Catching Value Added Tax evaders in Delhi using Machine Learning

...

Aprajit Mahajan (UC Berkeley)
Shekhar Mittal (UCLA)
Ofir Reich (Data Scientist, CEGA)

Tackle corruption with data

Using administrative data to pinpoint corruption

Not only research - enforcement

A general approach

This talk - tax evasion, Delhi

The problem - Value Added Tax evasion via bogus firms

National Capital Territory of **Delhi**

Bogus firms exist only on paper, and make money by falsely reporting transactions with genuine firms (more on this soon)

Media reports estimate annual revenue loss around $300 Million (₹2000 crore)

Hard to locate offenders, limited labor to inspect. When found their license revoked.

We show a way of better targeting inspections and finding bogus firms, and (the research design to) test the impact of finding those bogus firms on their trading partners and on tax revenue

The project in a nutshell

Data: Value Added Tax (VAT) returns of all registered private firms

Who sold to whom, for how much, what tax rate? Quarterly.

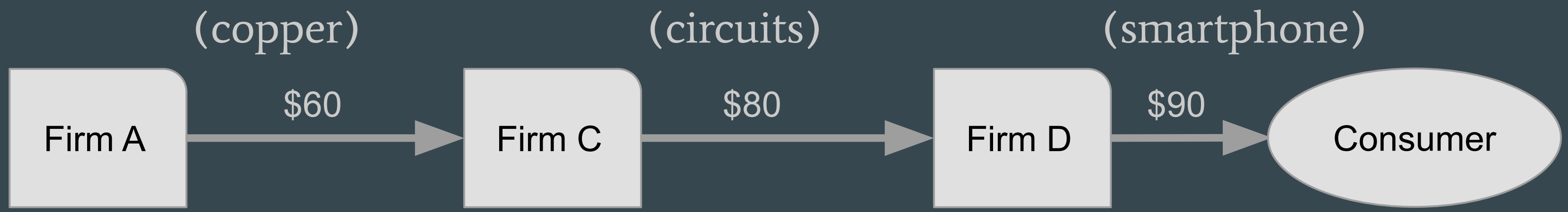
Automatically process the data and identify firms suspected of being bogus

Target them for physical inspections

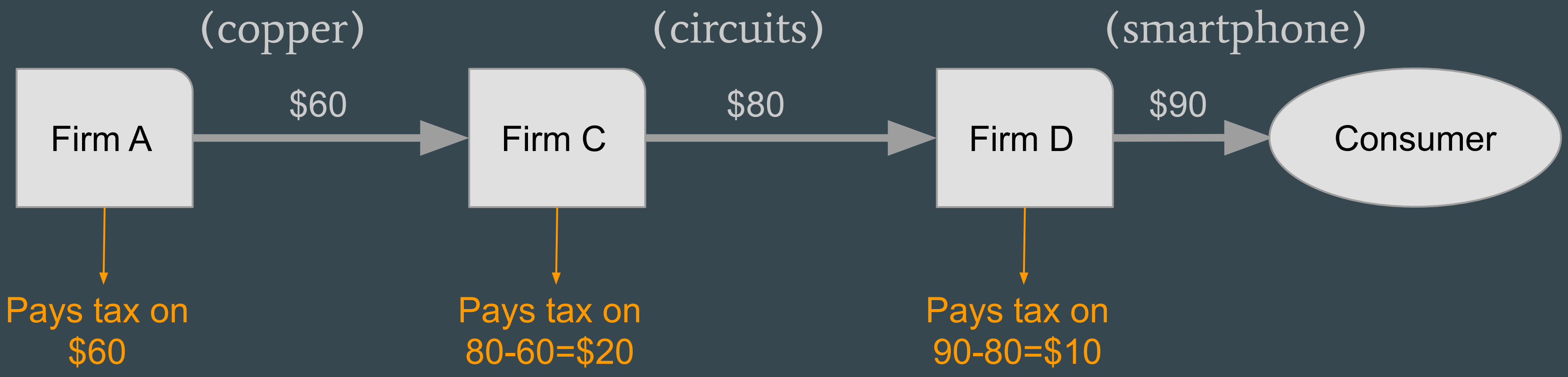
Machine Learning Approach:

Past bogus firms -> what is suspicious behavior -> similar behavior in present -> target

