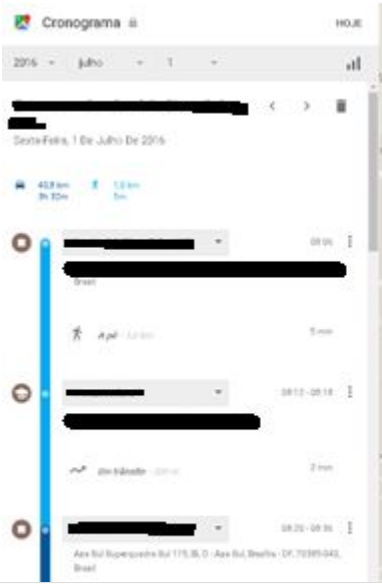
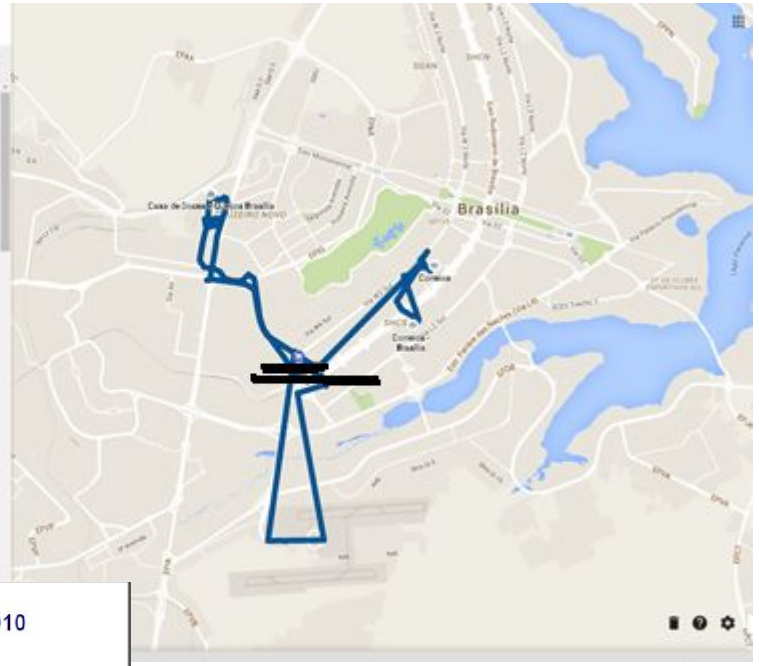
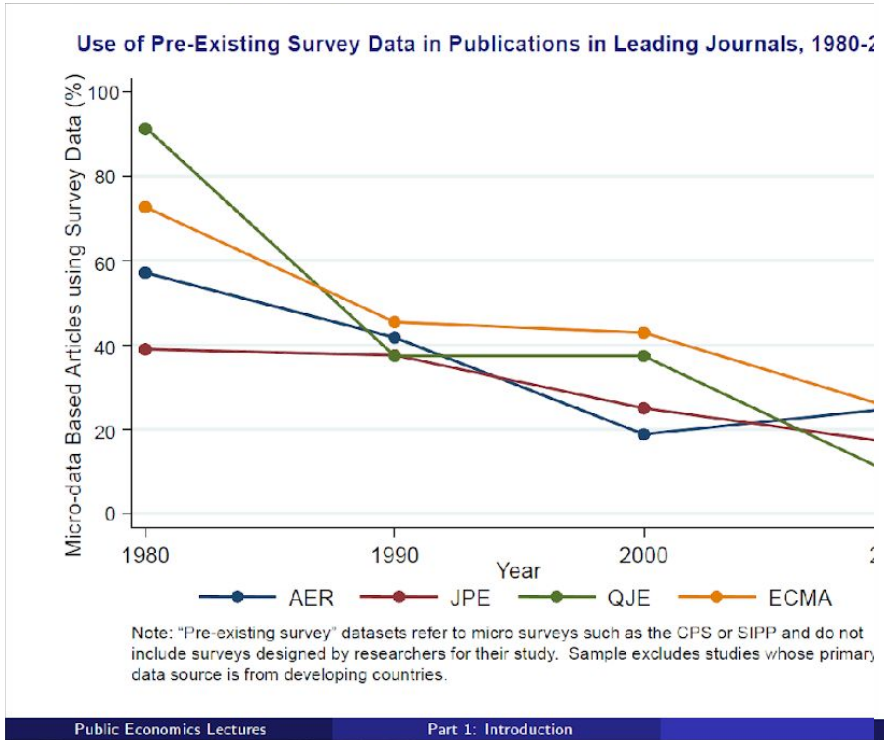
Ago/2024

