Uma experiência em Ciência de Dados aplicada a Projetos Sociais

Laura de Oliveira F. Moraes
PESC/COPPE/UFRJ

Data Science For Social Good

Summer Fellowship

THE UNIVERSITY OF CHICAGO

Data Science For Social Good Europe

Summer Fellowship 2018

CASCAIS NOVA NOVA SCHOOL OF BUSINESS & ECONOMICS THE UNIVERSITY OF CHICAGO

"The goal for the program was to find people who are interested in data and analytics and want to use those skills to help society" -- Rayid Ghani, Director of Data Science for Social Good Fellowship

https://www.uchicago.edu/features/tackling_citys_challenges_with_data/

Data Science For Social Good

Summer Fellowship

THE UNIVERSITY OF CHICAGO

Data Science For Social Good Europe

Summer Fellowship 2018

CASCAIS NOVA SCHOOL OF BUSINESS & ECONOMICS THE UNIVERSITY OF CHICAGO

"The goal for the program was to find people who are interested in data and analytics and want to use those skills to help society" -- Rayid Ghani, Director of Data Science for Social Good Fellowship

https://www.uchicago.edu/features/tackling_citys_challenges_with_data/

Educação	Saúde	Meio-ambiente	Justiça Criminal
Transporte	Segurança Pública	Serviços Sociais	Desenvolvimento Econ.

Rayid Ghani

- Cientista de Dados Chefe da campanha do Obama em 2012
- 10 anos como Pesquisador na Accenture

Rayid Ghani

- Cientista de Dados Chefe da campanha do Obama em 2012
- 10 anos como Pesquisador na Accenture

2013 — 2014 — 2015 — 2016 — 2017 — 2018

2018

~ **40** estudantes/ano
~ **12** semanas
~ **9** mentores
~ **12** projetos/ano

THE UNIVERSITY OF CHICAGO

NOVA SCHOOL OF BUSINESS & ECONOMICS


Treinar estudantes
Objetivos
Construir uma comunidade
Treinar parceiros

DATA SCIENCE FOR SOCIAL GOOD
ATLANTA 2016

Data Science For Social Good
Summer Fellowship

THE UNIVERSITY OF CHICAGO

KDD 2014
This year's special theme:
Data Science for Social Good

AI for Social Good

UNIVERSITY of WASHINGTON
eScience Institute
DATA SCIENCE FOR SOCIAL GOOD

IBM Social Good Fellowship

SoGood 2016

BAYES IMPACT

Volunteerism for the Data Generation:
Using Data Science Superpowers for Social Good

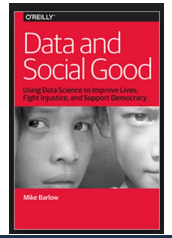
Statistics for Social Good

We're a group of Stanford students, researchers, and faculty exploring the potential to promote social good through effective data analysis.

NC DATA4GOOD
« data crunch for social good »

#Data4Good

Partnership for Social Good



DO GOOD DATA 2015

Data Science Initiative
UNC CHARLOTTE

DataKind
USING DATA IN THE SERVICE OF HUMANITY

DRIVENDATA

Data Science For Social Good Europe
Summer Fellowship 2018

CASCAIS
NOVA SCHOOL OF BUSINESS & ECONOMICS
THE UNIVERSITY OF CHICAGO

Problema importante

Problemas sociais crônicos com ações de impacto.

Dados

Dados estão disponíveis e acessíveis durante o projeto.

+

Parceiro comprometido

Possui a experiência diária e é especialista nos dados.

Problema desafiador e solucionável

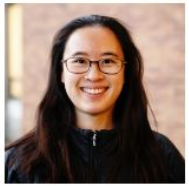
https://dssg.uchicago.edu/2015/11/04/what-makes-a-good-dssg-project/

Pranjal Bajaj

Quantitative Methods in the Social Sciences
Columbia University

Andrew Bell

Mathematical Sciences
Clemson University



Ruqian Chen

Mathematics
University of Washington

Bruno Del Papa

Physics/Neuroscience
Frankfurt Institute for Advanced Studies and Max Planck Institute for Brain Research, Goethe University Frankfurt

Anne Driscoll

Statistical Science
Duke University

Jordan Kupersmith

Masters of Information and Data Science
University of California at Berkeley

Indu Manickam

Electrical & Computer Engineering
Rice University

Kaushik Mohan

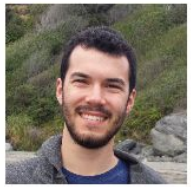
Applied Statistics for Social Science Research
New York University

