

Avaliação quantitativa e visualização de resultados experimentais da pesquisa

Flavia Bernardini

Professora Associada

Instituto de Computação

19/03/2019

Quem sou eu?

- Graduação em Computação – UNESP / MSc e DSc (2006) em Computação – ICMC/USP – Aprendizado de Máquina
- 2006 a 2009 – ADDLABS – Projetos de P&D usando IA em problemas de O&G
- Desde 2009 – UFF Rio das Ostras
- Dez 2017 – IC/UFF

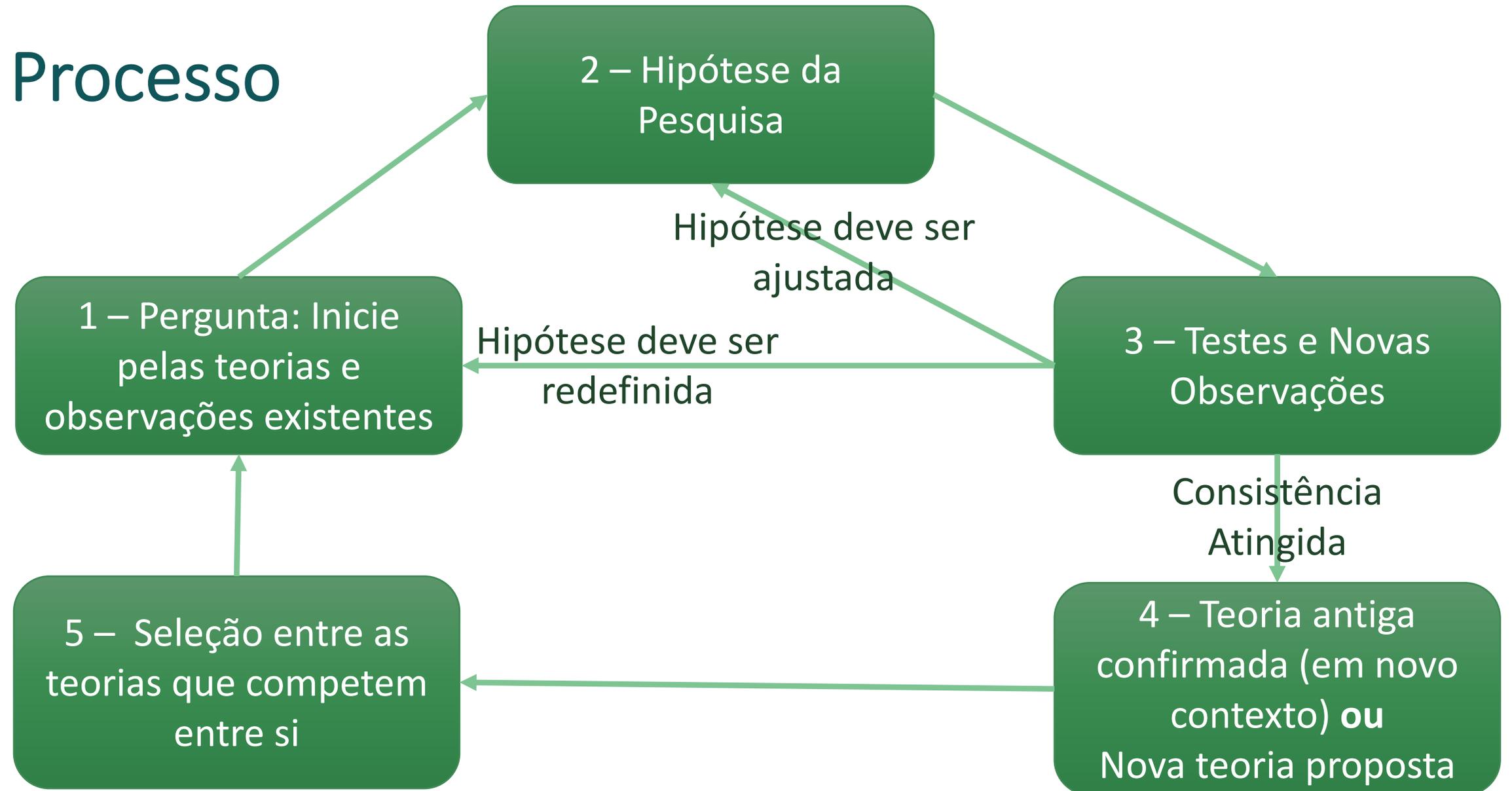


O Método Científico

O método científico

- Método científico é o esquema lógico usado pelos cientistas buscando respostas para as questões na ciência
- É usado para produzir teorias científicas, incluindo metáforas científicas (teorias sobre teorias), teorias usadas para projetar as ferramentas para a produção de teorias
 - Inclui instrumentos, ferramentas, algoritmos, etc

Processo



O método científico - Passos

1. Coloque a questão no contexto do conhecimento existente – teoria e observações

- Pode ser uma questão nova que velhas teorias são capazes de responder (geralmente o caso), ou uma pergunta que exige a formulação de uma nova teoria

2. Formule uma hipótese como uma tentativa de resposta

- Construa deduções possíveis a partir da hipótese para identificar a metodologia a ser utilizada para avaliar a hipótese

O método científico - Passos

3. Teste a hipótese em um campo específico de experimento / teoria

- A nova hipótese deve provar para se encaixar na visão de mundo existente (“ciência normal”)
- Caso a hipótese leve a contradições e demande mudanças radicais no contexto teórico existente, deve ser testado com especial cuidado
 - A nova hipótese tem que ser proveitosa e ofereça vantagens consideráveis, a fim de substituir o paradigma científico existente
 - Chamado de “Revolução científica” (Kuhn) e acontece muito raramente
- Em via de regra, o loop 2-3 é repetido com modificações da hipótese até que a concordância seja obtida, o que leva ao passo 4
 - Se grandes discrepâncias são encontradas o processo deve recomeçar a partir do passo 1

O método científico - Passos

4. Quando a consistência é obtida, a **hipótese torna-se uma teoria** e fornece um conjunto coerente de proposições que definem uma nova classe de fenômenos ou um novo conceito teórico
 - Resultados devem ser publicados
 - A teoria nesse estágio é objeto de processo de “seleção natural” entre teorias concorrentes (passo 5)
 - A teoria está então se tornando uma estrutura dentro da qual observações / fatos teóricos são explicados e previsões são feitas
 - O processo pode recomeçar do começo, mas o estado 1 mudou para incluir a nova teoria / teoria antiga melhorada

O que é CIÊNCIA da Computação?

O que é Ciência da Computação?

- Ciência da Computação é o estudo dos fenômenos relacionados aos computadores, (Newell, Perlis e Simon, 1967)
- A disciplina da computação é o estudo sistemático de processos algorítmicos que descrevem e transformam informações: teoria, análise, design, eficiência, implementação e aplicação (Computing Curricula 2001 – ACM)
- Ciência da Computação é o estudo das estruturas da **informação** (Wegner, 1968)
- Ciência da Computação é o estudo e gestão da complexidade (Dijkstra, 1969)

tradição empírica

tradição matemática

Grande complexidade de problemas de engenharia

Sub-áreas da computação

- Estruturas discretas
- Fundamentos de Programação
- Algoritmos e Complexidade
- Linguagens de Programação
- Arquitetura e Organização
- Sistemas Operacionais
- Computação Centrada em Rede
- Interação Humano-Computador
- Gráficos e Computação Visual
- Sistemas Inteligentes
- Gerenciamento de Informações
- Engenharia de Software
- Questões Sociais e Profissionais
- Ciência computacional e métodos numéricos

A Computação...

A Ciência da Computação não lida apenas com o uso do computador, tecnologia ou software – é uma **ciência** que engloba:

Pensamento matemático: encontrar **soluções para problemas**, ou provar que as soluções não existem

Engenharia: exige **habilidades para projetar sistemas de software complexos**

Métodos Científicos em Computação

Ciência da Computação Teórica

- Aderente à tradição da lógica e matemática
- Metodologia clássica de construção de teorias como sistemas lógicos com axiomas e regras
- Elementos chave:
 - Modelos formais e conceituais
 - Níveis de abstração
 - Eficiência
- Temas:
 - Iteração, Recursão e Indução
- Modelos de dados:
 - Modelos de dados em árvores
 - Modelos de dados em listas
 - Modelos de dados em conjuntos
 - Modelos de dados relacionais
 - Modelos de dados em grafos
 - Expressões regulares e padrões
- Resumindo: busca entender os limites de computação e o poder dos paradigmas computacionais. Os teóricos também desenvolvem abordagens gerais para solução de problemas

Métodos Científicos em Computação

Ciência da Computação Experimental

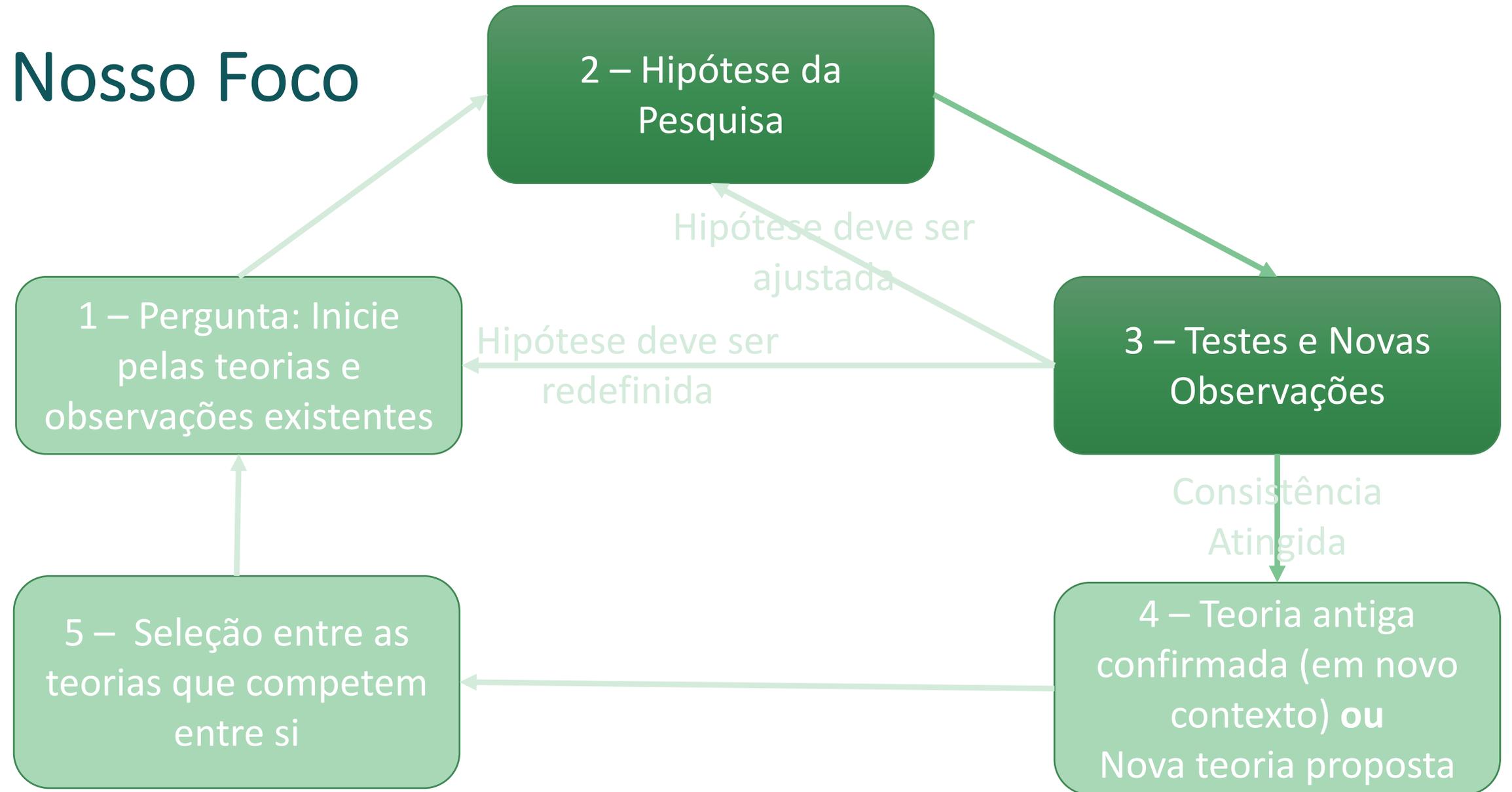
- Mais eficaz em problemas que exigem soluções de software complexas
 - Criação de software em ambientes de desenvolvimento
 - Organização de dados não tabulares
 - Construção de ferramentas para otimização específica a problemas
- A abordagem é em grande parte para identificar conceitos que facilitam soluções para um problema e, em seguida, avaliar as soluções por meio da construção de protótipos
- Os experimentos são tipicamente realizados em diferentes campos:
 - Prova automática de teoremas
 - Planejamento
 - Problemas NP-completos
 - Linguagem natural
 - Visão computacional
 - Jogos
 - Redes Neurais / Conexionismo
 - Aprendizado de Máquina

Métodos Científicos em Computação

Simulação Computacional

- Possibilita investigar regimes que estão além das capacidades experimentais atuais
- Estuda fenômenos que não podem ser replicados em laboratórios, como a evolução do universo
- Na ciência, simulações de computador são guiados por Teoria + Resultados experimentais
 - Resultados podem sugerir novas experiências e modelos teóricos
- **Caos e Sistemas Complexos:** Interesse em observar a complexidade em um modelo determinístico estruturalmente simples
 - Sistemas podem ser lineares ou não lineares
- **Realidade Virtual:** Interesse em imergir o analista no mundo simulado
 - Deve incorporar métodos para construir mundos digitais dinâmicos (virtuais) – típico em simulação de computador
- **Vida Artificial:** Desafia nossa definição do termo experimento
 - Um experimento em vida artificial: um programa de computador é escrito para simular formas de vida artificiais
 - Pode carregar metáforas como reprodução genética e mutação
- **Modelagem Baseada em Física e Animação por Computador:** Modelagem baseada na física – modelos baseados em restrições derivada das leis físicas

Nosso Foco



Introdução à Estatística

Introdução à estatística

- Um conjunto de princípios e métodos para auxiliar a tomada de decisão
- Princípios e métodos para:
 - Planejar coleta de dados
 - Sumarizar e interpretar os dados
- Qualquer área da Ciência e Tecnologia faz uso da estatística, principalmente para avaliação empírica de hipóteses

Áreas da Estatística

- Estatística Descritiva
 - Conjunto de técnicas para sumarizar os dados em tabelas, gráficos e medidas descritivas
 - Auxílio para extrair informações contidas nos dados
- Inferência Estatística
 - Conjunto de argumentos para permitir fazer afirmações sobre as características de uma população com base em informações dadas por amostras

O termo “científico”...

- Relacionado a uma investigação objetiva que assegura conclusões válidas a partir de um estudo experimental
- Experimentos devem ser planejados
- Conclusões são tiradas a partir dos dados experimentais usando a teoria estatística

O que é Variável Aleatória (V.A.)?

