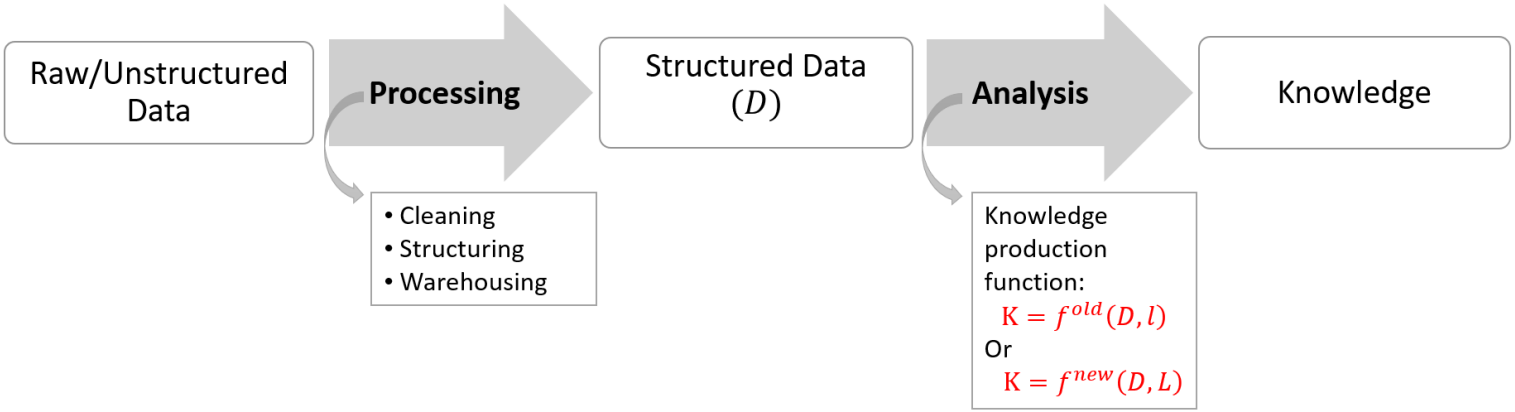
The Changing Economics of Knowledge Production

Simona Abis and Laura Veldkamp
Columbia University

Fall 2020

Motivation

- Claim: "Big data technologies are the industrialization of knowledge production."
  - Is this claim accurate? Let's measure it like industrialization and see.
- Key feature of industrialization: It changed the relative intensity of labor and capital (data).
  - Is AI doing the same?
  - How much is AI changing the labor intensity of knowledge production?
  - This matters for employment / labor income share / firm size and competition...
- Investment Management is a good lab because it's a knowledge industry.



Outline

A Model for Measurement

- Knowledge is produced using either the old technology or big data tech (AI). Same data can be used for both. Technologies have different rates of diminishing returns and use differently-skilled labor:

$$K_{it}^{AI} = A_t^{AI} {D_{it}}^{\alpha} {L_{it}}^{1-\alpha}, \tag{1}$$

$$K_{it}^{OT} = A_t^{OT} {D_{it}}^{\gamma} {I_{it}}^{1-\gamma}. \tag{2}$$

  A large $(\alpha - \gamma)$ = big revolution

- Data inputs are not raw data. They need to be structured, cleaned and machine-readable. This requires labor (λ) with diminishing marginal returns.
- New structured data is added to the existing stock of structured data. But data also depreciates at rate δ:

$$D_{i,t+1} = (1-\delta) D_{it} + A^{DM} \lambda_{it}^{1-\phi} \tag{3}$$

Maximization

- Firms maximize value function:

$$v(D_{it}) = \max_{\lambda_{it}, L_{it}, I_{it}} A_t^{AI} D_{it}^{\alpha} L_{it}^{1-\alpha} + A_t^{OT} D_{it}^{\gamma} I_{it}^{1-\gamma} - w_{L,t} L_{it} - w_{I,t} I_{it} - w_{\lambda,t} \lambda_{it} + \\ + \frac{1}{r} v(D_{i(t+1)}) \quad (4)$$

  where (3) holds.

- First Order Conditions:
  - L_{it}: $(1-\boldsymbol{\alpha}) K_{it}^{AI} - w_{L,t} L_{it} = 0$.
  - I_{it}: $(1-\boldsymbol{\gamma}) K_{it}^{OT} - w_{I,t} I_{it} = 0$.
  - λ_{it}: $\frac{(\alpha K_{it}^{AI} + \gamma K_{it}^{OT})(1-\boldsymbol{\phi})}{r-(1-\delta)} \frac{D_{i(t+1)} - (1-\delta) D_{it}}{D_{it}} - w_{\lambda,t} \lambda_{it} = 0$.

- These first order conditions allow us to identify $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$.

State Variable Evolution

- We have two challenges:

1. We don't observe firms' data stock (D_{it}) but we can infer it:

$$D_{it} = \frac{\left(\alpha \frac{w_{L,t} L_{i,t}}{(1-\alpha)} + \gamma \frac{w_{I,t} I_{i,t}}{(1-\gamma)}\right)(1-\phi)}{r-(1-\delta)} \frac{A^{DM} \lambda_{it}^{-\phi}}{w_{\lambda,t}}$$

$$= (1-\delta)^t D_{i0} + \sum_{s=1}^{t} (1-\delta)^{t-s} A^{DM} \lambda_{s-1}^{1-\phi}$$

   - lhs: Data must be optimal given wages paid to workers.
   - rhs: Data accumulates in proportion to data management hires.
   - We express the D_{i0} of each firm as a function of an average $\bar{D}_0$, proportional to each firm's cumulated hiring between 2000-2014 (burn-in period).

