MACHINE LEARNING IN HIGH-ENERGY PHYSICS

BALÁZS KÉGL

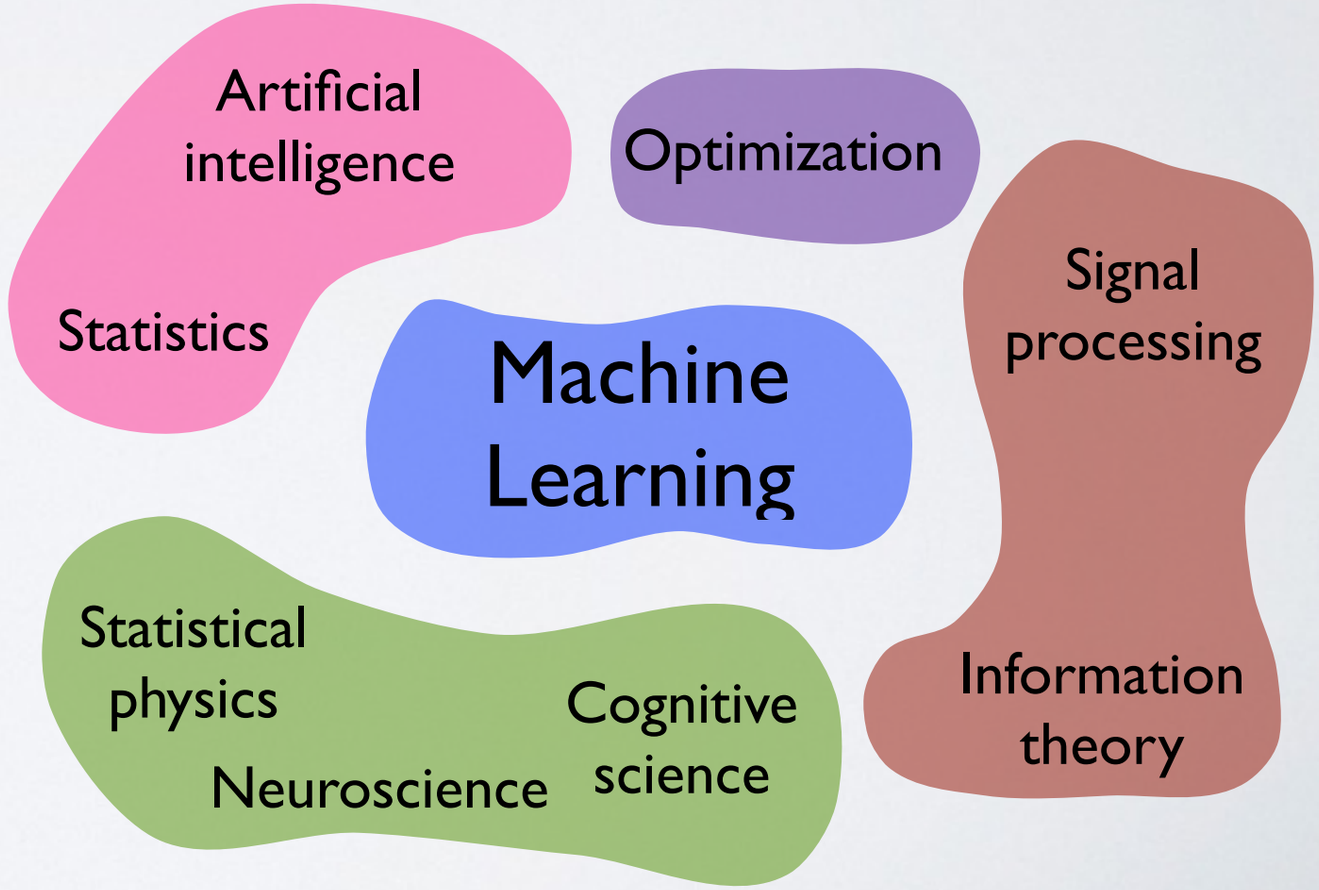
Linear Accelerator Laboratory and Computer Science Laboratory
CNRS/IN2P3 & University Paris-S{ud,aclay}

CERN, September 5, 2014

Outline

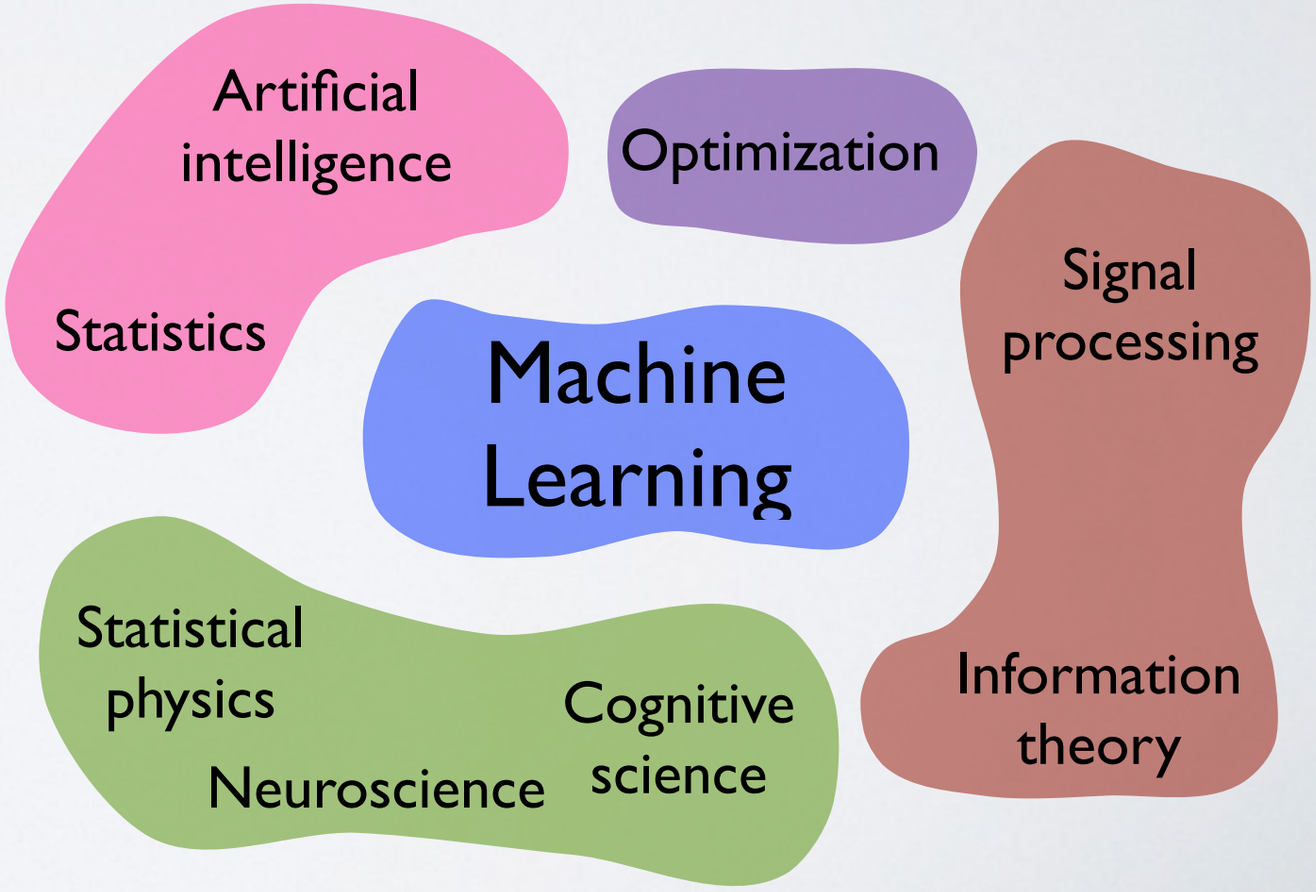
- What is machine learning/data science?
- Two projects to illustrate ML in HEP
  - budgeted learning for triggers (LHCb)
  - classification for discovery and the HiggsML challenge (ATLAS)

What is machine learning?



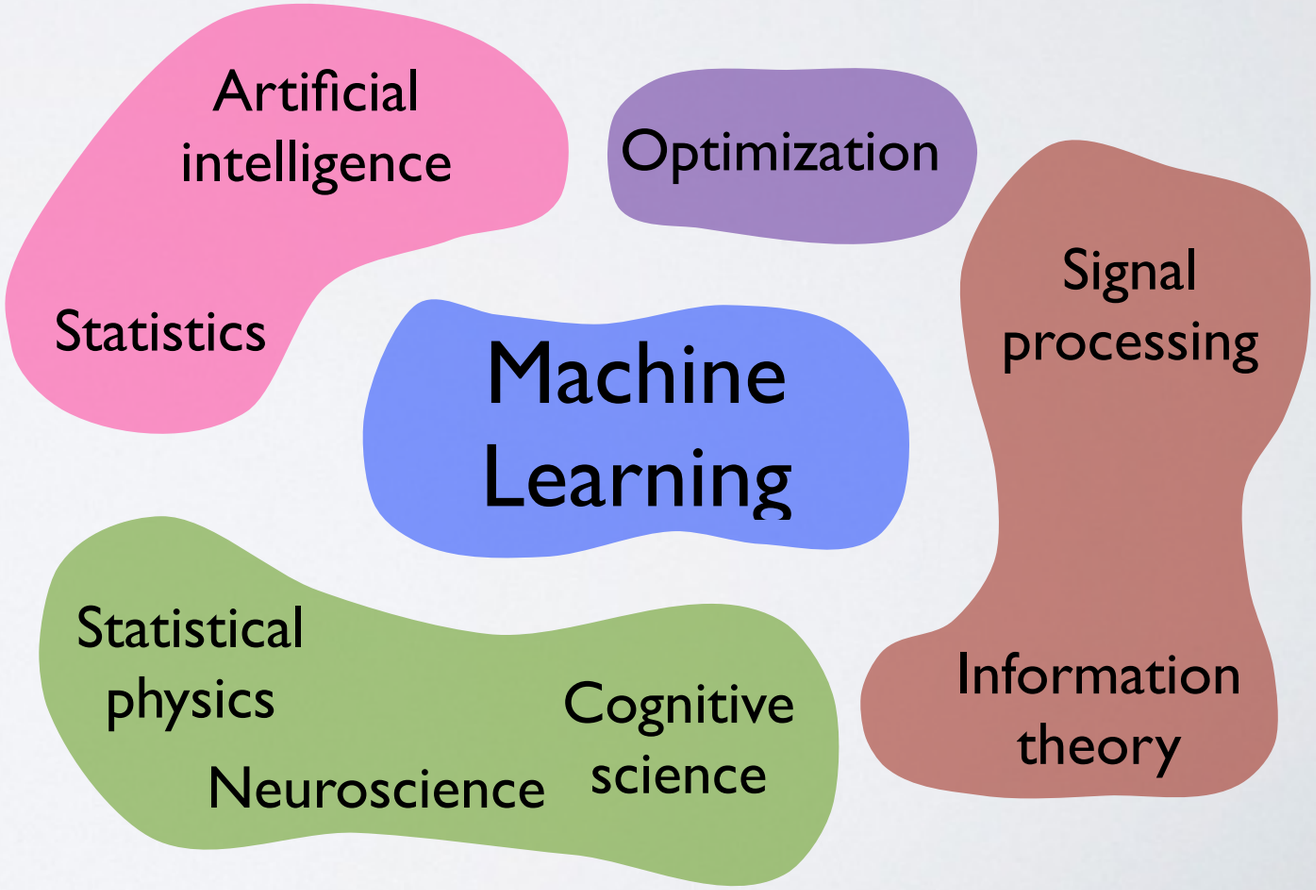
WHAT IS MACHINE LEARNING?

- "The science of getting computers to act without being explicitly programmed" - Andrew Ng (Stanford/Coursera)



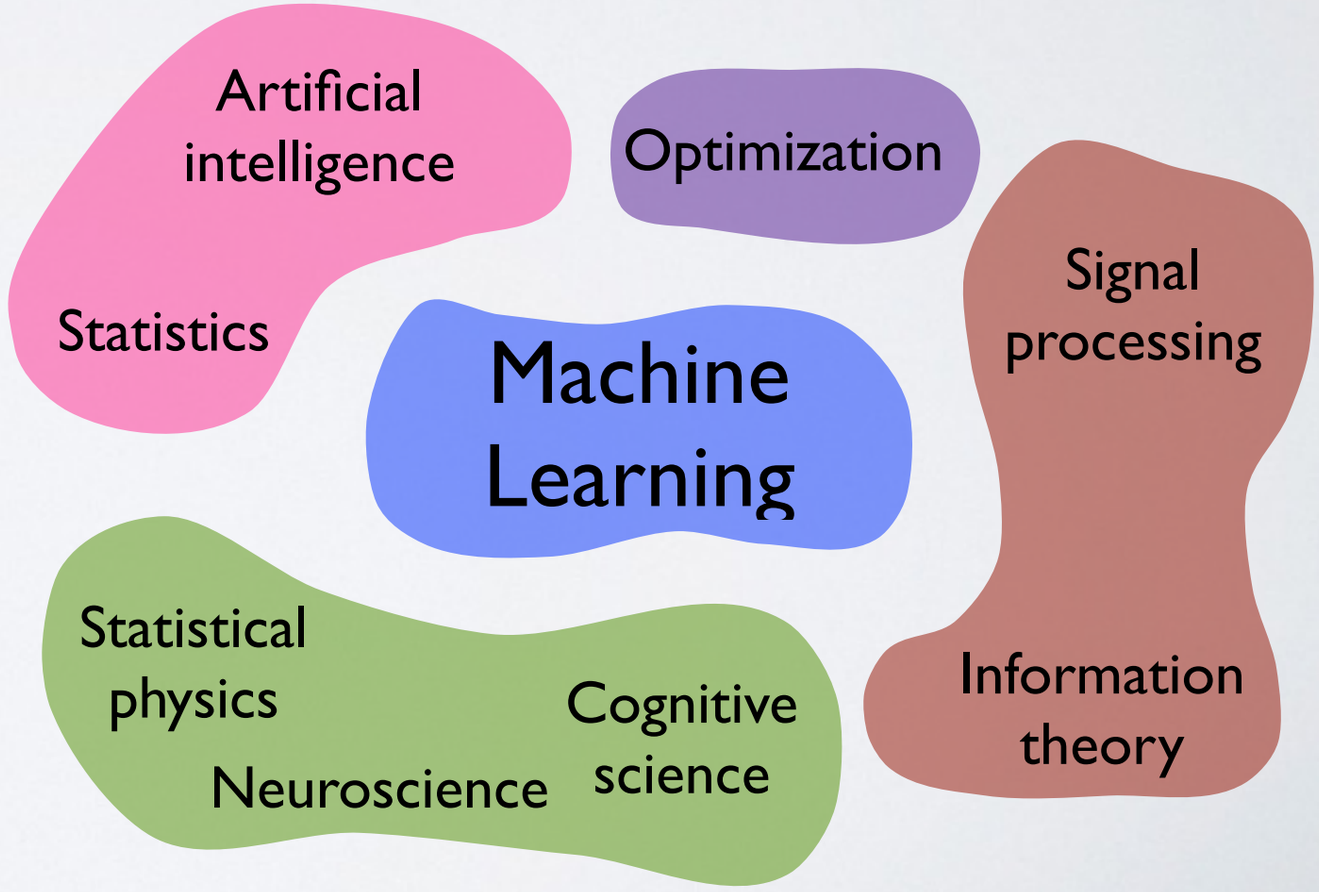
WHAT IS MACHINE LEARNING?

- "The science of getting computers to act without being explicitly programmed" - Andrew Ng (Stanford/Coursera)
  - part of standard computer science curriculum since the 90s



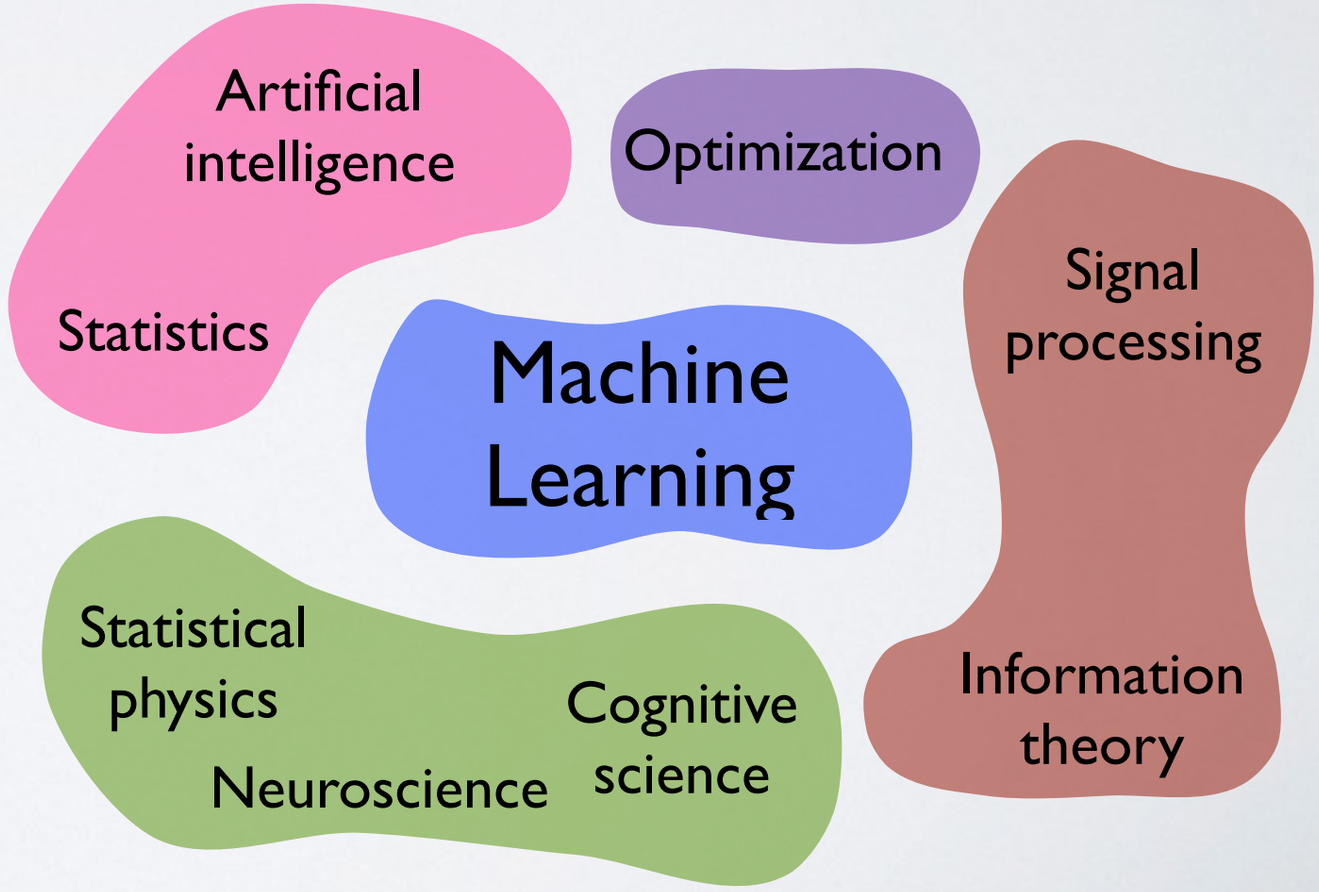
WHAT IS MACHINE LEARNING?

- "The science of getting computers to act without being explicitly programmed" - Andrew Ng (Stanford/Coursera)
  - part of standard computer science curriculum since the 90s
  - inferring knowledge from data



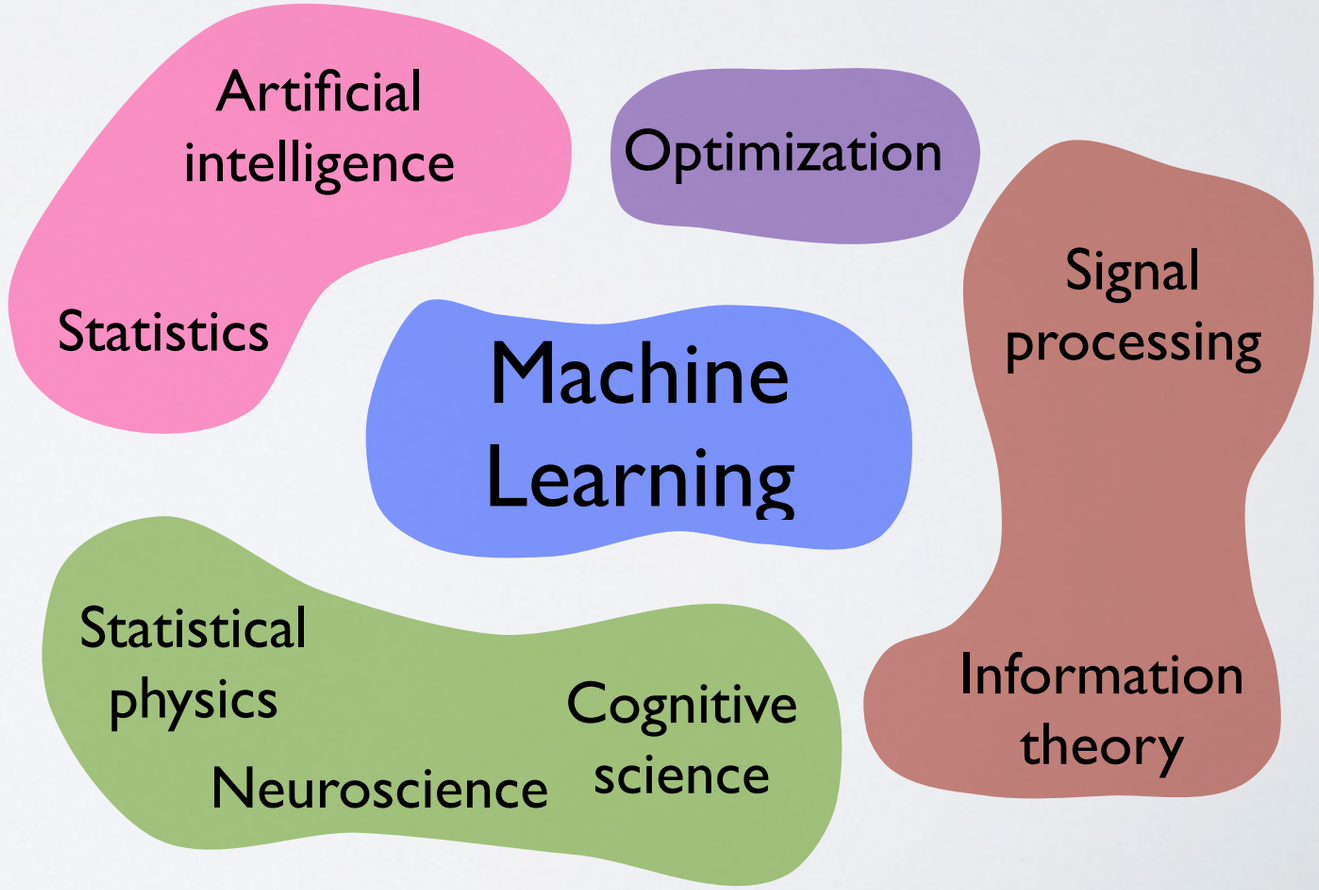
WHAT IS MACHINE LEARNING?

- "The science of getting computers to act without being explicitly programmed" - Andrew Ng (Stanford/Coursera)
  - part of standard computer science curriculum since the 90s
  - inferring knowledge from data
  - generalizing to unseen data



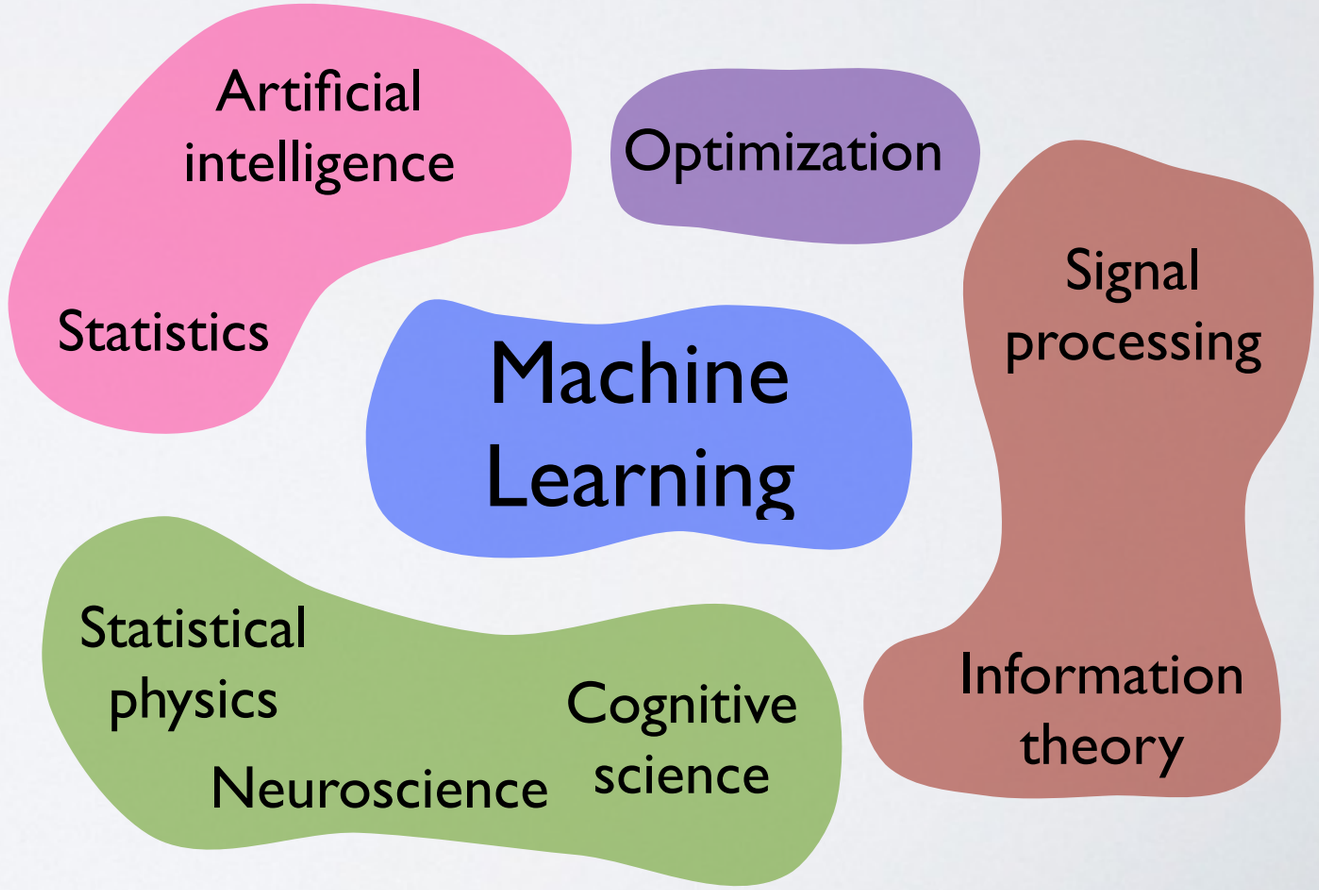
WHAT IS MACHINE LEARNING?

- "The science of getting computers to act without being explicitly programmed" - Andrew Ng (Stanford/Coursera)
  - part of standard computer science curriculum since the 90s
  - inferring knowledge from data
  - generalizing to unseen data
  - usually no parametric model assumptions



WHAT IS MACHINE LEARNING?