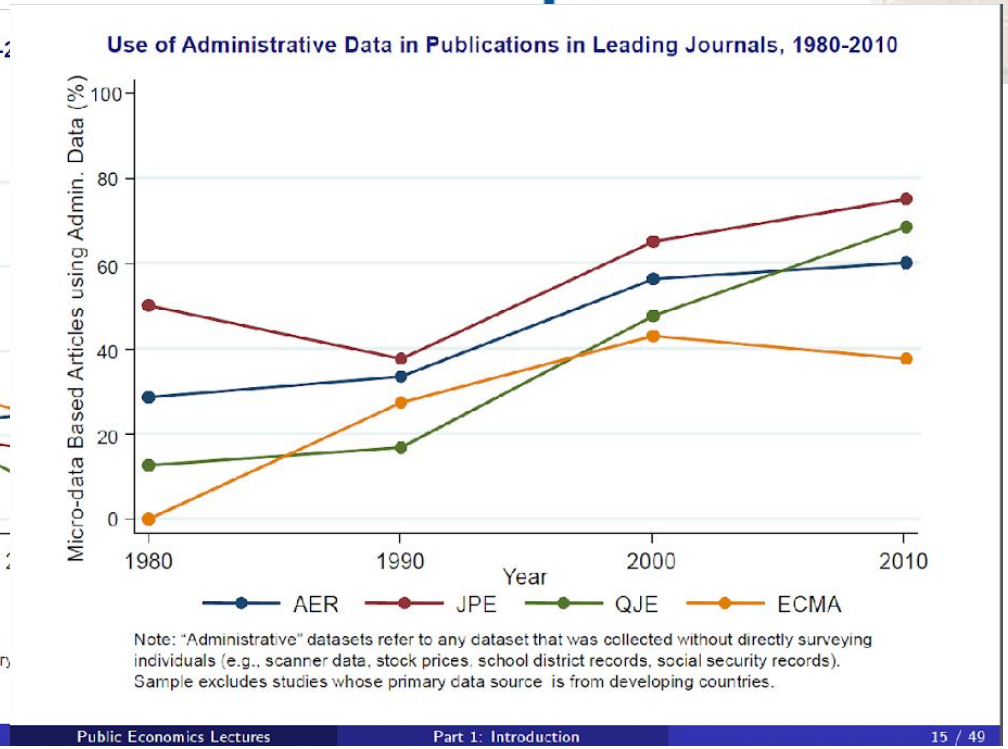
Integração de Registros Administrativos

Motivação


Cronograma
HOJE
2016
Julho
1


Brasília


Use of Pre-Existing Survey Data in Publications in Leading Journals, 1980-2
Micro-data Based Articles using Survey Data (%)
0
20
40
60
80
100
1980
1990
2000
Year
AER
JPE
QJE
ECMA
Note: "Pre-existing survey" datasets refer to micro surveys such as the CPS or SIPP and do not include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.
Public Economics Lectures
Part 1: Introduction


Use of Administrative Data in Publications in Leading Journals, 1980-2010
Micro-data Based Articles using Admin. Data (%)
0
20
40
60
80
100
1980
1990
2000
2010
Year
AER
JPE
QJE
ECMA
Note: "Administrative" datasets refer to any dataset that was collected without directly surveying individuals (e.g., scanner data, stock prices, school district records, social security records). Sample excludes studies whose primary data source is from developing countries.
Public Economics Lectures
Part 1: Introduction
15 / 49

Motivação

- Silos de dados

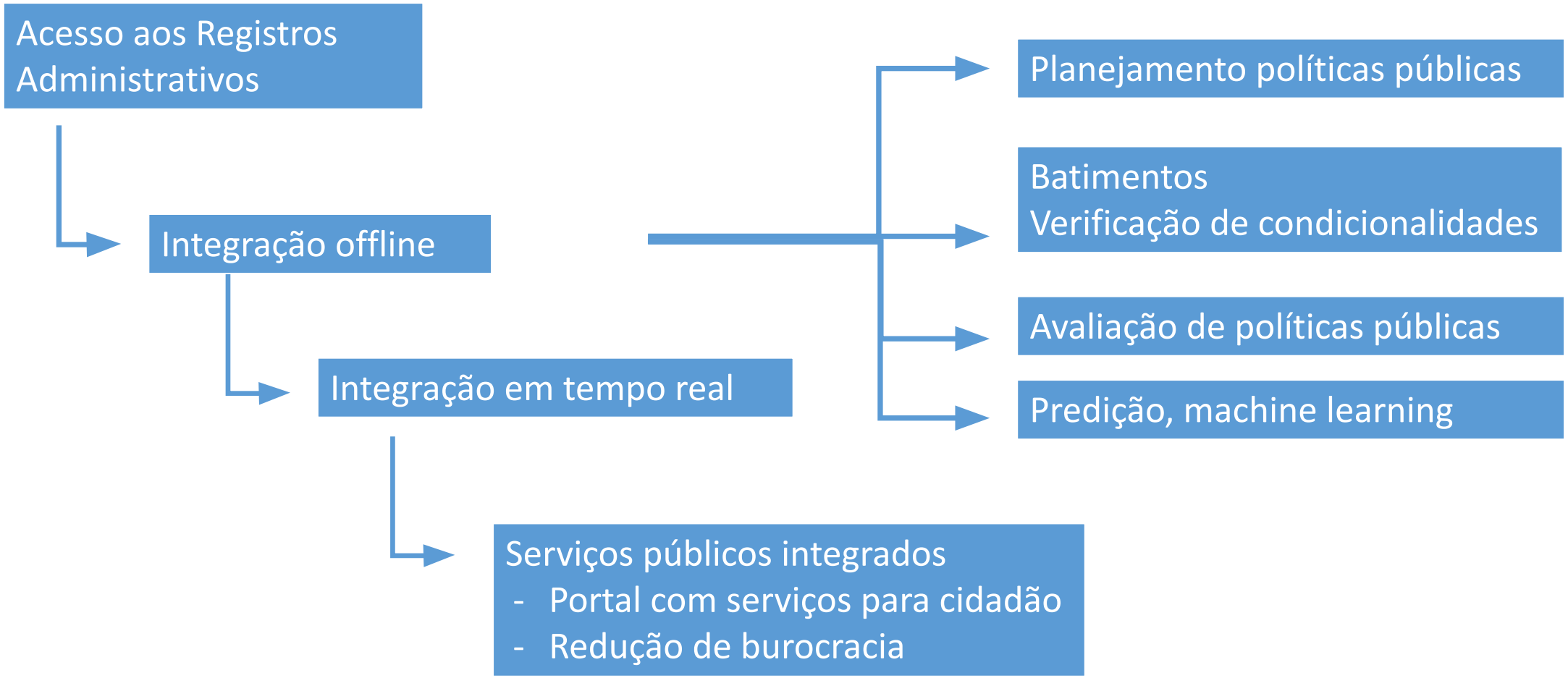
- Integração para
  - Melhorar a avaliação de políticas públicas

