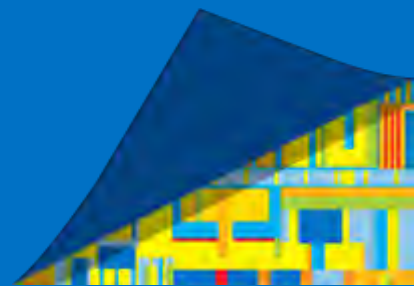
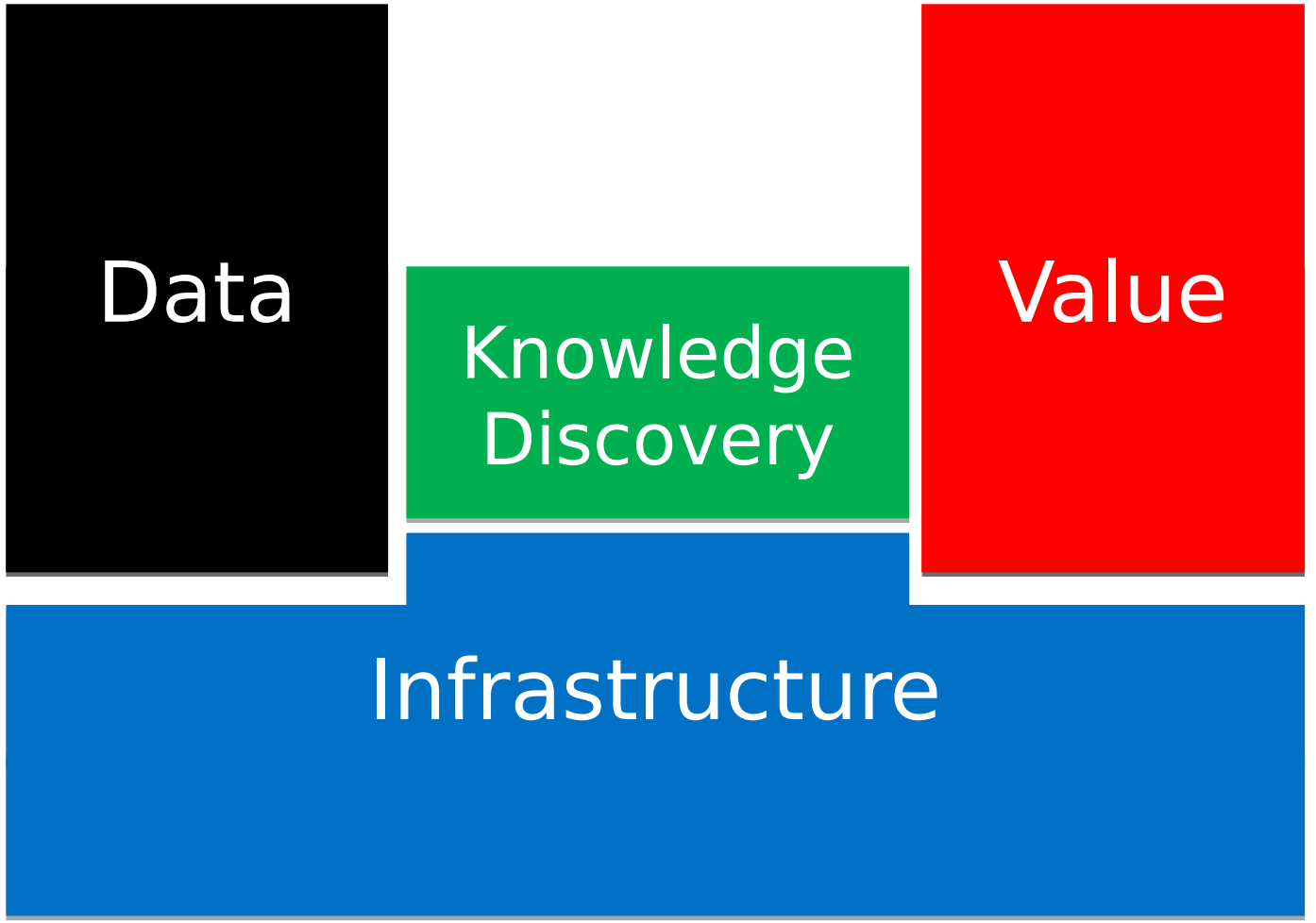
大数据分析师的卓越之道

