

Data Science e Machine Learning na Prática - Introdução e Aplicação na Indústria de Processos

Escola Piloto Prof.Giulio Massarani
Aula 0 – Introdução

Afrânia Melo

<http://instagram.com/engepolgrupo>
<http://kaggle.com/afrniomelo>
<http://afrjr.weebly.com>
afraeq@gmail.com

Laboratório de Modelagem, Simulação e Controle de Processos – LMSCP
Laboratório de Engenharia de Polimerização – EngePol
PEQ-COPPE-UFRJ

2020

1 Ciência de Dados e Aprendizado de Máquina

- Um pouco de contexto...
- Dados
- Ciência de Dados e Big Data
- Aprendizado de Máquina

2 Indústria 4.0 e transformação digital

- A Indústria 4.0
- Dados de processo

Bem-vindos(as)!



- Bem-vindas(os) à Escola Piloto Prof. Giulio Massarani, do Programa de Engenharia Química da COPPE-UFRJ!
- Aprenderemos neste curso os fundamentos da **Ciência de Dados e Aprendizado de Máquina**, sempre usando como pano de fundo aplicações na linguagem **Python** relacionadas a **problemas de aplicabilidade industrial**.
- As aulas se darão em um ambiente de programação conhecido como **Jupyter Notebook**, a ser apresentado mais adiante.
- Vamos começar nossa introdução!

Algumas notícias...

SAÚDE

Apple Watch salva a vida de brasileiro ao alertar taquicardia

Relógio inteligente avisou o usuário que sua frequência cardíaca estava muito acima do normal e, no hospital, ele foi diagnosticado

JULIANA CONTAIFER

07/01/2020 17:18, ATUALIZADO 07/01/2020 17:24

© DANIEL CANIBANO/UNSPLASH



Figura 1: Fonte: Metrópoles.

Algumas notícias...

Bezos, Zuckerberg e Musk ficaram US\$ 115 bilhões mais ricos só neste ano



O fundador e CEO da Amazon Jeff Bezos, homem mais rico do mundo

Imagen: Eric Baradat/AFP



Bloomberg

Ben Steverman e Sophie Alexander

30/07/2020 17h08

Figura 2: Fonte: UOL.

Algumas notícias...

Presidente da Cambridge Analytica confessa influência em eleições dos EUA

Suspenso pela consultoria política, Alexander Nix disse, em vídeo secreto divulgado nesta terça, que sua consultoria política teve papel decisivo na eleição de Trump

21/03/2018 | 09h34



■ Por Agências - Reuters



Figura 3: Fonte: Estadão.

Algumas notícias...

Alemanha já trabalha na regulamentação de carros autônomos

20/07/2020 às 15:30 • 1 min de leitura



Figura 4: Fonte: Tecmundo.

Algumas notícias...

18/04/2011 - 15h30

Irã acusa Siemens, EUA e Israel pelo vírus Stuxnet

DA REUTERS, EM TEERÃ

Um comandante militar iraniano acusou o conglomerado alemão Siemens de ajudar os Estados Unidos e Israel a lançar um ciberataque contra suas instalações nucleares, publicou o diário iraniano "Kayhan" nesta segunda-feira.

Gholamreza Jalali, diretor da defesa civil iraniana, disse que o vírus Stuxnet, que tomou por alvo o programa nuclear iraniano, é obra dos dois maiores inimigos do país, e que a companhia alemã deve arcar com parte da culpa.

"As investigações mostram que a fonte do vírus Stuxnet está nos Estados Unidos e no regime sionista", teria dito Jalali.

Jalali disse que o Irã considera a Siemens responsável pelos sistemas de controle usados para operar maquinaria industrial sofisticada --conhecidos como Supervisory Control and Data Acquisition (SCADA)-- que teriam sido atingidos pelo vírus.

Figura 5: Fonte: Folha.

Algumas frases...

- “Se os dados tivessem massa, a Terra seria um buraco negro” - Stephen Marsland, acadêmico;
- “A grande tendência da tecnologia é tornar os sistemas inteligentes, e para isso a matéria-prima são os dados ” – Amod Malviya, FlipKart CTO;
- “Aprender a partir dos dados é universalmente útil. Domine essa arte e você será bem-vindo em qualquer lugar” – John Elder, Elder Research CEO;
- “Guerra é 90% informação” – Napoleão Bonaparte, militar e monarca francês.

O que você irá aprender nesta aula?

- O que são **dados**?
- Qual a diferença entre **dado, informação, conhecimento e sabedoria**?
- O que é **Ciência de Dados**?
- O que é **Big Data**?
- O que é **Inteligência Artificial**?
- O que é **Aprendizado de Máquina**?
- O que são **padrões** em conjuntos de dados?
- O que são **modelos e algoritmos**? Como esses conceitos estão envolvidos nas tarefas de **aprendizado e predição** de padrões em dados?

- O que é o **treino** de um modelo?
- O que são **parâmetros e hiperparâmetros** de um modelo?
- Quais os principais **tipos de modelos e algoritmos** que existem?
- Quais as principais **dificuldades** que podem surgir em um projeto de Aprendizado de Máquina?
- O que é a **Indústria 4.0**? Quais são as principais tendências de sua aplicação na **indústria de processos**?
- Quais são as características particulares de **dados de processo** que devem ser levados em conta no âmbito de aplicações práticas?