Laura Moraes

Computer and Systems Engineering Program
Universidade Federal do Rio de Janeiro

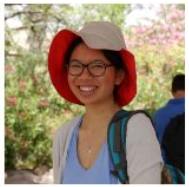
Harsh Nisar

Information Communication Technology ICT
Dhirubhai Ambani Institute Information Communication Technology (DA-IICT)



Alexander Rich

Psychology, New York University



Mélisande Teng

Applied Mathematics,
CentraleSupélec / ENS Paris-Saclay

Can Udomcharoenchaikit

Computer Engineering
Chulalongkorn University

Orsolya Vásárhelyi

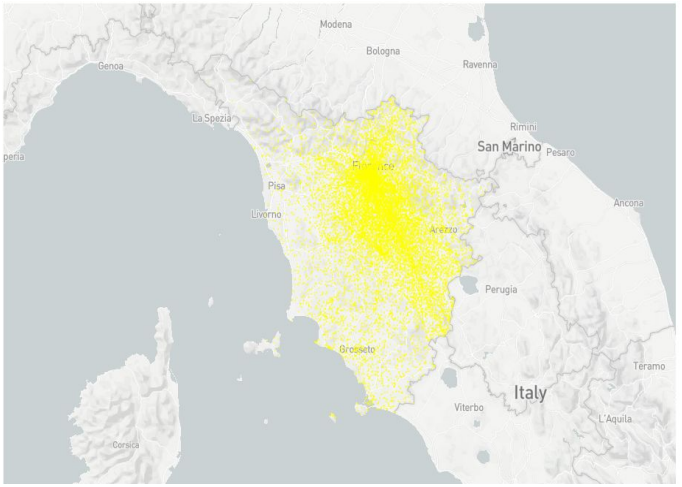
Center for Network Science
Central European University

Yanbing Wang

Agricultural Economics
Purdue University

15 estudantes

4 projetos


Modena
Bologna
Ravenna
Genoa
La Spezia
Rimini
San Marino
Pesaro
Pisa
Livorno
Ancona
Arezzo
Perugia
Grosseto
Teramo
Italy
Viterbo
L'Aquila
Corsica

