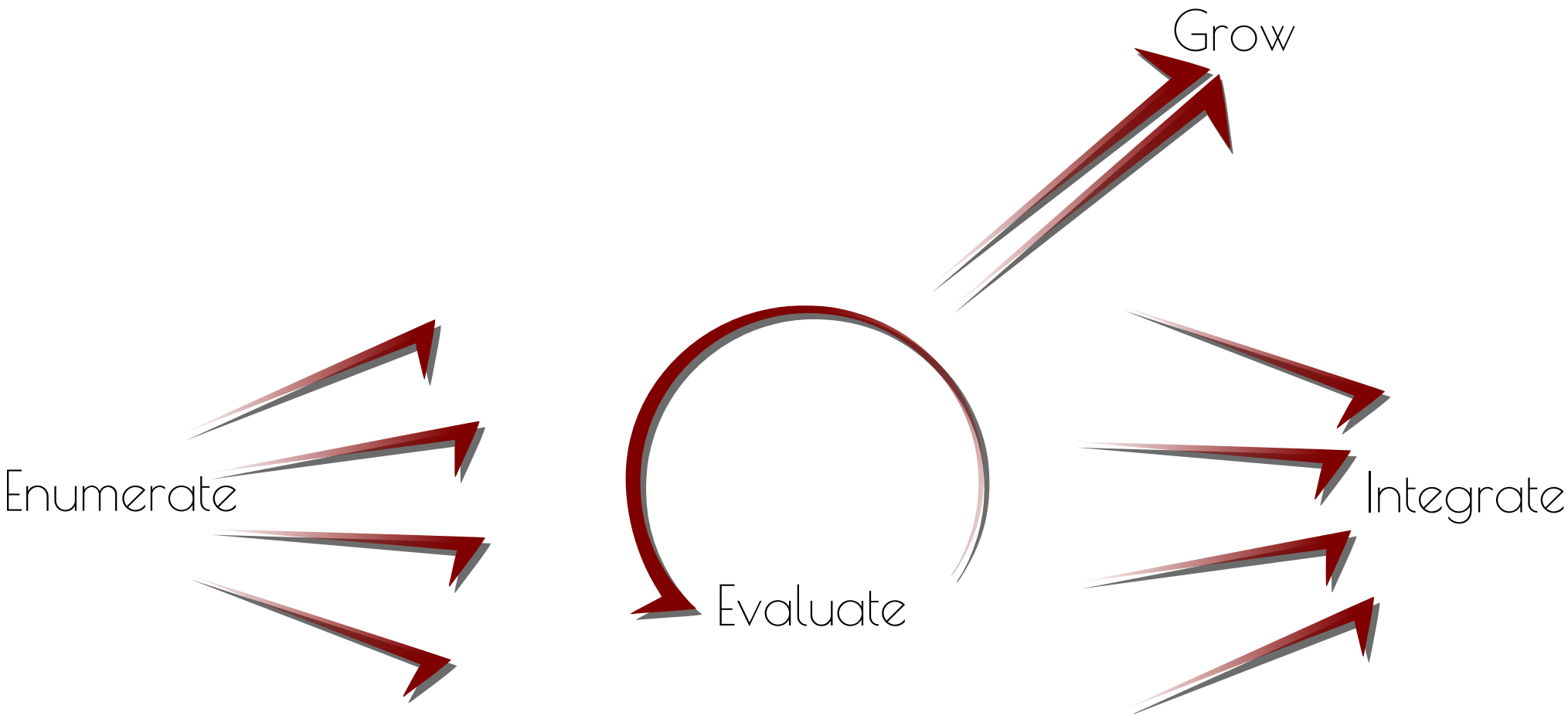
The Data Science Process

- Or how I contextualize all the work I do.