O que são dados?

Dados

- Nossos protagonistas são os *dados*!
- Vamos iniciar propondo uma definição para esse conceito?
- Gosto da seguinte^[1]:

“Dados são um conjunto de símbolos sem qualquer significado além de sua existência”.

[1] Wolfgang Marquardt et al. *OntoCAPE: a re-usable ontology for chemical process engineering*. RWTH edition. Heidelberg ; New York: Springer, 2010.

Pirâmide DIKW

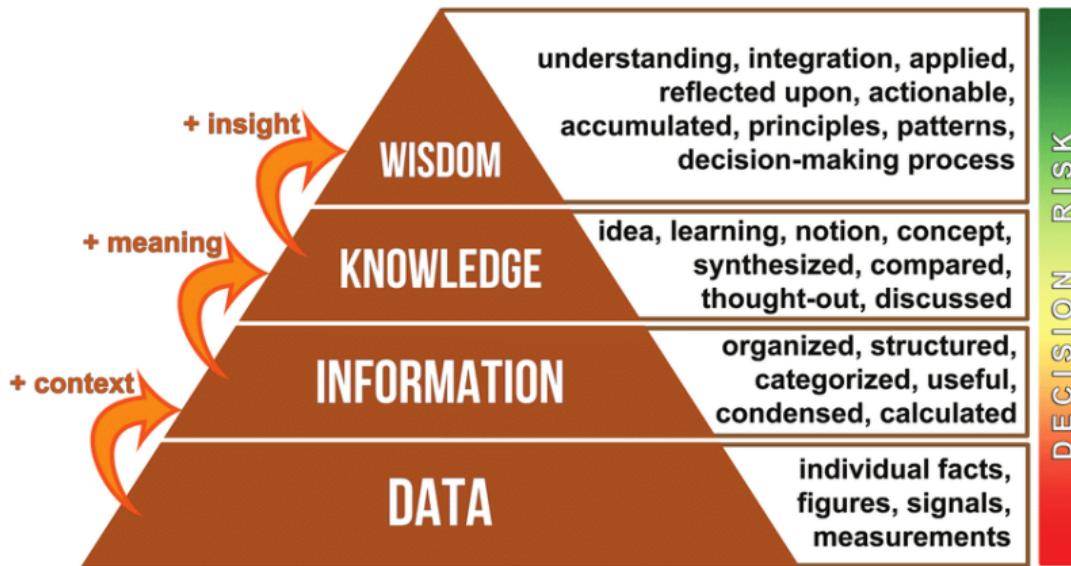


Figura 6: Famosa pirâmide DIKW^[2], que esquematiza o caminho hierárquico dos dados à sabedoria (por <https://researchgate.net/publication/332400827>).

[2] Jennifer Rowley. "The wisdom hierarchy: representations of the DIKW hierarchy". Em: *Journal of information science* 33.2 (2007), pp. 163–180.

O que é a Ciência de Dados?

- A pirâmide DIKW mostra os estágios da extração progressiva de *valor* a partir dos dados.
- Esses estágios são parte do processo cognitivo humano e da atividade científica desde tempos remotos.
- Então por que surgiu nos últimos anos o termo *Ciência de Dados*? Quais as inovações introduzidas por essa jovem disciplina?

O que é a Ciência de Dados?

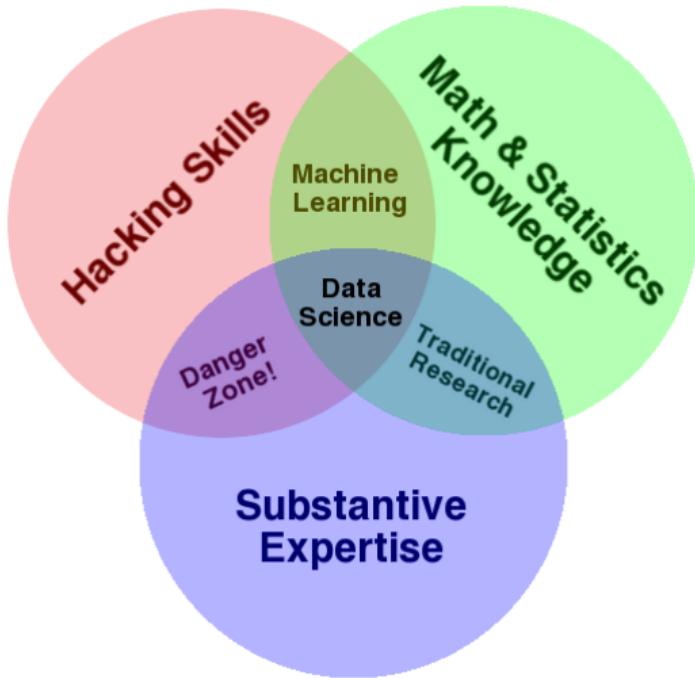
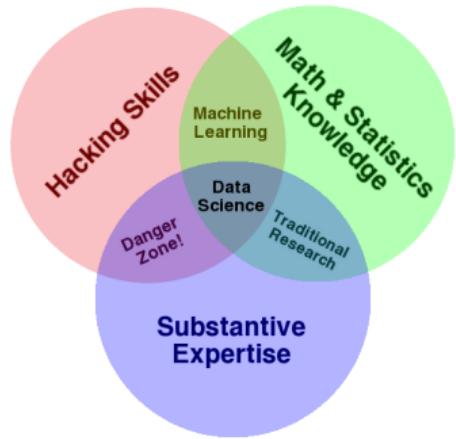


Figura 7: Diagrama de Venn definindo a Ciência de Dados como a interseção de três competências (por <http://drewconway.com>).