2. We need to compute the three productivity parameters:
   - A^{DM} is computed from the above equation, using time-series and cross-sectional averages
   - A_t^{OT} and A_t^{AI} are computed from the OT and AI FOCs respectively, using cross-sectional averages

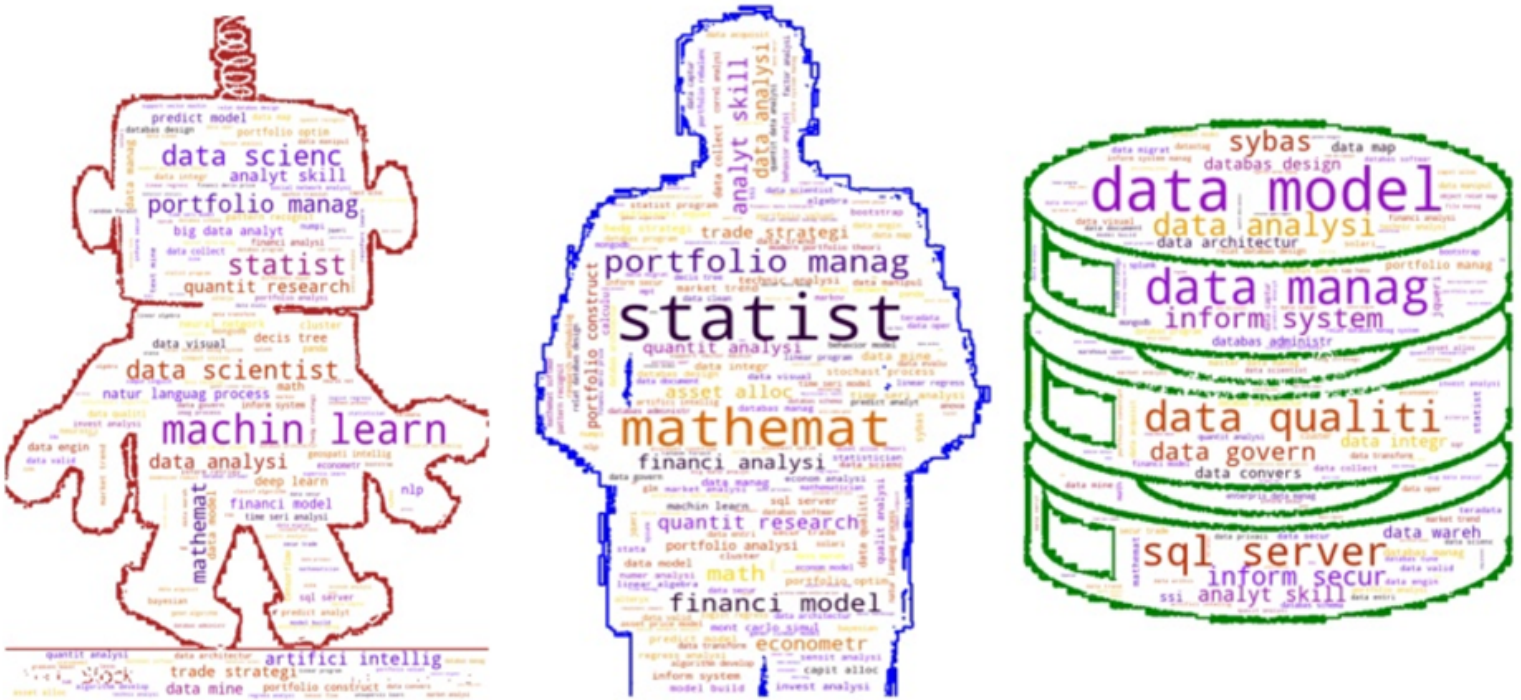
Outline

Labor Demand: Job Postings Sample

- Job postings & salaries: Burning Glass Technologies (BGT), 2010 – 2018.
  - More than 40,000 sources (e.g. job boards, employer sites, non-digital).
  - 60 – 80% of U.S. job vacancies, with the finance and technology industries having especially good coverage, Acemoglu et al. (2019).
- Filtering – we use job postings that are:
  - In the finance industry (based on NAICS, O*NET and BGT codes).
  - For whom $\geq$ 25% of analysis jobs require investment-management skills (based on BGT skills and skill clusters).
  - Hire $\geq$ 5 Old Tech or AI worker in 2010 – 2018 (based on full job text).
  - Our final sample comprises:
    - 507,971 job postings, 143,809 of which are AI, OldTech or DataMgmt.
    - 33,610 non-zero employer-month observations for 928 unique companies.
- Key innovations:
  - Focus on a specific industry to provide a more fine-grained division of labor.
  - Base categorization on full job text rather than BGT structured skills.

Labor Demand: Job Postings Categorization

1. Develop dictionaries of words and short phrases that indicate *data management* or *data analysis*.
2. Compute their relative frequency in each pre-processed job text to assign jobs to *data management* or *data analysis*.
3. Among *data analysis* keywords, identify those indicative of the *old* and *new* technologies.
4. Assign jobs to *OldTech* or *AI* depending on the relative frequency of words of the two types present in the posting.



An Example: Two Sigma - OldTech

Two Sigma – Aug 2010 – Quantitative Analyst:

We are looking for world-class quantitative modelers to join our highly motivated team. Quant candidates will have exceptional quantitative skills as well as programming skills, and will write production quality, high reliability, highly-tuned numerical code. Candidates should have: a bachelor's degree in mathematics and/or computer science from a top university; an advanced degree in hard science, computer science, or the equivalent (a field where strong math and statistics skills are necessary); 2 or more years of professional programming experience in Java and C, preferably in the financial sector; strong numerical programming skills; strong knowledge of computational numerical algorithms, linear algebra and statistical methods; and experience working with large data sets. (...)

- Keywords:
  - AI: None
  - OldTech: mathemat (x1), math (x1), statist (x2), algebra (x1)
  - DataMgmt: None

An Example: Two Sigma - AI

Two Sigma – Aug 2018 – Lead Data Scientist:

As machine learning and data-driven business intelligence have permeated industries, an abundance of new datasets and techniques have created opportunities for granular measurement of increasingly varied aspects of our economy. Two Sigma is looking to hire a highly creative & motivated Lead Data Scientist to further scale our long-standing efforts to leverage these advancements to measure and predict the world's financial outcomes.

Two Sigma's data engineering platform enables us to harness some of the world's most complex & challenging content, as we structure and integrate new datasets into a diverse ecosystem of syndicated financial and industry-specific data products. Two Sigma's data scientists are focused on joining, enriching, and transforming datasets into novel creative measures of economic activity. (...)

