Does Peer-Reviewed Research Help Predict Stock Returns?

Andrew Y. Chen (Federal Reserve Board)
Alejandro Lopez-Lira (University of Florida)
Tom Zimmermann (University of Cologne)

NBER SI AP - Cambridge – July 2024

Our question:

- Suppose a Ph.D. student says “I found a predictor with a t-stat > 2.0 and a sample mean return of 100 bps!”

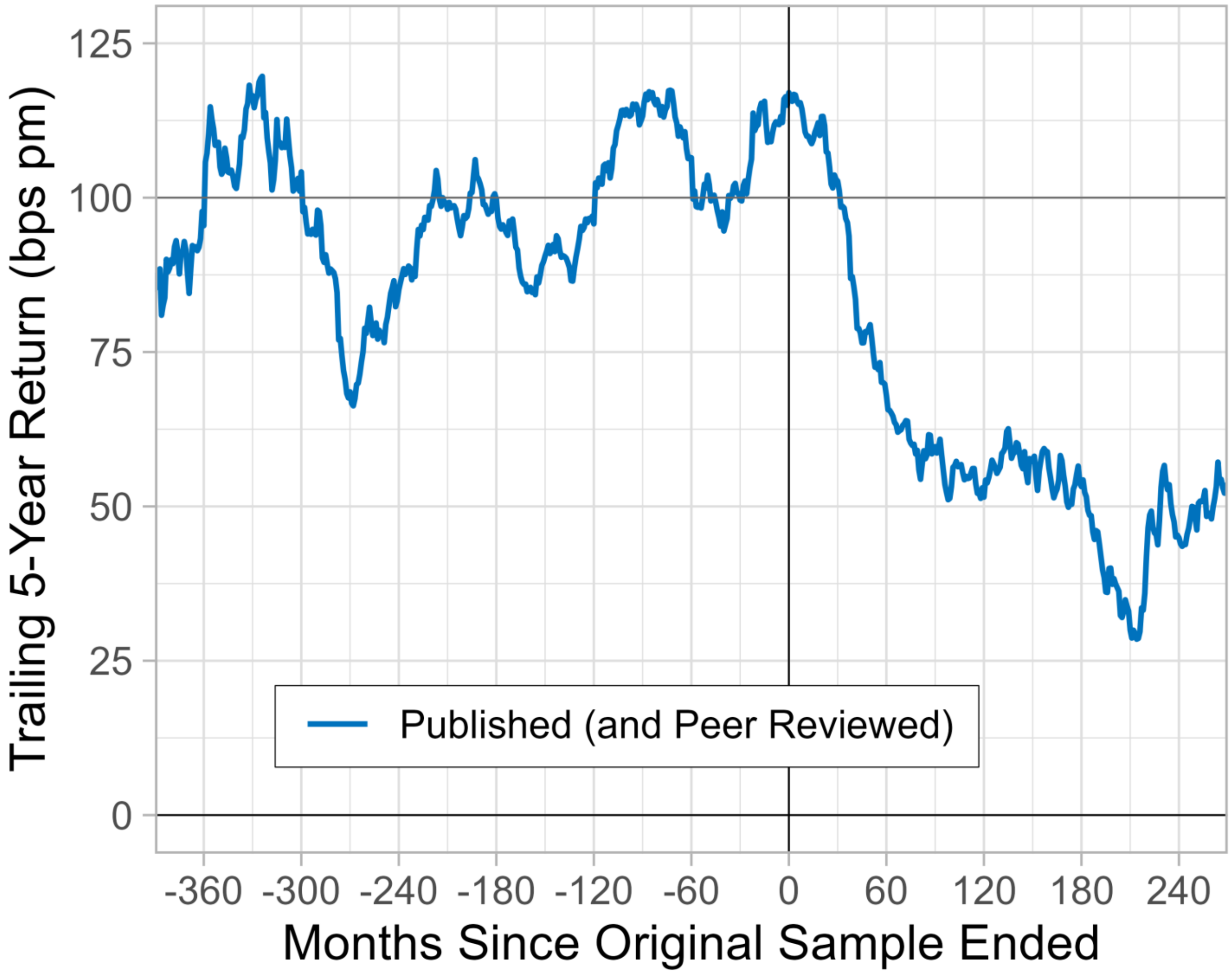
Our question:

- Suppose a Ph.D. student says “I found a predictor with a t-stat > 2.0 and a sample mean return of 100 bps!”
- You ask, “where does this predictor come from?”
  1. Was it based on an idea that is publishable in a top finance journal?
  2. Or did you just mine accounting ratios for t > 2?

Our question:

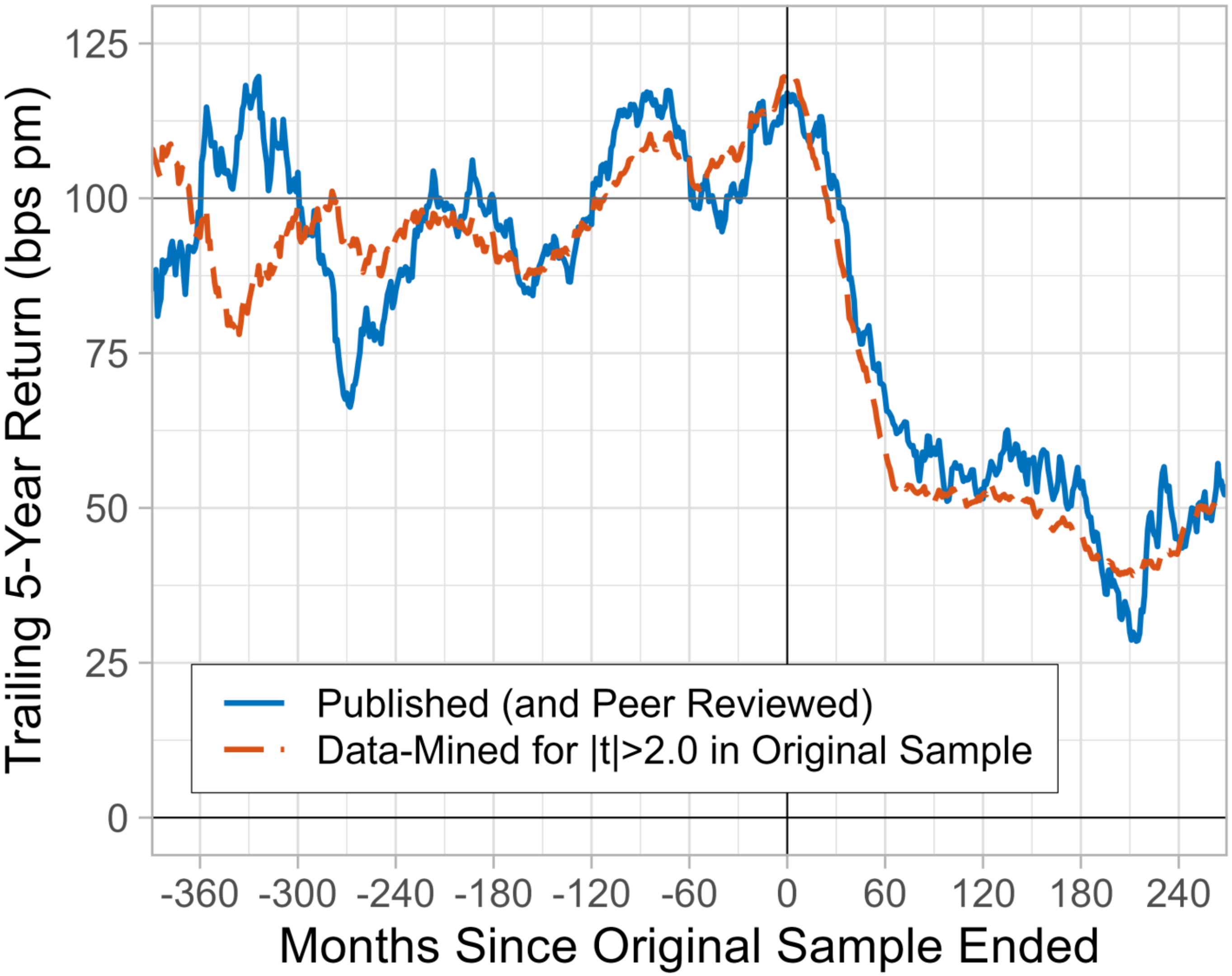
- Suppose a Ph.D. student says "I found a predictor with a t-stat > 2.0 and a sample mean return of 100 bps!"
- You ask, "where does this predictor come from?"
    1. Was it based on an idea that is publishable in a top finance journal?
    2. Or did you just mine accounting ratios for t > 2?
- **How should your expected out-of-sample return depend on his answer?**

Our answer:



Trailing 5-Year Return (bps pm)
125
100
75
50
25
0
-360 -300 -240 -180 -120 -60 0 60 120 180 240
Months Since Original Sample Ended
Published (and Peer Reviewed)

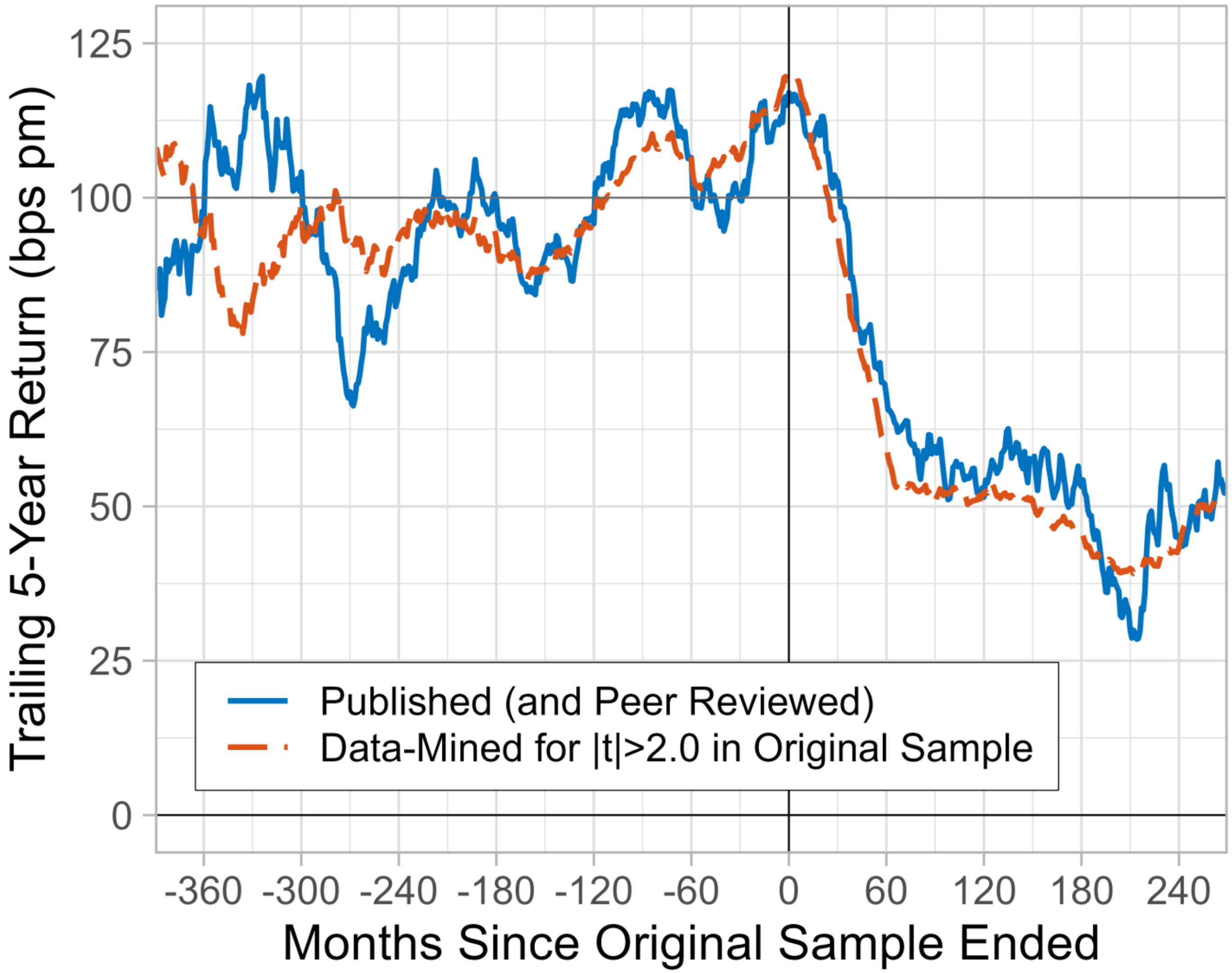
Our answer:



Trailing 5-Year Return (bps pm)
125
100
75
50
25
0
-360
-300
-240
-180
-120
-60
0
60
120
180
240
Months Since Original Sample Ended
Published (and Peer Reviewed)
Data-Mined for |t|>2.0 in Original Sample

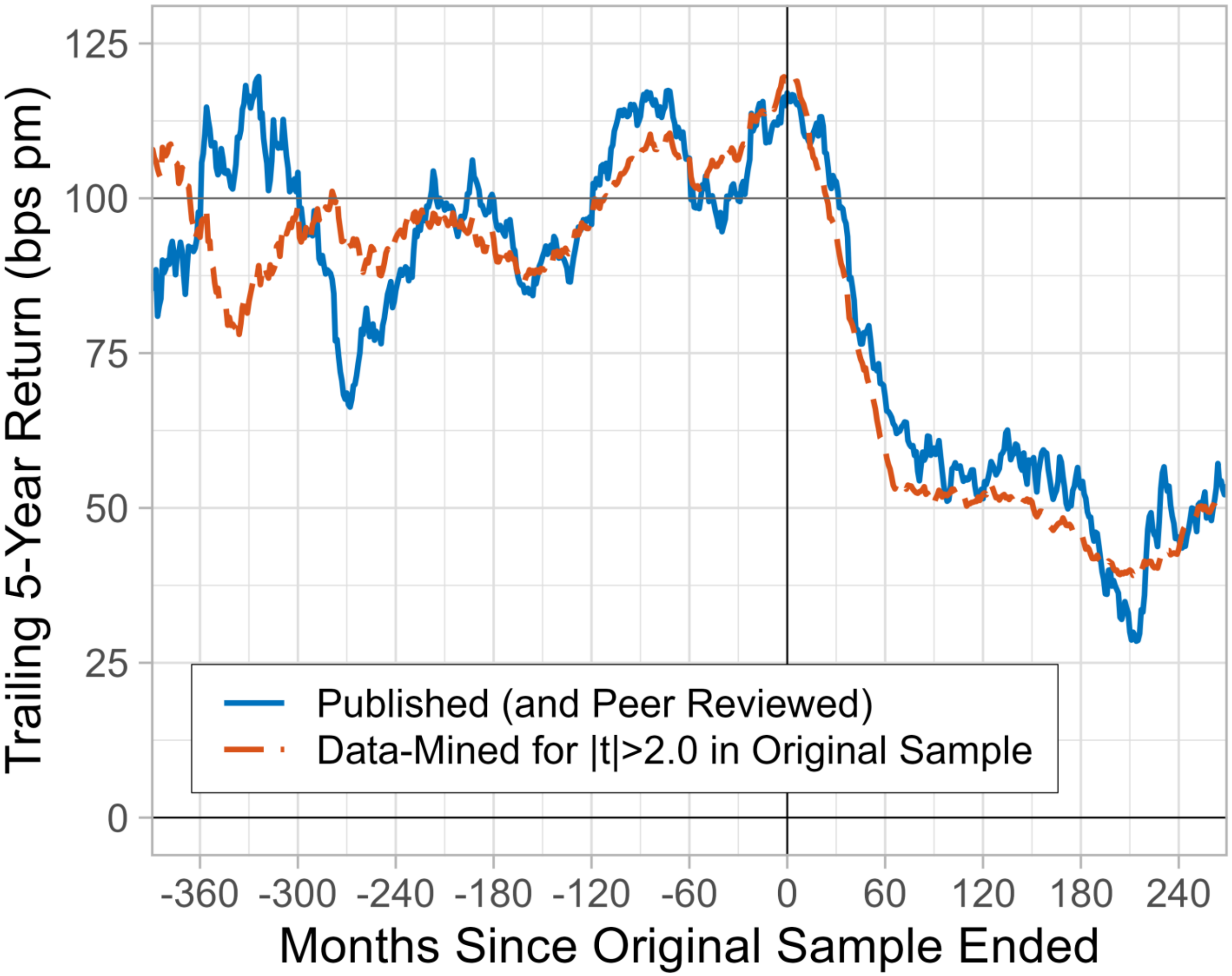
Our answer:

- **Publishable ideas outperform data mining by perhaps 2 bps per month**



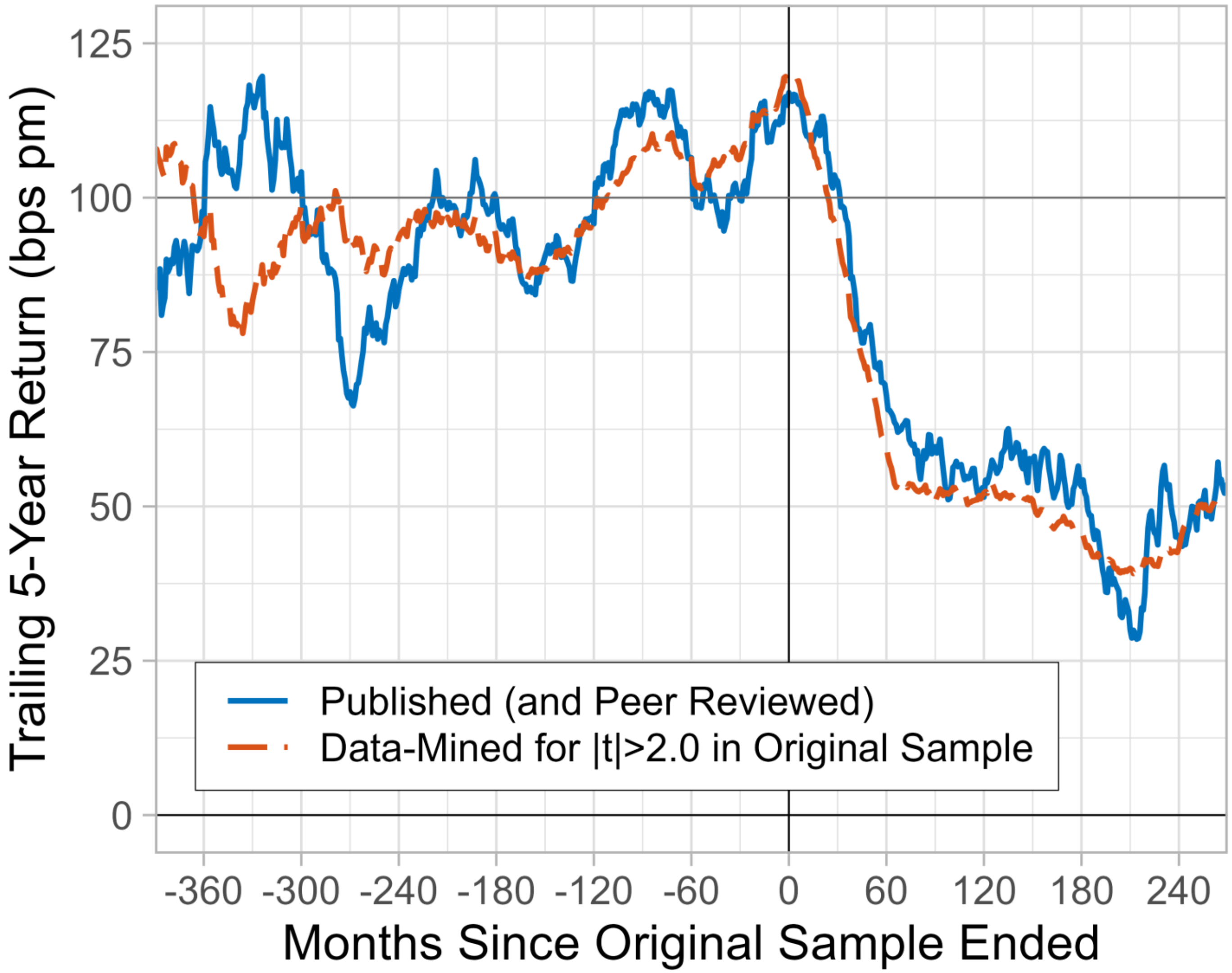
Our answer:

- **Publishable ideas outperform data mining by perhaps 2 bps per month**
- Focusing on publishable risk-based ideas does *not* help



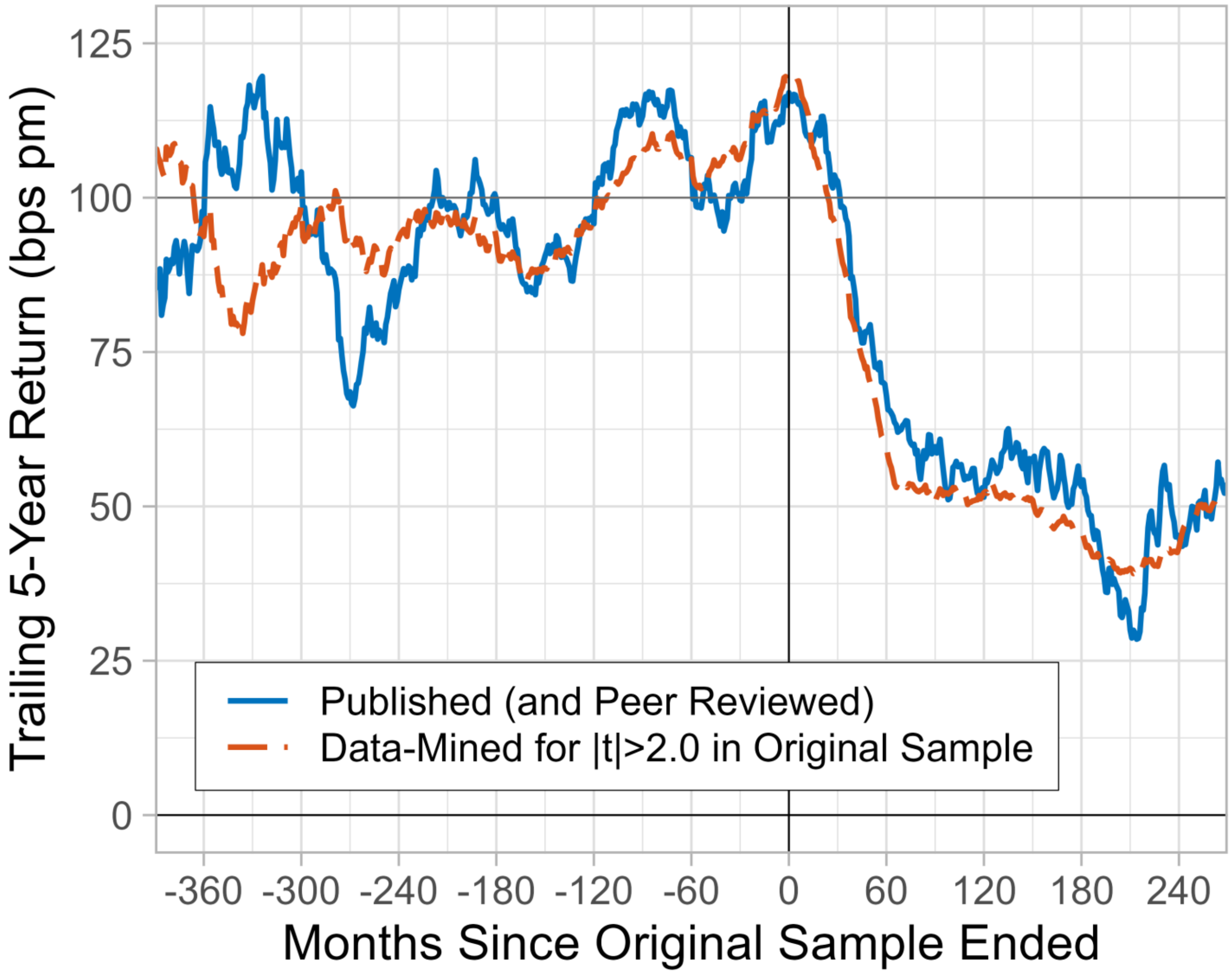
Our answer:

- **Publishable ideas outperform data mining by perhaps 2 bps per month**
- Focusing on publishable risk-based ideas does *not* help
- **On the bright side, data mining uncovers true predictability**



Our answer:

- **Publishable ideas outperform data mining by perhaps 2 bps per month**
- Focusing on publishable risk-based ideas does *not* help
- **On the bright side, data mining uncovers true predictability**
  - Reminiscent of data mining successes in language modeling (e.g. ChatGPT)



Data Mined Returns

Data-mined long-short strategies

- Two kinds of accounting ratios
  - Simple ratios: X/Y
  - Scaled first difference: ΔX/lag(Y)

Data-mined long-short strategies

- Two kinds of accounting ratios
  - Simple ratios: X/Y
  - Scaled first difference: ΔX/lag(Y)
- Where
  - X = one of 242 annual accounting vars (including market equity)
  - Y = one of the X's that is positive for > 25% of firms in 1963

Data-mined long-short strategies

- Two kinds of accounting ratios
  - Simple ratios: X/Y
  - Scaled first difference: ΔX/lag(Y)
- Where
  - X = one of 242 annual accounting vars (including market equity)
  - Y = one of the X's that is positive for > 25% of firms in 1963
- Yields 29,315 accounting ratios
- Using each ratio, form long-short decile strategies

Data-mined long-short strategies

- Two kinds of accounting ratios
  - Simple ratios: X/Y
  - Scaled first difference: ΔX/lag(Y)
- Where
  - X = one of 242 annual accounting vars (including market equity)
  - Y = one of the X's that is positive for > 25% of firms in 1963
- Yields 29,315 accounting ratios
- Using each ratio, form long-short decile strategies
- **Arguably no economics, no look-ahead bias**

Data mining generates large "out-of-sample" returns

In-Sample Bin	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	-59.3	-4.24	-49.4	16.7	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

Data mining generates large "out-of-sample" returns

In-Sample Bin	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	-59.3	-4.24	-49.4	16.7	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

- Each year, sort 29,000 strategies into bins based on past 30 year returns (IS), hold bin for one year (OOS)

Data mining generates large "out-of-sample" returns

In-Sample Bin	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	-59.3	-4.24	-49.4	16.7	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

- Each year, sort 29,000 strategies into bins based on past 30 year returns (IS), hold bin for one year (OOS)

Data mining generates large "out-of-sample" returns

In-Sample Bin	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	$\boxed{-59.3}$	-4.24	-49.4	16.7	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

- Each year, sort 29,000 strategies into bins based on past 30 year returns (IS), hold bin for one year (OOS)

Data mining generates large "out-of-sample" returns

In-Sample Bin	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	-59.3	-4.24	-49.4	16.7	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

- Each year, sort 29,000 strategies into bins based on past 30 year returns (IS), hold bin for one year (OOS)

Data mining generates large “out-of-sample” returns

In-Sample Bin	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	$\boxed{-59.3}$	-4.24	$\boxed{-49.4}$	$\boxed{16.7}$	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

- Each year, sort 29,000 strategies into bins based on past 30 year returns (IS), hold bin for one year (OOS)
- **Replicates + extends Yan-Zheng 2017 (underappreciated paper)**

Data mining generates large "out-of-sample" returns

In-Sample Bin	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	$\boxed{-59.3}$	-4.24	$\boxed{-49.4}$	$\boxed{16.7}$	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

- Each year, sort 29,000 strategies into bins based on past 30 year returns (IS), hold bin for one year (OOS)
- **Replicates + extends Yan-Zheng 2017 (underappreciated paper)**
- Contrasts with Harvey-Liu 2020, who find FDR ≈ 100%

Data mining generates large "out-of-sample" returns

In-Sample Bin	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	$\boxed{-59.3}$	-4.24	$\boxed{-49.4}$	$\boxed{16.7}$	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

- Each year, sort 29,000 strategies into bins based on past 30 year returns (IS), hold bin for one year (OOS)
- **Replicates + extends Yan-Zheng 2017 (underappreciated paper)**
- Contrasts with Harvey-Liu 2020, who find FDR ≈ 100%
  - *Consistent w/ Chen 2024: Harvey-Liu 2020 misinterprets FDR methods*

Data-mined strategies with |t|>2 are diverse

Covariance structure of long-short returns

Panel (a): Pairwise correlations

Quantiles	Q1	Q5	Q10	Q25	Q50	Q75	Q90	Q95	Q99
Equal-Weighted	-0.42	-0.23	-0.15	-0.04	0.05	0.16	0.29	0.38	0.56
Value-Weighted	-0.35	-0.20	-0.13	-0.05	0.04	0.14	0.25	0.32	0.51

Panel (b): PCA Explained Variance (%)

Number of PCs	1	5	10	20	30	40	50	60	70	80	90	100
Equal-Weighted	24	47	55	63	68	72	75	78	80	82	84	85
Value-Weighted	24	44	52	62	68	72	76	79	81	83	85	87

Data-mined strategies with |t|>2 are diverse

Covariance structure of long-short returns

Panel (a): Pairwise correlations

Quantiles	Q1	Q5	Q10	Q25	Q50	Q75	Q90	Q95	Q99
Equal-Weighted	-0.42	-0.23	-0.15	-0.04	0.05	0.16	0.29	0.38	0.56
Value-Weighted	-0.35	-0.20	-0.13	-0.05	0.04	0.14	0.25	0.32	0.51

Panel (b): PCA Explained Variance (%)

