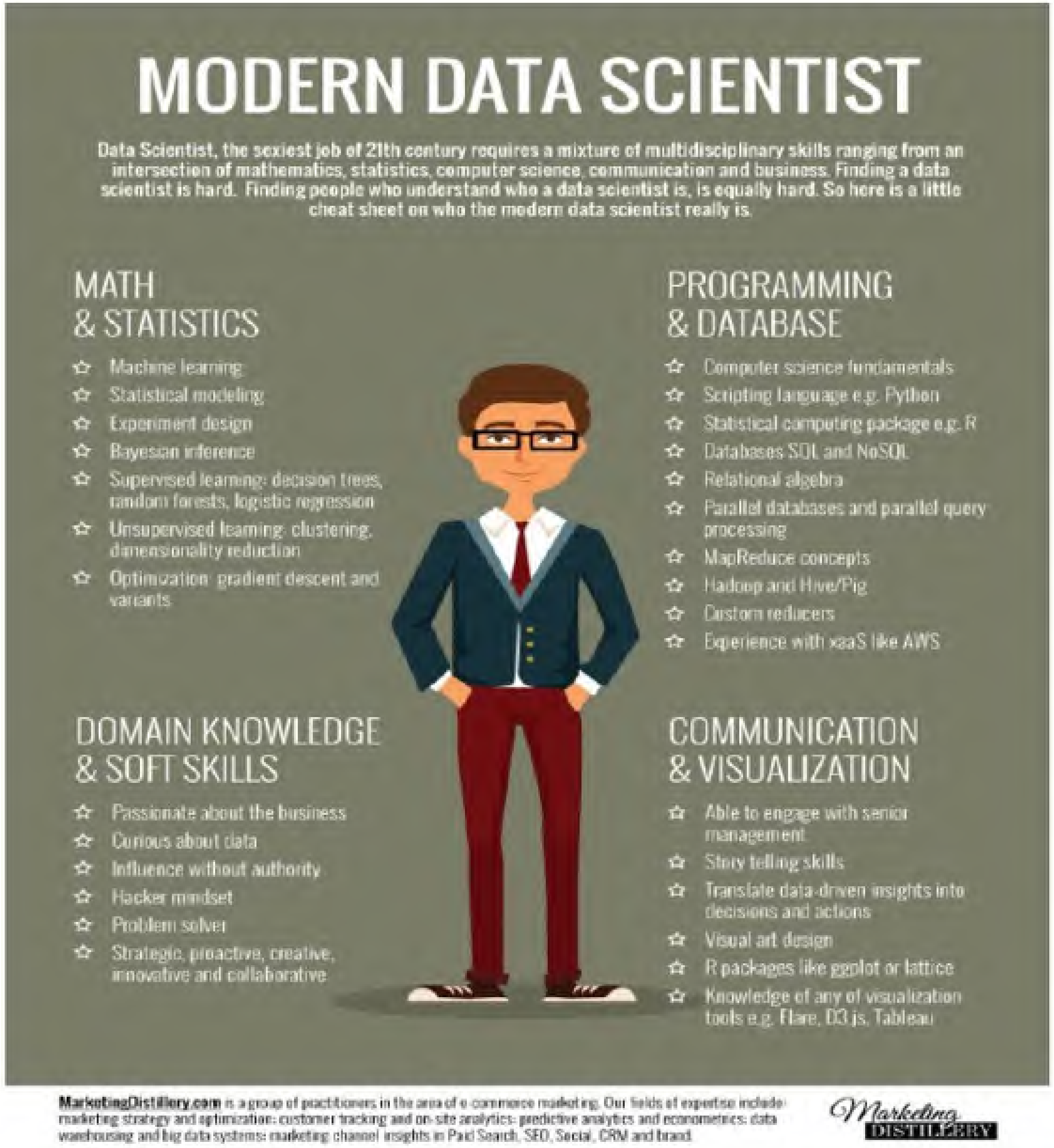
《数据科学R与Pyhon实践》

顾立平

来自2012年的“最性感”的未来职业



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