Allen Grimm
Grimm Science
allen@GrimmScience.com
#GrimmScientist


Grow
Enumerate
Evaluate
Integrate

Enumerate

Enumerate Points of Leverage

- This is a business problem.
- “Where does decision making and data intersect?”
- Examples:
  - Made-up bank: Fraud detection
  - UpSight: Churn Prediction
  - U.S. Bank: Uplift Modeling

Example: Fraud Detection

- Data Available:
  - Historic monthly sales data.
  - 1M Customers
- Business Problem:
  - Identity theft happens to 0.1% of users yearly
  - Avearge $5k in losses per attack
  - Results in annual damages of $500M damages
- Goal:
  - Early fraud detection

Example: Churn Prediction

- Data:
  - 5k integrated games
  - 130M unique users
  - 2.5B game sessions
- Business Problem:
  - User acquisition is expensive
- Goal:
  - Retain more of the current users by detecting and taking action on users likely to leave

Example: Uplift Modeling

- Data Available:
  - Banking data for all users
    - (currently) Over 15MM customers
    - (currently) Process 4B checks annually
    - (at the time) Held $282B in assets
- Business Problem:
  - Failing Marketing Campaigns
- Goal:
  - Better targeting of users


Evaluate

Evaluate Points of Leverage

- This is a machine learning problem
- Questions:
  - “How should the data be modeled?”
  - “What expected return can I expect on model application?”
- Deliverables:
  - Estimated model accuracy
  - Projected value of initiative

Example: Fraud Detection

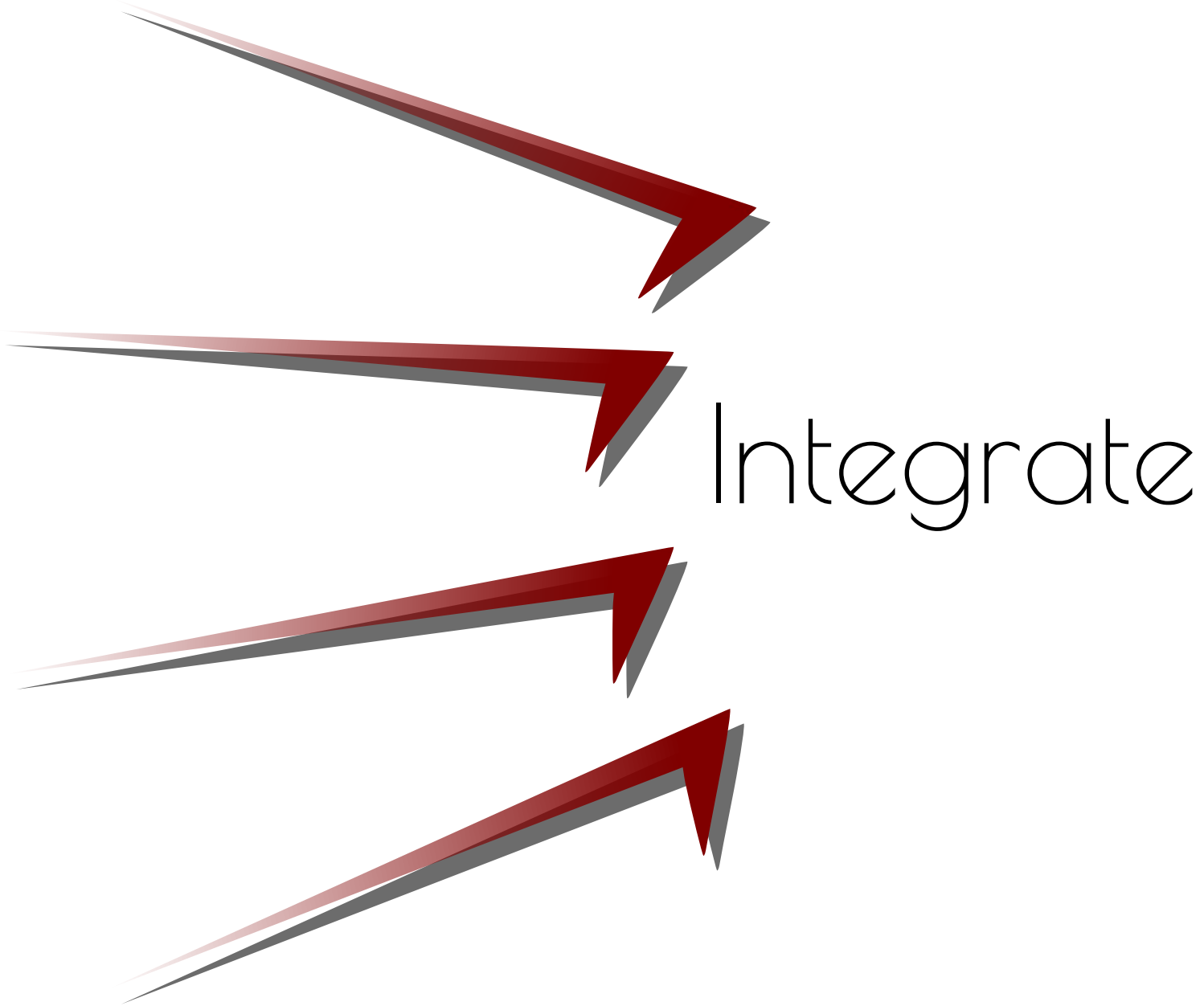
- Fraud detection is usually done by first learning “normal behavior” then looking for deviations from normal.
- 70% fraud detection within first 3 days
- Projected savings of 500M * .7 = 350M

