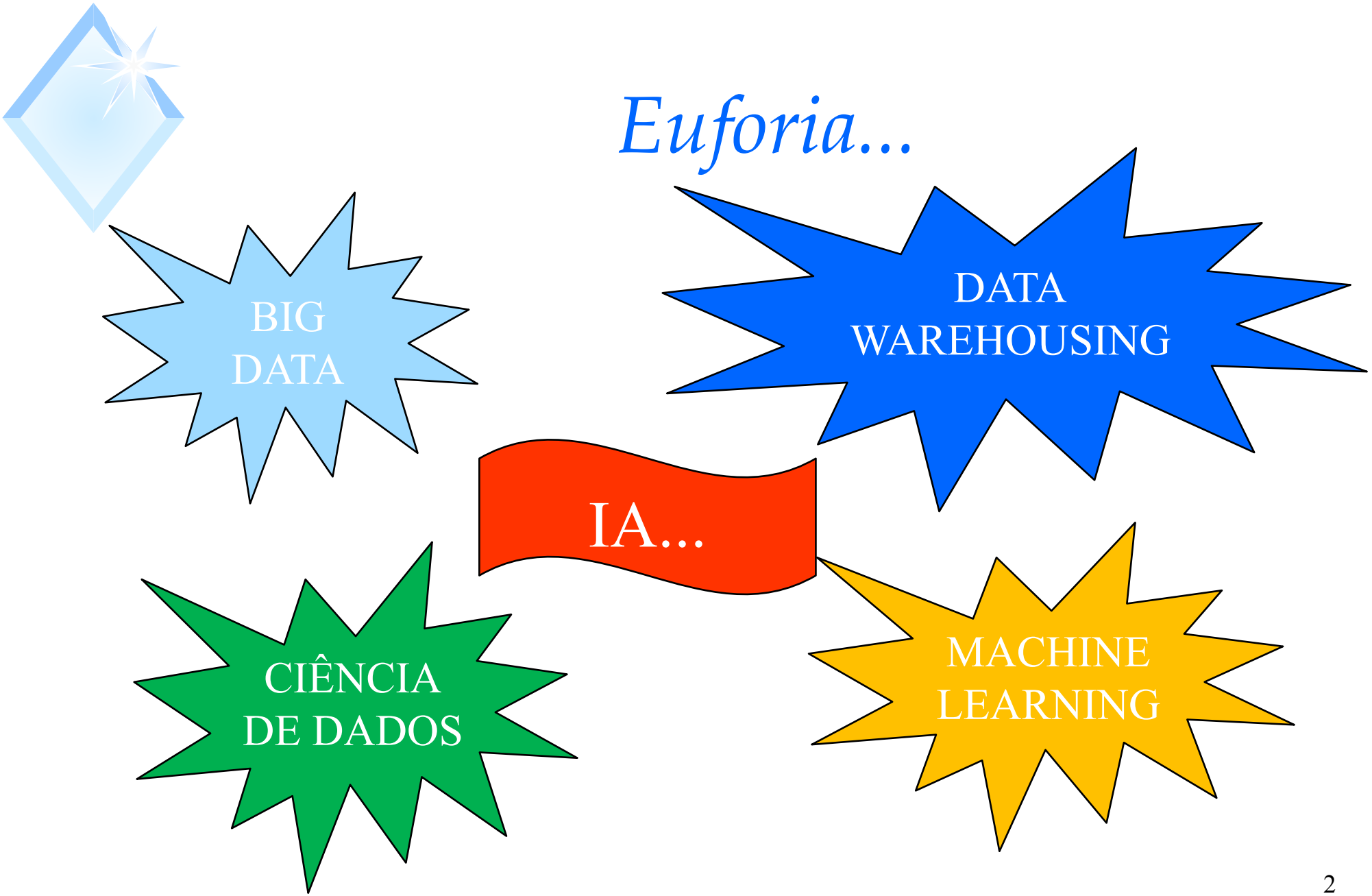
Estatística e Banco de Dados: Uma relação simbiótica

Professor Marcelo Menezes Reis
Departamento de Informática e Estatística
Centro Tecnológico
Universidade Federal de Santa Catarina
marcelo.menezes.reis@ufsc.br

Euforia...