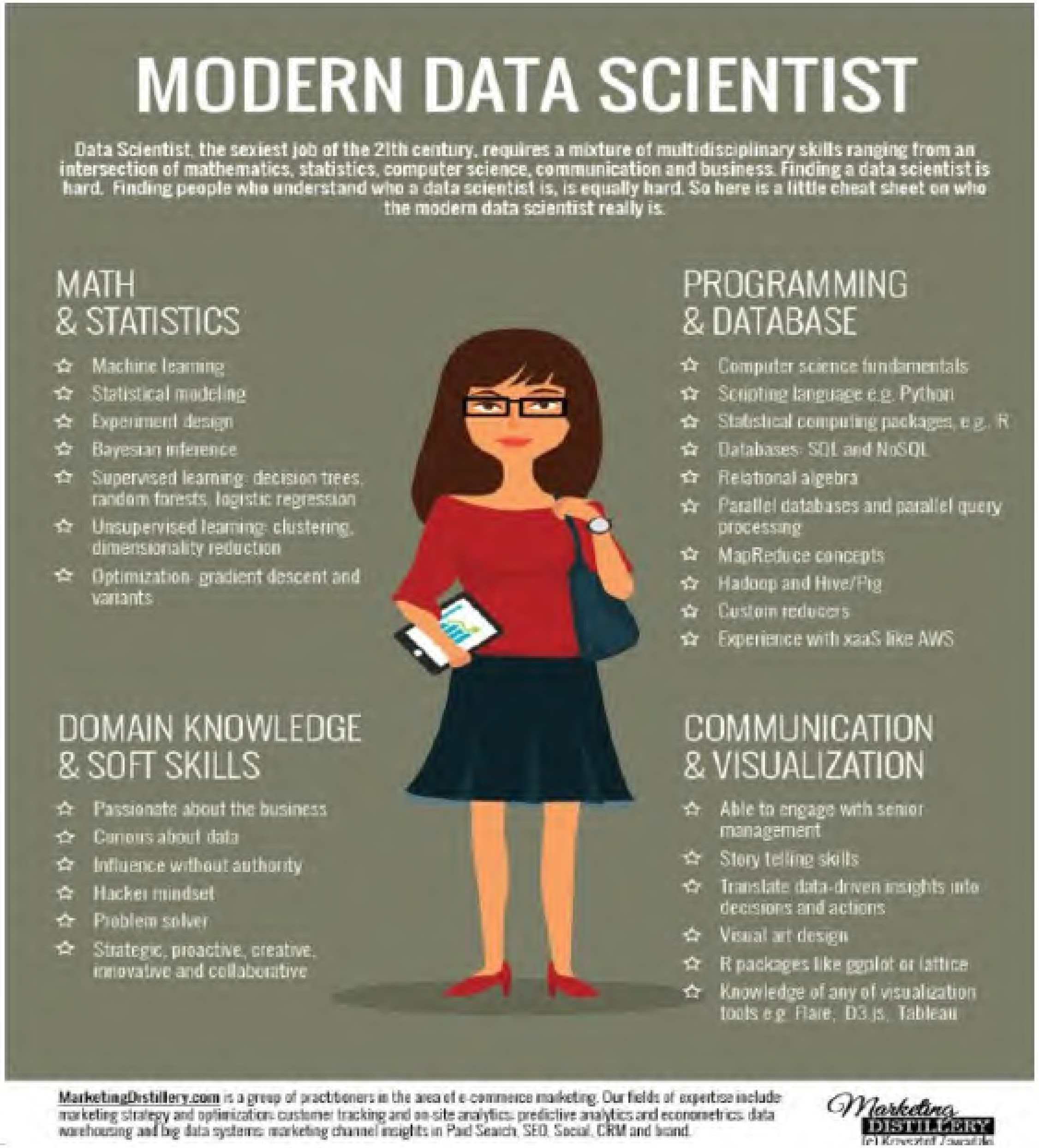
DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