Example: Churn Prediction

- Historic time series in, expected future time series out
- Technique: Discrete Multivariate Modeling
- Toolset:
  - Raw data pre-aggregated by Redshift
  - Model initially built and ran in pure python
  - Eventually started writing bottlenecks in hadoop streaming
- Model had over 75% accuracy

Example: Uplift Modeling

- Uplift Modeling Algorithm:
  1) Take sample of users
  2) Try various content on sampled users
  3) Apply most successful content to rest of users
- (Case study didn't give Evaluate details beyond their model being sufficient to Integrate)


Integrate

Integrate Initiative Into Product

- This is an engineering problem.
- Highest commitment, but also the point.
- “How can the prescribed model best be integrated into our system?”
- Deliverable:
  - Automated re-training and application of the model, as applied to the chosen business problem.

Example: Fraud Detection

- 30% attacks not caught averaging $5k in damages
- 70% attacks caught early averaging $1k in damages
- Say, 5MM expense of bulding infrastructure to freeze accounts.
- Net damages now 225M (down 50%)

Example: Churn Prediction

- Last I heard, not actually integrated.
- Intended to mark certain users as most likely to leave, allowing for a highly-targeted uplift modeling initiative. End goal was to increase user engagement.

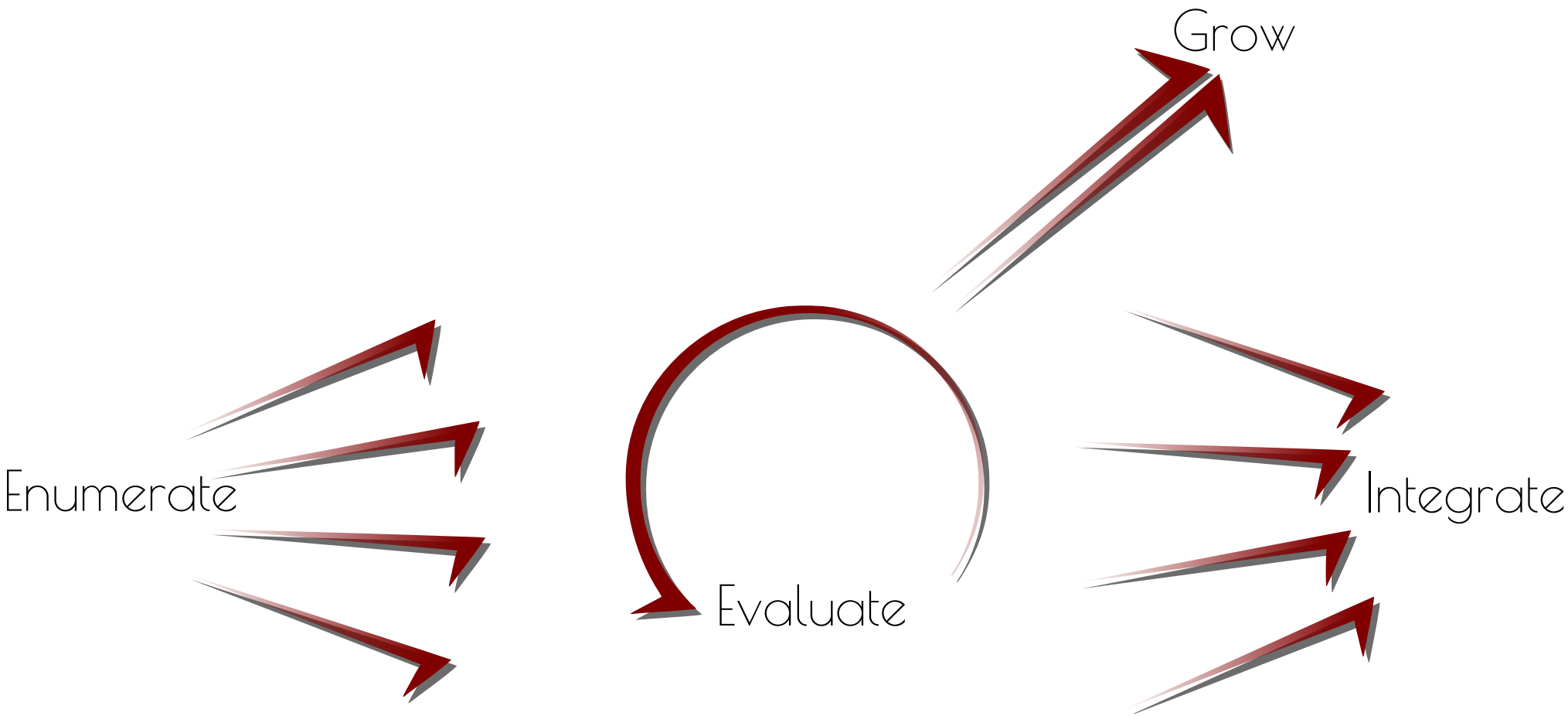
Example: Uplift Modeling