How VAT works

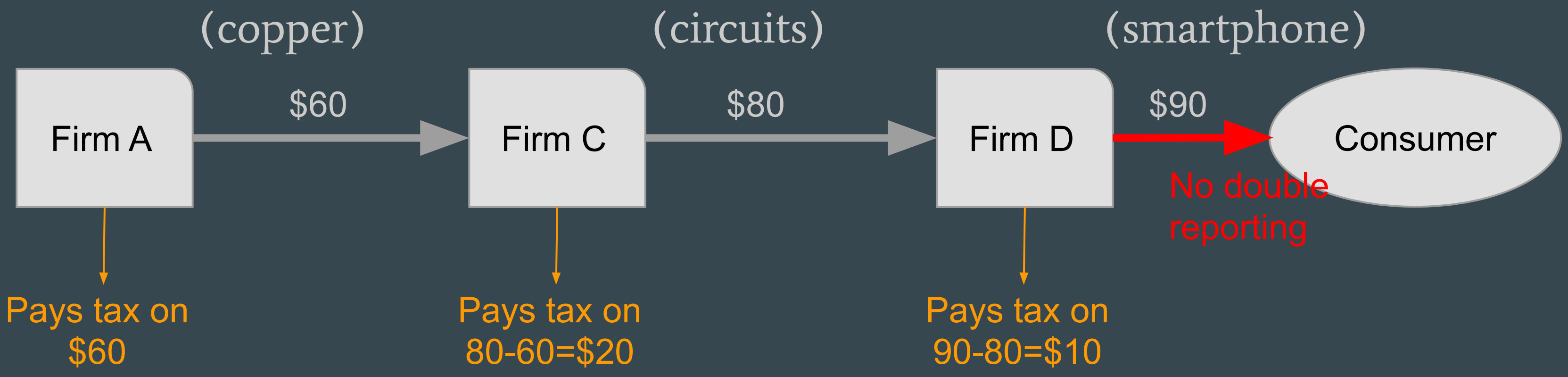


How VAT works



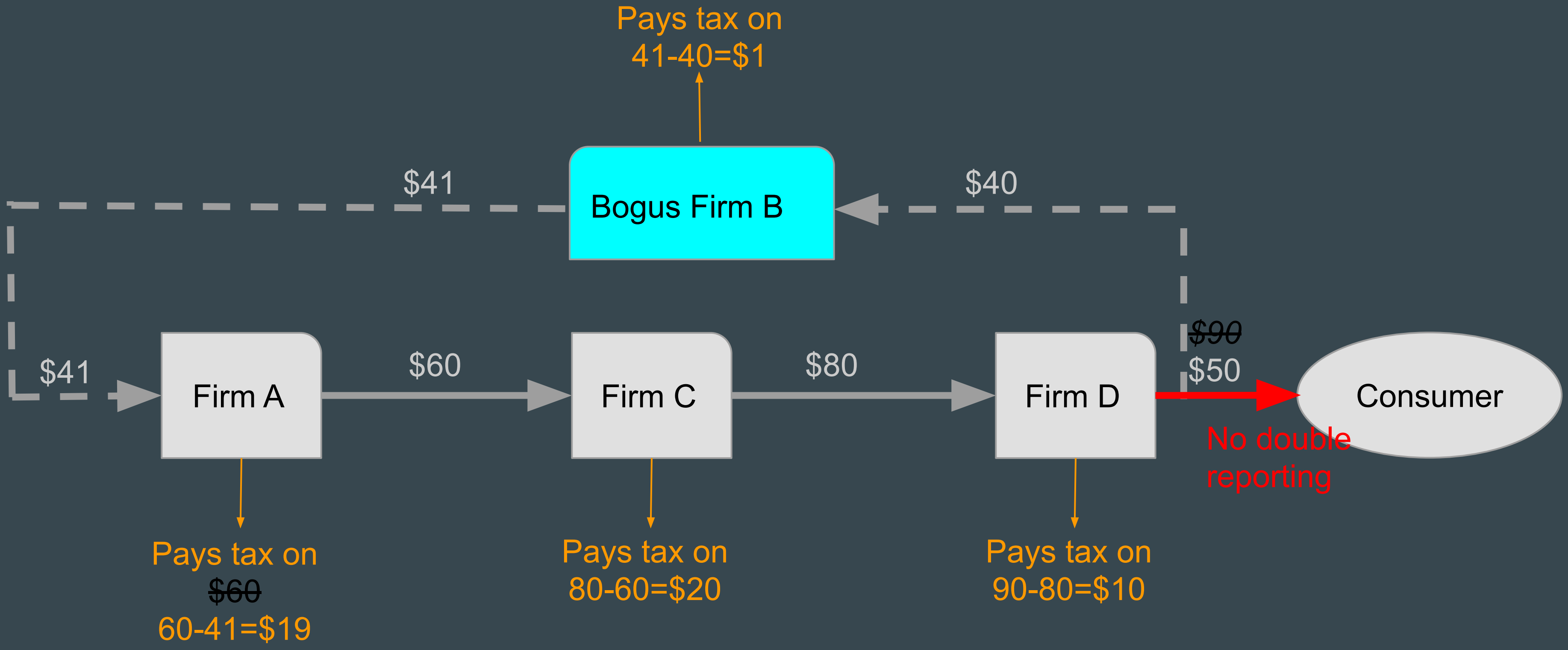
Government receives tax on $90 value added

How VAT works



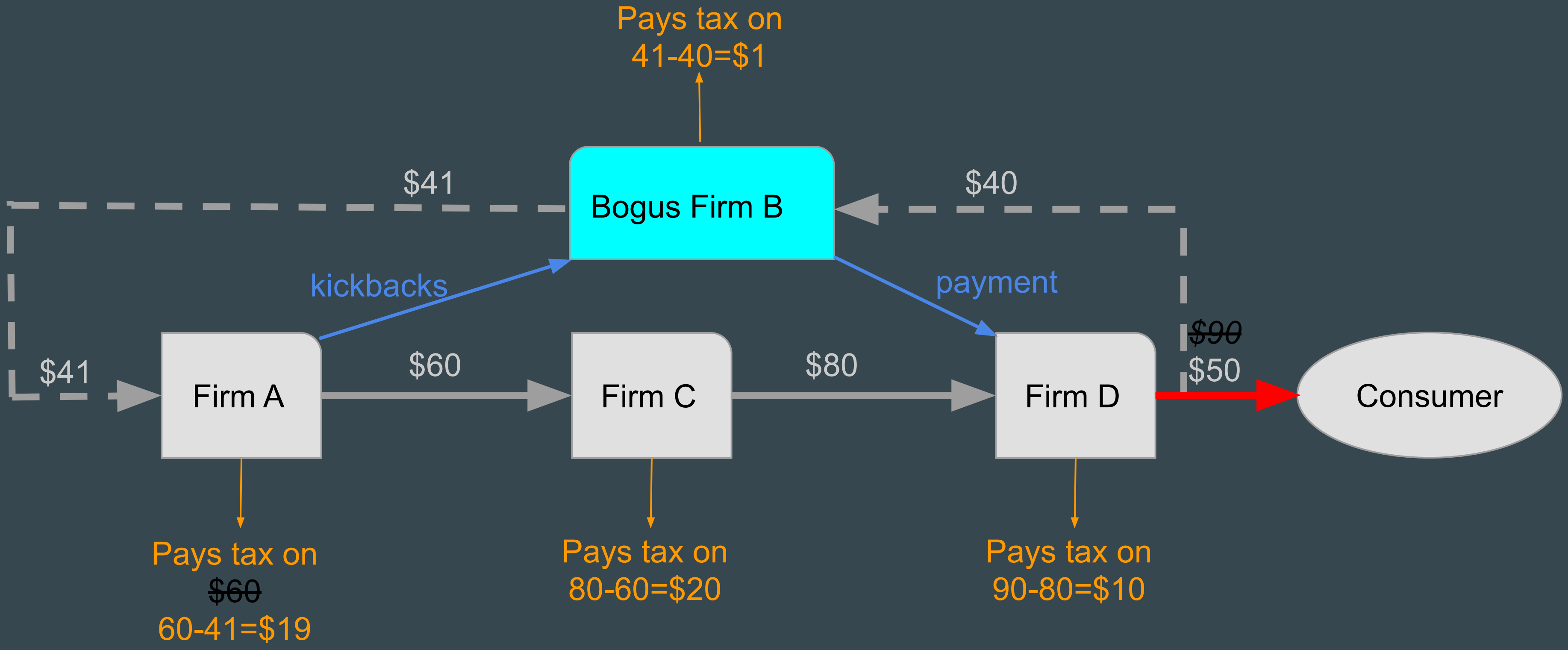
Government receives tax on $90 value added

How VAT evasion works



Government receives tax on $40 less value added

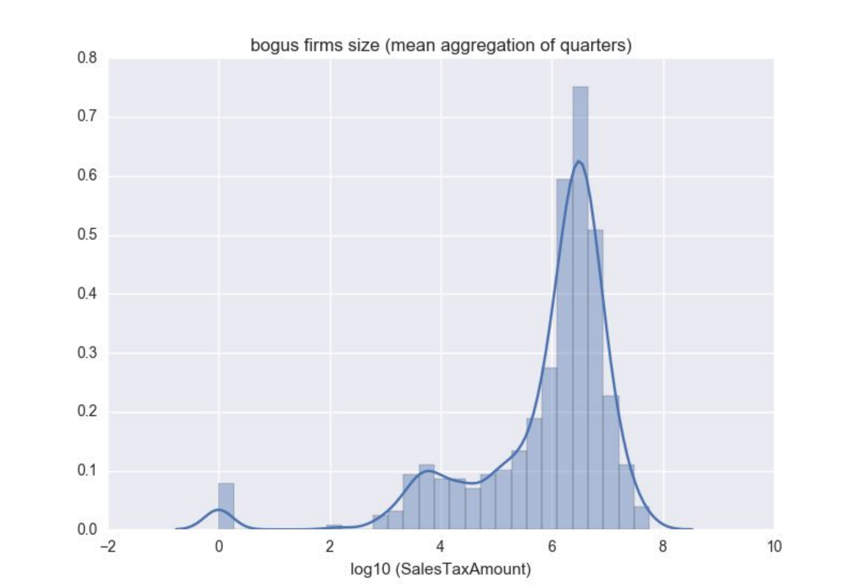
How VAT evasion works

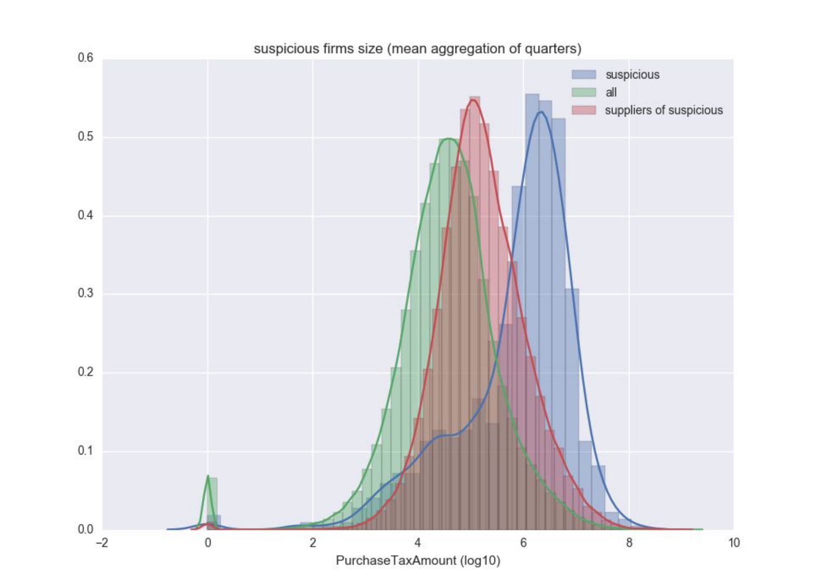


Government receives tax on $40 less value added. Surplus is divided between offenders.

