THE DATA JANITOR RETURNS

Daniel Molnar, Oberlo/Shopify

Build Stuff 2018
Where I'm coming from

• senior data analytics engineer,

• head of data and analytics,

• senior applied and data scientist,

• data analyst,

• or just data janitor.
Perspective

• rounded, not complete,
• slow, old, stupid and lazy and

there are only 2 types of software engineers:
1. Bitter, cynical, paranoid engineers
2. Junior engineers
22:11 - 2017 Mar 28
163 retweets 217 likes
tl;dr
tl;dr

• KISS is the philosophy,

• take the long view, invest in durable knowledge,

• strive for fast and good enough,

• just because you can doesn't mean you should.
tl;dr (new)

• KISS is the philosophy,
• take the long view, invest in durable knowledge,
• strive for fast and good enough,
• just because you can doesn't mean you should,
• figure what to worry about,
• you are not Google.
it used to be a hype
now this is a war
nobody's your friend
they want your money and data (preferably both locked in)
Things you worry about:

• machine learning,

• deep learning,

• GDPR.
Things you should really worry about:

• machine learning adblockers,
• deep learning ELT,
• GDPR, CRM (yes, CRM).
THE DATA SCIENCE HIERARCHY OF NEEDS

LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

CLEANING, ANOMALY DETECTION, PREP

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

AI, DEEP LEARNING
Don't skip leg day.
Do make programmatic KPI definitions.
Look at the *** data
Toolset Python, (P)SQL, Metabase.
Usual suspects
NPS

• one, simple number you can squint at,
• sampling is skewed,
• answer is unsure,
• easy to hack step function\(^1\),

**MONKEYPATCH:** look at the change of the distro.

---

\(^1\) Eve Rajca aka @EveTheAnalyst and Jacques Mattheij aka @jmattheij
Predictably wrong?
Google Analytics!
Curve-Fitting Methods
And the Messages They Send

Linear
"Hey, I did a regression."

Quadratic
"I wanted a curved line, so I made one with math."

Logarithmic
"Look, it's tapering off."

Exponential
"Look, it's growing uncontrollably."

Loess
"I'm sophisticated, not like those bumbling polynomial people."

Linear, no slope
"I'm making a scatter plot but I don't want to."

Logistic
"I need to connect these two lines, but my first idea didn't have enough math."

Confidence interval
"Listen, science is hard, but I'm a serious person doing my best."

Piecewise
"I have a theory, and this is the only data I could find."

Connecting lines
"I clicked 'smooth lines' in Excel."

Ad-hoc filter
"I had an idea for how to clean up the data. What do you think?"

House of cards
"As you can see, this model smoothly fits the... Wait, no ad don't extend it aaaaaaaa!"
Hero of the day

Martin Loetzsch
@martin_loetzsch

-=-

KPIs for e-commerce startups

Data Science in Early Stage Startups: the Struggle to Create Value

https://github.com/mara
LEARN & OPTIMIZE

SCIENCE

ARCH OF NEEDS

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRANSLATION

CLEANING, ANOMALY DETECTION, ...

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

OPTIMIZE
Half of the time when companies say they need "AI" what they really need is a SELECT clause with GROUP BY. You're welcome.

— Mat Velloso @matvelloso (Technical Advisor to CTO at Microsoft)
Don't do A/B tests

99% it will not worth doing it
... conversion rate is 2% ... detecting a relative change of 1% requires an experiment with 12 million users ...

— Simon Jackson (Booking.com)
R?

Shiny.
Usual suspects

- non-reproducible experiments and tests,
- R hodpepodge in production,
- beliefs hidden as implicits in models.
THE DATA SCIENCE HIERARCHY OF NEEDS

COLLECT

MOVE/STORE

EXPLORE/TRANSFORM

AGGREGATE/LABEL

LEARN/OPTIMIZE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

CLEANING, ANOMALY DETECTION, PREP

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

AI, DEEP LEARNING
SCIENCE

PRIORITY OF NEEDS

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

ANALYTICS, METRIC SEGMENTS, AGREGATES, FEATURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, PRE-PROCESSING

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSING, EXTERNAL DATA, USER GENERATED CONTENT
You don't have (enough) data.
Make your own data points!
Deploy good enough fast?

Man, Docker is being used for everything. I don't know how I feel about it. Story time!

Once, long ago, I wanted to use an old tablet as a wall display.

