

# Desafios na modelação do fluxo de estudantes no sistema de ensino Português

Luís Correia, Luís Cavique, Ana Paula Cláudio, Beatriz Carmo,  
Hugo Martiniano, Helena Aidos, Sara Madeira

Fórum Estatístico  
14 / nov / 2018

# ModEst

Modelação do fluxo de estudantes no sistema de ensino Português



**FCT** Fundação  
para a Ciência  
e a Tecnologia

**INVESTIGAÇÃO  
ADMINISTRAÇÃO  
PÚBLICA**  
Inteligência artificial e ciência de dados

# motivação

- DGEEC tem grandes quantidades de dados sobre o sistema de ensino português
  - 1,3 M estudantes
  - 5 k escolas
  - $O(100\text{ k})$  profissionais
- potencial de prospeção destes dados potencialmente elevado
  - impacto económico e social
- oportunidade pelo programa DSAIPA

# objetivo

- fazer prospeção dos dados da DGEEC isoladamente e integrados com dados socio-económicos
- obter **modelos setoriais** do sistema educacional
- obter um **modelo global**

IA na produção de conhecimento para a definição de políticas organizativas do sistema de educação e para tomar medidas específicas corretivas

# problema

- série de dados temporais
  - 70 anos ao nível da escola
  - 10 anos ao nível de aluno
- possibilidades da prospeção dos dados
  - encontrar modelos de previsão
    - do fluxo de alunos (entradas, progressões, abandonos,...)
  - gerar cenários
    - e testar os respetivos efeitos

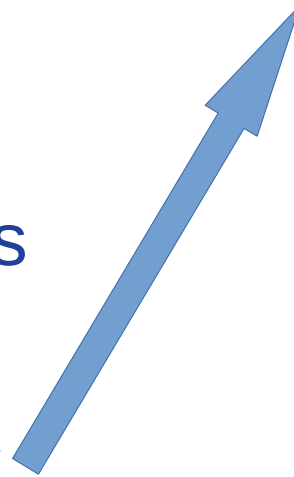
# prospecção de dados (*data mining*)

- extração de padrões a partir de grandes quantidades de dados  
  
e a partir dos padrões tentar obter conhecimento sob uma forma compreensível pelos humanos
  - regras, parâmetros de modelos pré-definidos, probabilidades de situações, ...
- envolve aprendizagem automática, estatística, bases de dados
  - e processamento de sinal (para séries temporais)

# padrões

fluxo estudantes

- macro-padrões:
  - dados agregados
- micro-padrões:
  - dados individuais



- perfis estudantes

- variáveis socio-económicas
- variáveis de comportamento

modelos de agentes

# dados individuais

- socio-económicos:

