Top Ten Data Science Blunders

(With Apologies to David Letterman)

DBg Database Group
MIT Computer Science and Artificial Intelligence Lab

My ML World

- Tamr principal
  - Data integration (data mastering: customers, parts, suppliers, …)
  - At scale (n = millions)
  - ML application – forest of trees
  - 200+ projects

- MIT projects

My ML World

- Structured enterprise data
  - No video
  - Little text
- Goal is always to run the enterprise better (save money, …
- No focus on recommendation systems, web advertising, …

Blunder #1
Belief that Data Scientists do Data Science

- Merck vignette
  - 4000+/- Oracle databases
  - Large data lake
  - Uncountable private databases and spreadsheets
  - Public data off the web

- Their data scientists spend ~95% of their time on data discovery and data integration (according to Mark Schrieber)

Blunder #1 (Continued)

- Irobot vignette
  - 90% of the time spent on data discovery and data integration
  - 90% of the remainder spent on fixing data cleaning/integration errors
  - 1% on data science
- Nobody says they spend less than 80%!!!!
- You do data integration not data science…..
- You should spend your effort correspondingly

Blunder #2
Belief that Data Lakes will Solve your Integration Issues

Conventional Wisdom: Just load all your data into a "data lake" and you will be able to correlate all data sets

Important Fact (Tattoo this on your Brain): Independently constructed data sets are never "plug compatible"

Blunder #2 (Why)

- Schemas don't match
  - You call it salary; I call it wages
- Units don't match
  - You use Euros; I use $$$
- Semantics don't match
  - My salaries are gross before taxes; yours are net after taxes with a lunch allowance

Blunder #2 (Continued)

- Time granularity doesn't match
  - You have annual data; I have monthly data
- Data is dirty
  - -99 means null (sometimes)
  - Null means "data missing" or "data not allowed" or …
- Duplicates must be removed
  - And there are no keys
  - I am Mike Stonebraker in one data set; M.R. Stonebreaker in a second one

Blunder #2
(The Net Result)

- Your analytics will be garbage
  - "GIGO"
- Your ML models will fail
  - I.e. produce garbage
  - Again "GIGO"

Blunder #3
Belief that Traditional Data Integration Techniques Will Solve Previous Issues

Extract Transform and Load
(Available from a variety of vendors)

Master Data Management
(Also available from the usual suspects)

Blunder #3
ETL

What's attempted:

- Decide what data sources to integrate (top down)
- Build a global data model (up front)
- For each data source
  - Send a programmer to interview the data set owner
  - He then builds an extractor, data cleaning routines (in a proprietary scripting language)
  - And loads data into the global schema

Why it doesn't work:

- I have never seen this technique work for more than 20 data sources
  - Too human intensive
- Building a global schema upfront is way too difficult at scale
  - Remember enterprise wide data models from 15-20 years ago…
- Most enterprises I know have way more than 20 data sources
  - Merck has 4000+/- Oracle data bases
  - A data lake
  - Countless files
  - And data from the web is also important

Blunder #3
MDM

- Once you have run ETL, you need “match/merge”
- MDM suggests building “golden records” by
  - Implementing match rules (e.g. two entities are the same if they have the same address)
  - Implementing merge rules (e.g. take the most recent value and ignore older ones)

Doesn’t Scale!

- GE classification problem: 20M spend transactions to be classified into a pre-built hierarchy
- 500 rules classified only 10% of the spend transactions

Blunder #3
So What to Do?

At scale, you need a solution that leverages ML and statistics

- Ok to use rules to generate training data
- That's what Tamr did on the GE problem

Blunder #4
Not Worrying About Scale

- If you have 1000 records, then run <<whatever>>
- If you have millions, then things get dicey
  - Toyoto Motor Europe wants to do “customer mastering” on ~30M raw records
  - Santander on 15M
  - Transamerica on 300M

Blunder #4
Not Worrying About Scale

- Training data is a HUGE problem
  - Manual tagging won't work
  - Have to use "rules" (or something else)
  - Deciding if training data is "good enough" is a challenge

- Always run a pilot first!!!!!

Blunder #4
And Algo Makes a Big Difference

- Carnival Cruise Lines -- 9 different lines (Celebrity, Holland American, Carnival, ...)
- 9 different spare parts depoting systems, each a data silo
- Carnival wants to share parts
- Need parts mastering of ~10M parts
  - No keys!

Blunder #4
And Algo Makes a Big Difference

- There are 9 parts classification hierarchies; all different
- Start with these “buckets”. Map the buckets to the new global hierarchy. Dedup these super buckets
- Start with the new global hierarchy. Classify raw records directly

- Tamr engineers had to decide which approach to use.

Blunder #5
Belief that Deep Learning is the Answer

- All the rage right now, but….
- Tamr has done ~200 projects for large enterprises – on structured data
- Nobody is interested in neural networks
  - Too much training data required!!!! (and it is always ”the high pole in the tent”)
  - Cannot be validated by non experts (Is Merck with a US address the same entity as Merck with a German address?)
  - Explanations often required – Someday neural networks will get there…..