- 300% increase to cross-sell revenue
- Over 1MM in increased revenue from initial campaigns
- 73% increase in direct deposit usage
- 40% less mail sent
- Model reported to be used elsewhere in the business, too


Grow

Grow Your Data Science Team

- This is a management problem.
- Enumerate:
  - Need to speak both business and data at the basic level
  - Experience with data structures and basic modeling techniques.
- Evaluate:
  - The broader the machine learning experience, the better for enabling the enumerate step.
  - The deeper the machine learning experience, the faster you'll generate nice, niche, models.
- Integrate:
  - Engineers with enough math to understand the Evaluate Reports?
  - Machine learners the programming chops to write product code? Your choice.


Grow
Enumerate
Evaluate
Integrate

Takehomes:

- Enumerate – Evaluate – Integrate – Grow
- Keep data science grounded in business
- Iterating on an already Evaluated problem is done at the expense of Iterating on a new problem
- For the new data scientist:
  - Find an interesting dataset (from your work?)
  - Install Anaconda (specialized Python dist.)
  - Play with your data and SKLearn
- PortlandDataScience.com

References

- Churn Prediction Presentation at PyData:
  - http://vimeo.com/79533999
- U.S. Bank Case Study:
  - Current stats: en.wikipedia.org/wiki/U.S._Bancorp
  - Case study: bit.ly/US_bank_uplift_modeling
- Anaconda:
  - http://continuum.io/downloads