- Uma Variável Aleatória é uma variável quantitativa, cujo resultado (valor) depende de fatores aleatórios
- Definição formal:
 - V.A. é uma função que associa a todo evento pertencente a uma partição do espaço amostral Ω um único número real

$$X: \Omega \rightarrow \mathbb{R}$$

- Ex: V.A. Altura – mapeia cada pessoa $\omega \in \Omega$ a sua altura $X(\omega)$

Tipos de Dados

- Dados Qualitativos ou Categóricos: Valores são rótulos para as categorias / não possuem relação de ordem – Ex:
 - Homem / Mulher
 - Tipo de Sangue: O / A / B / AB
 - Solteiro / Casado / Viúvo / Divorciado ...
- Dados Numéricos ou de Medidas (possuem relação de ordem)
 - Discretos (valores inteiros) – Ex: Número de Filhos em uma família; Número de Rodas em um veículo; etc
 - Contínuos – Ex: Peso; Altura; Salário; etc

Estatística Descritiva

- Ser humano tem dificuldade para interpretar tabelas com dados brutos
- Ver planilha [dados brutos.xlsx](#)

Dados Categóricos

- Para construir tabelas e gráficos – cálculo das frequências absoluta e relativa das categorias individuais

$$FA = FreqCat$$

$$FR = \frac{FreqCat}{TotalObs}$$

- FreqCat: Frequência na Categoria – número de vezes que o valor aparece na amostra
- TotalObs: Número Total de Observações

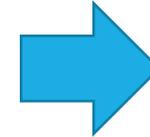
Exemplo

- Considerando a tabela [dados brutos.xlsx](#) – variável Estado Civil
 - Calcular a FA e FR para cada categoria
 - Montar a tabela com o resumo dos dados
 - Fazer um diagrama circular para os dados

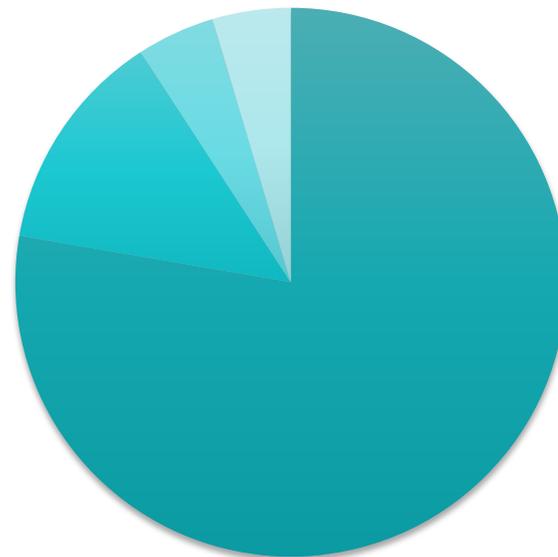
Exemplo

- Variável Estado Civil:

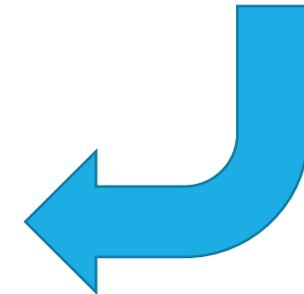
Estado Civil	FA	FR
Solteiro	17	0,77
Casado	3	0,14
Divorciado	1	0,05
Viúvo	1	0,05
Total	22	1,01



Estado Civil	FA	FR
Solteiro	17	0,77
Casado	3	0,13
Divorciado	1	0,05
Viúvo	1	0,05
Total	22	1,00



- Solteiro
- Casado
- Divorciado
- Viúvo



Exempnpo

- Outras variáveis categóricas:
 - Transporte
 - Procedência
 - Relação do trabalho com pesquisa em computação
 - Meio de Informação

Dados Discretos

- Para construir tabelas e gráficos – cálculo das frequências absoluta e relativa das categorias individuais

$$FA = FreqValorX$$

$$FR = \frac{FreqValorX}{TotalObs}$$

- FreqValorX: Frequência de um Valor x
- TotalObs: Número Total de Observações

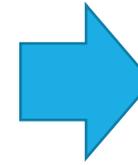
Exemplo

- Considerando a tabela [dados brutos.xlsx](#) – variável Núm. Irmãos
 - Calcular a FA e FR para cada categoria
 - Montar a tabela com o resumo dos dados calculados
 - Fazer um gráfico de barras verticais para os dados

Exemplo

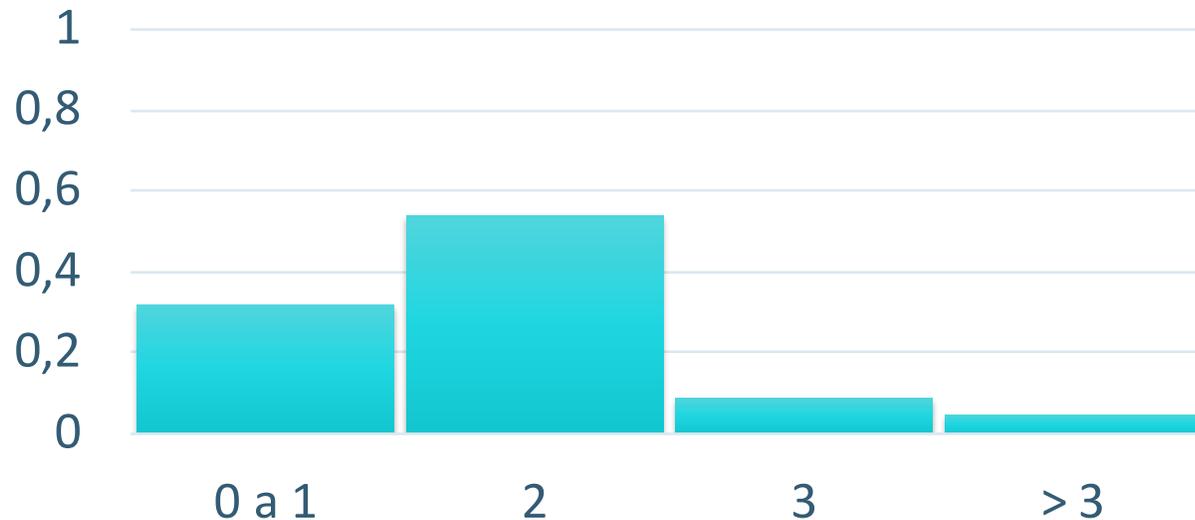
- Variável
Núm. Irmãos:

Núm. Irmãos	FA	FR
0	1	0,05
1	6	0,27
2	12	0,55
3	2	0,09
6	1	0,05
Total	22	1,01

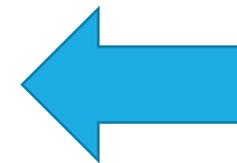


Núm. Irmãos	FA	FR
0	1	0,05
1	6	0,27
2	12	0,54
3	2	0,09
6	1	0,05
Total	22	1,00

Núm. Irmãos



OU



Núm. Irmãos	FA	FR
0 a 1	7	0,32
2	12	0,54
3	2	0,09
> 3	1	0,05
Total	22	1,00

Dados Contínuos

- Duas possibilidades de gráficos:
 - Diagramas de Pontos
 - Indicado para pequeno número de observações (valores de uma variável)
 - Histogramas
 - Indicado para grande número de observações

Distribuição de Frequência de Variáveis Contínuas

- Considerar intervalos fixos de mesmo comprimento para determinar as frequências relativas de cada intervalo – classe
- Passos:
 1. Achar máximo e mínimo da variável
 2. Escolher número de intervalos (10 é usual)
 3. Escolher intervalos de mesmo comprimento que cubra a amplitude entre mínimo e máximo – cada intervalo é denominado **classe**
 4. Contar número de observações da variável que pertence a cada intervalo
 5. Calcular as FRs de cada classe:

$$FR_C = \frac{FreqObsClasse}{TotalObs}$$

Histogramas

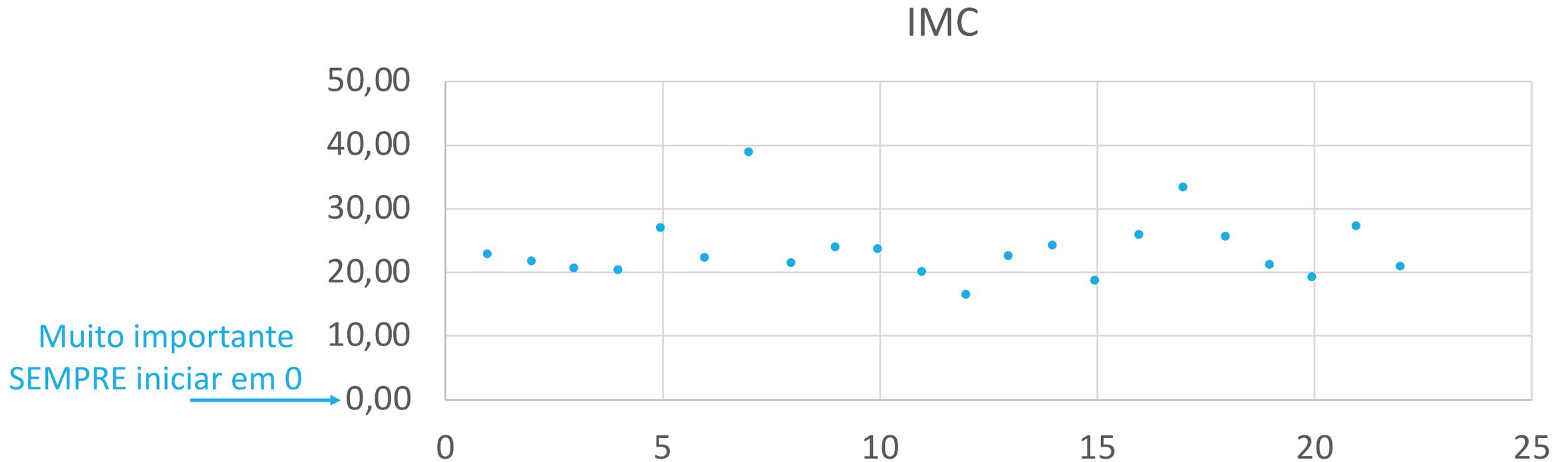
- São representações gráficas das distribuições de frequências dadas por retângulos
- Cada classe é representada por um retângulo
 - Base (largura) do retângulo: Intervalo da classe
 - Altura do retângulo:

$$A_C = \frac{FR_C}{L_C}$$

- A área total de um histograma é igual a 1 (UM)

Exemplo

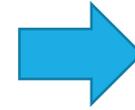
- Calcular FRs da variável IMC
- Plotar diagrama de pontos e histograma



Exemplo

- Variável IMC:

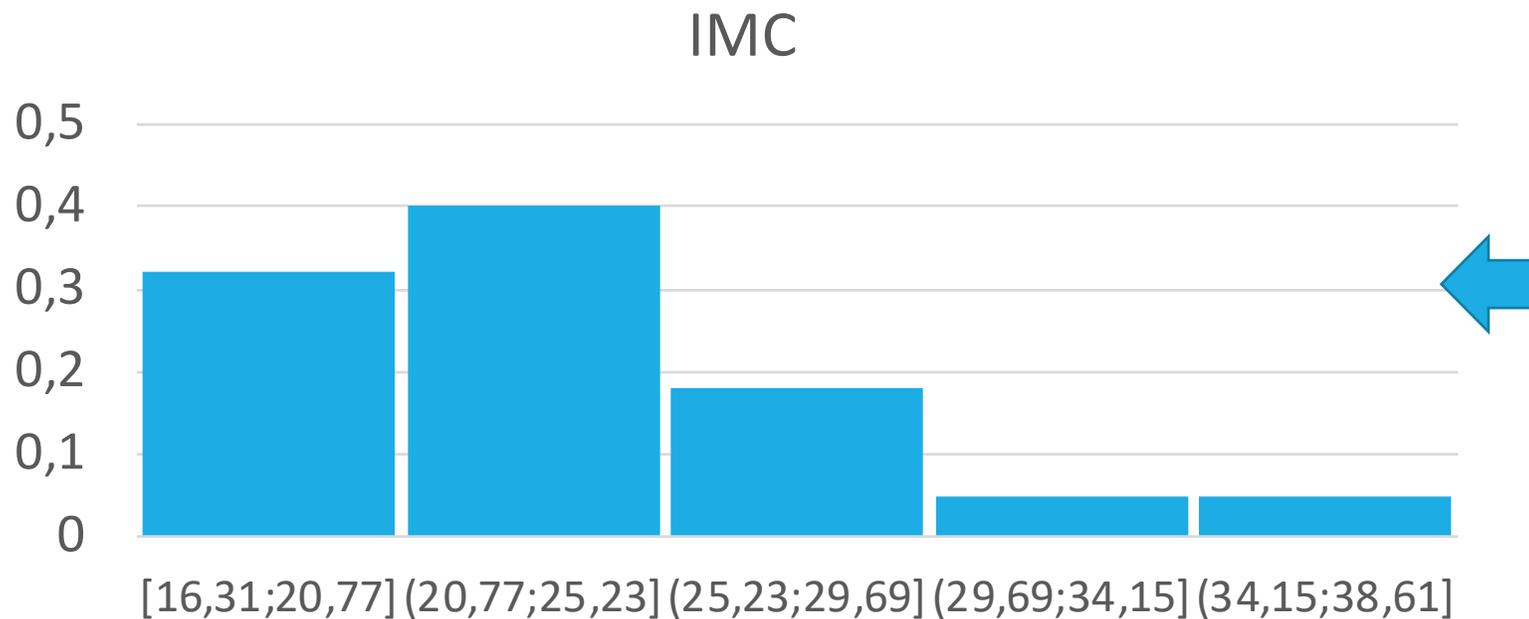
Caract. Histograma	
# Intervalos	5
Mínimo	16,31
Máximo	38,60
Tamanho de intervalo	4,458



Intervalos	FA	FR
[16,31;20,77]	7	0,32
(20,77;25,23]	9	0,41
(25,23;29,69]	4	0,18
(29,69;34,15]	1	0,05
(34,15;38,61]	1	0,05
Total	22	1,01



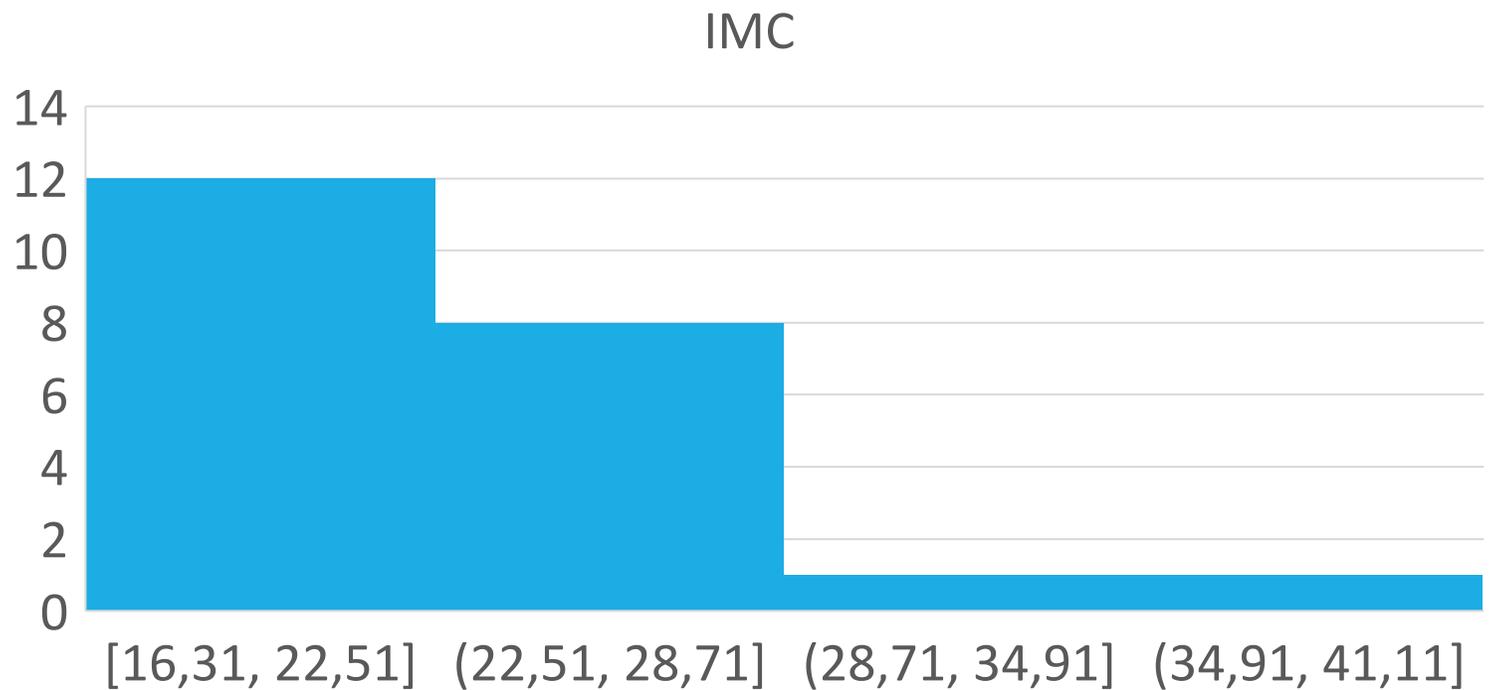
Intervalos	FA	FR
[16,31;20,77]	7	0,32
(20,77;25,23]	9	0,40
(25,23;29,69]	4	0,18
(29,69;34,15]	1	0,05
(34,15;38,61]	1	0,05
Total	22	1,00



Exemplo

- Variável
IMC:

Histograma construído
pelo Excel:



Medidas Descritivas

- Além dos gráficos, também podem ser utilizadas medidas descritivas para sumarizar as informações contidas nos dados em termos de:
 - Centralidade
 - Variabilidade
 - Formas das distribuições das frequências (estudadas mais adiante no curso)

Medidas de Centralidade (1)

- Média amostral

- Medida de centralidade dada pela média aritmética das observações

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- Mediana amostral

- É o valor de centralidade quando as observações são ordenadas
- Se n é par, calcula-se a média dos valores centrais

Medidas de Centralidade (2)

- Ex. 1: Produção (em toneladas/hectare) de milho em uma fazenda experimental:
 - 6,4; 10,5; 8,1; 9,2; 7,8
 - Média – 8,4 / Mediana – 8,1
- Ex. 2: Tempo de sobrevivência de 6 cobaias submetidas a um experimento médico:
 - 3; 15; 46; 64; 126; 623
 - Média – 146 / Mediana – 55

Medidas de Variabilidade (1)

- Além das medidas de centralidade, é importante conhecer o espalhamento ou **variabilidade** dos dados
- Ex: Um estudo de pessoas afetadas por certa doença revela que a maioria de pessoas afetadas são menores de 2 anos ou maiores de 70 anos
 - Sumarizando pela média, poderíamos obter que a média de idade das pessoas afetadas é de 30 anos – INADEQUADO
 - Necessidade de uma medida de variabilidade dos dados

Medidas de Variabilidade (2)

- Considerando um conjunto n de observações x_1, \dots, x_n , \bar{x} é a média amostral do conjunto e:

- Desvio:

- Obs:

$$DESvio_i = x_i - \bar{x}$$

$$\sum_{i=1}^n (DESvio_i) = \sum_{i=1}^n (x_i - \bar{x}) = 0$$

Medidas de Variabilidade (3)

- Variância Amostral S^2 :

$$S^2 = \frac{\sum (DESvio_i)^2}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

- Desvio Padrão Amostral S é a raiz quadrada da variância amostral:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Medidas de Variabilidade (4)

- Considerando os valores observados da variável x , na primeira coluna da tabela abaixo, complete a tabela:

x	$x - \bar{x}$	$(x - \bar{x})^2$
3	-3	9
5	-1	1
7	1	1
7	1	1
8	2	4
$\sum x = 6$	$\sum (x - \bar{x}) = 0$	$\sum (x - \bar{x})^2 = 16$

