Handling Concept Drift: An Application Perspective

Mykola Pechenizkiy, TU Eindhoven, the Netherlands

Indrė Žliobaitė, TU Eindhoven, the Netherlands

DHDHD 2010
Workshop

TU/e, Eindhoven
Aug. 20th, 2010

Learning from Evolving Data

Joao Gama
University of Porto, PORTUGAL

Ernestina Menasalvas
Technical Univ. of Madrid, SPAIN

Myra Spiliopoulou
University of Magdeburg, GERMANY

Athena Vakali
University of Thessaloniki, GREECE

and the Chairs of the HaCDAIS Workshop:

Mykola Pechenizkiy, Eidhoven Univ. of Technology, NETHERLANDS

Indrė Žliobaitė, Eidhoven Univ. of Technology, NETHERLANDS

UNIVERSIDADE DO PORTO

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

OTTO-VON-GUERICKE-UNIVERSITÄT MAGDEBURG

POLITÉCNICA
Ingeniamos el futuro

ECML-PKDD Conference Tutorial
Barcelona, Sept. 24th, 2010

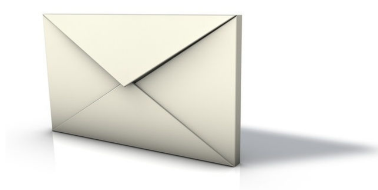
Concept Drift: Application Perspective

CD refers to non-stationary supervised learning problems

but there are different types of CD

and different types of applications

Personal recommenders, spam filters, fraud detection, navigation are affected by drifts coming from different sources



Motivation

View CD research from an application perspective

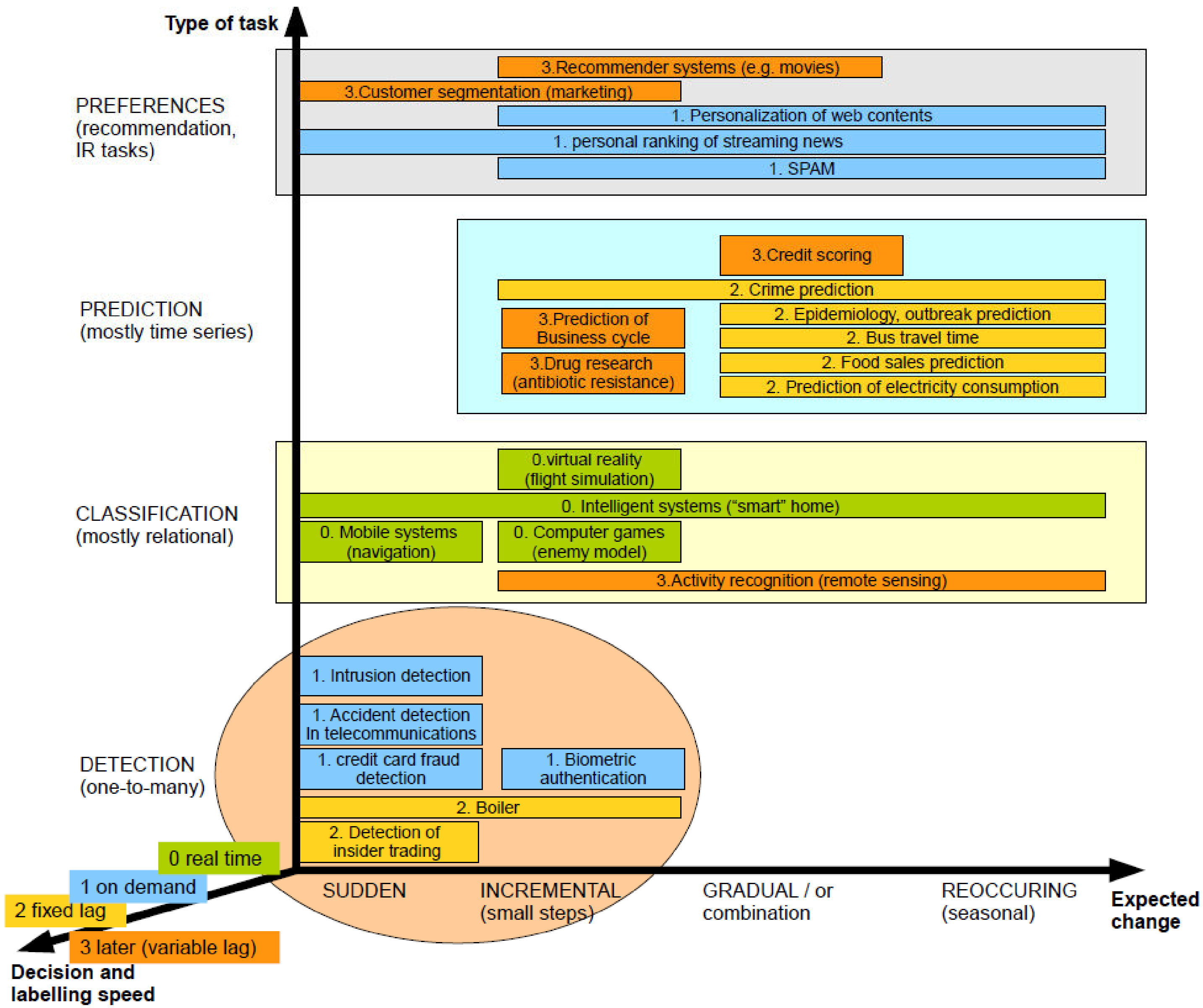
What is the match between the mainstream CD research assumptions and properties of the applications?

Identify promising future research directions from the application perspective

We will talk about

- Why changes appear in different applications?
- What are the properties of CD application tasks?
- How the application tasks can be categorized in terms of these basic properties?

Categorization of applications (Žliobaitė & Pechenizkiy, 2010)



Properties of the tasks

DATA task (detection, classification, prediction, ranking)

type (time series, relational, mix)

organization (stream/batches, data re-access, missing)

DRIFT change type (sudden, gradual, incremental, reoccurring)

source (adversary, interests, population, complexity)

expectation (unpredictable, predictable, identifiable)