- "The science of getting computers to act without being explicitly programmed" - Andrew Ng (Stanford/Coursera)
  - part of standard computer science curriculum since the 90s
  - inferring knowledge from data
  - generalizing to unseen data
  - usually no parametric model assumptions
  - emphasizing the computational challenges



MACHINE LEARNING TAXONOMY

MACHINE LEARNING TAXONOMY

- Supervised learning: non-parametric (model-free) input - output functions
  - classification (Trees, BDT, SVM, NN) - what you call MVA
  - regression (Trees, NN, Gaussian Processes)
- Unsupervised learning: non-parametric data representation
  - clustering (k-means, spectral clustering, Dirichlet processes)
  - dimensionality reduction (PCA, ISOMAP, LLE, auto-associative NN)
  - density estimation (kernel density, Gaussian mixtures, the Boltzmann machine)
- Reinforcement learning:
  - learning + dynamic control: learn to behave in an environment to maximize cumulative reward

MACHINE LEARNING TAXONOMY

- Supervised learning: non-parametric (model-free) input - output functions
  - classification (Trees, BDT, SVM, NN) - what you call MVA
  - regression (Trees, NN, Gaussian Processes)
- Unsupervised learning: non-parametric data representation
  - clustering (k-means, spectral clustering, Dirichlet processes)
  - dimensionality reduction (PCA, ISOMAP, LLE, auto-associative NN)
  - density estimation (kernel density, Gaussian mixtures, the Boltzmann machine)
- Reinforcement learning:
  - learning + dynamic control: learn to behave in an environment to maximize cumulative reward

MACHINE LEARNING RESEARCH

We make our living by inventing techniques

MACHINE LEARNING RESEARCH

We are interested in problems at the edges of our current methodological capacities

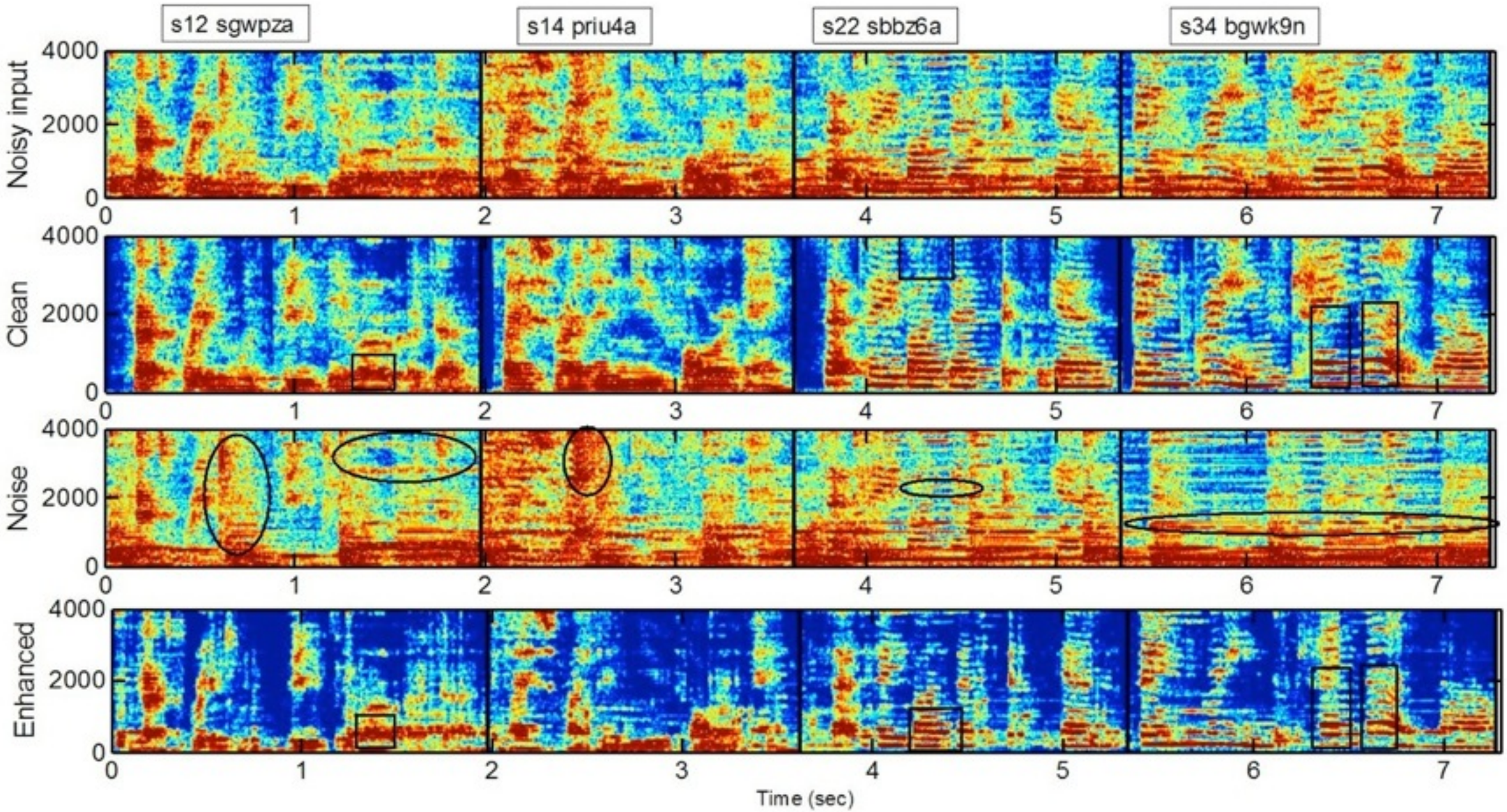
CLASSIFICATION

Character recognition

CLASSIFICATION

Emotion recognition

CLASSIFICATION



Speech recognition

CLASSIFICATION

- Input: a usually high dimensional vector x
- Output: a category (aka label, class) y
- Usually no parametric model
  - the classification function $y = g(x)$ is learned using a training set $D = \{(x_1,y_1), \ldots, (x_n,y_n)\}$
- Well-tested algorithms:
  - neural networks, support vector machines, boosting (BDTs)

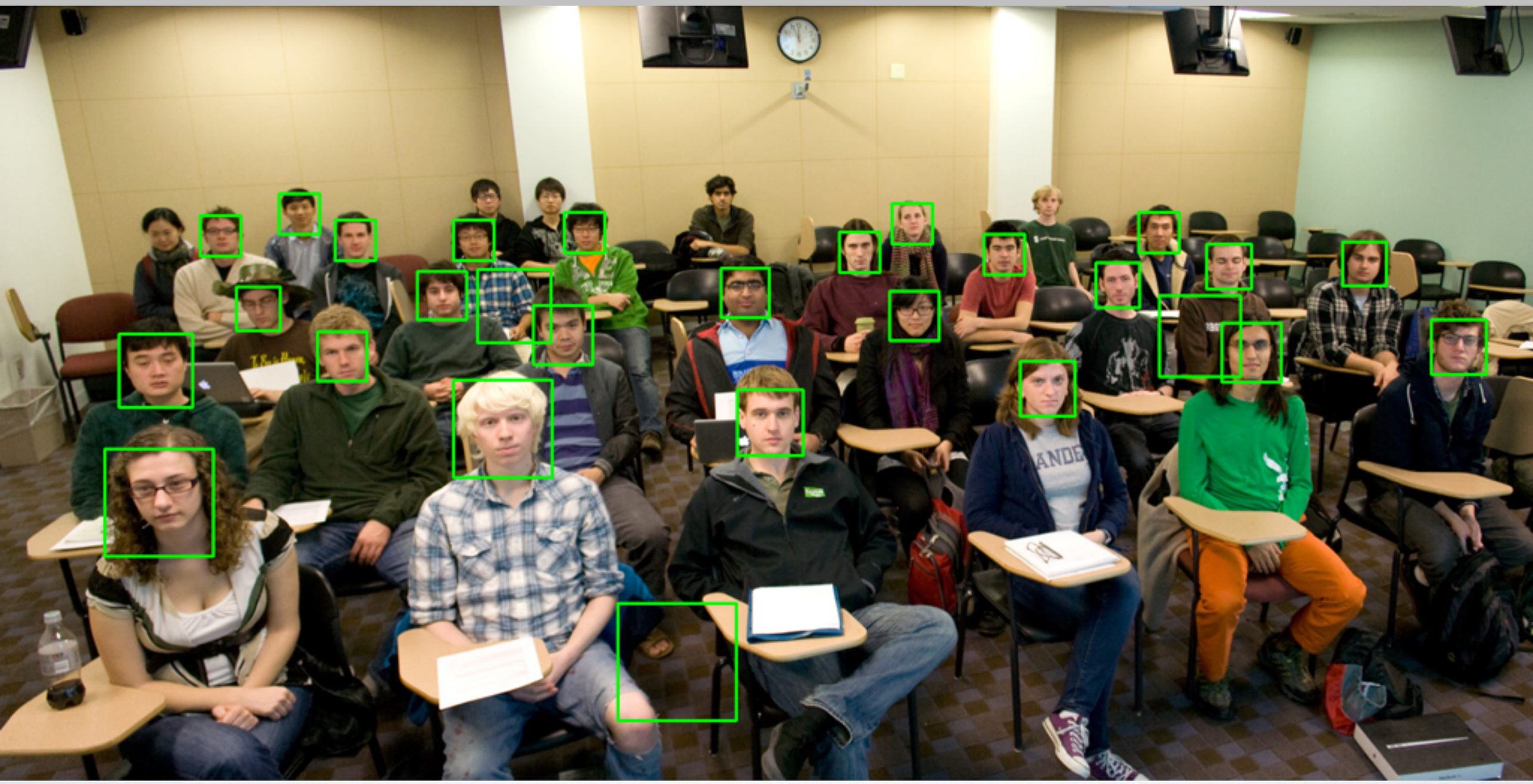
CLASSIFICATION

The only goal is a low probability of error

$$P(g(x) \neq y)$$

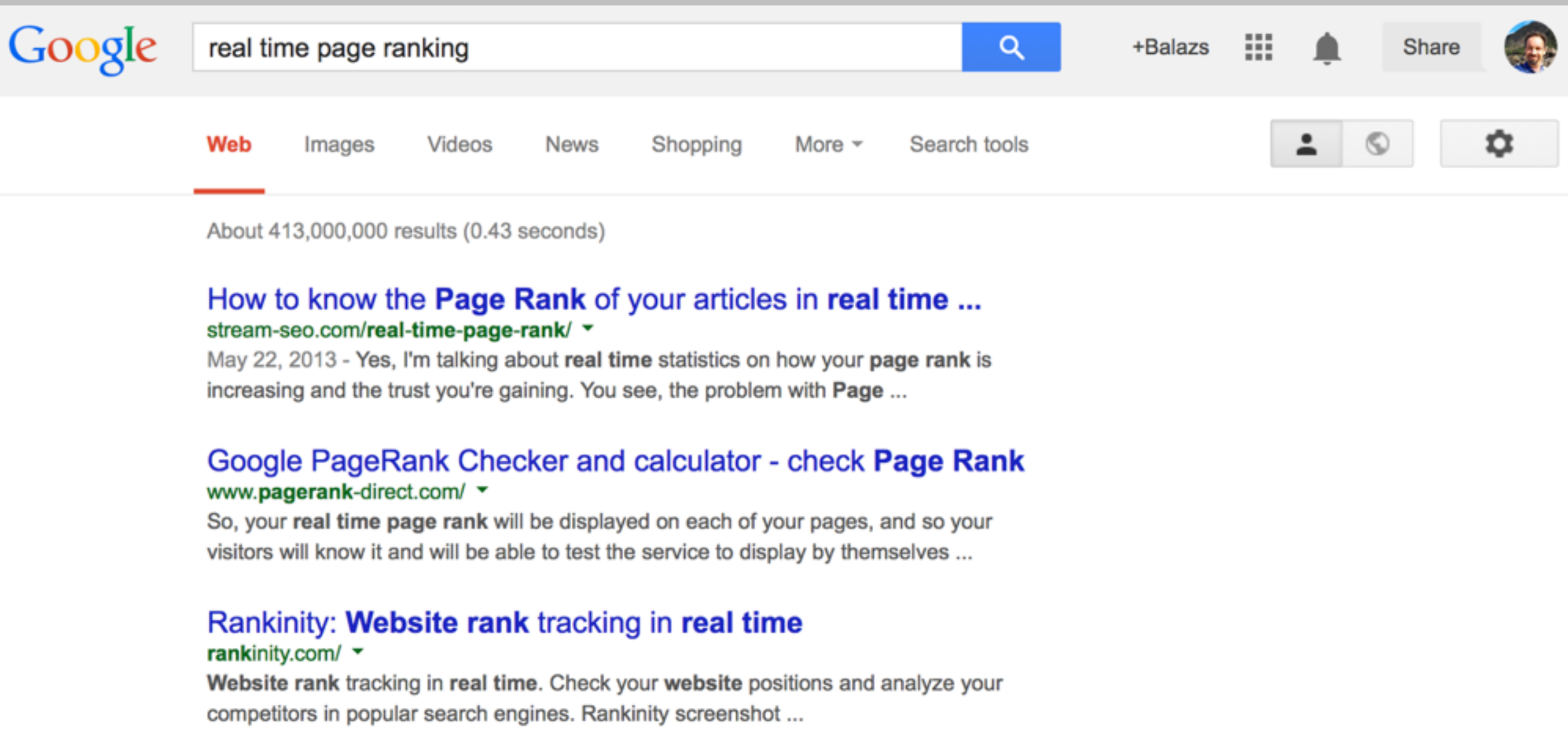
on previously unseen examples (x, y)

BUDGETED CLASSIFICATION



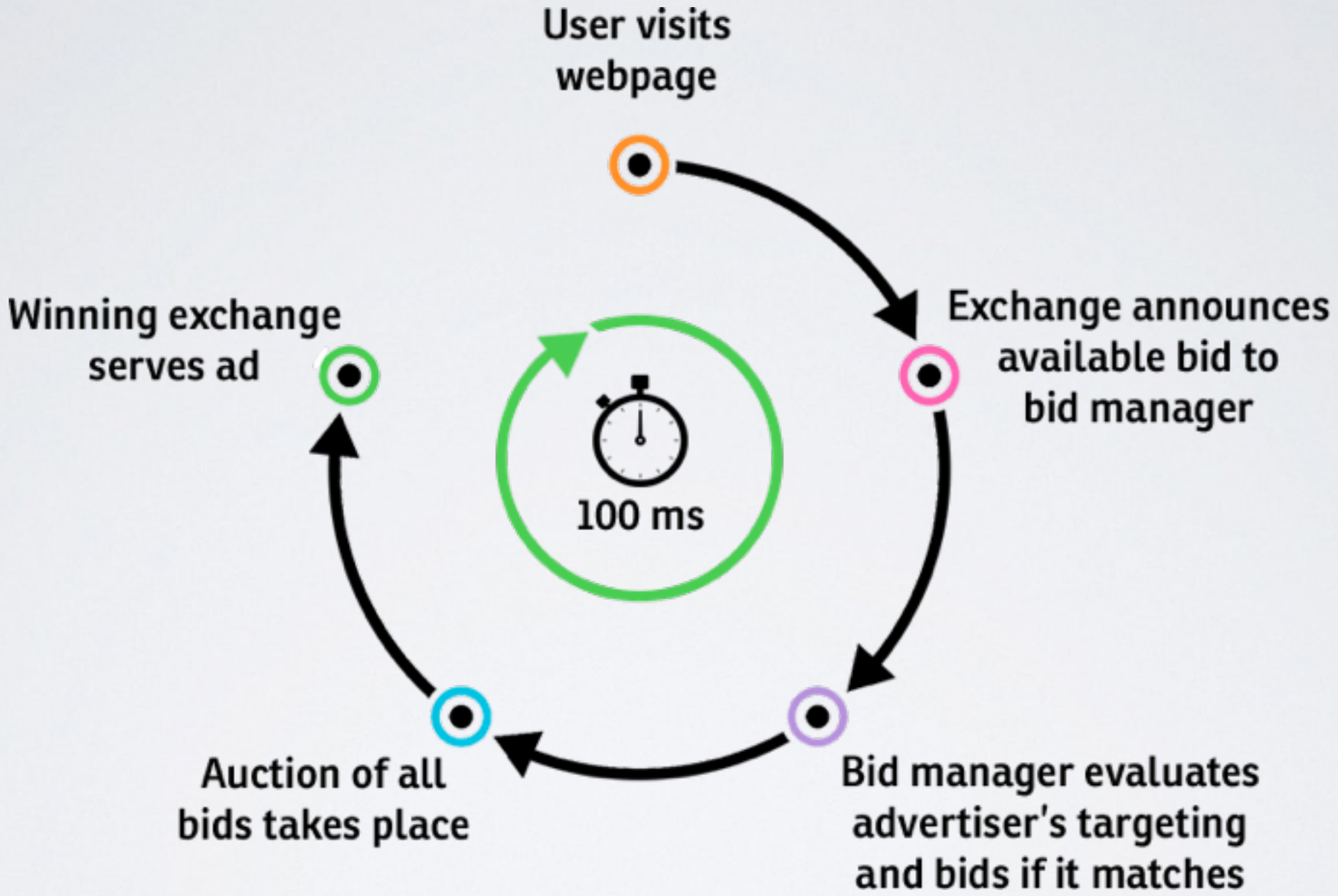
Real time face detection

BUDGETED CLASSIFICATION



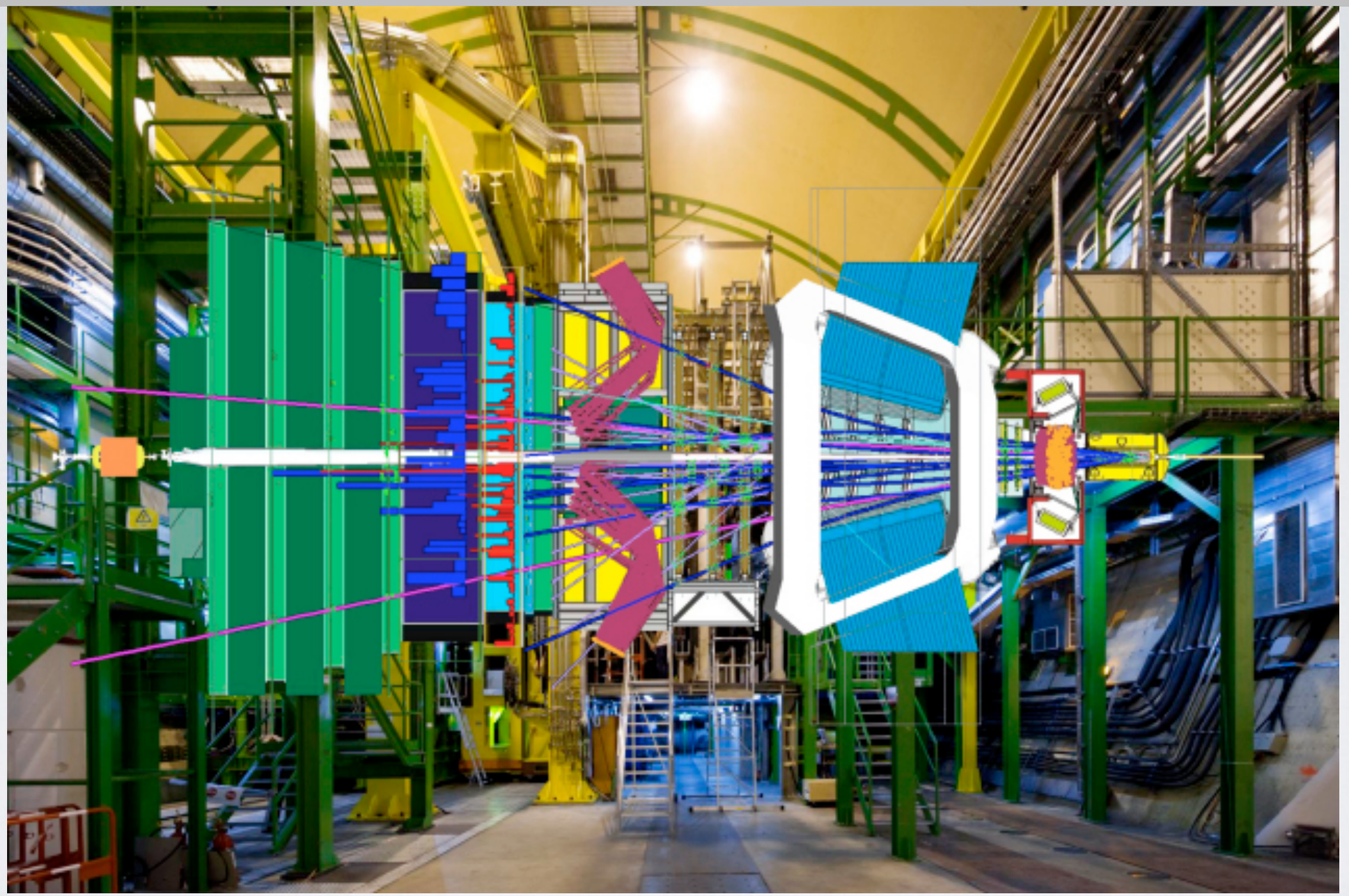
Real time web page ranking

BUDGETED CLASSIFICATION



Real time ad placement

BUDGETED CLASSIFICATION



Real time signal/background separation

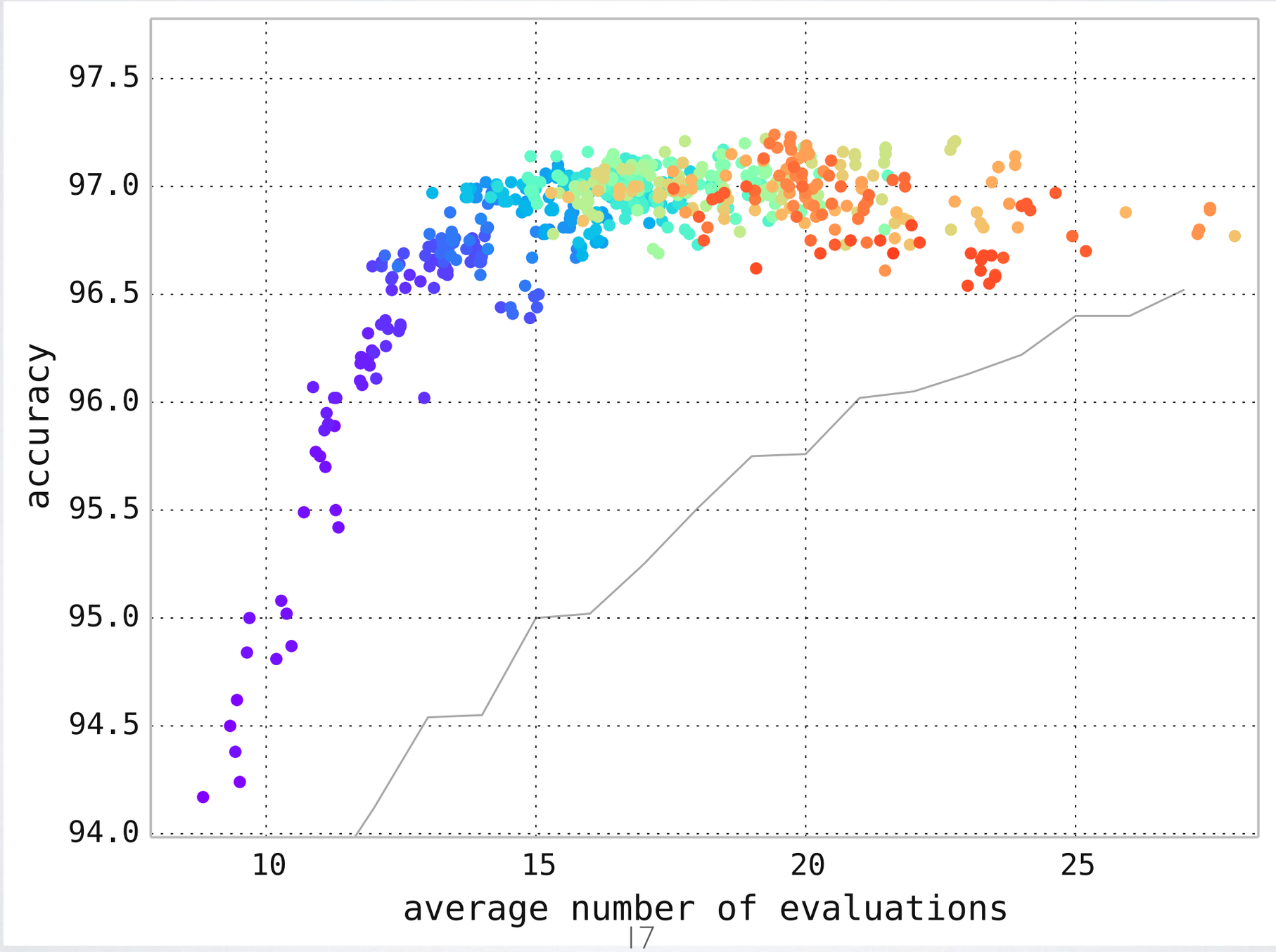
BUDGETED CLASSIFICATION

The second goal is the fast execution of

$$g(x)$$

BUDGETED CLASSIFICATION

Trade-off between quality and speed



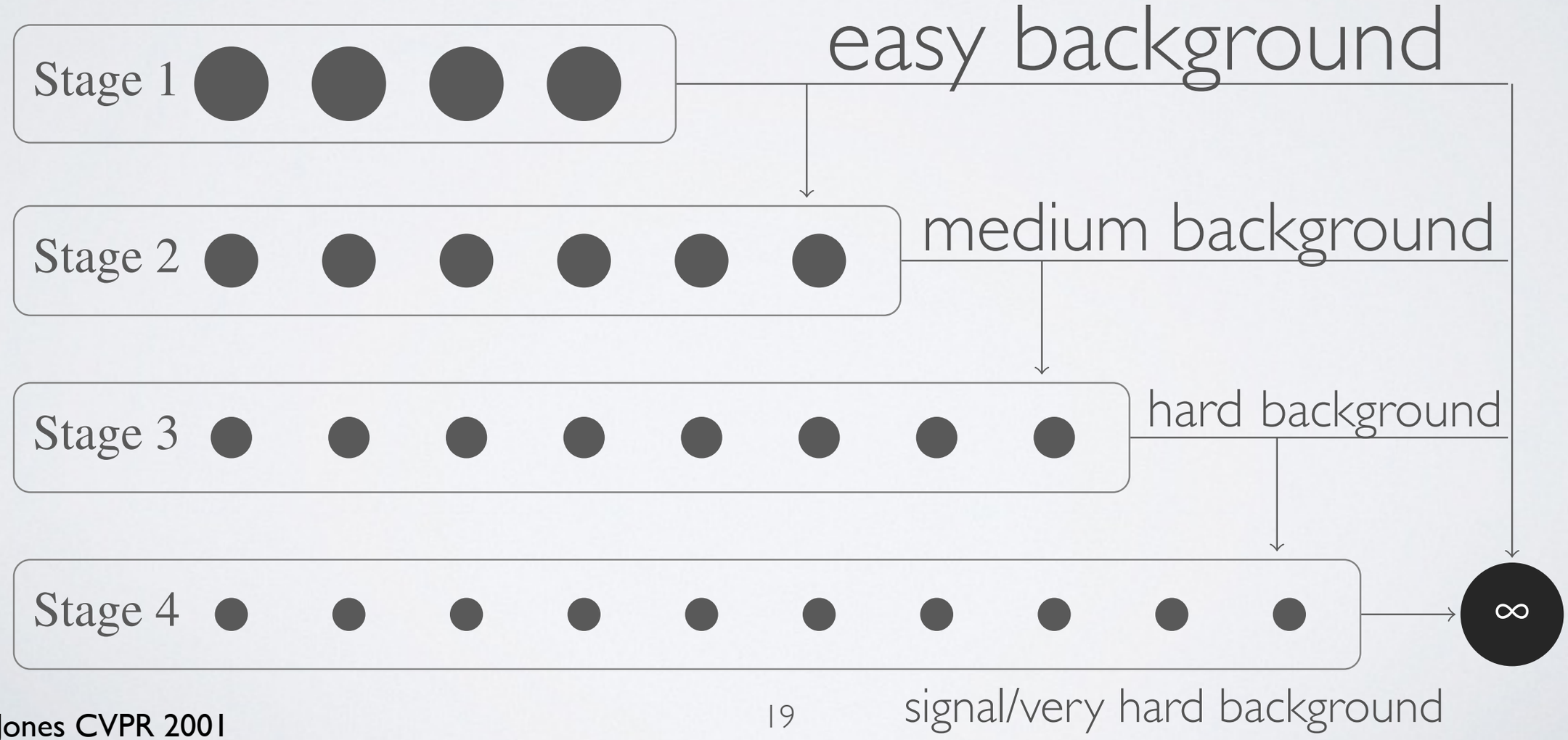
BUDGETED CLASSIFICATION

- Time constraints
- Memory constraints
- Consumption constraints
- Communication constraints

