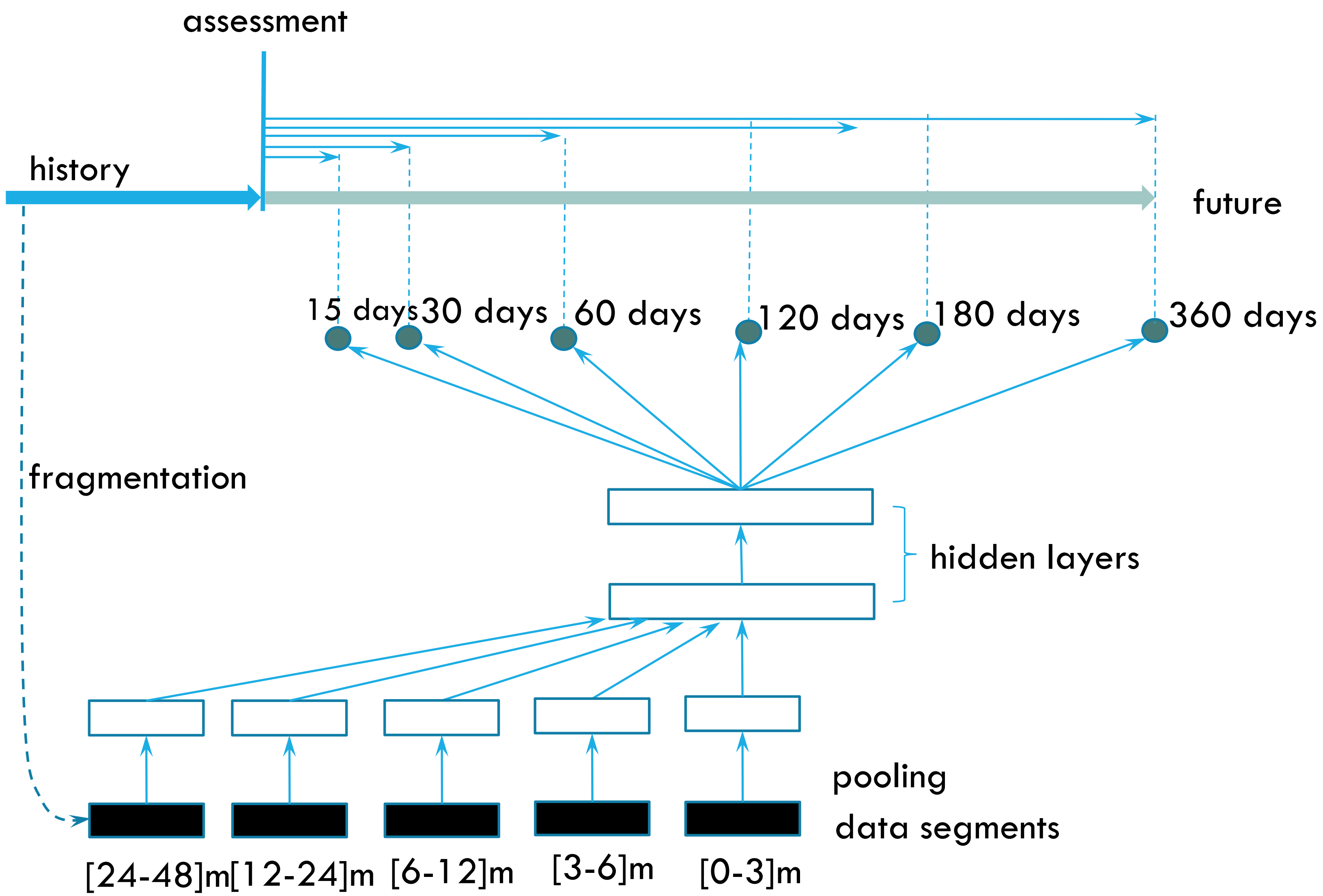


DNND

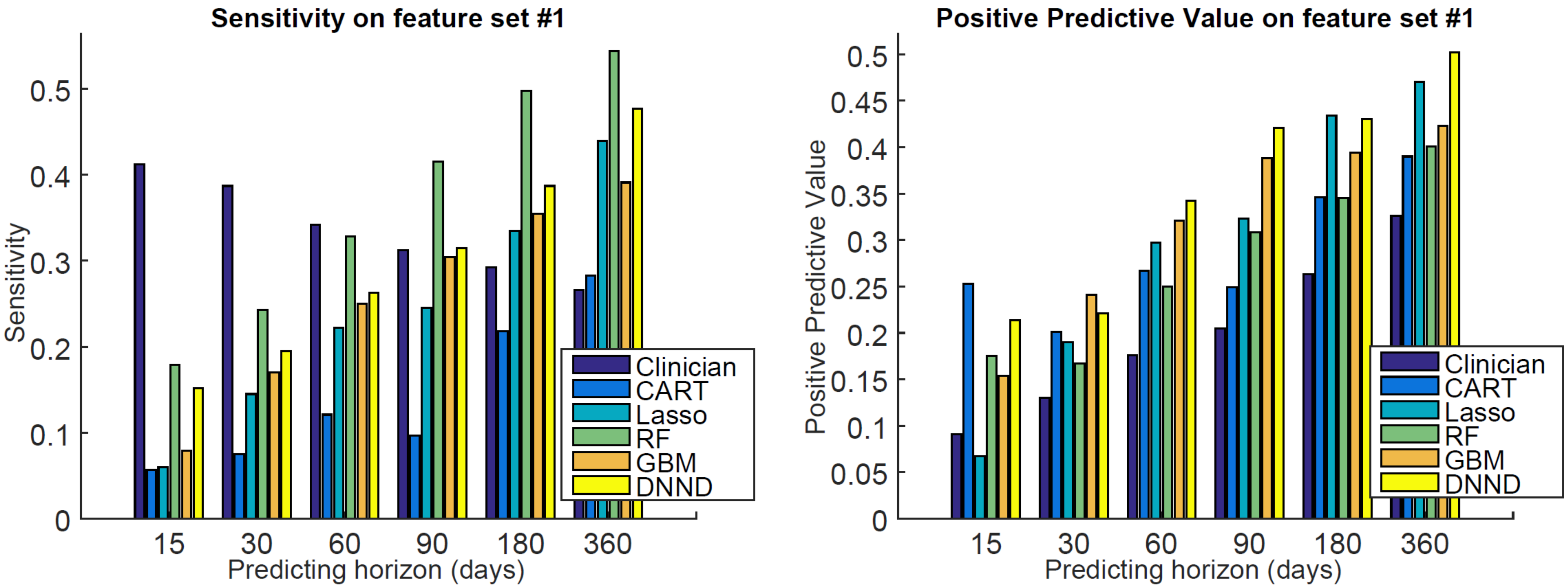
Input vector -- redundancy

Output vector – multiple prediction horizons

Tricks: Dropout



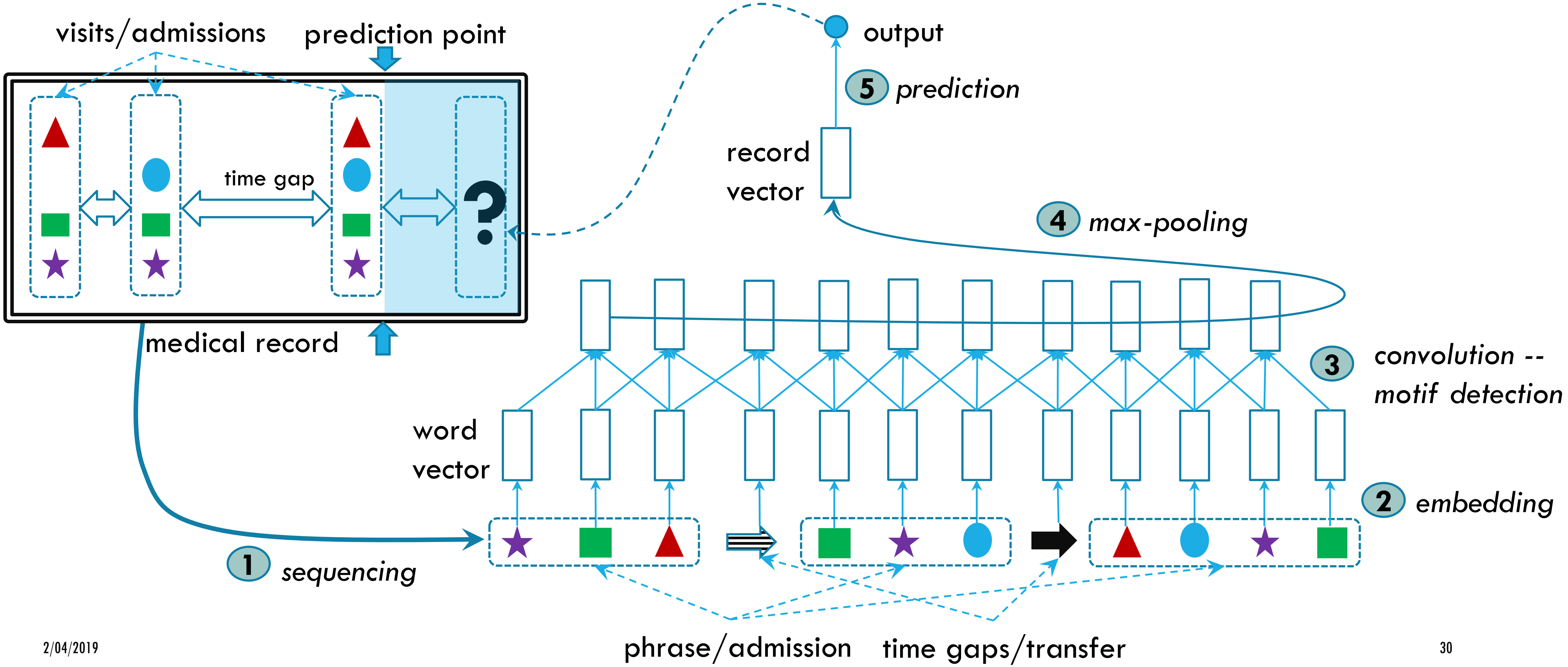
SUICIDE RISK PREDICTION: MACHINE VERSUS CLINICIAN