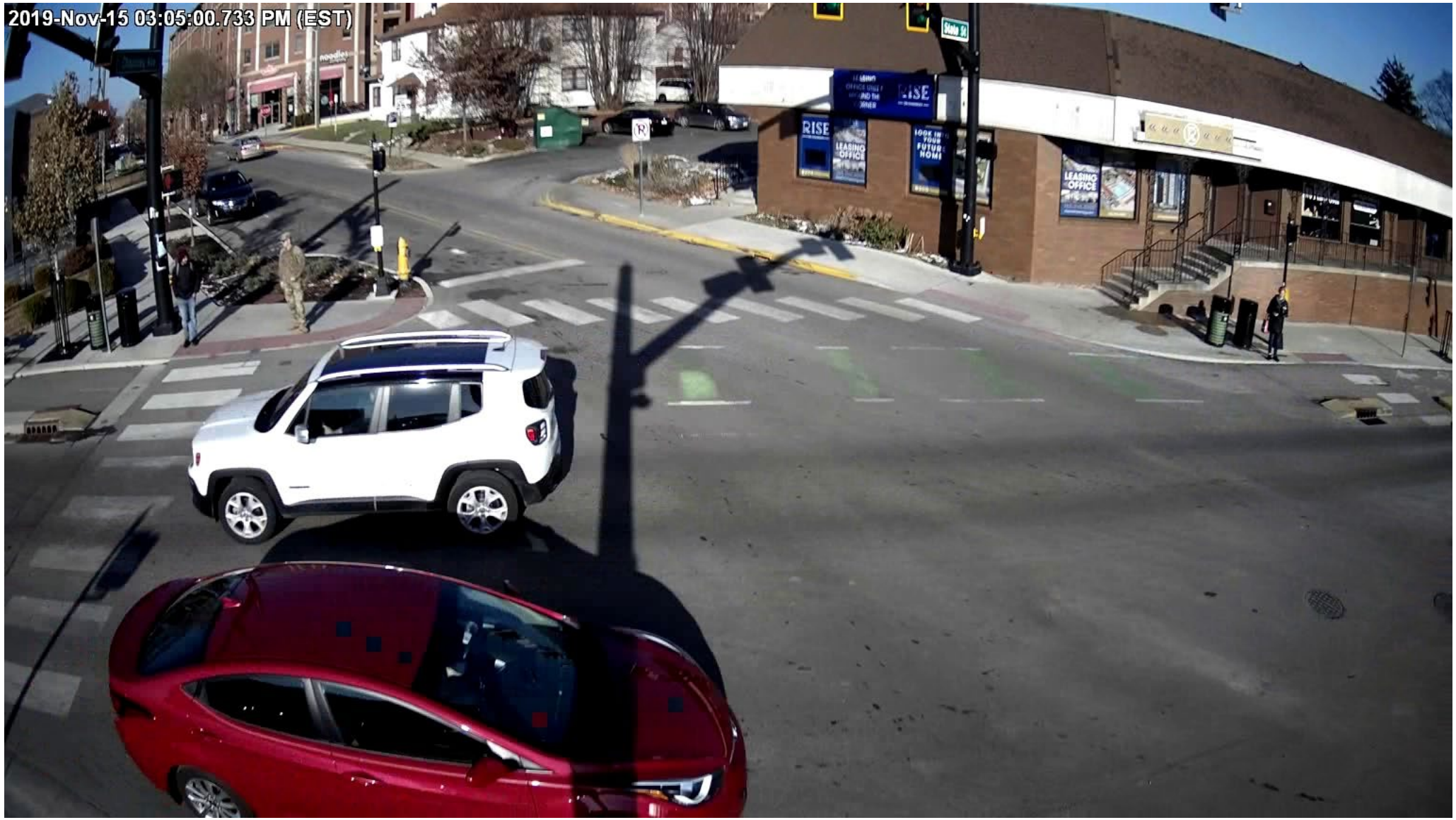
Blunder #6
Belief that Deep Learning is the Answer for Video

- We (MIT, Purdue) are doing a project with the West Lafayette police department
- Incident report appears (assault reported at time XXX in location YYY. Suspect is a tall male wearing jeans and a black shirt….)
- Cameras everywhere (in buses, police cars, body cameras on policemen, on lightpoles in the downtown areas….)
- Detectives say that video is the most effective investigative tool (and the suspect was, in fact, identified from city bus video)
- Police identified 31 features of interest (male, female, jeans, tattoos, …)

Video Is:

- Low res
- Objects are often far away
- Lighting is often poor
- It rains and snows

- Not posed teapots on the web


2019-Nov-15 03:05:00.733 PM (EST)
RISE
LEASING OFFICE

Blunder #6
Belief that Deep Learning is the Answer for Video

- Two students:
- Alice – Conventional deep learning by tagging frames
- Bob – Color analysis on segments of Yolo bounding boxes for “human”

- Result – stay tuned, but Bob is done and Alice is still tagging data

- Mantra: Use the simplest technique possible!!

Blunder #7
Not Worrying About the Details

- Bias
  - Face recognition works better on fair skinned people
- Skew
  - Often 90% of the data is in 10% of the space
  - Random sampling for training data will fail
- Model management
  - You will do hundreds to thousand of runs
  - Repository of training data, parameters, models, …

Blunder #8
Working for the Wrong Company

- You don’t have a CDO
  - Data is not important
- You don’t have a global catalog
  - How are you going to find the data you need?
- Your CDO does not have read access to everything
  - Otherwise, you will spend all your time on politics

Blunder #9
No Clear Way Forward

- “Do something interesting”
- “Show me some value”
- “Data science is supposed to be great – I started a group”

Blunder #10
Working for a Company That is not Trying to do Something about these Blunders

If you work for a company that is succumbing to (**even one**) of these blunders then:

1. You should be fixing it
   - Be part of the solution, not part of the problem

2. Or looking for a new employer
   - Tamr is hiring!