- Calcule a variância e o desvio padrão de x

$$\begin{aligned}\text{var}(x) &= 16 / (n-1) = 4 \\ \text{dp}(x) &= \text{raiz}(\text{var}(x)) = 2\end{aligned}$$

Medidas de Variabilidade (5)

- Observações:
 - Desvio padrão amostral é uma medida de variabilidade medida na mesma escala das observações consideradas
 - Forma alternativa de cálculo da variância amostral:

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 / n}{n - 1}$$

Sem necessidade de cálculo prévio da média

Dados Bivariados

- Suponha duas variáveis x e y
- Muitas vezes é interessante identificar possíveis relações existentes entre as duas variáveis
- Exs. de possíveis relações:
 - Fumante e câncer de pulmão
 - Quantidade de fertilizante e produção por hectare
- Possibilidade de 3 combinações de tipos de variáveis:
 - Duas variáveis categóricas
 - Duas variáveis numéricas
 - Uma variável de cada tipo

Dados Bivariados Categóricos (1)

- Construção de tabelas de frequência com dupla entrada
- Também conhecidas como tabela de contingência

Dados Bivariados Categóricos (2)

- Exemplo: Pesquisa de opinião entre 400 operários de uma indústria metalúrgica
 - Cada operário foi consultado sobre sua opinião quanto a uma greve local (Sim, Indiferente ou Não)
 - Também, foi consultado se pertence ou não ao sindicato local
 - Duas variáveis:
 - x – opinião do operário
 - y – situação sindical do operário

Dados Bivariados Categóricos (3)

	SIM	INDIFERENTE	NÃO	TOTAL
SIM	112	36	28	176
NÃO	84	68	72	224
TOTAL	196	104	100	400

Frequências Relativas de cada combinação de categoria:

	SIM	INDIFERENTE	NÃO	TOTAL
SIM	$112/400 = 0,28$	$36/400 = 0,09$	$28/400 = 0,07$	0,44
NÃO	$84/400 = 0,21$	$68/400 = 0,17$	$72/400 = 0,18$	0,56
TOTAL	0,49	0,26	0,25	1,00

Dados Bivariados Categóricos (4)

- Objetivo: comparar os dois grupos de pessoas (176 sindicalizadas e 224 não sindicalizadas) para verificar se as proporções para cada grupo são iguais
 - Cálculo de frequências em relação aos totais marginais (totais em cada classe)

	SIM	INDIFERENTE	NÃO	TOTAL
SIM	$112/176 = 0,636$	$36/176 = 0,205$	$28/176 = 0,159$	1,00
NÃO	$84/224 = 0,375$	$68/224 = 0,304$	$72/224 = 0,321$	1,00

- Conclusão: proporções são diferentes e, portanto, a situação sindical impacta na opinião quanto à greve

Dados Bivariados Contínuos

- Suponha duas variáveis contínuas
- Perguntas:
 - As variáveis são relacionadas?
 - Qual a forma desse relacionamento?
 - Como medimos esta relação?
 - Como prever o valor de uma variável a partir da outra?

Dados Bivariados Contínuos

- Coeficiente de Correlação Linear Amostral – Correlação de Pearson
 - Mede a intensidade da relação linear entre as variáveis quantitativas x e y com n observações em uma amostra
 - r = coeficiente de correlação amostral (estimativa para o verdadeiro valor r)

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}$$

- Problema: assume distribuição normal das variáveis x e y

Dados Bivariados Contínuos

- Coeficiente de Correlação Linear Amostral – Correlação de Postos de Spearman
 - Métrica não paramétrica
 - Pode ser utilizado para:
 - Quando não é conhecida a distribuição da V.A.
 - Quando há variáveis discretas

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

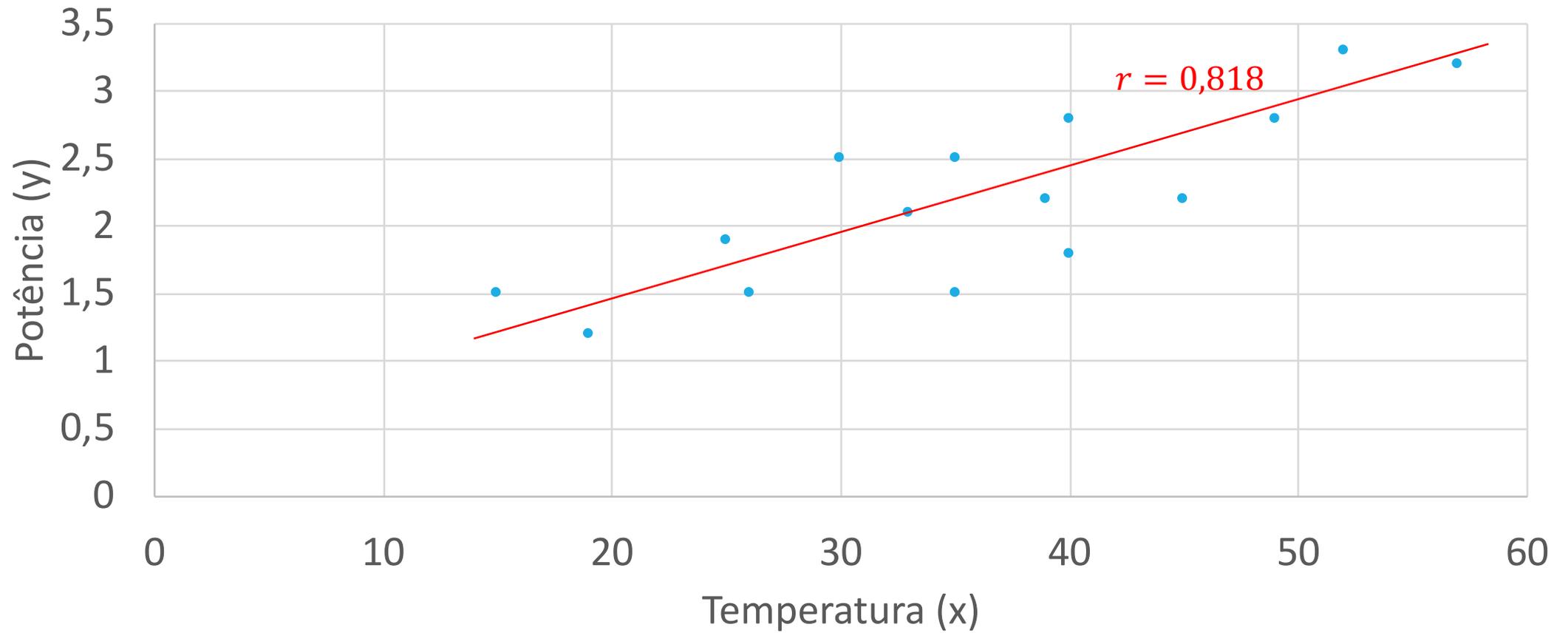
Exemplo

- Temperatura e Potência de motores medidas:

x	y	x	y	x	y
19	1,2	33	2,1	45	2,2
15	1,5	30	2,5	39	2,2
35	1,5	57	3,2	25	1,9
52	3,3	49	2,8	40	1,8
35	2,5	26	1,5	40	2,8

- Fazer um gráfico para tentar identificar relação
- Calcular correlações de Pearson e de Spearman

Exemplo



Exemplo

- Correlação de Spearman

Temp	Potência	Posto_T	Posto_P	d_i	d_i^2
15	1,5	1	3,5	-2,5	6,25
19	1,2	2	1	1	1
25	1,9	3	6	-3	9
26	1,5	4	3,5	0,5	0,25
30	2,5	5	10	-5	25
33	2,1	6	7	-1	1
35	1,5	7	3,5	3,5	12,25
35	2,5	8	11	-3	9
39	2,2	9	8,5	0,5	0,25
40	1,8	10	5	5	25
40	2,8	11	12,5	-1,5	2,25
45	2,2	12	8,5	3,5	12,25
49	2,8	13	12,5	0,5	0,25
52	3,3	14	15	-1	1
57	3,2	15	14	1	1
					105,75
				ρ	0,810

Como avaliar a correlação

- Correlação varia entre -1 e 1
- Interpretação:
 - 0 – não há correlação
 - $0 < |r| < 0.7$ – correlação moderada
 - $|r| > 0.8$ – correlação forte

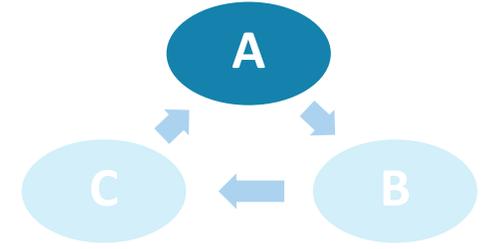
Planejamento de Experimentos

Planejamento de Experimentos

- Pode ser usado tanto no desenvolvimento do processo quanto na solução de problemas do processo
 - **Objetivo:** melhorar o desempenho ou obter um processo que seja robusto ou não-sensível a fontes externas de variabilidade
- Métodos de planejamento de experimentos podem ser úteis no estabelecimento do controle estatístico de um processo
 - Processo deve ter as várias variáveis de entrada controláveis
 - Métodos de planejamento experimental podem ser usados para identificar as variáveis influentes do processo
- Planejamento de experimentos é uma ferramenta de engenharia importante para melhorar um processo
 - Metodologia científica também é um processo

Planejamento de Experimentos

Procedimento Geral para Aplicação



1. Reconhecimento e relato do problema

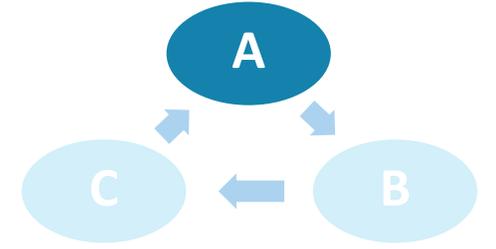
- Desenvolver todas as ideias sobre o problema e sobre os objetivos específicos do experimento
- Relato claro do problema e dos objetivos do experimento costuma contribuir substancialmente para uma melhor compreensão do que se espera avaliar

2. Definir objetivos do experimento

- A partir de uma boa definição do problema é mais natural a elaboração do objetivo do experimento
- Esse objetivo deve ser **não tendencioso, específico, mensurável e de resultado prático**

Planejamento de Experimentos

Procedimento Geral para Aplicação

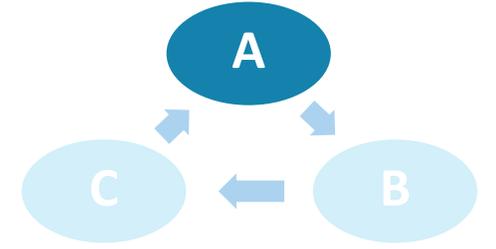


3. Escolha das variáveis resposta e controle

- Na seleção da variável resposta (**o que se quer medir**), o experimentador deve ter certeza de que aquela variável realmente fornece informação útil sobre o objeto em estudo
- Muitas vezes, a média ou o desvio padrão da característica medida será a variável de resposta (**Ex: taxa de erro de classificador**)
- Múltiplas variáveis não são raras
- A capacidade do medidor da variável é, também, um fator importante
 - Se capacidade do medidor baixa: apenas grandes efeitos serão detectados pelo experimento – quão difícil ou fácil é medir a variável?
- Como selecionar variáveis respostas e controle? Teoria e experimentos anteriores (revisão da literatura) ou de especialistas/experiência
- Onde esses experimentos se ajustam dentro do estudo?

Planejamento de Experimentos

Procedimento Geral para Aplicação

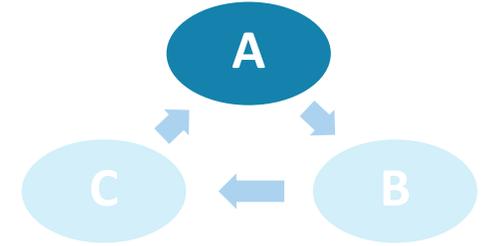


Importante:

- **Passos 2 e 3** são realizados simultaneamente, ou passo 3 pode ser feito antes
- **Passos 1 a 3:** Para o sucesso do experimento é vital que esses passos sejam realizados tão bem quanto possível.

Planejamento de Experimentos

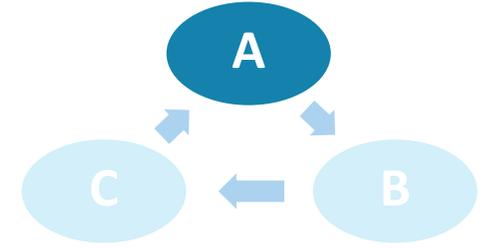
Procedimento Geral para Aplicação



4. **Listar para cada variável resposta** a precisão ou amplitude aos quais ela pode ser medida e como
5. **Listar para cada variável controle** a precisão ou amplitude a qual ela pode ser agrupada – são as **variáveis parametrizadas do experimento**, que você assume que não varia
 - Avaliar a finalidade da colocação da variável controle e o efeito de previsão que o cenário terá em cada variável resposta
6. **Escolha dos fatores e dos níveis** – quem conduz o experimento deve escolher:
 - Níveis específicos em que cada rodada de experimento será feita
 - Esse conhecimento é em geral, uma combinação de experiência prática e conhecimento teórico – definição do problema e revisão da literatura

Planejamento de Experimentos

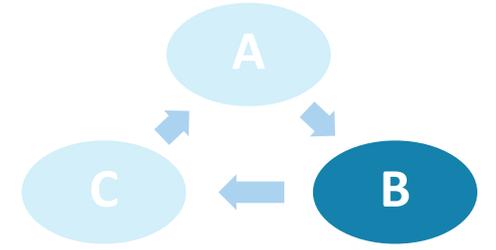
Procedimento Geral para Aplicação



7. **Listar e rotular** interações conhecidas ou supostas entre os parâmetros e as variáveis de resposta (o que se quer medir)
8. **Listar restrições** no experimento (**limitações do trabalho**):
 - facilidade de alterar a variável controle (parâmetros)
 - métodos de aquisição de dados
 - número de execuções
 - regiões experimentais irrelevantes ou não viável, limitação quando há aleatoriedade
 - custo de mudança no cenário da variável controle, etc
9. **Escolha do planejamento experimental**
 - Tamanho da amostra – pode implicar no número de replicações do experimento
 - Seleção de uma ordem adequada de execuções para as tentativas experimentais
 - Verificar restrições de aleatoriedade que podem estar envolvidas

Planejamento de Experimentos

Procedimento Geral para Aplicação

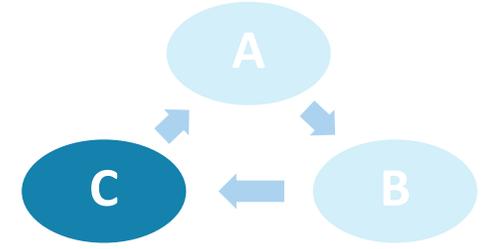


10. Realização do experimento

- Monitorar o processo, gerando **logs**
- Erros no procedimento experimental nesse estágio podem destruir a validade do experimento
- Planejamento geral, do início até o fim, é crucial para o sucesso – é fácil subestimar os aspectos logísticos e de planejamento em um ambiente complexo de experimentação

Planejamento de Experimentos

Procedimento Geral para Aplicação



11. Análise de dados

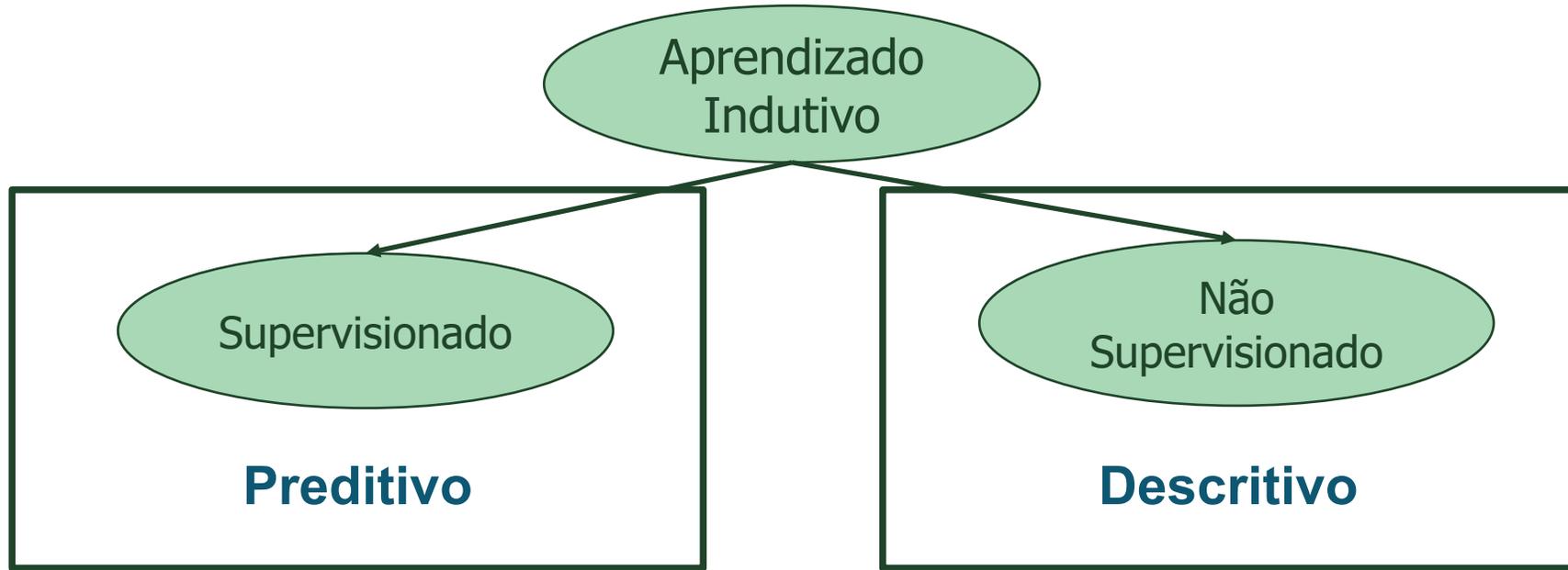
- Usar métodos estatísticos para analisar os dados – gráficos, ANOVA, regressão, testes de hipótese, etc
- Métodos gráficos simples têm papel importante na interpretação dos dados
- Resultados e conclusões devem ser objetivos
- Pacotes estatísticos estão disponíveis para ajudar na análise de dados