(a) Recall

(b) Precision

DEEPR: CNN FOR REPEATED MOTIFS AND SHORT SEQUENCES



DISEASE EMBEDDING & MOTIFS DETECTION

E11 I48 I50

Type 2 diabetes mellitus

Atrial fibrillation and flutter

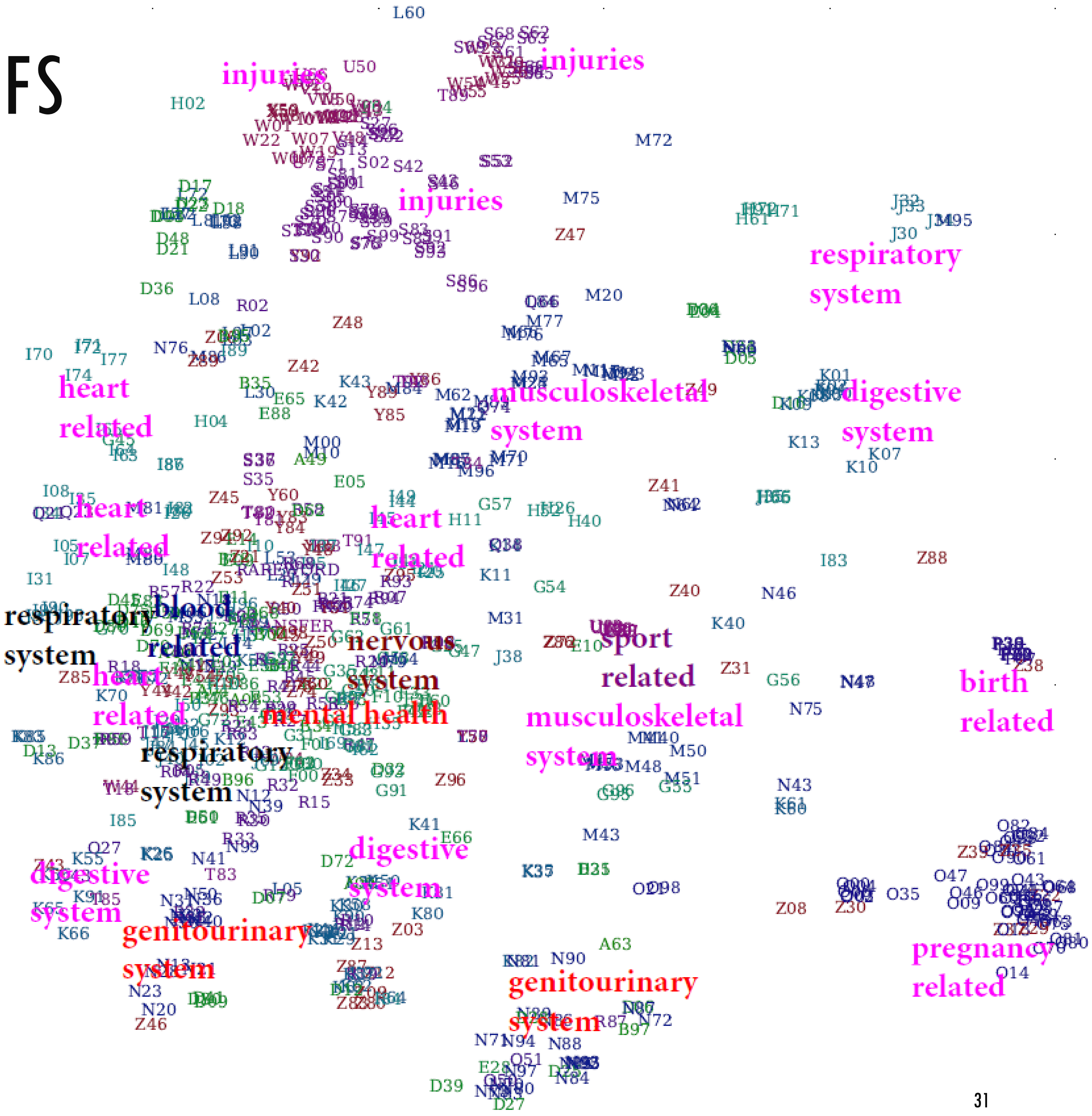
Heart failure

E11 I50 N17

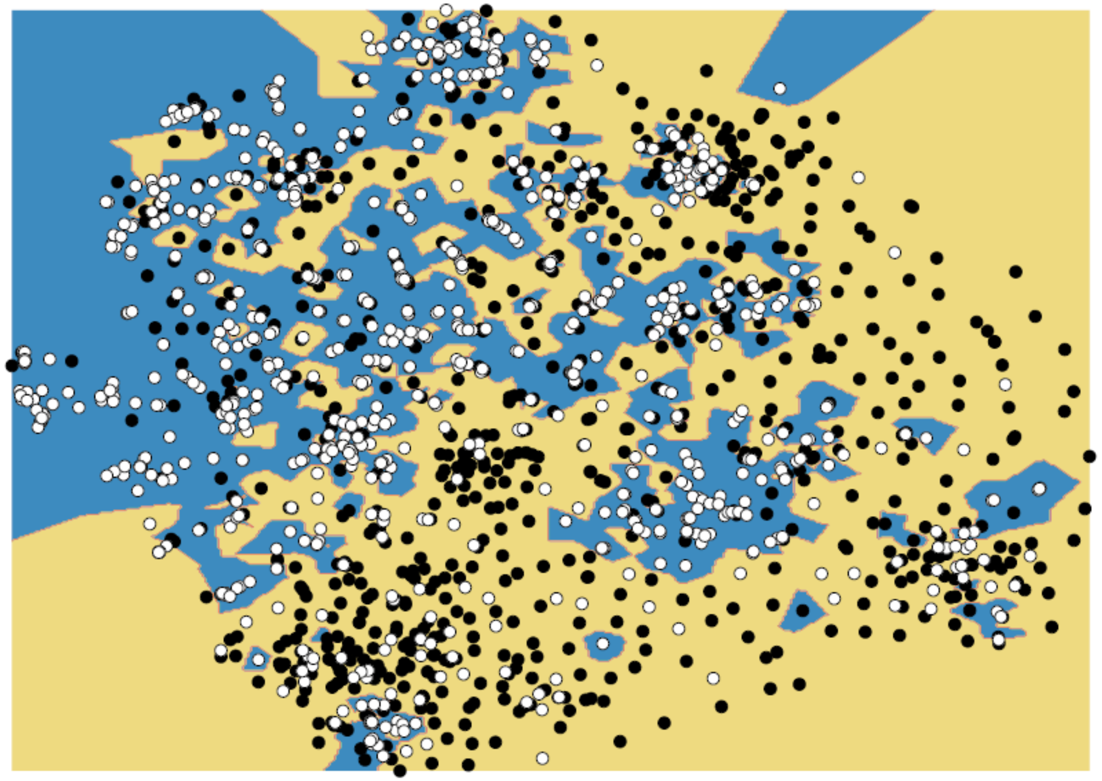
Type 2 diabetes mellitus

Heart failure

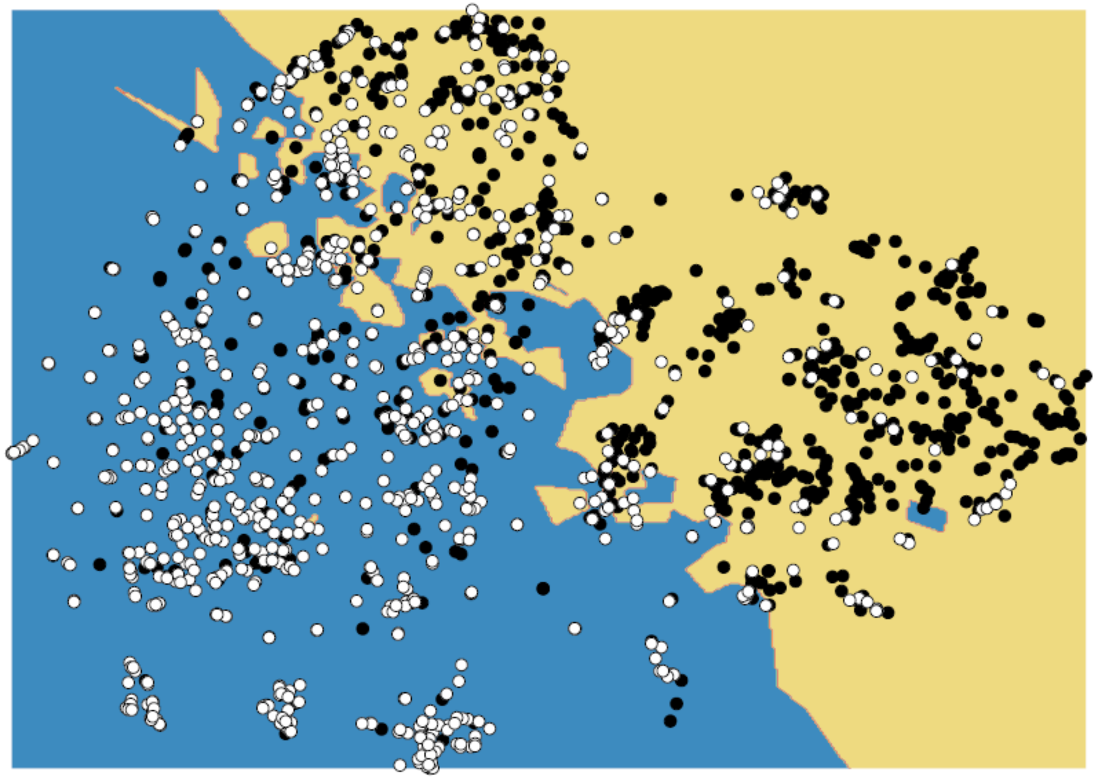
Acute kidney failure



EMBEDDING OF PATIENTS: LINEARIZING DECISION BOUNDARY