Number of PCs	1	5	10	20	30	40	50	60	70	80	90	100
Equal-Weighted	24	47	55	63	68	72	75	78	80	82	84	85
Value-Weighted	24	44	52	62	68	72	76	79	81	83	85	87

- More than 85% of correlations below 0.30 in absolute value

Data-mined strategies with |t|>2 are diverse

Covariance structure of long-short returns

Panel (a): Pairwise correlations									
Quantiles	Q1	Q5	Q10	Q25	Q50	Q75	Q90	Q95	Q99
Equal-Weighted	-0.42	-0.23	-0.15	-0.04	0.05	0.16	0.29	0.38	0.56
Value-Weighted	-0.35	-0.20	-0.13	-0.05	0.04	0.14	0.25	0.32	0.51

Panel (b): PCA Explained Variance (%)												
Number of PCs	1	5	10	20	30	40	50	60	70	80	90	100
Equal-Weighted	24	47	55	63	68	72	75	78	80	82	84	85
Value-Weighted	24	44	52	62	68	72	76	79	81	83	85	87

- More than 85% of correlations below 0.30 in absolute value
- 70 PCs are required to capture 80% of variance

Data-mined strategies with |t|>2 are diverse

Covariance structure of long-short returns

Panel (a): Pairwise correlations									
Quantiles	Q1	Q5	Q10	Q25	Q50	Q75	Q90	Q95	Q99
Equal-Weighted	-0.42	-0.23	-0.15	-0.04	0.05	0.16	0.29	0.38	0.56
Value-Weighted	-0.35	-0.20	-0.13	-0.05	0.04	0.14	0.25	0.32	0.51

Panel (b): PCA Explained Variance (%)												
Number of PCs	1	5	10	20	30	40	50	60	70	80	90	100
Equal-Weighted	24	47	55	63	68	72	75	78	80	82	84	85
Value-Weighted	24	44	52	62	68	72	76	79	81	83	85	87

- More than 85% of correlations below 0.30 in absolute value
- 70 PCs are required to capture 80% of variance
- **Data mining doesn't just pick up size, B/M, profitability**

Themes from mining the 1963-1980 sample

20 numerators and stock weights that produce largest t-stats

Themes from mining the 1963-1980 sample

20 numerators and stock weights that produce largest t-stats

Numerator (Stock Weight)	Pct Short	t-stat
ΔAssets (ew)	100	4.0
ΔIntangible assets (ew)	100	4.0
ΔPPE net (ew)	98	4.0
ΔPPE gross (ew)	98	3.8
ΔInvested capital (ew)	100	3.5
ΔCapital expenditure (ew)	100	3.2
ΔCommon stock (ew)	100	5.1
ΔLiabilities (ew)	100	4.7
ΔCapital surplus (ew)	100	4.1
ΔLong-term debt (ew)	100	3.6
ΔCapital surplus (vw)	98	3.0

Numerator (Stock Weight)	Pct Short	t-stat
ΔInventories (ew)	100	4.2
ΔNotes payable st (ew)	100	3.8
ΔReceivables (ew)	100	3.7
ΔDebt in current liab (ew)	100	3.7
ΔCurrent liabilities (ew)	100	3.7
ΔCost of goods sold (ew)	100	3.7
ΔOperating expenses (ew)	98	3.5
ΔSG&A (ew)	100	3.3
ΔInterest expense (ew)	98	3.3

Themes from mining the 1963-1980 sample

20 numerators and stock weights that produce largest t-stats

Numerator (Stock Weight)	Pct Short	t-stat
ΔAssets (ew)	100	4.0
ΔIntangible assets (ew)	100	4.0
ΔPPE net (ew)	98	4.0
ΔPPE gross (ew)	98	3.8
ΔInvested capital (ew)	100	3.5
ΔCapital expenditure (ew)	100	3.2
ΔCommon stock (ew)	100	5.1
ΔLiabilities (ew)	100	4.7
ΔCapital surplus (ew)	100	4.1
ΔLong-term debt (ew)	100	3.6
ΔCapital surplus (vw)	98	3.0

Numerator (Stock Weight)	Pct Short	t-stat
ΔInventories (ew)	100	4.2
ΔNotes payable st (ew)	100	3.8
ΔReceivables (ew)	100	3.7
ΔDebt in current liab (ew)	100	3.7
ΔCurrent liabilities (ew)	100	3.7
ΔCost of goods sold (ew)	100	3.7
ΔOperating expenses (ew)	98	3.5
ΔSG&A (ew)	100	3.3
ΔInterest expense (ew)	98	3.3

- **All top 20 numerators fit into themes from academic publications**

Themes from mining the 1963-1980 sample

20 numerators and stock weights that produce largest t-stats

Numerator (Stock Weight)	Pct Short	t-stat
Investment (Titman, Wei, Xie 2004)		
ΔAssets (ew)	100	4.0
ΔIntangible assets (ew)	100	4.0
ΔPPE net (ew)	98	4.0
ΔPPE gross (ew)	98	3.8
ΔInvested capital (ew)	100	3.5
ΔCapital expenditure (ew)	100	3.2
ΔCommon stock (ew)	100	5.1
ΔLiabilities (ew)	100	4.7
ΔCapital surplus (ew)	100	4.1
ΔLong-term debt (ew)	100	3.6
ΔCapital surplus (vw)	98	3.0

Numerator (Stock Weight)	Pct Short	t-stat
ΔInventories (ew)	100	4.2
ΔNotes payable st (ew)	100	3.8
ΔReceivables (ew)	100	3.7
ΔDebt in current liab (ew)	100	3.7
ΔCurrent liabilities (ew)	100	3.7
ΔCost of goods sold (ew)	100	3.7
ΔOperating expenses (ew)	98	3.5
ΔSG&A (ew)	100	3.3
ΔInterest expense (ew)	98	3.3

- **All top 20 numerators fit into themes from academic publications**

Themes from mining the 1963-1980 sample

20 numerators and stock weights that produce largest t-stats

Numerator (Stock Weight)	Pct Short	t-stat
Investment (Titman, Wei, Xie 2004)		
ΔAssets (ew)	100	4.0
ΔIntangible assets (ew)	100	4.0
ΔPPE net (ew)	98	4.0
ΔPPE gross (ew)	98	3.8
ΔInvested capital (ew)	100	3.5
ΔCapital expenditure (ew)	100	3.2
Ext Financing (Spiess/Affleck-Graves 1999)		
ΔCommon stock (ew)	100	5.1
ΔLiabilities (ew)	100	4.7
ΔCapital surplus (ew)	100	4.1
ΔLong-term debt (ew)	100	3.6
ΔCapital surplus (vw)	98	3.0

Numerator (Stock Weight)	Pct Short	t-stat
ΔInventories (ew)	100	4.2
ΔNotes payable st (ew)	100	3.8
ΔReceivables (ew)	100	3.7
ΔDebt in current liab (ew)	100	3.7
ΔCurrent liabilities (ew)	100	3.7
ΔCost of goods sold (ew)	100	3.7
ΔOperating expenses (ew)	98	3.5
ΔSG&A (ew)	100	3.3
ΔInterest expense (ew)	98	3.3

- **All top 20 numerators fit into themes from academic publications**

Themes from mining the 1963-1980 sample

20 numerators and stock weights that produce largest t-stats

Numerator (Stock Weight)	Pct Short	t-stat
Investment (Titman, Wei, Xie 2004)		
ΔAssets (ew)	100	4.0
ΔIntangible assets (ew)	100	4.0
ΔPPE net (ew)	98	4.0
ΔPPE gross (ew)	98	3.8
ΔInvested capital (ew)	100	3.5
ΔCapital expenditure (ew)	100	3.2
Ext Financing (Spiess/Affleck-Graves 1999)		
ΔCommon stock (ew)	100	5.1
ΔLiabilities (ew)	100	4.7
ΔCapital surplus (ew)	100	4.1
ΔLong-term debt (ew)	100	3.6
ΔCapital surplus (vw)	98	3.0

Numerator (Stock Weight)	Pct Short	t-stat
Accruals (Sloan 1996; Thomas-Zhang 2002)		
ΔInventories (ew)	100	4.2
ΔNotes payable st (ew)	100	3.8
ΔReceivables (ew)	100	3.7
ΔDebt in current liab (ew)	100	3.7
ΔCurrent liabilities (ew)	100	3.7
ΔCost of goods sold (ew)	100	3.7
ΔOperating expenses (ew)	98	3.5
ΔSG&A (ew)	100	3.3
ΔInterest expense (ew)	98	3.3

- **All top 20 numerators fit into themes from academic publications**

Themes from mining the 1963-1980 sample

20 numerators and stock weights that produce largest t-stats

Numerator (Stock Weight)	Pct Short	t-stat
Investment (Titman, Wei, Xie 2004)		
ΔAssets (ew)	100	4.0
ΔIntangible assets (ew)	100	4.0
ΔPPE net (ew)	98	4.0
ΔPPE gross (ew)	98	3.8
ΔInvested capital (ew)	100	3.5
ΔCapital expenditure (ew)	100	3.2
Ext Financing (Spiess/Affleck-Graves 1999)		
ΔCommon stock (ew)	100	5.1
ΔLiabilities (ew)	100	4.7
ΔCapital surplus (ew)	100	4.1
ΔLong-term debt (ew)	100	3.6
ΔCapital surplus (vw)	98	3.0

Numerator (Stock Weight)	Pct Short	t-stat
Accruals (Sloan 1996; Thomas-Zhang 2002)		
ΔInventories (ew)	100	4.2
ΔNotes payable st (ew)	100	3.8
ΔReceivables (ew)	100	3.7
ΔDebt in current liab (ew)	100	3.7
ΔCurrent liabilities (ew)	100	3.7
Earnings Surprise (Foster et. al 1984)		
ΔCost of goods sold (ew)	100	3.7
ΔOperating expenses (ew)	98	3.5
ΔSG&A (ew)	100	3.3
ΔInterest expense (ew)	98	3.3

- **All top 20 numerators fit into themes from academic publications**

Themes from mining the 1963-1980 sample

20 numerators and stock weights that produce largest t-stats

Numerator (Stock Weight)	Pct Short	t-stat
Investment (Titman, Wei, Xie 2004)		
ΔAssets (ew)	100	4.0
ΔIntangible assets (ew)	100	4.0
ΔPPE net (ew)	98	4.0
ΔPPE gross (ew)	98	3.8
ΔInvested capital (ew)	100	3.5
ΔCapital expenditure (ew)	100	3.2
Ext Financing (Spiess/Affleck-Graves 1999)		
ΔCommon stock (ew)	100	5.1
ΔLiabilities (ew)	100	4.7
ΔCapital surplus (ew)	100	4.1
ΔLong-term debt (ew)	100	3.6
ΔCapital surplus (vw)	98	3.0

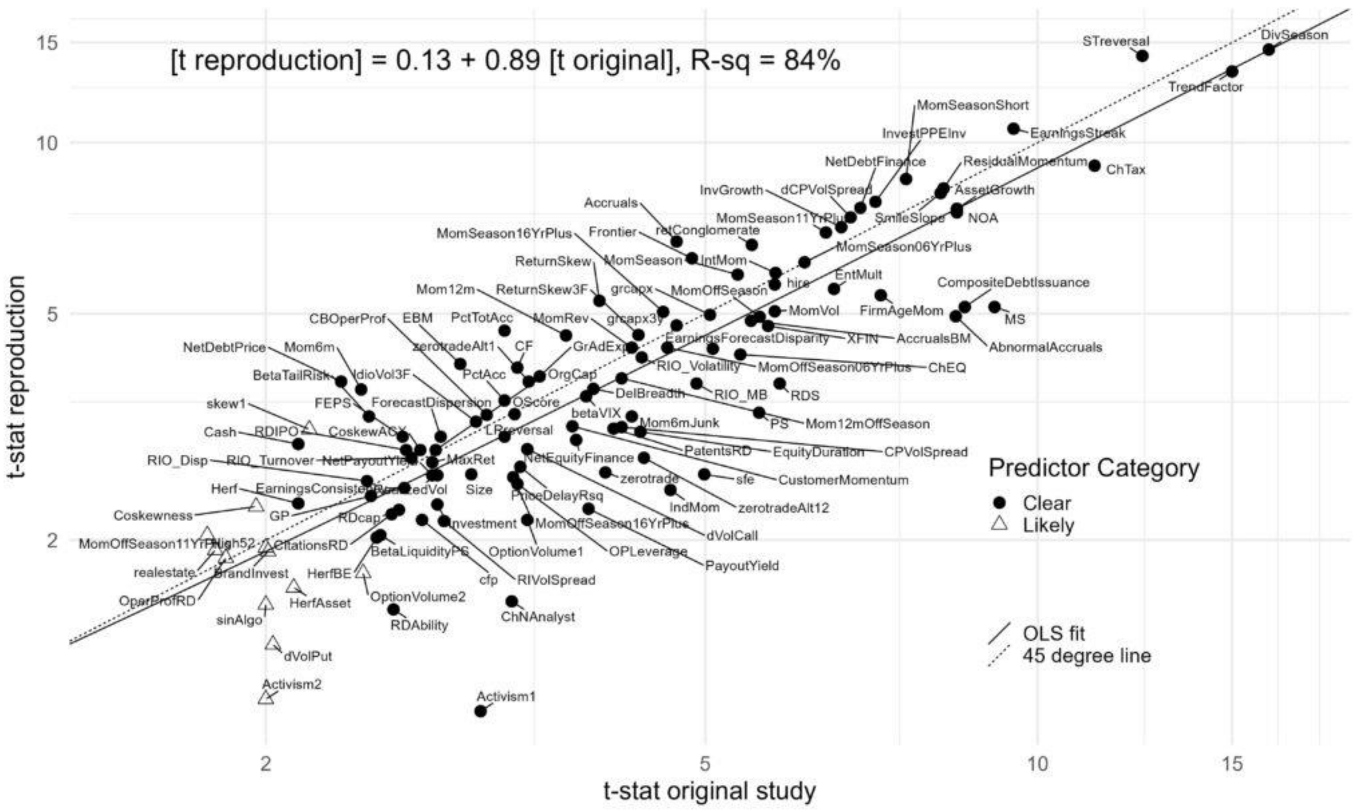
Numerator (Stock Weight)	Pct Short	t-stat
Accruals (Sloan 1996; Thomas-Zhang 2002)		
ΔInventories (ew)	100	4.2
ΔNotes payable st (ew)	100	3.8
ΔReceivables (ew)	100	3.7
ΔDebt in current liab (ew)	100	3.7
ΔCurrent liabilities (ew)	100	3.7
Earnings Surprise (Foster et. al 1984)		
ΔCost of goods sold (ew)	100	3.7
ΔOperating expenses (ew)	98	3.5
ΔSG&A (ew)	100	3.3
ΔInterest expense (ew)	98	3.3

- **All top 20 numerators fit into themes from academic publications**
- **But data mining can find the themes long before they are published**

Peer Review vs Data Mining

Peer-reviewed long-short strategies

- Chen-Zimmermann (2022) dataset
  - Dataset w/ most accurate reproductions of original tables
- Filter to have post-sample period $\geq$ 9 years
- Baseline data: 199 predictors



Data-mined return benchmarks

- For each published predictor,
  - Search the 29,000 accounting ratios for long-short $|t| > 2.0$

Data-mined return benchmarks

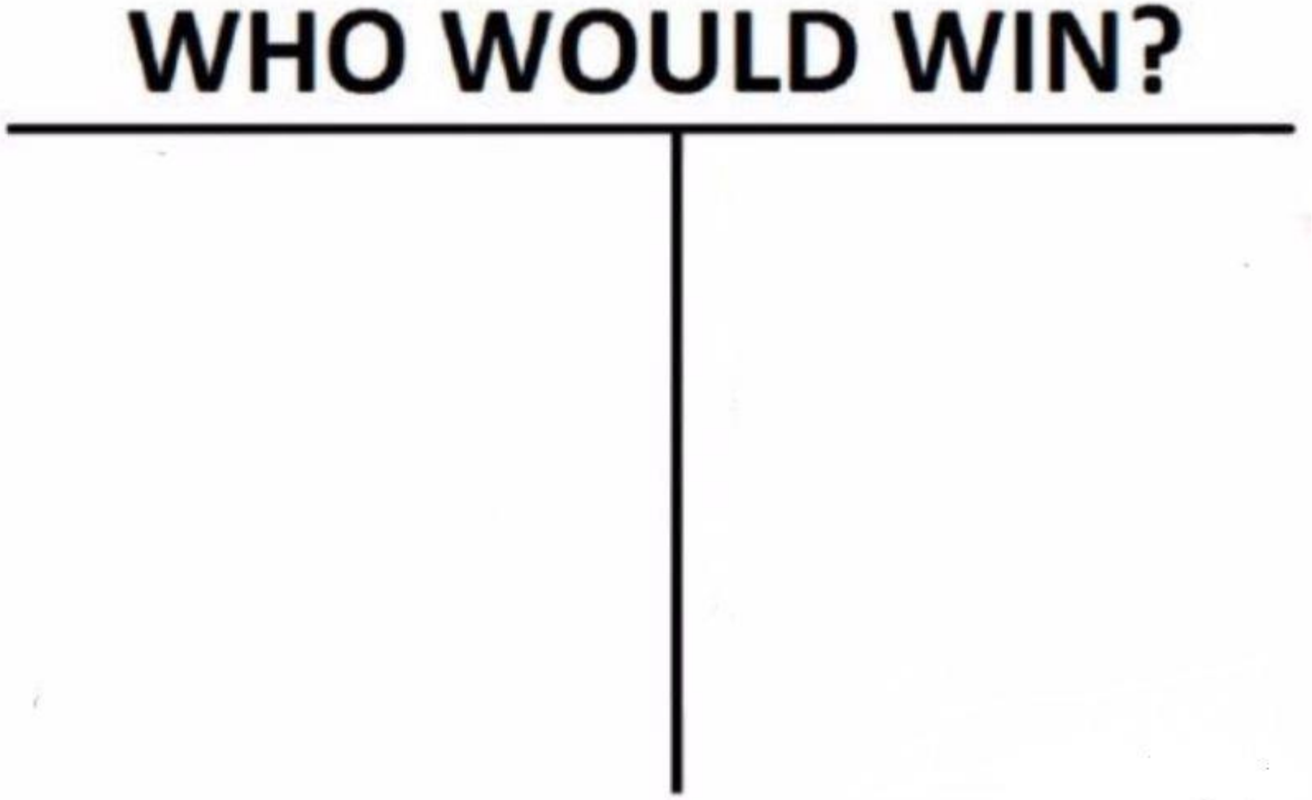
- For each published predictor,
  - Search the 29,000 accounting ratios for long-short $|t| > 2.0$
  - Using the original paper's
    - Sample period
    - Stock weighting (EW vs VW)

Data-mined return benchmarks

- For each published predictor,
  - Search the 29,000 accounting ratios for long-short $|t| > 2.0$
  - Using the original paper's
    - Sample period
    - Stock weighting (EW vs VW)
- Flip the long/short legs to have positive original-sample returns

Data-mined return benchmarks

- For each published predictor,
  - Search the 29,000 accounting ratios for long-short |t| > 2.0
  - Using the original paper's
    - Sample period
    - Stock weighting (EW vs VW)
- Flip the long/short legs to have positive original-sample returns

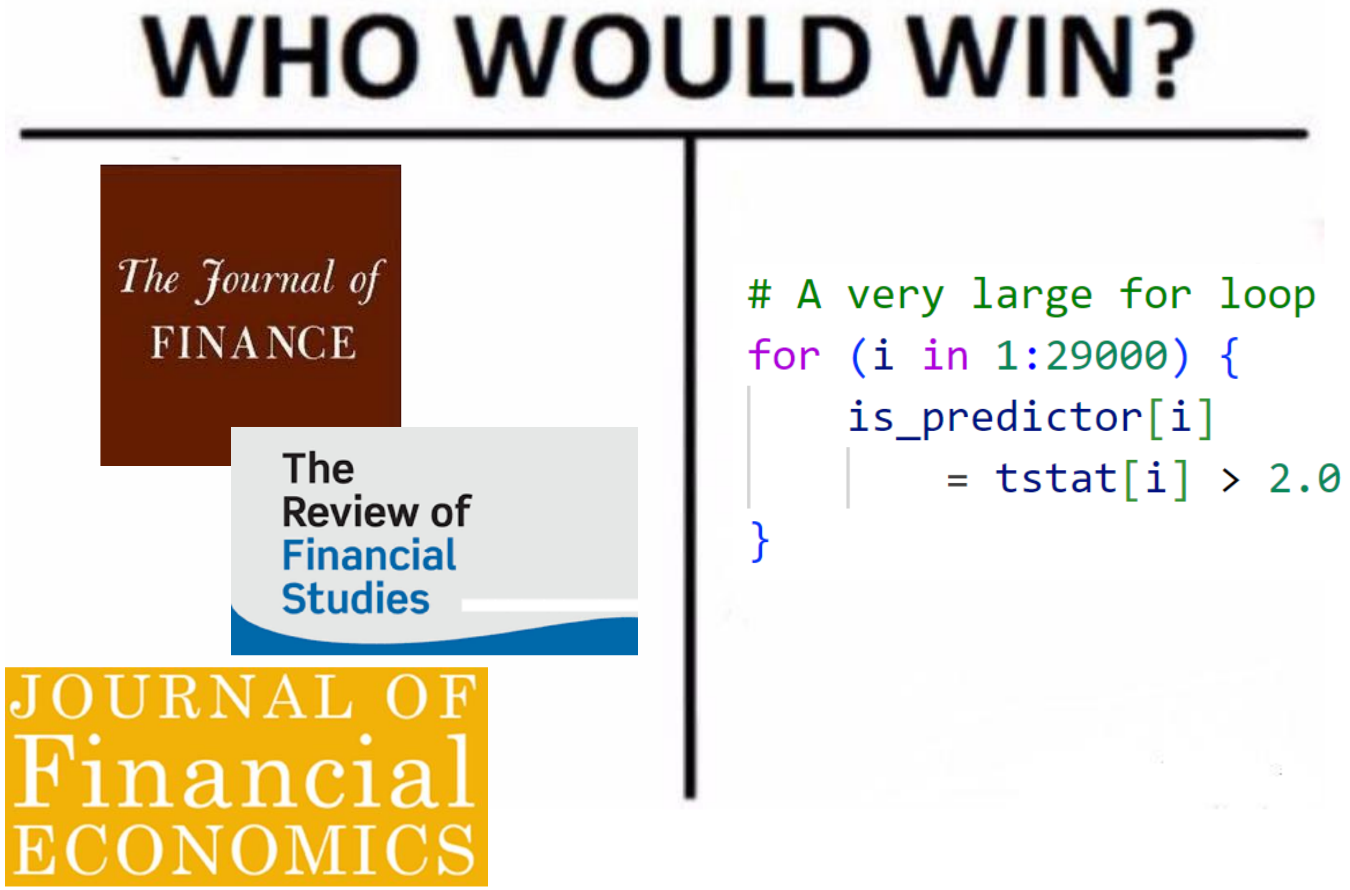


Data-mined return benchmarks

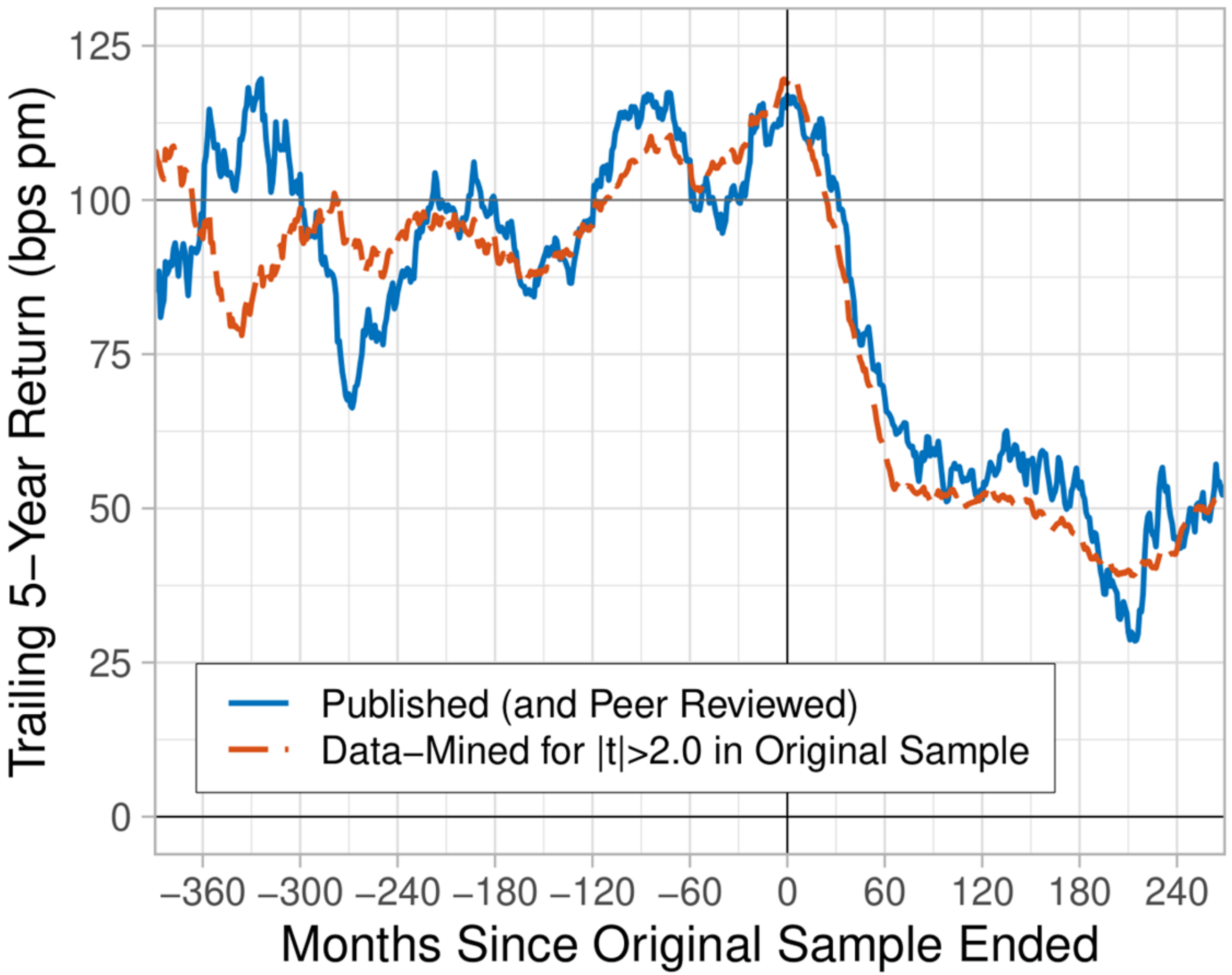
- For each published predictor,
  - Search the 29,000 accounting ratios for long-short |t| > 2.0
  - Using the original paper's
    - Sample period
    - Stock weighting (EW vs VW)
- Flip the long/short legs to have positive original-sample returns

Data-mined return benchmarks

- For each published predictor,
  - Search the 29,000 accounting ratios for long-short |t| > 2.0
  - Using the original paper's
    - Sample period
    - Stock weighting (EW vs VW)
- Flip the long/short legs to have positive original-sample returns

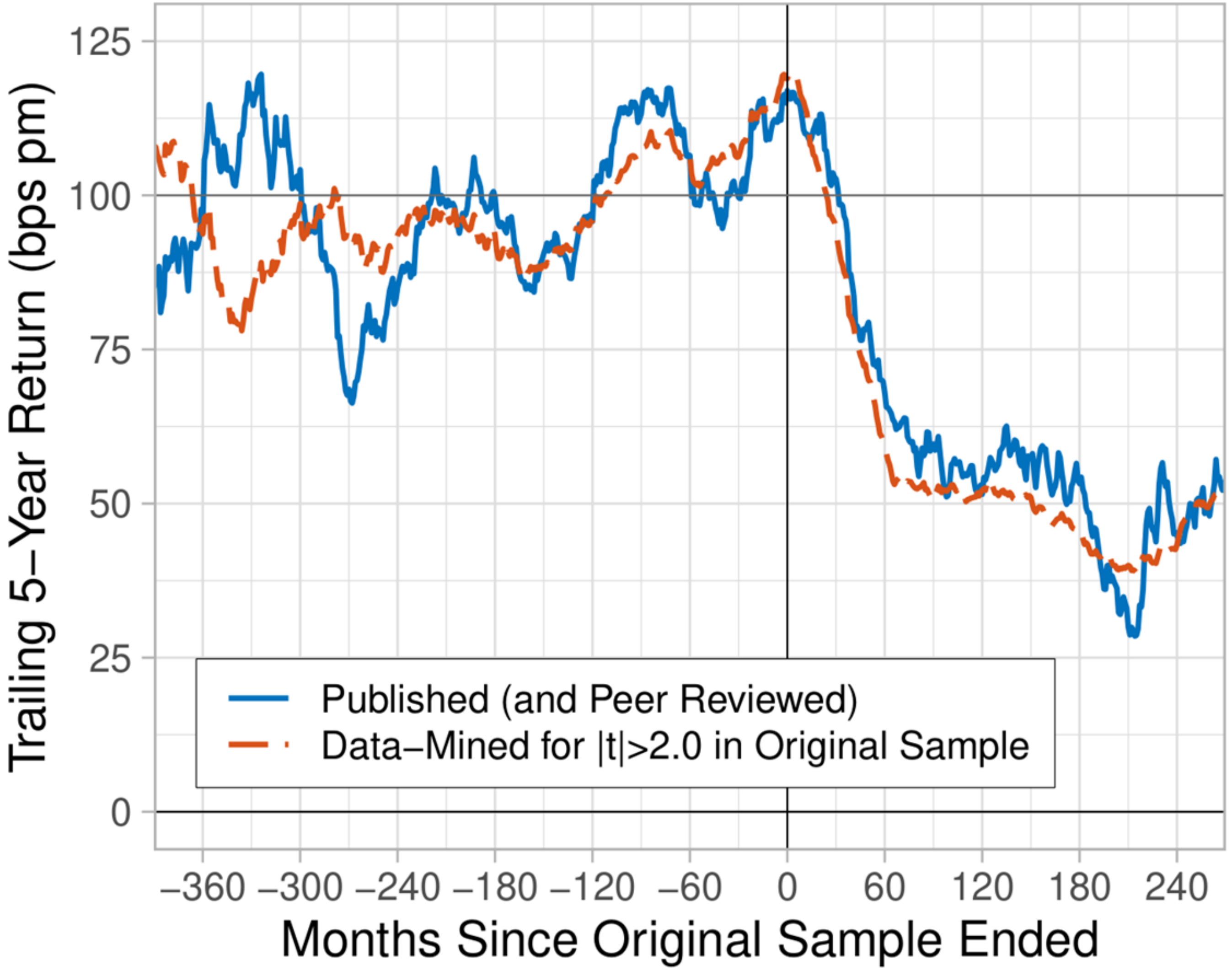


Does peer-reviewed research help predict the cross-section post-sample?