Is it a big problem?

Modal bogus firm pays taxes of ~ ₹10^7 = $150,000 per quarter.
There are hundreds of firms caught in the past.





How the current inspection system works

Irregular, opaque physical inspections to find bogus firms

Targeting: some heuristics about what is suspicious behavior (not validated with data)

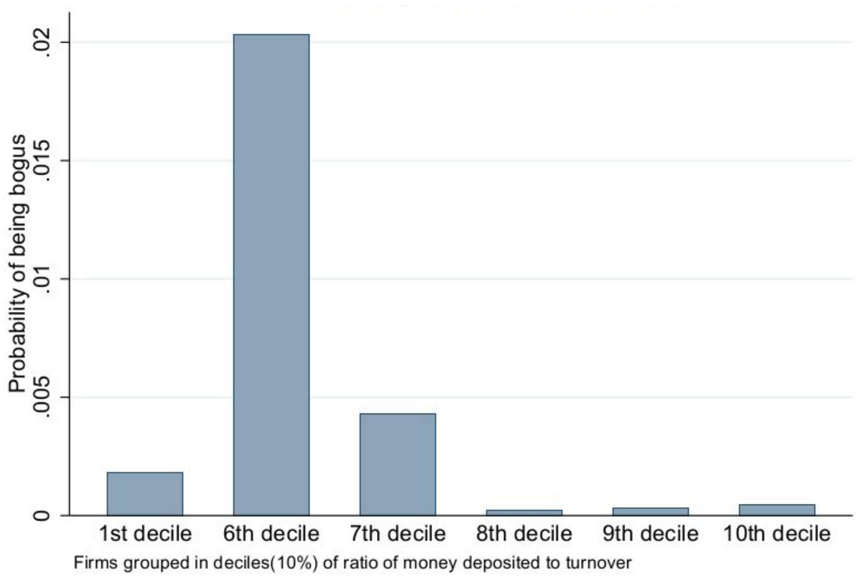
No organized recording of who was inspected, when and why, or of success rate

Only record canceled firms (whose license was revoked), not genuine firms

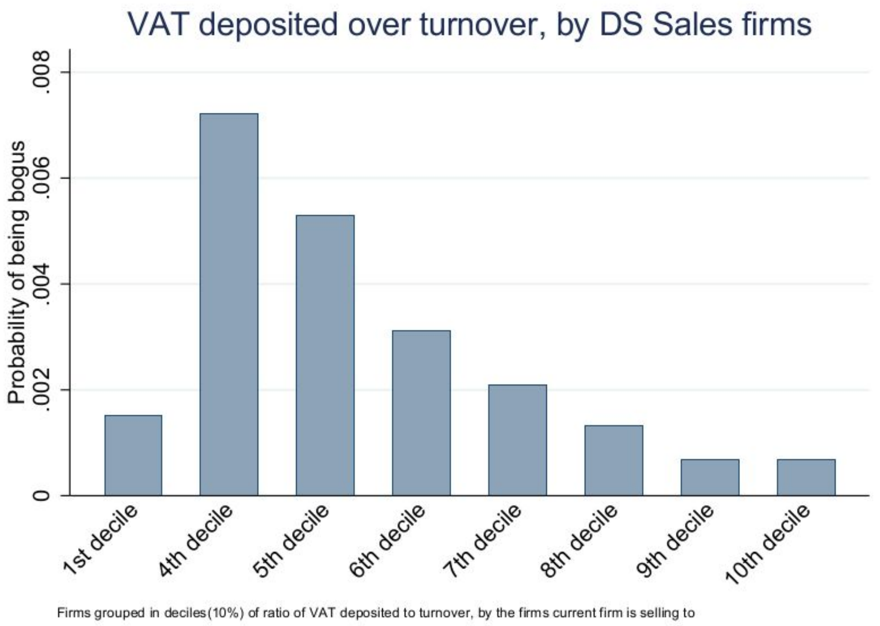
Our approach: rely on data

Past bogus firms -> what is suspicious behavior -> similar behavior in present -> target