吴甘沙

英特尔中国研究院



数据分析的典型场景



基础设施已经改朝换代

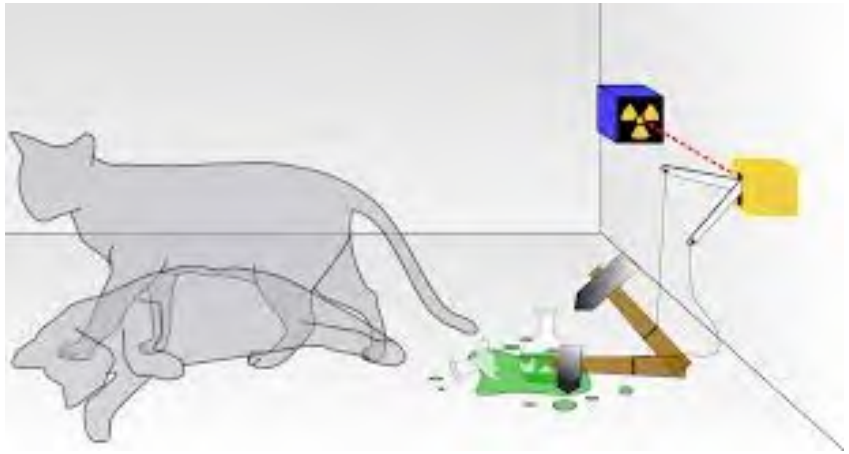
分析师也需要与时俱进

改变思维方式

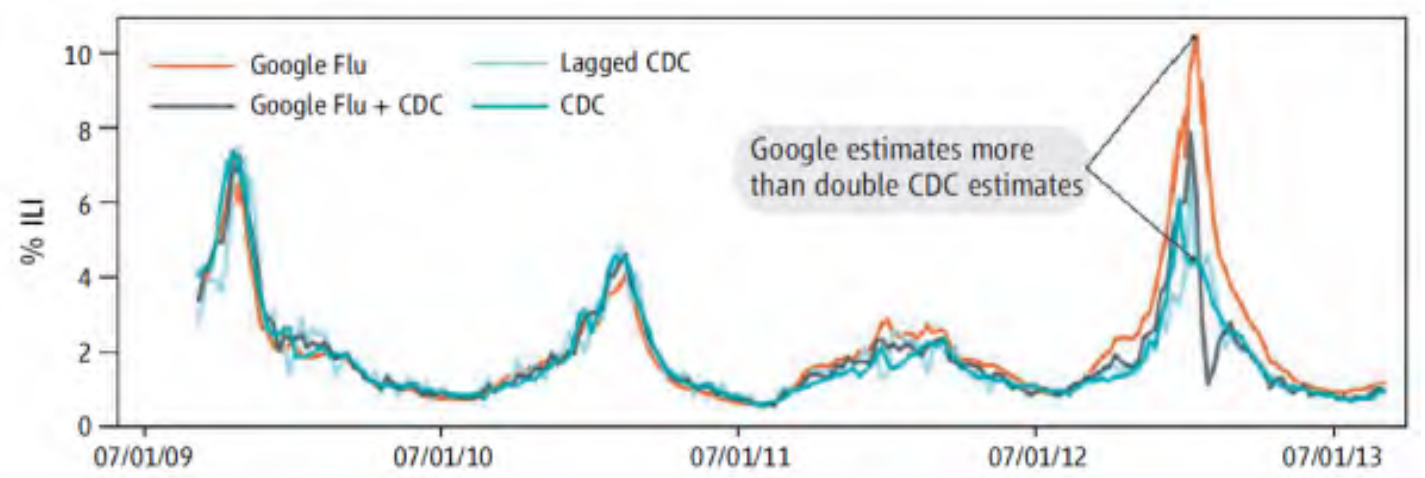
提高技术素养

丰富分析能力

新的世界观：不确定的世界

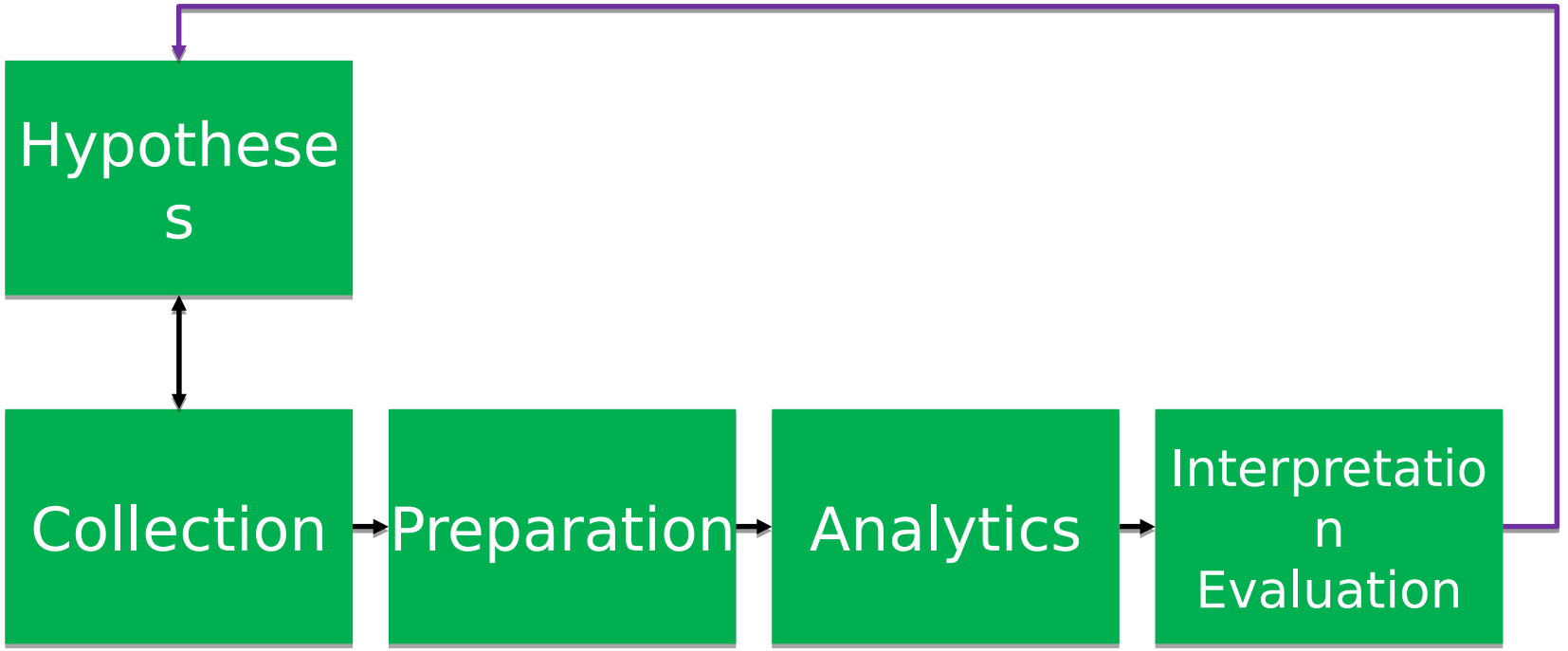


大数据的测不准

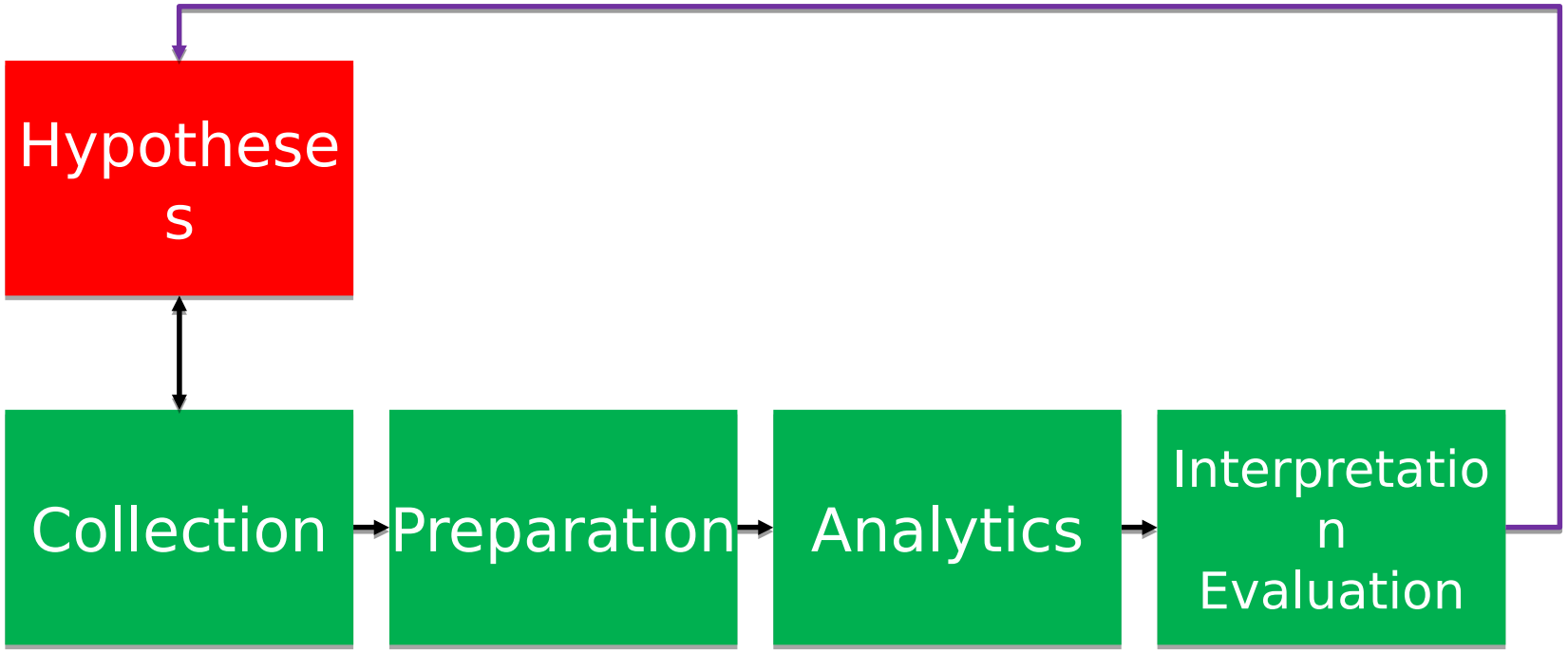


《自然》：测不准　　《科学》：大数据傲慢

数据分析方法论的升级


Hypotheses
Collection
Preparation
Analytics
Interpretation
Evaluation

数据分析方法论的升级


Hypothese
s
Collection
Preparation
Analytics
Interpretatio
n
Evaluation

Hypotheses

机械地发掘相关性和假设

直觉，拿侦探小说练手

