Získavanie a čistenie dát

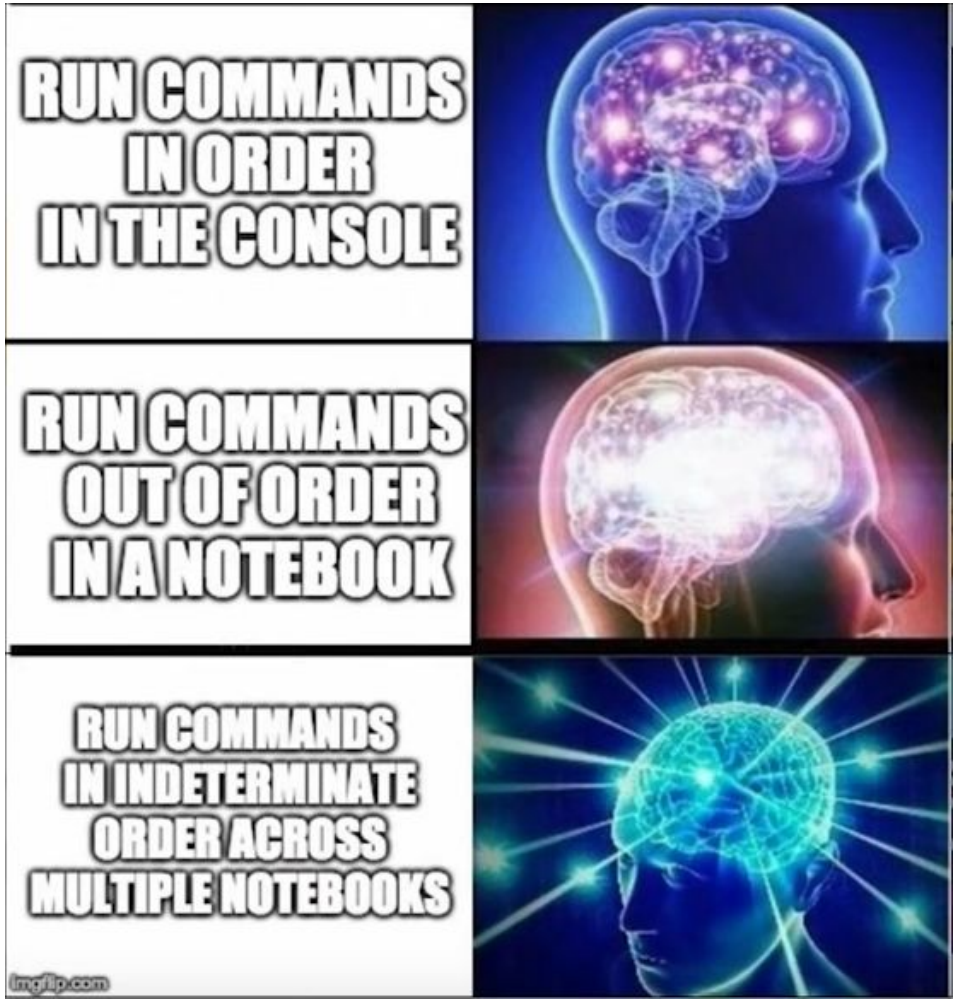
- Veľkú časť času strávi dátový vedec tým, že čistí, preusporiadava a preformátováva dátové súbory
- Takmer všetok zvyšok času strávi tým, že sa sťažuje na to, že vhodné dáta nie sú k dispozícii
- V angličtine termín “data munging” alebo “data wrangling”

Notebooky

- Dobre vytvorený notebook kombinuje dáta, výsledky výpočtov, aj text
  - Výsledky sú reprodukcibilné
  - Možno ľahko modifikovať
  - Obsahuje potrebnú dokumentáciu
- Všetky informácie relevantné k jednej analýze na jednom mieste
- Ľahká experimentácia s novými knižnicami
  - Posledný krok výpočtu sa dá skúšať v rôznych variantoch, až kým nefunguje
- Treba očakávať, že **analýzu bude treba prerobiť od začiatku do konca** (napríklad pri zmene základného parametru, pri zistení chyby v predspracovaní, ak vyplávajú počas analýzy artefakty)

Kedy NEpoužívať notebooky!