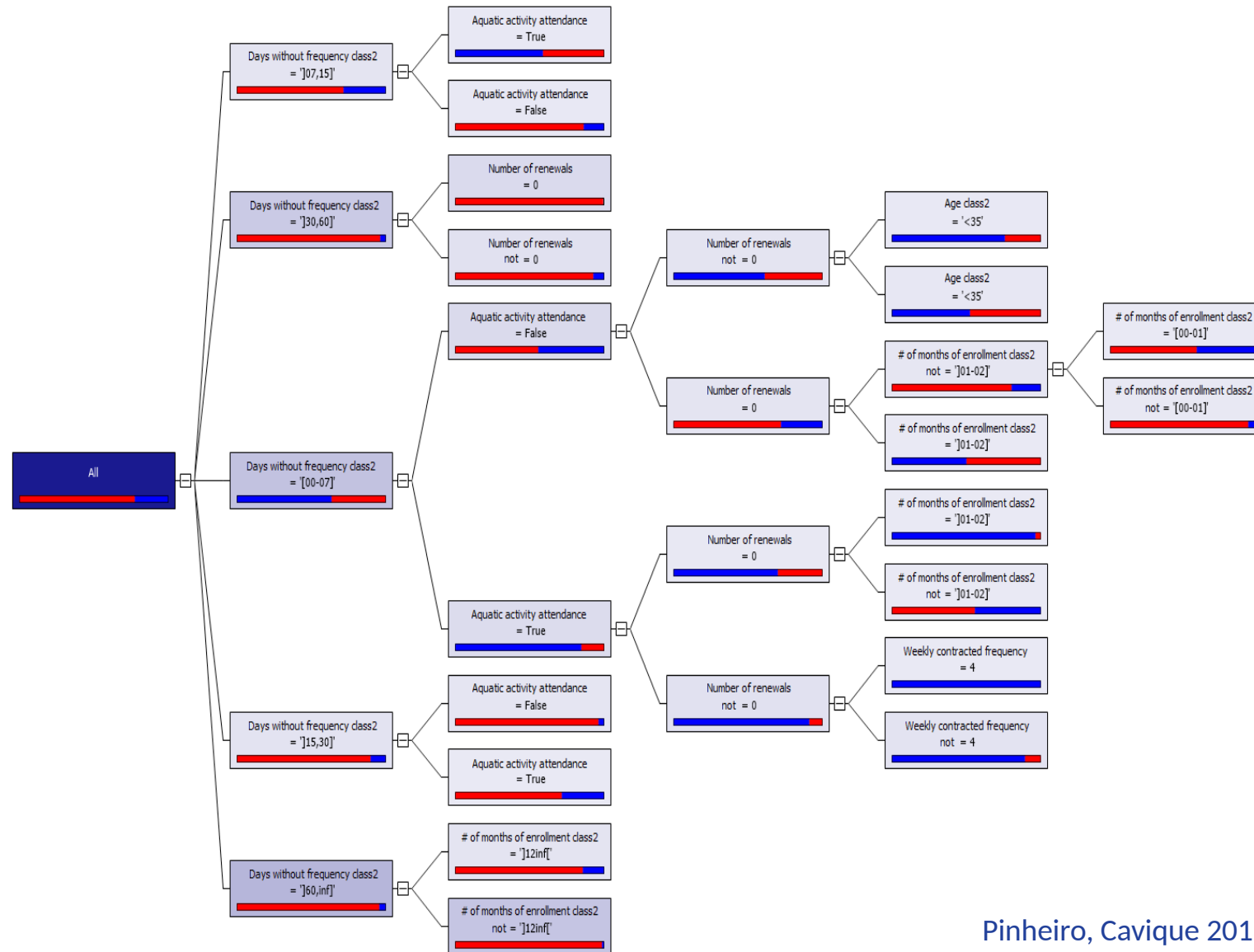
- género
- idade
- dados familiares
- rendimento
- ocupação pais
- classe social
- região