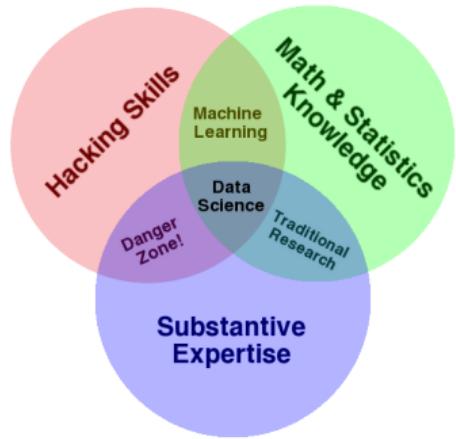
Afinal, o que é a Ciência de Dados?



Competência 1: habilidades computacionais

- Como os dados são bens armazenados e negociados eletronicamente, um cientista de dados precisa ter algumas habilidades computacionais, destacando-se:
 - pensamento algorítmico;
 - capacidade de manipular dados em diferentes formatos;
 - entendimento de operações vetorizadas.

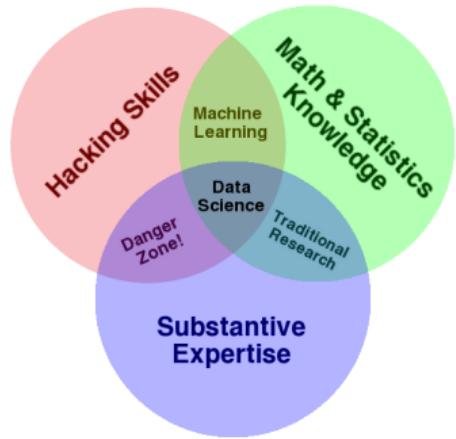
Afinal, o que é a Ciência de Dados?



Competência 2: conhecimento matemático

- Para extrair significado dos dados, é necessária a aplicação de métodos matemáticos e estatísticos.
- Isso requer conhecimento da base teórica das técnicas e de seus mecanismos de funcionamento.

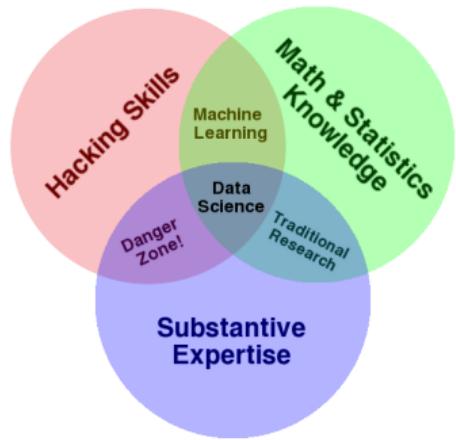
Afinal, o que é a Ciência de Dados?



Competência 3: conhecimento do domínio

- A terceira competência é aquela que justifica o uso do termo “ciência” em Ciência de Dados.
- Ter conhecimento do domínio (sobre Engenharia Química, Nanotecnologia, Fonoaudiologia, Geofísica, Sociologia, etc.) permite usar o significado extraído dos dados para gerar descobertas e construir conhecimento!

Afinal, o que é a Ciência de Dados?



Para pensar

- Reflita sobre as diferentes interseções entre os conjuntos no diagrama ao lado. Onde suas habilidades atuais se encaixam? Onde as habilidades da maioria das pessoas do mundo acadêmico se encaixam? E no mundo profissional? Há algum motivo para isso?

E o tal do *Big Data*?

Definindo *Big Data*

- O termo *Big Data* surgiu para descrever conjuntos de dados com as seguintes características^[3]:
 - **Volume**: há grande quantidade de dados (da ordem de vários TB);
 - **Velocidade**: dados são gerados a altas taxas;
 - **Varietade**: são diferentes formas nas quais e fontes das quais os dados são coletados (texto, som, vídeo, sensores, etc).
- Os três itens acima são conhecidos como os três V's do *Big Data*.
- Só é possível lidar com dados classificados como *Big Data* utilizando as novas metodologias introduzidas pela Ciência de Dados.
- Ultimamente vêm surgindo novas definições de *Big Data* com V's adicionais, como **Valor**, **Veracidade**, **Validade** e **Volatilidade**, por exemplo^[4].



[3] Leo Chiang, Bo Lu e Ivan Castillo. "Big Data Analytics in Chemical Engineering". Em: *Annual Review of Chemical and Biomolecular Engineering* 8.1 (2017), pp. 63–85. DOI: 10.1146/annurev-chembioeng-060816-101555.