Dados

11 years
of transactional history
from 2007 to 2017

64 million
interactions between
unemployed people and
IEFP

66
macroeconomic
indicators

170
features

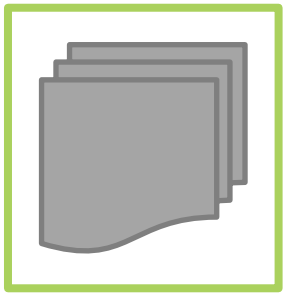
IEFP Data

Demographic | Professional

Geographical | IEFP Interactions

External Data

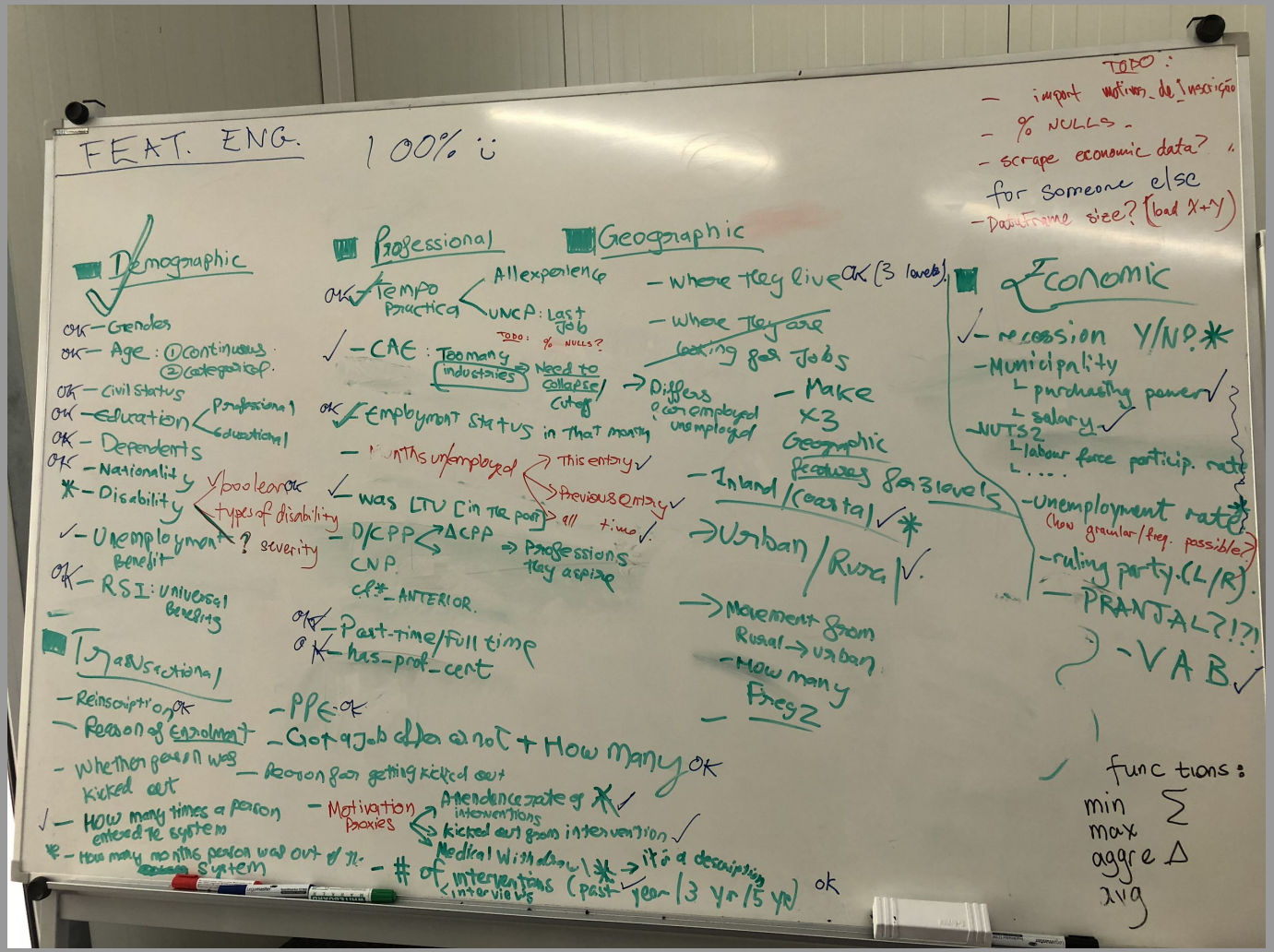
Macroeconomic Indicators



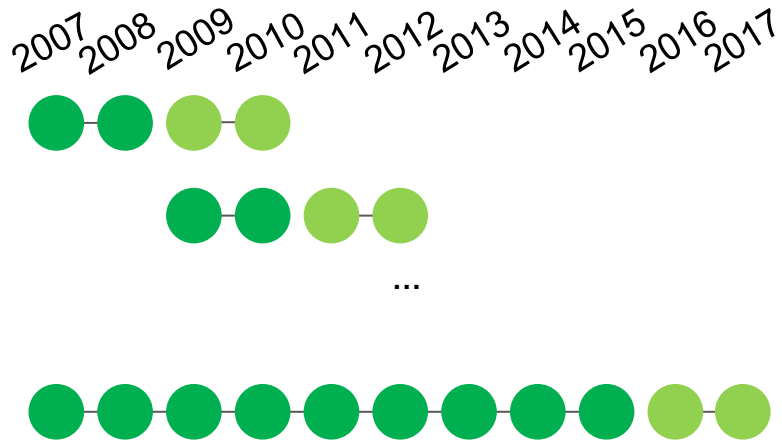
Data Story

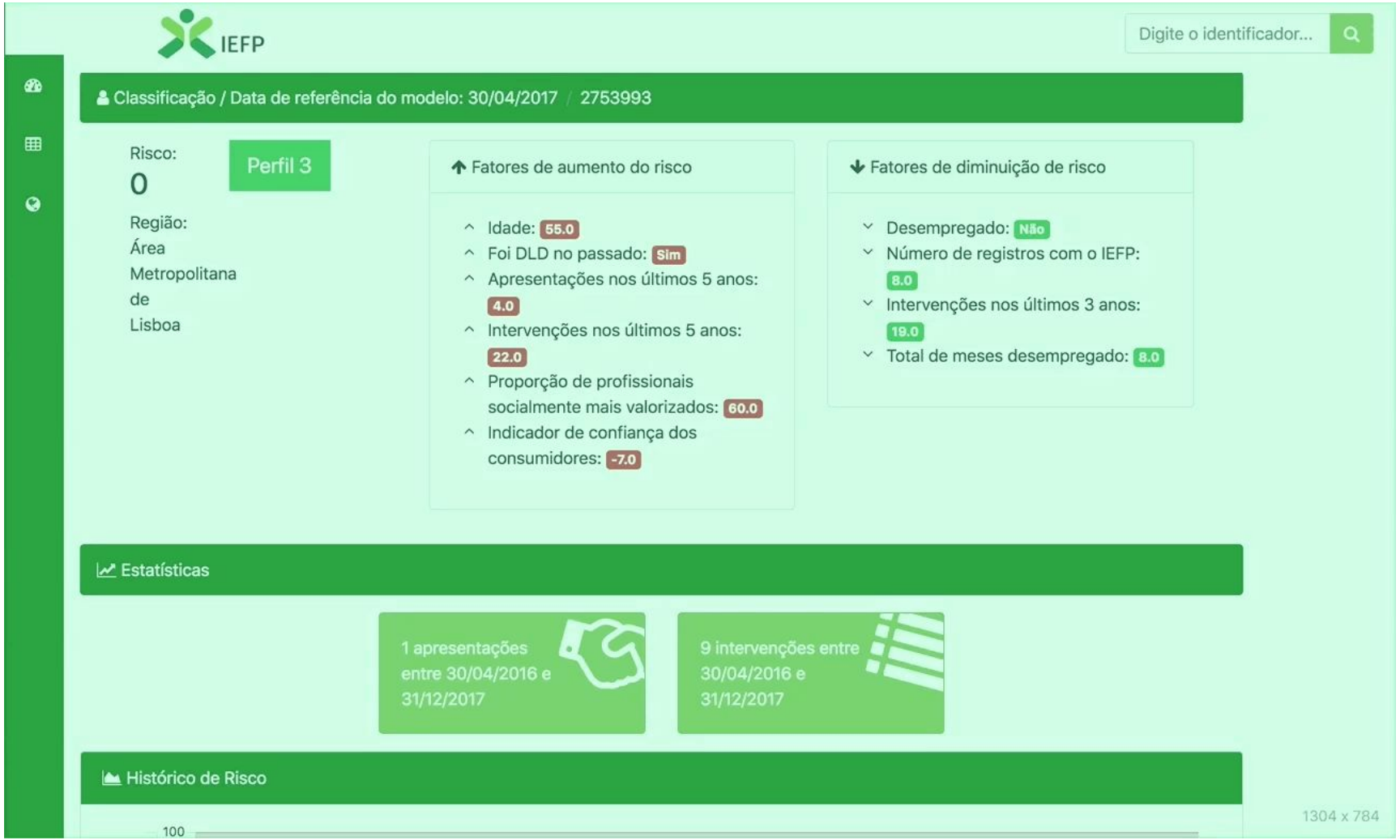

29.12.2006
01 | Desempregado in Rendufinho (New Job)
"Outros Motivos"
09.10.2007
38 | Convocado
Oferta Apresentado
15.10.2007
37 | Apresentação
Conseguida por Fafe
16.10.2007
Não conseguiu a vaga
Sem as qualificações desejadas
31.10.2007
31 | Falta controlo
Cancelado

Fizemos um *brainstorm* dos atributos...


FEAT. ENG. 100% :)
TODO:
- import motivos_de_inscrição
- % NULLS -
- scrape economic data?
for someone else
- Dataframe size? (bad X+Y)
Demographic
OK - Gender
OK - Age: ① continuous ② categorical
OK - Civil status
OK - Education
Professional
Educational
OK - Dependents
OK - Nationality
* - Disability
Yes/no or types of disability
✓ - Unemployment benefit ? severity
OK - RSI: Universal Benefit
Transactional
- Reinscription OK
- Reason of enrolment