- ...

- comportamento:

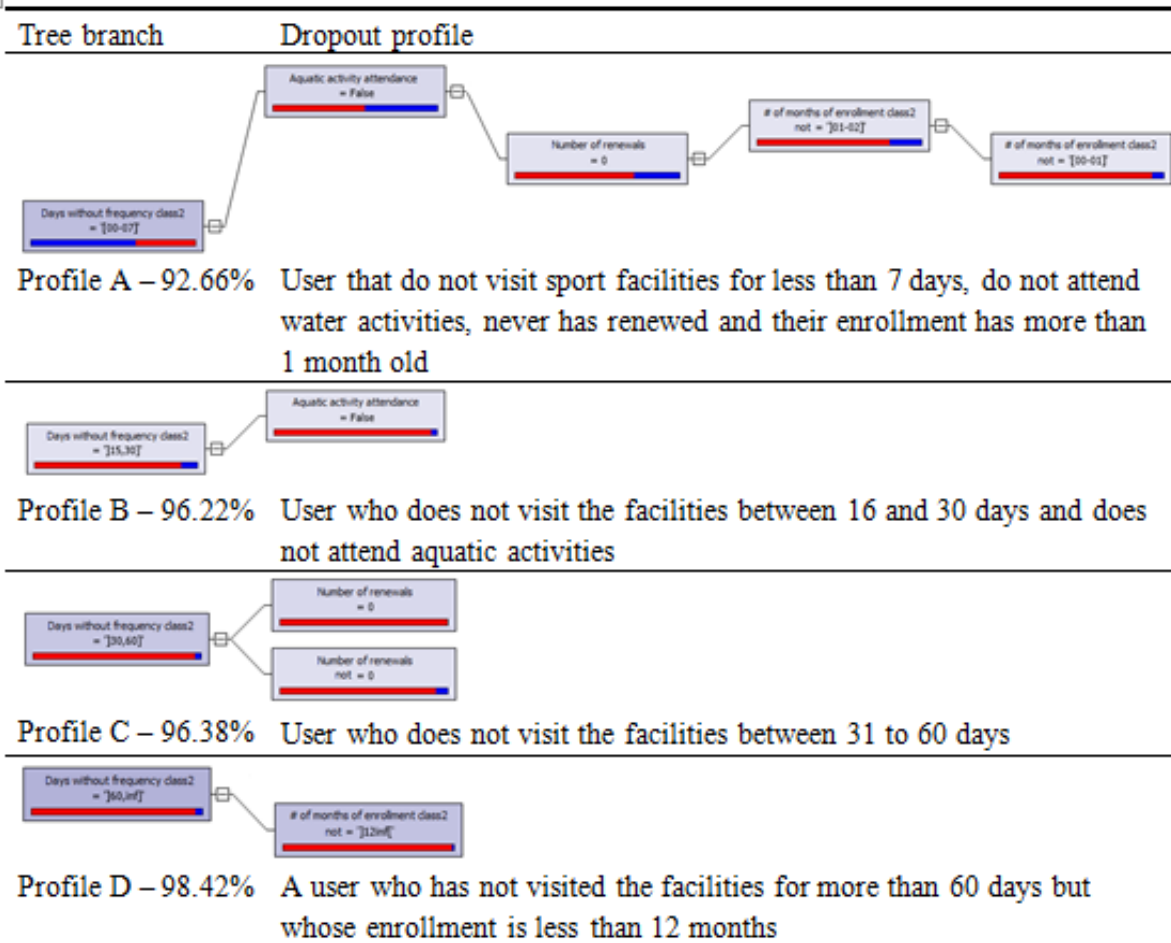
- classificação
- faltas justificadas
- faltas injustificadas
- ...