BOW+LR



CNN

DEEPCARE: LONG-TERM MEMORY

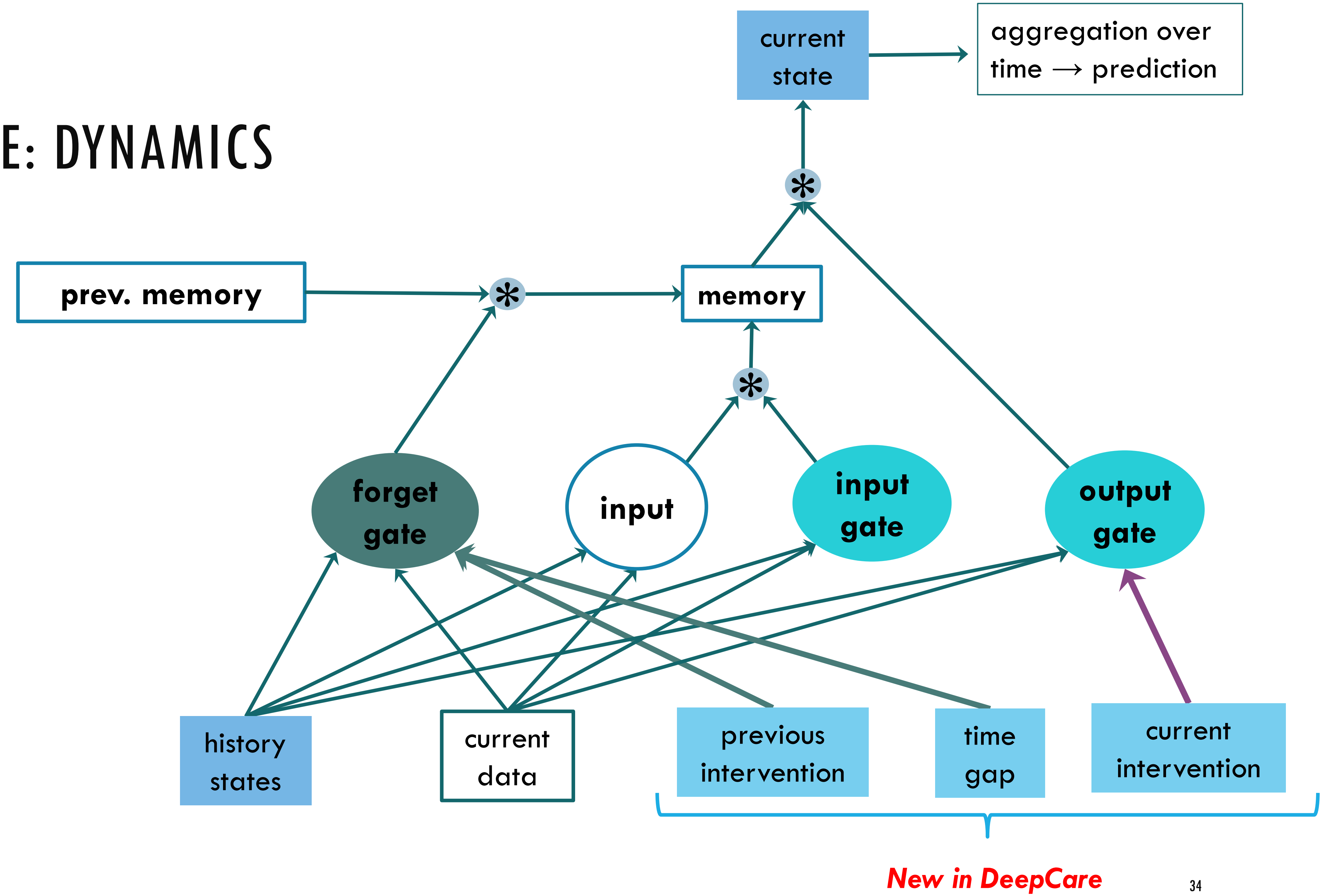
Illness states are a dynamic memory process → moderated by time and intervention

Discrete admission, diagnosis and procedure → vector embedding

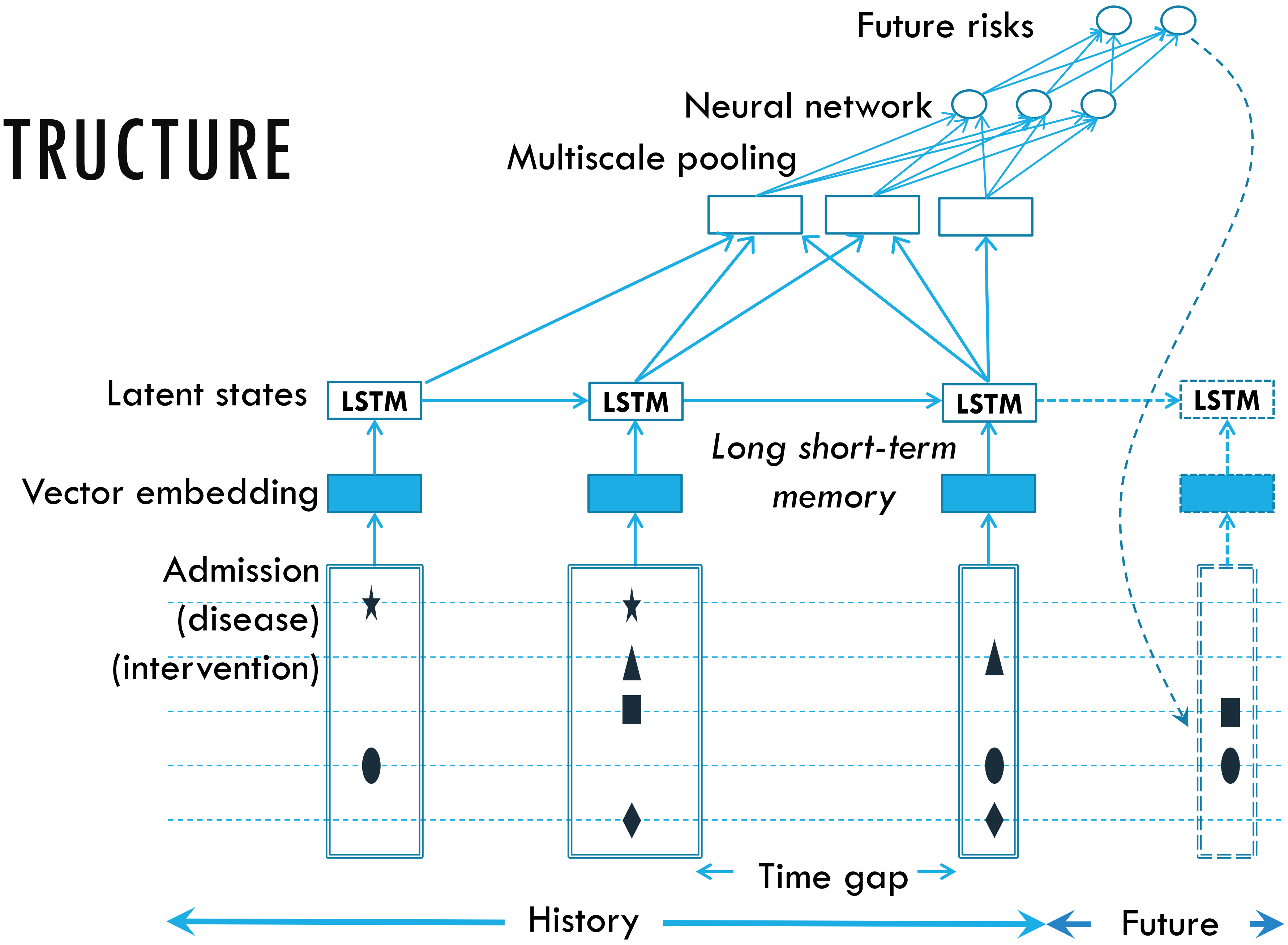
Time and previous intervention → “forgetting” of illness

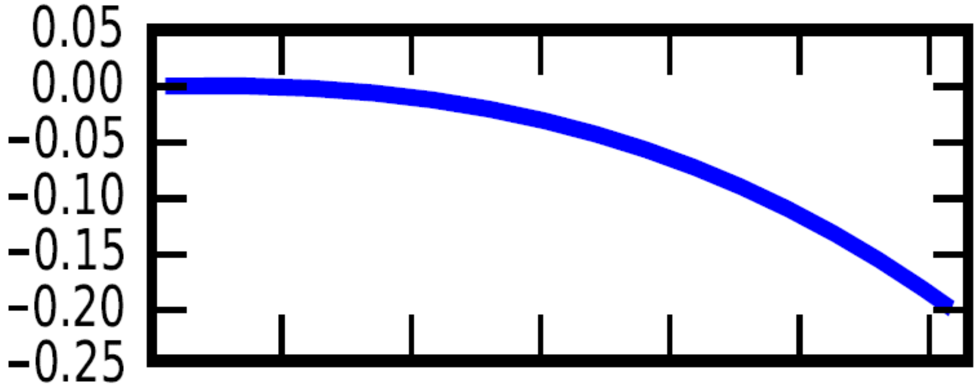
Current intervention → controlling the risk states