- Prostredie, v ktorom je **veľmi ľahké** streliť si do nohy
- Potrebujem bežať celú analýzu na iných dátach / s inými parametrami?
- Dáta sú **permanentne** naloadované v operačnej pamäti
  - zrýchľuje experiment
  - **pamäť je na zdieľaných prostrediach jeden z najdrahších / najvzácnejších zdrojov**



@joelgrus

Odkiaľ získavame dáta

- Privátne / komerčné dáta
- Open data iniciatívy (napr. vlády, samospráva)
- Akademické dáta
- Prehľadávanie webu
- Senzory
- Crowdsourcing
- Urob si sám
- Použi dáta na iný účel

Privátne / komerčné zdroje dát

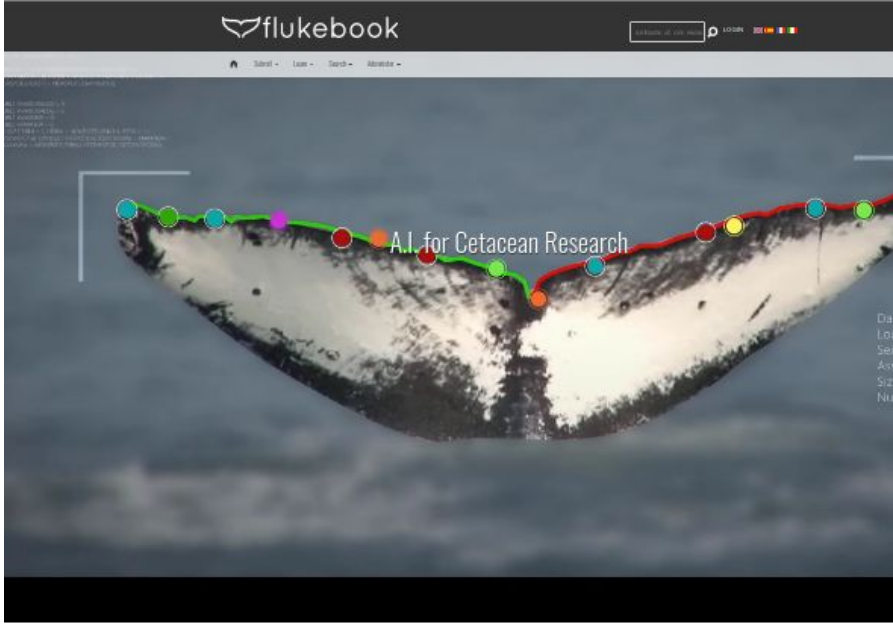
- Facebook, Google, Amazon - vynikajúce dáta o užívateľoch, transakciách, logy, a pod.
- Väčšina organizácií má množstvo interných dát týkajúcich sa rôznych interných aktivít
- Získať tieto dáta je takmer nemožné, pokiaľ nepracujete pre príslušnú organizáciu
- Niektoré organizácie vyrábajú ku dátovým zdrojom API s limitovanou frekvenciou dotazov

Open data initiatives (vláda, samosprávy)

- Zásada “keď si niečo platíme z daní, mali by sme mať prístup ku výsledkom v maximálnej možnej miere”
- Transparentnosť vlád, samospráv
- data.gov (>300,000 súborov z USA)
- data.europe.eu (>1.4 mil. súborov, 36 krajín)
- data.oecd.org (>2000 súborov týkajúcich sa krajín OECD)
- data.gov.sk (>3000 súborov slovenských verejných inštitúcií)
- opendata.bratislava.sk (cca 16500 súborov týkajúcich sa bratislavskej samosprávy)
- crz.gov.sk, portalvs.sk
- github.com/Institut-Zdravotnych-Analyz
- Vo veľa krajinách existuje analóg Zákona o slobodnom prístupe k informáciám
- Najčastejší problém: ochrana osobných údajov

Dáta z akademických zdrojov

- V mnohých oblastiach je povinnosťou voľne sprístupniť dáta získané v rámci výskumných grantov / podkladové dáta ku publikáciám
- NCBI / EMBL - systematické databázy biologických a medicínskych dát
- Najjednoduchšie začať vyhľadaním relevantných publikácií
- Open Science / Citizen Science iniciatívy



Prehľadávanie webu / scraping

- Čo sa nachádza na web stránkach sa dá väčšinou automatizovane posťahovať a preusporiadať do použiteľného tvaru
- Pozor na “Terms of Service”
- Ku veľa stránkam existujú otvorené pythonové knižnice
- … alebo aj oficiálne API

Logovanie senzorov

- “Internet of Things” (IoT)
- Verejne uploadované obrázky (aké bolo počasie?)
- Zemetrasenia cez akcelerometre v mobiloch
- Dopravné zápchy cez GPS v taxíkoch / sledovanie MHD

Disk je lacný, logujte!