# micro-padrões



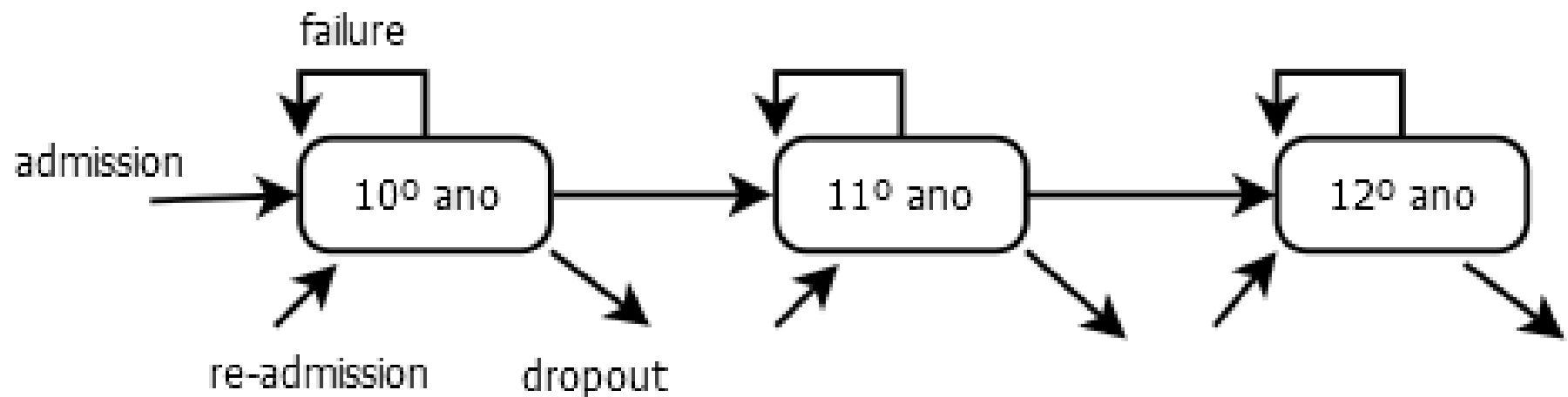
# micro-padrões: perfis



user with 93% probability of churn:

- user does not visit sport facilities in last 7 days
- &
- does not attend water activities
- &
- enrollment has more than 1 month old
- &
- has never renewed