BUDGETED CLASSIFICATION

The common design:

cascade classification = trigger with levels



Viola-Jones CVPR 2001

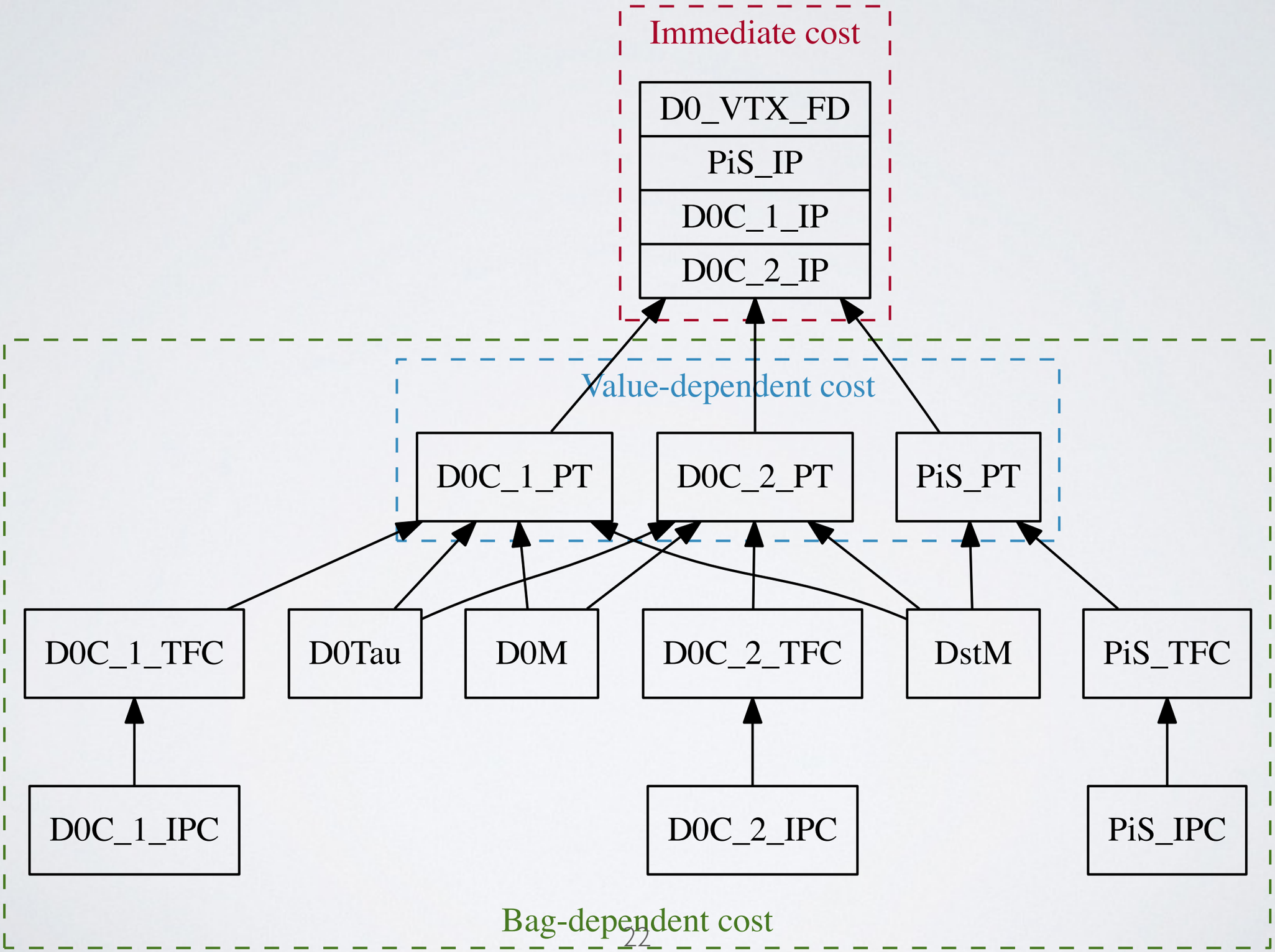
THE LHCB TRIGGER

- Collaboration with
  - Vava Gligorov (CERN)
  - Mike Williams (MIT)
  - Djalel Benbouzid (LAL)

THE LHCB TRIGGER

- A beautifully complex problem
- varying feature costs
- cost may depend on the value
- events are bags of overlapping candidates

THE LHCB TRIGGER

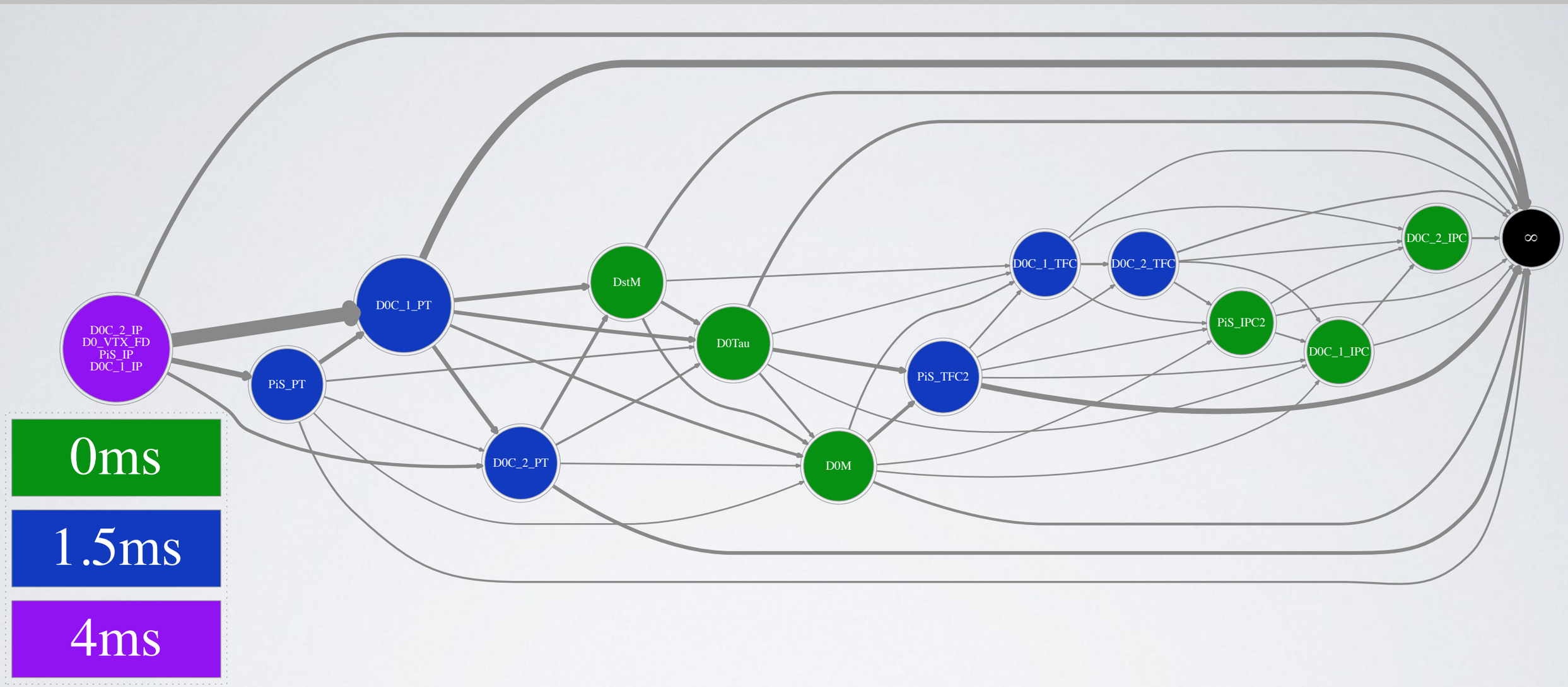


MDDAG: A SIGNAL/BCKG DECISION GRAPH

MDDAG: A SIGNAL/BCKG DECISION GRAPH

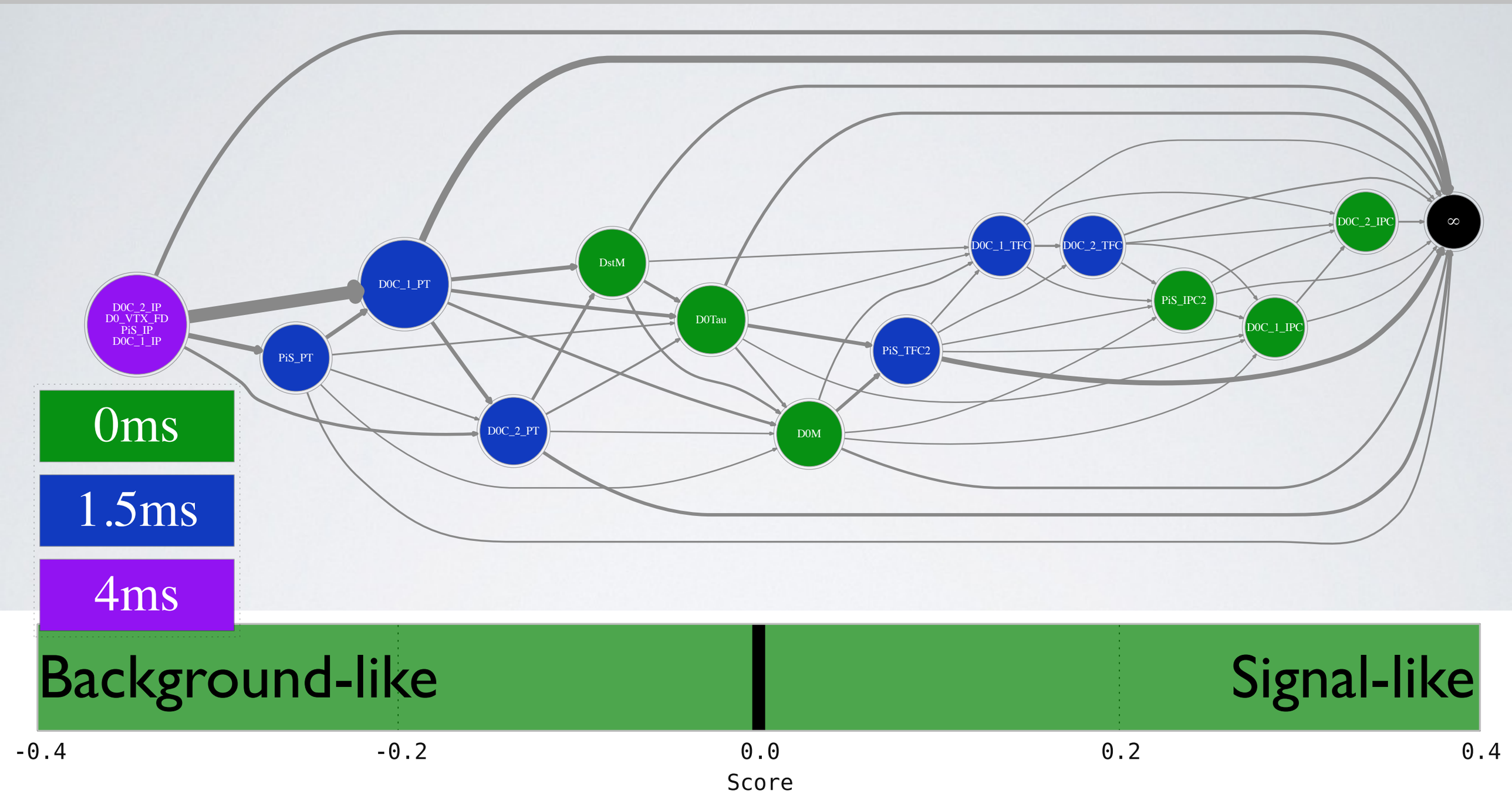
Benbouzid et al. ICML 2012

MDDAG: A SIGNAL/BCKG DECISION GRAPH



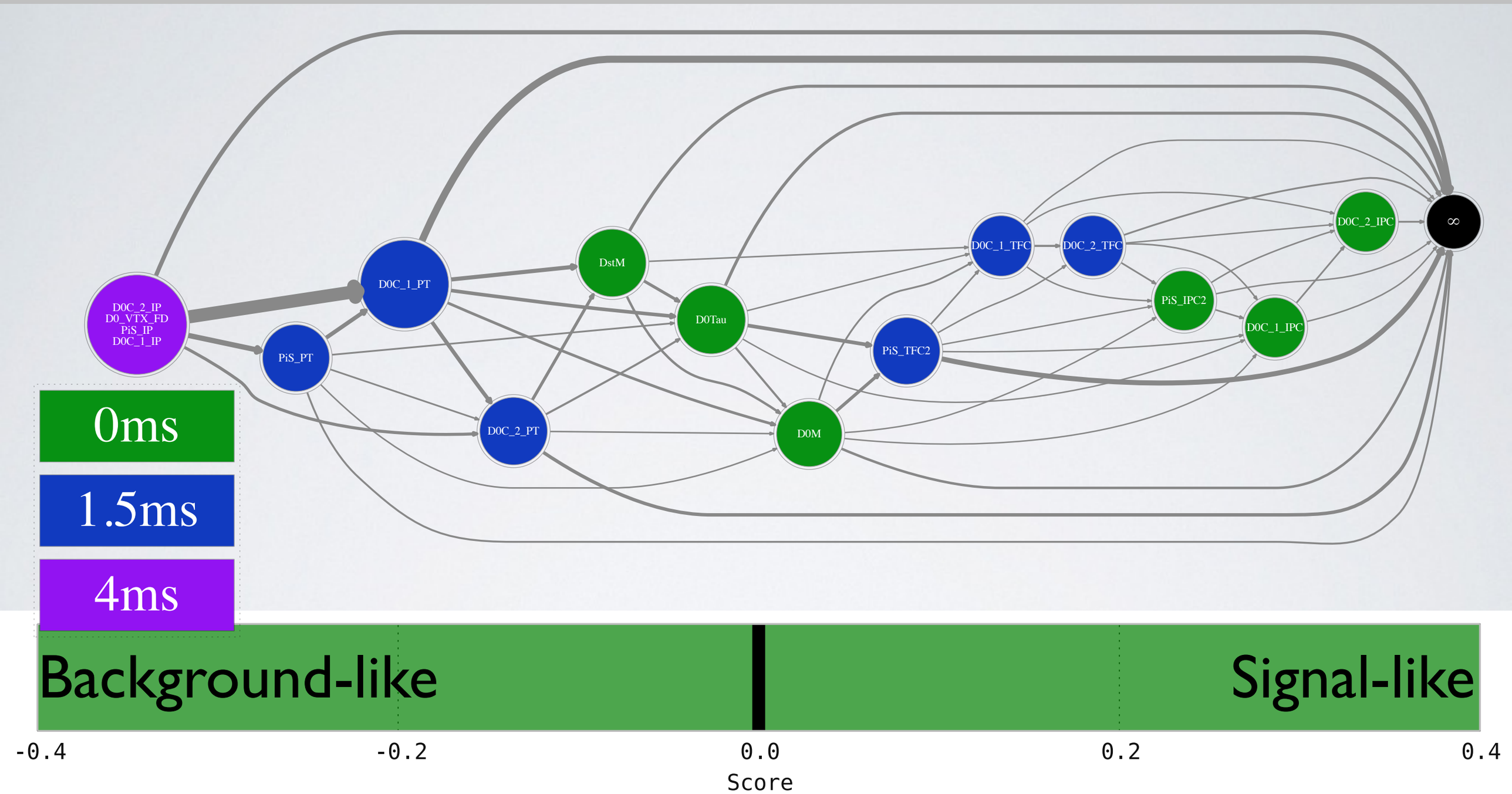
Benbouzid et al. ICML 2012

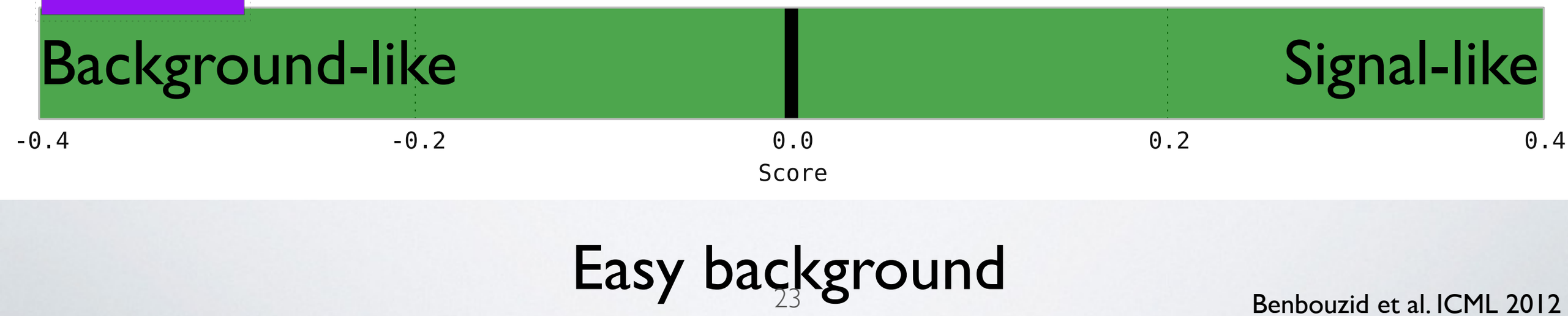
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Benbouzid et al. ICML 2012

MDDAG: A SIGNAL/BCKG DECISION GRAPH

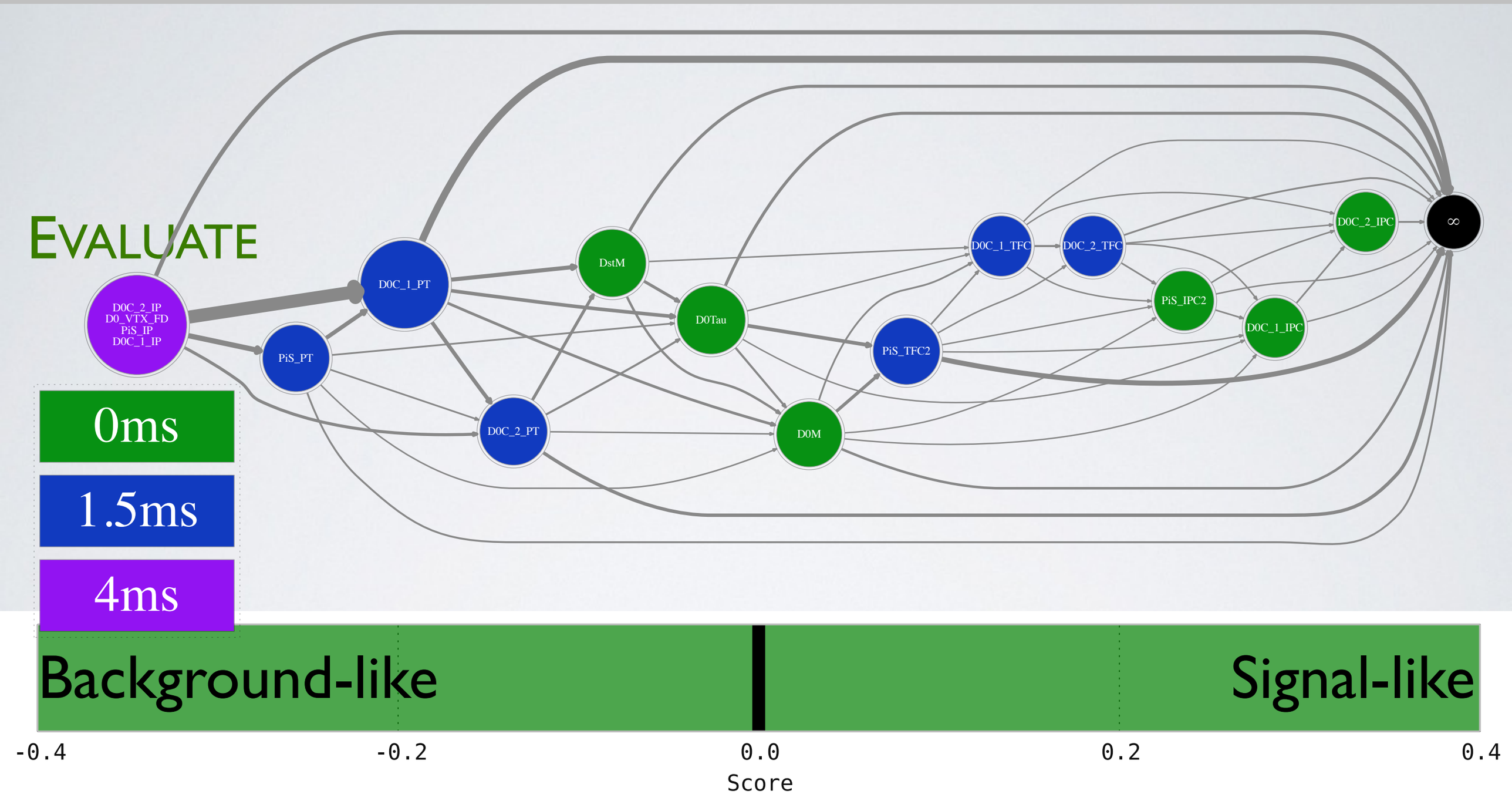




Easy background

Benbouzid et al. ICML 2012

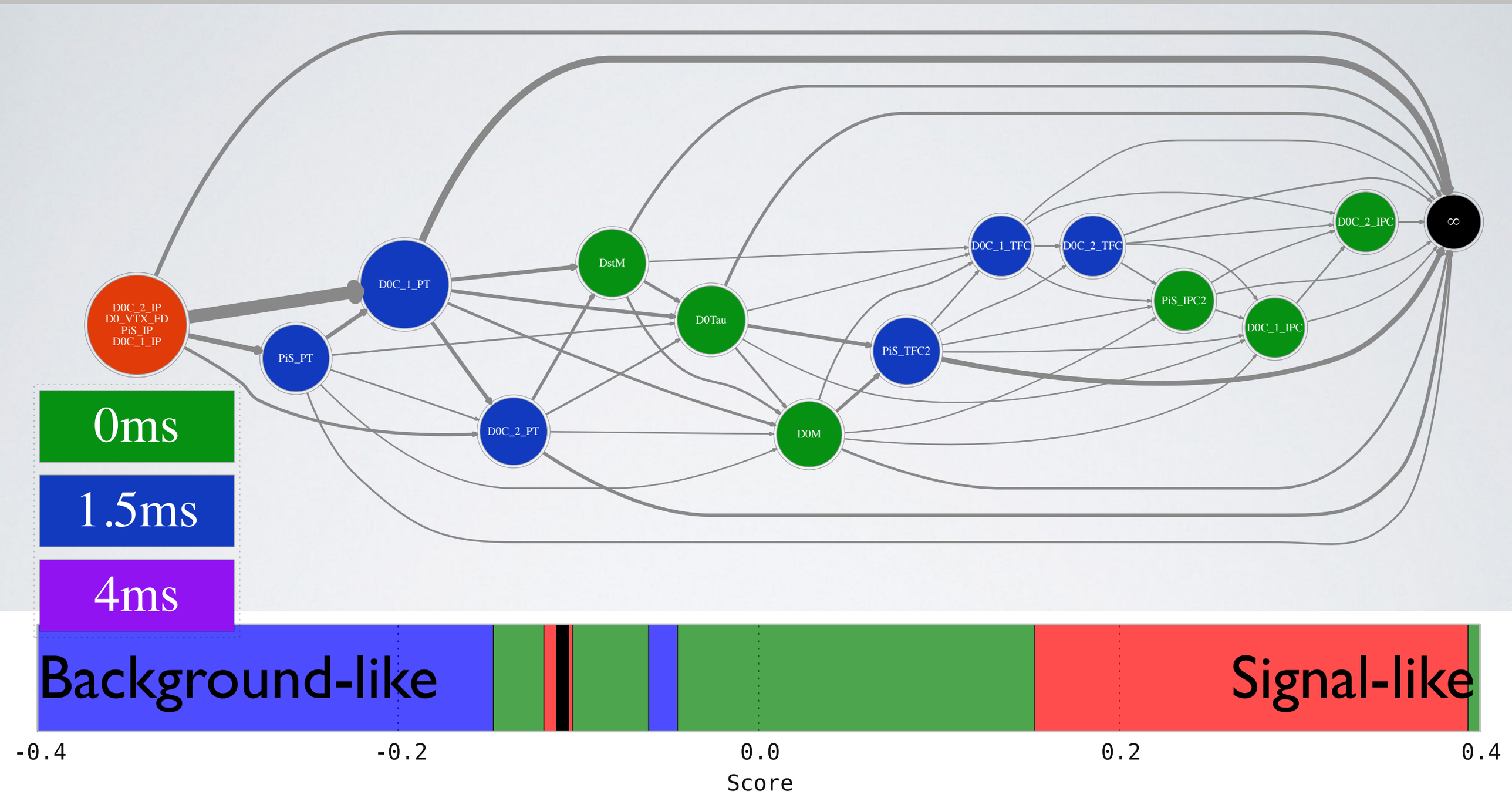
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy background

Benbouzid et al. ICML 2012

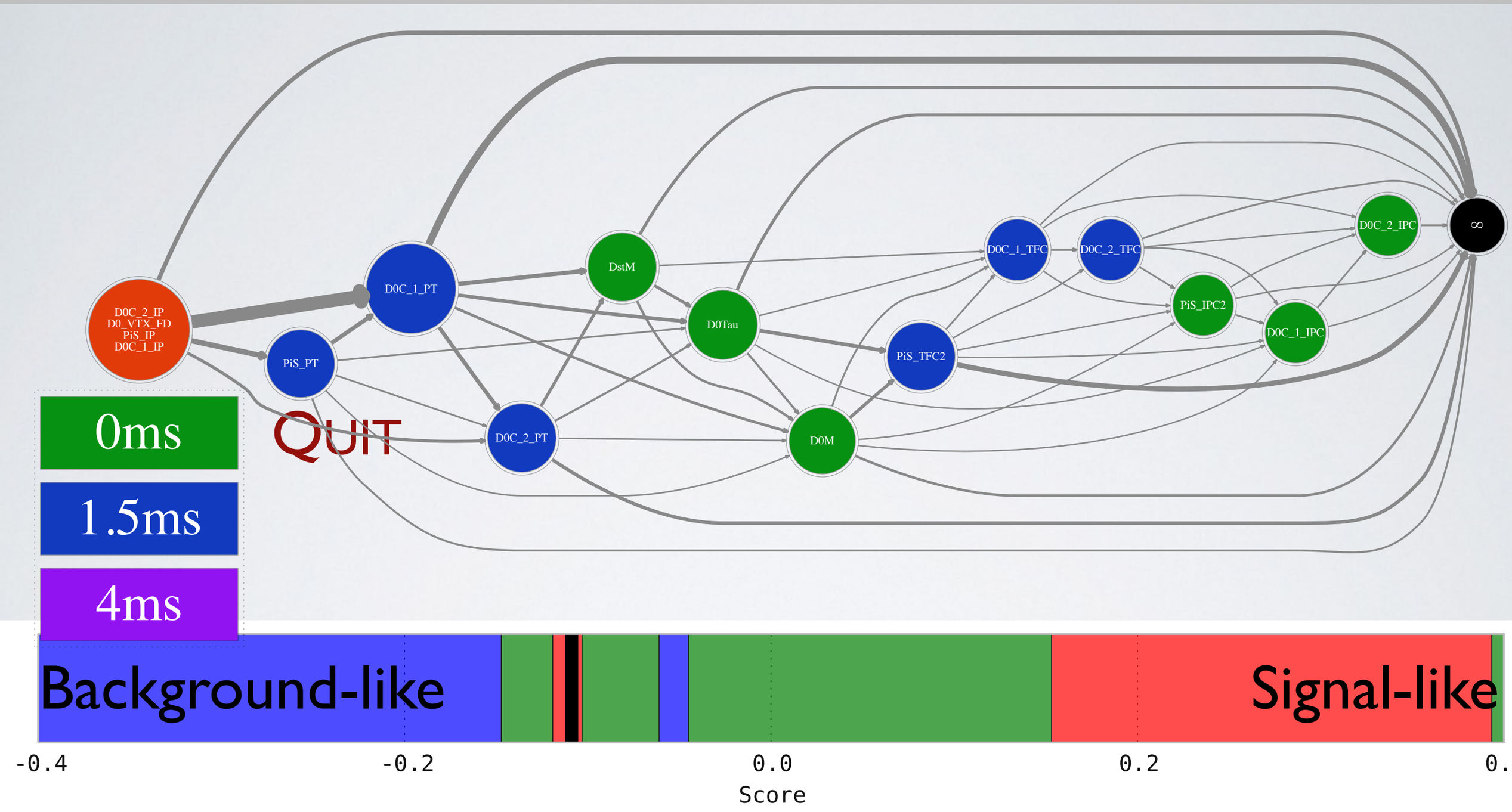
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy background