- Whether person was kicked out
✓ - How many times a person entered the system
* - How many months person was out of the system
Professional
OK - Tempo Practica
All experience
UNCP: Last Job
✓ - CAE: Too many industries
TODO: % NULLS?
Need to collapse/cutoff
OK ✓ Employment status in that month
- Months unemployed
This entry ✓
Previous entry ✓
all time ✓
✓ - was LTU [in the past]
- D/CPP ΔCPP
CNP
cf* - ANTERIOR.
⇒ Professions they aspire
OK - Part-time/full time
OK - has_prof_cert
- PPE: OK
- Got a job after or not + How many OK
- Reason for getting kicked out
- Motivation proxies
Attendance rate of ✓
interventions
kicked out from intervention ✓
Medical Withdraw * → it's a description
- # of interventions (past year / 3 yr / 5 yr) OK
<interviews
Geographic
- Where they live OK (3 levels)
- Where they are looking for jobs
→ Differs if employed/unemployed
- Make x3 Geographic features for 3 levels
- Inland / Coastal ✓ *
→ Urban / Rural ✓
→ Movement from Rural → Urban
- How many
Freq?
Economic
✓ - recession Y/N? *
- Municipality
└ purchasing power ✓
└ salary ✓
- NUTS2
└ labour force particip. rate
└ ....
- Unemployment rate *
(how granular/freq. possible?)
- ruling party (L/R)
- PRANTAL?!?
- VAB ✓
functions:
min Σ
max
aggre Δ
avg