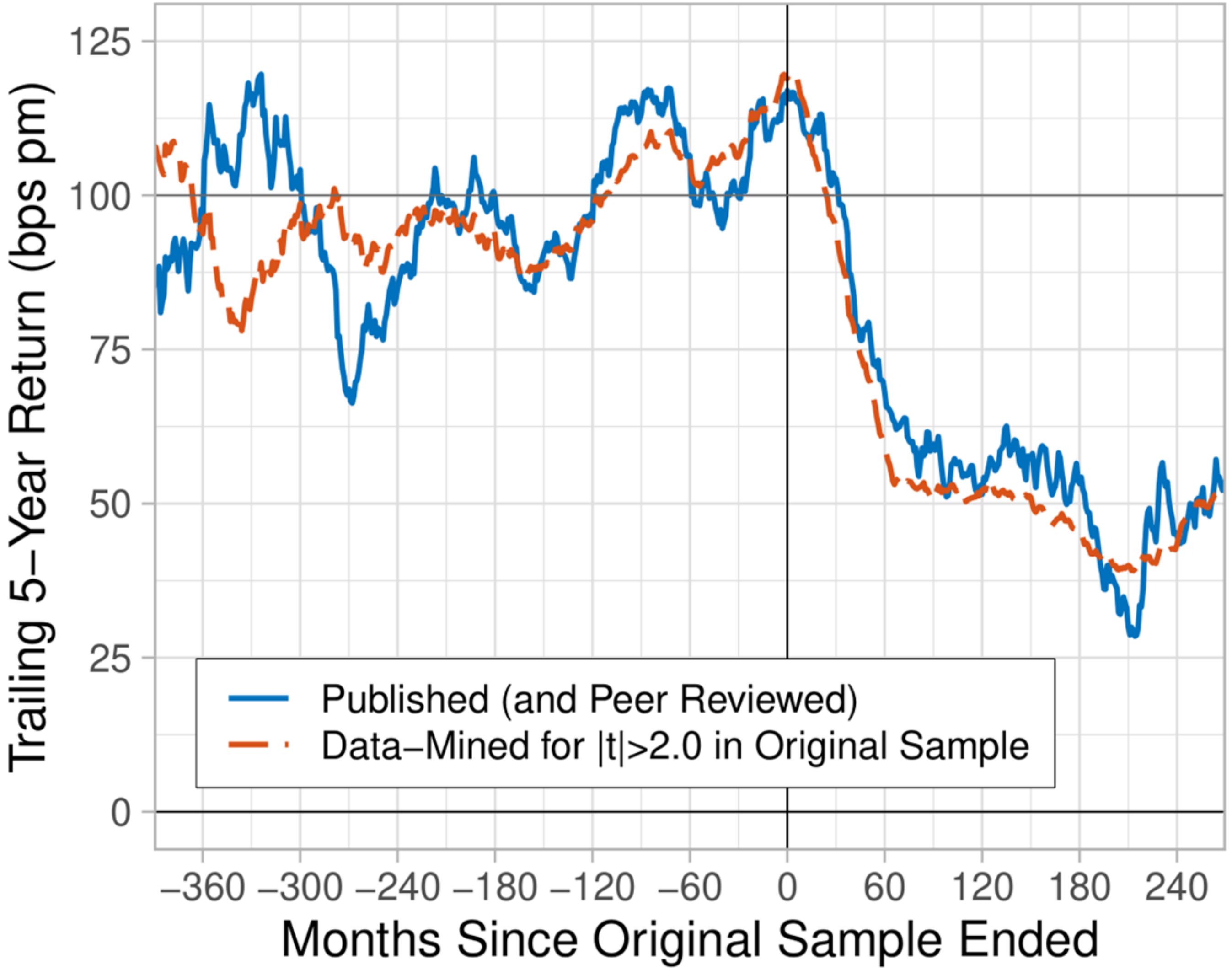
Does peer-reviewed research help predict the cross-section post-sample?

- Normalize so original sample return = 100 bps
  - For ease of interpretation



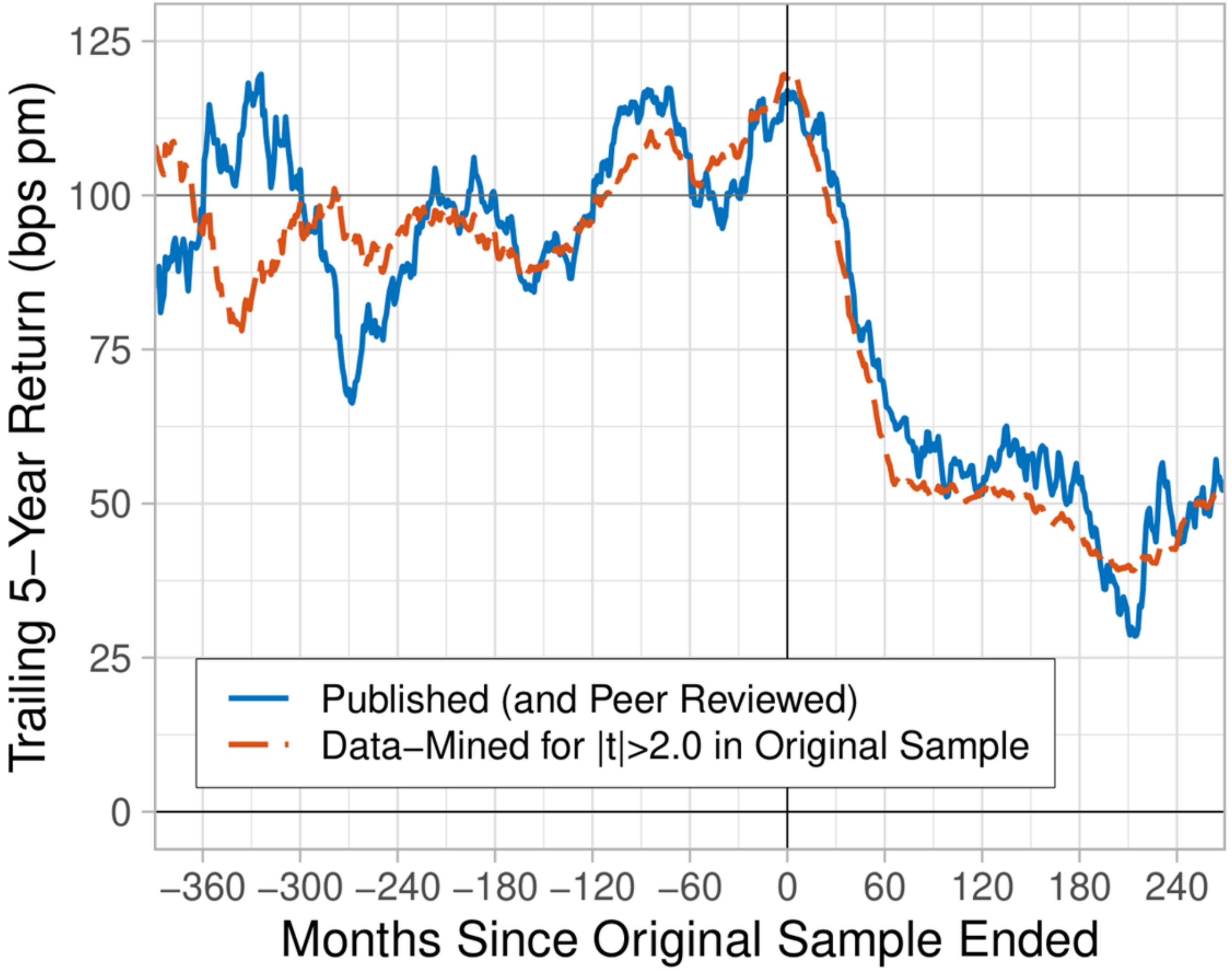
Does peer-reviewed research help predict the cross-section post-sample?

- Normalize so original sample return = 100 bps
  - For ease of interpretation
- 53% remains post-sample for published
  - (McLean-Pontiff 2016)



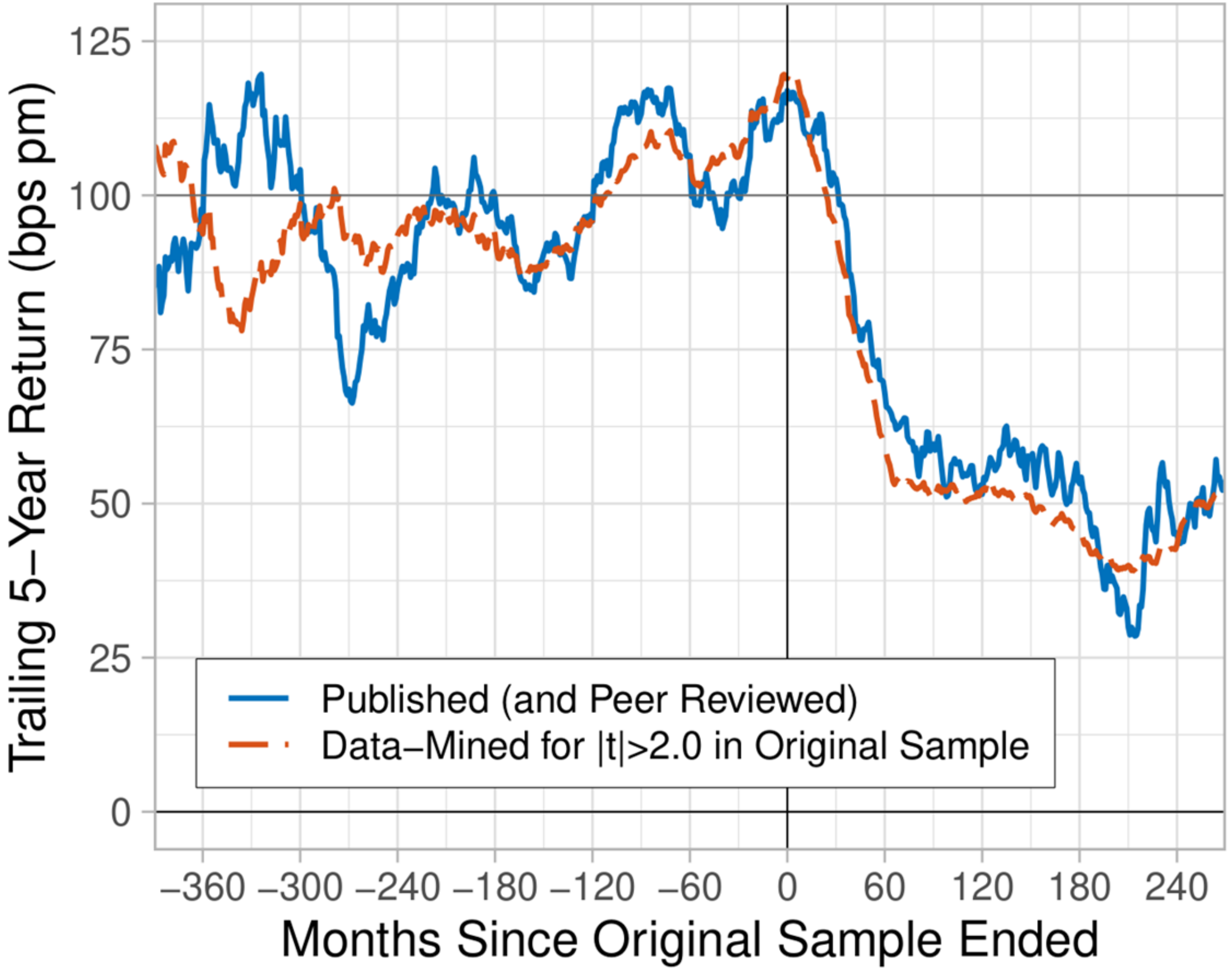
Does peer-reviewed research help predict the cross-section post-sample?

- Normalize so original sample return = 100 bps
  - For ease of interpretation
- 53% remains post-sample for published
  - (McLean-Pontiff 2016)
- **51% remains for data-mined benchmarks**
  - **(This paper)**



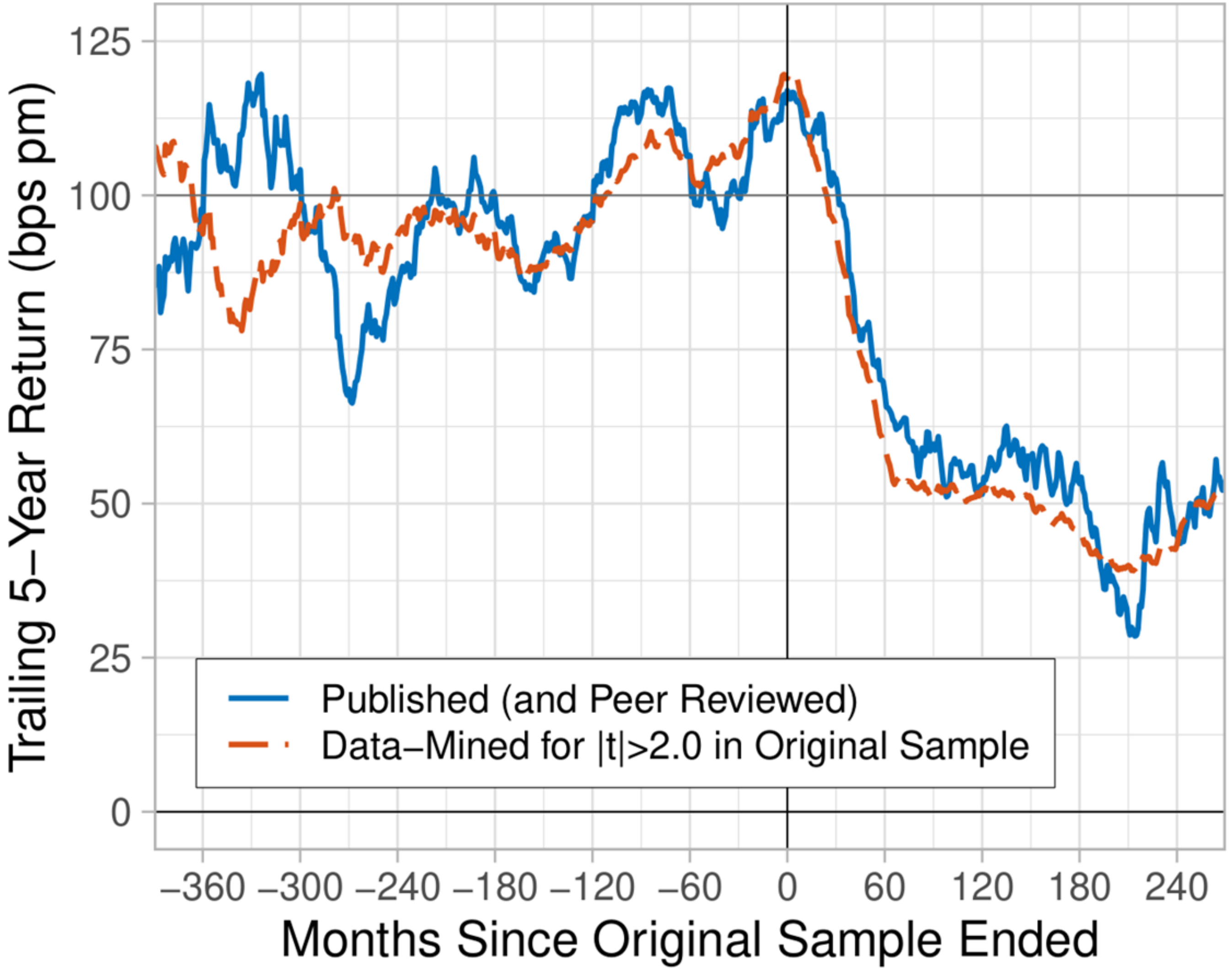
Does peer-reviewed research help predict the cross-section post-sample?

- **No, post-sample performance is similar to naïve back-testing**



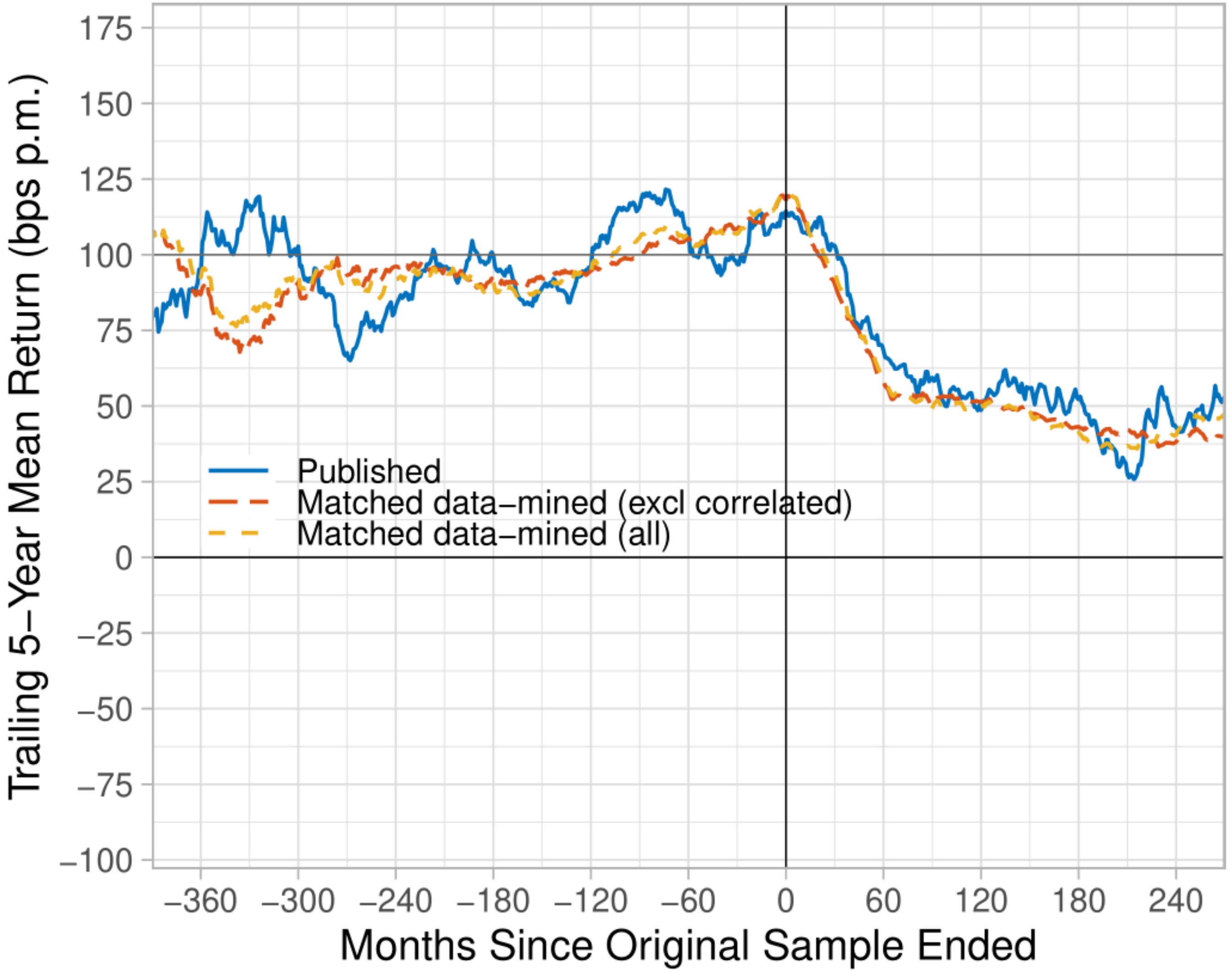
Does peer-reviewed research help predict the cross-section post-sample?

- **No, post-sample performance is similar to naïve back-testing**
  - Peer-reviewed motivations, supporting evidence, robustness tests, make little difference



Does peer-reviewed research help predict the cross-section post-sample?

- **No, post-sample performance is similar to naïve back-testing**
  - Peer-reviewed motivations, supporting evidence, robustness tests, make little difference
- Result robust to
  - Matching on in-sample returns and t-stats
  - Excluding correlated benchmarks



Do Risk-Based Explanations Help?

> *The best hope for finding pricing factors that are robust out of sample… …is to try to understand the fundamental macroeconomic sources of risk*
>
> –Cochrane 2005, Chapter 7

- Many papers take a different approach

> *The best hope for finding pricing factors that are robust out of sample… …is to try to understand the fundamental macroeconomic sources of risk*
>
> –Cochrane 2005, Chapter 7

- Many papers take a different approach
  - Banz 1981: “the size effect exists but it is not at all clear why it exists”

The best hope for finding pricing factors that are robust out of sample… …is to try to understand the fundamental macroeconomic sources of risk

–Cochrane 2005, Chapter 7

- Many papers take a different approach
  - Banz 1981: “the size effect exists but it is not at all clear why it exists”
  - De Bondt and Thaler 1985: “The empirical evidence… …is consistent with the overreaction hypothesis”

The best hope for finding pricing factors that are robust out of sample… …is to try to understand the fundamental macroeconomic sources of risk

–Cochrane 2005, Chapter 7

- Many papers take a different approach
  - Banz 1981: “the size effect exists but it is not at all clear why it exists”
  - De Bondt and Thaler 1985: “The empirical evidence… …is consistent with the overreaction hypothesis”
- **Do papers that follow Cochrane’s advice outperform data mining?**

The best hope for finding pricing factors that are robust out of sample… …is to try to understand the fundamental macroeconomic sources of risk

–Cochrane 2005, Chapter 7

- Many papers take a different approach
  - Banz 1981: “the size effect exists but it is not at all clear why it exists”
  - De Bondt and Thaler 1985: “The empirical evidence… …is consistent with the overreaction hypothesis”
- **Do papers that follow Cochrane’s advice outperform data mining?**
- **Method: Manually categorize explanations in original papers**
  1. Find summary passage
  2. Categorize passage as “risk,” “mispricing,” or “agnostic”
  3. Post passages and categories on GitHub, ask public for objections

Risk or Mispricing? According to Peer Review

Category	Num Predictors: Any Journal	Num Predictors: JF, JFE, RFS	Example Predictor	Example Passage
Risk	36	33	Real estate holdings (Tuzel 2010)	Firms with high real estate holdings are more vulnerable to bad productivity shocks and hence are riskier and have higher expected returns.
Mispricing	117	65	Share repurchases (Ikenberry, Lakonishok, Vermaelen 1995)	The market errs in its initial response and appears to ignore much of the information conveyed through repurchase announcements
Agnostic	46	25	Size (Banz 2981)	To summarize, the size effect exists but it is not at all clear why it exists
Total	199	123		

Risk or Mispricing? According to Peer Review

Category	Num Predictors: Any Journal	Num Predictors: JF, JFE, RFS	Example Predictor	Example Passage
Risk	36	33	Real estate holdings (Tuzel 2010)	Firms with high real estate holdings are more vulnerable to bad productivity shocks and hence are riskier and have higher expected returns.
Mispricing	117	65	Share repurchases (Ikenberry, Lakonishok, Vermaelen 1995)	The market errs in its initial response and appears to ignore much of the information conveyed through repurchase announcements
Agnostic	46	25	Size (Banz 2981)	To summarize, the size effect exists but it is not at all clear why it exists
Total	199	123		

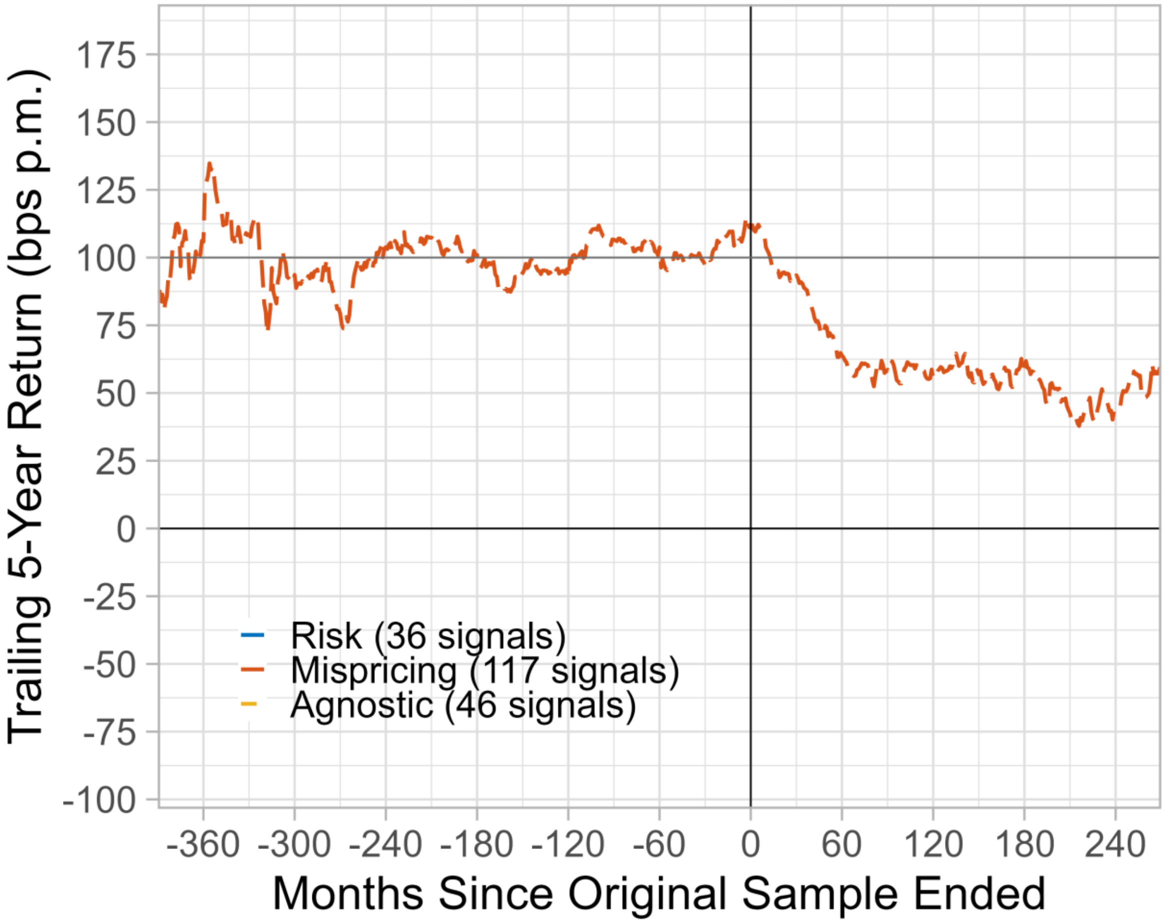
- **Only small minority 36/199= 18% are attributed to risk**

