

Introduction of "Data Pre-Processing"



About Data

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

xuhua@tsinghua.edu.cn

1

Leading Question



Data, Information and Knowledge?

2



Leading Question

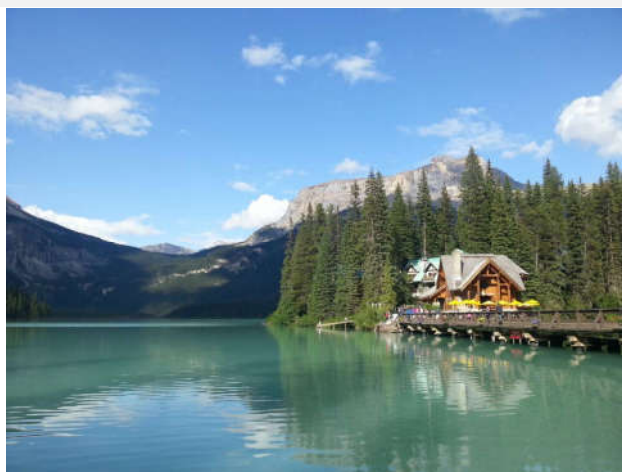


- ◉ **Data** : Number, Text, Figure, Audio, Video
- ◉ **Information** : Valuable Data
- ◉ **Knowledge** : non-trivial(非平凡的), implicit(隐含的), previously unknown(事先未知) and potentially useful(潜在用途) pattern or rules hidden in data

3



About Data (1)



Emerald Lake in Banff National Park of Canada (2014)

4



About Data (2)



2014 Beijing International Marathon
(Valuable Information)

5



About Data (3)



得看到的神翻译

汉译英：

菩提本无树
明镜亦非台
本来无一物
何处惹尘埃

谁
也是
醉了

Puti is not tree
Mirror is not table
It is empty at all
Why PM2.5 high ?

Confused Information (Dirty Data)

6



About Data (4)



转：某团队去年用核磁共振扫描了狗的大脑，发现狗是用左脑对语言进行处理的，并在Science上发了论文。然后今年才意识到，人是躺着进去的，而狗是趴着进去的。所以左右脑搞反了...
搞！反！了！！！！😂😂

Erratum
Erratum for the Report "Neural mechanisms for lexical processing in dogs" by A. Andics, A. Gábor, M. Gácsi, T. Farkas, D. Szabó, A. Miklósi

Science 27 Apr 2017

DOI: 10.1126/science.1251118

DOI: 10.1126/science.1251118

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

Article

令人无语的分析与数据

7



About Data (5)



Multi-form and Multi-modal Data



8



About Data (6)



◎ Text from one student' s diary

时间过得真快，一下就到半期考了，现在已经开始紧张的复习了，我必须要开始努力了，因为我如果不努力，成绩就上不去，我成绩上不去就会被家长骂。我被家长骂，就会失去信心，失去信心就会读不好书，读不好书就不能毕业，不能毕业就会找不到好工作，找不到好工作就赚不了钱，赚不了钱就会没钱纳税，没钱纳税，国家就难发工资给老师，老师领不到工资就会没心情教学，没心情教学，就会影响我们祖国的未来，影响了祖国的未来，中国就难以腾飞，中华民族就会退化成野蛮的民族。中华民族成了野蛮的民族，美国就会怀疑我国有大规模杀伤性武器，我国有大规模杀伤性武器，美国就会向中国开战，第三次世界大战就会爆发，第三次世界大战爆发其中一方必定会实力不足，实力不足就会动用核武器，动用核武器就会破坏自然环境，自然环境被破坏，大气层就会破个大洞，大气层破个大洞地球温度就会上升，两极冰山就会融化，冰山融化，地球水位就会上升，地球水位上升，全人类就会被淹死。因为这关系到全人类的生命财产安全，所以我要在剩下的几天里好好复习，考好成绩，不让悲剧发生。

9



About Data (7)



◎ Focused on Image



10



About Data (8)



11



About Data (9)



12



About Data (10)



13



About Data (11)



14



About Data (12)



15



About Data (13)



16



About Data (14)



17



About Data (15)



18



About Data (16)



当然，老外也模仿：
这次终于轮到别人抄我们。。。



Artfox Logo (美)



Redfox Logo (美)



法国房地产展销会 Logo



意大利内政部的标 Logo



波兰旅游宣传 Logo



Pilipinas Kay Ganda! 旅游 Logo (菲律宾)

抄的还是腾讯，我估计腾讯也没想到。。。

19



About Data (17)



细微的变化，实在有失水准



还有这个撞得天下皆知的。。。



INFINITI



CHERY

车标貌似是重灾区。。。



BENTLEY



20



About Data (18)



MDEC Logo



中秋节 Logo



09大连海事大学校庆 Logo



上海海洋大学校庆 Logo

争取把撞标撞向全世界



三龙创意 Logo



韩国丽水世博会中国馆 Logo

21



About Data (19)

GV车半年内
这个连颜色都不换一下



菲律宾银行MCC Logo



武汉市城市一卡通 Logo



还挺像模像样的



2016里约热内卢奥运会 Logo



众恒传播 Logo

22



About Data (20)



家电下乡? 旋转45°试试



23



About Data (21)



这个就不妄自下结论了, 自己琢磨吧!



OK吗? OK!



24



About Data (22)

仔细看还真有那么点血缘关系



墨西哥城地铁Logo



小米科技新Logo

将前者旋转180°试试?



vividways设计公司Logo (澳大利亚)



广发银行新Logo

截取了Vivid Ways的肢体部分



Vivid Ways的Logo



无线山西Logo

25



About Data (23) ——The emblem of the Tokyo Olympic Games



据《朝日新闻》报道，日前公布的2020年东京奥运会会徽被指与比利时一剧场的标志相似，该标志的设计师与剧场将以涉嫌侵犯知识产权为由向国际奥委会（IOC）提出申请，要求停止使用相关会徽。

26



About Data (24) —— Similarity = Joke



头条号 / 欧洲时报

V.S.



27



About Data (25) — Similarity = Joke



29



About Data (26) — Similarity = Joke



30



About Data (27) — Similarity = Joke



31



About Data (28) — Deeply Impressive



32



About Data (29) — Deeply Impressive



NASA地球照片



风云四号卫星画面

33



About Data (30)



◉ Data Labeling



Dissolving



Brushing



Scrubbing

34

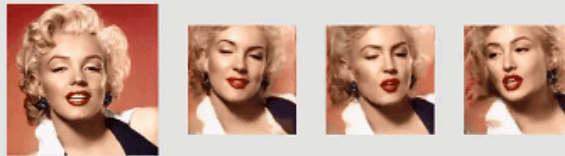


About Data (31)



◉ Augmented Data

Living portraits



35



Beginning from Basic and Abstract Numeric data



36





Thanks !