I had an app and a calendar webpage that I wanted to show side by side, but the OS didn't have split-screen support. So I decided to build my own app.

I downloaded the SDK and the IDE, registered as a developer, and started reading the language's docs.

...Then I realized it would be way easier to get two smaller phones on eBay and glue them together.

But you never learned to write software.

No, I just learned how to glue together stuff that I don't understand.

I...OK, fair.

On that day, I achieved software enlightenment.
Deep learn my ***

Do you really need it?

Tensorflow! ...

... so distributed deep learning can compress porn on the end device.

Will Compression Be Machine Learning’s Killer App?
Hero of the day

Szilard [Deeper than Deep Learning] @DataScienceLA

---

Better than Deep Learning: Gradient Boosting Machines (GBMs)

https://github.com/szilard/benchm-ml
Spark MLlib GBM implementation is 10x slower, uses 10x more memory and is buggy/lower accuracy. Total fucking garbage! 😂😭

— Szilard [Deeper than Deep Learning] @DataScienceLA
THE DATA SCIENCE HIERARCHY OF NEEDS

COLLECT

MOVE/STORE

EXPLORE/TRANSFORM

AGGREGATE/LABEL

LEARN/OPTIMIZE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

CLEANING, ANOMALY DETECTION, PREP

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

AI, DEEP LEARNING
SCIENCE

CHY OF NEEDS

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHM

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, I

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED OR UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
Q: Why are there so many programmers from Eastern Europe?
A: Slavic pessimism. Everything that can go wrong will go wrong. With such a mindset programming comes naturally.

– Martin Sustrik @sustrik (Creator of ZeroMQ, nanomsg, libdill.)
BEST PRACTICES IN APPLICATION ARCHITECTURE

TODAY: USE LAYERS TO DECOUPLE

ARCHITECTURE 2011

... AND EVERY YEAR WE MOUNT A NEW LAYER TO DECOUPLE US FROM THE CRAP WE'VE DONE THE YEAR BEFORE

ANNUAL RINGS
over engineering
you get an other machine
if you can use
one
Do embrace dirty reality.
Get cloud agnostic!

- AWS still leads the pack by far
- Azure will sell anyway, and all will cry,
- Google competes with the cheap and uncooked
ETL is #solved OMG

- Airflow is an overengineered underperforming nightmare,
- metl for source mappings in magnitude,
- Mara for generic e-commerce,
- night-shift for explicit minimalism.
Showdown
Hero of the day

Mark Litwintschik @marklit82

Summary of the 1.1 Billion Taxi Rides Benchmarks (500 GB uncompressed CSV)

https://tech.marksblogg.com
## Spark

<table>
<thead>
<tr>
<th>Setup</th>
<th>Query Median</th>
<th>QM per vCPU</th>
<th>Cost/hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 x m3.xlarge + HDFS</td>
<td>14.91</td>
<td>0.34</td>
<td>27.5</td>
</tr>
<tr>
<td>1 x i3.8xlarge + HDFS</td>
<td>26.00</td>
<td>0.81</td>
<td>2.5</td>
</tr>
<tr>
<td>21 x m3.xlarge + HDFS</td>
<td>32.00</td>
<td>0.38</td>
<td>5.67</td>
</tr>
<tr>
<td>5 x m3.xlarge + S3</td>
<td>466.50</td>
<td>23.33</td>
<td>1.35</td>
</tr>
<tr>
<td>3 x Raspberry Pi</td>
<td>1738.00</td>
<td>144.83</td>
<td></td>
</tr>
</tbody>
</table>

HDFS. RPi = 1/6 VCPU ~100 EUR. Linear scaling.
### Presto

<table>
<thead>
<tr>
<th>Setup</th>
<th>Query Median</th>
<th>QM per vCPU</th>
<th>Cost/hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 x n1-standard-4</td>
<td>7,00</td>
<td>0,04</td>
<td>9.50</td>
</tr>
<tr>
<td>21 x m3.xlarge</td>
<td>11,50</td>
<td>0,14</td>
<td>5.67</td>
</tr>
<tr>
<td>10 x n1-standard-4</td>
<td>16,00</td>
<td>0,36</td>
<td>2.09</td>
</tr>
<tr>
<td>1 x i3.8xlarge + HDFS</td>
<td>15,00</td>
<td>0,47</td>
<td>2.50</td>
</tr>
<tr>
<td>5 x m3.xlarge + HDFS</td>
<td>51,50</td>
<td>0,26</td>
<td>1.35</td>
</tr>
<tr>
<td>50 x m3.xlarge + S3</td>
<td>43,50</td>
<td>0,22</td>
<td>13.50</td>
</tr>
</tbody>
</table>