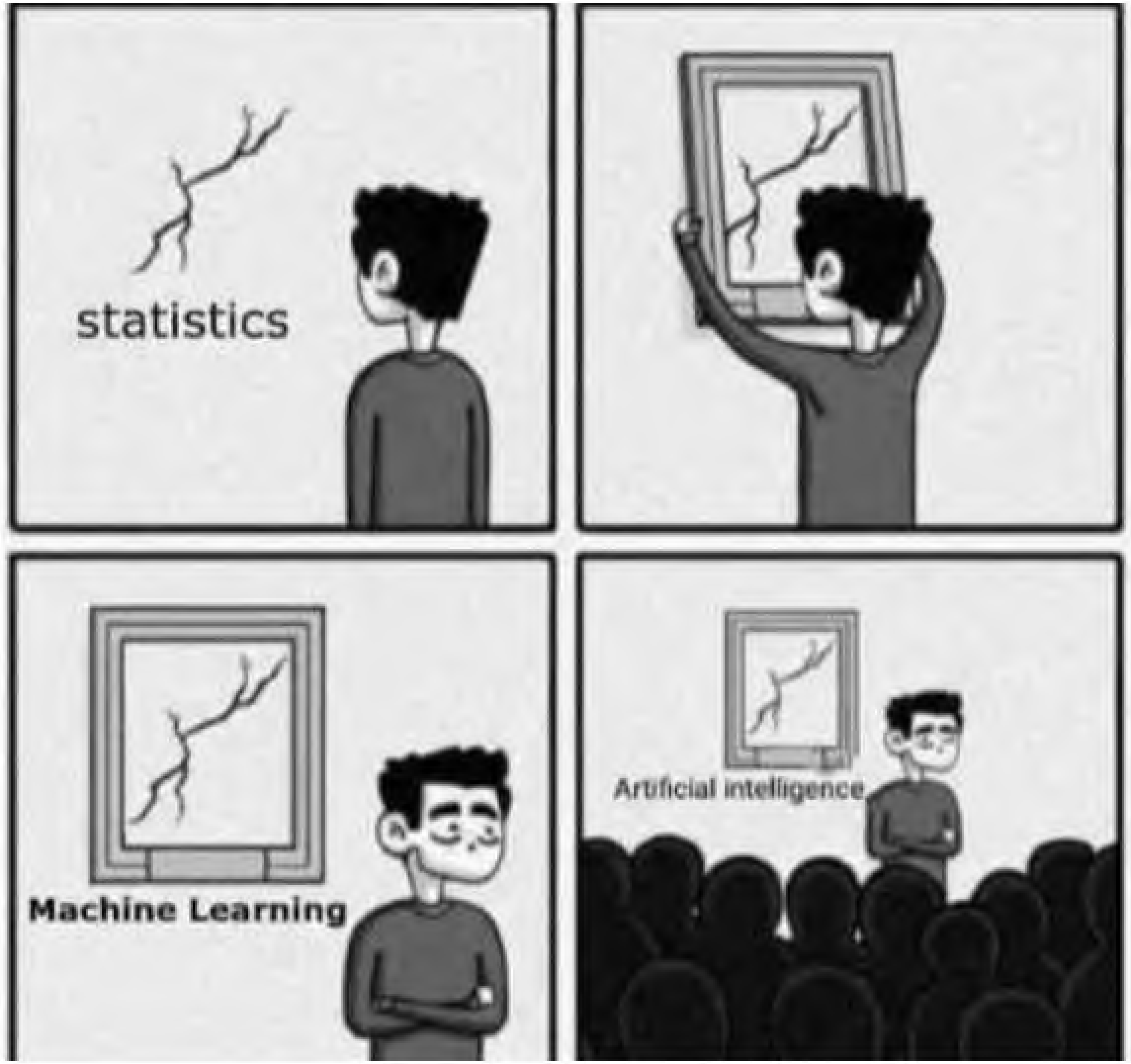
- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