- Keywords:
  - AI: data scienc (x4), data scientist (x5), machin learn (x1)
  - OldTech: statist (x2)
  - DataMgmt: data engin (x1), data sourc (x1), support data (x1)

An Example: TwoSigma - DataMgmt

Two Sigma – Dec 2013 – SQL Data Analyst:

*(...) Technology drives our business it's our main competitive advantage and as a result, software engineers play a pivotal role. They tackle the hardest problems through analysis, experimentation, design, and elegant implementation. Software engineers at Two Sigma build what the organization needs to explore data's possibilities and act on our findings to mine the past and attempt to predict the future. We create the tools at scale to enable vast data analysis; the technology we build enables us to engage in conversation with the data, and search for knowledge and insight. (...) You will be responsible for the following: * Capturing and processing massive amounts of data for thousands of different tradable instruments, including stocks, bonds, futures, contracts, commodities, and more; (...)*

- Keywords:
  - AI: None
  - OldTech: None
  - DataMgmt: explor data possibl, enabl vast data analysi, data specialist, data team

Cumulating Postings to Labor Stocks

s_t^{type}: separation rates by type-month (from BLS, match NAICS codes by type)
h_t^{type}: fraction of posted vacancies filled by type-month (BLS, same)
j_t^{type}: Burning Glass job postings rates by type-month

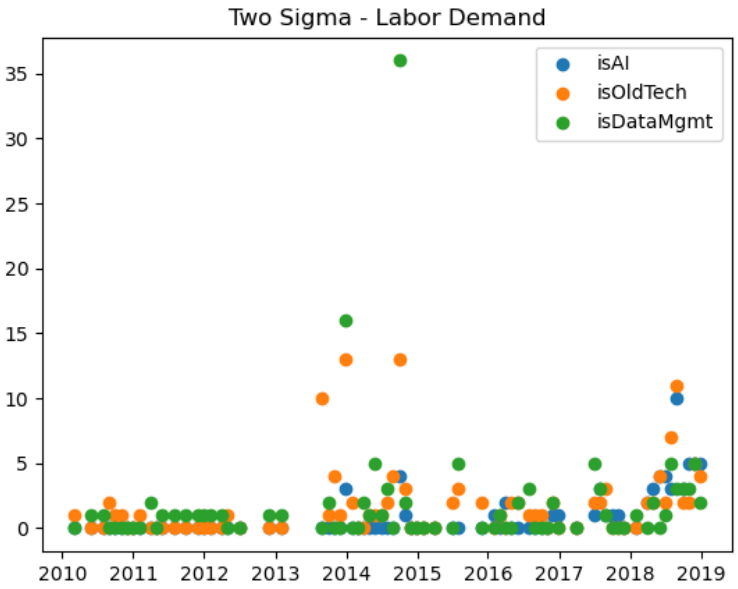
$$L_{it} = (1 - s_t^{AI}) L_{i(t-1)} + j_{it}^{AI} h_t^{AI}, \tag{5}$$

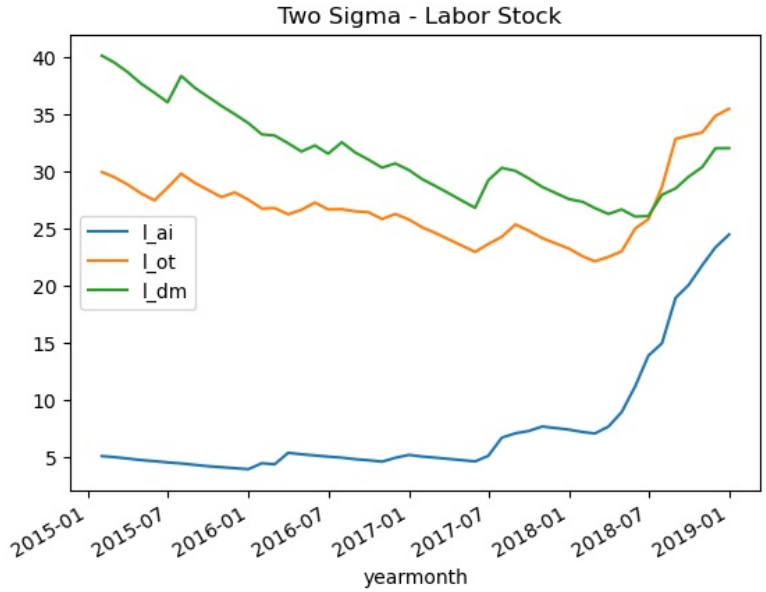
$$l_{it} = (1 - s_t^{OT}) l_{i(t-1)} + j_{it}^{OT} h_t^{OT}, \tag{6}$$

$$\lambda_{it} = (1 - s_t^{DM}) \lambda_{i(t-1)} + j_{it}^{DM} h_t^{DM}. \tag{7}$$

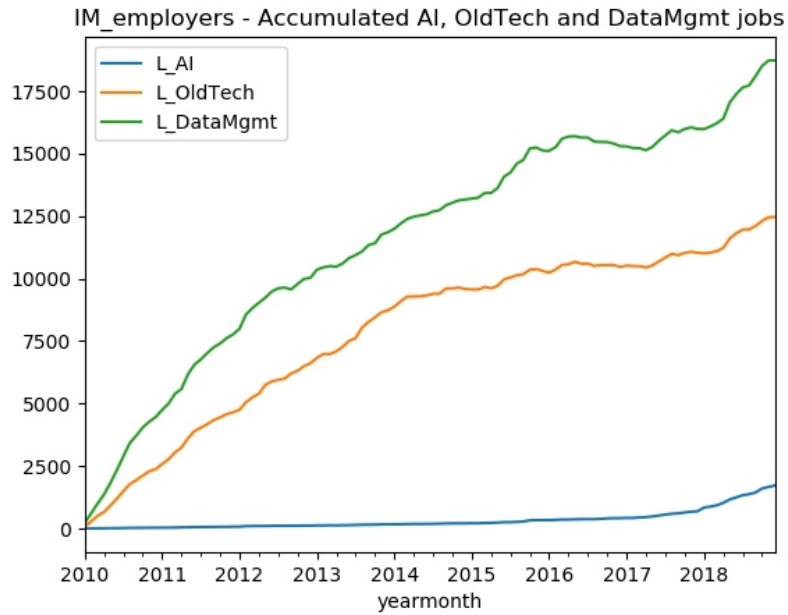
- Remaining question: What is the initial stock of labor? 2 possibilities:
  - Baseline: Start all initial employment at zero
  - Robustness: Assume that the sector in 2007 is in steady state. Then hiring is equal to the expected number of separations: $h_{i0} = s_t L_{i0}$ and $H_{i0} = S_t \lambda_{i0}$. Use initial hiring to impute initial stocks.
- In both cases, we use 2010-14 as burn-in and start estimation in 2015.