12. Conclusões

- Uma vez analisados os dados, o experimento deve levar a conclusões práticas sobre os resultados e recomendar uma ação (conclusões do trabalho)
- Sequências de acompanhamento e testes de confirmação podem ser realizados para validar as conclusões do experimento
- Apresentar as limitações do experimento – em geral, emergem os trabalhos futuros

Um Exemplo: Metodologia de Avaliação em AM e Métodos que envolvem Heurísticas/Metaheurísticas

Hierarquia de aprendizado



Importante: divisão não é rígida (modelo preditivo também provê descrição dos dados e modelo descritivo pode prover previsões após validado)

Aprendizado de Máquina Supervisionado

- Cada exemplo de treinamento é rotulado por um especialista do domínio / por sistemas de rotulação automática
- Rótulo pode ser pertencente a um conjunto de valores discretos ou contínuos

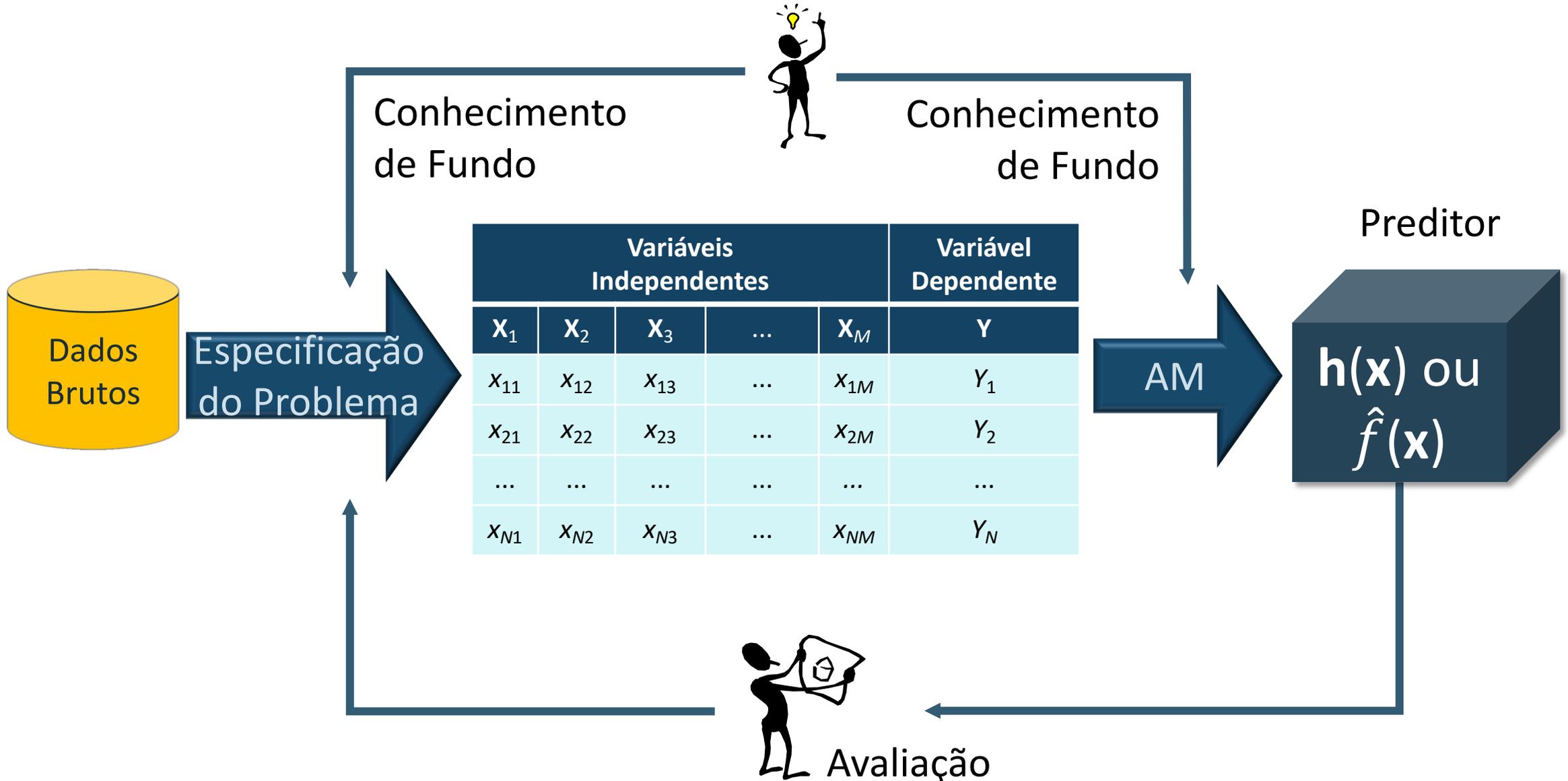
Conjunto de dados

- Hospital

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Meta: induzir modelo para fazer diagnósticos corretos para novos pacientes

Modelos Preditivos



Alguns algoritmos de AM

- Árvores de Decisão
- Redes Neurais
 - Redes Profundas
- SVMs
- Naive Bayes
- K-NN
- ...

Qual o melhor?
Qual o melhor conjunto de parâmetros?

Avaliação de Modelos Preditivos

- Não existe técnica de AM universal, que se saia melhor em qualquer tipo de problema
 - Teorema do No Free Lunch
 - Implica na necessidade de experimentos
 - O mesmo vale para áreas que envolvam algoritmos de otimização
- Características do problema e das técnicas pode auxiliar em alguns casos – ainda assim, diversos algoritmos e seus parâmetros podem ser candidatos

Avaliação de Modelos Preditivos

- Mesmo que um único algoritmo seja escolhido
 - Variações de parâmetros produzem diferentes modelos
- Domínio de AM: necessidade de experimentação
 - Experimentos controlados
 - Procedimentos que garantem a corretude e reproducibilidade dos experimentos

Avaliação de Modelos Preditivos

- Diferentes aspectos podem ser considerados:
 - Acurácia do modelo nas previsões
 - Compreensibilidade do conhecimento extraído
 - Tempo de aprendizado
 - Requisitos de armazenamento
 - Etc.

Concentraremos discussões a medidas de **desempenho preditivo**

Métricas de Erro

- Desempenho na rotulação de objetos
 - Métricas para classificação:
 - Taxa de erro
 - Acurácia
 - Métricas para regressão:
 - Erro quadrático médio
 - Distância absoluta média

Métricas para classificação

- Taxa de erro de um classificador f
 - De classificações incorretas

$$\text{err}(f) = (1/n) \sum_{i=1 \dots n} I(y_i \neq f(\mathbf{x}_i))$$

- Proporção de exemplos classificados incorretamente em um conjunto com n objetos
 - Comparação da classe conhecida com a predita
 - I é função identidade
 - = 1 se argumento é verdadeiro e 0 em caso contrário
 - Varia entre 0 e 1 e valores próximos de 0 são melhores

Métricas para classificação

- Taxa de acerto ou acurácia de um classificador f
 - Complemento da taxa de erro

$$ac(f) = 1 - err(f) = (1/n) \sum_{i=1 \dots n} I(y_i = f(\mathbf{x}_i))$$

- Proporção de exemplos classificados corretamente em um conjunto com n objetos
 - Varia entre 0 e 1 e valores próximos de 1 são melhores

Métricas para regressão

- Erro pode ser calculado pela distância entre o valor conhecido e o valor predito pelo modelo

Erro quadrático médio (*Mean Squared Error – MSE*)

$$\text{MSE}(f) = (1/n) \sum (y_i - f(\mathbf{x}_i))^2$$

Distância absoluta média (*Mean Absolute Distance MDA*)

$$\text{MDA}(f) = (1/n) \sum |y_i - f(\mathbf{x}_i)|$$

MSE e MAD são sempre não negativos;
valores mais baixos correspondem a melhores modelos

Métricas para regressão

- MSE normalizado
 - Para comparação de modelos/conjuntos de dados em diferentes escalas
 - Várias maneiras de normalizar
 - Exemplo:

$$\text{NMSE}(f) = \frac{\sum (y_i - f(\mathbf{x}_i))^2}{\sum (y_i - \bar{y})^2}$$

Amostragem

- Tem-se usualmente um único conjunto de n objetos
 - Deve ser usado para induzir e avaliar o preditor
 - Desempenho no conjunto de treinamento é otimista
 - Todos algoritmos tentam de alguma forma melhorar seu desempenho no conjunto de treinamento na fase indutiva
 - Avaliar modelo no conjunto de treinamento é conhecido como resubstituição
 - Produz taxa de erro/acerto aparente

Amostragem

- Métodos de amostragem: obter estimativas de desempenho mais confiáveis
 - Definindo subconjuntos disjuntos de:

Treinamento

- Dados empregados na **indução** e no **ajuste** do modelo
- Qualquer ajuste de parâmetros deve ser feito **nos dados de treinamento**

Teste

- Simulam a apresentação de **novos exemplos** ao preditor (não vistos em sua indução)
- **Somente avaliar** o modelo obtido

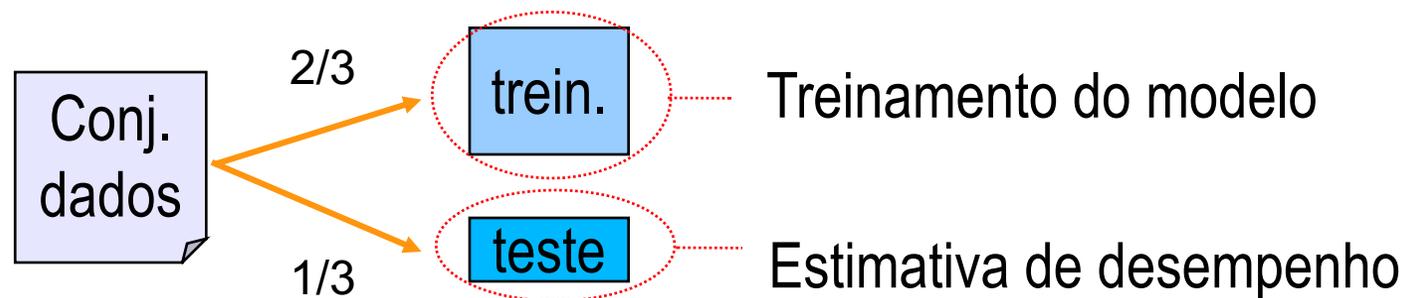
Em algumas situações, dados de treinamento são sub-divididos, gerando conjunto de **validação** dedicado ao ajuste de parâmetros

Amostragem

- Principais métodos de amostragem:
 - Holdout
 - Validação cruzada
 - Leave-one-out
 - Bootstrap

Holdout

- Método mais simples:
 - Divide conjunto de dados em proporção p para treinamento e $(1-p)$ para teste
 - Uma única partição
 - Valores típicos de p : $\frac{1}{2}$, $\frac{2}{3}$ ou $\frac{3}{4}$



Holdout

- Exemplo:

Objeto	Atributo 1	Atributo 2	Atributo 3	Classe
1	855	5142	2708	Safra 95
2	854	23155	2716	Safra 95
3	885	16586	2670	Safra 95
4	877	16685	2677	Safra 95
5	839	5142	2708	Safra 95
6	854	5005	2685	Safra 95
7	885	19455	2708	Safra 95
8	839	5027	2708	Safra 95
9	877	16823	2677	Safra 95
10	892	19180	2716	Safra 95
11	24628	39437	381	Safra 96
12	43183	39277	328	Safra 96
13	27871	39712	389	Safra 96
14	42329	40307	328	Safra 96
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96
17	33677	40375	328	Safra 96
18	33539	40078	335	Safra 96
19	34150	40353	358	Safra 96
20	34485	40742	358	Safra 96

Holdout

- Exemplo:

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo 3	Classe
4	877	16685	2677	Safra 95
6	854	5005	2685	Safra 95
8	839	5027	2708	Safra 95
2	854	23155	2716	Safra 95
10	892	19180	2716	Safra 95
1	855	5142	2708	Safra 95
6	854	5005	2685	Safra 95
18	33539	40078	335	Safra 96
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
12	43183	39277	328	Safra 96
17	33677	40375	328	Safra 96
20	34485	40742	358	Safra 96
11	24628	39437	381	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo 3	Classe
3	885	16586	2670	Safra 95
5	839	5142	2708	Safra 95
9	877	16823	2677	Safra 95
13	27871	39712	389	Safra 96
14	42329	40307	328	Safra 96
16	39399	40322	335	Safra 96

Holdout

- Indicado para grande quantidade de dados
 - Se pequena quantidade de dados
 - Poucos exemplos são usados no treinamento
 - Modelo pode depender da composição dos conjuntos de treinamento e teste
 - Quanto menor conjunto de treinamento, maior a variância do modelo
 - Quanto menor conjunto de teste, menos confiável a acurácia estimada para ele
- Muito usado para definir subconjuntos de validação

Holdout

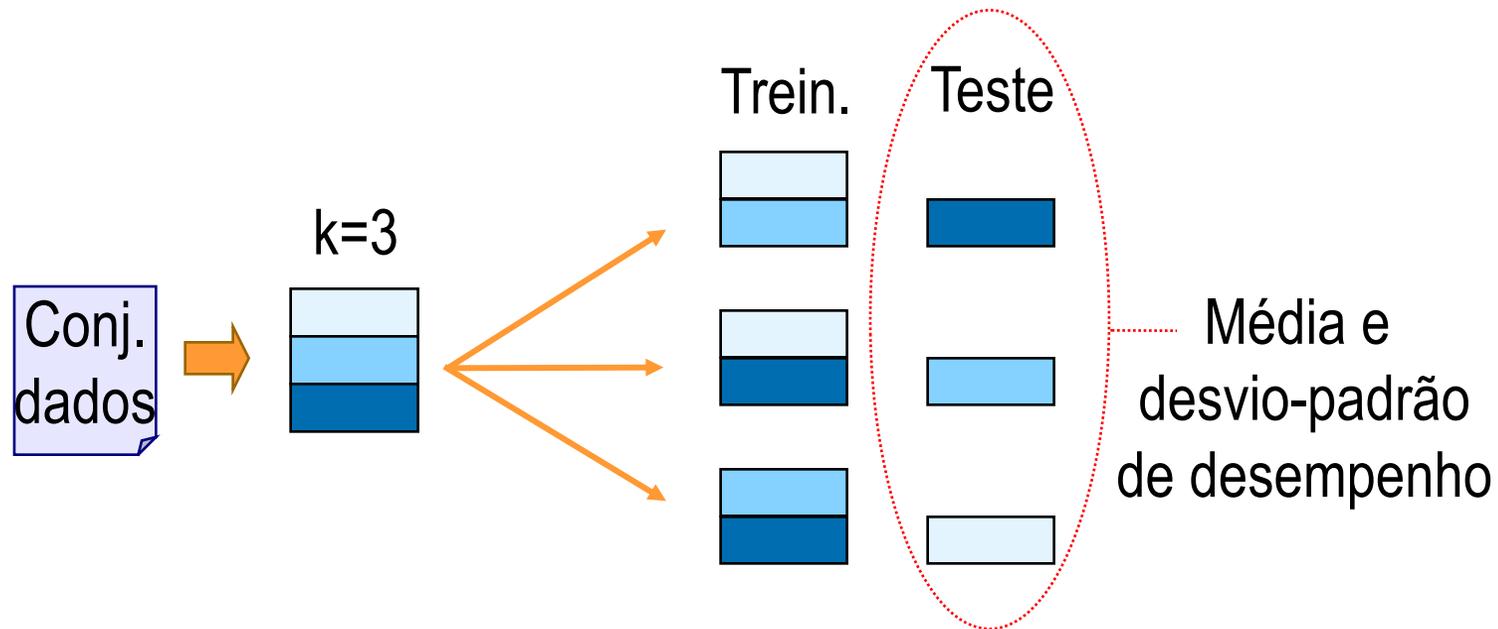
- Não avalia o quanto o desempenho de uma técnica varia
 - Quanto a diferentes combinações de exemplos de treinamento
 - É possível que uma divisão deixe no subconjunto de teste exemplos “mais fáceis”
 - Para tornar os resultados menos dependentes da partição feita: vários holdout
 - Random subsampling (amostragem aleatória)

Validação cruzada

- Método mais usado: k-fold cross validation
 - Conjunto é dividido em k partes de tamanho aproximadamente igual
 - Objetos de k-1 partes são usados no treinamento e a parte restante é usada para teste
 - Procedimento é repetido k vezes usando cada partição para teste
 - subconjuntos de teste são independentes entre si
 - Desempenho é dado por média
 - Valor típico de $k = 10$

Validação cruzada

- Ilustração para $k = 3$:



Validação cruzada

- Variação: k-fold cross validation estratificado
 - Manter a distribuição de classes em cada partição
 - Ex: se conjunto de dados original tem 20% na classe c1 e 80% na classe c2, cada partição também deve manter essa proporção
 - Distribuição de classes: proporção de exemplos em cada classe
 - Para cada classe c_j , $\text{dist}(c_j) = \text{número de exemplos que possuem a classe } c_j / \text{número total de exemplos}$

Distribuição de classes

- Ex.: conjunto de dados com 100 exemplos
 - 60 são da classe c1
 - 15 são da classe c2
 - 25 são da classe c3
 - A distribuição de classe é $\text{dist}(c1,c2,c3) = (0,60, 0,15, 0,25) = (60\%, 15\%, 25\%)$
 - A classe c1 é a classe majoritária ou prevalente
 - A classe c2 é a classe minoritária

Cross-validation estratificado

- Exemplo:
 - $r = 5$

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Cross-validation estratificado

- Ex.: Iteração 1

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96

Cross-validation estratificado

- Ex.: Iteração 2

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96

Cross-validation estratificado

- Ex.: Iteração 3

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96

Cross-validation estratificado

- Ex.: Iteração 4

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96

Cross-validation estratificado

- Ex.: Iteração 5

Conjunto de treinamento

Objeto	Atributo 1	Atributo 2	Atributo	Classe
4	877	16685	2677	Safra 95
9	877	16823	2677	Safra 95
18	33539	40078	335	Safra 96
11	24628	39437	381	Safra 96
1	855	5142	2708	Safra 95
3	885	16586	2670	Safra 95
14	42329	40307	328	Safra 96
20	34485	40742	358	Safra 96
7	885	19455	2708	Safra 95
10	892	19180	2716	Safra 95
15	41627	40032	335	Safra 96
19	34150	40353	358	Safra 96
6	854	5005	2685	Safra 95
2	854	23155	2716	Safra 95
17	33677	40375	328	Safra 96
12	43183	39277	328	Safra 96

Conjunto de teste

Objeto	Atributo 1	Atributo 2	Atributo	Classe
8	839	5027	2708	Safra 95
5	839	5142	2708	Safra 95
15	41627	40032	335	Safra 96
16	39399	40322	335	Safra 96

Leave-one-out

- Caso extremo de cross-validation com $k = n$
 - A cada ciclo exatamente um exemplo é separado para testes
 - Os $n-1$ restantes são usados no treinamento
 - Desempenho: soma dos desempenhos calculados para cada exemplo de teste individual
 - Produz estimativa mais fiel do desempenho preditivo
 - Mas é computacionalmente caro
 - Usado para conjuntos de dados pequenos

Leave-one-out

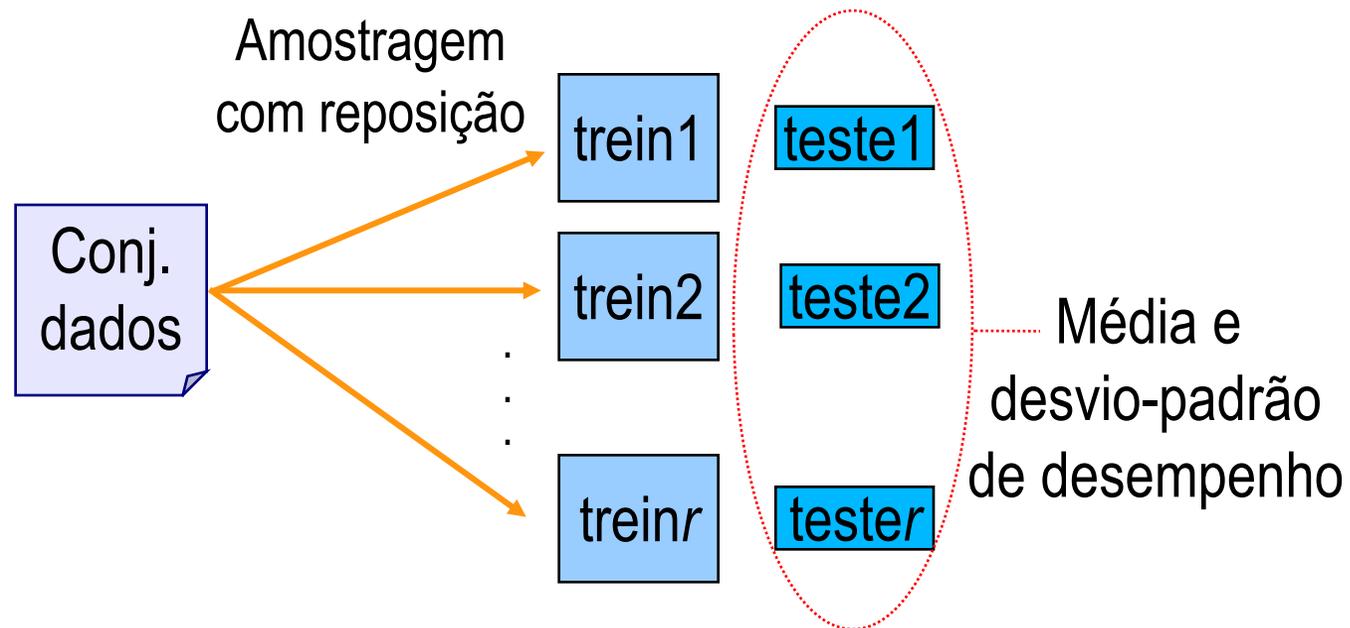
- Ex.: no caso do nosso exemplo a amostra tem 20 exemplos, então $k=20$
 - Ou seja, cada conjunto de treinamento será formado por 19 exemplos e o conjunto de teste por um único exemplo
 - Assim, o processo todo será repetido 20 vezes
 - E se o conjunto de dados tivesse 500 exemplos?

Bootstrap

- Baseado em amostragem com reposição
 - k subconjuntos de treinamento são amostrados, com reposição
 - Um exemplo pode estar presente mais de uma vez em um conjunto de treinamento
 - Exemplos não selecionados compõem conjuntos de teste
 - Desempenho: média dos desempenhos nos testes
 - Valor típico para $k = 100$ ou mais
 - É um procedimento custoso aplicado em conjuntos pequenos

Bootstrap

- Ilustração:



Bootstrap

- Há vários estimadores bootstrap, mais comum: e_0
 - Cada subconjunto de treinamento tem n exemplos
 - Cada exemplo tem probabilidade $1 - (1 - 1/n)^n$ de ser selecionado ao menos uma vez
 - Para n grande, tende a $1 - 1/e = 0,632$
 - Fração de exemplos não repetidos é de 63,2%
 - Exemplos remanescentes formam subconjunto de teste
 - Desempenho: média das iterações
 - Estimativa estiticamente equivalente a LOO, com menor variância

Amostragem

- Observações:
 - Para médias de desempenho, é importante reportar também os valores de desvio-padrão
 - Alto desvio padrão = alta variabilidade dos resultados
 - Indicativo de sensibilidade a variações nos dados de treinamento
 - Estimativas mais precisas também podem ser obtidas usando intervalos de confiança

Amostragem

- Ex. seja um dos métodos de amostragem
 - k-fold CV, por ser mais utilizado
 - Um indutor A gerará r hipóteses h_1, h_2, \dots, h_r
 - E cada hipótese terá uma medida de desempenho, na i-ésima partição
 - A média e o desvio-padrão do desempenho de A são então estimados por (considerando erro):

$$\text{média}(A) = 1/r \sum \text{erro}(h_i)$$

$$\text{desvPad}(A) = \sqrt{1/(r-1) \sum (\text{erro}(h_i) - \text{média}(A))^2}$$

$$\text{IC: } \text{média}(A) \pm t_{0,025,r-1} \frac{\text{desvPad}(A)}{\sqrt{r}}$$

Amostragem

- Exemplo:
 - Em 10-fold CV, algoritmo A obteve os erros:
 - (5,5; 11,4; 12,7; 5,2; 5,9; 11,3; 10,9; 11,2; 4,9; 11,0)
 - Temos então:

$$\text{média}(A) = 9,0$$

$$\text{desvPad}(A) = 3,17$$

$$\text{IC: } 9,0 \pm 2,26 * 3,17 / \sqrt{10} = 9,0 \pm 2,27 \\ (6,73; 11,27)$$

Testes de hipóteses

- Comparando desempenho de dois ou mais algoritmos
 - Há vários na literatura
 - Não há consenso sobre teste mais adequado
 - Embora se concorde que o ideal é comparar os desempenhos de diferentes algoritmos estatisticamente
 - Diferença entre eles é estatisticamente significativa? (a 95% de confiança)

Testes de hipóteses

- Comparação de algoritmos: todos devem ser testados em igualdade de condições
 - Mesmas partições dos dados
 - Desempenho medido nos mesmos objetos
 - Obtidas estimativas de desempenho – comparar com teste de hipóteses
 - Muitas vezes as diferenças não são significativas

Testes de hipóteses

- Hipótese estatística: alegação sobre o valor de um ou mais parâmetros
 - Ex. sejam m_1 e m_2 as médias de erro de dois algoritmos
 - Algumas hipóteses possíveis:
 - $m_1 = m_2$ (médias podem ser consideradas equivalentes)
 - $m_1 - m_2 > 0$ (média de 1 é superior à de 2)

Testes de hipóteses

- Suposições: normalmente há duas suposições contraditórias em avaliação
 - Ex. $H_0: m_1 = m_2$ vs $H_1: m_1 \neq m_2$
 - H_0 é chamada **hipótese nula**
 - É inicialmente assumida verdadeira
 - H_1 é chamada **hipótese alternativa**
 - Deve-se decidir qual das hipóteses é a correta
 - Aceitar ou rejeitar a hipótese nula

Hipótese nula é rejeitada em favor da hipótese alternativa se
Alguma evidência na amostra sugerir que H_0 é falsa

Testes de hipóteses

- Procedimento de teste: regra para decidir se H_0 é rejeitada
 - Procedimento de teste tem:
 - Estatística de teste: em função dos dados em que a decisão se baseia
 - Região de rejeição: conjunto de valores da estatística de teste para os quais H_0 é rejeitada
 - Hipótese nula é rejeitada se o valor da estatística de teste cai na região de rejeição

Testes de hipóteses

- Procedimento de teste podem cometer erros
 - Erro do tipo I: H_0 é rejeitada quando é verdadeira
 - Erro do tipo II: H_0 é falsa e não é rejeitada

Normalmente erros do tipo I são considerados mais sérios
⇒ maioria dos testes envolve controlar probabilidade de
Ocorrência de erro do tipo I, denotada por α
(calculam região de rejeição para manter α sob controle)

α é também chamado **nível de significância** do teste
Valor típico: 0,05 (resultado do teste possui nível de confiança
de 95% de não ter rejeitado hipótese nula quando ela é verdadeira)

Testes de hipóteses

- Testes para comparar modelos:
 - Em vários conjuntos de dados:
 - Dois algoritmos: Wilcoxon signed-rank
 - Vários algoritmos: Friedman
- Testes são pareados e não paramétricos
 - Wilcoxon e Friedman são baseados em ranqueamento
 - Permitem comparar qualquer medida de desempenho

Comparando dois modelos

- H0: desempenhos dos modelos são equivalentes
 - Ex. algoritmo padrão A vs novo algoritmo B
- Teste Wilcoxon signed-ranks:
 - Calcular diferenças das medidas de desempenho
 - Valores absolutos das diferenças são ranqueados
 - Ordem crescente
 - Em empates, atribui-se valores médios das posições
 - Comparam-se as posições das diferenças positivas e negativas entre os algoritmos

Comparando dois modelos

- *Teste Wilcoxon signed-ranks*:
 - Sejam dois algoritmos A e B
 - Tomando diferenças de B-A usando medida de desempenho em que maiores valores são melhores:
 - Diferenças **positivas**: melhor desempenho de B
 - Diferenças **negativas**: melhor desempenho de A

Conj. De Dados	Alg A	Alg B	Diferença	DifAbs	Posição
pulmão	0,583	0,583	0,000	0,000	1,5
fungo	0,583	0,583	0,000	0,000	1,5
atmosfera	0,882	0,888	0,006	0,006	3
mama	0,599	0,591	-0,008	0,008	4

Comparando dois modelos

- Teste Wilcoxon signed-ranks:
 - Sejam dois algoritmos A e B
 - R+: soma das posições em que B é melhor que A
 - R-: soma das posições em que B é pior que A
 - Diferenças nulas são repartidas igualmente

$$R+ = \sum_{d_i > 0} \text{pos}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{pos}(d_i)$$

$$R- = \sum_{d_i < 0} \text{pos}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{pos}(d_i)$$

- Seja S a menor dessas somas

Comparando dois modelos

- Teste Wilcoxon signed-ranks:
 - Sejam dois algoritmos A e B
 - Alguns livros apresentam tabelas com valores críticos exatos para S, até N = 25 conjuntos de dados
 - Para mais conjuntos de dados, estatística do teste é:

$$z = \frac{S - \frac{1}{4} N(N-1)}{\sqrt{(1/24) N(N+1)(2N+1)}}$$

Para $\alpha = 0,05$, a hipótese nula é rejeitada se $z < -1,96$

Comparando mais modelos

- Mais algoritmos sendo comparados
 - Múltiplos testes
 - Ex. $H_0: m_1 = m_2 = \dots = m_A$
 - Efeito da multiplicidade: probabilidade de um teste detectar diferença estatística quando ela não existe aumenta
 - Ajuste de nível de significância

Comparando mais modelos

- Teste de Friedman
 - Ranquea algoritmos pelo valor absoluto da medida de desempenho em cada conjunto de dados
 - Dos melhores para os piores
 - Empates: valores médios de posição são atribuídos
 - Seja r_{ij} a posição do desempenho do algoritmo j no conjunto de dados i
 - Compara ranqueamentos médios R_j dos algoritmos
 - H_0 : todos os algoritmos são equivalentes
 - Suas posições médias no ranqueamento são iguais

Comparando mais modelos

- Teste de Friedman
 - Estatística do teste:

$$F_F = \frac{(N-1) \chi^2}{N(A-1) - \chi^2}$$

$$\chi^2 = \frac{12N [R_j^2 - (A(A+1)^2)/4]}{A(A-1)}$$

A hipótese nula é rejeitada se $F_F > F_{A-1, (A-1)(N-1)}$
(distribuição F com A-1 e (A-1)(N-1) graus de liberdade)

Comparando mais modelos

- Teste de Friedman
 - Hipótese nula rejeitada – há diferenças entre os algoritmos
 - Mas não diz quais são
 - Necessário fazer um pós-teste
 - Pós-teste: par de algoritmos tem desempenho diferente caso a diferença entre os valores médios de posição deles seja maior ou igual a CD (critical difference)

$$CD = q_{\alpha} \sqrt{A(A+1)/N}$$

Comparando mais modelos

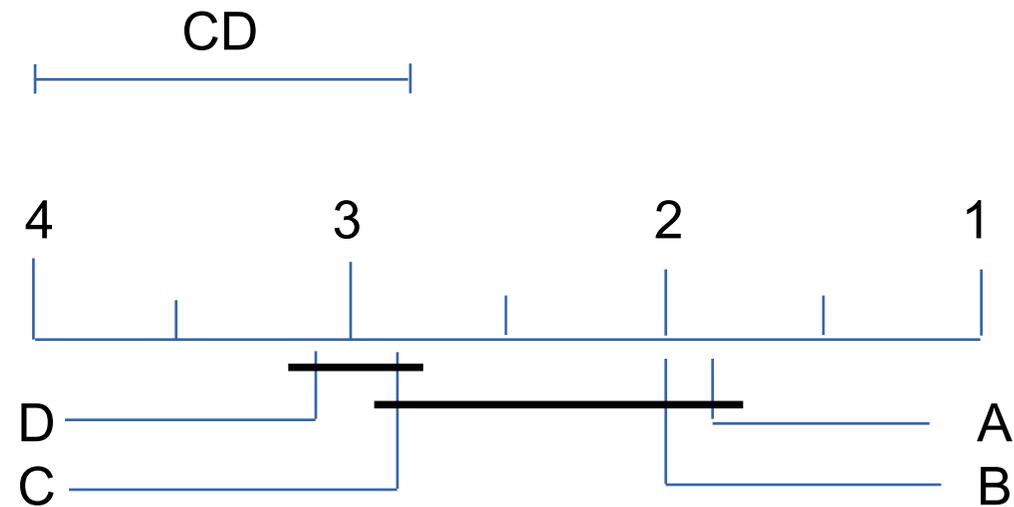
- Teste de Friedman: pós-teste
 - Comparando todos os algoritmos em pares: estatística de Nemenyi
 - Comparando vários a um único algoritmo: estatística de Bonferroni-Dunn

Valores de $q_{0,05}$

A	2	3	4	5	6	7	8	9	10
Nemenyi	1,96 0	2,34 3	2,56 9	2,728	2,850	2,949	3,031	3,102	3,164
Bonferroni-Dunn	1,96 0	2,24 1	2,39 4	2,498	2,576	2,648	2,690	2,724	2,773

Comparando mais modelos

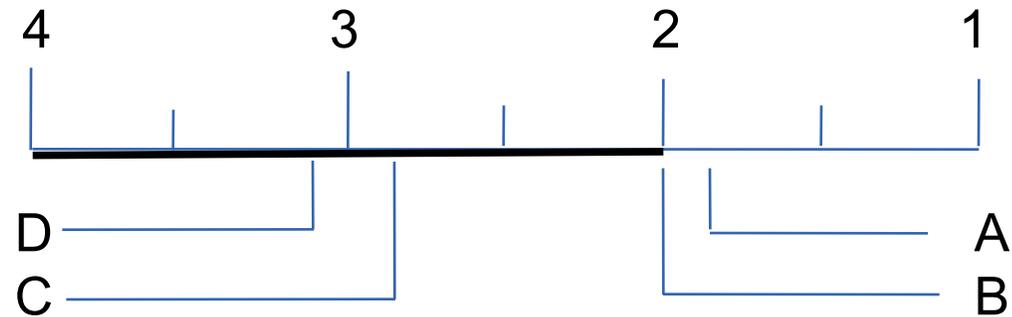
- Representando resultados dos pós-testes