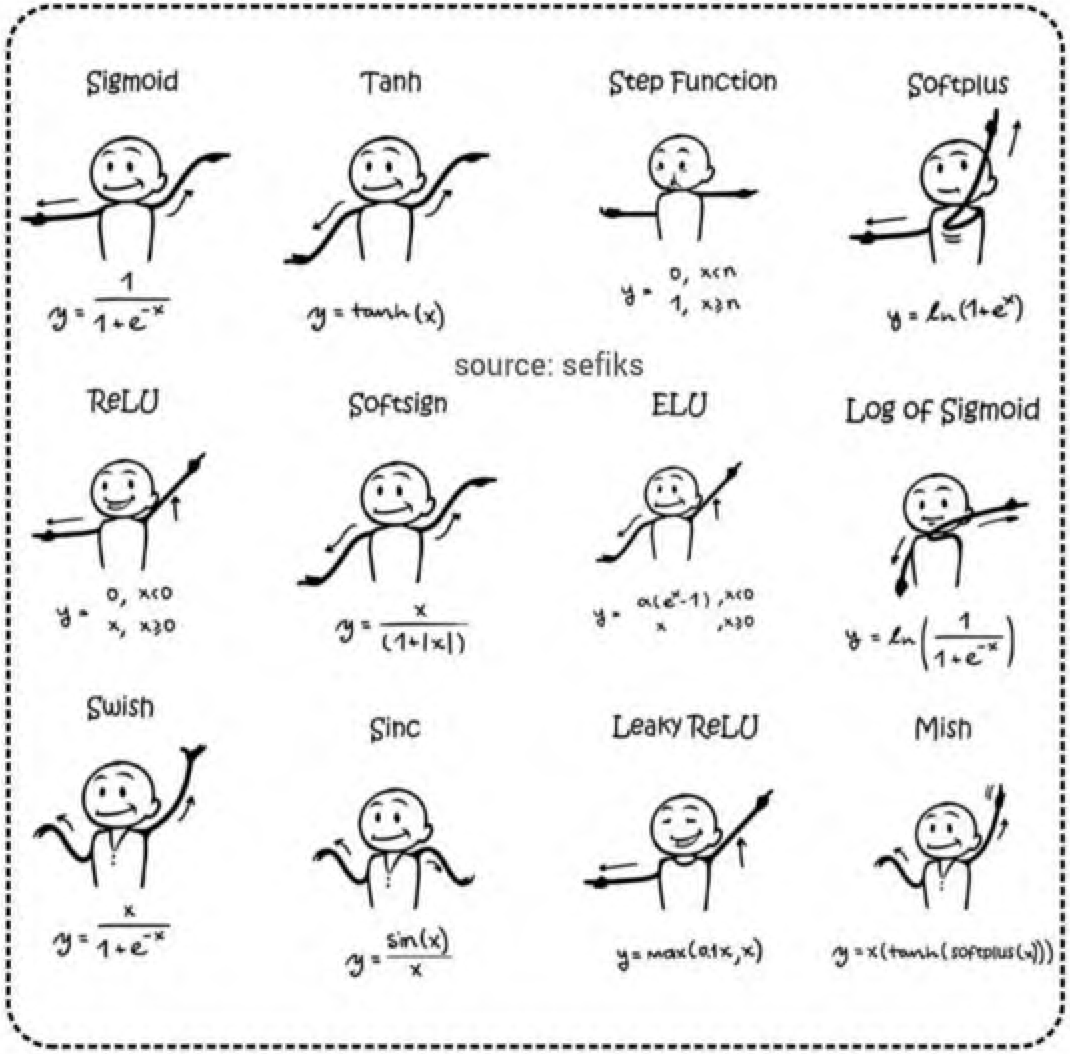
COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