DECISIONS and GROUND TRUTH

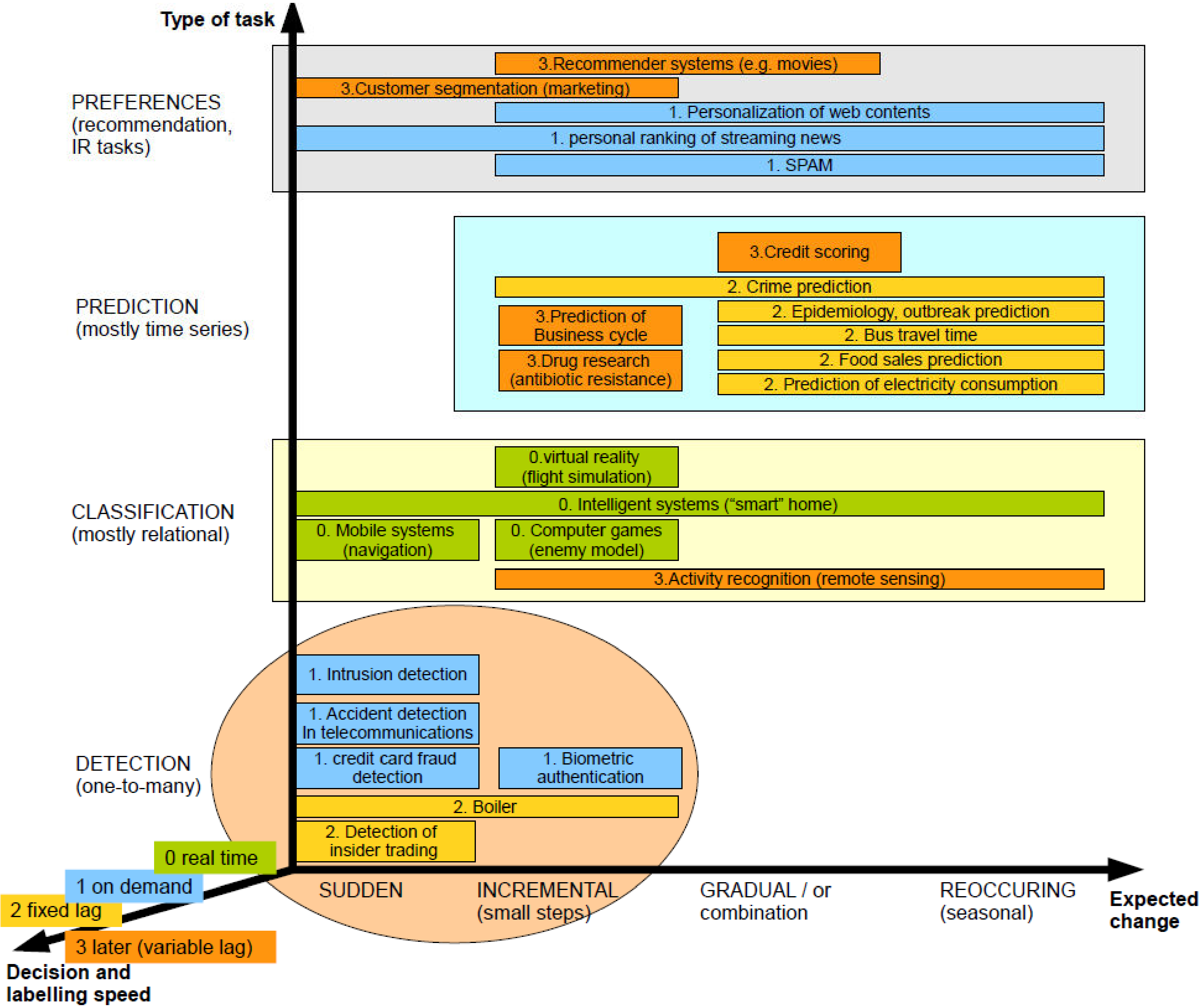
labels (real time, on demand, fixed lag, delay)

decision speed (real time, analytical)

ground truth labels (soft, hard)

costs of mistakes (balanced, unbalanced)

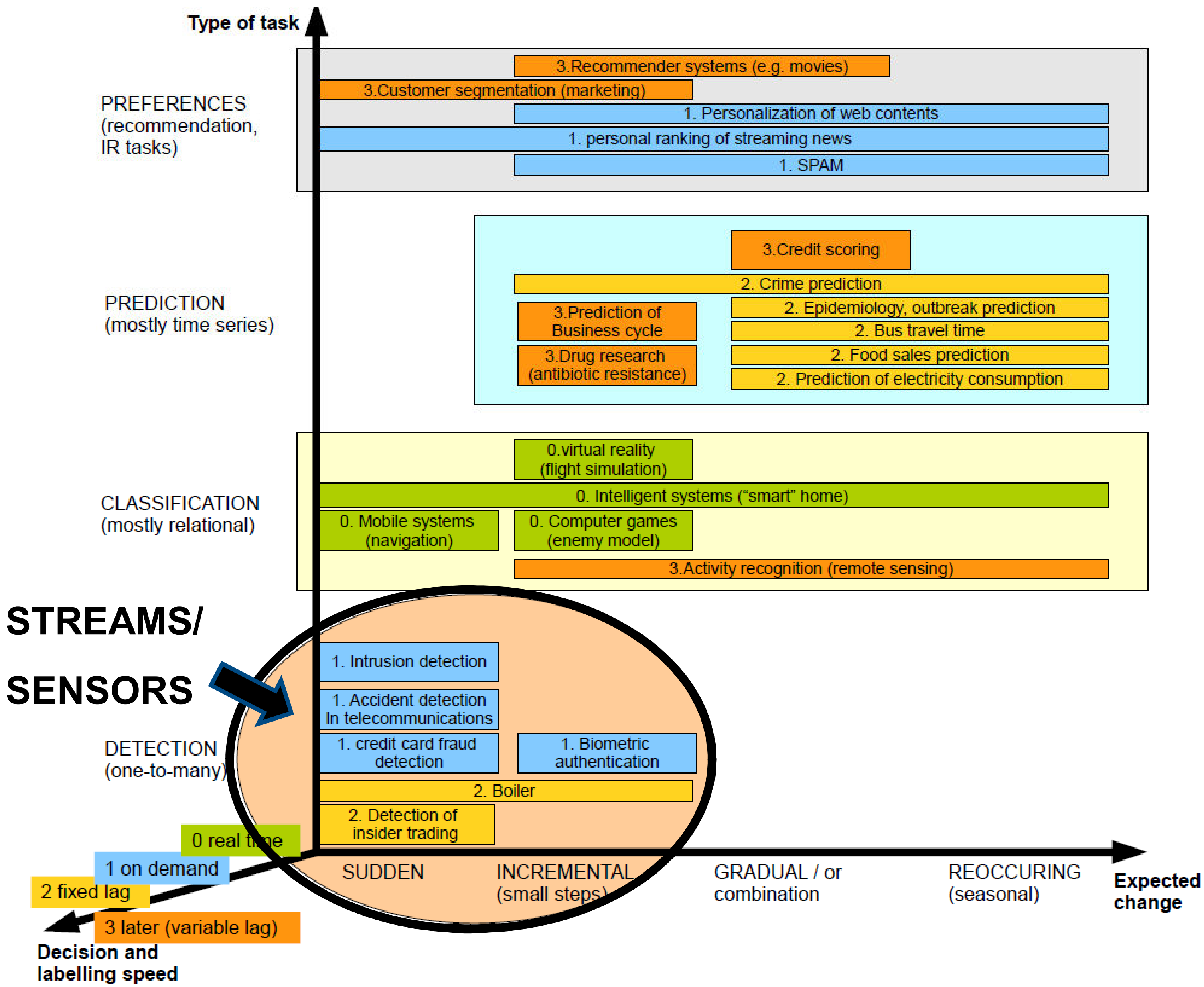
Categorization of applications (Žliobaitė & Pechenizkiy, 2010)