Risk or Mispricing? According to Peer Review

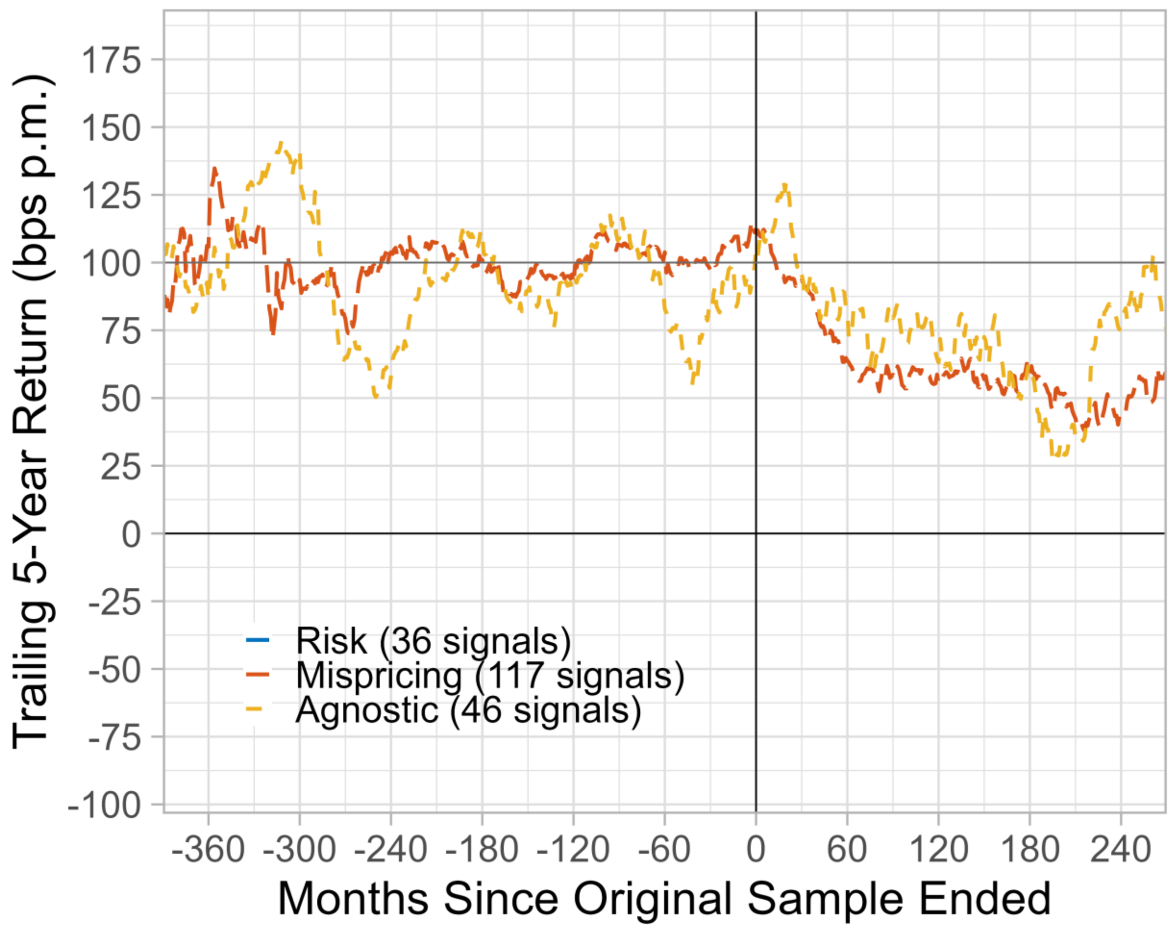
Category	Num Predictors: Any Journal	Num Predictors: JF, JFE, RFS	Example Predictor	Example Passage
Risk	36	33	Real estate holdings (Tuzel 2010)	Firms with high real estate holdings are more vulnerable to bad productivity shocks and hence are riskier and have higher expected returns.
Mispricing	117	65	Share repurchases (Ikenberry, Lakonishok, Vermaelen 1995)	The market errs in its initial response and appears to ignore much of the information conveyed through repurchase announcements
Agnostic	46	25	Size (Banz 2981)	To summarize, the size effect exists but it is not at all clear why it exists
Total	199	123		

- **Only small minority 36/199= 18% are attributed to risk**
  - **Top 3 Finance journals: 27% are risk**

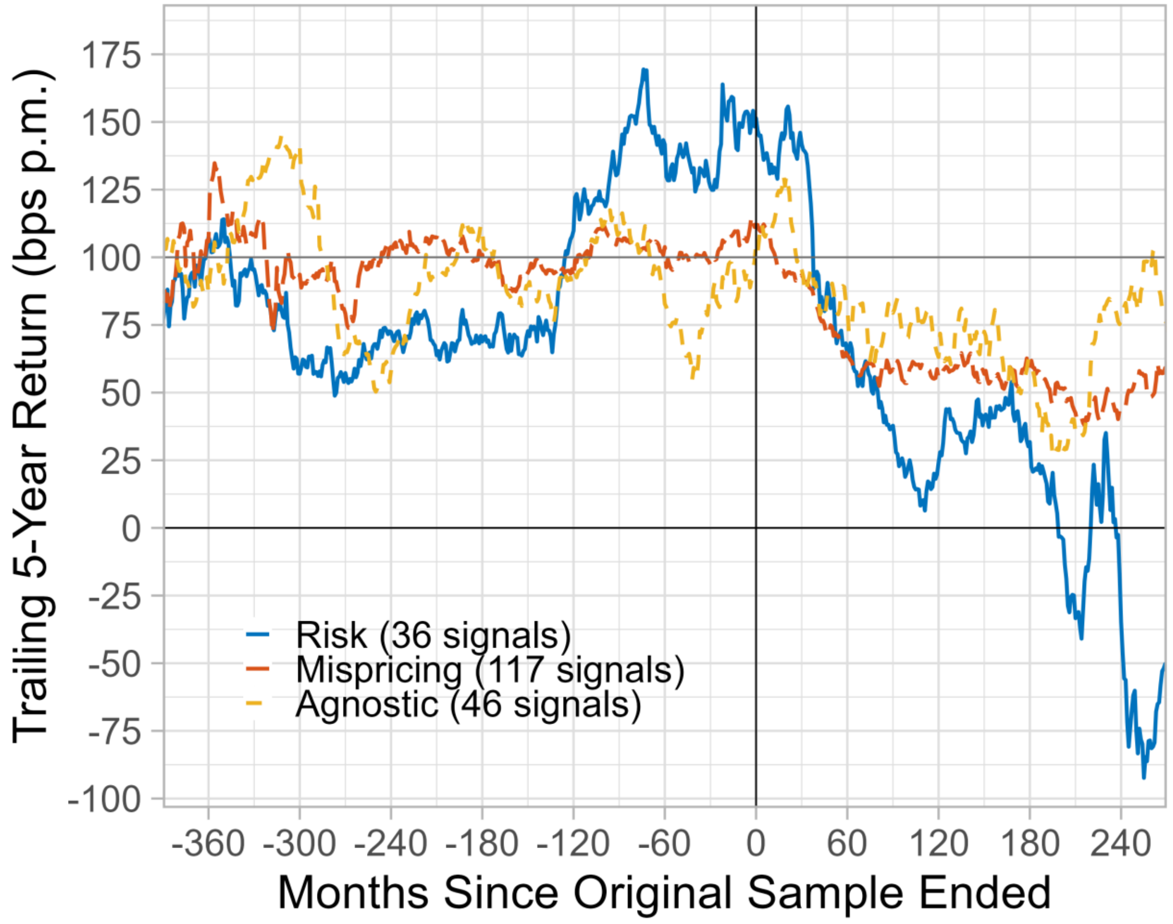
Post-sample decay: risk vs mispricing



Post-sample decay: risk vs mispricing

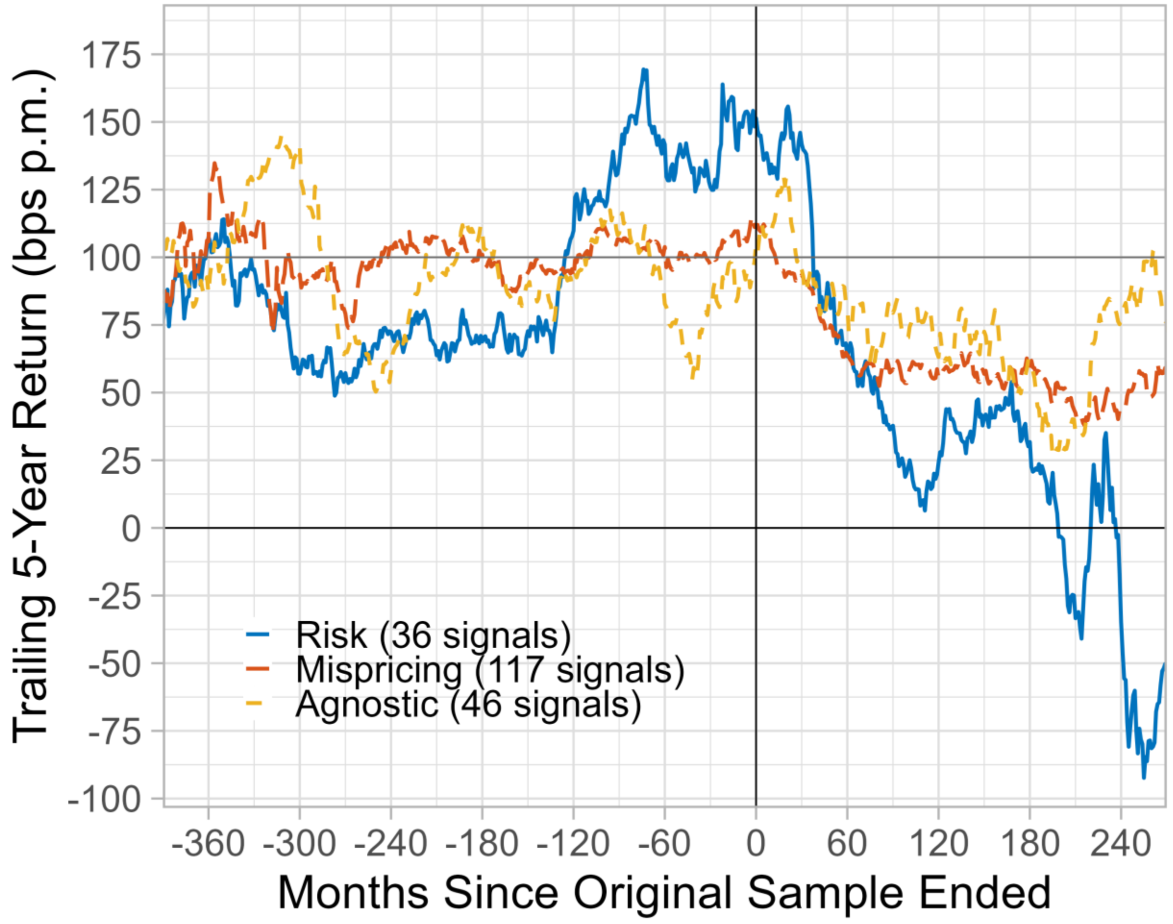


Post-sample decay: risk vs mispricing

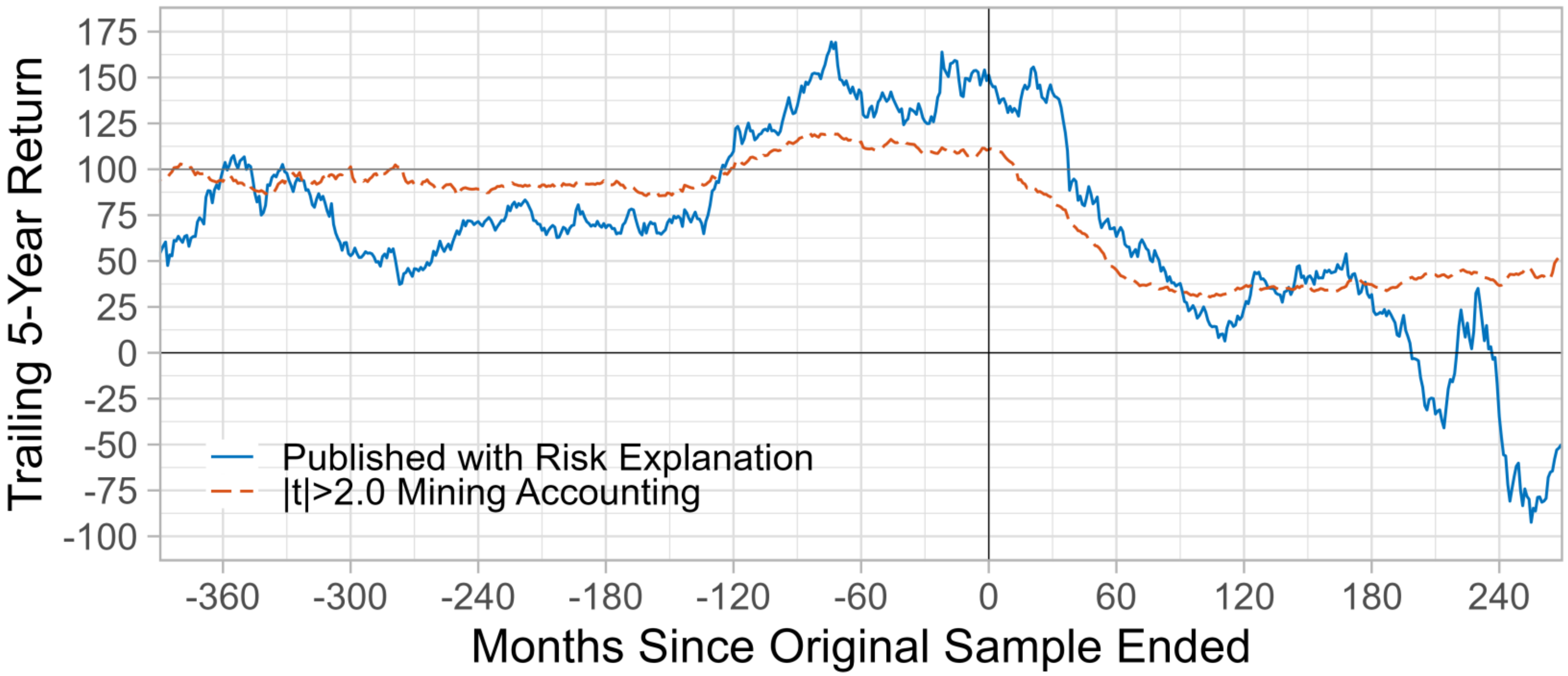


Post-sample decay: risk vs mispricing

- **No, publishable risk-based explanations do not help**
  - If anything, they lead to underperformance



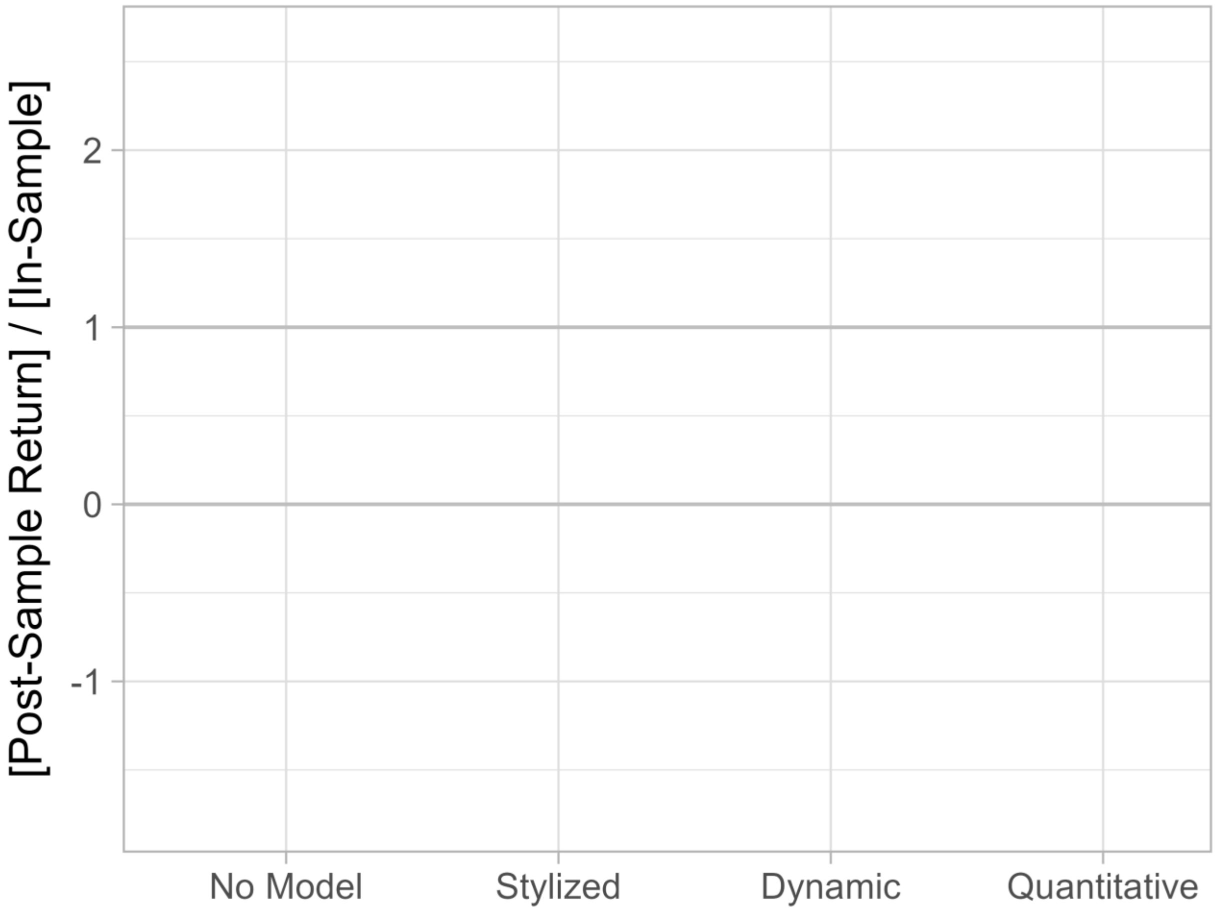
Risk vs data mining



- **Risk-based predictors fail to outperform data-mined benchmarks**
  - Data-mined benchmarks are exposed to the same market conditions

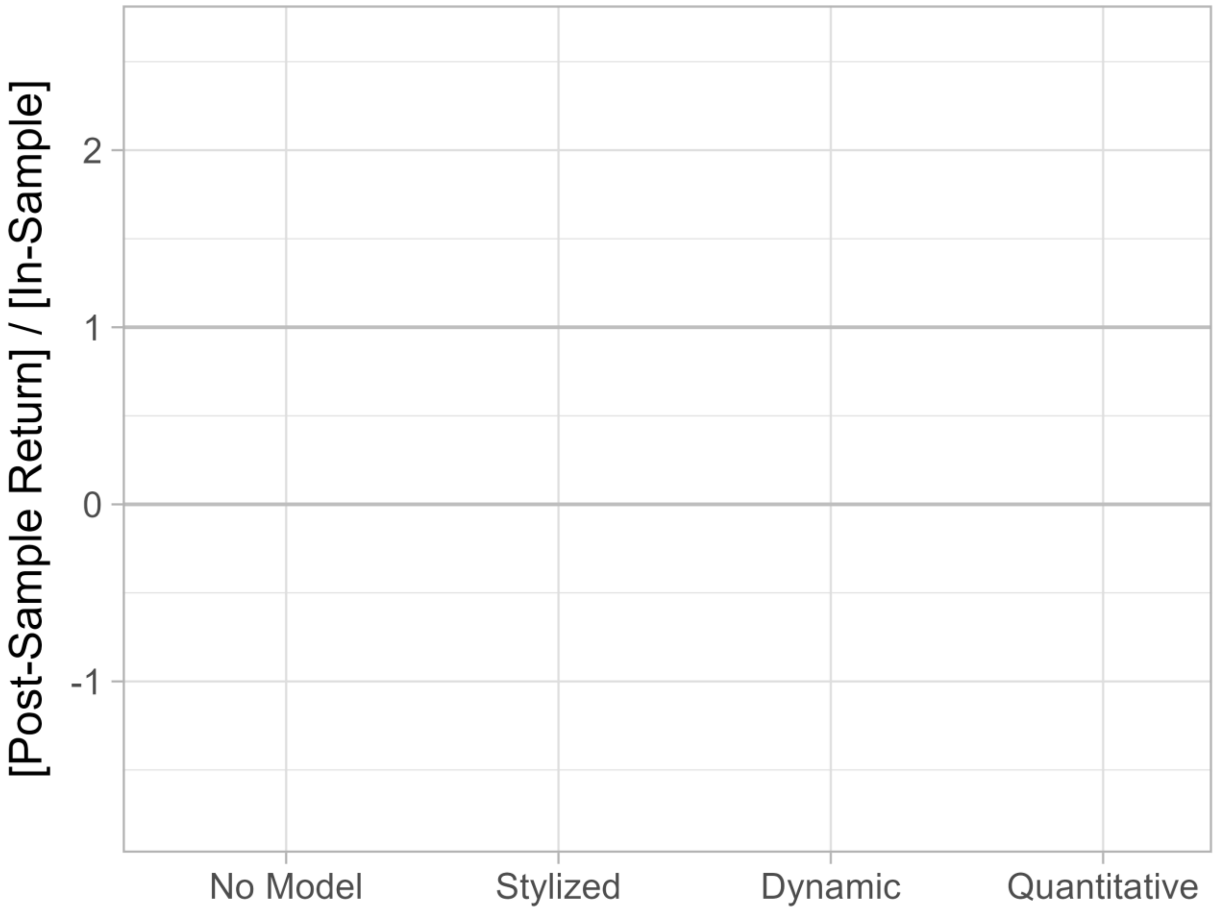
Robustness: Modeling Rigor

- Theory should help by disciplining the statistics (e.g. Fama French 2018)



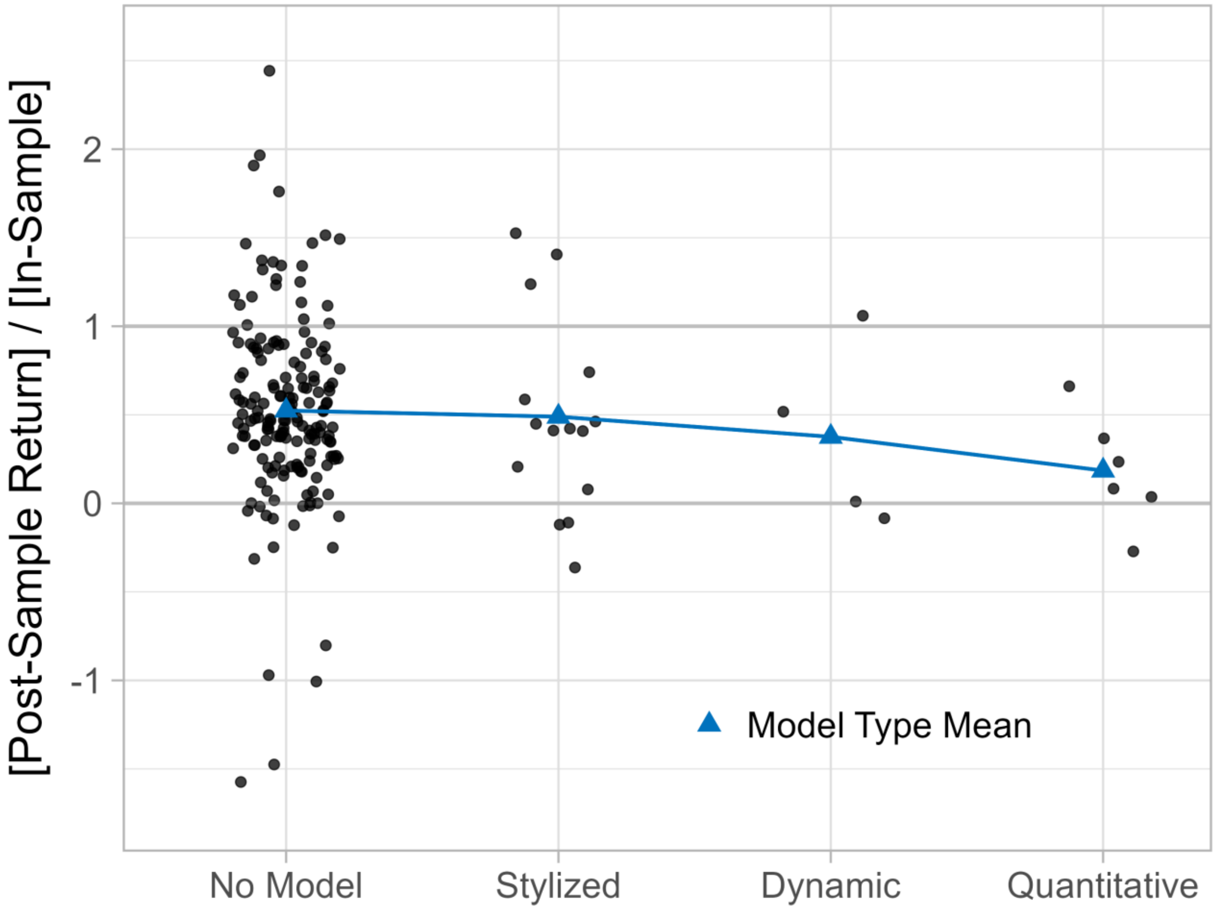
Robustness: Modeling Rigor

- Theory should help by disciplining the statistics (e.g. Fama French 2018)
- More rigorous theory ⇒ more discipline



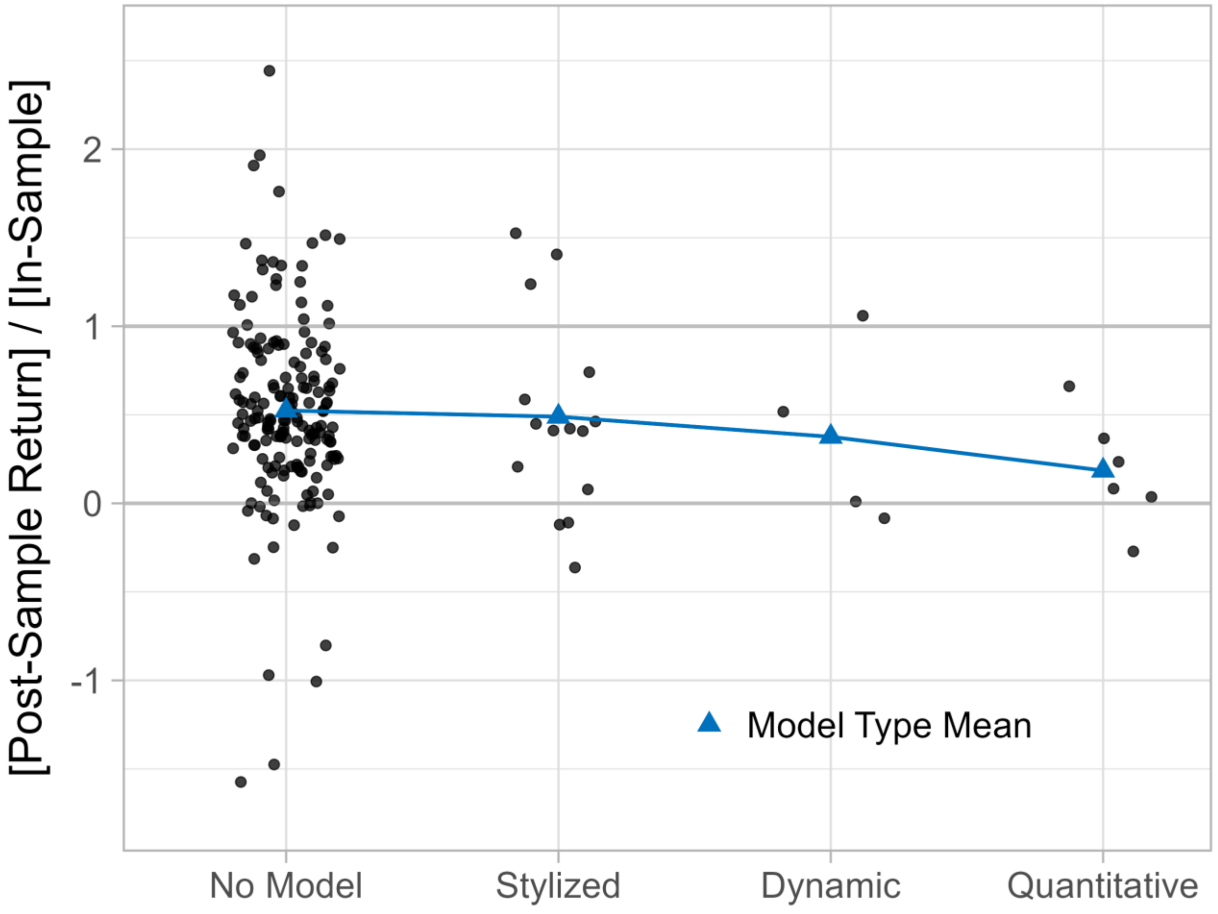
Robustness: Modeling Rigor

- Theory should help by disciplining the statistics (e.g. Fama French 2018)
- More rigorous theory ⇒ more discipline