# fluxo de estudantes - detalhe

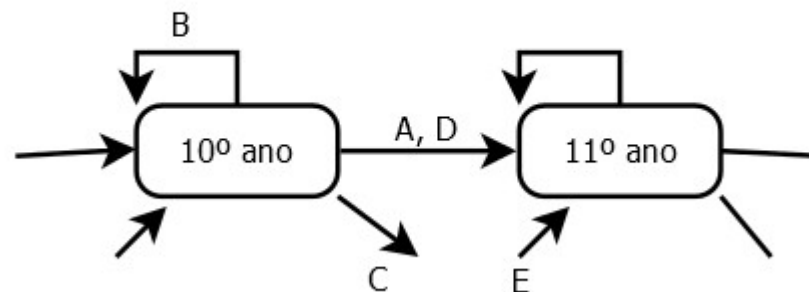


# prospecção de dados

	IdEstudante	anoLetivo	anoCurso
	A	2016	10
	A	2017	11
	A	2018	12
	B	2016	10
	B	2017	10
	B	2018	11
	C	2016	10
	D	2016	10
	D	2017	11
	E	2017	11
	E	2018	12
▶		0	0

pode  
inferir-se

2016-2017



## Situação Final do Aluno

Abandonou
Anulou Matrícula
CEF - Certificado Escolar
Conduiu
Em processo de avaliação
Excluído por Faltas
Falecido
Matriculado
Não Conduiu
Não prosseguiu
Não Transitou
Prosseguiu
Retido por Faltas
Transferido
Transitou

# técnicas de prospecção de dados

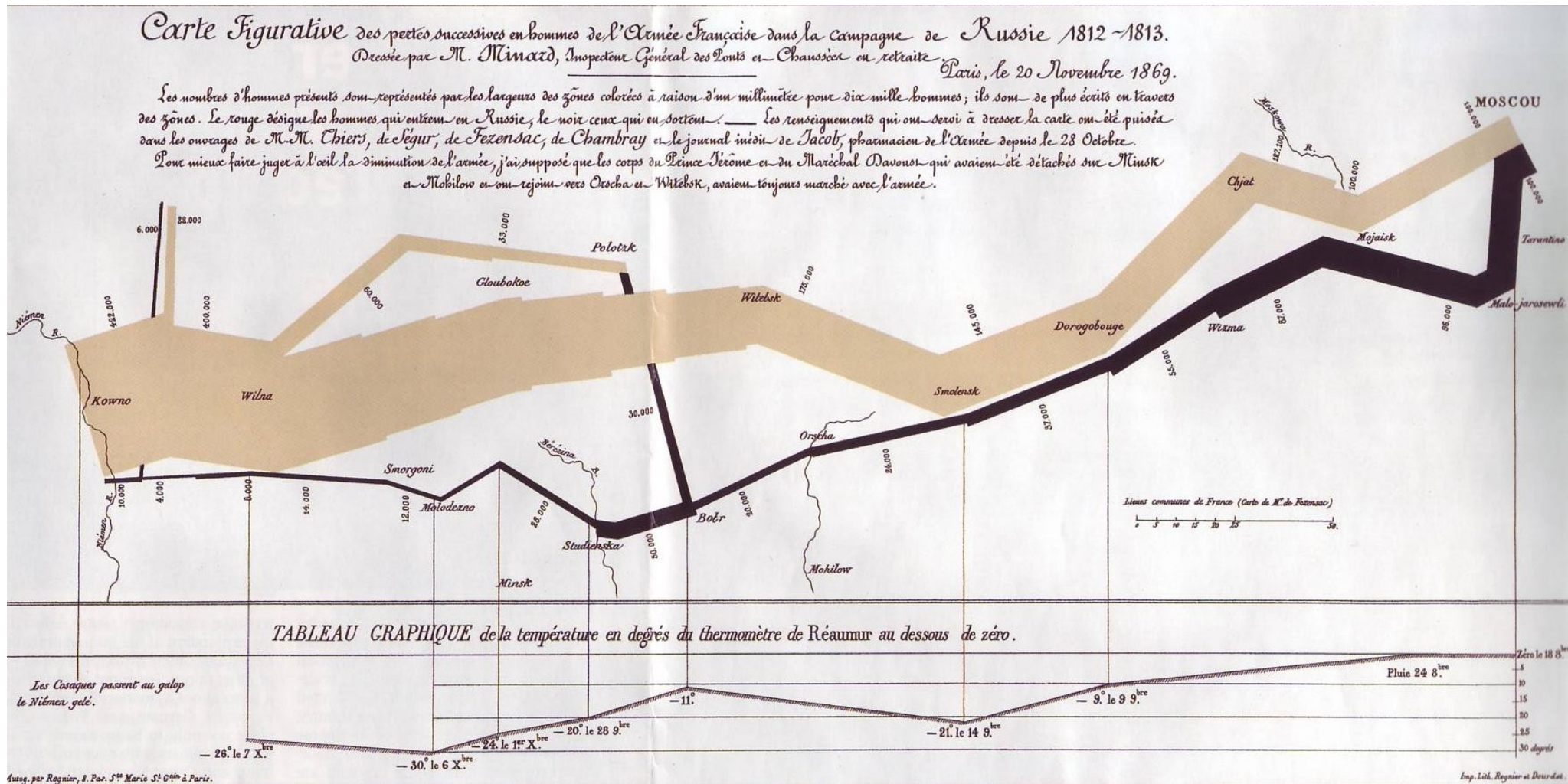
- estáticas
  - cada instância é um exemplo / ponto independente dos outros
  - instância tem atributos que não relacionam instâncias entre si
- temporais
  - cada instância é uma série de dados
  - cada elemento da série tem os atributos de uma instância estática e um **instante (índice)** associado
  - relação de ordem e possível dependência