Pós-teste Nemenyi comparando
quatro algoritmos A, B, C e D

Comparando mais modelos

- Representando resultados dos pós-testes



Pós-teste Bonferroni-Dunn comparando os algoritmos A, B e ao algoritmo D

Avaliação em AM – Considerações Finais

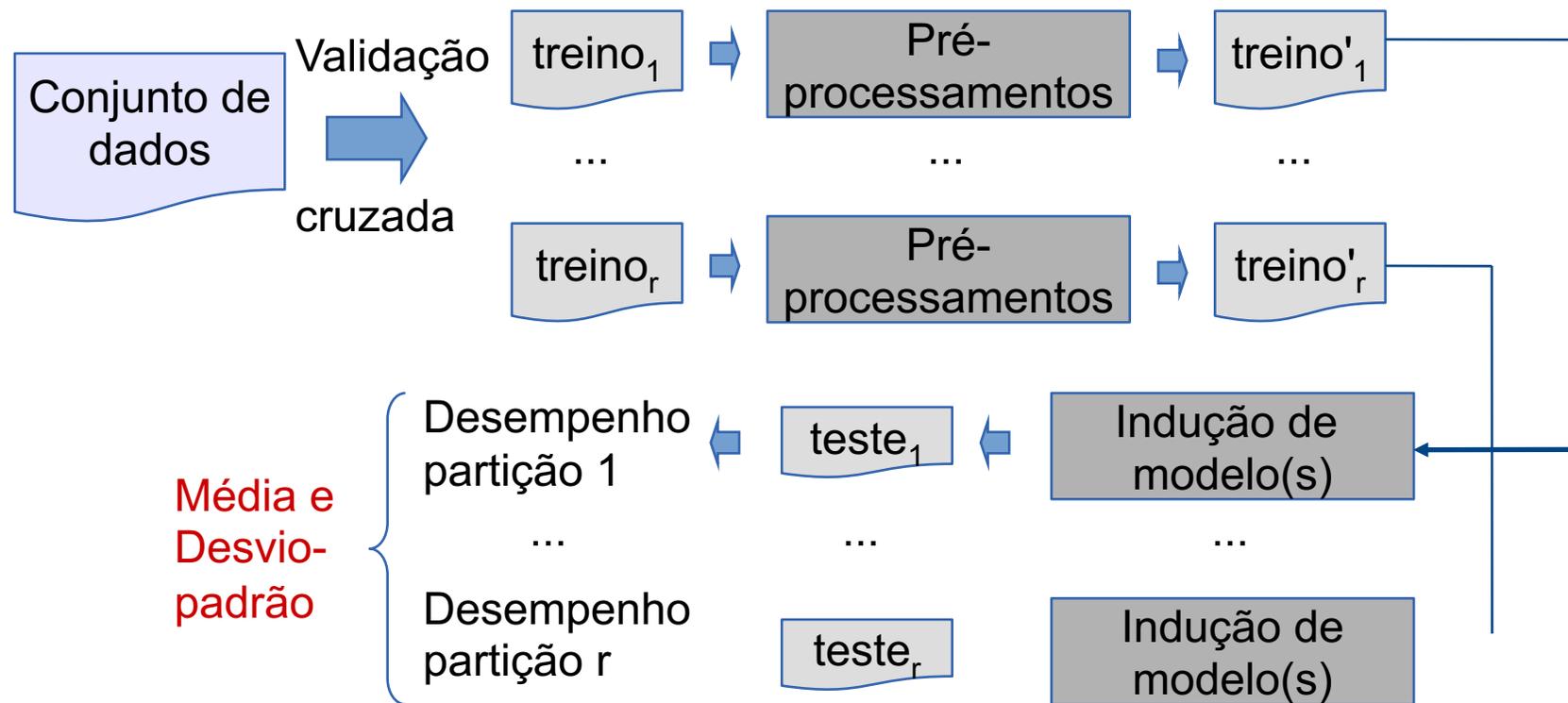
- Na prática, o mais usado é:
 - Amostragem do conjunto de dados:
 - 10 fold cross-validation
 - Estratificado para classificação
 - Leave-one-out para conjuntos de dados pequenos
 - Medidas de desempenho preditivo
 - Taxa de erro/acerto para classificação
 - Medicina: importante sensibilidade/especificidade, ROC
 - Recuperação de informação: medida F
 - MSE para regressão

Avaliação em AM – Considerações Finais

- Atenção:
 - Reportar média e desvio-padrão de desempenhos
 - Conjunto de teste deve ser somente para teste
 - Simular a chegada de dados totalmente novos ao modelo
 - Pré-processamentos, indução e ajustes de parâmetros devem ser feitos usando os dados de treinamento
 - Realizar testes estatísticos para comparar desempenhos de diferentes algoritmos
 - Diferenças podem não ser relevantes estatisticamente

Avaliação em AM – Considerações Finais

- Metodologia padrão:



Alguns exemplos

Transparency in practice: using visualization to enhance the interpretability of open data – d.go 2017

Raissa Barcellos, José Viterbo, Leandro Miranda,
Flavia Bernardini, Cristiano Maciel, Daniela Trevisan

- **Main question: Which type of data visualization is more suitable for each type of user and problem?**
- Preliminary results in understanding user interpretation on uni, bi and multivariate data visualization were presented at d.go 2017
- Dataset:
 - Socioeconomic indicators from Chicago city, available in their portal
 - Why this? Socioeconomic data is interesting for understanding how is situation of a city
- Research questions:
 - What types of visualization can be generated?
 - How they are interpreted?

Used Dataset

- The Socio-Chicago dataset
 - <http://data.cityofchicago.org>
- The portal offers a set of resources for data visualization, like columns, pizza, lines, area and tree map charts, beyond the tables
- Dataset: 7 features and 77 community areas

Table 1. Attributes from the Socio-Chicago Dataset.

Index	Feature Name
-	Community Area
-	Community Area Name
X_1	Birth Rate
X_2	Assault/Homicide
X_3	Below Poverty Level
X_4	Dependency
X_5	No High School Diploma
X_6	Per Capita Income
X_7	Unemployment

Public Health Statistics- Selected public health indicators by Chicago community area

View Data Download API Share ...

Health & Human Services

This dataset contains a selection of 27 indicators of public health significance by Chicago community area, with the most updated information available. The indicators are rates, percents, or other measures related to natality, mortality, infectious disease, lead poisoning, and economic status. See the full description at https://data.cityofchicago.org/api/assets/2107948F-357D-... More

Updated May 3, 2015 Data Provided by Illinois Department of Public Health (IDPH) and U.S. Census Bureau

About this Dataset

Updated **May 3, 2015**
Data Last Updated: May 30, 2013
Metadata Last Updated: May 3, 2015
Date Created: May 18, 2012

Views: **27.6K**
Downloads: **5,504**

Data Provided by: Illinois Department of Public Health (IDPH) and U.S. Census Bureau
Dataset Owner: Jamyia

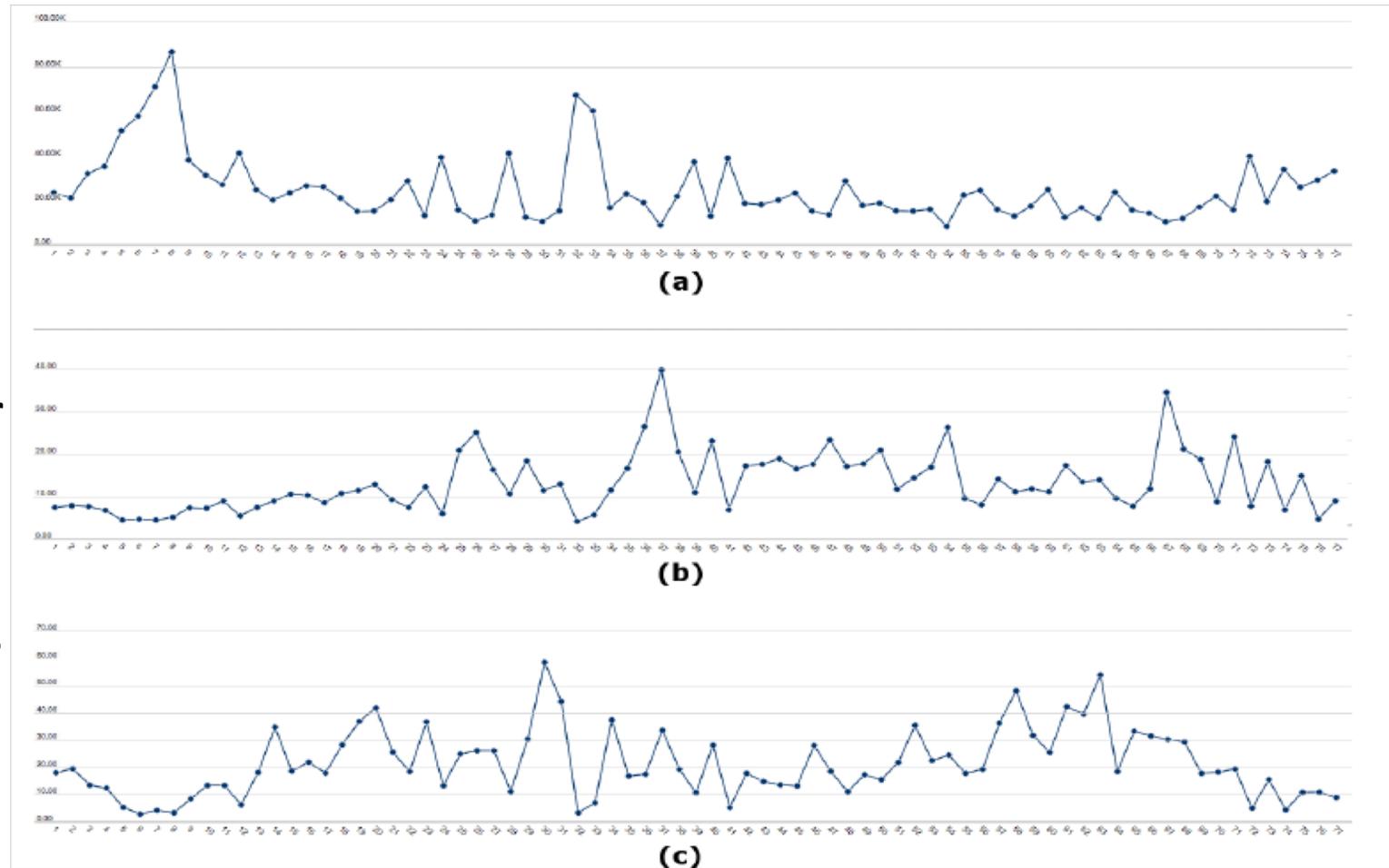
Metadata	
Last Updated Date via Automated Load	
Time Period	2005 - 2011
Data Owner	Epidemiology and Public Health Informatics, Chicago Department of Public Health (CDPH)
Frequency	Updated as new data becomes available

Attachments
[Dataset_Description_Selected_indicators_file_PORTAL.pdf](#)

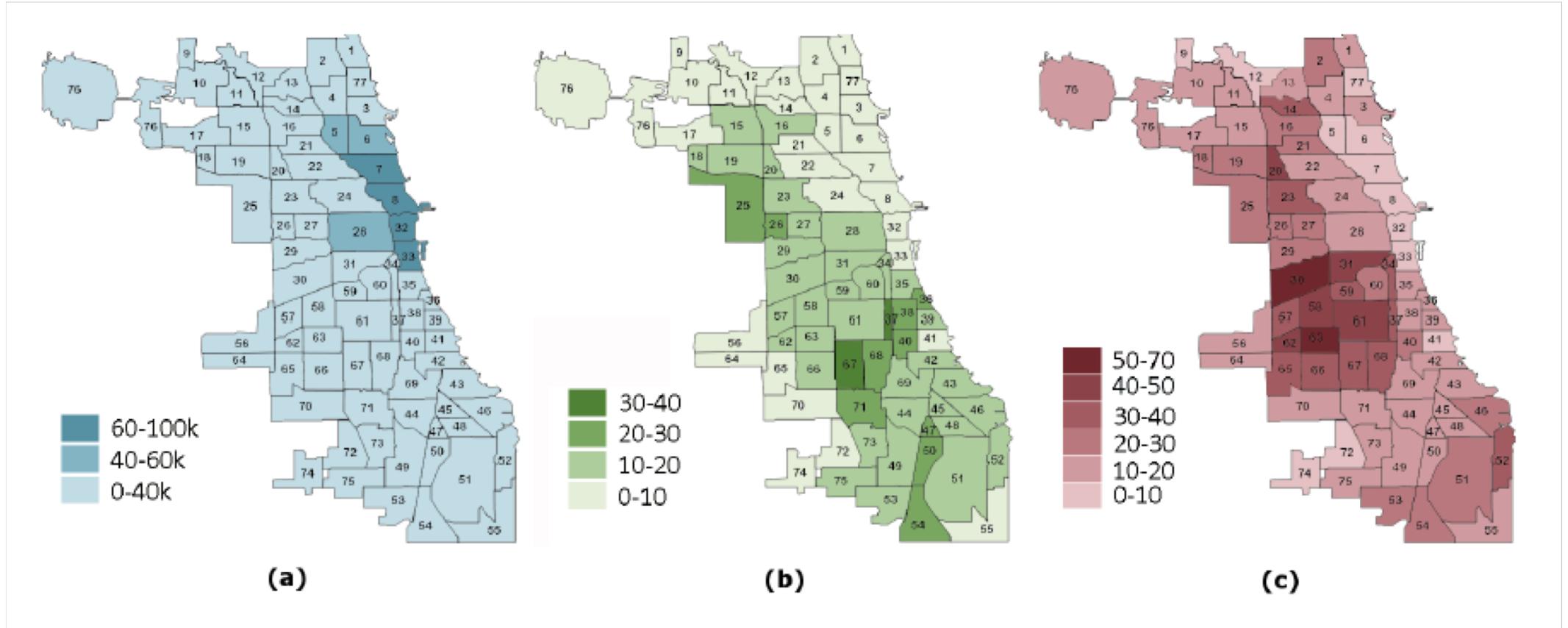
Topics

Data Visualizations Constructed

- Univariate data visualization
- We can observe:
 - a) High / low rates of per capita income
 - b) Unemployment
 - c) People without higher education (No high school Diploma)