BIG DATA
DATA WAREHOUSING
IA...
CIÊNCIA DE DADOS
MACHINE LEARNING

Big Data e Data Mining

"Descobrir padrões de significado prático em grandes arquivos de dados"

Técnicas de Inteligência Artificial

Técnicas de Aprendizagem de Máquina

Big Data e Data Mining

"Descobrir padrões de significado prático em grandes arquivos de dados"

"Não precisa de teoria"

"Não precisa de Estatística, Amostragem"

A QUANTIDADE de dados será suficiente: correlações

Big Data e Data Mining

Vamos ao passado e depois retornamos...

Big Data em 1936...

- EUA, Grande Depressão.
- Franklin Delano Roosevelt presidente em 1932:
  - Buscava a reeleição em 1936.
  - Candidato republicano: Alfred Landon.
- Revista *Literary Digest*:
  - Prever o resultado da eleição.
  - Usando *Big Data*.

Big Data em 1936

- *Literary Digest* enviou 10 milhões de cédulas pelo correio (25% do eleitorado).
  - Obtiveram 2,4 milhões de respostas.
  - "**Big Data**"
  - Landon venceria Roosevelt: 55% a 41%.
- George Gallup fez uma pesquisa com uma amostra *probabilística* do eleitorado.
  - "Apenas" 50 mil eleitores, Roosevelt venceria (56%)

Big Data em 1936

- Roosevelt venceu em 46 estados (61% a 37%).
- O que aconteceu? Era "Big Data"!
  - Viés na amostra: cédulas enviadas para pessoas com maior simpatia pelos Republicanos.
  - Baixa taxa de resposta (25%).
  - Viesamento dos respondentes: maioria dos que responderam preferiram Landon, maioria dos que *não* responderam escolheram Roosevelt.

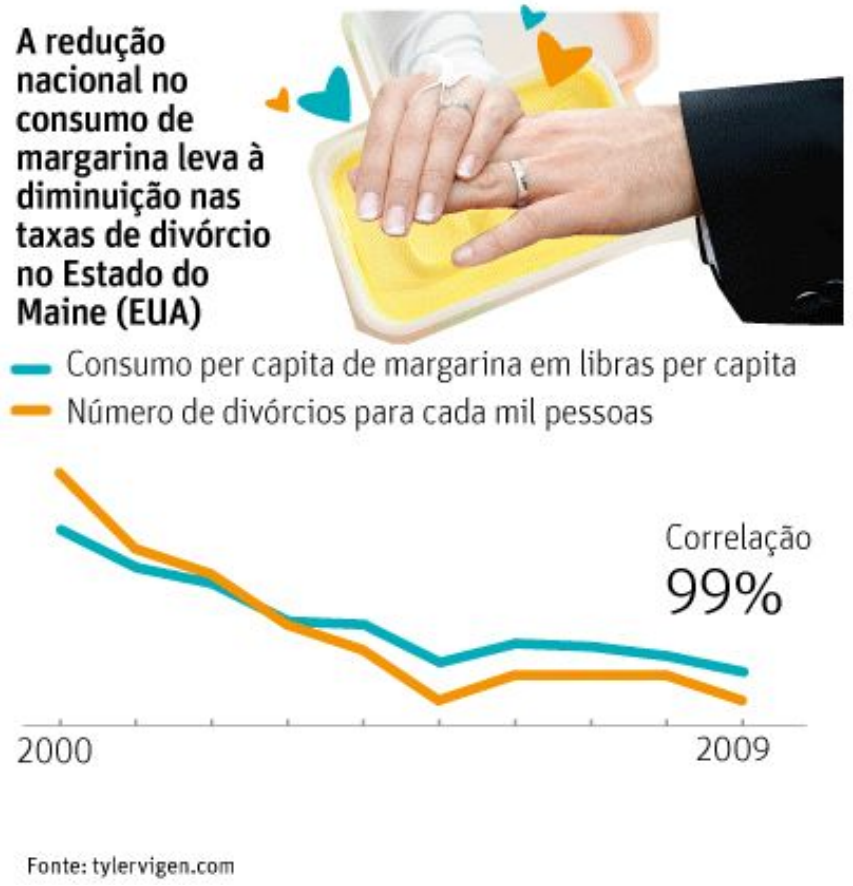
Fonte: Squire, P. 1988

Voltamos...

- Google Flu Trends
  - Prever o número de casos de gripe (Influenza) nos EUA mais rápido e melhor do que o CDC.
  - Base em buscas feitas no Google (“sintomas de gripe”, “farmácias perto de mim”).
  - “Livre de teoria”
    - Sem preocupação em definir uma teoria que ligasse o número real de casos com as buscas.