教学与实践

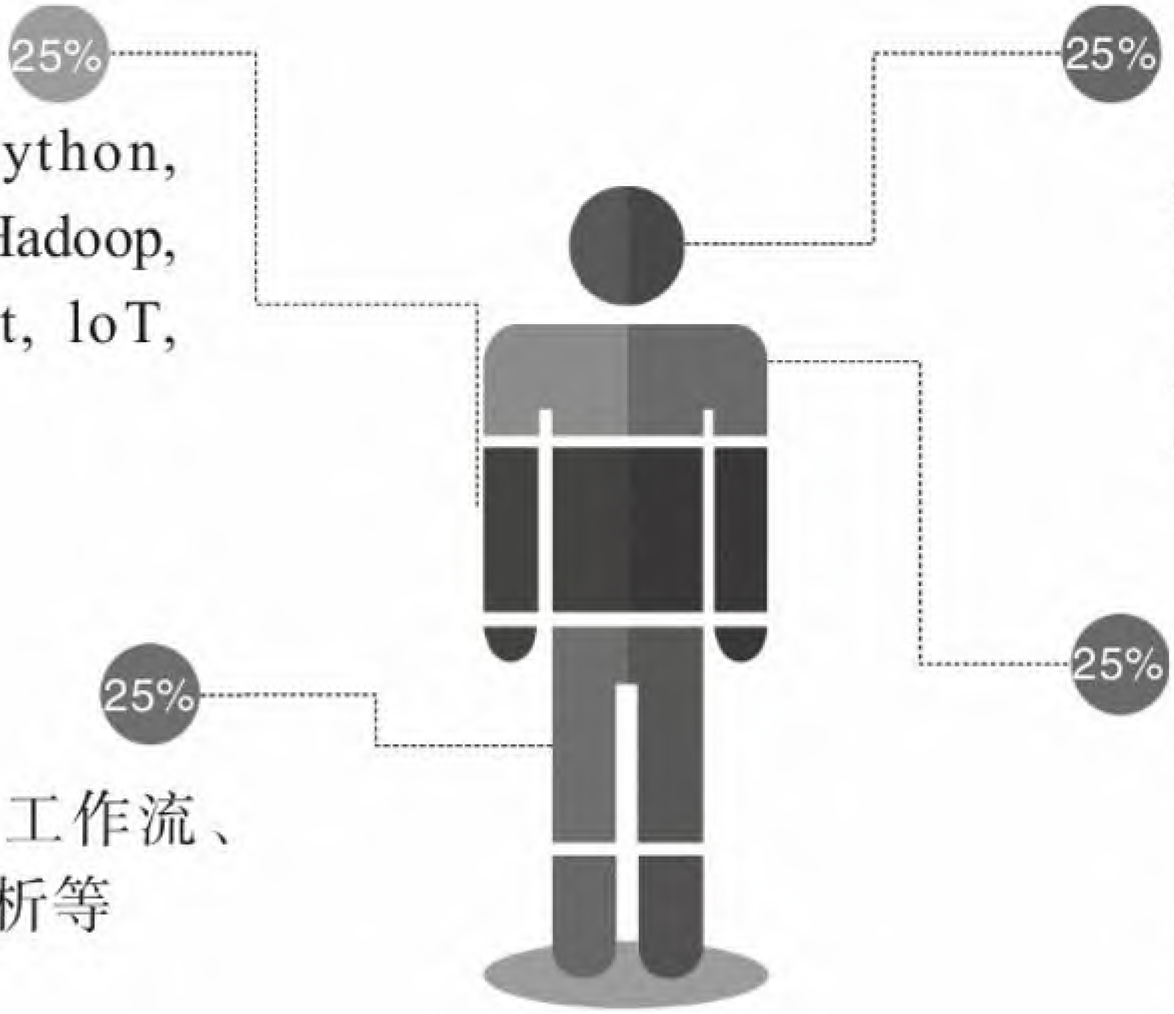




source: sefiks

难以企及和数据科学家

如今就业要求...



技术

JMP, SAS, R, Python, Perl, Excel, SQL, Hadoop, Java, JavaScript, loT, Event-Stream

方法论

数学和统计理论、定量和定性方法、试验（例如A/B 测试）、六西格玛质量管理方法、质量提升

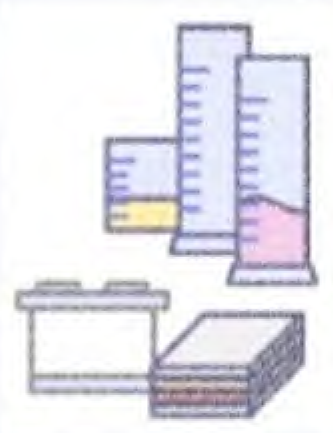
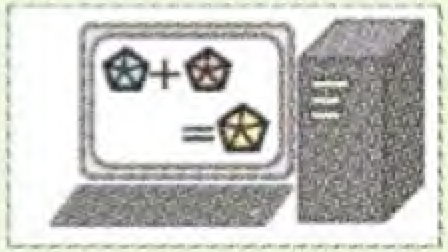
领域