  - Melhorar a provisão de serviços públicos

Interoperabilidade

- Iniciativa de interoperabilidade
  - Ministério do Planejamento: Secretaria Executiva, Seplan, STI
  - IPEA
  - Serpro
  - Dataprev
- Obtenção de bases de dados
- Cruzamentos de dados
- Apoio ao CMAP

Interoperabilidade



Iniciativas

Iniciativas de integração de registros administrativos

ipeaDATA lab

SERPRO

cidacs
Centro de Integração de Dados e Conhecimentos para a Saúde

TCU

LabContas

CGU

OBSERVATÓRIO da Despesa Pública

DATAPREV

GOVDATA


ipeaDATAlab

Objetivo

- Obter, organizar e integrar registros administrativos do Governo Federal
- BMAPP: Base para Monitoramento e Avaliação de Políticas Públicas
  - Acompanhar cidadãos e empresas, do nascimento à morte, nas diversas esferas da vida
    - Benefícios sociais recebidos
    - Saúde
    - Educação
    - Renda
    - Produtividade
  - Reconstruir Vínculos Familiares
- Usar a BMAPP para subsidiar as avaliações do Ipea, especialmente CMAPP

Objetivo

- Avaliação de impacto

$$y_{i,\ t} = \alpha + \beta * P_{i,\ t}$$

- Onde:
  - i: indivíduo/empresa
  - t: período

Política pública

Objetivo

- Avaliação de impacto

Controles indivíduo

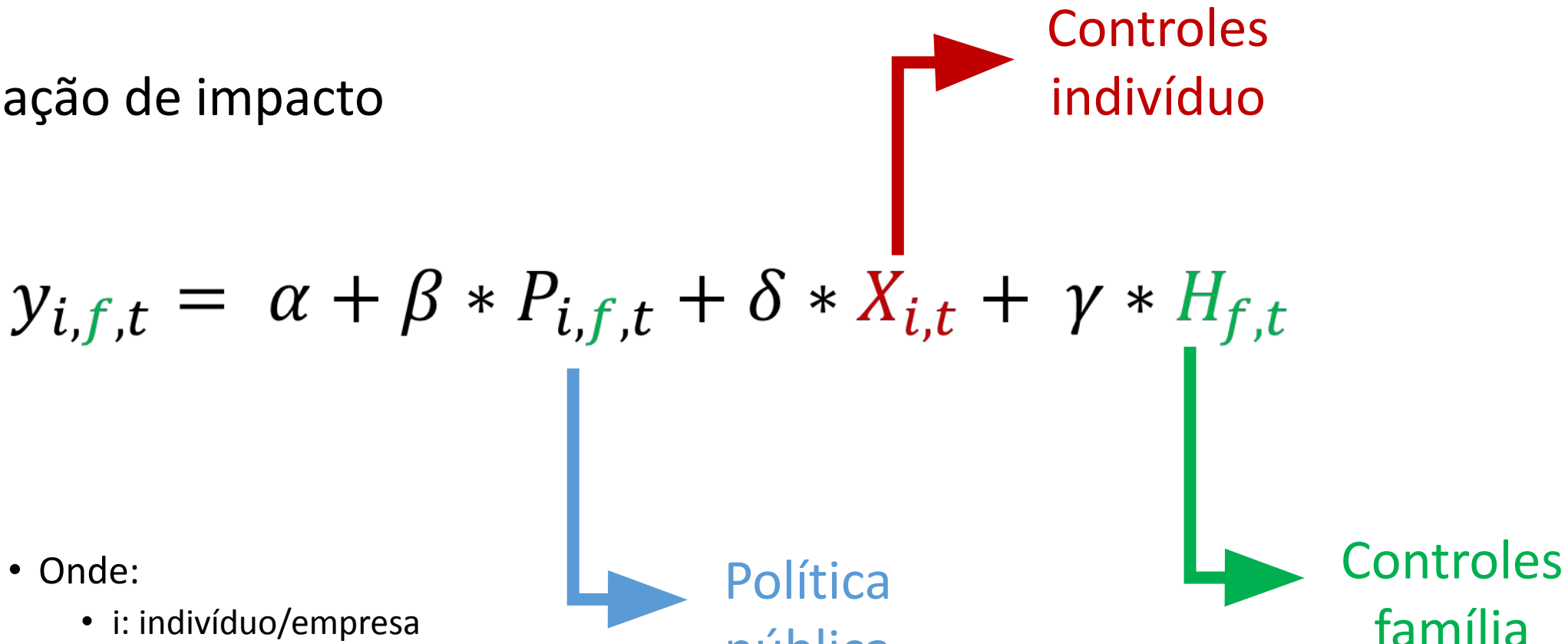
$$y_{i,\ t} = \alpha + \beta * P_{i,\ t} + \delta * X_{i,t}$$