Google Flu Trends

- Em 2011 – 2012: mais de 50% acima dos dados do CDC...
- Em 2012 – 2013 também...
  - Dezembro de 2012: notícias aterrorizantes sobre gripe na mídia.
    - Buscas na internet feitas por muitas pessoas *saudáveis*.



Fonte: Folha de São Paulo, 09/05/2015

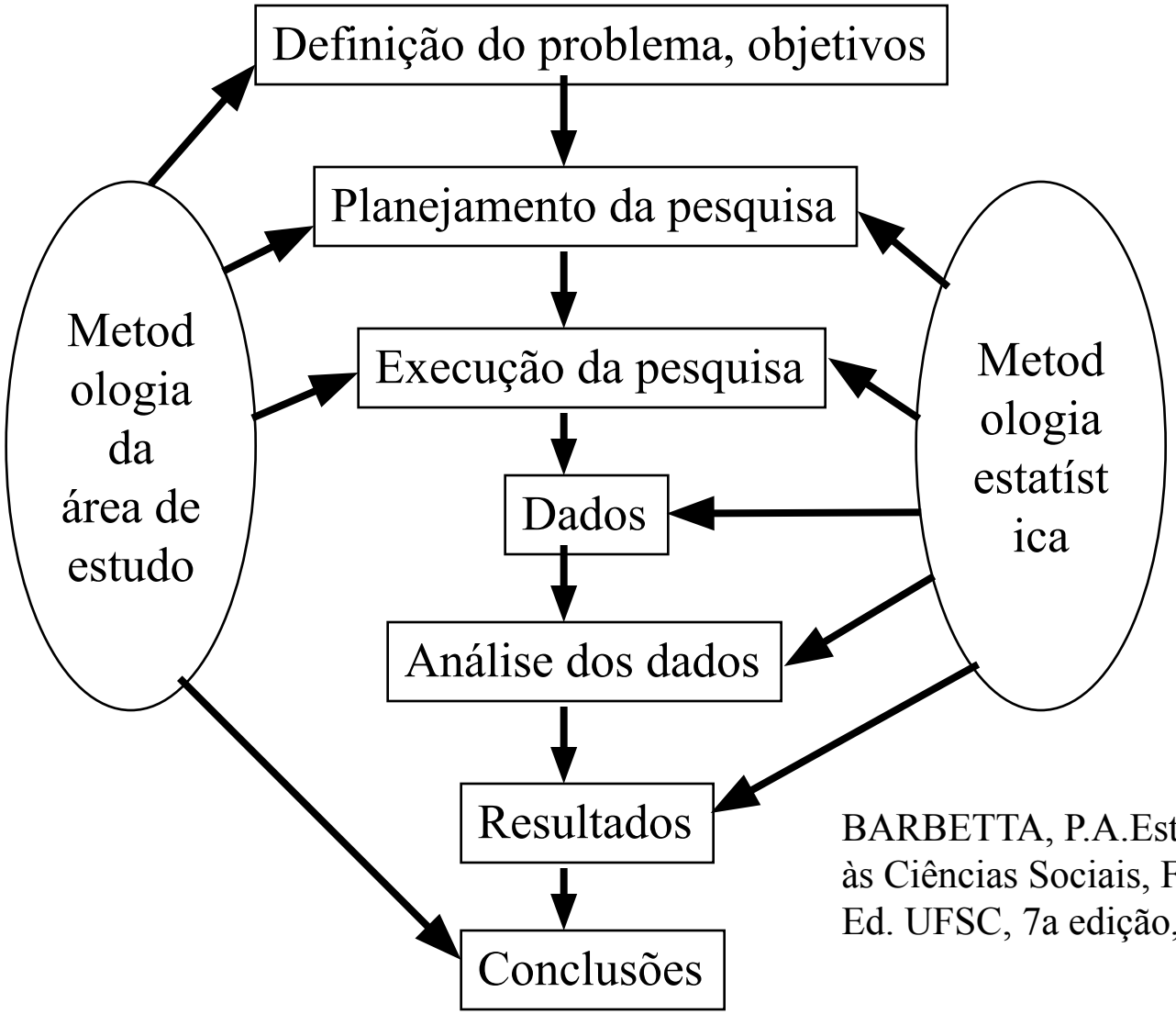
Estatística?

Conhecimento em Estatística é menor até do que em Matemática...

Estatística?

- “Estatística é a Ciência de obter conclusões a partir de dados” – Paul Velleman
  - Aprender a partir dos dados.
  - Medir, controlar e comunicar incerteza.
    - Incerteza registrada em Probabilidade.