Validação cruzada temporal




IEFP
Digite o identificador...
Classificação / Data de referência do modelo: 30/04/2017 / 2753993
Risco:
0
Perfil 3
Região:
Área
Metropolitana
de
Lisboa
Fatores de aumento do risco
Idade: 55.0
Foi DLD no passado: Sim
Apresentações nos últimos 5 anos: 4.0
Intervenções nos últimos 5 anos: 22.0
Proporção de profissionais socialmente mais valorizados: 60.0
Indicador de confiança dos consumidores: -7.0
Fatores de diminuição de risco
Desempregado: Não
Número de registros com o IEFP: 8.0
Intervenções nos últimos 3 anos: 19.0
Total de meses desempregado: 8.0
Estatísticas
1 apresentações entre 30/04/2016 e 31/12/2017
9 intervenções entre 30/04/2016 e 31/12/2017
Histórico de Risco
100
1304 x 784

IEFP

Digite o identificador...

Análise por Região

Região

Norte

Capacidade para Perfil 1

5000

Capacidade para Perfil 2

20000

Ver estatísticas

Valores limites mínimos de risco

Perfil 1: 49

Perfil 2: 33

Usar esses valores

1304 x 784

“So much of the computing agenda today is motivated by the problems that the big Internet companies face,” he observed. “But if all these kids are doing that, they are not working on the things that really matter.” -- Rayid

Obrigada!

Additional Material

DSSG 2018: General Overview of the Summer

Data Science For Social Good Europe
Summer Fellowship 2018
CASCAIS NOVA THE UNIVERSITY OF CHICAGO

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
May 28th-June 1st	June 4th- 8th	June 11th- 15th	June 18th - 22nd	June 25th - 29th	July 2nd - 6th	July 9th - 13rd	July 16th - 20th	July 23rd - 27th	July 30th - August 3rd	August 6th - 10th	August 13th - 17th
Orientation	**Project Scoping & Data Discovery**	**Project Scoping & Data Discovery**	**Data Discovery**	**Data Science Pipeline Development**	**Data Science Pipeline Development**	**Data Science Pipeline Development**	**Data Science Pipeline Development**	**Data Science Pipeline Development**	**Data Science Pipeline Development**	**Handover to Partner & Transition Planning**	**Presentations & Transition**
Make New Friends (both human and technical)	**Data Audit & Exploration**	**Finalize Project Scope & Data Stories**	**Pipeline & Technical Plan**	**Iteration 1** Build End-to-End Code Pipeline. Focus on end-to-end structure	**Iteration 1** Build End-to-End Code Pipeline. Focus on end-to-end structure	**Iteration 2** End-to-End Code Pipeline. Focus on feature development	**Iteration 2** End-to-End Code Pipeline. Focus on feature development	**Iteration 3** End-to-End Code Pipeline. Focus on evaluation, results and initial front-end demo	**Iteration 3** End-to-End Code Pipeline. Focus on evaluation, results and initial front-end demo	**Final Deliverable Acceptance & Handover to Partner**	**Final Event Presentations & Handover to Partner**
	Descriptive Statistics	Project Scope and Data Stories	Outline of Pipeline	Initial Pipeline First Version	Early Results	Prioritized Feature List	Interpretable Model(by Partner)	Results: Across models features & metrics	Actionable Deliverable (e.g. UI Demo)	Final Report	Poster
	Questions on Data for the Partner	Project Charter approved by partner with Project Plan	Mockups of actionable deliverable			Mid-Summer Partner Visit	Reproducible Data Science Pipeline			Signed off final deliverables	Presentation
	Data Dictionary	Presentation for Partner with initial insights and Data Stories	Technical Plan							Code developed during the project (e.g. clean Github repo)	