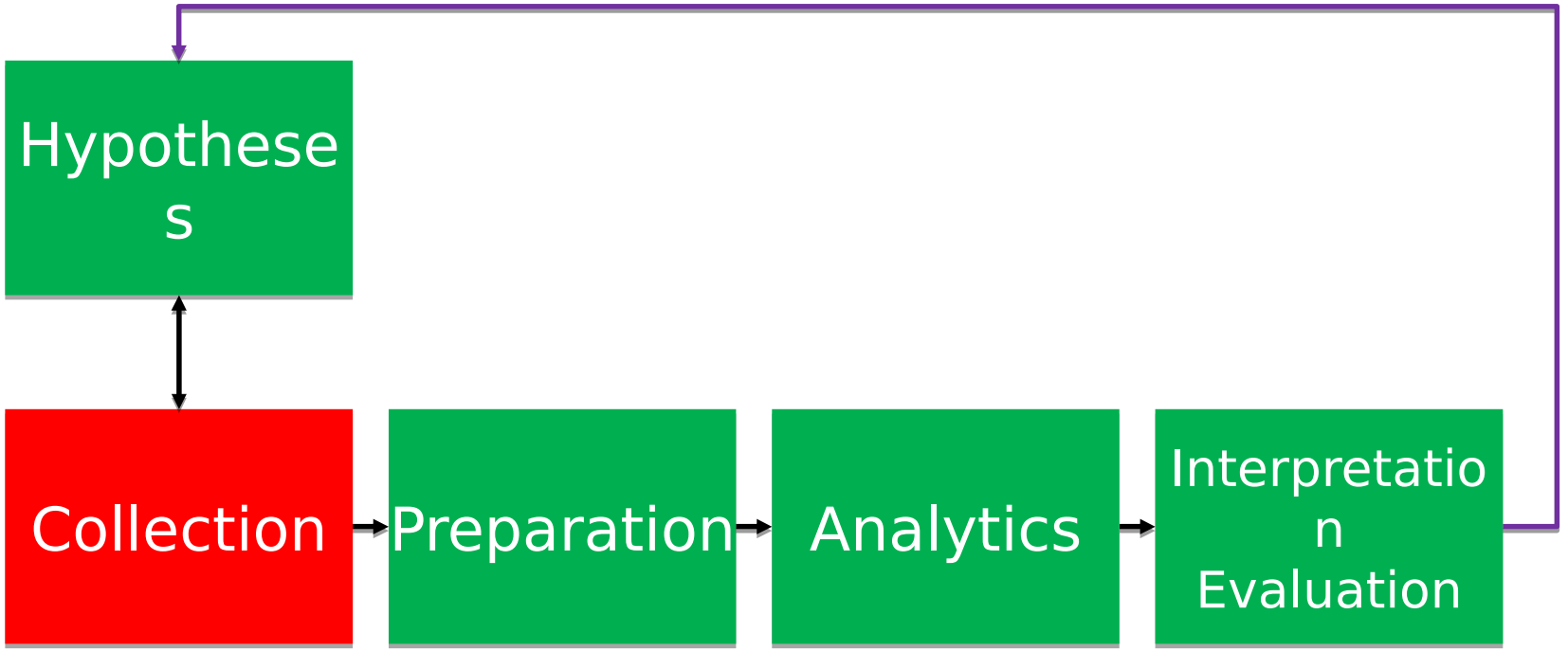
阅读广泛涉猎

跨界思维碰撞

融入业务部门

- 防止数据采集与分析、业务与数据分析的脱节

数据分析方法论的升级


Hypotheses
Collection
Preparation
Analytics
Interpretation
Evaluation

数据！数据！数据！

n=All !

Enterprise Data Warehouse ␣ Enterprise Data Hub/Data Lake

External data sources

Structured ␣ semi-structured ␣ unstructured

- Log analysis
- Text analysis
- Image/video
- Data with geo and temporal tags
- Networks and graphs

数据？数据？数据？

n=All ?

- More data vs. sampling
- “Raw data” is an oxymoron
- Signals and noises
- Sampling bias