# técnicas de prospecção de dados

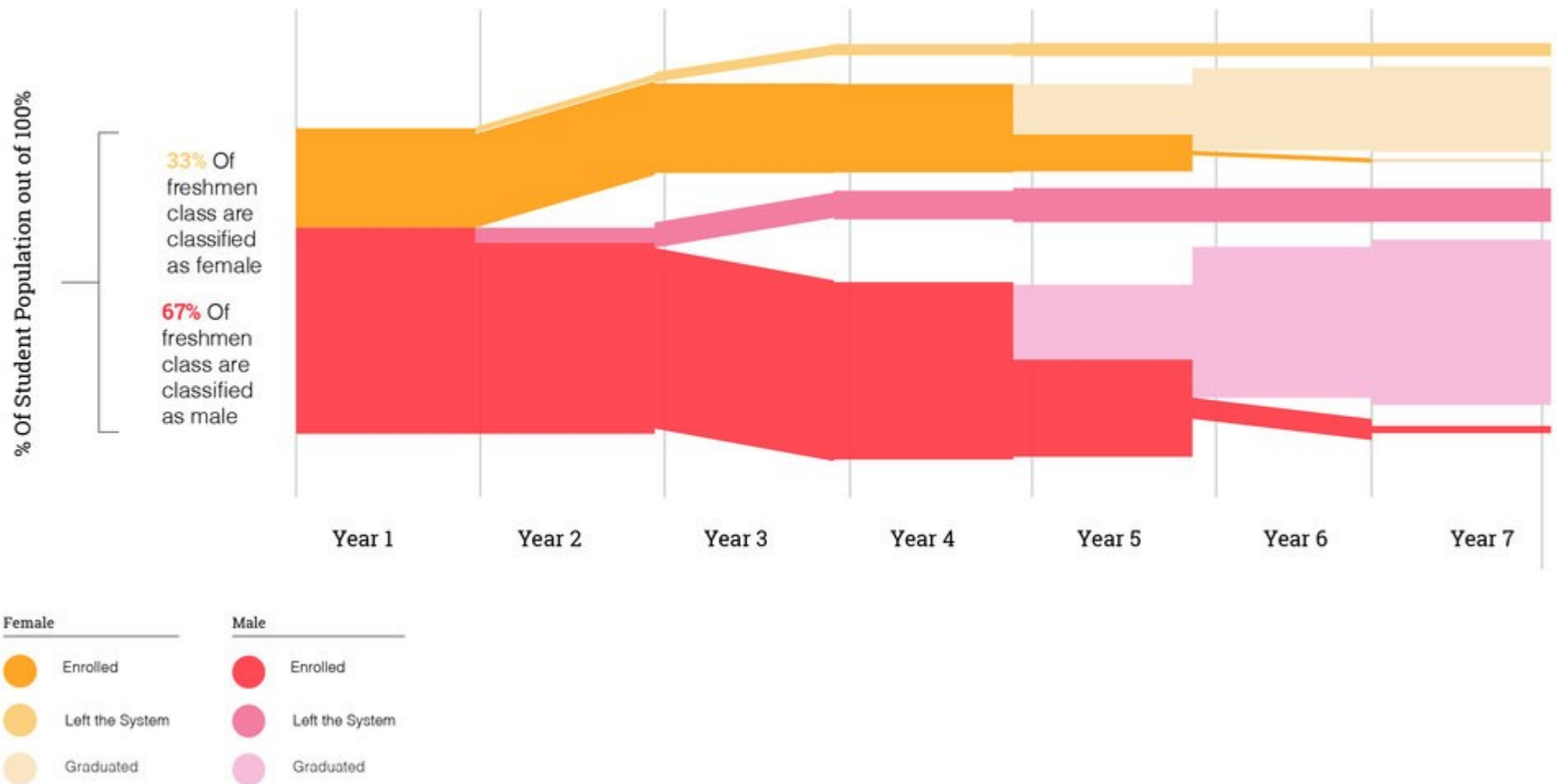
- estáticas (discretas)
  - árvore de decisão
  - rede Bayesiana
  - rede neuronal (RN)
  - ...
- temporais (discretas)
  - cadeia de Markov
  - RN temporal (c/ filtros)
  - RN com realimentação
  - ...