Data Visualizations Constructed



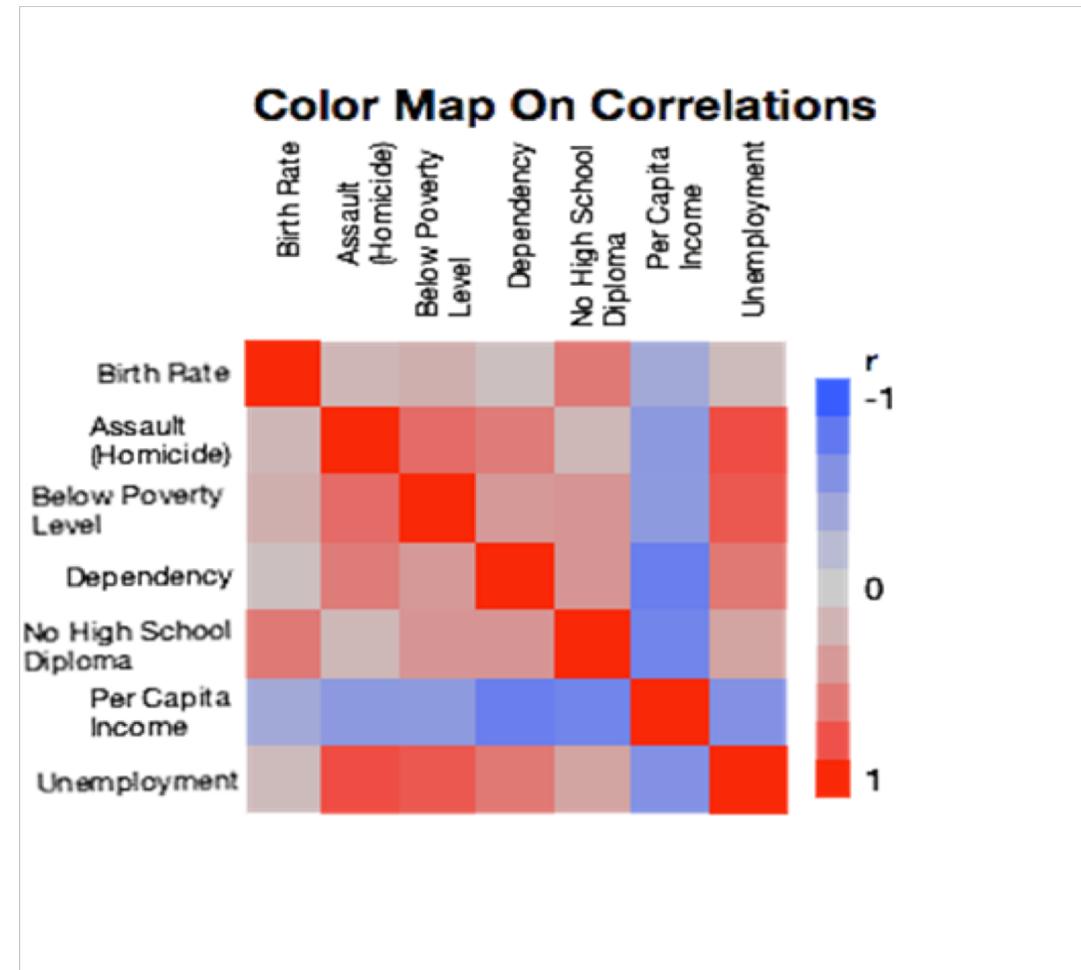
Per Capita Income

Unemployment

No High School Diploma

Data Visualizations Constructed

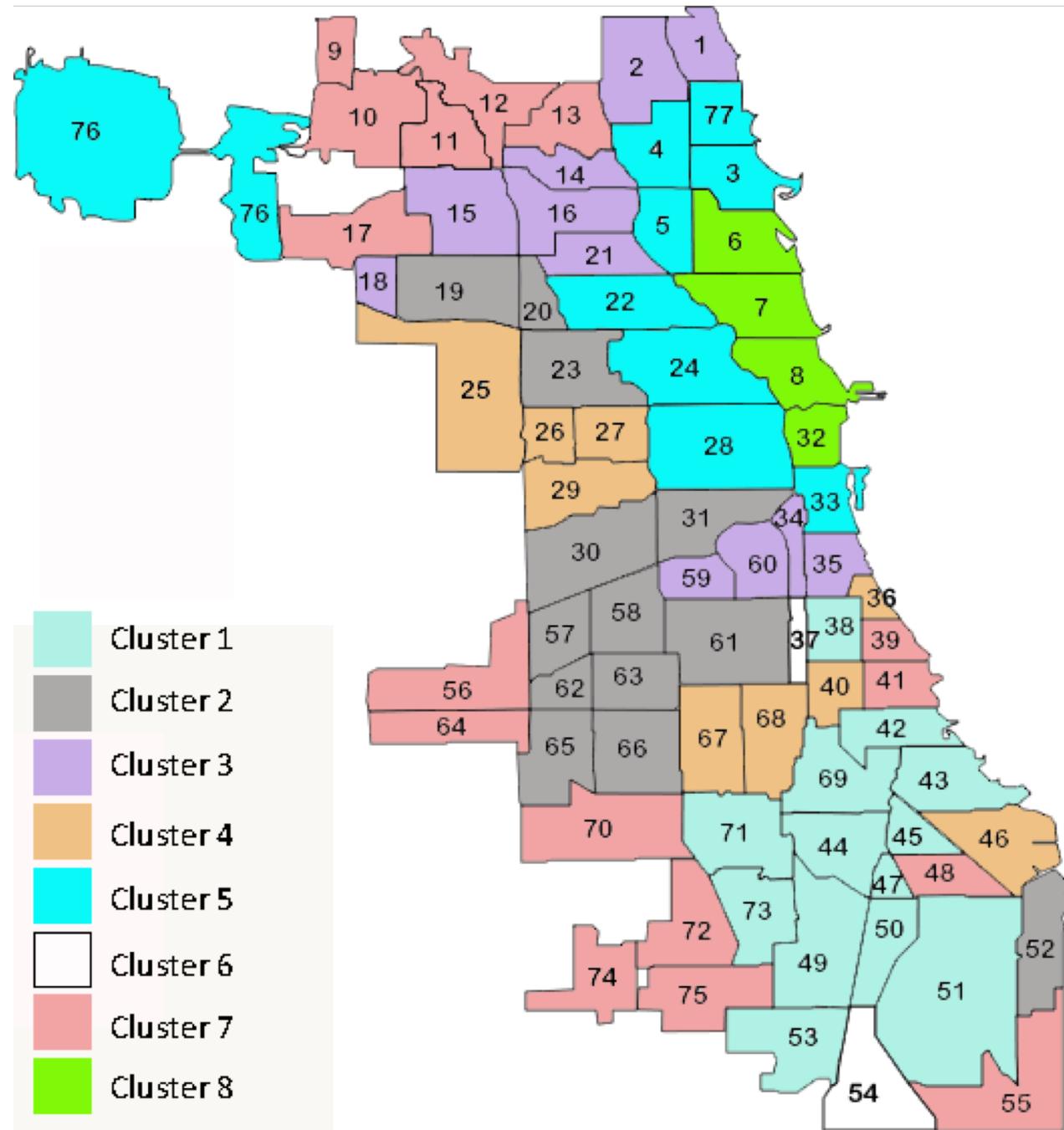
- Bivariate data visualization
- We calculate the Pearson's linear correlation matrix between pairs of data set attributes (Color Matrix).
- It lacks a possibility of discovery of knowledge about all the regions of Chicago.



Data Visualizations

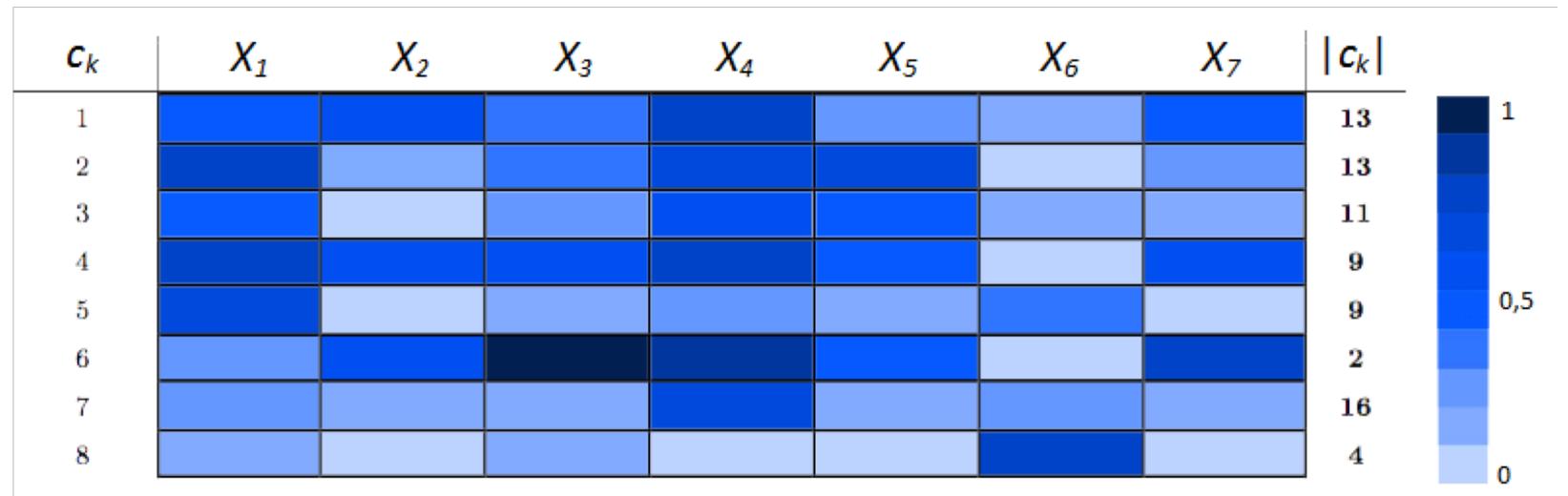
Constructed

- Multivariate data visualization
- Clustering (K-means)
- Cubic Clustering Criterion. (K=8)
- Example: Cluster 8 join richer areas of Chicago, with high per capita income and low unemployment.



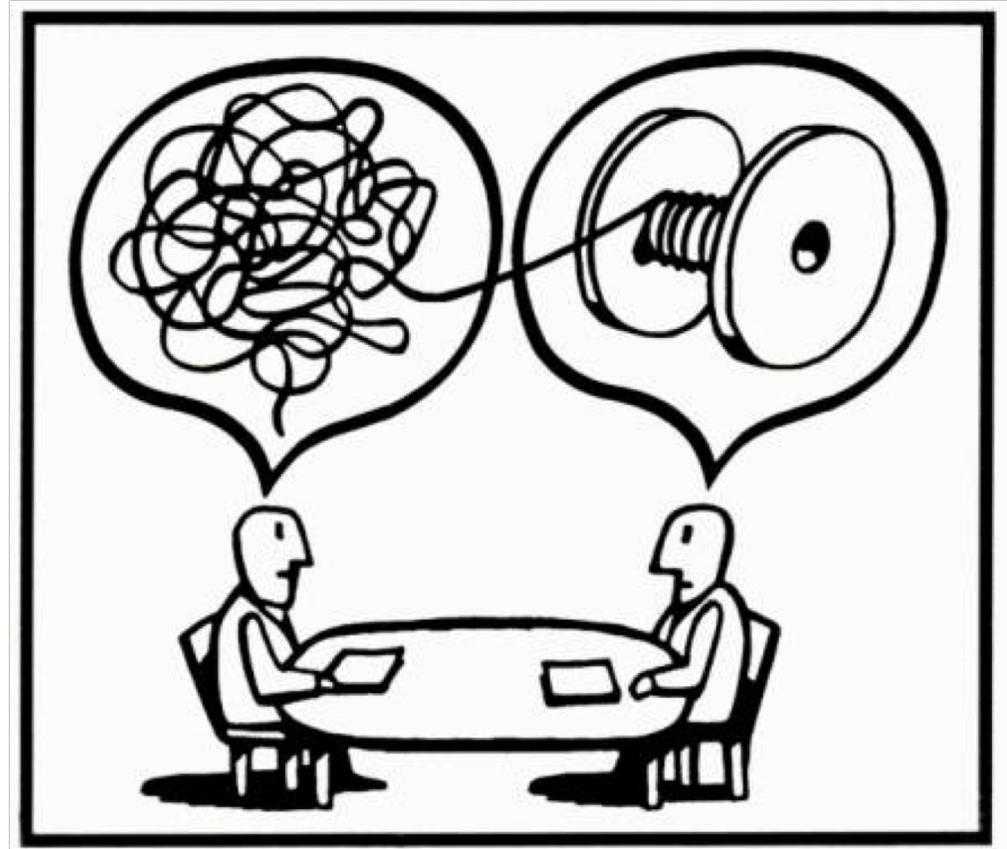
Data Visualizations Constructed

- Centroids per cluster constructed using K-means for $K = 8$ using the normalized version of Socio-Chicago dataset



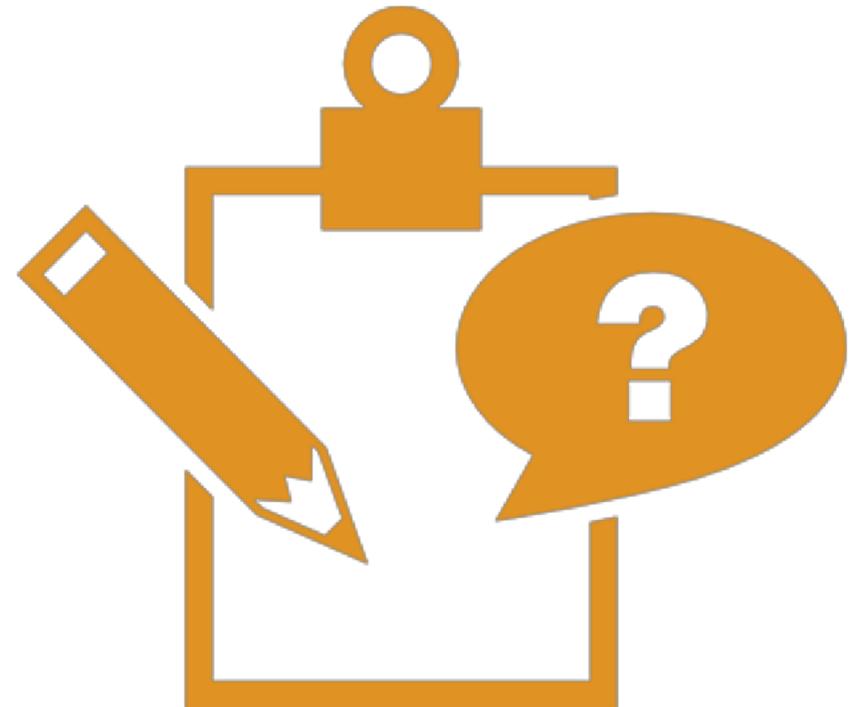
Sensemaking for evaluating interpretability

- Sensemaking evaluation of these visualizations to assess how they affect the users' understanding of the data.

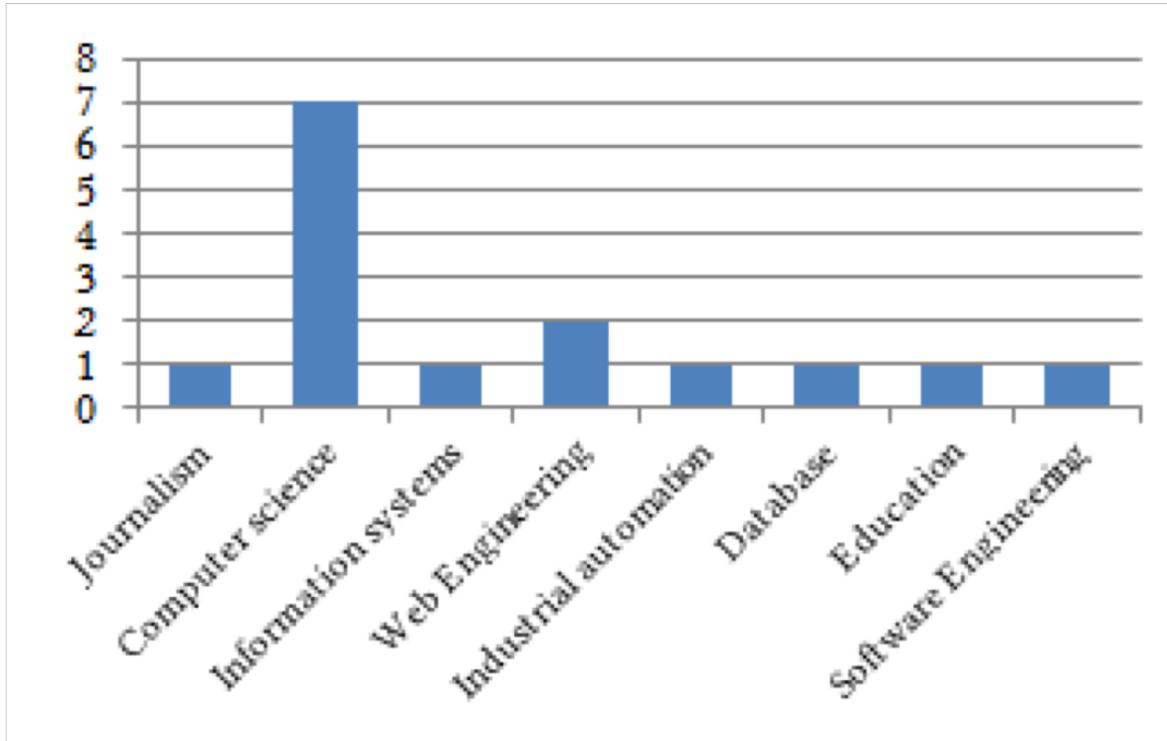


Sensemaking for evaluating interpretability

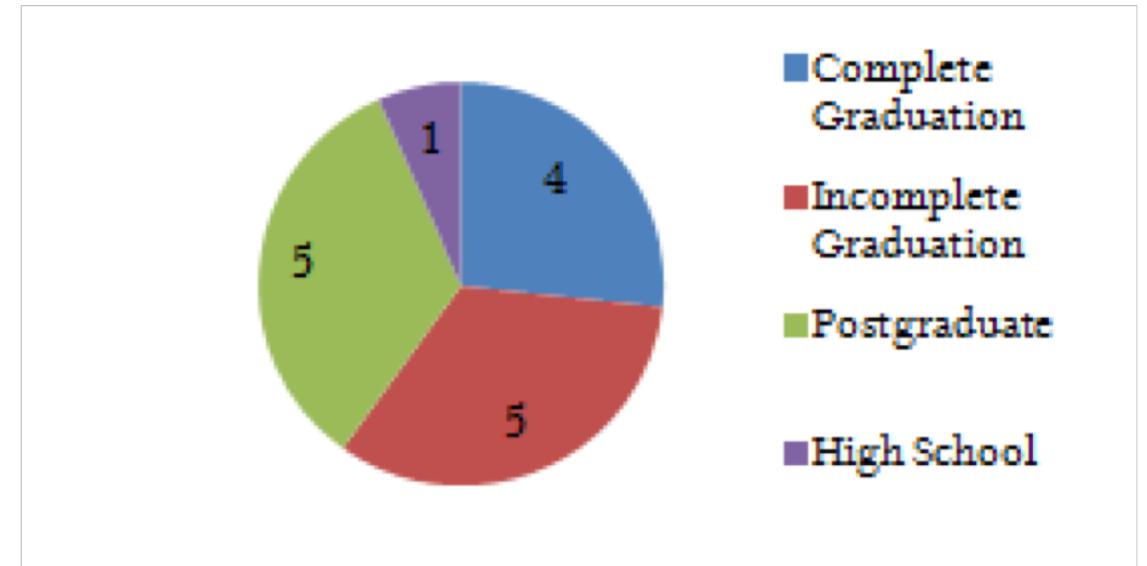
- Questionnaire constructed
- For each type of visualization (Graphs):
 - Direct questions evaluating if the right answer can be acquired from the visualization
 - Comparison to analysis of tabular data