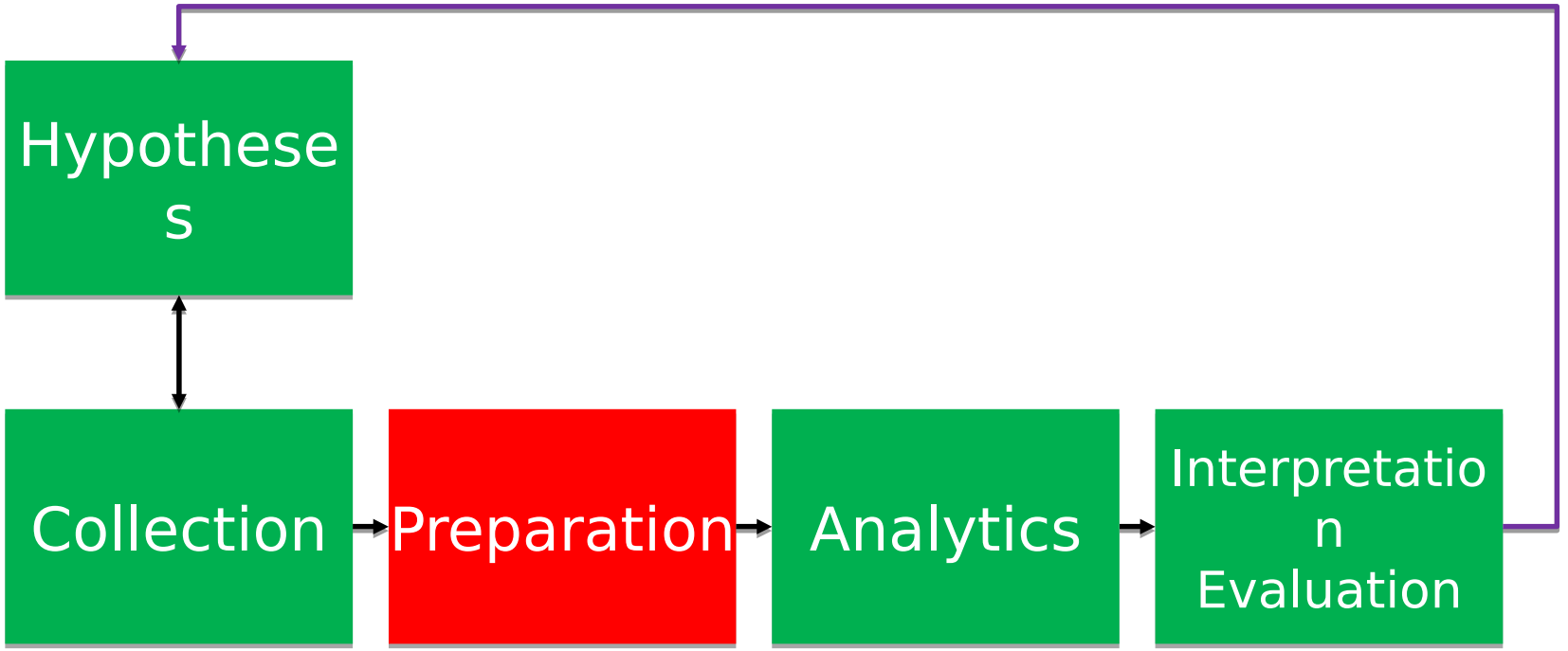
Data exchange and sharing

- Data rights, data pricing

Data lifecycle management

- Provenance capture, representation, and querying
- Sometimes data are not assets, but costs

数据分析方法论的升级


Hypotheses
Collection
Preparation
Analytics
Interpretation
Evaluation

数据质量：重中之重

Noisy, biased and polluted data are unavoidable

- Goal: models = components for noise + relatively complex models for signal

Cleansing, validation, …

- Can it start with a small subset? Can the process be automated?
- Work together with visualization, machine learning