Bogus firms tend to have low profit margins

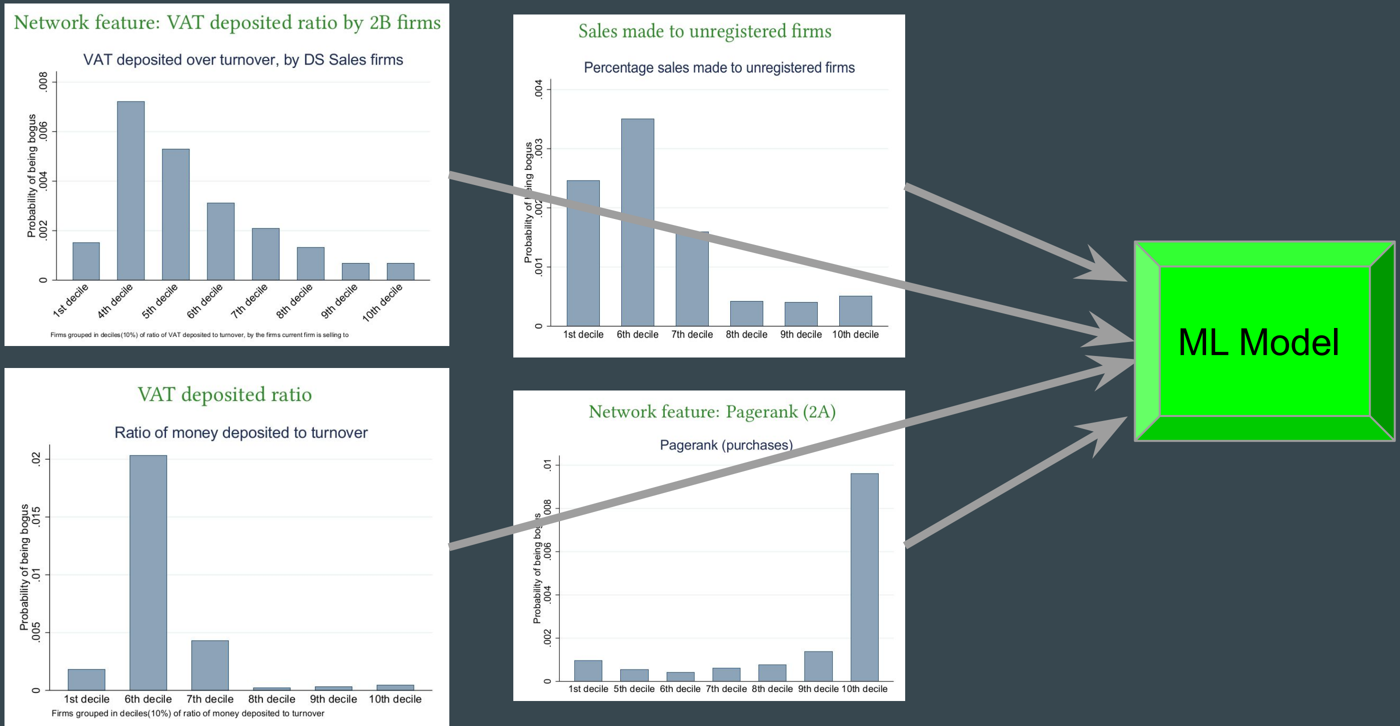


… and trading partners with low profit margins



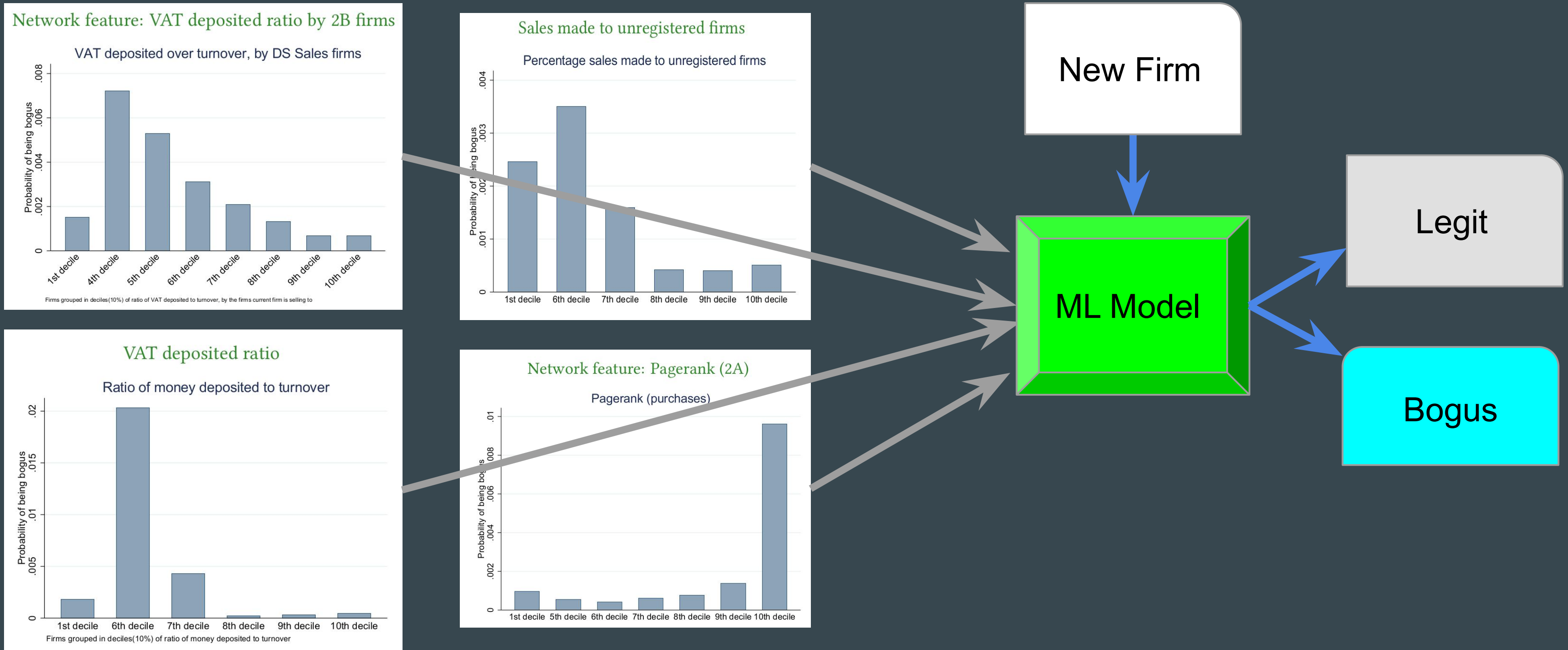
Our approach - Machine Learning

Past bogus firms -> **what is suspicious behavior -> similar behavior in present** -> target



Our approach - Machine Learning

Past bogus firms -> what is suspicious behavior -> **similar behavior in present -> target**



Our approach

Past bogus firms -> what is suspicious behavior -> **similar behavior in present -> target**

results of inspections

Target suspicious firms (by the model prediction) for inspection by the tax authority

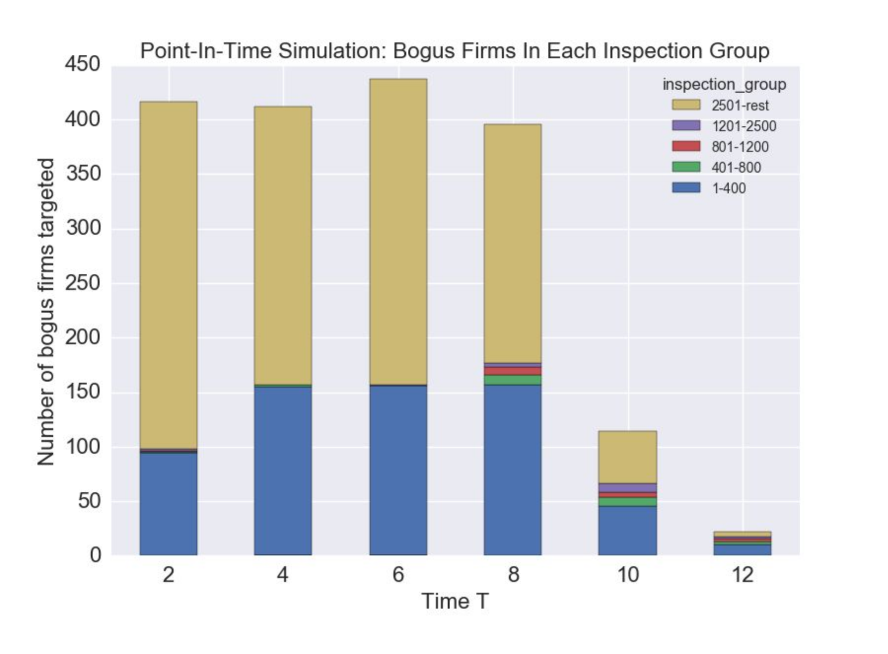
Inspection results -> feed back into the system, improve future prediction

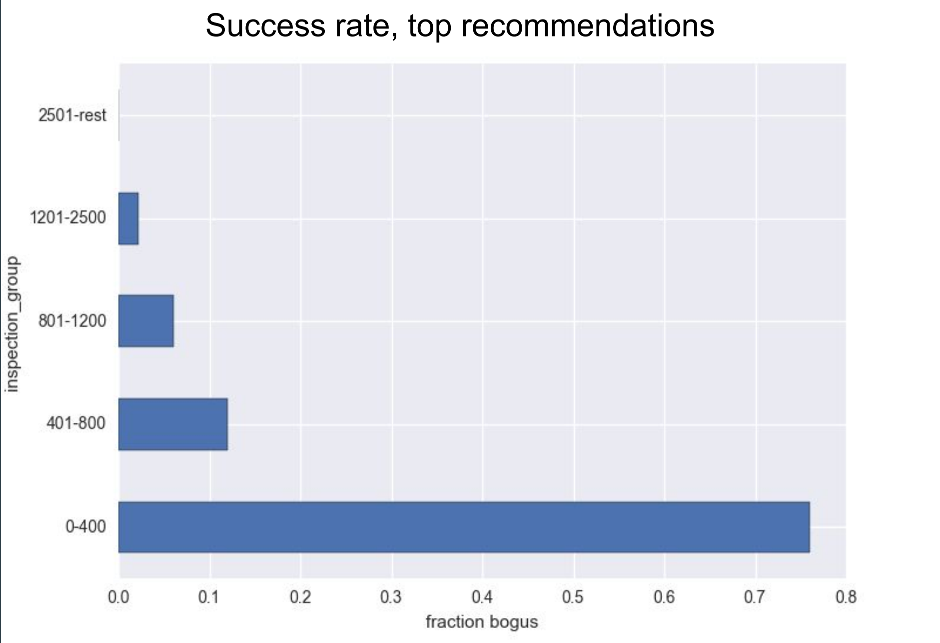
Evaluate impact

Added benefit: objective, fair targeting of inspections

Results

Of the top 400 suspicious firms our model finds, we expect at least 75% to be bogus.