Jano Suchal is feeling happy.
September 21 at 8:36 PM ·

IoT doma #4

Naposledy som zistil, že ani neviem, ale mám doma kopec smart zariadení. Konkrétne merače na radiátoroch a merače vody. Tak som našiel borcov, čo na to spravili čítací soft cez anténtu. Objednal som anténu, ale kedže som mal tušáka, že problém bude šifrovanie, tak som začal obvolávať firmu, ktorá to vyrába (Techem). Boli veľmi milí, pokecali sme, všeličo som sa dozvedel, ale povedali, že "šifrovacie kľúče nastavujú na centrále v Nemecku" a teda to vyzeralo zle, lebo z centrály zatiaľ ani bú, ani mú.

Dnes prišla anténa, pichol som to do USB, spustil, že si teda odskočím, kým to niečo nachytá (očakával som jeden ping za 15 minút), ale - deti moje zlaté - toto je elektrovoyerizmus. Všetko nešifrované, vidím komplet spotreby celého okolia. Susedov, všetko.

Čiže z problému "ako získam kľúče?" sa stal problém "ktoré z toho som do pekla ja?". Ukázalo sa, že to je ľahšie ako sa čakalo. Na radiátorových meračoch sú nejaké tri blikajúce číselká a tie sú presne to, čo to vysiela. A ešte to vysiela aj dve teploty (radiátora + okolia radiátora).

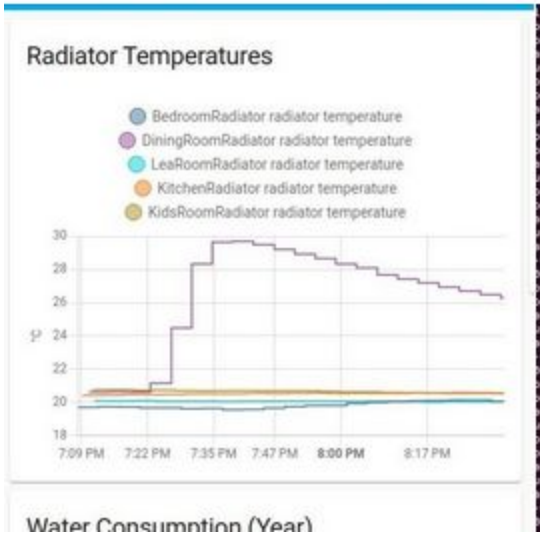
Vodné merače mi merajú teplú a studenú, ale trošku sklamanie, že presnosť je 0.1 metra kubického, čize sto litrov vody. Dám vedieť, keď sa to pohne.

Všetky senzory to vysielajú každé 4 (!!!) minúty.

No a samozrejme kedže to je celé opensource, tak som dorobil podporu na automatickú detekciu týchto senzorov v home assistant. Pánko, čo to spravuje to rovno za 20 minút schválil, čiže už to môže používať celá zemeguľa.

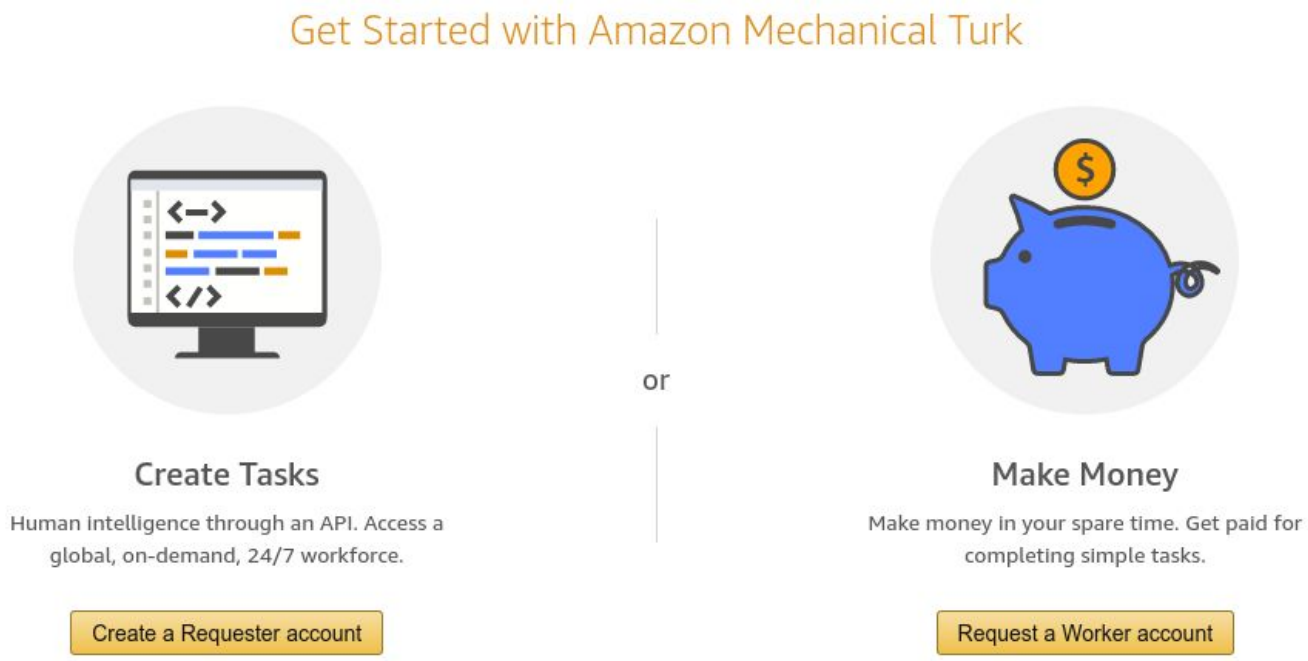
A ešte mi dnes prišiel merač na celkovú spotrebu energie. No malé vianoce.