专业知识包括：工作流、运营和可操作分析等

软技能

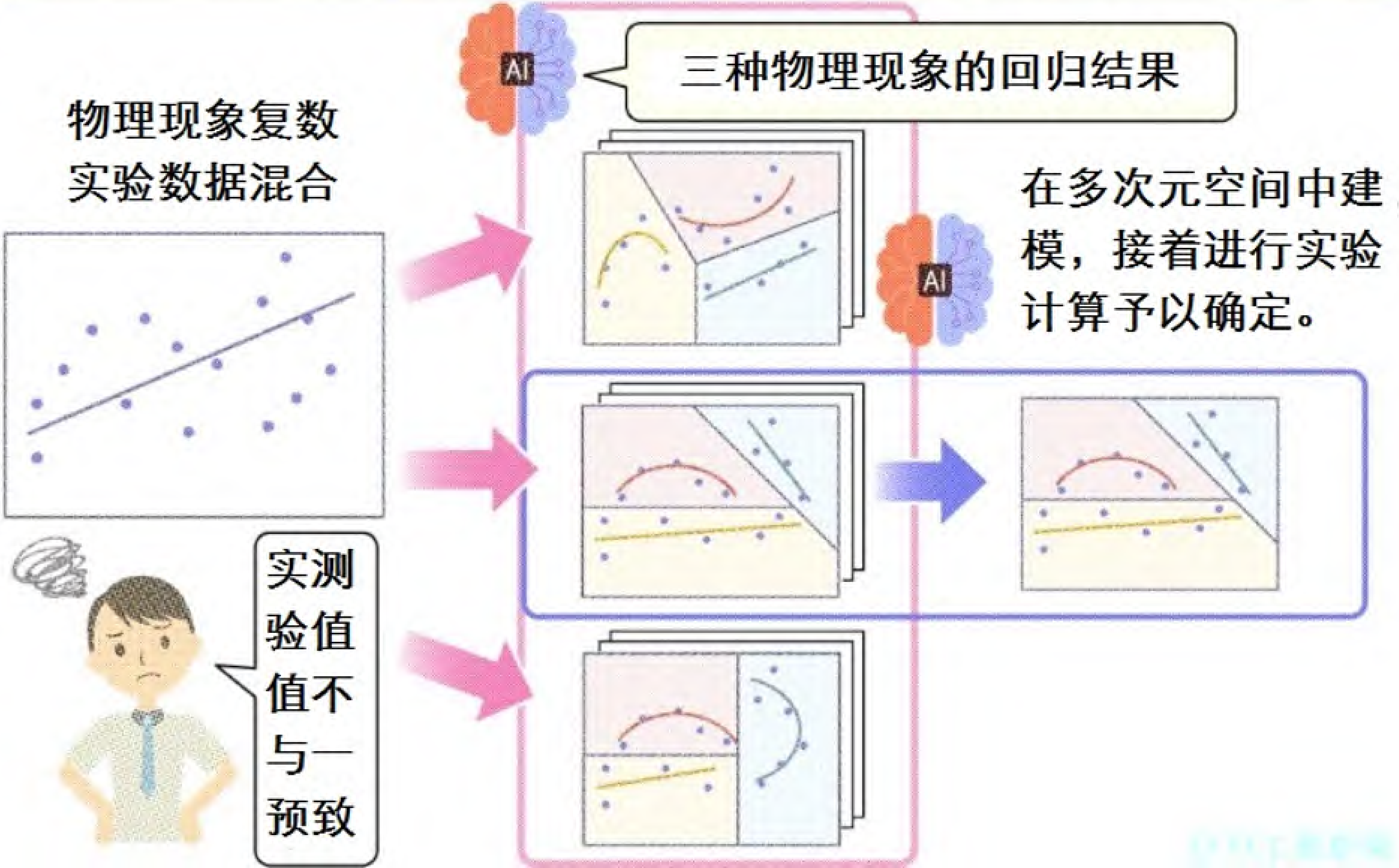
影响力，批判性思维、沟通交流能力，系统性思维，设计思维，可视化思维，设计

第四科学（理论、实验、计算、数据）

	演绎（理论、原理的预测）	归纳（数据的解释与探索）
人类社会	理论科学 理论构建 脑力竞争 	实验科学 经验解释 团队组织 
数字空间	计算科学 自动化 博弈胜负 	数据科学 人工智能 数据质量