aprendizagem profunda (RNs)

# visualização (ex. histórico)

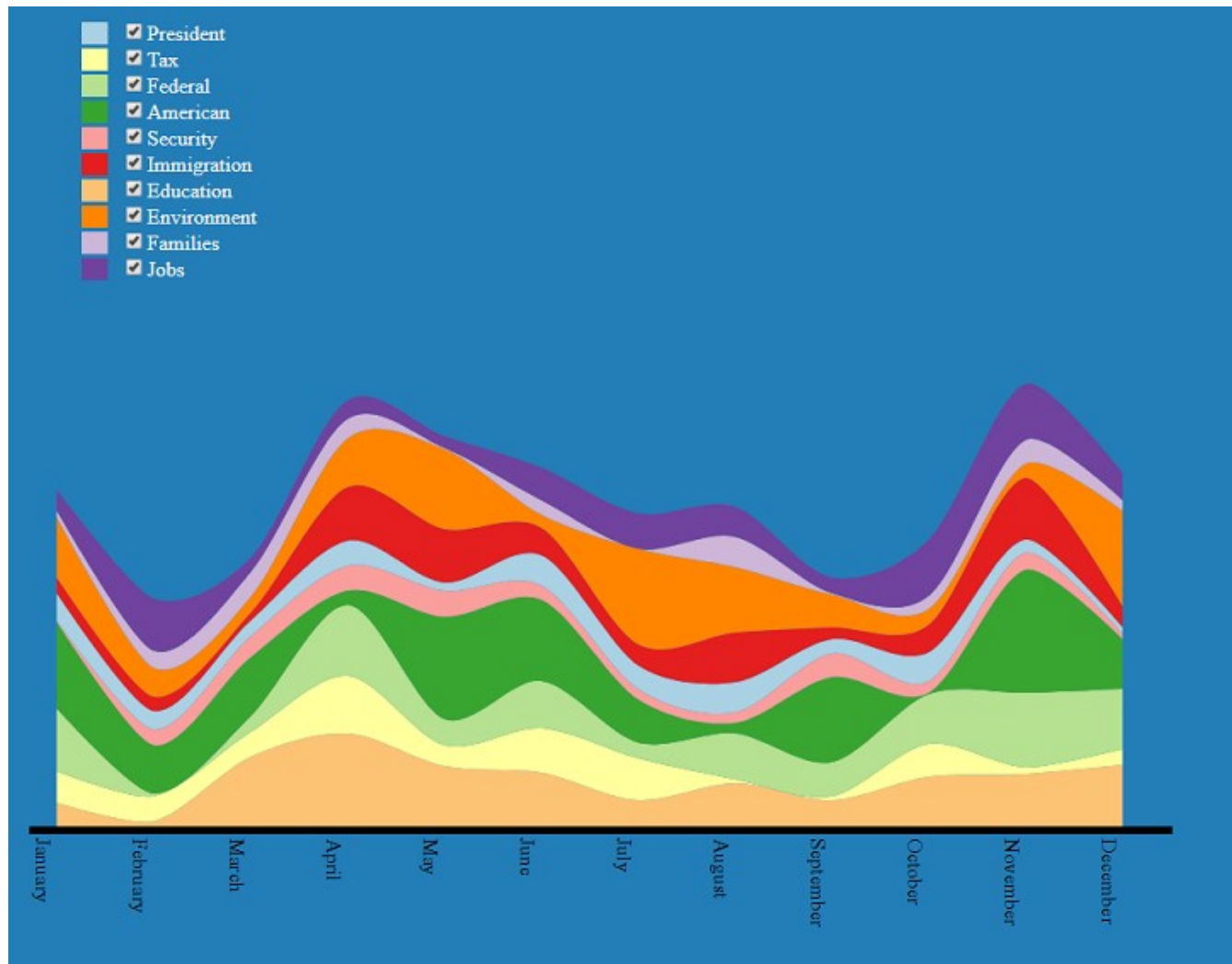


# visualização (ex. estudantes)





# visualização (ex. 2 leituras)



abordagem:

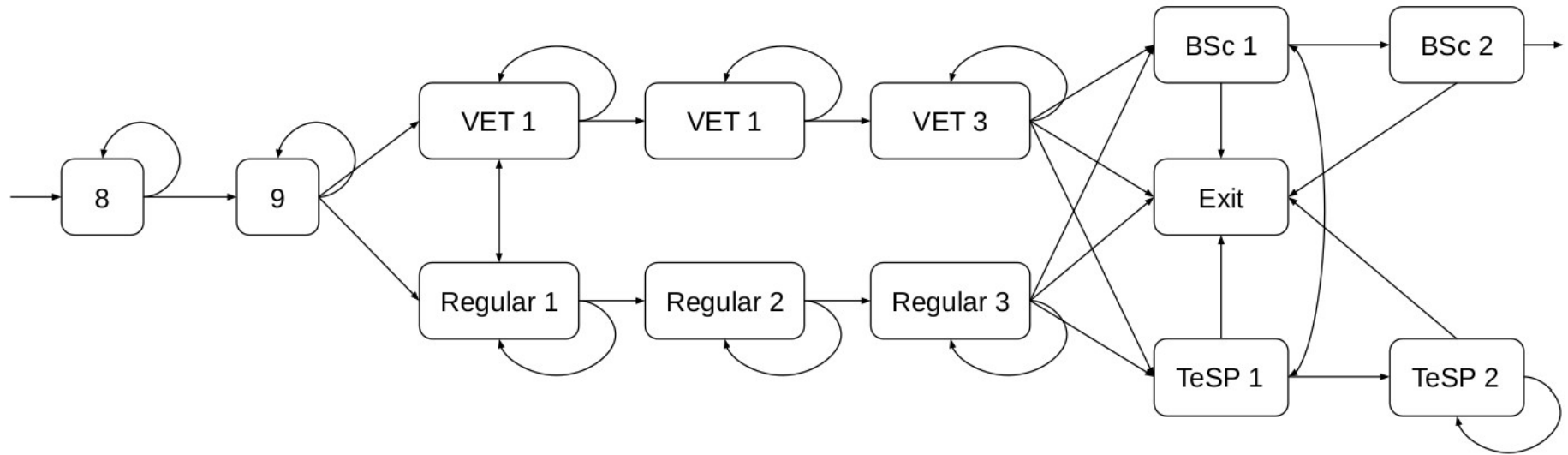
- estática  
12 pontos

- temporal  
1 série de  
12 pontos

# organização do ModEst

- projeto de 36 meses: início em janeiro de 2019
- tarefas de investigação
  - T1: Stationary Markov models
  - T2: Markov model with evolving probabilities
  - T3: Individual student model
  - T4: Graphical interface for interrogation and data visualisation