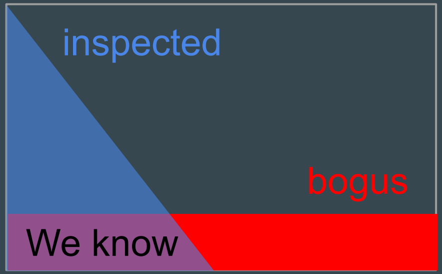


Part I:

Machine Learning

skip

Non-Standard Machine Learning Scenario



Base-rate unknown (class distribution in the population)

Stratify inspections on model score

One-sided labels: only firms that were (non-randomly) inspected and found bogus

Start with predicting (inspected & bogus) vs. the world

Probabilistic labels (via weighting)

Use cross-validation to predict for each unknown firm

Multiple time-periods for each firm

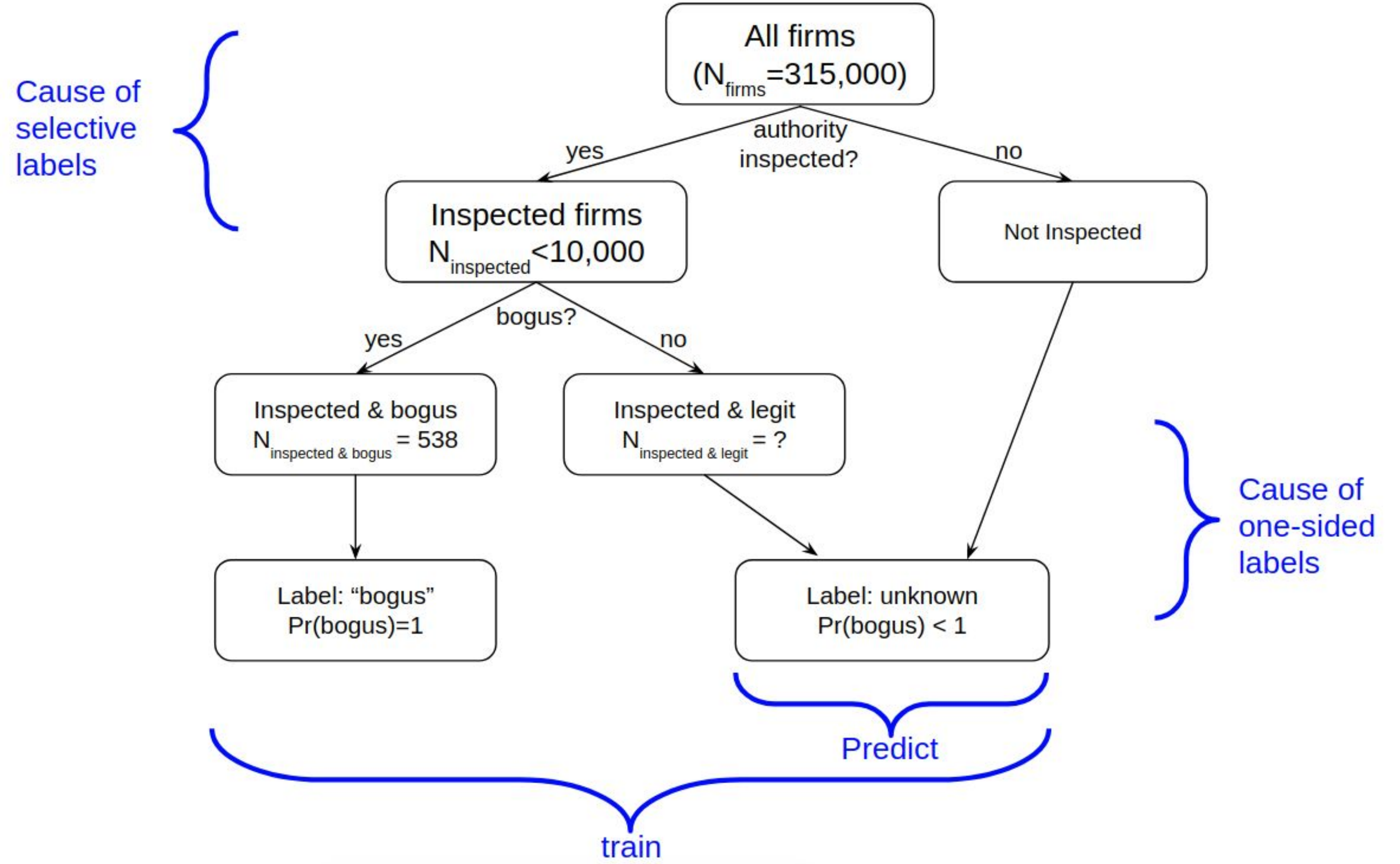
Train & predict for individual time period, perform aggregation. “Convolutional” approach.

Online Machine Learning, chosen targeted labeling

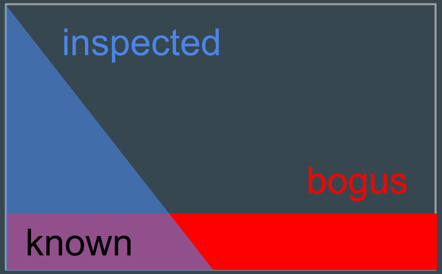
Improve predictions with each round of targeted inspections (target whom?)

We can choose which firms to target

Network Dataset & Features (danger of leakage)

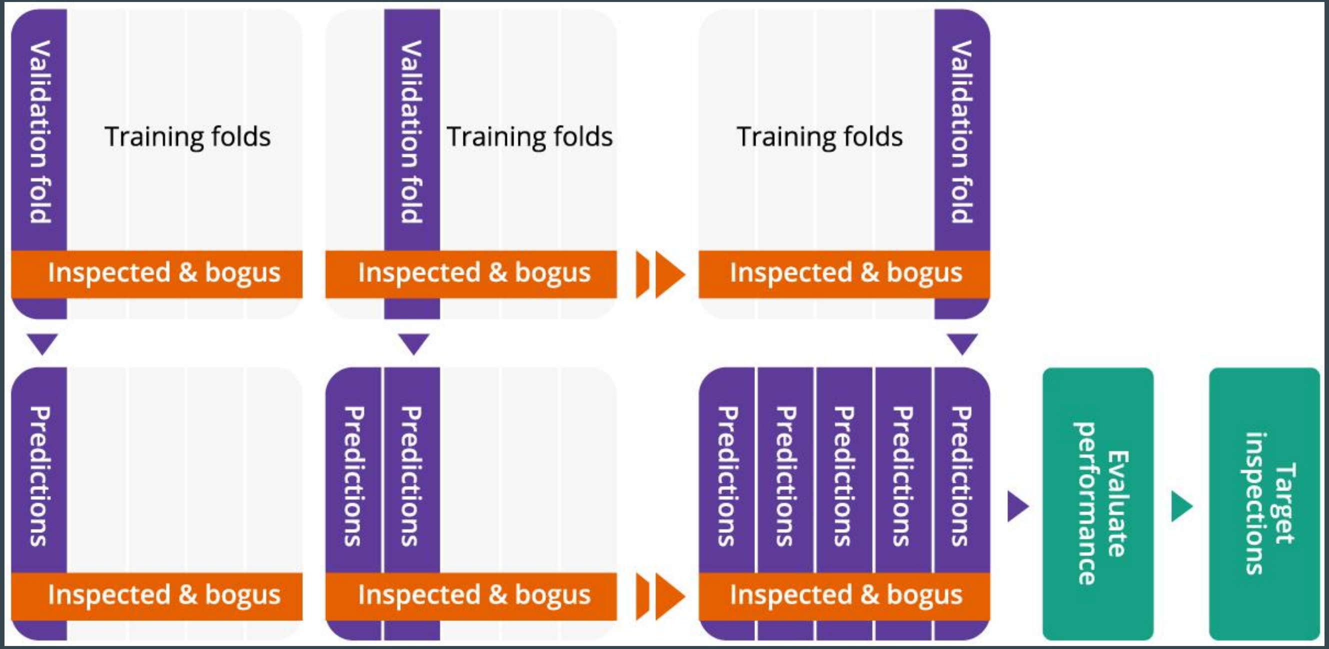

Cause of selective labels
All firms
(N_{firms}=315,000)
authority inspected?
yes
no
Inspected firms
$N_{inspected}$<10,000
Not Inspected
bogus?
yes
no
Inspected & bogus
$N_{inspected\ \&\ bogus}$ = 538
Inspected & legit
$N_{inspected\ \&\ legit}$ = ?
Cause of one-sided labels
Label: “bogus”
Pr(bogus)=1
Label: unknown
Pr(bogus) < 1
Predict
train