Benbouzid et al. ICML 2012

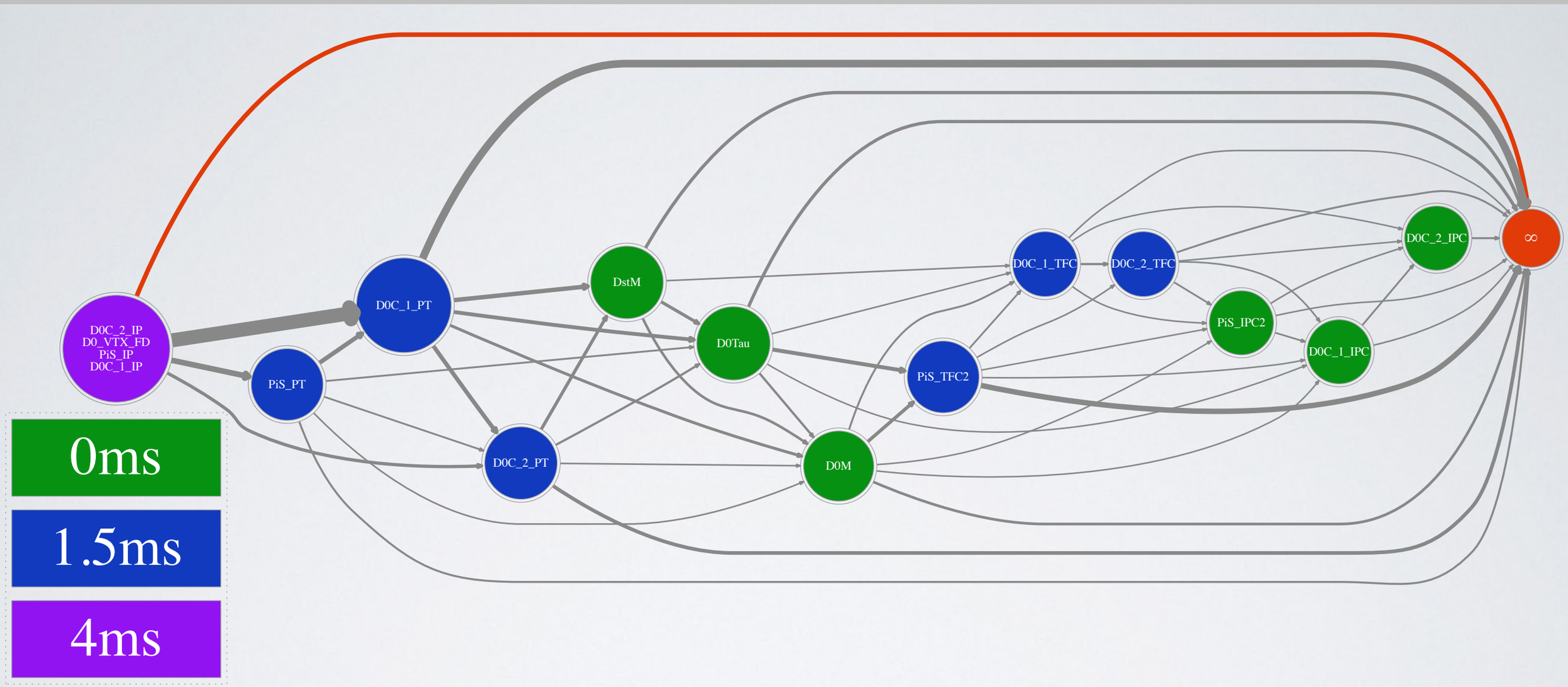
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy background

Benbouzid et al. ICML 2012

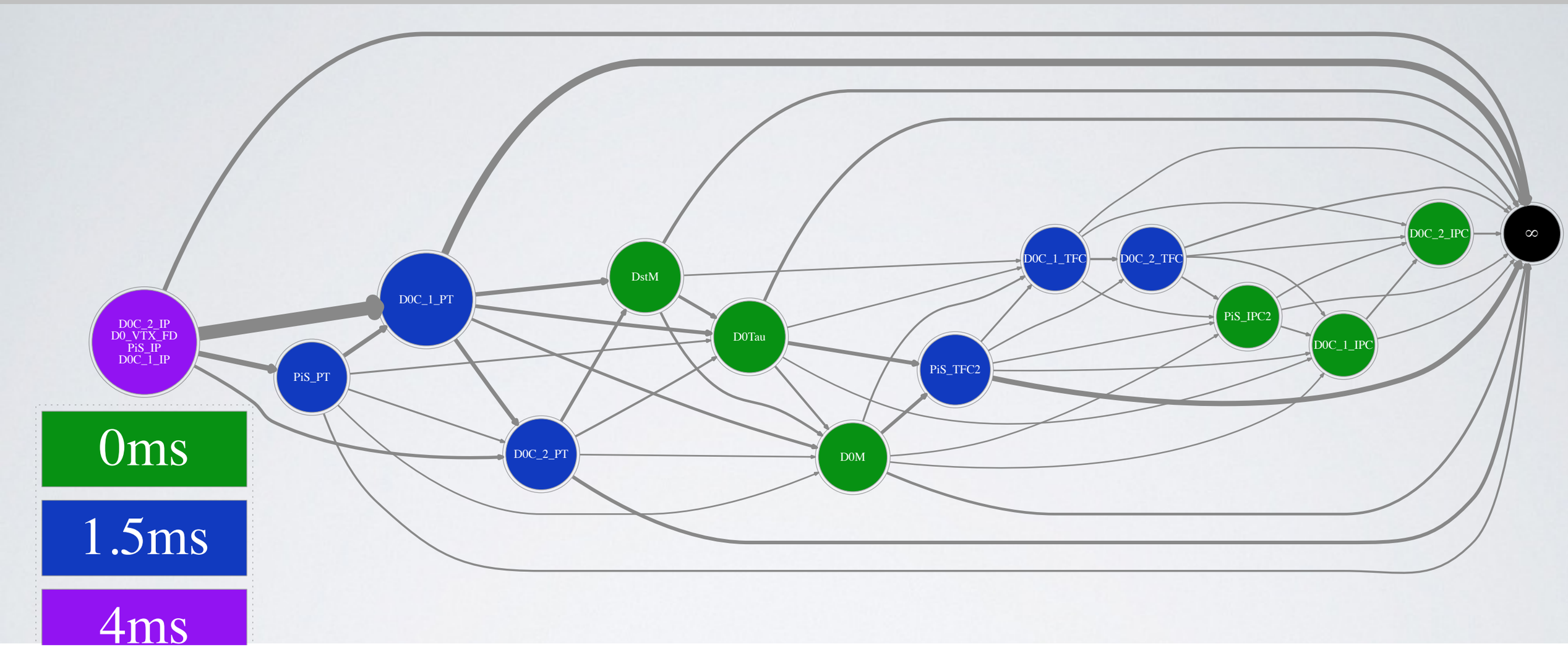
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy background

Benbouzid et al. ICML 2012

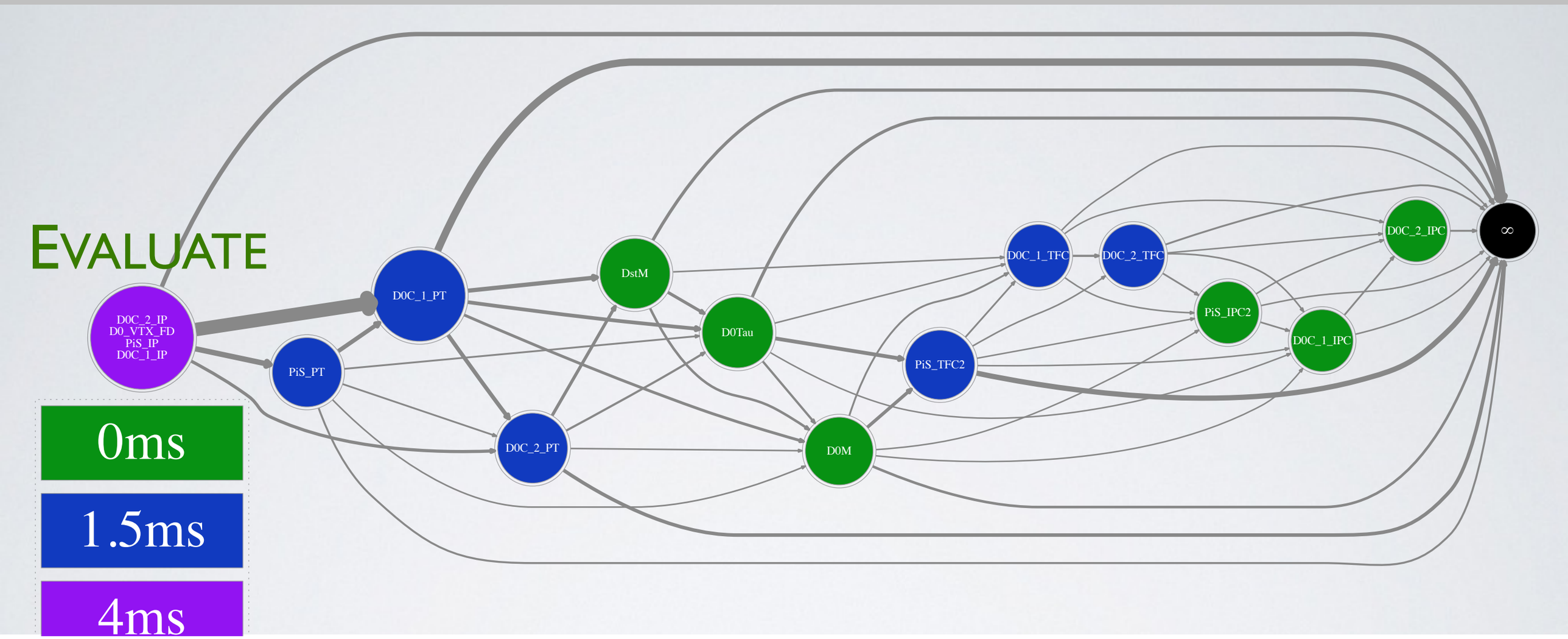
MDDAG: A SIGNAL/BCKG DECISION GRAPH

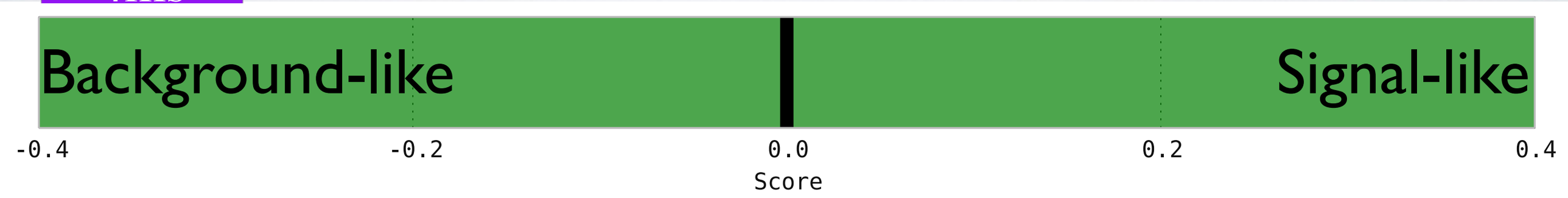


Hard background

Benbouzid et al. ICML 2012

MDDAG: A SIGNAL/BCKG DECISION GRAPH

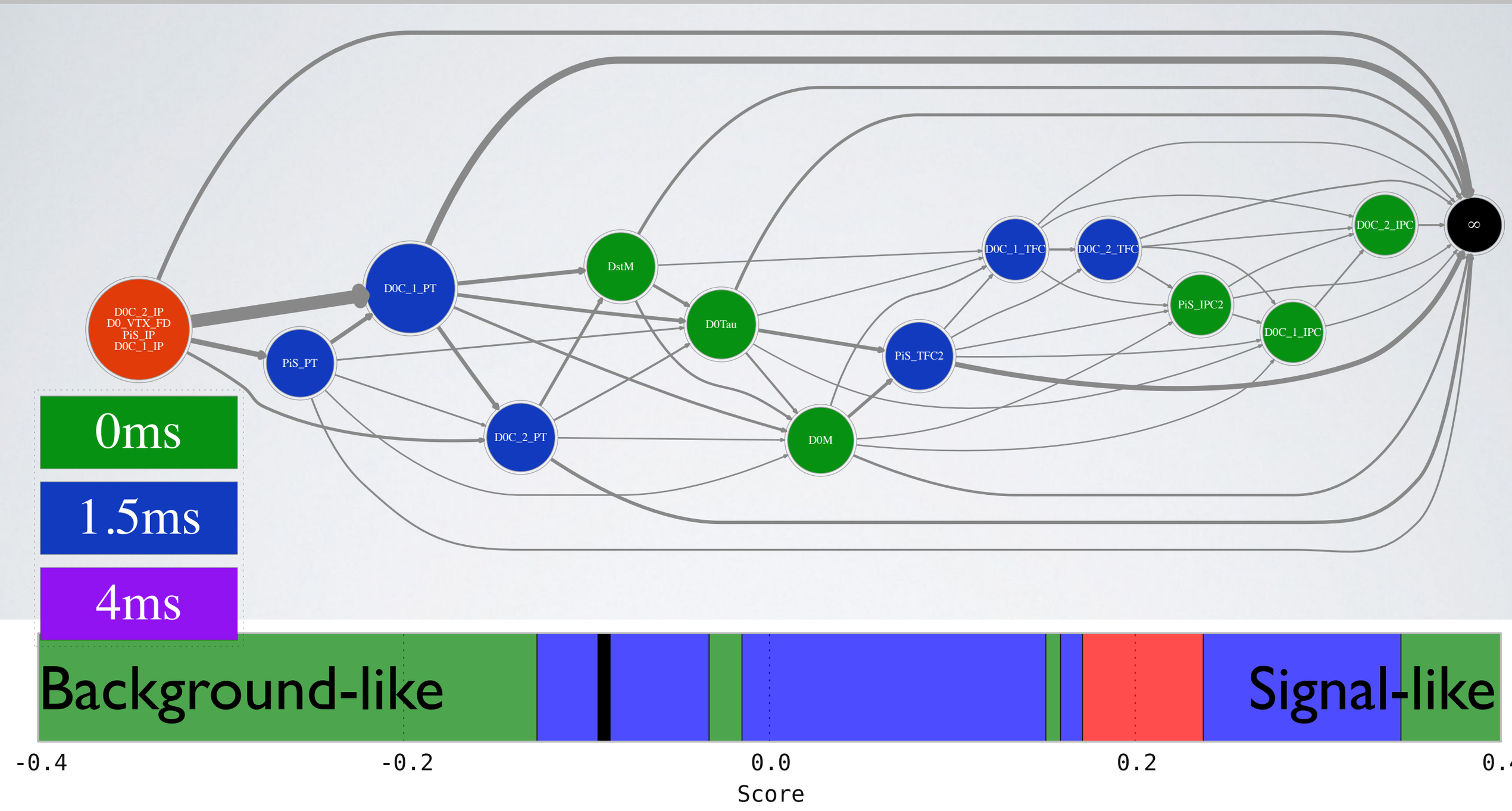




Hard background

Benbouzid et al. ICML 2012

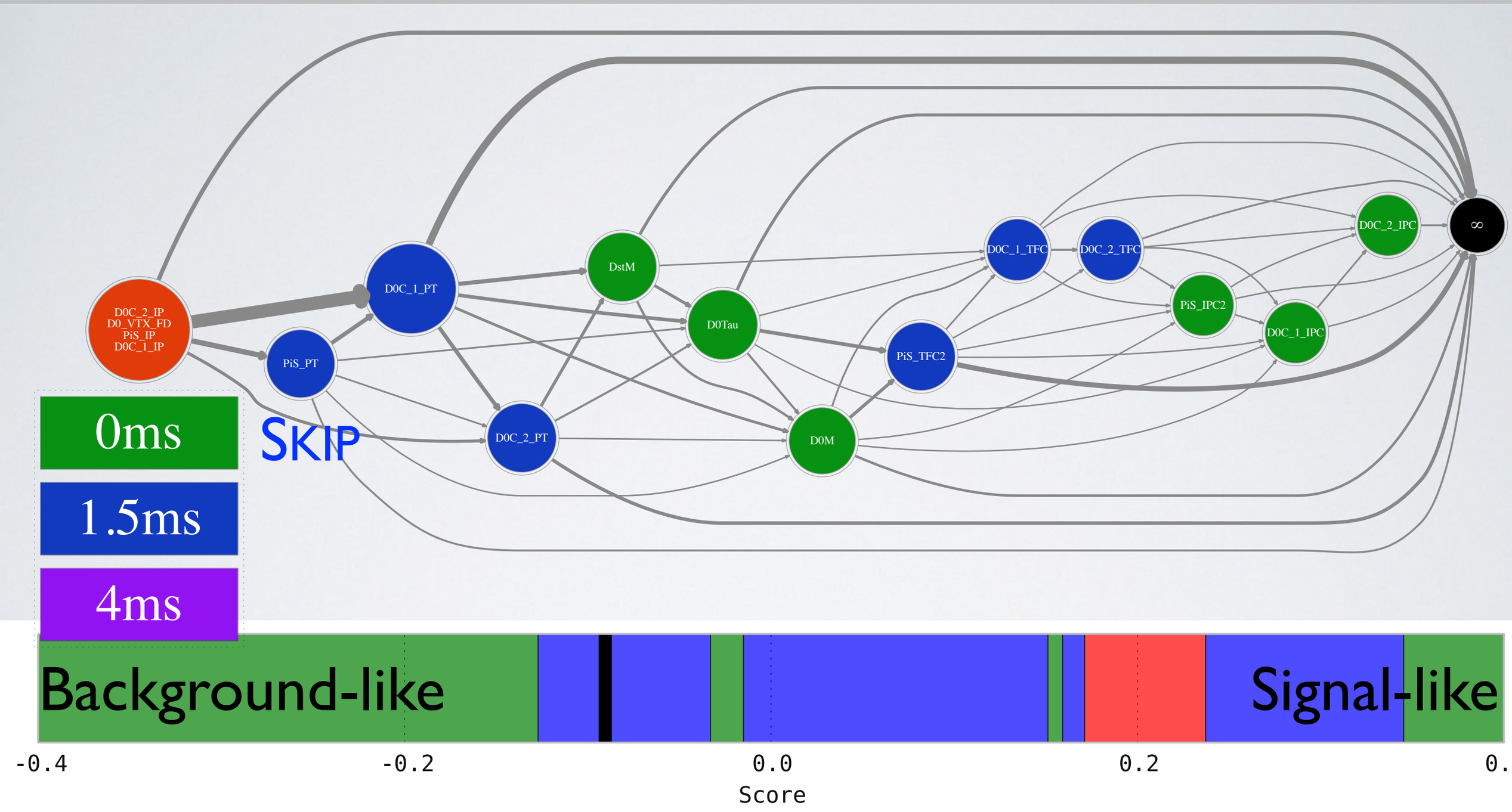
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

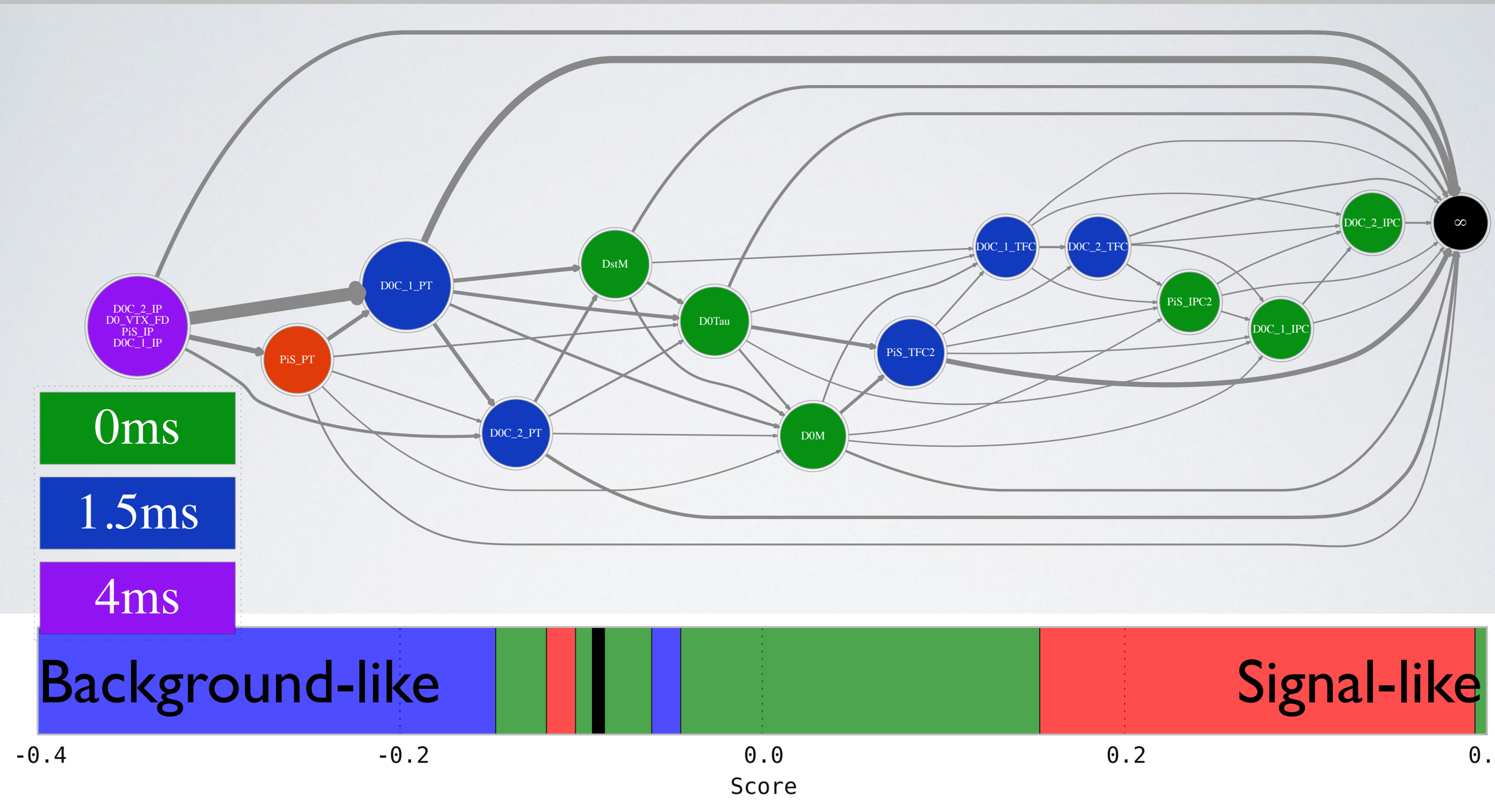
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

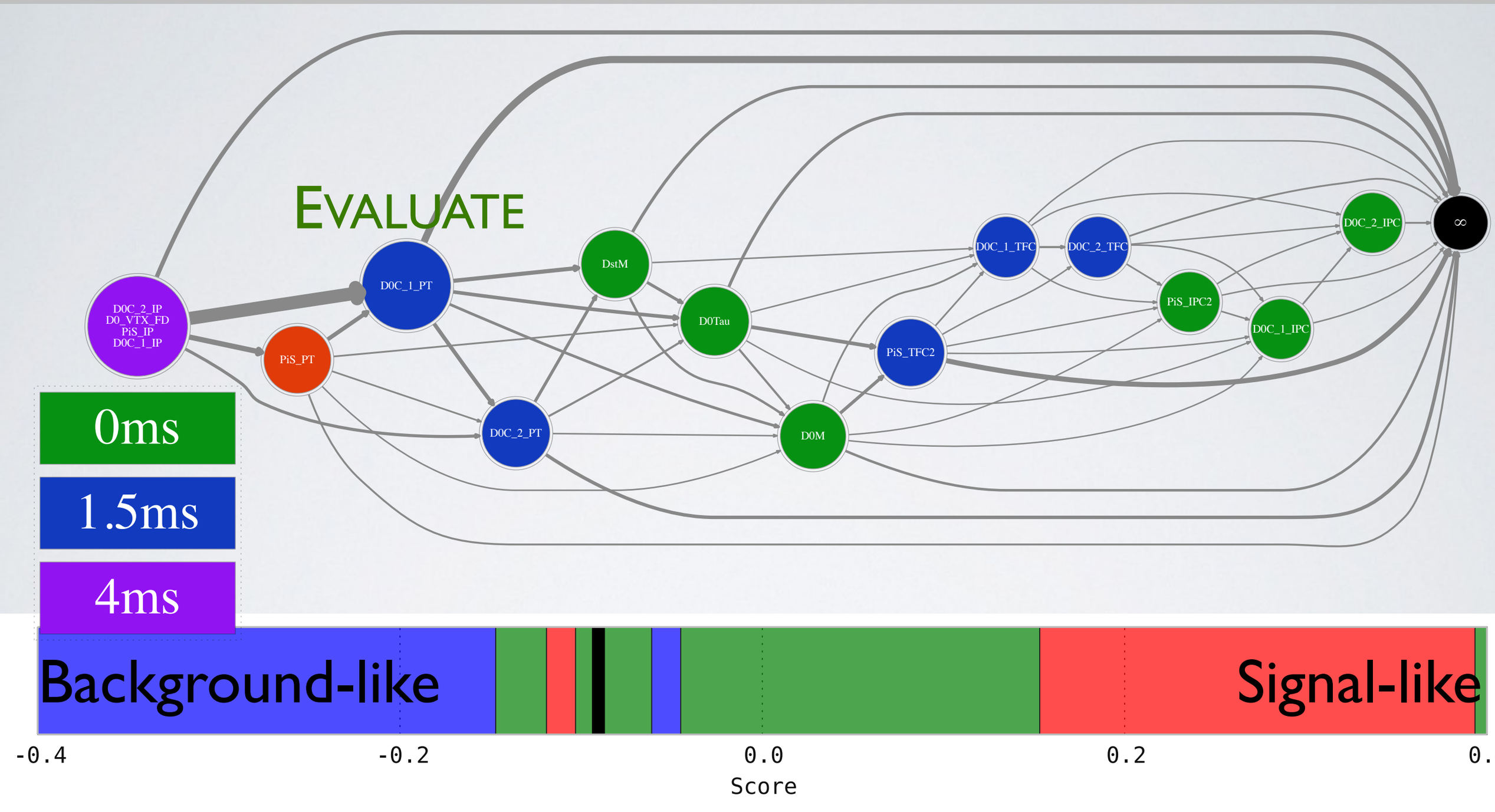
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

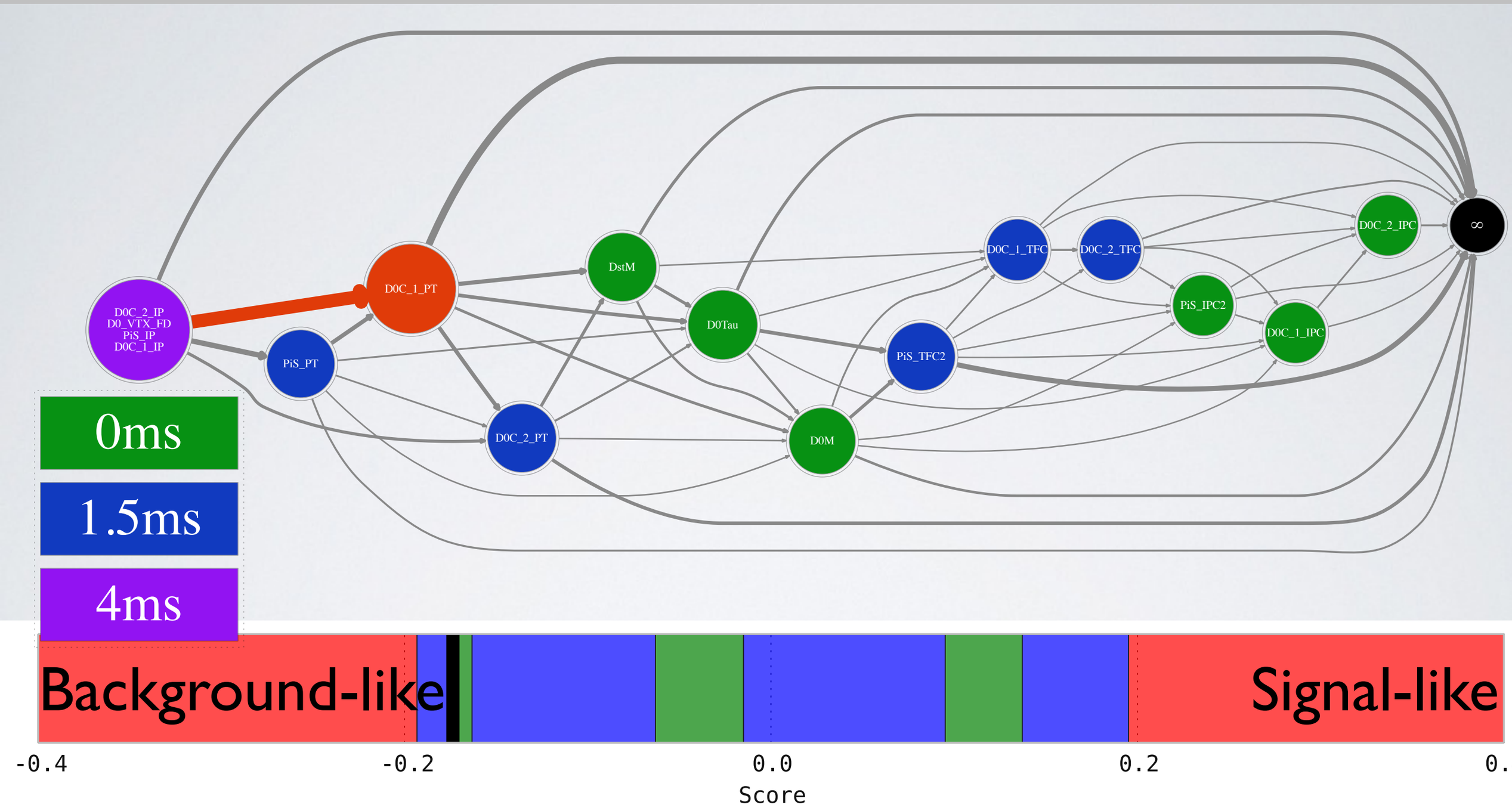
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