# T1: stationary Markov models



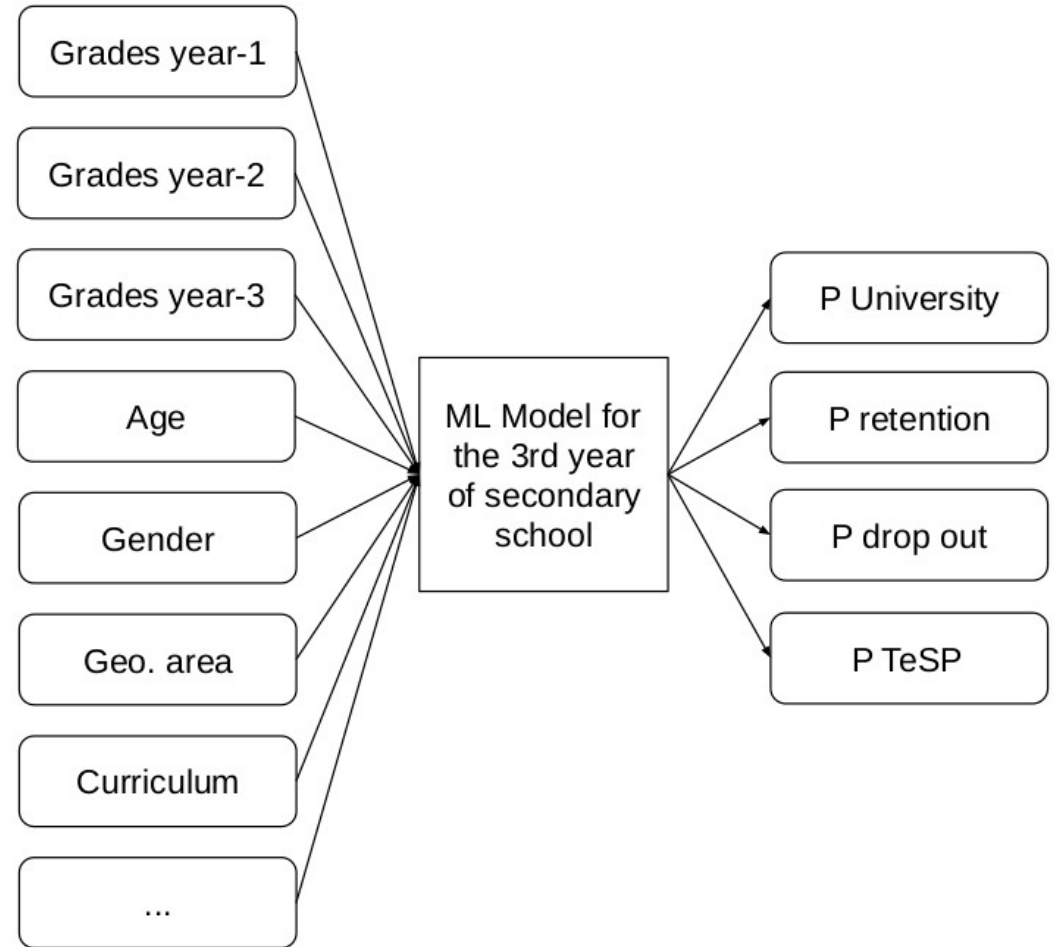
- preparação de dados
- elaboração de modelos de Markov estacionários
  - estimativa das probabilidades de transição
  - dados agregados

# T2: Markov model with evolving probabilities

- investigar modelos de Markov com probabilidades de transição que podem variar anualmente
- utilizar algoritmos de aprendizagem automática para estimar os valores das probabilidades de transição
  - tendo em conta variáveis socio-económicas
  - desenvolver mecanismos que permitam explicar a dependência das probabilidades de transição das variáveis de contexto
- investigar modelos alternativos (ex: profundos)

# T3: Individual student model

- usar dados individuais para a elaboração dos modelos
- escolha dos atributos
- análise de diferentes modelos preditivos
- identificar protótipos de alunos, e modelá-los com modelos de agente



# T4: Graphical interface for interrogation and data visualisation

- desenvolvimento da interface de utilizador da aplicação (protótipo) web a desenvolver
- desenvolvimento dos modelos de visualização de dados
- a interface deve permitir segmentações de dados e interrogações específicas aos modelos de previsão

# contribuição da DGEEC

- “cliente”
- conhecimento profundo dos dados
  - orientação sobre a respetiva semântica e tratamento
- familiaridade com pedidos habituais doutras entidades
- ideias sobre novo conhecimento que o ModEst pode proporcionar
  - prospeção dirigida
- requisitos da interface do protótipo

# contributo para a ciência

- modelos preditivos e descritivos dos fluxos e perfis dos alunos
- modelo de agente (aluno)
- modelo de visualização do fluxo de alunos
- *educational data mining / learning analytics*



# contributo para a sociedade

- previsões fiáveis dos fatores relevantes para descrever os fluxos e perfis dos alunos no sistema educativo Português
- protótipo de aplicação (alimentada automaticamente) centrada no utilizador, que permita à DGEEC extrair conhecimento relevante
- *melhoria da utilização de recursos do sistema educativo(!)*

obrigado!