Política pública

- Onde:
  - i: indivíduo/empresa
  - t: período

Objetivo

- Avaliação de impacto



- Onde:
  - i: indivíduo/empresa
  - t: período
  - f: família/grupo

Objetivo

- Avaliação de impacto

Educação
Saúde
Emprego
...

$$y_{i,f,t} = \alpha + \beta * P_{i,f,t} + \delta * X_{i,t} + \gamma * H_{f,t}$$

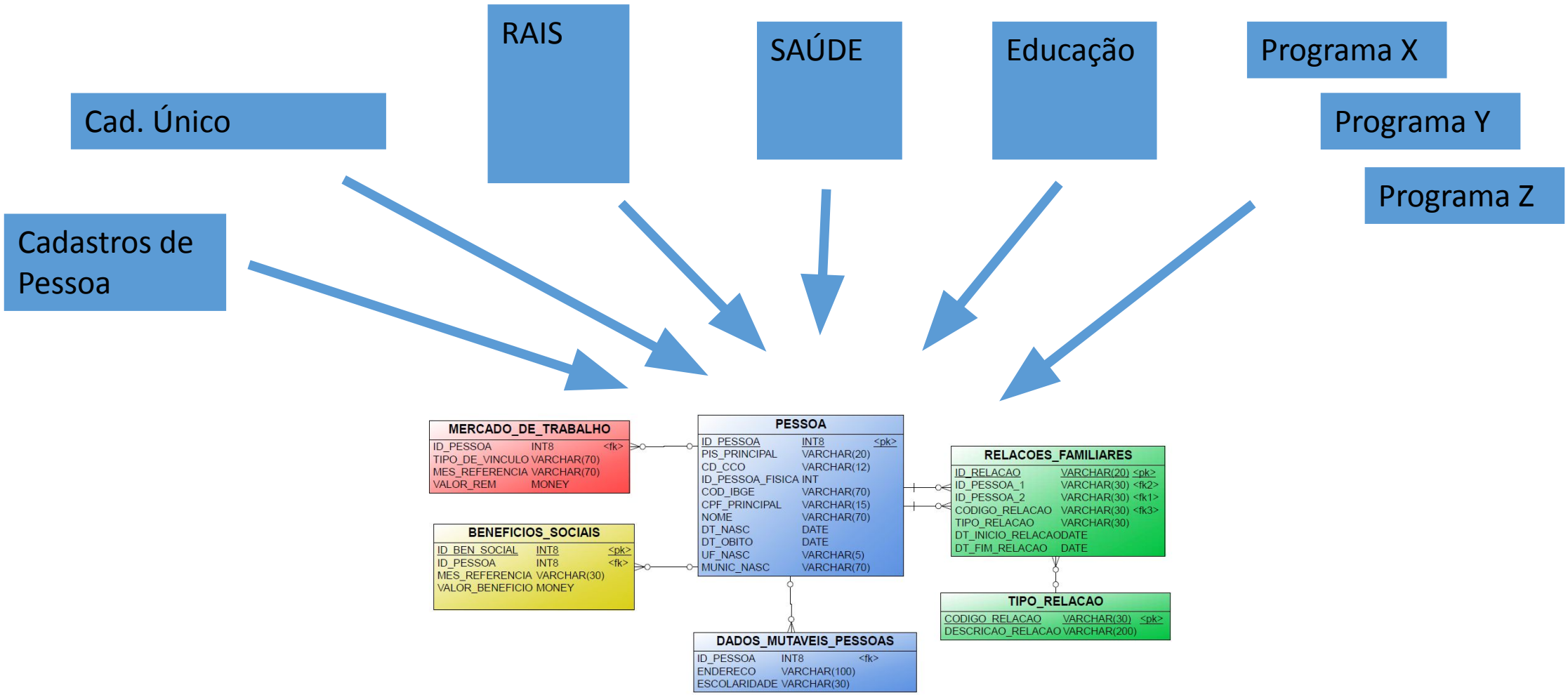
Controles indivíduo

Política pública

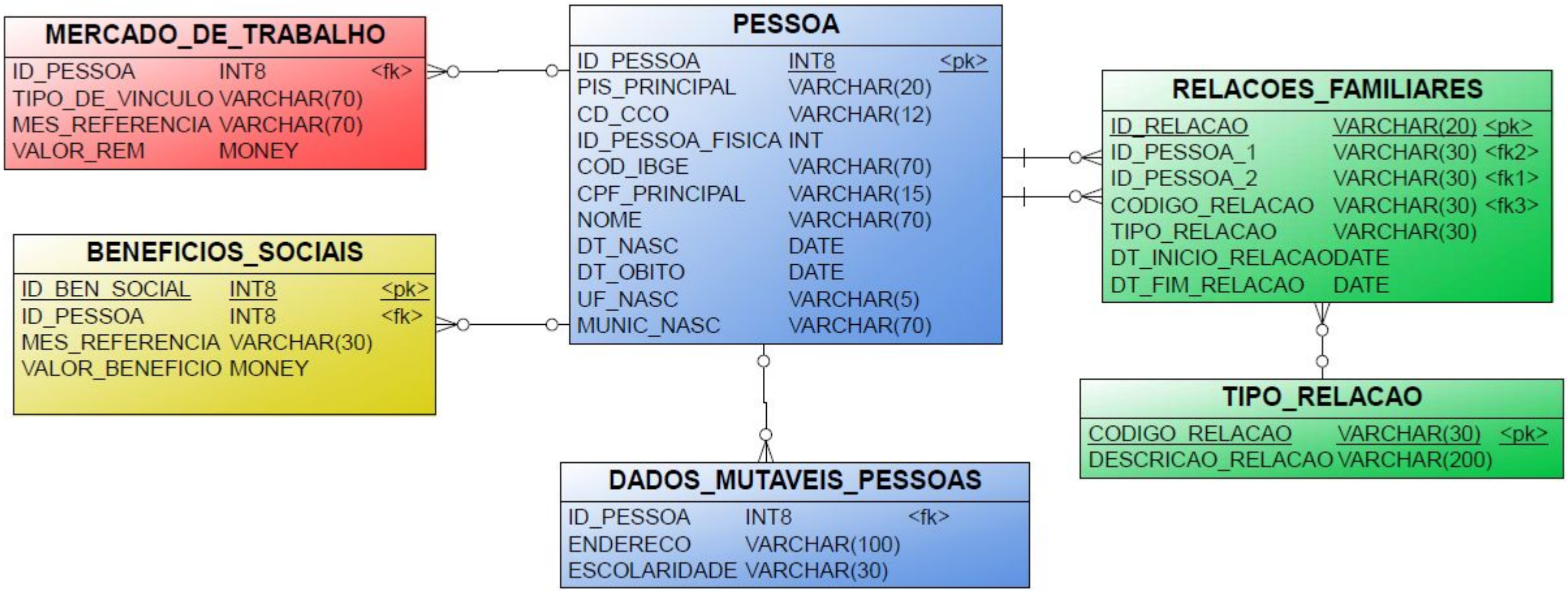
Controles família

- Onde:
  - i: indivíduo/empresa
  - t: período
  - f: família/grupo

Objetivo


Cad. Único
RAIS
SAÚDE
Educação
Programa X
Programa Y
Programa Z
Cadastros de Pessoa
MERCADO_DE_TRABALHO
ID_PESSOA INT8 <fk>
TIPO_DE_VINCULO VARCHAR(70)
MES_REFERENCIA VARCHAR(70)
VALOR_REM MONEY
BENEFICIOS_SOCIAIS
ID_BEN_SOCIAL INT8 <pk>
ID_PESSOA INT8 <fk>
MES_REFERENCIA VARCHAR(30)
VALOR_BENEFICIO MONEY
PESSOA
ID_PESSOA INT8 <pk>
PIS_PRINCIPAL VARCHAR(20)
CD_CCO VARCHAR(12)
ID_PESSOA_FISICA INT
COD_IBGE VARCHAR(70)
CPF_PRINCIPAL VARCHAR(15)
NOME VARCHAR(70)
DT_NASC DATE
DT_OBITO DATE
UF_NASC VARCHAR(5)
MUNIC_NASC VARCHAR(70)
DADOS_MUTAVEIS_PESSOAS
ID_PESSOA INT8 <fk>
ENDERECO VARCHAR(100)