Curation, Wrangling, …

- Automated learning to discover structure, resolve entities, and transform data

数据表示

Reduce compute and communication complexity

- Sparse, compressed data structure
- Approximate computation

Reduce statistical complexity

- Dimensionality reduction, clustering

UIMA: Unstructured Information Management Architecture

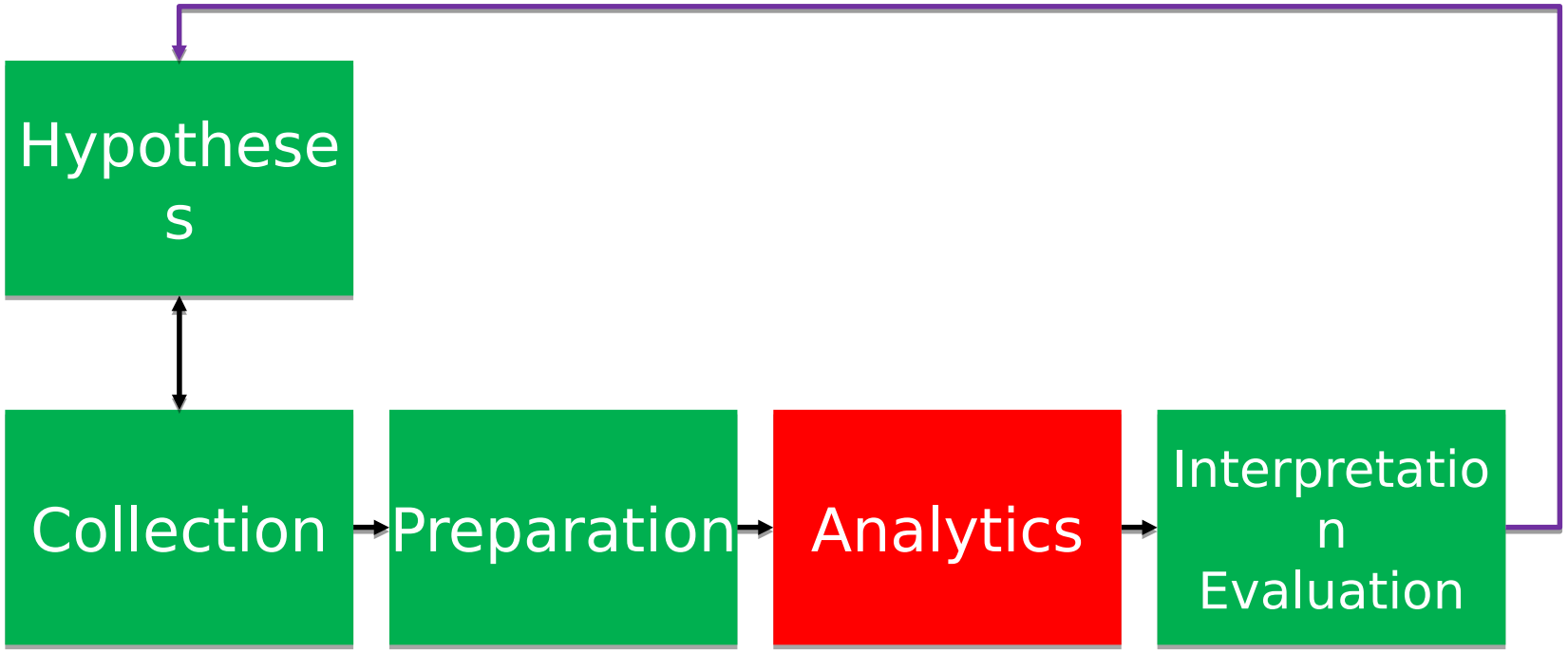
Sampling

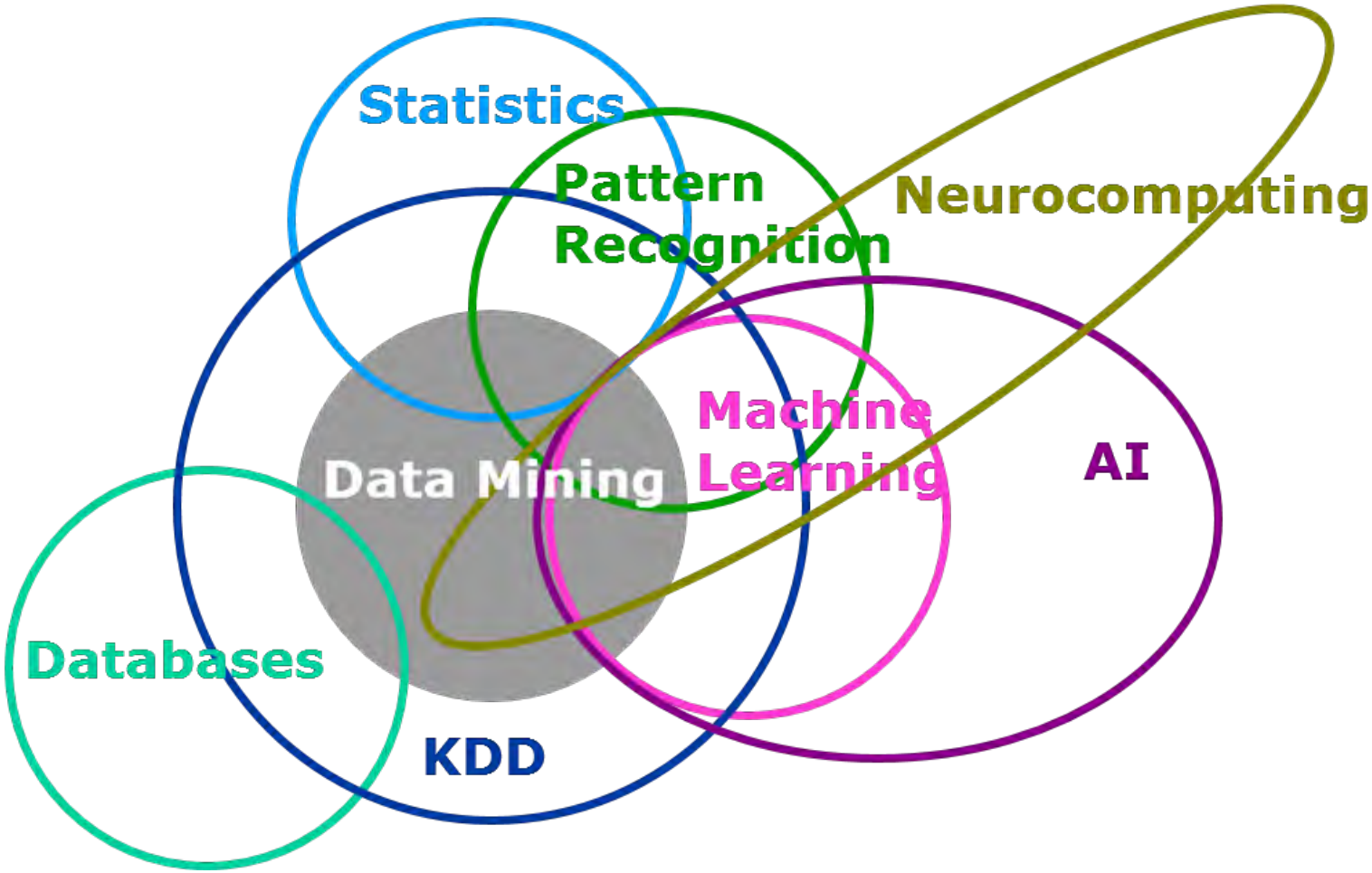
- Non-random sampling, compressive sensing, … …