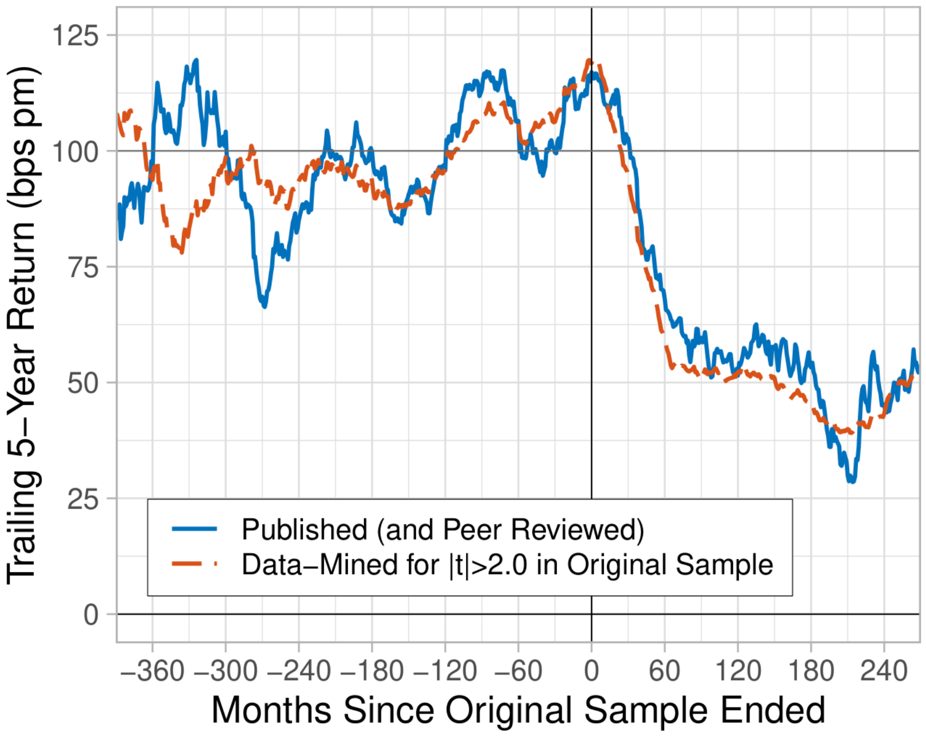
Robustness: Modeling Rigor

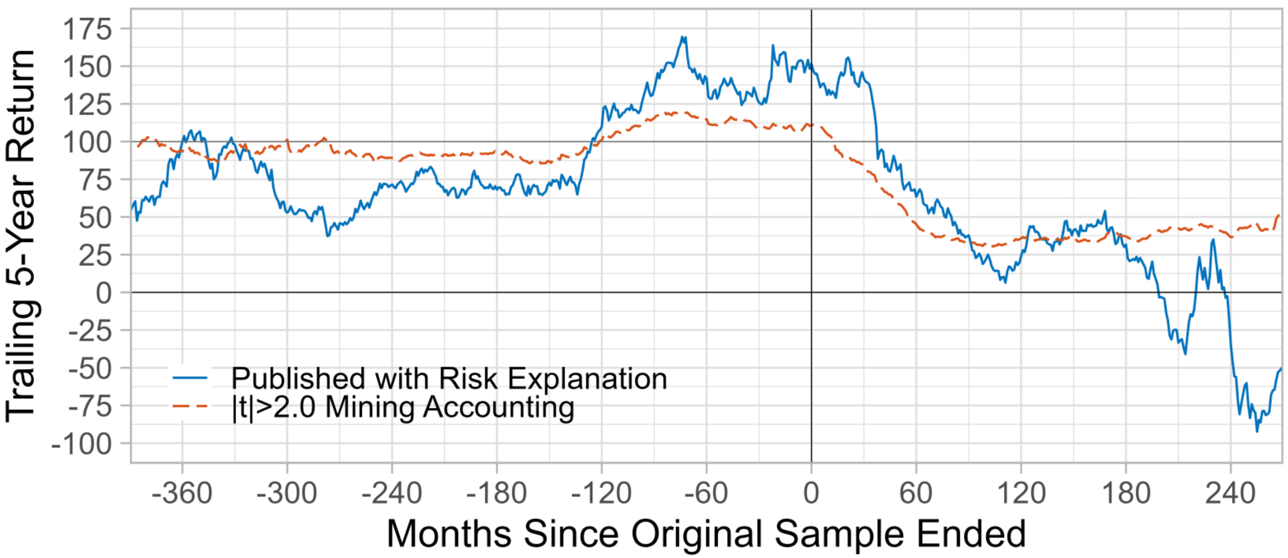
- Theory should help by disciplining the statistics (e.g. Fama French 2018)
- More rigorous theory ⇒ more discipline
- **Empirically: more discipline ⇒ less post-sample robustness**



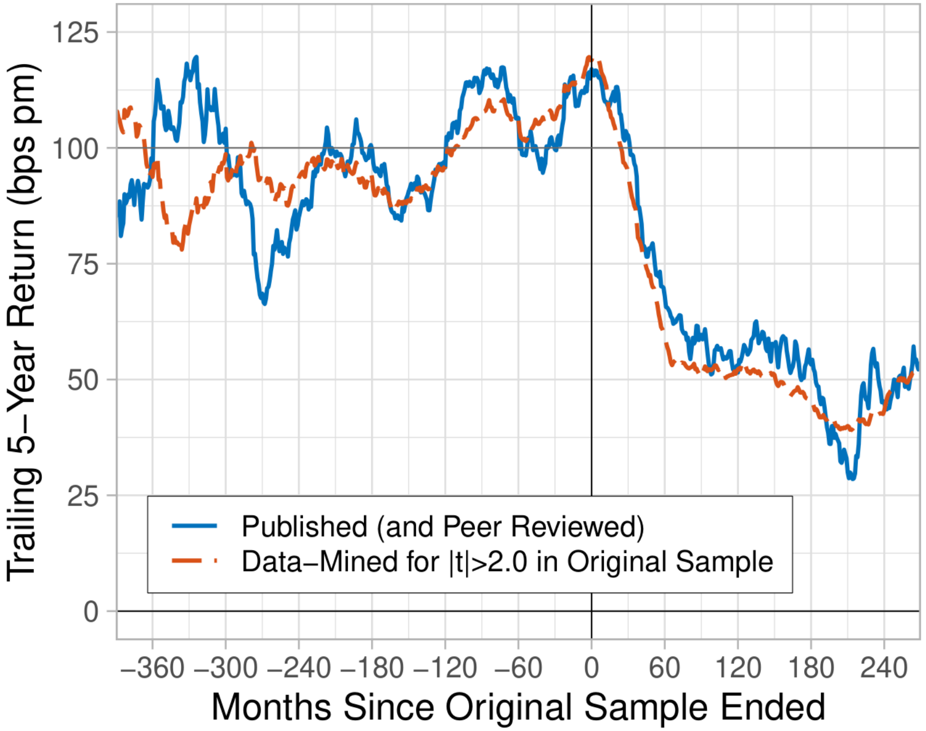
What do we make of this?

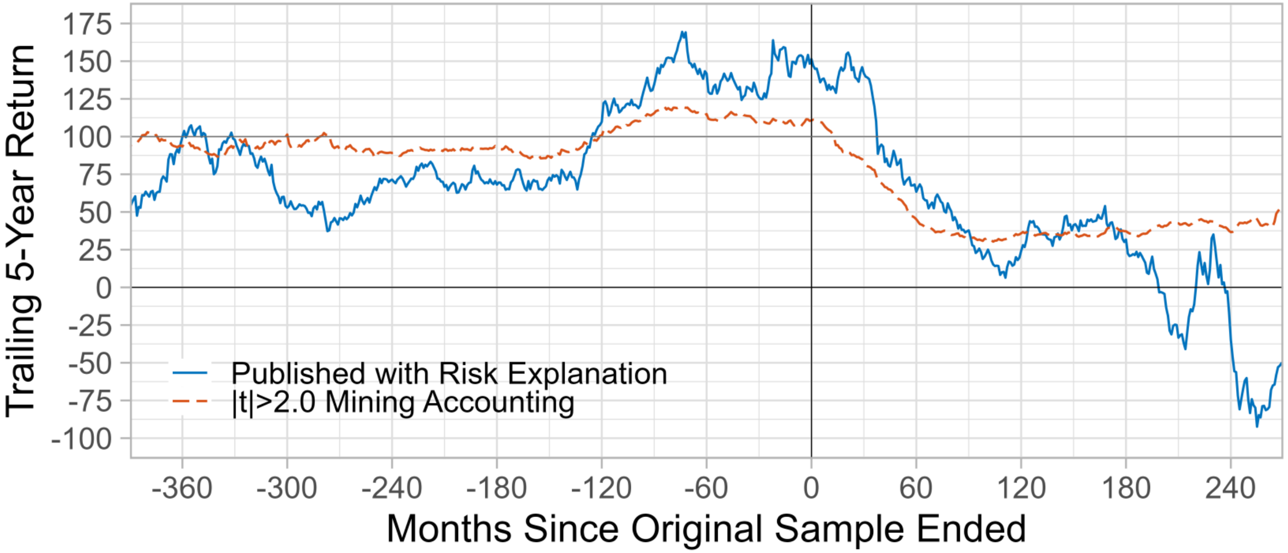
Peer reviewed predictability is similar to data mining---risk-based predictability is worse





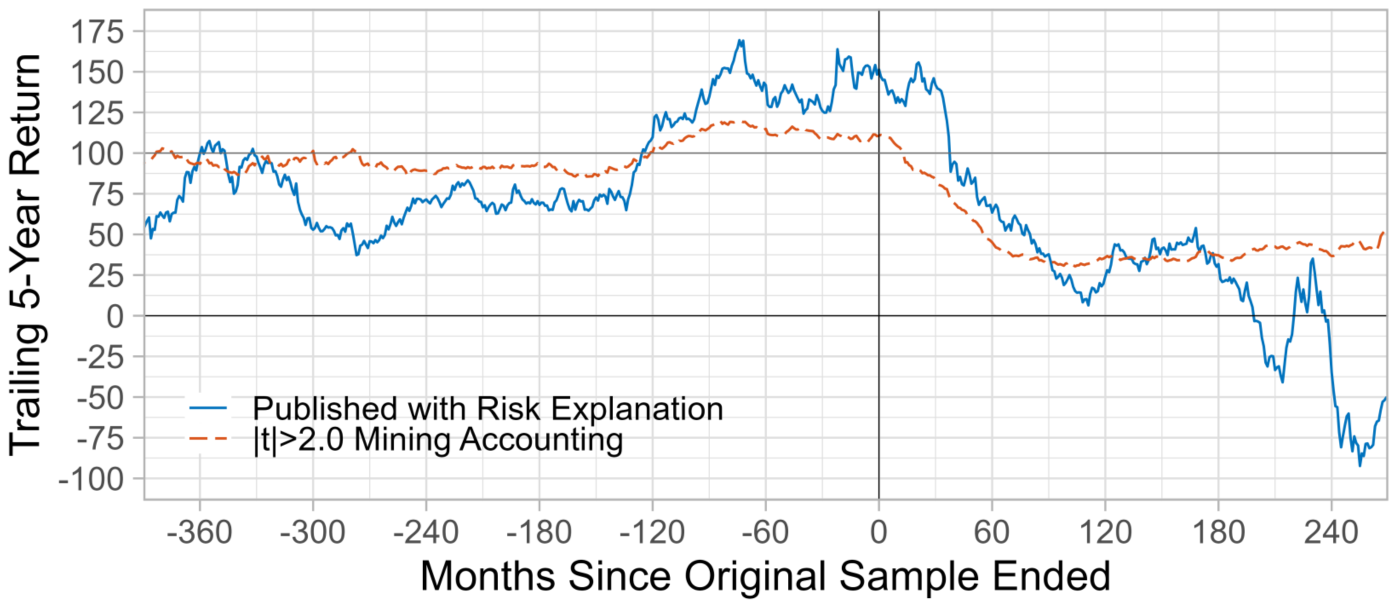
Peer reviewed predictability is similar to data mining---risk-based predictability is worse





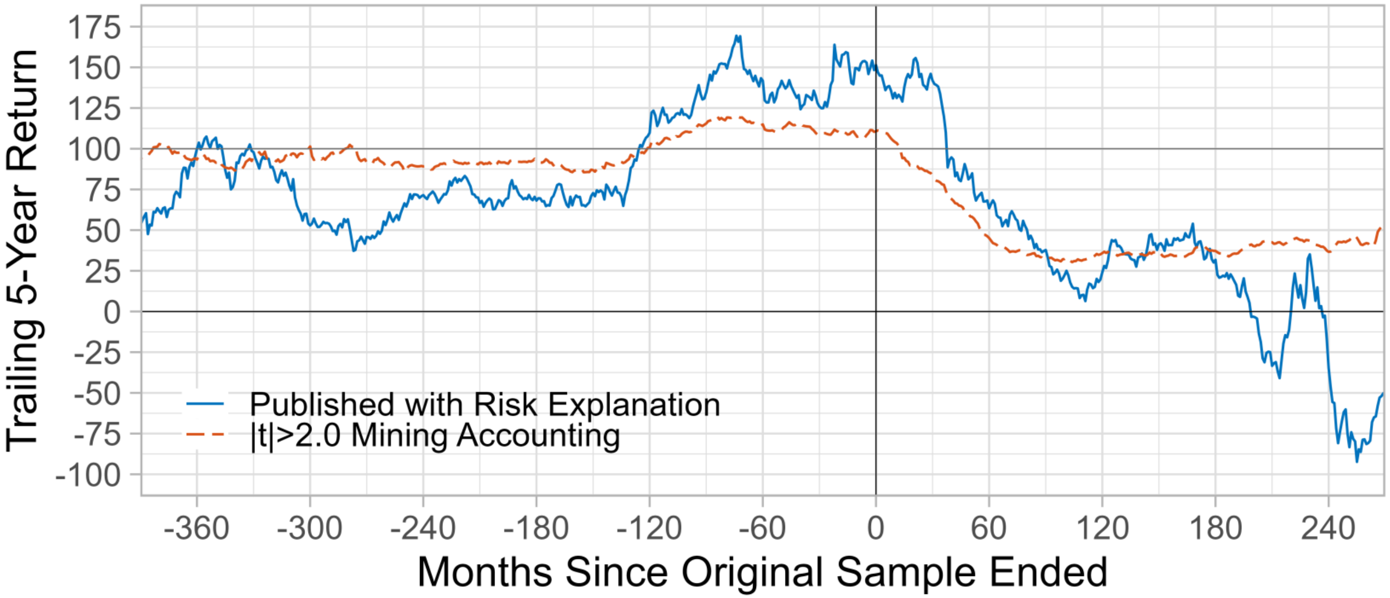
Two choices…

Choice 1: Cross-sectional stock predictability is not risk



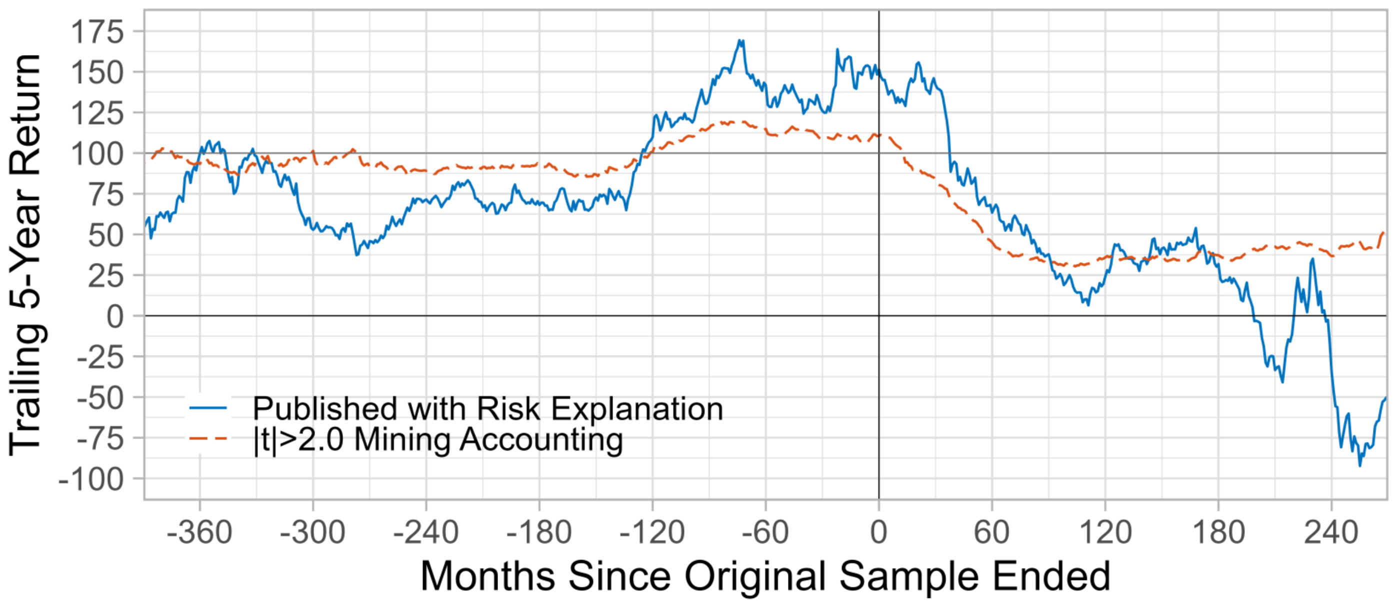
- Classical tests can only reject special cases of the class of risk theories

Choice 1: Cross-sectional stock predictability is not risk



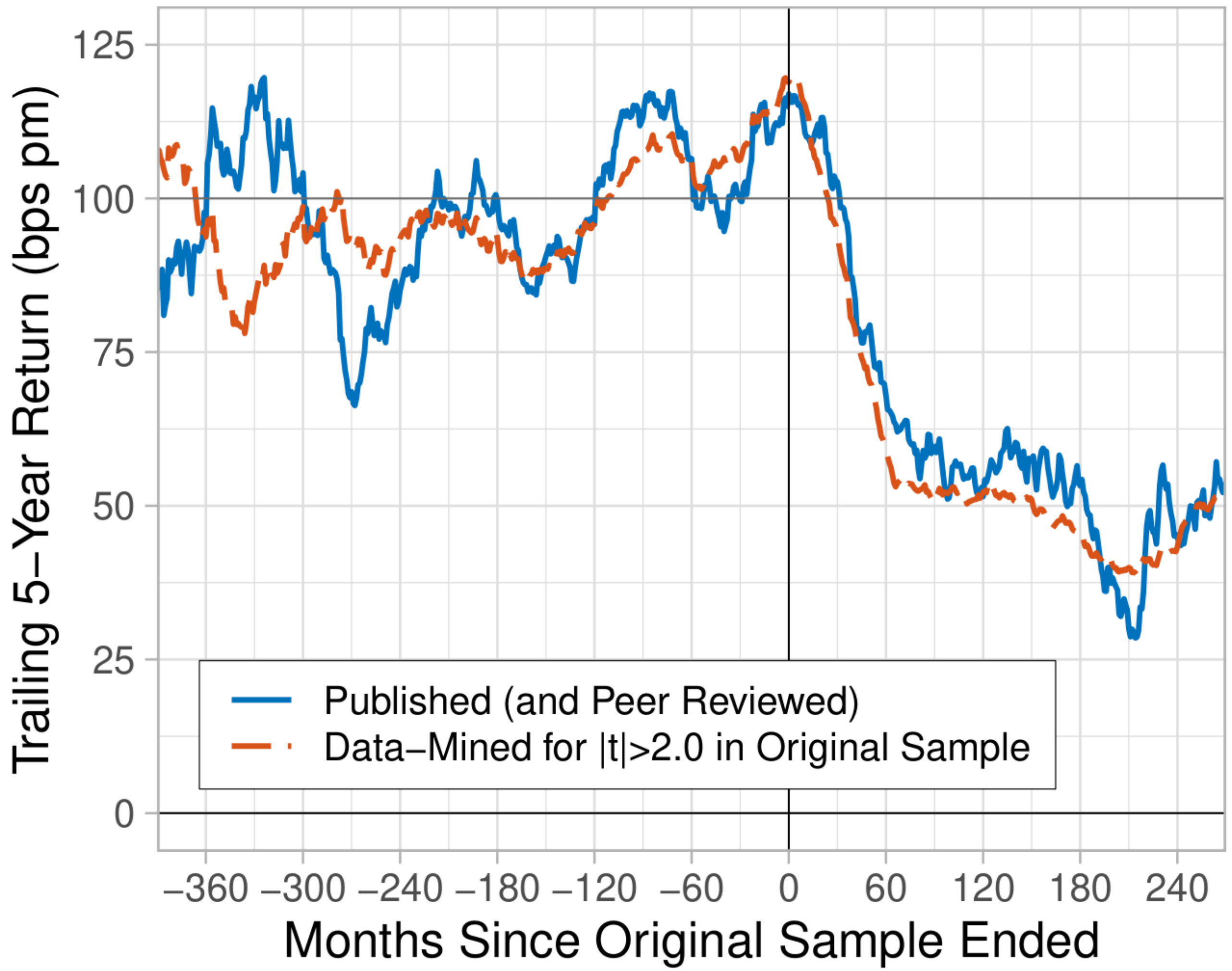
- Classical tests can only reject special cases of the class of risk theories
- But peer-review is a massive computer, designed to explore the full class

Choice 1: Cross-sectional stock predictability is not risk



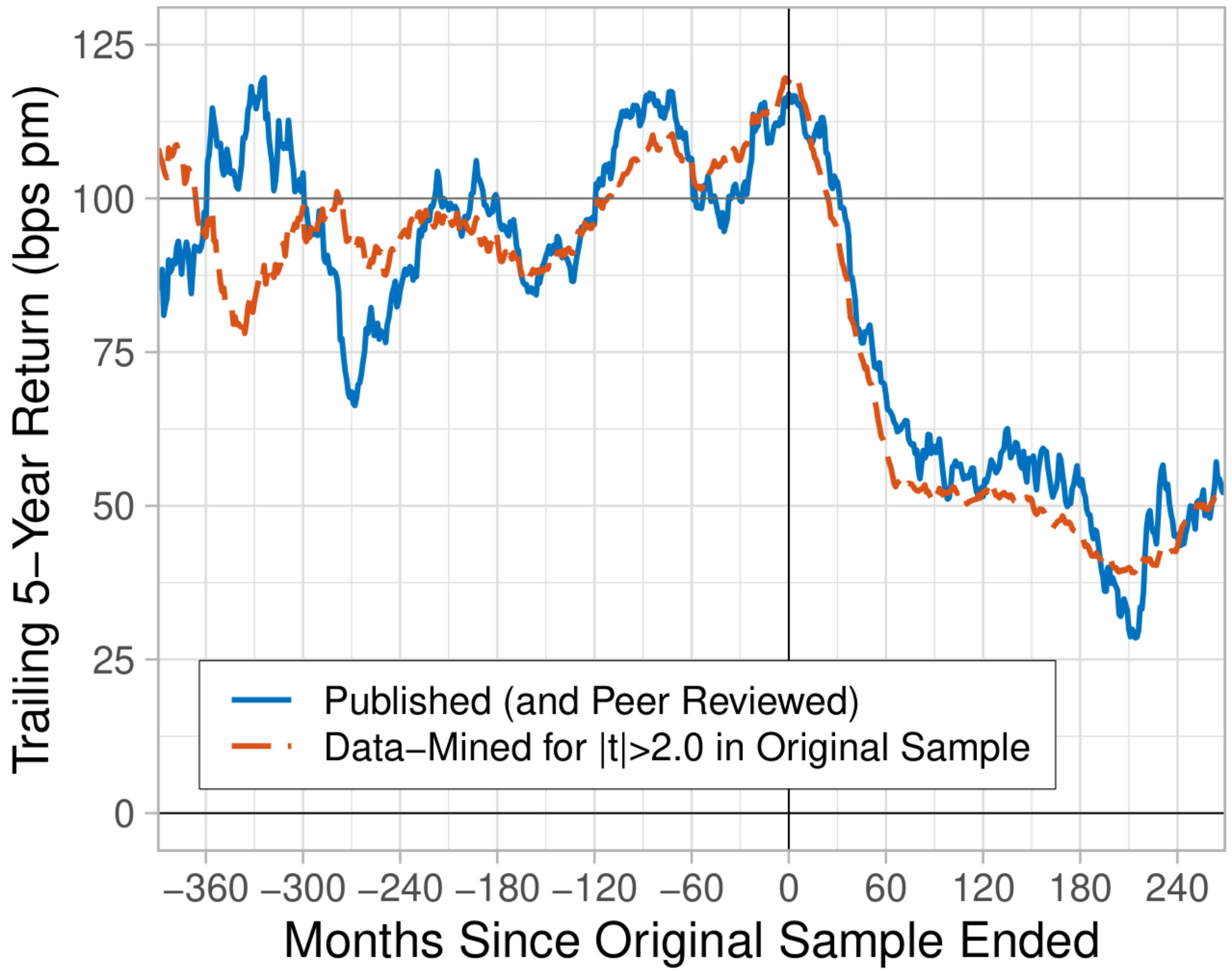
- Classical tests can only reject special cases of the class of risk theories
- But peer-review is a massive computer, designed to explore the full class
- **Over the past 40 years, this massive computer**
  - **Finds little risk**
  - **The "risk" it finds, decays out-of-sample, like data-mined predictability**

Or Choice 2: Peer review is not working properly



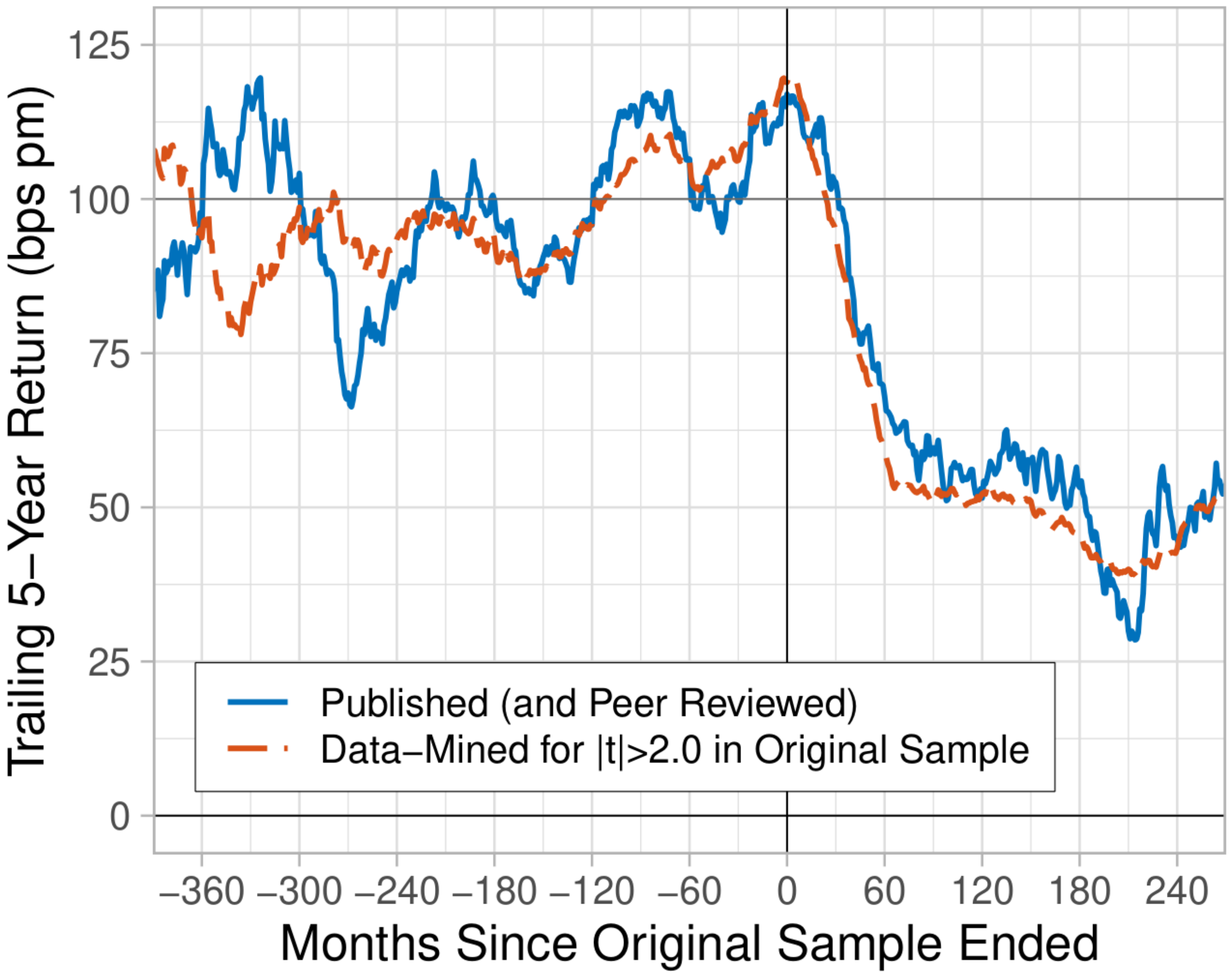
Or Choice 2: Peer review is not working properly

- Suppose passing peer review amounts to
  1. A long-short t-stat > 2
  2. **An economic parable unrelated to the real-world economy**



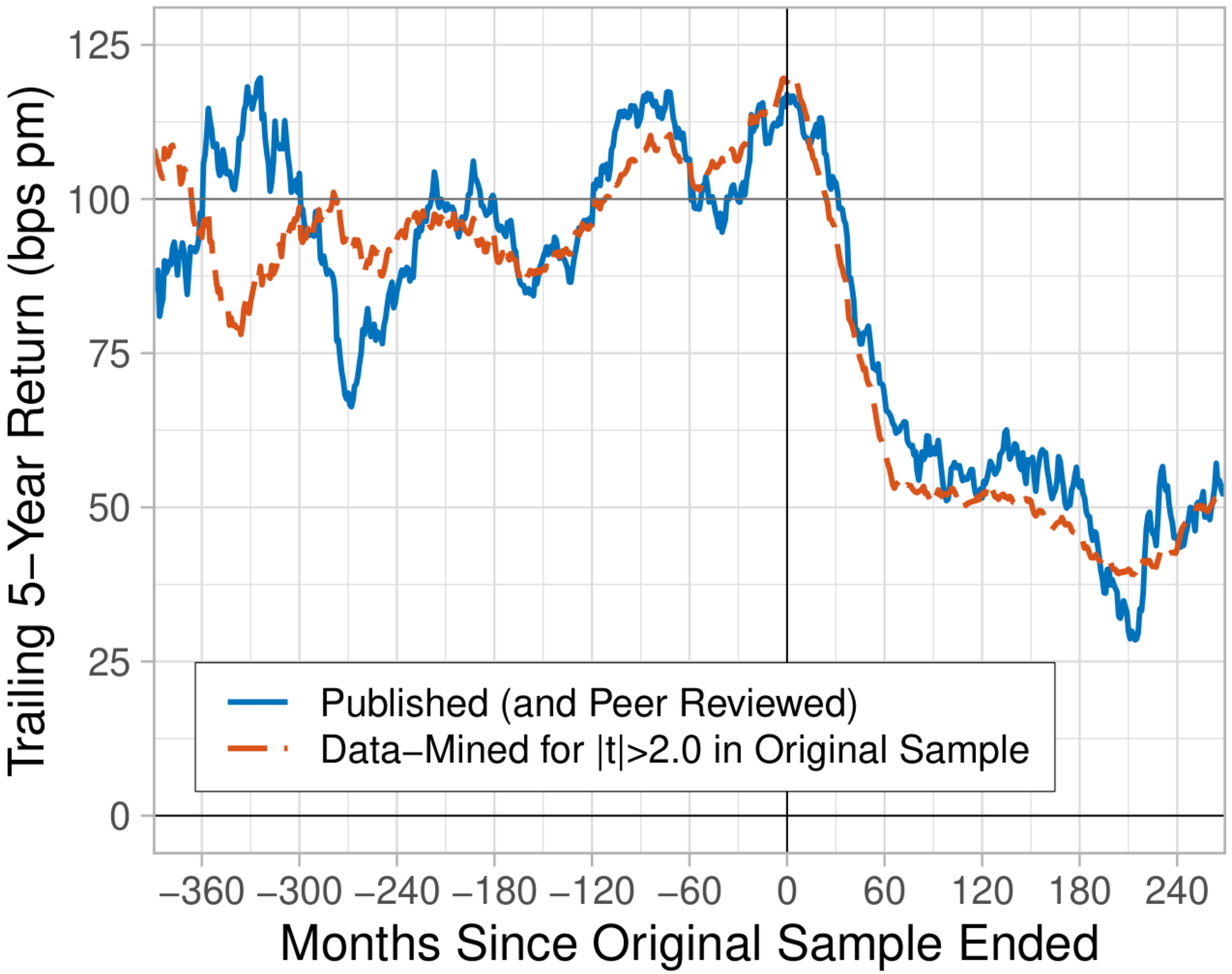
Or Choice 2: Peer review is not working properly