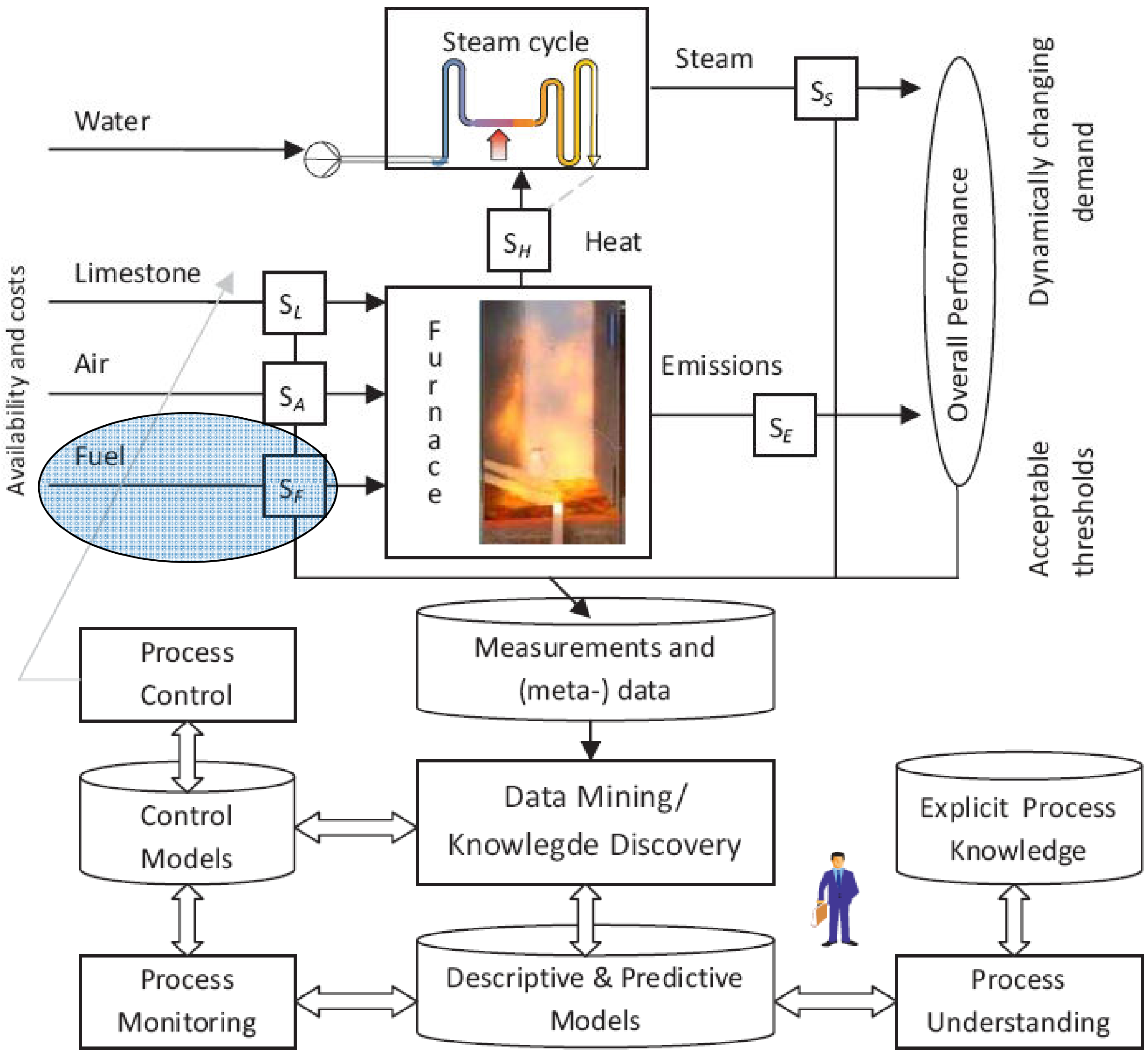
Landscape of applications

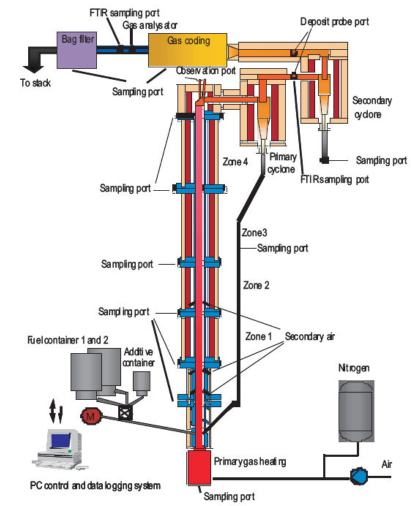
Types of apps / *Industries*	Monitoring/ control	Personal assistance/ personalization	Management and planning	Ubiquitous applications
Security, Police	Fraud detection, insider trading detection, adversary actions detection	--------	Crime volume prediction	Authentica-tion, Intrusion detection
Finance, Banking, Telecom, Credit Scoring, Insurance, Direct Marketing, Retail, Advertising, e-Commerce	Monitoring & management of customer segments, bankruptcy prediction	Product or service recommendation, including complimentary	Demand prediction, response rate prediction, budget planning	Location based services, related ads, mobile apps
Education (higher, professional, children, e-Learning) Entertainment, Media	Gaming the system, Drop out prediction	Music, VOD, movie, learning object recommendation, adaptive news access, personalized search	Player-centered game design, learner-centered education	Virtual reality, simulations
…	…	…	…	…

Categorization of applications (Žliobaitė & Pechenizkiy, 2010)



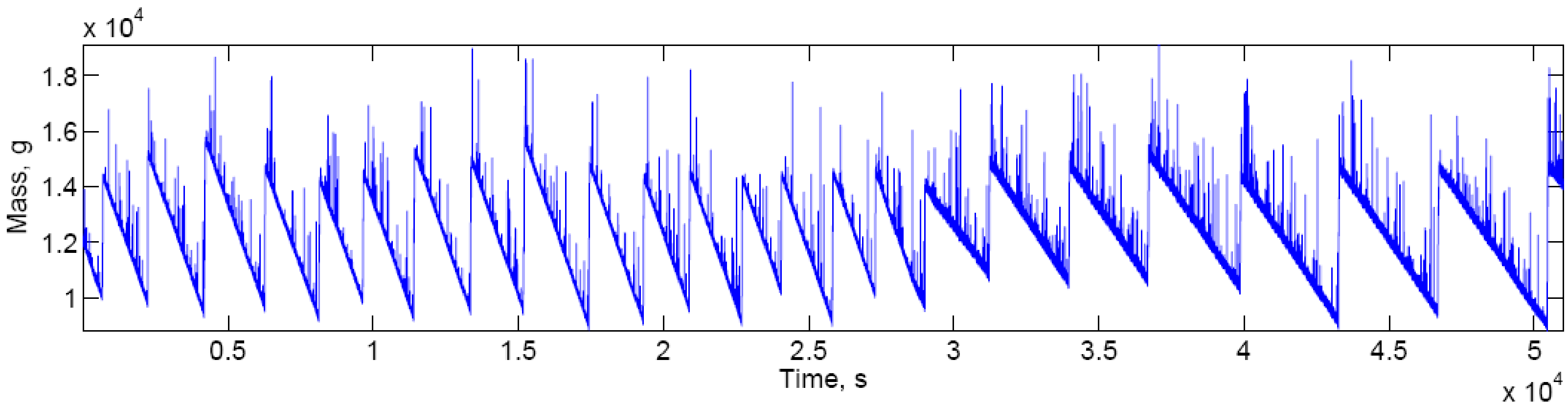
CFB Boiler Optimization





Online mass flow prediction in CFB boilers

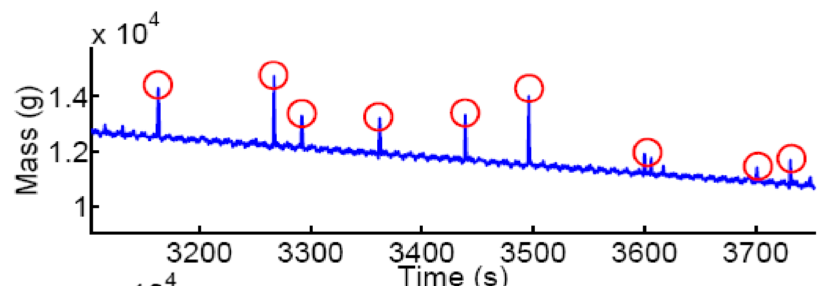
data collected from a typical experimentation with CFB boiler