Choose best representation for specific computational methods

- E.g. tables for data parallelism, networks/graphs for graph parallelism

数据分析方法论的升级


Hypotheses
Collection
Preparation
Analytics
Interpretation
Evaluation



Computational Science

Source: blogs.sas.com

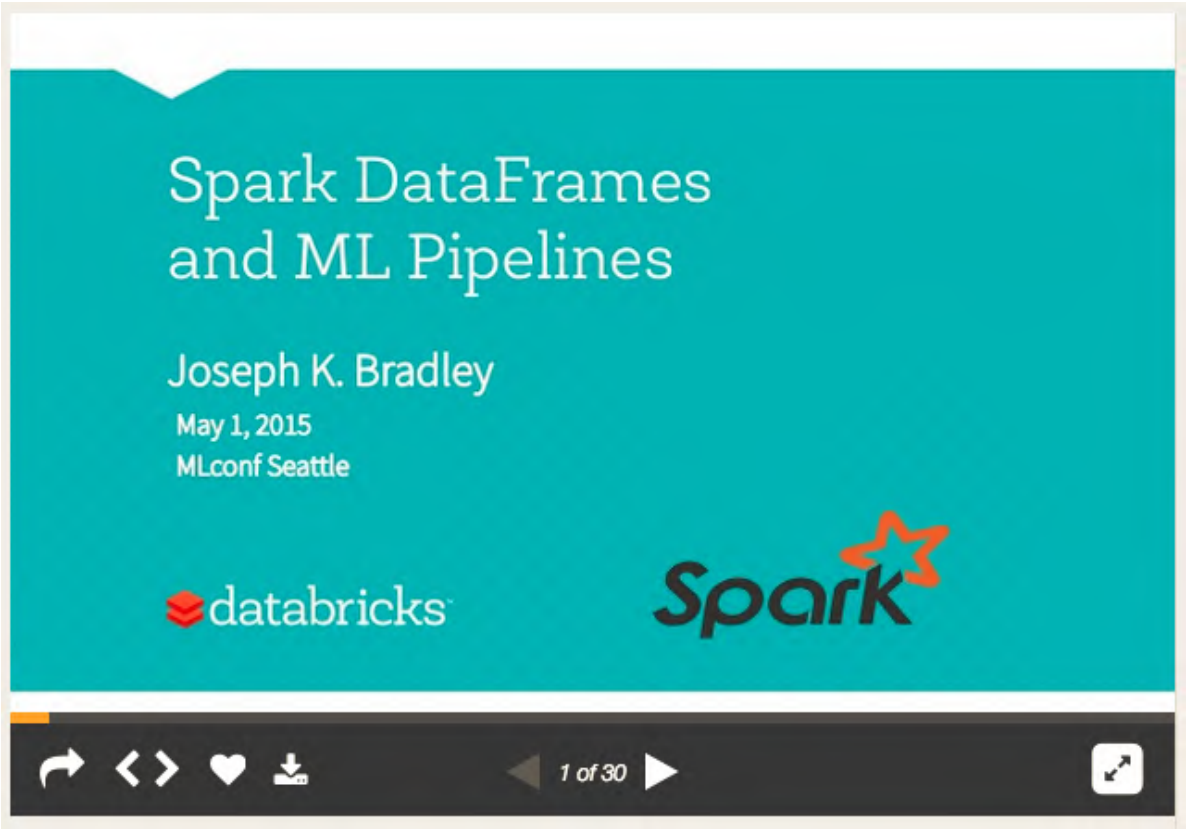
检查自身装备

检查自身装备

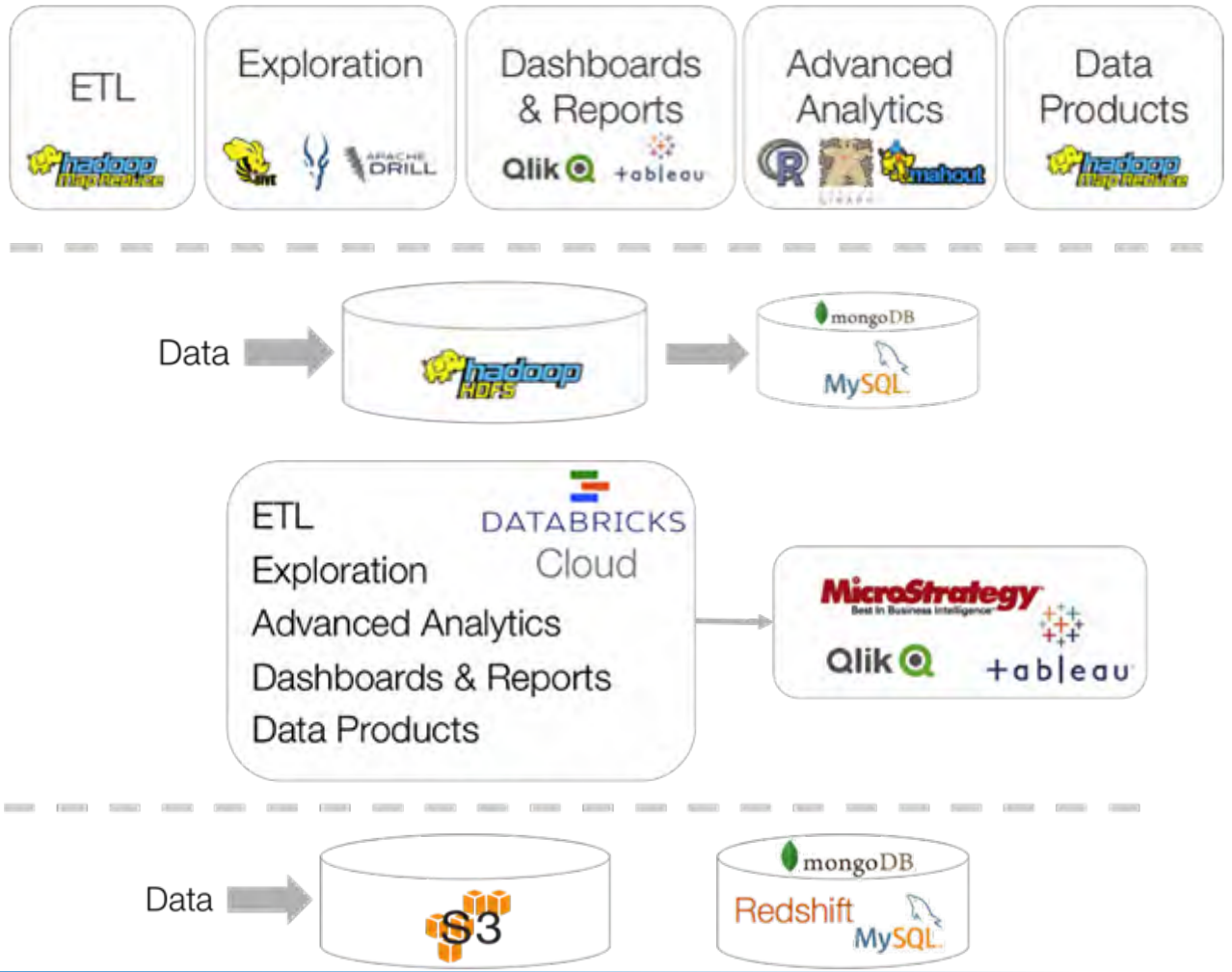

sas
R
SQL
python
cloudera
IMPALA
hadoop
MapReduce
TEZ
Spark
presto

ML Pipeline

Scikit-learn style pipelines



拥抱云的世界


ETL
Exploration
Dashboards & Reports
Advanced Analytics
Data Products
Data
mongoDB
MySQL
ETL
Exploration
Advanced Analytics
Dashboards & Reports
Data Products
DATABRICKS
Cloud
MicroStrategy
Qlik
tableau
Data
S3
mongoDB
Redshift
MySQL

all models are wrong, but some are useful

刺猬（一招鲜吃遍天） vs. 狐狸（一把钥匙开一把锁）

模型的复杂度与问题匹配：奥卡姆剃刀原理

如何做到数据越多、边际收益越大?

- 数据不可名状的功效：简单模型 + 大数据 > 复杂模型 + 小数据 ？
- Ensemble
- 混合模型
  - Non-parametric vs. Parametric
  - Linear vs. Non-Linear
  - Discriminative vs. Generative

感知长尾信号 Exponential assumption vs. long tail