KUONEN, D. - 2015



BARBETTA, P.A.Estatística Aplicada às Ciências Sociais, Florianópolis: Ed. UFSC, 7a edição, 2010

Estatística, Big Data

Estatística: coleta, análise, interpretação de dados, preocupando-se com as incertezas.
Análise top - down

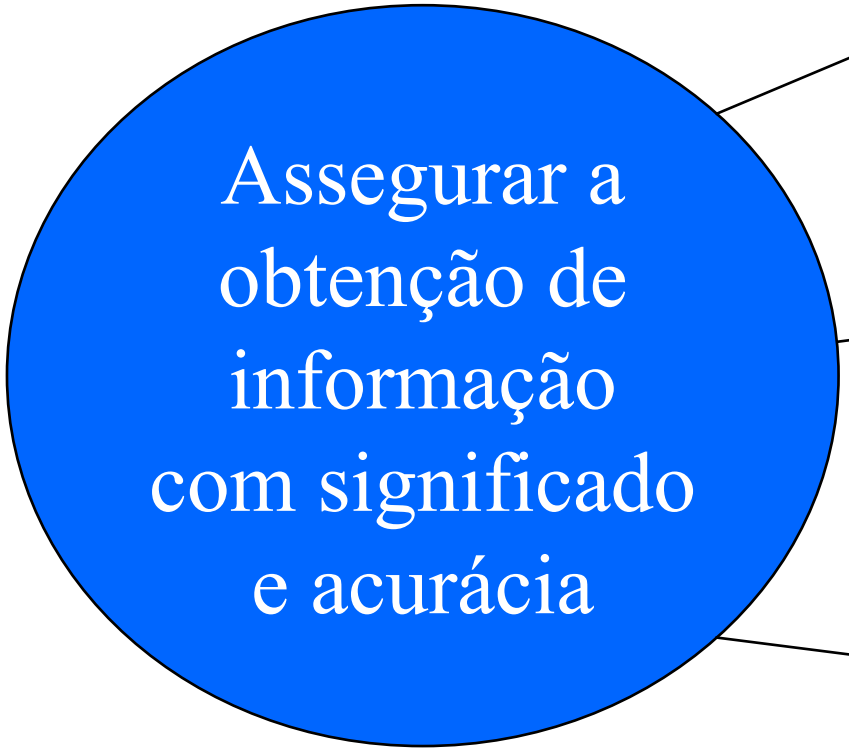
Big Data: coleta e análise de conjuntos complexos de dados em volume e variedade.
Análise bottom - up

Estatística e Big Data

“ Em Big Data, Estatística e a área de conhecimento do problema estão mais entrelaçadas do que nunca, e a metodologia estatística é absolutamente crítica para realizar inferências”

American Statistical Association

Importância da Estatística para Big Data e Banco de dados



Qualidade dos dados e dados perdidos => Pré-análise

Identificar/definir a natureza dos dados para saber o que perguntar.

Traduzir uma questão científica/ comercial em uma questão Estatística: viável

Pré-análise...