- Suppose passing peer review amounts to
  1. A long-short t-stat > 2
  2. **An economic parable unrelated to the real-world economy**
     - Perhaps, the parable confirms a referee's economic priors (Harvey 2017)



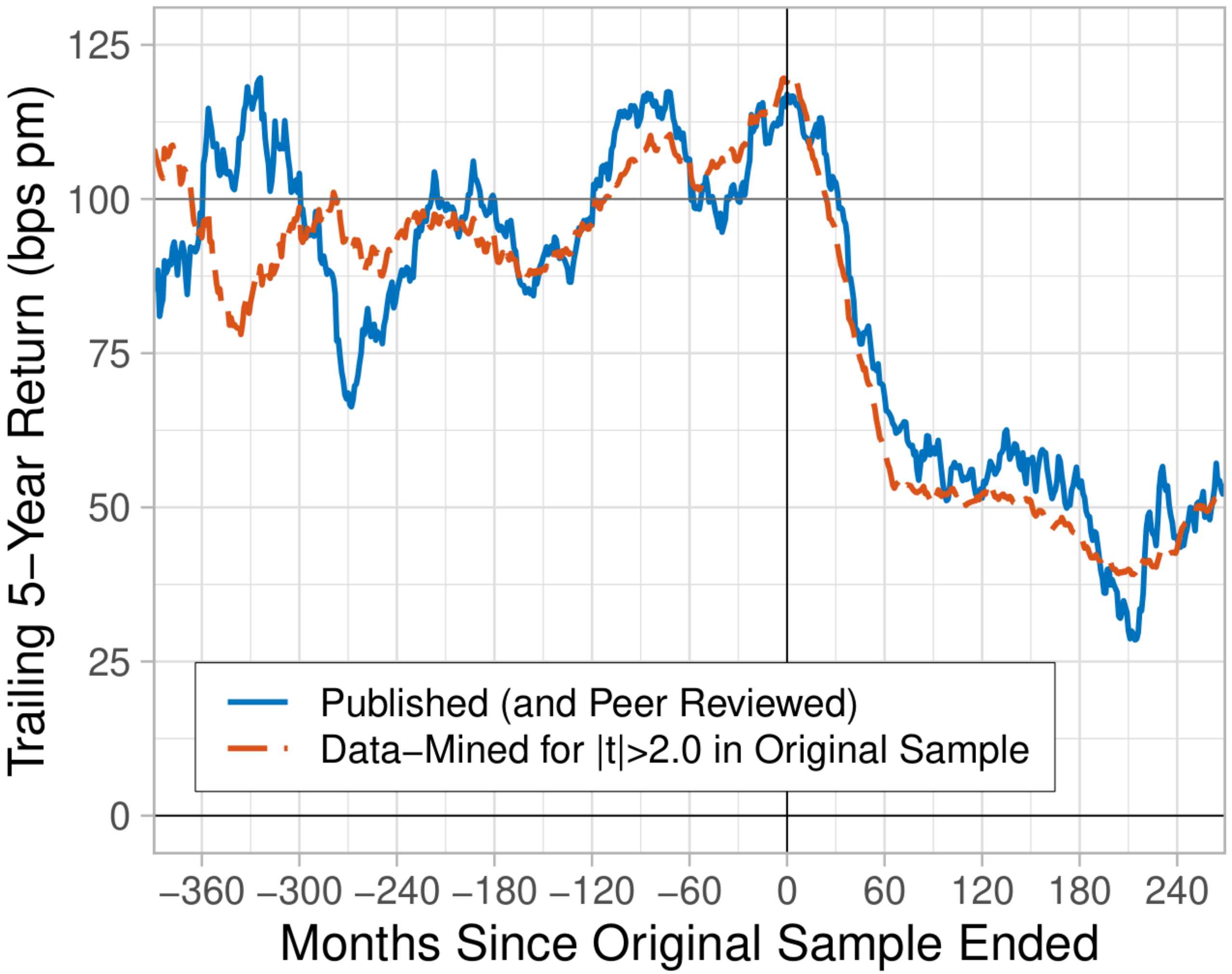
Or Choice 2: Peer review is not working properly

- Suppose passing peer review amounts to
  1. A long-short t-stat > 2
  2. **An economic parable unrelated to the real-world economy**
     - Perhaps, the parable confirms a referee's economic priors (Harvey 2017)
     - Or, it is written to boost strategic citations (Rubin-Rubin 2021 JPE)

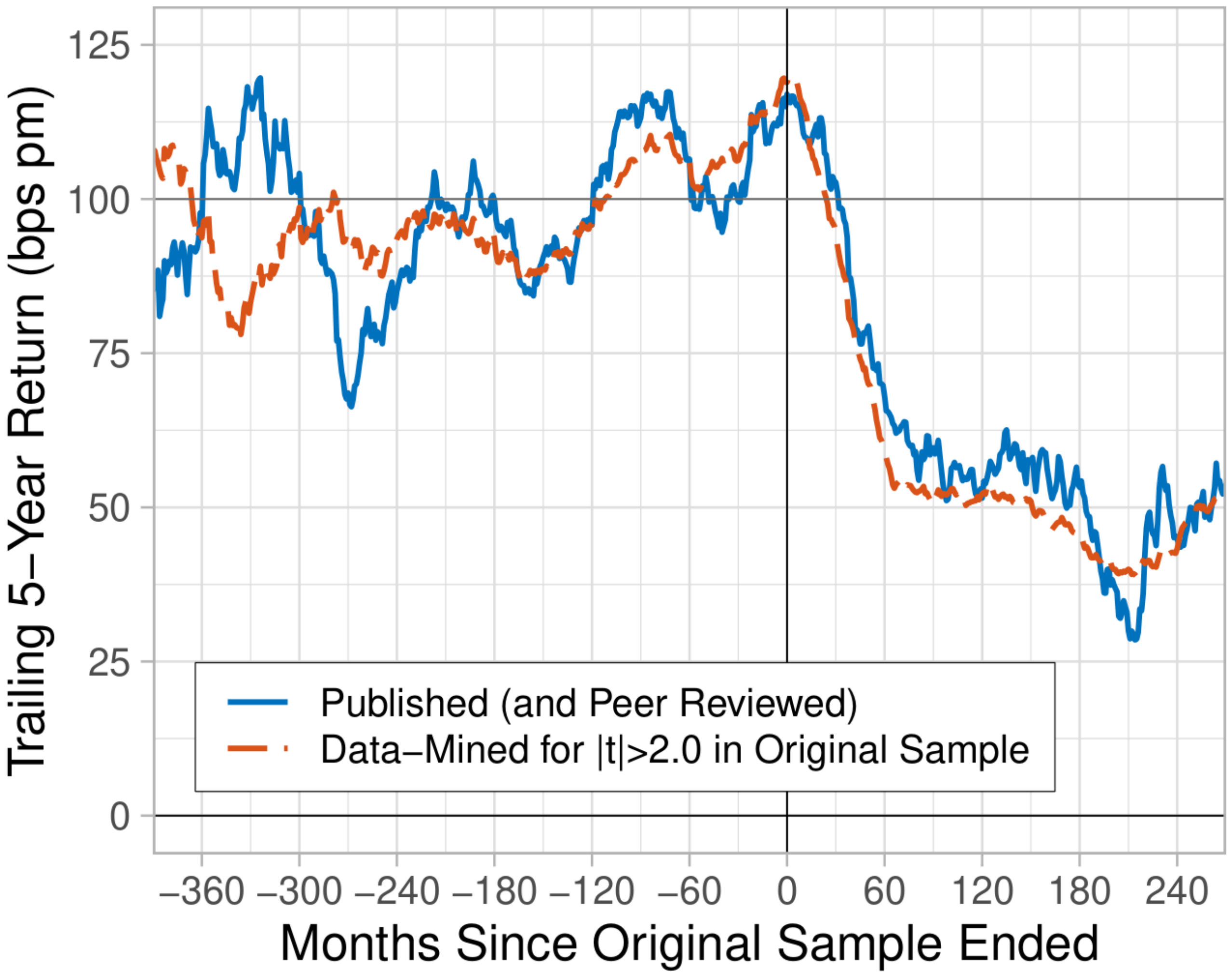


Or Choice 2: Peer review is not working properly

- Suppose passing peer review amounts to
  1. A long-short t-stat > 2
  2. **An economic parable unrelated to the real-world economy**
     - Perhaps, the parable confirms a referee's economic priors (Harvey 2017)
     - Or, it is written to boost strategic citations (Rubin-Rubin 2021 JPE)
- **We cannot reject this model**

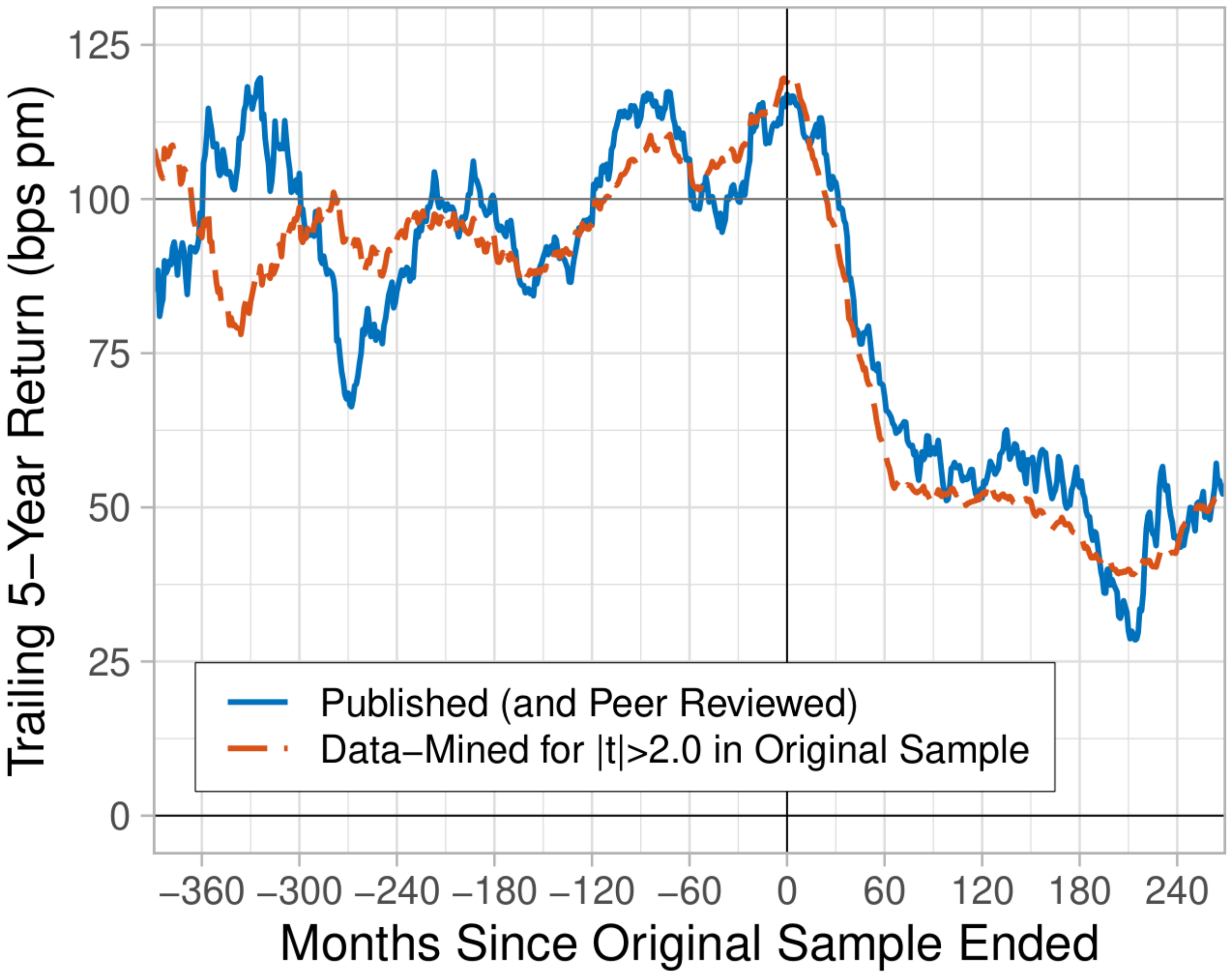


Regardless, data mining is clearly undervalued



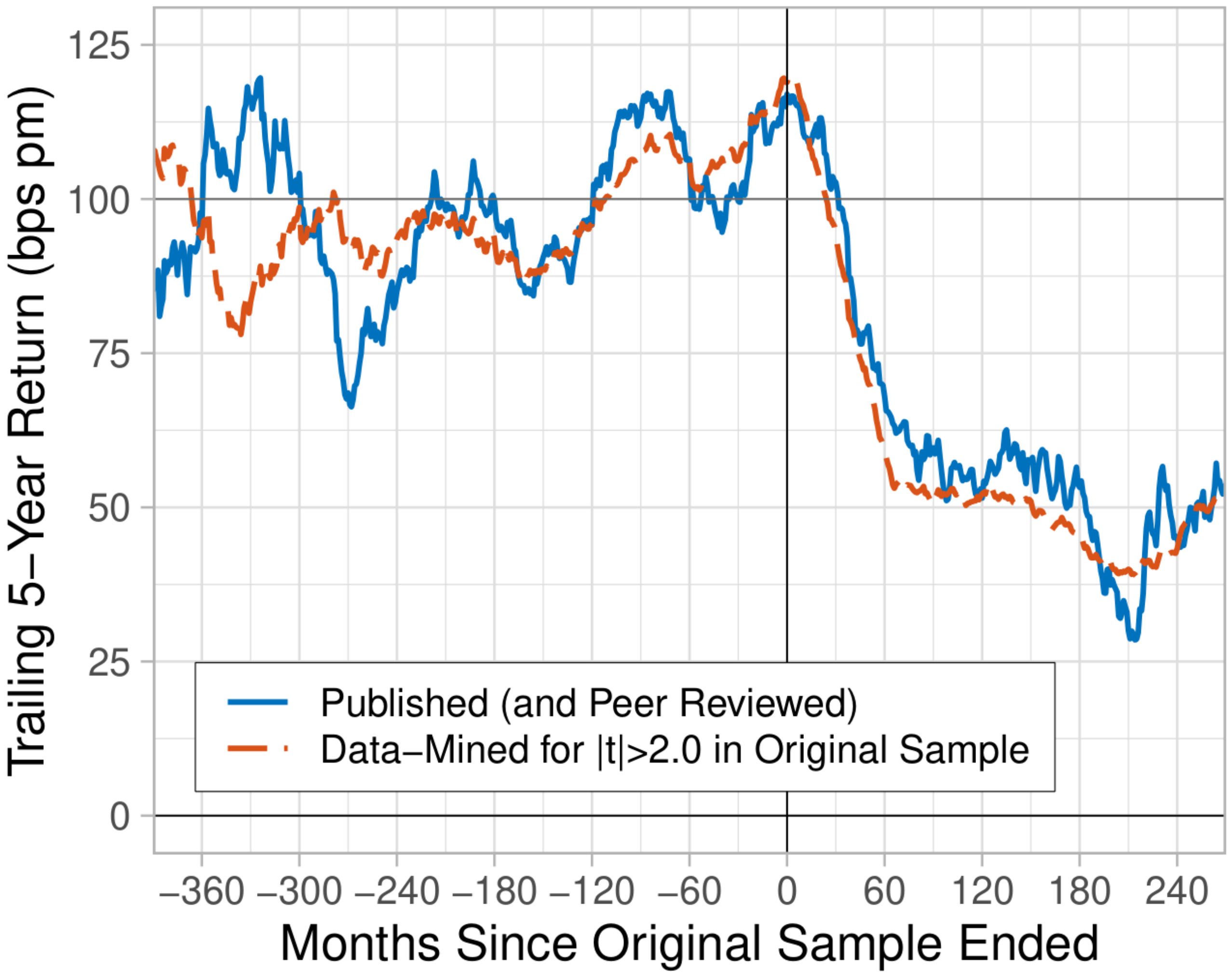
Regardless, data mining is clearly undervalued

- It uncovers true, out-of-sample predictability



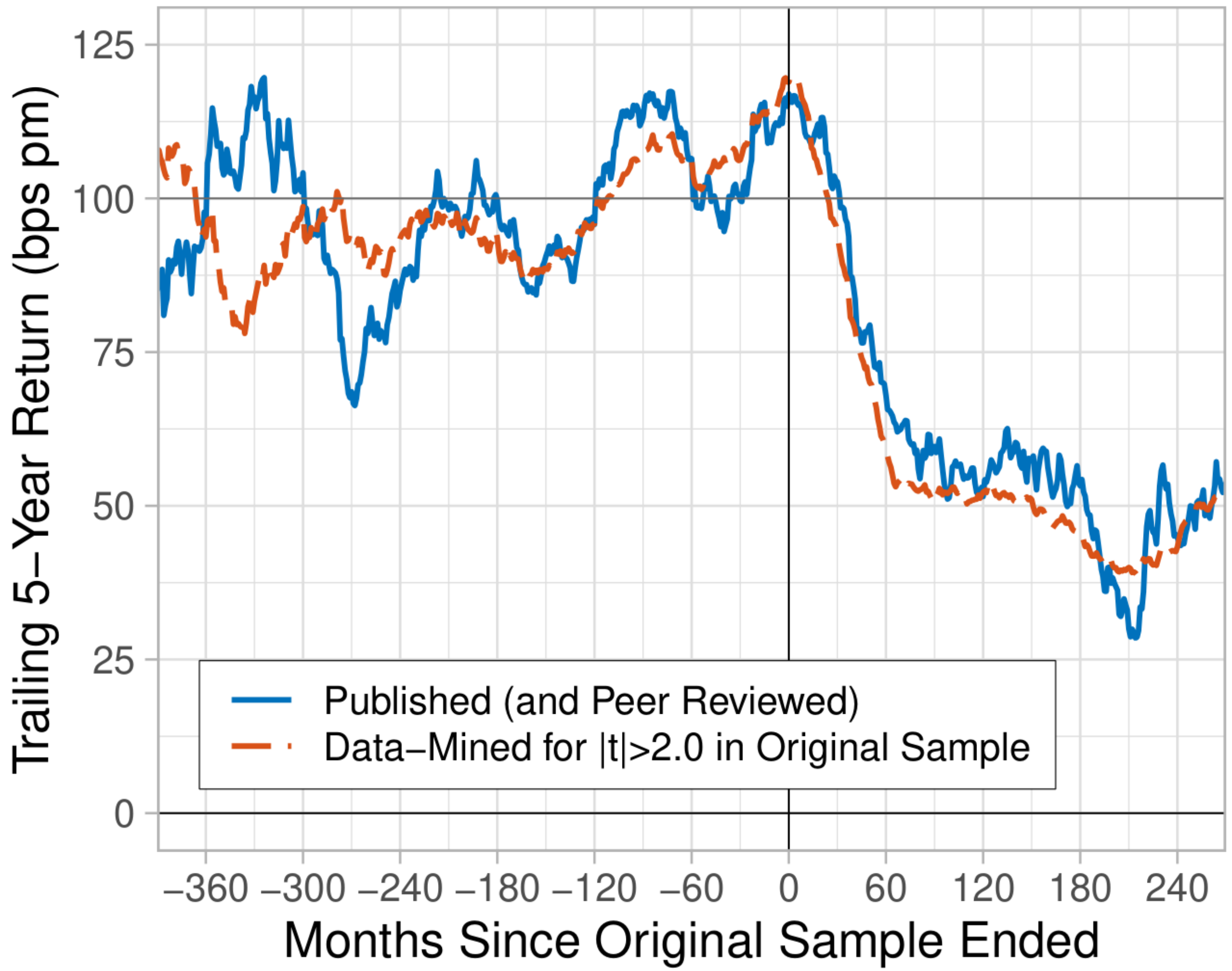
Regardless, data mining is clearly undervalued

- It uncovers true, out-of-sample predictability
- **It uncovers**
  - **the investment anomaly**



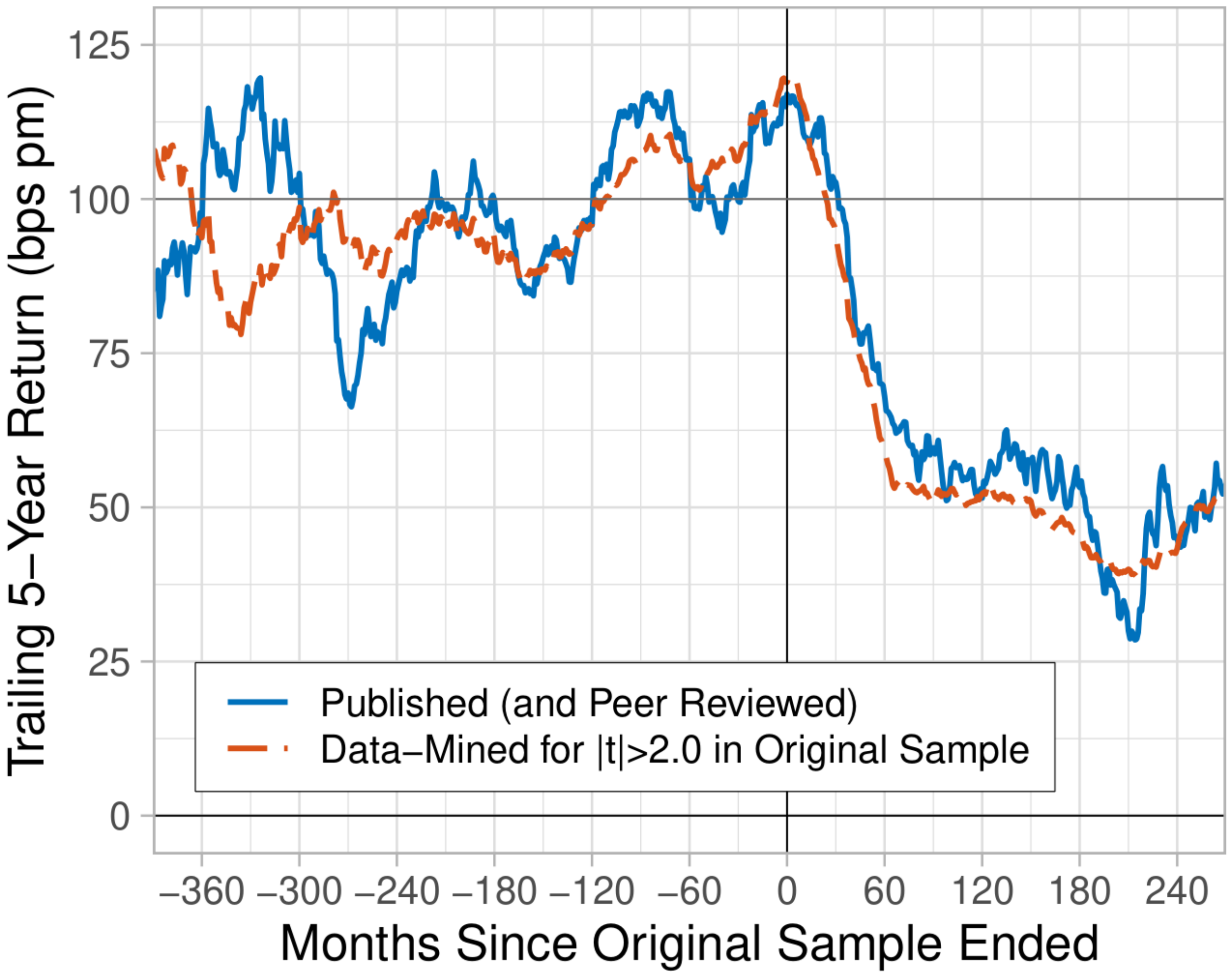
Regardless, data mining is clearly undervalued

- It uncovers true, out-of-sample predictability
- **It uncovers**
  - **the investment anomaly**
  - earnings surprise



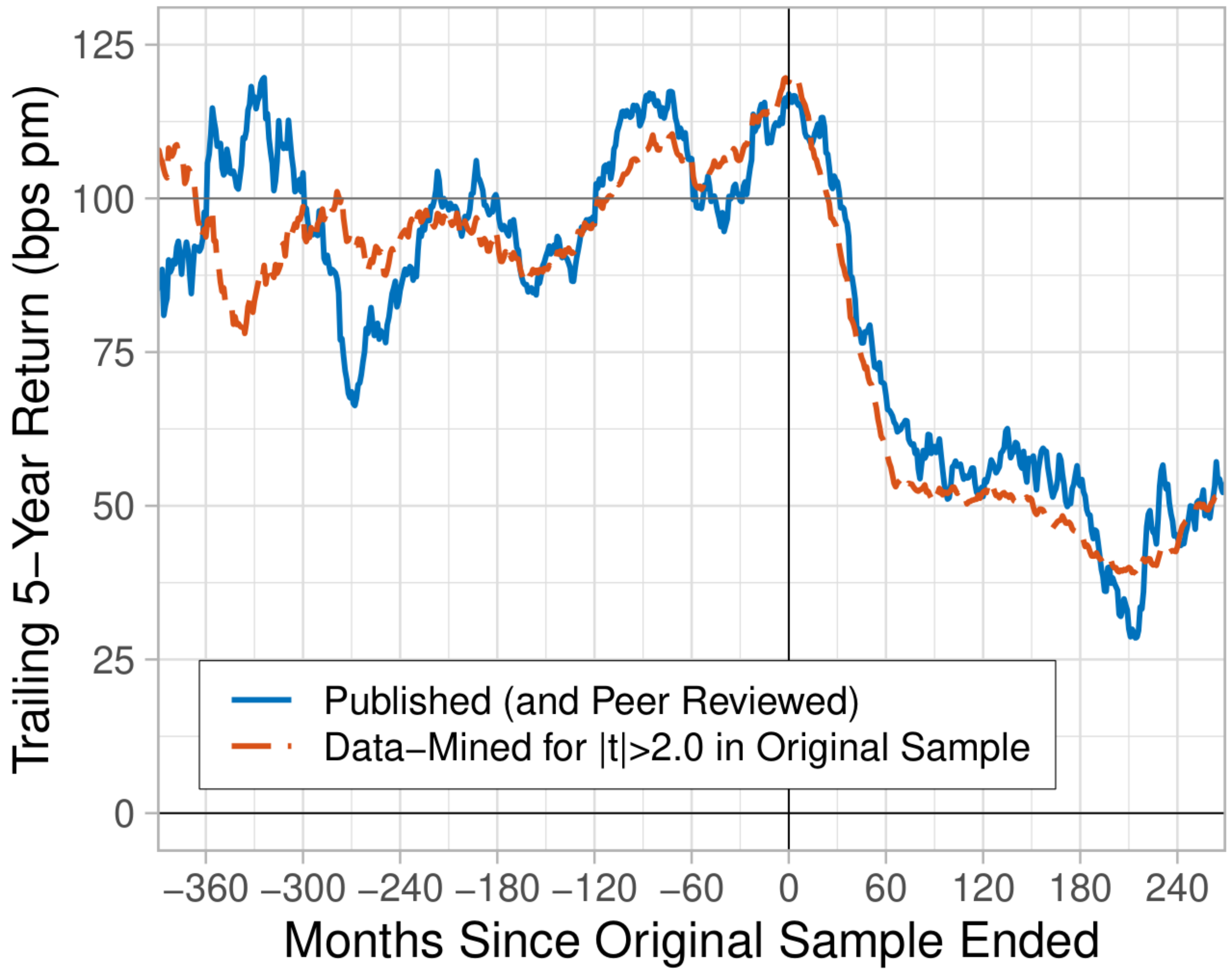
Regardless, data mining is clearly undervalued

- It uncovers true, out-of-sample predictability
- **It uncovers**
  - **the investment anomaly**
  - earnings surprise
  - accruals, inventory growth