One-sided labels - solution

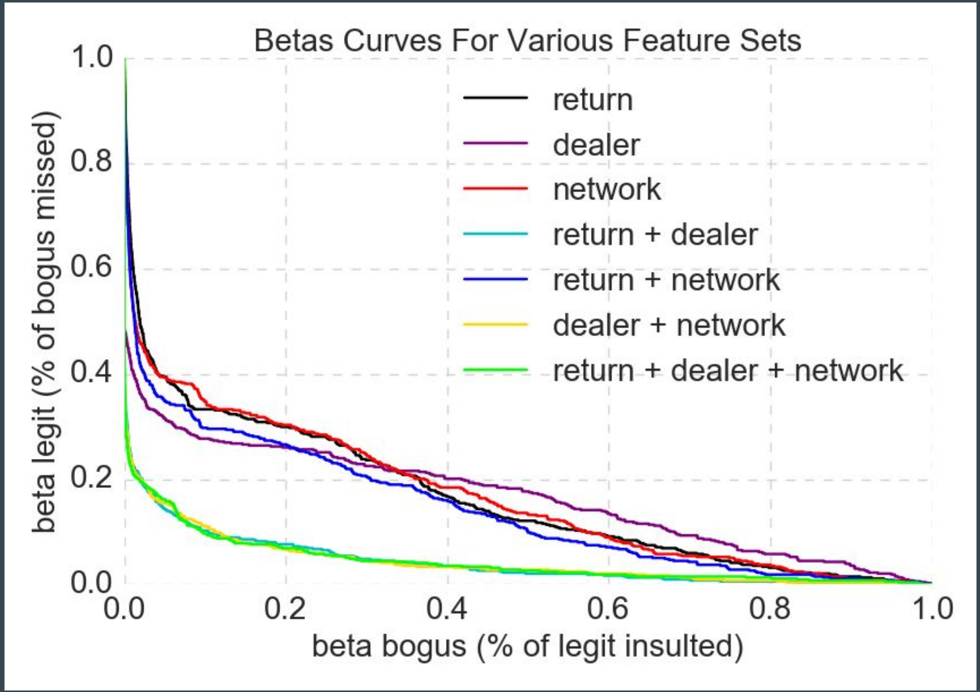


Start with predicting (inspected & bogus) vs. the world

- Use cross-validation to predict for each unknown firm (when it is in the validation fold)
- Evaluate performance



Sophisticated features help prediction



Point-in-time simulation

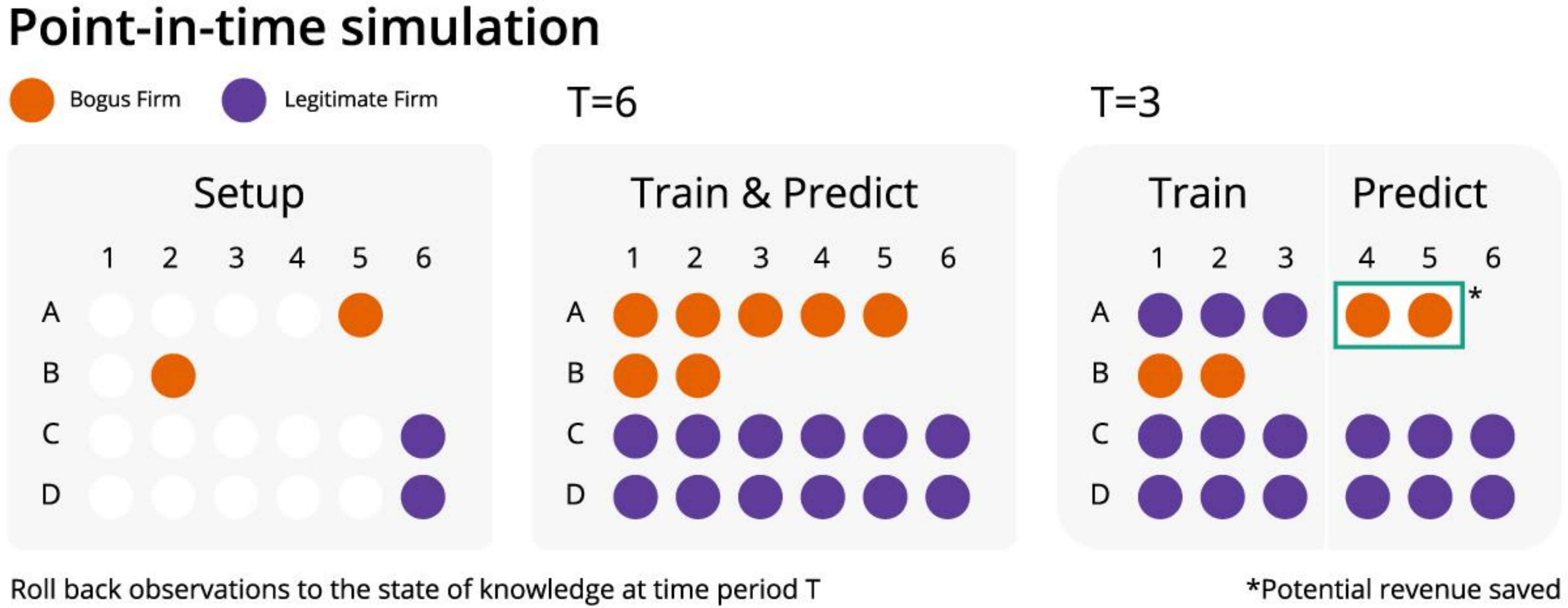
How to know not how well we do, but how well we would have done compared to the status quo?

Roll back observations to the state of knowledge at time period T:
- Drop all observations with time t > T.
- Firms canceled before/on time T are labeled “bogus”
- Firms canceled after time T (and those never canceled) are labeled “legit”

Run our training and prediction process on firms still operating at time T.

    How many do we catch that later turned out to be bogus?
    How many months of operations / lost revenue would we save? (displacement?)