Results achieved



Distribution of participants by graduation



Distribution of participants per educational level

Results achieved

Univariate visualizations	Hits (m ± sd)	Average effort (1-5)	Average self-conf. (1-5)
Raw data table	43% ± 0,26	2,73	2,3
Figure 1	73% ± 0,41	3,46	3,03
Figure 2	77% ± 0,37	3,03	2,33

Multivariate visualizations	Hits (m ± sd)	Average effort (1-5)	Average self-conf. (1-5)
Raw data table	90% ± 0,16	2,6	3
Figure 4	86% ± 0,60	2,6	3,1
Figure 4 + Figure 5	51% ± 0,35	2,6	2,9

Bivariate visualizations	Hits (m ± sd)	Average effort (1-5)	Average self-conf. (1-5)
Raw data table	18% ± 0,25	3,5	2,6
Figure 3	66,6% ± 0,38	2,2%	3,6

Sensemaking – Conclusions

- A lack of familiarity with new types of visualization is still a problem, although the number of hits for maps showing regions of the same group was high
- It is necessary to
 - Define a general methodology to assess interpretability
 - Extend the concept of interpretability of visualizations
 - Expand this study by increasing our surveyed population

How Cities Categorize Datasets in their Open Data Portals : an Exploratory Analysis

dgo 2018

Higor dos Santos Pinto

Flavia Bernardini

José Viterbo

Open Data

- Open data approach can increase transparency in public admin
- Rise of different economic and business opportunities
- Development of new applications and tools to improve cities service
- Several city governments have built open data portals

Urban Data Portals

- **Thousands of open datasets** published by governments were made available through portals on internet
- The **growth in the number of available datasets has increased users' difficulty** in obtaining useful information
- Users spend time **integrating the various datasets** and producing relationships between the chosen ones
- Often, administrations make datasets public by grouping them into categories or topics
 - Different portals may use distinct categories
 - There is semantic similarity between the categories

Categories

NYC OpenData

Datasets by Category



Business



City Government



Education



Environment



Health



**CHICAGO
DATA PORTAL**

BROWSE THE DATA CATALOG BY THE FOLLOWING CATEGORIES

 Administration & Finance

 Buildings

 Community

 Education

 Environment

 Ethics

 Events

 FOIA

 Facilities & Geo.
Boundaries

 Health & Human Services

 Historic Preservation

 Parks & Recreation

 Public Safety

 Sanitation

 Service Requests

 Transportation

BROWSE DATA BY CATEGORY

OPEN DATA HOUSTON



Geographic
Boundaries



Planning &
Development



Public Works &
Engineering



Permitting & Licensing



Public Health & Safety



Administration &
Regulatory Affairs



Neighborhood Services



Finance



Environmental



Property



Flood Hazards

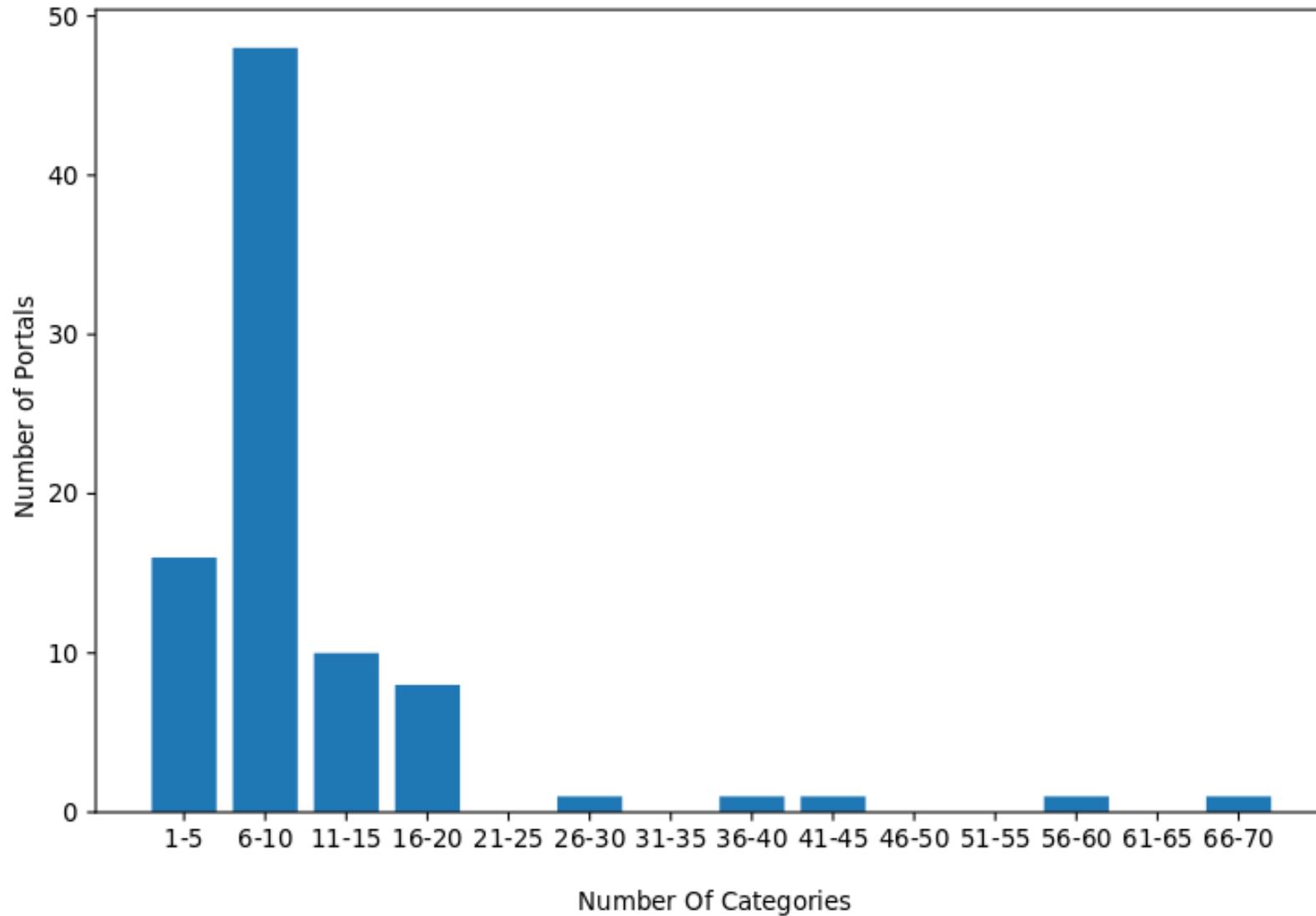


Parking

Objectives

- Identify the most significant categories used in urban data portals
- We conducted a study about how the 100 different portals of the most populous cities in the USA categorize their datasets
- We propose a method that creates a generic categorization for better organization of datasets in public portals

Portals Categorization



Number of portals
with a fixed number
of categories.

Most Significant Words

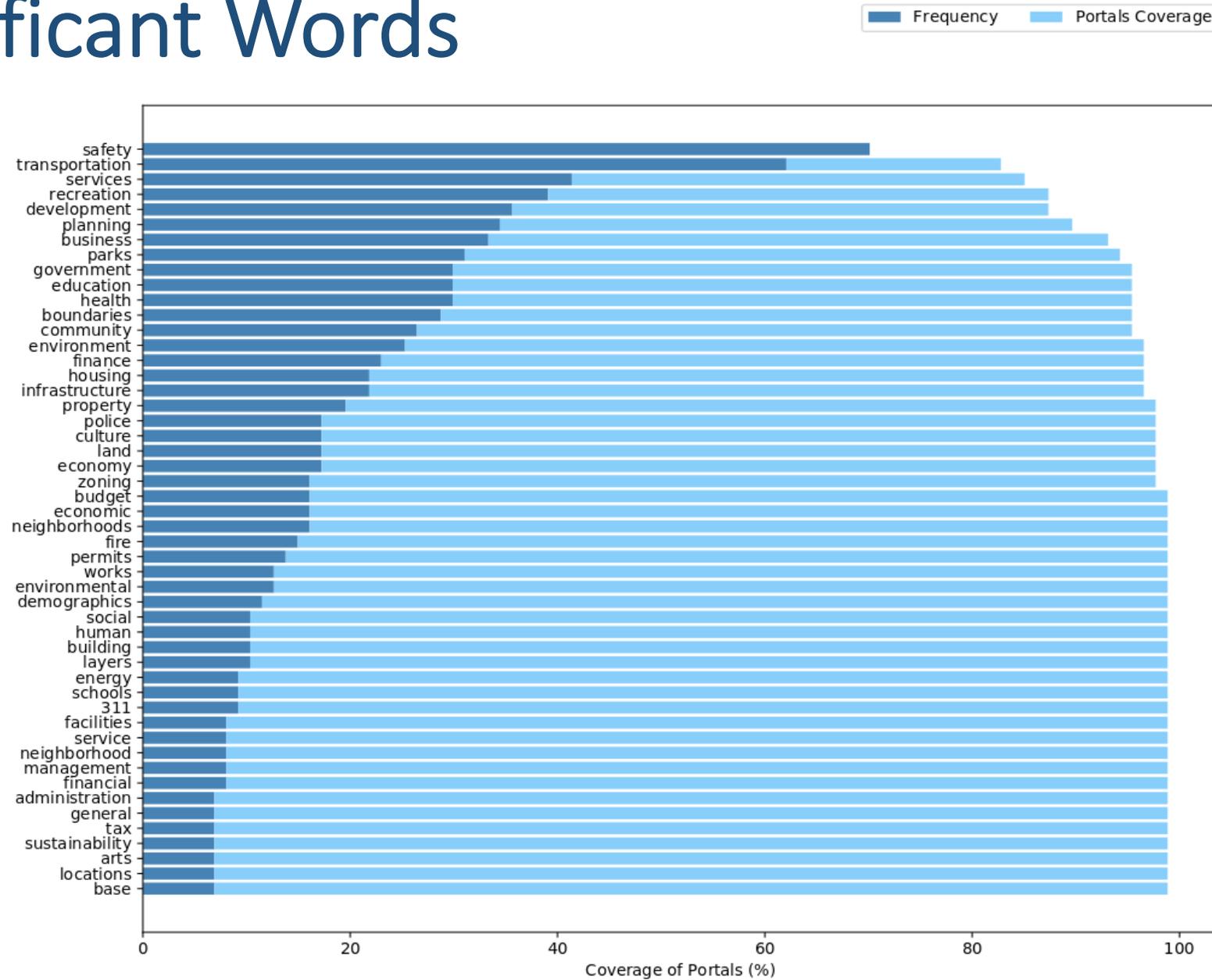


Fig. Shows the distribution of occurrence of the 50 most frequent words.

Most Significant Categories

- In addition, we discuss the most frequent categories of the word **community**, that are: **Community** and **Community Services**.
- As we can not distinguish the words related to the most frequent word we will use the category that maintains the most generic expression possible
- There are some repeated words among the categories of the obtained set, for instance, present the categories **Recreation**, **Parks & Recreation** and **Culture & Recreation**. And both **Business** and **Businesses & Budget**.
- Since they are categories with very similar semantic senses, we propose the category that is more frequent among those that have similar meaning. For this, we will evaluate how often the words that are part of the category appear together

Public Safety	Community
Transportation	Environment
City Services	Finance
Recreation	Housing
Economic Development	Infrastructure
Planning	Property
Business	Police
Government	Land Use
Education	Economy
Health	Zoning
Boundaries	

Most significant categories found in the 100 portals of American cities.

Spatiotemporal Anomaly Detection Applied to Flow Measurement Points in Natural Gas Production Plants

Hadriel Lima

Flavia Bernardini

Objective

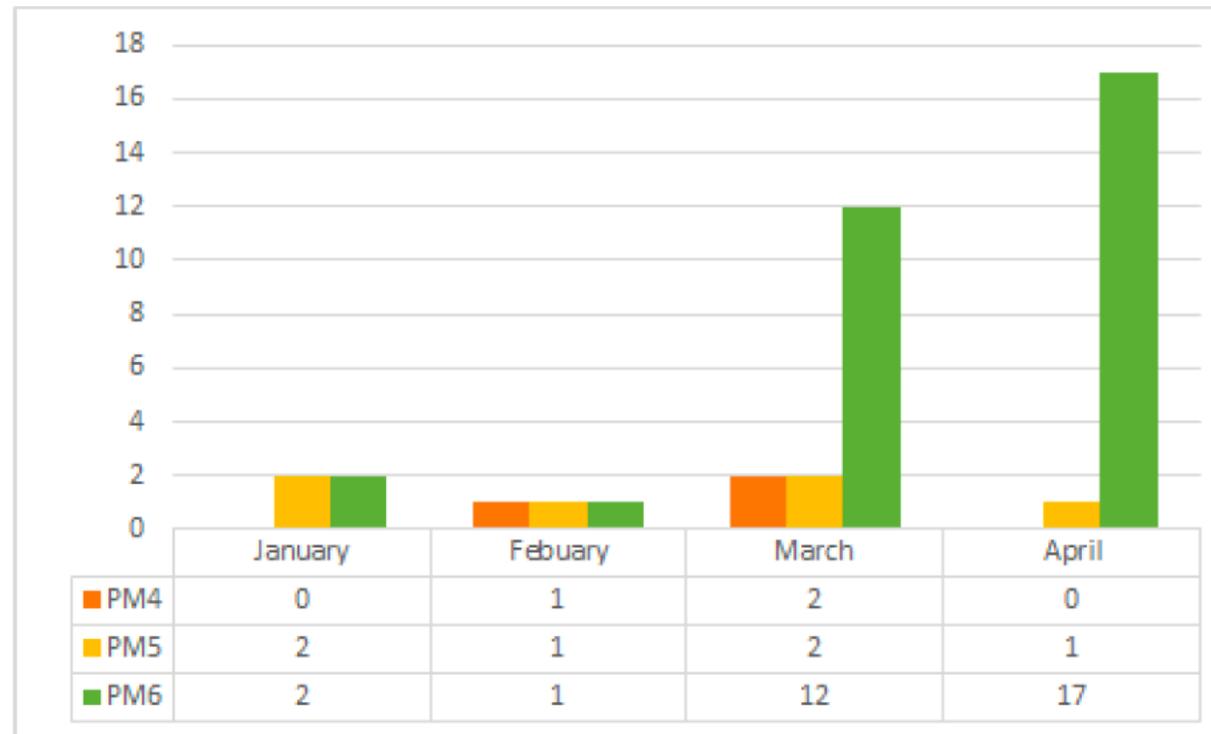
- “This article aims to propose a method for identifying anomalies in natural gas measurement points in production plants using Hybrid Bayesian Networks (HBN)”
- Problema: parametrização da Rede Bayesiana Híbrida

Results

Table III. Result of anomaly detection reported by Spatiotemporal Model

		Model Prediction		Accuracy	Precision	Recall
		Working	Fault			
PM1	Working	120	0	1	1	1
	Fault	0	0			
PM2	Working	120	0	1	1	1
	Fault	0	0			
PM3	Working	120	0	1	1	1
	Fault	0	0			
PM4	Working	115	3	0.975	1	1
	Fault	0	2			
PM5	Working	109	6	0.942	0.991	1
	Fault	1	4			
PM6	Working	82	32	0.733	1	1
	Fault	0	6			

Results



Results

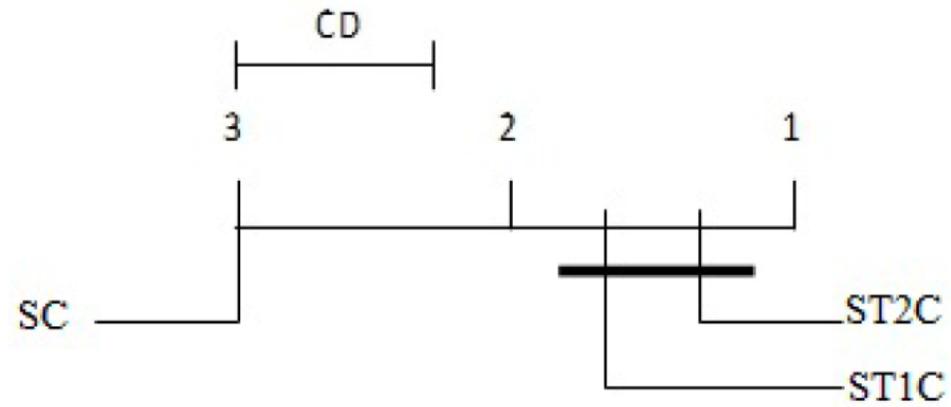


Fig. 10. Graphical representation of the post test of each Bayesian network model as a random variable

Conclusões

Como escolher a metodologia de avaliação?

- Verificar na literatura em trabalhos relacionados
- Pensar em quais são as variáveis de resposta e controle
 - Protocolo de pesquisa em planejamento de experimentos
- Uma boa visualização muitas vezes é construída após várias iterações e interações com o orientador/equipe
 - Importante o planejamento para entrega da versão ao orientador

Referências Bibliográficas

- FACELLI, K.; LORENA, A.C.; GAMA, J.; CARVALHO, A. Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. LTC, 2011
- JAPKOWICZ, N.; SHAH, M. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, 2011.
- DODIG-CRNKOVIC, G. Scientific Methods in Computer Science. Available at <http://poincare.math.rs/~vladaf/Courses/Matf%20MNSR/Literatura/Scientific%20Methods%20in%20Computer%20Science.pdf>.
- WERKEMA, M. C. C.; AGUIAR, S. (1996). Planejamento e análise de experimentos: como identificar as principais variáveis influentes em um processo. Belo Horizonte: Ufmg.

Obrigada!

fcbernardini@ic.uff.br