DEEPCARE: DYNAMICS



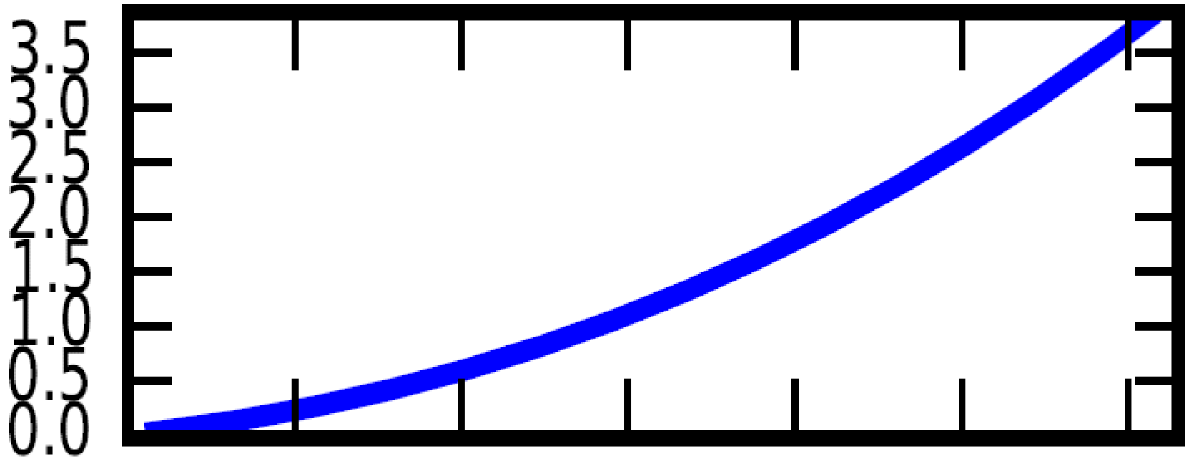
DEEPCARE: STRUCTURE





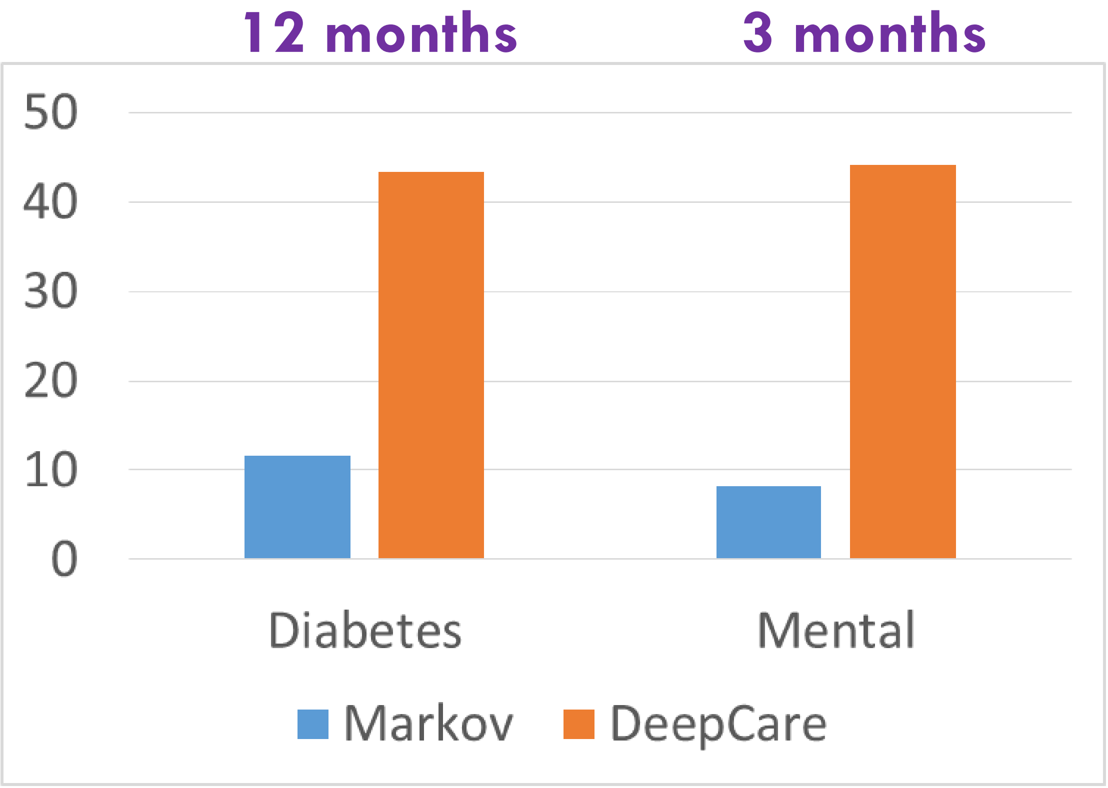
→ decreasing illness

→ Increasing illness

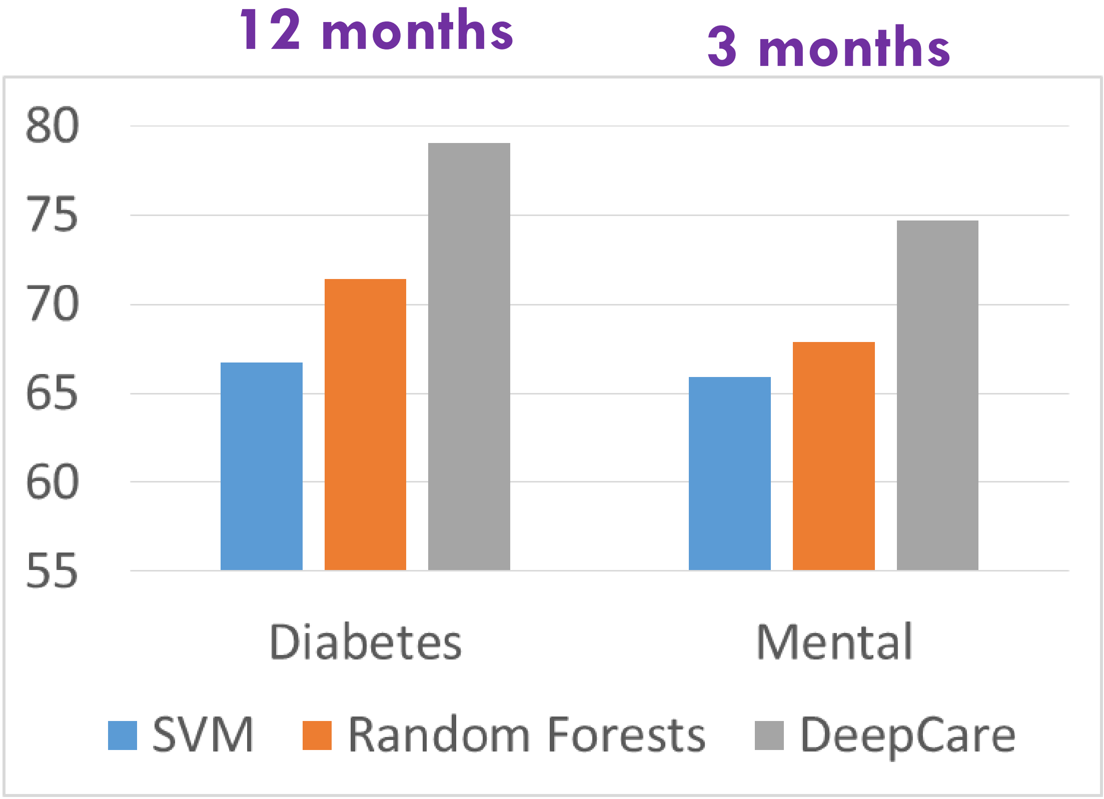


DEEPCARE: TWO MODES OF FORGETTING AS A FUNCTION OF TIME

DEEPCARE: PREDICTION RESULTS



Intervention recommendation (precision@3)



Unplanned readmission prediction (F-score)

AGENDA

Introduction to PRaDA

Introduction to deep learning

Deep learning for [X], where X (non-cognitive) is:

- Healthcare
- **Software engineering**
- Choice and ranking
- Anomaly detection
- Multi-relational databases
- Representation

The open room

X = SOFTWARE ANALYTICS

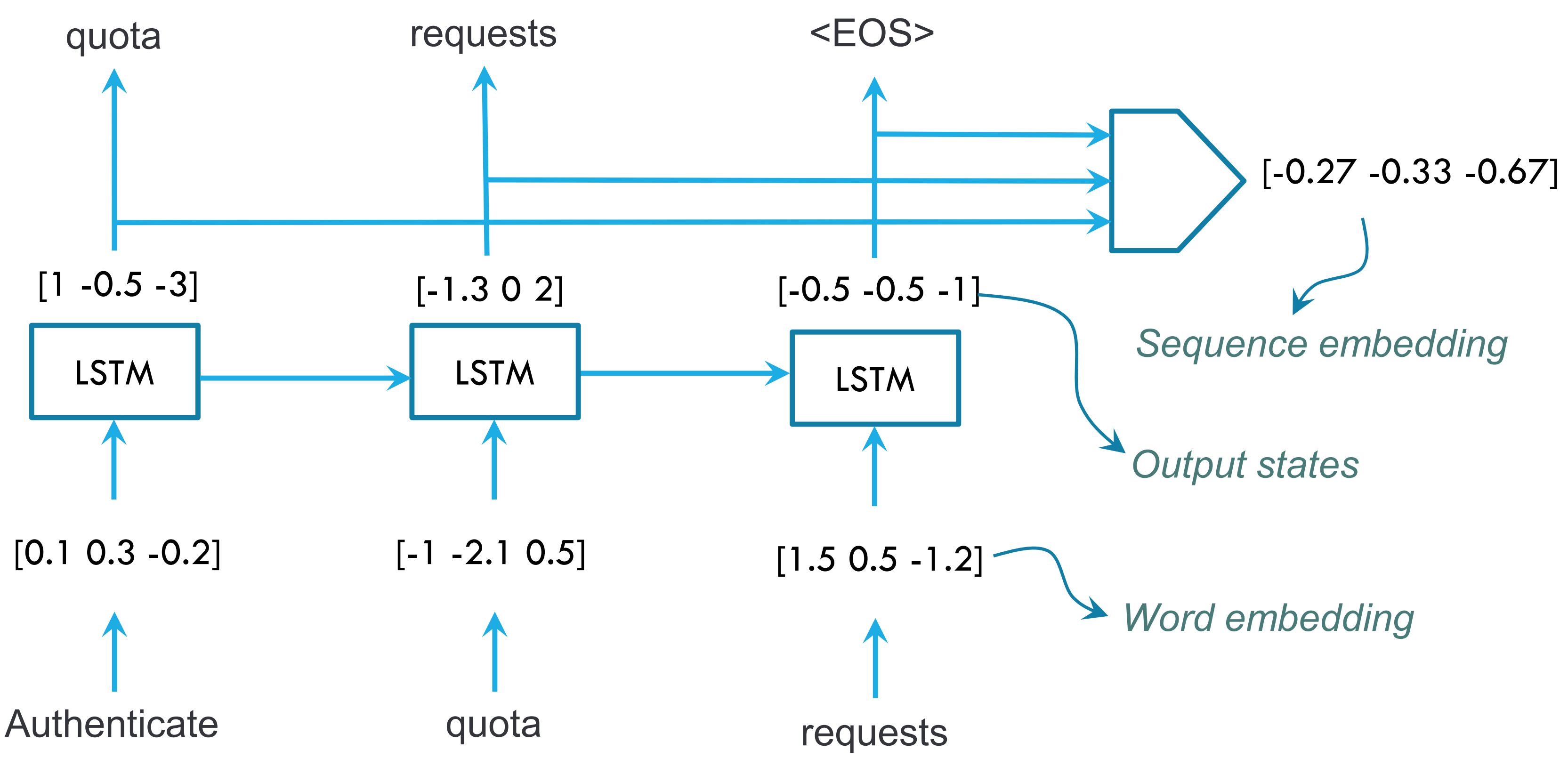
Goal: To model code, text, team, user, execution, project & enabled business process → answer any queries by developers, managers, users and business