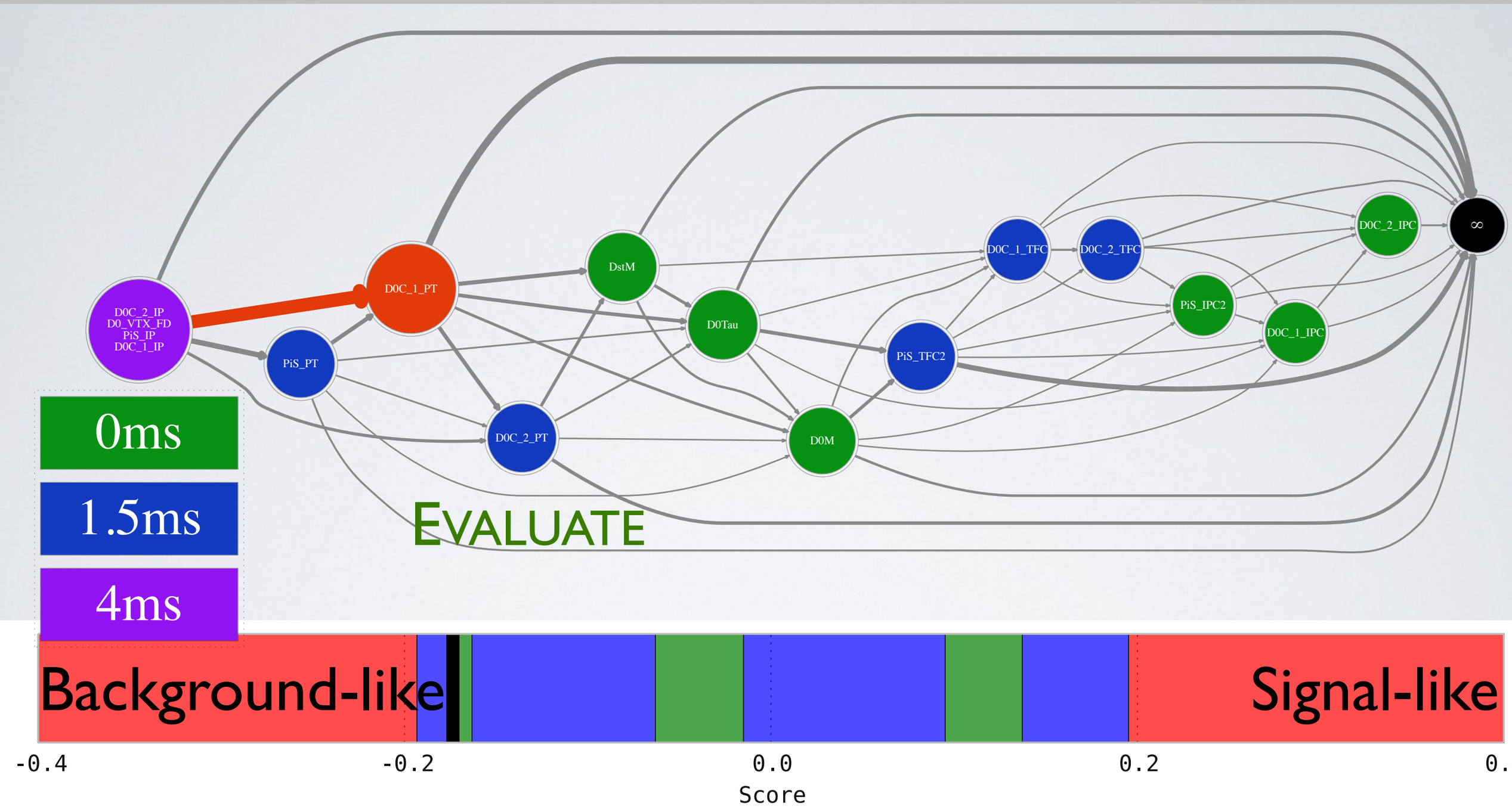
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

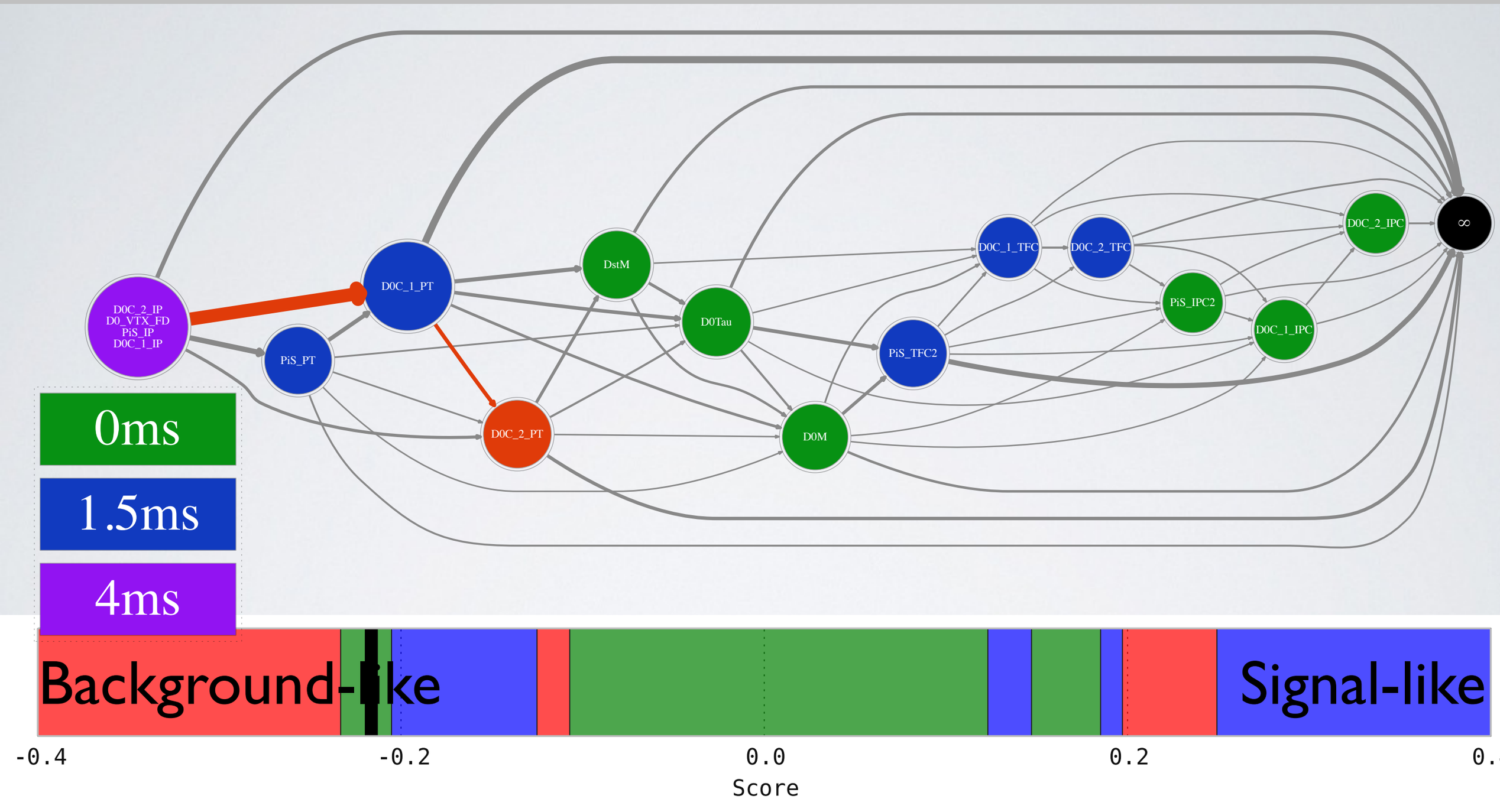
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

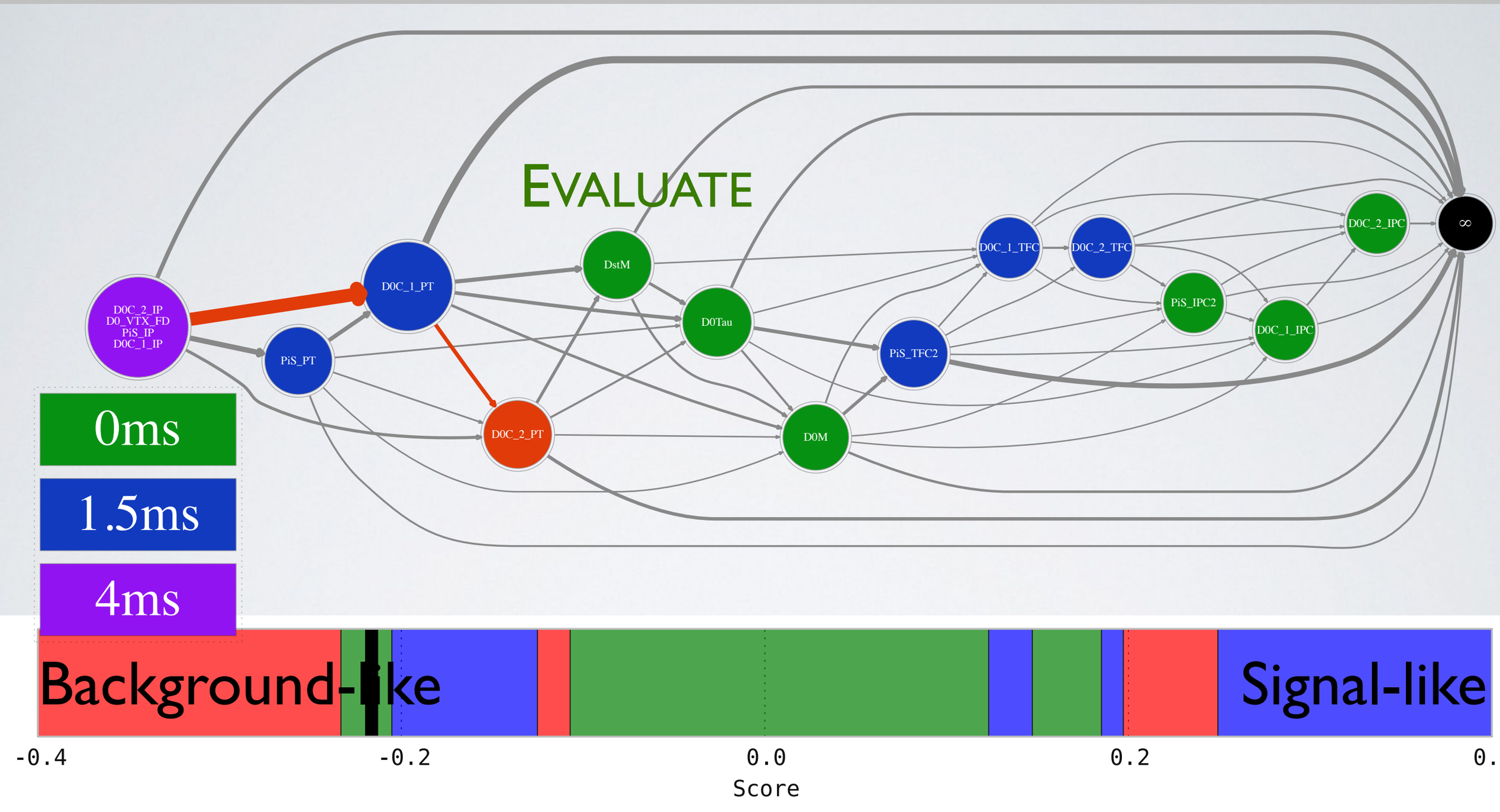
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

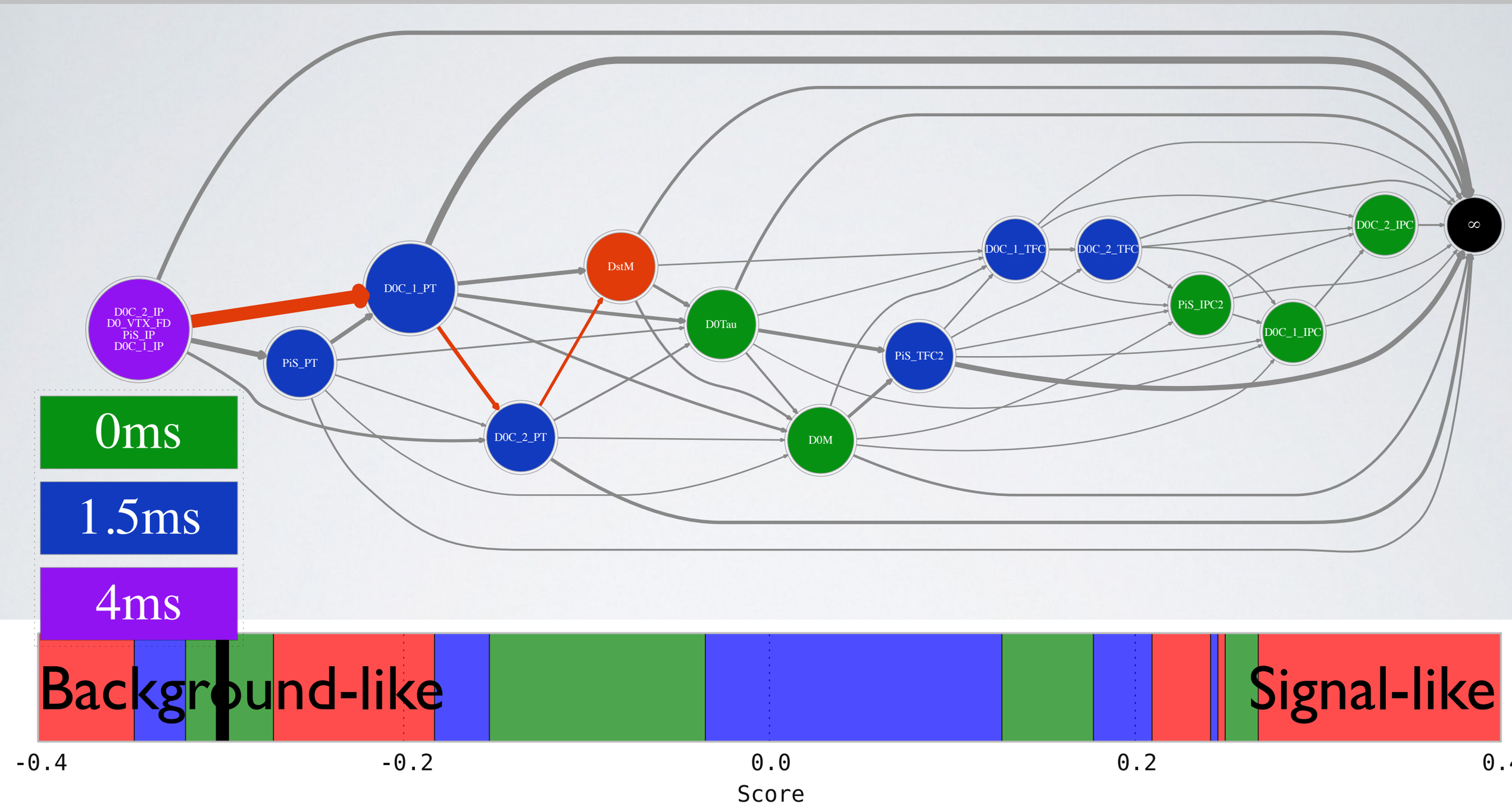
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

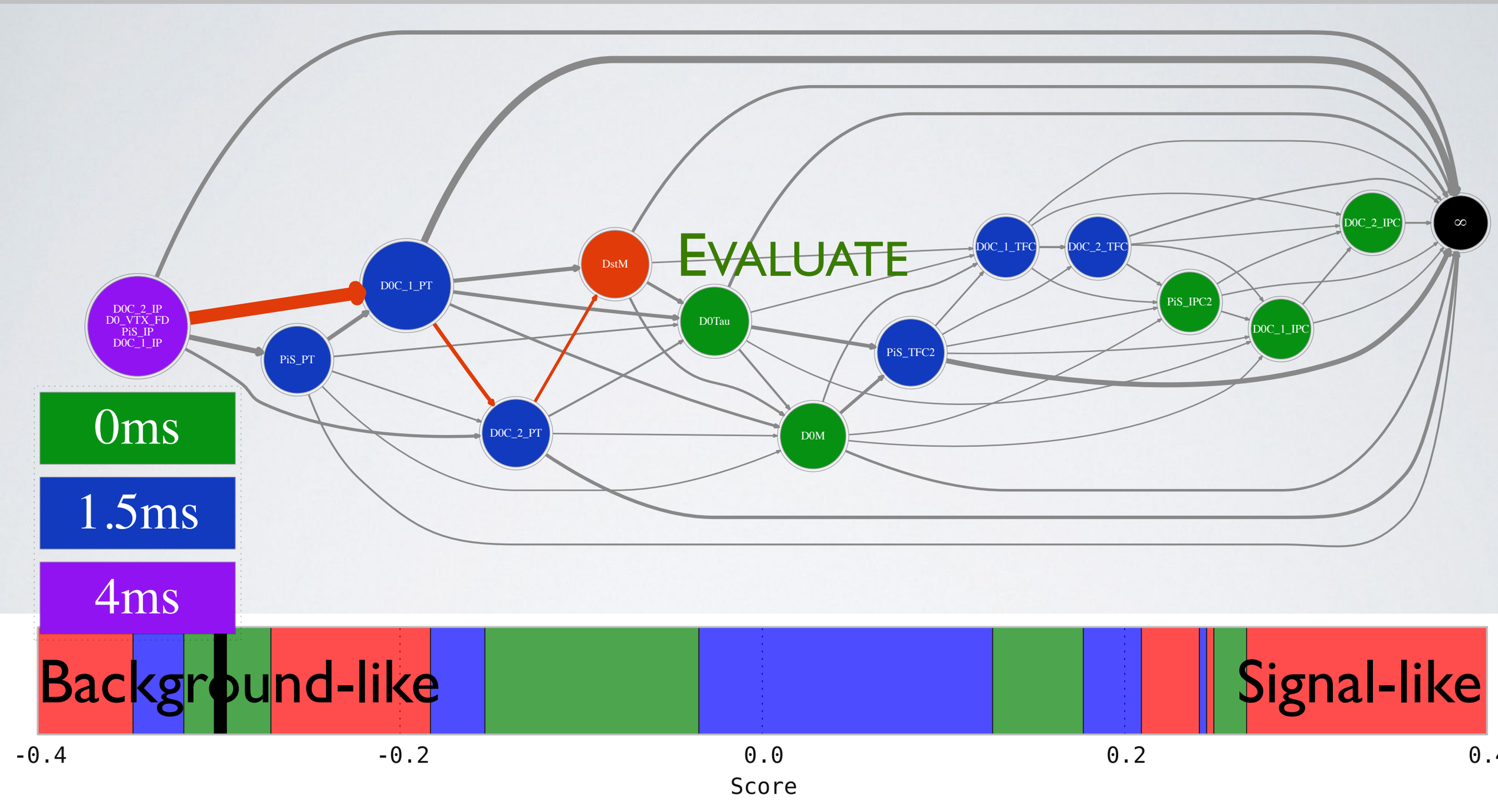
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

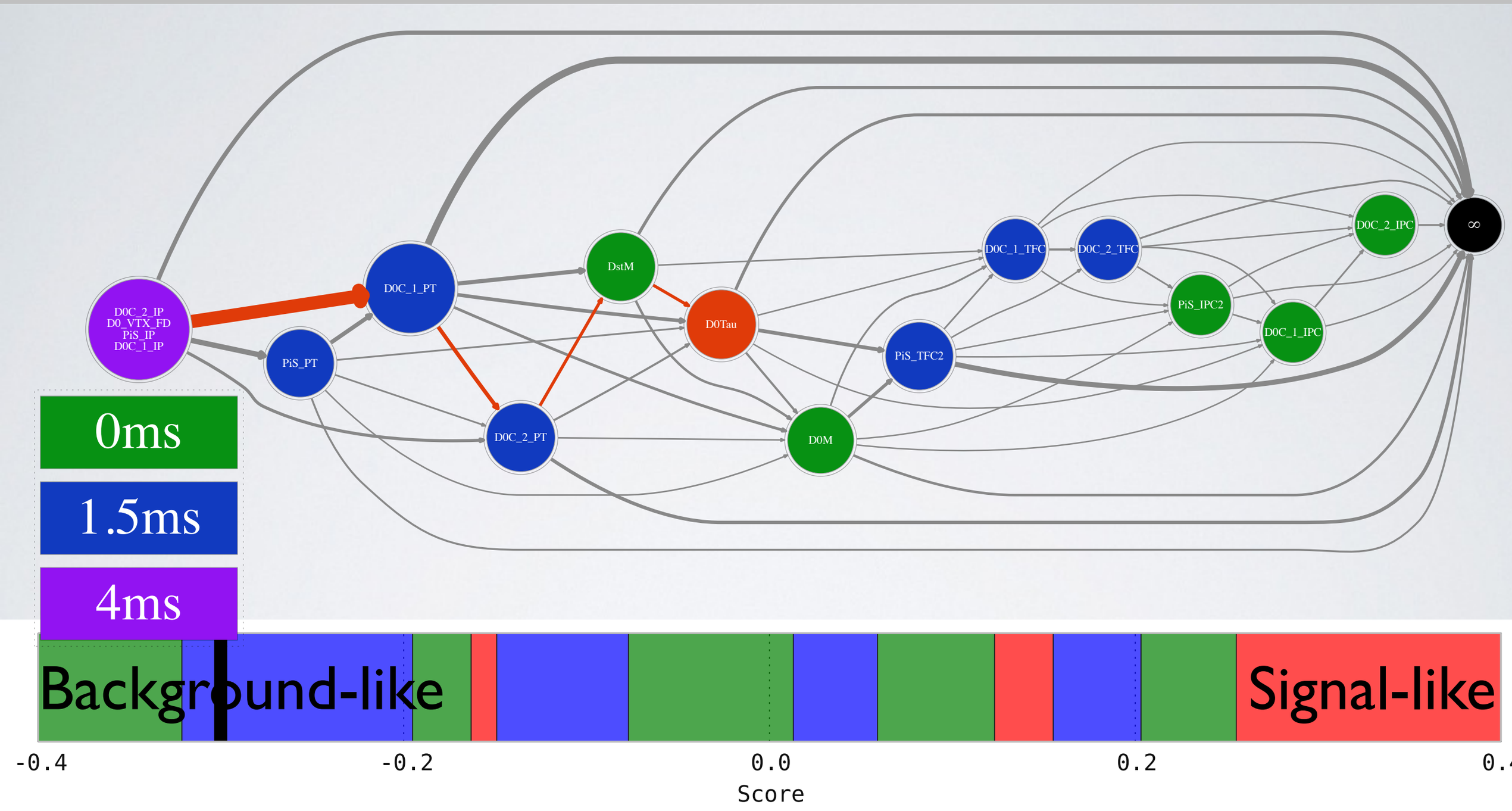
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

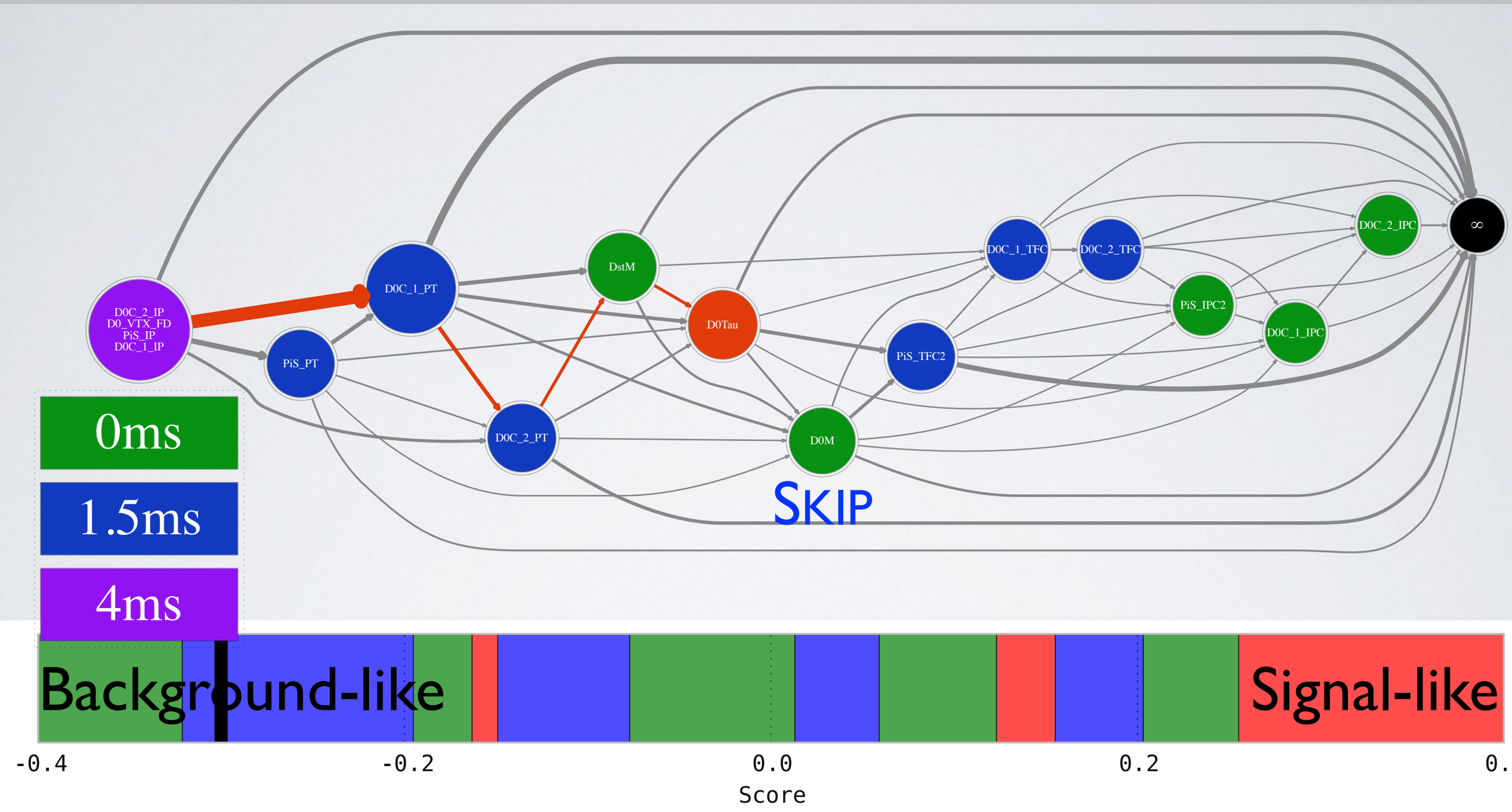
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