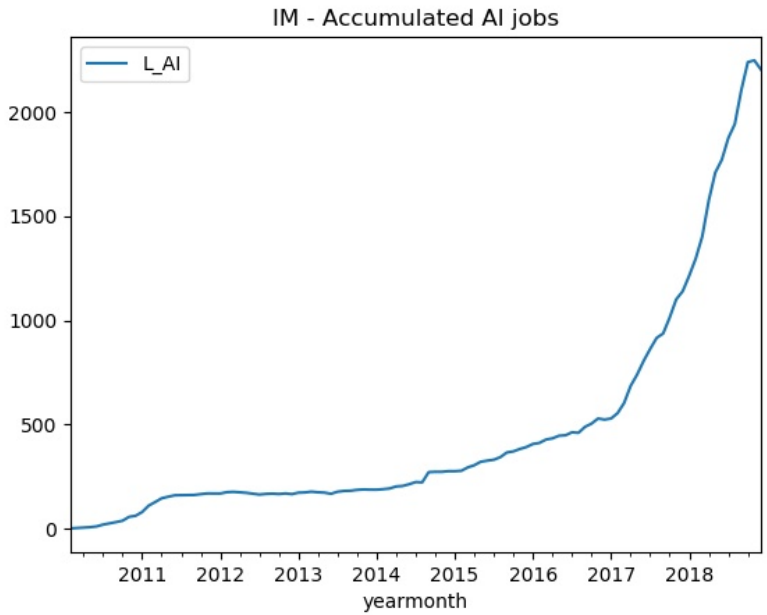
An Example: Two Sigma





Labor Stocks: Growth in AI employment





- In 2010 5% of employers in our sample hired AI workers, in 2018 30% did.
- AI employment has been growing at a faster rate since 2015.
  We use 2015-2018 as our estimation period.

Wages

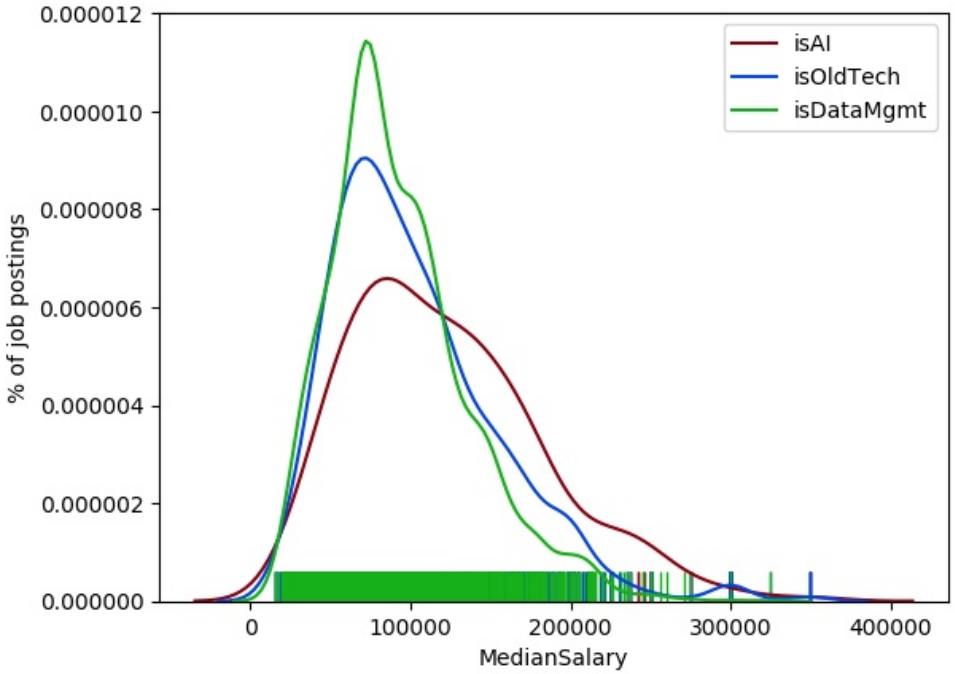


Figure: Distribution of wages for data managers, old technology analysts and machine learning analysts. Burning glass job postings, 2010-2018.

- AI analysts are paid $\$34,436$ more than OT analysts on average.
- Salaries vary by job-type and month.
- We are in the process of acquiring salaries with greater cross-firm variation.

Structural Estimation

- From the model's solution we have $4 \times 33,610$ equations to be set to zero.

- Procedure:
  - Non-linear least squares to iterate over different combinations of the diminishing returns parameters (α, γ and ϕ) and the average initial data stock ($\bar{D}_{i0}$).
  - In each iteration, we back out the production parameters from:
    - Cross-section and time series averages of the data process condition (A^{DM}).
    - Cross-sectional averages of the AI and OT FOCs (A_t^{AI} and A_t^{OT}).
  - We also use a grid search to check global convergence.

- Identification:
  - α, γ and ϕ:
    - Time-series variation in the growth rates of OldTech and AI employment.
    - Cross-firm covariation in Analysis and DataManagement employment.
  - $\bar{D}_{i0}$: Reconciles wages with the accumulation process of data.
  - Key: Data and productivity parameters are jointly estimated.

Outline

Main Results: Greater Productivity of Data

monthly data depreciation		$\delta = 1\%$	$\delta = 2.5\%$	$\delta = 10\%$
Data Management	ϕ	0.172	0.190	0.144
		(0.0025)	(0.0019)	(0.0022)
AI Analysis	α	0.806	0.734	0.613
		(0.0013)	(0.0026)	(0.0038)
Old Technology Analysis	γ	0.458	0.560	0.567
		(0.0024)	(0.0017)	(0.0006)

Table: The exponents α and γ represent the diminishing returns to data in the new and old technologies.