- End-to-end
- Compositional

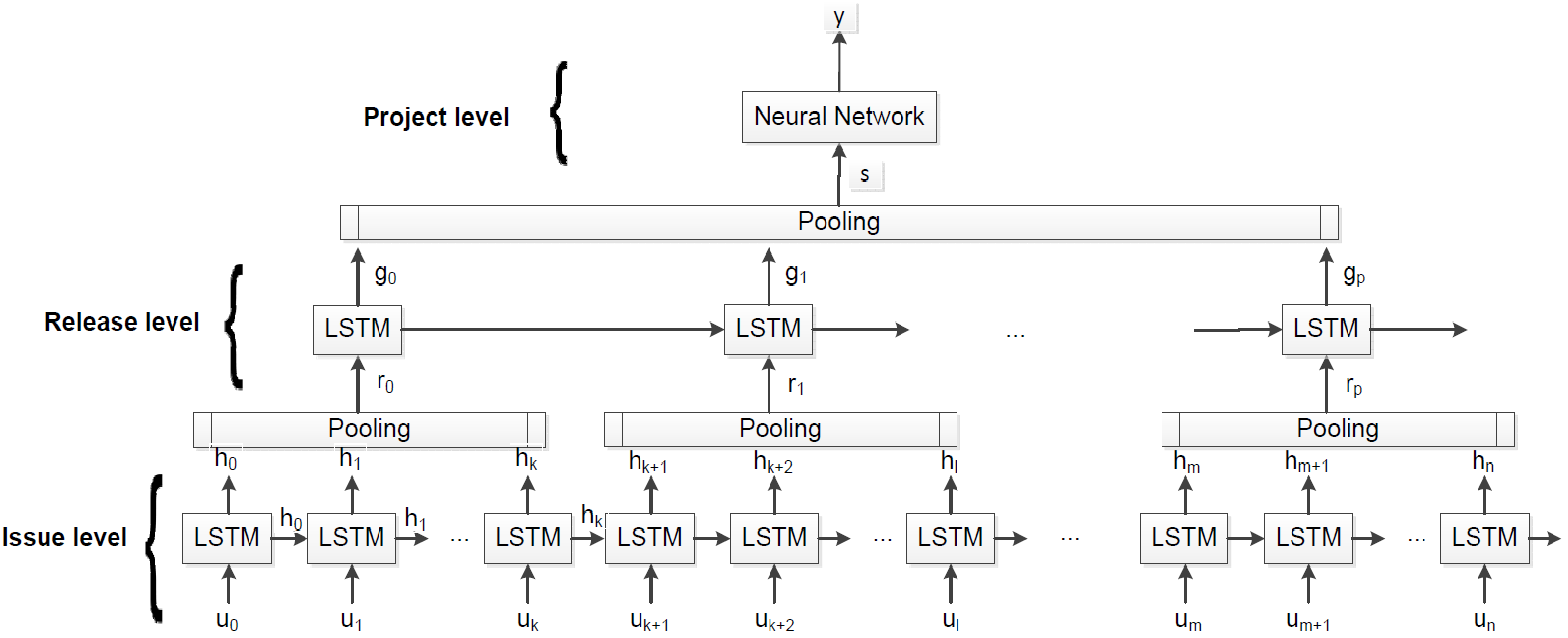
For now:

- LSTM for report representation
- DeepSoft vision paper
- Stacked/deep inference (later)

LONG SHORT-TERM MEMORY FOR TEXT REPRESENTATION



DEEPSOFT: COMPOSITIONAL DEEP NET FOR SW PROJECT



AGENDA

Introduction to PRaDA

Introduction to deep learning

Deep learning for [X], where X (non-cognitive) is:

- Healthcare
- Software engineering
- **Choice and ranking**
- Anomaly detection
- Multi-relational databases
- Representation

The open room

X = RANKING

- Raking web documents in search engines
- Movie recommendation
- Advertisement placement
- Tag recommendation
- Expert finding in a community network
- Friend ranking in a social network
- ???

*(these graphics were downloaded using image.google.com)

LEARNING-TO-RANK

- Web documents in search engines
  - query: *keywords*
- Movie recommendation
  - query: *an user*
- Advertisement placement
  - query: *a Web page*
- Tag recommendation
  - query: *a web object*
- Friend ranking in a social network
  - query: *an user*

Learn to rank responses to a query

A ML approach to Information Retrieval

- Instead of hand-engineering similarity measures, learn it

Two key elements

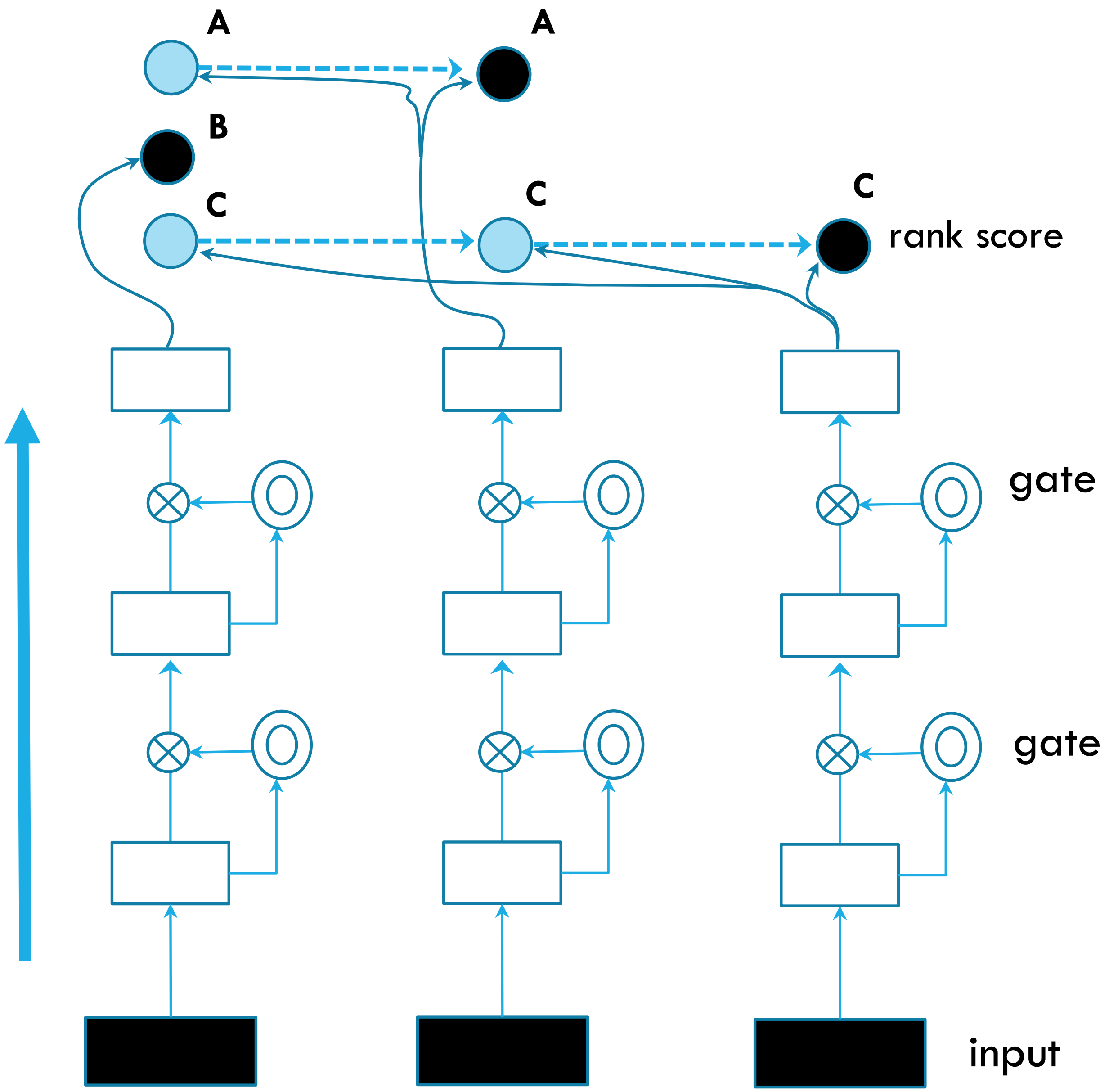
- Choice model → rank loss (how right/wrong is a ranked list?)
- Scoring function → mapping features into score (how good is the choice?)

CHOICE BY ELIMINATION

The networks represent the scoring function

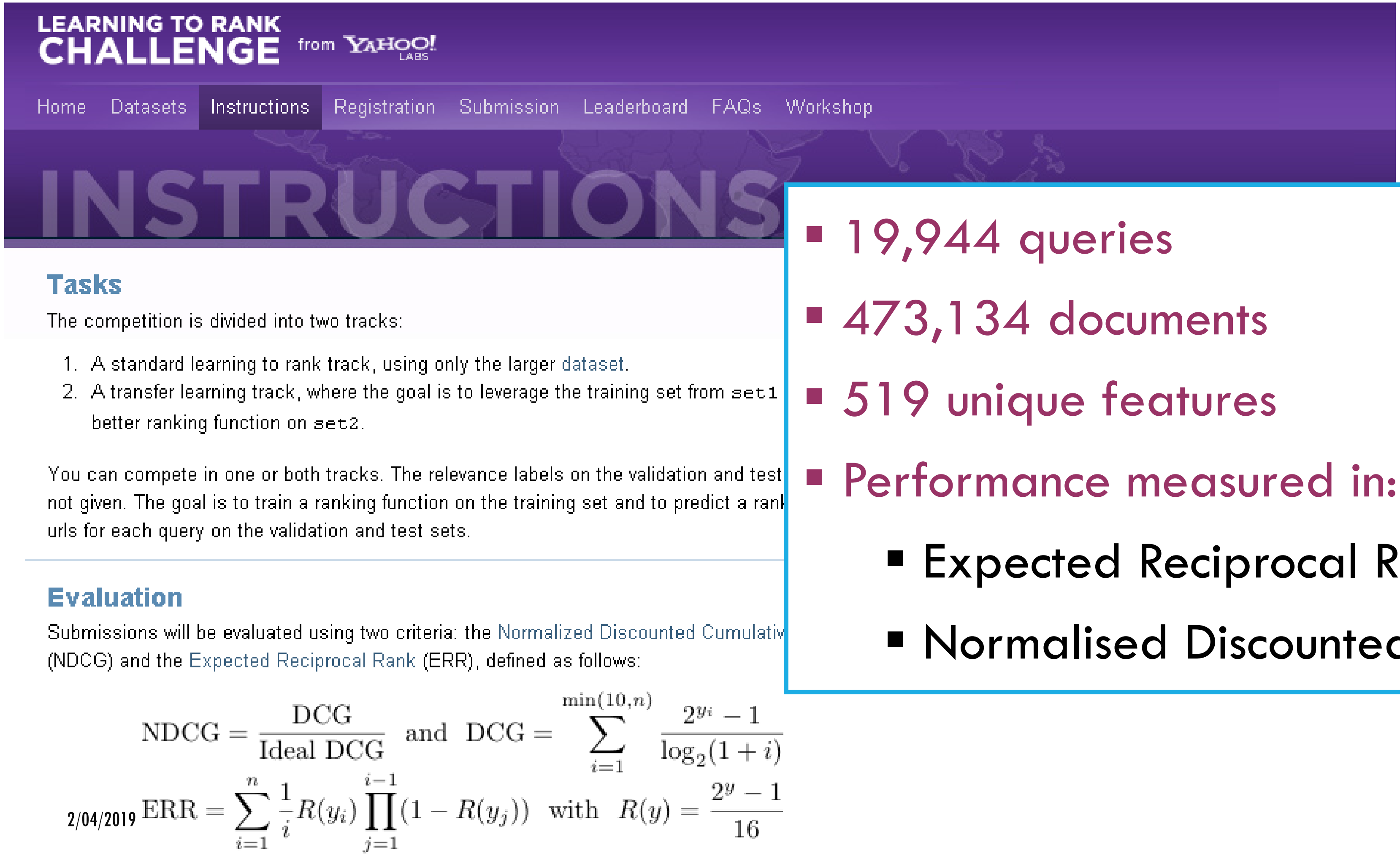
All networks are linked through the rank loss – neural choice by elimination

It is a structured output problem (permutation)



Recurrent highway networks

YAHOO! L2R CHALLENGE (2010)



LEARNING TO RANK CHALLENGE from YAHOO! LABS

Home Datasets Instructions Registration Submission Leaderboard FAQs Workshop

INSTRUCTIONS

Tasks

The competition is divided into two tracks:

1. A standard learning to rank track, using only the larger dataset.
2. A transfer learning track, where the goal is to leverage the training set from `set1` ... better ranking function on `set2`.