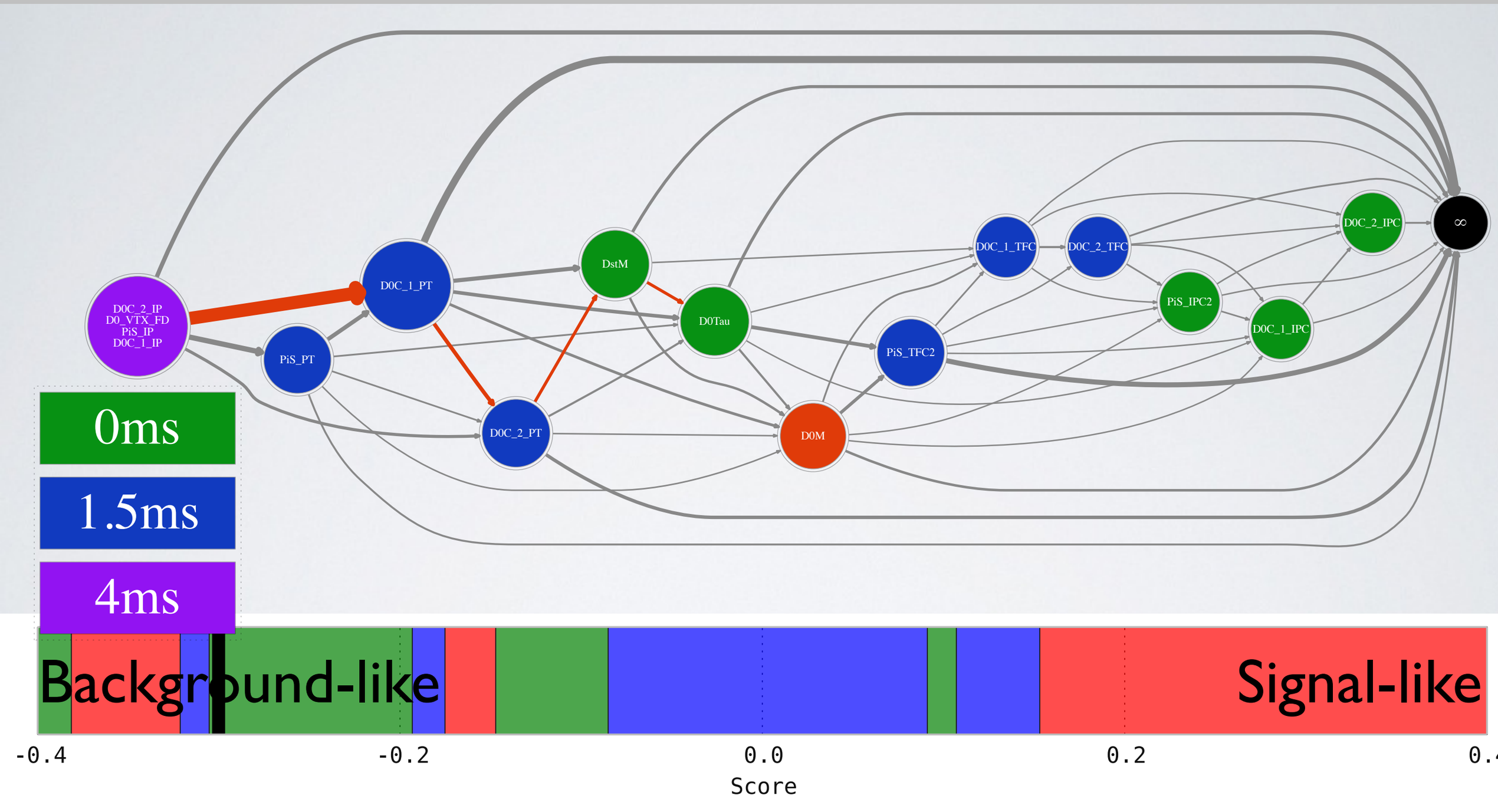
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

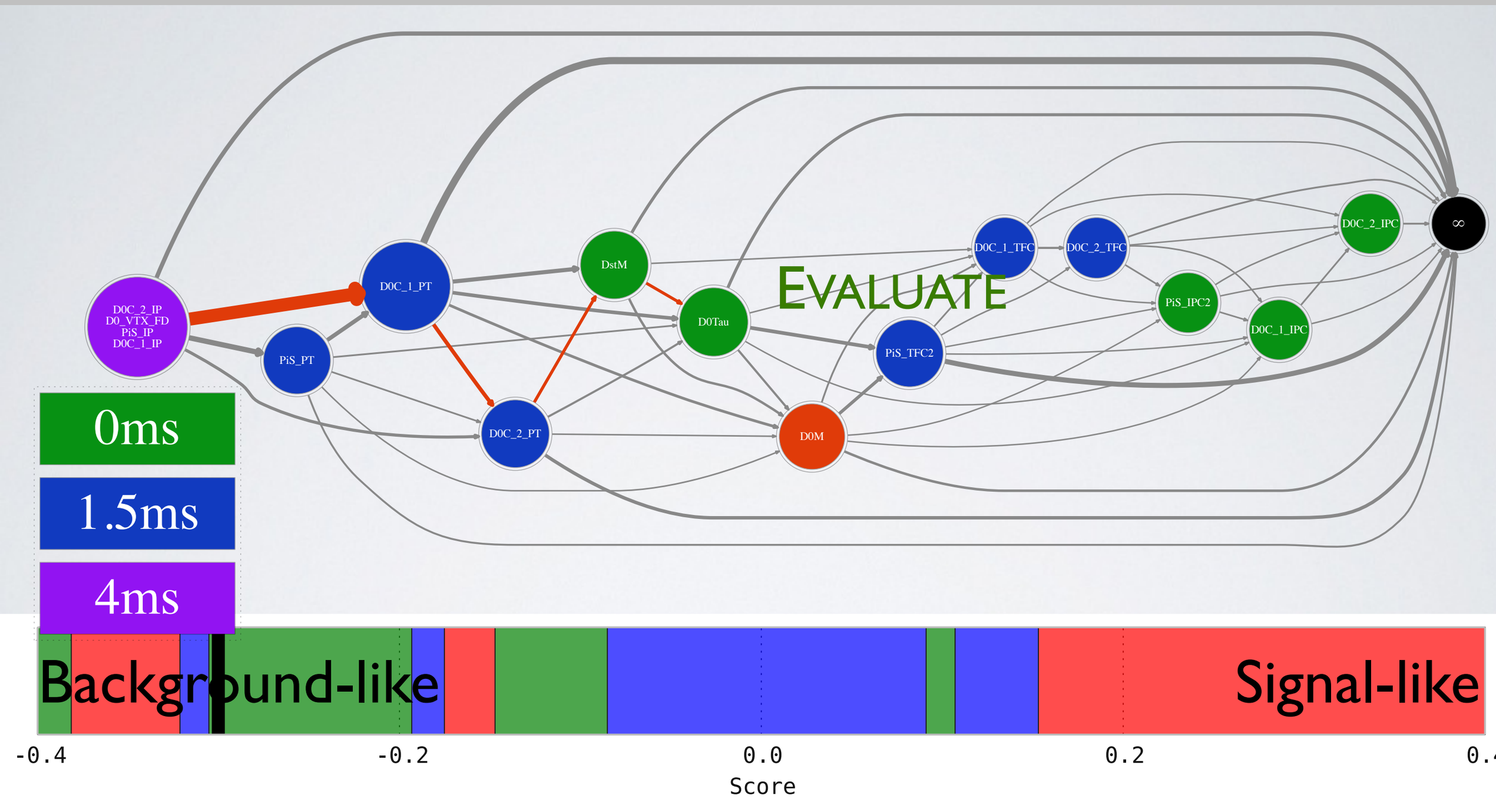
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

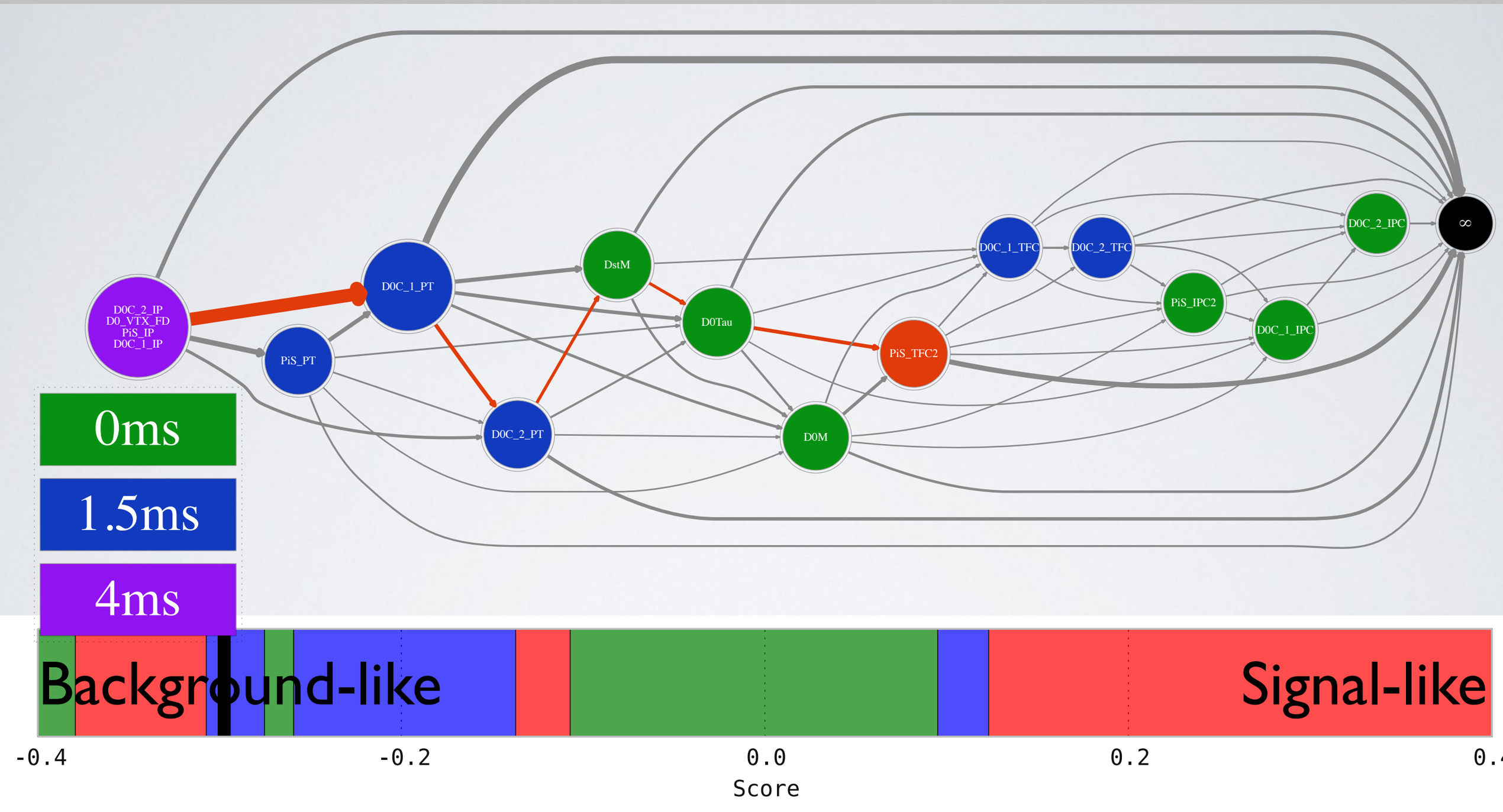
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

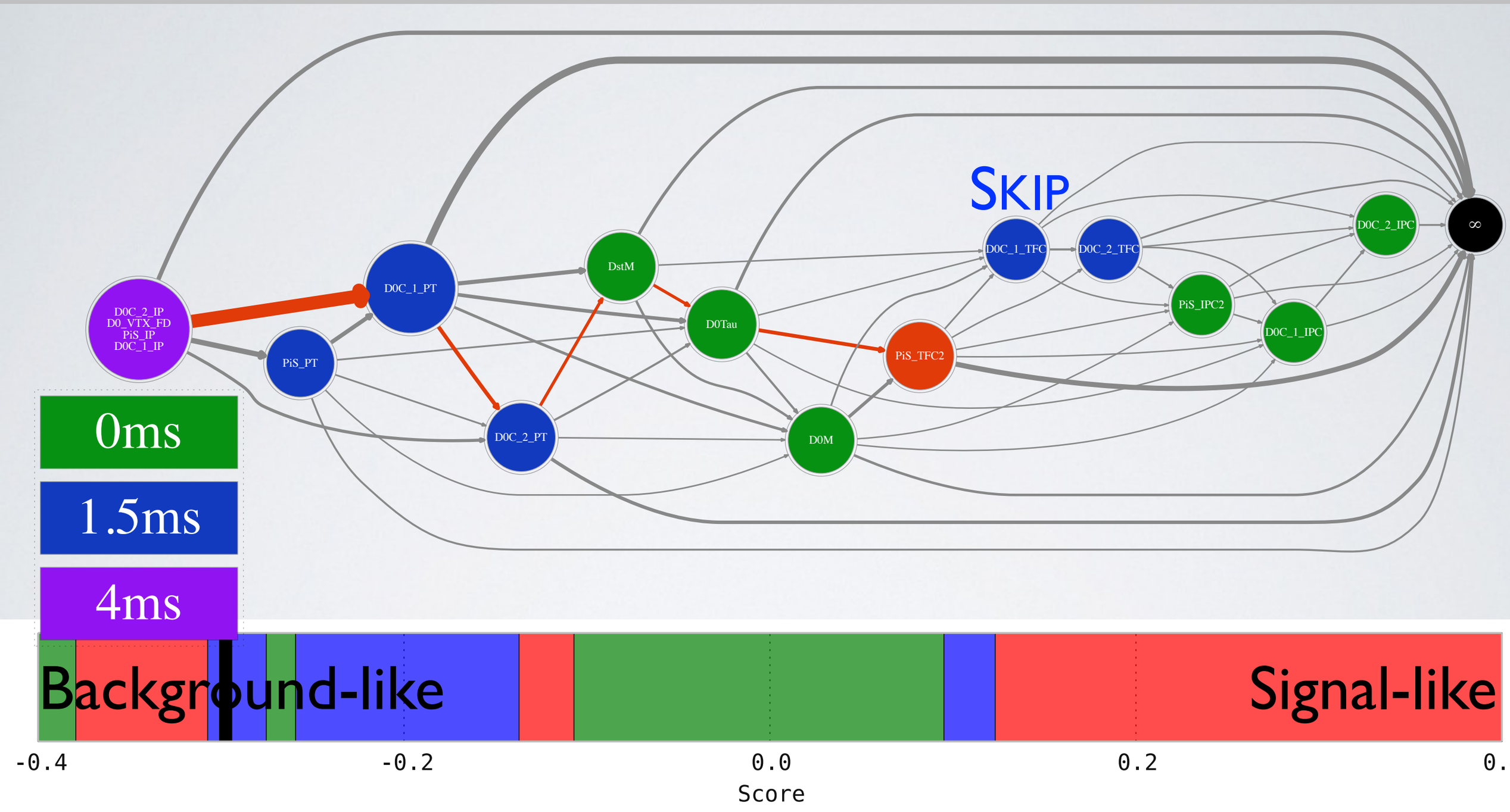
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

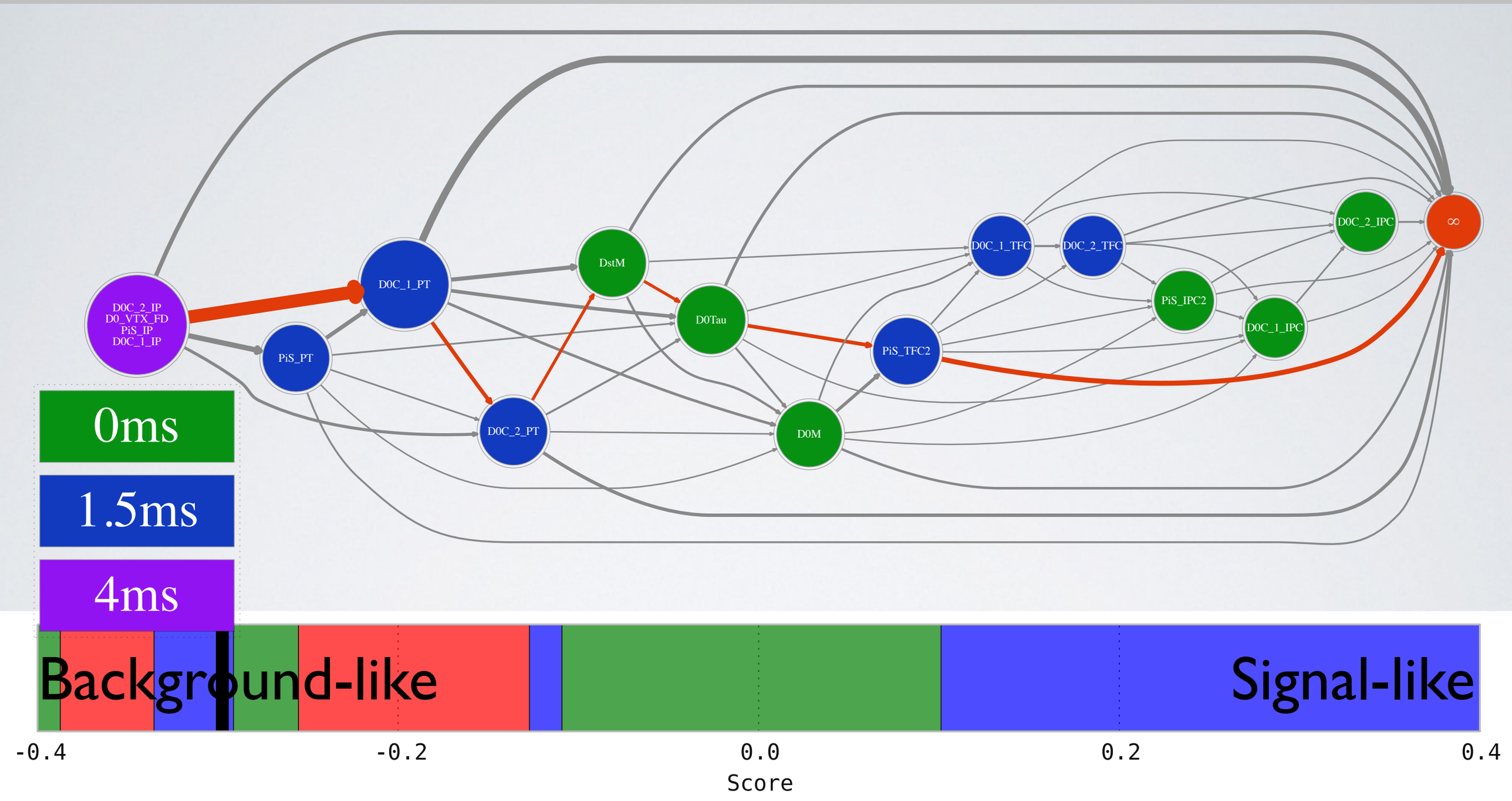
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

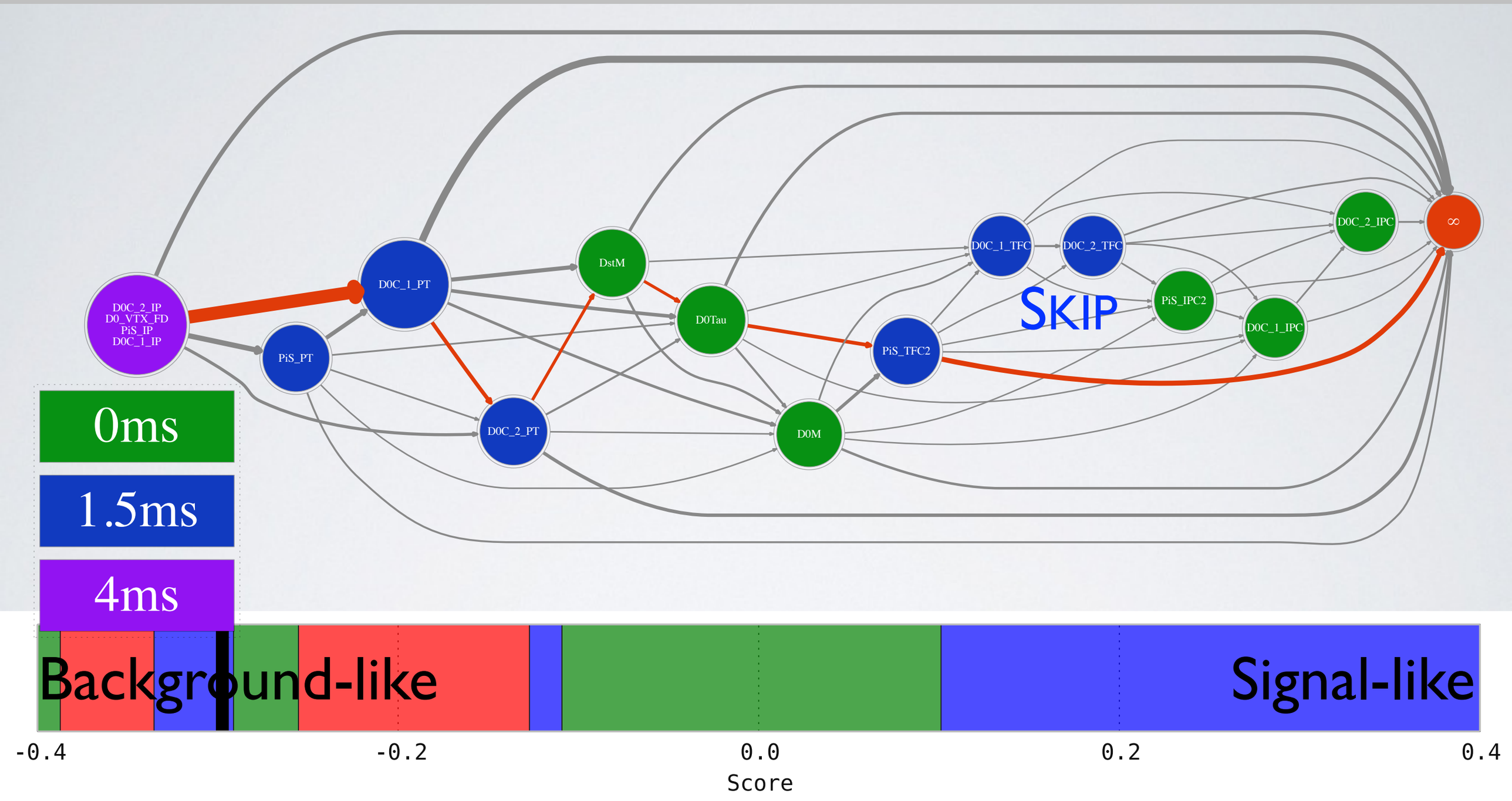
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

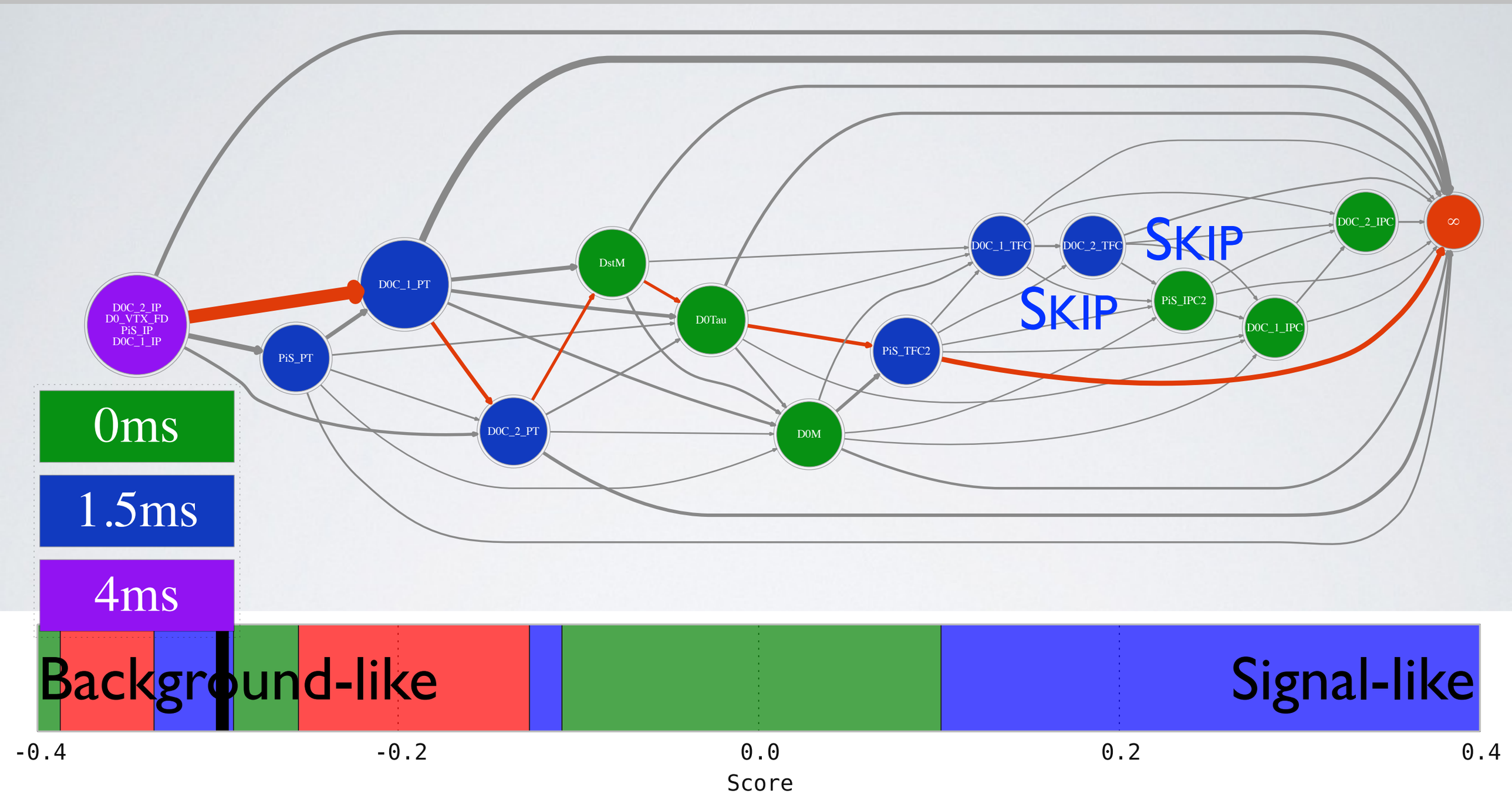
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

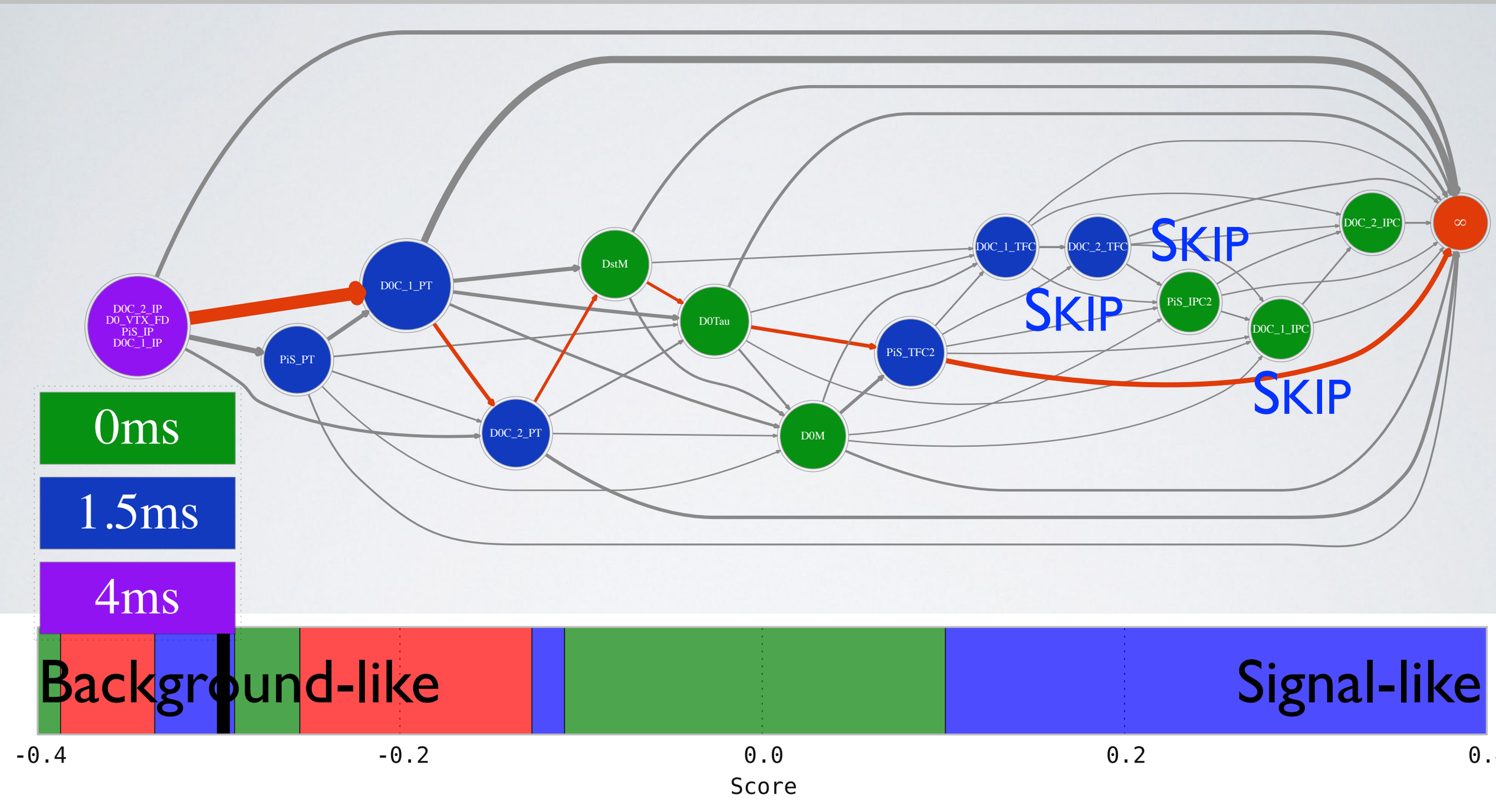
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