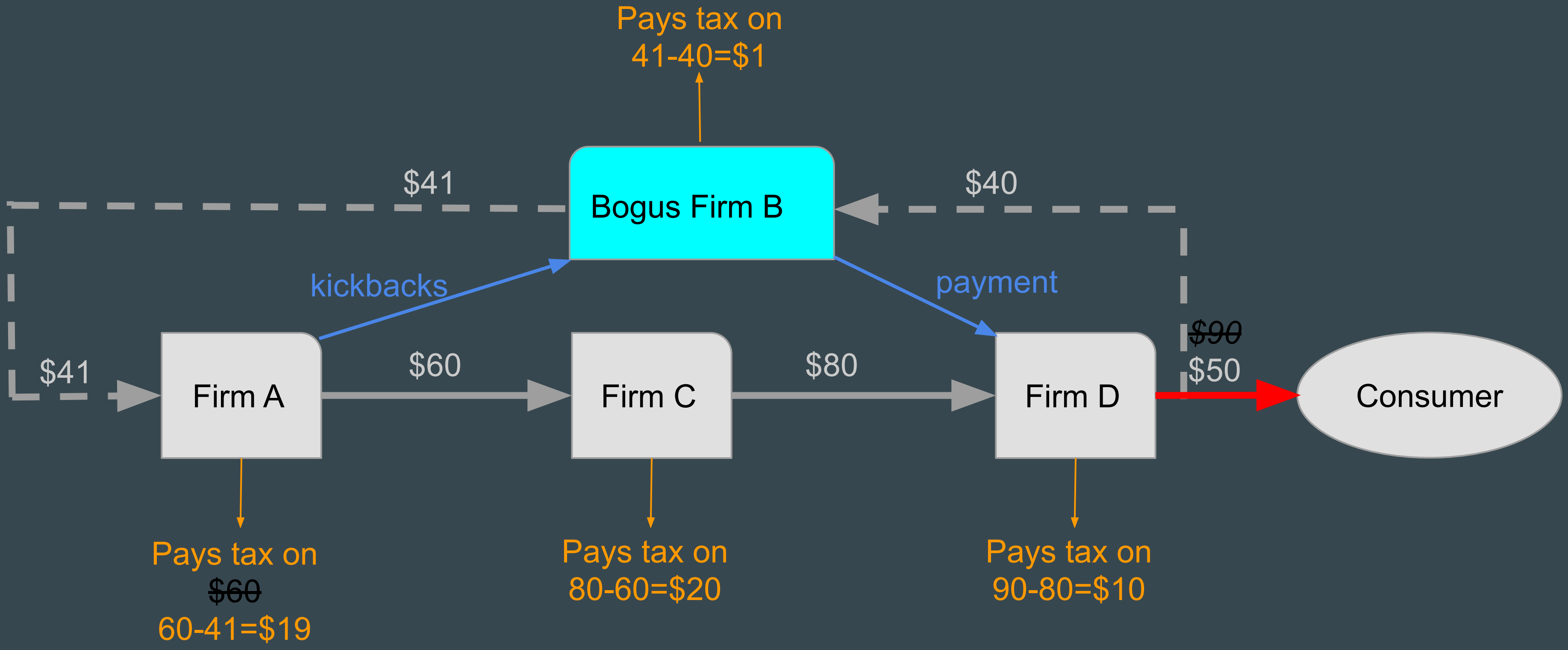
Point-in-time simulation



Part II:

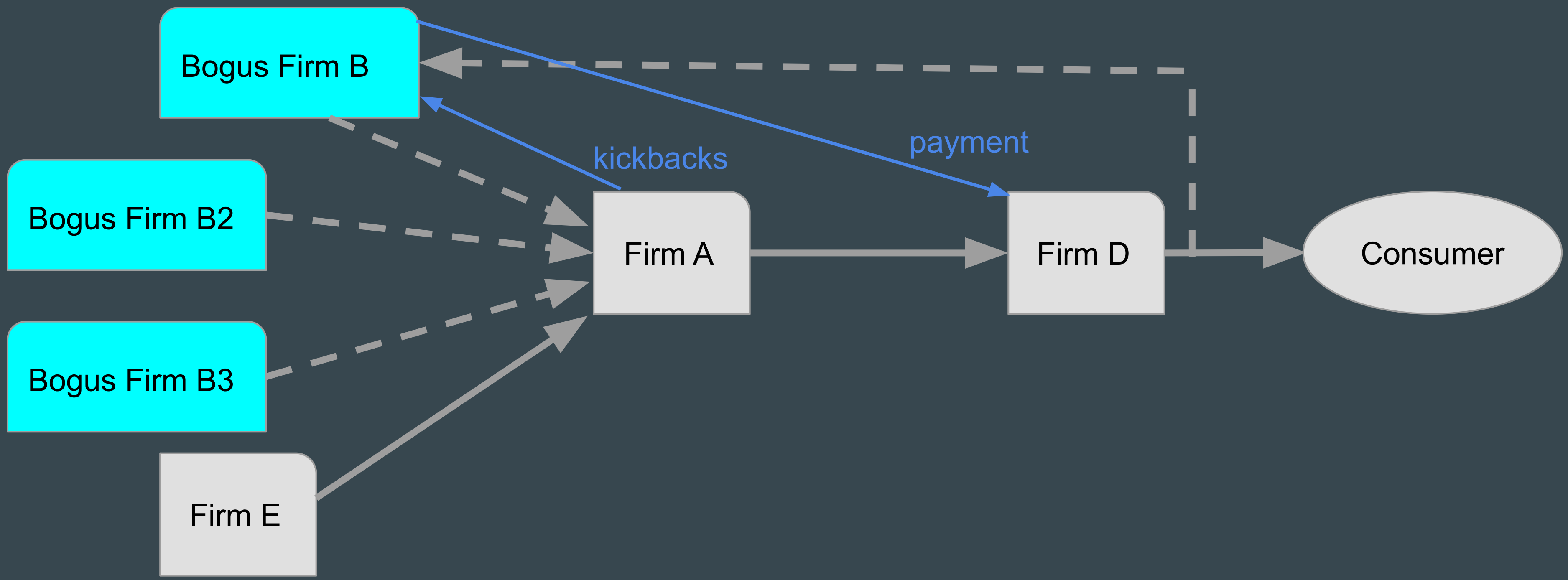
Economics

How VAT evasion works

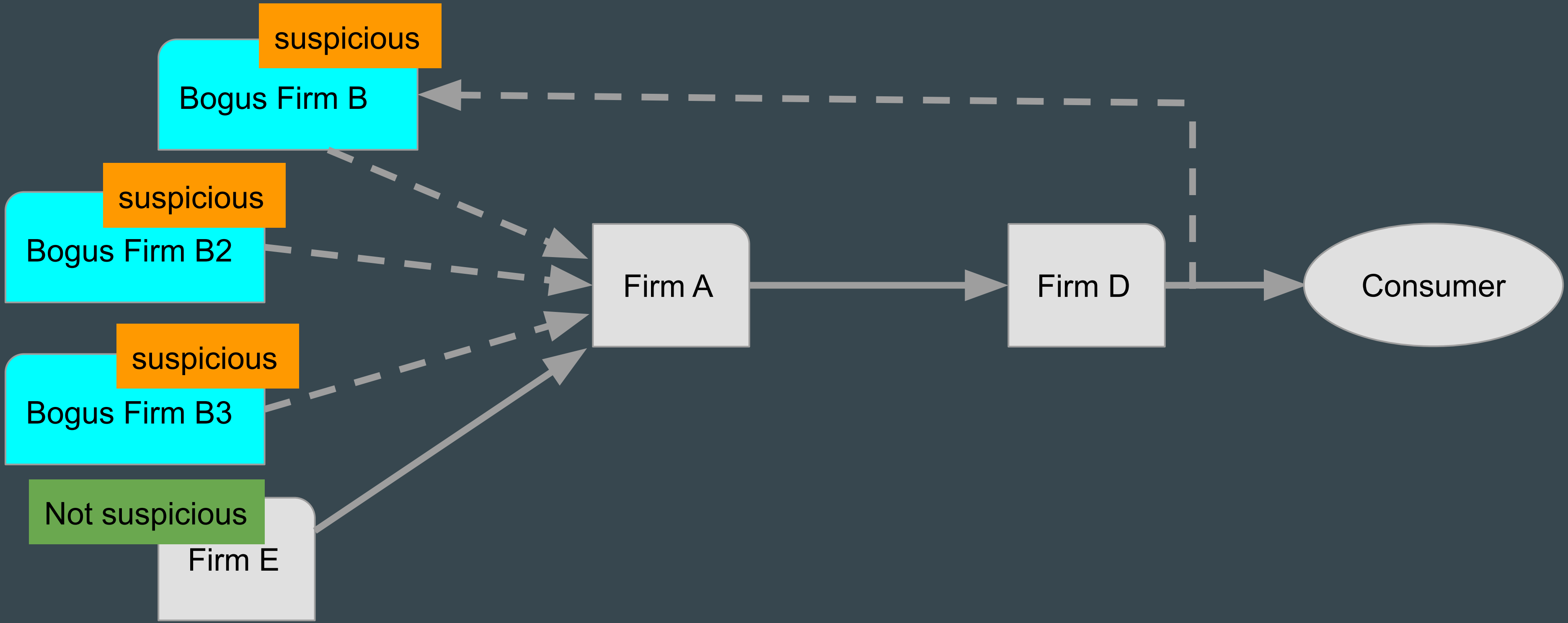


Government receives tax on $40 less value added. Surplus is divided between offenders.