ESCOLARIDADE VARCHAR(30)
RELACOES_FAMILIARES
ID_RELACAO VARCHAR(20) <pk>
ID_PESSOA_1 VARCHAR(30) <fk2>
ID_PESSOA_2 VARCHAR(30) <fk1>
CODIGO_RELACAO VARCHAR(30) <fk3>
TIPO_RELACAO VARCHAR(30)
DT_INICIO_RELACAO DATE
DT_FIM_RELACAO DATE
TIPO_RELACAO
CODIGO_RELACAO VARCHAR(30) <pk>
DESCRICAO_RELACAO VARCHAR(200)

Objetivo


MERCADO_DE_TRABALHO
ID_PESSOA INT8 <fk>
TIPO_DE_VINCULO VARCHAR(70)
MES_REFERENCIA VARCHAR(70)
VALOR_REM MONEY
PESSOA
ID_PESSOA INT8 <pk>
PIS_PRINCIPAL VARCHAR(20)
CD_CCO VARCHAR(12)
ID_PESSOA_FISICA INT
COD_IBGE VARCHAR(70)
CPF_PRINCIPAL VARCHAR(15)
NOME VARCHAR(70)
DT_NASC DATE
DT_OBITO DATE
UF_NASC VARCHAR(5)
MUNIC_NASC VARCHAR(70)
RELACOES_FAMILIARES
ID_RELACAO VARCHAR(20) <pk>
ID_PESSOA_1 VARCHAR(30) <fk2>
ID_PESSOA_2 VARCHAR(30) <fk1>
CODIGO_RELACAO VARCHAR(30) <fk3>
TIPO_RELACAO VARCHAR(30)
DT_INICIO_RELACAO DATE
DT_FIM_RELACAO DATE
BENEFICIOS_SOCIAIS
ID_BEN_SOCIAL INT8 <pk>
ID_PESSOA INT8 <fk>
MES_REFERENCIA VARCHAR(30)
VALOR_BENEFICIO MONEY
TIPO_RELACAO
CODIGO_RELACAO VARCHAR(30) <pk>
DESCRICAO_RELACAO VARCHAR(200)
DADOS_MUTAVEIS_PESSOAS
ID_PESSOA INT8 <fk>
ENDERECO VARCHAR(100)
ESCOLARIDADE VARCHAR(30)