- $\alpha > \gamma$
  - AI has significantly raised the productivity of analyzing larger data sets.
  - Labor share fell by 17% (for $\delta = 2.5\%$).
- Technological change is substantial.
  - Industrial revolution: capital exponent estimated to have risen of 0.122. We estimate an increase of 0.174 in the data exponent (for $\delta = 2.5\%$).
  - A fall in the labor share could mean fewer workers, or could mean more data. Which was it?

Results: Not a Labor Replacing Technology

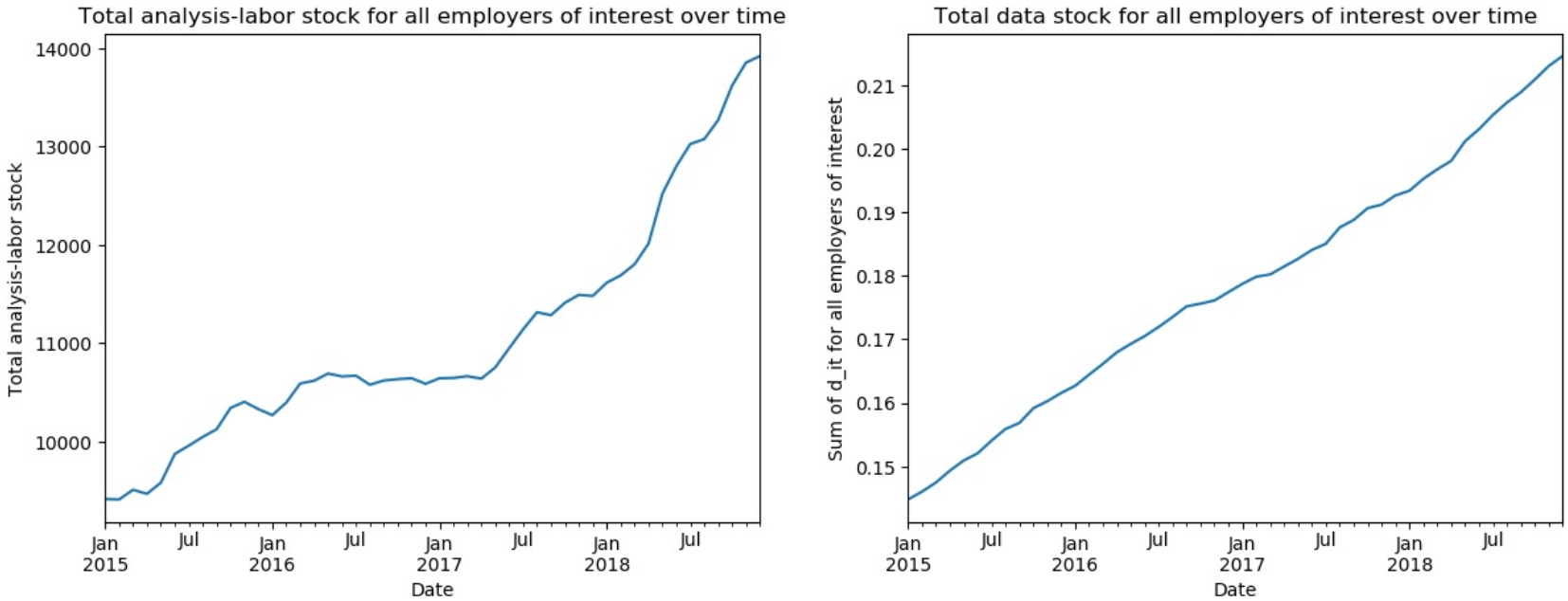


Figure: Data Stocks and Labor Stocks. Panel 1 displays the aggregate stock of analysis labor (AI and OldTech). Panel 2 is the sum of all data stocks, estimated for each firm in our sample ($\delta = 0.025$ per month).

- Both data and labor stock grow rapidly.
- Increase in labor split about evenly between AI and OldTech analysts

Results: Our Methodology Can Value Data

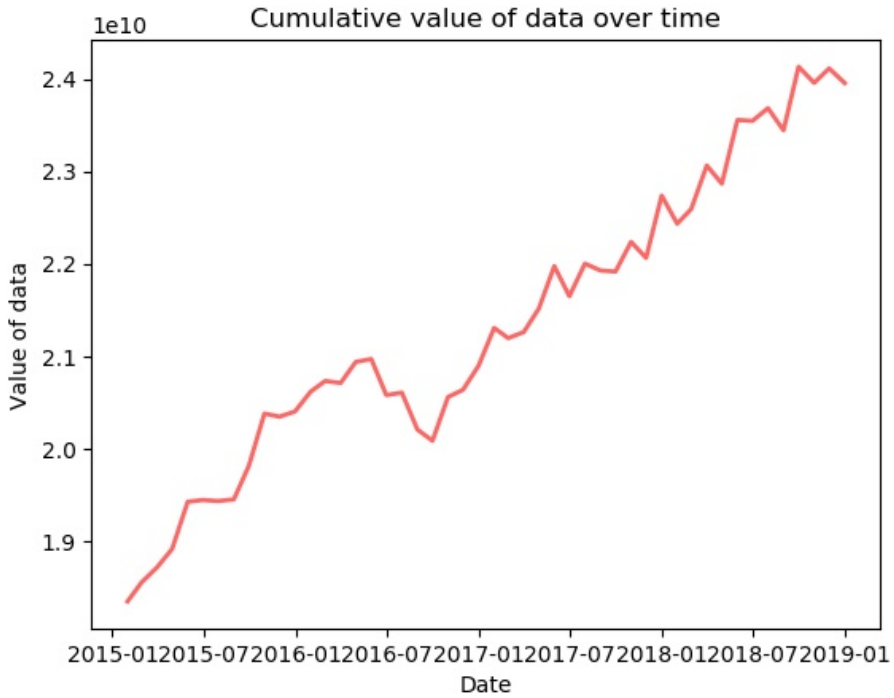


Figure: Estimated Value of the Aggregate Stock of Data, in billions of current U.S. dollars, 2015-2018.

- Once we have estimated production parameters and data stocks, we can put them back into our value function, and approximate the value of each firm's stock of data in each month.
- Data value for firms in our sample rose by 26% in 4 years.