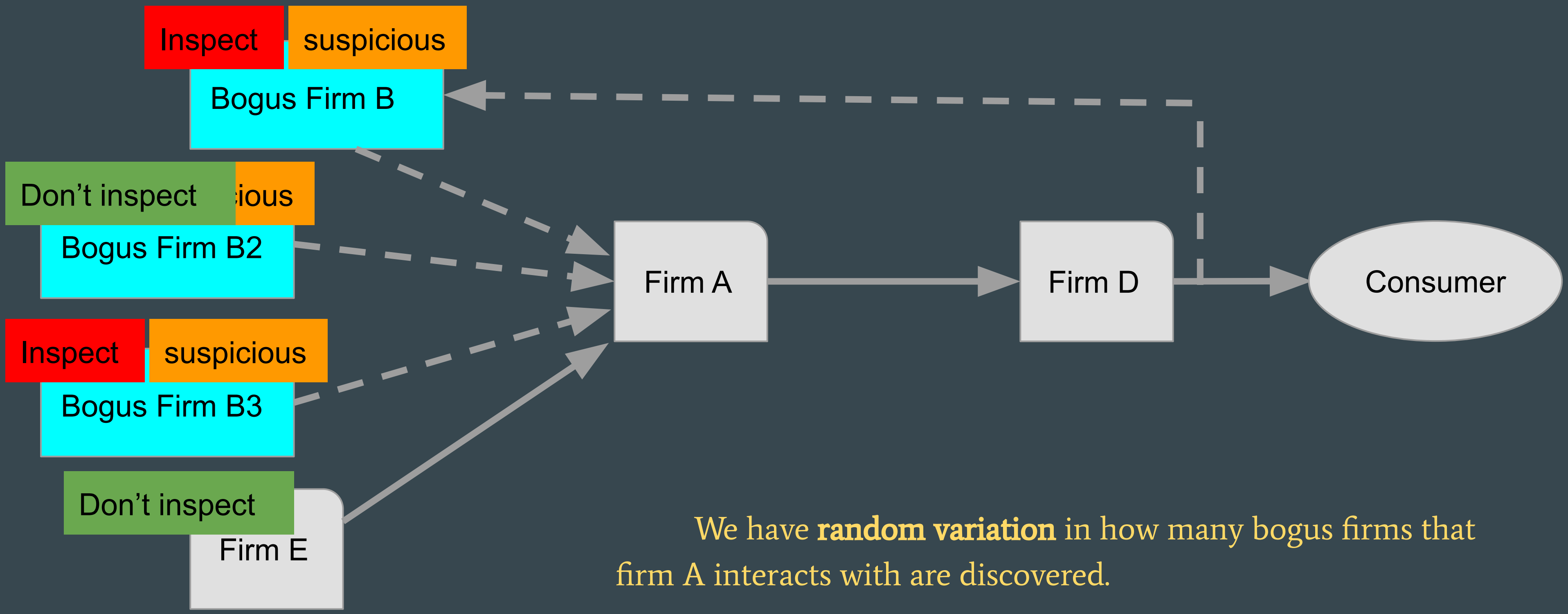
Economic research plans



Economic research plans - recommendations

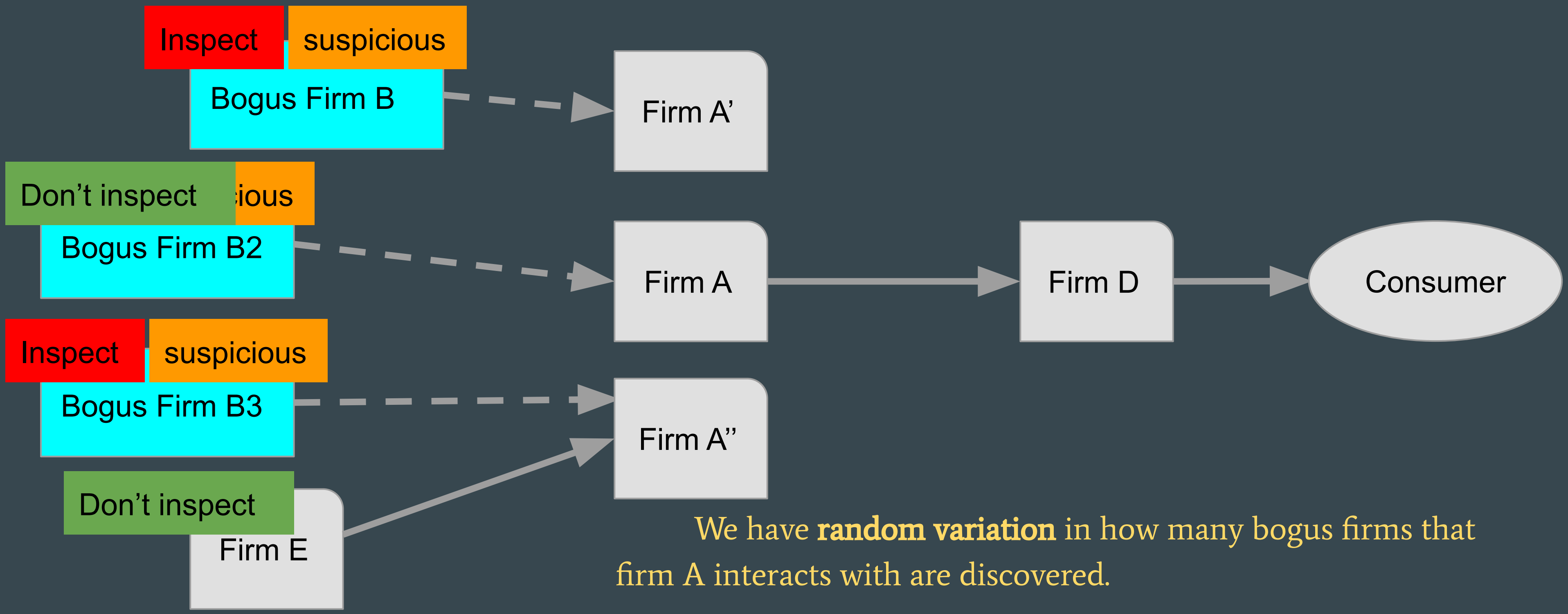


Economic research plans - random variation in enforcement



We have **random variation** in how many bogus firms that firm A interacts with are discovered.

Economic research plans - random variation in enforcement



We have **random variation** in how many bogus firms that firm A interacts with are discovered.

Research Questions - relevant to corruption

Tax revenue collected (if not, where does it break down?), cost-benefit

Deterrence or substitution? Firm A & Firm D

Long-term effects on firms:

- Deterrence effect of enforcement
- Spillovers to network, information propagation

Random variation in tax burden - deadweight loss, price effects, revenue

Recap

Use the universe of VAT returns in Delhi to find tax evaders ("bogus firms")

Machine Learning approach:
Past bogus firms -> what is suspicious behavior -> similar behavior in present -> target

Show high accuracy, millions of $ in potential revenue

E-auditing - a general anti-corruption approach

E-auditing:

<u>**Digital "paper trail" + ML => monitoring of service provision**</u>

Teacher attendance - mobile phone call records
Health workers give vaccines - electronic immunization cards/app
Welfare payments delivered - Aadhaar records
Collusion in public procurement - public records of auctions

. . .

Corruption + digital data = potential for e-auditing
(Past bogus firms -> what is suspicious behavior -> similar behavior in present -> target)

Thanks!

e.scribbler@gmail.com