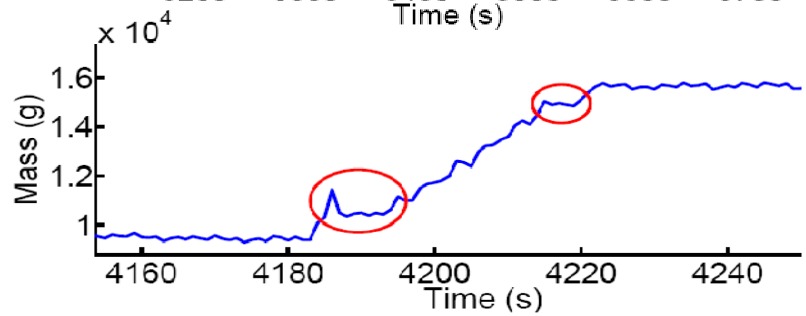


asymmetric nature of the outliers

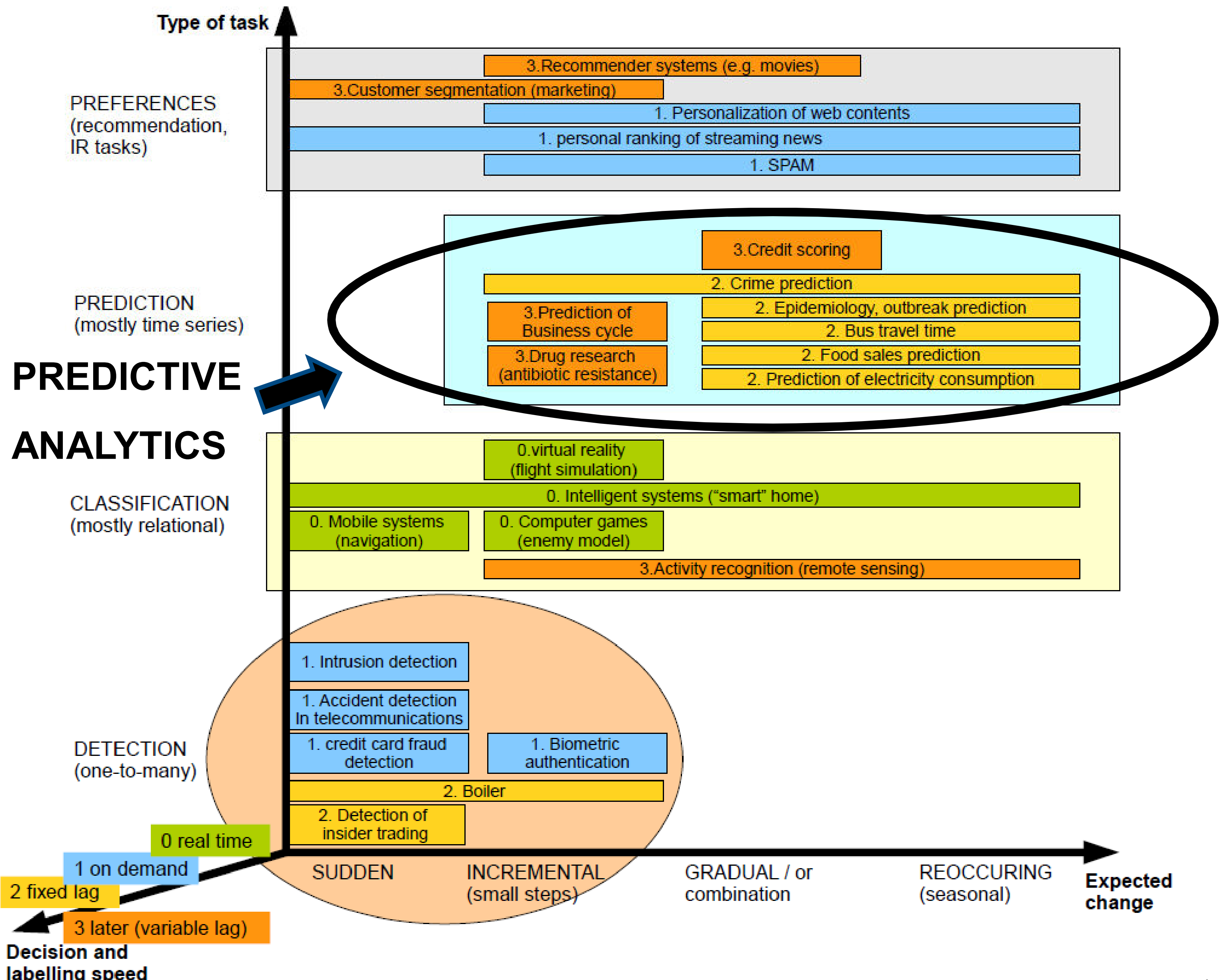
short consumption periods within feeding stages

Pechenizkiy et al. 2009



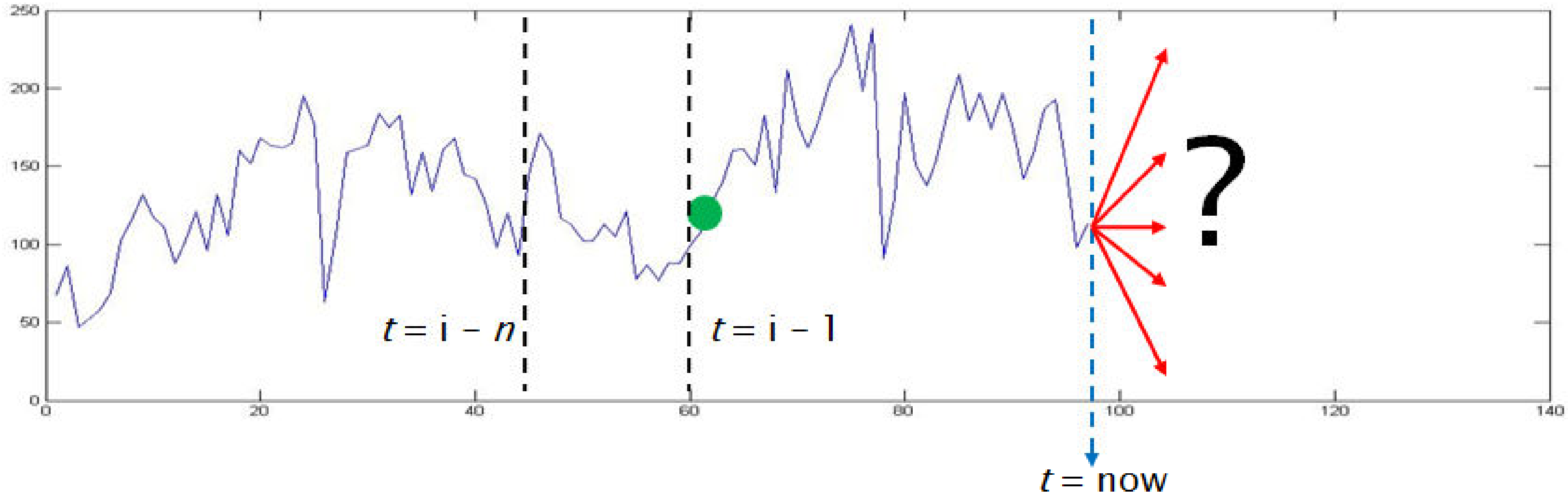


Categorization of applications (Žliobaitė & Pechenizkiy, 2010)