[4] M. Ali-ud-din Khan, Muhammad Fahim Uddin e Navarun Gupta. "Seven V's of Big Data understanding Big Data to extract value". Em: *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*. Bridgeport, CT, USA: IEEE, 2014. DOI: 10.1109/ASEEZone1.2014.6820689.

Aprendizado de Máquina - Definições



- “O *Aprendizado de Máquina* é o campo de estudo que fornece aos computadores a habilidade de aprender sem serem explicitamente programados^[5];”
- “Diz-se que um programa de computador aprende com a experiência E em relação a alguma tarefa T e alguma medida de desempenho P , se seu desempenho em T , medido por P , melhora com a experiência E ^[6].”

[5] Arthur L Samuel. "Some studies in machine learning using the game of checkers". Em: *IBM Journal of research and development* 3.3 (1959) pp. 210–229.

[6] Tom M. Mitchell. *Machine Learning*. 1^a ed. 49220. New York: McGraw-Hill Science/Engineering/Math, 1997. ISBN: 978-0-07-042807-2.

Aprendizado de Máquina - Definições

- “O *Aprendizado de Máquina* é essencialmente uma forma de estatística aplicada, com ênfase crescente no uso de computadores para estimar estatisticamente funções complicadas e menor ênfase na obtenção de intervalos de confiança em torno dessas funções^[7]”.



[7] Ian Goodfellow, Yoshua Bengio e Aaron Courville. *Deep Learning*.
deeplearningbook.org/.

Aprendizado de Máquina vs. Inteligência Artificial

Aprendizado de Máquina é sinônimo de Inteligência Artificial?

- Na verdade, o conceito de Inteligência Artificial (IA) é mais amplo e inclui^[8]:
 - raciocínio simbólico e prova de teoremas;
 - robótica;
 - visão computacional;
 - sistemas especialistas;
 - aprendizado de máquina;
 - etc...
- Uma definição famosa e abrangente de IA é:

“A Inteligência Artificial é o estudo de como fazer computadores realizar tarefas nas quais, no momento, pessoas são melhores^[9]”.
- O Aprendizado de Máquina pode ser visto como um dos mais importantes campos da Inteligência Artificial.

[8] Stuart J. Russell e Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3^a ed. Upper Saddle River: Prentice Hall, 2009. ISBN: 978-0-13-604259-4.

[9] Elaine Rich. *Artificial Intelligence*. New York: McGraw-Hill, 1983.

Modelos e Algoritmos

- Um **padrão** é qualquer característica regular de um conjunto de dados.
- O objetivo do Aprendizado de Máquina é processar dados para reconhecer padrões que possam ser usados para efetuar **previsões** em outros dados que sigam os mesmos padrões.
- Um **modelo** é uma representação de um ou mais desses padrões. Modelos podem estar expressos em diferentes formas, como veremos a seguir.
- Um **algoritmo** é o conjunto de passos a serem executados pelo computador para reproduzir os padrões capturados pelo modelo.

Modelos e Algoritmos

- O processo de aplicar o algoritmo a um conjunto de dados para gerar o modelo é conhecido como **treino** (em inglês, *train* ou *fit*). É a etapa em que o modelo **aprende** os padrões.
- **Parâmetros** são as variáveis da estrutura matemática do modelo que são ajustadas na etapa de treino para reproduzir os padrões desejados. As variáveis que *não* são ajustadas no treino são chamadas de **hiperparâmetros**.
- Após o treinamento do modelo, pode-se usar o algoritmo para fazer previsões em dados que ainda não lhe foram apresentados.

Tipos de algoritmos

Como se aprende?

- **Supervisionados:** o conjunto de treinamento possui a solução desejada. O algoritmo observa a solução e dela tenta encontrar um padrão para generalizá-la para outros conjuntos de dados.
- **Não-supervisionados:** o conjunto de treinamento não possui a solução desejada. O algoritmo tenta aprender o padrão sem um “professor”.
- **Semi-supervisionados:** alguns dados de treinamento possuem a solução desejada e outros não.
- **Por reforço:** o algoritmo observa o ambiente, seleciona e performa ações e ganha recompensas (ou sofre penalidades) em retorno. Com base nessas recompensas ou penalidades, o algoritmo decide o melhor caminho a seguir. Exemplo: <https://youtu.be/V1eYniJ0Rnk>.

Tipos de modelos

O que é aprendido?

Vários tipos de modelos podem ser desenvolvidos, a depender do problema que se deseja resolver:

- **Regressão:** predição de valores de variáveis contínuas. Usa algoritmos supervisionados.
- **Classificação:** predição de valores de variáveis discretas. Usa algoritmos supervisionados.
- **Clusterização:** separação de um conjunto de dados em diferentes grupos (*clusters*), cada um correspondendo a um ou mais padrões distintos. Usa algoritmos não-supervisionados.
- **Detecção de anomalias:** identificação de pontos com padrões distintos do padrão considerado como “normal”. Pode usar algoritmos supervisionados e não-supervisionados.