第五科学（AI+Data+Brain）

第五科学 （异种混合学习）



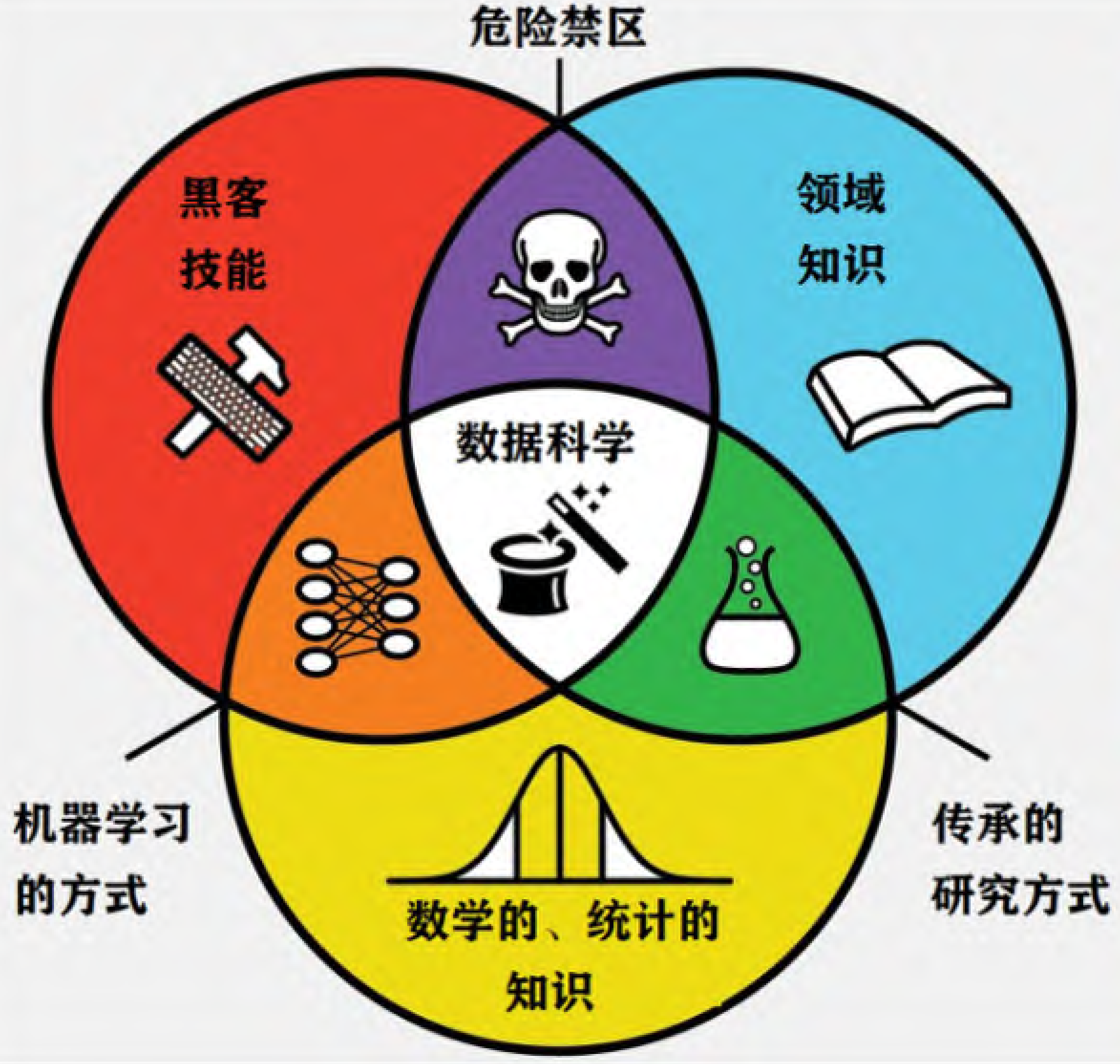
物理现象复数实验数据混合

三种物理现象的回归结果

在多次元空间中建模，接着进行实验计算予以确定。

实验值与预测值不一致

数据科学R与Python实践所需技能



数据科学因为跨学科性质，强调交叉能力：黑客技能、数学和统计知识以及学科专业的领域知识。

黑客技能有助于处理那些必须获取、清理和操纵的大量电子数据。

数学和统计知识有助于数据科学家选择适当的方法和工具，以便从数据中获得独到见解。

领域知识有助于探索有意义的研究问题及其假设建立，以及验证和解释随后的实验结果。

传承的研究方式有助于在数学和统计知识与领域知识之间进行新的方法遴选、使用和评估。

机器学习的方式有助于结合黑客技能与数学和统计知识，但是缺乏探索有意义的科学问题的科学动机。

危险禁区！如果没有严谨的科学精神和方法，黑客技能与领域知识相结合，很可能会导致不正确的分析。

谢谢您

顾立平

邮箱gulp@mail.las.ac.cn

手机/微信15801137067