Food sales prediction: utility of Belgium milk in Sep. 2009

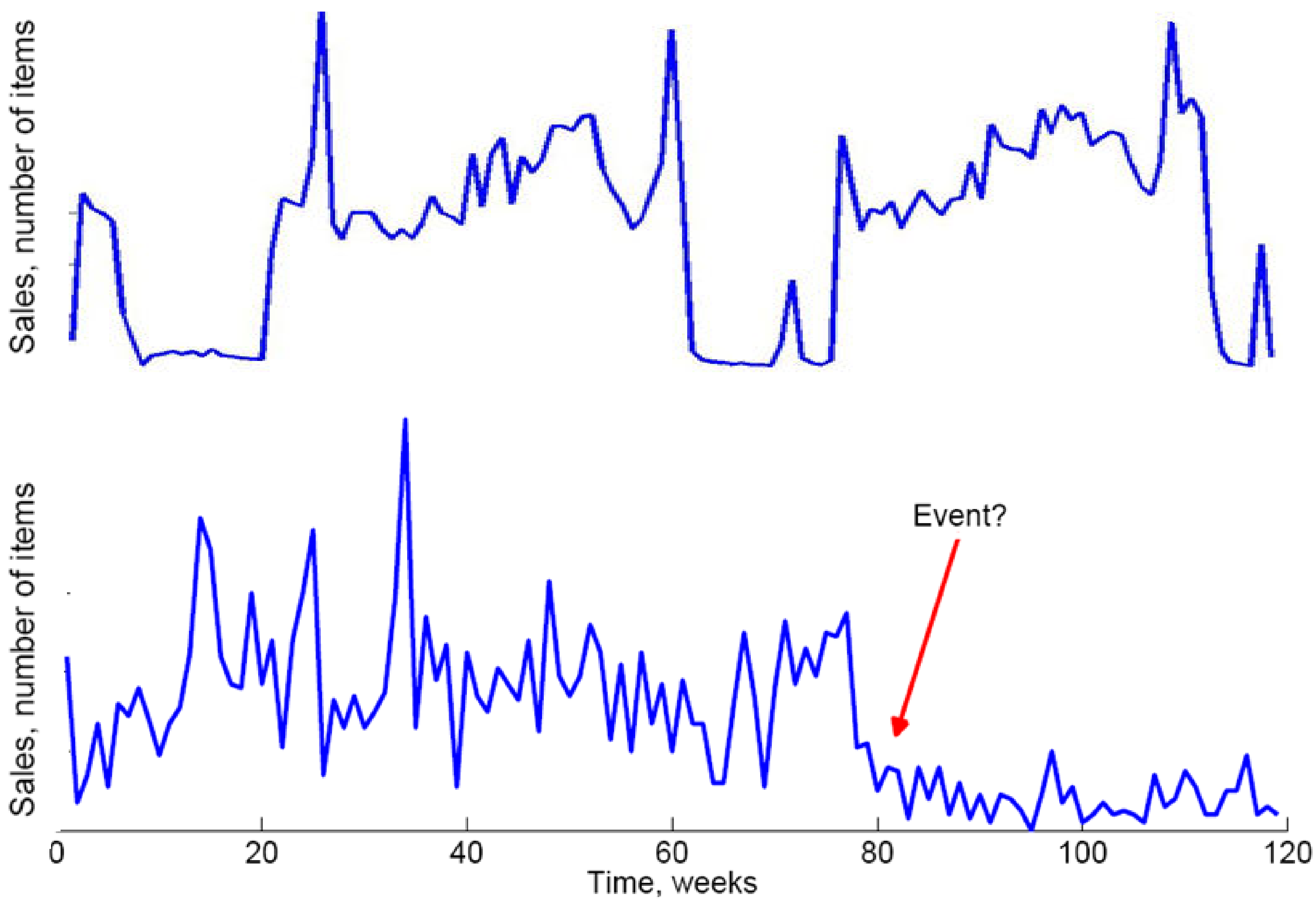
Challenges in food sales prediction (Zliobaite et al., 2009)



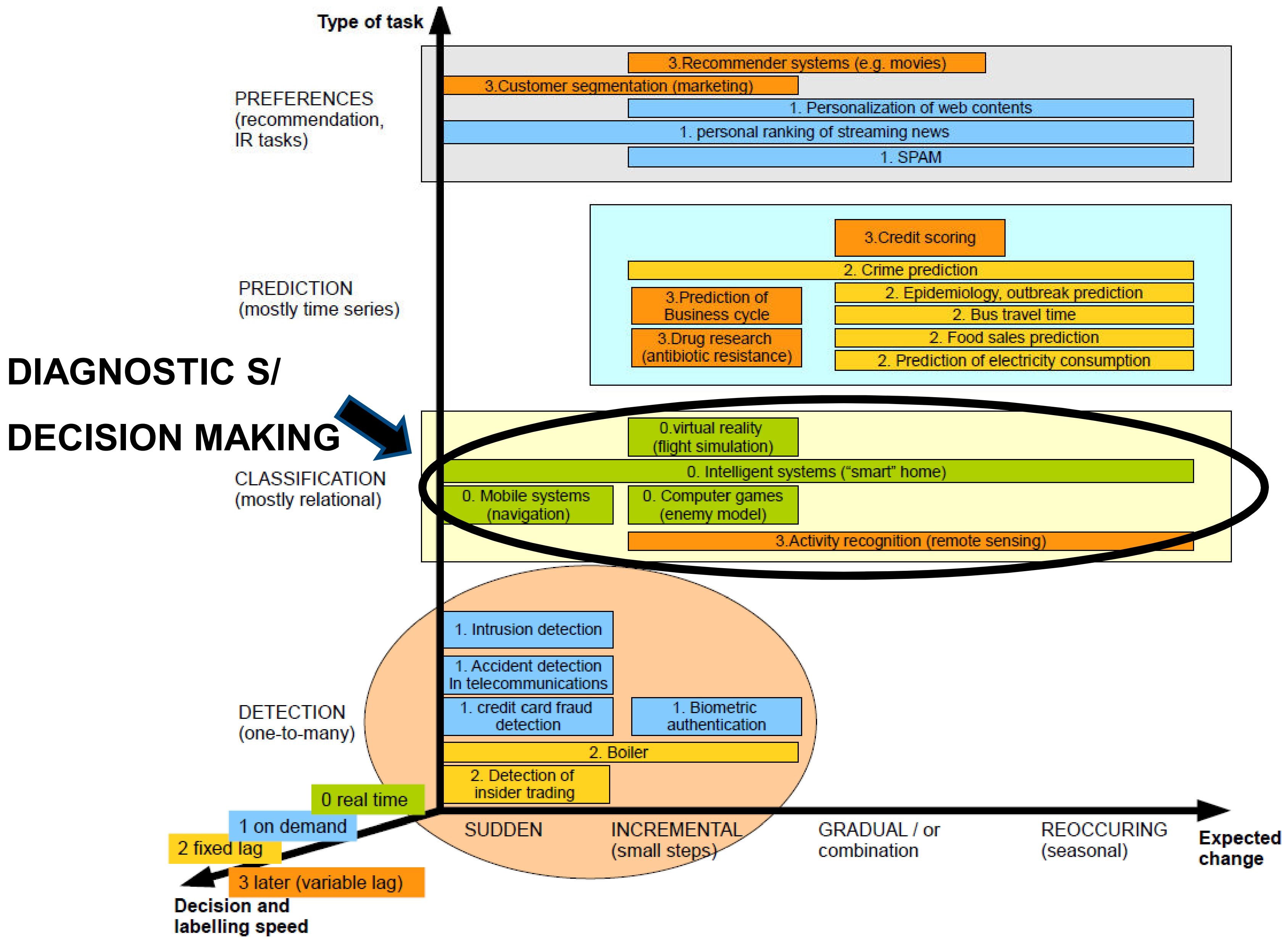
t	History	Temp	Holiday	Promo
1				
.				
i	{y(i-n) .. y(i-1)}			
.				
n				