You can compete in one or both tracks. The relevance labels on the validation and test ... not given. The goal is to train a ranking function on the training set and to predict a rank... urls for each query on the validation and test sets.

Evaluation

Submissions will be evaluated using two criteria: the Normalized Discounted Cumulativ... (NDCG) and the Expected Reciprocal Rank (ERR), defined as follows:

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{Ideal\ DCG}} \quad \text{and} \quad \mathrm{DCG} = \sum_{i=1}^{\min(10,n)} \frac{2^{y_i} - 1}{\log_2(1+i)}$$

$$\mathrm{ERR} = \sum_{i=1}^{n} \frac{1}{i} R(y_i) \prod_{j=1}^{i-1} (1 - R(y_j)) \quad \text{with} \quad R(y) = \frac{2^y - 1}{16}$$

- 19,944 queries
- 473,134 documents
- 519 unique features
- Performance measured in:
  - Expected Reciprocal Rank (ERR)
  - Normalised Discounted Cumulative Gain (NDCG)

RESULTS

As of 2011 – Forward selection + quadratic rank function

	ERR	NDCG@1	NDCG@5
Rank Regress	0.4882	0.683	0.6672
RankNet	0.4919	0.6903	0.6698
Ranking SVM	0.4868	0.6797	0.6662
ListMLE	0.4955	0.6993	0.6705
PairTies-D	0.4941	0.6944	0.6725
PairTies-RK	0.4946	0.6970	0.6716
PMOP-FD	0.5038	0.7137	0.6762
PMOP-Gibbs	0.5037	0.7105	**0.6792**
PMOP-MH	**0.5045**	**0.7139**	0.6790

Rank 41 out of 1500

As of 2016 – Backward elimination + deep nets

	Placket-Luce			Choice by elimination		
Rank function	ERR	NDCG@1	NDCG@5	ERR	NDCG@1	NDCG@5
SGTB	0.497	0.697	0.673	0.506	0.705	0.681
Neural nets	**0.501**	**0.705**	**0.688**	**0.509**	**0.719**	**0.697**

Rank?

AGENDA

Introduction to PRaDA

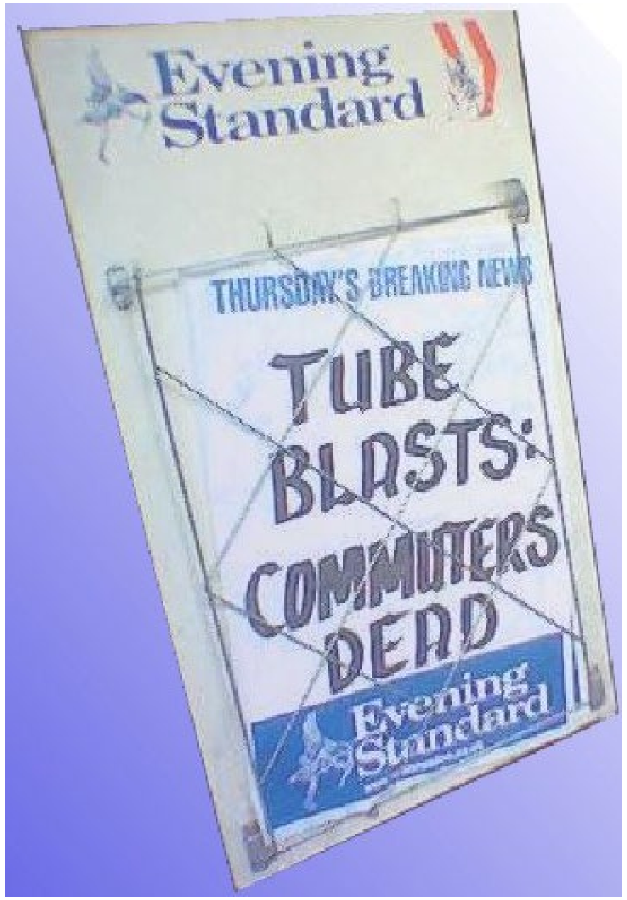
Introduction to deep learning

Deep learning for [X], where X =

- Healthcare
- Software engineering
- Choice and ranking
- **Anomaly detection**
- Multi-relational databases
- Representation