AI IS RAISING THE VALUE OF DATA

1. A larger data stock determines a higher cumulative value of data
2. More analysis workers make each data point more valuable
3. Firms are becoming more productive at using data:

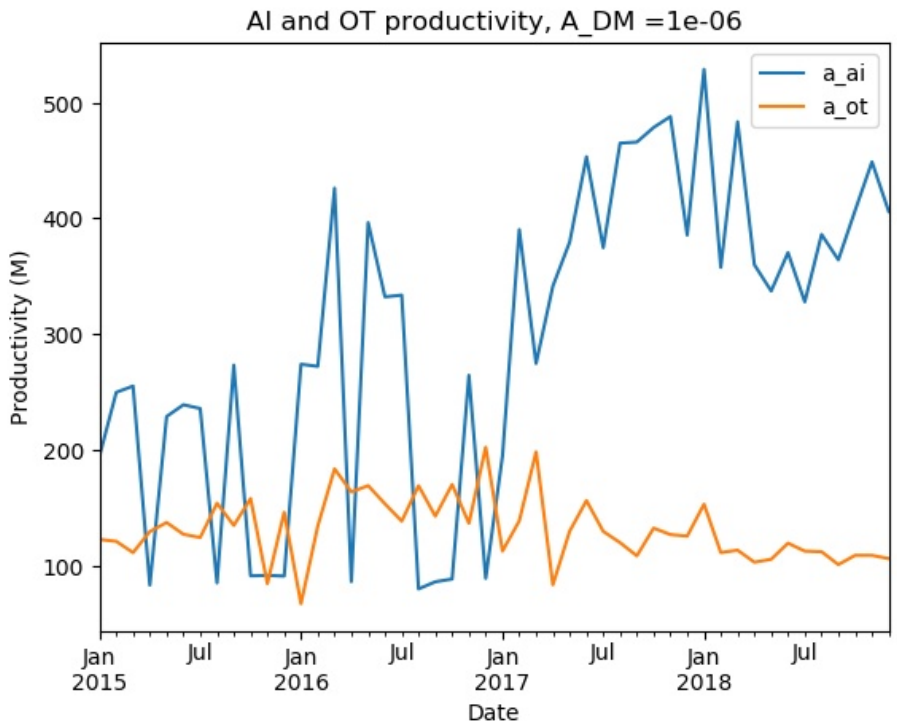


FIGURE: Productivity of Financial Data Analysis, reported for old tech and AI technologies, 2015-2018.

Outline

Key Take-Aways

- We infer how much data each firm has from data management hiring.
- We infer how much diminishing returns there is by asking what exponent would make their hiring patterns closest to optimal.
- The change in diminishing returns is large, with big implications for competition and firm size.