Modelos e Algoritmos

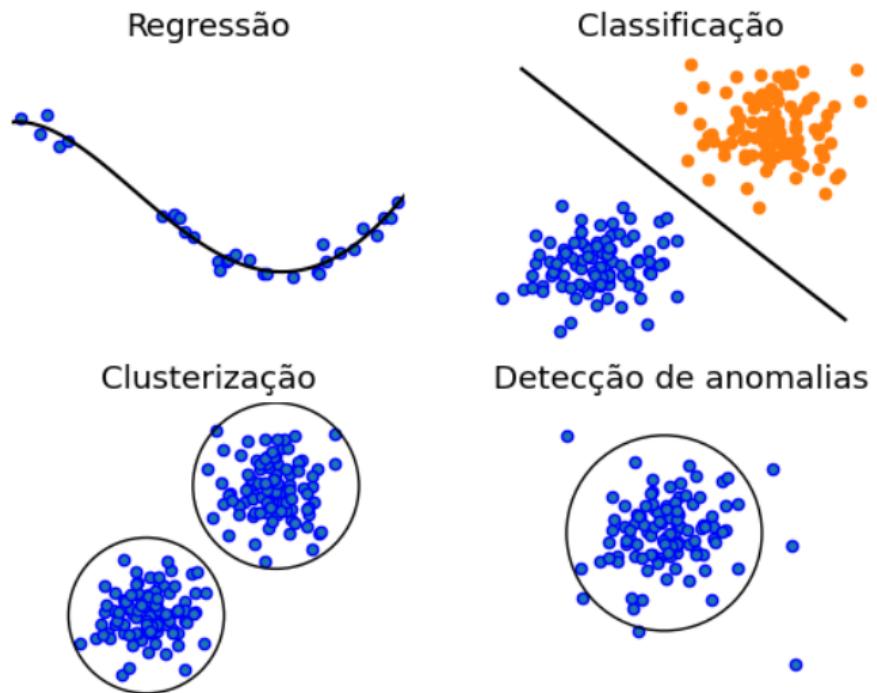


Figura 8: Representação visual 2D dos diversos tipos de modelos.

Dificuldades e problemas que podem ser enfrentados em um projeto de AM

- dados de treinamento não representativos e/ou de má qualidade:
 - se a amostra é muito pequena, pode ocorrer *ruído de amostragem*;
 - mesmo se a amostra é grande, pode ocorrer *viés de amostragem*;
 - se os dados forem medidos por sensores, pode ocorrer *ruído de medição*;
 - pode haver muitos dados faltantes;
 - algumas variáveis incluídas nos dados podem ser irrelevantes para o problema.
- problemas relacionados à modelagem:
 - métricas de desempenho mal-definidas;
 - subajuste dos dados;
 - sobreajuste dos dados (o mais ardiloso dos problemas!).

Subajuste e sobreajuste

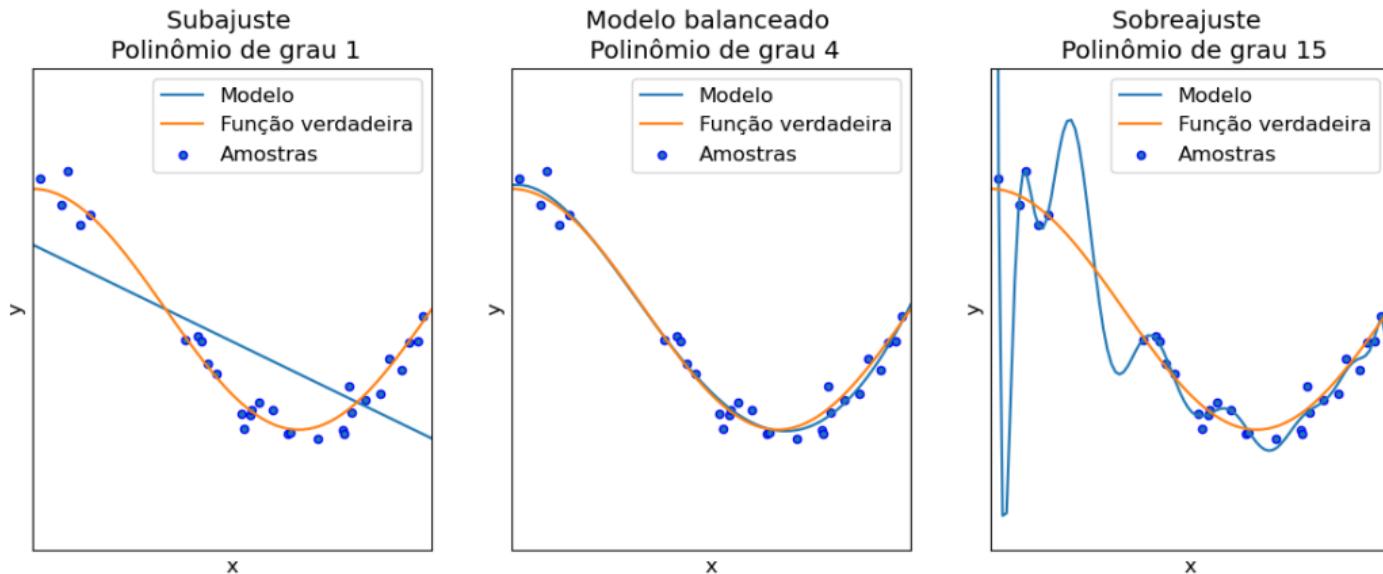


Figura 9: Exemplo, em um problema de regressão, dos fenômenos de subajuste (quando o modelo é simples demais para o padrão a ser aprendido) e sobreajuste (quando o modelo é complexo/flexível demais para o padrão a ser aprendido e acaba interpretando ruído como padrão). Função verdadeira: $f(x) = \cos(3\pi x/2)$. Adaptado de <http://scikit-learn.org>.