Princípios

- Segurança, privacidade
- Simplicidade, documentação
- Default: baseado em R (software livre)
- Desenvolvimento ágil
  - Controle de versão (git, gitlab)
  - Gerenciamento por issues (git, gitlab)
  - Prototipagem rápida

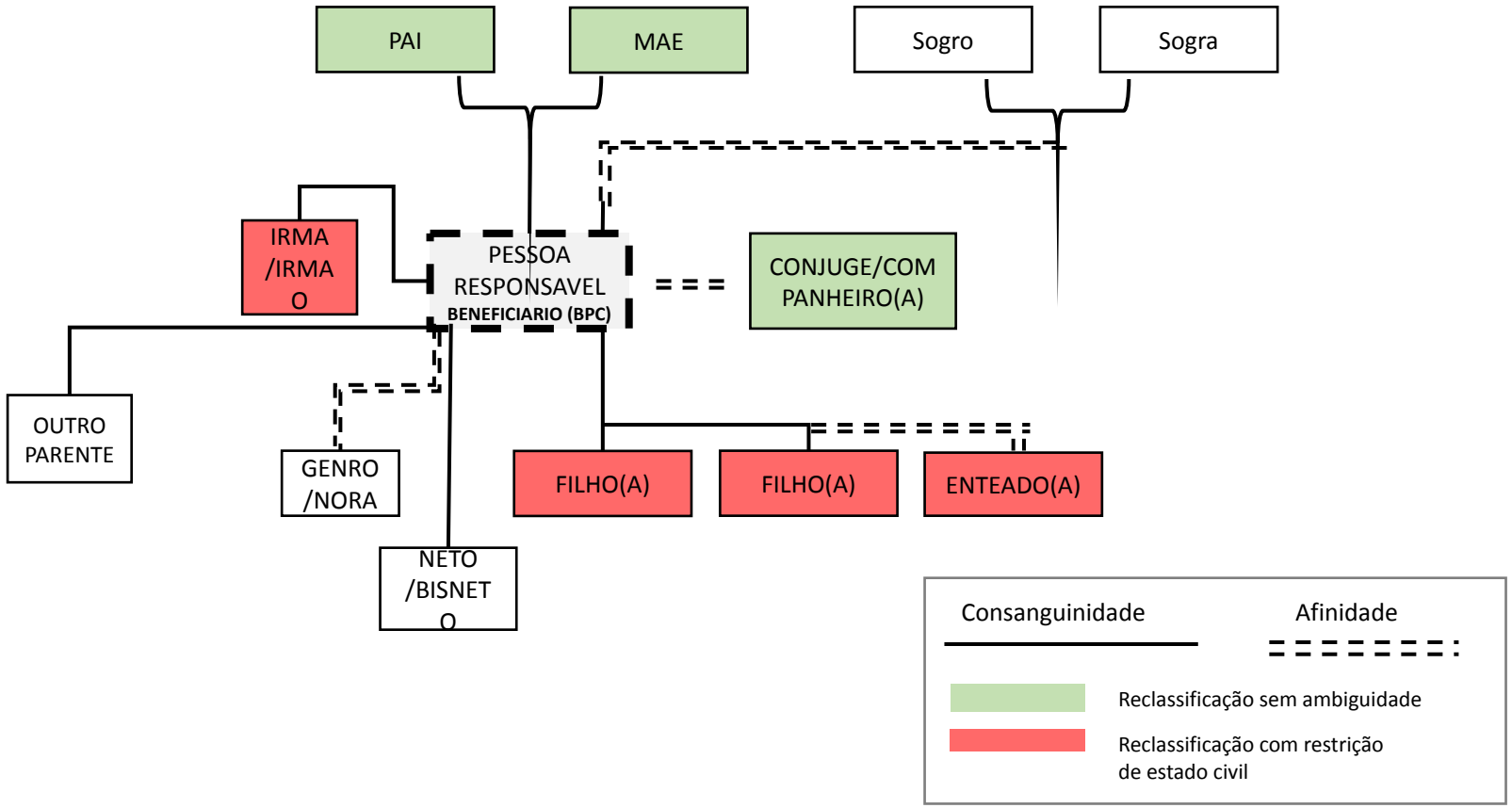
Equipe

- Equipe multidisciplinar, 12 pessoas
  - Economistas
  - Cientistas sociais
  - Cientistas da computação
  - Estatísticos