Backup Slides

Feasible Parameters: Positive Data

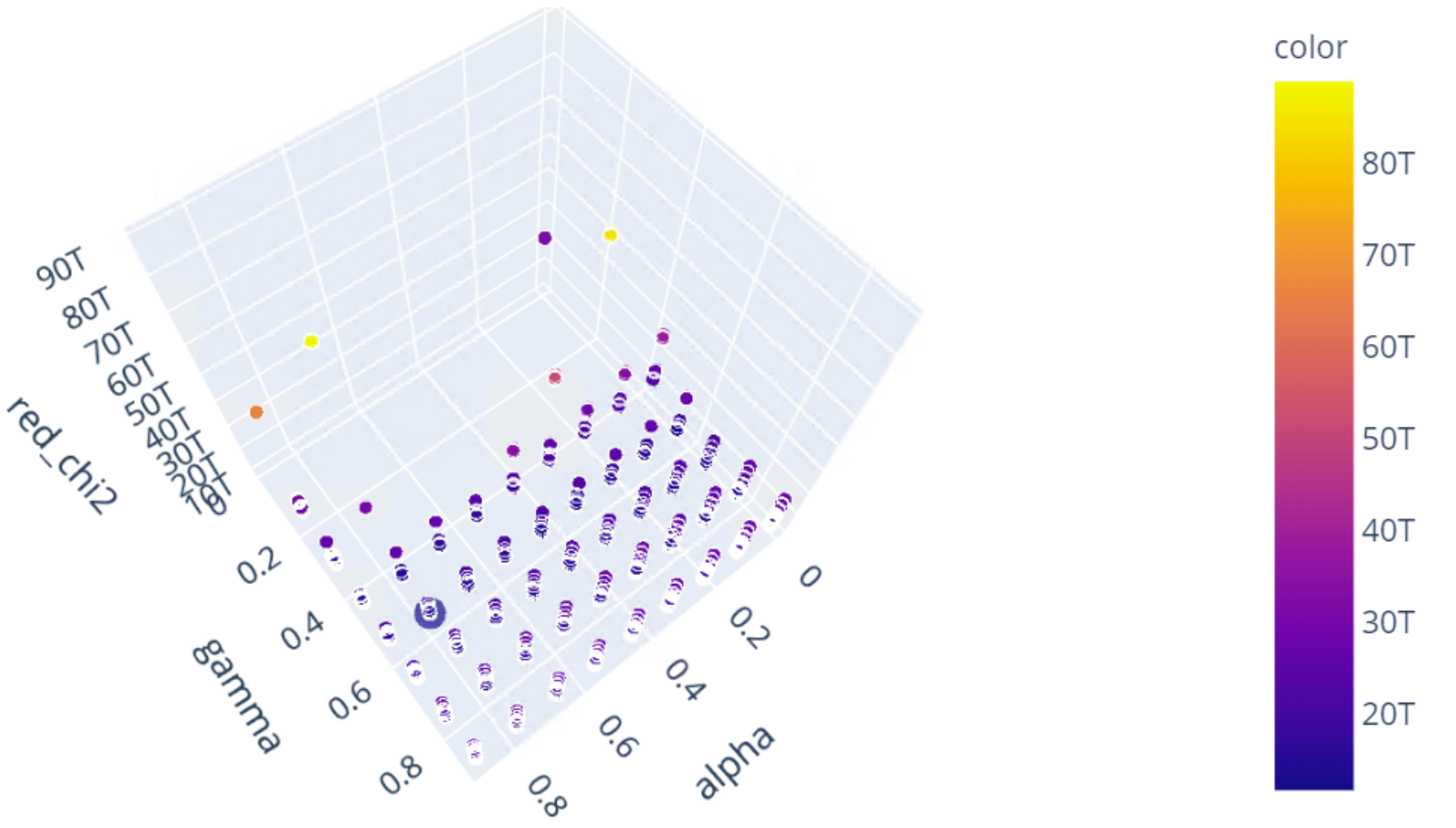


Figure: Grid Search 10x10x10x40: All feasible combinations of α and γ

Back

GRID SEARCH: α AND γ

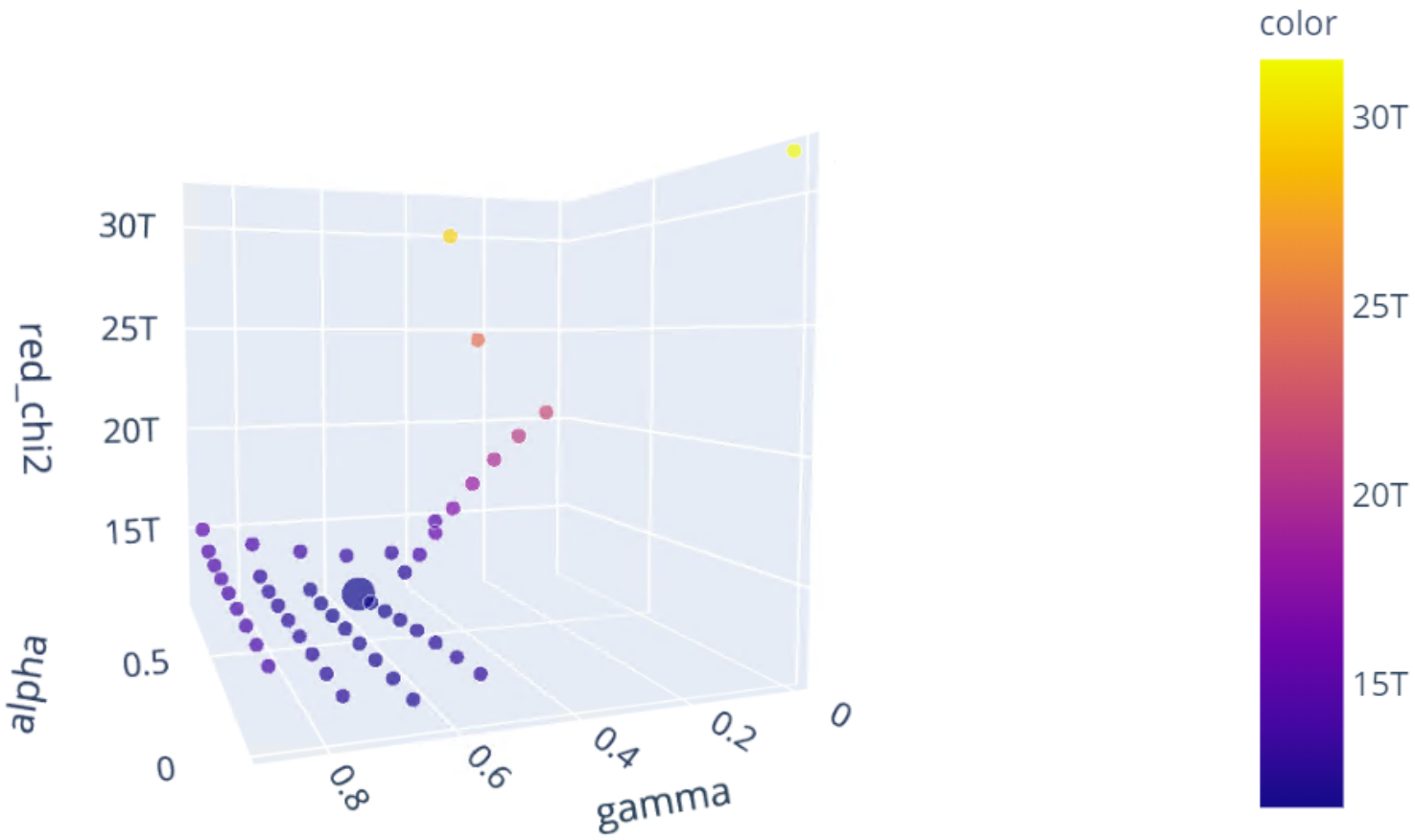


FIGURE: Grid Search 10x10x10x40: all feasible combinations of α and γ with lowest residual sum of squares