The open room

X = ANOMALY DETECTION

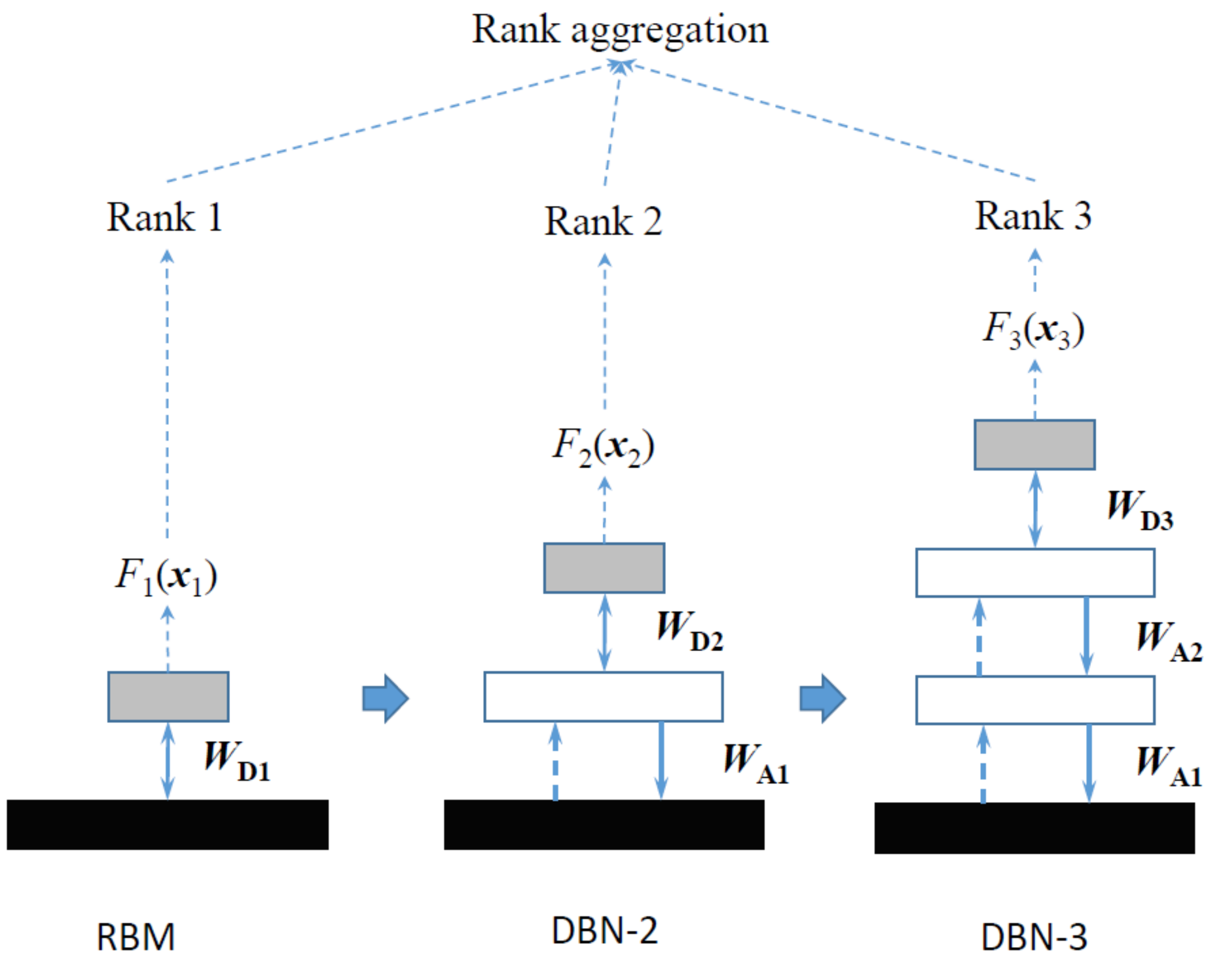


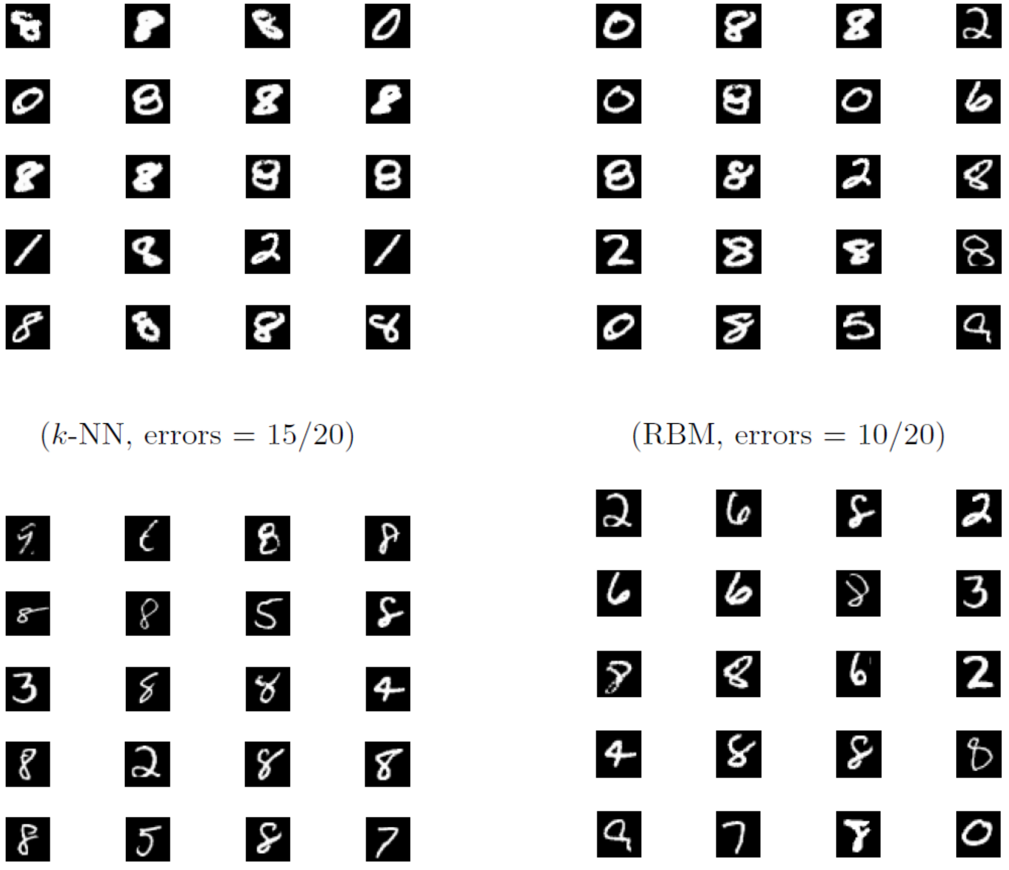
London, July 7, 2005

Real world - what are the operators monitoring?

Strategy: learn normality, anything does not fit in is abnormal

MAD: MULTILEVEL ANOMALY DETECTION





(k-NN, errors = 15/20) (RBM, errors = 10/20)

(DBN-L2, errors = 12/20) (MAD-L2p2, errors = 8/20)

AGENDA

Introduction to PRaDA

Introduction to deep learning

Deep learning for [X], where X =

- Healthcare
- Software engineering
- Choice and ranking
- Anomaly detection
- **Multi-relational databases**
- Representation

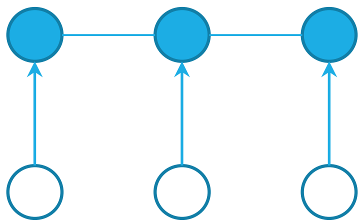
The open room

X =MULTI-RELATIONAL DATABASES

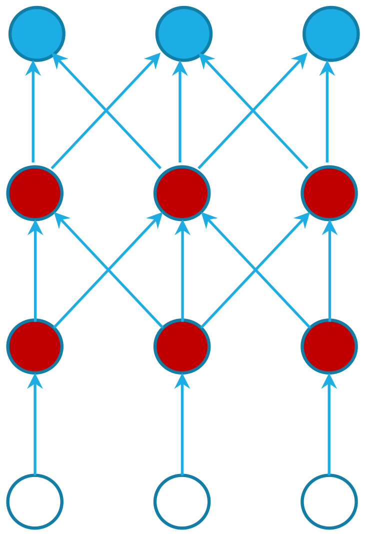
The world is multi-relational (e.g., friend, class-mate, collaborator, flat-mate).

Stacked inference (with Hoa & Morakot)

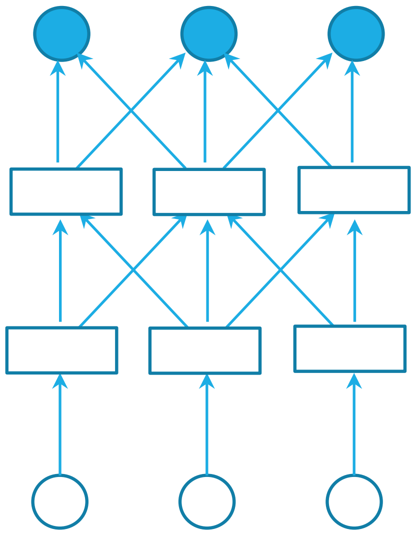
Deep inference



Shallow

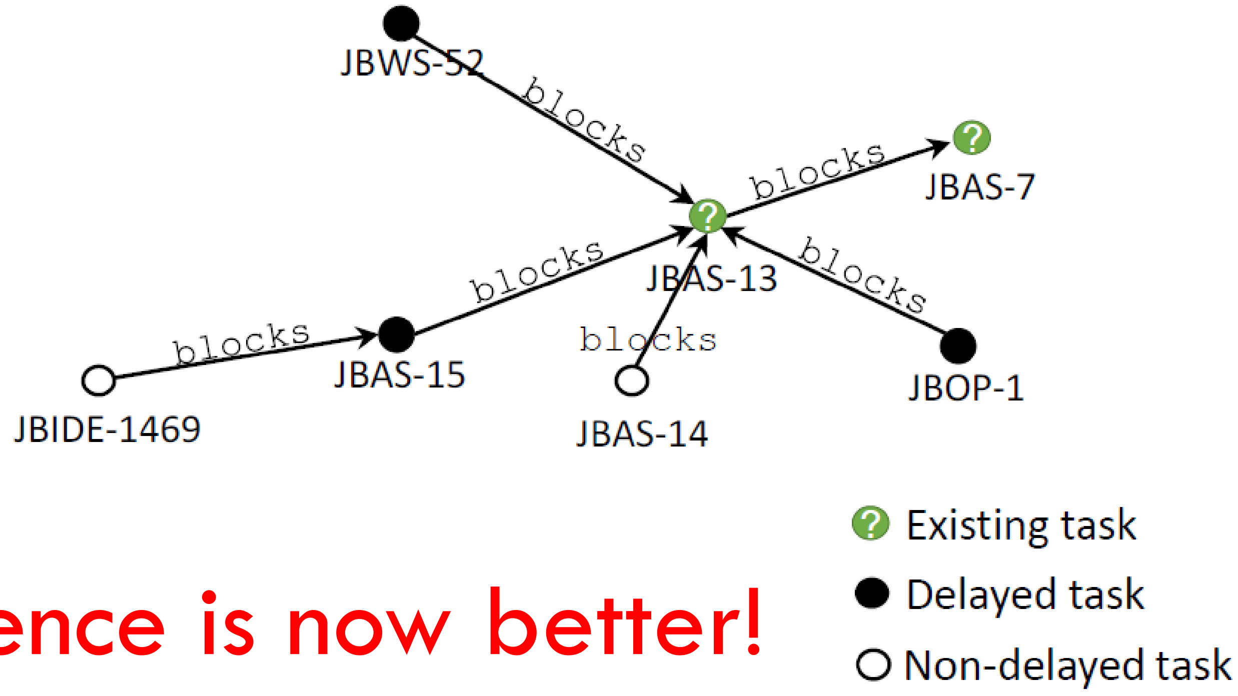


Stacked inference

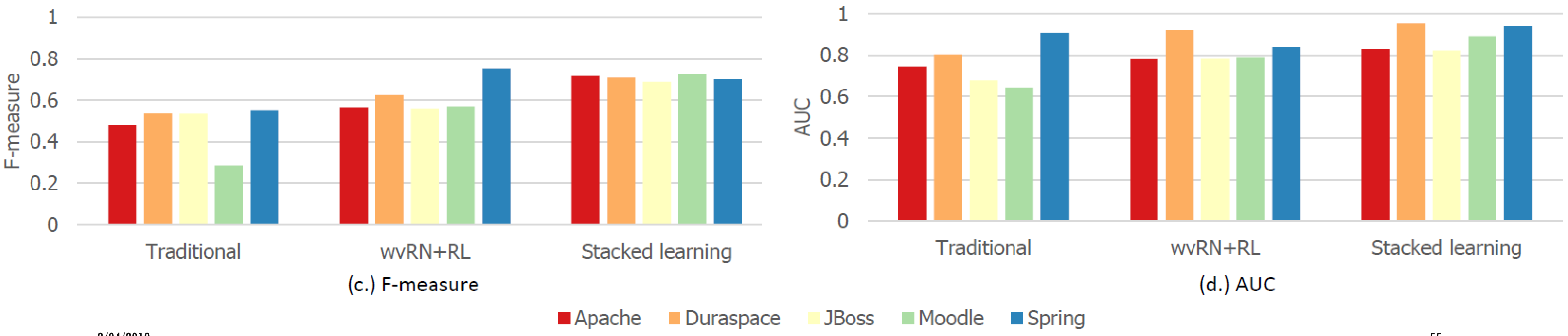


Deep inference

STACKED INFERENCE RESULT



Latest update: Deep Inference is now better!