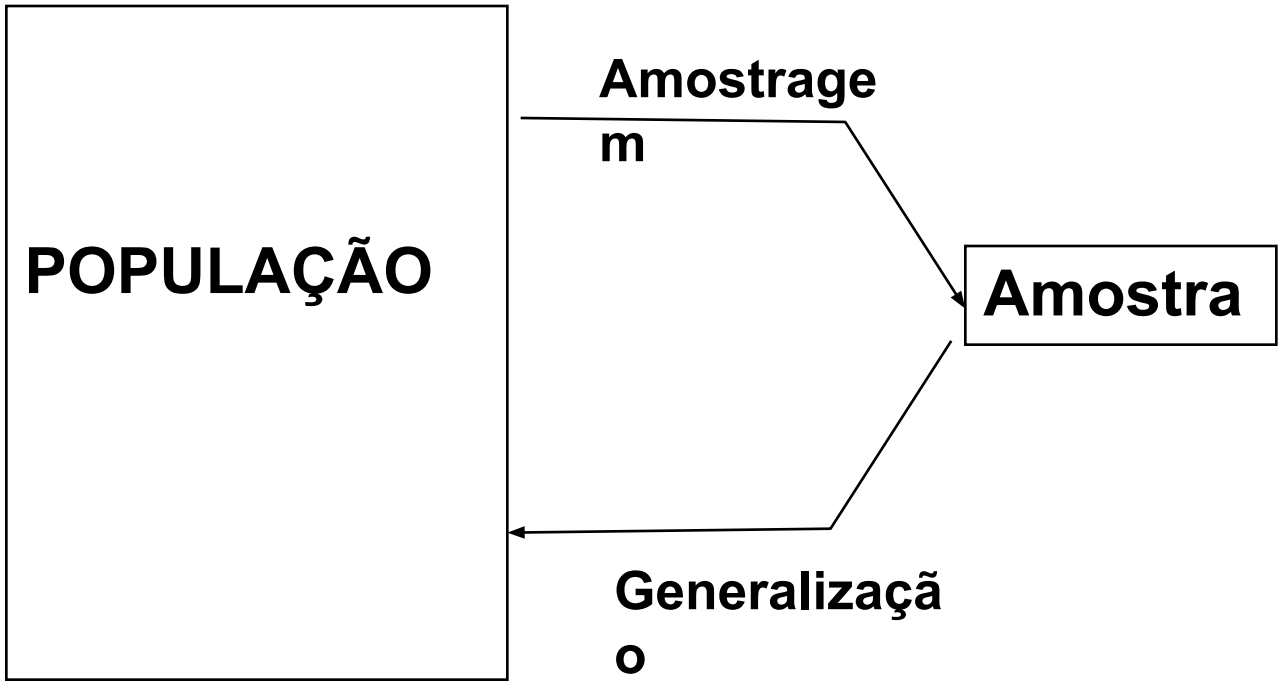
- Projeto de pesquisa, análise dos tempos de operação e tempos de reparo de linhas de transmissão, 1998 a 2006, milhares de ocorrências.
- Linhas identificadas por códigos:
  - AGV JAL, mas havia registros JAL AGV, AGV-JAL
  - OES CBAII, mas havia CBA2 OES, CBAII OES...

Pré-análise

- Alguma técnica de sumarização dos dados:
  - Mensurar quantidade de dados perdidos;
  - Identificar erros de registro;
- Processamento computacional pode ser um problema.
  - Escala da base de dados.
- Importância do Data Warehousing.

Amostragem

- Útil para selecionar um conjunto de dados para treinamento e outro para validação de técnica.



Amostragem

- Conhecimento da população.
- Representatividade: representar as subdivisões da população na amostra.
- Suficiência: garantir que haja quantidade de dados que retratem a variabilidade da população.
- Aleatoriedade: amostra coletada por sorteio não viciado (algoritmo de números pseudo-aleatórios).

Análise Estatística

- AED: sumarizar os dados para interpretação.
  - Tabelas, gráficos, medidas de síntese. SEMPRE!
- Inferência estatística: a partir de uma amostra aleatória de uma população estimar suas características.
  - Associação de probabilidade às conclusões.
- Modelagem estatística: obtenção de uma equação que mostre o relacionamento entre variáveis

Estatística e Aprendizado de Máquina

- Aprendizado de máquina supervisionado:
  - Regressão linear, regressão logística.
- Aprendizado de máquina não supervisionado:
  - Análise de agrupamentos em k-médias, análise de componentes principais, análise fatorial

Outras técnicas

- Modificação dos testes de hipóteses para Big Data:
  - Múltiplas comparações, método FDR;
  - Ou usar apenas Intervalos de Confiança;
- Métodos bayesianos modificados
- Técnicas de reamostragem: bootstrap, jacknife para estimação de parâmetros => com adaptações.

Desafio (KUONEN, D. 2015)

- Decisão baseada em dados e não mais em "faro".
- Grandes bancos de dados => Big Data
- Extrair valor dos dados => Estatística.
- Pensamento estratégico, conhecimento da área.
- Entender o processo que gerou os dados...
- Alerta para correlações espúrias e variáveis de confusão.

“A melhor maneira de perder dinheiro bem rápido é através de análise automática. Eu penso que devemos ter muito cuidado para não sair totalmente do controle e deixar as coisas acabarem mal”.

Thomas H. Davenport

Referências

- BARBETTA, P.A., REIS, M.M., BORNIA, A.C. Estatística para Cursos de Engenharia e Informática. 3ª ed. São Paulo: Atlas, 2010.
- HAIR Jr., J. F., BLACK, W.C., BABIN, B.J.; ANDERSON, R. E., TATHAM, R.L. Análise Multivariada de Dados. Porto Alegre: Bookman, 6ª edição, 2009.