A Quarta Revolução Industrial - Indústria 4.0

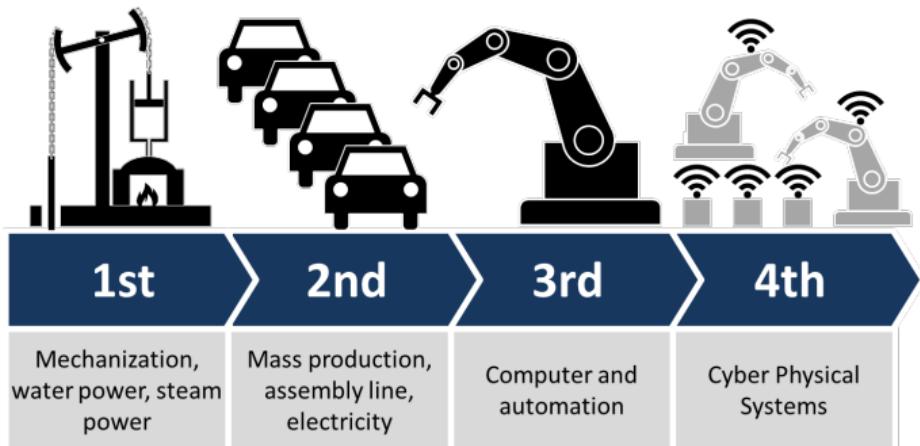
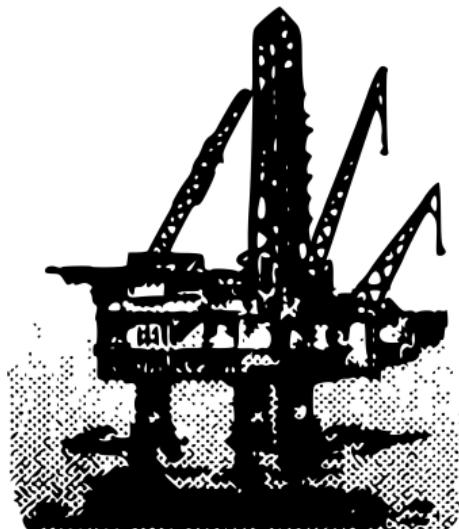


Figura 10: Evolução das revoluções industriais (em http://commons.wikimedia.org/wiki/File:Industry_4.0.png por Cristoph Roser).

A Quarta Revolução Industrial - Indústria 4.0

Pilares da Indústria 4.0^{[10][11]}

- Ciência de Dados e Aprendizado de Máquina;
- *Big Data*;
- Computação de alto desempenho;
- Computação em nuvem;
- Internet das coisas;
- Cibersegurança;
- Gêmeos digitais;
- Integração de sistemas e processos;
- Manufatura aditiva;
- Realidade aumentada;
- Sistemas autônomos;
- etc, etc, etc!



[10] Yang Lu. "Industry 4.0: A survey on technologies, applications and open research issues". Em: *Journal of Industrial Information Integration* 6 (2017), pp. 1–10. DOI: 10.1016/j.jiii.2017.04.005.

[11] Ercan Oztemel e Samet Gursoy. "Literature review of Industry 4.0 and related technologies". Em: *Journal of Intelligent Manufacturing* 31.1 (2020), pp. 127–182. DOI: 10.1007/s10845-018-1433-8.

A Quarta Revolução Industrial - Indústria 4.0

Aplicações da Ciência de Dados no âmbito da Indústria Química 4.0

- monitoramento de processos/detecção e diagnóstico de falhas^{[12][13]};
- sensores virtuais^[14];
- cibersegurança de processos^[15];
- desenvolvimento de novas moléculas, como fármacos ou catalisadores^{[16][17]};
- etc, etc, etc!

[12] Q. Peter He e Jin Wang. "Statistical process monitoring as a big data analytics tool for smart manufacturing". Em: *Journal of Process Control* 67 (2018), pp. 35–43. DOI: [10.1016/j.jprocont.2017.06.012](https://doi.org/10.1016/j.jprocont.2017.06.012).

[13] Yidan Shu et al. "Abnormal situation management: Challenges and opportunities in the big data era". Em: *Computers & Chemical Engineering* 91 (2016), pp. 104–113. DOI: [10.1016/j.compchemeng.2016.04.011](https://doi.org/10.1016/j.compchemeng.2016.04.011).