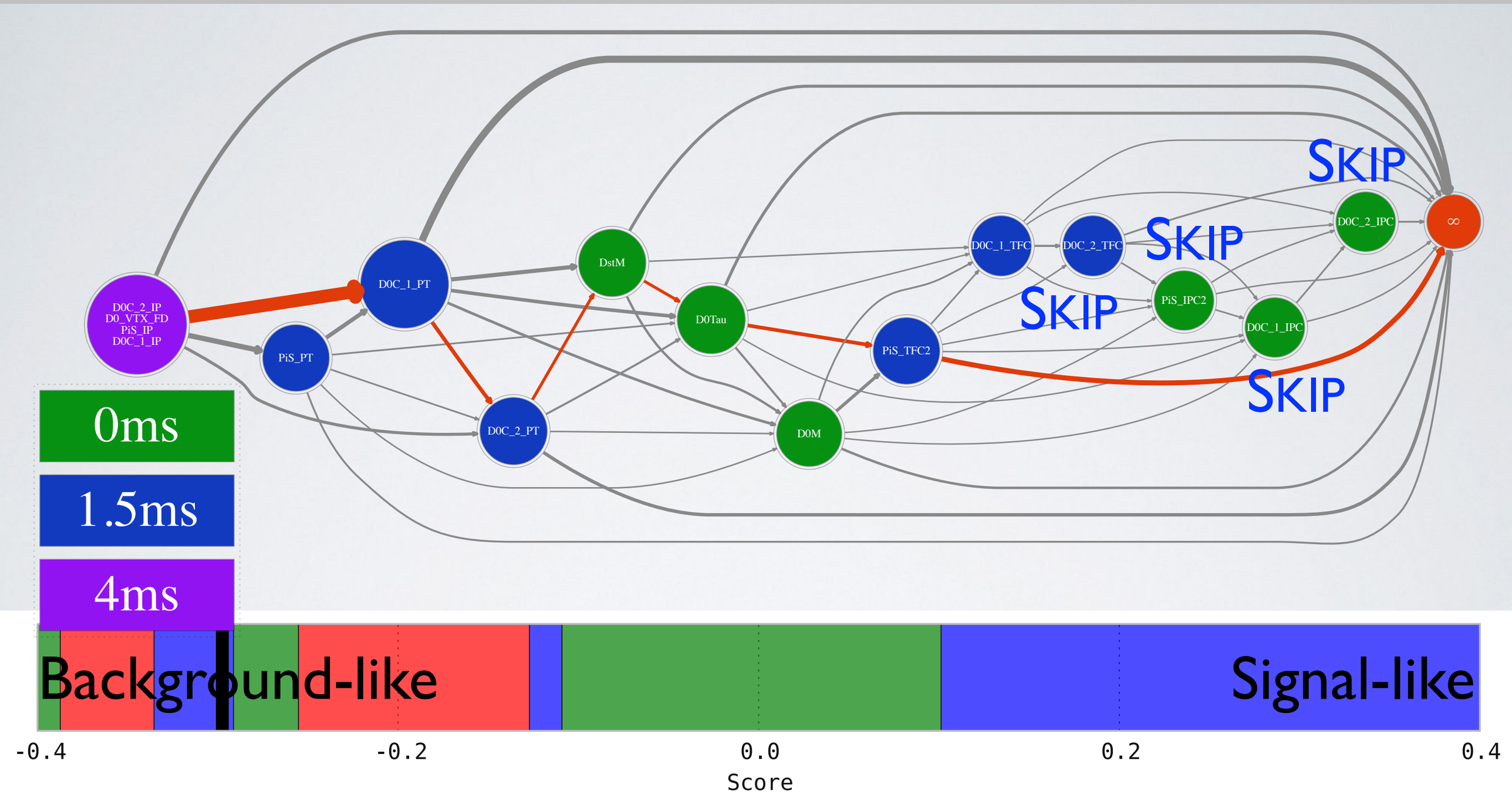
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

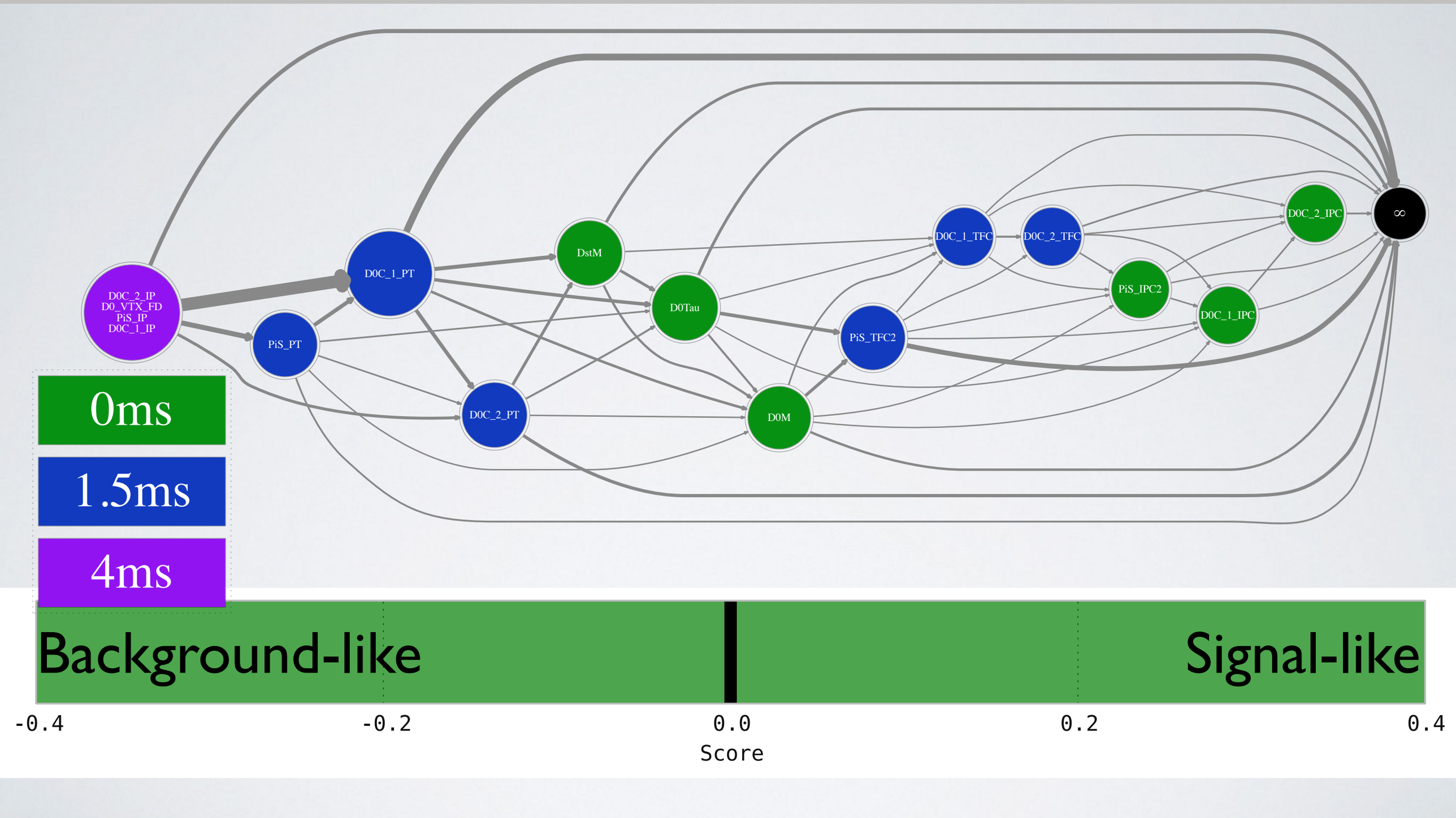
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard background

Benbouzid et al. ICML 2012

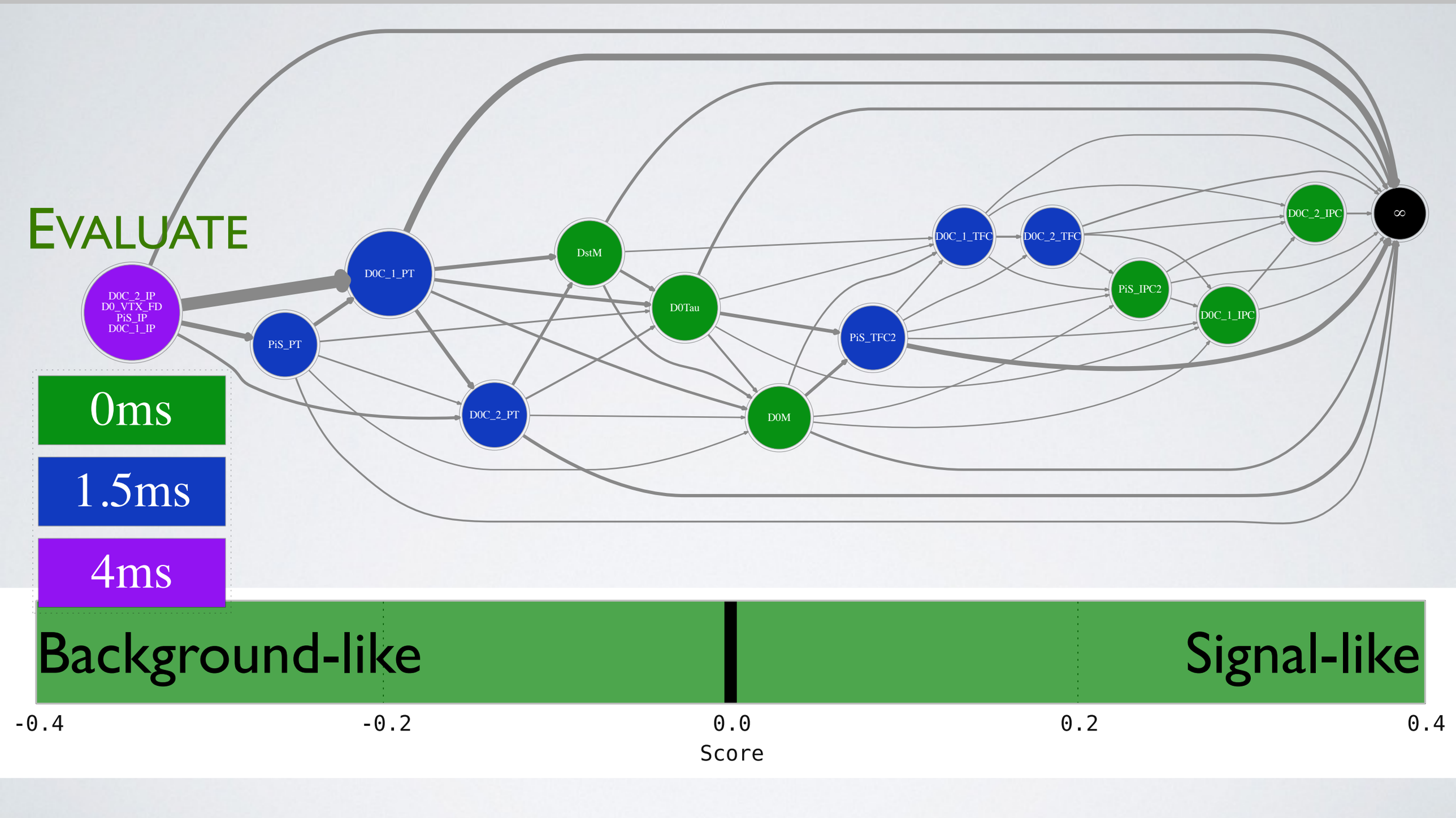
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy signal

Benbouzid et al. ICML 2012

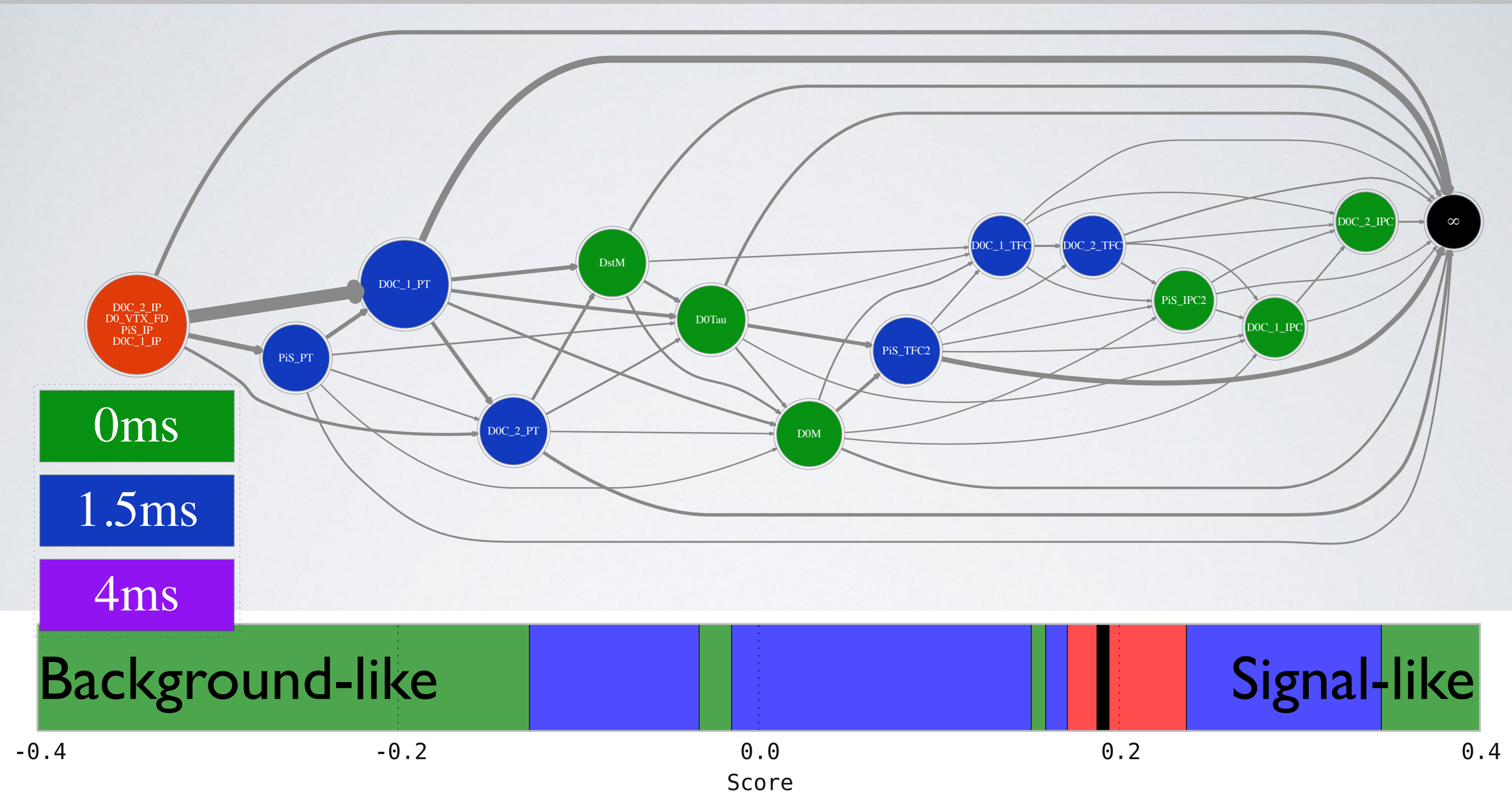
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy signal

Benbouzid et al. ICML 2012

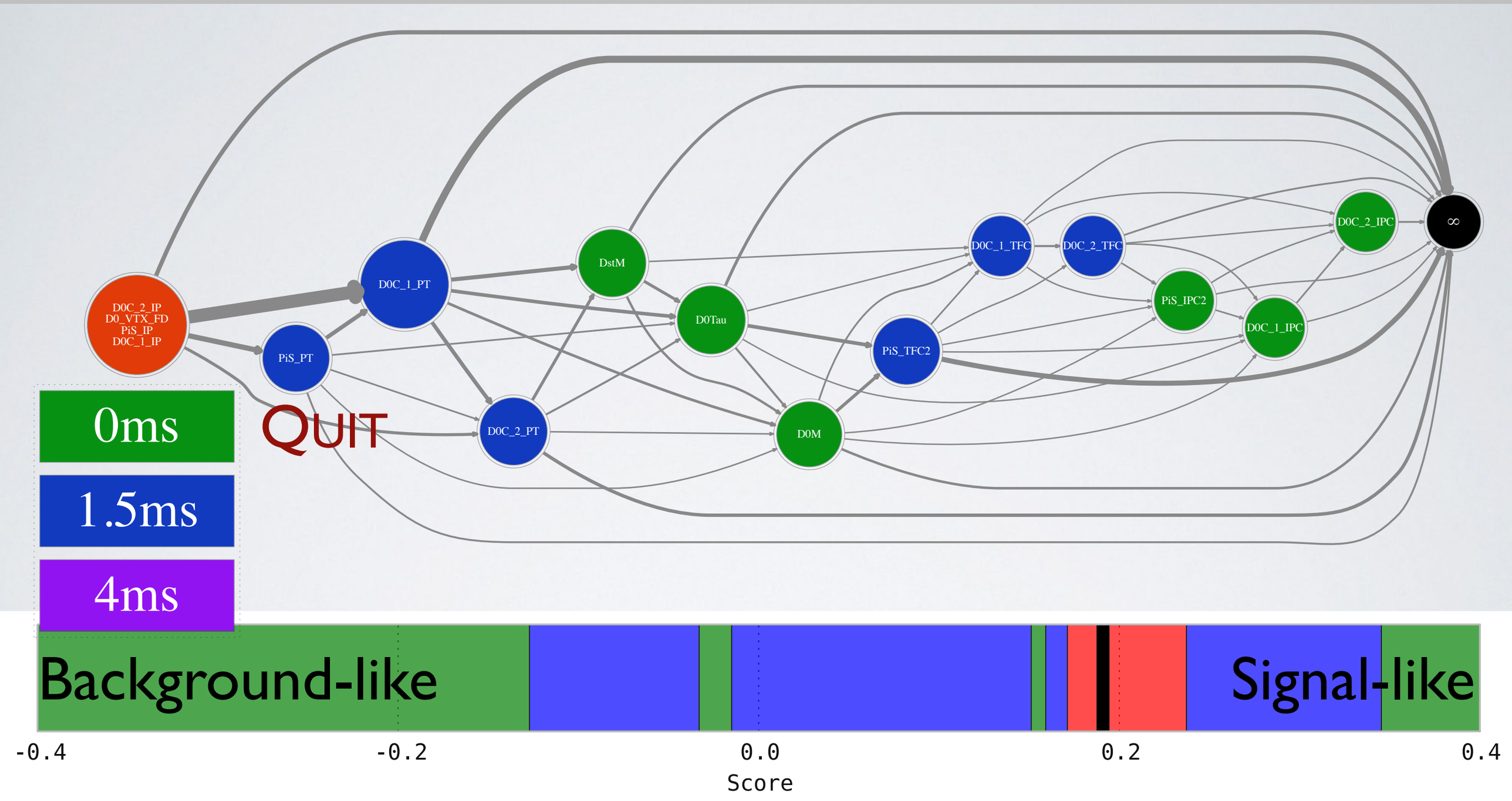
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy signal

Benbouzid et al. ICML 2012

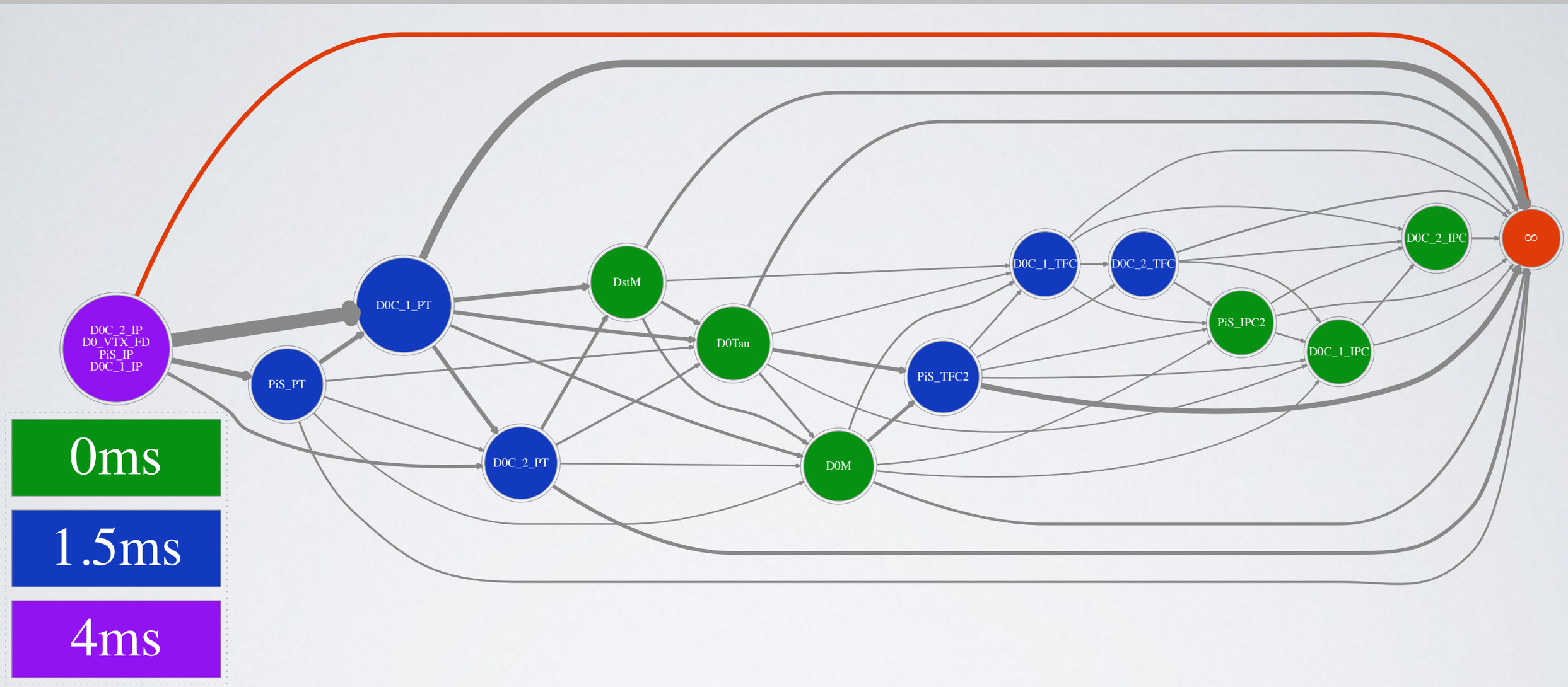
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy signal

Benbouzid et al. ICML 2012

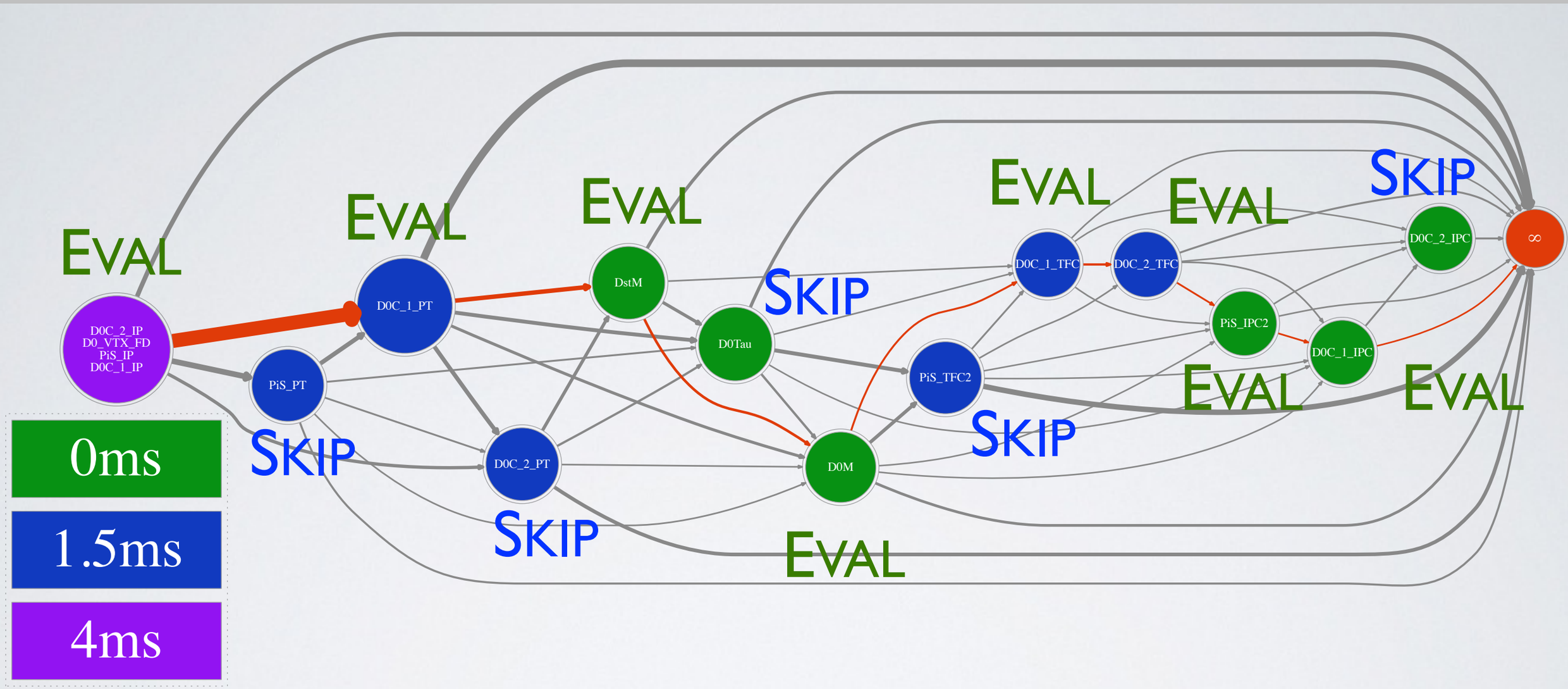
MDDAG: A SIGNAL/BCKG DECISION GRAPH



Easy signal

Benbouzid et al. ICML 2012

MDDAG: A SIGNAL/BCKG DECISION GRAPH



Hard signal

Benbouzid et al. ICML 2012

BUDGETED CLASSIFICATION

- Classification with test-time constraints
- An active research area due to IT applications
- To be exploited for trigger design

BUDGETED CLASSIFICATION

Take-home message:

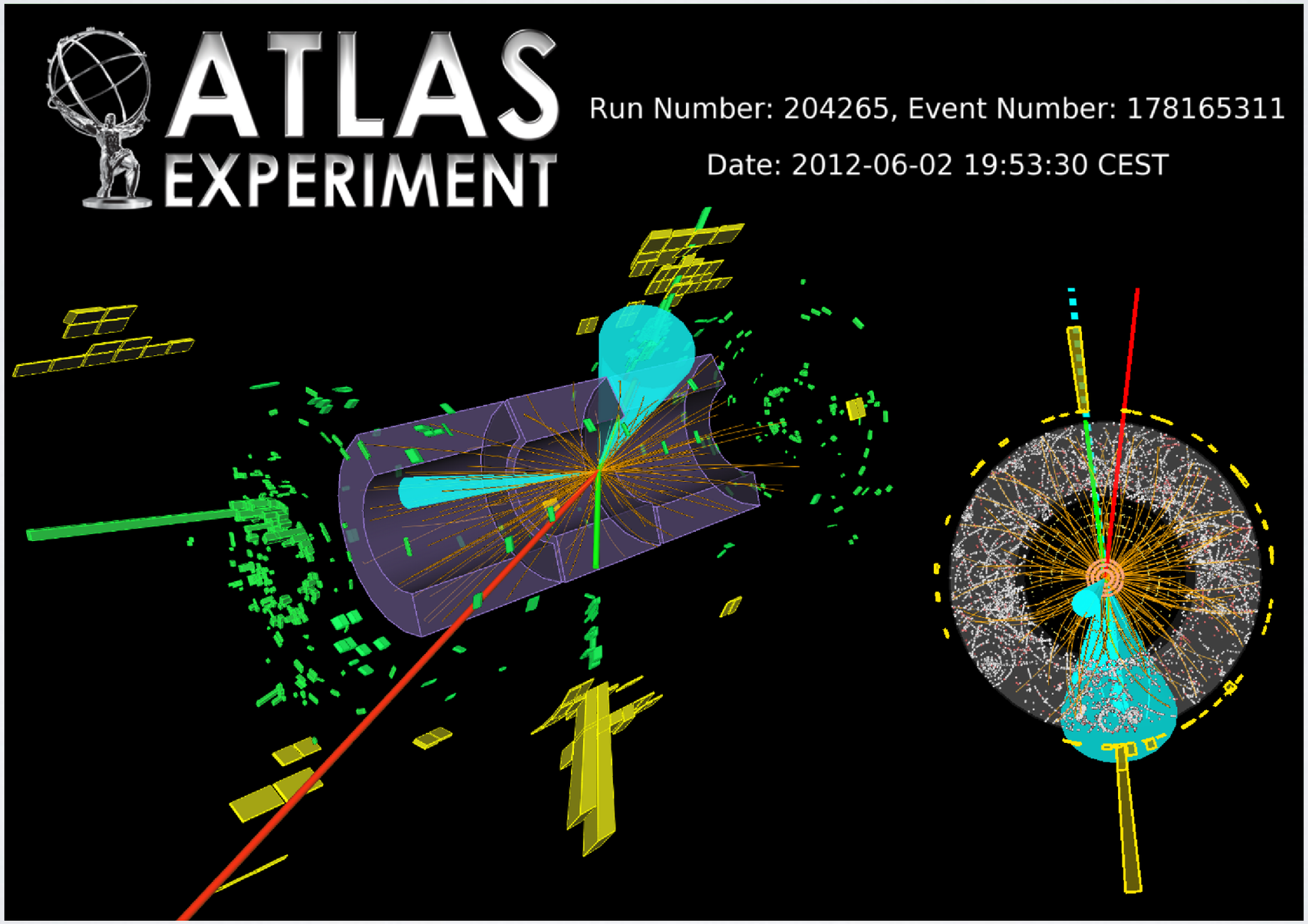
Once you go software trigger,
the set of possible solutions opens up

HEP-INSPIRED OBJECTIVES

- Low error probability ~ high accuracy ~ high efficiency
- Fast classification
- More exotic goals (or constraints):
  - easy control of detection efficiency in different signal or background classes
  - efficiency unbiased in certain variables
  - feeding the classifier into a statistical (counting) test: maximize test sensitivity
  - in the big data regime systematics are becoming more important than statistical efficiency

CLASSIFICATION FOR DISCOVERY

The HiggsML challenge



CLASSIFICATION FOR DISCOVERY

CLASSIFICATION FOR DISCOVERY

- In a nutshell

CLASSIFICATION FOR DISCOVERY

- In a nutshell
  - A vector ***x*** of variables is extracted from each event

CLASSIFICATION FOR DISCOVERY

- In a nutshell
  - A vector ***x*** of variables is extracted from each event
  - A classifier ***g(x)*** is trained to separate signal from background

CLASSIFICATION FOR DISCOVERY

- In a nutshell
  - A vector ***x*** of variables is extracted from each event
  - A classifier *g(x)* is trained to separate signal from background
  - The background *b* is estimated in the selection region
    $G = \{x : g(x) = s\}$

CLASSIFICATION FOR DISCOVERY

- In a nutshell
  - A vector ***x*** of variables is extracted from each event
  - A classifier *g(x)* is trained to separate signal from background
  - The background *b* is estimated in the selection region
    $G = \{x : g(x) = s\}$
  - Discovery is made when the number of real events *n* is significantly higher than *b*

CLASSIFICATION FOR DISCOVERY

CLASSIFICATION FOR DISCOVERY

- Exciting physics

CLASSIFICATION FOR DISCOVERY

- Exciting physics
  - The Higgs to tau-tau excess is not yet at five sigma
    Tech. Rep. ATLAS-CONF-2013-108

CLASSIFICATION FOR DISCOVERY

- Exciting physics
  - The Higgs to tau-tau excess is not yet at five sigma
    Tech. Rep. ATLAS-CONF-2013-108
- Exciting data science (statistics and machine learning)

CLASSIFICATION FOR DISCOVERY

- Exciting physics
  - The Higgs to tau-tau excess is not yet at five sigma
    Tech. Rep. ATLAS-CONF-2013-108
- Exciting data science (statistics and machine learning)
  - What is the theoretical relationship between classification and test sensitivity?