Segurança

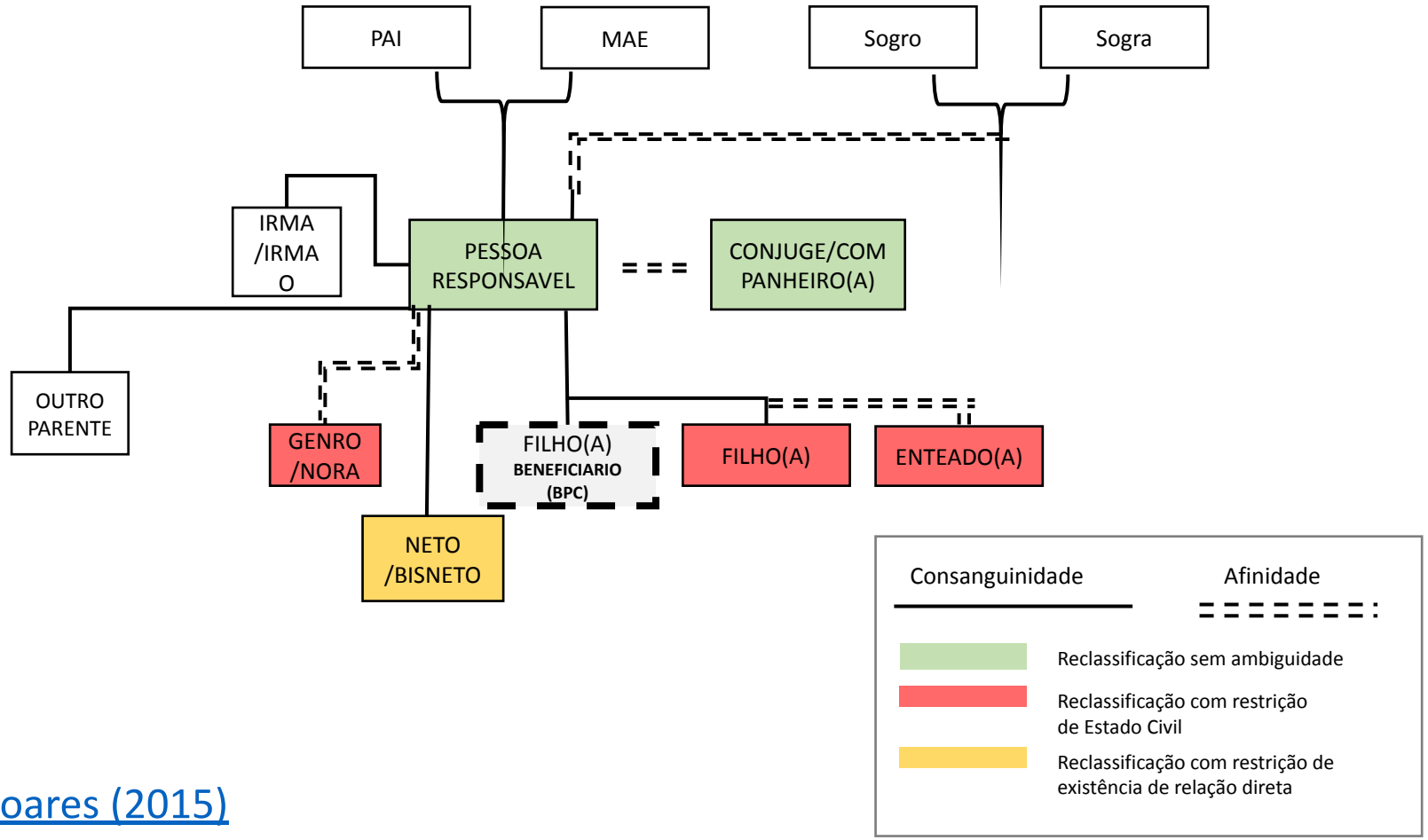
	Nível 1: Pastas restritas	**Nível 2 Servidor Restrito**	**Nível 3: Sala do Sigilo**
Termo de responsabilidade	✔	✔	✔
Entrada e saída controlada e rastreável		✔	✔
Acesso remoto em maquinas controladas		✔	✔
Isolamento lógico		✔	✔
Isolamento físico			✔
Controle biométrico de acesso			✔
Monitoramento: presencial e por vídeo			✔
Comitê de avaliação dos pedidos de liberação		✔	✔

Ex1: Família-BPC vs Família-CadÚnico



Fonte: [Mation e Moares (2015)]

Ex1: Família-BPC vs Família-CadÚnico



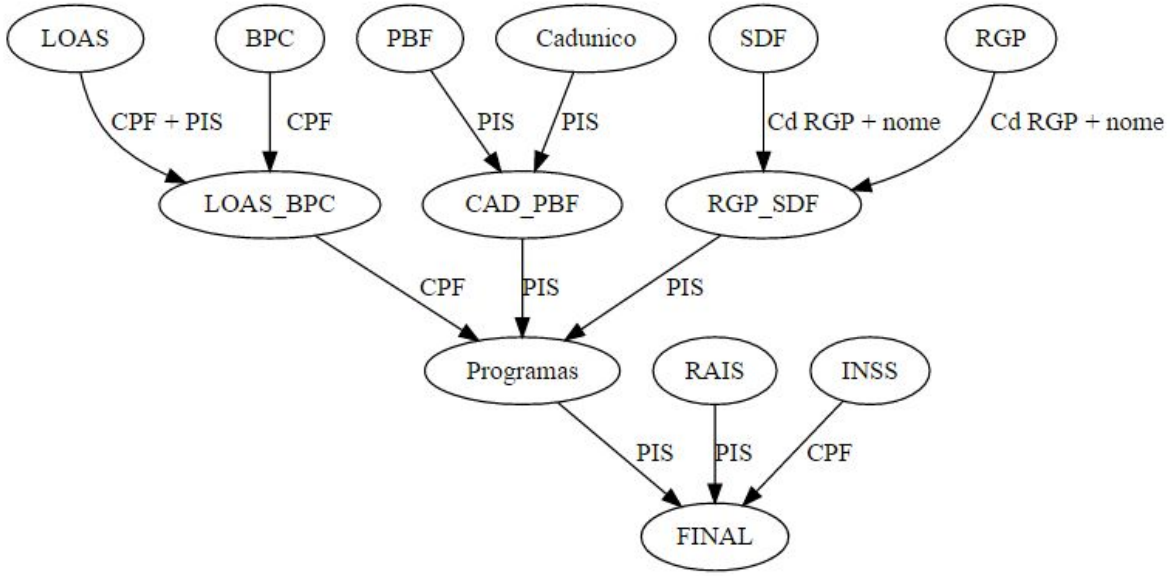
Fonte: Mation e Moares (2015)