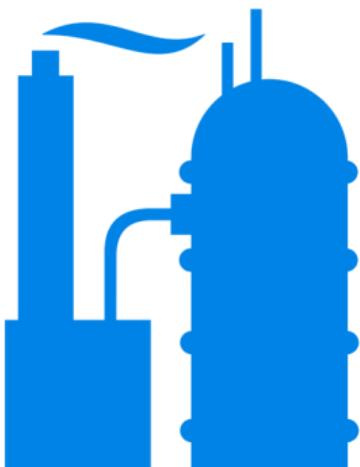
[14] Salvatore Graziani e Maria Gabriella Xibilia. "Deep Learning for Soft Sensor Design". Em: *Development and Analysis of Deep Learning Architectures*. Ed. por Witold Pedrycz e Shyi-Ming Chen. Springer International Publishing, 2020, pp. 31–59. ISBN: 978-3-030-31764-5. DOI: [10.1007/978-3-030-31764-5_2](https://doi.org/10.1007/978-3-030-31764-5_2).

[15] Scarlett Chen, Zhe Wu e Panagiotis D. Christofides. "Cyber-attack detection and resilient operation of nonlinear processes under economic model predictive control". Em: *Computers & Chemical Engineering* 136 (2020). DOI: [10.1016/j.compchemeng.2020.106806](https://doi.org/10.1016/j.compchemeng.2020.106806).

[16] Keith T. Butler et al. "Machine learning for molecular and materials science". Em: *Nature* 559.7715 (2018), pp. 547–555. DOI: [10.1038/s41586-018-0337-2](https://doi.org/10.1038/s41586-018-0337-2).

[17] Clémence Réda, Emilie Kaufmann e Andrée Delahaye-Duriez. "Machine learning applications in drug development". Em: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 241–252. DOI: [10.1016/j.csbj.2019.12.006](https://doi.org/10.1016/j.csbj.2019.12.006).

Dados de processo



Contextualização

- Plantas de processo possuem grandes volumes de dados históricos armazenados em bancos de dados, obtidos por meio de sensores que medem milhares de variáveis a frequências da ordem de segundos.
- Apesar de não ser uma prática generalizada, a exploração desses dados, no longo prazo, pode se mostrar um componente crítico da operação de um processo industrial.

Dados de processo

Características dos dados de processo^[18]

- São massivos em volume, da ordem de GB e TB.
- Resultam de medições altamente correlacionadas, apresentando muitas variáveis colineares (muitas vezes por conta da ação de controladores).
- Possuem baixa razão sinal/ruído, o que implica em baixa quantidade de informação por variável individual, tornando clara a importância de uma análise multivariada.
- Padrões associados à medições e incertezas podem variar no tempo (comportamento dinâmico).
- Podem apresentar muitos dados faltantes.

[18] T. Kourtzi. "Process analysis and abnormal situation detection: from theory to practice". Em: *IEEE Control Systems* 22 (2002), pp. 10–25. DOI: 10.1109/MCS.2002.1035214.

Dados de processo

Pré-tratamento

- Devido às características descritas, muitas vezes é difícil ajustar modelos de aprendizado a dados de processo.
- Etapas de pré-tratamento de dados mostram-se essenciais em grande parte dos casos práticos^[19].
- Exemplos de procedimentos de pré-tratamento:
 - preenchimento de valores faltantes;
 - detecção de *outliers*;
 - filtração do ruído;
 - normalização.

[19] Shu Xu et al. "Data cleaning in the process industries". Em: *Reviews in Chemical Engineering* 31.5 (2015). doi: 10.1515/revce-2015-0022.

Dados de processo

Evolução histórica

- Durante praticamente todo o século XX, a estratégia predominante para análise dos dados de processo foi o monitoramento individual de algumas variáveis importantes classificadas como *variáveis de qualidade*.
- A partir da década de 1990, migrou-se para uma estratégia multivariada baseada na aplicação de modelos estatísticos capazes de efetuar *redução de dimensionalidade* e *extração de variáveis latentes* de um conjunto grande de medições da planta, como PCA (*Principal Component Analysis*) e PLS (*Partial Least Squares*)^{[20][21]}.

[20] J.F. MacGregor et al. "Latent Variable Models and Big Data in the Process Industries". Em: *IFAC-PapersOnLine* 48.8 (2015), pp. 520–524. DOI: [10.1016/j.ifacol.2015.09.020](https://doi.org/10.1016/j.ifacol.2015.09.020).

[21] Ivan Miletic et al. "An industrial perspective on implementing on-line applications of multivariate statistics". Em: *Journal of Process Control* 14.8 (2004), pp. 821–836. DOI: [10.1016/j.jprocont.2004.02.001](https://doi.org/10.1016/j.jprocont.2004.02.001).

Dados de processo

Evolução histórica

- A partir de 2000, estratégias do domínio da Ciência de Dados e Aprendizado de Máquina passaram a ganhar terreno^[22].
- Mais recentemente, vêm surgindo abordagens livres de modelos parametrizados, baseadas totalmente em padrões e/ou assinaturas estatísticas dos dados.
- Essa é uma linha que vem sendo seguida com bastante ênfase pelo grupo LMSCP/Engepol, como demonstram alguns exemplos a seguir.

[22] S. Joe Qin e Leo H. Chiang. "Advances and opportunities in machine learning for process data analytics". Em: *Computers & Chemical Engineering* 126 (2019), pp. 465-473. DOI: 10.1016/j.compchemeng.2019.04.003.