- PCA/LDA/pLSA vs. 分级训练、模型组合、概率图模型、深度神经网络

Velocity

Interactive query

流计算

- 如时空数据

在线 / 流式 / 增量学习

- 增量训练，模型异步更新，快速部署

当又大又快时，你必须懂系统

- Big Learning System： 并行化 / 分布式化，系统调优

Deep Learning

从语音识别，到图像理解，到自然语言理解

- 领域特化：如医学图像分析

进军“非认知”任务

- 搜索，广告，推荐，……， visuo-motor skills, drug discovery
- Automated laboratory

Open Source ⊔ Collaborative Open Computer Science

- Pylearn2, Theano, Caffe
- GitXiv = arXiv + Github + Links + Discussion

Sparse Coding

A strong tool to deal with low SNR and poor veracity

We Have Successfully Applied Sparse Coding Based Prediction to a Number of Applications

2015:

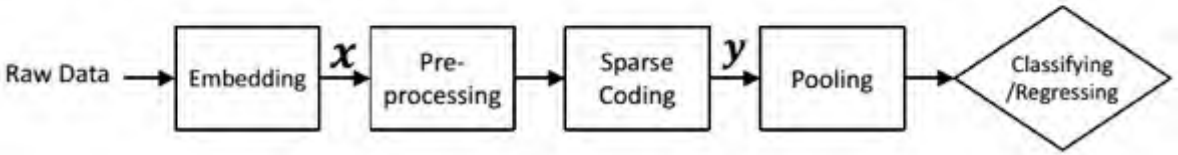
Twitter geolocation, facial emotion classification, wearable pulse sensing, link-layer adaptation to wireless fluctuations

2014:

Image object classification, chip power management, Wi-Fi traffic pattern inference, gait recognition, distributed spectrum sensing

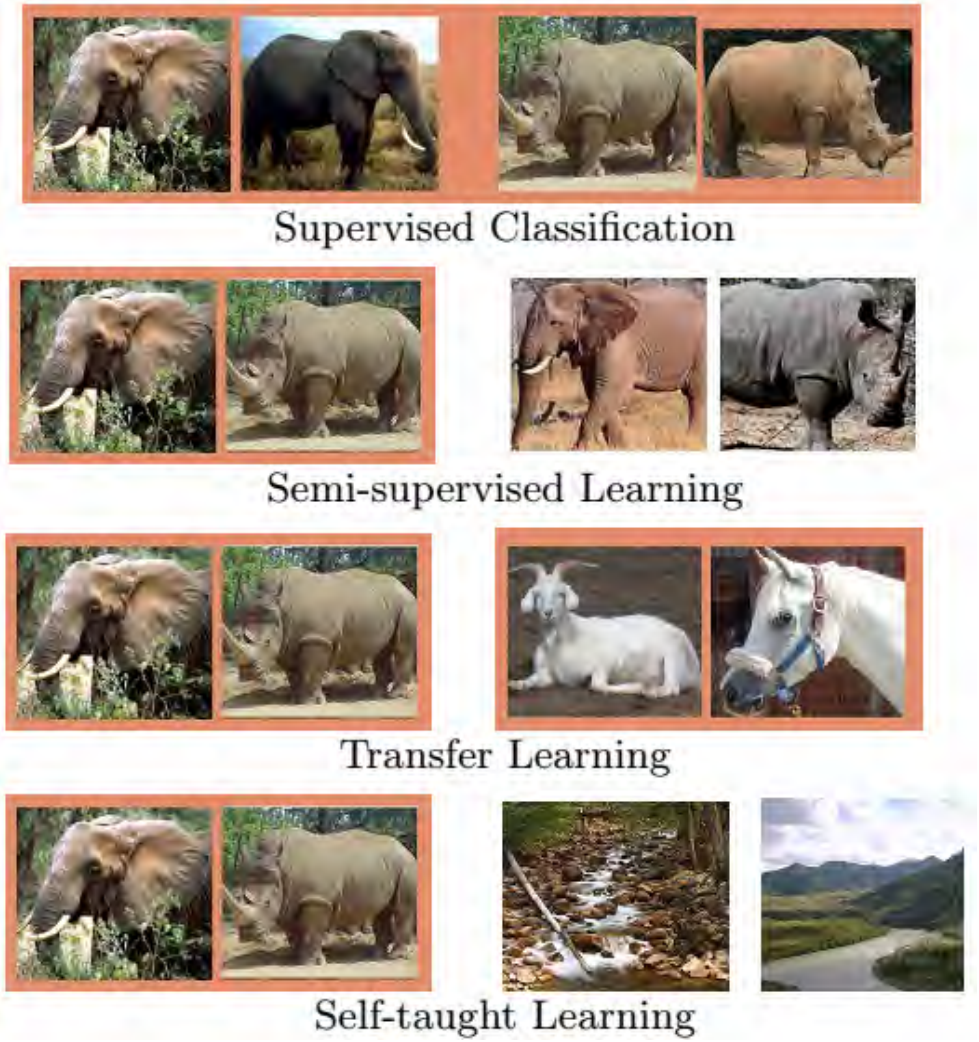
(See http://www.eecs.harvard.edu/htk/publications/)