Ex1: Família-BPC vs Família-CadÚnico

Merge	Número de observações
Só GRUFAM	4.668.301
Matched	2.289.602
Só CadÚnico Reclassificado	1.485.713
Total	**8.443.616**

Nota: Resultado após retirada das duplicatas nas chaves

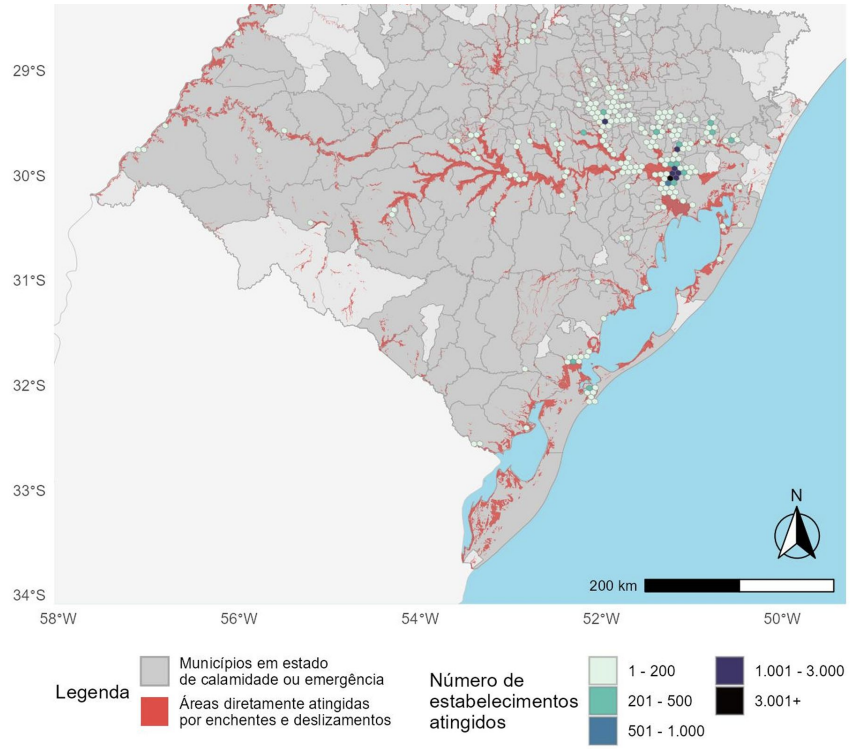
Reclassificações	N. de pessoas	(%)
Correta	2.279.966	99,6%
Errada	9.639	0,4%
Total	**2.289.605**	**100,0%**



Fonte: Mation e Moares (2015)

Ex2: Geocode

- Harmonização de endereços
  - Separação de campos (tipo, logradouro, N, complemento)
  - Harmonização por campo (ex: cs, casa, c.)
  - Correção de *typos*
- Geocode
  - Base proprietária
  - + CNEFE
- Aplicações:
  - Afetados enchentes RS



Limitações dos Registros Adm. Brasileiros

- Historicamente: Mundo PIS (Dataprev) vs Mundo CPF (Serpro)
- Falta de vínculos familiares explícitos

RAs: riscos e oportunidades

- Conhecimento substantivo de cada registro administrativo
  - Conhecer do significado das variáveis, motivação para coleta do dado
- Manutenção do acesso
  - Consolidar jurisprudência/entendimento da LGPD
- Multiplicidades de salas do sigilo

Fronteiras de pesquisa

- Registros Administrativos + pesquisas domiciliares
  - US Census, Statistics Canada, Norway, etc …
  - Mais relevante no Brasil. INFORMALIDADE
  - Valida informações da pesquisa
  - Mostra limites do registro administrativo
- Pareamento com arquivos históricos, genealógicos
  - Dados educacionais

Softwares Desenvolvidos

- microdadosBrasil
  - Facilita a leitura de microdados públicos brasileiros
  - branch microdadosBrasil-RA : facilita a leitura de registros administrativos
- validaRA
  - Facilita validação de dados de registros administrativos brasileiros
  - doc2integer64 : converte documentos para numérico retirando “-”, “.”, etc
  - validda_doc : verifica dígito verificador de CPF, PIS, CNPJ, CNS, etc...
  - relatorioDOC : gera relatório descrevendo a qualidade dos dados
    - missings, duplicidades de identificadores, histogramas
- geobr