Back

GRID SEARCH: α AND ϕ

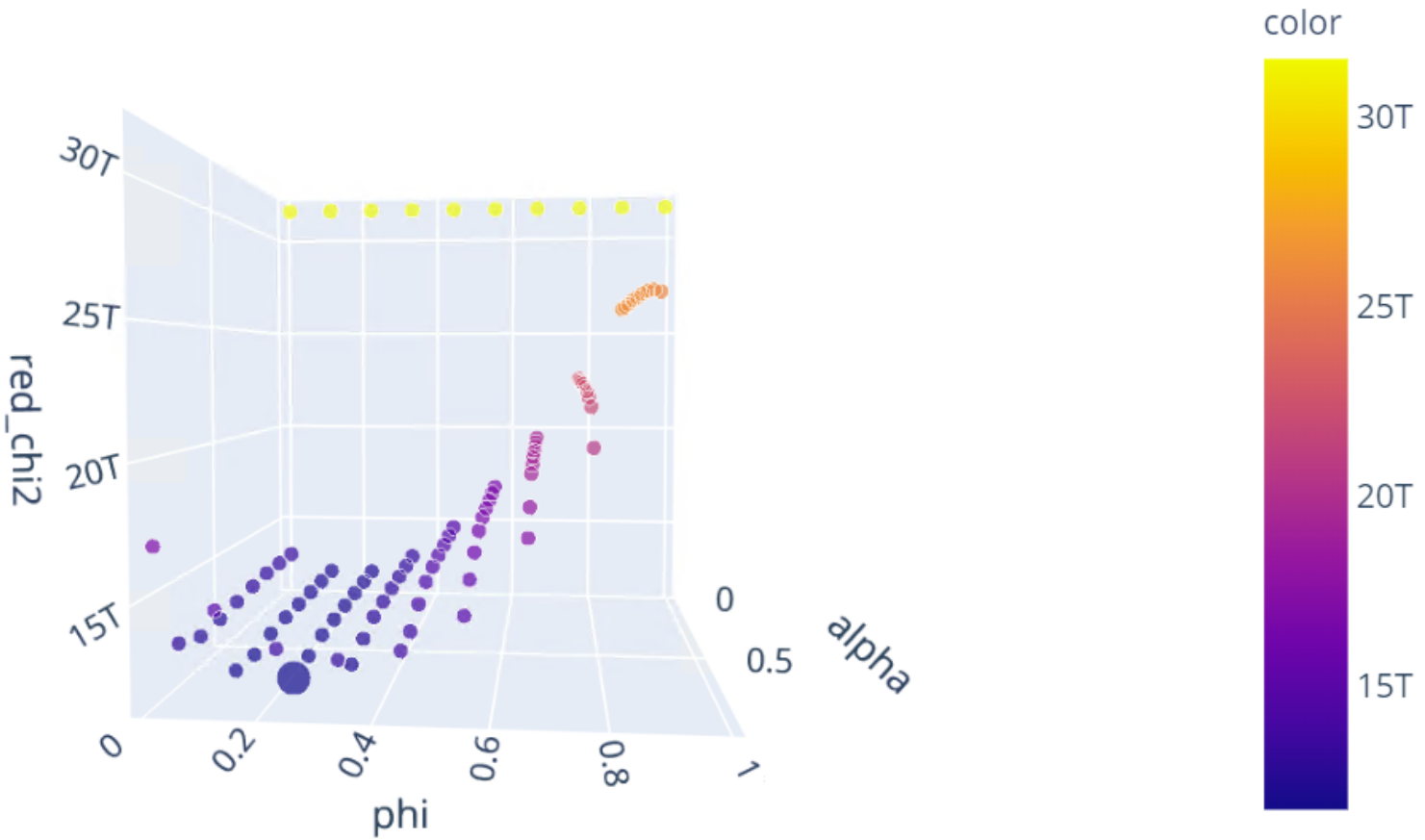


FIGURE: Grid Search 10x10x10x40: all feasible combinations of α and ϕ with lowest residual sum of squares

Back

GRID SEARCH: α AND $\bar{D}_0$

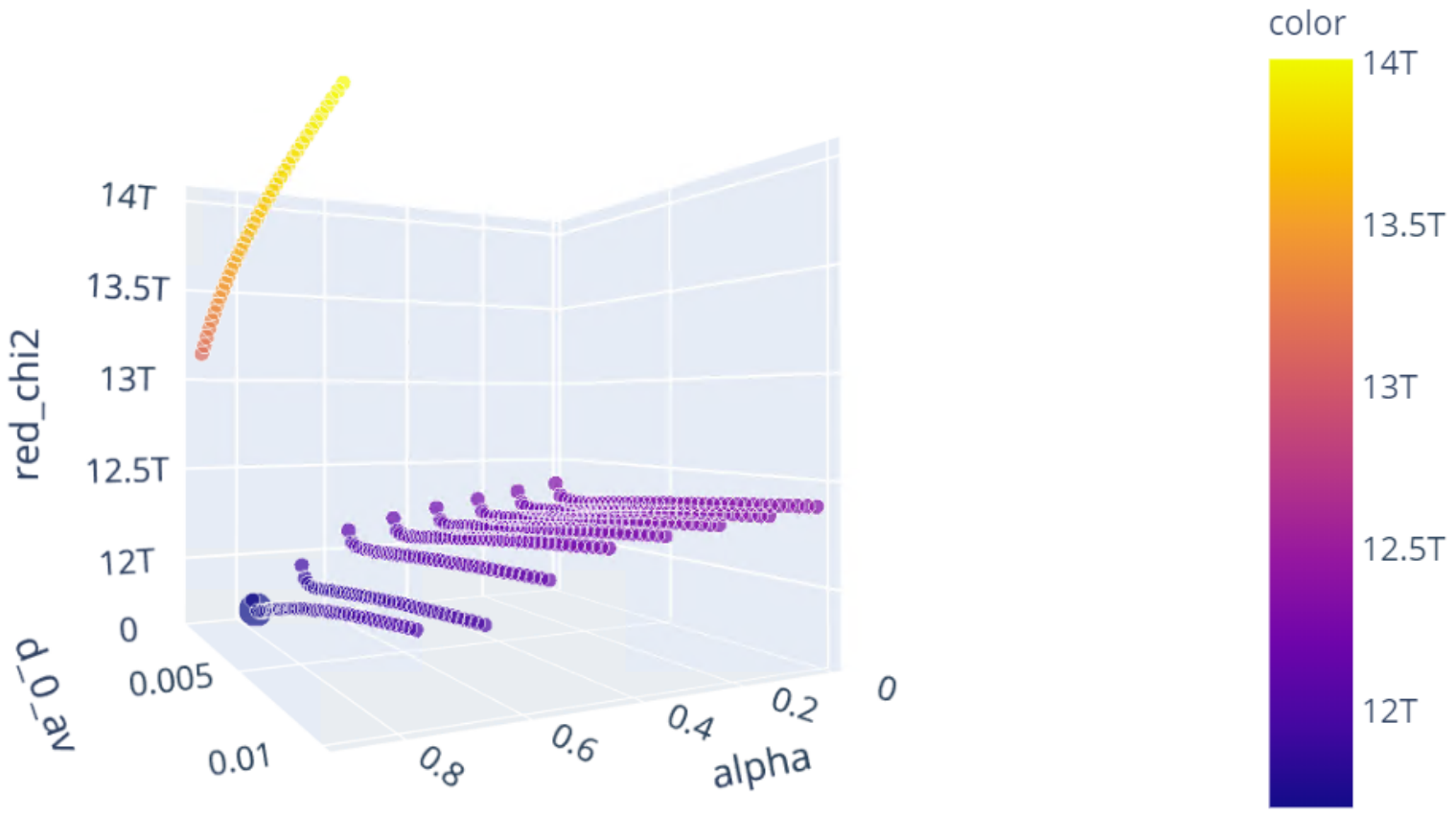


FIGURE: Grid Search 10x10x10x40: all feasible combinations of α and $\bar{D}_0$ with lowest residual sum of squares

Back