→ Predict label of each new instance

Reoccuring and suddent dritft in food sales



Categorization of applications (Žliobaitė & Pechenizkiy, 2010)



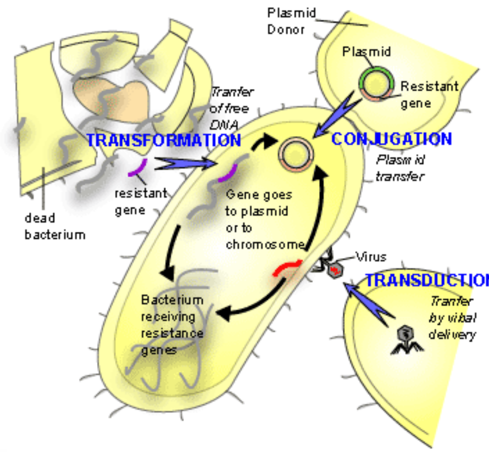
Antibiotic Resistance Prediction (Tsymbal et al., 2008)

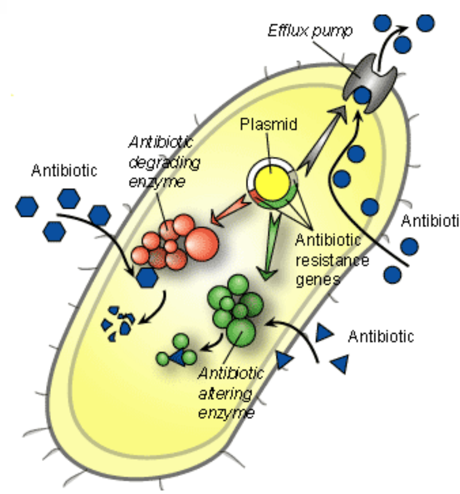
predict the sensitivity of a pathogen to an antibiotic based on data about the antibiotic, the isolated pathogen, and the demographic and clinical features of the patient.

date	sex	age	isNew	days_total	days_ICU	main_dept	pathogen data	antibiotic data	sensitivity
22.1.2002	m	25	1	171	81	9	...	...	3
22.1.2002	m	25	1	171	81	9	...	...	3
22.1.2002	m	25	1	171	81	9	...	...	3
22.1.2002	m	25	1	171	81	9	...	...	3
28.1.2002	f	61	0	261	52	3	...	...	3
28.1.2002	f	61	0	261	52	3	...	...	3
28.1.2002	f	61	0	261	52	3	...	...	3
28.1.2002	f	61	0	261	52	3	...	...	1
28.1.2002	f	61	0	261	52	3	...	...	1
28.1.2002	m	25	1	171	81	9	...	...	3
28.1.2002	m	25	1	171	81	9	...	...	3
30.1.2002	m	25	1	171	81	9	...	...	3
8.2.2002	m	30	0	209	209	9	...	...	3
8.2.2002	m	30	0	209	209	9	...	...	1
8.2.2002	m	30	0	209	209	9	...	...	1
11.2.2002	f	0	0	18	0	2	...	...	1
11.2.2002	f	0	0	18	0	2	...	...	1
11.2.2002	f	0	0	18	0	2	...	...	1
new data	...	...	...	...	...	...	...	...	?
new data	...	...	...	...	...	...	...	...	?
new data	...	...	...	...	...	...	...	...	?