All use **sparse coding based** prediction pipelines:



Source: HT Kung

缺乏标注数据下的学习



Source: Andrew Ng

人的角色 Human Machine Intelligence

人的工作不断被取代

- 特征工程
- 比如以 MLBase 和 VizDeck 为代表的自动化分析和可视化

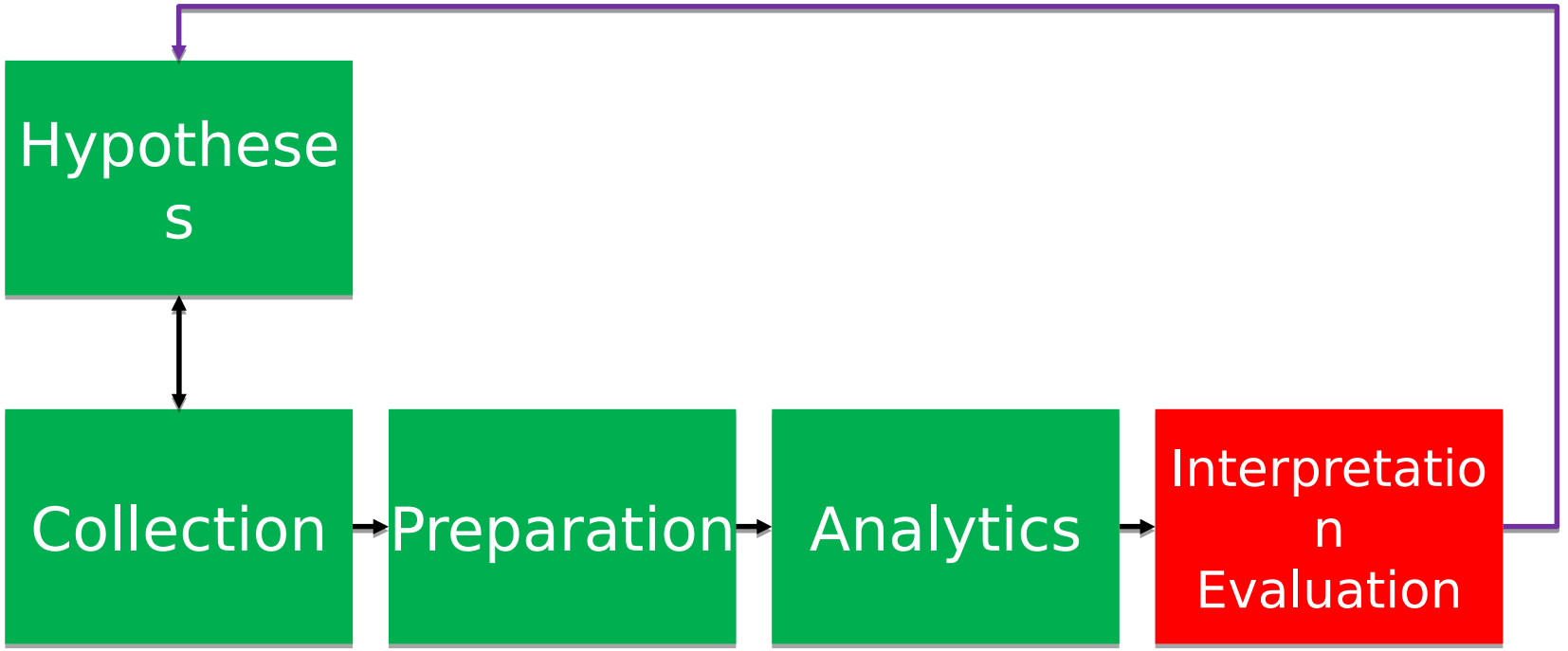
人与机器搭配获得最佳性能

- 比如 Exploratory analytics/visualization

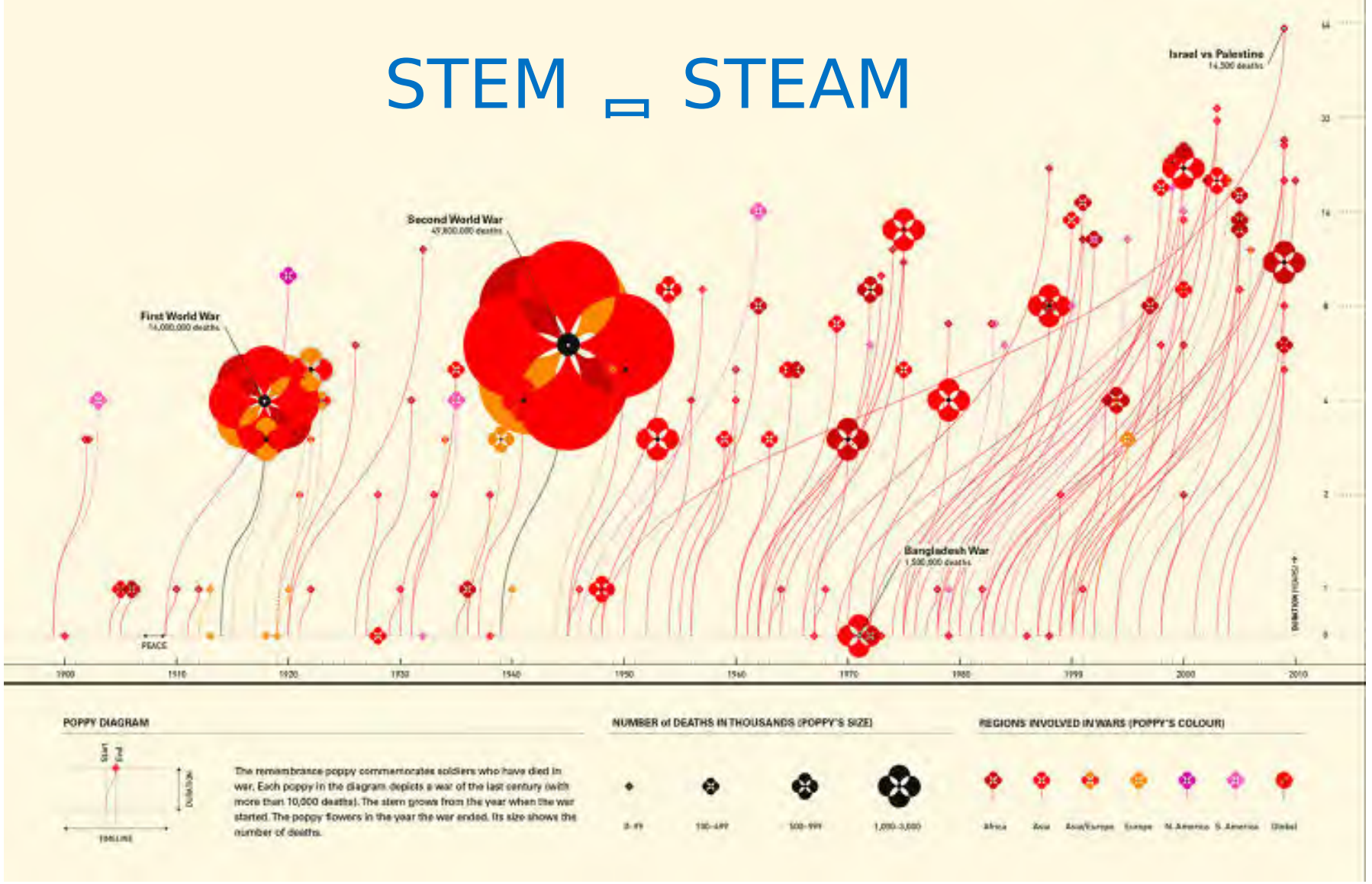
大规模人人、人机协作分析

- 外包 Kaggle，众包 CrowdDB，协作 DataHub，人类计算 Duolingo

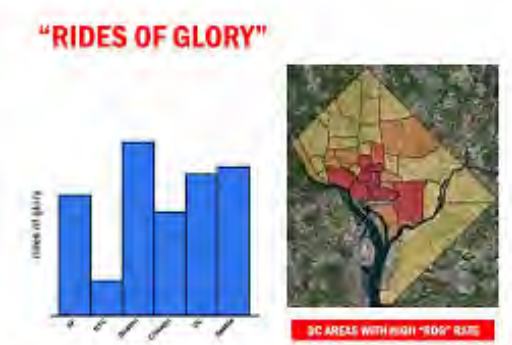
数据分析方法论的升级

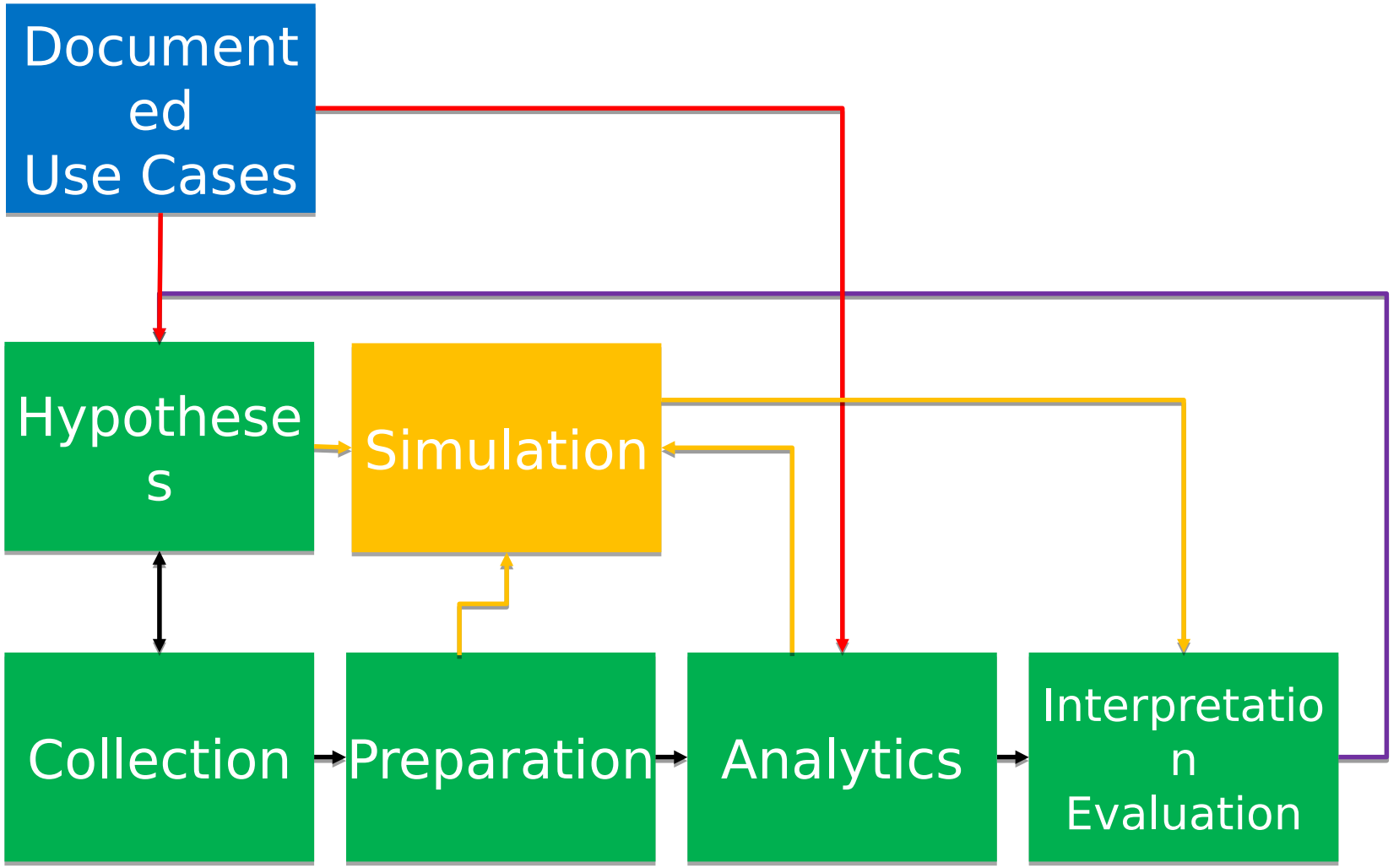

Hypothese
s
Collection
Preparation
Analytics
Interpretatio
n
Evaluation

STEM ⊂ STEAM


First World War
Second World War
Israel vs Palestine
14,500 deaths
Bangladesh War
PEACE
POPPY DIAGRAM
The remembrance poppy commemorates soldiers who have died in war. Each poppy in the diagram depicts a war of the last century (with more than 10,000 deaths). The stem grows from the year when the war started. The poppy flowers in the year the war ended. Its size shows the number of deaths.
NUMBER of DEATHS IN THOUSANDS (POPPY'S SIZE)
REGIONS INVOLVED IN WARS (POPPY'S COLOUR)

Story telling and “ideas worth spreading”




Documented Use Cases
Hypotheses
Simulation
Collection
Preparation
Analytics
Interpretation Evaluation

基础设施已经改朝换代

分析师也需要与时俱进

改变思维方式

提高技术素养

丰富分析能力

敬谢聆听