*Workhorse in favour. HDFS. 1 machine. Non-linear scaling.*
# Lazy Evaluation

<table>
<thead>
<tr>
<th>Setup</th>
<th>Query Median</th>
<th>Cost/hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redshift, 6 x ds2.8xlarge</td>
<td>1,91</td>
<td>40.80</td>
</tr>
<tr>
<td>BigQuery</td>
<td>2,00</td>
<td></td>
</tr>
<tr>
<td>Amazon Athena</td>
<td>6,30</td>
<td></td>
</tr>
<tr>
<td>Presto, 50 x n1-standard-4</td>
<td>7,00</td>
<td>9.50</td>
</tr>
<tr>
<td>Spark, 11 x m3.xlarge + HDFS</td>
<td>14,91</td>
<td>27.50</td>
</tr>
</tbody>
</table>

The human cost -- in both terms.
## One Machine

<table>
<thead>
<tr>
<th>Setup</th>
<th>Query Median</th>
<th>QM per vCPU</th>
<th>Cost/hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClickHouse</td>
<td>4,21</td>
<td>1,05</td>
<td></td>
</tr>
<tr>
<td>Elasticsearch tuned</td>
<td>13,14</td>
<td>3,29</td>
<td></td>
</tr>
<tr>
<td>Presto, 1 x i3.8xlarge + HDFS</td>
<td>15,00</td>
<td>0,47</td>
<td>2,50</td>
</tr>
<tr>
<td>Spark, 1 x i3.8xlarge + HDFS</td>
<td>26,00</td>
<td>0,81</td>
<td>2,50</td>
</tr>
<tr>
<td>Vertica</td>
<td>32,80</td>
<td>8,20</td>
<td></td>
</tr>
<tr>
<td>Elasticsearch</td>
<td>48,89</td>
<td>12,22</td>
<td></td>
</tr>
<tr>
<td>PSQL 9.5 + cstore_fdw</td>
<td>205,00</td>
<td>51,25</td>
<td></td>
</tr>
</tbody>
</table>

Intel Core i5 4670K VS i3.8xlarge (32 VCPUs). Desktop example costs <600 EUR.
Do you use adblocking?
Do you use Google Analytics?
9% of the events are lost to ~all third party trackers due to adblocking.
Sink > Sieve > Sort

ELT aka SQL on flat files with the minimum amount of code written.
THE DATA SCIENCE HIERARCHY OF NEEDS

COLLECT

MOVE/STORE

EXPLORE/TRANSFORM

AGGREGATE/LABEL

LEARN/OPTIMIZE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

CLEANING, ANOMALY DETECTION, PREP

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

AI, DEEP LEARNING
Who are you?

- Lip service provider.
- Fake news producer.
- Kingmaker.

Are you the fool or the grey eminent?
Don't believe the hype.
HR: good people leave.
Marketing

SO HOW DID OUR CAMPAIGN DO?  GREAT!
NEARLY 1/10TH OF 1% CLICKED ON OUR AD.
POSSIBLY SOME OF THEM ON PURPOSE.
WHICH BEAT ALL OUR BENCHMARKS.
AND EXCEEDED THE INDUSTRY AVERAGE.
DID IT ACTUALLY DRIVE ANY SALES?
YOUR GUESS IS AS GOOD AS MINE.

© marketoonist.com
Will this ever get better?

- adblocking,
- CPA silver bullets are gone,
- conversion & attribution are hard nuts,
- FB and GO are not your friends (the 900% on videos),
- but CRM is.
GDPR

• road to hell is paved with good intentions,
• it's about the process, matey,
• mostly fair,
• yes, you have to clean up your mess,
• dunno, wouldn't buy programmatic shares\(^1\).

\(^1\) Eve Rajca aka @EveTheAnalyst and Jacques Mattheij aka @jmattheij
Thank you!

@soobrosa

We're hiring!

visuals: @mrogati, @xkcd, @DorsaAmir, ˇCavin 🔗, thelearningcurvedotca, JD Hancock, Thomas Hawk, jonolist, Kalexanderson, Shopify Burst