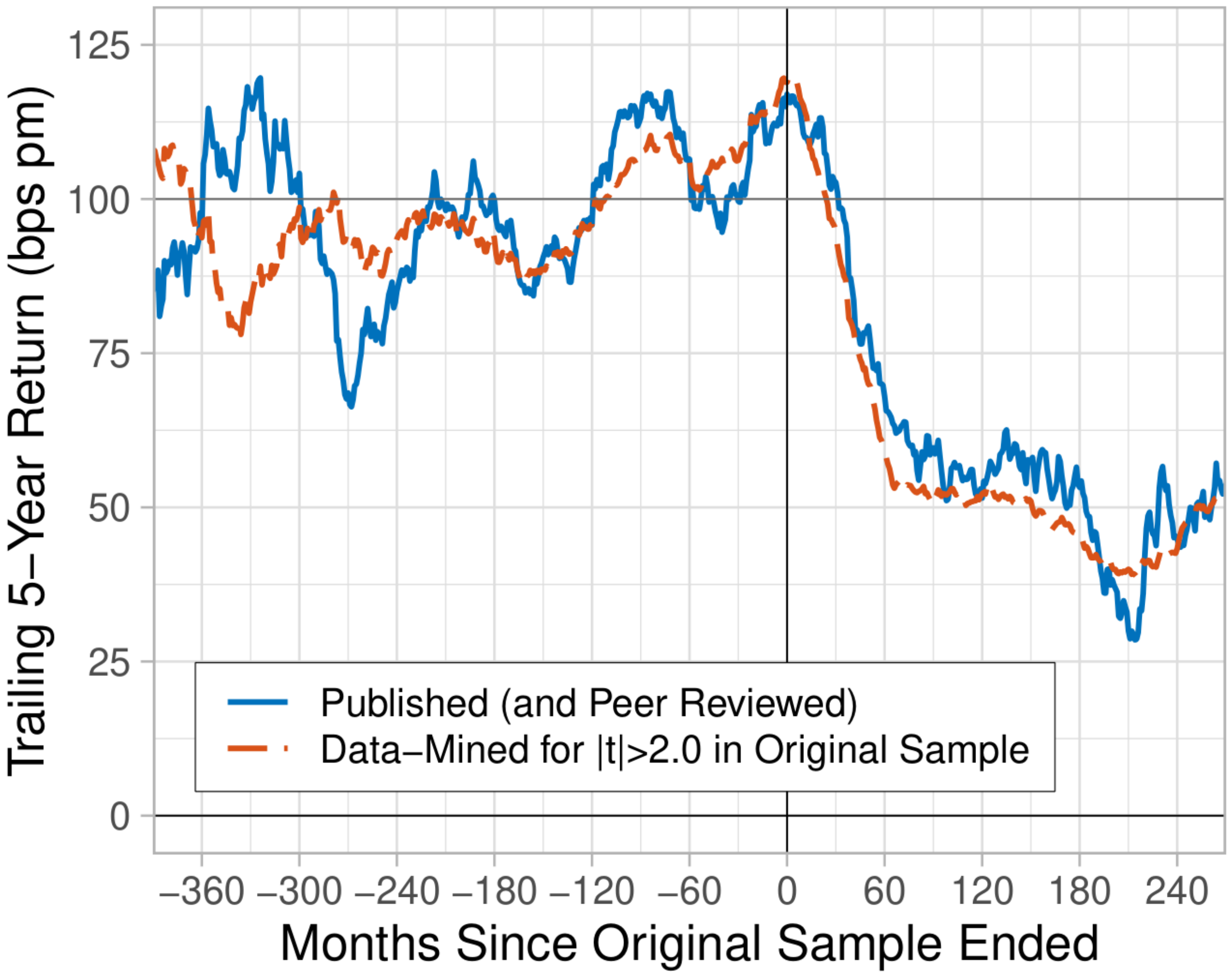
Regardless, data mining is clearly undervalued

- It uncovers true, out-of-sample predictability
- **It uncovers**
  - **the investment anomaly**
  - earnings surprise
  - accruals, inventory growth
  - stock issuance, debt issuance



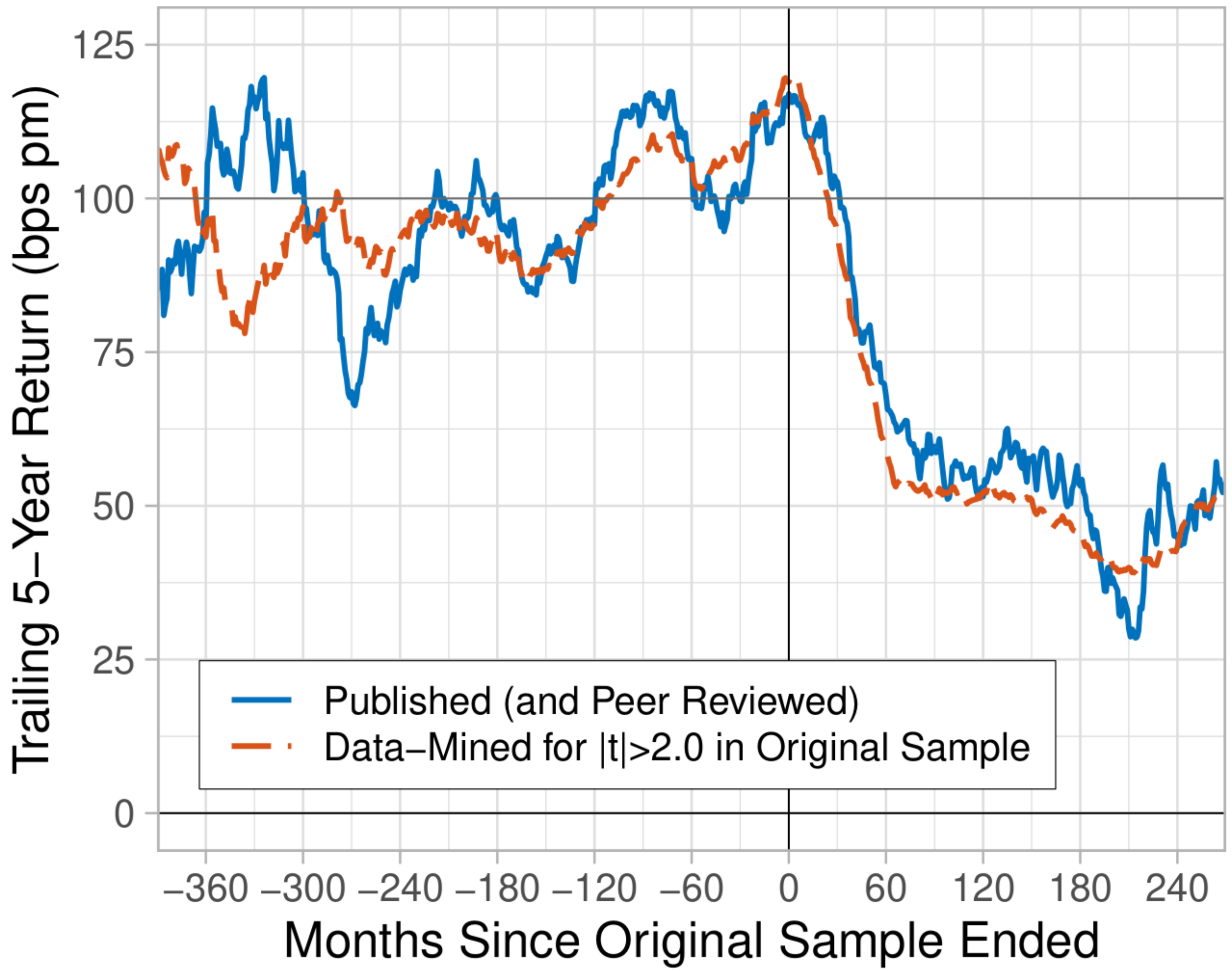
Regardless, data mining is clearly undervalued

- It uncovers true, out-of-sample predictability
- **It uncovers**
  - **the investment anomaly**
  - earnings surprise
  - accruals, inventory growth
  - stock issuance, debt issuance
  - **long before they are published**



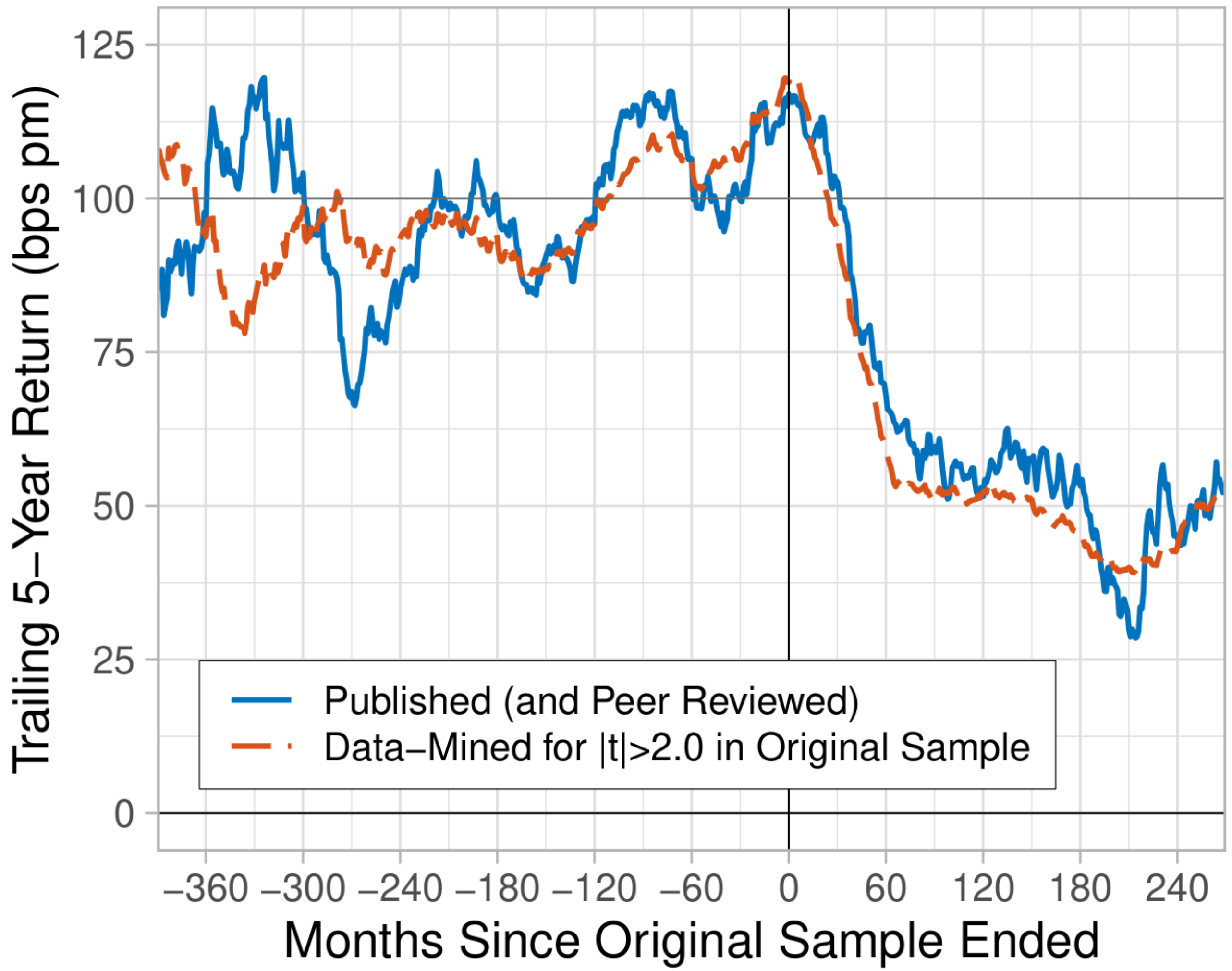
Regardless, data mining is clearly undervalued

- It uncovers true, out-of-sample predictability
- **It uncovers**
  - **the investment anomaly**
  - earnings surprise
  - accruals, inventory growth
  - stock issuance, debt issuance
  - **long before they are published**
- Multiple testing methods remove data-mining bias (Chen-Dim '24)



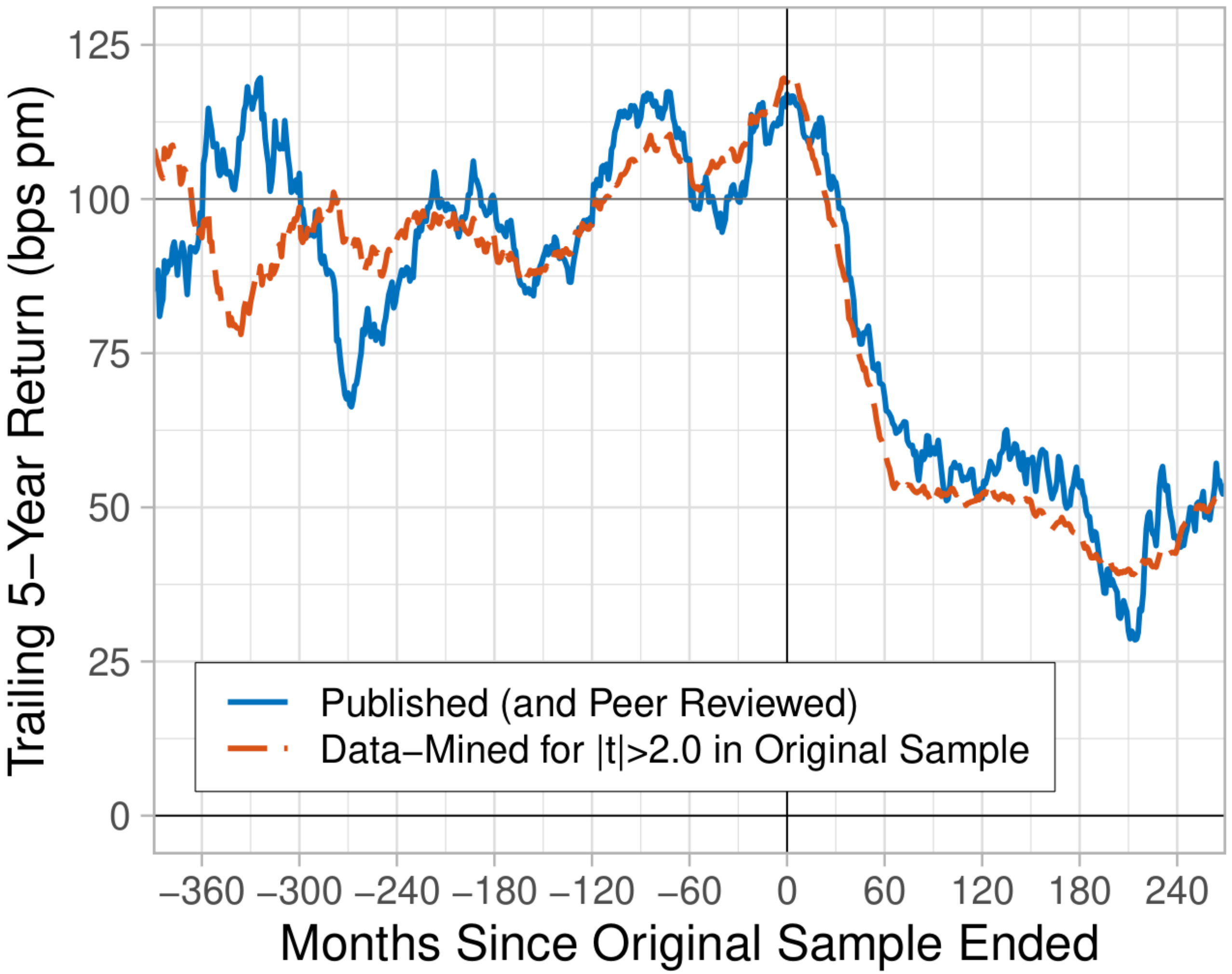
Regardless, data mining is clearly undervalued

- It uncovers true, out-of-sample predictability
- **It uncovers**
  - **the investment anomaly**
  - earnings surprise
  - accruals, inventory growth
  - stock issuance, debt issuance
  - **long before they are published**
- Multiple testing methods remove data-mining bias (Chen-Dim ‘24)
- Other fields have turned to data-centric methods (e.g. ChatGPT)



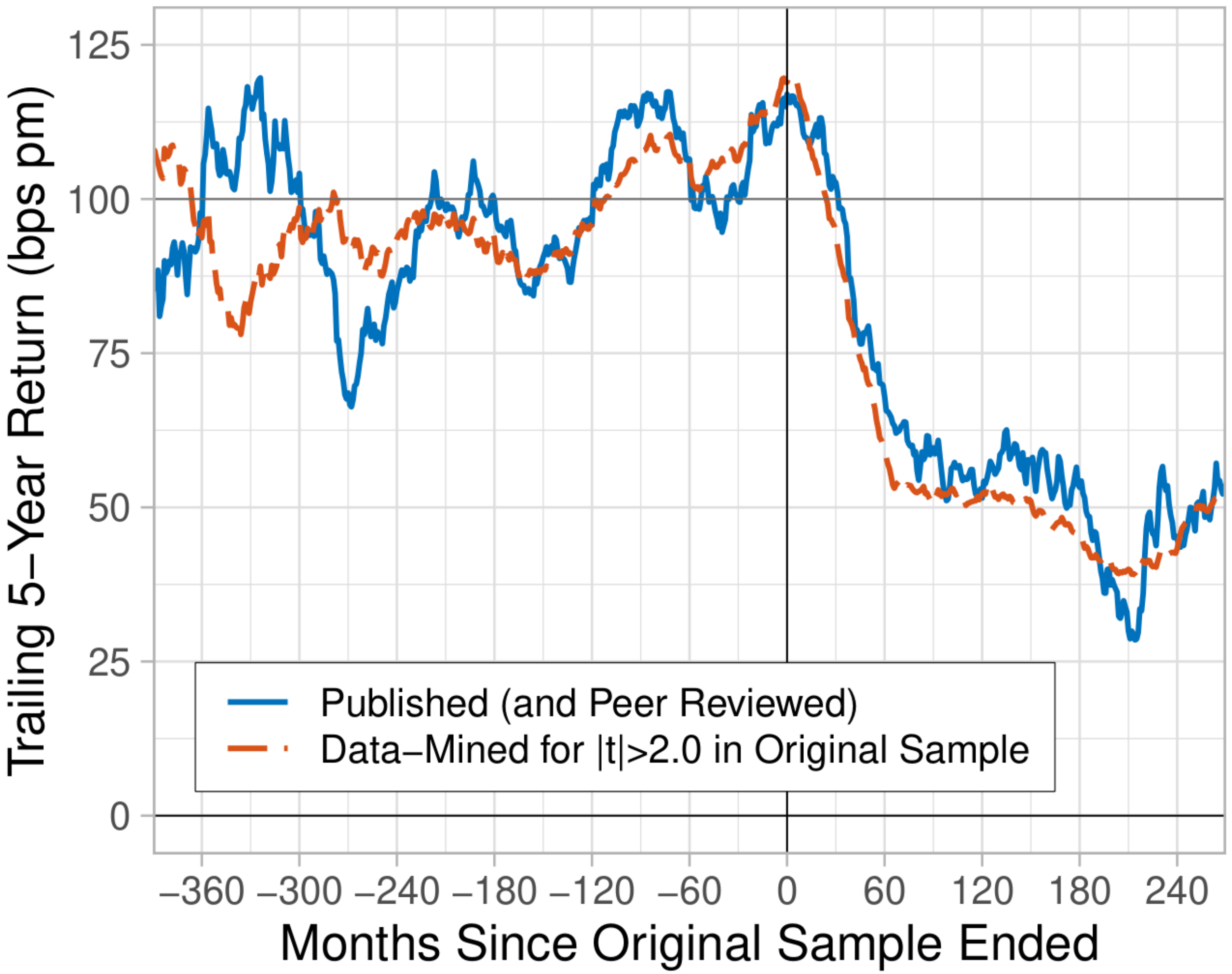
Regardless, data mining is clearly undervalued

- **Sutton's (2019) "Bitter Lesson"** from 70 years of AI research



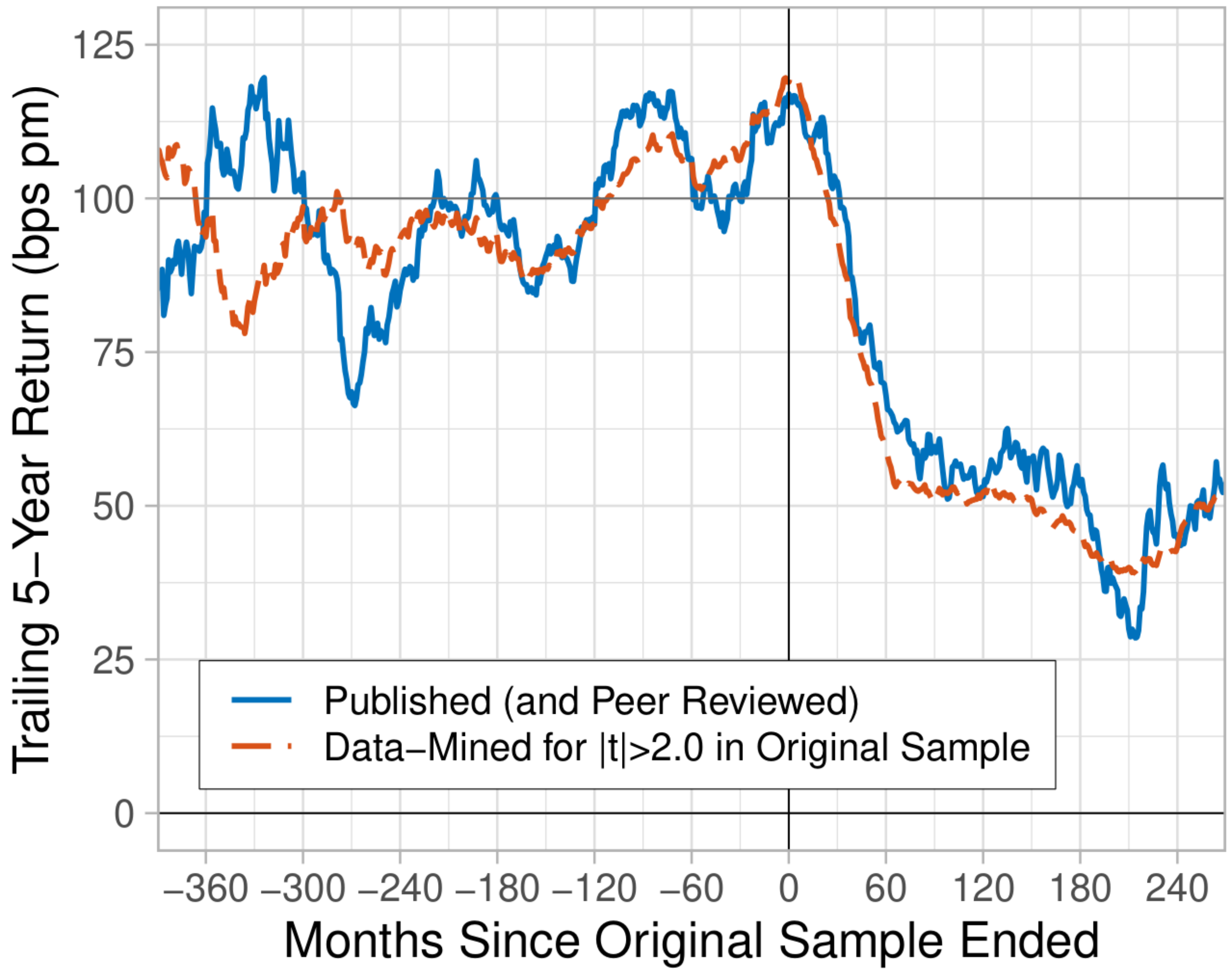
Regardless, data mining is clearly undervalued

- **Sutton's (2019) "Bitter Lesson"** from 70 years of AI research
  - Beloved, hand-crafted solutions end up "irrelevant, or worse"



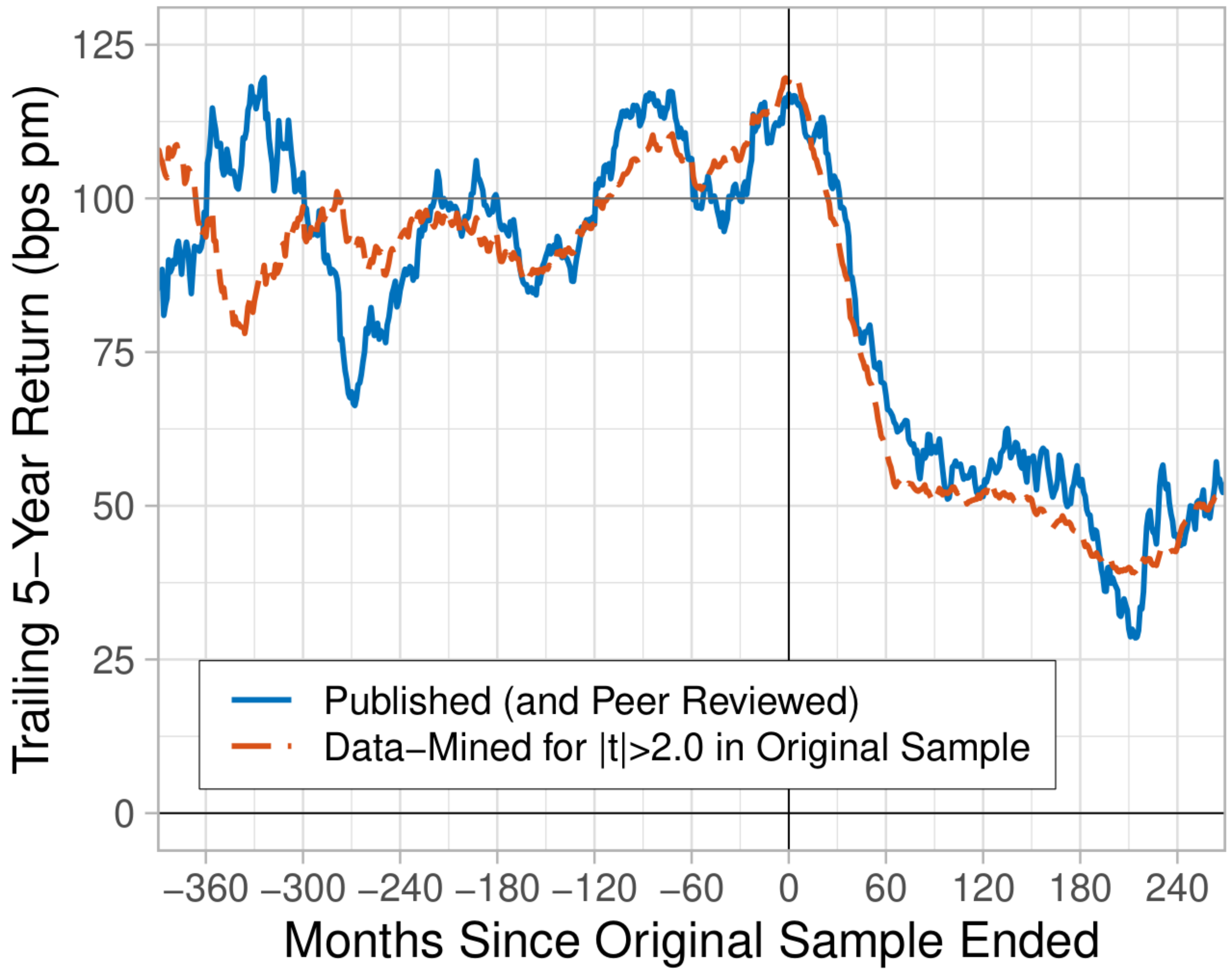
Regardless, data mining is clearly undervalued

- **Sutton's (2019) "Bitter Lesson"** from 70 years of AI research
  - Beloved, hand-crafted solutions end up "irrelevant, or worse"
  - Vast searches through huge datasets outperform