CLASSIFICATION FOR DISCOVERY

- Exciting physics
  - The Higgs to tau-tau excess is not yet at five sigma
    Tech. Rep. ATLAS-CONF-2013-108
- Exciting data science (statistics and machine learning)
  - What is the theoretical relationship between classification and test sensitivity?
  - What is the quantitative criteria to optimize?

CLASSIFICATION FOR DISCOVERY

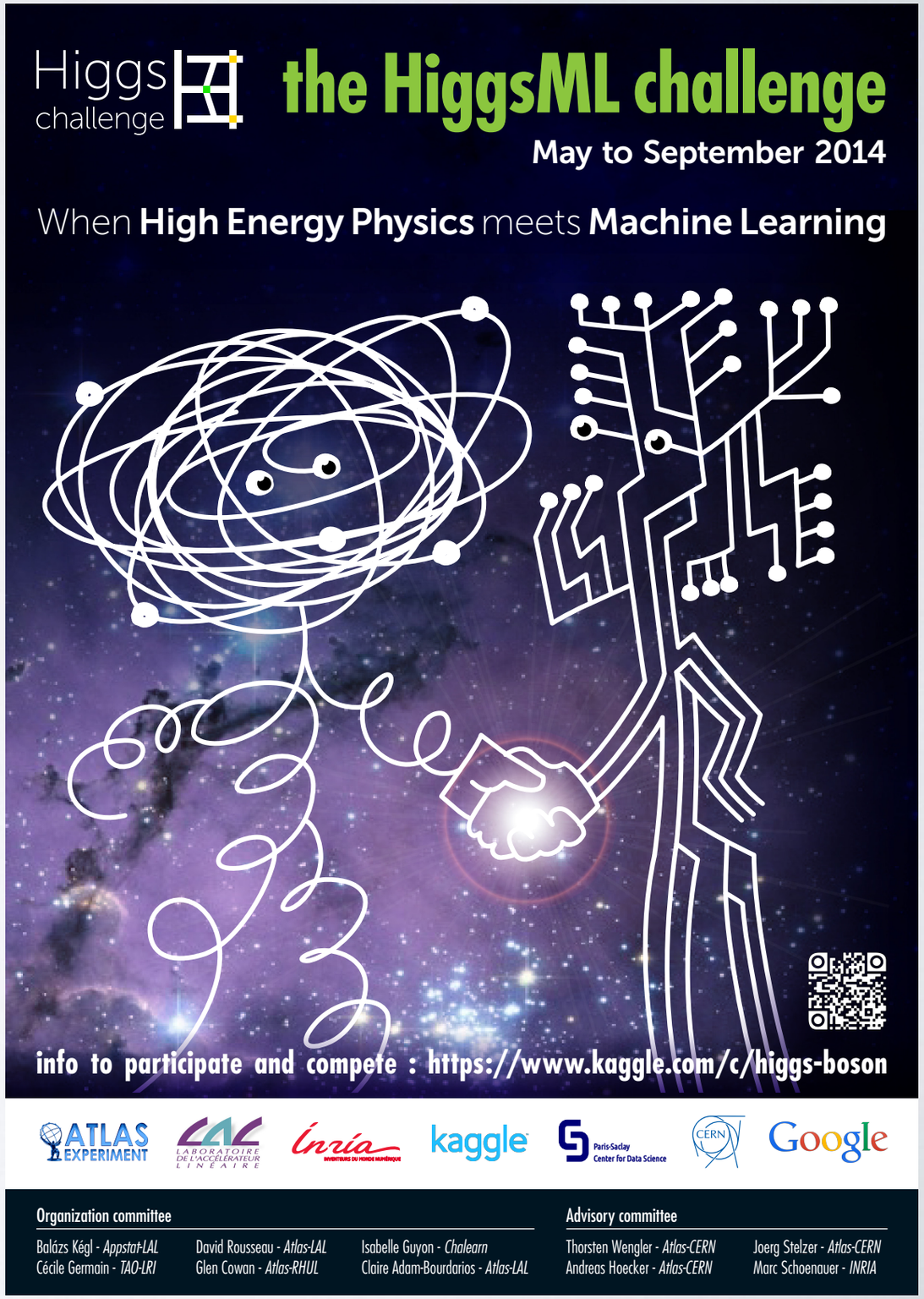
- Exciting physics
  - The Higgs to tau-tau excess is not yet at five sigma
    Tech. Rep. ATLAS-CONF-2013-108
- Exciting data science (statistics and machine learning)
  - What is the theoretical relationship between classification and test sensitivity?
  - What is the quantitative criteria to optimize?
  - How to formally include systematic uncertainties?

CLASSIFICATION FOR DISCOVERY

- Exciting physics
  - The Higgs to tau-tau excess is not yet at five sigma
    Tech. Rep. ATLAS-CONF-2013-108
- Exciting data science (statistics and machine learning)
  - What is the theoretical relationship between classification and test sensitivity?
  - What is the quantitative criteria to optimize?
  - How to formally include systematic uncertainties?
  - Can we redesign classical algorithms (boosting, SVM, neural nets) for optimizing this criteria?

CLASSIFICATION FOR DISCOVERY

We are running a
data challenge
to answer some of these
questions



CLASSIFICATION FOR DISCOVERY

The formal setup

- We simulate data: $\mathcal{D} = \{(\mathbf{x}_1, y_1, w_1), \ldots, (\mathbf{x}_n, y_n, w_n)\}$
  - $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector
  - $\mathbf{y}_i \in \{\mathsf{b}\text{ackground}, \mathsf{s}\text{ignal}\}$ is the label
  - $w_i \in \mathbb{R}^+$ is a non-negative weight (importance sampling)
  - let $\mathcal{S} = \{i : y_i = \mathsf{s}\}$ and $\mathcal{B} = \{i : y_i = \mathsf{b}\}$ be the index sets of signal and background events, respectively
- Maximize the Approximate Median Significance

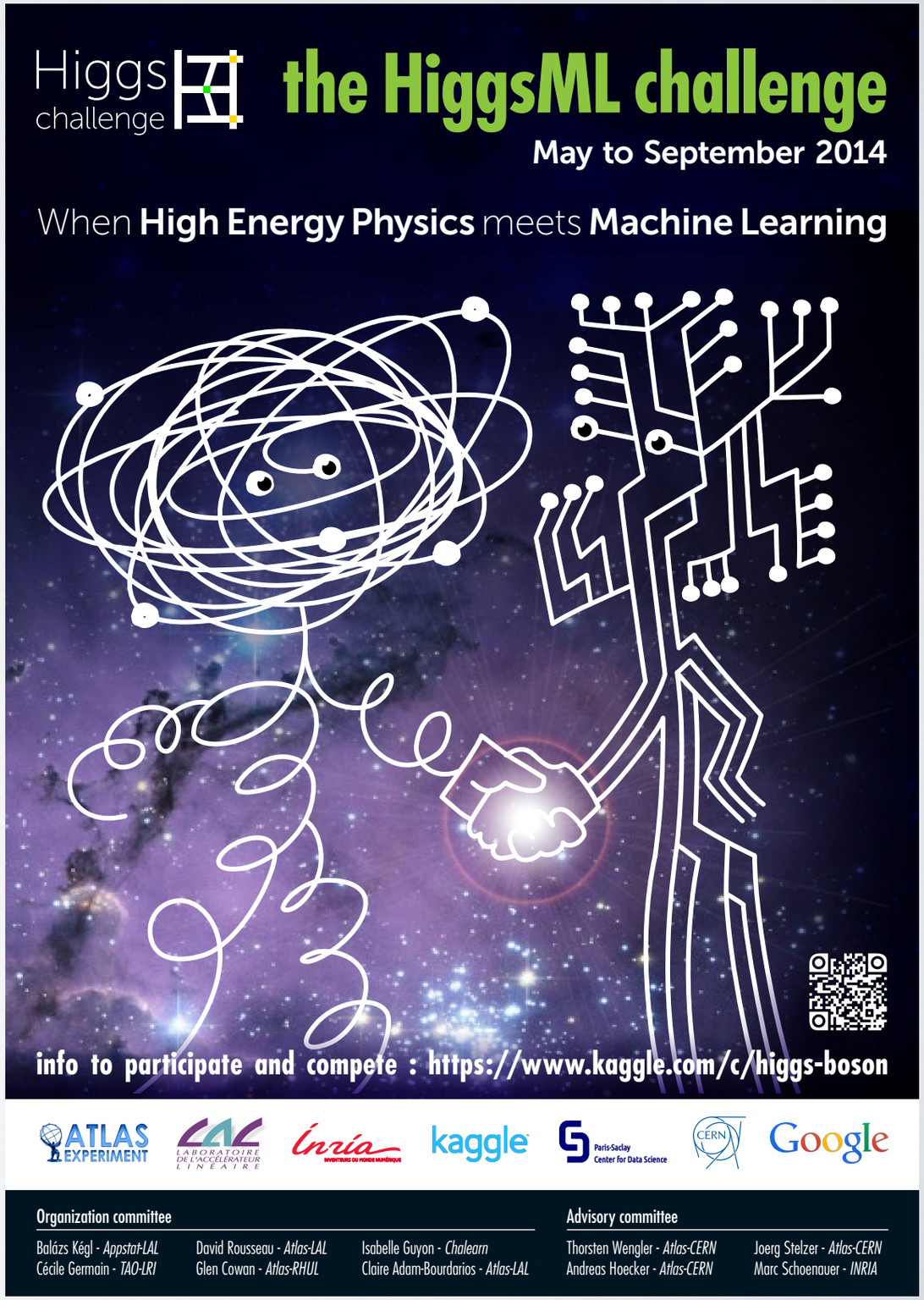
G. Cowan, K. Cranmer, E. Gross, and O. Vitells. EPJ C, 71:1554, 2011.

$$\text{AMS} = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)} \approx \frac{s}{\sqrt{b}}$$

- $\widehat{\mathcal{G}} = \{i : g(\mathbf{x}_i) = \mathsf{s}\}$
- $s = \sum_{i \in \mathcal{S} \cap \widehat{\mathcal{G}}} w_i$
- $b = \sum_{i \in \mathcal{B} \cap \widehat{\mathcal{G}}} w_i$

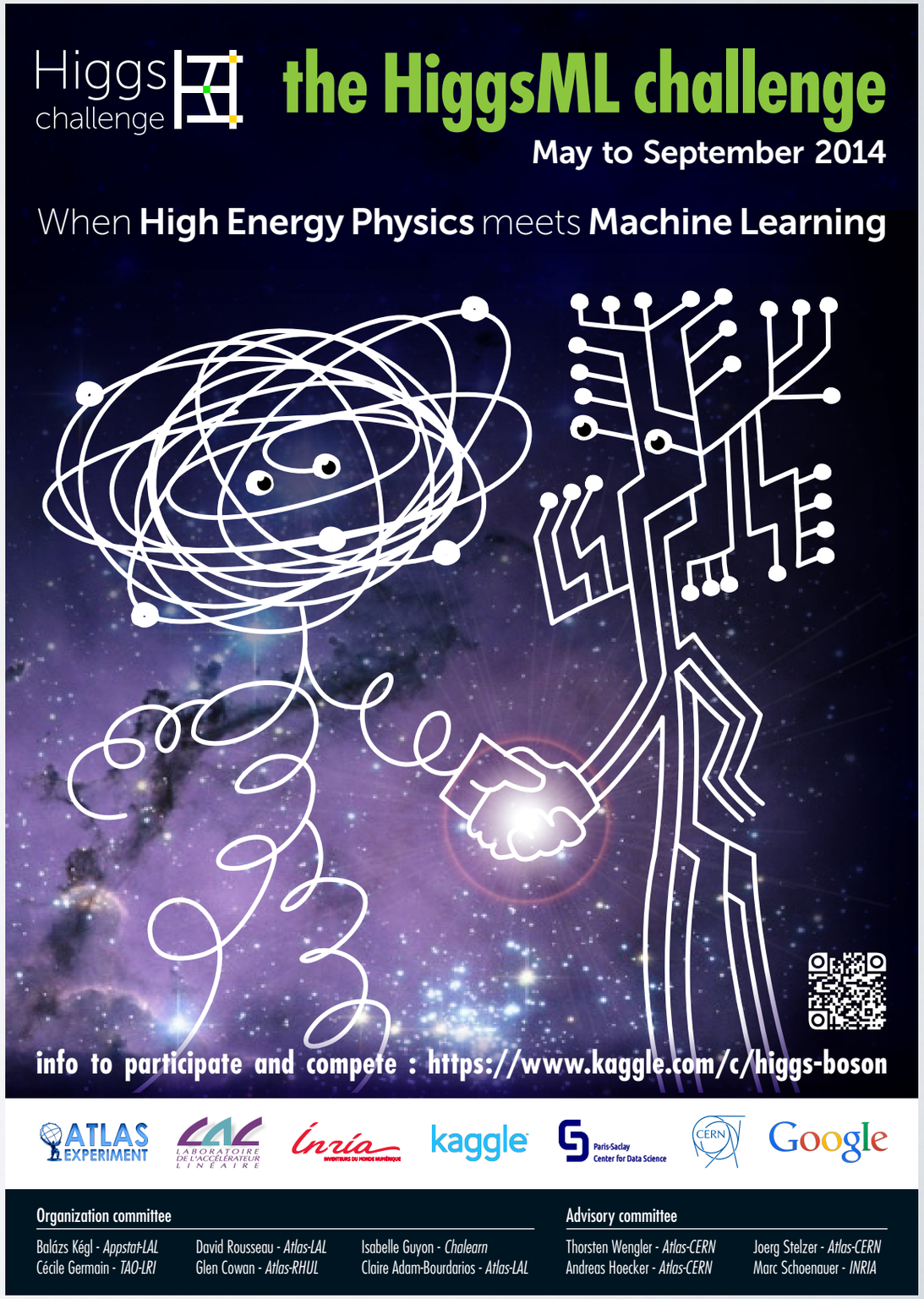
CLASSIFICATION FOR DISCOVERY

A tool for getting
the ML community excited
about your problem



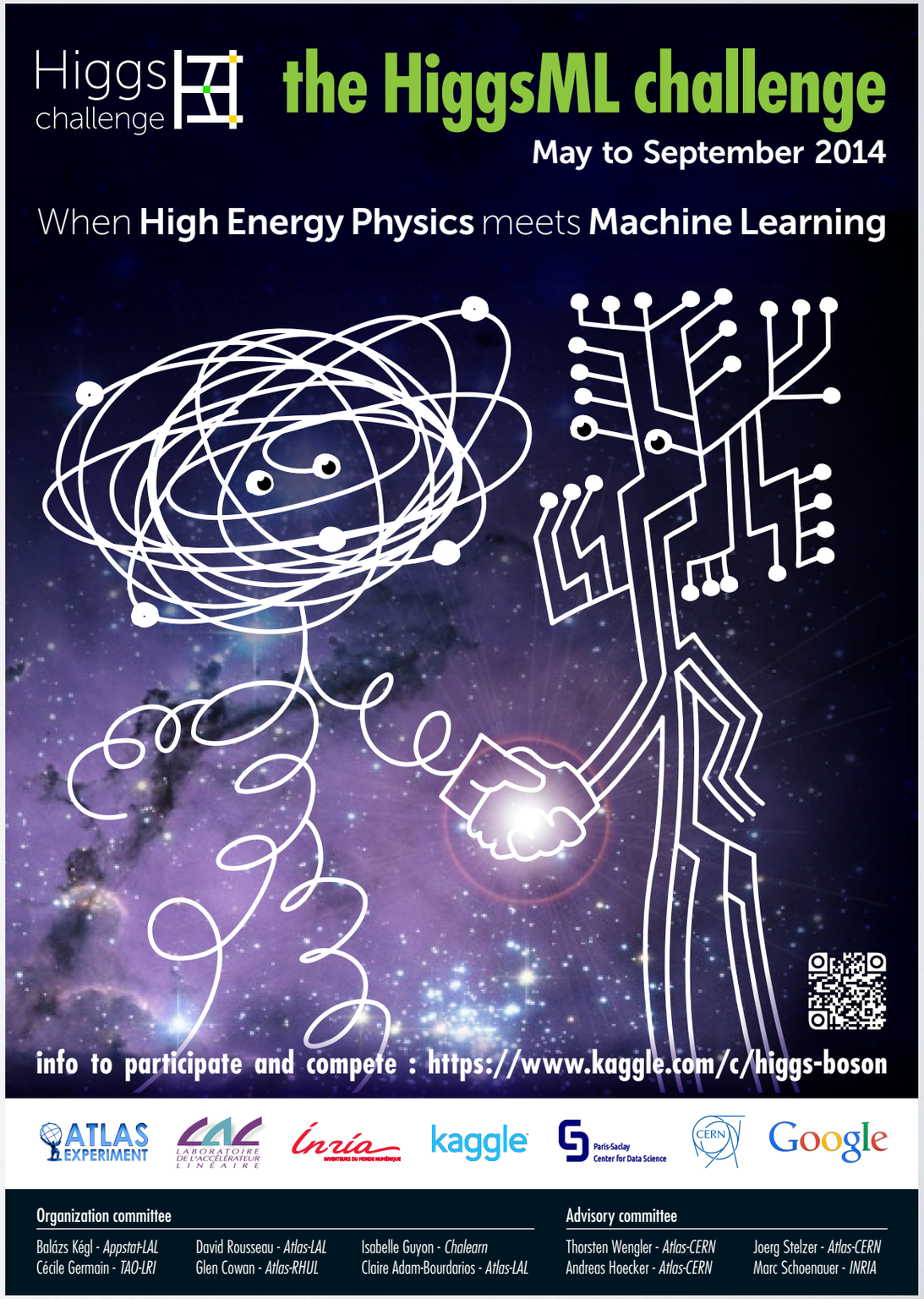
CLASSIFICATION FOR DISCOVERY

- Official ATLAS GEANT4 simulations
  - 30 features (variables)
  - 250K training: input, label, weight
  - 100K public test (AMS displayed real-time), only input
  - 450K private test (to determine the winner after the closing of the challenge), only input
  - public and private tests set are shuffled, participants submit a vector of 550K labels



CLASSIFICATION FOR DISCOVERY

- 16K$ prize pool
  - 7-4-2K$ for the three top participants
  - HEP meets ML award for the most useful model, decided by the ATLAS members of the organizing committee



Paris-Saclay Center for Data Science

CERN

ATLAS EXPERIMENT

LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE

Inria INVENTORS FOR THE DIGITAL WORLD

Google

CLASSIFICATION FOR DISCOVERY

kaggle

Customer Solutions | Competitions | Community ▾ | Sign up | Login

Higgs challenge

$13,000 • 1,627 teams

Higgs Boson Machine Learning Challenge

Enter/Merge by

Mon 12 May 2014 — Mon 15 Sep 2014 (11 days to go)

Dashboard

Leaderboard - Higgs Boson Machine Learning Challenge

This leaderboard is calculated on approximately 18% of the test data.
The final results will be based on the other 82%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ1w	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Gábor Melis *	3.85059	97	Thu, 04 Sep 2014 08:07:20 (-34d)
2	↑1	Luboš Motl's team *	3.84522	535	Thu, 04 Sep 2014 09:22:47 (-1.9h)
3	↓1	Tim Salimans *	3.84428	46	Tue, 12 Aug 2014 17:42:01 (-12.2d)

CLASSIFICATION FOR DISCOVERY

kaggle

Customer Solutions | Competitions | Community ▾ | Sign up | Login

Higgs challenge

$13,000 · 1,627 teams

Higgs Boson Machine Learning Challenge

Enter/Merge by

Mon 12 May 2014 — Mon 15 Sep 2014 (11 days to go)

Dashboard

Leaderboard - Higgs Boson Machine Learning Challenge

This leaderboard is calculated on approximately 18% of the test data.
The final results will be based on the other 82%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ1w	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Gábor Melis *	3.85059	97	Thu, 04 Sep 2014 08:07:20 (-34d)
2	↑1	Luboš Motl's team *	3.84522	535	Thu, 04 Sep 2014 09:22:47 (-1.9h)
3	↓1	Tim Salimans *	3.84428	46	Tue, 12 Aug 2014 17:42:01 (-12.2d)

CLASSIFICATION FOR DISCOVERY

#	Δ1w	Team Name ‡ model uploaded * in the MONEY	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Gábor Melis *	3.85059	97	Thu, 04 Sep 2014 08:07:20 (-34d)
2	↑1	Luboš Motl's team *	3.84522	535	Thu, 04 Sep 2014 09:22:47 (-1.9h)
3	↓1	Tim Salimans *	3.84428	46	Tue, 12 Aug 2014 17:42:01 (-12.2d)
4	—	nhlx5haze	3.80655	234	Thu, 28 Aug 2014 13:13:45 (-58.2d)
5	—	Opera Solutions	3.78818	187	Thu, 04 Sep 2014 11:46:44 (-3.7h)
6	—	Davut & Josef	3.76856	137	Thu, 04 Sep 2014 18:15:27 (-8.2d)
7	—	Roberto-UCIIIM	3.76560	246	Thu, 04 Sep 2014 14:15:52 (-32.6d)
834	↑197	Nemerle	3.25868	51	Thu, 04 Sep 2014 15:06:09 (-16.1h)
835	↓80	CIMFAVuv	3.25090	9	Wed, 23 Jul 2014 16:43:38 (-4.8d)
836	↓80	JeJe	3.25012	4	Sat, 21 Jun 2014 01:11:13
		simple TMVA boosted trees	3.24954		
837	↓80	Xiaohu SUN	3.24954	3	Tue, 03 Jun 2014 13:14:47
838	↓80	Pierre Boutaud	3.24954	10	Fri, 25 Jul 2014 15:25:07 (-30d)

CLASSIFICATION FOR DISCOVERY

#	Δ1w	Team Name ‡ model uploaded * in the MONEY	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Gábor Melis *	3.85059	97	Thu, 04 Sep 2014 08:07:20 (-34d)
2	↑1	Luboš Motl's team *	3.84522	535	Thu, 04 Sep 2014 09:22:47 (-1.9h)
3	↓1	Tim Salimans *	3.84428	46	Tue, 12 Aug 2014 17:42:01 (-12.2d)
4	—	nhlx5haze	3.80655	234	Thu, 28 Aug 2014 13:13:45 (-58.2d)
5	—	Opera Solutions	3.78818	187	Thu, 04 Sep 2014 11:46:44 (-3.7h)
6	—	Davut & Josef	3.76856	137	Thu, 04 Sep 2014 18:15:27 (-8.2d)
7	—	Roberto-UCIIIM	3.76560	246	Thu, 04 Sep 2014 14:15:52 (-32.6d)
834	↑197	Nemerle	3.25868	51	Thu, 04 Sep 2014 15:06:09 (-16.1h)
835	↓80	CIMFAVuv	3.25090	9	Wed, 23 Jul 2014 16:43:38 (-4.8d)
836	↓80	JeJe	3.25012	4	Sat, 21 Jun 2014 01:11:13
		simple TMVA boosted trees	3.24954		
837	↓80	Xiaohu SUN	3.24954	3	Tue, 03 Jun 2014 13:14:47
838	↓80	Pierre Boutaud	3.24954	10	Fri, 25 Jul 2014 15:25:07 (-30d)

CLASSIFICATION FOR DISCOVERY

#	Δ1w	Team Name ‡ model uploaded * in the MONEY	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Gábor Melis *	3.85059	97	Thu, 04 Sep 2014 08:07:20 (-34d)
2	↑1	Luboš Motl's team *	3.84522	535	Thu, 04 Sep 2014 09:22:47 (-1.9h)
3	↓1	Tim Salimans *	3.84428	46	Tue, 12 Aug 2014 17:42:01 (-12.2d)
4	—	nhlx5haze	3.80655	234	Thu, 28 Aug 2014 13:13:45 (-58.2d)
5	—	Opera Solutions	3.78818	187	Thu, 04 Sep 2014 11:46:44 (-3.7h)
6	—	Davut & Josef	3.76856	137	Thu, 04 Sep 2014 18:15:27 (-8.2d)
7	—	Roberto-UCIIIM	3.76560	246	Thu, 04 Sep 2014 14:15:52 (-32.6d)
834	↑197	Nemerle	3.25868	51	Thu, 04 Sep 2014 15:06:09 (-16.1h)
835	↓80	CIMFAVuv	3.25090	9	Wed, 23 Jul 2014 16:43:38 (-4.8d)
836	↓80	JeJe	3.25012	4	Sat, 21 Jun 2014 01:11:13
		simple TMVA boosted trees	3.24954		
837	↓80	Xiaohu SUN	3.24954	3	Tue, 03 Jun 2014 13:14:47
838	↓80	Pierre Boutaud	3.24954	10	Fri, 25 Jul 2014 15:25:07 (-30d)

CLASSIFICATION FOR DISCOVERY

CLASSIFICATION FOR DISCOVERY

- Take home message: running a data challenge is a great way to

CLASSIFICATION FOR DISCOVERY

- Take home message: running a data challenge is a great way to
- let data scientists know about your problem

CLASSIFICATION FOR DISCOVERY

- Take home message: running a data challenge is a great way to
- let data scientists know about your problem
- get meaningful solutions to difficult technical problems

CLASSIFICATION FOR DISCOVERY

- Take home message: running a data challenge is a great way to
- let data scientists know about your problem
- get meaningful solutions to difficult technical problems
- jump-start and nurture interdisciplinary collaborations