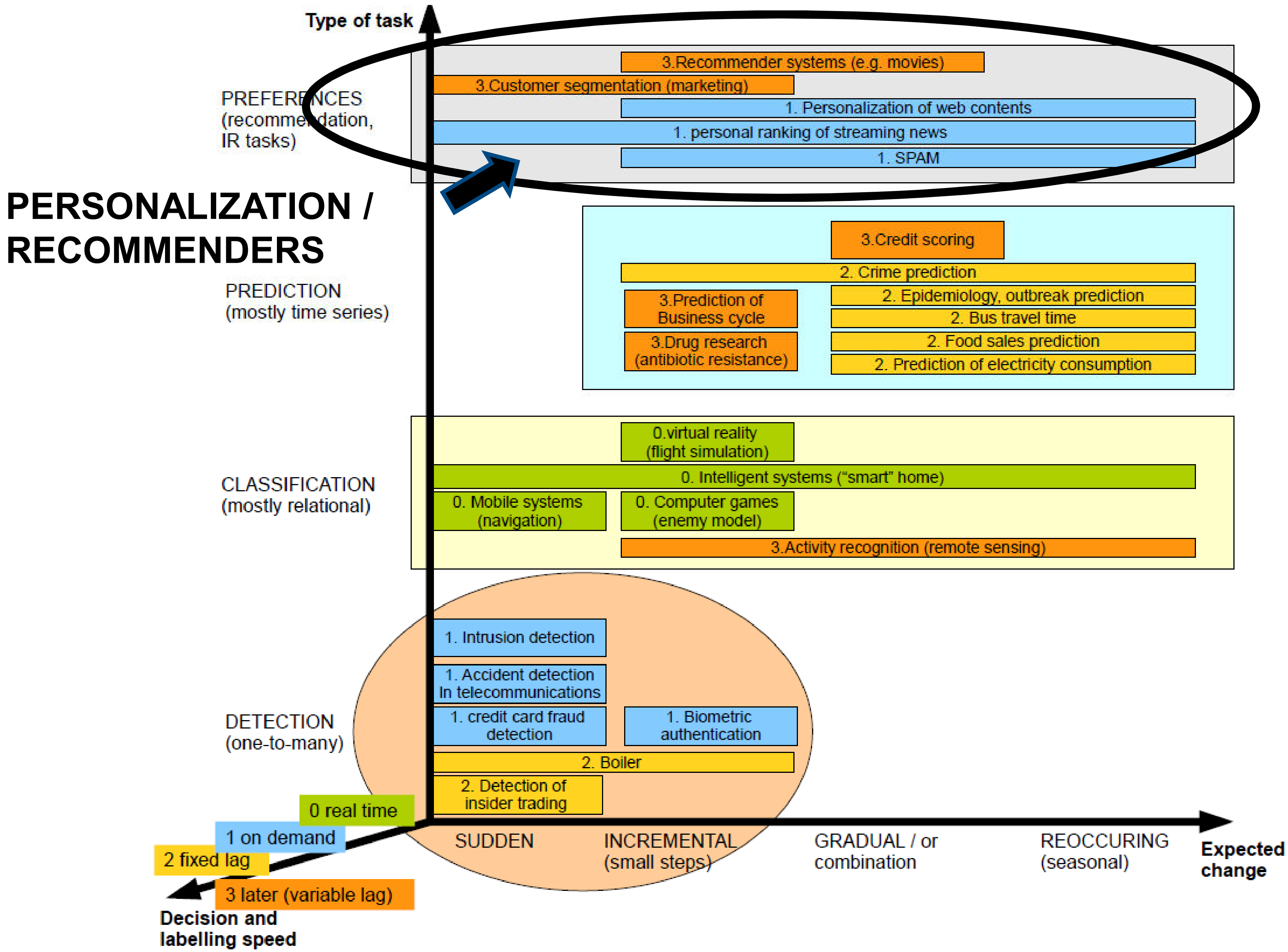
How Antibiotic Resistance Happens

It was on a short-cut through the hospital kitchens that Albert was first approached by a member of the Antibiotic Resistance.

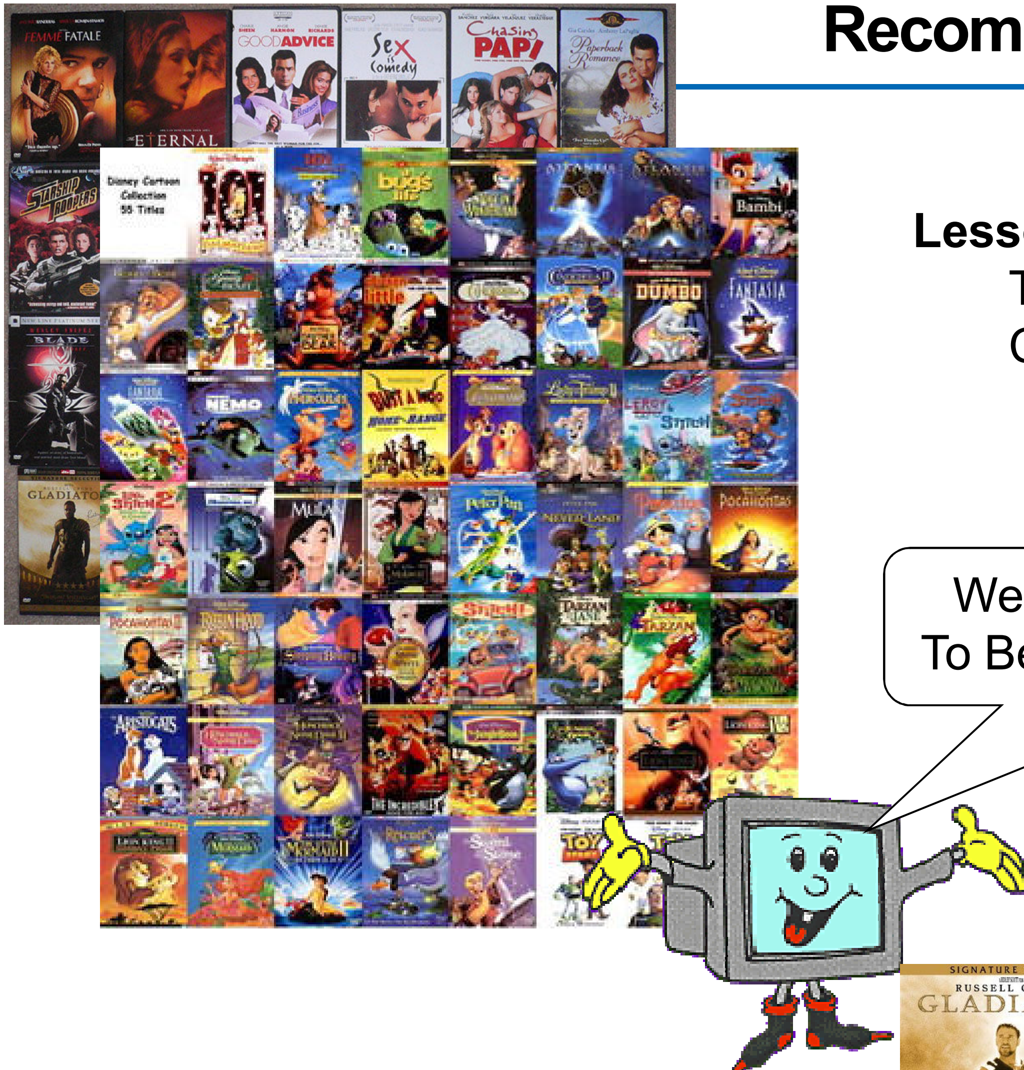




Categorization of applications (Žliobaitė & Pechenizkiy, 2010)



Recommender Systems



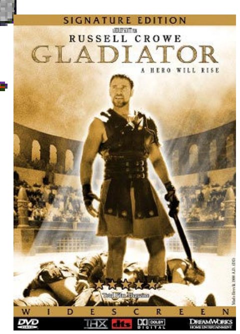
Lessons learnt from Netflix:

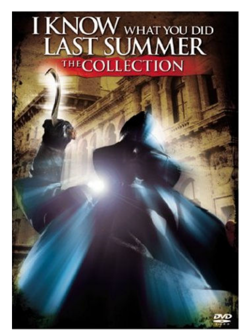
Temporal dynamics is important

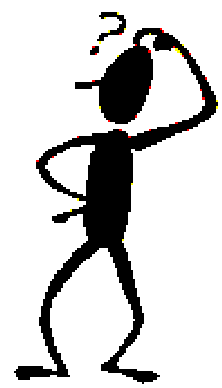
Classical CD approaches may not work

(Koren, SIGKDD 2009)