Pull request: https://github.com/weetmuts/wmbusmeters/pull/618



Crowdsourcing

- Stránky / databázy kolaboratívne vytvárané nadšencami (wikipedia, IMDB, …)
- Amazon Turk - možnosť zaplatiť si armádu ľudí, ktorí napr. anotujú obrázky
- Ľudská práca je často jednoduchšia / lacnejšia ako budovanie informačného systému



Získanie dát vlastnou prácou

- … keď to ináč nejde, treba získať dáta vlastnou prácou
- Množstvo historických dát len v papierových archívoch alebo nekvalitne naskenovaných PDF súboroch
- 1000 záznamov sa dá vytvoriť ručne za 2 pracovné dni!
  (frekvencia 1 záznam za minútu)
- Rozposielanie listov / e-mailov náhodným ľuďom

Použitie dát na iný účel

Metadáta v rôznych databázach

- názvy kníh v databáze ISBN čísel
- názvy tovarov v databáze EAN
- popisy obrázkov
- editovacia história na wikipédii

… vyžaduje kreatívne myslenie

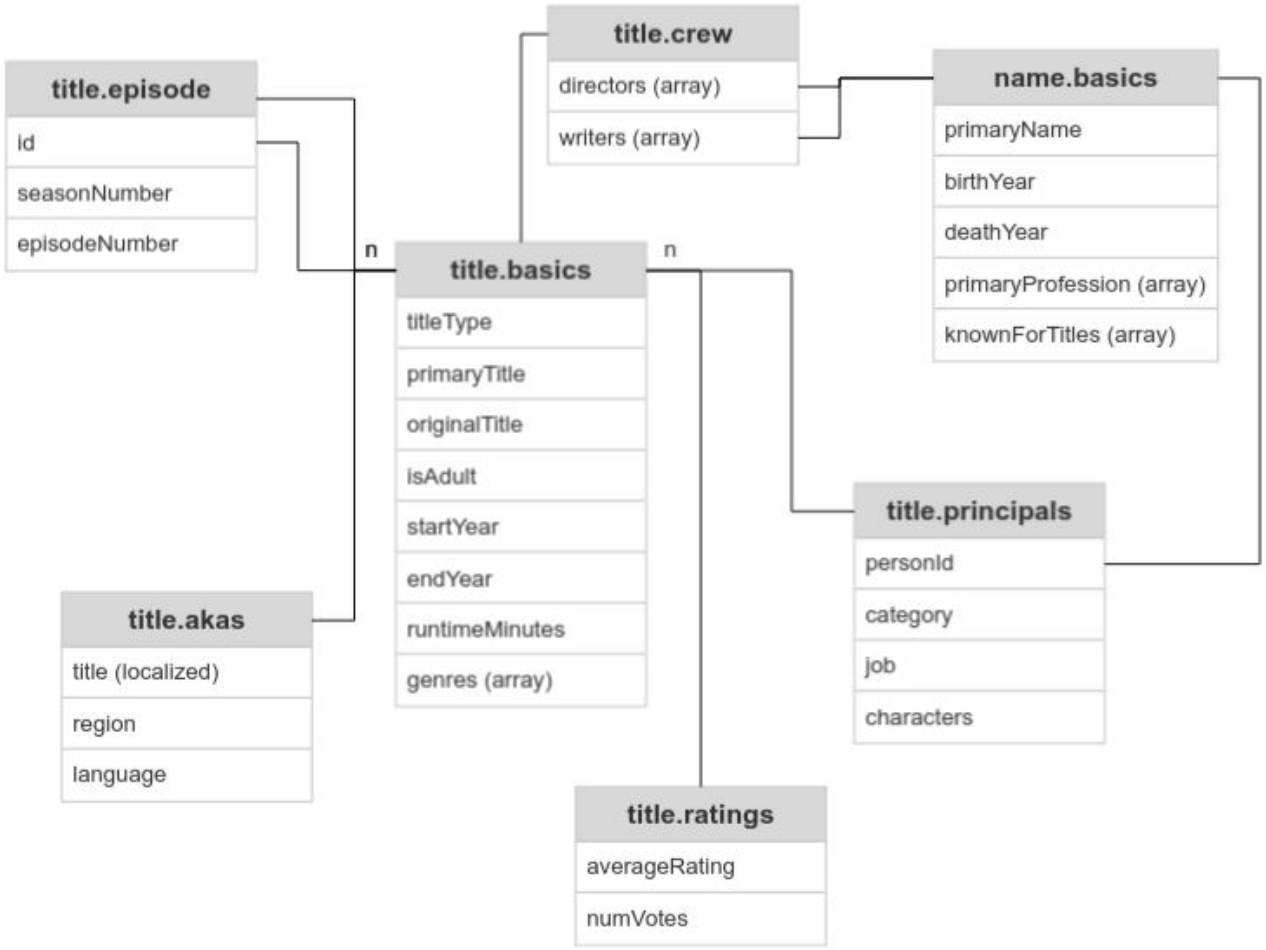
Dátové formáty

- tabuľkové formáty (CSV / TSV súbory, PARQUET, xls)
- SQL databázy
- XML súbory
- JSON (JavaScript Object Notation for APIs)

pandas

- výborný spôsob ako spracovávať “tabuľkové” súbory jednotným spôsobom
- zásadná nevýhoda: všetko sa deje v pamäti
- alternatívy:
  - jednoduchý kód prechádzajúci súbor po riadkoch
  - distribuované spracovanie veľkých komprimovaných súborov (napr. apache parquet)
  - SQL (a iné) databázy

imdb.com - štruktúra downloadovateľných TSV súborov



XML súbory

```
<Customer ID="C00-10101">
   <Name>John Hancock</Name>
   <Address>100 1st Street, San Francisco, CA 94118</Address>
   <Phone1>(858)555-1234</Phone1>
   <Phone2>(858)555-9876</Phone2>
   <Fax>(858)555-9999</Fax>