Exemplo - Espectro de variâncias

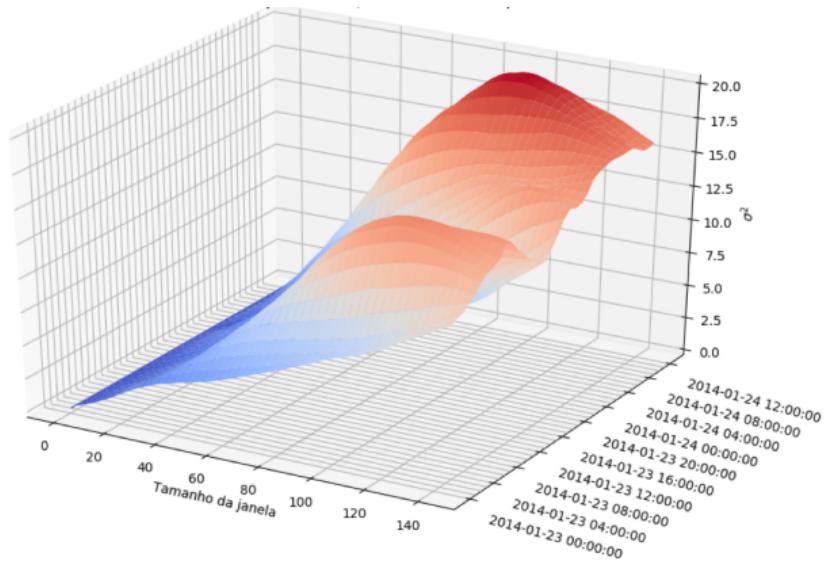


Figura 11: Dinâmica do espectro de variâncias^[23] de uma variável medida em poço de petróleo.

[23] Thiago Feital e José Carlos Pinto. "Use of variance spectra for in-line validation of process measurements in continuous processes". Em: *The Canadian Journal of Chemical Engineering* 93 (2015), pp. 1426–1437. DOI: 10.1002/cjce.22219.

Exemplo - Análise de recorrências

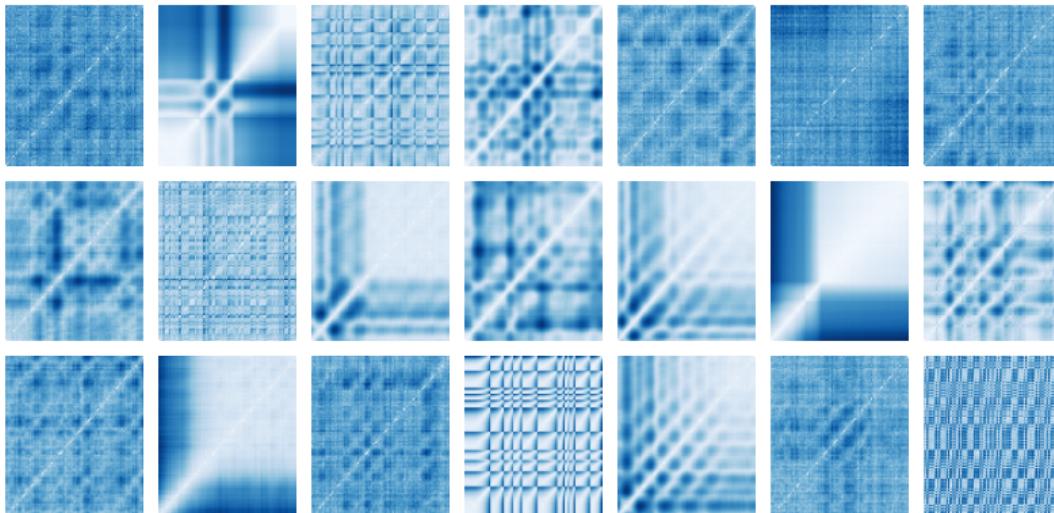


Figura 12: Matrizes de distâncias^[24] para cada um dos cenários de operação descritos no benchmark Tennessee Eastman^[25].

[24] Fernando Elias de Freitas Fadel. "Uma Avaliação Crítica Sobre Técnicas Baseadas em PCA para Detecção de Falhas em Processos da Indústria Química". Diss. de mestrado. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2018.

[25] L. H. Chiang, E. L. Russell e R. D. Braatz. *Fault Detection and Diagnosis in Industrial Systems*. London: Springer, 2001. ISBN: 978-1-85233-327-0.

Quem somos nós?

- Sou pesquisador de doutorado no Programa de Engenharia Química da COPPE sob orientação de José Carlos Pinto, realizando trabalhos no âmbito do **Laboratório de Modelagem, Simulação e Controle de Processos (LMSCP)** e **Laboratório de Engenharia de Polimerização (EngePol)**.



- Nossos interesses, dentre outros, incluem:
 - desenvolvimento e implementação de procedimentos para monitoramento de processos em tempo real;
 - desenvolvimento e aplicação a processos de técnicas que não assumam hipóteses prévias a respeito das características estatísticas dos dados.
- Estamos abertos e entusiasmados para conversas e parcerias sobre o tema!! :D

Quem somos nós?



Figura 13: Grupo LMSCP/EngePol na festa de fim de ano de 2019 :) (mal sabíamos o que estava por vir em 2020...)

Obrigado pela atenção!



@escolapilotopeq
@engepolgrupo