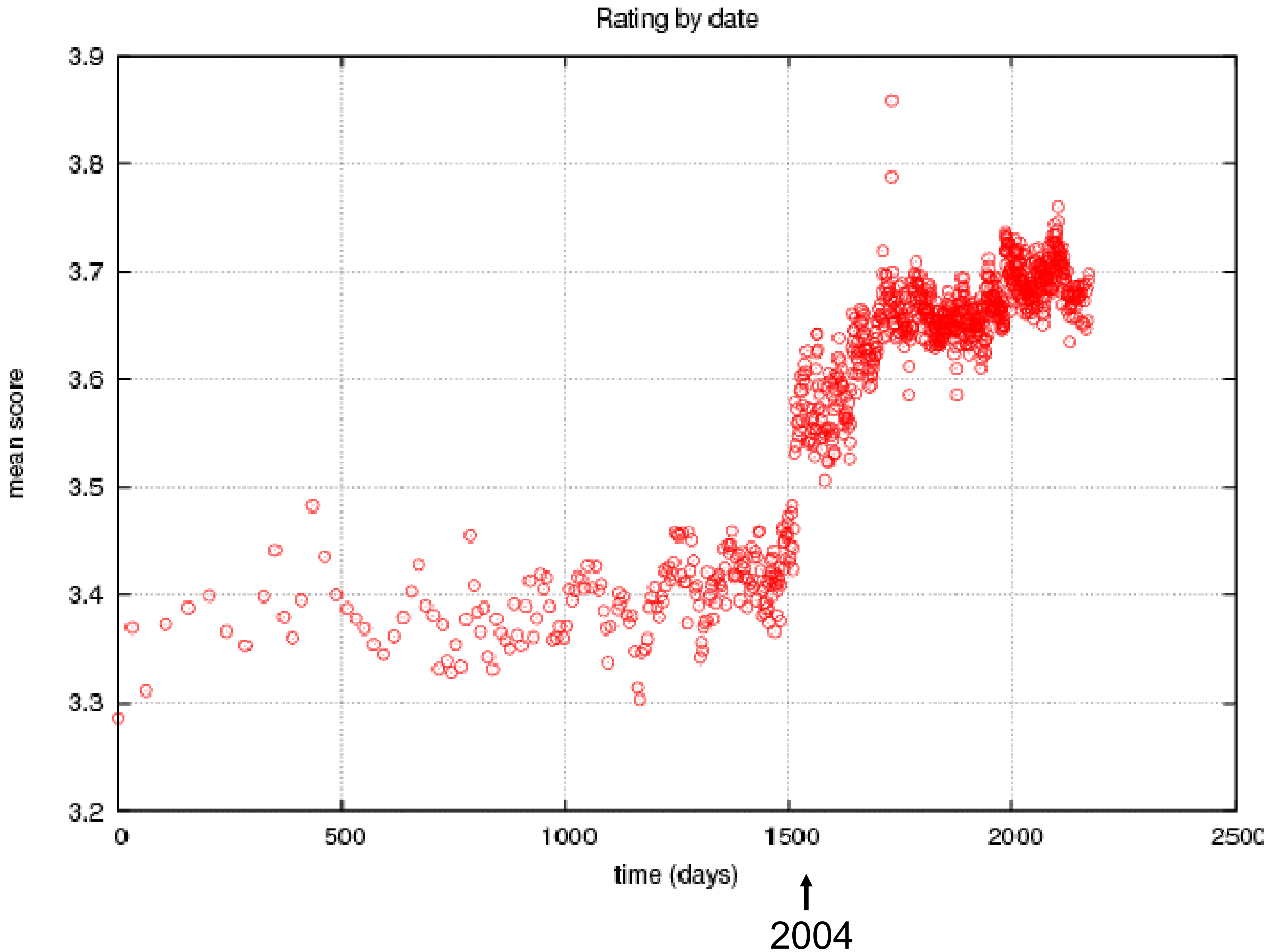
We Know What You Ought
To Be Watching This Summer



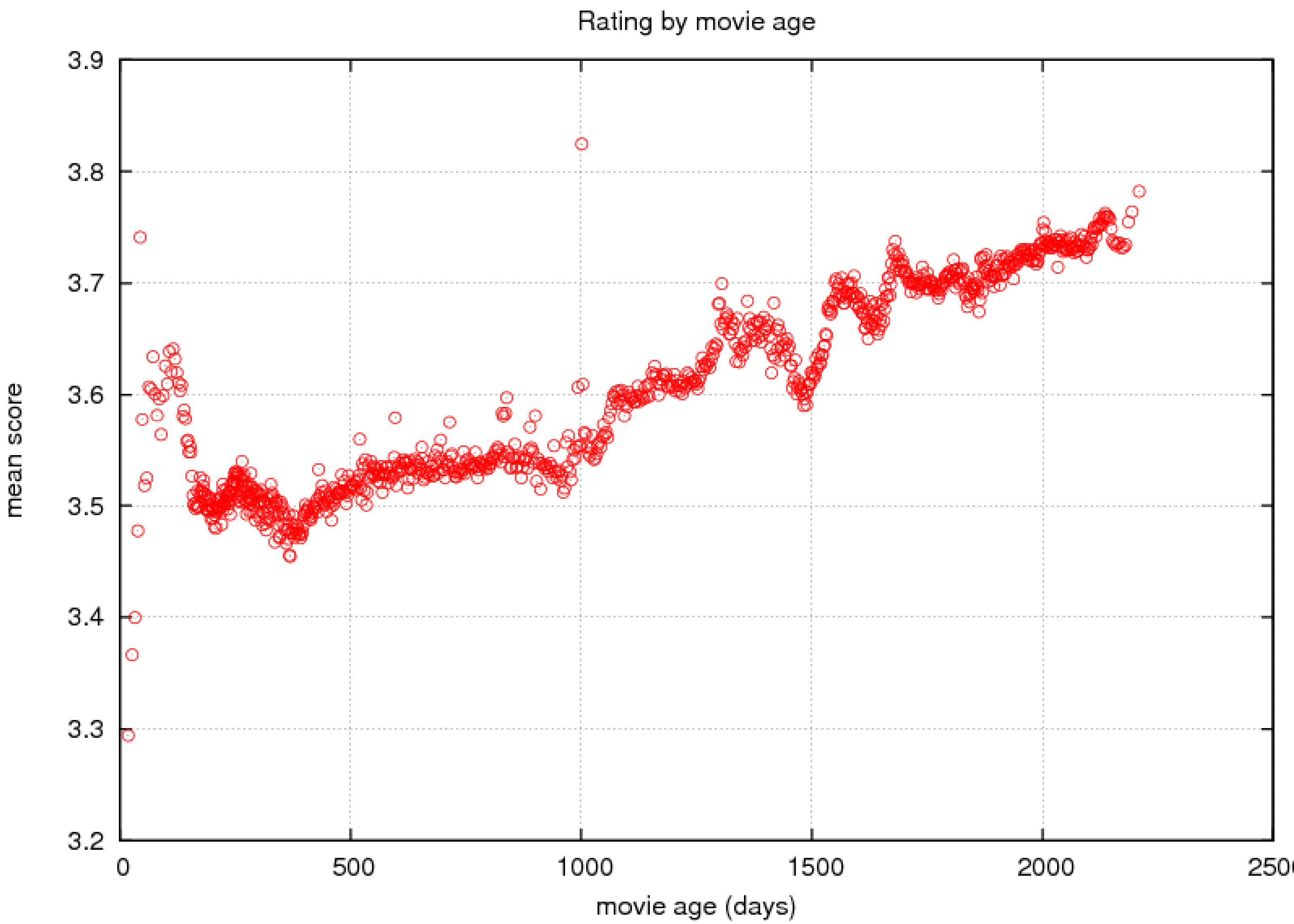




Something Happened in Early 2004…



Are movies getting better with time?



Multiple sources of temporal dynamics

Both items and users are changing over time

Item-side effects:

- Product perception and popularity are constantly changing
- Seasonal patterns influence items' popularity

User-side effects:

- Customers ever redefine their taste
- Transient, short-term bias; anchoring
- Drifting rating scale
- Change of rater within household

➔ Common "concept drift" methodologies won't hold.
E.g., underweighting older instances is unappealing

Outlook

From general methods to more specific application oriented problems like

delayed labeling,

label availability,

cost-benefit trade off of the model update.

Changing the focus to

change description,

prediction reoccurring contexts and

meta learning in addition to change detection.

Take Bowling Message

There is no uniform concept as ‘concept drift’