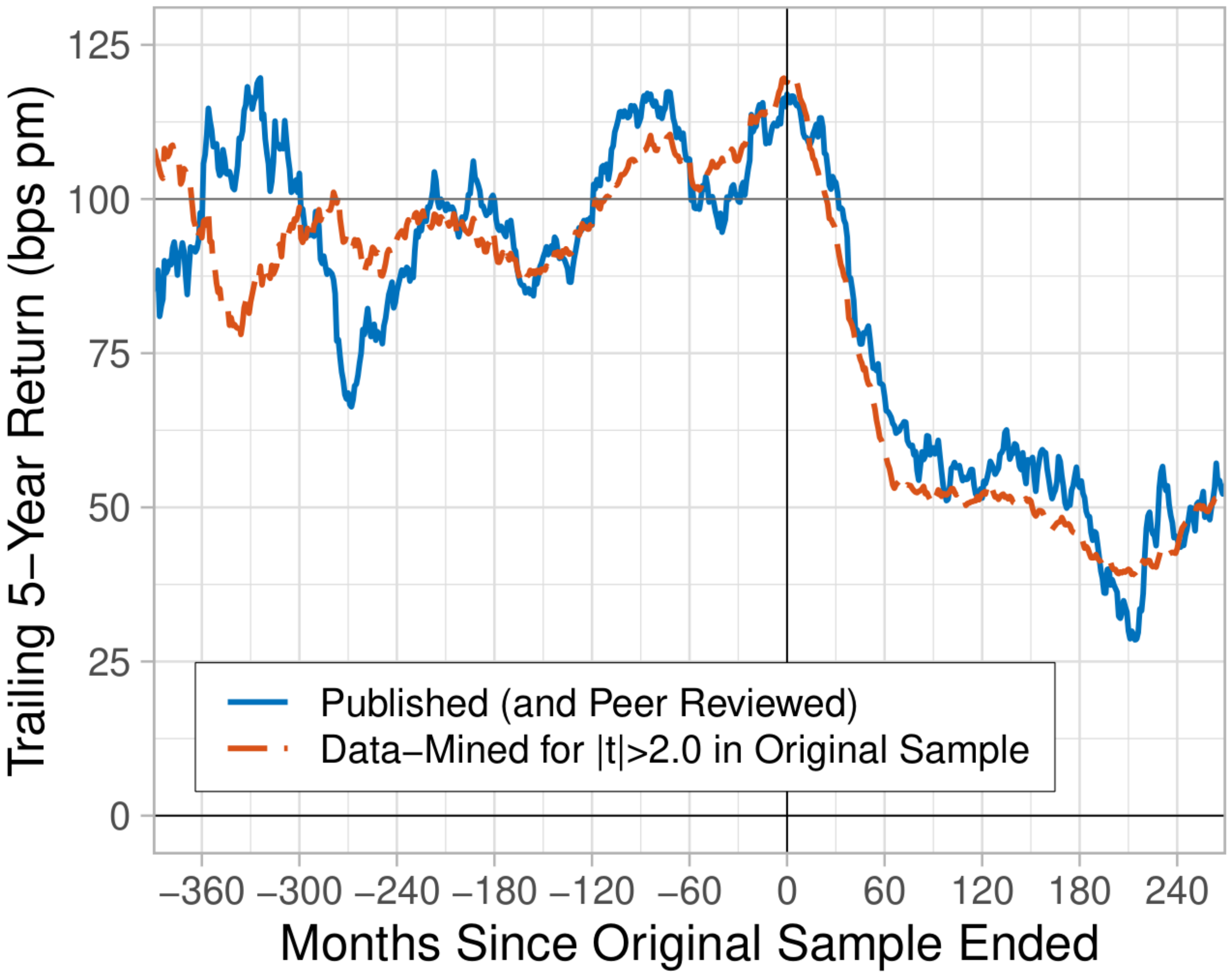
Regardless, data mining is clearly undervalued

- **Sutton's (2019) "Bitter Lesson"** from 70 years of AI research
  - Beloved, hand-crafted solutions end up "irrelevant, or worse"
  - Vast searches through huge datasets outperform
- **The real world is "tremendously, irredeemably complex"**



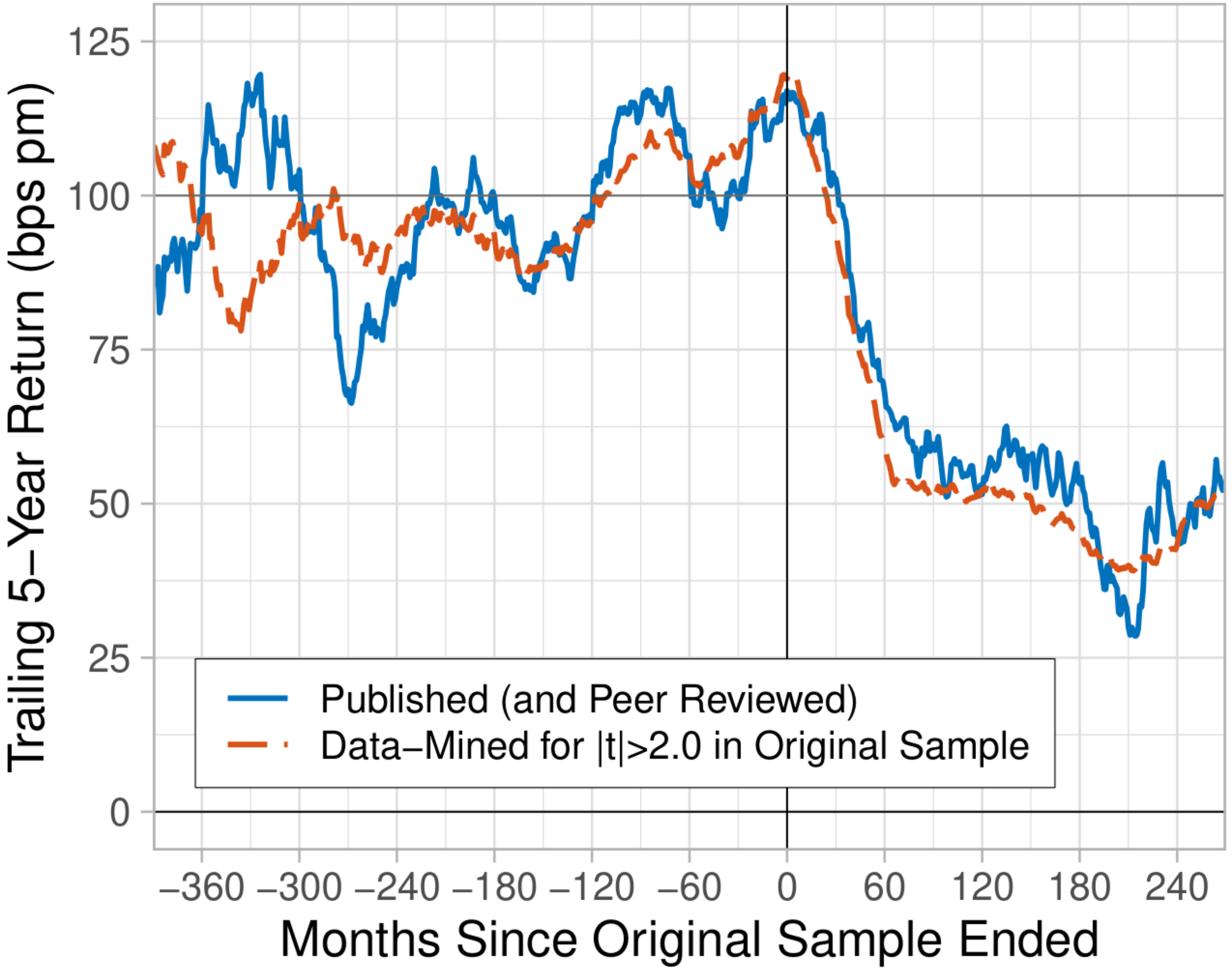
Regardless, data mining is clearly undervalued

- Economics is about beloved, hand-crafted parables



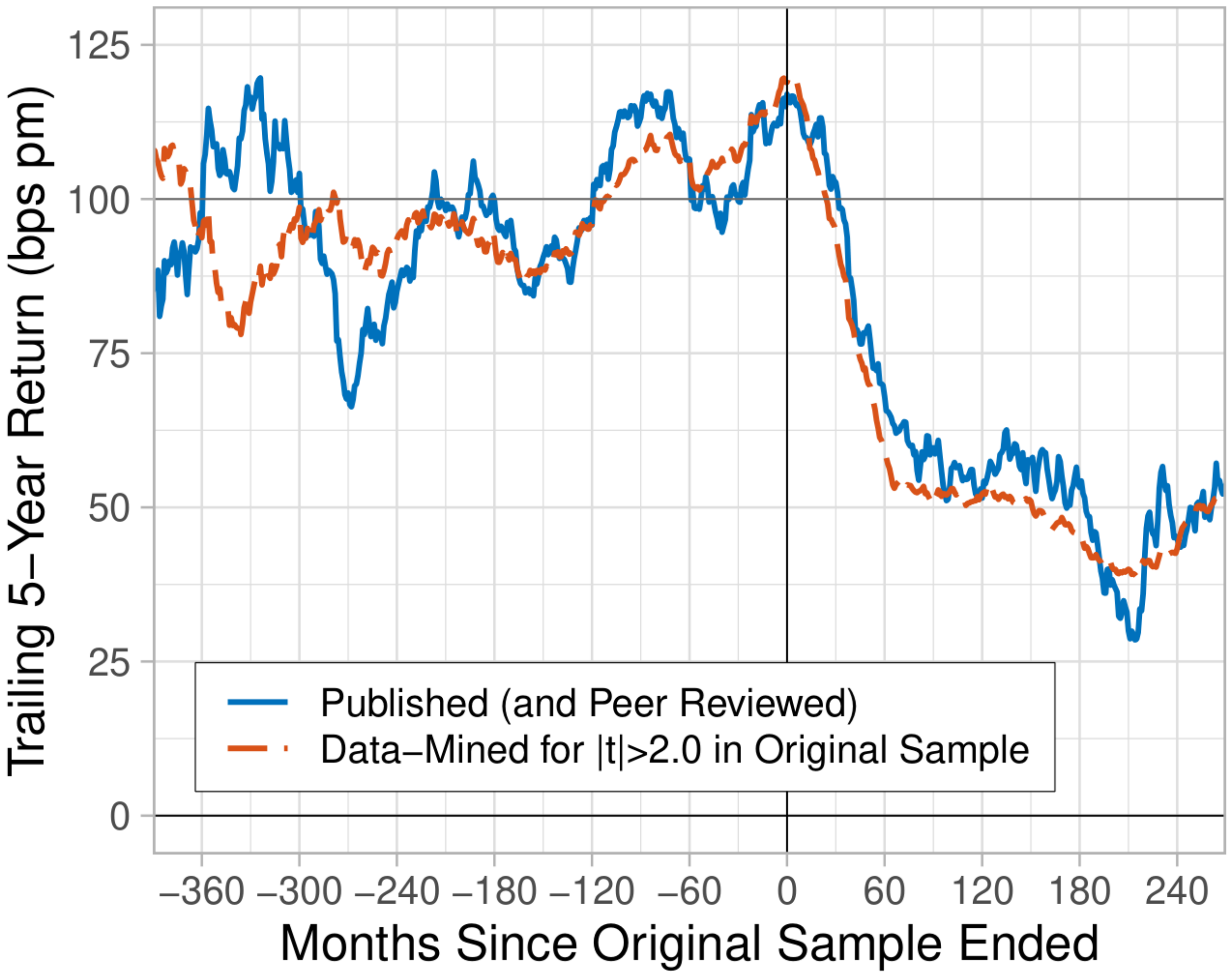
Regardless, data mining is clearly undervalued

- Economics is about beloved, hand-crafted parables
- **But perhaps if we fully explore the data...**
  - (embrace data mining)



Regardless, data mining is clearly undervalued

- Economics is about beloved, hand-crafted parables
- **But perhaps if we fully explore the data…**
  - (embrace data mining)
- **….we can produce parables that are closer to the tremendously, irredeemably complex real world**



Extra Slides

Regression of monthly returns on indicators

- Post-sample, returns decay 42% (McLean-Pontiff 2016)

	LHS: Long-Short Strategy Return (bps pm, scaled)				
RHS Variables	(1)	(2)	(3)	(4)	(5)
Intercept	100	100	100	100	102.3
	(6.4)	(6.4)	(6.4)	(6.4)	(6.8)
Post-Sample	-42.2	-25.1	-36.5	-24.4	0.7
	(8.7)	(11.7)	(10.3)	(15.3)	(14.6)
Post-Pub		-21.3		-14.9	
		(12.1)		(17.5)	
Post-Sample x Risk	-28.8	-18.8	-34.4	-19.5	-23.4
	(15.5)	(20.2)	(17.1)	(22.8)	(15.2)
Post-Pub x Risk		-14		-20.3	
		(27.2)		(30.2)	
Post-Sample x Mispricing			-8	-1	
			(7.8)	(15.5)	
Post-Pub x Mispricing				-9	
				(17.5)	
Post-2004					-59.6
					(16.7)
Null: Risk No Decay	< 0.1%	< 0.1%	< 0.1%	< 0.1%	< 0.1%

Regression of monthly returns on indicators

- Post-sample, returns decay 42% (McLean-Pontiff 2016)
- Predictors with risk explanations decay *more*

	LHS: Long-Short Strategy Return (bps pm, scaled)				
RHS Variables	(1)	(2)	(3)	(4)	(5)
Intercept	100	100	100	100	102.3
	(6.4)	(6.4)	(6.4)	(6.4)	(6.8)
Post-Sample	-42.2	-25.1	-36.5	-24.4	0.7
	(8.7)	(11.7)	(10.3)	(15.3)	(14.6)
Post-Pub		-21.3		-14.9	
		(12.1)		(17.5)	
Post-Sample x Risk	-28.8	-18.8	-34.4	-19.5	-23.4
	(15.5)	(20.2)	(17.1)	(22.8)	(15.2)
Post-Pub x Risk		-14		-20.3	
		(27.2)		(30.2)	
Post-Sample x Mispricing			-8	-1	
			(7.8)	(15.5)	
Post-Pub x Mispricing				-9	
				(17.5)	
Post-2004					-59.6
					(16.7)
Null: Risk No Decay	$<$ 0.1%	$<$ 0.1%	$<$ 0.1%	$<$ 0.1%	$<$ 0.1%

Regression of monthly returns on indicators

- Post-sample, returns decay 42% (McLean-Pontiff 2016)
- Predictors with risk explanations decay *more*
  - Even controlling for more recent publication dates

	LHS: Long-Short Strategy Return (bps pm, scaled)				
RHS Variables	(1)	(2)	(3)	(4)	(5)
Intercept	100	100	100	100	102.3
	(6.4)	(6.4)	(6.4)	(6.4)	(6.8)
Post-Sample	-42.2	-25.1	-36.5	-24.4	0.7
	(8.7)	(11.7)	(10.3)	(15.3)	(14.6)
Post-Pub		-21.3		-14.9	
		(12.1)		(17.5)	
Post-Sample x Risk	-28.8	-18.8	-34.4	-19.5	-23.4
	(15.5)	(20.2)	(17.1)	(22.8)	(15.2)
Post-Pub x Risk		-14		-20.3	
		(27.2)		(30.2)	
Post-Sample x Mispricing			-8	-1	
			(7.8)	(15.5)	
Post-Pub x Mispricing				-9	
				(17.5)	
Post-2004					-59.6
					(16.7)
Null: Risk No Decay	< 0.1%	< 0.1%	< 0.1%	< 0.1%	< 0.1%

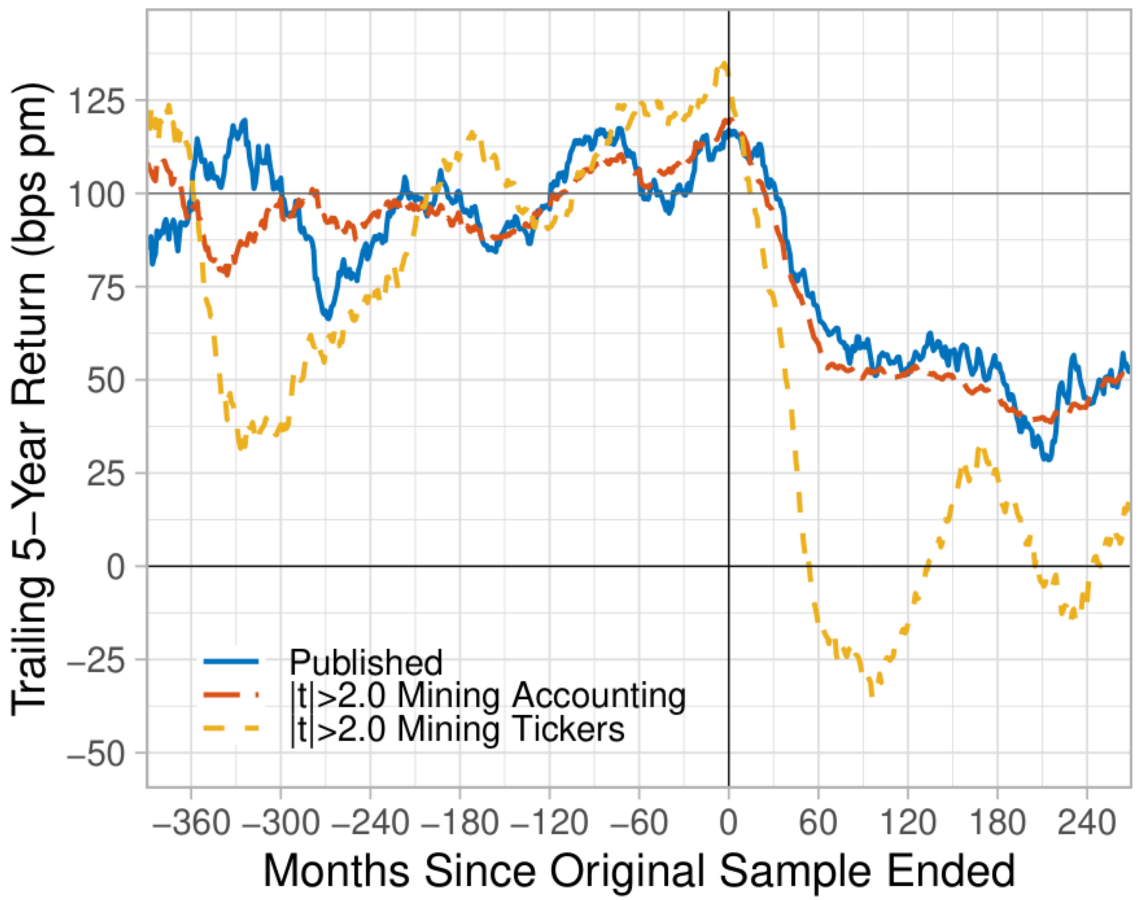
Regression of monthly returns on indicators

- Post-sample, returns decay 42% (McLean-Pontiff 2016)
- Predictors with risk explanations decay *more*
  - Even controlling for more recent publication dates
- **Does risk-based theory prevent out-of-sample decay?**
  - No, strongly reject

	LHS: Long-Short Strategy Return (bps pm, scaled)				
RHS Variables	(1)	(2)	(3)	(4)	(5)
Intercept	100	100	100	100	102.3
	(6.4)	(6.4)	(6.4)	(6.4)	(6.8)
Post-Sample	$\boxed{-42.2}$	-25.1	-36.5	-24.4	0.7
	$\boxed{(8.7)}$	(11.7)	(10.3)	(15.3)	(14.6)
Post-Pub		-21.3		-14.9	
		(12.1)		(17.5)	
Post-Sample x Risk	$\boxed{-28.8}$	-18.8	-34.4	-19.5	$\boxed{-23.4}$
	$\boxed{(15.5)}$	(20.2)	(17.1)	(22.8)	$\boxed{(15.2)}$
Post-Pub x Risk		-14		-20.3	
		(27.2)		(30.2)	
Post-Sample x Mispricing			-8	-1	
			(7.8)	(15.5)	
Post-Pub x Mispricing				-9	
				(17.5)	
Post-2004					-59.6
					(16.7)
Null: Risk No Decay	$\boxed{< 0.1\%}$	< 0.1%	< 0.1%	< 0.1%	< 0.1%

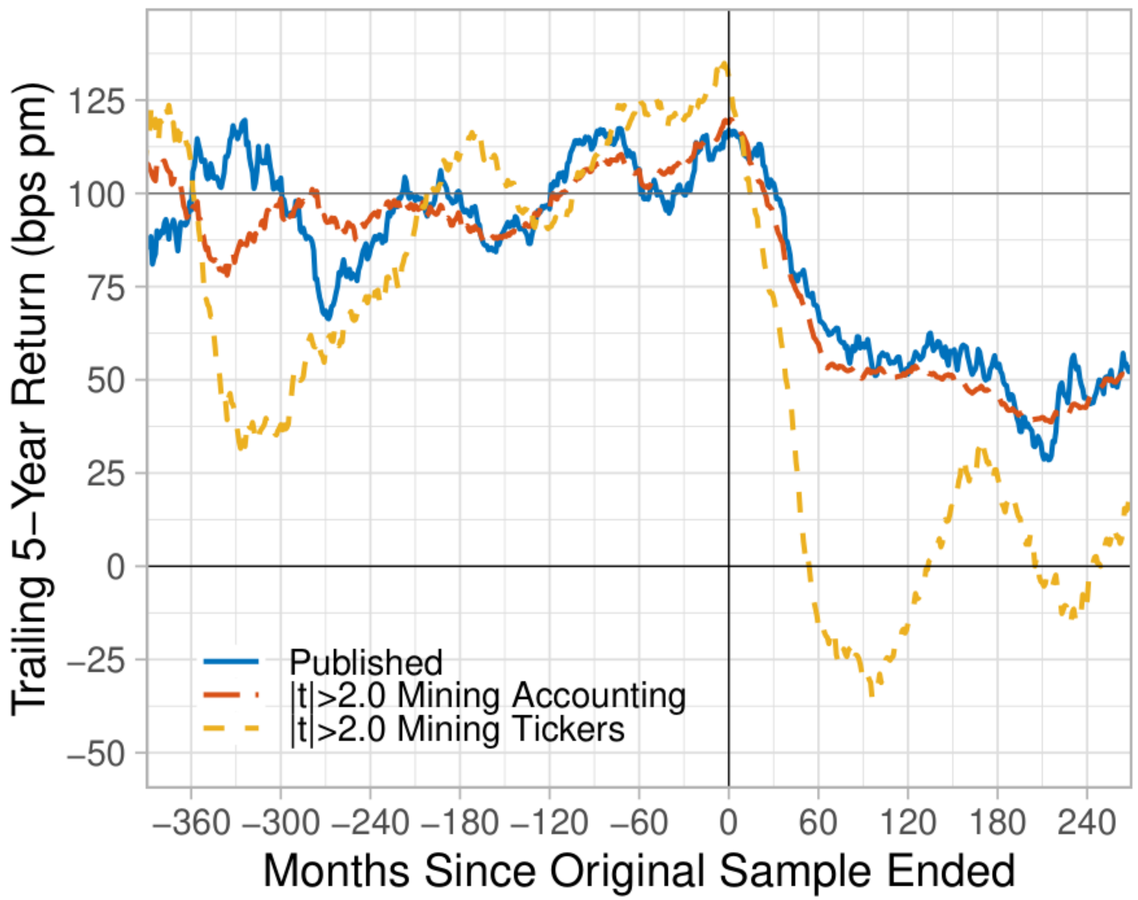
Robustness: Data mining procedure

- Construct 3,000 long-short portfolios based on letters of stock tickers
  - Suggested in Harvey (2017)
  - Far fewer than the 29,000 data-mined portfolios



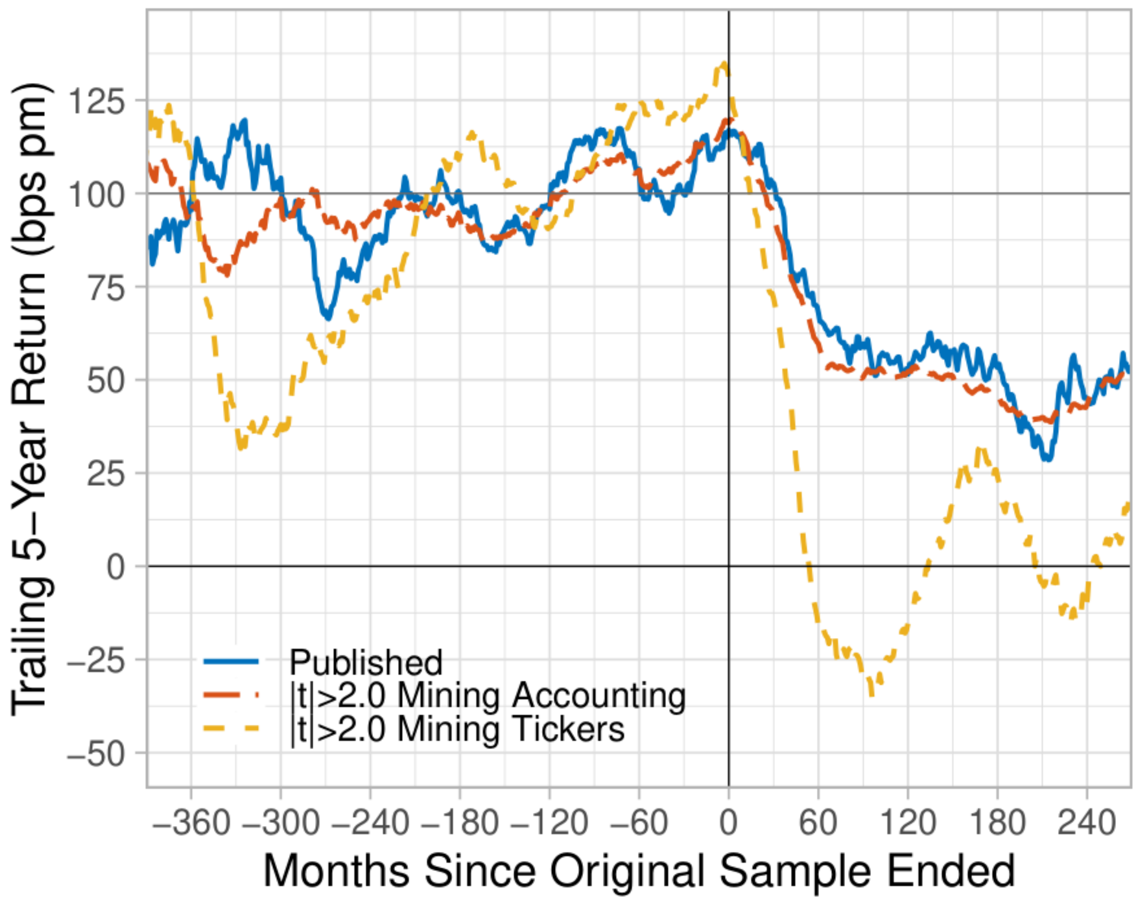
Robustness: Data mining procedure

- Construct 3,000 long-short portfolios based on letters of stock tickers
  - Suggested in Harvey (2017)
  - Far fewer than the 29,000 data-mined portfolios
- Mining tickers leads to mean zero returns post-sample (yellow)



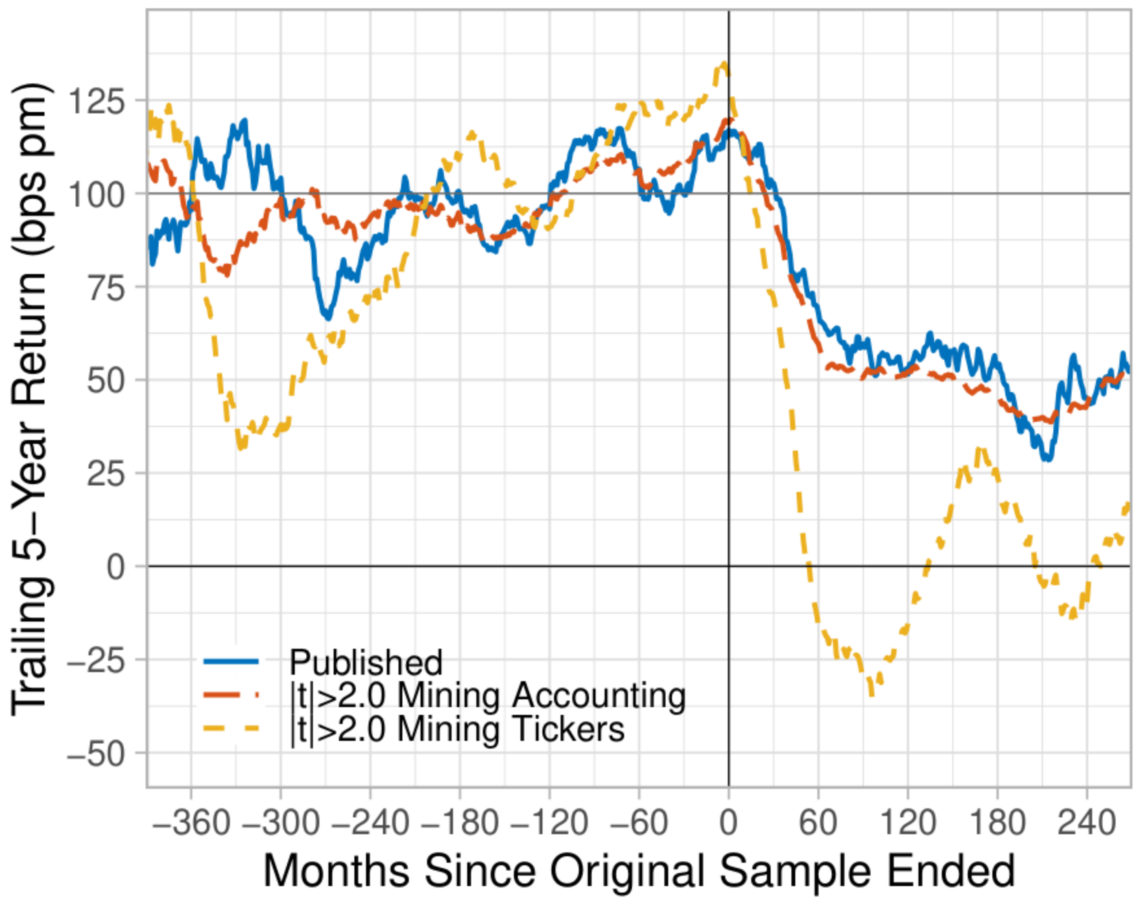
Robustness: Data mining procedure

- Construct 3,000 long-short portfolios based on letters of stock tickers
  - Suggested in Harvey (2017)
  - Far fewer than the 29,000 data-mined portfolios
- Mining tickers leads to mean zero returns post-sample (yellow)
- **2 Lessons**
  1. **The type of data being mined is important**



Robustness: Data mining procedure

- Construct 3,000 long-short portfolios based on letters of stock tickers
  - Suggested in Harvey (2017)
  - Far fewer than the 29,000 data-mined portfolios
- Mining tickers leads to mean zero returns post-sample (yellow)
- **2 Lessons**
  1. **The type of data being mined is important**
  2. **The amount of data mining is not**



Post-2004 pubs only