   <Email>John@somecompany.com</Email>
   <Order Number="NW-01-16366" Date="2012-02-28">
      <Contact>Mary Jane</Contact>
      <Phone>(987)654-3210</Phone>
      <ShipTo>Some company, 2467 Pioneer Road, San Francisco, CA - 94117</ShipTo>
      <SubTotal>434.99</SubTotal>
      <Tax>32.55</Tax>
      <Total>467.54</Total>
      <Item ID="001">
         <Quantity>10</Quantity>
         <PartNumber>F54709</PartNumber>
         <Description>Motorola S10-HD Bluetooth Stereo Headphones</Description>
         <UnitPrice>29.50</UnitPrice>
         <Price>295.00</Price>
      </Item>
      <Item ID="101">
         <Quantity>1</Quantity>
         <PartNumber>Z19743</PartNumber>
         <Description>Motorola Milestone XT800 Cell Phone</Description>
         <UnitPrice>139.99</UnitPrice>
         <Price>139.99</Price>
      </Item>
   </Order>
</Customer>
```

Query: /Customer/Order/Item

JSON


```
{       "id": "0001", "type": "donut", "name": "Cake", "ppu": 0.55,
        "batters":
                { "batter": [ { "id": "1001", "type": "Regular" }, { "id": "1002", "type": "Chocolate" },
                              { "id": "1003", "type": "Blueberry" }, { "id": "1004", "type": "Devil's Food" } ]
                 },

                "topping": [ { "id": "5001", "type": "None" }, { "id": "5002", "type": "Glazed" },
                             { "id": "5005", "type": "Sugar" }, { "id": "5007", "type": "Powdered Sugar" },
                             { "id": "5006", "type": "Chocolate with Sprinkles" },
                             { "id": "5003", "type": "Chocolate" },
                             { "id": "5004", "type": "Maple" } ]

}
```