(c.) F-measure

(d.) AUC

AGENDA

Introduction to PRaDA

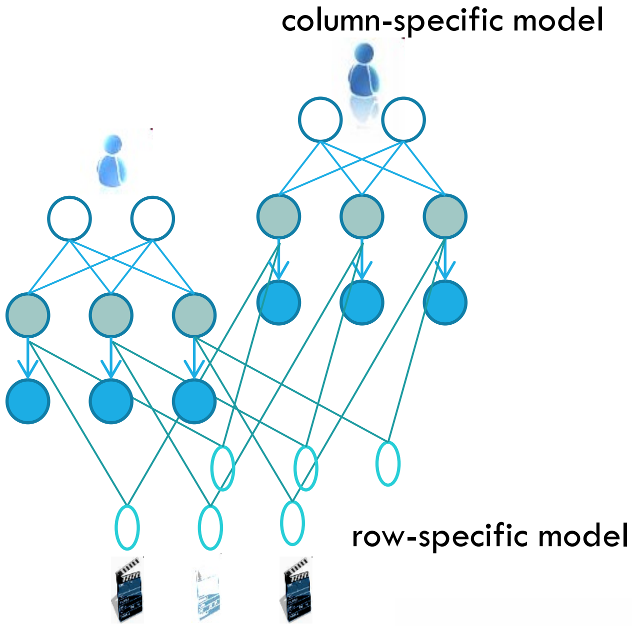
Introduction to deep learning

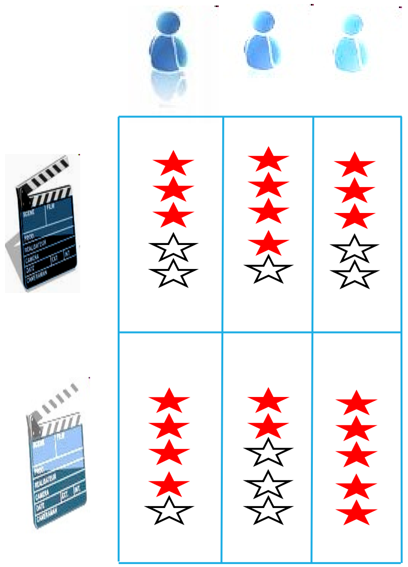
Deep learning for [X], where X =

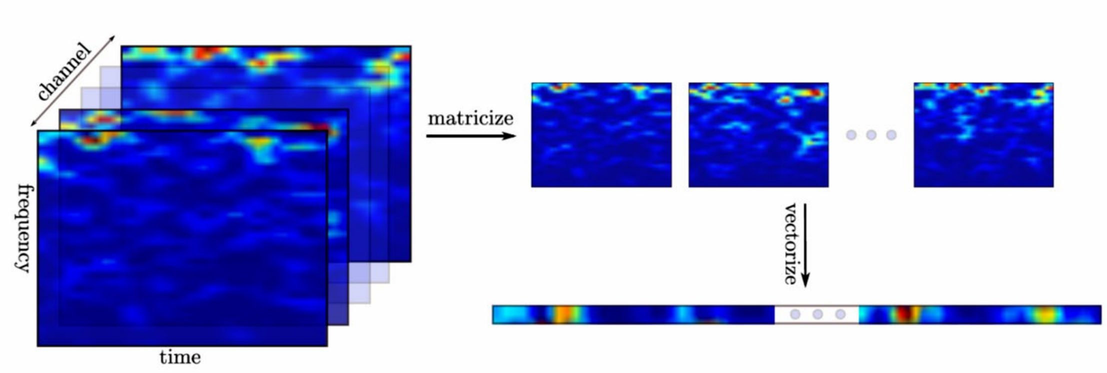
- Healthcare
- Software engineering
- Choice and ranking
- Anomaly detection
- Multi-relational databases
- **Representation**

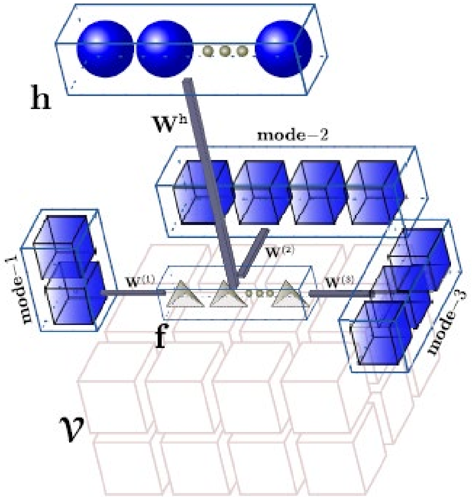
The open room

REPRESENTATION OF **MATRIX AND TENSORS**

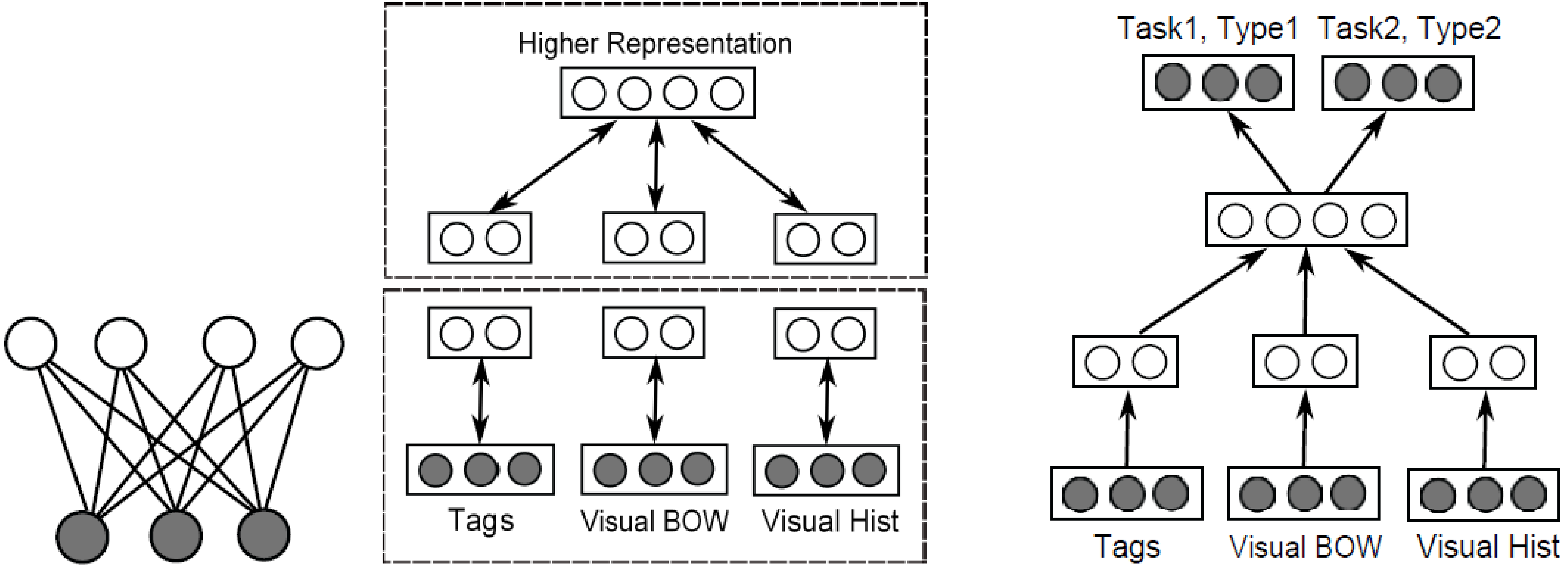








REPRESENTATION OF MIXED-TYPES

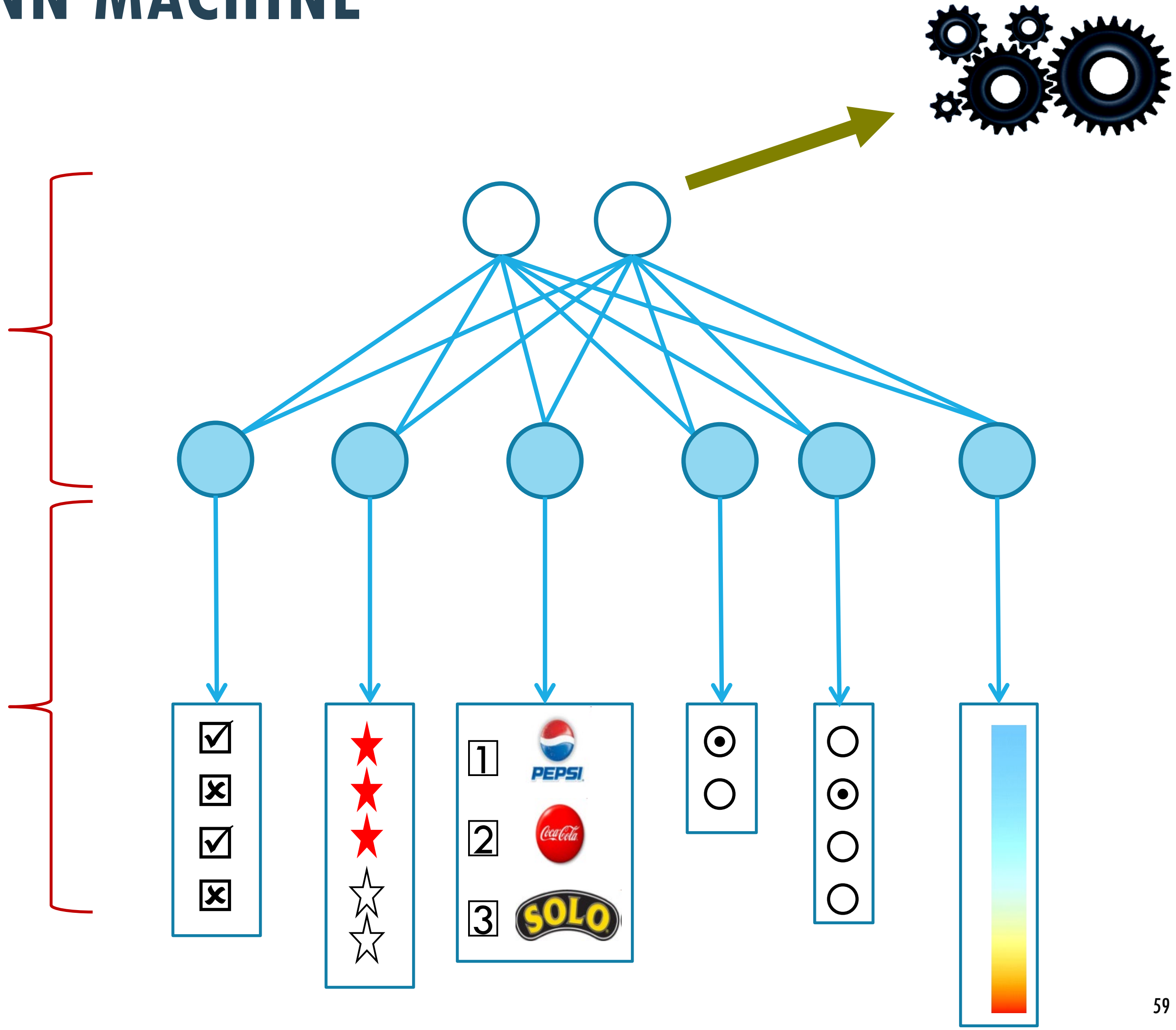


(A) Restricted Boltzmann machine (B) Layerwise training (C) Multitask multityped refinement

THURSTONIAN BOLTZMANN MACHINE

Gaussian RBM

$$b \leq_{\odot} Ax \leq_{\odot} c$$



AGENDA

Introduction to PRaDA

Introduction to deep learning

Deep learning for [X], where X =

- Healthcare
- Software engineering
- Choice and ranking
- Anomaly detection
- Multi-relational databases
- Representation

The open room

THE BEST STRATEGY TO PLAY THIS DEEP LEARNING GAME?

"[…] the dynamics of the game will evolve. In the long run, the right way of playing football is <u>to position yourself intelligently and to wait for the ball to come to you</u>. You'll need to run up and down a bit, either to respond to how the play is evolving or to get out of the way of the scrum when it looks like it might flatten you." (*Neil Lawrence*)

http://inverseprobability.com/2015/07/12/Thoughts-on-ICML-2015/

“A NEW IDEA IS JUST RE-PACKAGING OF OLD IDEAS”

DEEP (LEARNING) QUESTIONS

Is this just yet-another-toolbox or a way of thinking?

Is this a right approach to AI?